ANNUAL REVIEWS

*Annual Review of Biomedical Data Science*

# Toward Identification of Functional Sequences and Variants in Noncoding DNA

Remo Monti[1,2] and Uwe Ohler[1]

[1]Max Delbrück Center for Molecular Medicine (MDC), Helmholtz Association of German Research Centers, Berlin Institute for Medical Systems Biology (BIMSB), Berlin, Germany; email: uwe.ohler@mdc-berlin.de

[2]Digital Health–Machine Learning, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, Potsdam, Germany



Annu. Rev. Biomed. Data Sci. 2023. 6:191–210

First published as a Review in Advance on June 1, 2023

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-122120-110102



## Keywords

gene regulation, enhancer, transcription, sequence analysis, machine learning, deep learning, rare variants, whole genome sequencing, genome wide association studies, variant effect prediction

## Abstract

Understanding the noncoding part of the genome, which encodes gene regulation, is necessary to identify genetic mechanisms of disease and translate findings from genome-wide association studies into actionable results for treatments and personalized care. Here we provide an overview of the computational analysis of noncoding regions, starting from gene-regulatory mechanisms and their representation in data. Deep learning methods, when applied to these data, highlight important regulatory sequence elements and predict the functional effects of genetic variants. These and other algorithms are used to predict damaging sequence variants. Finally, we introduce rare-variant association tests that incorporate functional annotations and predictions in order to increase interpretability and statistical power.

## 1. INTRODUCTION

Less than 2% of the human genome is coding—i.e., translated into mRNA and transcribed into protein or short peptides. The remaining 98% is non-(protein-)coding and contains many regions that fulfill structural and regulatory functions. Within two decades after the release of the human genome sequence, next-generation sequencing (NGS) has fundamentally changed the way we study both the coding and noncoding parts of genomes. Experiments that capture and determine the sequence of functional nucleotide elements have generated rich annotations of both the human genome and those of model organisms (1).

While evolutionary sequence conservation across species and the categorization of regions based on experimental readouts emerged early as important directions of genome research (2), few could have foreseen the transformative role of machine learning, and specifically deep learning, in the field. Driven by advances in computer vision and natural language processing (3), deep learning models have become ubiquitous in the study of the molecular functions encoded in the genome (4, 5).

NGS is also driving progress in the study of human genetic variation. The falling costs of whole-genome sequencing (WGS) have made it possible to collect these data for hundreds of thousands of individuals (6, 7). Recently, the UK Biobank has provided one of the largest and most widely accessible WGS datasets to date, genotyping about 150,000 individuals (6). These data contain hundreds of millions of predominantly rare genetic variants, many of which have never been observed before.

Understanding the role of these rare variants in health and disease relies on our ability to distinguish neutral variants from those that disrupt the function of the genomic sequences that contain them. While variant effect prediction (VEP) for coding regions is relatively well developed (8) and recent algorithms are good at identifying damaging nonsynonymous variants (9), the prediction and evaluation of variant effects for noncoding regions are lagging behind. However, understanding noncoding variation is critical for variant prioritization in sequencing-based genome-wide association studies (seqGWAS) (10, 11) or clinical variant interpretation (12).

Here we present concepts and recent advances in the computational analysis of noncoding regions and short genetic variants therein. Most examples provided concern the regulation of transcription, although the concepts generalize to other mechanisms of gene regulation as well. Deep learning is covered as a method to extract important sequence features and predict the functional effects of genetic variants, but we defer to other articles for detailed descriptions of architectures and algorithms (4, 5). We cover ways to evaluate functional effect predictions and use them to predict damaging genetic variants. Finally, we describe the incorporation of functional annotations into rare-variant association tests in seqGWAS.

## 2. NONCODING MECHANISMS AND THEIR REPRESENTATIONS IN DATA

### 2.1. General Properties and Classes of Regulatory Elements

While the noncoding genome contains many classes of sequences, such as repetitive, structural regions and transposable elements, the analysis of noncoding regions is largely concerned with gene-regulatory regions. This entails the localization of regulatory DNA or RNA sequences, identifying their function, and understanding how their function is determined by the cell machinery. Established (albeit simplistic) models place short regions of DNA/RNA into one or more functional classes, where single class instances are typically hundreds to a few thousands of nucleotides long.

