ANNUAL REVIEWS

Annual Review of Genetics

Finding Needles in the Haystack: Strategies for Uncovering Noncoding Regulatory Variants

You Chen,[1,2,*] Mauricio I. Paramo,[1,2,*] Yingying Zhang,[1,2,*] Li Yao,[2,3,*] Sagar R. Shah,[1,2] Yiyang Jin,[1,2] Junke Zhang,[2,3] Xiuqi Pan,[1,2] and Haiyuan Yu[2,3]

[1] Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, USA

[2] Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, New York, USA; email: haiyuan.yu@cornell.edu

[3] Department of Computational Biology, Cornell University, Ithaca, New York, USA

Annu. Rev. Genet. 2023. 57:7.1–7.22

The *Annual Review of Genetics* is online at genet.annualreviews.org

https://doi.org/10.1146/annurev-genet-030723-120717

*These authors contributed equally to this article

Keywords

noncoding variant, enhancer, GWAS, rare-variant association test, machine learning, functional assay

Abstract

Despite accumulating evidence implicating noncoding variants in human diseases, unraveling their functionality remains a significant challenge. Systematic annotations of the regulatory landscape and the growth of sequence variant data sets have fueled the development of tools and methods to identify causal noncoding variants and evaluate their regulatory effects. Here, we review the latest advances in the field and discuss potential future research avenues to gain a more in-depth understanding of noncoding regulatory variants.

1. INTRODUCTION

The increasing use of whole-genome sequencing (WGS) in healthcare and research has enabled the identification of numerous variants in the noncoding regions, thus inspiring in recent years a growing interest in these noncoding variants and their biological implications. Accumulating evidence has suggested that functional noncoding variants can be the cause of missing heritability found in exome sequencing cohorts where large proportions of patients do not receive a molecular diagnosis (74). Notably, nearly 90% of disease-associated variants identified by genome-wide association studies (GWASs) lie in the noncoding regions, and they are enriched in transcriptional regulatory elements (TREs), presumably exerting effects by perturbing gene regulation (81).

Despite the critical role of noncoding variants in human diseases, the interpretation and prioritization of noncoding variants have long been hindered by our limited understanding of noncoding regions. Large consortia such as ENCODE (32) and FANTOM5 (5) and independent research groups have made tremendous progress in annotating potentially functional elements in this largely uncharted territory. In this review (**Figure 1**), we first discuss various annotations of the regulatory landscape and how these efforts can help decipher the biological impacts of noncoding variants. We then describe advances in bioinformatic tools to prioritize noncoding variants by integrating these functional annotations. Finally, we present a series of experimental assays to evaluate the regulatory potential of candidate variants.

2. ANNOTATIONS OF THE REGULATORY LANDSCAPE

2.1. Enhancer Annotation

While the noncoding genome contains a diversity of TREs, we limit the focus of this review to enhancers. Enhancers are *cis*-acting noncoding DNA sequences that activate the expression of target genes in an orientation-, position-, and distance-independent manner (116). Despite their importance in physiological and pathological states, the discovery of enhancers, including their defining properties and functions, remains incomplete.

In a continued effort to identify and functionally characterize TREs, a series of biochemical features, including chromatin accessibility, posttranslational histone marks, and noncoding transcriptional patterns, have served as proxies for active enhancers (116). With the advent of high-throughput sequencing technologies to synthesize and test DNA fragments episomally en masse, genome-wide elucidation of putative enhancers has been possible. Below, we briefly describe some of the widely accepted approaches and their advantages and limitations in the genome-wide search for active enhancers.

2.1.1. Chromatin accessibility. The structure of chromatin, including nucleosome positioning and spacing, determines the accessibility of DNA sequences to transcription factors (TFs). The open chromatin state of noncoding regions (i.e., depleted of nucleosomes and therefore accessible) is frequently used to identify potential enhancers (13, 113, 126, 129). To that end, several approaches, including DNase I or micrococcal nuclease (MNase) coupled with deep sequencing [DNase-seq (13) and MNase-seq (141), respectively], formaldehyde-assisted isolation of regulatory elements sequencing (FAIRE–seq) (41), and assay for transposase-accessible chromatin using sequencing (ATAC-seq) (16), have been developed to detect accessible regions to discern potential enhancer loci. Given that not all accessible regions harbor active enhancer elements, additional features are frequently used to refine such predictions.

2.1.2. Posttranslational histone marks. The biochemical properties of histone proteins in the flanking nucleosomes of open chromatin regions are frequently used for the identification

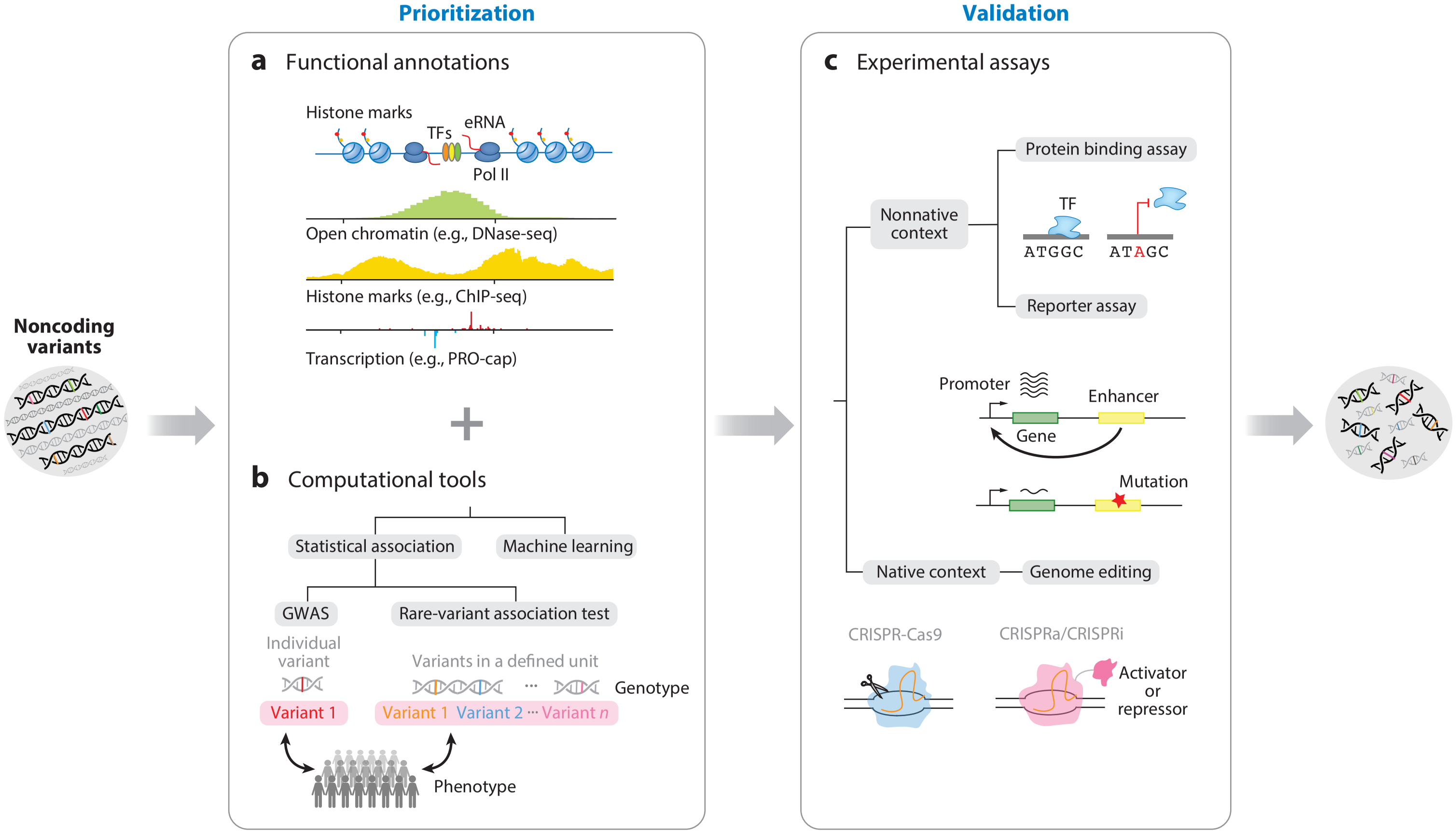

Figure 1

Overview of how noncoding regulatory variants are found. Computational tools with the integration of enriched functional annotations can aid in prioritizing noncoding variants, which can be further validated by a series of experimental assays. (*a*) Various markers, including open chromatin, histone marks, TF binding, and eRNA transcription, have been used to create a comprehensive map of enhancers. (*b*) Both statistical association tests and machine learning tools are frequently used to prioritize noncoding variants. GWASs can test the association between phenotypes and each individual variant, but this approach may not be suitable for evaluating the impacts of rare variants. To address this issue, rare-variant association tests can aggregate the impacts of variants in a defined region based on sliding window or functional annotations. (*c*) To validate the noncoding variants prioritized in the previous step, a variety of experimental assays have been developed. Protein binding assays can identify variants that disrupt key motifs for TF binding, while reporter assays can evaluate the impacts of variants on enhancer activity. Genome editing methods, which can assess the regulatory potential of noncoding variants in their native loci, are also commonly used. CRISPR-Cas9 can introduce mutations, while CRISPRi and CRISPRa can be achieved by fusing catalytically inactive dCas9 to corresponding effector domains. Abbreviations: ChIP-seq, chromatin immunoprecipitation followed by sequencing; CRISPRa, CRISPR-mediated activation; CRISPRi, CRISPR-mediated inhibition; dCas9, dead Cas9; DNase-seq, DNase I hypersensitive sites sequencing; eRNA, enhancer RNA; GWAS, genome-wide association study; Pol II, RNA polymerase II; TF, transcription factor.

of enhancers (10, 32, 66). Typically, histones that flank active enhancers are marked by histone H3 acetylation at lysine 27 (H3K27ac) and H3 monomethylated at lysine 4 (H3K4me1) (46). However, while this epigenomic pattern often correlates with genomic sites containing active enhancers, many enhancer loci lack these characteristic marks (128). In fact, the relationship between posttranslational histone modification and transcription is not entirely resolved.

