Title

Identifying gene-gene interactions that are highly associated with four quantitative lipid traits across multiple cohorts

Rishika De[1], Shefali S. Verma[2], Emily Holzinger[2], Molly Hall[2], Amber Burt[3], David S. Carrell[4], David R. Crosslin[5], Gail P. Jarvik[3,5], Helena Kuivaniemi[6], Iftikhar J. Kullo[7], Leslie A. Lange[8], Matthew B. Lanktree[9], Eric B. Larson[4], Kari E. North[10], Alex P. Reiner[11], Vinicius Tragante[12,13], Gerard Tromp[6], James G. Wilson[14], Folkert W. Asselbergs[12,15,16], Fotios Drenos[17,18], Jason H. Moore[19], Marylyn D. Ritchie[2*], Brendan Keating[20,21]*, Diane Gilbert-Diamond[22,23]*

[1]Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, NH, USA
[2]The Center for Systems Genomics, The Pennsylvania State University, University Park, PA, USA
[3]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA, USA
[4]Group Health Research Institute, Seattle, WA, USA
[5]Department of Genome Sciences, University of Washington, Seattle, WA, USA
[6]Division of Molecular Biology and Human Genetics, Department of Biomedical Sciences, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, 7505, South Africa
[7]Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN, USA
[8]Department of Genetics, University of North Carolina School of Medicine at Chapel Hill, Chapel Hill, NC, USA
[9]Departments of Medicine and Biochemistry, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario, Canada
[10]Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[11]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[12]Department of Cardiology, Division Heart and Lungs, University Medical Center Utrecht, Utrecht, The Netherlands
[13]Department of Medical Genetics, Biomedical Genetics, University Medical Center, Utrecht, The Netherlands
[14]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA
[15]Institute of Cardiovascular Science, University College London, 222 Euston Road, London, NW1 2DA, UK
[16]Durrer Center for Cardiogenetic Research, ICIN-Netherlands Heart Institute, Utrecht, The Netherlands
[17]MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Bristol, UK

[18]Centre for Cardiovascular Genetics, Institute of Cardiovascular Science, University College London, London, UK
[19]Institute for Biomedical Informatics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[20]University Medical Center Utrecht, Utrecht, The Netherlands
[21]The Children's Hospital of Philadelphia, Philadelphia, PA, USA
[22]Institute for Quantitative Biomedical Sciences at Dartmouth, Hanover, NH, USA
[23]Department of Epidemiology, Geisel School of Medicine at Dartmouth, Hanover, NH, USA

*Corresponding authors
mdr23@psu.edu, Ph: 570 214 7579 (MDR)
bkeating@mail.med.upenn.edu, Ph: 267 760 4507 (BK)
Diane.Gilbert-Diamond@dartmouth.edu, Ph: 603 653 3362 (DGD)

Abstract

Genetic loci explain only 25-30% of the heritability observed in plasma lipid traits. *Epistasis*, or gene-gene interactions may contribute to a portion of this *missing heritability*. Using genetic data from five NHLBI cohorts of 24,837 individuals, we combined the use of the quantitative multifactor dimensionality reduction (QMDR) algorithm with two SNP filtering methods to exhaustively search for SNP-SNP interactions that are associated with HDL cholesterol (HDL-C), LDL cholesterol (LDL-C), total cholesterol (TC) and triglycerides (TG). SNPs were filtered either on the strength of their independent effects (main effect filter) or the prior knowledge supporting a given interaction (Biofilter). After the main effect filter, QMDR identified 20 SNP-SNP models associated with HDL-C, 6 associated with LDL-C, 3 associated with TC, and 10 associated with TG (permutation P-value < 0.05). With the use of Biofilter, we identified 2 SNP-SNP models associated with HDL-C, 3 associated with LDL-C, 1 associated with TC and 8 associated with TG (permutation P-value < 0.05). In an independent dataset of 7,502 individuals from the eMERGE network, we replicated 14 of the interactions identified after main effect filtering: 11 for HDL-C, 1 for LDL-C and 2 for TG. We also replicated 23 of the interactions found to be associated with TG after applying Biofilter. Prior knowledge supports the possible role of these interactions in the genetic etiology of lipid traits. This study also presents a computationally efficient pipeline for analyzing data from large genotyping arrays and detecting SNP-SNP interactions that are not primarily driven by strong main effects.

Keywords

Epistasis, gene-gene interactions, lipid disorders, cholesterol, MDR

Introduction

Plasma lipid and lipoprotein levels are a major risk factor for cardiovascular disease (CVD), the leading cause of death in the world (Arsenault et al. 2011; World Health Organization 2014). In 2012, approximately one-third of all global deaths were caused by CVD (Deaton et al. 2011; World Health Organization 2014). Moreover, CVD no longer remains a disease associated with industrialized nations. With increasing urbanization around the world, 80% of global CVD-related deaths occur in low- and middle-income countries and the World Health Organization estimates global CVD-related deaths to reach 22.2 million by 2030 (World Health Organization 2014).

Although lipid levels can be influenced by factors such as age, sex, body mass index (BMI), environmental factors and lifestyle choices including diet, they can be influenced by genetic factors as well (Heller et al. 1993). Lipid traits such as high density lipoprotein cholesterol (HDL-C), low density lipoprotein cholesterol (LDL-C), total cholesterol (TC) and triglyceride (TG) levels have been shown to have heritability estimates ranging between 40% to 60% (Weiss et al. 2006).

Various genetic loci associated with lipid traits have been identified by genome-wide association studies (GWAS); however, these loci only explain 25-30% of the heritability observed in plasma lipid traits (Kathiresan et al. 2009; Teslovich et al. 2010). *Epistasis*, or interactions between genes, may help to explain a portion of the *missing heritability* of lipid traits (Manolio et al. 2009) and studies are needed to examine the genomic context of single nucleotide polymorphisms (SNPs) by specifically searching for non-linear interactions between them (Eichler et al. 2010).

Exhaustively searching for interactions between SNPs in large datasets generated with genotyping arrays leads to a prohibitive number of statistical tests and is computationally expensive (Moore et al. 2010). In this study, we addressed these bioinformatics challenges by applying SNP-filtering methods along with the quantitative multifactor dimensionality reduction (QMDR) machine learning algorithm to the analysis of lipid traits for the first time. We aimed to identify interactions between SNPs that are associated with four lipids traits (HDL-C, LDL-C, TC and TG) across five National Heart, Lung and Blood Institute (NHLBI) study cohorts. These SNPs have also been analyzed as part of a previous large-scale study aimed at identifying independent signals associated with multiple lipid traits using regression methods. This study identified 21 novel loci that had not been found to be associated with lipid traits before. The study also replicated a large number of previously implicated signals associated with lipid traits (Asselbergs et al. 2012).

Results

Main effect filter

The main effect filter resulted in a final list of 486 markers to be tested for SNP-SNP interactions for HDL-C, 462 markers for LDL-C, 571 markers for TC and 502 markers for TG. After QMDR analysis, at a permutation *P*-value < 0.05, we identified 20 SNP-SNP interaction models that were associated with HDL-C (Table 1), 6 SNP-SNP interaction models associated with LDL-C (Table 2), 3 SNP-SNP interaction models associated with TC (Table 3), and 10 SNP-SNP models associated with TG (Table 4).

