HHS Public Access

Author manuscript

Circ Genom Precis Med. Author manuscript; available in PMC 2019 February 01.

Published in final edited form as:

Circ Genom Precis Med. 2018 February ; 11(2): .

From Genotype to Phenotype: A Primer on the Functional Follow-up of Genome-Wide Association Studies in Cardiovascular Disease

Jennie Lin, MD, MTR[1,2] and **Kiran Musunuru, MD, PhD, MPH**[3]

[1]Division of Nephrology and Hypertension, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL

[2]Feinberg Cardiovascular Research Institute, Northwestern University Feinberg School of Medicine, Chicago, IL

[3]Division of Cardiovascular Medicine, Department of Medicine, Cardiovascular Institute, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA

Abstract

Genome-wide association studies (GWASs) have implicated many human genomic loci in the development of complex traits. The loci identified by these studies are potentially involved in novel pathways that contribute to disease pathophysiology. However, eventual therapeutic targeting of these pathways relies on bridging the gap between genetic association and function, a task that first requires validation of causal genetic variants, casual genes, and directionality of effect. Executing this task requires basic knowledge of interpreting GWAS results and prioritizing candidates for further study, in addition to understanding the experimental methods available for evaluating candidate variants. Here we review the basic genetic principles of genome-wide association studies, the computational and experimental tools used for identifying causal variants and genes, and salient illustrative examples of how cardiovascular loci have undergone functional investigation.

Keywords

genomics; genetics; association studies; genetic variation; genetic techniques; gene regulation

I. GENOME-WIDE ASSOCIATION STUDIES

Most cardiovascular diseases are complex, reflecting both genetic and environmental influences.[1-3] Contrasting with classic monogenic diseases, DNA variants at numerous genomic loci contribute to the development of a complex disease. Typically, each of these

Correspondence: Dr. Jennie Lin, Northwestern University, Feinberg School of Medicine, 303 E. Superior Street, Lurie 10-109, Chicago, IL 60611, Tel: 312-503-1787, Fax: 312-503-0137, jennie.lin@northwestern.edu.

Journal Subject Terms: Functional Genomics; Genetics; Gene Expression and Regulation; Genetic, Association Studies; Genetically Altered and Transgenic Models

Disclosures: None.

variants exerts a small effect on disease pathogenesis and does not singlehandedly determine whether a person will develop disease.[4] Instead, the incidence of disease reflects the combined contributions of the *small effects of numerous DNA variants*, which typically do not follow standard Mendelian inheritance patterns within families (autosomal dominant, autosomal recessive, etc.) and thus require a large sample size for detection.[5]

As such, complex diseases are better addressed with population-based association studies, in which large cohorts of unrelated individuals are assessed for associations between DNA variants and diseases. The most commonly used variants for this purpose are single-nucleotide polymorphisms (SNPs), although other types, e.g., copy number variants (CNVs), can also be tested. Genome-wide association studies (GWASs) employ arrays that can directly genotype up to millions of variants throughout a person's genome in a single experiment.[5]

A. Understanding minor allele frequencies and linkage disequilibrium

In addition to requiring careful genotyping and phenotyping of study participants, GWASs use minor allele frequency (MAF) and linkage disequilibrium (LD) data to calculate association between variant and a disease or trait. Most SNPs will have two different alleles within a population: a more common major allele and a less common minor allele. Because MAFs are calculated by population, they can vary widely among groups of different ethnicities due to their contrasting evolutionary histories.[6]

While leveraging MAFs for DNA variants, GWASs take advantage of LD among SNPs to define genomic loci within which there are variants that directly contribute to the pathogenesis of a disease.[7] LD can be calculated for any pair of SNPs. For example, if two SNPs are on different chromosomes, they will have no degree of linkage since the chromosomes will segregate independently during the process of meiosis, effectively making their inheritance independent of one another. If two SNPs are far apart on the same chromosome, they will be separated by numerous recombination hotspots, where crossover events during meiosis act to eliminate any linkage between the inherited alleles of the two SNPs. Such recombination hotspots typically lie tens to hundreds of kilobases apart on a chromosome. If two SNPs are close together on the same chromosome such that there is no recombination hotspot between them, there will be a high degree of linkage between the inherited alleles of the two SNPs: in other words, these SNPs exist in a state of LD. If within a population the major allele of the first SNP is always inherited with the major allele of the second SNP, and the minor allele of the first SNP is always inherited with the minor allele of the second SNP, they are in a state of perfect LD.[8]

All of the SNPs that lie between two recombination hotspots on a chromosome will have some degree of LD. These SNPs are considered to lie within a single locus whose boundaries are defined by the hotspots. In the context of a GWAS, if any one SNP within a locus were found to be associated with a particular disease, one would expect that other SNPs in the locus would have some degree of association with the same disease. In contrast, SNPs in neighboring loci would not automatically be expected to be associated with the disease.

B. Basic Principles of GWAS

The overarching goal of a GWAS is to determine which genomic loci are associated with the disease or trait of interest. If the disease phenotype is dichotomous (present versus absent), a GWAS asks whether a SNP's allele frequency differs between a cohort of people with disease (cases) and a cohort of people without the disease (controls). This query is performed systematically for all common SNPs across the genome. Although for the vast majority of SNPs there will be no difference in allele frequency, in a successful GWAS there will be a number of SNPs for which the allele frequency differs between the two groups. These differences may be quite small, reflecting that each individual SNP has only a small effect on the incidence of disease. Nonetheless, if the GWAS is adequately powered, which may require study cohorts with up to hundreds of thousands of individuals, allele frequency differences can be reliably detected at a robust statistical significance threshold. This threshold is typically $P < 5 \times 10^{-8}$, which is derived from Bonferroni correction (0.05/1,000,000) accounting for the fact that one million SNPs are being independently tested for disease association in a GWAS. If the disease phenotype is a continuous quantitative trait such as blood lipid levels, the query for any given SNP asks whether there are statistically significant differences in the phenotype among the groups of people with the three different genotypes at the SNP (major/major, major/minor, minor/minor).

Importantly, a SNP found to have a statistically significant association with the disease is not necessarily the causal DNA variant, i.e., a variant that has a direct pathogenic or protective effect. The association only signifies that the SNP's locus harbors a causal variant or variants in LD with the SNP identified by the GWAS. Thus, the original SNP—variously called the lead SNP, index SNP, or tag SNP—serves as a signpost defining an interval in the genome for which one must do follow-up studies to identify the causal variant(s). A causal variant may be a coding variant that alters the amino acid sequence of a gene that influences the disease phenotype. However, more often than not, a causal variant is non-coding and influences a gene's function from a distance via regulation of the gene's expression.[9,10] While the nature of LD dictates that a causal variant itself must lie within the locus defined by flanking recombination hotspots, a causal gene may lie outside of the locus, as far as hundreds of kilobases away from a causal variant due to complex chromatin looping, etc.

GWASs have been very successful in identifying SNPs in genomic loci associated with various cardiovascular diseases, in some cases identifying more than 100 loci for a phenotype. Identifying causal variants and causal genes at these loci has proven to be an arduous task, and only in a few cases have they been definitely established. In the next section, we will discuss the various tools and methods that are being used by investigators to follow up the results of GWASs in the hope of discovering new disease biology and therapeutic targets.

II. TOOLS AND METHODS FOR FINDING CAUSAL VARIANTS AND GENES

Pinpointing the causal variant underlying an association signal can be challenging due to LD among neighboring SNPs in a GWAS locus, especially if the locus harbors hundreds of potential candidates in LD with the lead SNP. Complicating the issue is the possibility that within a locus multiple SNPs in strong LD may all have functional effects and constitute a

set of causal variants. Identifying the causal genes and understanding the mechanism by which the causal variant(s) affect disease phenotype through gene function may lend important insights into identifying druggable targets for therapeutic benefit.

