Deciphering the Emerging Complexities of Molecular Mechanisms at GWAS Loci

Maren E. Cannon[1] and Karen L. Mohlke[1,*]

Genome-wide association studies (GWASs) have identified thousands of loci associated with hundreds of complex diseases and traits, and progress is being made toward elucidating the causal variants and genes underlying these associations. Functional characterization of mechanisms at GWAS loci is a multi-faceted challenge. Challenges include linkage disequilibrium and allelic heterogeneity at each locus, the noncoding nature of most loci, and the time and cost needed for experimentally evaluating the potential mechanistic contributions of genes and variants. As GWAS sample sizes increase, more loci are identified, and the complexities of individual loci emerge. Loci can consist of multiple association signals, each of which can reflect the influence of multiple variants, inseparable by association analyses. Each signal within a locus can influence the same or different target genes. Experimental studies of genes and variants can differ on the basis of cell type, cellular environment, or other context-specific variables. In this review, we describe the complexity of mechanisms at GWAS loci—including multiple signals, multiple variants, and/or multiple genes—and the implications these complexities hold for experimental study design and interpretation of GWAS mechanisms.

Introduction

Genome-wide association studies (GWASs) have identified thousands of loci associated with complex traits and diseases.[1,2] Converting GWAS findings into trait or disease insights includes elucidating both molecular mechanisms, by which genetic variants affect gene expression or function, and biological mechanisms, by which target genes affect a trait or disease. Progress is being made to identify candidate causal variants and genes underlying these associations, and complex molecular and biological mechanisms at GWAS loci are appearing. A recent review provides an excellent framework for the functional dissection of a genetic risk locus.[3] Here, we review the emerging complexities of molecular mechanisms at GWAS loci. After providing background to the challenges, we review three major questions: (1) How many association signals exist at a locus? (2) What are the candidate causal variant(s)? (3) What are the target gene(s)? In each section, we provide historical context to the question, methods available for addressing it, and evidence and observations from examples of GWAS loci that have been mechanistically characterized to date. Identifying mechanisms responsible for GWAS loci requires an accumulation of consistent evidence for the genes and variants that influence the trait or disease in humans (Figure 1). We conclude with future directions for researchers to consider in experimental design and interpretation of GWAS locus mechanisms.

Background

Complex genetic traits and diseases differ from monogenic traits and diseases. Monogenic diseases are caused by variation in single genes, whereas complex genetic traits are influenced by variation in multiple genes and environmental factors. GWASs have successfully identified thousands of genomic regions associated with hundreds of complex traits and diseases. GWAS publications typically report association results as a list of loci, distinguished from one another for counting purposes and labeled with a variant and one or more gene names as signposts. The variant named is typically the most strongly associated variant and is referred to as the lead, index, sentinel, or top variant (for other terminology, see Table 1). The gene names make referring to loci easier than using genome positions or variant labels, although the genes named in GWAS reports have variable evidence supporting their role in the trait or disease. Some GWAS reports simply indicate the nearest gene; others label loci with nearby gene(s) that have some annotation or experimental support. Early GWASs were performed with less densely spaced sets of variants, so the reported variant might not have been the strongest associated variant at a locus. More recent GWASs and GWAS meta-analyses are larger with sample sizes approaching one million for some traits, and although GWASs have often been performed in a single population, a growing number of trans-ancestry studies combine data across populations. For most identified loci, the molecular and biological mechanisms remain to be determined.

Much of the complexity of mechanisms at GWAS loci is due to allelic heterogeneity, in which multiple alleles act through the same gene to influence the same phenotype; allelic heterogeneity is common at monogenic disease-associated loci. For example, the Cystic Fibrosis Mutation Database includes >2,000 disease-causing mutations,[4] and at least 17 mutations can cause sickle cell disease.[5] As GWAS sample sizes become larger and we delve further into the mechanisms at GWAS loci, we are learning that allelic heterogeneity is also prevalent in complex genetic traits, and this heterogeneity influences both the design and interpretation of experimental studies.

Allelic heterogeneity of complex traits has been identified in studies of model organisms. Initially, quantitative

[1] Department of Genetics, University of North Carolina, Chapel Hill, NC, 27599, USA
*Correspondence: mohlke@med.unc.edu

https://doi.org/10.1016/j.ajhg.2018.10.001.

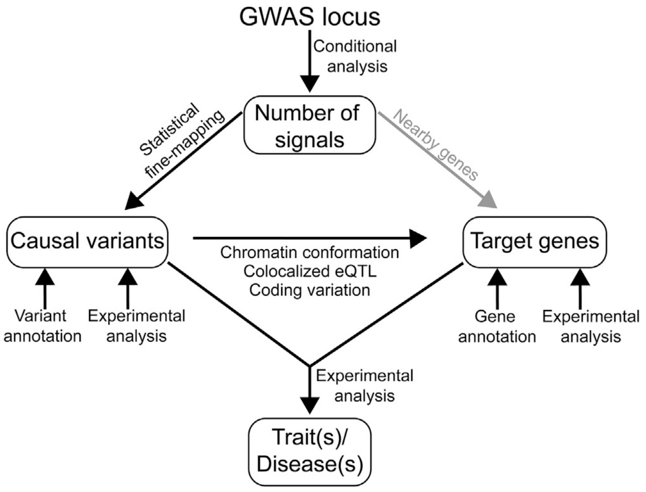

Figure 1. Process for Evaluating a GWAS Locus

Many approaches exist for identifying mechanisms at GWAS loci. In this review, we address three major questions at GWAS loci: (1) How many association signals exist at a locus? (2) What are the candidate causal variant(s)? (3) What are the target gene(s)? This flowchart shows how an accumulation of evidence can address these questions.

trait loci (QTLs) in model organisms were generally assumed to harbor one causal and multiple passenger alleles that affect a single causal gene; however, dissection of QTLs in inbred organisms has identified evidence of more than one gene in the same region.[6–8] In addition to identifying multiple genes at QTLs, fine-mapping efforts in model organisms have suggested multiple causal variants at a single locus.[8,9] Thus, genetic studies in model organisms suggest complex mechanisms that can involve multiple genes and variants at a single locus.

Experimental characterization of GWAS loci has lagged behind locus discovery because each locus presents a multi-faceted challenge. The location of many GWAS variants in noncoding regions[10] provides less straightforward hypotheses for mechanisms than variants within protein-coding regions. Each association signal typically consists of multiple variants in linkage disequilibrium (LD), and the sheer numbers of candidate variants can pose a challenge for interpreting annotations and performing experimental analyses; for example, 6,324 SNPs were reported to be in high LD ($r^2 > 0.5$) with 146 lead variants in 100 regions associated with prostate cancer.[11] Often, the cell type or tissue of action and the cellular state are unknown, and researchers must choose a cell type or model organism and potential stimuli to test mechanisms. It can be difficult to recapitulate the exact conditions of a trait or disease in model systems to determine the precise mechanistic effects in the human body. Identification of mechanisms, even in high throughput, requires a locus-by-locus interpretation, involving significant time and resources. Despite these challenges, significant progress has been made to identify molecular and biological mechanisms for GWAS loci across many complex diseases and traits.

How Many Association Signals Exist at a Locus?

Historical Context

Candidate-gene and early genome-wide studies identified multiple variants in the same gene in association with a complex trait. For example, four rare coding variants in *IFIH1* (MIM: 606951) were associated with lower type 1 diabetes risk,[12] and seven coding variants in *NOD2* (MIM: 605956) were associated with Crohn disease.[13,14] These examples represent allelic heterogeneity at complex-trait loci. In each example, the coding variants showed independent evidence of association with the disease or trait in larger GWAS analyses.[14]

Initial GWAS analyses identified genomic regions harboring variants associated with a given trait or disease as loci and typically defined distinct loci according to distance. When trait-associated variants at a locus do not exhibit strong pairwise LD with each other, they represent distinct association "signals." For example, Willer and colleagues[15] aligned GWAS loci for cholesterol and triglyceride levels to previously reported causal variants to demonstrate that the GWAS analysis had identified additional signals of association at these loci. Early studies had limited statistical power to detect loci with two or more significant signals.

Methods

Linkage Disequilibrium. To determine the number of signals at a locus, one strategy is to evaluate pairwise LD between a lead variant and other variants at the locus. GWAS analyses can define multiple signals within a genomic region on the basis of the LD measure r^2 in a selected reference population. GWASs have used different LD thresholds and different reference panels. For example, a schizophrenia GWAS used a threshold of $r^2 < 0.1$ to identify 128 independent signals at 108 loci,[16] and a coronary artery disease GWAS used a threshold of $r^2 < 0.2$ to identify 104 independent signals at 46 loci;[17] both used 1000 Genomes Project data as a reference panel.

Conditional Analysis. An additional strategy for identifying multiple signals at a locus is conditional association analysis. An initial lead associated variant is included as a covariate in association analyses testing other nearby variants; if a nearby variant remains significant, it is considered a conditionally distinct additional signal. When individual-level genotype and trait data are available, stepwise conditional analysis can be performed by including each newly identified lead variant as an additional covariate. When only summary-level association statistics from GWAS meta-analysis are available, a conditional and joint analysis option in the Genome-wide Complex Trait Analysis (GCTA) software[18] uses estimated LD from a provided reference sample to identify conditionally distinct signals. Significance thresholds for additional signals at a locus are typically set to account for the number of variants tested, either within a locus or genome-wide. Limitations of defining signals on the basis of LD in a reference panel are that reference panels missing variants analyzed in the association study can fail to detect signals and that

Table 1. Definitions of Terms Used in This Review

Monogenic disease or trait	a disease caused by or a trait influenced by variation in a single gene, although variants in multiple single genes can cause the same disease
Complex disease or trait	a disease or trait caused by a combination of variants in many genes; it can be influenced by behavioral and environmental factors
Molecular mechanism	a mechanism by which a genetic variant affects gene expression or function
Biological mechanism	a mechanism by which a gene affects a trait or disease
Allelic heterogeneity	the phenomenon by which multiple variants act on one gene to influence the same phenotype
Locus	a genomic region associated with a complex trait or disease; it is often defined by a distance, e.g., within 500 kb or 1 Mb of a reported variant
Signal	within a locus, a set of variants that are in strong pairwise LD with each other and are associated with a trait or disease; multiple signals can be conditionally distinct from each other, a subset of which are independent
Causal variant	a variant that affects a molecular or cellular process to have an impact on a trait or disease
Functional variant	a variant that shows evidence of allelic differences affecting gene regulation or function; variants can be functional but not affect a trait or disease
Target gene	a gene affected by a functional variant; also called an effector gene

population differences between the reference and GWAS samples can cause signals to be mis-characterized.

