4 Lipids

4.1 An introduction to lipids.

4.1.1 Biology and physiology circulating lipids

Lipids are a group of naturally occurring molecules that include fats, sterols, fat-soluble vitamins, triglycerides (TG), phospholipids, and others. The main biological functions of lipids include storing energy, signalling, and acting as structural components of cell membranes. The most familiar type of animal sterol is **cholesterol**, which is vital to animal cell membrane structure and function and a precursor to fat-soluble vitamins and steroid hormones. Cholesterol is transported inside **lipoproteins**. Lipoproteins are named based on their size and density; the lower the density, the larger the particle (Lusis and Pajukanta 2008, Ramasamy 2014). The density of lipoprotein is positively determined by the protein to lipid ratios. In order of increasing density, lipoproteins include chylomicrons, very-low-density lipoprotein (VLDL), LDL, intermediate-density lipoprotein (IDL), and high-density lipoprotein (HDL) (Olson 1998). Lipoproteins contain **apolipoproteins**, which bind to specific receptors on cell membranes and determine the starting and ending points of cholesterol transport. Chylomicrons, the least dense cholesterol transport molecules, carry fats from the intestine to muscle and other tissues in need of fatty acids for energy or fat production. Unused cholesterol remains in cholesterol-rich chylomicron remnants and is taken up to the bloodstream by the liver.

LDL particles are the major blood cholesterol carriers. Its molecule shells contain apolipoprotein B100, which is recognized by LDL receptors in peripheral tissues. The identification of the LDL receptor dramatically improved our understanding of cholesterol metabolism (Brown and Goldstein 1976). Excessive LDL molecules not bound by LDL receptors appear in blood circulation. When oxidized and taken up by macrophages, these LDL molecules become engorged and form foam cells, which often become trapped in the walls of blood vessels to form atherosclerotic plaques. HDL particles transport cholesterol back to the liver for excretion or for other tissues that synthesize hormones, in a process known as reverse cholesterol transport (RCT) (Lewis and Rader 2005). Because of the

function of HDL and LDL particles, the enzymatically measured HDL and LDL levels are often referred to as “good” and “bad” cholesterol, respectively.

TG is an ester derived from glycerol and three fatty acids, and it is the main constituents of vegetable oil (typically more unsaturated) and animal fats (typically more saturated). As a blood lipid, TG enables the bidirectional transference of adipose fat and blood glucose from the liver, playing an important role in metabolism as energy sources and transporters of dietary fat. Lipoprotein lipases on the walls of blood vessels break down TG into free fatty acids and glycerol so that it can pass through cell membranes. Fatty acids can then be taken up by cells via the fatty acid transporter.

4.1.2 Lipids as risk factors for CVD

TC and LDL as CVD risk factors

Large epidemiological studies have established serum level of total cholesterol (TC) especially LDL as major risk factors for CHD (Arsenault et al. 2011). This was later confirmed by MR studies (Cohen et al. 2006) and clinical trials (Shepherd et al. 1995, Downs et al. 1998, Heart Protection Study Collaborative 2002, Badimon et al. 2010). It was estimated that 1 mmol/L reduction in LDL level is associated with a 23% reduction in CHD events (Cholesterol Treatment Trialists et al. 2010), a 12% reduction in all-cause mortality, a 19% reduction in CHD-related mortality (Baigent et al. 2005). The association is log linear with no threshold below which benefit ceases. However, the association of TC or LDL with stroke is not as strong as that with CHD. One study reported that TC was weakly positively related to ischaemic and total stroke mortality in early middle age (40-59 years), and the association could be largely accounted for by the association between TC and blood pressure (Prospective Studies et al. 2007). The weak association with stroke could be due to the fact that stroke is a heterogeneous condition and various causes of ischemic stroke may have different associations with cholesterol (Amarenco et al. 2004, Amarenco and Steg 2007). Nevertheless, randomized trials of statin therapy have shown that reduction of LDL by about 1.5 mmol/L could reduce by about a third the incidence not only of ischemic heart disease but also of ischemic stroke, independently of age, BP or pre-randomization lipid concentrations (Baigent et al. 2005). Statin is the most widely used cholesterol lowering drug, developed

based on the discovery of the fungal metabolite ML-236A and ML-236B (Endo et al. 1976, Kuroda et al. 1979). These lipid modification therapies (LMTs) have revolutionised contemporary approaches to primary and secondary prevention of CVD (Webb et al. 2013).

The understanding that all cholesteryl esters transported by lipoproteins other than HDL (including LDL, VLDL, IDL, and chylomicron remnants) are atherogenic has led to the concept that non-HDL-c levels (TC minus HDL-c) might be more strongly associated with CVD risk than LDL-c alone (Robinson 2009). Several investigators have shown that the ratio between these particles predicts CVD risk better than isolated lipoprotein sub-fractions (Lemieux et al. 2001, Ingelsson et al. 2007, Kannel et al. 2008, Arsenault et al. 2009). The most widely used ratios including TC/HDL, followed by TG/HDL (Castelli 1988). In clinical trials, measuring Apo-B, or Apo-B/Apo-AI ratio also has advantages to assess the efficacy of lipid-lowering therapies.

HDL as CVD risk factors

The FHS first reported that HDL had an inverse association with the incidence of CHD (Gordon et al. 1977). This was later confirmed by other studies (Assmann et al. 1996, Goldbourt et al. 1997). It was estimated that 1 mg/dL increase of HDL is associated with a 1.9 to 2.3% reduction in cardiovascular risk in men and 3.2% in women. This relationship holds even for individuals with low level of LDL (Gordon et al. 1989). The atheroprotective effect of HDL has been mainly attributed to RCT. Over the past few years, other features of HDL have been suggested, including anti-inflammatory, immunomodulatory, antioxidant, antithrombotic, and endothelial cell repair effects (Choi et al. 2006, Ibanez et al. 2007, Badimon et al. 2010).

Although several lifestyle related approaches have demonstrated the ability to increase HDL and improve CVD outcomes (Choi et al. 2006), Mendelian randomization using variants associated with HDL at the *LCAT*, *CETP*, *APOA1*, *ABCA1*, *LIPC*, and *LIPG* loci have largely failed to support a strong causal relationship between HDL and risk of CAD (Frikke-Schmidt et al. 2008, Johannsen et al. 2009, Ridker et al. 2009, Haase et al. 2012, Voight et al. 2012). In clinical trials, Torcetrapib, an inhibitor for cholesteryl ester transfer protein (CETP), showed a significant increase in HDL-c levels but also led to an increase in cardiovascular events and total mortality (Barter et al. 2007, Barter 2009). Small peptides that mimic some of the properties of apolipoprotein A-I (Apo-AI) have been shown to improve HDL function and reduce atherosclerosis without altering overall HDL levels (Navab et al. 2011). It was reasoned that the quality of HDL, rather than the quantity, may influence its

atheroprotective effects. In a more recent clinical trial, a high dose of quinazoline molecule RVX-208 was used to stimulate increased synthesis of endogenous Apo-AI and provided some encouraging results (Nicholls et al. 2011). Detailed proteomic and lipidomic analyses are needed to provide further new insights into the heterogeneous efforts of various HDL compositions. Novel pharmaco-therapeutic strategies directed at HDL include augmenting Apo-AI levels directly and indirectly, mimicking the functionality of Apo-AI, and enhancing steps in the RCT pathways (Degoma and Rader 2011).

TG as CVD risk factor

Serum TG level has been reported for positive association with incidence of CVD (Bansal et al. 2007, Nordestgaard et al. 2007, Sarwar et al. 2007). In 2009, a large meta-analysis based on more than 300,000 individual from 68 long-term prospective studies reported that TG was no longer an independent risk factor for CVD (including non-fatal MI, CHD death, stroke) after adjustment for other risk factors (Emerging Risk Factors et al. 2009). This study indicated that CVD outcomes might be influenced by correlates of TG (such as non-HDL, HDL, or LDL) and TG is a marker instead of a risk factor for CVD. In the same year, another meta-analysis of 31 studies reported a positive association between TG and stroke, with a note for the need for additional large prospective studies especially in stroke subtypes to firmly establish the independent nature of the effect (Labreuche et al. 2009).

There is more evidence for a causal role of TG from MR studies. In 2010, the Triglyceride Coronary Disease Genetics Consortium and Emerging Risk Factors Collaboration first showed a causal association between triglyceride-mediated pathways and coronary heart disease (Triglyceride Coronary Disease Genetics et al. 2010). The instrumental variable used in this study is a single SNP in the promoter of the *APOA5* gene (-1131T>C，rs662799), which directly affects TG metabolism while is only indirectly associated with other lipid parameters including LDL. Another MR study included 185 common variants in a model that accounted for effects on HDL and LDL and also concluded the causal role of TG (Do et al. 2013). A recent WES study for early-onset MI found that carriers of rare non-synonymous mutations in *APOA5* had higher plasma TG and increased risk for MI (Do et al. 2014). Rare mutations that disrupt *APOC3*, a gene in close proximity to and functionally related to *APOA5*, were also associated with a lower level of TG and a reduced risk for CHD (Tg et al. 2014) and ischemic CVD (Jorgensen et al. 2014). These

evidences support that disordered metabolism of TG-rich lipoproteins contributes to CVD risk.

4.1.3 Genetic determinants of lipids levels

Disruptions in the lipoprotein metabolism can cause many different kinds of dyslipidemias depending on the particle or enzyme that is affected. Most of these lipid related syndromes are caused by a mutation in a single gene, i.e., monogenic, and are inherited based on Mendelian laws. There are two major groups of lipid related syndromes: hyperlipidemias and lipoprotein deficiency disorders. Hyperlipidemias are syndromes where lipoprotein levels are elevated in blood and are further classified into different categories (Fredrickson and Lees 1965). It is estimated that genetic and environmental factors have a roughly equal impact on the variation of plasma levels of lipids, with heritability around 50% (Beekman et al. 2002, Pilia et al. 2006, Weiss et al. 2006, Goode et al. 2007). The discovery of genetic factors influencing or even causing lipid level variations is very important for translational medical advances. For example, low-frequency coding variants in *PCSK9* were found to play a causal role in lowering LDL level and protecting against risk of CHD (Abifadel et al. 2003, Allard et al. 2005), which led to the development of a new class of drugs for lowering plasma LDL level (Stein et al. 2012).

Findings from candidate gene and linkage analysis

So far, a total of 26 monogenic genes with causative mutations for dyslipidemia were reported (Kuivenhoven and Hegele 2014) (**Table 4.1**). About half of these were discovered through candidate gene studies with *a priori* knowledge of the protein products. Another ~ 20% of causative gene mutations for monogenic dyslipidemias were found using genetic mapping approaches such as linkage analysis. The availability of patients and families with extreme dyslipidemia is essential in these studies. High throughput approaches including WES have confirmed the role of previously established genes and identified a small number of new causes of monogenic dyslipidemias. Out of 20 loci for genes causing severe changes in lipid metabolism, 16 have also shown association in GWAS, and four of these overlapping loci include genes that are known drug targets (**Figure 4.1**).

Findings from first generation GWAS

Since 2007, a total of 34 GWAS studies have been conducted to discover genetic variations underlying lipids, most of them are based on individuals of European ancestry (**Table 4.2**). The two biggest one are published in 2010 (Teslovich et al. 2010) and in 2013 (Global Lipids Genetics et al. 2013). The former reported 95 loci in total while the latter added 62 more loci with nearly ~200,000 samples, leading to a total of 157 loci. Among the 62 new loci, 32 have some previous connection within lipoprotein metabolism. Among the 157 GWAS loci, 65 show significant associations with two or more of the four main lipid traits, four of which (*CETP, TRIB1, FADS1-2-3, APOA1*) show associations with all lipids traits. However, there is still an overall lack of new knowledge of lipids, given the adequate power of these studies. The phenotypic variation explained by these new GWAS loci is also low, with ~2% of the variation explained by the 62 new loci, which increases the total explained by all GWAS loci to ~15% (Global Lipids Genetics et al. 2013). Nevertheless, further functional studies have begun to emerge and showed promising results. Besides reporting the largest number of novel lipids loci based on statistical significance, the Global Lipids Genetics study also conducted further functional analyses including association with mRNA expression levels and pathway analyses to uncover relationships between lipids loci and those of genes and other functional elements in the genome. The results provided direction for biological and therapeutic research into risk factors for CAD.

Findings from next generation sequencing

Next generation sequencing (on both DNA and RNA) are yielding tremendous successes for discovering novel genes and novel mutations underlying single gene syndromic disorders across a wide range of disease entities and disciplines (Boycott et al. 2013). For lipids, sequencing studies on candidates genes revealed a burden of rare missense or nonsense variants for individuals with low plasma HDL-c levels in the general population (Cohen et al. 2004) and patients with hypertriglyceridemia (Johansen et al. 2010). Next generation sequencing especially WES was first applied to patients with familial dyslipidemia, but has thus far mostly confirmed already known loci instead of finding novel mutations (**Table 4.3**). A recent WES study on 2,005 individuals including 554 with extreme levels of LDL identified significant associations of rare or low frequency variants in known LDL modifying genes such as *PCSK9, LDLR,* and *APOB,* as well as for a novel gene *PNPLA5*. This study

reported that the effect sizes for the burden of rare variants for each associated gene were substantially higher than those observed for individual SNPs identified from GWASs (Lange et al. 2014). Exome chip is a cost-effective alternative to WES. An exome-chip based study with > 200,000 low-frequency and rare coding sequence variants in 56,538 individuals identified new low-frequency variants in four known genes with large effects on HDL-C and/or triglycerides (Peloso et al. 2014). None of these four variants was associated with risk for CHD, suggesting that examples of low-frequency coding variants with robust effects on both lipids and CHD will be limited. Another recent exome-chip based study with ~80,000 coding variants in 5,643 individuals identified a variant that encodes p.Glu167Lys for association with TC and the risk of MI. It is within a locus previously known as *NCAN-CILP2-PBX4* or 19p13 (Holmen et al. 2014).

