CellPress

Trends in
Genetics

Review

Functional characterization of human genomic variation linked to polygenic diseases

Tania Fabo[1,2,3,4] and Paul Khavari[1,2,3,4,5,*]

The burden of human disease lies predominantly in polygenic diseases. Since the early 2000s, genome-wide association studies (GWAS) have identified genetic variants and loci associated with complex traits. These have ranged from variants in coding sequences to mutations in regulatory regions, such as promoters and enhancers, as well as mutations affecting mediators of mRNA stability and other downstream regulators, such as 5′ and 3′-untranslated regions (UTRs), long noncoding RNA (lncRNA), and miRNA. Recent research advances in genetics have utilized a combination of computational techniques, high-throughput *in vitro* and *in vivo* screening modalities, and precise genome editing to impute the function of diverse classes of genetic variants identified through GWAS. In this review, we highlight the vastness of genomic variants associated with polygenic disease risk and address recent advances in how genetic tools can be used to functionally characterize them.

Highlights

Genome-wide association studies (GWAS) have localized variants to diverse classes of genomic elements, including coding sequences, introns, promoters, enhancers, 5′ and 3′-untranslated regions (UTRs), long noncoding RNA (lncRNA), and miRNAs.

Annotation of diverse GWAS-identified loci from precise variant to biologic function to pathogenic impact requires understanding how different classes of genomic elements are affected by genetic variants and appropriately adapting a wide range of genomic tools to uncover variant function and effect.

A variety of high-throughput screening methods and informatics tools have been adapted to capture the specific ways in which different variant classes affect gene expression and/or function.

Precision gene-editing approaches that produce isogenic cells and tissues that differ only at the variant of interest comprise the gold standard for characterizing the effects of a GWAS variant in its native context.

Introduction

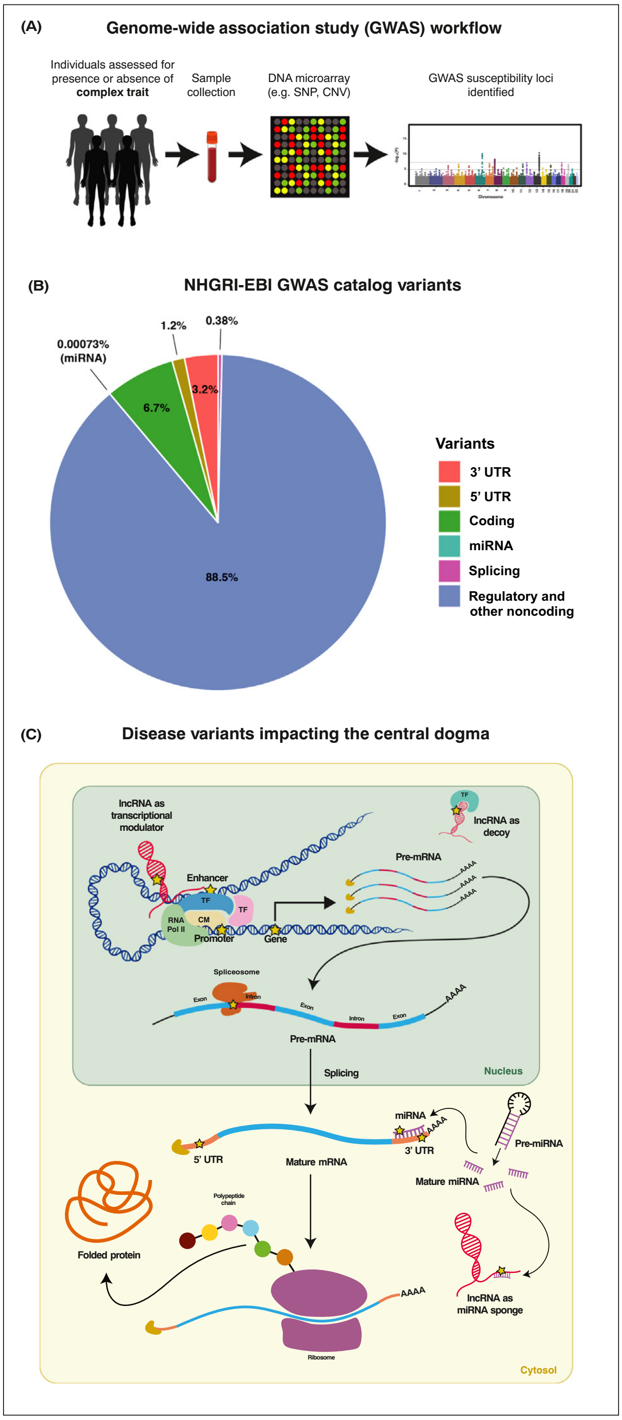
The burden of human diseases is dominated by complex, polygenic disease rather than by monogenic disorders associated with single, highly penetrant, genetic mutations. Since the mid-2000s, GWAS have been the preferred tool for identifying associations between genetic variants and complex traits [1]. While the association between genomic variation and phenotype has long been understood, GWAS have greatly expanded the landscape of genomic variants associated with traits and disease [2]. The NHGRI-EBI GWAS Catalog has almost 6000 GWAS entries published since 2005 [3]. From these studies have emerged over 400 000 associations between genetic variants and human traits. These trait- and disease-associated variants lie in a variety of genomic locales, including coding and noncoding regions (Figure 1), and illuminate the complex biological pathways that regulate the formation of a functional protein product from a gene and how variant-mediated perturbations in relevant pathways can promote the pathogenesis of polygenic disorders (Figures 1 and 2).

An important limitation of the GWAS approach is that the findings are just that: associations. A GWAS hit does not specify the identity of the precise causal variant; neither does it reveal the causal trait-associated gene or the perturbed biological process [2]. Rather, GWAS only identify associations between traits and the lead or index SNPs present in the assays used in these studies. Thus, for GWAS hits to be translated into any meaningful clinical impact, the true functional variants need to be identified and their impact on gene expression and trait development interrogated. In this review, we discuss efforts by researchers over the years to understand the diverse types of genomic variants associated with traits and disease by developing experimental techniques and tools tailored to the unique function and impact of the classes of variants being studied. We focus particularly on recent developments in high-throughput screening methodologies and computational tools, which have greatly facilitated ongoing efforts to functionally annotate and determine causal genes for the broad catalog of trait-associated genetic variants identified through GWAS.

[1]Program in Epithelial Biology, Stanford University, Stanford, CA, USA
[2]Stanford Cancer Institute, Stanford University, Stanford, CA, USA
[3]Graduate Program in Genetics, Stanford University, Stanford, CA, USA
[4]Stanford University School of Medicine, Stanford University, Stanford, CA, USA
[5]Veterans Affairs Palo Alto Healthcare System, Palo Alto, CA, USA

*Correspondence:
khavari@stanford.edu (P. Khavari).

https://doi.org/10.1016/j.tig.2023.02.014



Trends in Genetics

(See figure legend at the bottom of the next page.)

Understanding the variants

Coding variants

Long before the advent of GWAS, the connection between coding variation and a disease or trait had been well understood, particularly for Mendelian disease. However, since the early 2000s, greater efforts have been directed toward understanding the genes that drive complex traits and diseases. In their 2002 article, Glazier *et al.* provided a framework for identifying genes that drive complex traits, which has come to reflect efforts undertaken since to uncover coding variants that drive complex disease [4]. They suggest first performing association studies to identify genomic intervals (i.e., loci) associated with a complex (or quantitative) trait; that is, its quantitative trait loci (QTLs). This should be followed by statistical fine mapping to reduce the size of the candidate genomic interval, including linkage disequilibrium (LD) studies to determine how genetic markers across the locus associate with the trait of interest. Given that GWAS typically identify lead or index SNPs associated with disease rather than the causal SNP itself, statistical fine-mapping tools are frequently used to attempt to nominate the true causal SNP (or SNPs) for downstream functional studies. Statistical fine mapping typically entails identifying the region of interest, exploring the region for its LD structure and the genes and SNPs contained in that region, partitioning the region into independent regions that each affect the trait, and then using statistical fine-mapping methods (e.g., simple heuristic methods, penalized regression models, and Bayesian methods) to identify the SNP in the region most strongly associated with the trait or disease [5]. In their framework, Glazier *et al.* also noted the importance of sequencing the region of interest to identify coding variants, which further improves the power of statistical mapping. Lastly, they stressed the importance of performing functional assays to determine the role of a gene, especially as it relates to the trait, and the impact of a variant in that gene. These steps highlight the critical connection that must be made between QTL, variant, and gene to discover coding genetic variants that impact a trait. As researchers have endeavored over the past two decades to uncover genetic variants in coding regions that contribute to complex traits, varied approaches have been taken to connect coding variants and phenotype.

Researchers studying coding variation in complex traits before the development of GWAS commonly used a gene → variant approach, rather than the locus → variant → gene approach outlined by Glazier *et al.* [4]. Studies often focused on the identification and characterization of single nucleotide variants (SNVs)/SNPs and other genetic variants at gene loci already known to be involved in a particular trait or disease (Box 1). By contrast, GWAS is a locus-centric approach: GWAS identify a susceptibility locus associated with a trait, and further work is done to identify the genes and causal variants driving the trait. Thus, a GWAS approach allows for the discovery of novel genes and pathways associated with a trait, thereby enabling insight into the complicated puzzle of genetically complex polygenic traits. It is also most aligned with the QTL–variant–gene triad that Glazier *et al.* proposed for identifying genes that drive complex traits [4]. Studies that take this approach have often discovered and functionally characterized novel genes that may have a role in a trait. For example, Kim *et al.* used prior GWAS data on serum acylcarnitine levels to identify a susceptibility locus of interest [6]. Fine mapping of the locus determined acylcarnitine

Figure 1. Genome-wide association studies (GWAS) identify a diverse catalog of variants that affect gene expression and protein function across the central dogma. (A) Workflow for GWAS. (B) Distribution of GWAS-identified variants across variant classes (based on the 2022 GWAS Catalog), highlighting a large proportion of noncoding variants. 'Regulatory and Other Noncoding' includes variants designated as regulatory, intronic, and otherwise noncoding in the GWAS catalog. (C) Variants found to be associated with disease can disrupt gene expression and protein function through diverse mechanisms. Yellow stars highlight common genetic elements and mechanisms impacted by trait- and disease-associated variants. The diverse mechanisms highlight the need for diverse tools to functionally annotate variants. Abbreviations: CM, chromatin modifier; CNV, copy number variant; lncRNA, long noncoding RNA; TF, transcription factor; UTR, untranslated region.

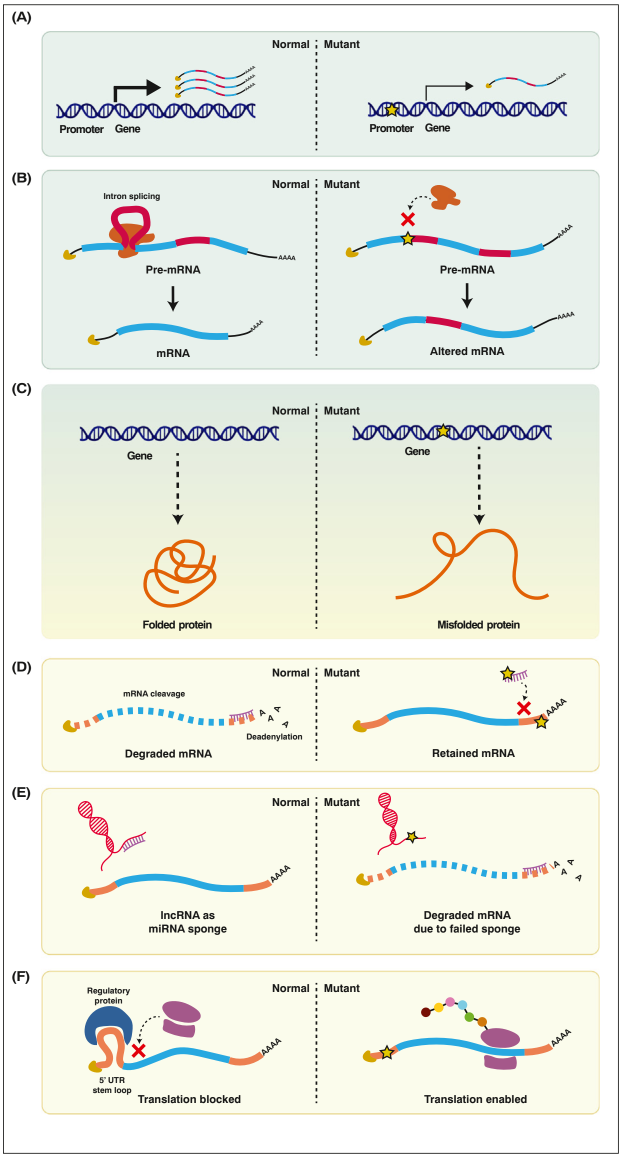


Figure 2. Variants in various classes of genomic elements have diverse effects on gene expression and protein translation. (A) Promoter and enhancer variants, as well as some types of long noncoding (lnc)RNA variants, disrupt the regulation of gene expression, altering transcription. (B) Splice site variants can result in abnormally spliced transcripts and the production of altered protein. (C) Coding variants can disrupt protein function by various mechanisms, including protein misfolding. (D) Variants in miRNA or 3′-untranslated region (UTR) sequences can disrupt miRNA-mediated regulation of mRNA transcript abundance, resulting in mRNA retention. (E) Variants in some types of lncRNAs can disrupt the ability of lncRNAs to function as an miRNA sponge, resulting in mRNA degradation. (F) 5′-UTR variants can alter binding of regulatory proteins, resulting in altered translation.

levels to be associated with variants in *SLC22A1*, which encodes a membrane transporter in the liver with no previously described role in acylcarnitine transport. Functional characterization of the gene and its human variants identified a novel role of SLC22A1 in mediating acylcarnitine efflux from the liver, and reduced protein levels and efflux activity resulting from genetic variants. In another study, Chang *et al.* performed exome-wide association analysis (similar to GWAS except for

Box 1. Pre-GWAS era characterization of coding variants

Pre-GWAS gene-centric studies often focused on the identification and characterization of SNPs and other genetic variants in genes already known to be involved in a particular trait or disease. For example, work by Blaisdell *et al.*, Dai *et al.*, and Zhang *et al.* identified novel SNPs in genes previously described to have a role in drug metabolism, *CYP2C19*, *CYP3A4*, and *PXR* [210–212]. Similarly, Cargill *et al.* compiled a list of 106 genes encoding proteins thought to have a role in common diseases, such as type 2 diabetes mellitus, schizophrenia, and coronary artery diseases, and identified 392 coding-region SNPs that might influence these diseases [213]. Sharon *et al.* focused their studies on the olfactory receptor locus and identified 21 novel coding SNPs [214]. The SNPs identified in these studies are often further characterized to determine their effect on the genes of interest, using methods such as site-directed mutagenesis to generate plasmids expressing the mutant proteins and study variant effects on DNA binding affinity, catalytic activity, expression stability, hydroxylation activity, and other functional assays relevant to the protein of interest [210–212]. What these studies have in common is their gene-centric approach, requiring *a priori* knowledge of the genes thought to have an important role the trait or disease. GWAS provides the advantage of nominating novel trait- and disease-associated loci and genes, and, thus, novel coding variants.

its exclusive focus on the human exome) of colorectal cancer cases and controls and identified a novel missense variant in *TCF7L2*, a transcription factor (TF) that activates transcription of *MYC*, a proto-oncogene [7]. The authors showed that the variant in *TCF7L2* decreases its activity, leading to decreased *MYC* expression and lower risk for colorectal cancer.