Haplotype Association. When individual-level genotype and trait data are available, haplotype association analysis can help interpret the inheritance patterns of signals that are not fully independent from one another. Variant alleles that are present on the same or different homologous copies of a genomic region can have differing physiological consequences. Several methods are available for estimating haplotypes from genotype data.[19] Haplotypes are then tested for association with a trait either via one test per haplotype or through a global test for all haplotypes in a region.[20,21] Estimates of haplotype effect sizes and directions can aide interpretation of multiple nearby signals.[22–24]

Multiple Signals at GWAS Loci

Loci that appear to consist of only one signal can be found to harbor multiple signals after denser genotyping, conditional analysis, or the inclusion of more samples or ancestry groups. Using 1000 Genomes instead of HapMap as an imputation reference panel to analyze the same individuals detected additional loci and different lead variants,[25] and genome sequencing by the UK10K Consortium identified additional low-frequency and rare alleles that had not been detected by array-based genotyping.[26] Conditional analysis in larger sample sizes of GWAS meta-analyses has detected additional signals at established loci for many phenotypes; for example, an early report identified additional signals at 26 loci for lipids,[27] and a more recent study of lipids in a multi-ancestry cohort identified 121 new association signals, including 15 additional signals at established loci.[28] In another example, a blood pressure analysis accounting for smoking behavior identified nine novel signals at known loci, including eight that were detected only through the inclusion of data from study subjects of all ancestries.[29]

As more signals are identified, the definition of a "locus" can change. Signals can span distances of hundreds of kilobases, and a single 500 kb locus containing two signals could easily be defined as two separate loci with narrower spans. A study of 36 blood cell traits reported signals without reporting loci; investigators used stepwise multiple regression to identify 3,755 conditionally distinct associations that clustered into 2,706 LD-group signals.[30]

Given the growing number of signals identified for some traits, determining whether association signals are new or previously reported is becoming more challenging. Some studies have aimed to identify additional new signals by simultaneously conditioning on all known signals. Such "global" conditional analyses can include hundreds of variants as covariates in a GWAS. For example, an age-related macular degeneration GWAS using a modified global conditional analysis identified 52 signals across 24 loci,[31] and a GWAS of lipoprotein particle subclasses conditioned on 885 variants at 157 previously identified lipid loci and identified five novel signals.[32]

Identifying multiple signals at a locus can help explain some of the "missing heritability" of complex diseases and traits. A stepwise conditional analysis that identified seven signals at three loci associated with fetal hemoglobin levels increased the explained heritable variation in fetal hemoglobin from 38.6% to 49.5%.[33] Re-sequencing at five low-density lipoprotein cholesterol (LDL-C) loci identified additional signals and increased the estimate of heritability from 3.1% to 6.5%,[34] and including additional ancestries further increased lipid heritability estimates from 1.3- to 1.8-fold across all signals and traits.[35] Similar patterns were identified in expression QTL (eQTL) associations—9% of *cis*-eQTL loci showed evidence of a secondary signal, resulting in a 31% average increase in explained phenotypic variance.[22]

Identifying additional signals enables additional mechanisms to be characterized. For example, variants in three enhancers representing association signals bind three different transcription factors to influence expression of *RET* (MIM: 164761), leading to Hirschsprung disease.[36] In addition, the *TCF7L2* (MIM: 602228) locus for

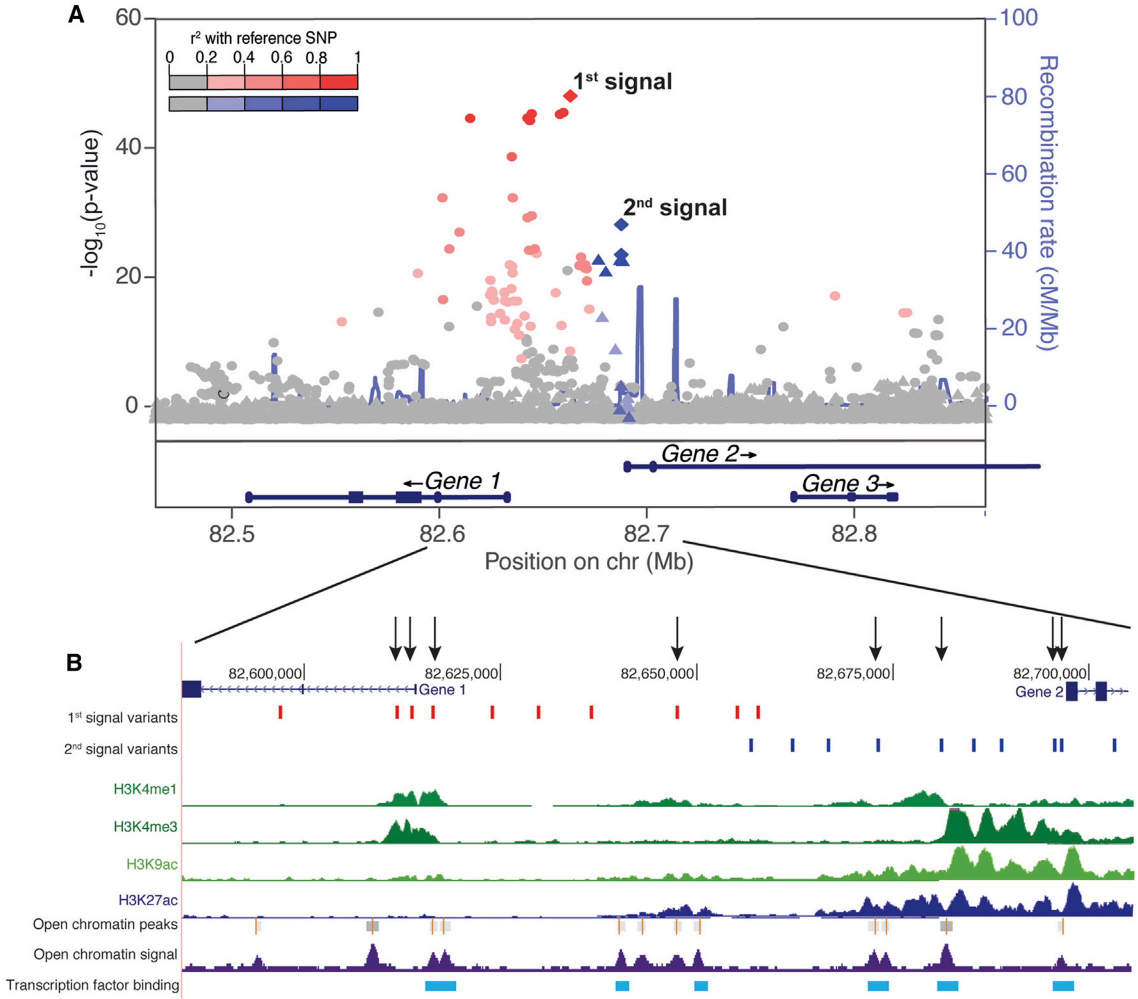

Figure 2. Hypothetical GWAS Locus with Two Signals that Affect Two Genes
(A) Plot of association for two signals within 100 kb at a single GWAS locus. The first signal is shown by red circles, and the second is shown by blue triangles. The intensity of color corresponds to the strength of LD between the lead variant and other variants in the signal.
(B) Hypothetical regulatory marks overlapping the positions of candidate variants. Arrows point to variants that overlap predicted regulatory regions: four for signal 1 and four for signal 2. Signal 1 variants could target gene 1, and signal 2 variants could target gene 2 because variants are located in each respective promoter.

type 2 diabetes initially appeared to consist of a single signal, and early variant characterization suggested that rs7903146 affected islet enhancer activity.[37,38] Now, eight signals at the *TCF7L2* locus have been reported to be associated with diabetes risk, and several do not overlap islet regulatory elements.[39] One or more of the new signals could affect other mechanisms of *TCF7L2* regulation, including alternative splicing, expression in other tissues, or both.[40] In these examples, the additional signals could target the same candidate gene, but signals could also target different nearby genes. In a recent analysis of eQTLs at GWAS loci, Gamazon et al. observed more than one colocalized gene and one tissue at more than 50% of signals.[41] Nearby signals that target different genes or transcripts, possibly with different mechanisms across cell types, could be especially common in gene-dense regions (Figure 2).

Haplotype analysis can aid the interpretation of multiple signals. Identifying shared haplotypes between alleles of multiple signals can help explain why a variant with low initial evidence of association becomes much more significant after being conditioned on a nearby variant and why a variant with strong initial evidence of association becomes less significant but still meets a significance threshold.[22] Haplotype analysis can also help interpret the mechanistic consequences of regulatory and coding variants at the same locus.[23,24] In a study of *G6PC2* (MIM: 612108) missense variants associated with fasting glucose, single-variant association results showed an apparent discrepancy with results of cellular functional studies. Haplotype analysis explained the discrepancy by showing that the glucose-lowering allele of the coding variant was always inherited with the glucose-raising allele of a more common noncoding GWAS signal.[42] Haplotype analysis can be especially relevant to identifying the functional consequences of variants at multi-signal loci.

Conclusions

As GWAS sample sizes increase, the observable complexity of association signals at individual GWAS loci is increasing. Multiple signals exist at many GWAS loci, and a pattern is emerging whereby the strongest GWAS loci are often influenced by multiple nearby association signals. These multiple signals represent more of the disease or trait heritability than initial signals, and the additional candidate variants can have distinct mechanisms affecting the associated trait or disease, such as variants in different regulatory elements that regulate different genes. Alleles at distinct, but not completely independent, signals can act together through haplotypes. We encourage researchers to consider the possibility that more than one signal contributes to a GWAS locus as a valuable step in accurately delineating the mechanisms at GWAS loci. Considering more than one signal can be especially helpful when the direction of effect of a signal appears inconsistent with other data. When complex relationships between signals are identified, consider the potential contribution of gene regulation, and embrace the opportunity to identify potentially novel interacting genetic mechanisms.

What Are the Candidate Causal Variant(s)?

Historical Context

Identifying candidate causal variants underlying GWAS signals is valuable in helping to identify target genes for GWAS loci because the gene(s) responsible for an association are often not clear, and identifying causal links allows variant-gene links to be validated. The value of identifying candidate causal variants can be questioned; however, specific candidate causal variants provide a data-driven link to genes rather than relying on assumptions about the connections to specific genes and their directions of effect to alter the amount or function of gene products. At noncoding GWAS loci, identifying candidate causal variants further informs our understanding of gene regulation, especially distances at which variants and regulatory regions can act and mechanisms by which variants affect protein-coding RNAs directly and via antisense, long noncoding, or other molecular moieties.

For monogenic disorders, disease-associated chromosomes are typically found to contain a single causal variant. For complex traits, a single causal variant per GWAS signal is the simplest explanation, and although such mechanisms have been described,[43–47] it is becoming increasingly evident that multiple variants can be causal at a single GWAS signal.[48–50] That is, for complex traits, allelic heterogeneity exists at the level of both the locus (multiple signals) and the signal (multiple causal variants).