Promoters, enhancers, and insulators [most prominently CCCTC-binding factor (CTCF) binding sites] are classes that regulate transcription (13, 14). 3′ and 5′ untranslated regions (UTRs)

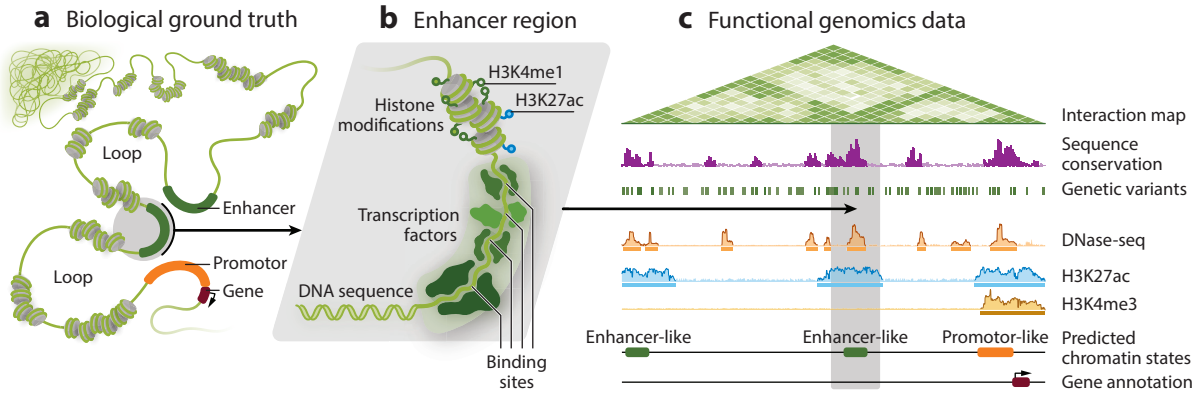


**Figure 1**

Model of transcriptional regulation and its observation through functional genomics data. (*a*) Two enhancers are brought close to a promoter, where they activate transcription of a gene. The DNA bends and forms loops. (*b*) Close-up of an enhancer bound by transcription factors. The flanking histones bare typical modifications (H3K27ac, H3K4me1). The DNA is nucleosome-free and, therefore, accessible. (*c*) Functional genomics data measure processes depicted in panels *a* and *b*. The reference sequence provides the coordinate system (*x*-axis). A 2D interaction map shows interaction frequencies between regions of the DNA. Contacts between enhancers and promoters appear as dark regions on the off-diagonal elements. Sequence conservation and the location of genetic variants are shown together with functional genomics data tracks. Peak calls appear as bars underneath continuous signal tracks. The enhancer highlighted in gray is conserved, is accessible (measured by DNase-seq), and has elevated H3K27ac (measured by ChIP-seq). Chromatin states are inferred from functional genomics data. Abbreviations: ChIP-seq, chromatin immunoprecipitation and sequencing; DNase-seq, DNase I hypersensitive sites sequencing.

and splice-regulatory regions influence posttranscriptional processes, such as RNA processing, localization, stability, and translation, which are regulated by RNA-binding proteins and RNAs (e.g., microRNAs) (15). All of these classes constitute *cis*-regulatory elements (CREs), where "*cis*" indicates interaction with nearby genes (i.e., discrete, functional transcribed regions) encoded on the same chromosome. The functions of CREs are established by the sequence-specific binding of regulators, which recruit or direct catalytic macromolecular complexes such as RNA polymerase II (Pol-II) for transcription or the RISC (RNA-induced silencing complex) for RNA degradation.

## 2.2. Enhancers, Promoters, and the *Cis*-Regulatory Code

A promoter is the DNA sequence surrounding and immediately upstream of transcription start sites (TSSs). The core promoter is the region directly at the TSS that serves as a docking platform for Pol-II recruitment; the proximal promoter is the region directly upstream of the TSS. Promoters interact with enhancers in order to activate transcription (13) (**Figure 1**).

The specificity of enhancers and promoters depends on the binding of selectively expressed transcription factors (TFs). A single TF can only bind compatible short sequence elements [~8–20 base pairs (bp)] and might interact with other TFs, for example, through cooperative or competitive binding. Typical transcriptional regulatory regions contain several binding sites for the same or a small subset of TFs (16). Through these mechanisms, the underlying DNA sequence determines the conditions under which a regulatory region becomes active.