A. Computational approaches

i. Fine mapping—In the early days of GWAS investigations, the available catalogs of human genetic variation, especially common SNPs, were incomplete. Upon identification of a lead SNP in a GWAS locus, a necessary next step was to perform fine mapping of the locus. This process first involved extensive resequencing of the locus in hundreds of individuals from the study population to discover and annotate all existing common DNA variants in the locus. All of the common variants, both known and newly discovered, would then be directly genotyped in the study population and assessed for association with the phenotype.

Ideally, fine mapping of a disease-associated locus would allow investigators to identify the causal variant as the single SNP with the strongest phenotypic association. In practice, instead of a single SNP, a group of SNPs in very strong LD harbor the strongest degree of association with the phenotype. Often the list of candidate causal SNPs within the LD block remains a manageable number, and each SNP can then be individually interrogated by some of the functional methods described below. In some cases, the list of candidate SNPs remains too long, and other approaches are needed to hone in on the causal variant(s).

Presently, efforts such as the 1000 Genomes Project have cataloged variants from enough genomes that virtually all common SNP variants and LD patterns in major ethnic populations have been identified, making fine mapping in the traditional sense no longer necessary in most situations.[11] Now it is straightforward to perform a combination of direct genotyping and imputation of 1000 Genomes data to identify the alleles of all common SNPs in the human genome simultaneously, followed by association analyses and immediate narrowing to the most likely causal variants.

ii. Trans-ethnic fine mapping—Because of their distinct evolutionary histories, different ethnic populations can display quite varied MAFs for the same SNPs. Furthermore, the locations of recombination hotspots can vary significantly among ethnic populations due to naturally occurring genetic variation affecting the proteins that determine the hotspots.[12] For example, loci defined by hotspots tend to be smaller in the genomes of individuals of African descent compared to the genomes of individuals of European descent. In combination, these phenomena can create quite distinct, ethnicity-specific patterns of LD among SNPs in an area of the genome: two SNPs that are in strong LD in one ethnic group may be in weak or no LD in another ethnic group. Although this could theoretically pose challenges in GWAS analyses, distinct LD patterns among populations can potentially be leveraged to narrow down a list of candidate causal variants in a GWAS locus. For example, fine mapping of a lead GWAS SNP in two different ethnic populations would yield two distinct sets of SNPs based on ethnicity-specific LD patterns (one with the strongest disease associations in Population A, the other with the strongest disease associations in Population B).[13] Assuming that the causal variant underlying a GWAS association in Population A is

also a causal variant in Population B—i.e., affects disease pathogenesis by the same mechanism in both populations—then in principle, the overlap subset of the two SNP sets should include the causal variant. Data from additional ethnic populations might narrow the SNP subset even further. Ideally, this would allow investigators to converge on a single overlap SNP that represents a prime candidate causal variant.

iii. Epigenomics data—Whereas a causal SNP in coding DNA, which most likely alters an amino acid in the downstream protein, is fairly straightforward to interpret, a causal SNP in non-coding DNA presents a challenge in determining how it might be influencing gene function. If it lies within a gene's promoter region, 5′ untranslated region, or 3′ untranslated region, it may affect the transcription, translation, or stability of the mRNA. If it lies near an exon/intron junction, it could alter splicing of the mRNA.[14,15] If it lies at a distance far away from any coding sequences, it may alter a noncoding RNA transcript such as a microRNA or a long noncoding RNA (lncRNA),[16] which in turn modulates the function of coding genes. However, the most plausible explanation is that the SNP affects a regulatory element, such as an enhancer or a repressor, upon which transcription factors bind and assemble in order to modulate the transcription of genes up to hundreds of kilobases away from the SNP.[9,17]

In addressing a list with up to hundreds of candidate causal variants in non-coding DNA, even after fine mapping, one approach is to assess whether any of the SNPs lie in regions with regulatory potential, as implied by the configuration of chromatin at the regions. A variety of methods have been developed to identify regions of "open" chromatin where transcription factors can access the DNA: micronuclease sequencing (MNase-seq),[18] DNase I hypersensitivity sequencing (DNase-seq),[19] formaldehyde-assisted isolation of regulatory elements and sequencing (FAIRE-seq),[20] and assay for transposase-accessible chromatin and sequencing (ATAC-seq).[21] More precise determinations of the types of regulatory activities that occur at these regions can be undertaken with chromatin immunopreciptation sequencing (ChIP-seq) experiments with antibodies that can distinguish different types of chromatin modifications. For example, active enhancers are often marked by histone H3 acetylated at lysine 27 (H3K27ac) and H3 monomethylated at lysine 4 (H3K4me1), whereas active promoters are often marked by H3K27ac and H3 trimethylated at lysine 4 (H3K4me3).[22] ChIP-seq can also be used to identify regions bound by specific transcription factors, which can provide clues as to which factors are involved in the regulatory activity of SNP sites.

All of these types of epigenomic data have been extensively cataloged in a variety of cell types through the efforts of consortia such as Encyclopedia of DNA Elements (ENCODE)[23] and the Roadmap Epigenomics Project,[24] both funded by the U.S. National Institutes of Health (NIH). The data is easily accessible to investigators and can provide useful insights to guide post-GWAS studies. It should be stressed that epigenomics data cannot directly establish causality for SNPs, nor does the lack of data for SNPs necessarily rule out their being causal; rather, epigenomics data provides a basis on which to prioritize certain SNPs for further functional studies.

iv. Expression quantitative trait loci—As previously discussed, transcriptional regulation of gene expression is thought to underlie the associations of many SNPs with diseases. An association-based method of assessing whether any particular SNP operates through this mechanism is termed expression quantitative trait locus (eQTL) mapping.[25] The genotype of the SNP is compared to the transcript levels of nearby genes (termed *cis* regulation) or distant genes on the same chromosome or other chromosomes (termed *trans* regulation). A statistically robust association suggests a causal relationship between a variant in strong LD with the SNP and expression of the gene. Importantly, this does not necessarily signify that a SNP that is causal for an eQTL vis-à-vis a particular gene is also causal for a disease phenotype interrogated in GWAS.[26] At a minimum, one would expect that the lead GWAS SNP should be in strong LD with the eQTL SNP. Even if this is the case, there are implicit assumptions that (1) the causal variant operates through transcriptional regulation, rather than some other mechanism and (2) the eQTL gene is a causal gene for the phenotype in question. Thus, the existence of a strong eQTL SNP in a GWAS locus needs further validation, requiring follow-up functional experimentation to validate the presumed causal mechanism. Additional conditional analyses and causal inference testing can help refine the eQTL signals to prioritize candidates for functional validation.[27-29]

One can perform global analyses for eQTLs if there is a combination of genome-wide genotyping data and genome-wide gene expression data [obtained with a platform such as RNA sequencing (RNA-seq)] in a sufficiently powered collection of samples of a tissue of interest. In principle, eQTLs can be mapped in any tissue type of interest. As a practical matter, some tissues are more difficult to collect than others. Whereas it is straightforward to amass a large collection of blood samples and perform eQTL studies with peripheral blood cells, tissues relevant to cardiovascular diseases such as myocardium, vascular tissues from specific vascular beds, etc., need to be collected either during surgical procedures or post-mortem. It is only ethically permissible for invasive procedures to be performed if medical indications exist, meaning that patient tissue donors have medical co-morbidities that may confound the eQTL studies. With post-mortem tissues, the collection needs to occur as soon after death as possible in order to preserve their physiological properties. The NIH-funded Genotype-Tissue Expression (GTEx) project is in the process of collecting a large variety of different tissues from hundreds of post-mortem donors, with the intent of generating a global eQTL database as a resource for the scientific community.[30]

