HHS Public Access

Author manuscript

Circ Res. Author manuscript; available in PMC 2017 February 18.

Published in final edited form as:

Circ Res. 2014 August 15; 115(5): 478–480. doi:10.1161/CIRCRESAHA.114.304693.

Study of Exonic Variation Identifies Incremental Information Regarding Lipid-Related and Coronary Heart Disease Genes

Themistocles L. Assimes and **Thomas Quertermous**

Department of Medicine, Cardiovascular Research Institute, Stanford University School of Medicine, Stanford, CA

Abstract

Recently, a modest-sized population-based study of exonic variants facilitated the identification of the causal gene, *TM6SF2*, in a gene-rich locus on 19p1 previously associated with cholesterol levels in blood. The study also provided compelling functional validation of the locus and evidence at the population level that interference with the function of this gene may substantially reduce the risk of coronary artery disease. The study highlights the potential utility of large-scale studies of coding variants but also hints toward the need of much larger studies to provide insight at other loci. Conducting such studies in parallel with association studies of variation in well-annotated regulatory regions is likely to ultimately yield the highest returns.

Genome-wide association studies (GWAS) conducted over the past decade have identified many regions of the genome harboring common susceptibility variants for complex traits.[1] For many traits, the overall proportion of the genetic variance explained by GWAS to date remains modest.[2] A few exceptions exist including blood levels of lipids where large meta-analyses have successfully identified >150 loci for total cholesterol (TC), low-density lipoprotein (LDL), high-density lipoprotein (HDL), and triglycerides.[3,4] Remarkably, these discoveries seem to explain more than one quarter of the genetic variance of each trait.[3,4]

A majority of these GWAS loci are novel and have not been previously linked to lipid biology.[3,4] For many, the lead single-nucleotide polymorphism (SNP) falls in noncoding regions of the genome and the association interval stretches over several kilobases often overlapping several candidate genes.[3,4] This situation makes it extremely challenging to determine the causal gene. Several population genetic approaches can be used to help identify the causal gene. First, one can examine the same risk interval in multiple racial groups with the hope that differences in linkage disequilibrium among groups will narrow the interval significantly.[5] However, this approach may not be practical if very large samples

Correspondence to Thomas Quertermous, MD, Division of Cardiovascular Medicine, Stanford University, 300 Pasteur Dr, Stanford, CA 94305. tomq1@stanford.edu.

The opinions expressed in this Commentary are not necessarily those of the editors or of the American Heart Association.

Commentaries serve as a forum in which experts highlight and discuss articles (published here and elsewhere) that the editors of *Circulation Research* feel are of particular significance to cardiovascular medicine.

Commentaries are edited by Aruni Bhatnagar.

Disclosures

None.

of other racial groups either do not exist or have not been genotyped. Furthermore, this approach may not be feasible if adequate haplotype diversity is lacking across racial groups in the region of interest.[5] A second more straightforward method involves the careful survey of all exonic SNPs that are highly correlated with the index GWAS SNP. Most of these SNPs will by necessity be common and already known, allowing them to be reliably imputed if they have not already been genotyped directly. However, a small fraction of such SNPs may have escaped detection to date particularly if they are population-specific. Lastly, systematic assessment of coding variation in the region through targeted sequencing or genotyping of less common and rare variants in a population sample may point to the causal gene through the identification of ≥1 novel coding variants that are strongly associated with the phenotype.[6] Such SNPs are much more likely to have a lower frequency and to be statistically independent of the index GWAS SNP allowing the causal gene to emerge.

Holmen et al[7] recently leveraged the last 2 approaches to gain insight into the mechanism of association at several loci related to blood lipids. In this work, investigators applied Illumina's HumanExome BeadChip array on 5771 white participants of the Norwegian Nord-Trondelag Health (HUNT) study to examine exonic variation in subjects with and without a history of myocardial infarction.[8] The chip allowed for the cost-effective examination of >80 000 exonic variants identified in previous large-scale exome sequencing projects with >80% of these having a minor allele frequency <5%.[7] The investigators used low-pass whole-genome sequencing in a small subset of samples to estimate that ≈70.9%, 77.4%, and 78.0% of rare, low-frequency, and common coding variants had been captured by the HumanExome BeadChip array.[7] The sequencing also confirmed the overall high quality of the genotype calls for the SNPs on the array irrespective of allele frequency.

