HHS Public Access

Author manuscript

Arterioscler Thromb Vasc Biol. Author manuscript; available in PMC 2016 October 01.

Published in final edited form as:

Arterioscler Thromb Vasc Biol. 2015 October ; 35(10): 2079–2080. doi:10.1161/ATVBAHA.115.306366.

From locus association to mechanism of gene causality: the devil is in the details

Clint L. Miller[#], **Milos Pjanic**[#], and **Thomas Quertermous**

Department of Medicine, Cardiovascular Research Institute, Stanford University School of Medicine, 300 Pasteur Drive, Stanford CA 94305

[#] These authors contributed equally to this work.

Arguably one of the greatest advances in biomedical research in the past decades has been the association of human allelic variation with complex human diseases and phenotypes. Certainly initially these findings were met with skepticism, given the underpowered and misleading number of candidate gene association studies that preceded the genome wide association (GWA) era. Also, there was considerable incredulity when it became clear that identified variation in reasonably powered genome wide association studies could not address the bulk of attributable genetic risk for the majority of diseases. However, scientists in academia and industry are now increasingly recognizing the importance of studying these loci for better understanding of disease pathways and developing new therapeutics. This is perhaps most significant for atherosclerotic coronary heart disease (CAD), the primary source of mortality and morbidity worldwide, for which no single drug has yet been developed to target the primary disease process in the vessel wall.

GWA studies have identified common variation throughout the genome that associates with specific diseases, but these single nucleotide polymorphisms (SNPs) provide little information regarding the mechanism for this association. Lead variants reported in GWA studies are commonly tag SNPs chosen to represent regions of linkage disequilibrium that can be hundreds of thousands of nucleotides in length. The next era of GWA studies will be focused on finding the causal variation in these loci, using this information to identify the causal gene, and then elucidating the mechanisms behind disease risk susceptibility. Deciphering precise molecular mechanisms will involve both state-of-the-art computational analysis and in vitro and in vivo experimental validation (Figure). Our future understanding of coronary artery disease and development of novel treatments for patients will ultimately depend on how we approach this daunting task.

In a paper published in the current volume of ATVB, Braenne and colleagues[1] have taken advantage of an extensive array of existing datasets to develop comprehensive annotation for 159 CAD loci[2]. Importantly, in this approach they have incorporated several layers of selection and filtering to prioritize candidate genes based on the identification of variants in linkage disequilibrium (LD) with lead SNPs that are located in coding region of exons,

Corresponding author: Thomas Quertermous MD, Division of Cardiovascular Medicine, Stanford University, 300 Pasteur Drive, Stanford, California 94305, tomq1@stanford.edu, Tel/Fax: (650) 723-5013/(650) 725-2178.

Disclosures: None

represent expression quantitative trait loci (eQTL), and reside in regulatory regions (epigenetic features of transcriptional activity). The three main criteria for the filtering were: (i) non-synonymous amino-acid change, (ii) eQTL effect and (iii) overlap with the regulatory region. From the initial 159 lead CAD SNPs only 33 were exonic (22 non-synonymous variants associated with the amino-acid change) while the majority of variants reside within regulatory elements or heterochromatic regions. Sixty-six CAD lead SNPs had eQTL associations with genes located less than one million base pairs (1Mb) away. These findings bolster previous studies suggesting that the principal mechanism for the majority of non-coding variants at CAD loci involves regulating local gene expression. Consistent with this hypothesis, promoter SNPs had up to three additional nearby genes as eQTLs. Moreover, this study identifies CAD SNPs as eQTLs for genes located up to 0.5Mb away. For instance the variant, rs2895811, is located in the intron of the *HHIPL1* gene but is instead associated with variance in *YY1* expression levels. They also note the complexity of deleterious protein-coding variants such as the lead SNP, rs867187, in the *PROCR* gene, which is in high LD with another deleterious variant in the *MYH7B* gene. While we have commonly annotated lead variants in relation to their nearest coding gene, this type of analysis highlights the structural complexity of the genome and the need for more systematic approaches that may disentangle the interacting regulatory architecture in regions of disease-associated loci (Figure).

While the majority of post-GWAS efforts have focused on transcriptional regulation of candidate regulatory variants, this study also emphasizes the importance of miRNA regulation of causal gene expression as another mechanism of CAD associations. Regulation of *TCF21* has been previously linked to miR-224 mediated interaction with the 3' UTR lead SNP[3], but this putative causal mechanism has not been systematically investigated. Here, the authors reveal that 55 CAD SNPs that reside in the 3' UTR of 33 genes are predicted to disrupt binding of 254 unique miRNA core binding sequences. Not surprisingly, they note that 23 of these miRNAs are predicted to target multiple CAD genes, and the miR-SNPs are also in high LD with promoter SNPs. These predicted interactions at the transcriptional and post-transcriptional level may explain causal regulatory mechanisms for multiple disease associations. Importantly, several CAD SNP-eQTL associations were highlighted as being likely due to disruption of miRNA binding misregulation. It will be critical to validate these associations with changes in the endogenous upstream transcription factors and miRNAs in the appropriate context.

By considering primarily eQTLs and non-synonymous amino acid changes the authors provisionally identify 151 candidate CAD genes from 159 SNPs, among which 98 represent genes not previously linked to the pathology of CAD. A literature-based approach to prioritization of SNPs from this list yielded only few genes with maximum scores, that have been extensively described in CAD-related publications, while 31% of CAD SNPs could be linked to CAD solely using a data-driven approach, e.g. the *REST* gene being among them, a strong phenotypic modulator of vascular smooth muscle cells.