An alternative approach is to use induced pluripotent stem cells (iPSCs) from a diverse group of individuals who represent a broad distribution of genotypes. The National Heart, Lung, and Blood Institute (NHLBI) funded an $80 million effort to generate a variety of cohorts of iPSC lines, each numbering in the dozens to hundreds.[31] Several of these cohorts have been used for differentiation into cell types such as cardiomyocytes, vascular endothelial cells, and hepatocytes in order to perform unbiased eQTL studies.[26,32,33] Notably, stem cell-based studies have discovered eQTLs not identified in the corresponding primary tissue cohorts of the GTEx project,[26] suggesting that the two types of eQTL studies will prove to have complementary value. Of note, but beyond the scope of this review, iPSCs have their own set of limitations, including their relatively immature phenotype that may

affect transcriptional profiles, heterogeneity in differentiation processes, and donor-specific epigenetic signatures that may persist depending on reprogramming protocols.[34-37]

v. Other types of studies to link variants to genes—Noncoding variants can act through mechanisms outside of regulating gene transcription. One such mechanism is altering messenger RNA splicing patterns. With RNA-seq data paired with genotype data, it is feasible to perform splicing quantitative trait locus (sQTL) mapping to assess for associations between SNPs and differential alternative splicing of transcripts from nearby genes.[38] Similarly, the ability to assess for methylation status of cytosines in CpG dinucleotide sites throughout the genome (a common form of epigenetic alteration that generally silences gene expression) allows for methylation quantitative trait locus (meQTL) mapping, i.e., association analyses between SNP genotypes and methylation of CpG sites within nearby genes.[39]

B. Experimental approaches

Computational analyses provide important information on which SNPs within a trait-associated LD block could be causal variants. However, as mentioned previously, these analyses do not provide definitive proof of causality but rather bring forward strong candidates to undergo functional validation at the bench. In addition, cellular and *in vivo* experiments would be needed to prove that the gene(s) regulated by a causal variant do indeed lead to disease-relevant phenotypes if perturbed.

i. Reporter assays—A mainstay of the study of putative transcriptional regulatory elements, such as those implicated by eQTL studies, is the reporter assay. Typically a DNA region of interest will be subcloned either upstream or downstream of a promoter-reporter gene cassette in a plasmid. The plasmid is then transfected into a cell type of interest, e.g., cultured hepatoma cells if the putative regulatory element is believed to be active in hepatocytes. Activity of the protein product of the reporter gene is then measured either from the cells themselves or from the media in which the cells have been grown, depending on whether the protein is intracellular or secreted. Commonly used reporters include luciferase, for which routine and scalable assays are available, and fluorescent proteins such as GFP, which allows for measurement by flow cytometry or imaging-based techniques.

Reporter assays allow for direct comparison of versions of a DNA region that differ only with respect to a particular SNP or SNPs, thus assaying for a gene expression effect modulated by the SNPs. While straightforward and relatively quick to perform, reporter assays do have a substantial limitation when used to interrogate candidate causal variants, in that they test for regulatory activity of DNA sequences removed from their endogenous genomic and epigenomic contexts. If the complete regulatory element stretches across kilobases, subcloning of a few hundred basepairs around a SNP might not allow for accurate modeling of the SNP's effect on regulatory activity. Furthermore, reporter plasmids transfected into cells for short-term experiments lack the complete chromatin structure of endogenous genomic DNA. Thus, reporter assays are unable to capture the effects of short-range or long-range chromatin interactions that can significantly impact gene expression.

ii. Massively parallel reporter assays—Although traditional individual reporter gene experiments such as luciferase assays can be used to test whether SNPs modulate gene transcription, testing a large number of SNPs would require a more high-throughput, scalable, and efficient platform. Massively parallel reporter assays (MPRAs) were developed as a means to interrogate the regulatory activity of hundreds or even thousands of sequences simultaneously.[40,41] Thus, an MPRA can be readily applied to screen a large number of candidate causal variants identified in GWASs, assuming that the causal variants act by modulating the expression of nearby genes.[26,42,43]

The MPRA approach uses an array to synthesize a large number of oligonucleotides in parallel, followed by cleavage of the oligonucleotides off of the array to form a pool in solution. Each oligonucleotide contains a short genomic DNA sequence of interest coupled to a unique barcode tag. The oligonucleotides are subcloned into a fixed reporter plasmid backbone with the genomic DNA sequences placed in a position to influence reporter gene expression while the barcode tags are positioned to be transcribed with the reporter gene (e.g., in the 3′ untranslated region).[41] Once generated, the pool of reporter plasmids is introduced into cultured cells or into a tissue *in vivo* in an animal model. After incubation for a sufficient time to allow the reporters to be fully expressed, high-throughput RNA sequencing and counting of the barcode tags allows for the calculation of the relative quantities of reporter transcripts. This facilitates the determination of the relative regulatory activities of all of the genomic DNA sequences in a single experiment. Barcode tags that are enriched in the final RNA transcript pool relative to the original DNA plasmid pool identify sequences with enhancer activity, and barcode tags that are depleted identify sequences with repressor activity.[41]

To test specific variants, short DNA sequences containing the major and minor alleles of candidate SNPs would be paired with unique barcode tags and subcloned into a reporter plasmid pool, with the goal of identifying SNPs in which allelic variation changes regulatory activity, e.g., the major allele version of a sequence has enhancer activity while the minor allele version has no activity. Ideally, this experiment would identify one or a few SNPs that stand out from the others. While this approach would not unequivocally establish causality, it would serve to prioritize SNPs for further functional testing.

As with individual reporter assay experiments, the MPRA design has the shortcoming that it tests for regulatory activity of DNA sequences removed from their endogenous genomic contexts. The new set of technologies collectively known as genome editing provides a means to test SNPs for regulatory activity directly within the genome.

iii. CRISPR-Cas9 genome editing—In recent years, various genome-editing tools have come into widespread use, including zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and clustered regularly interspaced short palindromic repeats (CRISPR)-CRISPR-associated protein 9 (Cas9).[44] CRISPR-Cas9 systems have become particularly popular because of their efficacy and ease of use compared to other tools, so here we will focus on the CRISPR-Cas9 approach.

Overview of CRISPR-Cas9 technology: CRISPR-Cas9 systems, which comprise both protein and RNA components, are based on naturally occurring bacterial adaptive immune systems that have evolved to destroy foreign genomic material. As an endonuclease, the Cas9 protein can bind specific DNA and RNA sequences and catalyze a double-strand break (DSB) in DNA. In the simplified *Streptococcus pyogenes* CRISPR-Cas9 system, the RNA component approximately 100 nucleotides in length is termed the guide RNA. Cas9 binds to this guide RNA, the first 20 nucleotides of which can hybridize to one strand of double-strand DNA. This 20-nucleotide sequence is termed the protospacer. Cas9 also binds to several adjacent DNA nucleotides (termed the protospacer-adjacent motif, or PAM). Thus, a triple complex of protein, RNA, and DNA is formed. By changing the protospacer sequence in the guide RNA, one can redirect the protein-RNA complex to bind a different sequence and thus target a different site in the genome. This feature of CRISPR-Cas9 technology allows for ease of targeting multiple different genomic sequences without complicated engineering. Once bound, the complex will generate a DSB in the DNA.