Edwards et al.[10] noted that "… variants are unlikely to act alone, and the importance of combinatorial effects should be considered." In one study, the authors suggested that "throughout the evolution of species it is haplotype blocks, rather than individual genes and mutations, that serve as the fundamental unit of inheritance" on the basis of highly inbred yeast strains.[8] These observations further support the potential role for multiple variants at a single locus. Variants can act in concert in promoters, enhancers, repressors, and coding regions, adding to the complexity of determining mechanisms at GWAS loci.

Methods

Statistical Fine-Mapping. As a first step in identifying candidate causal variants in humans, the strongest GWAS variants and/or variants in LD with the most significantly associated variant at each signal are considered candidates. Typically, an LD r^2 threshold of 0.7 or 0.8 is used for determining candidate variants at a GWAS locus; however, various studies use different, less stringent thresholds (e.g., $r^2 > 0.5$) given the strong contribution of allele frequency to r^2 estimates or population differences between samples and reference panels.

Statistical fine-mapping uses association statistics to predict which variants are more likely to be causal. Analyses fit in two broad categories: (1) prioritizing variants on the basis of association statistics and LD and (2) Bayesian methods that assign posterior probabilities of causality to each variant (Table 2), as reviewed recently.[52] Methods differ in the required input (such as original genotype-trait data versus summary statistics), assumptions (such as one versus more than one potential causal variant), and output. Some methods, such as CAVIAR,[53] allow for an arbitrary number of candidate causal variants and return a list, or "credible set" of variants. FINEMAP explores a set of the causal configurations of variants in a region.[54] PAINTOR,[55] fGWAS,[56] and deterministic approximation of posteriors (DAP)[57] incorporate functional annotations in the analysis. Performing fine-mapping across diverse populations (e.g., with MANTRA[51]) can further refine a shared signal on the basis of differing LD between the populations; however, this approach assumes a shared variant between populations, which might or might not reflect the true causal variants within each population. In most cases, statistical fine-mapping predicts a set of variants that are likely to be causal and can then be further examined for functional effects.[39,58] A notable limitation of these approaches is that they analyze only the variants provided; other variants not included as a result of failed genotyping, poor imputation, or other exclusions (e.g., indels and triallelic variants) cannot be considered candidates. The approaches can be sensitive to differences in evidence of association as a result of sample-size variation or inaccurate imputation, and some methods are limited by computation, reducing the number of variants that can be analyzed simultaneously.

Variant Annotation. The effects of coding variants can be interpreted more directly than noncoding variants and are thus frequently the first variants considered for causality. The effects of coding variants can be predicted by computational algorithms, including SIFT,[59] PolyPhen-2,[60] MutationTaster2,[61] CADD,[62] MAPPIN,[63] and others.[64] These methods consider sequence conservation, protein structure, and amino acid properties to predict the effect of missense coding variants. Some methods, such as MutationTaster2 and CADD, also integrate functional

Table 2. Toolbox for Identifying and Annotating Candidate GWAS Variants

Purpose	Tool[a]	URL or Reference
Identification of LD proxies	LDlink	https://ldlink.nci.nih.gov/
Statistical fine-mapping	BIMBAM	http://stephenslab.uchicago.edu/software.html
	MANTRA	Morris[51]
	CAVIARBF	https://bitbucket.org/Wenan/caviarbf
	CAVIAR	http://genetics.cs.ucla.edu/caviar/
	PAINTOR	https://github.com/gkichaev/PAINTOR_V3.0
	fGWAS	https://github.com/joepickrell/fgwas
	PICS	https://pubs.broadinstitute.org/pubs/finemapping/pics.php
	funciSNP	https://doi.org/doi:10.18129/B9.bioc.FunciSNP
	FINEMAP	http://www.christianbenner.com/
	GenoWAP	http://genocanyon.med.yale.edu/GenoWAP
	DAP	https://github.com/xqwen/dap
Annotation of splice variants	HEXplorer	http://www2.hhu.de/rna/html/hexplorer_score.php
	ASSA	http://splice.uwo.ca/
	MaxEntScan	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html
	GeneSplicer	http://ccb.jhu.edu/software/genesplicer/
	Human Splicing Finder	http://www.umd.be/HSF3/
Annotation of coding variants	PolyPhen2	http://genetics.bwh.harvard.edu/pph2/
	SIFT	http://provean.jcvi.org/index.php
	CONDEL	http://bbglab.irbbarcelona.org/fannsdb/
	MAPPIN	https://doi.org/10.6084/m9.figshare.4639789
	dbNSFP	https://sites.google.com/site/jpopgen/dbNSFP
	MutationTaster2	http://www.mutationtaster.org/
Annotation of noncoding and coding variants	CADD	https://cadd.gs.washington.edu/
	PredictSNP2	https://loschmidt.chemi.muni.cz/predictsnp2/
	FATHMM-MKL	http://fathmm.biocompute.org.uk/
	EIGEN	http://www.columbia.edu/~ii2135/eigen.html
Annotation of noncoding variants	GWAVA	https://www.sanger.ac.uk/sanger/StatGen_Gwava
	ARVIN	https://github.com/gaolong/arvin
	SNVrap	http://jjwanglab.org/snvrap
	DANN	https://cbcl.ics.uci.edu/public_data/DANN/
	DanQ	https://github.com/uci-cbcl/DanQ
	SNPDelScore	https://www.ncbi.nlm.nih.gov/research/snpdelscore
	deltaSVM	http://www.beerlab.org/deltasvm/
	DeepSEA	http://deepsea.princeton.edu/job/analysis/create/
	3DSNP	http://cbportal.org/3dsnp/
	HaploReg	http://pubs.broadinstitute.org/mammals/haploreg/haploreg.php
	SNiPA	http://snipa.helmholtz-muenchen.de/snipa3/
	GREGOR	https://genome.sph.umich.edu/wiki/GREGOR
	GARFIELD	https://www.ebi.ac.uk/birney-srv/GARFIELD/
	RegulomeDB	http://www.regulomedb.org/

(*Continued on next page*)

Table 2. *Continued*

Purpose	Tool[a]	URL or Reference
Regulatory datasets	ENCODE	https://www.encodeproject.org/
	Roadmap Epigenomics	http://www.roadmapepigenomics.org/
	Fantom5	http://fantom.gsc.riken.jp/5/
	VISTA enhancer	https://enhancer.lbl.gov/
	cistromeDB	http://cistrome.org/db/#/
	Blueprint	http://www.blueprint-epigenome.eu/
Databases of transcription factor binding motifs	JASPAR	http://jaspar.genereg.net/
	HOCOMOCO	http://hocomoco11.autosome.ru/
	PWMtools	https://ccg.vital-it.ch/pwmtools/

[a]These example tools are intended as a starting point for researchers to identify available tools.

genomic data (e.g., DNase I hypersensitivity). The MAPPIN algorithm additionally includes post-translational modifications of proteins, biological networks, and allele frequency to determine whether a variant is predicted to be deleterious or not. Cell or organism models can show the effect of a coding variant in a model system. Variants within or near the conserved splice site between exons can affect splicing, and deleterious effects can be predicted with tools, as reviewed by Jian et al.[65]

Noncoding variants can be annotated by colocalization with genomic regulatory regions in relevant cell types. Large consortium efforts, including the Encyclopedia of DNA Elements (ENCODE),[66] the Roadmap Epigenomics Project,[67] and the International Human Epigenome Consortium,[68] have created robust datasets for many cell and tissue types to describe regions characteristic of regulatory activity. These datasets include chromatin immunoprecipitation sequencing (ChIP-seq) of histone marks often observed at enhancers, promoters, and insulators and transcription factors and open chromatin profiles generated by DNase hypersensitivity sequencing (DNase-seq), formaldehyde-assisted isolation of regulatory elements sequencing (FAIRE-seq), and the assay for transposase-accessible chromatin using sequencing (ATAC-seq). Algorithms can be used to assess which annotations are enriched in association results (e.g., GREGOR[69] and GARFIELD[70]), to annotate variants at a signal (e.g., HaploReg and RegulomeDB), and predict functional consequence of noncoding variants (e.g., CADD and GWAVA) (Table 2).

Sequencing data generated from ChIP-seq, DNase-seq, FAIRE-seq, or ATAC-seq can be used for identifying heterozygous sites of allelic imbalance in transcription factor binding or chromatin accessibility. A site of allelic imbalance occurs when a sample has disproportionate sequencing reads for each allele in comparison with the expected 50:50. An important aspect of evaluating allelic imbalance is to align reads by using a strategy that avoids bias toward the reference genome allele at the expense of an alternate allele.[71] Numerous methods have been described for the identification of allelic imbalance in high-throughput sequence data.[72–74] These strategies require sufficient reads to detect imbalances and require the sequenced sample to be heterozygous at a position of interest.

Experimental Analysis. Experimental analysis can determine whether variants show allelic differences in gene function or regulation, such as allele-specific effects on gene expression levels.[47,75] Experimental analysis of coding and splicing variants requires gene-specific assays that include examining protein function and downstream phenotypes in cells and model organisms.[76] Many approaches also exist for testing regulatory variants in functional experiments.[77]

Transcriptional reporter assays test variant alleles located in regulatory regions for differences in transcriptional activity. When an individual regulatory region is analyzed, the regions surrounding an associated variant are cloned into a vector containing a reporter gene, usually luciferase or GFP, and transfected into a cell line or transiently expressed in a model organism. The activity of reporter genes is measured, and the variant alleles are compared for the detection of any allelic differences in transcriptional activity. Transcriptional reporter assays can also be performed in a high-throughput manner with massively parallel reporter assays (MPRAs), in which hundreds of regulatory regions are tested simultaneously.[78] Reporter assays require the presence of the transcription factors that drive allelic differences in activity, so selection of cell type and context, such as differentiation state and stimuli, is important. The limitations of reporter assays in low or high throughput are that (1) the size of the cloned segment and the location of the variant in the segment and in relation to the transcriptional start site can affect detection of allelic differences, leading to at least false-negative results and possibly false positives, and (2) the variant is removed from the chromosomal context; however, this latter limitation can be somewhat remedied by lentiviral transduction and recombineering technologies that allow the construct to incorporate into the genome.

Protein binding assays—including electrophoretic mobility shift assays (EMSAs), DNA-affinity pull-downs, and ChIP experiments—are used to identify variant alleles that bind transcription factors differentially. EMSAs are an *in vitro* approach that visualizes nuclear protein complexes that bind to ~20–100 bp DNA probes surrounding a candidate variant. The DNA-protein complexes are visualized on a gel, and the identity of the transcription factor can be determined with antibodies to transcription factors predicted by conserved transcription factor binding motifs or identified by ChIP-seq datasets. DNA-affinity pulldowns are similar to EMSAs: all DNA-protein complexes are captured by a probe including a candidate variant allele and visualized on a gel. Proteins in allele-specific bands are identified by mass spectrometry. Allelic differences in transcription factor binding can also be evident in differences in ChIP-seq or ChIP-qPCR between samples of different genotypes.[48] Protein binding assays can also be performed in high throughput.[79,80]

An increasingly popular method for determining the function of regulatory variants and elements is to use CRISPR-Cas9 genome editing to delete the regulatory region, create the alternate allele of a variant, or alter the epigenome and regulatory regions with CRISPRi.[81,82] After genome editing, assays of gene expression and/or gene function are performed. Deleting a regulatory element or substituting the alternate allele allows for direct observation of phenotypic effects in the native chromatin context. Challenges that arise in optimizing CRISPR-Cas9 genome editing include the following: it is difficult to edit many disease-specific cell types, targeting can create mutations at unwanted sites, the deletion of a regulatory element might not have the same effect as that of a single-nucleotide change, and editing in model organisms can create mosaicism.

