Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Genetics & Development

Insights into blood lipids from rare variant discovery

Ellen M Schmidt[1] and Cristen J Willer[1,2,3]

Large-scale genome wide screens have discovered over 160 common variants associated with plasma lipids, which are risk factors often linked to heart disease. A large fraction of lipid heritability remains unexplained, and it is hypothesized that rare variants of functional consequence may account for some of the missing heritability. Finding lipid-associated variants that occur less frequently in the human population poses a challenge, primarily due to lack of power and difficulties to identify and test them. Interrogation of the protein-coding regions of the genome using array and sequencing techniques has led to important discoveries of rare variants that affect lipid levels and related disease risk. Here, we summarize the latest methods and findings that contribute to our current understanding of rare variant lipid genetics.

Addresses
[1] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA
[2] Department of Internal Medicine, Division of Cardiovascular Medicine, University of Michigan, Ann Arbor, MI 48109, USA
[3] Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA

Corresponding author: Willer, Cristen J (cristen@umich.edu)

Current Opinion in Genetics & Development 2015, **33**:25–31

This review comes from a themed issue on **Molecular and genetic bases of disease**

Edited by **Dan E Arking** and **Johanna M Rommens**

http://dx.doi.org/10.1016/j.gde.2015.06.008

Introduction

Cardiovascular disease is one of the leading causes of death in the US and throughout the world, representing a significant human health burden [1]. Genetic studies of lipid levels, known risk factors for heart disease with heritability ranging from 40% to 60% in humans [2], are logical targets in efforts to prevent and treat heart disease. Modulation of these quantitative lipid traits, which include low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides (TG), and total cholesterol (TC), can be effective therapeutically. Lipid traits are not independent, making it challenging to untangle the effects of specific lipids on disease risk. LDL-C and TC generally act in the same direction since the majority of TC is composed of LDL-C. In addition, high TG is associated with high LDL-C and low HDL-C, while LDL-C and HDL-C are positively and inversely associated with heart disease risk, respectively [3].

We can harness the information from single nucleotide polymorphisms (SNPs) in the human population to gain knowledge about the genetic contribution to plasma lipids. Early genome wide association studies (GWAS) with modest sample sizes (<10,000) uncovered common variants (MAF > 5%) with large effect sizes. The power to discover new lipid-associated variants increases with sample size [4–6]. In the largest lipids GWAS to date, researchers jointly meta-analyzed data from a total sample size of ~180,000, to uncover 157 independent genetic loci containing lipid-associated variants [7••]. Most of the associated variants are non-protein-coding however, suggesting a regulatory role [8]. An illustration of the regulatory role of a noncoding LDL-C-associated variant is evident at the *SORT1* locus. Researchers demonstrated in human-derived hepatocytes that variant rs12740374 at an LDL-C-associated GWAS locus creates a C/EBP transcription factor binding site, causing altered expression of the nearby *SORT1* gene [9]. Still, common variants generally have limited functional consequence. Understanding the underlying biology of noncoding variation and translating these findings into clinical practice remains a universal challenge.

Common variants identified by GWAS explain only a fraction (10–15%) of the trait variance for lipids [6]. In effort to explain some of the 'missing heritability', we turn to rare (MAF $\leq$ 0.5%) and low frequency (0.5% < MAF $\leq$ 5%) variation. Protein-coding variants with deleterious function are likely to be rarer in the human population due to natural selection acting against them, and have mostly arisen recently in evolutionary history [10]. The Common Disease-Rare Variant hypothesis proposes that the combined effect of a number of low frequency variants with large effect sizes accounts for some of the missing heritability [11]. Indeed, early sequencing studies of candidate genes supported the contribution of multiple rare alleles to plasma HDL-C levels [12]. This rare variation cannot be found with traditional GWAS single-variant approaches due to poor coverage on arrays and difficulties with imputing.

Mendelian family studies involving large pedigrees are valuable for rare variant genetic studies. In this design, the co-segregation of variants among affected family members can be traced. Investigators recently used

whole-exome sequencing in a multi-generation family to uncover a rare variant in a highly conserved codon of *SLC25A40* that is associated with TG, giving insight into a previously unknown biological mechanism of hypertriglyceridemia [13]. In this review, our subsequent focus will be on the rare variant findings from large-scale array and sequencing studies for complex lipid traits.