The DSB activates the cell's natural DNA repair machinery, which operates in one of two ways. The default repair pathway is non-homologous end-joining (NHEJ), in which the free ends from the DSB are rejoined. NHEJ is an error-prone process that occasionally results in the insertion or deletion of basepairs, collectively termed indels. If the DSB is directed to the coding sequence of a gene, NHEJ can introduce frameshift mutations, thereby knocking out the gene. If the DSB occurs within a non-coding regulatory region, NHEJ can disrupt the regulatory element with potential functional consequences. Indel sizes resulting from NHEJ of a single DSB can range from one basepair to thousands of basepairs (though smaller indels are more frequent) and often are difficult to predict. This can be circumvented through the introduction of two nearby DSBs on the same chromosome, which will often result in a clean deletion bordered by the DSBs, lending some measure of control to NHEJ.

The other method by which the cell can repair a DSB is homology-directed repair (HDR). Via homologous recombination, a repair template allows the area of the DSB to be repaired with a high level of accuracy. Ordinarily, the cell will use a sister chromatid as the repair template. If a custom-made DNA template with homology arms flanking the desired alteration is introduced into the cells, HDR can use this template and stably incorporate the alteration into the genome. HDR operates in proliferating cells, limited to the S and G2 phases, during which the DNA content of the cell is double the usual amount, due to each chromosome having two sets of chromatids as a result of DNA replication in S phase. An important consequence of this limitation of HDR is that it occurs less frequently than NHEJ in proliferating cells and not at all in non-proliferating cells. Because NHEJ occurs in all cells during all cell cycle phases, it is more feasible to knock out a gene or regulatory element than it is to cleanly knock in a specific desired alteration, such as a variant SNP allele, into the genome.

Application of CRISPR-Cas9 to validate variants as causal: CRISPR-Cas9 provides ways in which to functionally test whether a variant is causal for a change in gene function. The most rigorous approach is to use CRISPR-Cas9 to knock in alternate SNP alleles into cellular or animal models, thereby generating wild-type and variant models that are isogenic, i.e., of matched genetic background. Since the only difference between the models is with

respect to the SNP, in principle any phenotypic differences observed between the models can be attributed to the SNP.[45] An important limitation of this approach is that HDR-mediated knock-in of alternate SNP alleles can be very inefficient, and so it might not be straightforward to generate the desired models. Other potential limitations of CRISPR-Cas9 technology in introducing variants are reviewed elsewhere.[44] In addition, knocking in a human SNP into a non-human genetic background may affect study results due to confounding genetic factors. Nonetheless, if the models can be generated, especially as a complete allelic series (e.g., major/major, major/minor, and minor/minor), they would constitute the gold standard for testing whether a variant is causal for a phenotype.

An alternative approach that takes advantage of the increased efficiency of NHEJ-mediated disruption compared to HDR is to knock out the SNP site. This can be done in an imprecise way—introduction of an indel of semi-random size into the SNP site using a single guide RNA, which will likely have the effect of disrupting either gene function (if the variant is coding) or a regulatory element (if the variant is non-coding) regardless of the exact size of the indel.[44] This can also be done in a more precise way—deletion of a small sequence around the SNP site using two guide RNAs that generate flanking DSBs.[26] Either strategy will generate isogenic models that can be compared for phenotypic differences. Importantly, any such differences cannot be directly attributed to variation of the SNP itself; rather they can only be attributed to the sequence containing the SNP allele. In other words, the experiment would not necessarily prove that SNP variation affects gene regulatory activity, but rather that the sequence containing the SNP site has gene regulatory activity. Nonetheless, a well-designed series of experiments can prove the former. For example, two different cell lines that have alternate genotypes of a SNP (e.g., homozygous major and homozygous minor) can be engineered with the same dual-guide-RNA SNP sequence deletions. If the homozygous major cell line shows a phenotype upon SNP sequence deletion, but the homozygous minor cell line does not, then it would argue that SNP variation is in fact causal for the phenotype and that the SNP major allele-bearing sequence has gene regulatory activity.

Application of CRISPR-Cas9 to validate genes as causal: To prove that a gene implicated for a disease or trait by GWAS is indeed causal, one must perform functional studies for that gene. CRISPR-Cas9 technology can be harnessed to generate knockout cell lines and animal models through NHEJ-mediated indel formation. In the context of GWAS functional follow-up, knockout models are useful to establish (1) whether a coding variant is a loss-of-function or gain-of function mutation and (2) whether a gene implicated by eQTL analysis is causal for the human phenotype. For cell lines, CRISPR-Cas9 technology facilitates permanent and efficient knockout of genes of interest, with fewer off-target effects than knockdown through traditional RNA interference (e.g., siRNA), which tends to be less efficient in reducing gene expression. For animal models, the CRISPR-Cas9 approach can knock out a gene without the limitations presented by the insertion of an antibiotic resistance cassette into the genome for positive selection, which includes possible production of aberrant mRNA and truncated protein products that might exert unanticipated effects. In addition, knockout models can be produced relatively quickly using CRISPR-Cas9, often within weeks rather than months to years.

Establishing through experimental evidence whether a candidate gene affects phenotype can accelerate understanding of the mechanistic underpinnings of a GWAS trait or disease as well as identify new druggable targets. The functional follow-up of a lipid GWAS candidate variant, discussed in section III, highlights the importance of correctly identifying the causal gene.

iv. CRISPR interference and activation—CRISPR-Cas9 technology can extend beyond genome editing; recently it has been adapted to study transcriptional regulation. The Cas9 protein can be altered so that it is catalytically dead (i.e., cannot generate a double-strand break in the DNA) while still forming a protein-RNA-DNA complex with specifically engineered guide RNAs. This so-called dCas9 protein can serve as a customizable, sequence-specific DNA-binding domain to which other functional domains can be attached. For example, addition of a transcriptional activator domain leads to increased expression of a target gene when the complex is directed by a guide RNA to bind at regulatory regions, a phenomenon known as CRISPR activation.[46] Addition of a transcriptional repressor domain can suppress gene expression and thus leads to CRISPR interference (CRISPRi), by analogy to RNA interference.[47] Typically CRISPR activation and CRISPRi are transient because the Cas9 protein does not make permanent alterations to the genomic target sequence.

CRISPRi can be employed to assess whether sequences containing SNPs have gene regulatory activity.[26] The simple act of positioning dCas9, with or without additional domains, at the site of a SNP can inhibit regulatory activity via steric interference, i.e., by competing with endogenous regulatory factors for binding to the site. For example, if dCas9 is positioned at the site of a SNP variant that has enhancer activity, it should reduce expression of the gene regulated by the enhancer. An advantage of this approach compared to standard genome editing is that transient transfection of the CRISPRi reagents (protein and guide RNA) is sufficient to produce the interference effect and, potentially, the desired phenotypic consequence. A CRISPRi experiment can be completed within a few days (versus the weeks to months required to generate pure genome-edited cell lines).