Based on limited studies reported so far, applying NGS to general healthy population did not yield many novel findings either. Nevertheless, the effect sizes from the burden of rare variants are substantially higher than those from single marker based analysis, therefore supporting a strategy for rare variants aggregation tests. WGS study on lipids was first reported in 2013, with ~1,000 samples with 6X coverage sequencing (Morrison et al. 2013). This study estimated that common and low frequency variation contributes more to heritability of HDL levels (61.8%) than rare variation (7.8%). It also highlighted the value of regulatory and non-protein-coding regions of the genome in addition to protein-coding regions.

Table 4.1 Gene discovery in monogenic dyslipidemias

This table is adopted from (Kuivenhoven and Hegele 2014), listing the single gene causes for the main dyslipidemia states encountered in the clinic, subdivided according to the primary lipid disturbance.

Gene	Discovery	References
Elevated LDL		
ABCG5/G8	Linkage mapping	(Berge et al. 2000)
APOB	A priori knowledge of protein	(Soria et al. 1989)
LDLRAP1	Linkage mapping	(Garcia et al. 2001)
LDLR	A priori knowledge of protein	(Lehrman et al. 1985)
LIPA	WES plus a priori knowledge of protein	(Stitziel et al. 2013)
PCSK9	Linkage analysis	(Abifadel et al. 2009)
Depressed LDL		
ANGPTL3	Mouse studies plus WES	(Musunuru et al. 2010)
APOB	A priori knowledge of protein	(Young et al. 1987)
PCSK9	Linkage analysis plus sequencing	(Cohen et al. 2005)
MTTP	A priori knowledge of protein	(Sharp et al. 1993)
SAR1B	Linkage mapping	(Jones et al. 2003)
MYLIP (IDOL)	In vitro studies (Zelcer et al. 2009)	(Sorrentino et al. 2013)
Elevated HDL		
CETP	A priori knowledge of protein	(Brown et al. 1989)
LIPC	A priori knowledge of protein	(Hegele et al. 1991)
Depressed HDL		
APOA1	A priori knowledge of protein	(von Eckardstein et al. 1989)
LCAT	A priori knowledge of protein	(Funke et al. 1991)
ABCA1	Linkage mapping	(Rust et al. 1999)
Elevated TG		
APOA5	Bioinformatics	(Marcais et al. 2005)
APOC2	A priori knowledge of protein	(Cox et al. 1978)
APOE	A priori knowledge of protein	(Cladaras et al. 1987)
GPD1	Linkage mapping	(Basel-Vanagaite et al. 2012)
GPIHBP1	mutant mouse	(Beigneux et al. 2009)
LMF1	mouse study	(Peterfy et al. 2007)
LPL	A priori knowledge of protein	(Emi et al. 1990)
SLC25A49	Linkage studies plus WES	(Rosenthal et al. 2013)
Depressed TG		
APOC3	GWAS in isolate	(Pollin et al. 2008)

Figure 4.1 Lipids loci overlap between candidate gene studies and GWAS

This figure is modified and updated from (Kathiresan and Srivastava 2012)

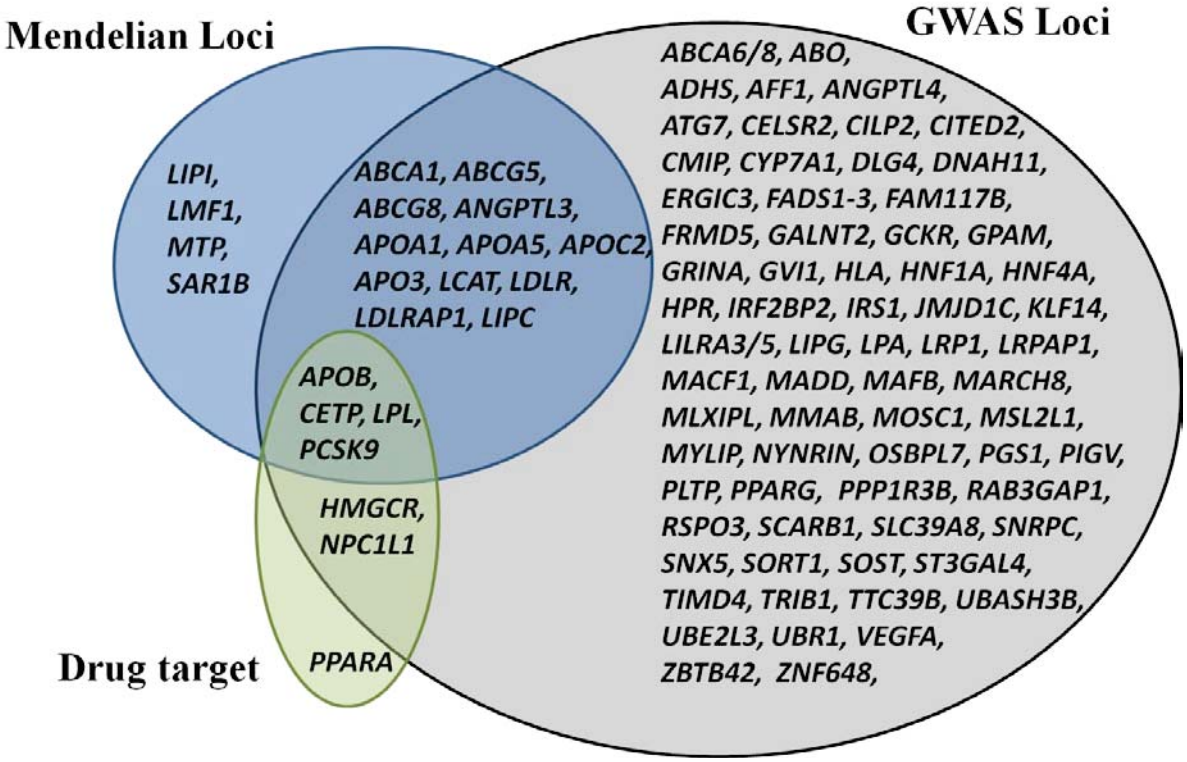


Table 4.2 GWAS studies of lipids

Date is for publication date. Samples are all European ancestry unless explicitly specified otherwise: FIN for Finnish, CHN for Chinese, KOR for Korean, JAP for Japanese, AA for African American, MEX for Mexican, HIS for Hispanics. The sample size before "+" is for discovery while the sample size after "+" is for replication.

Date	Sample size	Main findings	Reference
2007-04	1464 T2D +1467	A locus in *GCKR* with TG	(Saxena et al. 2007)
2007-09	1,087 + ~8,100	No replicated associations	(Kathiresan et al. 2007)
2008-01	1,955 + 2,033	Replicated PSRC1 and CELSR2	(Wallace et al. 2008)
2008-01	8,656+11,437	11 known loci	(Willer et al. 2008)
2008-01	2,758+18,544	6 new loci	(Kathiresan et al. 2008)
2008-01	1,005+6,827	A missense SNP in MLXIPL for TG	(Kooner et al. 2008)
2008-02	11,685+4,979	2 novel variants for LDL	(Sandhu et al. 2008)
2008-09	2,346 Kosrae	3 SNPs in HMGCR for LDL	(Burkhardt et al. 2008)
2008-10	4,274+15,873	CETP and LPL for HDL	(Heid et al. 2008)
2008-10	6,382 + 970	5 novel loci for lipids	(Chasman et al. 2008)
2008-12	19,840+20,623	30 loci including 11 novel	(Kathiresan et al. 2009)
2008-12	4,763 FIN	9 novel loci	(Sabatti et al. 2009)
2008-12	21,848 and 714	6 novel and 16 known for lipids	(Aulchenko et al. 2009)
2008-12	809 + 698 Amish	A null mutation in APOC3	(Pollin et al. 2008)
2009-02	18,245	SNPs at CETP predicts MI risk	(Ridker et al. 2009)
2009-04	900 + 1,810 JAP	variants at CETP for HDL	(Hiura et al. 2009)
2009-11	17,296 + 2700	10 novel loci for lipids	(Chasman et al. 2009)
2010-01	656 + 3,282	2 novel loci	(Igl et al. 2010)
2010-02	8,993 JAP	46 novel loci for blood and lipids traits	(Kamatani et al. 2010)
2010-04	6,078 + 1,231	2 novel loci for lipids	(Ma et al. 2010)
2010-08	100,184	59 novel and 36 known loci	(Teslovich et al. 2010)
2010-09	17,723 + 37,774	4 novel loci for lipids	(Waterworth et al. 2010)
2011-09	12,545+30,395 KOR	10 novel loci for metabolic traits	(Kim et al. 2011)
2011-11	32,225 + 11,509	1 new locus for TC	(Surakka et al. 2011)
2011-12	1,999+1,496 CHN	1 novel locus	(Tan et al. 2012)
2012-01	8,330 FIN	11 novel loci for metabolic traits	(Kettunen et al. 2012)
2012-08	1867 EMR based	A strong protective variant in APOE	(Rasmussen-Torvik et al. 2012)
2012-12	1,720 + 1,261 twins	1 locus related to variability of HDL	(Surakka et al. 2012)
2013-03	2,240 + 2,121 MEX	A novel locus for TG	(Weissglas-Volkov et al. 2013)
2013-05	7,917 AA, 3,506 HIS	striking similarities across populations	(Coram et al. 2013)
2013-09	1,782 + 1,719 FIL	2 known loci: APOE, APOA5	(Wu et al. 2013)
2013-09	839+5,248 Sorbs	1 novel locus	(Keller et al. 2013)
2013-10	94,595 + 93,982	62 novel and 95 known loci	(Willer et al. 2013)
2013-12	3,451 + 8,830 CHN	Replicated 8 known loci	(Zhou et al. 2013)

Table 4.3 NGS studies on lipids

There are five small scale sequencing studies on patients with familial dyslipidemia and three studies on healthy populations with relatively large sample size. WES, WGS, and exome-chip technologies were used for each of the three studies on healthy population. Samples are all European ancestry unless explicitly specified otherwise.

Date	Sample size	Main findings	Reference
Familial dyslipidemia			
2010-10	WES on 2	ANGPTL3 mutations for familial combined hypolipidemia	(Musunuru et al. 2010)
2010-11	WGS of 1	two nonsense mutations in ABCG5 caused sitosterolemia	(Rios et al. 2010)
2012-03	WES on 1 family	novel APOB mutation for ADH	(Motazacker et al. 2012)
2012-10	WES on 14	heterozygous in-frame deletion in the APOE gene for ADH	(Marduel et al. 2013)
2013-09	WES on 3	a homozygous splicing mutation in LIPA for hypercholesterolemia	(Stitziel et al. 2013)
Healthy population			
2013-06	WGS of 962	HDL Heritability mainly explained by common variants	(Morrison et al. 2013)
2014-01	WES of 2,005	LDL and the burden of rare variants in PNPLA5	(Lange et al. 2014)
2014-03	X-chip of 5,771	causal variant in *TM6SF2* influencing TC and MI	(Holmen et al. 2014)

.

4.1.4 Aims of this study

Under the framework of the UK10K project (The UK10K Consortium 2015), this study aims to identify novel genetic variants that are associated with plasma lipids levels and also fine map known lipids loci with WGS data. The current study is by far the largest WGS based association study of lipids, with up to 3,210 WGS samples and more than 22,000 samples with WGS imputed data. I first analyse the WGS samples aiming to discover rare and low frequency variants with large effect sizes. Then I analyse a much larger group of cohorts with imputed data to discover novel associations across the full MAF spectrum. Besides single marker based genome-wide scan, this study is able to fine map known loci and investigate the association and contribution of rare variants to serum lipids variance. This work will not only contribute to the understanding of the allelic architecture of lipid variation in healthy population but also provide a good reference for using WGS data to study complex traits in general.

4.2 Methods

4.2.1 Cohorts & phenotype measurements

There were a total of 14 cohorts included for the expanded discovery, including both WGS and the SNP-array imputed samples for TwinsUK and ALSPAC, plus 10 other cohorts where genome-wide SNP data and raw lipids phenotypes were made available (**Table 4.4**). There were 11 more cohorts included for stage-1 replication. Some of them had genome-wide results as well, but only the top hits from the expanded discovery were queried from the replicate data. For the final few replicated variants, I used the WHI data for a further replication. The details of these cohorts were given in chapter 2.