Table 1 Results from QMDR association analysis of Main effect filter SNPs for HDL Cholesterol levels

Rank	Model	SNP1	Chr:bp	Gene1	SNP2	Chr:bp	Gene2	Permuted P-Value
1	rs4783961,rs1800775	rs4783961	16:55552395	*CETP*	rs1800775	16:55552737	*CETP*	< 0.00001
2	rs12720918,rs158477	rs12720918	16:55551713	*CETP*	rs158477	16:55565111	*CETP*	< 0.00001
3	rs4783961,rs1864163	rs4783961	16:55552395	*CETP*	rs1864163	16:55554734	*CETP*	< 0.00001
4	rs12720918,rs4783961	rs12720918	16:55551713	*CETP*	rs4783961	16:55552395	*CETP*	< 0.00001
5	rs1864163,rs4784744	rs1864163	16:55554734	*CETP*	rs4784744	16:55568686	*CETP*	< 0.00005
6	rs12708967,rs820299	rs12708967	16:55550712	*CETP*	rs820299	16:55557785	*CETP*	< 0.00024
7	rs12447924,rs9939224	rs12447924	16:55551693	*CETP*	rs9939224	16:55560233	*CETP*	< 0.00031
8	rs4783961,rs158477	rs4783961	16:55552395	*CETP*	rs158477	16:55565111	*CETP*	< 0.00087
9	rs1864163,rs158477	rs1864163	16:55554734	*CETP*	rs158477	16:55565111	*CETP*	< 0.00104
10	rs1864163,rs820299	rs1864163	16:55554734	*CETP*	rs820299	16:55557785	*CETP*	0.00405
11	rs4783961,rs9939224	rs4783961	16:55552395	*CETP*	rs9939224	16:55560233	*CETP*	< 0.00525
12	rs1800775,rs820299	rs1800775	16:55552737	*CETP*	rs820299	16:55557785	*CETP*	< 0.0083
13	rs12744291,rs1010554	rs12744291	1:66135559	*PDE4B*	rs1010554	3:52517959	*STAB1*	< 0.00873
14	rs230541,rs4935047	rs230541	4:103716823	*NFKB1*	rs4935047	10:54200073	*MBL2*	0.01116
15	rs12976922,rs2952101	rs12976922	19:60562163	*COX6B2*	rs2952101	X:14768515	*FANCB*	< 0.0119
16	rs9644636,rs7013777	rs9644636	8:19869176	*LPL*	rs7013777	8:19922636	*LPL*	< 0.01719
17	rs9939224,rs4784744	rs9939224	16:55560233	*CETP*	rs4784744	16:55568686	*CETP*	0.01843
18	rs599839,rs2952101	rs599839	1:109623689	*PSRC1*	rs2952101	X:14768515	*FANCB*	0.02954
19	rs12708967,rs158477	rs12708967	16:55550712	*CETP*	rs158477	16:55565111	*CETP*	0.03023
20	rs3870336,rs6641322	rs3870336	3:49532861	*DAG1*	rs6641322	X:149494622	*IDS*	< 0.04418

20 signals reached a permutation P-value < 0.05. SNPs have been mapped to their corresponding genes using dbSNP (build 139). SNP1 and SNP2 indicate the individual SNPs within a given SNP-SNP interaction model. Chromosomal location of SNPs is noted in the following format - Chromosome:Base pair. P-values were calculated from a distribution built from 1000 permutations.

Table 2 Results from QMDR association analysis of Main effect filter SNPs for LDL cholesterol levels

Rank	Model	SNP1	Chr:bp	Gene1	SNP2	Chr:bp	Gene2	Permuted P-Value
1	rs157580,rs439401	rs157580	19:50087106	*TOMM40*	rs439401	19:50106291	*APOE*	< 0.00522
2	rs17435152,rs3764261	rs17435152	7:40568630	*C7orf10*	rs3764261	16:55550825	*CETP*	0.00743
3	rs157580,rs405509	rs157580	19:50087106	*TOMM40*	rs405509	19:50100676	*APOE*	0.00781
4	rs12811752,rs1469713	rs12811752	12:20469072	*PDE3A*	rs1469713	19:19389806	*GATAD2A*	< 0.01293
5	rs480780,rs2965174	rs480780	13:32505319	*KL*	rs2965174	19:49936855	*BCL3*	0.02482
6	rs625619,rs3764261	rs625619	1:55290754	*PCSK9*	rs3764261	16:55550825	*CETP*	0.03809

6 signals reached a permutation P-value < 0.05. SNPs have been mapped to their corresponding genes using dbSNP (build 139). SNP1 and SNP2 indicate the individual SNPs within a given SNP-SNP interaction model. Chromosomal location of SNPs is noted in the following format - Chromosome:Base pair. P-values were calculated from a distribution built from 1000 permutations.

Table 3 Results from QMDR association analysis of Main effect filter SNPs for total cholesterol levels

Rank	Model	SNP1	Chr:bp	Gene1	SNP2	Chr:bp	Gene2	Permuted P-Value
1	rs693,rs661665	rs693	2:21085700	*APOB*	rs661665	2:21118646	*APOB*	< 0.00835
2	rs12898801,rs953065	rs12898801	15:56585846	*LIPC*	rs953065	15:87203929	*ACAN*	< 0.01297
3	rs10744777,rs749767	rs10744777	12:110717401	*ALDH2*	rs749767	16:31031908	*BCKDK*	0.01975

3 signals reached a permutation P-value < 0.05. SNPs have been mapped to their corresponding genes using dbSNP (build 139). SNP1 and SNP2 indicate the individual SNPs within a given SNP-SNP interaction model. Chromosomal location of SNPs is noted in the following format - Chromosome:Base pair. P-values were calculated from a distribution built from 1000 permutations.

Table 4 Results from QMDR association analysis of Main effect filter SNPs for triglyceride levels

Rank	Model	SNP1	Chr:bp	Gene1	SNP2	Chr:bp	Gene2	Permuted P-Value
1	rs2075295,rs6589568	rs2075295	11:116133611	*BUD13*	rs6589568	11:116175948	*APOA5*	< 0.00001
2	rs4938303,rs180327	rs4938303	11:116090197	*BUD13*	rs180327	11:116128869	*BUD13*	< 0.00001
3	rs180327,rs2075295	rs180327	11:116128869	*BUD13*	rs2075295	11:116133611	*BUD13*	< 0.00007
4	rs180327,rs10750097	rs180327	11:116128869	*BUD13*	rs10750097	11:116169250	*APOA5*	< 0.00027
5	rs11216129,rs10750097	rs11216129	11:116125466	*BUD13*	rs10750097	11:116169250	*APOA5*	< 0.00516
6	rs609526,rs12257915	rs609526	1:228375529	*GALNT2*	rs12257915	10:90982709	*LIPA*	< 0.03321
7	rs2075295,rs10750097	rs2075295	11:116133611	*BUD13*	rs10750097	11:116169250	*APOA5*	< 0.03351
8	rs4938303,rs6589568	rs4938303	11:116090197	*BUD13*	rs6589568	11:116175948	*APOA5*	< 0.03379
9	rs174455,rs689243	rs174455	11:61412693	*FADS3*	rs689243	11:116227903	*KIAA0999*	< 0.03789
10	rs180327,rs618923	rs180327	11:116128869	*BUD13*	rs618923	11:116159369	*ZNF259*	< 0.03833

10 signals reached a permutation P-value < 0.05. SNPs have been mapped to their corresponding genes using dbSNP (build 139). SNP1 and SNP2 indicate the individual SNPs within a given SNP-SNP interaction model. Chromosomal location of SNPs is noted in the following format - Chromosome:Base pair. P-values were calculated from a distribution built from 1000 permutations.

In the case of HDL-C, a large number of the identified SNP-SNP models represent intra-genic interactions within *CETP*. Fig. 1 shows the underlying LD structure of these interactions. None of the interacting SNPs were in strong LD ($r^2 > 0.8$). Moreover, none of the identified pairwise interactions for each of the quantitative lipid traits exhibited strong LD (Figs. S1-3).

Biofilter

The Biofilter procedure resulted in a final list of 1,811 markers (22,487 SNP-SNP models) for HDL-C, 1,812 markers (22,491 SNP-SNP models) for LDL-C, 1,812 markers (22,454 SNP-SNP models) for TC and 1,811 markers (22,487 SNP-SNP models) for TG. QMDR analysis identified 14 significant SNP-SNP models with a permutation *P*-value < 0.05: 2 SNP-SNP models associated with HDL-C, 3 SNP-SNP models associated with LDL-C, 1 SNP-SNP model associated with TC and 8 SNP-SNP models associated with TG (Table 5). None of the interacting SNPs were found to be in strong LD in this case as well (Figs. S4-7).