As with SNP sequence deletion, the CRISPRi approach does not directly interrogate whether SNP variation has phenotypic consequences; rather, it assesses whether the sequence bearing the SNP allele has gene regulatory activity. However, testing of CRISPRi in cell lines with alternate genotypes at the SNP and finding differential effects of dCas9 on a phenotype can serve as a means to assess whether a variant is causal without having to undertake the laborious process of genome editing to generate an allelic series of isogenic cell lines. Even if testing of cell lines with alternate genotypes is not practical (e.g., it may not be easy to find a cell line that is homozygous minor for SNPs in a GWAS locus, since by definition it is the rarest genotype), CRISPRi provides a means by which one can potentially screen through a large number of SNPs in a locus to obtain evidence for regulatory activity at a subset of the SNPs and thus prioritize those SNPs for further study.

v. Massively parallel gene expression assays—As described above, an important limitation of standard reporter assays and massively parallel reporter assays is that they interrogate candidate regulatory DNA sequences taken out of their native genomic context. CRISPR-Cas9 offers alternative methodologies that can test candidate regulatory DNA

sequences throughout a locus of interest, within their endogenous context, in a massively parallel fashion.[48-50] In this approach, it is necessary to measure the expression or activity of a gene product in a highly quantitative fashion in a large number of cells at once, and then to cleanly separate higher-expressing cells from lower-expressing cells. In some cases, the nature of the protein product (e.g., a protein with an extracellular domain for which there exists a sensitive antibody) allows for straightforward quantitative measurement and separation using a standard technique such as fluorescence-activated cell sorting (FACS). In other cases, the target gene needs to be endogenously tagged with a reporter, such as a fluorescent protein, which serves as a proxy for the expression level of that gene and is amenable to a technique like FACS. A reporter can be inserted into the endogenous locus using CRISPR-Cas9 genome editing.[51]

Once a suitable cell line is available, Cas9 along with a guide RNA library that targets a set of candidate SNP sequences or regulatory elements is introduced into the cell line. The intent is that on average each cell receives one guide RNA, which potentially disrupts its individual target sequence in the cell's genome via NHEJ. If the target sequence harbors enhancer activity with respect to the gene of interest, the guide RNA should reduce the gene's expression; conversely, if the target sequence harbors repressor activity, the gene's expression will be increased. The entire pool of cells is subjected to a technique like FACS to isolate the highest gene/reporter-expressing cells and/or the lowest gene/reporter-expressing cells, and the guide RNAs within the isolated cells are subjected to deep sequencing and counted.[50] Guide RNAs that are enriched or depleted in the isolated cells relative to the original library will directly point to regulatory elements in the genomic DNA. By plotting out the degree of enrichment or depletion of each guide RNA, one can map the transcriptional regulatory landscape of the locus with respect to the gene of interest. In principle, CRISPRi could be used as an alternative approach to generate a map of the transcriptional regulatory landscape. Whether genome editing or CRISPRi is used, it would be important to use cell lines with alternative SNP genotypes within the locus of interest in order to comprehensively determine which alleles confer regulatory activity.

Although the massively parallel gene expression assay strategy has not yet been applied to a specific case of screening through a list of candidate causal variants in a GWAS locus, it has been used to perform saturation mutagenesis of entire loci in order to more broadly define all of the endogenous regulatory elements within the loci.[40-42,52]

vi. Electrophoretic mobility shift assays and chromatin immunoprecipitation— Upon validation of a causal DNA variant, particularly those that appear to be involved in gene regulation, a natural question is which factor (e.g., transcription factor) binds at the site of the variant. While computational methods can *predict* which factors bind to any given genomic sequence, these methods simply generate hypotheses that then must be tested by functional experiments. A common approach to assessing whether a factor binds a sequence is the electrophoretic mobility shift assay (EMSA), more colloquially known as gel shift assays, in which a labeled (whether radioactive or non-radioactive), synthetic version of the sequence is mixed and incubated with lysate from a relevant cell type. The mix is then subjected to electrophoresis on a gel, which separates DNA-protein complexes from unbound DNA. To prove that any observed DNA-protein complex involves a given candidate

protein, a mix of sequence, lysate, and protein-specific antibody is also subjected to electrophoresis, with the expectation that the DNA-protein complex will either be inhibited by the antibody or, possibly, “supershifted” by the antibody (i.e., a triple complex will form and migrate differently on the gel).[53]

An important limitation of EMSAs is that they are *in vitro* experiments that may not faithfully reflect what happens with endogenous DNA in the nucleus of a cell. Chromatin immunoprecipitation (ChIP) entails crosslinking of proteins and DNA within the nucleus. This is followed by isolation of the candidate protein with a specific antibody, retrieval of any DNA sequences crosslinked to the purified protein, and sequencing of the DNA (either in an unbiased fashion or with a specific sequence in mind). To establish that a DNA variant affects the binding of a candidate factor, one should demonstrate with ChIP that the factor is preferentially binding to a sequence with one allele compared to the alternate allele.[54]

A complementary approach to establishing a functional relationship between a regulatory factor and a DNA sequence harboring a causal variant is to overexpress or knock down the activity of the factor and demonstrate an effect on the target gene. Even more compelling is a demonstration that modulation of the factor affects the target gene differently depending on which genotype of the causal variant is present in the cell line tested.

III. EXAMPLES OF FUNCTIONAL FOLLOW-UP FOR CORONARY ARTERY DISEASE

Coronary artery disease (CAD) has the distinction of being one of the very first phenotypes to which the GWAS methodology was applied. Three independent GWASs for CAD in individuals of European descent were published simultaneously in 2007, and each study reported the same novel locus on chromosome 9p21 as being by far the most strongly associated with the disease.[55-57] One of the studies also reported several other novel loci, including a locus on chromosome 1p13, which harbors the *SORT1* gene.[55] In the years since, increasingly large GWAS meta-analyses have been performed for individuals of European descent, and smaller studies have been performed in several other ethnic populations. As of the time of this writing, at least 95 loci with SNPs meeting the statistical significance threshold of $P < 5 \times 10^{-8}$ for association with CAD have been reported.[58] About a quarter of these loci harbor genes involved in the regulation of blood lipid levels, and a handful harbor genes involved in the regulation of blood pressure. For the most part, the remaining loci do not appear to be linked in any way to traditional risk factors for CAD, suggesting novel mechanisms contributing to the pathogenesis of the disease.

The *SORT1* and chromosome 9p21 loci represent two examples for which extensive follow-up functional genomic studies have been undertaken with varying degrees of success, so we have chosen to use these examples to illustrate the application of the various tools and methods described in the previous section.

A. The *SORT1* locus

SNPs in the *SORT1* locus were found not only to be strongly associated with CAD but also to be strongly associated with low-density lipoprotein cholesterol (LDL-C) in separate

GWASs of blood lipid levels.[59-61] Indeed, they were the most strongly associated with LDL-C out of all of the tested SNPs in the human genome. Alternative genotypes of lead SNPs in the locus are associated with up to a 10%–15% difference in LDL-C and up to a 30%–40% change in CAD risk. Because of the broadly accepted evidence that LDL-C is a causal risk factor for CAD, it is natural to conclude that the *SORT1* locus influences CAD risk through the modulation of blood LDL-C levels. There are several genes within 100 kb of the lead SNPs; at the time of the discovery of the locus, there was little prior evidence to suggest a biological link between any of the genes and lipoprotein metabolism.

A critical clue to the determination of both the causal variant and the causal gene in the locus was provided by eQTL studies.[10,60,62] Several eQTL studies documented a robust relationship between the genotype of a lead SNP in the locus and the expression level of three of the genes in the locus—*SORT1*, *PSRC1*, and *CELSR2*—in human liver. Liver samples from individuals homozygous for the minor allele of the lead SNP displayed several-fold higher expression of the genes than liver samples from individuals homozygous for the major allele. Notably, minimal or no eQTL relationships with these genes were observed in other tissues such as adipose and blood, suggesting a liver-specific regulatory effect of the causal SNP on gene expression.