With methodological advances, the tools available to functionally characterize candidate genes identified in GWAS susceptibility loci and their coding variants have improved. *In silico* tools, such as the Ensembl Variant Effect Predictor, PredictSNP, dbNSFP, Revel, and Envision, have been used to predict the deleterious effects of candidate coding variants on protein structure and function [8–14]. Furthermore, various functional assays have been carried out to characterize candidate genes and the effect of variants in *in vitro* and *in vivo* contexts. In a high-throughput effort to functionally screen candidate genes, Chapuis *et al.* performed a genome-wide siRNA screen for genes that alter amyloid-β precursor protein metabolism and found that eight of the top siRNA hits overlapped with genes located in GWAS-identified susceptibility loci for Alzheimer's disease [15]. *In vivo* knockouts of candidate genes in mice and other animals are useful for interrogating the function of these genes in a whole organism [6,16]. Candidate variants may also be introduced into genes, using CRISPR/Cas9 homology-directed repair (HDR) or other tools for site-directed mutagenesis, and their effects on protein stability and function studied. Functional studies can also include reporter assays of the activity of variants of a protein known to be a TF [7,9]. Other functional studies are particularly tailored to the protein of interest, such as studying the effects of *SLC22A1* variants on carnitine transport [6], screening *BRCA1* variants for their ability to repair DNA [17], or colony-formation assays to test the pathogenicity of *TP53* mutations [9]. HDR may also be used *in vivo* to study gene variant effects, as was done by Zhu *et al.*, who edited a human variant of GWAS-identified *STXBP5* into the corresponding mouse gene and found that the variant caused a decreased thrombotic phenotype in the mice [18].

In their 2019 study, Keller *et al.* provided an example of how to use the integrated QTL–SNP–gene approach to identify candidate genetic variants that drive disease [19]. Their study sought to identify genes associated with the modulation of insulin secretion in a nondiabetic state. They did so by isolating islet cells from nondiabetic mice fed a Western diet and measuring insulin secretion, while also conducting genetic screens of the mice and performing QTL analysis to identify loci associated with insulin secretion. These loci were integrated with SNP hits found in human GWAS for diabetes-related traits to identify genes and causal variants that may modulate insulin secretion and risk for type 2 diabetes mellitus (T2DM). Three of these genes, *Ptpn18*, *Hunk*, and *Zfp148*, were further characterized *in vivo*, including mice with HDR-inactivated *Ptpn18*, which revealed the role of the gene in regulating insulin sensitivity. Given that coding variants linked to pathogenic risk for specific polygenic disorders commonly affect protein function, they are particularly valuable for providing insight into disease pathogenesis.

Lastly, it is important to highlight the role of rare variants [minor allele frequency (MAF) <1%] in advancing our understanding of genetic disease risk. It has been shown that GWAS variants, which are typically common =(MAF >5%) with small effect sizes, do not completely explain disease heritability and risk [20]. Rare variants may explain some of this 'missing heritability'. Rare coding variants in particular are more likely to be deleterious and have larger effect sizes, thus helping narrow down causal genes and allowing for better elucidation of disease mechanisms to improve therapeutic targeting. Recent rare variant-focused GWAS have shown how rare variants can identify associations for genes not previously known to have a role in disease. For example, in their analysis of very rare variants (MAF <0.1%) in the UK Biobank, Cirulli *et al.* identified novel associations between *STAB1* and brain structure measurements [21]. Studies such as these emphasize the importance of continuing work on rare variant detection, because they will continue to have an important role in functional annotation of GWAS signals.

Variants that affect splicing

While coding variants are perhaps the best-characterized examples linking changes in protein sequence to impacts on protein function in polygenic disorders, an often underappreciated class of variants that also alter protein structure and function comprises those impacting pre-mRNA splicing, the process by which introns are removed and exons linked together to form mature mRNA. Splice site variants can generate mature mRNA transcripts containing introns or missing exons (Figure 2). Such mutations have been extensively investigated on a gene-by-gene basis, uncovering genetic variants causing alternative splicing in physiologically important genes/proteins, such as *CYP3A5* (drug metabolism) [22], tau protein (Alzheimer's disease) [23], α-galactosidase A (Fabry disease) [24], and others. GWAS efforts have catalyzed further discovery of such variants that affect pre-mRNA splicing. However, while GWAS identify associations between a SNP locus and a trait, they do not provide information on the effect of the variant, including whether it impacts alternative splicing.

Thus, efforts have been undertaken to colocalize (i.e., overlap) the GWAS QTL map with a functional map of genomic loci; that is, to determine which GWAS loci overlap with loci known to have some effect on gene expression [25]. This colocalization approach, used for several types of noncoding variants discussed in this article, is an extension of the approach outlined by Glazier *et al.* Once colocalization has identified the most promising functional SNP loci associated with a trait, fine mapping and functional characterization may proceed accordingly. In the case of splicing variants, researchers have worked to generate splicing QTL (sQTL) maps of various normal and abnormal (e.g., cancerous) tissue types [26–29]. sQTLs contain SNPs and other genomic variants that affect alternative splicing and, thus, alter the transcripts produced from a gene. These sQTLs reside not only in canonical splicing sites at the exon–-intron junction, but also across the whole genome, including exons and 5′ and 3′-UTRs [30]. Such non-canonical splicing sites are referred to as cryptic splice sites. While variants in canonical splice sites lead to abnormal whole-intron inclusion or exon skipping, cryptic splice variants, due to their location, cause partial inclusion or deletion of the intron or exon, respectively.

sQTL maps for specific tissue types can be colocalized with GWAS data for tissue-relevant diseases to identify potential causal splicing variants, which may modulate disease risk. An example of this approach was illustrated by Takata *et al.*, who began by analyzing prefrontal cortex RNA-sequencing (Seq) data matched with SNP genotyping for 206 individuals to identify sQTLs in the prefrontal cortex [30]. The authors followed this with enrichment analyses of the sQTL SNPs and the genomic loci found to be associated with disease through GWAS and found that the prefrontal cortex sQTL SNPs were most significantly enriched in schizophrenia-associated loci. Brotman *et al.* modeled a similar approach, beginning by sequencing and analyzing the adipose tissue of 426 individuals to identify sQTLs [26]. Colocalization analysis revealed gene sQTLs that overlapped with GWAS loci for several cardiometabolic traits, including lipid levels and body mass index (BMI).

Noncoding variants

Promoter/enhancer variants

While the classes of variants discussed thus far act primarily by changing the quality of the transcripts and proteins produced, there are several other variant classes, including those present in noncoding DNA, which can modulate the quantity of gene expression (Figure 2A). These noncoding variants comprise more than 90% of all GWAS variants, consistent with their large proportion (>98%) of the human genome [3] (Figure 1B). Thus, characterization of noncoding variants and their target genes is important for understanding the diverse contributors to complex traits and diseases. An important group of noncoding variants resides in gene promoters and enhancers, noncoding regulatory elements in the genome that modulate gene expression. A promoter is a 100–1000-base pair (bp)-long region of DNA located immediately upstream of the transcription start site of a gene, to which proteins bind to initiate transcription of that gene. An enhancer is a 50–1500 bp-long region of DNA to which TFs bind that distally regulate transcription of a gene. Enhancers can be located up to 1 Mbp upstream or downstream from the genes they regulate, and they can be found in intergenic or even intronic sequences, making their potential locations and reach expansive. What promoters and enhancers have in common is that they are bound by TFs, which act to regulate transcription of their *cis*-target gene. Additionally, enhancers and promoters often form loops with each other to fine-tune gene expression. GWAS efforts have identified many noncoding variants, which frequently overlap with putative promoter and enhancer regions, as contributors to complex disease. Thus, in recent years, research on genetic variation has shifted toward understanding the consequences of these noncoding variants and their impacts on their putative target genes and disease processes.

Variants in loci containing promoters and enhancers associated with a complex trait or disease (QTLs) can be nominated by GWAS. However, to confidently assign these noncoding loci to a dysregulated target gene, they must be colocalized with data sets that orthogonally connect the genomic regulatory region where the variant resides with a putative target gene. This is commonly undertaken using tissue-specific expression QTL (eQTL) data, which, analogous to their sQTL counterparts, track associations between SNPs in a given genetic locus and changes in expression of genes within its genomic region. Noncoding variant linkage to putative target genes can also be assigned via DNA looping data, such as Hi-C and HiChIP, which identify physical contacts between putative enhancers and specific gene promoters, as well as via CRISPR studies, which can disrupt activity of a regulatory locus to identify its potential target gene(s). The latter is commonly initially achieved by CRISPR interference (CRISPRi) using catalytically inactive Cas9 fused to a mediator of gene silencing, which can impair function of enhancers and promoters, thereby nominating target genes as those the expression levels of which are modulated by inhibition of a variant-containing regulatory locus. To identify and validate potential causal SNPs, fine mapping and prioritization of SNPs can also be coupled with reporter assays to test SNP activity compared with its non-risk control. Ultimately, the most likely causal GWAS-derived variant and its proposed target gene can be linked using gene-editing tools, such as CRISPR HDR. Much debate in recent years has arisen over how best to determine the target gene of a locus, particularly for long-range and often pleiotropic enhancers. The aforementioned tools and other methods that have been developed to improve locus target predictions are discussed and compared in more detail in a later section.

Application of the colocalization approach to the characterization of GWAS noncoding variants associated with various complex traits and polygenic diseases has led to the identification of novel genes that may contribute to disease phenotypes. Gu *et al.* colocalized kidney disease and function GWAS with kidney eQTL data to identify *MANBA* as a novel gene the low expression of which is associated with increased risk for chronic kidney disease [16]. The authors validated *MANBA* as a novel CKD-associated gene with phenome-wide association analysis of almost 41 000 patients, finding loss-of-function coding variants in *MANBA* associated with renal failure, as well as using a mouse *Manba*-knockout study to show that *Manba* loss increased susceptibility to kidney fibrosis. Doke *et al.* and Gupta *et al.* similarly focused their studies on one particular locus [31,32]. Similar to Gu *et al.,* Doke *et al.* colocalized GWAS hits for eGFR (a measure of kidney function) with kidney eQTL data to nominate *CASP9* as a novel kidney disease risk gene [31]. To identify the causal SNP, they analyzed kidney single nuclear assay for transposase-accessible chromatin sequencing (ATAC-Seq) data to identify variants located in open chromatin and defined a tubule-specific SNP in a locus that regulates *CASP9* expression, as validated by gene knockout. Similar again to Gu *et al.*, they used a mouse model to demonstrate the mechanism by which loss of *Casp9* protects against renal disease and fibrosis. Gupta *et al.* performed similar studies, focusing on a locus associated with vascular disease that modulates *EDN1* expression, using CRISPR HDR to help validate potential causal SNPs [32].

However, with recent shifts to more high-throughput approaches, other researchers have looked for ways to screen and prioritize multiple GWAS loci concurrently. Some have made use of massively parallel reporter assays (MPRAs), a pooled assay in which an enhancer sequence modulates activity of a minimal promoter, which drives expression of a reporter. If reference and alternative sequences of an enhancer for a particular SNP are included, MPRA can be used to determine which GWAS-linked SNPs have differential functional activity. MPRAs have been used to screen SNPs in loci associated with diseases such as dementia [33,34], T2DM [35], melanoma [36], and chronic obstructive pulmonary disease [37]. MPRA-identified functional SNPs can then be further characterized using familiar techniques, including CRISPRi targeting SNP loci to identify responsive genes [33], eQTL colocalizations [33,35,36], HiChIP data linking the SNP locus with candidate gene promoters [36], as well as testing of gene function using *in vivo* animal models [35].

Others have used alternative approaches to prioritize loci of interest, rooted in the predicted mechanistic impacts of variants in these loci. For example, after identifying 23 400 unique SNPs across 18 schizophrenia-associated loci identified through GWAS, Huo *et al.* used data from 30 ChIP-Seq experiments in brain tissue to derive TF DNA-binding motifs and, from the 23 400 SNPs, identified 132 SNPs that were predicted to disrupt TF binding [38]. These SNPs were then colocalized with eQTL data to identify target genes and a subset validated with a luciferase reporter assay to test for differential activity of the reference versus alternative allele. By contrast, Zhang *et al.* focused on allele-specific open chromatin rather than on TF-binding site (TFBS) disruption [39]. They began by identifying 20 people heterozygous for GWAS index SNPs at 70 schizophrenia susceptibility loci and made iPSCs from these individuals. These iPSCs were differentiated into various neuronal cell types, followed by ATAC-Seq and RNA-Seq to identify allele-specific open chromatin (ASoC) SNPs, which were significantly enriched in promoter and enhancer regions. Candidate target genes included genes nominated using colocalization with brain eQTLs and brain and neuronal Hi-C loops, as well as those differentially expressed in CROP-Seq (a CRISPR-based screening tool with a single-cell transcriptome readout) targeting ASoC sites. TF-binding analysis identified a subset of ASoC SNPs predicted to disrupt a TF motif, matching the allelic direction of the allele-specific chromatin accessibility differences. CRISPR HDR-mediated introduction of a subset of ASoC SNPs validated the correct target genes of those nominated by eQTL, Hi-C, and CROP-Seq, as well as the mechanism by which SNPs act. The authors demonstrated that one schizophrenia-associated risk variant, rs2027349, was associated with allele-specific chromatin accessibility and TF footprinting. Zhao *et al.* took an alternative approach to prioritize loci, beginning by performing a GWAS of white matter microstructure, as measured by diffusion magnetic resonance imaging (dMRI), followed by colocalization of dMRI GWAS data with GWAS data for various brain-related complex traits and diseases, such as schizophrenia, depression, stroke, and glioma [40]. MAGMA and FUMA, two prediction tools, were then used to nominate target genes and identified overlap between white matter microstructure-associated genes and those targeted by common nervous system drugs. This method of colocalizing GWAS loci for complex diseases/traits with GWAS loci linked to a physiological phenotype proposed to be associated with the disease helps unite candidate causal genes under one common physiological pathway.

Variants in 5′ and 3′-untranslated regions

5′ and 3′-UTRs are another frequent site of noncoding genetic variation. UTRs are regions present either immediately upstream (5′) or downstream (3′) of the mature coding mRNA and have an important role in regulating post-transcriptional gene expression. 5′ and 3′-UTRs are able to regulate translation through a wide array of mechanisms, and it is these mechanisms that may be impacted in disease-causing mutations. For example, 5′-UTRs contain various regulatory components, including upstream open reading frames (uORFs), which, when translated, can inhibit expression of the downstream primary ORF, secondary structures (such as stem loops), which regulate translation initiation, and iron response elements (IREs), which modulate translation in response to iron levels [41]. 3′-UTRs also contain IREs, as well as miRNA response elements, which bind an miRNA and lead to transcript degradation, AU-rich response elements (AREs), which bind to ARE-binding proteins (ARE-BPs) to modulate translation, and sequences that regulate polyadenylation [42].