Moreover, there is no evidence that these biochemical marks are necessary or sufficient for enhancer activation. Since no consensus histone modification profile exists to robustly predict active enhancers, a combinatorial approach is often used to improve enhancer assignments. For example, a multivariate hidden Markov model that explicitly considers the presence or absence of each chromatin mark using chromatin immunoprecipitation followed by sequencing (ChIP-seq) data sets of various histone modifications is frequently used to designate putative enhancer loci systematically (34, 35). Likewise, the ENCODE consortium has systematically integrated DNA accessibility and chromatin modification data to create a categorized registry of likely enhancers. Together, these two approaches offer a step forward in genome-wide curating of putative enhancers.

2.1.3. Sequence and regulator binding profiles. Features such as evolutionary sequence conservation and the presence of TF binding and their motifs or certain enhancer-associated proteins such as the histone acetyltransferase p300 and CREB-binding protein (CBP) have been used to predict putative enhancer loci (43). However, the presence of apparently nonfunctional or neutral binding events limits the precise segmentation of the genome. Thus, inferring sequence patterns and protein binding profiles, which do not accurately capture functionality, cannot be the only method used to annotate enhancers.

2.1.4. Enhancer activity. In addition to biochemical and sequence features, scientists have also screened enhancers by directly measuring the enhancer activity of fragmented whole-genome sequences through massively parallel reporter assays (MPRAs) and self-transcribing active regulatory region sequencing (STARR-seq).

Major differences between MPRAs and STARR-seq arise from construct design, particularly in the location of the candidate enhancer cloning site, known to have a significant influence on the sensitivity and specificity of enhancer calls (62). Canonical MPRAs typically clone candidate enhancers upstream of the reporter gene protomers. Given the known widespread initiation of transcription at active enhancers, this construction may be confounded by measuring promoter potential rather than enhancer activity, as a result of spurious initiation occurring at candidate enhancers, leading to read-through transcription of reporter transcripts. In addition, most canonical MPRAs use oligonucleotide-synthesized candidate sequences that consist of <200 bp, limiting both the size and number of candidate sequences to be tested (92).

STARR-seq-based assays clone elements within $3'$ UTR of the reporter transcript such that sequencing of the element itself allows for the direct measurement of activity without the need for a barcode. Due to the direct coupling of candidate sequences to enhancer activity, STARR-seq-based assays can test millions of DNA sequences from arbitrary sources, making genome-wide identification of enhancers possible (7). However, since the human genome is large and highly complex, deeper sequencing depth is required to increase both genome-wide coverage of candidate sequences and accuracy to detect and quantify enhancers. Moreover, there is an observed strand bias inherent to the assay; the cloning of exogenous sequences within $3'$ UTR may inadvertently lead to strand-specific messenger RNA (mRNA) instability, thus confounding enhancer activity quantifications (7, 128).

A general limitation of these reporter assays is the testing of elements outside of their endogenous context, in particular for plasmid-based systems that rely on nucleofection of synthetic constructs into host cells. While the use of chromosomal integration offers some moderation in the artificial environment in which elements are tested, integrated reporters still intrinsically lack the element-specific genomic background of the native locus, making them likely prone to a high rate of false-positive and false-negative results. This is evident in the substantial discrepancies observed between integration- and episomal-based assays (52).

2.1.5. Transcription. While the abovementioned approaches are informative, mounting evidence points to enhancer RNAs (eRNAs) as a critical mark for detecting active enhancers (128). Genome-wide studies have revealed widespread RNA polymerase II–mediated divergent transcription initiation from enhancer regions (27, 60). While different RNA sequencing (RNA-seq)-based technologies have been utilized to detect eRNAs, these approaches suffer from low sensitivity and specificity.

Given the low abundance and short half-lives of eRNAs, nascent RNA-seq offers an optimal strategy to identify actively transcribed enhancers. In fact, a systematic comparison of various experimental assays for genome-wide identification of active enhancers indicates that nuclear run-on with cap selection and sequencing assays (GRO-cap and its variant, PRO-cap), have advantages in enhancer RNA detection and active enhancer identification (136, 138). Another advantage of these nuclear run-on-based methods, especially PRO-cap, is that they precisely delineate enhancer boundaries, facilitating a high-resolution mapping of all active enhancers genome-wide (128).

Current efforts are underway to comprehensively annotate enhancers genome-wide across the human body at base-pair resolution, which will potentiate the study of noncoding variants in many biological contexts.

2.2. Single-Cell and Spatial Transcriptomics

With the advancement of sequencing technologies, transcriptomic analysis has progressed beyond bulk-based samples, enabling researchers to analyze enhancers and noncoding variants at a much higher resolution. Cutting-edge techniques, such as single-cell sequencing and spatial transcriptomics, allow for the mapping of enhancers and noncoding variants at the single-cell level or specific locations within tissues. This higher-resolution view provides a more comprehensive understanding of how noncoding regulatory variants lead to various developmental and disease phenotypes.

2.2.1. Single-cell transcriptomics. When applied to tissue samples, the bulk-based sequencing methods described in Section 2.1 might be biased toward specific major cell types within the sample, and the heterogeneity of different cell types may not be well captured. As enhancer activities can be highly specific across different cell types and cell states, single cell–based assays have been developed to identify enhancers in relevant biological contexts.

Traditional poly(T) oligonucleotide-based single-cell RNA-seq methods cannot capture eRNAs as they are not poly(A)-tailed. Random displacement amplification sequencing (RamDA-seq) is the first RNA-seq method to sequence all RNA species, including eRNAs, to full length at the single-cell level using not-so-random primers (44). However, this method is unable to pinpoint the 5′ end of transcripts. To resolve this issue, C1 cap-analysis gene expression (CAGE) makes use of the C1 cell-sorting system to perform CAGE at a single-cell scale (65). C1 CAGE can precisely map eRNA transcription start sites, thus enabling enhancer identification at single-cell resolution. Although RamDA-seq and C1 CAGE can detect enhancers within single cells, their sensitivity needs further improvement, as most of the signals are detected from coding regions. Given the advantages of PRO-cap assay in eRNA detection that were mentioned in Section 2.1, statistical deconvolution of bulk PRO-cap data and the development of single-cell PRO-cap assays are ongoing areas of research.

Multiple studies have employed single-cell RNA-seq to map expression quantitative trait loci (eQTLs) in various cell types at different developmental stages of peripheral blood mononuclear cells (PBMCs). Single-cell eQTLs have demonstrated how cell type–specific genetic variation contributes to autoimmune diseases (139); how genetic variation leads to expression changes in coexpressed genes (131); and how the same eQTLs may have opposing effects on gene expression, depending on the cell state (95). Although the application of single-cell eQTLs is limited to PBMCs due to data availability, the framework is potentially generalizable to other tissue types in the human body and will be particularly useful when applied to disease tissue samples (e.g., tumors) that encompass a heterogeneous population of cells.

2.2.2. Spatial transcriptomics. Despite the ability to dissect cell subpopulations within tissues, single-cell assays are not able to capture their spatial distribution and intercellular networks. To address this issue, spatial identification of enhancers across tissues has also been made possible recently.

Spatially transcriptomic mapping can be based on high-plex imaging or spatial barcoding. High-plex imaging methods work by encoding individual RNA species through error-robust barcodes, imprinting the barcodes physically onto RNAs using combinatorial oligonucleotide labels, and measuring each barcode through sequential rounds of imaging (87). A high-plex RNA imaging–based method, multiplexed error-robust fluorescence in situ hybridization (MERFISH), can be modified to incorporate spatial profiling of single-cell epigenomic features to map putative enhancers, and it has been successfully applied to mouse brains (82).

For spatial barcoding, spatially barcoded poly(T) oligonucleotides on a slide can capture poly(A)-tailed transcripts across tissue cross sections, and RNA expression patterns can be assigned to the cross section images after detachment, deep sequencing, and demultiplexing steps (80). Spatial total RNA-seq (STRS) was recently developed by adding a poly(A) tail to the full spectrum of RNAs for efficient capture through the poly(T) oligonucleotides on the slide. This method can potentially be used to map eRNA transcription spatially across different tissue samples (85).