Lipids measurement methods were as following: for **ALSPAC**, plasma levels of TC, HDL and TG were measured with enzymatic colorimetric assays (Roche) on a Hitachi Modular P Analyser. LDL was derived from the following formula: TC- (HDL+TG/2.19); for **TwinsUK,** Enzymatic colorimetric assays were used to measure serum levels of TC, HDL and TG were measured using three analysing devices (Cobas Fara; Roche Diagnostics, Lewes, UK; Kodak Ektachem dry chemistry analysers (Johnson and Johnson Vitros Ektachem machine, Beckman LX20 analysers, Roche P800 modular system)); for **1958BC**, serum TG, TC and HDL were measured in serum by Olympus model AU640 autoanalyser in a central lab in Newcastle. Enzymatic colorimetric determination GPO-PAP method was used to determine TG, CHOD-PAP method for TC and for HDL; for **INGI-VB**, lipids were measured using HITACHI 917 ROCHE and Unicel Dx-C 800 BECKMAN devices; for **INGI-FVG** and **INGI-Carl**, lipids were measured using BIOTECNICA BT-3000 TARGA chemistry analyser; for **INCIPE**, enzymatic determination of TC and TG was performed on Dimension RxL apparatus (Siemens Diagnostics). HDL cholesterol was determined by the homogeneous method; LDL cholesterol by the Friedewald formula (Friedewald et al. 1972); for **LURIC**, TC and TG were obtained by ß-quantification from serum and measured enzymatically using WAKO reagents on a WAKO 30R analyser (Neuss, Germany). LDL and HDL were measured after separating lipoproteins with a combined ultracentrifugation-precipitation method; for **HELIC Manolis** and **HELIC Pomak** and **Teenage**, TC, HDL, TG were assessed using enzymatic colorimetric assays and while LDL levels were calculated according to Friedewald equation (Friedewald et al. 1972). For WHI, HDL, LDL, and TG

measurements were performed at the University of Minnesota by standard biochemical methods on the Roche Modular P Chemistry analyzer (Roche Diagnostics): HDL was measured in serum by the HDL-C plus third generation direct method; TG was measured in serum by Triglyceride GB reagent, and total cholesterol (TC) was measured in serum by a cholesterol oxidase method. LDL was calculated in serum specimens having a TG value < 400 mg/dl according to the formula of Friedewald et al. [Based on the LDL-lowering effects of statins, we estimated the pretreatment LDL value for individuals on lipid-lowering medication by dividing treated LDL values by 0.75.

For phenotype harmonization, extra care was given to the TwinsUK cohorts given there was random efforts of different dates of visits and different instrumental measurements (**Table 4.5)**. For ALSPAC and other cohorts in expanded discovery and replication, the same phenotype protocol was used. Inverse normal transformation was applied to all cohorts. For each cohort, the residuals with confounding variables regressed out were standardized so that the phenotype had a mean of 0 and a standard deviation of 1.

Table 4.4 Characteristics of participating cohorts

All cohorts are population based, except for TwinsUK. Imputation was conducted with the 1000G and UK10K combined reference panel, unless otherwise specified. Age is in mean (range). Traits values are in the format of mean (SD). For each trait of each cohort, the residuals with confounding variables regressed out were standardized so that the phenotype has a mean of 0 and a standard deviation of 1.

	Study	N	Country	Age	% Female	HDL	LDL	TG	TC
discovery	ALSPAC WGS	1,497	UK	10 (9-11)	50.3	1.40 (0.01)	2.31 (0.01)	1.14 (0.01)	4.24 (0.02)
	TwinsUK WGS	1,713	UK	56 (17-85)	100.0	1.79 (0.01)	3.16 (0.02)	1.12 (0.01)	5.48 (0.03)
	ALSPAC GWA	2,820	UK	10 (9-12)	49.2	1.40 (0.01)	2.36 (0.01)	1.14 (0.01)	4.28 (0.01)
	TwinsUK GWA	1,896	UK	50 (16-83)	81.1	1.51 (0.01)	3.33 (0.03)	1.18 (0.02)	5.38 (0.03)
	1958 BC	5,493	UK	44 (44-44)	52.3	1.56 (0.01)	3.42 (0.01)	2.07 (0.02)	5.88 (0.01)
	INGI-Carl	413	Italy	50 (18-83)	60.0	--	--	1.48 (0.04)	5.30 (0.06)
	INGI-FVG	1,394	Italy	52 (18-92)	58.2	1.38 (0.01)	3.71 (0.03)	1.30 (0.02)	5.69 (0.03)
	INGI-VB	1,776	Italy	55 (18-102)	56.3	1.52 (0.01)	3.23 (0.02)	1.19 (0.02)	5.3 (0.03)
	INCIPE1	653	Italy	60 (35-89)	54.4	1.49 (0.01)	3.49 (0.03)	1.18 (0.03)	5.52 (0.04)
	INCIPE2	1,382	Italy	58 (26-95)	50.9	1.49 (0.01)	3.39 (0.02)	1.10 (0.02)	5.39 (0.03)
	LURIC-Ctrl	983	Germany	61 (17-91)	60.8	1.07 (0.01)	3.21 (0.03)	1.82 (0.04)	5.22 (0.03)
	HELIC MANOLIS	1,264	Greece	62 (18-99)	57.2	1.32 (0.01)	3.22 (0.03)	1.56 (0.03)	5.57 (0.08)
	HELIC POMAK	999	Greece	43 (13-87)	72.1	1.15 (0.01)	3.15 (0.03)	1.52 (0.03)	5.01 (0.03)
	TEENAGE	557	Greece	13 (11-18)	55.9	1.44 (0.01)	2.33 (0.02)	0.67 (0.01)	4.09 (0.03)
replication	LOLI-EW610	905	UK	56 (35-75)	26.8	1.42 (0.01)	3.46 (0.03)	1.54 (0.04)	5.57 (0.03)
	LOLI-EWA	566	UK	55 (23-75)	13.1	1.30 (0.01)	3.16 (0.04)	1.70 (0.05)	5.21 (0.05)
	LOLI-EWP	610	UK	56 (32-67)	0.0	1.26 (0.01)	3.06 (0.04)	1.83 (0.06)	5.13 (0.04)
	RS-1	2981	NL	69 (48-75)	41.2	1.06 (0.01)	3.21 (0.04)	1.262 (0.06)	6.06 (0.04)
	RS-2	1823	NL	67 (51-75)	47.7	1.29 (0.01)	3.22 (0.03)	1.23 (0.03)	6.12 (0.04)
	GoT2D	2076	UK	NA	NA	NA	NA	NA	NA
	InChianti	621	Italy	56 (47-71)	56.3	1.53 (0.01)	3.36 (0.03)	1.28 (0.02)	4.99 (0.03)
	FinRisk	817	Finland	56 (47-68)	46.8	1.4 (0.03)	3.11 (0.05)	1.68 (0.04)	5.78 (0.05)
	Fenland	8701	UK	65 (47-77)	46.2	1.43 (0.01)	3.21 (0.02)	1.65 (0.02)	5.12 (0.03)
	UCLEB-BRHS	2742	UK	69 (58-81)	0.0	1.15 (0.01)	3.89 (0.02)	2.05 (0.03)	6.36 (0.02)
	UCLEB-BWHHS	3309	UK	71 (60-81)	100.0	1.62 (0.01)	4.14 (0.03)	1.91 (0.02)	6.62 (0.03)
	WHI	10,999	US	51 (44-69)	100.0	1.36 (0.02)	3.11 (0.04)	1.93 (0.06)	5.27 (0.05)

Table 4.5 Phenotype harmonization protocol for lipids traits

Analyser was tested as a random effect variable, while the others including age and age^2 are tested as fixed effect covariates.

Dataset	Trait	Transformation	Gender stratified	Co-variates tested	Filter	Analyser
ALSPAC WGS+GWA	HDL	inverse normal	yes	age, age^2	5 SD	--
TwinsUK GWA	HDL	inverse normal	yes	age,age^2,analyser	4 SD	yes
TwinsUK WGS	HDL	inverse normal	--	age, age^2	5 SD	yes
ALSPAC WGS+GWA	LDL	inverse normal	yes	age, age^2	5 SD	--
TwinsUK GWA	LDL	inverse normal	yes	age,age^2,analyser	4 SD	yes
TwinsUK WGS	LDL	inverse normal	--	age, age^2	5 SD	yes
ALSPAC WGS+GWA	TC	inverse normal	yes	age, age^2	5 SD	--
TwinsUK GWA	TC	inverse normal	yes	age,age^2,analyser	4 SD	yes
TwinsUK WGS	TC	inverse normal	--	age, age^2	5 SD	yes
ALSPAC WGS+GWA	TG	inverse normal	yes	age, age^2	5 SD	--
TwinsUK GWA	TG	inverse normal	yes	age,age^2,analyser	4 SD	yes
TwinsUK WGS	TG	inverse normal	--	age, age^2	5 SD	yes

4.2.2 Single marker based discovery and follow-up

For single marker tests, I first fitted linear models on standardised trait residuals to test associations of allele dosages with 13,074,236 SNVs and 1,122,542 biallelic InDels (MAF≥0.1%) in the two WGS samples (TwinsUK and ALSPAC), using SNPTEST. Then I run the same analysis for 12 more cohorts with imputed data to identify novel variants across the allele frequency spectrum with a much larger sample size and increased power. Among the 12 additional cohorts, SNPTEST was used for population based samples while GEMMA was used for genetic isolates and cohorts with family structure. Meta-analyses were performed using GWAMA v2.1 (Magi and Morris 2010), assuming a fixed effect model adjusted genomic control to the summary statistics for both input and output data. Meta-analysis was first run for two WGS cohorts, to generate the WGS only based "2-way" results. Meta-analyses were then run for all 14 cohorts with genome-wide association results, leading to "14-way" results as an expanded discovery. Given the poor imputation quality and weak statistical power for rare variants, I chose to exclude the variants that did not pass a low allele frequency threshold (MAF<0.1%). For imputed cohorts, the variants with INFO score <0.4 were also excluded.

Given a large number of lipids loci already reported by previous GWAS with much larger sample size than this study, a rigorous loci selection was conducted to select putative novel loci that are statistically truly novel. The core of this loci selection process was a step-wise conditional analysis as described in chapter 2. Initially, GWAS Catalog and literature review were used to identify known variants. For those variants that survived the conditional tests, they were further checked against the full genome-wide results of the two largest GWAS (Teslovich et al. 2010, Global Lipids Genetics et al. 2013) (available at http://csg.sph.umich.edu/locuszoom/) to ensure their true novelty. As described in chapter 2, I excluded those variants that did not survive the step-wide conditional analyses or those having modest to high LD (r2>0.1) with known variants. For putative novel variants discovered from above, I conducted meta-analysis for replication cohorts and further performed a joint meta-analysis that calculated the statistics of all discovery and replication cohorts combined together.

4.2.3 Rare variant aggregation based discovery and follow-up

I first evaluated the associations of rare variants by considering genes as functional units of analysis. I applied two separate statistical models with different properties to rare variants (MAF<1%): SKAT and burden tests, both implemented in a unified software SKAT-O. As described in chapter 2, in *naïve* tests, all variants in exons, untranslated regions (UTRs) and essential splice sites were considered, and were given equal weight of being causal (50,214 windows for 35,709 genes, mean=35 variants, median=38 variants per window). In functional tests, only loss of function (LoF) and predicted functional variants were included (15,528 gene windows with ≥ 5 variants, mean=18, median=14 variants per gene). Finally, I run the locus-based analysis genome-wide in an agonistic fashion, by constructing ~1.8 million windows of 3 kb each, overlapping by half (median 35 SNVs/window, MAF<1%), assigning an equal weight to all variants.

For replication of locus based top hits, we used rareMetal (Feng et al. 2014) to reconstruct gene-level test statistics from single marker score statistics (Liu et al. 2014). The single maker score statistics were calculated with the Cochran-Mantel-Haenszel method. RareMetal works for meta-analysis of results from burden tests as well as SKAT tests. The windows with *P*<1E-5 in GW and *P* <1E-4 for EW based were taken forward for replication. Replications were conducted in three cohorts: GoT2D, FinRisk, InChianti. Finally, for those replicated loci, I explored a “drop-one” approach to determine whether the aggregation association was mainly driven by a single contributing variant. This worked by sequentially dropping one variant at a time and re-run SKAT-O for the same region with the same parameters. A variant was found to be contributing to the SKAT signal when dropping it causes a significant change of the SKAT-O P, usually from significant to non-significant. When more than one variant were found to be contributing, LD patterns were examined to evaluate the independence of those variants. In cases where a single variant with main effect could explain the association, usually the single marker was not sufficiently powered to detect an association in the same region.

4.2.4 Fine-mapping of known loci

For lipids, there were a total of 157 known loci reported. Many of those loci were significant in multiple lipids traits. I identified a total of 282 trait-specific regions for carrying out fine-mapping analysis to assess the probability of each variant being causal given other variants in the region. Within each signal I included SNPs in high LD (defined as all variants having $r^2 \geq 0.8$ with the most associated variants in the region), apart for *APOE* where an extended analysis interval was considered. As described in chapter 2, for each lipids trait I first created a list of fine-mapping regions based on HapMap estimates of recombination rates. I then analysed each region separately for each of the 14 participating cohort using Bayesian linear additive models, by accounting for covariates as in the general single point association analyses. At the end, the resulting BFs for each variant were multiplied to obtain a joint BF measure of association, with the assumption that each cohort is independent. These BFs were then used to calculate posterior probabilities, based on the assumption that there was exactly one causal SNP in each region. In addition, 95% and 99% credible sets were constructed in order to assess the uncertainty of the fine-mapping analysis.

The fine-mapped variants were further overlapped with four liver-essential TFBS data(Ballester et al. 2014). In brief, the genome-wide occupancy of four transcription factors (HNF4A, CEBPA, ONECUT1, and FOXA1) was determined in primary liver in five species (*Homo sapiens, Macaca mulatta, Canis familiaris, Mus musculus,* and *Rattus norvegicus*) using chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq). After mapping and peak calling, the regions of the genomes with the various combinations of the transcription factor binding events were analysed to determine the extent that binding events are shared across species and the characteristics of the shared and non-shared binding sites.

4.4 Results

4.4.1 Novel loci and novel variants from single marker analysis

WGS for low frequency and rare variants

The assessment of associations based on imputation or WES has been incomplete. I thus sought to investigate if additional low-frequency or rare variants with strong effects could be detected from the WGS dataset. I first tested association results using solely the WGS dataset in order to identify whether these variants existed. Associations were carried out in 13,074,236 SNVs and 1,122,542 biallelic InDels (MAF≥0.1%) using linear regression and data from the two WGS cohorts was meta-analysed.