Finally, model organisms can be used to model the effect of candidate causal variants. Reporter assays can be performed in model organisms, often in mice or zebrafish, for assessing variant effects in the context of multiple cell types and controllable environmental conditions. Genome editing in organisms allows more complex drug- or chemically inducible effects to be characterized. Model organisms have the benefit of providing the context of an entire biological system but can sometimes show phenotypic effects that are not consistent in humans.

Candidate Variants at GWAS Signals

Fine-mapping and computational approaches often detect multiple candidate causal variants at individual GWAS loci. For example, in a recent GWAS meta-analysis of type 2 diabetes, only 18 of 380 signals resulted in a single-variant credible set.[39] Examination of autoimmune GWAS loci suggested a “multiple enhancer variant” hypothesis, whereby multiple associated variants in clusters of enhancers work together to alter gene expression.[49] CAVIAR fine-mapping of skin pigmentation GWAS loci resulted in multiple predicted causal variants at most loci.[58] Fine-mapping analyses can reduce the list of candidate variants at a GWAS locus.

Statistical fine-mapping methods use different models and can generate inconsistent results. Methods make different assumptions about the contributions of underlying variants and apply different strategies incorporating genetic association, LD, and functional annotation. At the *ANGPTL8* (MIM: 616223) locus, associated with high-density lipoprotein (HDL) cholesterol across populations, three fine-mapping methods (MANTRA, CAVIAR, and PAINTOR) of prioritizing candidate variants generated differing results;[83] MANTRA identified a credible set of ten variants, CAVIAR identified a credible set of 24 variants by using Finnish association data and a credible set of two variants by using African American association data, and PAINTOR identified ten, seven, and five variants in Finnish, African American, and the combined studies, respectively. Of the 39 variants identified by least one method, only 12 were identified in at least two analyses, and only four were identified in all three analyses. A candidate causal variant that was identified by all three methods showed significant allelic differences in two assays of regulatory function; however, further experiments are needed to determine the full molecular mechanism.

Great progress has been made in identifying regulatory mechanisms at GWAS loci. Among numerous examples, at the *PHACTR1* (MIM: 608723) locus (associated with vascular disease), rs9349379 was identified as a regulator of *EDN1* (MIM: 131240) expression by evidence from enhancer signatures in heart tissue and genome editing.[47] At the 6q22.1 locus (associated with prostate cancer), rs339331 risk alleles showed increased expression of *RFX6* (MIM: 612659), increased binding of HOXB13, increased enhancer histone mark H3K4me2, and altered cell morphology and adhesion.[84] At a *PARP1* (MIM: 173870) locus associated with melanoma, the risk allele showed increased expression of *PARP1* and decreased RECQL binding. Interestingly, RECQL overexpression in melanoma cells resulted in significant allelic differences in transcriptional activity, whereas basal levels of RECQL resulted in no allelic differences, suggesting that cellular context can be important for identifying functional effects of GWAS variants.[75]

Multiple variants within a single GWAS signal have shown evidence of functional effects. For example, at the *MFSD12* (MIM: 617745) locus (associated with skin pigmentation), six variants showed allelic differences in transcriptional reporter luciferase assays.[58] Similarly, at the *GALNT2* (MIM: 602274) locus (associated with HDL cholesterol), at least two regulatory variants were described as having effects on transcriptional reporter luciferase assays, *in vitro* protein binding, and allelic imbalance in ChIP-seq and DNase reads.[48] Four regulatory variants could contribute to the mechanism at the *HKDC1* (MIM: 617221) locus (associated with gestational hyperglycemia).[85] MPRAs identified 32 functional variants at 23 GWAS loci associated with red blood cell traits, and targeted genome editing showed that three functional variants at one locus affect transcription of *SMIM1* (MIM: 615242), *RBM38* (MIM: 612428), and *CD164* (MIM: 603356) and that *RBM38* is involved in erythropoiesis.[86] Multiple variants can be tested together in transcriptional reporter assays, but the distance between variants can present a problem for cloning, and plasmids do not provide the native chromatin context that is most likely important for the regulatory regions to act together. As genome-editing methods are optimized for cell types of interest, it will be important to test multiple variants together *in vivo* to determine the effects of haplotypes and to truly delineate the complex mechanisms. These functional-mechanism success stories highlight the great progress in understanding genetic contribution to disease in the post-GWAS era.

The *TERT* (MIM: 187270) locus (associated with breast cancer, ovarian cancer, and telomere length) provides an example of both multiple signals and multiple variants. This GWAS locus consists of at least three signals; signals 2 and 3 are distinct but not completely independent ($r^2 = 0.33$).[87] At the first signal, three regulatory variants located in the *TERT* promoter decrease transcriptional activity in breast and ovarian cancer cell lines. At the second signal, one variant in a *TERT* intron increases transcriptional activity. At the third signal, a variant alters *TERT* splicing, resulting in a premature stop codon and truncated protein. The mechanistic connections for multiple variants provide strong evidence for *TERT* as a plausible target gene.

Conclusions

Multiple candidate causal variants at a single GWAS locus adds complexity to delineating the molecular mechanism at the locus. If one functional variant is identified without full evaluation of all potential candidates, additional variants could contribute, and part of the mechanistic impact on a gene or genes could be missed. The effect of individual variants could be small and might not be observed in functional experiments, perhaps because the effect is observed in only a specific cellular environment or in combination with other variants. Additionally, a variant that affects a gene in functional assays does not necessarily demonstrate that it causes trait variation or a disease; further evidence is needed to prove this connection definitively. Many new tools and methods need further vetting to determine their effectiveness for a given situation. Modeling a single variant in some assays and systems could fail to exhibit a sufficient genetic or physiological consequence if multiple candidate causal variants act together to affect the gene of its function. Identifying different variants in different assays (e.g., transcriptional activity and chromatin interaction) can lead to seemingly inconsistent evidence when assay variability or missing data are responsible. In these cases, we encourage researchers to identify a consistent direction of allelic effect across multiple experiments, which together can provide strong conclusions about candidate causal variants at GWAS loci.

What Are the Target Gene(s)?

Historical Context

Identifying the target gene(s) at a GWAS locus is a fundamental part of elucidating the molecular mechanism because these genes provide a key to understanding the pathogenic processes and provide potential new targets for drug development. However, the target genes remain largely unknown for most GWAS loci. Although some early studies of complex genetic traits were designed on the basis of expecting perhaps tens of contributing genes, GWASs have found that hundreds or thousands of genes might contribute to complex genetic traits.

A large number of genes contributing to a trait or disease is consistent with the possibility that multiple susceptibility genes are located relatively close to each other, even at a single locus (Figure 2). GWAS loci can have multiple genes that appear to be good functional candidates on the basis of gene function and expression, coding variants, chromosome interactions, and/or literature review. Although one candidate causal gene per locus is perhaps still the most likely scenario, GWAS variants can affect multiple genes, perhaps at multi-signal loci or in gene-dense regions.

Methods

Colocalized Variant Association with Gene Expression. One method of identifying target genes for GWAS signals is identifying colocalized eQTL associations. eQTL analysis identifies variants associated with the RNA levels of genes (or transcripts, isoforms, or exons), usually in a single tissue or cell type. Whereas some eQTLs are shared across tissues, others are tissue specific.[88] Typically, representative GWAS variants are examined in eQTL datasets for the identification of GWAS variants that show significant association with the expression level of one or more genes. However, the GTEx Consortium reported that a remarkable 92.7% of common variants tested show nominal association ($p < 0.05$) with the expression level of at least one gene in at least one tissue.[88] In addition, given the very strong associations that can be observed between variants and gene expression level, noncausal variants in only moderate LD with candidate causal regulatory variants can also show significant association with gene expression levels. Hence, further analyses are required to determine whether the same variant(s) underlying the GWAS trait association are also likely to be the variant(s) that affect the RNA levels of the eQTL gene.

In the simplest scenario, a lead GWAS variant is also the lead variant associated with a gene in an eQTL study using samples from the same ancestral population. In this setting, eQTL and GWAS associations are considered to be colocalized through a comparison of the lead GWAS and eQTL variants. When the lead GWAS and eQTL variants are identical or LD r^2 between them is high (e.g., $r^2 > 0.8$ or 0.9), then the signals can be colocalized. A recent analysis of 3,718 independent GWAS signals found that 58.0% were in LD ($r^2 > 0.8$) with at least one eQTL and that 27.8% were in LD with the best eQTL variant for a gene.[41] Conditional analyses in the eQTL dataset can provide further support for colocalization; if the association between the lead eQTL variant and gene expression is no longer significant after conditioning on the GWAS variant and if the association between the lead GWAS variant and gene expression is no longer significant after conditioning on the lead eQTL variant, then the signals are typically considered colocalized.

Other methods more fully assess colocalization of GWAS and eQTL signals by using additional statistical tests, and some methods can be applied to summary association statistics and/or incorporate variant annotation. For example, COLOC applies a Bayesian procedure to estimate the posterior probabilities that a variant is causal in both GWASs and eQTL studies,[89] and eCAVIAR applies a Bayesian procedure that allows for more than one candidate causal variant.[90] Summary Mendelian randomization in conjunction with a test for heterogeneity in dependent instruments (HEIDI) tests whether gene expression and a trait are associated because of a shared candidate causal variant and can distinguish that model from two or more distinct genetic variants that are in LD and independently affect the gene expression level and the trait.[91] Limitations of using eQTLs to identify target genes include that eQTL datasets can still be underpowered to detect associations or unavailable in appropriate cell types or contexts, the LD structure might not be identical between available GWAS and eQTL datasets, and colocalization approaches can be computationally intensive and not robust to the presence of multiple eQTL signals, leading to potential false-positive and -negative colocalization. Finally, evidence of an eQTL colocalized with a GWAS signal does not necessarily mean that expression of that gene mediates the effect of the signal on the trait.

Similar to eQTL associations, allelic imbalance in gene expression can identify potential target genes. Allelic expression imbalance (AEI) analysis involves examining the cDNA or RNA-sequencing reads of genes of interest in individuals heterozygous for a transcribed variant. If the ratio of reads from each allele deviates from 1:1, the correlated alleles of a noncoding and transcribed variant can determine *cis*-acting variants. Advantages of this approach are that the two variant alleles are assayed in the same environment and that significant differences can be detected in smaller sample sizes than needed for eQTLs. Disadvantages of AEI are that some genes do not contain common variants in available samples and that especially in small sample sizes, a variant can exhibit AEI merely through moderate LD with a candidate causal regulatory variant.