Moreover, researchers can also computationally map single-cell transcriptomics and epigenomics data on a virtual tissue template and identify spatially differentiated enhancer activity across the tissues (14). Thus, spatial profiling of enhancers can further help decipher the role of noncoding regulatory variants during development and pathogenesis of complex diseases.

3. PRIORITIZATION OF NONCODING VARIANTS

3.1. Genome-Wide Association Studies

Over the past decade, GWASs have yielded numerous genotype–phenotype associations and substantial molecular insights into common traits and complex diseases. To further refine the identified variants and their functional effects, researchers have developed methods for GWAS fine-mapping and GWAS-eQTL colocalization analysis. GWAS fine-mapping can help identify candidate causal variants, while GWAS-eQTL colocalization can provide additional evidence that the identified variants have the potential to affect gene expression and may serve as regulatory variants.

3.1.1. Functionally informed fine-mapping. Since GWAS resolution is limited by correlations between nearby variants, it remains a great challenge to pinpoint the actual causal variants at risk loci. There are often tens to hundreds of variants in high linkage disequilibrium (LD) with the reported associated single-nucleotide polymorphisms (SNPs) that can be potentially causal (49).

In a continued effort to prioritize the causal variants, three main fine-mapping strategies have been developed: heuristic approaches, penalized regression models, and Bayesian methods (110).

A common heuristic approach is to retain SNPs with r^2 (a measure of pairwise LD with the lead SNP) above an arbitrary threshold (42). Despite its ease of use, it overlooks the joint effects of SNPs on the trait and cannot quantitatively measure the confidence of causality. Penalized regression models, including lasso (127), elastic net (24), minimax concave penalty (MCP) (15), and normal exponential γ (NEG) (48), can jointly model the simultaneous effects of multiple SNPs and shrink small effect estimates toward zero. However, they tend to result in sparse models that can reduce the chance of selecting causal variants (110).

Bayesian methods account for the joint effects of SNPs, measure the probability of including an SNP as causal in any of the models [i.e., posterior inclusion probability (PIP)], and determine an α credible set by ranking SNPs from largest to smallest PIPs and taking the cumulative sum of PIPs until it reaches α. Bayesian methods have been demonstrated to perform better than the other two approaches (23, 130).

To further prioritize candidates for functional validation, an ad hoc review of genomic annotations is often applied to SNPs selected by fine-mapping methods, which can be cumbersome and biased. An alternative approach is to integrate functional annotations as prior information for Bayesian methods. For instance, a computationally scalable framework, PolyFun, has significantly improved fine-mapping accuracy compared to nonfunctionally informed counterparts by leveraging a broad set of coding, conserved, regulatory, minor allele frequency, and LD-related annotations genome-wide (133). Their regulatory annotations include DNase-seq and ChIP-seq for various histone marks (H3K27ac, H3K4me1, H3K4me3, and H3K9ac), as well as enhancer and promoter annotations from large consortia (e.g., FANTOM5 and ENCODE). Given the cell-type specificity of TREs, another fine-mapping tool, RefMap, has utilized epigenetic profiling (ATAC-seq; ChIP-seq for H3K27ac, H3K4me1, and H3K4me3) of induced pluripotent stem cell–derived motor neurons, the key cell type for the pathogenesis of amyotrophic lateral sclerosis (ALS), to pinpoint causal variants at ALS GWAS risk loci specifically (142).

Well-informed prior probabilities can improve the power and resolution of fine-mapping, while misspecified prior probabilities can result in misleading results. Thus, getting accurate functional annotations is critical. As discussed in Section 2.1, functional annotations based on epigenomic data have their limitations. Incorporating GWAS results with better molecular profiling (e.g., nascent transcription) of disease-relevant cell types and single-cell sequencing readouts promises to further increase fine-mapping power and shed light on the pathogenesis of complex diseases.

3.1.2. Colocalizing genome-wide association studies with expression quantitative trait locus mapping. Since the majority of GWAS hits lie in the noncoding regions, these noncoding variants are widely assumed to affect gene expression via disruption of regulatory element activity. A series of colocalization methods have been developed to test whether the overlap between the GWAS hits and eQTLs is statistically significant (18). For instance, a Bayesian method, eCAVIAR, applied fine-mapping to GWASs and eQTLs independently and estimated the posterior probability for each variant as the product of probabilities that this variant is causal in the GWAS and eQTL mapping (50).

However, despite these efforts, only a limited number of GWAS hits colocalize with eQTLs, raising the concern of missing regulation—the missing link between genetic association and regulatory function (26). To test the model of noncoding GWAS signals acting as eQTLs, Connally et al. (26) constructed a positive set of genes that are found in GWAS loci associated with a complex trait and also harbor coding variants known to be causative for a related Mendelian trait or the same complex trait. The colocalization of trait and eQTL associations is only found, using colocalization tools, in 18 out of 220 (8%) genes. They attributed this inconsistency to context dependency (e.g., cell type, developmental timing, cell state, or environment), nonlinearity (e.g., expression below a certain threshold may not manifest phenotype), and the use of the steady-state expression model (expression can be stochastic and dynamic).

Pritchard and colleagues (90) also looked into the issue of limited overlap of GWAS hits and eQTLs, and they showed that these two assays are systematically biased toward different types of variants. eQTLs are clustered around gene transcription start sites, while GWAS hits are usually farther away from genes and enriched in numerous functional annotations such as enhancers. They agreed that increasing data in more biological contexts may help bridge the gap. Meanwhile, they proposed that other types of molecular QTL assays (e.g., chromatin accessibility, DNA methylation, and chromatin acetylation) and other orthogonal methods such as reporter assays or genome editing tools may help elucidate the role of gene regulation in complex traits.

Given the small effect size of GWAS variants and the enrichment of SNP heritability in enhancers, it may be more informative to measure the impacts on eRNA transcription level instead of target gene expression level. In addition, colocalization of GWASs and eQTLs is mainly used to link noncoding variants to their target genes; however, if GWAS hits are indeed located in enhancers, there are other computational tools to identify target genes. For instance, the Engreitz group (94) generated enhancer–gene maps in 131 human cell types and tissues using the activity-by-contact model, a model based on epigenomic and Hi-C data (39). They utilized these maps to interpret the molecular and cellular functions of GWAS variants, and demonstrated better performance compared to other approaches, including colocalization methods.

3.2. Rare-Variant Association Test

A single-variant association test employed by GWAS design to test the effects of rare variants individually is typically underpowered. Rare-variant association tests are thus developed to measure their effects in aggregate across shared functional units. Such tests require careful consideration of several factors, including the selection of qualified variants and the choice of testing unit, statistical models, and significance threshold (**Figure 2**).

3.2.1. Selection of qualified variants.

Simulation studies performed under realistic scenarios have found that rare-variant association tests often lack power (12, 29, 69, 124). An important driver of power is the ratio of causal to noncausal variants in the studied unit. To increase this ratio, qualified variants are usually chosen based on allele frequency and predicted variant effects.

Filtering based on allele frequency often leverages large-scale human genomic variation databases such as the Genome Aggregation Database (gnomAD) (56) and the 1000 Genomes Project (1). Variants with allele frequencies $<10^{-3}$ or $<10^{-4}$ are often included for further analyses. In some other cases, researchers have relied on the frequency of alleles present within their own cohorts. Typically, these cohorts consist of a considerably smaller number of genomically sequenced samples. Consequently, in such studies, allele frequency cutoffs are often 1% or 5%.

To select qualified variants based on predicted consequences, Combined Annotation-Dependent Depletion (CADD) (61) is among the most commonly used tools. Other computational tools have also been developed recently to predict the pathogenicity of noncoding variants, including deep learning–based sequence analyzer (DeepSEA) (149), delta support vector machine (deltaSVM) (72), and ExPecto (148). A more detailed overview of these prediction tools appears in Section 3.3. These tools need to be benchmarked for their performance in selecting qualified rare variants for association tests.