The genome-wide patterns of binding sites constitute the so-called *cis*-regulatory code. The logic behind this code imposes evolutionary constraints, leading to conservation of regulatory elements and TF binding preferences across millions of years of evolution (17). Noncoding patterns of sequence conservation differ from protein-coding conservation: Constraints may be

placed not on the precise arrangement of binding sites, but rather on their overall presence and encoded binding affinity within a certain neighborhood. This often leads to binding site turnover, meaning that the function of nucleotides that trace back to the same ancestral nucleotide may differ, resulting in low colinear sequence similarity and the frequent emergence of new regulatory regions (18).

## 2.3. Loops and Topologically Associating Domains

On the linear genome sequence, enhancers can be found a few thousand (proximal) to hundreds of thousands (distal) base pairs away from the promoters with which they interact. Inside the nucleus, active enhancers are brought into physical proximity of promoters, and the frequency and duration of these enhancer–promoter contacts are correlated with the strength of transcription (activity by contact) (19, 20). When enhancer–promoter contacts are established, the nuclear DNA forms loops, with loop formation in mammals regulated by CTCF (21). A gene may be regulated solely by its promoter (e.g., in the case of some housekeeping genes), or by dozens of enhancers that sometimes cluster together as so-called superenhancers [e.g., in the case of developmental master regulators (22)]. In a specific condition, only a subset of enhancers will contact the promoter, and each enhancer encodes only part of the global expression pattern of a gene (e.g., its activity in one specific cell type). Large regions of the genome in which active sequences frequently contact each other are called topologically associating domains (23). These domains are separated by domain boundaries, across which contact frequencies are low, but which are dynamically shaped over the course of development (24, 25).

## 2.4. The Histone Code and DNA Methylation

TFs compete with histones for DNA binding. Histones are multimeric proteins involved in the packing of DNA as part of DNA–protein complexes called nucleosomes (26). The complex of DNA together with nucleosomes is referred to as chromatin. DNA that is occupied by nucleosomes is not directly accessible to other DNA-binding proteins, which limits its regulatory activity. The functions of histones are tuned by posttranslational modifications. Histones flanking promoters and enhancers are often acetylated at lysine 27 (H3K27ac) (27), histones flanking promoters are methylated at lysine 4 (H3K4me3) (28, 29), and H3K4me1 is often found at enhancers (30) (**Figure 1**). Other histone modifications indicate active repression (H3K27me3) or tight packing in inactive regions called heterochromatin (H3K9me3) (31, 32).

The DNA itself can also be chemically modified. The methylation of GpG-dinucleotides has been linked to repression, while demethylation has been linked to activation (33). As methylated residues are prone to mutational changes, promoters at active regions contain a disproportionally large number of these dinucleotides (34). Both the chemical modification of histones and the DNA affect the recruitment of factors that activate or repress transcription. In recent years, a plethora of chemical modifications of RNA have been identified as well, reportedly influencing diverse processes ranging from processing to stability and translation (35).

## 2.5. Functional Genomics Assays and Data

NGS-based assays provide noisy readouts of the gene-regulatory processes and interactions mentioned above, typically genome- or transcriptome-wide. These assays biochemically enrich for functional nucleotide fragments, which are then sequenced (36–42). The raw sequencing data constitute millions of so-called reads, i.e., sequences of typically a few hundred nucleotides (composed of letters A, C, G, and T) corresponding to the enriched fragments. The reads are mapped back to the reference genome in order to determine their source of origin. Specialized

experimental techniques further allow assigning reads to single cells (43, 44). For simplicity, this review focuses on bulk experiments that lack this level of resolution.

ChIP-seq (chromatin immunoprecipitation and sequencing) enriches for fragments that are bound by specific proteins of interest (36, 37), such as TFs or histones with specific modifications. DNase-seq (DNase I hypersensitive sites sequencing) (38) and ATAC-seq (assay for transposase-accessible chromatin using sequencing) (45) enrich for accessible (i.e., nucleosome-free) sequences such as enhancers, promoters, and CTCF-binding sites. RNA-seq (RNA sequencing) quantifies expressed transcripts (39), and CAGE (cap analysis of gene expression) measures transcription initiation (46). Bisulfite-seq measures CpG methylation (47).

After mapping to the reference genome, the number of reads originating from all positions is quantified. These per-position counts, also called coverage, are visualized as histogram-like tracks along the genome (**Figure 1*c***). Statistical methods are used to determine regions with significantly elevated read counts (peaks) (48), which can indicate, for example, accessible regions (in the case of DNase-seq or ATAC-seq) or regions bound by specific proteins (ChIP-seq). Other postprocessing steps might include converting discrete counts into relative enrichment over background, smoothing, and corrections for sequence biases (49).