Methods for rare variant association testing

Because of the rare nature of most variants with functional consequence, studies carefully designed to uncover rare variant associations are crucial [14,15••]. Recent advances in exome sequencing and exome array technologies have facilitated larger and more accurate studies for interrogating the protein-coding 1–2% of the genome. However, single variant association tests commonly used by GWAS carry a heavy multiple testing burden and still lack power when applied to rare variants of high impact. Thus, aggregation-based tests that group multiple variants by a single gene or functional unit have become standard for rare variant association analysis.

Several regression-based approaches have been developed in recent years to optimize rare variant discovery. In a simple burden test, multiple rare variants are collapsed into a genetic score representing the cumulative effect of those variants in a single unit. Then, the score is tested for association with a trait or disease. This idea, which has been implemented by numerous investigators [16–20], assumes that all variants in a single unit are causal and that all alleles affect the phenotype with the same magnitude and direction of effect. More robust modifications of a simple burden test introduce adaptive weights or thresholds [21–26]. We can account for the protective or deleterious impact of alleles on phenotype by considering the magnitude and direction of effect using variance-component tests [27–29]. Finally, we can combine burden and component tests [30•,31,32] or score statistics [33] to achieve more robust power. Choosing the optimal strategy for grouping rare variants is flexible and may depend on the genetic architecture of a particular trait [34].

Meta-analysis is a powerful tool to jointly analyze datasets from multiple studies, especially when individual-level data are not available. Traditional Fisher's method [35] for combining *P*-values is straight-forward, but does not weight by sample size or consider magnitude or direction of effect. When studies provide score statistics for each variant and a variance–covariance matrix, a fixed-effects meta-analysis will achieve improved power [36•,37,38]. Both genotyping and sequencing-based replication strategies are appropriate for confirmation of rare variant association findings [39].

Rare coding variation implicated in lipids

Interrogating the protein-coding genome through sequencing coding regions [40] and whole-exome sequencing [41] is hypothesized to reveal rare mutations with a large effect on phenotype. A splice variant in *APOC3* associated with TG was identified using whole-genome sequencing, representing one of the first rare variants of large effect to be found using this sequencing approach at the population scale [42]. Sequencing the exome is more cost-effective than whole-genome sequencing, however, and fewer statistical tests are performed, reducing the multiple testing burden. The potential of exome sequencing has resulted in studies powered for the discovery of novel rare variation implicated in blood lipids. For example, investigators of the NHLBI Exome Sequencing Project (ESP) used exome sequencing to identify the burden of rare variants in four genes (*PNPLA5*, *PCSK9*, *LDLR* and *APOB*) significantly associated with LDL-C [43•] (Table 1). By contrast to modest effect sizes observed from individual SNPs identified by GWAS, the burdens of rare variants in these genes have substantially higher effect sizes (Figure 1).

Although it is clear that increasingly large sample sizes will improve power in rare variant discovery, this raises the question of how many individuals need to be sequenced to assess the true impact of rare variation on lipid traits and related diseases. LDL-C is a reasonably heritable quantitative trait, and associated variants have relatively high effect sizes. Given the genetic architecture of lipids, more than just a few thousand samples will likely need to be sequenced to comprehend the full impact of rare variation. A more cost-effective study design involves screening millions of genetic variants by 1000 Genomes Project imputation. Association testing of genotyped and successfully imputed SNPs has led to novel insights into the impact of rare variants on lipids [44•].

Additionally exome chip, a custom genotyping array, allows for large-scale efficient genotyping of low frequency coding variants with large effect sizes. Exome wide association studies for lipids and related diseases revealed several significant variants at both established and previously unknown lipid loci. Rare variants at *ANGPTL4*, *LIPC* and *LIPG*, for example, were found to be associated with TG and HDL-C [45] (Table 1). In addition, a more common variant in the protein-coding gene, *TM6SF2*, was found to be associated with TC and myocardial infarction risk [45,46]. Functional follow-up revealed that modulation of Tm6sf2 in mice alters lipid levels, providing the causal gene at a GWAS locus that was previously intractable for follow-up due to a large number of genes in the associated region. These findings illustrate the importance of interrogating changes in the exome in guiding our knowledge of the functional gene at lipid loci.

Translation of plasma lipid levels to disease risk

Understanding the relationship between plasma lipid concentrations and disease risk is paramount in human