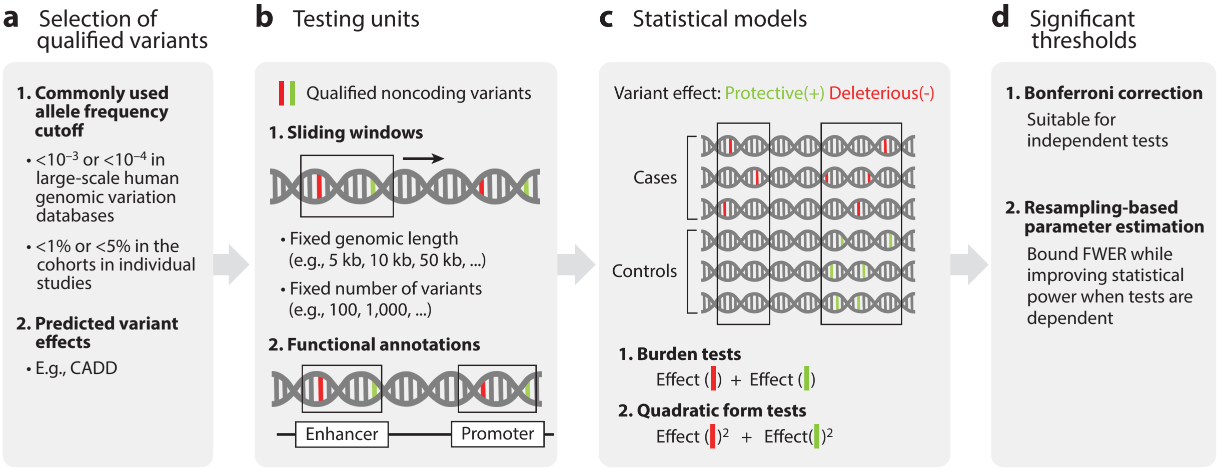
a Selection of qualified variants
1. Commonly used allele frequency cutoff
• <10⁻³ or <10⁻⁴ in large-scale human genomic variation databases
• <1% or <5% in the cohorts in individual studies
2. Predicted variant effects
• E.g., CADD
b Testing units
Qualified noncoding variants
1. Sliding windows
• Fixed genomic length (e.g., 5 kb, 10 kb, 50 kb, ...)
• Fixed number of variants (e.g., 100, 1,000, ...)
2. Functional annotations
Enhancer
Promoter
c Statistical models
Variant effect: Protective(+) Deleterious(-)
Cases
Controls
1. Burden tests
Effect ( ) + Effect ( )
2. Quadratic form tests
Effect ( )² + Effect( )²
d Significant thresholds
1. Bonferroni correction
Suitable for independent tests
2. Resampling-based parameter estimation
Bound FWER while improving statistical power when tests are dependent

Figure 2

Key considerations for rare-variant association tests. (*a*) Allele frequency and prediction tools are commonly employed to prioritize potentially relevant variants and increase the signal-to-noise ratio. (*b*) To better capture signals, appropriate testing units are selected using sliding-window strategies or functional annotations. (*c*) Burden tests are particularly effective when rare variants in the testing unit have consistent effects on the phenotype, while quadratic form test statistics can capture complex relationships between genomic variants and disease risks in the presence of both deleterious and protective variants. (*d*) Another important consideration is selecting an appropriate significance threshold that balances the need to minimize the FWER while maximizing statistical power. Abbreviations: CADD, Combined Annotation-Dependent Depletion; FWER, family-wise error rate.

3.2.2. Testing units. When testing for rare-variant associations in noncoding regions, selecting the appropriate testing unit is crucial for achieving high statistical power. Unlike coding variants, noncoding regions lack natural functional units, making this task particularly challenging.

Currently, two methods are commonly used. The first is the sliding-window strategy, where the genome is scanned by fixed-size windows or a fixed number of variants, and the variants within each window are treated as a group. Several studies used sliding-window strategies to analyze WGS data or candidate regions (45, 79, 89, 106, 132). The main advantage of this method is that it does not require prior knowledge of functional annotations and therefore may discover association signals that uncover new regulatory biology. However, the size of the sliding window can significantly impact results, making it difficult to choose an optimal size that balances the trade-off between specificity and sensitivity.

The second method defines testing units based on available functional annotations generated by large-scale efforts such as ENCODE. For example, Cochran et al. (25) used a gene-centric approach to WGS data by grouping together coding variants in each gene and noncoding variants in their associated TREs. Werling et al. (134) integrated genomic annotations at the level of nucleotides, genes, and TREs and defined 51,801 annotation categories to perform rare-variant association tests. The main advantage of this method is that it provides a more biologically meaningful testing unit and thus can improve the statistical power. However, it may miss signals of association if the underlying biology is not well understood.

3.2.3. Statistical approaches. Statistical approaches used for rare-variant association tests include burden tests and quadratic form tests. For burden tests, some approaches calculate the sum of effects of all qualified variants in the region of interest and contrast the cases versus controls (100, 134), while some others obtain null distributions by permuting phenotypes and then calculate *p*-values based on the observed data (45).

Burden tests are powerful when the rare variants in the testing unit exert effects in the same direction (either deleterious or protective), but they have lower statistical power when a mix of deleterious and protective variants is present (12). By contrast, a quadratic form test statistic combines the effects of individual variants and thus can capture complex relationships between genomic variants and disease risks (45, 96, 137). When attempting to determine the effects of individual variants, one common method is to treat every qualified variant as equal. However, more sophisticated methods assign weights to each variant based on various factors such as its pathogenicity score or functional annotations of the genome (59, 78, 89).

3.2.4. Significance threshold. It is important to appropriately determine the significance threshold when interpreting the results of rare-variant association tests. A common approach for accounting for multiple comparisons is to use the Bonferroni correction, which has been widely used by researchers (89, 109). This method is straightforward and applicable to a wide range of studies. If the tests are independent, the Bonferroni correction can effectively control the family-wise error rate (FWER).

However, the assumption of independence can be violated in many cases. For instance, when using sliding windows, there are typically overlaps between adjacent windows. Moreover, variants may be in LD, and thus the corresponding test statistics can be correlated. Additionally, functional prediction scores of the same variant in different tissues may also be dependent. While using Bonferroni correction can provide an upper bound for the FWER, it may not be the most powerful method in such cases. Therefore, researchers have proposed alternative methods to optimize the significance threshold for improving statistical power while controlling the FWER. These methods typically combine (*a*) a closed form of significance thresholds and (*b*) parameter estimation by resampling algorithms to determine an appropriate significance threshold (45, 79).

3.3. Machine Learning

Computational scores such as CADD have long been used to evaluate the potential impact of genetic variants. These scores can aid in estimating prior probabilities for Bayesian fine-mapping or selecting qualified variants for rare-variant association tests. Additionally, they can be used on their own to assess the deleteriousness and molecular impact of variants. In the following section, we will explain the methodology used to calculate these in silico scores.

3.3.1. Direct prediction of variant consequences. In addition to the statistical methods described in the above sections, machine learning methods have also been used to infer the phenotypic effects of noncoding variants (19, 51, 53, 61, 72, 75, 91, 105, 107, 120) (**Figure 3**). These tools work by finding the hyperplanes that partition the benign and deleterious variants based on a combination of feature values. The invention of these tools can be decomposed into three processes: feature engineering, reference compiling, and model training. Here, we focus on feature engineering and reference compiling. (For a detailed overview of model training methods, see Reference 11.)

3.3.1.1. Feature engineering. In the feature-engineering process, researchers select feature values that have the potential to distinguish two classes of variants and apply necessary numerical transformations to increase the robustness of predictions.

Evolutionary conservation-based features give satisfying performances when used to evaluate the deleterious effects of coding variants (2, 119); naturally, they also play essential roles in almost all noncoding variant dissection tools (19, 51, 53, 61, 72, 75, 91, 105, 107, 120). The most commonly used conservation scores include PhastCons (118), phyloP (101), and genomic evolutionary rate profiling (GERP)++ (28). The maturation of high-throughput sequencing techniques