Based on the meta-analysis of two UK10K WGS cohorts, there were a total of 267 trait-specific associations reaching the generally used genome-wide significance *P*<5.0E-08. All but two of these associations were previously reported, mapped to five known loci (*PCSK9, CELSR2, SID2, CETP, APOE)* (**Figure 4.2**). The first putative novel association is rs1505058, an intergenic variants on chromosome 5, for association with HDL (MAF=0.1%, beta=2.26, *P*=2.9E-09). The second putative novel association is rs185450930, an intronic variant within *SEMA3A* on chromosome 7, for association with TG (MAF=0.1%, beta=2.92, *P*=2.3E-08).

To look at suggestive associations, I used a less stringent threshold and discovered 117 more variants (a total of 384) having *P*<1E-6. Among all 384 variants, 90 variants have MAF between 0.1% and 5% and 22 are independent of known variants, i.e., either having no positive controls within 1Mb or surviving the conditional analysis and LD pruning with known variants within 1Mb. This list of 22 variants included the two variants with *P*<5E-08 described above, and are considered putative novel variants based on the two WGS cohorts. One de-novo genotyped cohort (Fenland) and three external WGS cohorts included in the expanded discovery (GoT2D, InChianti, FinRisk) were used as replication datasets for these 22 putative novel variants based on UK10K WGS, although not all these four cohorts have association results for these 22 variants. Their association summary statistics and replication results for these 22 variants were given in **Table 4.6**. The replication results for each of the four individual cohorts were given in **Table 4.7**. Based on the limited replication, only one variant within *LDLR* (rs72658867, EAF=1.2% (A), beta=-0.584) was replicated with a

consistent and comparable effect size (beta=-0.471, *P*=4.8E-12). Of note, the rare splice variant (rs138326449) in the *APOC3* gene was recently reported by us and others as associated with TG and coronary artery disease risk (Timpson et al. , Jorgensen et al. 2014, The TG and HDL Working Group of the Exome Sequencing Project 2014), therefore, it is viewed as a positive control instead of a novel locus.

Give the low power of single marker based replication for variants with low to rare frequency, the rare variants based tests (implemented in SKAT-O) were conducted for the 21 windows that include 21 variants except the variant on chromosome X (**Table 4.8).** Ten windows have SKAT-O *P* <2.3E-3 (i.e., 0.05/22), much more than expected. For all these 21 windows, the SKAT-O *P* is not much more significant than SKAT *P*, indicating that the signals are mainly driven by SKAT test instead of burden test. Indeed, for each of those five windows with SKAT *P*<1E-5, the SKAT signal was found to be driven by a single variant through a drop-one SKAT-O analysis.

Figure 4.2. Single point association results of lipids on WGS samples

X-axis is for chromosome and positions (build 37). Y-axis is for –log10(*P*). Variants passing threshold of 5E-08 and 1E-06 are shown in red and blue, respectively. For those passing threshold of 5E-08, known loci were marked in green text while putative novel loci were marked in red text.

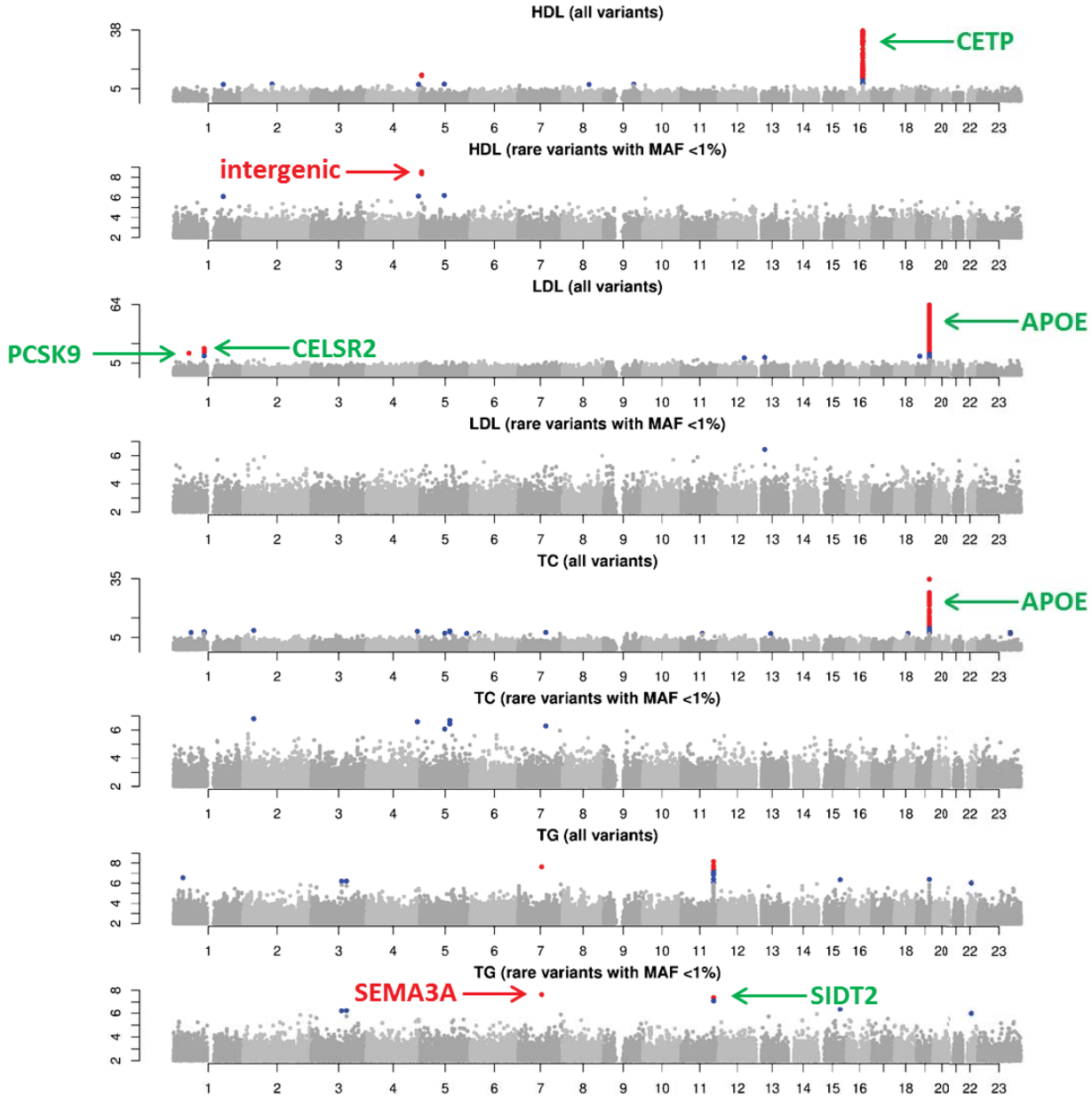


Table 4.6 Putative novel variants of low or rare frequency from UK10K WGS

WGS variants (P<1E-6) either have no positive controls within 1Mb or are independently significant from known variants.

Six have low frequency (MAF between 1-5%) and could be imputed with fair accuracy.

					UK10K WGS						Replication (WGS, de novo)			
Low frequency (6 variants)														
trai	**rsID**	**CHR**	**POS**	**Gene**	**EA**	**NEA**	**EAF**	**beta**	**SE**	**P**	**Beta**	**SE**	**P**	**N**
HD	rs72831743	2	103,690,744	Intergenic	A	C	0.012	-0.572	0.115	6.4E-07	-	0.05	7.8E-01	12161
TC	rs139029427	7	98,664,474	*SMURF1*	C	T	0.010	-0.643	0.128	5.1E-07	-	0.06	3.3E-01	12233
HD	rs150103869	8	94,349,833	*LINC00535*	A	C	0.010	0.632	0.128	8.1E-07	-	0.06	8.9E-03	12159
LD	rs77198522	12	91,641,075	Intergenic	T	C	0.030	-0.384	0.076	4.8E-07	0.004	0.03	9.1E-01	11948
LD	rs72658867	19	11,231,203	*LDLR*	A	G	0.012	-0.584	0.112	1.7E-07	-	0.06	4.8E-12	12215
TC	chrX:117293318	X	117,293,318	Intergenic	G	GGA	0.013	-0.869	0.172	4.6E-07	-	0.35	1.3E-02	614
Rare (16 variants)														
HD	rs184490209	1	178,071,554	*RASAL2*	A	G	0.001	1.958	0.396	8.0E-07	0.266	0.27	3.3E-01	2750
TC	chr2:37882057	2	37,882,057	*CDC42EP3*	GA	G	0.007	-0.752	0.143	1.6E-07	0.206	0.30	5.0E-01	2247
TC	rs143755400	2	37,883,627	*CDC42EP3*	A	G	0.007	-0.747	0.142	1.6E-07	-	0.09	8.0E-01	11618
TG	rs147039106	3	108,844,173	*MORC1*	C	T	0.007	0.799	0.160	6.2E-07	0.008	0.05	8.7E-01	12332
TG	chr3:126360068	3	126,360,068	*TXNRD3*	C	T	0.003	-1.138	0.228	6.0E-07	-	0.13	1.6E-01	12438
TC	chr4:182413170	4	182,413,170	*RP11-433O3.1*	A	G	0.001	-1.803	0.350	2.6E-07	0.001	0.17	9.9E-01	10818
HD	chr4:186058963	4	186,058,963	*SLC25A4*	G	T	0.004	-1.044	0.210	7.4E-07	-	0.11	2.4E-01	11758
HD	rs1505058	5	6,558,466	Intergenic	C	A	0.001	2.258	0.379	2.9E-09	-	0.20	7.3E-01	8776
HD	chr5:87396789	5	87,396,789	Intergenic	T	C	0.001	-1.887	0.378	6.5E-07	0.039	0.19	8.4E-01	10878
TC	rs183893710	5	88,977,348	Intergenic	G	C	0.005	-0.872	0.177	8.6E-07	-	0.07	6.3E-01	12467
TC	chr5:107200309	5	107,200,309	*FBXL17*	T	C	0.001	-2.571	0.495	2.1E-07	--	--	--	--
TG	rs185450930	7	83,755,035	*SEMA3A*	A	G	0.001	2.923	0.523	2.3E-08	--	--	--	--
TG	chr11:117053959	11	117,053,959	*SIDT2*	A	G	0.003	-1.359	0.248	4.2E-08	--	--	--	--
LD	chr13:31087680	13	31,087,680	*HMGB1*	C	T	0.002	-1.378	0.271	3.7E-07	-	0.19	9.9E-01	10755
TG	chr15:78513033	15	78,513,033	*ACSBG1*	T	C	0.001	-2.851	0.564	4.4E-07	--	--	--	--
TG	rs191808700	22	30,633,306	*LIF*	G	A	0.001	2.458	0.500	9.0E-07	--	--	--	--

* chr11:117053959 is in close proximity with the *APOC3* variants rs138326449 (chr11:116701354), with modest LD (r2 = 0.644), and is not independent significant based on conditional analysis.

Table 4.7 Replication results of WGS top hits

GoT2D, InChianti, and FinRisk used WGS, while Fenland used de-novo genotyping. For each set of results, the effect allele frequency (EAF), beta, standard deviation (SE), *P* value, and the total sample size were presented. Records with $P < 0.05$ are highlighted in red text.

		GoT2D					InChianti					FinRisk					Fenland				
trait	**rsID**	**EAF**	**Beta**	**SE**	**P**	**N**	**EAF**	**Beta**	**SE**	**P**	**N**	**EAF**	**beta**	**SE**	**P**	**N**	**EAF**	**beta**	**SE**	**P**	**N**
HDL	rs72831743	0.009	0.145	0.159	3.6E-01	2129	0.006	-0.197	0.380	6.0E-01	621	0.006	0.298	0.282	2.9E-01	856	0.012	-0.037	0.084	6.6E-01	5760
TC	rs139029427	0.010	-0.069	0.156	6.6E-01	2247	0.011	-0.377	0.280	1.8E-01	614	0.012	0.032	0.210	8.8E-01	856	0.010	-0.042	0.094	6.5E-01	5729
HDL	rs150103869	0.006	-0.396	0.214	6.3E-02	2129	0.010	-0.370	0.280	1.9E-01	621	0.003	-0.011	0.417	9.8E-01	856	0.012	-0.057	0.084	5.0E-01	5764
LDL	rs77198522	0.070	0.041	0.062	5.1E-01	2076	0.022	-0.259	0.197	1.9E-01	621	0.084	0.064	0.086	4.6E-01	817	0.035	0.021	0.052	6.9E-01	5653
LDL	rs72658867	0.006	-0.426	0.203	3.5E-02	2076	0.013	-0.579	0.252	2.2E-02	621	0.001	-0.024	0.697	9.7E-01	817	0.010	-0.473	0.076	4.9E-10	8701
TC	chrX:117293318	--	--	--	--	--	0.007	-0.882	0.354	1.3E-02	614	--	--	--	--	--	--	--	--	--	--
HDL	rs184490209	0.002	-0.126	0.536	8.1E-01	2129	0.008	0.404	0.318	2.0E-01	621	--	--	--	--	--	--	--	--	--	--
TC	chr2:37882057	0.003	0.206	0.303	5.0E-01	2247	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
TC	rs143755400	0.002	0.223	0.322	4.9E-01	2247	0.005	0.305	0.410	4.6E-01	614	--	--	--	--	--	0.006	-0.066	0.100	5.1E-01	8757
TG	rs147039106	0.033	-0.021	0.089	8.2E-01	2190	0.002	-0.013	0.708	9.9E-01	614	0.047	0.116	0.110	2.9E-01	856	0.011	-0.020	0.074	7.9E-01	8672
TG	chr3:126360068	0.002	-0.335	0.415	4.2E-01	2190	0.002	-0.466	0.578	4.2E-01	614	0.002	0.018	0.545	9.7E-01	856	0.003	-0.165	0.149	2.7E-01	8778
TC	chr4:182413170	0.001	0.759	0.470	1.1E-01	2247	--	--	--	--	--	--	--	--	--	--	0.002	-0.122	0.190	5.2E-01	8571
HDL	chr4:186058963	0.002	-0.093	0.406	8.2E-01	2129						0.001	-0.126	0.658	8.5E-01	856	0.004	-0.139	0.122	2.5E-01	8773
HDL	rs1505058	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	0.001	-0.072	0.209	7.3E-01	8776
HDL	chr5:87396789	0.001	-1.660	0.619	7.4E-03	2129	--	--	--	--	--	--	--	--	--	--	0.001	0.232	0.209	2.7E-01	8749
TC	rs183893710	0.003	0.184	0.283	5.2E-01	2247	0.008	0.006	0.319	9.8E-01	614	0.003	0.027	0.437	9.5E-01	856	0.009	-0.057	0.079	4.7E-01	8750
TC	chr5:107200309	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
TG	rs185450930	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
TG	chr11:117053959	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
LDL	chr13:31087680	0.001	-0.286	0.576	6.2E-01	2076	--	--	--	--	--	--	--	--	--	--	0.001	0.035	0.209	8.7E-01	8679
TG	chr15:78513033	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
TG	rs191808700	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Table 4.8 SKAT results for single point test top hits