In the pre-GWAS era, mutations in 5′ and 3′-UTRs had already been shown to have important roles in disease, particularly Mendelian disorders, by disrupting the aforementioned mechanisms of transcriptional regulation (Figure 2D,F and Box 2). The advent of GWAS led to the discovery of additional 5′ and 3′-UTR variants that contribute to the pathophysiology of complex diseases. However, similar to other variants, particularly noncoding variants, the challenge lies in functionally annotating the variants associated with disease, which in part requires linking the variant to a target gene that is mediating its effects. In some cases, UTRs have been linked to genes using eQTL data, because, similar to promoters and enhancers, they can affect transcript abundance through, for example, miRNA-mediated transcript degradation [43,44]. However, eQTL analysis is insufficient for linking UTR variants with target genes, given that UTR mutations affect not only transcript levels, but also the rate of translation, alternative polyadenylation, and other processes not captured by expression analysis. Thus, others have performed alternative analyses to predict the effects of loci variation on genes, using readouts that capture additional functions of UTRs. For example, Li *et al.* used RNA-Seq data generated for GTEx (an eQTL database) to identify tissue-specific 3′-UTR alternative polyadenylation (APA) QTLs (3′aQTLs); that is, loci associated with the generation of different polyadenylation sites on an mRNA, resulting in 3′-UTRs of varying length with differing susceptibilities to translational regulation [45]. They found that these loci are significantly enriched in 3′-UTRs. From this atlas of 3′aQTLs, CLIP-Seq was performed to identify RNA-binding protein (RBP) binding sites interrupted by 3′aQTL SNPs. 3′aQTLs were found to differ significantly from both eQTLs and sQTLs in their genomic element enrichments, highlighting the importance of mechanism-specific loci–gene associations. Others have described similar prediction and identification of APA-specific QTLs [46,47]. In another effort, Wei *et al.* used PIVar, an algorithm that predicts variants that cause post-transcriptional impairment of a gene

Box 2. Pre-GWAS era characterization of UTR variants

In the pre-GWAS era, mutations in 5′ and 3′-UTRs had already been shown to have important roles in disease, particularly in Mendelian disorders, by disrupting mechanisms of transcript regulation. For example, myotonic dystrophy, a genetic disorder that causes progressive muscle wasting and weakness, is driven by expanded CTG trinucleotide repeats in the 3′-UTR of *DMPK* [215]. These repeats are recognized by an RNA-binding protein (RBP), resulting in transcripts being retained in cell nuclei and not translated [215–217]. Relatedly, a 3′-UTR variant in *FMR1*, which causes Fragile X syndrome, does so by disrupting binding of an RBP, resulting in loss of FMR1 translation [218]. In mantle cell lymphoma, the t(11;14)(q13;q32) translocation, which drives the disease, leads to deletions at the 3′ end of *CCND1* [215,219]. These deletions disrupt AREs, which normally reduce the lifespan of the *CCND1* mRNA, resulting in less mRNA degradation and more CCND1 translation. Mutations in one of the 5′-UTR uORFs of thrombopoietin (THPO), a hormone that regulates platelet production, have been shown to cause loss of uORF-mediated repression, resulting in increased THPO translation and hereditary thrombocythemia (a clotting disorder caused by excess platelets) [220,221]. Similarly, mutations in the 5′-UTR of L-ferritin, an iron storage protein, disrupt the IRE and impair the interaction between the IRE and iron-regulatory proteins [220,222]. This leads to increased L-ferritin translation, causing hereditary hyperferritinemia/cataract syndrome, a disorder characterized by abnormally high blood ferritin levels. Additional 5′-UTR variants that alter the efficiency of translation efficiency of various genes, including *BRCA1* (breast cancer) and *MYC* (multiple myeloma), have been identified [220,223–225].

(piSNVs), to identify 3′-UTR piSNVs [48]. They identified piSNVs that overlapped with GWAS SNPs to find the potential GWAS causal gene and, similar to Li *et al.*, used CLIP-Seq to predict RBP-binding site disruption by 3′-UTR piSNVs. Protein QTL (pQTL) analysis is another downstream alternative to eQTLs, measuring variant-dependent differences in protein abundance. Given the ability of UTRs to modulate protein abundance by modulating translation, pQTL analysis can be a valuable tool for identifying variants that impact translation, although the study of UTR pQTLs has been limited [49–51].

Functional annotation of a variant entails not only linking the GWAS-associated locus with a candidate target gene, but also identifying and studying the causal variant in that locus. The use of MPRAs has been described to screen enhancer sequences and their SNPs for differential activity by linking the enhancer to a minimal promoter that drives expression of a reporter. These traditional MPRA methods have helped some researchers identify functional UTR variants with differential activity [52,53]. However, similar to the limitations of using eQTLs for UTRs, the MPRA assay, which solely measures the ability of a sequence to increase transcript abundance, can miss diverse mechanisms whereby UTRs and their variants can function in disease. Griesemer *et al.* addressed this challenge by creating MPRAu, an MPRA designed to specifically test 3′-UTR sequences and variants [54]. Using MPRAu, Griesemer *et al.* identified and characterized two functional 3′-UTR SNPs, one that interrupts the 3′-UTR miRNA binding site of *TRIM14* and another that changes *PILRB* transcript levels through an unknown mechanism. Other UTR MPRA adaptations have included measuring the impact of APA variants in 3′-UTRs and studying how 5′-UTR sequences affect ribosomal loading [55,56].

Noncoding RNA variants

Unlike coding genes, which get translated into protein, ncRNA genes are a special class of genes that, as the name implies, do not get translated but rather get transcribed and can carry out their regulatory functions as RNA molecules. There are several types of ncRNAs, which carry out a wide array of functions. These include lncRNAs and miRNAs, which have both been shown to be altered in not only Mendelian, but also complex diseases.

Similar to UTRs, ncRNAs carry out diverse, complex functions, which are important for understanding the functional impact of variants in these RNA genes. LncRNAs in particular carry out a variety of functions, including acting as genome guides for regulatory factors, decoys for regulatory factors to prevent them from acting on a gene, sponges for miRNAs to block their inhibitory functions, precursors for miRNAs, and scaffolds for chromatin looping [57]. miRNAs work by binding to an mRNA transcript, usually in the 3′-UTR, and inhibiting its translation by inducing mRNA cleavage, speeding up deadenylation, or blocking ribosome-mediated translation [58]. In addition to having important and previously underappreciated roles in regulating gene expression, ncRNAs are also widely expressed and often located in regions previously thought to be 'junk'. For example, it has been estimated that lncRNAs comprise 68% of expressed genes, with disease-associated SNPs overlapping with 7% of lncRNAs [59].

Given that over 80% of GWAS-discovered variants are predicted to lie in intronic and other noncoding regions, it is important to functionally annotate ncRNA variants to understand their role in complex disease (Figure 2D,E). However, efforts to functionally annotate disease-associated ncRNA variants have lagged compared with other variant classes thus far. Most of the work looking at GWAS-identified ncRNA variants has studied variants that lie in ncRNA gene promoters and enhancers, affecting ncRNA expression levels, rather than the RNA sequence itself. For example, Wu *et al.* colocalized GWAS for steroid-associated phenotypes with eQTLs across several tissues to validate the role of *lincNORS*, a long intergenic ncRNA (lincRNA), in sterol homeostasis [60]. They identified SNPs in the promoter of the lincRNA that are associated with steroid-linked traits (e.g., age of menarche, age of first facial hair, and others), and used a luciferase reporter assay to validate the effect of the SNPs on promoter activity. GWAS SNPs affecting the expression of a lncRNA, often by interrupting TF binding, have been identified and characterized in several other diseases, including colorectal cancer [61], prostate cancer [62], and renal cell carcinoma [63]. While these SNPs do not impact the lncRNA sequence, they are nonetheless useful for identifying disease-relevant lncRNAs and the pathways they regulate. By contrast, others have identified SNPs in lncRNAs genes that impact the sequence of the ncRNA itself and, thus, its function. For example, Feng *et al.* identified a lncRNA SNP in a GWAS susceptibility locus associated with non-small-cell lung cancer and found that overexpression of one allele caused reduced proliferation of cancer cells [64]. To understand the mechanism of action of the SNP, they used lncRNASNP, a bioinformatic tool that predicts the ability of a SNP to create/disrupt an miRNA-binding site, and found that the SNP in the lncRNA *LOC146880* lies in a putative miRNA binding site for *miR-539-5p*. They validated this finding with a luciferase assay, showing differential miRNA binding to the reference and alternative allele. Further mechanistic studies revealed that the low-risk allele results in increased *miR-539-5p* binding to *LOC146880*, preventing phosphorylation of ENO1, which is important for activating downstream cancer-linked genes.

The landscape of miRNA research has similarly identified SNPs that impact miRNA expression as well as those that impact their function by disrupting binding through mutations in miRNAs themselves or miRNA response elements (primarily 3′-UTRs). As an example of the former, Nikpay *et al.* used miRNA-seq data to identify miRNA-eQTLs (miQTLs) in blood; that is, regulatory eQTLs containing SNPs that explain variations in miRNA levels [65]. These miQTLs were then colocalized with GWAS data for cardiometabolic traits to identify trait-associated SNPs that might act through altered miRNA levels. Larson *et al.* similarly focused on diseases impacted by miRNA regulation by performing a transcriptome-wide association study to find associations between miRNA expression levels and prostate cancer, and integrating these with GWAS data to predict prostate cancer-associated SNPs that regulate levels of top candidate miRNAs [66]. Others identified SNPs that affect not miRNA levels, but the processing, stability, and function of the miRNA itself [67–69]. Ghanbari *et al.* identified SNPs associated with open-angle glaucoma that lie in miRNA genes and used the Vienna RNAfold algorithm to predict which SNPs affect the stem-loop structure of precursor miRNAs (pre-miRNAs) [67,70]. Using this method, they identified two glaucoma-associated variants, one predicted to affect miRNA secondary structure and another predicted to reduce the interaction between the miRNA and its target genes. The authors then used TargetScan and miRDB, two computational tools, to predict target genes of the candidate miRNAs and performed luciferase assays (with target 3′-UTR sequences linked to the reporter) to validate the miRNA SNP effects [67,71,72]. Mens *et al.* used a multi-omics approach to find miRNAs associated with disease [68]. They utilized GWAS data to find variants in miRNA genes associated with cardiometabolic traits and then layered this with miRNA CpG site data (epigenomic) and miRNA expression data (transcriptomic) to nominate candidate miRNAs altered in cardiometabolic traits.

As with other noncoding variant classes, some of the key challenges in functionally annotating SNPs affecting ncRNAs includes understanding the mechanism of impact, identifying the target gene(s), and validating the effect of the SNP on the gene and proposed disease-relevant pathway. While some novel associations between noncoding RNAs and disease have been identified, the limited depth of research in this area, particularly the limited use of high-throughput functional screens to annotate SNPs and ncRNA genes, suggests that there is much still to discover connecting ncRNA variants with complex disease.

Copy number variants (CNVs)

While most GWAS have focused on identifying single base pair SNPs associated with disease, there is another class of variants that have been shown to be linked to disease: structural variants. Structural variants are those that, rather than a single nucleotide change, involve larger alterations to the structure of a chromosome. CNVs are a type of structural variant that in particular comprise a large proportion of genomic variants and have been the target of GWAS because they are thought to have an important role in complex disease [73,74]. CNVs are duplications, deletions, and other alterations of genomic segments between 50 bases and 5 Mbp [75]. There are several reasons for studying CNV associations with complex disease alongside SNPs. While there are more individual SNPs than CNVs in the genome, CNVs account for 1.2% of genomic variation from the human reference sequence, whereas SNPs account for 0.1%, highlighting the potentially important role of CNVs in driving human phenotypic variation in traits and disease [76–78]. Additionally, while many GWAS-identified CNV loci overlap with SNP loci for the same trait, CNV-specific GWAS have also identified several novel disease susceptibility loci [73,77,79]. Furthermore, in susceptibility loci that contain both a SNP and a CNV for a trait, identifying and functionally annotating both is important because the presence of a CNV can distort the haplotype and, therefore, the significance of a colocalized SNP [80].

CNV GWAS have identified loci containing those genomic alterations that associate with complex traits and disease, including autoimmune diseases [73,81], schizophrenia and other psychiatric illnesses [81,82], hematological traits [79], anthropometric traits [79,83], cardiometabolic traits [73,81], and others. As previously noted, the loci identified are often colocalized with SNP loci identified in previous trait-matched GWAS to support the role of the CNV locus in the phenotype. Of particular interest, however, are those studies that identify novel loci associated with the trait, highlighting the unique contribution of CNVs to complex traits and disease. For example, Macé *et al.* discovered five novel associations between CNV loci and anthropometric traits, while Marshall *et al.* identified novel CNV loci associated with schizophrenia. Marshall *et al.* further investigated these loci, performing pathway analysis of genes in the loci, and found an enrichment of genes associated with synaptic function and mouse neurobehavioral phenotypes. Nonetheless, critically missing from studies of disease-associated CNVs thus far have been extensive characterization and annotation of the CNV locus. Given that CNVs are structural alterations that can impact multiple genes and other regions with potential regulatory function, an important next step, as has been done with SNPs, is identifying the causal genomic segments driving the phenotype. While these efforts have been limited, Collins *et al.* have begun to make headway toward this goal, using associations between rare CNVs and several disorders to construct a machine learning model that predicts the dosage sensitivity (i.e., haploinsufficiency or triplosensitivity) of individual genes in connection to disease [84]. Hujoel *et al.* similarly developed a computational approach, HI-CNV, to advance our understanding of how rare CNVs affect complex traits by taking advantage of haplotype sharing in human biobanks [85]. Future work will require making use of advances in *in silico* and experimental tools to further and fully characterize CNV loci associated with traits and disease.

Understanding the tools

Efforts to functionally annotate genomic variants connected to disease through GWAS have led to the development of a variety of tools to predict variant effects. These have included computational *in silico* prediction, high-throughput variant screens, and, for noncoding variants in particular, putative target gene prediction.

Computational/*in silico* tools

Due to their minimal upfront costs, computational tools have long been used to perform *in silico* functional annotation and characterization of genomic variants (Table 1). A plethora of tools to predict the functional effect of missense and nonsynonymous variants in the coding sequence of a gene have been developed, including PredictSNP [8], Ensembl Variant Effect Predictor [14], Sorting Tolerant from Intolerant (SIFT) [86], Polymorphism Phenotyping v2 (PolyPhen-2) [87], Protein Variation Effect Analyzer (PROVEAN) [88], SNPs&GO [89], and others [90–96]. Other tools have gone further by performing structural analysis of the effect of SNPs, such as surface accessibility with NetSurfP-2.0 [97], secondary structure with SOPMA [98], protein stability with I-mutant 3.0 [99], and others [100–102]. These tools have aided the characterization of coding SNPs in various disease-associated genes, including the androgen receptor gene [103], *CYP2U1* [104], *RETN* [10], *PPT1* [105], *BRCA2* (in male breast cancer) [106], and *PNMT* (in neurodegenerative disorders) [107]. Structural characterization of coding SNPs in *BRCA2* associated with male breast cancer even led to the discovery of a therapeutic hit with potential anticancer properties [106].