A large fraction (127 of 157) of lead GWAS SNPs for blood lipids was also included on the array.[7] These SNPs yielded genome-wide significant associations with their respective lipid measures at 7 loci and nominal associations at 45 loci (both fractions much higher than expected by chance). Robust correlations were also observed for the direction and size of effects of these SNPs with those observed in larger GWAS studies. A subset of 51 453 missense and loss-of-function variants with ≥6 copies of the minor allele observed in the sample set were then carefully examined for association with lipid phenotypes, and 18 of these variants with $P<2\times10^{-5}$ and minor allele frequency <10% were taken forward for replication in 4666 additional Norwegian participants of the population-based Tromsø study. After a joint analysis of both sample sets, a total of 16 variants in 11 genes reached genome-wide significance. Most of these mapped to known lipid loci where there is no ambiguity of the causal gene. However, 2 mapped to genes (*RNF111* and *TM6SF2*) that have not been previously clearly implicated in blood lipid levels. Importantly, only 2 of these variants had a minor allele frequency <1%, and only 3 had a minor allele frequency <2%.

The genome-wide significant variants within the 9 established lipid loci revealed a range of mostly predictable relationships to the corresponding GWAS index SNPs.[7] At some loci, the exonic variant was found to be either identical or statistically indistinguishable to the GWAS index SNP (eg, *APOB* variant and LDL, *LPL* variant and triglycerides). At other loci, the exonic SNPs served as additional strong signals independent of the GWAS index SNPs (eg, *ABCG5/ABCG8* variants and LDL, *ANGPTL4* variant and triglycerides, *CETP/LIPC/LIPG*

variants and HDL). At yet another locus, the coding variants appeared to be a shadow of the association of the lead GWAS SNP (*APOA5* and triglycerides). The identification for the first time of an association between a rare coding variant in *LIPC* (p.Thr405Met, rs113298164) and HDL in a population sample was arguably the most interesting observation within the established lipid loci because this variant had previously only been linked to HDL levels in families with hepatic lipase deficiency. The second rare variant in *RNF111* initially suggested the discovery of a new locus for HDL, but further careful and thoughtful examination by the investigators of the allele distribution of this SNP in other populations combined with chromosomal-level conditional analyses indicated that this SNP was simply a population-specific long-range shadow of rs113298164 in *LIPC*.

A primary focus of the work presented in this article was the finding involving *TM6SF2*.[7] This gene falls within the gene-rich *NCAN-CILP2-PBX4* (or 19p13) cholesterol locus.[4,9] The p.Glu167Lys variant (rs58542926) in *TM6SF2* had the lowest *P* value of association with total cholesterol and also had the highest linkage disequilibrium with the GWAS index SNP (r^2=0.97). Consequently, the SNP was statistically indistinguishable to the GWAS index SNP, suggesting the latter may simply be a noncausal proxy of the causal coding SNP. The SNP with the next highest linkage disequilibrium with the GWAS index SNP was a missense SNP in *NCAN*, but its *P* value was substantially lower than that of rs58542926. *NCAN* was not considered a good candidate given that it is primarily expressed in the brain. Furthermore, the investigators observed very strong in silico replication for TC, LDL, and triglyceride associations with the TM6SF2 coding variant in ≈92 000 subjects genotyped with the Metabochip.[3] These findings provide compelling evidence that the causal gene in this region is *TM6SF2*.

The investigators then undertook functional studies to provide further evidence that *TM6SF2* is the likely causal gene in this locus.[7] First, they showed that this gene is expressed at both the mRNA and protein levels in the liver, linking its possible function to hepatic metabolism. They then performed gain- and loss-of-function studies in the mouse and correlated lipid levels as the phenotypic marker to gene expression. Tail vein injection of recombinant adenovirus was used to target expression of the human *TM6SF2* mRNA specifically to the liver and comparison made to a lacZ reporter gene expressing adenovirus. With this approach they were able to achieve a 2.4-fold increase in *TM6SF2* protein levels, and this increased expression was associated with increased levels for TC (2.3-fold), LDL (5.8-fold), and triglycerides (1.13-fold). The same approach was used to deliver adenovirus expressing short-hairpin RNAs targeting the endogenous mouse *Tm6SF2* gene. With this approach they were able to achieve a mean 49% decrease in TM6SF2 protein levels in the liver, which was associated with a significant 18.2% decrease in TC levels. These data showing that TC is directly regulated by *TM6SF2* expression, in conjunction with the observation that the human minor allele is associated with decreased *TM6SF2* expression, suggested to the authors that the substitution of a positively charged lysine residue at codon 167 for the major allele encoding glutamic acid (p.Glu167Lys) results in decreased function of TM6SF2, consistent with the probably damaging assessment by the Polyphene 2 algorithm.[10]