For each of the 22 top hits based on WGS single marker analysis, the selected SKAT-O window included the index variant. For genome-wide SKAT-O analysis with overlapping windows, when there are two windows include a variant, the one with the lower *P* value is listed. For SKAT-O test, *P* <2.3E-3 (i.e., 0.05/22) are shown in red.

trait	rsID	GW SKAT region	GW SKAT	GW SKATO	EW SKAT Region	EW SKAT	EW SKATO
HDL	rs72831743	chr2:103690501-103693500	2.66E-01	4.11E-01	--	--	--
TC	rs139029427	chr7:98664001-98667000	1.11E-06	2.70E-06	SMURF1.w3	8.25E-01	1
HDL	rs150103869	chr8:94348501-94351500	7.23E-01	5.06E-01	--	--	--
LDL	rs77198522	chr12:91641001-91644000	2.53E-01	4.02E-01	--	--	--
LDL	rs72658867	chr19:11230501-11233500	4.47E-02	3.04E-02	LDLR.w3	2.51E-01	3.75E-01
TC	chrX:117293316	--	--	--	--	--	--
HDL	rs184490209	chr1:178071001-178074000	5.32E-03	9.23E-03	RASAL2.w1	7.06E-01	8.66E-01
TC	chr2:37882057	chr2:37881001-37884000	4.01E-06	9.74E-06	CDC42EP3.w4	2.51E-02	7.73E-03
TC	rs143755400	chr2:37882501-37885500	6.53E-04	1.30E-03	CDC42EP3.w4	2.51E-02	7.73E-03
TG	rs147039106	chr3:108843001-108846000	9.09E-07	2.69E-06	--	--	--
TG	chr3:126360068	chr3:126360001-126363000	1.21E-06	2.90E-06	TXNRD3.w3	4.58E-01	6.46E-01
TC	chr4:182413170	chr4:182412001-182415000	1.18E-04	2.57E-04	--	--	--
HDL	chr4:186058963	chr4:186058501-186061500	2.12E-03	4.56E-03	--	--	--
HDL	rs1505058	chr5:6558001-6561000	3.77E-05	7.67E-05	--	--	--
HDL	chr5:87396789	chr5:87396001-87399000	7.51E-02	1.27E-01	--	--	--
TC	rs183893710	chr5:88977001-88980000	8.74E-07	2.46E-06	--	--	--
TC	chr5:107200309	chr5:107199001-107202000	6.36E-02	1.12E-01	FBXL17.w2	3.95E-01	1.38E-01
TG	rs185450930	chr7:83754001-83757000	5.13E-02	9.32E-02	SEMA3A.w3	2.01E-01	3.24E-01
TG	chr11:117053959	chr11:117052501-117055500	1.66E-03	3.58E-03	SIDT2.w2	6.64E-01	8.62E-01
LDL	chr13:31087680	chr13:31087501-31090500	2.77E-04	5.72E-04	HMGB1.w3	8.59E-01	2.04E-01
TG	chr15:78513033	chr15:78513001-78516000	8.70E-03	1.67E-02	ACSBG1.w4	8.30E-01	3.89E-01
TG	rs191808700	chr22:30633001-30636000	2.33E-03	4.25E-03	--	--	--

Meta-analysis for identifying novel variants of all allele spectrums

Given the enhanced imputation quality with the UK10K WGS reference panel as demonstrated in chapter 3, I included 12 more cohorts with imputed data for an expanded discovery, to increase power for discover variants across all allele frequency spectrum. As mentioned earlier in the methods section, variants with MAF <0.1% or imputation INFO <0.4 were not included. This effort yielded 5,306 variants with *P* <1E-07, 5,023 of which reached genome-wide significant threshold (*P* <5E-08) (**Figure 4.3**). I carried out step-wise conditional analysis to identify putative novel associations, as described in chapter 2 and the methods section of this chapter. All but four associations did not survive the novelty test, i.e, either association singles going away after conditional on known variants or in modest to high LD with known variants (r2>0.1). Two of these associations don't have positive controls within 1Mb. For the other two with position controls within 1Mb, chr16: 66926255 is conditioned on the four known variants (chr16:67708897, chr16:67902070, chr16:68013471, chr16:68024995) and its conditional *P* is 1.2E-07; rs72658867 is conditioned on four known variants (chr19:11195030, chr19:11202306, chr19:11224265, chr19:11227602) and its conditional *P* is 6.2E-10.

The four putative novel variants were taken forward in two rounds of replications that included genotypes from WGS, imputation and *de novo* genotyping. The association results including discovery and two rounds of replications for these four variants were reported in **Table 4.9**. The cohort specific results for these four variants were given in **Table 4.10**. The first variant is a common variant (MAF of 16.5%, rs57367316) on chromosome 2, for association with TG. It did not survive the first round of replication. Its best proxy rs4404266 (chr2:107712732, 12,462bp apart, r^2=0.63) has *P*=0.91 in the Global lipids study (Global Lipids Genetics et al. 2013). As shown in **Table 4.10**, this variant is only marginally significant in one replication cohort (FinRisk, *P*=0.046) but with an opposite effect size. Therefore, this variant is most likely to be false positive. The second variant chr16: 66926255 has an overall MAF of 0.003 and *P*=6.9E-08. However, this variant did not show evidence for replication either. Upon further inspection, the signal in the expanded discovery was mostly driven by a single cohort (HELIC-Manolis, beta (SE) = 1.491(0.236), *P*=9.7E-10), a genetic isolate of Greek origin, where its MAF is much higher (0.009) than the remaining cohorts. Failure to replicate this variant may be due to either a false positive in the Greek discovery cohort, or insufficient power in the non-isolate cohorts where the variant has low MAF. The third novel association detected was with variant rs72658867 within *LDLR*,

associated with LDL levels. This variant is annotated to be in a splice region, with MAF of 0.01 and meta-analysis *P*=1.49E-10. This variant is replicated in both rounds of replication, with *P*= 2.9E-11 and *P*=2.5E-02 respectively (**Table 4.9**). The combined meta-analysis result is: EAF=0.10 (A), beta (SE) =-0.326 (0.035), *P*=1.50E-20, N=51,757. This variant is independent of (LD r2<0.01) neighboring variants previously reported for association with CHD or lipids phenotypes (**Figure 4.4**). Previously, this variant was annotated as in intron 14 of *LDLR* under the name of "2140+5G>A", reported to have no effect on plasma cholesterol levels (Whittall et al. 2002) in a control sample with ~700 subjects. The fourth novel association, a common, X-linked variant associated with LDL (rs5985471, chrX:109703961, MAF=0.403, beta=0.050, *P*= 7.37E-08). This association is also replicated in two rounds of replication, with *P*= 6.6E-05 and *P*=2.8E-04 respectively. The combined meta-analysis result is: EAF=0.40 (T); beta (SE) = -0.042 (0.005), *P*= 2.02E-14, N=50,929. A sex-stratified analysis based on two cohorts with large number of males and females (ALSPAC and 1985BC) found that this association is significant in both males and females, therefore, not sex-specific. Within +/-500kb of rs5985471, there are two known associations, both of which are in high LD with rs5985471 (r2>0.8). The first one is rs5943057 (chrX:109939205), previously reported for association with CAD (*P* = 8.66E-07) in the C4D study (Coronary Artery Disease Genetics 2011). The minor allele for rs5985471 in this study is associated with a decreased level of LDL, i.e., protective. In the C4D study, the minor allele of rs5943057 is associated with a decreased level of CAD. The other known variant in strong LD is rs1573036 (chrX:109820068), previously reported for association with sex hormone-binding globulin levels (Coviello et al. 2012).

Figure 4.3 Association results of 14-way meta-analysis of the four main lipid traits

X-axis is for chromosome and positions (build 37). Y-axis is for –log10(*P*). Variants passing threshold of 5E-08 and 1E-07 are shown in red and blue, respectively. For those passing threshold of 5E-08, known loci were marked in green text while putative novel loci were marked in red text.

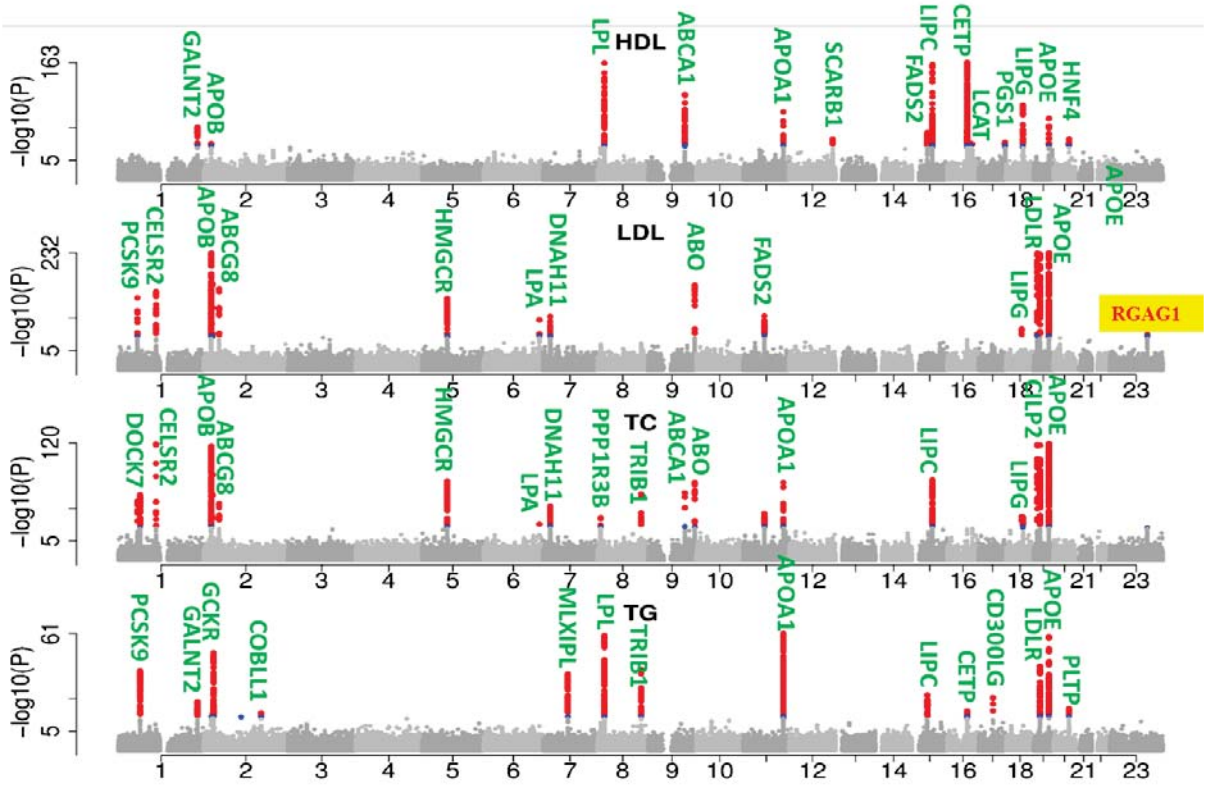


Table 4.9 Expanded discovery(14-way meta-analysis) top hits

This table shows the results of the expanded discovery meta-analysis (i.e., 14-way), followed by the two round of replications. For each set of results, the effect allele frequency (EAF), beta, standard deviation (SE), *P* value, and the total sample size were presented.