Computational tools have also been developed that can predict the effects of a wide array of noncoding variants associated with disease. For example, SNPs that are predicted to lie in the 5′ or 3′-UTR of a gene can be further analyzed for disruption of a TFBS or miRNA seeding region. RegulomeDB is one tool that can perform such functional analysis, integrating data from public data sets, such as ENCODE, with information on DNase hypersensitivity sites (chromatin accessibility), ChIP-Seq, and TFBS, and other DNA regulatory elements [108]. UTRscan is a tool that similarly predicts the functional effects of SNPs, with a specific focus on 5′ and 3′-UTR elements, and was used by Elkhattabi *et al.* to identify SNPs associated with obesity and insulin resistance that disrupt the polyadenylation signal in the 3′-UTR of *RETN* [10,109,110]. Additional tools for identifying UTRs and their variants include ExUTR, which predicts 3′-UTR sequences from RNA-Seq data, and UTRannotator, a 5′-UTR variant effect prediction tool [111,112]. Furthermore, several tools for miRNA variant effect prediction have been developed, including SubmiRine [113], miR2GO [114], and others [115,116]. However, tools for predicting the effects of variants in lncRNAs have lagged behind and present an opportunity for future computational tool development. Other types of noncoding SNPs, such as those predicted to lie in promoter and enhancer regions, can be similarly computationally annotated using tools that predict disruption of TFBS or other epigenetic alterations. For example, motifbreakR is an R/bioconductor package that uses TF-binding motif information from public data sets to predict whether a variant will break the motif [117]. It has proved useful in predicting the TF-binding disruption of SNPs in regulatory elements associated with Parkinson’s disease [118], colorectal cancer [119], ovarian

Table 1. Computational/*in silico* tools for annotation and characterization of genomic variants

Variant type	Tool	Description	Refs
Coding	Ensembl	Predicting effect of variant/amino acid substitution	[14]
	Meta-SNP		[92]
	PANTHER		[93]
	PolyPhen-2		[87]
	PredictSNP 2.0		[8,90]
	PROVEAN		[88]
	SIFT		[86]
	SNPs&GO		[89]
	SNPsnap		[94]
	SuSPect		[96]
	UMD-Predictor		[95]
	NetSurf2.0	Surface accessibility effects	[97]
	SOPMA	Secondary structure effects	[98]
	I-mutant-3.0	Protein stability effects	[99]
	ConSurf	Evolutionary conservation	[100]
	HOPE	3D structure effects	[101]
	SWISS-MODEL	Homology modeling	[102]
Promoter/enhancer	MotifBreakR	TFBS disruption	[117]
	RegulomeDB	Regulatory DNA variant prediction	[108]
	Basenji	Deep learning	[123]
	DeepSea		[121]
	COLOC	Colocalization	[124]
	eCaviar		[25]
	FUSION		[127]
	MetaXcan		[125,126]
Splicing	GeneSplicer	Splicing site and variant effect prediction	[131]
	Human Splicing Finder		[130]
	NetGene2		[132,133]
	RegSNPs-intron		[128]
	SpliceAI		[129]
5′/3′ UTRs	ExUTR	3′-UTR sequence prediction	[111]
	RegulomeDB	UTR variant effect prediction	[108]
	UTRannotator	5′-UTR variant effect prediction	[112]
	UTRscan	UTR sequence prediction	[109,110]
miRNA/lncRNA	CPSS 2.0	miRNA and lncRNA sequence prediction	[116]
	MicroSNiPer	miRNA variant effect prediction	[115]
	miR2GO		[114]
	SubmiRine		[113]

cancer [120], and several other diseases. However, a key limitation of motifbreakR is that it requires *a priori* knowledge of TF-binding motif sequences, thus limiting the analysis to motifs that have already been discovered.

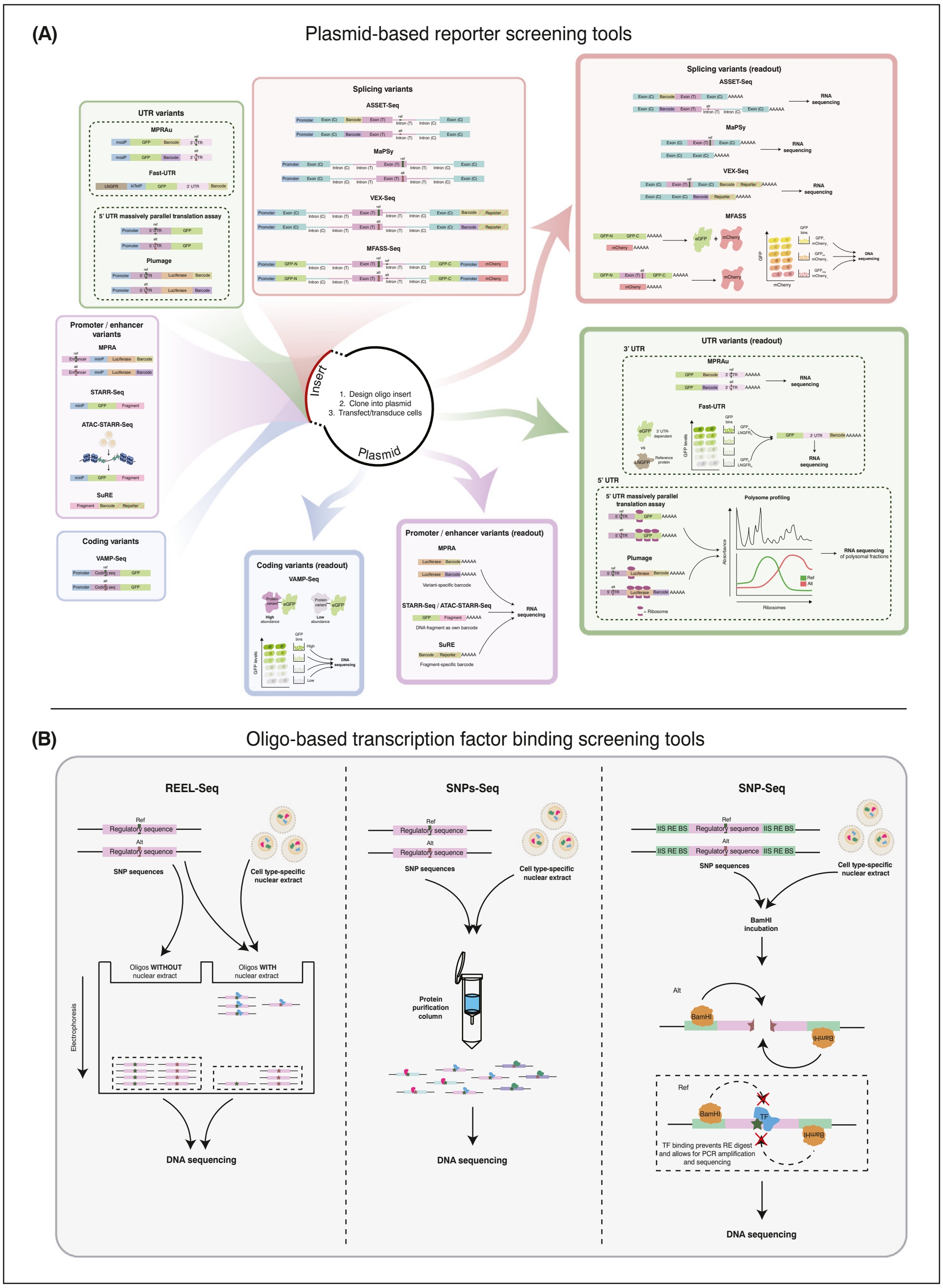
Alternatively, deep learning tools, such as DeepSEA and Basenji, can be trained on tissue-specific regulatory features and used to predict the epigenetic state of a novel DNA sequence and the chromatin effects of variants [121–123]. Other tools have been identified to colocalize GWAS SNPs with tissue-specific eQTLs to nominate disease-associated genes, including COLOC [124], eCaviar [25], MetaXcan [125,126], and FUSION [127]. Computational tools have also been developed that predict how SNPs might affect splicing. For example, Lin *et al.* developed RegSNPs-intron, a tool for predicting the effect of intronic SNPs on splicing and protein structure [128]. Jaganathan *et al.* developed SpliceAI, a tool for predicting splice junctions (particularly cryptic splice sites) from a given pre-mRNA sequence, allowing it to effectively predict how variants affect splicing [129]. Other computational tools for splice site prediction include Human Splicing Factor [130], GeneSplicer [131], NetGene2 [132,133], and MLCsplice [134]. One crucial limitation of *in silico* prediction tools is the quality of the data from which the tools are built, because they are often developed from public epigenomic and transcriptomic data sets with a wide variety of data quality, thus limiting their predictive power. Improved standards for data quality, such as high signal-to-noise ratio and high replicate correlation, will be important moving forward to address this issue [123]. Furthermore, given the known cell type specificity of genomic regulation, it will be important to collect more single cell sequencing data for better cell type resolution, and thus better cell type-specific regulation and variant prediction.

High-throughput functional reporter assays

While computational tools are useful for nominating the most likely causal variants, the function of a SNP must ultimately be experimentally validated to help make accurate predictions about its effect in human disease. Given the high volume of candidate SNPs that emerge from GWAS, high-throughput screening assays have been developed by researchers to screen thousands of disease-associated genomic variants for some functional effect. Such assays have been particularly used in the study of variants in noncoding regions, such as promoters, enhancers, introns, and UTRs, with some downstream reporters serving as a readout for differential activity (Figure 3A).

One such assay is the MPRA, first developed to study the activity of transcriptional regulatory elements by having the candidate sequences control a minimal promoter that drives expression of a reporter and element-specific tag [135]. Given the use of the minimal promoter, MPRA is most useful for identifying sequences that increase baseline gene expression, rather than those that decrease it. MPRA was later adapted to study how variants in enhancer regions affect regulatory activity by introducing reference and alternative alleles into the enhancer sequences [136]. This approach has been used widely to screen GWAS-identified disease-associated SNPs for regulatory function, covering diseases including neurodegenerative diseases [33], T2DM [35], melanoma [36], chronic obstructive pulmonary disease [37], and several others. Integration of MPRA profiles with eQTL data through colocalization allows predicted causal variants to be linked to gene expression, as shown by Abell *et al.* [137]. Kircher *et al.* adapted MPRA in a different manner by performing saturation mutagenesis of 20 disease-associated promoters and enhancers and cloning the tagged sequences into a reporter construct, such that the different tested sequences drove expression of a luciferase reporter and their unique tags [138]. This adapted MPRA protocol is particularly useful for investigating enhancers and promoters commonly associated with disease, because it generates single nucleotide resolution information about the effect of variants on regulatory function and, thus, can be integrated with GWAS SNP data to nominate functional SNPs.

Other parallel reporter assays have been developed and adapted to study the functional effects of SNPs in regulatory elements. STARR-Seq is a reporter assay in which, rather than driving



Trends in Genetics

Figure 3. Overview of high-throughput screening tools used to test variant function. (A) Plasmid-based reporter screening tools have been adapted for various types of variants, with readouts specifically tailored to the variant effects on function. These include massively parallel reporter assays (MPRAs) (enhancer), MPRAu (3′ UTR), VAMP-Seq (coding), and ASSET-Seq (splicing). (B) Oligo-based transcription factor (TF) binding screening tools used to test the ability of variants to disrupt TF binding. REEL-Seq relies on differential migration oligos bound versus unbound to TF. SNPs-Seq uses protein purification columns to enrich for oligo SNPs bound to TFs and uses restriction enzyme digests to test the ability of a variant to disrupt TF binding. Abbreviations: C, constant; T, test.

expression of a barcoded reporter from upstream the promoter, the enhancer is located downstream of the promoter and adjacent to the reporter ORF, such that the enhancer transcribes itself and the strength of an enhancer is reflected by its transcript abundance [139]. More recently, STARR-Seq has been adapted into the method HiDRA, a combination of experimental regulatory element testing (ATAC-STARR-Seq) and computational modeling to identify driver regulatory nucleotides (SHARPR-RE) [140]. ATAC-STARR-Seq involves selecting and amplifying accessible regions of fragmented whole genome and cloning these fragments (enriched for regulatory elements) into a STARR-Seq plasmid construct, such that each DNA fragment serves as its own barcode. Overlap between different fragments allowed for the development of SHARPR-RE, a machine learning model that can predict regulatory driver nucleotides and look for overlap with disease-associated genetic variants. Adapting HiDRA to cell types of interest could lead to the discovery of unique driver nucleotides located in key regulatory elements, which can be colocalized with GWAS SNP data to nominate most likely causal variants. Another reporter technology, Survey of Regulatory Elements (SuRE), similarly relies on a fragmented whole genome, which is inserted into a promoter-less barcoded plasmid, thus testing the ability of the fragment to initiate transcription [141]. van Arensbergen *et al.* performed SuRE with genomes from four individuals of different ethnic backgrounds to effectively test 5.9 million SNPs for their impact on transcription, which were then linked to cell type-relevant GWAS traits and diseases [142].

While the aforementioned parallel reporter assays test the downstream effect of a regulatory SNP, that is, increased or decreased expression, another class of high-throughput assays tests the upstream event of differential TF binding (Figure 3B). Several assays have been developed that involve incubation of synthesized oligos containing regulatory elements and their SNPs with nuclear extract from a disease-relevant cell type of interest and then assaying SNPs for differential TF binding. REEL-Seq uses an electrophoresis mobility shift assay to isolate and sequence oligos not bound to TFs and identify SNPs affecting TF-binding affinity [143]. By contrast, SNPs-Seq passes incubated oligos through a protein purification column to isolate and sequence oligos that are bound to TFs to identify SNPs [144]. Another method, SNP-Seq, takes advantage of sequence-independent cleavage by Type IIS restriction enzymes (REs) and situates the regulatory element containing the SNP at the cleavage site of the RE, such that, after incubation with nuclear extract, SNPs that bind to a regulatory protein are protected from RE cleavage and can be PCR amplified and analyzed [145]. Another oligo-based TF-binding tool is SNP-SELEX, a high-throughput plate-based assay that combines SNP-containing oligos with recombinantly expressed TFs to screen millions of unique TF-DNA combinations to analyze TF-binding specificity and how disease-linked SNPs disrupt TF affinity for DNA [146]. A recently developed non-high-throughout assay, PROBER, complements these approaches by providing an unbiased approach to identify TFs and associated DNA-associated proteins that differentially bind variants of interest in living cells [147]. The PROBER assay is applicable to diverse cell types and enables quantitation of regulatory SNPs that act as TF-binding QTLs (bQTLs).