Predicted Gene Expression Association Studies. A complementary approach to prioritizing candidate genes and direction of effect is predicted expression association studies. In this approach, GWASs and eQTL studies are integrated to identify disease associations on the basis of sets of variants that influence gene expression. The portion of gene expression due to genetic variants is estimated from reference eQTL studies, and these estimates are used for predicting gene expression in larger GWAS, where the imputed gene expression can be correlated to the trait for the identification of candidate genes.[92–94] These strategies increase power to detect genes that exhibit differences in genotype-dependent expression patterns, although power is reduced by pleiotropy, and false positives can be identified as a result of LD and the complex genetic architecture of GWAS loci.[57,95] Nonetheless, these studies can be used to help identify candidate genes for complex traits.[96,97]

Chromatin Conformation. Another approach to identifying target genes is identifying contact between GWAS variants and the promoters and transcription start sites of target genes by using chromatin conformation assays. Assays such as 3C, Hi-C, and Capture-C identify physical interactions and loops in the genome,[98,99] and these interactions can differ by cell and tissue type. Locus-specific assays can be performed to test interactions between GWAS variant regions and promoters of nearby target genes.[100–103] Genome-wide datasets can be used as a resource for detecting the physical interactions between GWAS variants and nearby gene promoters or established enhancers. Genome-wide chromatin conformation datasets have been generated in multiple cell types[104–106] and can be visualized with tools such as the 3D Genome Browser,[107] the Hi-C Unifying Genomic Interrogator (HUGIn),[108] or Juicebox.[109] Resolution of the chromosomal regions involved in the interaction depends on the assay and/or sequencing depth.[105,110] Cell type is an important consideration given that chromatin conformation assays identify active regulatory elements, which can differ between cell types. Limitations of using chromatin contacts to identify candidate genes are that significance thresholds are not well established, the strategy is uninformative when the candidate variants are located in close proximity to a target gene as a result of the large number of chance contacts between adjacent genome sites, chromatin loops can be general mechanisms of gene regulation and not relevant to GWAS mechanisms, and physical interaction of chromatin does not guarantee a consequence of that interaction.

Coding Variation. Coding variants in genes that are associated with a complex disease or trait can point to target genes at a locus. Numerous resources exist for annotating variants as nonsense, frameshift, and splice altering, all of which are expected to lead to loss of function of a gene,[111] and for predicting which missense variants are most likely to alter protein function according to evolutionary, biochemical, and structural information. In addition, multiple, typically rare, coding variants can independently associate with the same phenotype as a GWAS signal. For example, at a GWAS locus for type 2 diabetes, a study of coding variants in multiple populations identified 12 rare protein-truncating variants in *SLC30A8* (MIM: 611145), associated with decreased disease risk,[112] and of 31 rare (MAF $<$ 1%) coding or splice-site variants newly identified to be independently associated with hematological traits, 30 mapped to loci previously implicated in hematopoiesis by GWASs.[113] However, coding variants are not necessarily causal and can be in LD with variants affecting other genes.[114,115] When mechanistic consequences are considered, coding variants located at GWAS loci can help recognize biologically relevant genes.

Functional Studies. Functional studies of genes at GWAS loci can provide insight into target genes. Studies including overexpression, knockdown, or knockout of a target gene in cells or model organisms with the use of plasmids, viruses, oligos, or CRISPR-Cas9 approaches can point to genes with relevant biology.[116] For example, functional studies in mice identified a role of *CDKAL1* (MIM: 611259) in insulin secretion, and variants near *CDKAL1* are associated with type 2 diabetes.[117] Studies of gene function can provide insight into the biochemical pathways, protein interactions, and relevant cell or tissue types. Of note, whereas recapitulating some phenotypes in model systems is straightforward, recapitulating many others, especially neurological phenotypes, is very difficult. Deletion or substitution of a noncoding region containing candidate variants and subsequent analysis of gene expression and/or function can also point to target genes.[46,47] Another approach is to epigenetically alter a regulatory element by using a "dead" Cas9 with no nuclease activity tethered to an "epigenetic switch" consisting of chromatin modifying domains to mimic either an enhancer or a repressor.[103,116] Studies in human cells are of particular interest when a gene function is cell autonomous, and experiments in model organisms are especially informative when the biological process or pathway and disease context is well represented in the organism.

Candidate Genes at GWAS Signals

Evaluation of eQTLs colocalized with GWAS signals can point to one or more than one target gene. For example, a study of seven autoimmune diseases found colocalized eQTLs in immune cell types for 91 unique GWAS loci, including more than one gene (up to as many as 12) at 31 loci.[118] A study of cardiometabolic traits found colocalized subcutaneous adipose tissue eQTLs for 109 GWAS loci, including more than one gene (up to as many as five) at 25 loci.[119] Analyses of GTEx data found that among the GWAS signals that co-localized with an eQTL in one or more tissues, up to 62% of the signals co-localized with more than one gene.[41,88] In addition, recent eQTL studies have identified conditionally distinct eQTL signals, including some that colocalize with GWAS signals.[88,120–122] Although the associated changes in gene expression levels might not all have an effect on phenotype, the many GWAS loci with more than one colocalized eQTL provide many examples where more than one gene can be causal.

Although chromatin interactions often suggest multiple plausible target genes, they can provide stronger evidence when observed with colocalized eQTLs or functional evidence. At the *ARL15* locus (associated with insulin resistance traits and type 2 diabetes), HiC data in multiple cell types show significant interaction between the GWAS variant region and the *FST* (MIM: 136470) promoter,[108] ~500 kb away, and these data support evidence that the GWAS signals exhibit significant colocalized eQTLs for *FST*,[83] although functional studies support a role for *ARL15* in insulin secretion.[123] At a schizophrenia-associated locus, variants were linked by both HiC and genome editing to *FOXG1* (MIM: 164874) >700 kb away,[124] and at the *TMEM106B* (MIM: 613413) GWAS locus (associated with dementia), a variant in a CTCF binding site altered chromatin architecture to change *TMEM106B* expression and cell toxicity.[101] Recent analysis with ATAC-seq open-chromatin data identified over 15,000 interactions between distal regions of open chromatin, and allele-specific effects were inferred from the read data, resulting in improved fine-mapping of eQTL signals.[125]

Chromatin interactions with a GWAS variant region can also suggest multiple candidate genes. For example, at a breast-cancer-associated locus at 11q13, the GWAS variant region interacted with two long noncoding RNAs, *LINC01488* (MIM: 617696) and *CUPID2* (MIM: 617697), in allele-specific 3C experiments.[103] At a locus associated with schizophrenia and bipolar disorder, an enhancer variant interacted with two nearby miRNAs by 3C, supported by transcriptional activity and transcription factor binding assays.[100] At the *FTO* (MIM: 610966) locus (associated with obesity), the GWAS variants showed physical chromatin interaction with eight genes, at least two of which (*IRX3* [MIM: 612985] and *IRX5* [MIM: 606195]) showed genotype-expression association in preadipocytes.[102] Although the chromatin interactions might not all lead to an effect on phenotype, the GWAS loci with more than one interacting promoter provide examples where more than one gene can be causal.

Coding variants in genes can point to target genes, and rare coding variants can provide evidence for the target genes of common noncoding variants at the same locus. At the fasting-glucose-associated GWAS locus near *G6PC2* and *ABCB11* (MIM: 603201), for which the target gene was unknown, exome array data from ~33,000 individuals identified three rare *G6PC2* coding variants associated with fasting glucose,[42] suggesting that *G6PC2* might also be altered by the noncoding variants. *In vitro* assays confirmed the effect of the coding variants and established *G6PC2* as an effector gene and likely target gene of the noncoding variants at this GWAS locus.

Experiments in model organisms also identify multiple target genes at GWAS loci. At the *Agtrap-Plod1* locus (associated with blood pressure), six nearby potential target genes exist. Flister et al. performed mutagenesis in mice to create mutant strains for all six of these genes. They found that five genes showed an effect on blood pressure or renal function, suggesting that there might be multiple target genes at this GWAS locus.[126] One important consideration is that gene function is dependent on cell type and cellular context and might not be replicated in humans versus model organisms.

Genome and epigenome editing of regulatory regions pinpoints target genes. A noncoding region near *PHACTR1* (MIM: 608723) and *EDN1* (MIM: 131240) is associated with five vascular diseases—coronary artery disease, migraine, dissection, fibromuscular dysplasia, and hypertension. Deletion of 88 bp surrounding a fine-mapped SNP resulted in increased expression of *EDN1*, but not *PHACTR1* or four other nearby genes,[47] providing initial evidence of *EDN1* as a target gene. *EDN1* has known roles in the physiology of vasculature, suggesting a mechanism for the associated diseases.[47] In a second example, at the

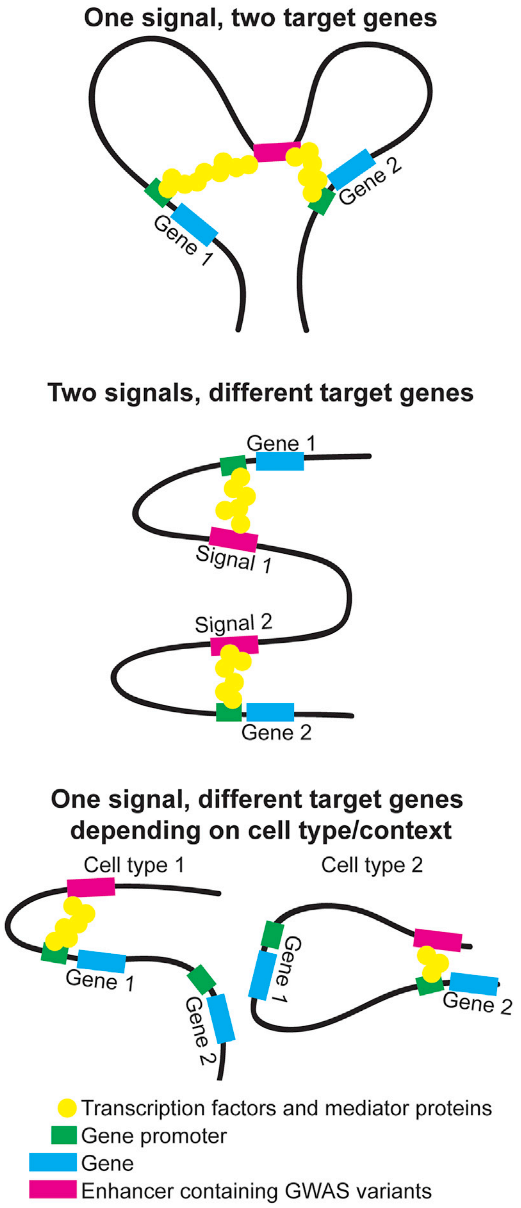
One signal, two target genes
Gene 1
Gene 2
Two signals, different target genes
Gene 1
Signal 1
Signal 2
Gene 2
One signal, different target genes depending on cell type/context
Cell type 1
Cell type 2
Gene 1
Gene 2
Gene 1
Gene 2
● Transcription factors and mediator proteins
🟩 Gene promoter
🟦 Gene
🟥 Enhancer containing GWAS variants

Figure 3. Hypotheses for Multiple Target Genes at a GWAS Locus
Multiple target genes can be present at a single GWAS locus. Three examples show how multiple genes might be affected. At a locus with one GWAS signal, an enhancer containing GWAS variants could target two genes simultaneously, or different genes could be targeted depending on cell type or cellular context. At a locus with two signals, each signal could target different genes. Other mechanisms could exist.