						14-way				
Trait	**rsID**	**CHR**	**POS**	**Gene**	**EA**	**EAF**	**Beta**	**SE**	**P**	**N**
TG	rs57367316	2	107,725,194	intergenic	A/G	0.165	0.074	0.014	6.9E-08	22,727
HDL	16:66926255	16	66,926,255	PDP2	T/A	0.003	-0.556	0.102	6.9E-08	22,385
LDL	rs72658867	19	11,231,203	LDLR	A/G	0.010	-0.342	0.053	1.5E-10	22,013
LDL	rs5985471	X	109,703,961	RGAG1	T/C	0.406	-0.047	0.009	7.4E-08	20,217

		Stage 1 replication					Stage 2 replication				
Trait	**rsID**	**EAF**	**Beta**	**SE**	**P**	**N**	**EAF**	**Beta**	**SE**	**P**	**N**
TG	rs57367316	0.156	-0.016	0.012	0.175	25599	--	--	--	--	--
HDL	16:66926255	0.002	-0.438	0.304	1.5E-01	4941	--	--	--	--	--
LDL	rs72658867	0.008	-0.390	0.059	2.9E-11	19099	0.010	-0.185	0.077	2.5E-02	10645
LDL	rs5985471	0.393	-0.034	0.008	6.6E-05	20066	0.406	-0.055	0.014	2.8E-04	10646

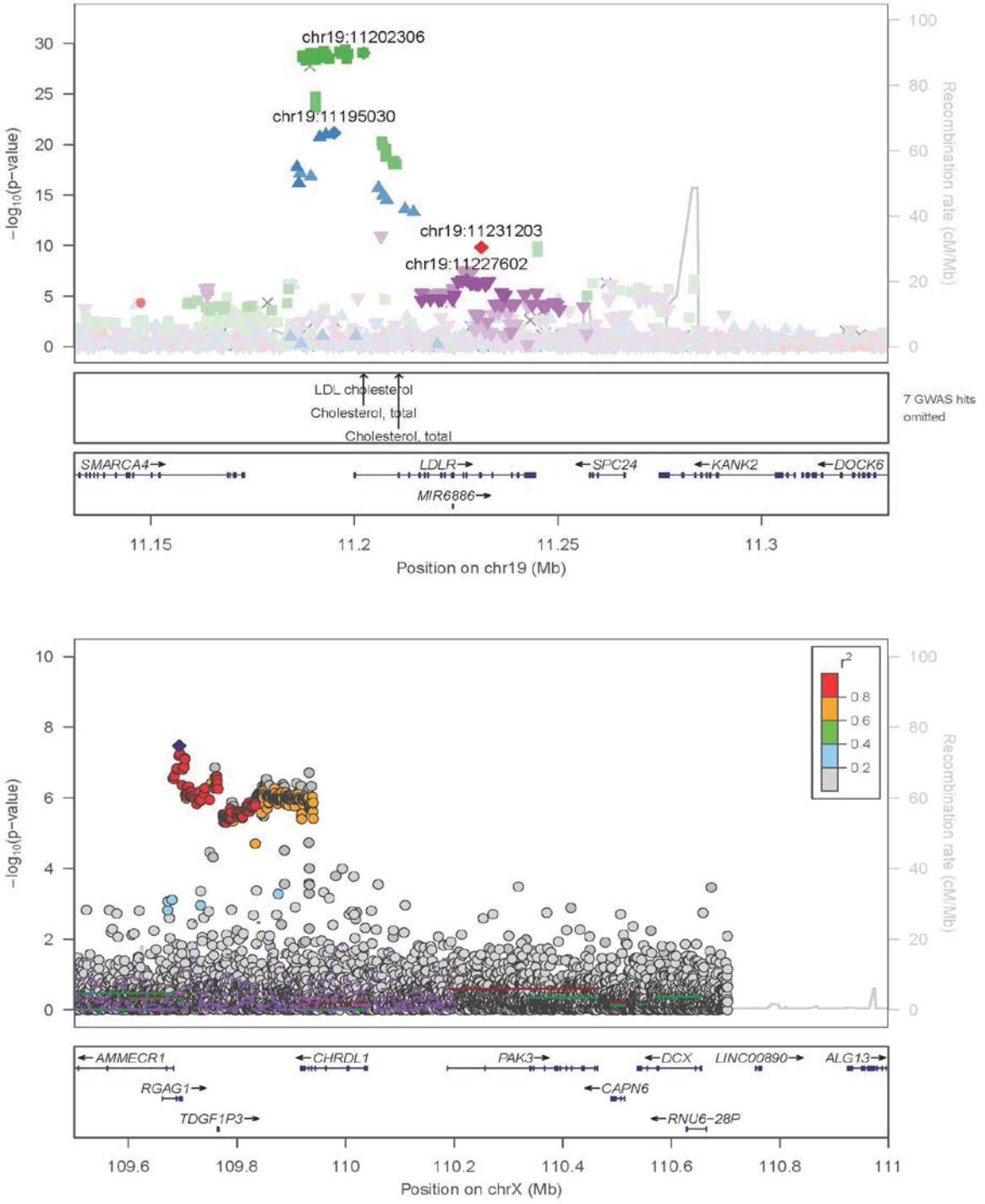
Table 4.10 Cohort specific results for four top variants based on 14-way meta-analysis

For each set of results, the effect allele frequency (EAF), beta, standard deviation (SE), *P* value, sample size (N), and imputation INFO score were presented. Records with *P* < 0.05 are highlighted in red text.

	rs57367316, TG						chr16: 66926255, HDL						rs72658867, LDL						rs5985471, LDL					
Cohort	**EAF**	**Beta**	**SE**	**P**	**N**	**Info**	**EAF**	**Beta**	**SE**	**P**	**N**	**Info**	**EAF**	**Beta**	**SE**	**P**	**N**	**Info**	**EAF**	**Beta**	**SE**	**P**	**N**	**Info**
ALSPAC WGS	0.154	0.113	0.052	3.0E-02	1497	0.99	0.002	-1.233	0.498	1.3E-02	1497	0.88	0.012	-0.713	0.157	5.8E-06	1495	0.99	0.404	-0.049	0.029	9.7E-02	1495	1.00
TwinsUK WGS	0.156	0.124	0.047	8.3E-03	1705	1.00	0.002	-0.838	0.372	2.5E-02	1713	0.88	0.012	-0.452	0.159	4.5E-03	1696	0.96	0.399	0.049	0.035	1.6E-01	1696	1.00
ALSPAC GWA	0.154	0.077	0.039	5.1E-02	2820	0.90	0.003	-0.191	0.328	5.6E-01	2820	0.63	0.009	-0.524	0.157	8.6E-04	2815	0.83	0.392	-0.077	0.023	6.4E-04	2815	1.00
TwinsUK GWA	0.154	0.002	0.048	9.6E-01	1882	0.92	0.003	-0.581	0.331	8.0E-02	1896	0.63	0.009	-0.245	0.189	1.9E-01	1870	0.83	0.406	-0.051	0.031	9.8E-02	1870	1.00
1958BC	0.154	0.081	0.028	3.8E-03	5485	0.91	0.003	0.080	0.214	7.1E-01	5493	0.65	0.011	-0.360	0.102	4.1E-04	5186	0.83	0.394	-0.087	0.016	1.5E-07	5186	1.00
INGI-CARL	0.143	0.077	0.121	5.3E-01	412	0.78	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
INGI-FVG	0.168	0.156	0.058	7.6E-03	1377	0.84	--	--	--	--	--	--	0.005	-0.049	0.343	8.9E-01	1377	0.61	0.452	-0.041	0.032	2.2E-01	1377	0.96
INGI-VB	0.178	0.006	0.053	9.1E-01	1776	0.78	0.004	-0.106	0.458	8.2E-01	1776	0.44	0.014	-0.238	0.193	2.2E-01	1775	0.59	0.379	-0.033	0.034	3.3E-01	1775	0.80
HELIC-A	0.238	0.018	0.052	7.4E-01	1245	0.86	0.009	-1.491	0.236	9.7E-10	1247	0.89	0.017	-0.206	0.205	3.2E-01	1253	0.60	0.384	0.007	0.035	8.5E-01	1253	1.00
HELIC-P	0.151	-0.033	0.071	6.4E-01	964	0.88	0.002	-0.953	0.559	8.9E-02	976	0.91	0.002	-0.904	0.512	7.9E-02	976	0.78	0.393	0.082	0.042	5.3E-02	976	0.99
INCIPE-1	0.191	0.092	0.071	2.0E-01	653	0.88	0.003	-0.358	0.634	5.7E-01	653	0.63	0.015	0.056	0.290	8.5E-01	653	0.69	0.462	-0.079	0.048	9.7E-02	653	0.99
INCIPE-2	0.175	0.010	0.056	8.6E-01	1382	0.80	0.004	-0.413	0.312	1.9E-01	1382	0.89	0.012	0.284	0.213	1.8E-01	1380	0.69	0.442	-0.018	0.031	5.6E-01	1380	0.99
LURIC	0.177	0.167	0.063	8.2E-03	983	0.85	0.002	-0.101	0.634	8.7E-01	983	0.52	0.009	-0.255	0.289	3.8E-01	983	0.71	0.409	-0.038	0.037	3.0E-01	960	1.00
Teenage	0.171	0.236	0.087	6.8E-03	551	0.85	0.004	0.231	0.616	7.1E-01	557	0.56	0.007	-0.304	0.520	5.6E-01	557	0.50	0.413	-0.065	0.051	2.1E-01	557	0.99
Fenland	0.162	-0.017	0.021	4.2E-01	8660	1.00	--	--	--	--	--	--	0.010	-0.473	0.076	4.9E-10	8701	1.00	0.392	-0.041	0.013	1.5E-03	8590	1.00
FinRisk	0.130	-0.133	0.067	4.6E-02	856	1.00	--	--	--	--	--	--	0.001	-0.024	0.697	9.7E-01	817	1.00	--	--	--	--	--	--
GoT2D	0.151	-0.007	0.042	8.7E-01	2190	--	--	--	--	--	--	--	0.006	-0.426	0.203	3.5E-02	2076	--	--	--	--	--	--	--
InChianti	0.204	-0.035	0.074	6.4E-01	614	1.00	--	--	--	--	--	--	0.013	-0.579	0.252	2.2E-02	621	1.00	0.383	-0.129	0.048	7.0E-03	621	1.00
Lolipop EW610	0.167	-0.064	0.065	3.2E-01	927	0.90	--	--	--	--	--	--	0.016	-0.166	0.200	4.1E-01	905	0.91	--	--	--	--	--	--
Lolipop EWA	0.148	-0.065	0.085	4.4E-01	582	0.89	--	--	--	--	--	--	0.004	-0.360	0.598	5.5E-01	566	0.66	--	--	--	--	--	--
Lolipop EWP	0.160	-0.024	0.084	7.7E-01	642	0.83	--	--	--	--	--	--	0.013	0.125	0.267	6.4E-01	610	0.89	--	--	--	--	--	--
RS-1	0.157	-0.021	0.037	5.6E-01	3108	0.90	0.001	0.102	0.582	8.6E-01	3081	0.48	0.005	-0.578	0.226	1.1E-02	2981	0.72	0.396	-0.010	0.022	6.4E-01	2981	1.00
RS-2	0.158	0.026	0.048	5.9E-01	1847	0.89	0.003	-0.640	0.356	7.2E-02	1861	0.69	0.006	0.199	0.269	4.6E-01	1823	0.59	0.381	0.009	0.028	7.5E-01	1823	0.99
UCLEB BRHS	0.149	0.025	0.038	5.1E-01	2785	1.00	--	--	--	--	--	--	--	--	--	--	--	--	0.399	-0.049	0.020	1.3E-02	2742	1.00
UCLEB BWHS	0.146	-0.018	0.034	6.0E-01	3388	1.00	--	--	--	--	--	--	--	--	--	--	--	--	0.397	-0.020	0.025	4.3E-01	3309	1.00
WHI garnet	0.163	0.048	0.032	1.4E-01	3755	0.93	0.003	0.078	0.240	7.5E-01	3781	0.75	0.011	-0.136	0.128	2.9E-01	3726	0.77	0.400	-0.066	0.023	4.5E-03	3726	0.99
WHI hipfx	0.149	-0.039	0.075	6.0E-01	799	0.89	--	--	--	--	--	--	0.012	0.177	0.290	5.4E-01	639	0.80	0.401	-0.046	0.057	4.2E-01	639	0.99
WHI mopmap	0.164	0.120	0.071	9.2E-02	768	0.94	--	--	--	--	--	--	0.006	-0.100	0.410	8.1E-01	745	0.63	0.380	-0.023	0.053	6.7E-01	745	0.99
WHI whims	0.163	0.038	0.027	1.6E-01	5546	0.92	0.003	0.152	0.223	5.0E-01	5580	0.71	0.011	-0.268	0.104	9.8E-03	5537	0.79	0.415	-0.052	0.019	6.7E-03	5537	1.00

Figure 4.4 Regional plots of two loci with replicated novel associations

The top plot is for association with LDL in the *LDLR* region. The bottom plot is for the novel locus on chromosome X. Both are for association with LDL and *P* values are based on the 14-way meta-analysis. For the LDLR locus, the novel variant is shown in red text, while the SNPs tagged by previously reported variants are known in other colors. For the chromosome X region, there were no previously reported variants.



4.3.2 Fine mapping of known and novel loci

To fine-map lipid-associated regions, I implemented the method of Maller et al. (Maller et al. 2012), as described in chapter 2 and the Methods section above. For 41 out of a total of 282 regions examined, there are sufficient resolution to limit the number of possible causal variants to a small informative set (log10BF>5 and # of variants <20). The distribution of the number of causal variants within these 41 loci is shown in **Figure 4.5**.

To further characterize the predicted functional consequence of the FM variants, the fine-mapping regions were overlapped with four liver-essential TFBS data (Ballester et al. 2014). Ten variants that are in the 95% credible set of these 41 fine-mapped regions also overlapped with a TFBS (**Table 4.11**). These 10 variants should be considered as good candidates for further functional and causality studies. By further overlapping these 10 variants with liver expression of quantitative trait loci (eQTL) data on GTEx (http://www.gtexportal.org/), I identified two variants have significant eQTL signal (eQTL *P*<5E-08). The first one is rs12740374 in *SORT1*, which was previously identified as causal (Musunuru K, et. al. 2010, Nature). The second one is rs10438978 (A/G alleles) close to *LIPG*, with eQTL *P*=1.96E-10 and motif change of CTCF_disc3. The discovery of a causal variant in *SORT1* locus demonstrated the proof-of-concept for this approach.