Parallel reporter assays have also been adapted for screening UTRs and their variants. Griesemer *et al.* developed the MPRA for 3′-UTRs (MPRAu), in which a moderately strong promoter controls the expression of a reporter gene and a library of 3′-UTRs [54]. Using MPRAu, they screened thousands of GWAS 3′-UTR variants and identified several potentially functional causal variants across various traits and diseases, including prostate cancer, triglycerides, and coffee consumption. Zhao *et al.* developed fast-UTR, another parallel reporter assay that uses a bidirectional promoter to drive constant expression of a reference protein in one direction and 3′-UTR-dependent expression of eGFP in the other direction [148]. Cells are sorted via flow cytometry into GFP-high and GFP-low populations to identify UTR segments that increase or decrease reporter production. While not originally developed to screen UTR variants, fast-UTR could easily be adapted to screen SNP or indel variants in 3′-UTR sequences of interest. Sample *et al.* adapted MPRA for screening 5′-UTR variants, designing a vector in which a strong promoter drives expression of a 5′-UTR and eGFP reporter, which is then transfected into cells that subsequently undergo polysome profiling to identify 5-UTR variants that affect ribosomal loading [56]. Lim *et al.* took a similar approach when developing PLUMAGE, a high-throughput 5′-UTR screening tool in which a promoter drives expression of variable 5′-UTR sequences and a luciferase reporter gene [149].

Assays that test the effects of variants on splicing have also been developed, all of which in some way test for the inclusion or exclusion of an exon of interest. ASSET-Seq was developed to test whether intronic SNPs of interest affect splicing [128]. The test exon and intron (containing the SNP) are inserted into a plasmid between two universal exons downstream of a unique barcode and the resulting transcripts from transfected cells are sequenced to determine SNP effects on splicing variant abundance. By contrast, MaPSy and VEX-Seq are two alternative assays that test exonic SNPs and, similarly, involve inserting the test exon with mutations of interest between two universal exons and comparing variant abundance [150,151]. MFASS is another high-throughput splicing assay in which the test exon and intron are inserted within the GFP-coding sequence, such that skipping of the middle exon results in reconstitution of GFP expression and cells can be sorted into GFP-high and GFP-low populations to identify SNPs that drive exon skipping and inclusion, respectively [152]. Fewer MPRAs have been developed for screening coding variants, but one such assay is VAMP-Seq. This is a fluorescence-reporter system developed to assay the effects of thousands of coding variants on protein stability by linking the coding sequence to a GFP reporter, applied to study the pathogenicity of poorly characterized *PTEN* and *TPMT* variants [153].

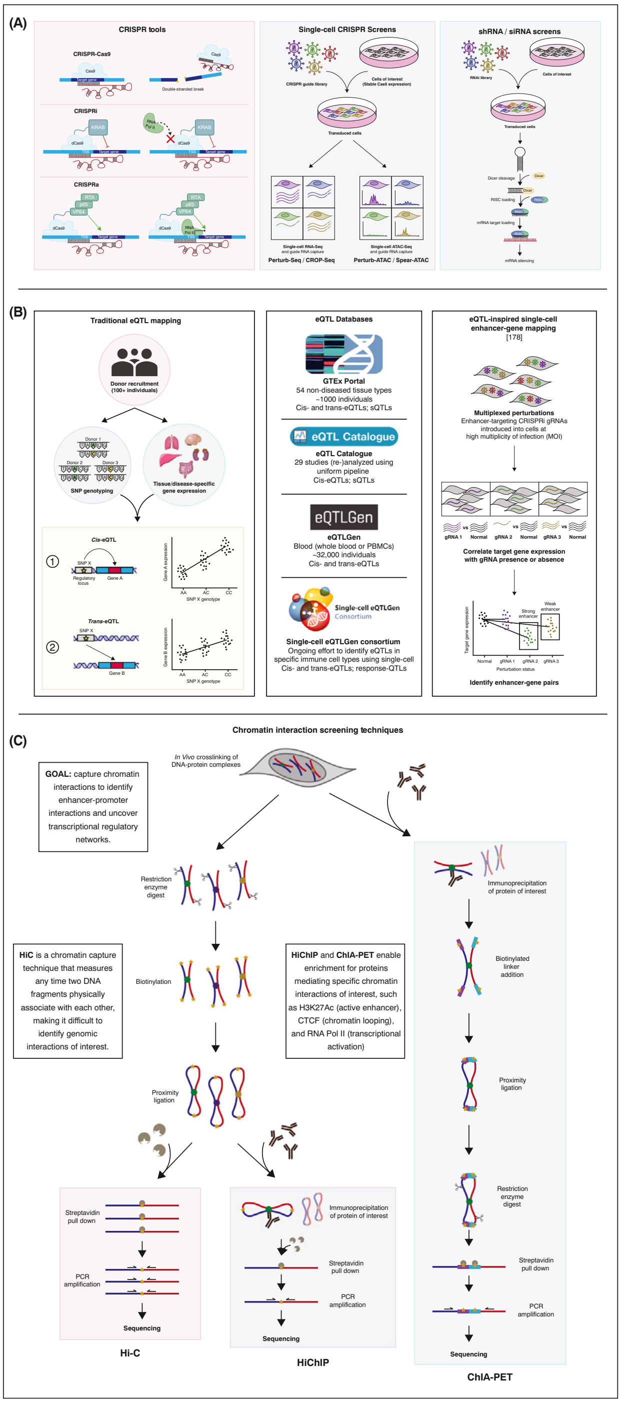
One important limitation of these parallel reporter assays is that they separate the genetic event being tested (splicing, transcription regulation, etc.) from its native context, thus limiting the biological validity of the findings. To truly validate the effect of a coding or noncoding variant on biological pathways, the genomic element and its variant must be assessed in its native genomic context.

Genomic element characterization

An important aspect of functional annotation of GWAS-identified variants is characterizing the genomic elements affected by variants, including coding genes, ncRNA genes, and regulatory elements, in their native genomic contexts to predict the disease-relevant biological effects of the variants. For coding and noncoding genes, this entails using tools that can perturb the expression and function of the genes to determine downstream effects and identify associated pathways. For regulatory elements, particularly those such as enhancers and distal promoters, which act at a distance from the target gene, it is important to first identify the genes linked to the regulatory elements of interest to accurately predict the effect of the variant in disease. Importantly, it is critical to assess the element not only in its native genomic context, but also in a system reflective of the native biological context and disease of interest, because the activity of genes and proteins, and their regulation, is often highly dependent on cell/tissue type and state.

The tool most commonly used to study the function of a candidate risk gene is CRISPR/Cas9, a gene-editing technology that uses a Cas protein coupled with a guide RNA (gRNA) to introduce double-stranded breaks (DSBs) into a sequence of interest, effectively blocking gene activity [154] (Figure 4A). Single cell methods, such as Perturb-Seq [155,156] and CROP-Seq [157], have been developed that allow for screening of a large pool of gRNAs, coupled with single cell RNA-Seq to determine the effects of individual gRNAs on RNA expression, thus linking genes of interest with potential biological pathways. While most of these single cell CRISPR studies are performed in cell culture, Jin *et al.* advanced this technology further by performing Perturb-Seq *in vivo* in the developing mouse neocortex to screen autism risk genes in their biological context [158]. Alternative readouts to RNA expression can be used; for example, Rubin *et al.* developed Perturb-ATAC, coupling CRISPR screening with single-cell ATAC-Seq, which they used to screen TFs, chromatin-modifying factors, and ncRNAs for effects on chromatin accessibility [159]. CRISPR/Cas9 has also been modified into tools including CRISPRi and CRISPR activation (CRISPRa), which provide alternative ways to test the function of a gene without generating DNA breaks (Figure 4A). CRISPRi uses a catalytically dead Cas9 (dCas9), which is targeted to the sequence of interest using a gRNA and blocks transcription [160]. CRISPRa similarly uses dCas9, but it is fused to transcriptional activators such that, when targeted to the gene of interest, it can activate transcription [161]. Both CRISPRi and CRISPRa have been used to screen large gene sets for the consequences of inhibition and activation, respectively, and are useful tools for evaluating the function of genes nominated by GWAS [155,162–164]. Other tools for silencing gene expression that have been used to screen GWAS-identified genes include short hairpin RNAs (shRNAs), used by Nandakumar *et al.* to study genes linked to blood traits, and siRNAs, used by Chapuis *et al.* to screen genes in 19 Alzheimer's disease susceptibility loci [15,165] (Figure 4A). Both are forms of RNA interference (RNAi), which, unlike CRISPR tools that target DNA, bind and target mRNA for degradation and can thus be designed to target specific gene transcript variants of interest.

As has already been discussed, a critical component of the functional annotation of SNPs in regulatory regions involves predicting the genes that these regions control, because this will determine the downstream disease-relevant impacts of the variant. While regulatory elements are often presumed to act on the nearest gene, many elements, such as enhancers or distal promoters, can act on genes from a distance and be located up to 1 Mbp from their target genes [166]. Furthermore, target genes are often context-specific and, thus, a great challenge in this work is to identify the precise regulatory networks that are active in the tissue and cell state of interest [167]. Nonetheless, a plethora of tools and analysis techniques have been developed to link regulatory elements with genes. One commonly used tool that we have already discussed is eQTL mapping, which identifies genomic loci that explain a portion of variation in mRNA levels, based on associations between SNPs in these loci and gene expression differences [168,169] (Figure 4B). These eQTLs can be colocalized with GWAS susceptibility loci to identify genes the expression of which is most likely affected by variants at the GWAS loci [25,170]. Recent eQTL mapping work has shifted to single cell eQTL mapping to uncover cell type-specific gene regulatory networks lost during bulk RNA-Seq [171–173]. Importantly, eQTL mapping requires hundreds of samples at a minimum from different donors to identify SNPs and have the power necessary to detect eQTLs [174]. Thus, rather than performing their own eQTL analyses, many researchers use databases, such as eQTLGen and GTEx, which have identified SNP eQTLs in blood (eQTLGen) and other diverse tissue types (GTEx) [175–177] (Figure 4B). Other forms of QTLs, such as sQTLs, miQTLs, and pQTLs, are also cataloged in such databases to capture the diverse effects of variants. However, the limitation of these databases is that the tissue states captured in their data may not be trait- or disease-concordant. One alternative approach has been to adapt the eQTL approach to a single cell level. Gasperini *et al.* developed an eQTL-inspired framework for identifying enhancer–gene pairs in a single cell line by using multiplexed CRISPRi enhancer perturbations coupled with single cell RNA-Seq [178] (Figure 4B). Rather than requiring hundreds or thousands of people with tissue-level transcriptomes and SNP variants between them, this approach only required thousands of cells with single cell-level transcriptomes and multiple CRISPRi perturbations (i.e., variants) per cell. It is an approach that can be adapted to identifying disease context-specific enhancer–gene pairs for GWAS-identified regulatory loci across various cell types.



GOAL: capture chromatin interactions to identify enhancer-promoter interactions and uncover transcriptional regulatory networks.

HiC is a chromatin capture technique that measures any time two DNA fragments physically associate with each other, making it difficult to identify genomic interactions of interest.

HiChIP and **ChIA-PET** enable enrichment for proteins mediating specific chromatin interactions of interest, such as H3K27Ac (active enhancer), CTCF (chromatin looping), and RNA Pol II (transcriptional activation)

Trends in Genetics

(See figure legend at the bottom of the next page.)

Alternative non-eQTL tools have also been developed that can match regulatory regions with putative target genes (Figure 4C). Hi-C is a chromatin capture technique that involves crosslinking two DNA fragments that are physically associated, ligating them together, and sequencing to identify interacting genomic fragments [179]. Hi-ChIP is a derivative of Hi-C that pairs the technique with ChIP to enrich for genomic interactions occurring near a regulatory protein of interest [180]. H3K27Ac HiChIP in particular has been used to capture enhancer–promoter loops to identify genes linked to enhancers of interest, because H3K27Ac is an epigenetic marker associated with active enhancers [181]. This technique is similar to ChIA-PET, another protein-centered chromatin capture technique that can enrich for CTCF- and cohesin-associated chromatin loops, although HiChIP has been shown to require much lower cell input with greater read output [181–183]. These methods have been able to identify links between regulatory regions and genes that are not revealed by eQTL mapping [184,185]. NEAT-Seq is an alternative protein-centered approach developed by Chen *et al.* that entails simultaneous profiling of enhancer-associated TFs (using antibody-derived tags), open chromatin peaks (ATAC-Seq), and gene expression (RNA-Seq) to interrogate gene regulatory mechanisms [186]. The authors showed that the technique is capable of linking SNPs in TF motifs to putative target gene expression, highlighting its potential use for GWAS target gene nomination. In a reversal of the previously outlined methods, if a putative target gene of interest has been identified, CRISPRi-FlowFISH and activity-by-contact (ABC) modeling have been shown to accurately predict enhancer–gene connections [187]. Developed by Fulco *et al.*, CRISPRi-FlowFISH entails performing CRISPRi with gRNAs against enhancer regions of interest, using RNA fluorescence *in situ* hybridization (FISH) to label single cells based on expression of the gene of interest, sorting the labeled cells into bins based on RNA abundance using fluorescence-activated cell sorting (FACS), and sequencing each bin to identify guides enriched in low or high target gene expressing populations. Naturally, a key limitation of this technique is that it requires *a priori* knowledge on the putative target gene to link it to the appropriate enhancers.

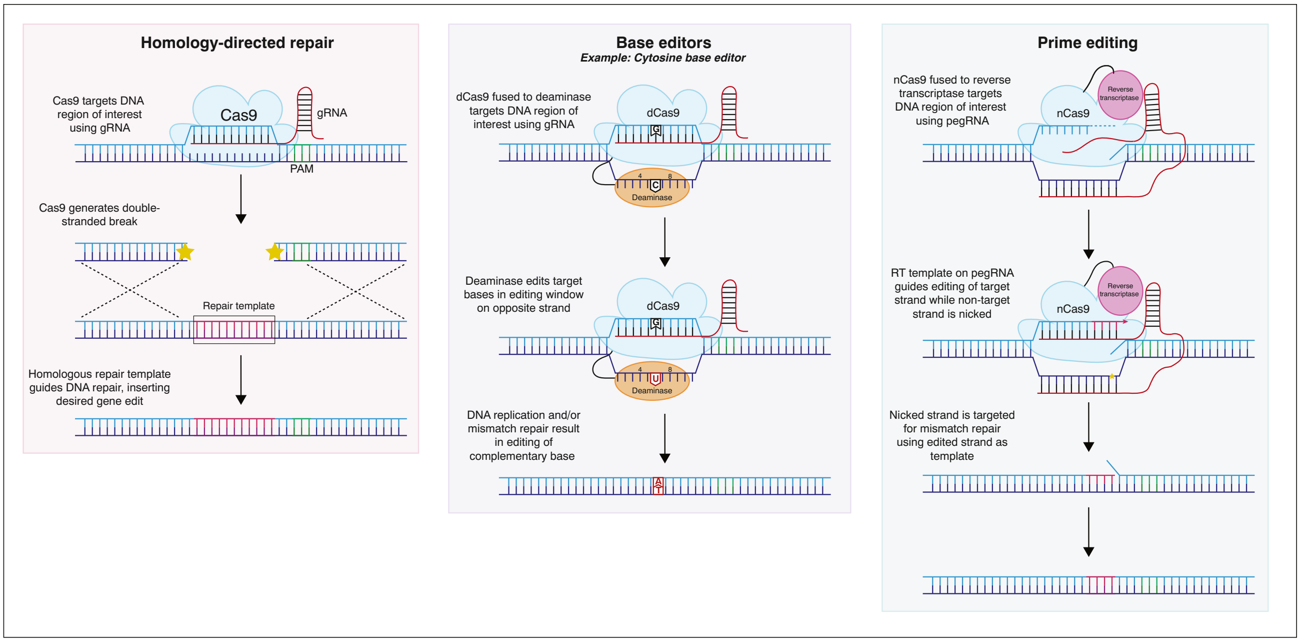
Variant effect characterization

By far the most rigorous method for functionally annotating a GWAS-identified variant (SNP, insertion, deletion, CNV, etc.) involves introducing that precise variant into its native genomic context and studying the functional effects in a relevant cell and tissue context. This approach is particularly powerful when the variant is compared with its non-disease-associated counterpart allele in an otherwise genetically identical isogenic context, which allows isolation of any impacts to the variant in question. This approach is laborious but recent advances in the efficiency of gene editing have facilitated its application to human organoid tissue to study acquired somatic SNVs in collagen genes, which established a role for common *COL11A1* mutations as functionally active promoters of neoplastic invasion [188]. While the tools described up to this point can all predict the effect of a variant, only such precise gene editing with isogenic controls can definitively determine the disease-relevant biological perturbations resulting from a genetic variant and, thus, this approach represents the gold standard for disease-linked variant characterization.