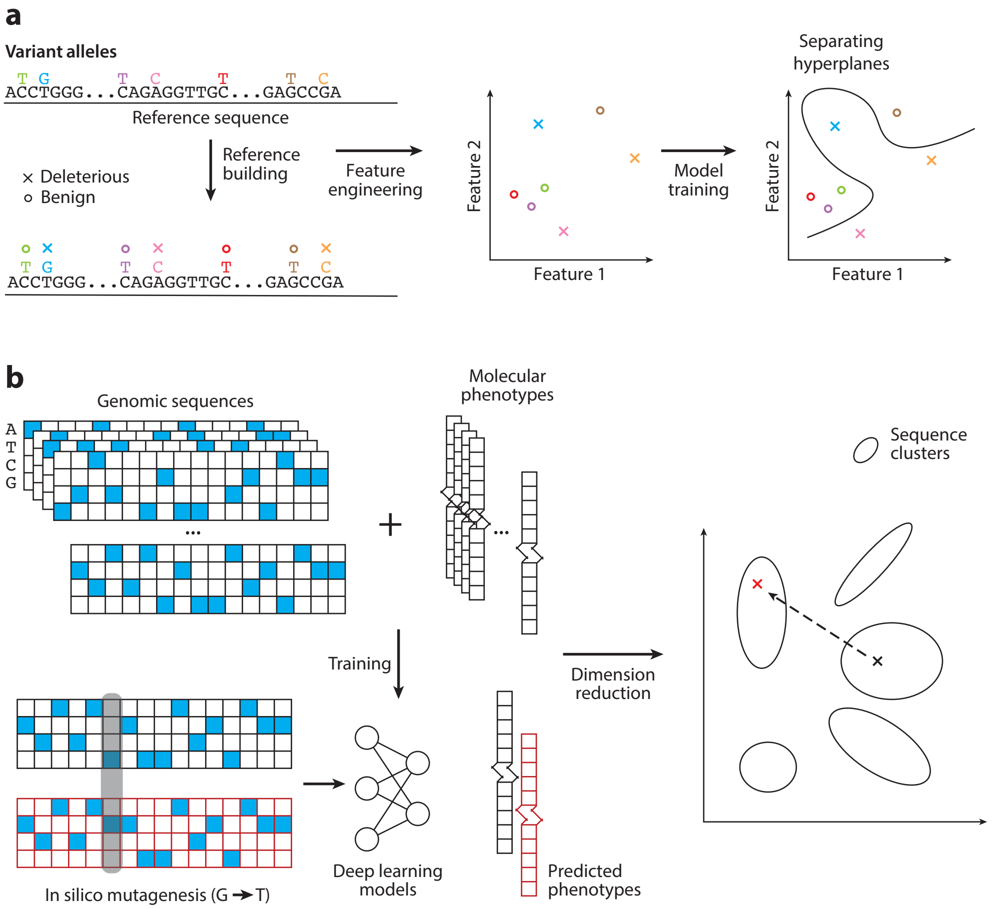

Figure 3

Strategies of machine learning methods to predict the pathogenicity of noncoding variants. Two strategies are commonly used to infer the phenotypic effects of noncoding variants using machine learning methods. (*a*) The first strategy involves training models to find the separating hyperplanes between benign and deleterious variants based on a combination of feature values. (*b*) The second strategy typically involves a two-step process, starting with training a deep learning model to predict molecular phenotypes from sequences and then clustering the molecular phenotypes. Changes in cluster labels can then be used to evaluate the functional outcomes induced by in silico mutagenesis.

enabled the rapid accumulation of sequencing data, which led to the creation of comprehensive catalogs of recent and ongoing natural selection, as reflected in the 1000 Genomes Project and gnomAD databases. Tools such as regulatory Mendelian mutation (ReMM)-Genomiser (120) and NCBoost (19) leveraged allele frequencies from these sources to model the corresponding variants. Meanwhile, with these large-scale sequencing libraries, new conservation metrics tailored for noncoding regions, such as the context-dependent tolerance score (CDTS) (30), are also introduced. Variant classification tools, such as CADD (105), NCBoost (19), and Functional Identification of Noncoding Sequences Using Random Forests (FINSURF) (91), included this new metric as part of their conservation feature set.

Distal TREs are known to have higher evolutionary turnover (111), so in addition to conservation scores, epigenomic signals have frequently been included in the feature set (61, 107). For instance, Genome-Wide Annotation of Variants (GWAVA) (107), a pioneer tool for noncoding variant classification, incorporated information on chromatin accessibility, the binding status of 124 TFs, and 12 histone-modification statuses. More recently, comprehensive catalogs of transcribed enhancers are becoming more accessible (5, 138). Several tools have leveraged enhancer annotations from such sources and reported performance gains (19, 51, 120).

Additionally, features that recapitulate the sequence context of variants are commonly included, such as GC content (19, 61, 115, 120), CpG density (19, 61, 107, 120), and relative position to genes (19).

3.3.1.2. Compiling references. Compiling references is another important step for training supervised models. Most of the aforementioned tools use annotated damaging mutations from public databases, such as the Human Gene Mutation Database (HGMD) (121), or in-house curated pathogenic noncoding variants associated with specific diseases (19, 115, 120) as the positive set (i.e., the deleterious variants).

In order to build the negative sets (i.e., the benign variants), different studies have employed different strategies. For instance, FINSURF used ClinVar (71) variants that do not have known medical impacts (91). CADD (61, 105) and ReMM-Genomiser (120) used noncoding nucleotides that have diverged in humans compared with the inferred ancestral primate genome sequence. Common variants identified by the 1000 Genomes Project or dbSNP (114) variants that do not have clinical assertions are also widely used (19, 107, 115, 120).

One challenge raised by this practice is that the negative set is much larger than the positive set (studies can have up to 32,572-fold more negatives). If unaddressed, the heavily imbalanced training set will significantly bias the model. One work-around for this problem is to downsample or partition the negative set to form multiple balanced training sets and then train reassembled models (19, 91, 107, 115, 120). The other work-around is to enlarge the training set. For instance, CADD simulated equal numbers of de novo variants free of selective pressure (105). However, a considerable fraction of these simulated variants can still be neutral. An alternative way to circumvent these challenges is to characterize variants using unsupervised learning methods (53).

3.3.2. Prediction of molecular phenotypes. Aside from directly predicting deleterious variants, researchers have also devised methods that predict them in a two-step manner: They first determine if the variant introduces any molecular phenotypic changes and then predict the deleterious variants that lead to changes at the molecular level.

Traditional machine learning methods, such as the gapped *k*-mer support vector machine (gkm-SVM) (72), contributed insights on this route. Furthermore, a lot of effort has been focused on leveraging the power of deep learning methods, especially convolutional neural networks (CNNs), to extract flexible DNA syntax and predict epigenomic signals such as chromatin histone modification (57, 144, 149), TF binding (4, 9, 144, 149), chromatin accessibility (57, 58, 93), transcription expression (3, 57, 148), and chromatin interaction (38, 147). In addition to the CNN architecture, tools based on graph convolutional networks and transformer architectures are emerging with promising performances (8, 70, 73, 143).

After learning the syntax between DNA sequences and epigenomic signals, researchers can introduce mutations to the input DNA sequences via a process called in silico mutagenesis and then compare the differences in epigenomic predictions between the wild-type and mutated inputs. Tools such as DeepSEA feed the differences in epigenomic signals into a logistic regression model to discriminate between deleterious and benign variants. The changes in predictions can be used directly for variant prioritization as well. For example, ExPecto (148) and Xpresso (3) predict transcript expression levels, and variants that lead to substantial predicted expression changes largely overlap with causal SNPs identified from GWASs. Dimension reduction and clustering techniques were also used to investigate the potential molecular effects of variants. In this case, the input sequences are first converted to numerical representations (referred to as embeddings) by using the trained deep learning models, and then dimension reduction and clustering are applied to these embeddings to see if specific variants affect the cluster membership of the corresponding sequences (21).

4. FUNCTIONAL CHARACTERIZATION OF NONCODING REGULATORY VARIANTS

While the application of the methods discussed above has led to the identification of countless noncoding regulatory variants implicated in numerous human traits and diseases, functional characterization of the mechanisms by which these variants confer regulatory impact to associated phenotypes remains a central challenge. Extraordinary complexity arises from noncoding variants exerting regulatory influence in a cell type– and cell state–dependent manner—biological contexts that are typically challenging to recapitulate in controlled experimental conditions. Thus, important consideration for intrinsically relevant physiological variables should be taken when designing rigorous functional studies.

Experimental approaches used to derive biological insight for noncoding regulatory variants are categorized into three major classes: protein binding assays, reporter assays, and genome editing (reviewed in greater detail in Reference 103). Here, we summarize general use cases and highlight key considerations for designing informative functional studies.

4.1. Protein Binding Assays

Transcriptional enhancers modulate the spatiotemporal expression of target genes via dynamic DNA-binding patterns of regulatory TFs (116). Models hypothesize that noncoding variants influence transcriptional regulation via disruption and/or stabilization of protein binding by altering TF binding motifs within phenotypically relevant TREs (116). Numerous methods developed to interrogate DNA–protein interactions have been applied to examine alterations in TF binding dynamics mediated by noncoding regulatory variants.

Protein binding assays such as electrophoretic mobility shift assays (EMSAs), ChIP–quantitative polymerase chain reaction (qPCR), and ChIP-seq (54, 98) can provide qualitative and/or quantitative assessment of differential TF binding patterns induced by noncoding regulatory variants (103). Protein binding assays, however, are often limited in that they are performed in vitro and outside of their endogenous genomic context and thus are burdened by high rates of false-positive and false-negative results. While ChIP-qPCR and ChIP-seq are performed closer to the native state due to in vivo DNA–protein cross-linking, they typically require a priori knowledge of TF binding partners in order to directly assess functional impacts.

High-throughput protein binding assays offer considerable advantages. Such methods include those that utilize direct affinity measurements [e.g., microfluidic (83), surface plasmon resonance (17, 99, 117), and microarrays such as ChIP-chip (67) and DNA immunoprecipitation with microarray detection (DIP-chip) (77)], in vitro selection [e.g., high-throughput systematic evolution of ligands by exponential enrichment (SELEX) (108)], bacterial one-hybrid systems (86), and unbiased high-throughput screens [e.g., SNP sequencing (SNP-seq) (76)], which have previously been reviewed in great detail (103, 122). The evaluation of protein binding on a large scale has made the screening of the protein binding effects of thousands of candidate noncoding regulatory variants possible, offering a powerful means for the prioritization of potentially consequential variants. However, large-scale screens typically fall short of providing detailed information into the molecular mechanisms and relevant cellular pathways in which noncoding regulatory variants function and thus require further follow-up studies to recapitulate and explore initial findings. Nevertheless, as protein binding assays allow for the determination of specific TF binding alterations as well as the assessment of global TF binding patterns, these methods provide powerful screening potential, leading to the prioritization of candidate noncoding regulatory variants, especially when performed using relevant cell types and under relevant cell states.