Fine mapping of the locus in individuals of European descent narrowed the list of candidate causal variants to six SNPs, which were essentially in perfect LD.[10] The six SNPs were in close proximity, contained within a non-coding DNA segment about 6 kb in length. Because it had been hypothesized that the causal variant altered a transcriptional regulatory element, the entire 6-kb segment (harboring either the major alleles of six candidate SNPs or the minor alleles of the SNPs) was subcloned in luciferase reporter constructs, which were transfected into hepatocyte-like cultured human hepatoma cells. The minor allele version of the construct expressed substantially higher luciferase than the major allele version of the construct, consistent with the human liver eQTL findings. Alteration of the alleles of each of the six candidate SNPs one-by-one in the reporters constructs established that one SNP, rs12740374, was responsible for the difference in luciferase expression. Remarkably, fine mapping of the locus in African American individuals converged on a single SNP having the strongest association with LDL-C in that population, and that SNP too was rs12740374. The concordance of the two independent lines of evidence strongly supported rs12740374 as the single causal variant responsible for the locus' associations with LDL-C and CAD.[10]

The mechanism by which rs12740374 influences gene expression came to light with the recognition that the rs12740374 minor allele sequence closely matches the consensus binding sequence of the C/EBP family of transcription factors, whereas the major allele disrupts the sequence.[10] EMSA experiments demonstrate preferential binding of a factor in hepatocyte lysates to the minor allele sequence compared to the major allele sequence, with supershifting in the presence of C/EBP antibodies. ChIP-PCR experiments with cells homozygous for the minor allele documented binding of C/EBP with the endogenous SNP site. Finally, knockdown of C/EBP activity in hepatoma cells reduced luciferase expression from the minor allele-bearing reporter construct, but not from the major allele-bearing reporter construct; similarly, knockdown of C/EBP activity reduced endogenous *SORT1*

expression in hepatoma cells heterozygous at rs12740374, but not in hepatoma cells homozygous for the major allele.

The determination of the causal gene in the locus, as well as the mechanism by which the causal gene affects blood LDL-C levels and CAD risk, primarily employed two different model systems: mice and human pluripotent stem cells. *SORT1* stood out as a candidate gene because it had the strongest eQTL relationship with rs12740374 and because its protein product, sortilin, a transmembrane protein responsible for shuttling other proteins to different cellular compartments, had biological plausibility. Of note, the promoter of *SORT1* lies 120 kb away from rs12740374. AAV-mediated overexpression of the murine *Sort1* gene in mouse liver decreased blood LDL-C levels, whereas siRNA-mediated knockout of endogenous *Sort1* expression in mouse liver increased blood LDL-C levels.[10] Modulation of other genes in the locus in mouse liver did not alter blood lipid levels, suggesting *SORT1* to be the single causal gene. Further experiments in mice, primary mouse hepatocytes, and cultured hepatoma cells established two models by which *SORT1* influences lipoprotein metabolism in hepatocytes: (1) sortilin binds directly to apolipoprotein B (apoB) in nascent very-low-density lipoprotein (VLDL) particles in the Golgi apparatus and shunts the particles to the endolysosomal compartment for degradation, which reduces the number of VLDL particles secreted into the bloodstream; and (2) sortilin binds to apoB in mature LDL particles in the bloodstream and facilitates endocytosis of the particles into the cells through a mechanism independent of the LDL receptor.[10,63] Either mechanism should act to reduce blood LDL-C levels.

Confirming the role of *SORT1* in lipoprotein metabolism in human hepatocytes, the gene was knocked out in human pluripotent stem cells with the use of TALENs, followed by differentiation into hepatocytes.[64] Compared to wild-type hepatocytes, *SORT1* knockout hepatocytes displayed a substantial reduction in secreted apoB mass. Lentiviral reconstitution of *SORT1* expression in the knockout hepatocytes restored the secreted apoB mass to wild-type levels. The experiments were independently performed in two human pluripotent stem cell lines with different genetic backgrounds and had highly consistent results.

Together, these observations establish an overarching model in which a single nucleotide change influences clinically important phenotypes: rs12740374 major alleles result in low baseline hepatic *SORT1* expression, whereas minor alleles enable C/EBP-mediated transactivation of *SORT1* expression, decreased hepatocyte VLDL particle secretion and increased LDL particle uptake, and thus decreased blood LDL-C levels and reduced risk of CAD.

B. The chromosome 9p21 locus

Determining the mechanism by which the chromosome 9p21 locus influences CAD has been challenging because the locus is a so-called “gene desert.” The minimal locus (as defined by recombination hotspots) is ~58 kilobases in individuals of European descent and harbors no coding genes, leaving it unclear how the causal DNA variant(s) in the locus influence disease. Within this locus, a lncRNA called *ANRIL*, whose function remains uncertain, is transcribed;[65] this lncRNA is not conserved across species. The nearest coding

genes, *CDKN2A* and *CDKN2B*, encode cyclin-dependent kinase inhibitors, which are involved in cell-cycle regulation. These genes lie a long distance away from the lead GWAS SNPs in the chromosome 9p21 locus. The lead SNPs are not associated with any of the traditional risk factors for CAD, although they are associated with a variety of vascular phenotypes such as abdominal aortic aneurysm, intracranial aneurysm, peripheral arterial disease, and platelet reactivity.[57,66]

Given the strength of association between the lead 9p21 SNPs and CAD, multiple groups have attempted to elucidate the mechanisms by which 9p21 contributes to CAD risk. Long-range chromatin capture experiments have shown that enhancers within the locus interact with chromatin near *CDKN2A* and *CDKN2B* as well as *MTAP*, with possible modulation of transcriptional activity by the interferon family of proteins.[67] Although the role of interferon signaling in trascriptional regulation of this region remains unclear given conflicting results from other studies,[68,69] linked 9p21 SNPs rs10811656 and rs4977757 were shown through EMSAs and luciferase reporter studies to disrupt TEAD transcription factor binding, while knockdown of the factors decreased *CDKN2A* expression in cultured human vascular smooth muscle cells heterozygous for CAD risk alleles.[70] In a separate study, a different 9p21 SNP, rs1537373, was found to be a significant *cis*-eQTL for *CDKN2B*,[71] which further implicates the *CDKN2A/B* family in 9p21 biology.

Complementing these data, other groups have shown that *CDKN2B* is involved in atherogenesis. For example, *Cdnk2b* deletion in mice with an *Apoe* knockout background led to the development of larger necrotic core lesions in atherosclerotic plaques compared to those seen in *Apoe* knockout mice with wild-type *Cdnk2b*.[72] *Cdnk2b* deletion in mice also disrupts vascular repair mechanisms,[73] with implications for vascular biology that may extend beyond atherogenesis. However, mice in which a non-coding region of 70 kilobases corresponding to the human 9p21 locus was deleted showed changes in *Cdkn2b* and *Cdkn2a* expression but no difference in atherosclerosis, suggesting that the causal 9p21 mechanism in humans might not be conserved in mice.[74]

Based on computational analyses, 9p21 has more than one candidate causal player in disease pathophysiology requiring functional follow-up.[65,71,75] For example, outside of protein-coding genes, the CAD risk association of 9p21 may involve the expression of the lncRNA *ANRIL*, which has primate-specific Alu repeats that have been shown to regulate *ANRIL* target genes in *trans*.[65] Interestingly, this lncRNA can exist in linear form and in circular form, with the circular form expressed in whole blood and peripheral blood mononuclear cells associated with the protective 9p21 haplotype.[16] When CRISPR-Cas9 was used to delete the majority of *ANRIL* exons in cultured HEK293 cells, decreased apoptosis and increased cellular proliferation were seen; these effects were partially reversed with reconstitution of the cells with circular *ANRIL*, which induces p53 activation by preventing ribosomal RNA maturation.[16] The effects of *ANRIL* were also found to be independent of *CDKN2A* and *CDKN2B*, suggesting that 9p21 modulates CAD risk through more than one mechanism.

Although the 9p21 locus still requires further work in terms of identifying all causal players in its CAD risk association, the work performed so far on the *CDKN2A/B* family and

ANRIL has provided functional validation that this complex locus is involved in biological processes relevant to atherogenesis.