Figure 4.1 Number of putative causal variants within fine-mapped loci

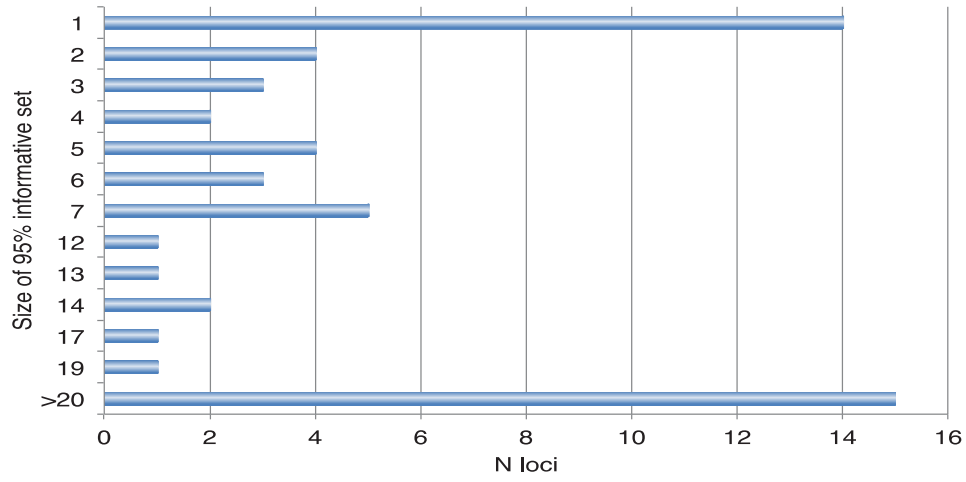


Table 4.11 Predictive causal variants based on fine mapping

This table lists 10 putative causal variants within the 41 fine-mapped regions that overlap with a TFBS.

BF: bayes factor, PP: posterior probability

Trait	SNP	Chr	Pos	log10BF	PP	gene
LDL	rs12740374	1	109,817,590	24.33	0.15	CELSR2 3_prime
LDL	rs4245791	2	44,074,431	7.41	0.30	ABCG8 intron
HDL	rs4100654	9	107,669,241	9.13	0.71	ABCA1 intron
HDL	rs1077834	15	58,723,479	25.83	0.10	LIPC:upstream
HDL	rs1800588	15	58,723,675	26.26	0.25	LIPC:upstream
HDL	rs2070895	15	58,723,939	26.36	0.33	LIPC:upstream
HDL	rs10438978	18	47,158,186	10.72	0.18	LIPG
HDL	rs9304381	18	47,158,234	10.93	0.29	LIPG
LDL	rs58542926	19	19,379,549	25.24	0.15	TM6SF2 missense
TG	rs483082	19	45,416,178	15.76	0.26	APOE upstream

			WGS				14-way			
Trait	SNP	EA	EAF	beta	SE	P	EAF	beta	SE	P
LDL	rs12740374	T	0.211	-0.178	0.030	3.2E-09	0.218	-0.139	0.012	2.9E-32
LDL	rs4245791	T	0.658	-0.033	0.025	1.9E-01	0.663	-0.080	0.010	4.3E-15
HDL	rs4100654	C	0.098	-0.205	0.042	1.3E-06	0.096	-0.128	0.017	1.4E-14
HDL	rs1077834	C	0.204	0.136	0.031	1.5E-05	0.215	0.146	0.012	7.6E-35
HDL	rs1800588	T	0.202	0.137	0.031	1.3E-05	0.211	0.148	0.012	1.9E-35
HDL	rs2070895	A	0.204	0.134	0.031	2.0E-05	0.215	0.147	0.012	2.4E-35
HDL	rs10438978	C	0.819	0.048	0.033	1.5E-01	0.835	0.098	0.013	5.2E-14
HDL	rs9304381	T	0.819	0.048	0.033	1.5E-01	0.836	0.099	0.013	3.5E-14
LDL	rs58542926	T	0.073	-0.140	0.048	3.4E-03	0.074	-0.190	0.018	6.6E-25
TG	rs483082	T	0.242	0.126	0.029	1.6E-05	0.213	0.130	0.012	2.7E-27

4.3.3 Novel loci based on rare variants aggregation test

The above are for single marker base tests, which has limited power to detect associations for low frequency and rare variants given the current number of samples with WGS. Here I show association results based on rare variants aggregation tests. As stated in the Methods section, three types of SKAT-O analyses were run: genome-wide sliding window, exome-wide gene based, and exome-wide with only functional variants. Overall, the statistics of these tests follow the expected distribution assuming a NULL association, where the lambda is close to 1 and the tail does not significantly deviate from the expected (**Figure 4.6**). Of note, the QQ plots are not based on SKAT-O *P* value because that is a statistic after comparing two tests (SKAT and burden). The genome-wide significance thresholds are predefined as 6.8E-08, 1.2E-06, 1E-05 respectively for genome-wide, exome-wide, and functional variants based SKAT-O. There are four loci surpassing these significance thresholds (**Figure 4.7)**. These four windows and another 103 windows with *P*<1E-5 in GW and *P* <1E-4 for EW based were taken forward for replication in three cohorts (GoT2D, FinRisk, InChianti). At the most liberal threshold of replication P<0.05, 19 windows have evidence for replication by either SKAT or burden statistics. However, only the *APOC3* region has an adequate replication (*P* < 0.0005) that survived the multiple tests on 107 windows, with combined SKAT *P*=1.36E-08. The only other window with a combined SKAT *P*<5E-08 is chr4:110946001-110949000 for TG (SKAT *P* =2.23E-08). As shown in **Figure 4.8**, the peak of the SKAT signal lies between the *EGF* and *ELOVL6* gene. The full name for *ELOVL6* is ELOVL Fatty Acid Elongase 6, whose function is to catalyze the synthesis of saturated and monounsaturated fatty acids. It is certainly a plausible gene for impacting circulating lipids levels. The best single marker variant within this region is rs184358074, AF=0.6%, *P*=5.3E-04, which would be considered non-significant based on the pre-defined threshold. Drop-one anlaysis confirmed that this signal is not driven by any single variants that were included in the SKAT-O analysis.

Figure 4.6 QQ plots of SKAT tests for lipids

The four columns are for HDL LDL TC TG; each pairs of rows are for genome-wide, exome-wide, and functional variants.

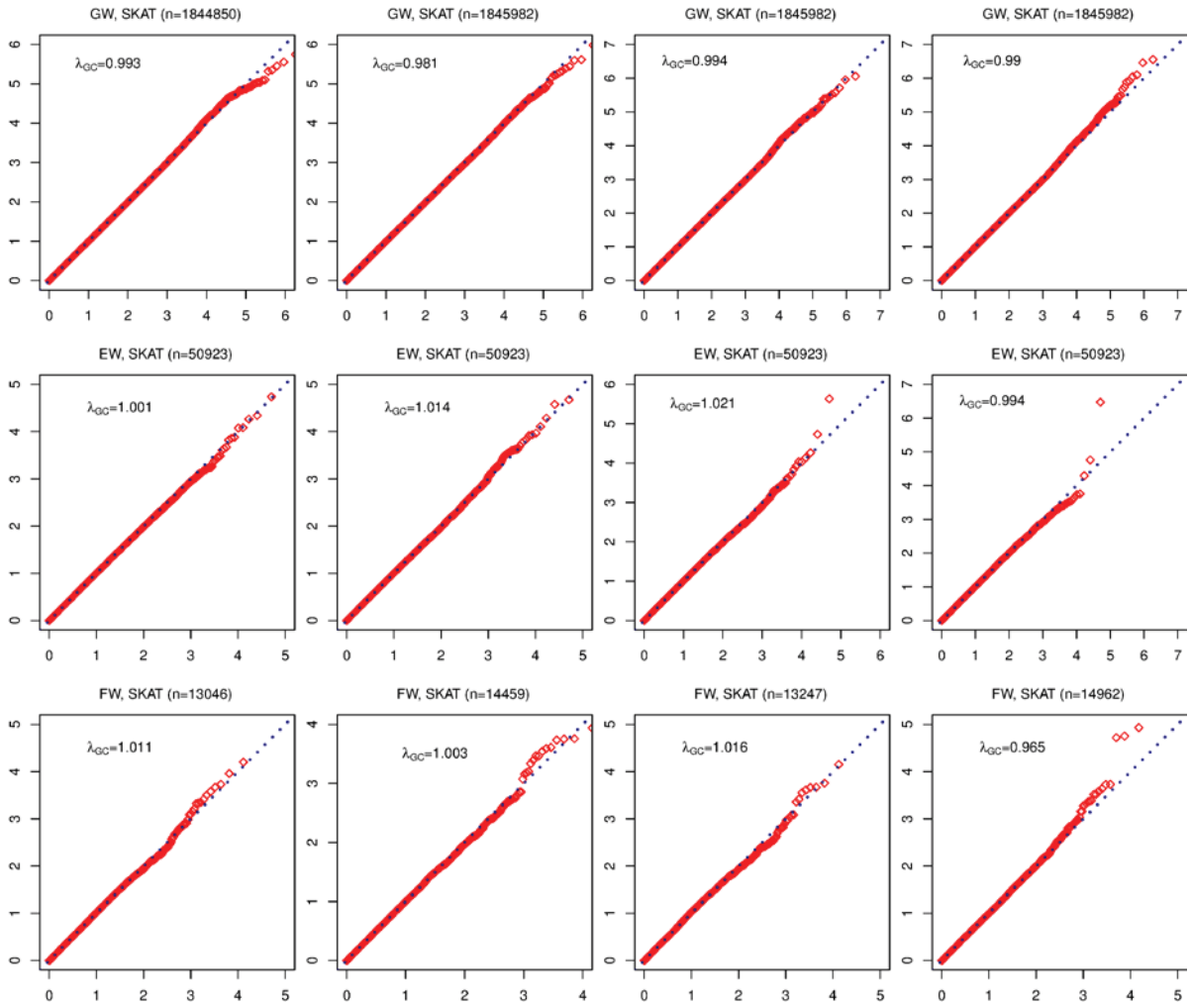


Figure 4.7 Rare variants aggregation test results for lipids

The genome-wide significant signals are shown in red, with threshold of $P < 6.8E-08$, 1.2E-06, 1E-05 respectively for genome-wide, exome-wide, and functional variants based SKAT-O. Suggestive signals are shown in blue, with threshold of $P < 1E-05$, 1E-04, 1E-04 respectively for genome-wide, exome-wide, and functional variants based SKAT-O.