Table 5 Results from QMDR association analysis of Biofilter SNPs for lipid traits

Rank	Model	SNP1	Chr:bp	Gene1	SNP2	Chr:bp	Gene2	Permuted P-Value
HDL Cholesterol Levels								
1	rs17496549,rs615672	rs17496549	6:32517686	*HLA-DRA*	rs615672	6:32682149	*HLA-DRB1*	< 0.01178
2	rs549888,rs7240326	rs549888	6:33660180	*GGNBP1*	rs7240326	18:59068331	*BCL2*	0.0404
LDL Cholesterol Levels								
1	rs39499,rs751919	rs39499	8:90839744	*RIPK2*	rs751919	16:49333246	*CYLD*	0.03262
2	rs12693591,rs8072566	rs12693591	2:191568747	*STAT1*	rs8072566	17:37729889	*STAT3*	0.04211
3	rs2066795,rs8074524	rs2066795	2:191560142	*STAT1*	rs8074524	17:37723124	*STAT3*	< 0.04788
Total Cholesterol Levels								
1	rs4725431,rs10875915	rs4725431	7:151104112	*PRKAG2*	rs10875915	12:47716361	*MLL2*	< 0.04276
Triglyceride Levels								
1	rs9521510,rs2860184	rs9521510	13:109224872	*IRS2*	rs2860184	19:7238748	*INSR*	< 0.00079
2	rs9521510,rs6510976	rs9521510	13:109224872	*IRS2*	rs6510976	19:7217944	*INSR*	< 0.00289
3	rs2075110,rs4789172	rs2075110	7:55186653	*EGFR*	rs4789172	17:70853307	*GRB2*	< 0.00385
4	rs4773088,rs4804404	rs4773088	13:109219885	*IRS2*	rs4804404	19:7169382	*INSR*	< 0.00401
5	rs7999797,rs8109559	rs7999797	13:109224001	*IRS2*	rs8109559	19:7122629	*INSR*	0.01758
6	rs4771646,rs4804404	rs4771646	13:109225180	*IRS2*	rs4804404	19:7169382	*INSR*	< 0.02007
7	rs1729409,rs11216162	rs1729409	11:116178978	*APOA5*	rs11216162	11:116233487	*KIAA0999*	< 0.02063
8	rs7999797,rs7252268	rs7999797	13:109224001	*IRS2*	rs7252268	19:7121505	*INSR*	< 0.03164

Signals reached a permutation P-value < 0.05. SNPs have been mapped to their corresponding genes using dbSNP (build 139). SNP1 and SNP2 indicate the individual SNPs within a given SNP-SNP interaction model. Chromosomal location of SNPs is noted in the following format - Chromosome:Base pair. P-values were calculated from a distribution built from 1000 permutations.

Replication analyses

After following an identical QMDR analysis procedure, we were able to replicate SNP-SNP models in the eMERGE dataset at a permutation *P*-value threshold of 0.05. Eleven main effect filtered SNP-SNP models were replicated for HDL-C, 1 main effect filtered SNP-SNP model for LDL-C and 2 such models for TG (Table 6). Additionally, 23 Biofilter SNP-SNP models replicated for TG (Table 6).

Table 6 Results from QMDR association analysis of main effect and Biofilter SNP-SNP models replicated in eMERGE dataset

Rank	Model	SNP1	Chr:bp	Gene1	SNP2	Chr:bp	Gene2	Permuted P-Value
Main effect filter: HDL Cholesterol Levels								
1	rs4783961,rs1800775	rs4783961	16:55552395	*CETP*	rs1800775	16:55552737	*CETP*	< 0.00001
2	rs4783961,rs3816117	rs4783961	16:55552395	*CETP*	rs3816117	16:55553659	*CETP*	< 0.00001
3	rs4783961,rs1532624	rs4783961	16:55552395	*CETP*	rs1532624	16:55562980	*CETP*	< 0.00001
4	rs4783961,rs1532625	rs4783961	16:55552395	*CETP*	rs1532625	16:55562802	*CETP*	< 0.00001
5	rs4783961,rs7205804	rs4783961	16:55552395	*CETP*	rs7205804	16:55562390	*CETP*	< 0.00001
6	rs4783961,rs711752	rs4783961	16:55552395	*CETP*	rs711752	16:55553712	*CETP*	< 0.00001
7	rs4783961,rs708272	rs4783961	16:55552395	*CETP*	rs708272	16:55553789	*CETP*	< 0.00001
8	rs1864163,rs289717	rs1864163	16:55554734	*CETP*	rs289717	16:55566889	*CETP*	< 0.00004
9	rs1864163,rs4784744	rs1864163	16:55554734	*CETP*	rs4784744	16:55568686	*CETP*	< 0.00004
10	rs1864163,rs291044	rs1864163	16:55554734	*CETP*	rs291044	16:55568953	*CETP*	< 0.00004
11	rs4783961,rs1864163	rs4783961	16:55552395	*CETP*	rs1864163	16:55554734	*CETP*	< 0.00229
Main effect filter: LDL Cholesterol Levels								
1	rs157580,rs405509	rs157580	19:50087106	*TOMM40*	rs405509	19:50100676	*APOE*	< 0.00488
Main effect filter: Triglyceride Levels								
1	rs180327,rs618923	rs180327	11:116128869	*BUD13*	rs618923	11:116159369	*ZNF259*	< 0.31483
2	rs180326,rs618923	rs180326	11:116129913	*BUD13*	rs618923	11:116159369	*ZNF259*	< 0.33204
Biofilter: Triglyceride Levels								
1	rs9521510,rs6510976	rs9521510	13:109224872	*IRS2*	rs6510976	19:7217944	*INSR*	< 0.01014
2	rs35612086,rs6510976	rs35612086	13:109244865	*IRS2*	rs6510976	19:7217944	*INSR*	< 0.01155
3	rs36092351,rs6510976	rs36092351	13:109246741	*IRS2*	rs6510976	19:7217944	*INSR*	< 0.01163
4	rs2117455,rs6510976	rs2117455	13:109241895	*IRS2*	rs6510976	19:7217944	*INSR*	< 0.01166
5	rs9521517,rs6510976	rs9521517	13:109245638	*IRS2*	rs6510976	19:7217944	*INSR*	< 0.01166
6	rs1414316,rs6510976	rs1414316	13:109248190	*IRS2*	rs6510976	19:7217944	*INSR*	< 0.01168
7	rs9521510,rs6510975	rs9521510	13:109224872	*IRS2*	rs6510975	19:7217878	*INSR*	< 0.01259