V. CONCLUSIONS

In the age of precision medicine, finding the biological link between genotype and disease phenotype will yield novel insights for the development of novel therapeutics. As more human genetic data become available and refined through GWASs, exome sequencing studies, and rare variant studies, it will become increasingly important for investigators to understand these studies and to develop and execute pipelines for validating which variants and which associated genes are causal for disease pathophysiology. Harnessing the methods we discussed above, such work has become feasible to perform and will elucidate further the genetic and mechanistic underpinnings of complex human diseases.

References

1. Bhatnagar A. Environmental Determinants of Cardiovascular Disease. Circulation Research. 2017; 121:162–180. [PubMed: 28684622]
2. LeBlanc M, Zuber V, Andreassen BK, Witoelar A, Zeng L, Bettella F, et al. Identifying Novel Gene Variants in Coronary Artery Disease and Shared Genes With Several Cardiovascular Risk Factors. Circulation Research. 2016; 118:83–94. [PubMed: 26487741]
3. CARDIoGRAMplusC4D Consortium. Deloukas P, Kanoni S, Willenborg C, Farrall M, Assimes TL, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. Nat Genet. 2013; 45:25–33. [PubMed: 23202125]
4. Devuyst O, Pattaro C. The UMOD Locus: Insights into the Pathogenesis and Prognosis of Kidney Disease. Journal of the American Society of Nephrology. 2017
5. Musunuru K, Hickey KT, Al-Khatib SM, Delles C, Fornage M, Fox CS, et al. Basic concepts and potential applications of genetics and genomics for cardiovascular and stroke clinicians: a scientific statement from the American Heart Association. Circ Cardiovasc Genet. 2015; 8:216–242. [PubMed: 25561044]
6. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Mexican American Type 2 Diabetes (MAT2D) Consortium, Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples (T2D-GENES) Consortium. Mahajan A, et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet. 2014; 46:234–244. [PubMed: 24509480]
7. The International HapMap Consortium. A haplotype map of the human genome. Nature. 2005; 437:1299–1320. [PubMed: 16255080]
8. Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. The American Journal of Human Genetics. 2001; 69:1–14. [PubMed: 11410837]
9. Harmston N, Lenhard B. Chromatin and epigenetic features of long-range gene regulation. Nucleic Acids Research. 2013; 41:7185–7199. [PubMed: 23766291]
10. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature. 2010; 466:714–719. [PubMed: 20686566]
11. 1000 Genomes Project Consortium. Durbin RM, Kang HM, McVean GA, Lehrach H, Wilson RK, et al. A global reference for human genetic variation. Nature. 2015; 526:68–74. [PubMed: 26432245]
12. Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, et al. The landscape of recombination in African Americans. Nature. 2011; 476:170–175. [PubMed: 21775986]

13. Buyske S, Wu Y, Carty CL, Cheng I, Assimes TL, Dumitrescu L, et al. Evaluation of the Metabochip Genotyping Array in African Americans and Implications for Fine Mapping of GWAS-Identified Loci: The PAGE Study. PLoS ONE. 2012; 7:e35651–7. [PubMed: 22539988]
14. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015; 347:1254806. [PubMed: 25525159]
15. Lin J, Hu Y, Nunez S, Foulkes AS, Cieply B, Xue C, et al. Transcriptome-Wide Analysis Reveals Modulation of Human Macrophage Inflammatory Phenotype Through Alternative Splicing. Arteriosclerosis, Thrombosis, and Vascular Biology. 2016; 36:1434–1447.
16. Stahringer A, Sass K, Pichler G, Kulak NA, Wilfert W, Kohlmaier A, et al. Circular non-coding RNA ANRIL modulates ribosomal RNA maturation and atherosclerosis in humans. Nat Comms. 2016; 7:1–14.
17. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: Illuminatingthe Dark Road from Association to Function. The American Journal of Human Genetics. 2013; 93:779–797. [PubMed: 24210251]
18. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, et al. High-Resolution Profiling of Histone Methylations in the Human Genome. Cell. 2007; 129:823–837. [PubMed: 17512414]
19. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell. 2008; 132:311–322. [PubMed: 18243105]
20. Gaulton KJ, Nammo T, Pasquali L, Simon JM, Giresi PG, Fogarty MP, et al. A map of open chromatin in human pancreatic islets. Nat Genet. 2010; 42:255–259. [PubMed: 20118932]
21. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Meth. 2013; 10:1213–1218.
22. Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. Nat Rev Genet. 2016; 17:487–500. [PubMed: 27346641]
23. Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
24. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–330. [PubMed: 25693563]
25. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015; 16:197–212. [PubMed: 25707927]
26. Pashos EE, Park Y, Wang X, Raghavan A, Yang W, Abbey D, et al. Large, Diverse Population Cohorts of hiPSCs and Derived Hepatocyte-like Cells Reveal Functional Genetic Variation at Blood Lipid-Associated Loci. Stem Cell. 2017; 20:558–570.e10.
27. Civelek M, Wu Y, Pan C, Raulerson CK, Ko A, He A, et al. Genetic Regulation of Adipose Gene Expression and Cardio-Metabolic Traits. Am J Hum Genet. 2017; 100:428–443. [PubMed: 28257690]
28. Jansen R, Hottenga J-J, Nivard MG, Abdellaoui A, Laport B, de Geus EJ, et al. Conditional eQTL analysis reveals allelic heterogeneity of gene expression. Hum Mol Genet. 2017; 26:1444–1451. [PubMed: 28165122]
29. Ackermann M, Sikora-Wohlfeld W, Beyer A. Impact of natural genetic variation on gene expression dynamics. PLoS Genet. 2013; 9:e1003514. [PubMed: 23754949]
30. The GTEx Consortium. Ardlie KG, DeLuca DS, Segre AV, Sullivan TJ, Young TR, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015; 348:648–660. [PubMed: 25954001]
31. Sweet DJ. iPSCs Meet GWAS: The NextGen Consortium. Stem Cell. 2017; 20:417–418.
32. Panopoulos AD, D'Antonio M, Benaglio P, Williams R, Hashem SI, Schuldt BM, et al. iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types. Stem Cell Reports. 2017; 8:1086–1100. [PubMed: 28410642]
33. Carcamo-Orive I, Hoffman GE, Cundiff P, Beckmann ND, D'Souza SL, Knowles JW, et al. Analysis of Transcriptional Variability in a Large Human iPSC Library Reveals Genetic and Non-

genetic Determinants of Heterogeneity. Cell Stem Cell. 2017; 20:518–532.e9. [PubMed: 28017796]