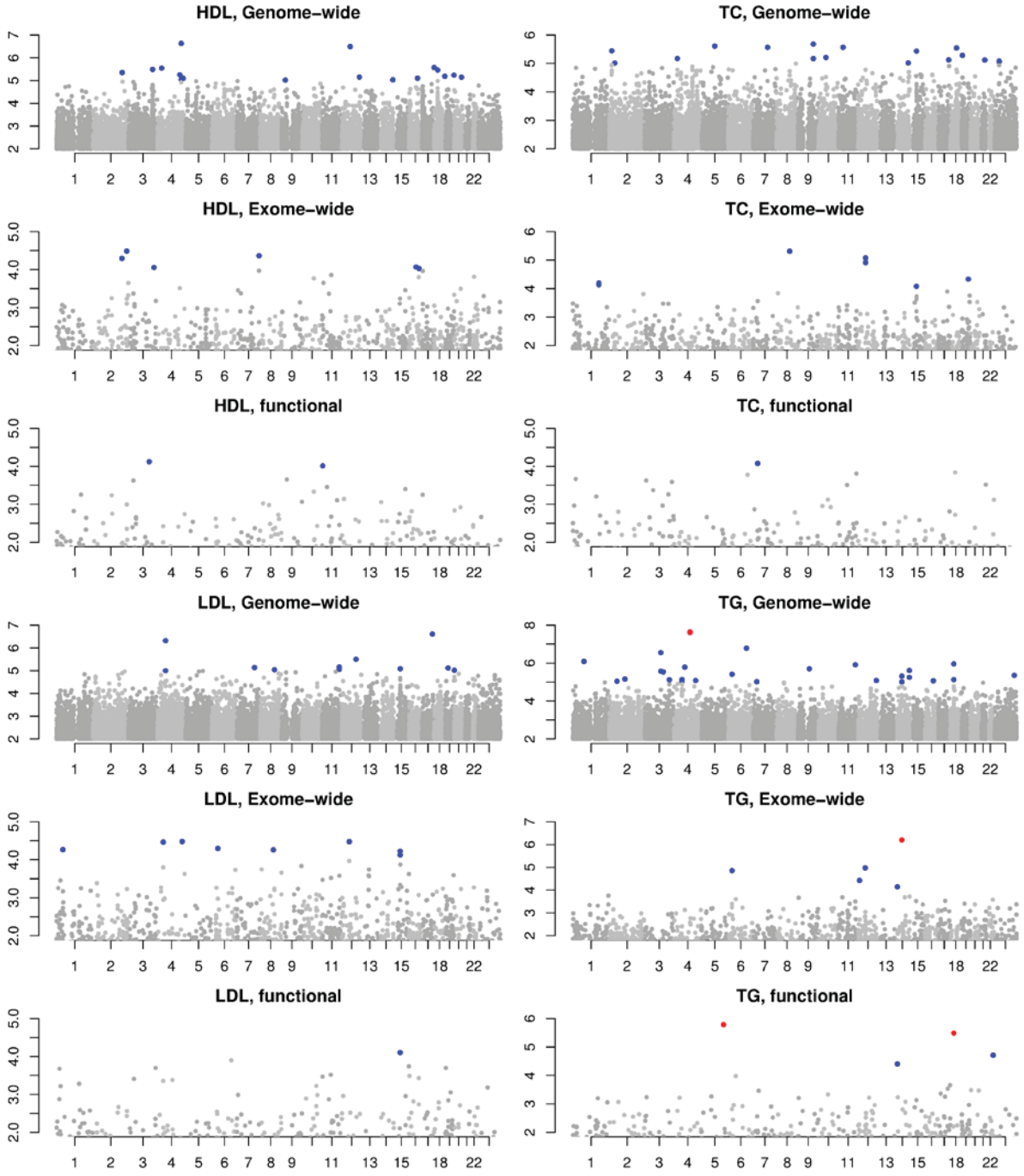
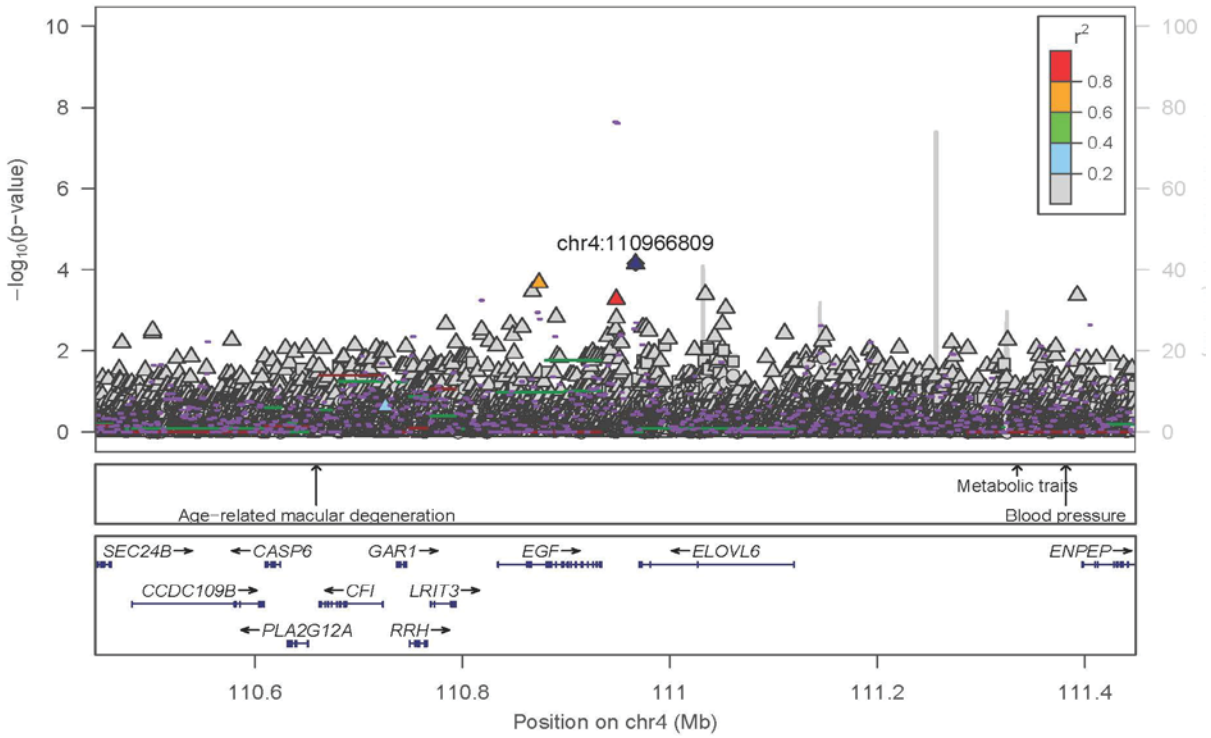


Figure 4.8 Regional plot of SKAT-O locus *EGF-ELOVL6*

The UK10K WGS single marker results are shown in points, where circle, cube, and triangle are used for common, low frequency, and rare variants. The UK10K SKAT-O results are shown in horizontal lines, where purple, green, brown are used for genome-wide SKAT, exome-wide SKAT, and functional variants exome-SKAT.



4.4 Conclusion & Discussion

4.4.1 Summary of main findings

This is by far the largest genome-wide scan on identifying genetic variants of plasma lipids using WGS data. Although the total sample size is much smaller than that in Global lipids study, the sequencing generated data and WGS imputed data provide an unprecedented opportunity to uncover rare and causal variants and their associations, as demonstrated by the example of *APOC3*, *LDLR*, and the novel locus on chromosome X. Although the clinical relevance of the *LDLR* variant (rs72658867) is yet to be confirmed, the *APOC3* variant (rs138326449, IVS2+1G→A) was already reported to be strongly associated with reduced CHD risk. In two studies that established the causality of rare variants within *APOC3*, one used high-depth WES (Tg et al. 2014) and the other used targeted re-sequencing (Jorgensen et al. 2014). The UK10K data is the first low-coverage WGS data that discovered this variant through both single marker based test and rare variant aggregation test.

Recently, there was an exome-array based study reported four rare variants for association with HDL or TG with large effect sizes (Peloso et al. 2014). But only one variant, rs186808413 within *PAFAH1B2*, is marginally significant in the UK10K WGS based results, P=0.018. This variant is in low LD with the reported splice variant within *APOC3* (rs138326449), r2=0.18, 341kb apart. Another WES based study reported an association between LDL and the burden of rare and low-frequency variants in *PNPLA5* (Lange et al. 2014). However, this result is not replicated in our exome-wide based SKAT-O test (P >0.05).

4.4.2 Interpretation of results

A wealth of novel lipid loci have been identified through a variety of approaches focused on common and low-frequency variation and collaborative meta-analyses in multi-ethnic populations. Despite progress in identification of loci, the task of determining causal variants remains challenging. This work will undoubtedly be enhanced by improved understanding of regulatory DNA at a genome-wide level as well as new methodologies for interrogating the relationships between noncoding SNPs and regulatory regions. Equally

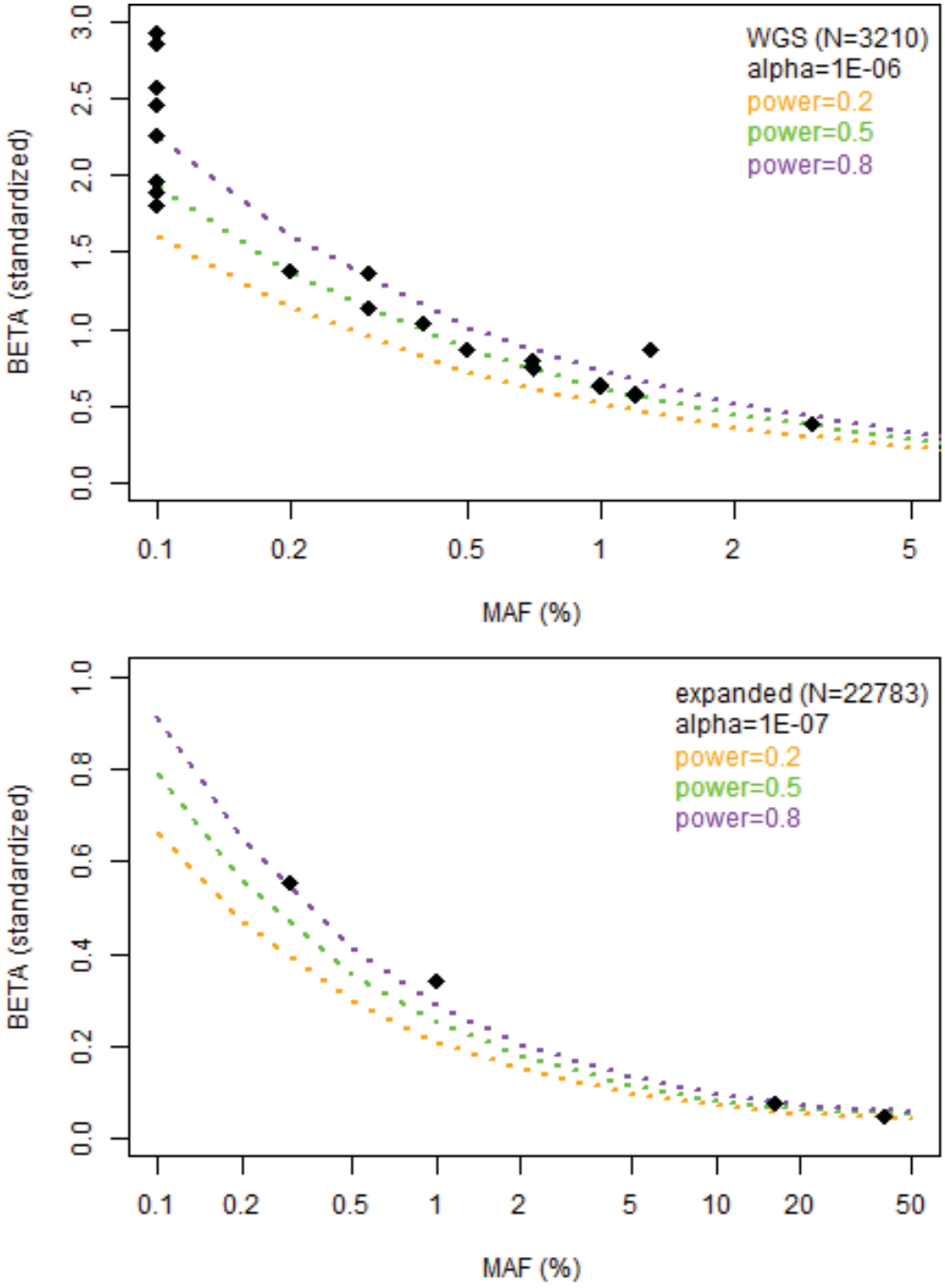
challenging is the identification of causal genes at novel loci. Additional insights will be gleaned from focusing on low-frequency and rare coding variation at candidate loci in large populations.

The single marker association testing of four lipids follows closely the expected relationship between EAF and effect size (beta) as dictated by study power (Park et al. 2011), as shown in **Figure 4.9**. Low frequency alleles of very high penetrance (beta ~1 SD) are unlikely to exist within this allelic space in the general European-ancestry population. Examples such as the rare *APOC3* or *LDLR* variants, with sufficient individual effect sizes to be clinically informative, are beginning to emerge (Flannick et al. 2012), but these findings are likely to be exceptions rather than a paradigm. Greater power than the current study will be required for capturing a greater proportion of missing heritability through either increases in sample size or genotyping accuracy and SNV density. The assessment of rare variants using a range of single-marker, exome-based and genome-based tests suggests that naïve and even functional scans were broadly underpowered to detect associations with high certainty, requiring extensive follow-up replication studies (Zuk et al. 2014). Deep sequencing will be needed to discover and fully assess this frequency range, which contains highly penetrant, potentially clinically important variants not accessible through imputation.

Finally, based on **Table 4.1** and **Figure 4.1**, there are five genes that were discovered by both linkage analysis and GWAS: ABCA1, ABCG5, ABCG8, LDLRAP1, PCSK9. However, none of these gene regions is significant based on exome-wide SKAT-O analyses (P >0.05). In single variant based analysis, there are no variants with MAF <5% in these genes have a P-value that surpassed the pre-defined threshold of 1.0E-07. This could be very likely due to the limited power of the current study to detect association signals for low frequency and rare variants.

Figure 4.9 Statistical power and novel variants from single marker analysis

The top and bottom plots are for WGS samples and expanded discovery samples respectively. Y-axis is a variant's effect, expressed in standard deviation units. X-axis is MAF of effect alleles. Colored lines indicate 20%, 50%, and 80% power. Alpha is set at *P*<1E-06 for WGS and *P*<1E-07 for expanded discovery respectively. The 16 putative novel WGS variants are shown in the top power plot for WGS, and the four putative novel variants from expanded discovery are shown in the bottom power plot for expanded discovery.



4.4.3 Future direction

Presently, there are still challenges in applying statistical methods to rare variants based analysis, especially when the sample size is small. During phenotype harmonization, samples with values that are more than three standard deviation of the mean are excluded. This is justifiable given that the focus of this study is on quantitative traits in healthy populations. However, this approach might have prevented the identification of a small group of individuals who carry rare variants with large effects that are linked with Mendelian conditions, as that reported by the Morrison study (Morrison et al. 2013).

As the field of lipid genetics moves beyond GWAS to focusing on identification of causal variants, causal loci, and biological mechanisms underlying novel genes, the study of low frequency and rare variants with large sample sizes and integrating genomic data with functional data would be critical. For common noncoding variants that are within (or in high LD with) defined promoter or known regulatory regions of nearby genes, one could assess the underlying effects of them through gene reporter assays, binding affinity for specific transcription factors, and related functional approaches. Such efforts have been done for a limited number of lipid-associated variants, such as for the causal role of *SORT1* to LDL and CVD risk (Musunuru et al. 2010), where the minor allele of the causal variant within a cis-regulatory region was found to create a de novo C/EBP TFBS that caused C/EBP-dependent upregulation of expression of the nearby genes. Another approach is to overlay GWAS variants with regions with chromatin marks or regions of DNase I hypersensitivity, suggesting open chromatin and active transcription (Maurano et al. 2012). Finally, in vivo overexpression or knockdown of candidate genes at a locus in animal models would provide most convincing causal evidence. The large lipids GWAS in 2010 reported such work for three candidate genes influencing HDL: *GALNT2*, *PPP1R3B* and *TTC39B* (Teslovich et al. 2010).