8	rs9515119,rs6510975	rs9515119	13:109207337	*IRS2*	rs6510975	19:7217878	*INSR*	< 0.01643
9	rs9515119,rs6510976	rs9515119	13:109207337	*IRS2*	rs6510976	19:7217944	*INSR*	< 0.01657
10	rs2289046,rs6510975	rs2289046	13:109205907	*IRS2*	rs6510975	19:7217878	*INSR*	< 0.01657
11	rs2289047,rs6510975	rs2289047	13:109205816	*IRS2*	rs6510975	19:7217878	*INSR*	< 0.01657
12	rs2289046,rs6510976	rs2289046	13:109205907	*IRS2*	rs6510976	19:7217944	*INSR*	0.01680
13	rs2289047,rs6510976	rs2289047	13:109205816	*IRS2*	rs6510976	19:7217944	*INSR*	0.01680
14	rs4771649,rs6510976	rs4771649	13:109248514	*IRS2*	rs6510976	19:7217944	*INSR*	< 0.01709
15	rs35612086,rs6510975	rs35612086	13:109244865	*IRS2*	rs6510975	19:7217878	*INSR*	< 0.02214
16	rs36092351,rs6510975	rs36092351	13:109246741	*IRS2*	rs6510975	19:7217878	*INSR*	< 0.02220
17	rs2117455,rs6510975	rs2117455	13:109241895	*IRS2*	rs6510975	19:7217878	*INSR*	< 0.02223
18	rs9521517,rs6510975	rs9521517	13:109245638	*IRS2*	rs6510975	19:7217878	*INSR*	0.02223
19	rs1414316,rs6510975	rs1414316	13:109248190	*IRS2*	rs6510975	19:7217878	*INSR*	< 0.02225
20	rs2075110,rs4789172	rs2075110	7:55186653	*EGFR*	rs4789172	17:70853307	*GRB2*	< 0.02793
21	rs4771649,rs6510975	rs4771649	13:109248514	*IRS2*	rs6510975	19:7217878	*INSR*	0.03304
22	rs2075109,rs4789172	rs2075109	7:55186397	*EGFR*	rs4789172	17:70853307	*GRB2*	0.04293
23	rs9521518,rs6510976	rs9521518	13:109251997	*IRS2*	rs6510976	19:7217944	*INSR*	0.04882

Shown here are models that reached a permutation P-value < 0.05 in the replication dataset. SNPs have been mapped to their corresponding genes using dbSNP (build 139). SNP1 and SNP2 indicate the individual SNPs within a given SNP-SNP interaction model. . Chromosomal location of SNPs is noted in the following format - Chromosome:Base pair. P-values were calculated from a distribution built from 1000 permutations.

Discussion

Although many researchers acknowledge the need for embracing the complexity of the genotype-phenotype relationship by studying gene-gene interactions, exploring epistasis in large genotyping arrays presents a biostatistical and computational challenge. These challenges call for new computational methods since more traditional approaches such as general linear models may have limited power when modeling high-dimensional data. The use of SNP-filtering methods has been presented as a suitable solution to ease the computational burden of exhaustively searching for all possible interactions between large numbers of SNPs (Moore et al. 2010).

In our analyses, we combined genotypic and phenotypic information for four quantitative lipid traits – HDL-C, LDL-C, TC and TG – for 24,837 individuals from five study cohorts. We reduced the number of interactions tested by filtering SNPs either based on the strength of their independent effects or the strength of relevant prior biological knowledge. Filtered SNPs were tested for two-way SNP-SNP interactions associated with each quantitative lipid trait using QMDR.

Below we highlight the potential biological functions for several genes at or near the identified interacting SNPs. Further details regarding the biological roles and functions associated with all genes at or near these interactions are listed in Table S1.

HDL-C

In our analysis, the 25 total SNPs that we identified to interact significantly with each other, were in or near 15 genes. Ten SNPs were located at or near *CETP*, which is involved in the transfer of cholesteryl ester from HDL to other lipoproteins (Barter et al. 2003). Moreover, we identified 12 intra-genic interactions between independent SNPs within the *CETP* region (Table 1). Three of the original intra-genic *CETP* interactions (interactions 1, 9 and 11 in Table 6) and 8 additional LD expanded interactions were replicated in the eMERGE dataset. Though the impact of these intra-genic regions on HDL-C is unknown, they may act through regulatory or epigenetic mechanisms (Soto-Ramírez et al. 2013; Olsson et al. 2014).

~~Some of the other identified genes also have well studied roles in lipid and cholesterol metabolism such as *LPL* and *PSRC1* (Brown et al. 1989; Kuivenhoven et al. 1997). Both genes were identified in interactions in the discovery dataset only. Two SNPs were at or near *LPL*; mutations in *LPL* are linked to various disorders of lipoprotein metabolism and have been previously reported to alter HDL-C levels (Reymer et al. 1995; Wittekoek et al. 1998). We also identified one intra-genic interaction within *LPL* (Table 1). Lastly, 1 SNP was near *PSRC1*. Variations within *PSRC1* have also been shown be associated with cholesterol traits in previous GWA studies (Kathiresan et al. 2008a; Ma et al. 2010; Voight et al. 2012).~~

The IMP network of genes represented in main effect filtered SNP-SNP models associated with HDL-C, includes genes from replicated and non-replicated interactions.

The interaction between *PSRC1* with *BRCA1* via *AURKA* in this network, highlights a potentially interesting biological connection between dyslipidemia and breast cancer (Fig. 2). High cholesterol has been highlighted as a risk factor for breast cancer and various mechanisms linking the two diseases have been hypothesized and studied (Nelson et al. 2014). The interaction between *PSRC1* and *AURKA* in this network reflects their well-known role in spindle organization. There was also strong support for the interaction between *AURKA* and *BRCA1*. This is not surprising since, AURKA is a known to be an activator of Akt (Yao et al. 2009) – a kinase involved in tumor cell growth (Paplomata and O'Regan 2014).

LDL-C

We identified 16 total SNPs to interact significantly with each other, located in or near 13 genes. These findings included two SNP-SNP models representing interactions between the genes *TOMM40* and *APOE*. One of these interaction models replicated in the eMERGE dataset (Table 6). The *TOMM40/APOE-C1-C2-C4* gene cluster has been shown to affect LDL-C levels previously (Klos et al. 2008; Middelberg et al. 2011). In the IMP network built from SNP-SNP models identified after main effect filtering, there is strong support for the interaction between *TOMM40* and *FARSA*, which encodes for the alpha subunit of a phenylalanyl-tRNA synthetase (Fig. 3). FARSA is also involved in a protein-protein interaction with the ECSIT signaling integrator, which in turn interacts with APOE. Furthermore, *APOE* interacts with *LDLR* in the network, highlighting their shared role in sterol transport and cholesterol homeostasis (Fig. 3). *LDLR*, which encodes for the LDL receptor, also interacts with *PCSK9* in the network since both genes share a role in cholesterol homeostasis (Fig. 3). PCSK9 binds LDLR and promotes degradation of the LDL receptor either in the lysosome or in the liver (Cao et al. 2011). Hence, due to its inhibitory role with LDLR, it has emerged as an attractive drug target for hypercholesterolemia (Akram et al. 2010).

The IMP network also had an enrichment of processes such as - cholesterol and lipid homeostasis, cholesterol transport, regulation of plasma lipoprotein particle levels, plasma lipoprotein particle clearance, and low density lipoprotein particle receptor catabolic process. The genes involved in these processes included *APOE, CETP* and *PCSK9*. SNPs within these genes have been previously found to be associated with LDL-C (Talmud et al. 2009).

Total Cholesterol

There were 3 SNP-SNP interactions that were significantly associated with TC after main effect filtering. An intra-genic interaction within *APOB* was most significantly associated with TC after main effect filtering, although it did not replicate in the eMERGE dataset. Mutations within *APOB* can cause familial defective apolipoprotein B-100 (FDB) – an inherited form of hypercholesterolemia (Hooper et al. 2005). ~~The protein encoded by this gene forms the building block for various types of low-density lipoproteins. It is also involved in cholesterol homeostasis and sterol transport.~~

~~Researchers have also found a polymorphism on this gene to increase LDL-C levels (Benn et al. 2005).~~

There was one significant SNP-SNP interaction associated with TC after Biofilter filtering which also did not replicate in the eMERGE dataset. ~~The gene-gene interaction between *MLL2* and *PRKAG2* highlights biological processes such as histone methylation, protein alkylation and protein methylation (Wong et al. 2012).~~ *MLL2* codes for a mixed-lineage leukemia histone methylase which contributes to the activation of SR-B1 – a class-B type-1 scavenger receptor responsible for maintaining blood cholesterol levels (Ansari et al. 2012). ~~*PRKAG2* encodes for the regulatory γ2 subunit of an AMP-activated protein kinase.~~ Homozygotes of an intronic SNP within *PRKAG2* have been found to have elevated serum concentrations of TC and TG (Xu et al. 2005).