ADCY5 (MIM: 600293) GWAS locus (associated with type 2 diabetes), deletion of the orthologous associated regulatory region in rat pancreatic islet cells resulted in decreased *ADCY5* expression and reduced insulin secretion, supporting other evidence that *ADCY5* is a plausible target gene at the GWAS signal.[46] At the 11q13 GWAS locus (associated with breast cancer), epigenetically silencing a regulatory region with nuclease-inactivated Cas9 fused to the Kruppel-associated (KRAB) repressor reduced expression levels of three targeting genes: *LINC01488*, *CUPID2*, and *CCND1* (MIM: 168461).[103] These experiments provide evidence that genome and epigenome editing can validate target genes *in vivo*.

Emerging molecular QTL associations, such as histone, splice, methylation, metabolite, and other endophenotype QTLs, will aid in the identification of target genes and biological mechanisms. For example, an analysis that combined eQTLs, histone QTLs, splicing QTLs, and methylation QTLs to annotate 41 diseases and complex traits found that these QTLs are strongly enriched with disease heritability and provide complementary information about disease.[127] Together, these molecular associations can suggest disease mechanisms.

Conclusions

Each approach to identifying genes at a GWAS signal can provide evidence supporting a potential contribution of more than one gene. As expression datasets increase in size; the number of tissues, cell types, and contexts; and gene, isoform, and exon specificity, more colocalized eQTLs are being identified. Similarly, as chromatin-interaction datasets are generated in additional tissues, cell types, and contexts, more GWAS signals can be connected to one or more genes. Functional assays can suggest different genes depending on cell type, cellular environment, or other factors. In addition, two signals at a locus can act on the same or different genes; variants could target multiple genes via chromatin looping or different genes via tissue-specific enhancers (Figure 3). If multiple genes show strong candidacy, researchers should consider pursuing both genes in functional experiments because they could both be true target genes. Lack of support for a gene from any one approach could reflect that data are not available in the appropriate cell type or environmental state. Given limitations in concluding causality, multiple lines of genetic, bioinformatic, and experimental evidence supporting the role of a gene strengthen its candidacy.

Concluding Remarks

In this review, we have outlined three important aspects of evaluating GWAS loci (Figure 1). Generally, multiple pieces of evidence supporting a gene or variant that affects a complex trait can show a consistent direction and a single mechanism. However, given the contributions of multiple genes and variants at complex-trait loci and the imperfect nature of experimental systems, some evidence might not fit a simple model. When interpreting the results of a computational or experimental analysis, especially unexpected results, researchers should consider that additional signals might exist at a locus and that variants not considered candidates according to LD might nonetheless contribute to the mechanism of the locus. When searching for target genes, consider that variants might act through more than one nearby gene to influence disease. When identifying a variant that exhibits allelic effects on a gene, consider that additional variants might also have functional effects. Finally, when evaluating the biological effects of genes on disease, consider that cell type, cellular context, and multiple molecular mechanisms acting together can affect disease pathogenesis.

As GWASs are performed in more samples and additional populations, more loci with multiple signals and variants will be identified. The future is bright, given that progress is being made more quickly with high-throughput assays and with genome-editing experiments in the native chromatin context. Better statistical methods are continually being developed for identifying and localizing loci, signals, genes, and variants. The field can look forward to a better understanding of gene regulation, biological mechanisms, and disease pathways by closely examining GWAS loci.

Acknowledgments

The authors thank James Davis, Cassandra Spracklen, Chelsea Raulerson, and Kevin Currin for helpful discussion and feedback. This work was supported by National Institutes of Health grants F31 HL127984, U01 DK105561, R01 DK 093757, and R01 DK072193.

References

1. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS discovery: Biology, function, and translation. Am. J. Hum. Genet. *101*, 5–22.
2. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. *45* (D1), D896–D901.
3. Gallagher, M.D., and Chen-Plotkin, A.S. (2018). The post-GWAS era: From association to function. Am. J. Hum. Genet. *102*, 717–730.
4. Cutting, G.R. (2015). Cystic fibrosis genetics: from molecular understanding to clinical application. Nat. Rev. Genet. *16*, 45–56.
5. Rees, D.C., Williams, T.N., and Gladwin, M.T. (2010). Sickle-cell disease. Lancet *376*, 2018–2031.
6. Legare, M.E., Bartlett, F.S., 2nd, and Frankel, W.N. (2000). A major effect QTL determined by multiple genes in epileptic EL mice. Genome Res. *10*, 42–48.
7. Dixit, S., Kumar Biswal, A., Min, A., Henry, A., Oane, R.H., Raorane, M.L., Longkumer, T., Pabuayon, I.M., Mutte, S.K., Vardarajan, A.R., et al. (2015). Action of multiple intra-QTL genes concerted around a co-localized transcription factor underpins a large effect QTL. Sci. Rep. *5*, 15183.
8. She, R., and Jarosz, D.F. (2018). Mapping causal variants with single-nucleotide resolution reveals biochemical drivers of phenotypic change. Cell *172*, 478–490.e15.
9. Logan, R.W., Robledo, R.F., Recla, J.M., Philip, V.M., Bubier, J.A., Jay, J.J., Harwood, C., Wilcox, T., Gatti, D.M., Bult, C.J., et al. (2013). High-precision genetic mapping of behavioral traits in the diversity outbred mouse population. Genes Brain Behav. *12*, 424–437.
10. Edwards, S.L., Beesley, J., French, J.D., and Dunning, A.M. (2013). Beyond GWASs: illuminating the dark road from association to function. Am. J. Hum. Genet. *93*, 779–797.
11. Thibodeau, S.N., French, A.J., McDonnell, S.K., Cheville, J., Middha, S., Tillmans, L., Riska, S., Baheti, S., Larson, M.C., Fogarty, Z., et al. (2015). Identification of candidate genes for prostate cancer-risk SNPs utilizing a normal prostate tissue eQTL data set. Nat. Commun. *6*, 8653.
12. Nejentsev, S., Walker, N., Riches, D., Egholm, M., and Todd, J.A. (2009). Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science *324*, 387–389.
13. Hugot, J.P., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J.P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C.A., Gassull, M., et al. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature *411*, 599–603.
14. Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., et al.; National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC); United Kingdom Inflammatory Bowel Disease Genetics Consortium; and International Inflammatory Bowel Disease Genetics Consortium (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. *43*, 1066–1073.
15. Willer, C.J., Sanna, S., Jackson, A.U., Scuteri, A., Bonnycastle, L.L., Clarke, R., Heath, S.C., Timpson, N.J., Najjar, S.S., Stringham, H.M., et al. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat. Genet. *40*, 161–169.
16. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.
17. Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T.L., Thompson, J.R., Ingelsson, E., Saleheen, D., Erdmann, J., Goldstein, B.A., et al.; CARDIoGRAMplusC4D Consortium; DIAGRAM Consortium; CARDIOGENICS Consortium; MuTHER Consortium; and Wellcome Trust Case Control Consortium (2013). Large-scale association analysis identifies new risk loci for coronary artery disease. Nat. Genet. *45*, 25–33.
18. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
19. Tewhey, R., Bansal, V., Torkamani, A., Topol, E.J., and Schork, N.J. (2011). The importance of phase information for human genomics. Nat. Rev. Genet. *12*, 215–223.
20. Schaid, D.J., Rowland, C.M., Tines, D.E., Jacobson, R.M., and Poland, G.A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am. J. Hum. Genet. *70*, 425–434.
21. Datta, A.S., and Biswas, S. (2016). Comparison of haplotype-based statistical tests for disease association with rare and common variants. Brief. Bioinform. *17*, 657–671.
22. Wood, A.R., Hernandez, D.G., Nalls, M.A., Yaghootkar, H., Gibbs, J.R., Harries, L.W., Chong, S., Moore, M., Weedon, M.N., Guralnik, J.M., et al. (2011). Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. Hum. Mol. Genet. *20*, 4082–4092.
23. Khetarpal, S.A., Edmondson, A.C., Raghavan, A., Neeli, H., Jin, W., Badellino, K.O., Demissie, S., Manning, A.K., DerOhannessian, S.L., Wolfe, M.L., et al. (2011). Mining the LIPG allelic spectrum reveals the contribution of rare and common regulatory variants to HDL cholesterol. PLoS Genet. *7*, e1002393.
24. Castel, S.E., Cervera, A., Mohammadi, P., Aguet, F., Reverter, F., Wolman, A., Guigo, R., Iossifov, I., Vasileva, A., and

Lappalainen, T. (2018). Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. Nat. Genet. *50*, 1327–1334.