Specialized methods such as chromatin conformation capture (3C), 4C (circular 3C), and Hi-C (high-throughput 3C) measure contact frequencies between DNA sequences in the nucleus (40, 42). The readouts from these experiments are 2D and symmetric, with both axes corresponding to locations on the reference. These contact-count matrices are visualized as heatmaps. Again, post-processing steps are typically applied for visualization and statistical testing (e.g., down-weighting expected short-range contacts).

Large consortia have made efforts to systematically annotate the genome of model organisms and humans by collecting functional genomics data across many tissues and time-points in development. By combining the signals from many experiments, these efforts have identified millions of candidate CREs (1). Numerous statistical methods have been developed to integrate experiments and define genome-wide chromatin states (i.e., locations with stereotypical combinations of histone modifications, accessibility, or methylation) corresponding to the classes of regulatory elements introduced above (e.g., enhancer-like states) (50, 51).

Reporter assays are used to validate functions of regulatory regions in vivo (in native context within a living organisms) or in vitro (in isolation from regular biological context, e.g., in a test tube). Typically, this requires cloning or barcoding fragments of interest next to a reporter gene whose activity can then be measured (e.g., by sequencing or fluorescence imaging) and compared across the fragments assayed. For example, reporter assays have been used to identify single developmental enhancers in mouse embryos (17). Massively parallel reporter assays measure the activities of many elements at once (52, 53). STARR-seq (self-transcribing active regulatory region sequencing) is an in vivo assay that inserts enhancer fragments in gene bodies and hence uses the RNA-seq readout directly as evidence for functionality (54). Reporter assays that include mutagenesis measure the effects of genetic perturbations on regulatory elements (55). However, typically only a limited set of regulatory elements and cell types can be investigated by a single experiment. Readouts from reporter assays have been used to train models for tissue-specific enhancer prediction (56) or to train and validate sequence models (57), as introduced in the next sections.

## 3. HUMAN GENETIC VARIATION IN HEALTH AND DISEASE

Genetic variants naturally occur throughout the genome. They are defined by their location relative to the reference genome and the reference and alternative sequence (i.e., the reference and alternative alleles). The more common sequence is defined as the major allele, and the less common sequence as the minor allele. The average human genome contains over four million

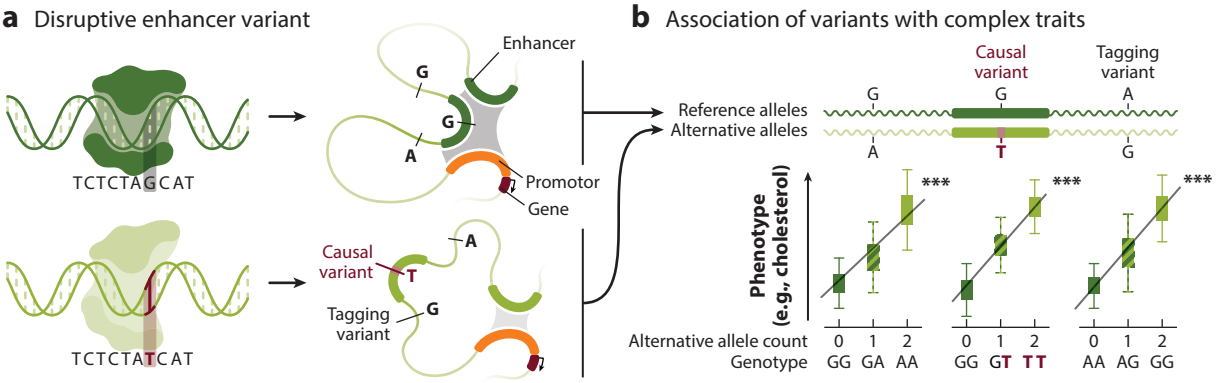


**Figure 2**

Functional enhancer variants and their observation through genome-wide association studies (GWAS). (*a*) An enhancer variant disrupts the binding of a transcription factor. This leads to fewer contacts with the promoter and, therefore, lower expression of a gene involved in cholesterol metabolism. The causal enhancer variant is inherited together with nearby noncausal tagging variants. (*b*) A GWAS measures the correlation of variants' alternative allele counts with cholesterol in a large population, shown by allele dosage plots with regression lines (effect sizes are exaggerated for clarity). All variants are significantly correlated with the trait (***$p$-value $< 5 \times 10^{-8}$), which hinders identification of the causal variant.