Triglycerides

We found an interaction between *BUD13* and *ZNF259*. This interaction and an LD-expanded SNP-SNP model representing it were replicated in the eMERGE dataset (Table 6). An interaction between variants on these two genes has been found to be associated with TG and TC before (Aung et al. 2014). Moreover, many studies have found polymorphisms within *BUD13* to be associated with TG (Kathiresan et al. 2008b; Waterworth et al. 2010; Aung et al. 2014). *BUD13* encodes for the BUD13 homolog protein. It is part of the RES complex that was originally identified as a splicing factor in yeast and shown to affect nuclear pre-mRNA retention (Brooks et al. 2009).

Six of the eight SNP-SNP interactions associated with TG after Biofilter filtering, represent an interaction between the genes *INSR* and *IRS2*. Twenty-one models representing this SNP-SNP interaction were identified in the eMERGE dataset as well (Table 6). This included two of the original SNP-SNP interactions between these genes (interactions 1 and 20 in Table 6). *INSR* encodes for the insulin receptor, which works with the IRS2 molecule in hepatic insulin signaling. Insulin is also known to activate lipogenesis within the liver. Moreover, an inverse relationship between *IRS2* and *SREBP-1* gene expression has been demonstrated (Ide et al. 2004). SREBPs are transcription factors that are involved in the expression of various genes involved in the synthesis of triglycerides (Horton et al. 2002).

Strengths and Limitations

Despite the computational and biostatistical challenges of investigating gene-gene interactions in datasets from large genotyping arrays, we have established an efficient analytic framework to overcome the limited power of traditional statistical methods when modeling high-dimensional data. The use of knowledge-based filtering methods within our framework improved our ability to identify biologically relevant interactions in the context of lipid phenotypes.

However, our methods are limited by the strength of the knowledge of gene functions available in public databases. Additionally, our replication sample was small which could have decreased our power to replicate the interactions we identified in our discovery dataset. Lastly, although the use of SNP-filtering methods have been suggested as a favorable solution for reducing the computational burden of studying epistasis in such large datasets, they do introduce their own biases into the study, which have been discussed previously (Ritchie 2011).

The use of traditional statistical methods focusing on main effects has been able to explain only a portion of the heritability of lipid traits. We performed a comprehensive analysis by examining gene-gene interactions within four quantitative lipid traits – HDL-C, LDL-C, TC and TG, from five study cohorts. With the use of machine learning algorithms such as QMDR, a targeted gene-centric genotypic chip and SNP-filtering methods, we identified multiple gene-gene interactions associated with these lipid traits. Existing knowledge suggests potentially important roles for these genes in the pathobiology of lipid traits. Ultimately, the true effect of these interactions will have to be validated through targeted functional studies at the bench.

Subjects and Methods

Participating Cohorts

The overall study design is shown in Fig. 4. Genotype and phenotype information was combined from the following studies: Atherosclerosis Risk in Communities (ARIC) (Hill et al. 1989); Coronary Artery Risk Development in Young Adults (CARDIA) (Friedman et al. 1988); Cardiovascular Health Study (CHS) (Fried et al. 1991); Framingham Heart Study (FHS) (Dawber et al. 1951); and Multi-Ethnic Study of Atherosclerosis (MESA) (Bild et al. 2002) (Table S2), resulting in an initial sample size of 24,837 individuals of self-reported European ancestry.

The eMERGE I-660 dataset was used for replication analyses (McCarty et al. 2011). This dataset was imputed using data from the 1000 Genomes Project (Verma et al. 2014). Detailed information regarding the replication dataset is presented in Table S3.

Phenotypic outcomes measured

HDL-C, LDL-C, TC and TG levels were measured from baseline or first measurement blood samples. All measurements were converted to mmol/L. LDL-C was calculated according to Friedewald's formula (Friedewald et al. 1972):

$$L \sim C - H - kT$$

where C is total cholesterol, H is HDL, L is LDL, T is triglycerides, and k is 0.45 for mmol/L. If TG values were > 4.51 mmol/L, then LDL was treated as a missing value. Additionally, TG values were transformed for normality.

Genotyping and quality control

Study participants in the discovery dataset were genotyped using the gene-centric ITMAT-Broad-CARe (IBC) array. The IBC array contains 47, 451 SNPs and it was designed to test ~ 2,100 loci that have been implicated in various cardiovascular, metabolic and inflammatory phenotypes (Keating et al. 2008). SNPs with a genotype missing rate greater than 95%, with an exact test of Hardy-Weinberg equilibrium *P*-value $< 1.0 \times 10^{-7}$ or a minor allele frequency (MAF) < 0.05 were excluded. Samples with a genotype missing rate greater than 90% were also excluded. This reduced our dataset to 24, 837 individuals and 44,570 SNPs.

Non-founder individuals were also removed from the study population. To check for relatedness between individuals, identity-by-descent (IBD) estimates were calculated using PLINK (Purcell et al. 2007). For each pair of individuals with a $\hat{\pi} > 0.3$, one individual was removed.

Finally, individual datasets with no missing phenotype data were created for each of the lipid outcomes measured. Within each of the datasets, SNPs were further tested for linkage disequilibrium (LD) – a SNP was removed from each pair of SNPs that had an LD $(r^2) \geq 0.6$. Genotypes were also imputed, to ensure there was no missing genotype information. The most common genotype for a given marker was used as the imputed genotype. Further details of the number of SNPs and individuals in each of these datasets can be found in Fig. 4.

Study participants in the replication dataset were from the eMERGE network. The eMERGE network is a consortium of institutions with DNA from biorepositories linked to data from patient electronic medical records (EMR) (Gottesman et al. 2013). The eMERGE set was genotyped with the Illumina660W GWAS platform and further imputed using 1000 Genomes project data, as described previously (Verma et al. 2014). The replication set consisted of data from the Marshfield Clinic, Northwestern University, Group Health Cooperative, Mayo Clinic, and Vanderbilt University. After QC, the final eMERGE sample size was n=7,502 for all lipid traits. Details on quality control and phenotype extractions from the EMR have been published previously (Rasmussen-Torvik et al. 2012). Briefly, each cohort tested for population stratification and relatedness, adjusting accordingly. The minimum variant and sample call rate threshold for all replication cohorts was 0.95 and 0.90, respectively. A Hardy-Weinberg equilibrium test *P*-value threshold of at least $P < 1 \times 10^{-6}$ was applied by each group.

Marker Selection

To reduce the computational time burden and multiple hypothesis testing, additional parallel SNP filtering steps (main effect filter and Biofilter) were employed. These strategies have been implemented by other studies as two powerful options for gene-gene interaction analysis in large-scale genotype datasets (Sun et al. 2014).

Main Effect Filter

SNPs were tested for their independent association with the continuous lipid outcome using linear regression (Asselbergs et al. 2012). SNPs with a main effect *P*-value < 0.01 were selected for further analysis.

Biofilter

SNPs were also analyzed using Biofilter 2.0, a knowledge-based software package that enables the analysis and identification of multi-SNP models in large datasets (Bush et al. 2009; Pendergrass et al. 2013). It has previously been used to identify predictive SNP-SNP models for traits such as age-related cataract (Hall et al. 2015), multiple sclerosis (Bush et al. 2011), HIV pharmacogenetics (Grady et al. 2011) and HDL cholesterol (Turner et al. 2011). The software combines information from various online public knowledge databases to identify genes and SNPs that are most likely to interact with each other through their mutual participation in biological processes, signaling pathways and protein-protein interactions. Biofilter also provides an *implication index*, which measures the strength of the knowledge-based support for a putative interaction model. This is indicated by the sum of the number of supporting data sources for each of the genes in a given interaction. In our analyses, we included models if they were supported by at least five sources. This was a slightly more stringent *implication index* cut-off than those used in previous studies (Turner et al. 2011).