25. Huang, J., Ellinghaus, D., Franke, A., Howie, B., and Li, Y. (2012). 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. Eur. J. Hum. Genet. *20*, 801–805.
26. Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., Lawson, D., et al.; UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.
27. Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature *466*, 707–713.
28. Hoffmann, T.J., Theusch, E., Haldar, T., Ranatunga, D.K., Jorgenson, E., Medina, M.W., Kvale, M.N., Kwok, P.Y., Schaefer, C., Krauss, R.M., et al. (2018). A large electronic-health-record-based genome-wide study of serum lipids. Nat. Genet. *50*, 401–413.
29. Sung, Y.J., Winkler, T.W., de Las Fuentes, L., Bentley, A.R., Brown, M.R., Kraja, A.T., Schwander, K., Ntalla, I., Guo, X., Franceschini, N., et al.; CHARGE Neurology Working Group; COGENT-Kidney Consortium; GIANT Consortium; and Lifelines Cohort Study (2018). A large-scale multi-ancestry genome-wide study accounting for smoking behavior identifies multiple significant loci for blood pressure. Am. J. Hum. Genet. *102*, 375–400.
30. Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A., et al. (2016). The allelic landscape of human blood cell trait variation and links to common complex disease. Cell *167*, 1415–1429.e19.
31. Fritsche, L.G., Igl, W., Bailey, J.N., Grassmann, F., Sengupta, S., Bragg-Gresham, J.L., Burdon, K.P., Hebbring, S.J., Wen, C., Gorski, M., et al. (2016). A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nat. Genet. *48*, 134–143.
32. Davis, J.P., Huyghe, J.R., Locke, A.E., Jackson, A.U., Sim, X., Stringham, H.M., Teslovich, T.M., Welch, R.P., Fuchsberger, C., Narisu, N., et al. (2017). Common, low-frequency, and rare genetic variants associated with lipoprotein subclasses and triglyceride measures in Finnish men from the METSIM study. PLoS Genet. *13*, e1007079.
33. Galarneau, G., Palmer, C.D., Sankaran, V.G., Orkin, S.H., Hirschhorn, J.N., and Lettre, G. (2010). Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. Nat. Genet. *42*, 1049–1051.
34. Sanna, S., Li, B., Mulas, A., Sidore, C., Kang, H.M., Jackson, A.U., Piras, M.G., Usala, G., Maninchedda, G., Sassu, A., et al. (2011). Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. PLoS Genet. *7*, e1002198.
35. Wu, Y., Waite, L.L., Jackson, A.U., Sheu, W.H., Buyske, S., Absher, D., Arnett, D.K., Boerwinkle, E., Bonnycastle, L.L., Carty, C.L., et al. (2013). Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained. PLoS Genet. *9*, e1003379.
36. Chatterjee, S., Kapoor, A., Akiyama, J.A., Auer, D.R., Lee, D., Gabriel, S., Berrios, C., Pennacchio, L.A., and Chakravarti, A. (2016). Enhancer variants synergistically drive dysfunction of a gene regulatory network in hirschsprung disease. Cell *167*, 355–368.e10.
37. Lyssenko, V., Lupi, R., Marchetti, P., Del Guerra, S., Orho-Melander, M., Almgren, P., Sjögren, M., Ling, C., Eriksson, K.F., Lethagen, A.L., et al. (2007). Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes. J. Clin. Invest. *117*, 2155–2163.
38. Gaulton, K.J., Nammo, T., Pasquali, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuis, T.M., Mieczkowski, P., Secchi, A., Bosco, D., et al. (2010). A map of open chromatin in human pancreatic islets. Nat. Genet. *42*, 255–259.
39. Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J., Steinthorsdottir, V., Scott, R.A., Grarup, N., et al. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat. Genet. Published online October 8, 2018. https://doi.org/10.1038/s41588-018-0241-6.
40. Mondal, A.K., Das, S.K., Baldini, G., Chu, W.S., Sharma, N.K., Hackney, O.G., Zhao, J., Grant, S.F., and Elbein, S.C. (2010). Genotype and tissue-specific effects on alternative splicing of the transcription factor 7-like 2 gene in humans. J. Clin. Endocrinol. Metab. *95*, 1450–1457.
41. Gamazon, E.R., Segrè, A.V., van de Bunt, M., Wen, X., Xi, H.S., Hormozdiari, F., Ongen, H., Konkashbaev, A., Derks, E.M., Aguet, F., et al.; GTEx Consortium (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nat. Genet. *50*, 956–967.
42. Mahajan, A., Sim, X., Ng, H.J., Manning, A., Rivas, M.A., Highland, H.M., Locke, A.E., Grarup, N., Im, H.K., Cingolani, P., et al.; T2D-GENES consortium and GoT2D consortium (2015). Identification and functional characterization of G6PC2 coding variants influencing glycemic traits define an effector transcript at the G6PC2-ABCB11 locus. PLoS Genet. *11*, e1004876.
43. Gregory, A.P., Dendrou, C.A., Attfield, K.E., Haghikia, A., Xifara, D.K., Butter, F., Poschmann, G., Kaur, G., Lambert, L., Leach, O.A., et al. (2012). TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis. Nature *488*, 508–511.
44. Allen, E.K., Randolph, A.G., Bhangale, T., Dogra, P., Ohlson, M., Oshansky, C.M., Zamora, A.E., Shannon, J.P., Finkelstein, D., Dressen, A., et al. (2017). SNP-mediated disruption of CTCF binding at the IFITM3 promoter is associated with risk of severe influenza in humans. Nat. Med. *23*, 975–983.
45. Sankaran, V.G., Ludwig, L.S., Sicinska, E., Xu, J., Bauer, D.E., Eng, J.C., Patterson, H.C., Metcalf, R.A., Natkunam, Y., Orkin, S.H., et al. (2012). Cyclin D3 coordinates the cell cycle during differentiation to regulate erythrocyte size and number. Genes Dev. *26*, 2075–2087.
46. Roman, T.S., Cannon, M.E., Vadlamudi, S., Buchkovich, M.L., Wolford, B.N., Welch, R.P., Morken, M.A., Kwon, G.J., Varshney, A., Kursawe, R., et al.; National Institutes of Health Intramural Sequencing Center (NISC) Comparative Sequencing Program (2017). A type 2 diabetes-associated functional regulatory variant in a pancreatic islet enhancer at the *ADCY5* locus. Diabetes *66*, 2521–2530.
47. Gupta, R.M., Hadaya, J., Trehan, A., Zekavat, S.M., Roselli, C., Klarin, D., Emdin, C.A., Hilvering, C.R.E., Bianchi, V., Mueller, C., et al. (2017). A genetic variant associated with

five vascular diseases is a distal regulator of endothelin-1 gene expression. Cell *170*, 522–533.e15.

48. Roman, T.S., Marvelle, A.F., Fogarty, M.P., Vadlamudi, S., Gonzalez, A.J., Buchkovich, M.L., Huyghe, J.R., Fuchsberger, C., Jackson, A.U., Wu, Y., et al. (2015). Multiple hepatic regulatory variants at the GALNT2 GWAS locus associated with high-density lipoprotein cholesterol. Am. J. Hum. Genet. *97*, 801–815.
49. Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal lari, R., Lupien, M., Markowitz, S., and Scacheri, P.C. (2014). Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. *24*, 1–13.
50. Stadhouders, R., Aktuna, S., Thongjuea, S., Aghajanirefah, A., Pourfarzad, F., van Ijcken, W., Lenhard, B., Rooks, H., Best, S., Menzel, S., et al. (2014). HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. J. Clin. Invest. *124*, 1699–1710.
51. Morris, A.P. (2011). Transethnic meta-analysis of genomewide association studies. Genet. Epidemiol. *35*, 809–822.
52. Spain, S.L., and Barrett, J.C. (2015). Strategies for fine-mapping complex traits. Hum. Mol. Genet. *24* (R1), R111–R119.
53. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. Genetics *198*, 497–508.
54. Benner, C., Spencer, C.C., Havulinna, A.S., Salomaa, V., Ripatti, S., and Pirinen, M. (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics *32*, 1493–1501.
55. Kichaev, G., Yang, W.Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. *10*, e1004722.
56. Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573.
57. Wen, X., Lee, Y., Luca, F., and Pique-Regi, R. (2016). Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. Am. J. Hum. Genet. *98*, 1114–1129.
58. Crawford, N.G., Kelly, D.E., Hansen, M.E.B., Beltrame, M.H., Fan, S., Bowman, S.L., Jewett, E., Ranciaro, A., Thompson, S., Lo, Y., et al.; NISC Comparative Sequencing Program (2017). Loci associated with skin pigmentation identified in African populations. Science *358*, eaan8433.
59. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1081.
60. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
61. Schwarz, J.M., Cooper, D.N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. Nat. Methods *11*, 361–362.
62. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.
63. Gosalia, N., Economides, A.N., Dewey, F.E., and Balasubramanian, S. (2017). MAPPIN: a method for annotating, predicting pathogenicity and mode of inheritance for nonsynonymous variants. Nucleic Acids Res. *45*, 10393–10402.
64. Eilbeck, K., Quinlan, A., and Yandell, M. (2017). Settling the score: variant prioritization and Mendelian disease. Nat. Rev. Genet. *18*, 599–612.
65. Jian, X., Boerwinkle, E., and Liu, X. (2014). In silico tools for splicing defect prediction: a survey from the viewpoint of end users. Genet. Med. *16*, 497–503.
66. Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113.
67. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol. *28*, 1045–1048.
68. Stunnenberg, H.G., Hirst, M.; and International Human Epigenome Consortium (2016). The International Human Epigenome Consortium: A blueprint for scientific collaboration and discovery. Cell *167*, 1145–1149.
69. Schmidt, E.M., Zhang, J., Zhou, W., Chen, J., Mohlke, K.L., Chen, Y.E., and Willer, C.J. (2015). GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. Bioinformatics *31*, 2601–2606.
70. Iotchkova, V., Ritchie, G.R.S., Geihs, M., Morganella, S., Min, J.L., Walter, K., Timpson, N.J., UK10K Consortium, Dunham, I., Birney, E., et al. (2018). GARFIELD - GWAS analysis of regulatory or functional information enrichment with LD correction. bioRxiv. https://doi.org/10.1101/085738.
71. Degner, J.F., Marioni, J.C., Pai, A.A., Pickrell, J.K., Nkadori, E., Gilad, Y., and Pritchard, J.K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics *25*, 3207–3212.
72. Reddy, T.E., Gertz, J., Pauli, F., Kucera, K.S., Varley, K.E., Newberry, K.M., Marinov, G.K., Mortazavi, A., Williams, B.A., Song, L., et al. (2012). Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. Genome Res. *22*, 860–869.
73. Buchkovich, M.L., Eklund, K., Duan, Q., Li, Y., Mohlke, K.L., and Furey, T.S. (2015). Removing reference mapping biases using limited or no genotype data identifies allelic differences in protein binding at disease-associated loci. BMC Med. Genomics *8*, 43.
74. Harvey, C.T., Moyerbrailean, G.A., Davis, G.O., Wen, X., Luca, F., and Pique-Regi, R. (2015). QuASAR: quantitative allele-specific analysis of reads. Bioinformatics *31*, 1235–1242.
75. Choi, J., Xu, M., Makowski, M.M., Zhang, T., Law, M.H., Kovacs, M.A., Granzhan, A., Kim, W.J., Parikh, H., Gartside, M., et al. (2017). A common intronic variant of PARP1 confers melanoma risk and mediates melanocyte growth via regulation of MITF. Nat. Genet. *49*, 1326–1335.
76. MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. Nature *508*, 469–476.
77. Ward, L.D., and Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. Nat. Biotechnol. *30*, 1095–1106.
78. Starita, L.M., Ahituv, N., Dunham, M.J., Kitzman, J.O., Roth, F.P., Seelig, G., Shendure, J., and Fowler, D.M. (2017). Variant

interpretation: Functional assays to the rescue. Am. J. Hum. Genet. *101*, 315–325.

79. Stormo, G.D., and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. Nat. Rev. Genet. *11*, 751–760.

80. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327–339.

81. Zhang, F., Wen, Y., and Guo, X. (2014). CRISPR/Cas9 for genome editing: Progress, implications and challenges. Hum. Mol. Genet. *23* (R1), R40–R46.

82. Thakore, P.I., D'Ippolito, A.M., Song, L., Safi, A., Shivakumar, N.K., Kabadi, A.M., Reddy, T.E., Crawford, G.E., and Gersbach, C.A. (2015). Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. Nat. Methods *12*, 1143–1149.