The primary limitation of this study is the lack of additional datasets that support the identification of causal variants and point to new mechanisms of disease association. New approaches promise to provide insights into the native chromatin architecture in disease-

relevant cell-types, and the underlying *cis*-regulatory mechanisms of the associations. For instance, the recently developed Assay of Transposase Accessible Chromatin (ATAC-Seq) method provides access to critical information regarding locus anatomy and can be conducted on primary cultured individual cells or individual cells that can be harvested from human disease lesions by microdissection, tissue dissolution and single cell capture[4]. Analogous to eQTL approaches, allele-specific expression (ASE) data derived from limited numbers of primary cultured cells or residential lesion cells would be highly informative for identification of causal variants and causal genes[5]. These studies would require fewer numbers of individuals to detect significant expression changes using heterozygous exonic SNPs as a surrogate, and may unravel the dynamics of intercellular heterogeneity. These types of data would also provide much needed insights into the mechanisms by which causal variants located outside known transcription factor motifs or miRNA binding sites mediate changes in gene expression. Estimates from this study show that ~50% of CAD SNPs reside outside both the ENCODE regulatory elements and gene-coding regions and mechanisms behind these associations remain elusive.

The findings from the integrative approach reviewed here have honed the pool of candidate genes to be further validated and studied with in vitro and in vivo functional studies to investigate the mechanism of their association (Figure). Through continued development of multi-omic datasets from relevant cells and tissues, and painstaking identification of the causal genes and their biologically meaningful functions, we can significantly advance our understanding of disease-linked pathways to ultimately develop therapeutics targeted to the vessel wall. To paraphrase Admiral Hyman G. Rickover, “The devil is in the details, but so is the solution.”

Acknowledgements

Acknowledgements: None

Sources of funding: This work has been supported by NIH grants HL103635 (TQ), U01HL107388 (TQ), HL109512 (TQ), R21HL120757 (TQ), HL125912 (CLM) and a grant from the LeDucq Foundation.

References

1. Braenne I, Kanoni S, Willenborg C, et al. Prediction of causal candidate genes in coronary artery disease loci. ATVB. 2015 in press.
2. Deloukas P, Kanoni S, Willenborg C, et al. Nat Genet. 2012; 45:25–33. [PubMed: 23202125]
3. Miller CL, Haas U, Diaz R, Leeper NJ, Kundu RK, Patlolla B, Assimes TL, Kaiser FJ, Perisic L, Hedin U, Maegdefessel L, Schunkert H, Erdmann J, Quertermous T, Sczakiel G. Coronary heart disease-associated variation in TCF21 disrupts a mir-224 binding site and mirna-mediated regulation. PLoS Genet. 2014; 10:e1004263. [PubMed: 24676100]
4. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015; 523:486–490. [PubMed: 26083756]
5. Conde L, Bracci PM, Richardson R, Montgomery SB, Skibola CF. Integrating gwas and expression data for functional characterization of disease-associated snps: An application to follicular lymphoma. Am J Hum Genet. 2013; 92:126–130. [PubMed: 23246294]

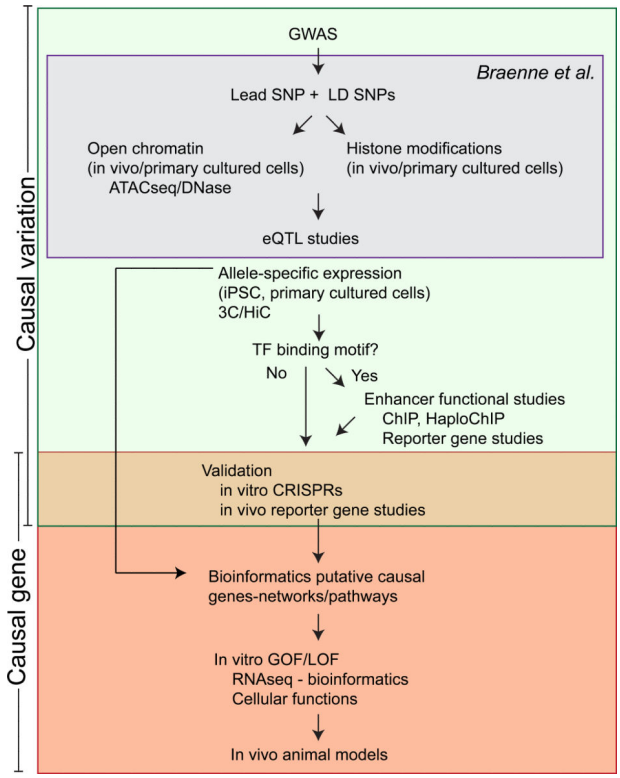

Figure.

Comprehensive experimental approach to identification and study of causal variation and causal genes in CAD GWA study loci. Initial studies investigate the causal variant which is usually required to identify the causal gene. The gray box highlights the analyses used by Braenne et al[1] to identify causal variants and genes. Abbreviations: DNase, DNase 1 hypersensitivity method to identify open chromatin; 3H, chromosomal conformation capture; HiC, high-throughput adaptation of 3C; ChIP, chromatin immunoprecipitation; HaploChIP, differential ChIP and comparison of binding on two alleles; CRISPRs, (clustered regularly interspaced short palindromic repeats), system for genome editing; GOF, gain of function; LOF, loss of function; RNA-Seq, high-throughput sequencing of RNA to identify differential gene expression.