4.2. Reporter Assays

Orthogonal to assessing a variant's impact on TF binding is directly measuring its influence on transcriptional activity. In ectopic-based reporter assays, TREs are cloned into heterologous reporter constructs such that when introduced into a cellular host system, they modulate the expression of a reporter gene. Genomic integration–based reporter assays follow the same principle; however, they are assessed within a genomic context introduced via either random or site-directed integration into a host-cell genome. Both ectopic- and integration-based assays can differ in experimental read-out to obtain activity measurements. Imaging-based readouts quantify activity using the enzymatic (e.g., luciferase and β-galactosidase) or fluorescent [e.g., green fluorescent protein (GFP)] activity of protein products, though they can be confounded by variables acting at the level of posttranscription and/or translation. As an alternative, reporter RNA transcript readouts using reverse transcriptase qPCR (RT-qPCR) can circumvent many of these potential limitations, assuming fixed reporter RNA stability.

In addition to applications in enhancer screening, as described in Section 2.1.4, MPRAs and STARR-seq assays are also frequently used to test the regulatory effects of noncoding variants (125). Testing in nonnative contexts as well as limitations in our understanding of fundamental architectural and logical properties of enhancer elements, such as element unit boundaries and enhancer–promoter compatibility rules, may, in part, explain the observed inconsistency common to reporter assays (52). Moreover, reporter assays may lack the dynamic range to detect the small effect sizes of variant impacts as well as the ability to detect additive and/or multiplicative effects of multiple variants acting in concert on regulation. Developing improved reporter constructs and analytical pipelines that reduce experimental bias in addition to testing with diverse representative sets of promoters and cell types under relevant cell states may partially address some of these limitations. Furthermore, we anticipate that the recently developed single-cell STARR-seq (84) and single-cell MPRA (146) will be widely used to evaluate candidate variants for cell type– and cell state–specific *cis*-regulatory effects, particularly in native tissue contexts.

4.3. Genome Editing

The major limitation of both protein binding and reporter assays is the testing of variants outside of their native genomic context, resulting in high rates of false positives and false negatives. Genome editing methods are thus promising given their ability to investigate the functions of noncoding regulatory variants within their native locus under endogenous physiological conditions.

Genome editing has diverse applications due to the development of a host of different strategies, allowing for the evaluation of variants under native biological contexts, including via the use of in vitro primary cell culture or in vivo animal models. This makes investigating variant influence on regulatory activity and characterization of target genes and relevant molecular pathways feasible, which are essential steps for determining the impact on associated phenotypes (40). Broadly, genome editing technologies rely on programmable sequence-specific nucleases (SSNs) to induce targeted DNA breaks (31). Early genome editing methods, including meganucleases (33), zinc-finger nucleases (ZFNs) (20), and transcription activator-like effector nucleases (TALENs)

(55), required the design of sequence-specific DNA-binding protein domains, making them exceptionally difficult to produce. However, the discovery and subsequent rapid improvement of the low-cost and easy-to-use CRISPR-Cas9 system has made earlier methods less favorable. Instead of relying on protein engineering, CRISPR-Cas9 utilizes the Cas9 nuclease, which is directed by a programmable guide RNA to a specific location in the genome where it can then cut the DNA and induce the desired change. The mechanisms that underlie genome editing have been previously reviewed (140).

CRISPR-Cas9 has been used to introduce small insertions and deletions (indels) at or near variants of interest as well as induce larger genomic deletions including or surrounding variants to disrupt regulatory activity (64). Inactivation of the Cas9 nuclease resulting in complete loss of DNA cleavage activity [dead Cas9 (dCas9)] and fusion of dCas9 enzymes to effector domains enable efficient transcriptional regulation, including CRISPR-mediated inhibition (CRISPRi) and activation (CRISPRa) (102). CRISPRi and CRISPRa have allowed for the dynamic spatiotemporal control of gene expression and have been applied to study noncoding regulatory variants via epigenetic control of the local regulatory region (102). Additionally, dCas enzymes can be fused with other enzymatic domains, such as the base modification enzymes cytidine deaminase, to create cytosine base editors that can convert C • G to T• A base pairs, and adenine deaminase, to create adenine base editors that convert A •T to G • C base pairs. Recently, glycosylase base editors, capable of inducing C • G to G • C and C • G to A •T transitions (68, 145), have been developed. Base editing can generate precise point mutations in the genome and is a powerful tool for studying noncoding regulatory variants, allowing for allelic substitution of variants of interest in functional studies. Base editors and their applications have been reviewed extensively (47, 88, 104).

Although base editing can be used to perform the four transition mutations and some transversion mutations, it is limited in performing all eight transversions and inducing small indels. Prime editors are composed of a Moloney murine leukemia virus (M-MLV) reverse transcriptase fused to an RNA-programmable nickase (nCas9), and a prime editing guide RNA guides the prime editor to the target site, whereby it can perform all 12 possible conversions, and small indels into target DNA sites (6). Prime editors also offer advantages in their ability to induce base substitutions in more regions with fewer bystander mutations at the targeted locus (6). However, the experimental design of prime editors is much more complex than other CRISPR methods.

Despite its immense potential, genome editing still suffers from technical difficulties. An increase in efficiency, specificity, and targetability in genome editing technologies remains a challenge that will need to be overcome before its full potential can be realized. In addition, the identification and further development of recently described CRISPR-targeted transposases and recombinases represent an exciting area of research in genome editing that may enable more precise targeting of loci (22, 63, 123). Therefore, we speculate that genome editing will accelerate the translation of genomic information to therapeutic strategies.

4.4. Other Considerations

As the precise characterization of variant function relies on robust experimental design, important considerations should be taken when designing and executing functional experiments in order to gain mechanistic insights. Given the highly specific cell type– and cell state–dependent nature of gene regulation, particularly as it relates to enhancer function (97, 112), noncoding regulatory variants are likely to influence target gene expression and the associated phenotype only under highly specific conditions. Thus, recapitulating the parameters that influence variant function, including cell type, environmental conditions, and transient perturbations, must be considered, as variants may only show a phenotypic response to such settings. Integration of tissue-specific gene expression and genomic annotations of candidate regulatory variants, in particular at single-cell resolution, can allow for the prioritization of associated loci in specific cell types and states.

Despite the immense utility of experiments performed in human immortalized cell line models, the demonstration of a variant's impact on an altered phenotype following allelic substitution, either in vitro in primary cell culture or in vivo in animal models, is favorable to precisely evaluate the function of noncoding regulatory variants. The use of patient-derived primary cells is a powerful system given that they mimic the exact genomic and cellular background naturally observed in the patient. However, although genome editing can be performed in cell line models, it is particularly challenging to perform in primary cells given their difficulty to culture. As an alternative, human induced pluripotent stem cells (hiPSCs), which can be differentiated into diverse cell types (135), are an elegant system to study the molecular mechanisms of genetic variants, especially during cell transition states that may be consequential to various phenotypes such as developmental conditions. Additionally, despite differences in their genome architecture, mammalian animal models are also attractive systems due to their anatomical and physiological conservation with humans. However, as specific phenotypes may not be recapitulated in model organisms, patient-derived xenografts can allow for the study of human cells in an animal setting, although they may still lack the ability to recapitulate the physiology of native tissues.

5. CONCLUSION

The ever-growing computational toolbox (GWAS, rare-variant association tests, and machine learning) with the integration of large-scale data sets of functional annotations has enabled the prioritization of noncoding variants in relevant cell types and tissues. Scalable functional assays have significantly furthered our understanding of how noncoding regulatory variants may influence associated phenotypes. The insights gained from these studies have immense potential to drive translational advances that may enable more effective disease prevention and treatment, such as gene therapy strategies (36, 37).

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health (UM1 HG009393, R01 DK115398, and R01 AG077899) to H.Y.

LITERATURE CITED

1. 1000 Genomes Proj. Consort., Auton A, Brooks LD, Durbin RM, Garrison EP, et al. 2015. A global reference for human genetic variation. *Nature* 526(7571):68–74
2. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. 2010. A method and server for predicting damaging missense mutations. *Nat. Methods* 7(4):248–49
3. Agarwal V, Shendure J. 2020. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Rep.* 31(7):107663
4. Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33(8):831–38
5. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* 507(7493):455–61

6. Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, et al. 2019. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* 576(7785):149–57
7. Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339(6123):1074–77
8. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, et al. 2021. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* 18(10):1196–203
9. Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, et al. 2021. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* 53(3):354–66
10. Bell O, Tiwari VK, Thomä NH, Schübeler D. 2011. Determinants and dynamics of genome accessibility. *Nat. Rev. Genet.* 12(8):554–64
11. Bishop CM. 2006. *Pattern Recognition and Machine Learning*. New York: Springer
12. Bocher O, Génin E. 2020. Rare variant association testing in the non-coding genome. *Hum. Genet.* 139(11):1345–62
13. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, et al. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132(2):311–22
14. Bravo González-Blas C, Quan X-J, Duran-Romaña R, Taskiran II, Koldere D, et al. 2020. Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics. *Mol. Syst. Biol.* 16(5):e9438
15. Breheny P, Huang J. 2009. Penalized methods for bi-level variable selection. *Stat. Interface* 2(3):369–80
16. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10(12):1213–18
17. Campbell CT, Kim G. 2007. SPR microscopy and its applications to high-throughput analyses of biomolecular binding events and their kinetics. *Biomaterials* 28(15):2380–92
18. Cano-Gamez E, Trynka G. 2020. From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases. *Front. Genet.* 11:424
19. Caron B, Luo Y, Rausell A. 2019. NCBoost classifies pathogenic non-coding variants in Mendelian diseases through supervised learning on purifying selection signals in humans. *Genome Biol*. 20(1):32
20. Carroll D. 2011. Genome engineering with zinc-finger nucleases. *Genetics* 188(4):773–82
21. Chen KM, Wong AK, Troyanskaya OG, Zhou J. 2022. A sequence-based global map of regulatory activity for deciphering human genetics. *Nat. Genet.* 54(7):940–49
22. Chen SP, Wang HH. 2019. An engineered Cas-Transposon system for programmable and site-directed DNA transpositions. *CRISPR J.* 2(6):376–94
23. Chen W, Larrabee BR, Ovsyannikova IG, Kennedy RB, Haralambieva IH, et al. 2015. Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics* 200(3):719–36
24. Cho S, Kim H, Oh S, Kim K, Park T. 2009. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. *BMC Proc*. 3(Suppl. 7):S25
25. Cochran JN, Geier EG, Bonham LW, Newberry JS, Amaral MD, et al. 2020. Non-coding and loss-of-function coding variants in TET2 are associated with multiple neurodegenerative diseases. *Am. J. Hum. Genet.* 106(5):632–45
26. Connally NJ, Nazeen S, Lee D, Shi H, Stamatoyannopoulos J, et al. 2022. The missing link between genetic association and regulatory function. *eLife* 11:e74970
27. Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. 2014. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* 46(12):1311–20
28. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLOS Comput. Biol.* 6(12):e1001025
29. Derkach A, Lawless JF, Sun L. 2014. Pooled association tests for rare genetic variants: a review and some new results. *Stat. Sci*. 29(2):302–21
30. di Iulio J, Bartha I, Wong EHM, Yu H-C, Lavrenko V, et al. 2018. The human noncoding genome defined by genetic diversity. *Nat. Genet.* 50(3):333–37

31. Doudna JA, Charpentier E. 2014. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* 346(6213):1258096
32. ENCODE Proj. Consort. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74
33. Epinat J-C, Arnould S, Chames P, Rochaix P, Desfontaines D, et al. 2003. A novel engineered meganuclease induces homologous recombination in yeast and mammalian cells. *Nucleic Acids Res.* 31(11):2952–62
34. Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* 9(3):215–16
35. Ernst J, Kellis M. 2017. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* 12(12):2478–92
36. Esrick EB, Lehmann LE, Biffi A, Achebe M, Brendel C, et al. 2021. Post-transcriptional genetic silencing of *BCL11A* to treat sickle cell disease. *N. Engl. J. Med.* 384(3):205–15
37. Frangoul H, Altshuler D, Cappellini MD, Chen Y-S, Domm J, et al. 2021. CRISPR-Cas9 gene editing for sickle cell disease and β-thalassemia. *N. Engl. J. Med.* 384(3):252–60
38. Fudenberg G, Kelley DR, Pollard KS. 2020. Predicting 3D genome folding from DNA sequence with Akita. *Nat. Methods* 17(11):1111–17
39. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, et al. 2019. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* 51(12):1664–69
40. Gallagher MD, Chen-Plotkin AS. 2018. The post-GWAS era: from association to function. *Am. J. Hum. Genet.* 102(5):717–30
41. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* 17(6):877–85
42. Gong J, Mei S, Liu C, Xiang Y, Ye Y, et al. 2018. PancanQTL: systematic identification of *cis*-eQTLs and *trans*-eQTLs in 33 cancer types. *Nucleic Acids Res.* 46(D1):D971–76
43. Hardison RC, Taylor J. 2012. Genomic approaches towards finding *cis*-regulatory modules in animals. *Nat. Rev. Genet.* 13(7):469–83
44. Hayashi T, Ozaki H, Sasagawa Y, Umeda M, Danno H, Nikaido I. 2018. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat. Commun.* 9(1):619
45. He Z, Xu B, Buxbaum J, Ionita-Laza I. 2019. A genome-wide scan statistic framework for whole-genome sequence data analysis. *Nat. Commun.* 10(1):3018
46. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39(3):311–18
47. Hess GT, Tycko J, Yao D, Bassik MC. 2017. Methods and applications of CRISPR-mediated base editing in eukaryotic genomes. *Mol. Cell* 68(1):26–43
48. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ. 2008. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLOS Genet.* 4(7):e1000130
49. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. 2014. Identifying causal variants at loci with multiple signals of association. *Genetics* 198(2):497–508
50. Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JWJ, et al. 2016. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* 99(6):1245–60
51. Huang Y-F, Gulko B, Siepel A. 2017. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* 49(4):618–24
52. Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, et al. 2017. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* 27(1):38–52
53. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. 2016. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* 48(2):214–20
54. Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316(5830):1497–502
55. Joung JK, Sander JD. 2013. TALENs: a widely applicable technology for targeted genome editing. *Nat. Rev. Mol. Cell Biol.* 14(1):49–55

56. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581(7809):434–43
57. Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. 2018. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* 28(5):739–50
58. Kelley DR, Snoek J, Rinn JL. 2016. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* 26(7):990–99
59. Kim T, Wei P. 2016. Incorporating ENCODE information into association analysis of whole genome sequencing data. *BMC Proc.* 10(Suppl. 7):9
60. Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465(7295):182–87
61. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46(3):310–15
62. Klein JC, Agarwal V, Inoue F, Keith A, Martin B, et al. 2020. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods* 17(11):1083–91
63. Klompe SE, Vo PLH, Halpin-Healy TS, Sternberg SH. 2019. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* 571(7764):219–25
64. Komor AC, Badran AH, Liu DR. 2017. CRISPR-based technologies for the manipulation of eukaryotic genomes. *Cell* 168(1–2):20–36
65. Kouno T, Moody J, Kwon AT-J, Shibayama Y, Kato S, et al. 2019. C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. *Nat. Commun.* 10(1):360
66. Kouzarides T. 2007. Chromatin modifications and their function. *Cell* 128(4):693–705
67. Kurdistani SK, Grunstein M. 2003. In vivo protein–protein and protein–DNA crosslinking for genomewide binding microarray. *Methods* 31(1):90–95
68. Kurt IC, Zhou R, Iyer S, Garcia SP, Miller BR, et al. 2021. CRISPR C-to-G base editors for inducing targeted DNA transversions in human cells. *Nat. Biotechnol.* 39(1):41–46
69. Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CMT, Richards JB. 2012. The empirical power of rare variant association methods: results from Sanger sequencing in 1,998 individuals. *PLOS Genet.* 8(2):e1002496
70. Lanchantin J, Qi Y. 2020. Graph convolutional networks for epigenetic state prediction using both sequence and 3D genome data. *Bioinformatics* 36(Suppl. 2):i659–67
71. Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, et al. 2020. ClinVar: improvements to accessing data. *Nucleic Acids Res.* 48(D1):D835–44
72. Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, et al. 2015. A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* 47(8):955–61
73. Lee D, Yang J, Kim S. 2022. Learning the histone codes with large genomic windows and three-dimensional chromatin interactions using transformer. *Nat. Commun.* 13(1):6678
74. Lee H, Deignan JL, Dorrani N, Strom SP, Kantarci S, et al. 2014. Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* 312(18):1880–87
75. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536(7616):285–91
76. Li G, Martínez-Bonet M, Wu D, Yang Y, Cui J, et al. 2018. High-throughput identification of noncoding functional SNPs via type IIS enzyme restriction. *Nat. Genet.* 50(8):1180–88
77. Liu X, Noll DM, Lieb JD, Clarke ND. 2005. DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome Res.* 15(3):421–27
78. Liu Y, Liang Y, Cicek AE, Li Z, Li J, et al. 2018. A statistical framework for mapping risk genes from *de novo* mutations in whole-genome-sequencing studies. *Am. J. Hum. Genet.* 102(6):1031–47
79. Li Z, Li X, Liu Y, Shen J, Chen H, et al. 2019. Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. *Am. J. Hum. Genet.* 104(5):802–14
80. Longo SK, Guo MG, Ji AL, Khavari PA. 2021. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat. Rev. Genet.* 22(10):627–44
81. Loos RJF. 2020. 15 years of genome-wide association studies and no signs of slowing down. *Nat. Commun.* 11(1):5900