Statistical Analyses

Covariate Adjustment

Quantitative lipid outcome values were regressed on age, sex, BMI, use of medications for lowering lipids, first ten principal components addressing population substructure, type II diabetes status and smoking status. The residual lipid outcome values from this regression model were then used as the continuous phenotypic outcome variable in QMDR analysis. Principal components were computed using the EIGENSTRAT software (Price et al. 2006))

Association Analysis using QMDR

SNPs obtained from the filtering procedures described above were tested for association with the corresponding continuous lipid outcome using QMDR. QMDR is an extension of the two-class MDR algorithm used to detect and characterize multi-SNP interactions in the context of a quantitative trait (Ritchie et al. 2001; Gui et al. 2013).

Originally, the MDR algorithm was designed as a data reduction method to enable the identification of multi-locus genotype combinations that are associated with high or low risk of a disease (Ritchie et al. 2001). For a dataset of *m* SNPs, *k* SNPs can be selected to study a *k*-order interaction. Next, a contingency table is constructed and case-control ratios are calculated for each of the possible multi-locus genotypes for these *k* SNPS. The case-control ratio for each multi-locus genotype is then compared to the case-control ratio for the whole dataset. If the genotype-specific case-control ratio exceeds the case-control ratio for the dataset, it is considered to be *high-risk*, otherwise it is considered to be *low-risk*.

However, in the case of QMDR, the algorithm compares the mean value of the phenotype for a specific multi-locus genotype, to the overall mean of the phenotype within the entire dataset. Consequently, a genotype combination is considered *high-level* if its mean phenotype value is larger than the overall mean of the phenotype. Otherwise, it is considered *low-level*. Finally, QMDR combines the *'high-level'* and *'low-level'* genotypes into separate groups and compares the phenotypic outcomes between these two groups using a T-test.

QMDR also involves a 10-fold cross-validation procedure similar to the original MDR algorithm. The data is divided into 10 portions – 9 portions are used as a training dataset and the remaining portion is used as a testing dataset. The algorithm repeats the procedure described above and calculates the overall mean of the phenotype separately for the training and the testing dataset. The training t-statistic is calculated for each *k*-way interaction in the training dataset. Next, the *k*-way model with the best training score is used to predict the case-control status in the testing dataset. The training t-statistic score is used to choose the best *k*-order interaction model and the highest testing t-statistic is used to select the best interaction model for the dataset.

In our analyses, we used QMDR to analyze filtered SNPs for all possible SNP-SNP interaction models that are associated with a given continuous lipid outcome (HDL-C, LDL-C, TC and TG) based on their training t-statistic scores. Amongst these models, the 100 best overall SNP-SNP models were selected using their testing t-statistic scores. Additionally, we used linear regression to adjust for the main effect of each SNP within a SNP-SNP model tested by QMDR. This was performed to increase our ability to identify pairwise interactions that are not primarily driven by the strong independent effects of the participating SNPs within a model.

Permutation testing to assess statistical significance

We also performed 1000 permutations to establish a null distribution and determine the threshold for an α=0.05 significance level. Identical to our analysis procedure, the 100 best SNP-SNP models were selected based on their t-statistic training and testing values for each permuted dataset. The null distribution built from the 100 best SNP-SNP models from all permutations and their corresponding t-statistic values was utilized to calculate *P*-values.

Mapping SNPs to genes

SNPs within the statistically significant pairwise interactions for each quantitative lipid trait were mapped to a corresponding gene using dbSNP (build 139) and SCANdb (www.scandb.org).

Integrated Multi-Species Prediction (IMP) web server

We also used the Integrated Multi-Species Prediction (IMP) web server to query genes represented by the SNPs within identified interactions (Wong et al. 2012). IMP integrates biological evidence from multiple information sources such as experimentally verified information from gene expression studies, IntAct, MINT, MIPS, and BioGRID databases. The software mines empirical data to provide a probability score that two genes are involved in a functional and biological relationship.

Replication Analyses

SNP-SNP models with a permutation *P*-value < 0.05 were chosen for replication in the eMERGE dataset (McCarty et al. 2011). We also identified all SNP-SNP models that were in LD with the identified significant models. SNPs that are in high LD ($r^2 > 0.8$) with the SNPs in the interaction models, were identified using SNAP (Johnson et al. 2008). This data was used to generate a list of ‘proxy’ SNP-SNP models representing the original significant interaction models. Both the statistically significant original models and the proxy models representing them were tested for replication. Table S4 shows the number of models tested per lipid quantitative trait. Additional details of the number of LD expanded models generated and tested for each original model are presented in Table S5. The same QMDR analysis procedure was performed as described earlier.

Acknowledgements

CARe acknowledges the support of the National Heart, Lung and Blood Institute and the contributions of the research institutions, study investigators, field staff, and study participants in creating this resource for biomedical research (NHLBI contract number HHSN268200960009C). The IBC array data (also known as 'Cardiochip' or 'CVDSNP55v1_A' from the National Heart, Lung and Blood Institute (NHLBI) Candidate Gene Association Resource (CARe) was downloaded with appropriate permissions from the database of Genotypes and Phenotypes (dbGaP) (www.ncbi.nlm.gov/gap). The imputed genotype data for eMERGE-I and eMERGE-II can be downloaded from the database of Genotypes and Phenotypes (dbGaP) (www.ncbi.nlm.gov/gap).

Funding Statement

This work was supported by National Institutes of Health grants: NLM R01 grants (LM0l0098, LM011360, LM009012), GMS P20 grants (GM103506, GM103534 and GM104416), and F31 HG008588. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

eMERGE Network (Phase II – Year 1) Acknowledgement

The eMERGE Network was initiated and funded by the National Human Genome Research Institute (NHGRI) through the following grants: U01HG006389 (Essentia

Institute of Rural Health, Marshfield Clinic Research Foundation and Pennsylvania State University); U01HG006382 (Geisinger Clinic); U01HG006375 (Group Health Cooperative/University of Washington); U01HG006379 (Mayo Clinic); U01HG006380 (Icahn School of Medicine at Mount Sinai); U01HG006388 (Northwestern University); U01HG006378 (Vanderbilt University Medical Center); and U01HG006385 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG004438 (CIDR) and U01HG004424 (the Broad Institute) serving as Genotyping Centers.

eMERGE Network (Phase I) Acknowledgement

The eMERGE Network was initiated and funded by the National Human Genome Research Institute (NHGRI), in conjunction with additional funding from the National Institute of General Medical Sciences (NIGMS) through the following grants: U01-HG-004610 (Group Health Cooperative/University of Washington); U01-HG-004608 (Marshfield Clinic Research Foundation and Vanderbilt University Medical Center); U01-HG-04599 (Mayo Clinic); U01HG004609 (Northwestern University); U01-HG-04603 (Vanderbilt University Medical Center, also serving as the Administrative Coordinating Center); U01HG004438 (CIDR) and U01HG004424 (the Broad Institute) serving as Genotyping Centers.

Competing Interests

The authors declare that no competing interests exist.

References

Akram ON, Bernier A, Petrides F, et al (2010) Beyond LDL Cholesterol, a New Role for PCSK9. Arterioscler Thromb Vasc Biol 30 :1279–1281. doi: 10.1161/ATVBAHA.110.209007

Ansari KI, Kasiri S, Hussain I, et al (2012) MLL Histone Methylases Regulate Expression of HDLR-SR-B1 in Presence of Estrogen and Control Plasma Cholesterol in Vivo. Mol Endocrinol 27:92–105. doi: 10.1210/me.2012-1147

Arsenault BJ, Boekholdt SM, Kastelein JJP (2011) Lipid parameters for measuring risk of cardiovascular disease. Nat Rev Cardiol 8:197–206.