34. Siller R, Greenhough S, Park I-H, Sullivan GJ. Modelling human disease with pluripotent stem cells. Curr Gene Ther. 2013; 13:99–110. [PubMed: 23444871]
35. Patterson M, Chan DN, Ha I, Case D, Cui Y, Handel Ben Van, et al. Defining the nature of human pluripotent stem cell progeny. Cell Research. 2011; 22:178–193. [PubMed: 21844894]
36. Kim K, Doi A, Wen B, Ng K, Zhao R, Cahan P, et al. Epigenetic memory in induced pluripotent stem cells. Nature. 2010; 467:285–290. [PubMed: 20644535]
37. Kim K, Zhao R, Doi A, Ng K, Unternaehrer J, Cahan P, et al. Donor cell type can influence the epigenome and differentiation potential of human induced pluripotent stem cells. Nat Biotechnol. 2011; 29:1117–1119. [PubMed: 22119740]
38. Zhang X, Joehanes R, Chen BH, Huan T, Ying S, Munson PJ, et al. Identification of common genetic variants controlling transcript isoform variation in human whole blood. Nat Genet. 2015; 47:345–352. [PubMed: 25685889]
39. Hedman AK, Mendelson MM, Marioni RE, Gustafsson S, Joehanes R, Irvin MR, et al. Epigenetic Patterns in Blood Associated With Lipid Traits Predict Incident Coronary Heart Disease Events and Are Enriched for Results From Genome-Wide Association Studies. Circ Cardiovasc Genet. 2017; 10:e001487. [PubMed: 28213390]
40. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, et al. Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol. 2012; 30:265–270. [PubMed: 22371081]
41. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 2012; 30:271–277. [PubMed: 22371084]
42. Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, et al. Direct Identification of Hundreds of Expression- Modulating Variants using a Multiplexed Reporter Assay. Cell. 2016; 165:1519–1529. [PubMed: 27259153]
43. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, et al. Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. Cell. 2016; 165:1530–1545. [PubMed: 27259154]
44. Lin J, Musunuru K. Genome engineering tools for building cellular models of disease. FEBS J. 2016; 283:3222–3231. [PubMed: 27218233]
45. Jones PD, Kaiser MA, Ghaderi Najafabadi M, McVey DG, Beveridge AJ, Schofield CL, et al. The Coronary Artery Disease-associated Coding Variant in Zinc Finger C3HC-type Containing 1 (ZC3HC1) Affects Cell Cycle Regulation. Journal of Biological Chemistry. 2016; 291:16318–16327. [PubMed: 27226629]
46. Mali P, Aach J, Stranges PB, Esvelt KM, Moosburner M, Kosuri S, et al. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nature Publishing Group. 2013; 31:833–838.
47. Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, et al. CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. Cell. 2013; 154:442–451. [PubMed: 23849981]
48. Canver MC, Smith EC, Sher F, Pinello L, Sanjana NE, Shalem O, et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature. 2015; 527:192–197. [PubMed: 26375006]
49. Rajagopal N, Srinivasan S, Kooshesh K, Guo Y, Edwards MD, Banerjee B, et al. High-throughput mapping of regulatory DNA. Nat Biotechnol. 2016; 34:167–174. [PubMed: 26807528]
50. Korkmaz G, Lopes R, Ugalde AP, Nevedomskaya E, Han R, Myacheva K, et al. Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. Nat Biotechnol. 2016; 34:192–198. [PubMed: 26751173]
51. He X, Tan C, Wang F, Wang Y, Zhou R, Cui D, et al. Knock-in of large reporter genes in human cells via CRISPR/Cas9-induced homology-dependent and independent DNA repair. Nucleic Acids Research. 2016; 44:e85–e85. [PubMed: 26850641]

52. Kitzman JO, Starita LM, Lo RS, Fields S, Shendure J. Massively parallel single-amino-acid mutagenesis. Nat Meth. 2015; 12:203–206.

53. Dey B, Thukral S, Krishnan S, Chakrobarty M, Gupta S, Manghani C, Rani V. DNA–protein interactions: methods for detection and analysis. Mol Cell Biochem. 2012; 365:279–299. [PubMed: 22399265]

54. Mundade R, Ozer HG, Wei H, Prabhu L, Lu T. Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. Cell Cycle. 2014; 13:2847–2852. [PubMed: 25486472]

55. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, et al. Genomewide association analysis of coronary artery disease. N Engl J Med. 2007; 357:443–453. [PubMed: 17634449]

56. McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, et al. A common allele on chromosome 9 associated with coronary heart disease. Science. 2007; 316:1488–1491. [PubMed: 17478681]

57. Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, et al. A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science. 2007; 316:1491–1493. [PubMed: 17478679]

58. Klarin D, Zhu QM, Emdin CA, Chaffin M, Horner S, McMillan BJ, et al. Genetic analysis in UK Biobank links insulin resistance and transendothelial migration pathways to coronary artery disease. Nat Genet. 2017; 49:1392–1397. [PubMed: 28714974]

59. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet. 2008; 40:161–169. [PubMed: 18193043]

60. Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, Rieder MJ, 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]

61. Sandhu MS, Waterworth DM, Debenham SL, Wheeler E, Papadakis K, Zhao JH, et al. LDL-cholesterol concentrations: a genome-wide association study. Lancet. 2008; 371:483–491. [PubMed: 18262040]

62. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 2008; 6:e107. [PubMed: 18462017]

63. Strong A, Ding Q, Edmondson AC, Millar JS, Sachs KV, Li X, et al. Hepatic sortilin regulates both apolipoprotein B secretion and LDL catabolism. J Clin Invest. 2012; 122:2807–2816. [PubMed: 22751103]

64. Ding Q, Lee Y-K, Schaefer EAK, Peters DT, Veres A, Kim K, et al. A TALEN Genome-Editing System for Generating Human Stem Cell-Based Disease Models. Stem Cell. 2013; 12:238–251.

65. Holdt LM, Hoffmann S, Sass K, Langenberger D, Scholz M, Krohn K, et al. Alu Elements in ANRIL Non-Coding RNA at Chromosome 9p21 Modulate Atherogenic Cell Functions through Trans-Regulation of Gene Networks. PLoS Genet. 2013; 9:e1003588–12. [PubMed: 23861667]

66. Musunuru K, Post WS, Herzog W, Shen H, O'Connell JR, McArdle PF, et al. Association of Single Nucleotide Polymorphisms on Chromosome 9p21.3 With Platelet Reactivity: A Potential Mechanism for Increased Vascular Disease. Circ Cardiovasc Genet. 2010; 3:445–453. [PubMed: 20858905]

67. Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, et al. 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. Nature. 2011; 470:264–268. [PubMed: 21307941]

68. Almontashiri NAM, Fan M, Cheng BLM, Chen H-H, Roberts R, Stewart AFR. Interferon-γ Activates Expression of p15 and p16 Regardless of 9p21.3 Coronary Artery Disease Risk Genotype. J Am Coll Cardiol. 2013; 61:143–147. [PubMed: 23199516]

69. PhD CE, RN JGB, MSc PSB, MD NJS. The 9p21 Locus Does Not Affect Risk of Coronary Artery Disease Through Induction of Type 1 Interferons. J Am Coll Cardiol. 2013; 62:1376–1381. [PubMed: 23933542]

70. Almontashiri NAM, Antoine D, Zhou X, Vilmundarson RO, Zhang SX, Hao KN, et al. 9p21.3 Coronary Artery Disease Risk Variants Disrupt TEAD Transcription Factor-Dependent

Transforming Growth Factor β Regulation of p16 Expression in Human Aortic Smooth Muscle Cells. Circulation. 2015; 132:1969–1978. [PubMed: 26487755]

71. Miller CL, Pjanic M, Wang T, Nguyen T, Cohain A, Lee JD, et al. Integrative functional genomics identifies regulatory mechanisms at coronary artery disease loci. Nat Comms. 2016; 7:12092.
72. Kojima Y, Downing K, Kundu R, Miller C, Dewey F, Lancero H, et al. Cyclin-dependent kinase inhibitor 2B regulates efferocytosis and atherosclerosis. J Clin Invest. 2014; 124:1083–1097. [PubMed: 24531546]
73. Nanda V, Downing KP, Ye J, Xiao S, Kojima Y, Spin JM, et al. CDKN2B Regulates TGFβ Signaling and Smooth Muscle Cell Investment of Hypoxic Neovessels. Circulation Research. 2016; 118:230–240. [PubMed: 26596284]
74. Visel A, Zhu Y, May D, Afzal V, Gong E, Attanasio C, et al. Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. Nature. 2010; 464:409–412. [PubMed: 20173736]
75. Cunnington MS, Santibanez Koref M, Mayosi BM, Burn J, Keavney B. Chromosome 9p21 SNPs Associated with Multiple Disease Phenotypes Correlate with ANRIL Expression. PLoS Genet. 2010; 6:e1000899. [PubMed: 20386740]