82. Lu T, Ang CE, Zhuang X. 2022. Spatially resolved epigenomic profiling of single cells in complex tissues. *Cell* 185(23):4448–64.e17
83. Maerkl SJ, Quake SR. 2007. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* 315(5809):233–37
84. Mangan RJ, Alsina FC, Mosti F, Sotelo-Fonseca JE, Snellings DA, et al. 2022. Adaptive sequence divergence forged new neurodevelopmental enhancers in humans. *Cell* 185(24):4587–603.e23
85. McKellar DW, Mantri M, Hinchman MM, Parker JSL, Sethupathy P, et al. 2022. Spatial mapping of the total transcriptome by in situ polyadenylation. *Nat. Biotechnol.* **https://doi.org/10.1038/s41587-022-01517-6**
86. Meng X, Brodsky MH, Wolfe SA. 2005. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat. Biotechnol.* 23(8):988–94
87. Moffitt JR, Hao J, Wang G, Chen KH, Babcock HP, Zhuang X. 2016. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *PNAS* 113(39):11046–51
88. Molla KA, Yang Y. 2019. CRISPR/Cas-mediated base editing: technical considerations and practical applications. *Trends Biotechnol*. 37(10):1121–42
89. Morrison AC, Huang Z, Yu B, Metcalf G, Liu X, et al. 2017. Practical approaches for whole-genome sequence analysis of heart- and blood-related traits. *Am. J. Hum. Genet.* 100(2):205–15
90. Mostafavi H, Spence JP, Naqvi S, Pritchard JK. 2022. Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. bioRxiv 2022.05.07.491045. **https://doi.org/10.1101/2022.05.07.491045**
91. Moyon L, Berthelot C, Louis A, Nguyen NTT, Roest Crollius H. 2022. Classification of non-coding variants with high pathogenic impact. *PLOS Genet*. 18(4):e1010191
92. Muerdter F, Boryń ŁM, Arnold CD. 2015. STARR-seq—principles and applications. *Genomics* 106(3):145–50
93. Nair S, Kim DS, Perricone J, Kundaje A. 2019. Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. *Bioinformatics* 35(14):i108–16
94. Nasser J, Bergman DT, Fulco CP, Guckelberger P, Doughty BR, et al. 2021. Genome-wide enhancer maps link risk variants to disease genes. *Nature* 593(7858):238–43
95. Nathan A, Asgari S, Ishigaki K, Valencia C, Amariuta T, et al. 2022. Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature* 606(7912):120–28
96. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, et al. 2011. Testing for an unusual distribution of rare variants. *PLOS Genet*. 7(3):e1001322
97. Nott A, Holtman IR, Coufal NG, Schlachetzki JCM, Yu M, et al. 2019. Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. *Science* 366(6469):1134–39
98. Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10(10):669–80
99. Paul S, Vadgama P, Ray AK. 2009. Surface plasmon resonance imaging for biosensing. *IET Nanobiotechnol*. 3(3):71–80
100. Persyn E, Karakachoff M, Le Scouarnec S, Le Clézio C, Campion D, et al. 2017. DoEstRare: a statistical test to identify local enrichments in rare genomic variants associated with disease. *PLOS ONE* 12(7):e0179364
101. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*. 20(1):110–21
102. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, et al. 2013. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* 152(5):1173–83
103. Rao S, Yao Y, Bauer DE. 2021. Editing GWAS: experimental approaches to dissect and exploit disease-associated genetic variation. *Genome Med*. 13(1):41
104. Rees HA, Liu DR. 2018. Base editing: precision chemistry on the genome and transcriptome of living cells. *Nat. Rev. Genet.* 19(12):770–88
105. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. 2019. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. 47(D1):D886–94

106. Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, et al. 2020. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* 578(7793):102–11

107. Ritchie GRS, Dunham I, Zeggini E, Flicek P. 2014. Functional annotation of noncoding sequence variants. *Nat. Methods* 11(3):294–96

108. Roulet E, Busso S, Camargo AA, Simpson AJG, Mermod N, Bucher P. 2002. High-throughput SELEX–SAGE method for quantitative modeling of transcription-factor binding sites. *Nat. Biotechnol.* 20(8):831–35

109. Sarnowski C, Satizabal CL, DeCarli C, Pitsillides AN, Cupples LA, et al. 2018. Whole genome sequence analyses of brain imaging measures in the Framingham Study. *Neurology* 90(3):e188–96

110. Schaid DJ, Chen W, Larson NB. 2018. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* 19(8):491–504

111. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* 328(5981):1036–40

112. Schoenfelder S, Fraser P. 2019. Long-range enhancer–promoter contacts in gene expression control. *Nat. Rev. Genet.* 20(8):437–55

113. Schones DE, Cui K, Cuddapah S, Roh T-Y, Barski A, et al. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell* 132(5):887–98

114. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, et al. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29(1):308–11

115. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, et al. 2015. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 31(10):1536–43

116. Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* 15(4):272–86

117. Shumaker-Parry JS, Aebersold R, Campbell CT. 2004. Parallel, quantitative measurement of protein binding to a 120-element double-stranded DNA array in real time using surface plasmon resonance microscopy. *Anal. Chem.* 76(7):2071–82

118. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15(8):1034–50

119. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* 40(W1):W452–57

120. Smedley D, Schubach M, Jacobsen JOB, Köhler S, Zemojtel T, et al. 2016. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. *Am. J. Hum. Genet.* 99(3):595–606

121. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.* 133(1):1–9

122. Stormo GD, Zhao Y. 2010. Determining the specificity of protein–DNA interactions. *Nat. Rev. Genet.* 11(11):751–60

123. Strecker J, Ladha A, Gardner Z, Schmid-Burgk JL, Makarova KS, et al. 2019. RNA-guided DNA insertion with CRISPR-associated transposases. *Science* 365(6448):48–53

124. Sung YJ, Korthauer KD, Swartz MD, Engelman CD. 2014. Methods for collapsing multiple rare variants in whole-genome sequence data. *Genet. Epidemiol.* 38(Suppl. 1):S13–20

125. Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, et al. 2016. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* 165(6):1519–29

126. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* 489(7414):75–82

127. Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.* 58(1):267–88

128. Tippens ND, Liang J, Leung AK-Y, Wierbowski SD, Ozer A, et al. 2020. Transcription imparts architecture, function and logic to enhancer units. *Nat. Genet.* 52(10):1067–75

129. Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A. 2011. Determinants of nucleosome organization in primary human cells. *Nature* 474(7352):516–20

130. van de Bunt M, Cortes A, IGAS Consort., Brown MA, Morris AP, McCarthy MI. 2015. Evaluating the performance of fine-mapping strategies at common variant GWAS loci. *PLOS Genet*. 11(9):e1005535
131. van der Wijst MGP, Brugge H, de Vries DH, Deelen P, Swertz MA, et al. 2018. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* 50(4):493–97
132. Vecchio-Pagán B, Blackman SM, Lee M, Atalar M, Pellicore MJ, et al. 2016. Deep resequencing of *CFTR* in 762 F508del homozygotes reveals clusters of non-coding variants associated with cystic fibrosis disease traits. *Hum. Genome Var*. 3:16038
133. Weissbrod O, Hormozdiari F, Benner C, Cui R, Ulirsch J, et al. 2020. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* 52(12):1355–63
134. Werling DM, Brand H, An J-Y, Stone MR, Zhu L, et al. 2018. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.* 50(5):727–36
135. Wilson KD, Wu JC. 2015. Induced pluripotent stem cells. *JAMA* 313(16):1613–14
136. Wissink EM, Vihervaara A, Tippens ND, Lis JT. 2019. Nascent RNA analyses: tracking transcription and its regulation. *Nat. Rev. Genet.* 20(12):705–23
137. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89(1):82–93
138. Yao L, Liang J, Ozer A, Leung AK-Y, Lis JT, Yu H. 2022. A comparison of experimental assays and analytical methods for genome-wide identification of active enhancers. *Nat. Biotechnol.* 40(7):1056–65
139. Yazar S, Alquicira-Hernandez J, Wing K, Senabouth A, Gordon MG, et al. 2022. Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. *Science* 376(6589):eabf3041
140. Yeh CD, Richardson CD, Corn JE. 2019. Advances in genome editing through control of DNA repair pathways. *Nat. Cell Biol.* 21(12):1468–78
141. Yuan G-C, Liu Y-J, Dion MF, Slack MD, Wu LF, et al. 2005. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* 309(5734):626–30
142. Zhang S, Cooper-Knock J, Weimer AK, Shi M, Moll T, et al. 2022. Genome-wide identification of the genetic basis of amyotrophic lateral sclerosis. *Neuron* 110(6):992–1008.e11
143. Zhang Z, Feng F, Liu J. 2022. Characterizing collaborative transcription regulation with a graph-based deep learning approach. *PLOS Comput. Biol.* 18(6):e1010162
144. Zhang Z, Park CY, Theesfeld CL, Troyanskaya OG. 2021. An automated framework for efficiently designing deep convolutional neural networks in genomics. *Nat. Mach. Intell.* 3(5):392–400
145. Zhao D, Li J, Li S, Xin X, Hu M, et al. 2021. Glycosylase base editors enable C-to-A and C-to-G base changes. *Nat. Biotechnol.* 39(1):35–40
146. Zhao S, Hong CKY, Myers CA, Granas DM, White MA, et al. 2023. A single-cell massively parallel reporter assay detects cell-type-specific gene regulation. *Nat. Genet.* 55(2):346–54
147. Zhou J. 2022. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat. Genet.* 54(5):725–34
148. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. 2018. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* 50(8):1171–79
149. Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* 12(10):931–34