Asselbergs FW, Guo Y, Van Iperen EP a, et al (2012) Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci. Am J Hum Genet 91:823–838. doi: 10.1016/j.ajhg.2012.08.032

Aung L-H-H, Yin R-X, Wu J-Z, et al (2014) Association between the MLX Interacting Protein-Like, BUD13 Homolog and Zinc Finger Protein 259 Gene Polymorphisms and Serum Lipid Levels.

Barter PJ, Brewer HB, Chapman MJ, et al (2003) Cholesteryl Ester Transfer Protein: A Novel Target for Raising HDL and Inhibiting Atherosclerosis. Arterioscler Thromb Vasc Biol 23:160–167. doi: 10.1161/01.ATV.0000054658.91146.64

Benn M, Nordestgaard BG, Jensen JS, et al (2005) Polymorphism in APOB associated with increased low-density lipoprotein levels in both genders in the general population. J Clin Endocrinol Metab 90:5797–5803. doi: 10.1210/jc.2005-0974

Bild DE, Bluemke DA, Burke GL, et al (2002) Multi-ethnic study of atherosclerosis:

objectives and design. Am J Epidemiol 156:871–81.

Brooks MA, Dziembowski A, Quevillon-Cheruel S, et al (2009) Structure of the yeast Pml1 splicing factor and its integration into the RES complex. Nucleic Acids Res 37:129–143. doi: 10.1093/nar/gkn894

Brown ML, Inazu A, Hesler CB, et al (1989) Molecular basis of lipid transfer protein deficiency in a family with increased high-density lipoproteins. Nature 342:448–451. doi: 10.1038/342448a0

Bush WS, Dudek SM, Ritchie MD (2009) Biofilter: A Knowledge-Integration System for the Multi-Locus Analysis of Genome-Wide Association Studies. Pacific Symp Biocomput 368–379.

Bush WS, McCauley JL, DeJager PL, et al (2011) A knowledge-driven interaction analysis reveals potential neurodegenerative mechanism of multiple sclerosis susceptibility. Genes Immun 12:335–340. doi: 10.1038/gene.2011.3

Cao A, Wu M, Li H, Liu J (2011) Janus kinase activation by cytokine oncostatin M decreases PCSK9 expression in liver cells. J Lipid Res 52:518–530. doi: 10.1194/jlr.M010603

Dawber TR, Meadors GF, Moore FE (1951) Epidemiological approaches to heart disease: the Framingham Study. Am J Public Health Nations Health 41:279–81.

Deaton C, Froelicher ES, Wu LH, et al (2011) The global burden of cardiovascular disease. Eur J Cardiovasc Nurs 10:S5–S13. doi: 10.1016/S1474-5151(11)00111-3

Eichler EE, Flint J, Gibson G, et al (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446–50. doi:

10.1038/nrg2809

Fried LP, Borhani NO, Enright P, et al (1991) The Cardiovascular Health Study: design and rationale. Ann Epidemiol 1:263–76.

Friedewald WT, Levy RI, Fredrickson DS (1972) Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clin Chem 18:499–502.

Friedman GD, Cutter GR, Donahue RP, et al (1988) CARDIA: study design, recruitment, and some characteristics of the examined subjects. J Clin Epidemiol 41:1105–16.

Gottesman O, Kuivaniemi H, Tromp G, et al (2013) The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med 15:761–771. doi: 10.1038/gim.2013.72

Grady BJ, Torstenson ES, McLaren PJ, et al (2011) Use of biological knowledge to inform the analysis of gene-gene interactions involved in modulating virologic failure with efavirenz-containing treatment regimens in ART-naive ACTG clinical trials participants. Pac Symp Biocomput 253–264.

Gui J, Moore JH, Williams SM, et al (2013) A Simple and Computationally Efficient Approach to Multifactor Dimensionality Reduction Analysis of Gene-Gene Interactions for Quantitative Traits. PLoS One 8:e66545. doi: 10.1371/journal.pone.0066545

Hall MA, Verma SS, Wallace J, et al (2015) Biology-Driven Gene-Gene Interaction Analysis of Age-Related Cataract in the eMERGE Network. Genet Epidemiol 39:376–384. doi: 10.1002/gepi.21902

Heller DA, de Faire U, Pedersen NL, et al (1993) Genetic and Environmental Influences on Serum Lipid Levels in Twins. N Engl J Med 328:1150–1156. doi: 10.1056/NEJM199304223281603

Hill C, Gerardo D, James F, et al (1989) The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. Am J Epidemiol 129:687–702.

Hooper AJ, van Bockxmeer FM, Burnett JR (2005) Monogenic hypocholesterolaemic lipid disorders and apolipoprotein B metabolism. Crit Rev Clin Lab Sci 42:515–545. doi: 10.1080/10408360500295113

Horton JD, Goldstein JL, Brown MS (2002) SREBPs: activators of the complete program of cholesterol and fatty acid synthesis in the liver. J Clin Invest 109:1125–1131. doi: 10.1172/JCI15593

Ide T, Shimano H, Yahagi N, et al (2004) SREBPs suppress IRS-2-mediated insulin signalling in the liver. Nat Cell Biol 6:351–357.

Johnson AD, Handsaker RE, Pulit SL, et al (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24:2938–9. doi: 10.1093/bioinformatics/btn564

Kathiresan S, Melander O, Guiducci C, et al (2008a) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet 40:189–197.

Kathiresan S, Melander O, Guiducci C, et al (2008b) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet 40:189–197. doi: 10.1038/ng.75

Kathiresan S, Willer CJ, Peloso GM, et al (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet 41:56–65. doi: 10.1038/ng.291

Keating BJ, Tischfield S, Murray SS, et al (2008) Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. PLoS One 3:e3583. doi: 10.1371/journal.pone.0003583

Klos K, Shimmin L, Ballantyne C, et al (2008) APOE/C1/C4/C2 hepatic control region polymorphism influences plasma apoE and LDL cholesterol levels. Hum Mol Genet 17 :2039–2046. doi: 10.1093/hmg/ddn101

Kuivenhoven JA, de Knijff P, Boer JMA, et al (1997) Heterogeneity at the CETP Gene Locus : Influence on Plasma CETP Concentrations and HDL Cholesterol Levels . Arterioscler Thromb Vasc Biol 17 :560–568. doi: 10.1161/01.ATV.17.3.560

Ma L, Yang J, Runesha HB, 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. doi: 10.1186/1471-2350-11-55

Manolio TA, Collins FS, Cox NJ, et al (2009) Finding the missing heritability of complex diseases. Nature 461:747–53. doi: 10.1038/nature08494

McCarty CA, Chisholm RL, Chute CG, et al (2011) The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics 4:13. doi: 10.1186/1755-8794-4-13

Middelberg RPS, Ferreira MAR, Henders AK, et al (2011) Genetic variants in LPL, OASL and TOMM40/APOE-C1-C2-C4 genes are associated with multiple cardiovascular-related traits. BMC Med Genet 12:123. doi: 10.1186/1471-2350-12-123

Moore JH, Asselbergs FW, Williams SM (2010) Bioinformatics challenges for genome-wide association studies. Bioinformatics 26:445–55. doi: 10.1093/bioinformatics/btp713

Nelson ER, Chang C, McDonnell DP (2014) Cholesterol and breast cancer pathophysiology. Trends Endocrinol Metab 25:649–55. doi: 10.1016/j.tem.2014.10.001

Olsson AH, Volkov P, Bacos K, et al (2014) Genome-Wide Associations between Genetic and Epigenetic Variation Influence mRNA Expression and Insulin Secretion in Human Pancreatic Islets. PLoS Genet 10:e1004735. doi: 10.1371/journal.pgen.1004735

Paplomata E, O’Regan R (2014) The PI3K/AKT/mTOR pathway in breast cancer: targets, trials and biomarkers. Ther Adv Med Oncol 6:154–166. doi: 10.1177/1758834014530023