83. Cannon, M.E., Duan, Q., Wu, Y., Zeynalzadeh, M., Xu, Z., Kangas, A.J., Soininen, P., Ala-Korpela, M., Civelek, M., Lusis, A.J., et al. (2017). *Trans*-ancestry fine mapping and molecular assays identify regulatory variants at the *ANGPTL8* HDL-C GWAS locus. G3 (Bethesda) *7*, 3217–3227.

84. Spisák, S., Lawrenson, K., Fu, Y., Csabai, I., Cottman, R.T., Seo, J.H., Haiman, C., Han, Y., Lenci, R., Li, Q., et al.; GAME-ON/ELLIPSE Consortium (2015). CAUSEL: an epigenome- and genome-editing pipeline for establishing function of noncoding GWAS variants. Nat. Med. *21*, 1357–1363.

85. Guo, C., Ludvik, A.E., Arlotto, M.E., Hayes, M.G., Armstrong, L.L., Scholtens, D.M., Brown, C.D., Newgard, C.B., Becker, T.C., Layden, B.T., et al. (2015). Coordinated regulatory variation associated with gestational hyperglycaemia regulates expression of the novel hexokinase HKDC1. Nat. Commun. *6*, 6069.

86. Ulirsch, J.C., Nandakumar, S.K., Wang, L., Giani, F.C., Zhang, X., Rogov, P., Melnikov, A., McDonel, P., Do, R., Mikkelsen, T.S., and Sankaran, V.G. (2016). Systematic functional dissection of common genetic variation affecting red blood cell traits. Cell *165*, 1530–1545.

87. Bojesen, S.E., Pooley, K.A., Johnatty, S.E., Beesley, J., Michailidou, K., Tyrer, J.P., Edwards, S.L., Pickett, H.A., Shen, H.C., Smart, C.E., et al.; Australian Cancer Study; Australian Ovarian Cancer Study; Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer (kConFab); Gene Environment Interaction and Breast Cancer (GENICA); Swedish Breast Cancer Study (SWE-BRCA); Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON); Epidemiological study of BRCA1 & BRCA2 Mutation Carriers (EMBRACE); and Genetic Modifiers of Cancer Risk in BRCA1/2 Mutation Carriers (GEMO) (2013). Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. Nat. Genet. *45*, 371–384, e1–e2.

88. Battle, A., Brown, C.D., Engelhardt, B.E., Montgomery, S.B.; GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration & Visualization—EBI; Genome Browser Data Integration & Visualization—UCSC Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis & Coordinating Center (LDACC); NIH program management; Biospecimen collection; Pathology; and eQTL manuscript working group (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

89. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. *10*, e1004383.

90. Hormozdiari, F., van de Bunt, M., Segrè, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL signals detects target genes. Am. J. Hum. Genet. *99*, 1245–1260.

91. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., and Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. *48*, 481–487.

92. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., Cox, N.J., Im, H.K.; and GTEx Consortium (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.

93. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.

94. Manor, O., and Segal, E. (2013). Robust prediction of expression differences among human individuals using only genotype information. PLoS Genet. *9*, e1003396.

95. Veturi, Y., and Ritchie, M.D. (2018). How powerful are summary-based methods for identifying expression-trait associations under different genetic architectures? Pac. Symp. Biocomput. *23*, 228–239.

96. Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H.K., Reshef, Y., Song, L., Safi, A., McCarroll, S., Neale, B.M., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2018). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nat. Genet. *50*, 538–548.

97. Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., and Pasaniuc, B. (2017). Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. Am. J. Hum. Genet. *100*, 473–487.

98. Belmont, A.S. (2014). Large-scale chromatin organization: the good, the surprising, and the still perplexing. Curr. Opin. Cell Biol. *26*, 69–78.

99. Risca, V.I., and Greenleaf, W.J. (2015). Unraveling the 3D genome: genomics tools for multiscale exploration. Trends Genet. *31*, 357–372.

100. Duan, J., Shi, J., Fiorentino, A., Leites, C., Chen, X., Moy, W., Chen, J., Alexandrov, B.S., Usheva, A., He, D., et al.; Molecular Genetics of Schizophrenia collaboration; and Genomic Psychiatric Cohort consortium (2014). A rare functional noncoding variant at the GWAS-implicated MIR137/MIR2682 locus might confer risk to schizophrenia and bipolar disorder. Am. J. Hum. Genet. *95*, 744–753.

101. Gallagher, M.D., Posavi, M., Huang, P., Unger, T.L., Berlyand, Y., Gruenewald, A.L., Chesi, A., Manduchi, E., Wells, A.D.,

Grant, S.F.A., et al. (2017). A dementia-associated risk variant near TMEM106B alters chromatin architecture and gene expression. Am. J. Hum. Genet. *101*, 643–663.

102. Claussnitzer, M., Dankel, S.N., Kim, K.H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., et al. (2015). FTO obesity variant circuitry and adipocyte browning in humans. N. Engl. J. Med. *373*, 895–907.

103. Betts, J.A., Moradi Marjaneh, M., Al-Ejeh, F., Lim, Y.C., Shi, W., Sivakumaran, H., Tropée, R., Patch, A.M., Clark, M.B., Bartonicek, N., et al. (2017). Long noncoding RNAs CUPID1 and CUPID2 mediate breast cancer risk at 11q13 by modulating the response to DNA damage. Am. J. Hum. Genet. *101*, 255–266.

104. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

105. Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

106. Montefiori, L.E., Sobreira, D.R., Sakabe, N.J., Aneas, I., Joslin, A.C., Hansen, G.T., Bozek, G., Moskowitz, I.P., McNally, E.M., and Nóbrega, M.A. (2018). A promoter interaction map for cardiovascular disease genetics. eLife *7*, e35788.

107. Wang, Y., Song, F., Zhang, B., Zhang, L., Xu, J., Kuang, D., Li, D., Choudhary, M.N.K., Li, Y., Hu, M., et al. (2018). The 3D Genome Browser: A web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol. *19*, 151.

108. Martin, J.S., Xu, Z., Reiner, A.P., Mohlke, K.L., Sullivan, P., Ren, B., Hu, M., and Li, Y. (2017). HUGIn: Hi-C Unifying Genomic Interrogator. Bioinformatics *33*, 3793–3795.

109. Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., and Aiden, E.L. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. *3*, 99–101.

110. Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290–294.

111. Cooper, G.M., and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. *12*, 628–640.

112. Flannick, J., Thorleifsson, G., Beer, N.L., Jacobs, S.B., Grarup, N., Burtt, N.P., Mahajan, A., Fuchsberger, C., Atzmon, G., Benediktsson, R., et al.; Go-T2D Consortium; and T2D-GENES Consortium (2014). Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. Nat. Genet. *46*, 357–363.

113. Mousas, A., Ntritsos, G., Chen, M.H., Song, C., Huffman, J.E., Tzoulaki, I., Elliott, P., Psaty, B.M., Auer, P.L., Johnson, A.D., et al.; Blood-Cell Consortium (2017). Rare coding variants pinpoint genes that control human hematological traits. PLoS Genet. *13*, e1006925.

114. Huyghe, J.R., Jackson, A.U., Fogarty, M.P., Buchkovich, M.L., Stančáková, A., Stringham, H.M., Sim, X., Yang, L., Fuchsberger, C., Cederberg, H., et al. (2013). Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat. Genet. *45*, 197–201.

115. Mahajan, A., Wessel, J., Willems, S.M., Zhao, W., Robertson, N.R., Chu, A.Y., Gan, W., Kitajima, H., Taliun, D., Rayner, N.W., et al.; ExomeBP Consortium; MAGIC Consortium; and GIANT Consortium (2018). Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. Nat. Genet. *50*, 559–571.

116. Tak, Y.G., and Farnham, P.J. (2015). Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. Epigenetics Chromatin *8*, 57.

117. Wei, F.Y., and Tomizawa, K. (2011). Functional loss of Cdkal1, a novel tRNA modification enzyme, causes the development of type 2 diabetes. Endocr. J. *58*, 819–825.

118. Chen, L., Ge, B., Casale, F.P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., et al. (2016). Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell *167*, 1398–1414.e24.

119. Civelek, M., Wu, Y., Pan, C., Raulerson, C.K., Ko, A., He, A., Tilford, C., Saleem, N.K., Stančáková, A., Scott, L.J., et al. (2017). Genetic regulation of adipose gene expression and cardio-metabolic traits. Am. J. Hum. Genet. *100*, 428–443.

120. Dobbyn, A., Huckins, L.M., Boocock, J., Sloofman, L.G., Glicksberg, B.S., Giambartolomei, C., Hoffman, G.E., Perumal, T.M., Girdhar, K., Jiang, Y., et al.; CommonMind Consortium (2018). Landscape of conditional eQTL in dorsolateral prefrontal cortex and co-localization with schizophrenia GWAS. Am. J. Hum. Genet. *102*, 1169–1184.

121. Jansen, R., Hottenga, J.J., Nivard, M.G., Abdellaoui, A., Laport, B., de Geus, E.J., Wright, F.A., Penninx, B.W.J.H., and Boomsma, D.I. (2017). Conditional eQTL analysis reveals allelic heterogeneity of gene expression. Hum. Mol. Genet. *26*, 1444–1451.

122. Zhernakova, D.V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., van 't Hof, P., Mei, H., van Dijk, F., Westra, H.J., et al. (2017). Identification of context-dependent expression quantitative trait loci in whole blood. Nat. Genet. *49*, 139–145.

123. Thomsen, S.K., Ceroni, A., van de Bunt, M., Burrows, C., Barrett, A., Scharfmann, R., Ebner, D., McCarthy, M.I., and Gloyn, A.L. (2016). Systematic functional characterization of candidate causal genes for type 2 diabetes risk variants. Diabetes *65*, 3805–3811.

124. Won, H., de la Torre-Ubieta, L., Stein, J.L., Parikshak, N.N., Huang, J., Opland, C.K., Gandal, M.J., Sutton, G.J., Hormozdiari, F., Lu, D., et al. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. Nature *538*, 523–527.

125. Kumasaka, N., Knights, A., and Gaffney, D. (2018). High resolution genetic mapping of causal regulatory interactions in the human genome. bioRxiv. https://doi.org/10.1101/227389.

126. Flister, M.J., Tsaih, S.W., O'Meara, C.C., Endres, B., Hoffman, M.J., Geurts, A.M., Dwinell, M.R., Lazar, J., Jacob, H.J., and Moreno, C. (2013). Identifying multiple causative genes at a single GWAS locus. Genome Res. *23*, 1996–2002.

127. Hormozdiari, F., Gazal, S., van de Geijn, B., Finucane, H.K., Ju, C.J., Loh, P.R., Schoech, A., Reshef, Y., Liu, X., O'Connor, L., et al. (2018). Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. Nat. Genet. *50*, 1041–1047.