Figure 4. Overview of tools to characterize diverse types of genomic loci. (A) CRISPR tools, including CRISPR interference (CRISPRi) and CRISPR activation (CRISPRa), have been used to characterize coding and noncoding regions. CRISPR screens, such as Perturb-Seq, can be leveraged to study the effect of a CRISPR perturbation on gene expression in a high-throughput, single cell manner. Short hairpin (sh)RNA and siRNA screens can similarly be used to test for the effect of gene silencing, this time at the RNA level, on a phenotype. (B) Colocalization of genome-wide association study (GWAS) susceptibility loci with expression quantitative trait loci (eQTLs) is a common strategy for identifying target genes for noncoding variants/loci. Traditional eQTL mapping uses hundreds of individuals to identify associations between a variant and a change in gene expression. eQTL databases include eQTLGen and GTEx. Single cell eQTL mapping using high MOI CRISPR perturbations can bypass the need for hundreds of human samples [178]. (C) Hi-C, HiChIP, and ChIA-PET are chromatin capture techniques used to identify physical contacts between DNA to predict enhancer–promoter interactions. HiChIP and ChIA-PET allow for capture of specific proteins to enrich for enhancer–promoter loops. Abbreviations: gRNA, guide RNA; PBMC, peripheral blood mononuclear cell.

While a wide array of genome-editing tools has historically been used to introduce genetic variation *in situ*, including zinc-finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs), CRISPR-based tools have become the favorite for gene editing and several iterations of this technology have been developed to improve the accuracy of gene editing (Figure 5). CRISPR/Cas9 was first adapted for precise genome editing by taking advantage of the innate HDR pathway by introducing a DNA template (containing the mutation) into cells along with the CRISPR/Cas9 construct. This DNA sequence can serve as a repair template after CRISPR-mediated cutting to introduce the mutation into the genome of the cell [189,190]. CRISPR HDR has been used to study the effects of both coding and noncoding variants, as well as CNVs, *in vitro* and *in vivo* [37,191–193]. However, there are several drawbacks to CRISPR HDR that limit its use, particularly in a more high-throughput manner. When the Cas9 protein generates DSBs, both HDR and non-homologous end-joining (NHEJ) pathways compete to repair the breaks, resulting in a high frequency of insertions and deletions (indels) near the target site and limiting the precision of the edits [189]. Furthermore, HDR is restricted to the G2 and S phases of the cell cycle and, thus, HDR-mediated editing not efficient in non-dividing cells [194]. Alternatives to CRISPR HDR have been developed that aim to improve the efficiency of genetic editing.

One such alternative is base editors, which use dCas9 and a gRNA fused to a cytidine deaminase (C to T), an adenosine deaminase (A to G), or a uracil DNA glycosylase (C to G) to generate the appropriate edit to the bases in its 5-bp activity window [195–197]. Unlike CRISPR HDR, base editing does not generate DSBs and, thus, indels are not found in edited cells. However, similar to any technology, base editors have limitations. If, for example, there are two of the same target


Homology-directed repair
Cas9 targets DNA region of interest using gRNA
Cas9
gRNA
PAM
Cas9 generates double-stranded break
Repair template
Homologous repair template guides DNA repair, inserting desired gene edit
Base editors
Example: Cytosine base editor
dCas9 fused to deaminase targets DNA region of interest using gRNA
dCas9
G
4
8
C
Deaminase
Deaminase edits target bases in editing window on opposite strand
dCas9
G
4
8
U
Deaminase
DNA replication and/or mismatch repair result in editing of complementary base
A
Prime editing
nCas9 fused to reverse transcriptase targets DNA region of interest using pegRNA
Reverse transcriptase
nCas9
RT template on pegRNA guides editing of target strand while non-target strand is nicked
Reverse transcriptase
nCas9
Nicked strand is targeted for mismatch repair using edited strand as template

Trends in Genetics

Figure 5. CRISPR-based precision gene-editing tools. Homology-directed repair (HDR), base editing, and prime editing are all tools for introducing a precise genetic edit of a variant [SNP, indel, or copy number variant (CNV)] at its native genomic context. HDR involves the generation of double-stranded breaks using CRISPR/Cas9, followed by HDR using a supplied repair template with homology to the DNA region of interest. Base editors use a catalytically dead Cas9 (dCas9) fused to a cytidine deaminase, adenosine deaminase, or uracil DNA glycosylase, which can introduce a single base change. Prime editing uses nicking (n) Cas9 and reverse transcriptase to insert a precise edit included in the prime editing guide (peg)RNA template. Abbreviation: RT, reverse transcription.

base within the activity window, both will get edited, generating unwanted bystander edits; additionally, base editors can sometimes generate unwanted edits even outside the activity window [198]. Furthermore, the scope of edits is limited to those for which base editors have been developed, with ~30% of pathogenic mutations not covered [199]. Nonetheless, base editors have increasingly been used to introduce disease-associated point mutations of interest [196,200–202]. Furthermore, separate work by Hanna *et al.* and Coelho *et al.* recently demonstrated the ability to introduce variants into their endogenous loci in a massively parallel manner using base editors, thus opening the door for further high throughput endogenous variant screens [203].

Prime editors are a novel gene-editing tool that aims to overcome some of the limitations of HDR and base editors. They use a Cas9 nickase (which generates single-stranded nicks instead of DSBs) fused to a prime editing gRNA (pegRNA) and a reverse transcriptase [204]. The pegRNA guides the construct to the target sequence, Cas9 nicks one strand, binding and template-guided repair occur, and, finally, the non-edited strand is nicked so that it is targeted for repair, with the edited strand serving as the template. Prime editors can be used to insert not only point mutations, but also up to 44-bp insertions and 80-bp deletions. Similar to base editors, prime editors do not generate DSBs, resulting in fewer indels. Additionally, while gRNA target sequences for both CRISPR HDR and base editing must be located within ~15 nucleotides of a protospacer adjacent motif (PAM), a short DNA recognition sequence used by CRISPR proteins, the PAM sequence of prime editors can be over 30 nucleotides from the edit site, thus expanding the scope of targetable regions. Finally, although still a relatively new technology, prime editors have been shown to generate higher efficiency edits compared with both HDR and base editors, supporting further exploration and optimization of this tool [204,205].

Once genomic edits have been introduced, familiar tools are available to explore the effect of the variant on gene expression, protein stability, molecular pathways, and other downstream effects. For example, RNA-Seq and ATAC-Seq can be used to explore variant effects on gene expression through mRNA levels and chromatin accessibility at transcription start sites and enhancer loci [39]. ChIP-Seq might be used to explore how the variant impacts TF binding [7]. *In vivo* editing can demonstrate how the variant affects organism function and biological pathways [206–209]. Understanding the mechanism of the effect of the variant and how that is connected to its associated disease can deliver the ultimate holy grail of GWAS: a variant the resulting deficit of which can be corrected to either alleviate or completely eliminate disease.

Concluding remarks and future perspectives

GWAS efforts have broadly expanded our appreciation for the diverse types of genomic element variation that can contribute to the pathogenesis of polygenic disease. Far from being limited to coding genes, disease-associated variants can lie in promoters, enhancers, ncRNAs, and other genomic elements that can perturb gene expression and molecular pathways through diverse mechanisms. As a result, a diverse array of tools, from computational to high throughput to gene editing, have been developed to identify and interrogate these diverse element-specific mechanisms, extending insight from GWAS-identified susceptibility loci to functional variants to putative causal genes and how their functions are related to the trait or disease (Figure 6). The great task of the coming years will be to continue to improve upon and apply these tools to uncover the wide array of variants that contribute in some way to a trait- or disease-causing pathway.

There are several feasible next steps in GWAS functional annotation with the tools we currently have, as well as important next directions in tool development (see Outstanding questions). Characterization of noncoding variants, including introns, lncRNAs, UTRs, and others, has lagged behind that of coding variants, despite comprising over 90% of GWAS-identified variants. While the

Outstanding questions

While the importance of coding variation disease is appreciated, how can understanding the genes and pathways linked to disease through noncoding variation help us understand, prevent, and treat disease?

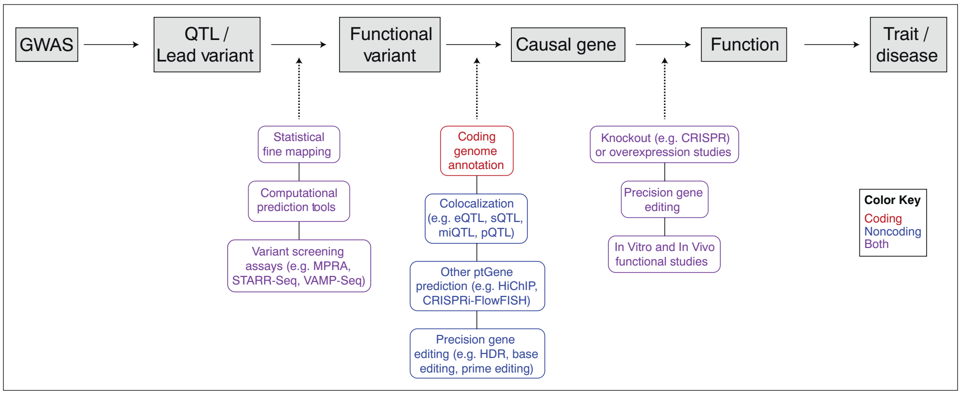
How can genomic tools, including high-throughput screening modalities, be adapted to study the functional effects of CNVs in disease?

How well do high-throughput reporter screens, such as MPRA and STARR-Seq, capture the activity of a variant in its native locus?

How can the different enhancer-gene pairing tools, from eQTL mapping to DNA looping analysis via HiC and HiChIP, be best integrated to nominate the most likely target gene for an enhancer in the trait- or disease-relevant context?

What depth of specificity is necessary in eQTL mapping (e.g., whole-tissue vs. single cell sequencing) to capture true locus/variant–gene relationships that are relevant to GWAS traits/diseases?

How can precision gene-editing tools be optimized for high-throughput variant-editing tools to allow for massive screening of GWAS variants for function in their native genomic context?


GWAS
QTL / Lead variant
Functional variant
Causal gene
Function
Trait / disease
Statistical fine mapping
Computational prediction tools
Variant screening assays (e.g. MPRA, STARR-Seq, VAMP-Seq)
Coding genome annotation
Colocalization (e.g. eQTL, sQTL, miQTL, pQTL)
Other ptGene prediction (e.g. HiChIP, CRISPRi-FlowFISH)
Precision gene editing (e.g. HDR, base editing, prime editing)
Knockout (e.g. CRISPR) or overexpression studies
Precision gene editing
In Vitro and In Vivo functional studies
Color Key
Coding
Noncoding
Both

Trends in Genetics

Figure 6. Flowchart from a genome-wide association study (GWAS) to a disease/trait-relevant function. A diverse array of computational, high-throughput screening, gene editing, and other tools can be used to annotate coding and/or noncoding variants. In all cases, the goal is to identify the causal gene driving GWAS disease heritability and characterize the disease- or trait-relevant function to provide better understanding of the disease and guide future therapeutics. Abbreviations: eQTL, expression QTL; miQTL, miRNA-eQTL; pQTL, protein QTL; QTL, quantitative trait locus; sQTL, splicing QTL.

effects of noncoding variants are often less clear due to complex regulatory mechanisms, we now have tools that aid us further in untangling this web. Given that *in silico* prediction tools have an important role in noncoding variant functional annotation in particular, improved data quality for better modeling is critical.

Additionally, because the importance of CNVs in disease heritability is now increasingly appreciated, further efforts should be directed at understanding the mechanisms of CNV disease-associated perturbations. Furthermore, further tool development will be important to improve the accuracy of GWAS functional annotation results. For example, there is currently no consensus on which enhancer–gene pairing strategy is the most accurate, which makes definitively connecting an enhancer variant to a causal gene and disease pathway difficult. Additionally, the efficiency of precision editing tools must be improved to be more widely applied to screening variants in their endogenous loci, the current gold standard for predicting variant effects. Relatedly, further development of high-throughput variant-editing screens would improve our ability to screen large numbers of variants with high confidence. Lastly, as diverse contributors to complex diseases continue to be discovered and functionally characterized, it will be important to explore how best to target the multiple genes and interwoven pathways found to be causal in disease to develop the most effective treatments, as is the ultimate goal of the GWAS.

Acknowledgments

This work was supported by AR076965 from NIAMS/NIH, by U24HG010856 from NHGRI/NIH, by USVA Merit Review BX001409 to PAK and by the generous support of K. Fields and G. Rayant. T.F. is supported by the P.D. Soros Fellowship for New Americans, the Knight-Hennessy Scholarship, and the Stanford Medical Scientist Training Program.

Declaration of interests

No interests are declared.