Pendergrass SA, Frase A, Wallace J, et al (2013) Genomic analyses with biofilter 2.0: knowledge driven filtering, annotation, and model development. BioData Min 6:25. doi: 10.1186/1756-0381-6-25

Price AL, Patterson NJ, Plenge RM, et al (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–9. doi: 10.1038/ng1847

Purcell S, Neale B, Todd-Brown K, et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–75. doi: 10.1086/519795

Rasmussen-Torvik LJ, Pacheco JA, Wilke RA, et al (2012) High density GWAS for LDL cholesterol in African Americans using electronic medical records reveals a strong protective variant in APOE. Clin Transl Sci 5:394–399. doi: 10.1111/j.1752-8062.2012.00446.x

Reymer PW, Gagne E, Groenemeyer BE, et al (1995) A lipoprotein lipase mutation (Asn291Ser) is associated with reduced HDL cholesterol levels in premature atherosclerosis. Nat Genet 10:28–34. doi: 10.1038/ng0595-28

Ritchie MD (2011) Using Biological Knowledge to Uncover the Mystery in the Search for Epistasis in Genome-Wide Association Studies. Ann Hum Genet 75:172–182. doi: 10.1111/j.1469-1809.2010.00630.x.Using

Ritchie MD, Hahn LW, Roodi N, et al (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138–47. doi: 10.1086/321276

Soto-Ramírez N, Arshad SH, Holloway JW, et al (2013) The interaction of genetic variants and DNA methylation of the interleukin-4 receptor gene increase the risk of asthma at age 18 years. Clin Epigenetics 5:1. doi: 10.1186/1868-7083-5-1

Sun X, Lu Q, Mukheerjee S, et al (2014) Analysis pipeline for the epistasis search - statistical versus biological filtering. Front Genet 5:106. doi: 10.3389/fgene.2014.00106

Talmud PJ, Drenos F, Shah S, et al (2009) Gene-centric Association Signals for Lipids and Apolipoproteins Identified via the HumanCVD BeadChip. Am J Hum Genet 85:628–642. doi: 10.1016/j.ajhg.2009.10.014

Teslovich TM, Musunuru K, Smith A V, et al (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466:707–713. doi: 10.1038/nature09270

Turner SD, Berg RL, Linneman JG, et al (2011) Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks. PLoS One 6:e19586. doi: 10.1371/journal.pone.0019586

Verma SS, de Andrade M, Tromp G, et al (2014) Imputation and quality control steps for combining multiple genome-wide datasets. Front Genet 5:1–15. doi: 10.3389/fgene.2014.00370

Voight BF, Peloso GM, Orho-Melander M, et al (2012) Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet 380:572–80. doi: 10.1016/S0140-6736(12)60312-2

Waterworth DM, Ricketts SL, Song K, et al (2010) Genetic variants influencing circulating lipid levels and risk of coronary artery disease. Arterioscler Thromb Vasc Biol 30:2264–2276. doi: 10.1161/ATVBAHA.109.201020

Weiss LA, Pan L, Abney M, Ober C (2006) The sex-specific genetic architecture of quantitative traits in humans. Nat Genet 38:218–222. doi: 10.1038/ng1726

Wittekoek ME, Pimstone SN, Reymer PWA, et al (1998) A Common Mutation in the Lipoprotein Lipase Gene (N291S) Alters the Lipoprotein Phenotype and Risk for Cardiovascular Disease in Patients With Familial Hypercholesterolemia. Circ 97 :729–735. doi: 10.1161/01.CIR.97.8.729

Wong AK, Park CY, Greene CS, et al (2012) IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. Nucleic Acids Res 40:W484–90. doi: 10.1093/nar/gks458

World Health Organization (2014) Global Status Report On Noncommunicable Diseases 2014.

Xu M, Li X, Wang J-G, et al (2005) Glucose and lipid metabolism in relation to novel polymorphisms in the 5’-AMP-activated protein kinase gamma2 gene in Chinese. Mol Genet Metab 86:372–378. doi: 10.1016/j.ymgme.2005.06.012

Yao J, Yan M, Guan Z, et al (2009) Aurora-A down-regulates IkappaBalpha via Akt activation and interacts with insulin-like growth factor-1 induced phosphatidylinositol 3-kinase pathway for cancer cell survival. Mol Cancer 8:95. doi: 10.1186/1476-4598-8-95

Figure Captions

Fig. 1 Main effect filter analysis - underlying linkage disequilibrium (LD) structure of SNPs within pairwise interactions (P-value < 0.05) associated with HDL cholesterol level. LD diagram was generated using Haploview. Interactions between SNPs are shown with dotted lines. SNPs were mapped to corresponding genes using dbSNP (build 139) and SCANdb. (rs2952101 and rs6641322 on chromosome X are not shown here)

Fig. 2 Functional relationship network generated from Integrated Multi-Species Prediction (IMP) from SNP-SNP interactions associated with HDL-C after main effect filtering (P-value < 0.05). SNPs were mapped to their respective genes and used to query IMP. Nodes in the network represent genes. Orange nodes are the genes that were queried. Edges between nodes represent a functional relationship between two genes. The color of the edge signifies the strength of the relationship confidence. Known relationships are highlighted in gold

Fig. 3 Functional relationship network generated from Integrated Multi-Species Prediction (IMP) from SNP-SNP interactions associated with LDL-C after main effect filtering (P-value < 0.05). SNPs were mapped to their respective genes and used to query IMP. Nodes in the network represent genes. Orange nodes are the genes that were queried. Edges between nodes represent a functional relationship between two genes. The color of the edge signifies the strength of the relationship confidence. Known relationships are highlighted in gold

Fig. 4 Schematic design of study for the QMDR lipid traits association analysis

Supporting Information

Fig. S1 Main effect filter analysis - underlying linkage disequilibrium (LD) structure of SNPs within pairwise interactions (P-value < 0.05) associated with LDL cholesterol level. LD diagram was generated using Haploview

Fig. S2 Main effect filter analysis - underlying linkage disequilibrium (LD) structure of SNPs within pairwise interactions (P-value < 0.05) associated with total cholesterol level. LD diagram was generated using Haploview

Fig. S3 Main effect filter analysis - underlying linkage disequilibrium (LD) structure of SNPs within pairwise interactions (P-value < 0.05) associated with triglyceride level. LD diagram was generated using Haploview

Fig. S4 Biofilter analysis - underlying linkage disequilibrium (LD) structure of SNPs within pairwise interactions (P-value < 0.05) associated with HDL cholesterol level. LD diagram showing r2 values was generated using Haploview

Fig. S5 Biofilter analysis - underlying linkage disequilibrium (LD) structure of SNPs within pairwise interactions (P-value < 0.05) associated with LDL cholesterol level. LD diagram showing r2 values was generated using Haploview

Fig. S6 Biofilter analysis - underlying linkage disequilibrium (LD) structure of SNPs within pairwise interactions (P-value < 0.05) associated with total cholesterol level. LD diagram showing r2 values was generated using Haploview

Fig. S7 Biofilter analysis - underlying linkage disequilibrium (LD) structure of SNPs within pairwise interactions (P-value < 0.05) associated with triglyceride level. LD diagram showing r2 values was generated using Haploview

Table S1 Known biological roles of genes identified within SNP-SNP interactions associated with each lipid trait. Gene information found using GeneCards database (www.genecards.org, Accessed March 28, 2015)

Table S2 Information for cohorts providing individual level data

Table S3 Information of eMERGE cohorts providing individual level data for replication analyses

Table S4 Number of original (non-proxy) and LD-expanded (proxy) SNP-SNP models tested for replication in eMERGE dataset. Numbers are shown for each lipid trait after using both filtering methods

Table S5 Number of LD-expanded (proxy) SNP-SNP models generated for each original discovered SNP-SNP model. Also shown are the number of SNP-SNP models tested for replication in eMERGE dataset per signal. Numbers are shown for each lipid trait after using both main effect and Biofilter filtering methods