The investigators also highlighted the therapeutic potential of the *TM6SF2* locus by demonstrating that rs58542926 was associated with myocardial infarction within the same 2

cohorts in which it was associated with cholesterol.[7] This association was in the expected direction with the cholesterol-lowering allele also being associated with a decreased risk of myocardial infarction (odds ratio, 0.87; $P=5\times10^{-3}$). A similar association between the GWAS index SNP and the broader outcome of coronary artery disease was observed in the Coronary Artery Disease Genome-Wide Replication and Meta-Analysis (CARDIoGRAM) consortium meta-analysis involving >20 000 cases and >60 000 controls (odds ratio, 0.90; $P=2\times10^{-4}$). These findings in combination with the functional studies imply that a drug developed to block the biological effects of *TM6SF2* will not only lower cholesterol levels but also protect against the development of clinical coronary artery disease. Unfortunately, the authors did not investigate whether perturbation of *Tm6sf2* expression in mice in the setting of hyperlipidemia affects the development of atherosclerosis. *Tm6sf2*-knockout alleles in embryonic stem cells are available through the Knockout Mouse Project (www.komp.org). Such follow-up experiments would add significantly to considerations of therapeutic targeting and will hopefully be the subject of future studies by this or another interested group.

The investigative team should be highly commended for their systematic approach to discovery and thoughtful follow-up and interpretation of their findings. Surveying the remaining coding variants in their sample could rule out the small chance that the *TM6SF2* variant is actually a proxy of another coding variant in the region. Nevertheless, the study demonstrates the potential usefulness of examining all exonic SNPs in a sample for the identification of causal genes within gene-rich GWAS loci. In these regions, even intronic SNPs may not necessarily point to the causal gene given they may lie within regulatory regions of adjacent genes.[11] The ultimate yield of this approach remains to be determined and will depend on how many of these loci harbor ≥1 pathogenic and protective low-frequency coding variants. In this study of ≈10 000 subjects with ≈70% to 80% coverage of all coding variants, the data yielded novel insights on the causal gene in only 1 out of 127 lipid loci examined.[7] Somewhat paradoxically, this insight did not involve a low-frequency variant. These observations suggest challenges ahead when paired with the limited discoveries of large-scale exome sequencing projects to date for cardiometabolic disorders.[12] Hopefully, such challenges will be overcome, at least in part, with larger sample sizes. However, ambitious projects such as the Encyclopedia of DNA Elements, the Roadmap Epigenomics Project, and the Genotype-Tissue Expression Program should not be forgotten.[11,13,14] These and other studies are painstakingly annotating the regulatory regions of human genes in multiple human tissues. Surveying genetic variation within such regulatory regions and documenting its association with complex traits may ultimately be as productive and possibly more productive than exome sequencing in identifying causal genes.[15] Hopefully, for many loci, evidence from both approaches will converge and decrease the time it takes to confidently identify the culprit variation as well as increase the pace of translation to novel therapeutic interventions.

Acknowledgments

Sources of Funding

T.L. Assimes is supported by an NIH career development award K23DK088942. T. Quertermous is supported by a grant from the LeDucq Foundation and NIH grants R01HL103635, U01HL107388, R01HL109512, R21HL120757.

References

1. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014; 42:D1001–D1006. [PubMed: 24316577]
2. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012; 90:7–24. [PubMed: 22243964]
3. Willer CJ, Schmidt EM, Sengupta S, et al. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013; 45:1274–1283. [PubMed: 24097068]
4. Teslovich TM, Musunuru K, Smith AV, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010; 466:707–713. [PubMed: 20686565]
5. Twee-Hee Ong R, Wang X, Liu X, Teo YY. Efficiency of trans-ethnic genome-wide meta-analysis and fine-mapping. Eur J Hum Genet. 2012; 20:1300–1307. [PubMed: 22617345]
6. Wang Z, Liu X, Yang BZ, Gelernter J. The role and challenges of exome sequencing in studies of human diseases. Front Genet. 2013; 4:160. [PubMed: 24032039]
7. Holmen OL, Zhang H, Fan Y, et al. Systematic evaluation of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk. Nat Genet. 2014; 46:345–351. [PubMed: 24633158]
8. Grove ML, Yu B, Cochran BJ, et al. Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. PLoS One. 2013; 8:e68095. [PubMed: 23874508]
9. Kathiresan S, Melander O, Guiducci C, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet. 2008; 40:189–197. [PubMed: 18193044]
10. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7:248–249. [PubMed: 20354512]
11. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 2011; 9:e1001046. [PubMed: 21526222]
12. Peloso GM, Auer PL, Bis JC, et al. NHLBI GO Exome Sequencing Project. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. Am J Hum Genet. 2014; 94:223–232. [PubMed: 24507774]
13. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013; 45:580–585. [PubMed: 23715323]
14. Chadwick LH. The NIH Roadmap Epigenomics Program data resource. Epigenomics. 2012; 4:317–324. [PubMed: 22690667]
15. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: illuminating the dark road from association to function. Am J Hum Genet. 2013; 93:779–797. [PubMed: 24210251]