References

1. Visscher, P.M. *et al.* (2012) Five years of GWAS discovery. *Am. J. Hum. Genet.* 90, 7–24
2. Visscher, P.M. *et al.* (2017) 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22
3. Buniello, A. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47, D1005–D1012
4. Glazier, A.M. *et al.* (2002) Finding genes that underlie complex traits. *Science* 298, 2345–2349

5. Schaid, D.J. *et al.* (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* 19, 491–504
6. Kim, H.I. *et al.* (2017) Fine mapping and functional analysis reveal a role of SLC22A1 in acylcarnitine transport. *Am. J. Hum. Genet.* 101, 489–502
7. Chang, J. *et al.* (2018) A rare missense variant in TCF7L2 associates with colorectal cancer risk by interacting with a GWAS-identified regulatory variant in the MYC enhancer. *Cancer Res.* 78, 5164–5172
8. Bendl, J. *et al.* (2014) PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput. Biol.* 10, e1003440
9. Doffe, F. *et al.* (2021) Identification and functional characterization of new missense SNPs in the coding region of the TP53 gene. *Cell Death Differ.* 28, 1477–1492
10. Elkhattabi, L. *et al.* (2019) In silico analysis of coding/noncoding SNPs of human RETN gene and characterization of their impact on resistin stability and structure. *J. Diabetes Res.* 2019, e4951627
11. Gray, V.E. *et al.* (2018) Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell Syst.* 6, 116–124
12. Ioannidis, N.M. *et al.* (2016) REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* 99, 877–885
13. Liu, X. *et al.* (2020) dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* 12, 103
14. McLaren, W. *et al.* (2016) The Ensembl Variant Effect Predictor. *Genome Biol.* 17, 122
15. Chapuis, J. *et al.* (2017) Genome-wide, high-content siRNA screening identifies the Alzheimer's genetic risk factor FERMT2 as a major modulator of APP metabolism. *Acta Neuropathol. (Berl.)* 133, 955–966
16. Gu, X. *et al.* (2021) Kidney disease genetic risk variants alter lysosomal beta-mannosidase (MANBA) expression and disease severity. *Sci. Transl. Med.* 13, eaaz1458
17. Starita, L.M. *et al.* (2018) A multiplex homology-directed DNA repair assay reveals the impact of more than 1,000 BRCA1 missense substitution variants on protein function. *Am. J. Hum. Genet.* 103, 498–508
18. Zhu, Q.M. *et al.* (2017) Novel thrombotic function of a human SNP in STXBP5 revealed by CRISPR/Cas9 gene editing in mice. *Arterioscler. Thromb. Vasc. Biol.* 37, 264–270
19. Keller, M.P. *et al.* (2019) Gene loci associated with insulin secretion in islets from nondiabetic mice. *J. Clin. Invest.* 129, 4419–4432
20. Gibson, G. (2012) Rare and common variants: twenty arguments. *Nat. Rev. Genet.* 13, 135–145
21. Cirulli, E.T. *et al.* (2020) Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nat. Commun.* 11, 542
22. Kuehl, P. *et al.* (2001) Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. *Nat. Genet.* 27, 383–391
23. Varani, L. *et al.* (1999) Structure of tau exon 10 splicing regulatory element RNA and destabilization by mutations of frontotemporal dementia and parkinsonism linked to chromosome 17. *Proc. Natl. Acad. Sci. U. S. A.* 96, 8229–8234
24. Ishii, S. *et al.* (2002) Alternative splicing in the α-galactosidase A gene: increased exon inclusion results in the Fabry cardiac phenotype. *Am. J. Hum. Genet.* 70, 994–1002
25. Hormozdiari, F. *et al.* (2016) Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* 99, 1245–1260
26. Brotman, S.M. *et al.* (2022) Subcutaneous adipose tissue splice quantitative trait loci reveal differences in isoform usage associated with cardiometabolic traits. *Am. J. Hum. Genet.* 109, 66–80
27. Garrido-Martín, D. *et al.* (2021) Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat. Commun.* 12, 727
28. Tian, J. *et al.* (2019) CancerSplicingQTL: a database for genome-wide identification of splicing QTLs in human cancer. *Nucleic Acids Res.* 47, D909–D916
29. Zhang, Y. *et al.* (2020) Regional variation of splicing QTLs in human brain. *Am. J. Hum. Genet.* 107, 196–210
30. Takata, A. *et al.* (2017) Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nat. Commun.* 8, 14519
31. Doke, T. *et al.* (2021) Genome-wide association studies identify the role of caspase-9 in kidney disease. *Sci. Adv.* 7, eabi8051
32. Gupta, R.M. *et al.* (2017) A genetic variant associated with five vascular diseases is a distal regulator of endothelin-1 gene expression. *Cell* 170, 522–533
33. Cooper, Y.A. *et al.* (2022) Functional regulatory variants implicate distinct transcriptional networks in dementia. *Science* 377, eabi8654
34. Myint, L. *et al.* (2020) A screen of 1,049 schizophrenia and 30 Alzheimer's-associated variants for regulatory potential. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 183, 61–73
35. Khetan, S. *et al.* (2021) Functional characterization of T2D-associated SNP effects on baseline and ER stress-responsive β cell transcriptional activation. *Nat. Commun.* 12, 5242
36. Choi, J. *et al.* (2020) Massively parallel reporter assays of melanoma risk variants identify MX2 as a gene promoting melanoma. *Nat. Commun.* 11, 2718
37. Castaldi, P.J. *et al.* (2019) Identification of functional variants in the FAM13A chronic obstructive pulmonary disease genome-wide association study locus by massively parallel reporter assays. *Am. J. Respir. Crit. Care Med.* 199, 52–61
38. Huo, Y. *et al.* (2019) Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk. *Nat. Commun.* 10, 670
39. Zhang, S. *et al.* (2020) Allele-specific open chromatin in human iPSC neurons elucidates functional disease variants. *Science* 369, 561–565
40. Zhao, B. *et al.* (2021) Common genetic variation influencing human white matter microstructure. *Science* 372, eabf3736
41. Leppek, K. *et al.* (2018) Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat. Rev. Mol. Cell Biol.* 19, 158–174
42. Mayr, C. (2019) What are 3′ UTRs doing? *Cold Spring Harb. Perspect. Biol.* 11, a034728
43. Flynn, E.D. *et al.* (2022) Transcription factor regulation of eQTL activity across individuals and tissues. *PLoS Genet.* 18, e1009719
44. Levran, O. *et al.* (2019) A 3′ UTR SNP rs885863, a cis-eQTL for the circadian gene VIPR2 and lincRNA 689, is associated with opioid addiction. *PLoS One* 14, e0224399
45. Li, L. *et al.* (2021) An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nat. Genet.* 53, 994–1005
46. Mariella, E. *et al.* (2019) The length of the expressed 3′ UTR is an intermediate molecular phenotype linking genetic variants to complex diseases. *Front. Genet.* 10, 714
47. Shulman, E.D. and Elkon, R. (2020) Systematic identification of functional SNPs interrupting 3′UTR polyadenylation signals. *PLoS Genet.* 16, e1008977
48. Wei, W. *et al.* (2022) Comprehensive characterization of posttranscriptional impairment-related 3′-UTR mutations in 2413 whole genomes of cancer patients. *NPJ Genomic Med.* 7, 34
49. Emilsson, V. *et al.* (2022) Coding and regulatory variants are associated with serum protein levels and disease. *Nat. Commun.* 13, 481
50. He, B. *et al.* (2020) Genome-wide pQTL analysis of protein expression regulatory networks in the human liver. *BMC Biol.* 18, 97
51. Yao, C. *et al.* (2018) Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.* 9, 3268
52. Klein, J.C. *et al.* (2019) Functional testing of thousands of osteoarthritis-associated variants for regulatory activity. *Nat. Commun.* 10, 2434
53. Schrode, N. *et al.* (2019) Synergistic effects of common schizophrenia risk variants. *Nat. Genet.* 51, 1475–1485
54. Griesemer, D. *et al.* (2021) Genome-wide functional screen of 3′ UTR variants uncovers causal variants for human disease and evolution. *Cell* 184, 5247–5260.e19
55. Bogard, N. *et al.* (2019) A deep neural network for predicting and engineering alternative polyadenylation. *Cell* 178, 91–106.e23

56. Sample, P.J. *et al.* (2019) Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nat. Biotechnol.* 37, 803–809
57. Sweta, S. *et al.* (2019) Importance of long non-coding RNAs in the development and disease of skeletal muscle and cardiovascular lineages. *Front. Cell Dev. Biol.* 7, 228
58. O'Brien, J. *et al.* (2018) Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Front. Endocrinol.* 9, 402
59. Iyer, M.K. *et al.* (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* 47, 199–208
60. Wu, X. *et al.* (2020) Regulation of cellular sterol homeostasis by the oxygen responsive noncoding RNA lincNORS. *Nat. Commun.* 11, 4755
61. Tian, J. *et al.* (2020) Risk SNP-mediated enhancer–promoter interaction drives colorectal cancer through both FADS2 and AP002754.2. *Cancer Res.* 80, 1804–1818
62. Hua, J.T. *et al.* (2018) Risk SNP-mediated promoter-enhancer switching drives prostate cancer through lncRNA PCAT19. *Cell* 174, 564–575
63. Wang, J. *et al.* (2021) SNP-mediated lncRNA-ENTPD3-AS1 upregulation suppresses renal cell carcinoma via miR-155/HIF-1α signaling. *Cell Death Dis.* 12, 672
64. Feng, T. *et al.* (2020) A SNP-mediated lncRNA (LOC146880) and microRNA (miR-539-5p) interaction and its potential impact on the NSCLC risk. *J. Exp. Clin. Cancer Res.* 39, 157
65. Nikpay, M. *et al.* (2019) Genome-wide identification of circulating-miRNA expression quantitative trait loci reveals the role of several miRNAs in the regulation of cardiometabolic phenotypes. *Cardiovasc. Res.* 115, 1629–1645
66. Larson, N.B. *et al.* (2022) A microRNA transcriptome-wide association study of prostate cancer risk. *Front. Genet.* 13, 836841
67. Ghanbari, M. *et al.* (2017) A genome-wide scan for microRNA-related genetic variants associated with primary open-angle glaucoma. *Invest. Ophthalmol. Vis. Sci.* 58, 5368–5377
68. Mens, M.M.J. *et al.* (2020) Multi-omics analysis reveals microRNAs associated with cardiometabolic traits. *Front. Genet.* 11, 110
69. Rhead, B. *et al.* (2019) miRNA contributions to pediatric-onset multiple sclerosis inferred from GWAS. *Ann. Clin. Transl. Neurol.* 6, 1053–1061
70. Lorenz, R. *et al.* (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.* 6, 26
71. McGeary, S.E. *et al.* (2019) The biochemical basis of microRNA targeting efficacy. *Science* 366, eaav1741
72. Wong, N. and Wang, X. (2015) miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res.* 43, D146–D152
73. Craddock, N. *et al.* (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 464, 713–720
74. McCarroll, S.A. (2008) Extending genome-wide association studies to copy-number variation. *Hum. Mol. Genet.* 17, R135–R142
75. Carvalho, C.M.B. and Lupski, J.R. (2016) Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* 17, 224–238
76. MacDonald, J.R. *et al.* (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* 42, D986–D992
77. Pang, A.W. *et al.* (2010) Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* 11, R52
78. Sherry, S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311
79. Auwerx, C. *et al.* (2022) The individual and global impact of copy-number variants on complex human traits. *Am. J. Hum. Genet.* 109, 647–668
80. Liu, J. *et al.* (2018) The coexistence of copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) at a locus can result in distorted calculations of the significance in associating SNPs to disease. *Hum. Genet.* 137, 553–567
81. Li, Y.R. *et al.* (2020) Rare copy number variants in over 100,000 European ancestry subjects reveal multiple disease associations. *Nat. Commun.* 11, 255
82. Marshall, C.R. *et al.* (2017) Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* 49, 27–35
83. Macé, A. *et al.* (2017) CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. *Nat. Commun.* 8, 744
84. Collins, R.L. *et al.* (2022) A cross-disorder dosage sensitivity map of the human genome. *Cell* 185, 3041–3055.e25
85. Hujoel, M.L.A. *et al.* (2022) Influences of rare copy-number variation on human complex traits. *Cell* 185, 4233–4248.e27
86. Vaser, R. *et al.* (2016) SIFT missense predictions for genomes. *Nat. Protoc.* 11, 1–9
87. Adzhubei, I.A. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249
88. Choi, Y. (2012) A fast computation of pairwise sequence alignment scores between a protein and a set of single-locus variants of another protein. In *BCB '12: Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pp. 414–417, Association for Computing Machinery
89. Calabrese, R. *et al.* (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.* 30, 1237–1244
90. Bendl, J. *et al.* (2016) PredictSNP2: a unified platform for accurately evaluating SNP effects by exploiting the different characteristics of variants in distinct genomic regions. *PLoS Comput. Biol.* 12, e1004962
91. Capriotti, E. *et al.* (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* 22, 2729–2734
92. Capriotti, E. *et al.* (2013) Collective judgment predicts disease-associated single nucleotide variants. *BMC Genomics* 14, S2
93. Mi, H. *et al.* (2019) PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* 47, D419–D426
94. Pers, T.H. *et al.* (2015) SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinformatics* 31, 418–420
95. Salgado, D. *et al.* (2016) UMD-Predictor: a high-throughput sequencing compliant system for pathogenicity prediction of any human cDNA substitution. *Hum. Mutat.* 37, 439–446
96. Yates, C.M. *et al.* (2014) SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *J. Mol. Biol.* 426, 2692–2701
97. Klausen, M.S. *et al.* (2019) NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning. *Proteins Struct. Funct. Bioinforma.* 87, 520–527
98. Geourjon, C. and Deléage, G. (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Bioinformatics* 11, 681–684
99. Capriotti, E. *et al.* (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* 33, W306–W310
100. Ashkenazy, H. *et al.* (2016) ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* 44, W344–W350
101. Venselaar, H. *et al.* (2010) Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinforma.* 11, 548
102. Waterhouse, A. *et al.* (2018) SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303
103. Adiba, M. *et al.* (2021) In silico characterization of coding and non-coding SNPs of the androgen receptor gene. *Inform. Med. Unlocked* 24, 100556
104. Akhtar, A. *et al.* (2021) In silico computation of functional SNPs of CYP2U1 protein leading to hereditary spastic paraplegia. *Inform. Med. Unlocked* 24, 100610
105. Thirumal Kumar, D. *et al.* (2022) Computational and structural investigation of Palmitoyl-Protein Thioesterase 1 (PPT1) protein causing Neuronal Ceroid Lipofuscinoses (NCL). *Adv. Protein Chem. Struct. Biol.* 132, 89–109
106. Shinde, S.D. *et al.* (2022) Computational biology of BRCA2 in male breast cancer, through prediction of probable nsSNPs, and hit identification. *ACS Omega* 7, 30447–30461

107. Saxena, S. *et al.* (2022) In-silico analysis of deleterious single nucleotide polymorphisms of PNMT gene. *Mol. Simul.* 48, 1411–1425
108. Boyle, A.P. *et al.* (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 22, 1790–1797
109. Pesole, G. and Liuni, S. (1999) Internet resources for the functional analysis of 5′ and 3′ untranslated regions of eukaryotic mRNAs. *Trends Genet.* 15, 378
110. Pesole, G. *et al.* (2002) UTRdb and UTRsite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. Update 2002. *Nucleic Acids Res.* 30, 335–340
111. Huang, Z. and Teeling, E.C. (2017) ExUTR: a novel pipeline for large-scale prediction of 3′-UTR sequences from NGS data. *BMC Genomics* 18, 847
112. Zhang, X. *et al.* (2021) Annotating high-impact 5′untranslated region variants with the UTRannotator. *Bioinformatics* 37, 1171–1173
113. Maxwell, E.K. *et al.* (2015) SubmiRine: assessing variants in microRNA targets using clinical genomic data sets. *Nucleic Acids Res.* 43, 3886–3898
114. Bhattacharya, A. and Cui, Y. (2015) miR2GO: comparative functional analysis for microRNAs. *Bioinformatics* 31, 2403–2405
115. Barenboim, M. *et al.* (2010) MicroSNiPer: a web tool for prediction of SNP effects on putative microRNA targets. *Hum. Mutat.* 31, 1223–1232
116. Wan, C. *et al.* (2017) CPSS 2.0: a computational platform update for the analysis of small RNA sequencing data. *Bioinformatics* 33, 3289–3291
117. Coetzee, S.G. *et al.* (2015) motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* 31, 3847–3849
118. Schilder, B.M. and Raj, T. (2022) Fine-mapping of Parkinson's disease susceptibility loci identifies putative causal variants. *Hum. Mol. Genet.* 31, 888–900
119. Leberfarb, E.Y. *et al.* (2020) Potential regulatory SNPs in the ATXN7L3B and KRT15 genes are associated with gender-specific colorectal cancer risk. *Pers. Med.* 17, 43–54
120. Jones, M.R. *et al.* (2020) Ovarian cancer risk variants are enriched in histotype-specific enhancers and disrupt transcription factor binding sites. *Am. J. Hum. Genet.* 107, 622–635
121. Zhou, J. and Troyanskaya, O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* 12, 931–934
122. Zhou, J. *et al.* (2018) Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* 50, 1171–1179
123. Kelley, D.R. *et al.* (2018) Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* 28, 739–750
124. Wallace, C. (2021) A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet.* 17, e1009440
125. Barbeira, A.N. *et al.* (2018) Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* 9, 1825
126. Barbeira, A.N. *et al.* (2019) Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet.* 15, e1007889
127. Grishin, D. and Gusev, A. (2022) Allelic imbalance of chromatin accessibility in cancer identifies candidate causal risk variants and their mechanisms. *Nat. Genet.* 54, 837–849
128. Lin, H. *et al.* (2019) RegSNPs-intron: a computational framework for predicting pathogenic impact of intronic single nucleotide variants. *Genome Biol.* 20, 254
129. Jaganathan, K. *et al.* (2019) Predicting splicing from primary sequence with deep learning. *Cell* 176, 535–548
130. Desmet, F.-O. *et al.* (2009) Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* 37, e67
131. Pertea, M. *et al.* (2001) GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.* 29, 1185–1190
132. Brunak, S. *et al.* (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* 220, 49–65
133. Hebsgaard, S.M. *et al.* (1996) Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.* 24, 3439–3452
134. Liu, H. *et al.* (2022) Performance evaluation of computational methods for splice-disrupting variants and improving the performance using the machine learning-based framework. *Brief. Bioinform.* 23, bbac334
135. Melnikov, A. *et al.* (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* 30, 271–277
136. Tewhey, R. *et al.* (2016) Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* 165, 1519–1529
137. Abell, N.S. *et al.* (2022) Multiple causal variants underlie genetic associations in humans. *Science* 375, 1247–1254
138. Kircher, M. *et al.* (2019) Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* 10, 3583
139. Arnold, C.D. *et al.* (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339, 1074–1077
140. Wang, X. *et al.* (2018) High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat. Commun.* 9, 5380
141. van Arensbergen, J. *et al.* (2017) Genome-wide mapping of autonomous promoter activity in human cells. *Nat. Biotechnol.* 35, 145–153
142. van Arensbergen, J. *et al.* (2019) High-throughput identification of human SNPs affecting regulatory element activity. *Nat. Genet.* 51, 1160–1169
143. Zhao, Y. *et al.* (2020) A sequential methodology for the rapid identification and characterization of breast cancer-associated functional SNPs. *Nat. Commun.* 11, 3340
144. Zhang, P. *et al.* (2018) High-throughput screening of prostate cancer risk loci by single nucleotide polymorphisms sequencing. *Nat. Commun.* 9, 2022
145. Li, G. *et al.* (2018) High-throughput identification of noncoding functional SNPs via type IIS enzyme restriction. *Nat. Genet.* 50, 1180–1188
146. Yan, J. *et al.* (2021) Systematic analysis of binding of transcription factors to noncoding variants. *Nature* 591, 147–151
147. Mondal, S. *et al.* (2022) PROBER identifies proteins associated with programmable sequence-specific DNA in living cells. *Nat. Methods* 19, 959–968
148. Zhao, W. *et al.* (2014) Massively parallel functional annotation of 3′ untranslated regions. *Nat. Biotechnol.* 32, 387–391
149. Lim, Y. *et al.* (2021) Multiplexed functional genomic analysis of 5′ untranslated region mutations across the spectrum of prostate cancer. *Nat. Commun.* 12, 4217
150. Adamson, S.I. *et al.* (2018) Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. *Genome Biol.* 19, 71
151. Soemedi, R. *et al.* (2017) Pathogenic variants that alter protein code often disrupt splicing. *Nat. Genet.* 49, 848–855
152. Cheung, R. *et al.* (2019) A multiplexed assay for exon recognition reveals that an unappreciated fraction of rare genetic variants cause large-effect splicing disruptions. *Mol. Cell* 73, 183–194
153. Matreyek, K.A. *et al.* (2018) Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* 50, 874–882
154. Jinek, M. *et al.* (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816–821
155. Adamson, B. *et al.* (2016) A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* 167, 1867–1882.e21
156. Dixit, A. *et al.* (2016) Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 167, 1853–1866
157. Datlinger, P. *et al.* (2017) Pooled CRISPR screening with single-cell transcriptome read-out. *Nat. Methods* 14, 297–301

158. Jin, X. *et al.* (2020) In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science* 370, eaaz6063
159. Rubin, A.J. *et al.* (2019) Coupled single-cell CRISPR screening and epigenomic profiling reveals causal gene regulatory networks. *Cell* 176, 361–376
160. Larson, M.H. *et al.* (2013) CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protoc.* 8, 2180–2196
161. Perez-Pinera, P. *et al.* (2013) RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nat. Methods* 10, 973–976
162. Horlbeck, M.A. *et al.* (2016) Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *eLife* 5, e19760
163. Joung, J. *et al.* (2017) Genome-scale activation screen identifies a lncRNA locus that regulates a gene neighborhood. *Nature* 548, 343–346
164. Liu, S.J. *et al.* (2017) CRISPRi-based genome-scale identification of functional long non-coding RNA loci in human cells. *Science* 355, aah7111
165. Nandakumar, S.K. *et al.* (2019) Gene-centric functional dissection of human genetic variation uncovers regulators of hematopoiesis. *eLife* 8, e44080
166. Chepelev, I. *et al.* (2012) Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Res.* 22, 490–503
167. Nott, A. *et al.* (2019) Brain cell type-specific enhancer–promoter interactome maps and disease-risk association. *Science* 366, 1134–1139
168. Mackay, T.F.C. *et al.* (2009) The genetics of quantitative traits: challenges and prospects. *Nat. Rev. Genet.* 10, 565–577
169. Michaelson, J.J. *et al.* (2010) Data-driven assessment of eQTL mapping methods. *BMC Genomics* 11, 502
170. Wu, Y. *et al.* (2019) Colocalization of GWAS and eQTL signals at loci with multiple signals identifies additional candidate genes for body fat distribution. *Hum. Mol. Genet.* 28, 4161–4172
171. Perez, R.K. *et al.* (2022) Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science* 376, eabf1970
172. van der Wijst, M. *et al.* (2020) The single-cell eQTLGen consortium. *eLife* 9, e52155
173. Yazar, S. *et al.* (2022) Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* 376, eabf3041
174. Schliekelman, P. (2008) Statistical power of expression quantitative trait loci for mapping of complex trait loci in natural populations. *Genetics* 178, 2201–2216
175. GTEx Consortium (2017) Genetic effects on gene expression across human tissues. *Nature* 550, 204–213
176. GTEx Consortium (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330
177. Võsa, U. *et al.* (2021) Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* 53, 1300–1310
178. Gasperini, M. *et al.* (2019) A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* 176, 377–390
179. Lieberman-Aiden, E. *et al.* (2009) Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science* 326, 289–293
180. Mumbach, M.R. *et al.* (2016) HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* 13, 919–922
181. Mumbach, M.R. *et al.* (2017) Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.* 49, 1602–1612
182. Fullwood, M.J. *et al.* (2009) An oestrogen-receptor-α-bound human chromatin interactome. *Nature* 462, 58–64
183. Tang, Z. *et al.* (2015) CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* 163, 1611–1627
184. Giambartolomei, C. *et al.* (2021) H3K27ac HiChIP in prostate cell lines identifies risk genes for prostate cancer susceptibility. *Am. J. Hum. Genet.* 108, 2284–2300
185. Wang, W. *et al.* (2021) Functional interrogation of enhancer connectome prioritizes candidate target genes at ovarian cancer susceptibility loci. *Front. Genet.* 12, 646179
186. Chen, A.F. *et al.* (2022) NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. *Nat. Methods* 19, 547–553
187. Fulco, C.P. *et al.* (2019) Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* 51, 1664–1669
188. Lee, C.S. *et al.* (2021) Mutant collagen COL11A1 enhances cancerous invasion. *Oncogene* 40, 6299–6307
189. Lin, S. *et al.* (2014) Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *eLife* 3, e04766
190. Ran, F.A. *et al.* (2013) Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* 8, 2281–2308
191. Lee, M.N. *et al.* (2014) Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* 343, 1246980
192. Long, C. *et al.* (2014) Prevention of muscular dystrophy in mice by CRISPR/Cas9-mediated editing of germline DNA. *Science* 345, 1184–1188
193. Soldner, F. *et al.* (2016) Parkinson-associated risk variant in distal enhancer of α-synuclein modulates target gene expression. *Nature* 533, 95–99
194. Cox, D.B.T. *et al.* (2015) Therapeutic genome editing: prospects and challenges. *Nat. Med.* 21, 121–131
195. Gaudelli, N.M. *et al.* (2017) Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* 551, 464–471
196. Komor, A.C. *et al.* (2016) Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 533, 420–424
197. Kurt, I.C. *et al.* (2021) CRISPR C-to-G base editors for inducing targeted DNA transversions in human cells. *Nat. Biotechnol.* 39, 41–46
198. Antoniou, P. *et al.* (2021) Base and prime editing technologies for blood disorders. *Front. Genome Ed.* 3, 618406
199. Rees, H.A. and Liu, D.R. (2018) Base editing: precision chemistry on the genome and transcriptome of living cells. *Nat. Rev. Genet.* 19, 770–788
200. Hu, Y. *et al.* (2021) Whole-genome sequencing association analysis of quantitative red blood cell phenotypes: the NHLBI TOPMed program. *Am. J. Hum. Genet.* 108, 874–893
201. Yuan, J. *et al.* (2018) Genetic modulation of RNA splicing with a CRISPR-guided cytidine deaminase. *Mol. Cell* 72, 380–394
202. Zeng, Y. *et al.* (2018) Correction of the Marfan syndrome pathogenic FBN1 mutation by base editing in human cells and heterozygous embryos. *Mol. Ther.* 26, 2631–2637
203. Hanna, R.E. *et al.* (2021) Massively parallel assessment of human variants with base editor screens. *Cell* 184, 1064–1080.e20
204. Anzalone, A.V. *et al.* (2019) Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* 576, 149–157
205. Gao, P. *et al.* (2021) Prime editing in mice reveals the essentiality of a single base in driving tissue-specific gene expression. *Genome Biol.* 22, 83
206. Böck, D. *et al.* (2022) In vivo prime editing of a metabolic liver disease in mice. *Sci. Transl. Med.* 14, eabl9238
207. Cai, Y. *et al.* (2019) In vivo genome editing rescues photoreceptor degeneration via a Cas9/RecA-mediated homology-directed repair pathway. *Sci. Adv.* 5, eaav3335
208. Koblan, L.W. *et al.* (2021) In vivo base editing rescues Hutchinson–Gilford progeria syndrome in mice. *Nature* 589, 608–614
209. Xu, L. *et al.* (2021) Efficient precise in vivo base editing in adult dystrophic mice. *Nat. Commun.* 12, 3719
210. Blaisdell, J. *et al.* (2002) Identification and functional characterization of new potentially defective alleles of human CYP2C19. *Pharmacogenet. Genomics* 12, 703–711

211. Dai, D. *et al.* (2001) Identification of variants of CYP3A4 and characterization of their abilities to metabolize testosterone and chlorpyrifos. *J. Pharmacol. Exp. Ther.* 299, 825–831

212. Zhang, J. *et al.* (2001) The human pregnane X receptor: genomic structure and identification and functional characterization of natural allelic variants. *Pharmacogenet. Genomics* 11, 555–572

213. Cargill, M. *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* 22, 231–238

214. Sharon, D. *et al.* (2000) Identification and characterization of coding single-nucleotide polymorphisms within a human olfactory receptor gene cluster. *Gene* 260, 87–94

215. Conne, B. *et al.* (2000) The 3′ untranslated region of messenger RNA: A molecular 'hotspot' for pathology? *Nat. Med.* 6, 637–641

216. Davis, B.M. *et al.* (1997) Expansion of a CUG trinucleotide repeat in the 3′ untranslated region of myotonic dystrophy protein kinase transcripts results in nuclear retention oftranscripts. *Proc. Natl. Acad. Sci.* 94, 7388–7393

217. Lu, X. *et al.* (1999) Cardiac Elav-type RNA-binding protein (ETR-3) binds to RNA CUG repeats expanded in myotonic dystrophy. *Hum. Mol. Genet.* 8, 53–60

218. Suhl, J.A. *et al.* (2015) A 3′ untranslated region variant in FMR1 eliminates neuronal activity-dependent translation of FMRP by disrupting binding of the RNA-binding protein HuR. *Proc. Natl. Acad. Sci. U. S. A.* 112, E6553–E6561

219. Rimokh, R. *et al.* (1994) Rearrangement of CCND1 (BCL1/PRAD1) 3′ untranslated region in mantle- cell lymphomas and t(11q13)-associated leukemias. *Blood* 83, 3689–3696

220. Chatterjee, S. and Pal, J.K. (2009) Role of 5′- and 3′-untranslated regions of mRNAs in human diseases. *Biol. Cell.* 101, 251–262

221. Kondo, T. *et al.* (1998) Familial essential thrombocythemia associated with one-base deletion in the 5′-untranslated region of the thrombopoietin gene. *Blood* 92, 1091–1096

222. Allerson, C.R. *et al.* (1999) Clinical severity and thermodynamic effects of iron-responsive element mutations in hereditary hyperferritinemia-cataract syndrome. *J. Biol. Chem.* 274, 26439–26447

223. Chappell, S.A. *et al.* (2000) A mutation in the c-myc-IRES leads to enhanced internal ribosome entry in multiple myeloma: a novel mechanism of oncogene de-regulation. *Oncogene* 19, 4437–4440

224. Evans, J.R. *et al.* (2003) Members of the poly (rC) binding protein family stimulate the activity of the c-myc internal ribosome entry segment in vitro and in vivo. *Oncogene* 22, 8012–8020

225. Signori, E. *et al.* (2001) A somatic mutation in the 5′UTR of BRCA1 gene in sporadic breast cancer causes down-modulation of translation efficiency. *Oncogene* 20, 4596–4600