genetic variants, primarily consisting of single-nucleotide polymorphisms (i.e., variants that change a single base) and short insertions or deletions. The vast majority of these variants (>99%) lie in noncoding regions, including roughly 300,000 variants in possible enhancers (58). On the population level, the majority (>90%) of variants are rare, with minor allele frequencies below 0.1%, and over 40% are only observed in single individuals at current sample sizes (singletons) (6, 7). However, on the level of the individual, common variants outnumber rare variants by a large margin (>95% have minor allele frequencies above 0.5%) (58). Most are inconsequential, but some variants alter biomolecular function (**Figure 2*a***). While these molecular changes often go unobserved, they manifest themselves as measurable differences in traits at the population level (**Figure 2*b***). Here we briefly introduce population genetics, reviewed by Karczewski & Martin (59), and genome-wide association studies (60) through the lens of the functional analysis of noncoding regions. We defer to other reviews for a discussion of large structural variants (61).

## 3.1. Databases of Human Genetic Variation

Increasing access to genetic sequencing has inspired efforts to catalog the observed genetic variation in humans and its distribution across populations. In 2010, the 1000 Genomes Project set out to generate the first freely available reference of human genetic variation (58). While initial research focused on whole-exome sequencing (i.e., strongly enriching for coding regions), access to WGS data is increasing, driven by biobanks and other large-scale collaborative efforts (6, 7, 62, 63). The online database gnomAD provides statistics on observed variants across populations (64). Noncoding variants are placed into potential gene-regulatory contexts through overlap with functional annotations [e.g., UTRs or regions of accessible chromatin (**Figure 1*c***)] or by functional variant effect prediction, introduced below.

## 3.2. Genome-Wide Association Studies

Genome-wide association studies (GWAS) quantify the correlation of genetic sequence variants with heritable traits and disease (i.e., they measure the association of genotype and phenotype).

Biobanks that systematically collect genotypes and other data of hundreds of thousands of individuals have become major resources for GWAS (62, 65).

Due to cost and ease of processing, genotypes have mainly been measured using microarrays, which allow for predefined sets of typically hundreds of thousands of common variants. These are used to statistically impute millions of variants using variant reference panels (66). Imputation is possible because neighboring variants are inherited together, and therefore are highly correlated, a phenomenon termed as linkage disequilibrium (LD).

Phenotypes are derived from physical measurements, blood tests, imaging, questionnaires, or health records. Quantitative trait locus (QTL) studies are GWAS that use readouts from molecular assays as their phenotypes. For example, expression QTL (eQTL) studies, reviewed by Flynn & Lappalainen (67), correlate gene expression with genetic variants. However, even large eQTL studies like the GTEx (Genotype-Tissue Expression) project (68) have been limited in their sample sizes (<1,000 individuals).

The most common approach in GWAS independently tests each variant that reaches certain inclusion criteria (e.g., frequency and imputation quality) for its association with the phenotype. Specialized software allows these association tests to be rapidly performed, while correcting for covariates like age or sex and confounding by population stratification (e.g., ancestry) (69–71). GWAS summary statistics list the strength, direction, and *p*-values of associations between variants and the phenotype and can be visualized along the genome, similar to functional genomics tracks (72). However, LD prevents resolving association signals to the scale of regulatory elements (**Figure 2*b***). Statistical fine-mapping, reviewed by Schaid et al. and Cano-Gamez & Trynka (73, 74), is necessary to narrow down groups of correlated variants to one or a few candidate causal variants responsible for the associations at a genetic region. Many of these fine-mapping methods use functional genomics data because causal variants are enriched in regulatory regions like enhancers or promoters.

Rare-variant association studies that use sequencing-based genotyping suffer less from issues related to LD (e.g., signal localization). However, they face challenges with statistical power due to the large number of variants with very low frequencies and singletons. These issues are alleviated by variant aggregation and prioritization using variant effect predictions (see below).

The growing number of GWAS has inspired efforts to collect summary statistics and make them accessible through portals such as the NHGRI-EBI (National Human Genome Research Institute–European Bioinformatics Institute) GWAS Catalog (75). GWAS have revealed that common traits and diseases depend on many variants with small effects scattered throughout the genome (polygenicity) (59), which complicates downstream applications like the identification of relevant genes, cell types, or pathways through integration with functional genomics data (74). Besides providing biological insights, GWAS results are used for the construction of polygenic risk scores, as reviewed by Torkamani et al. and Lambert et al. (76, 77).

## 4. LEARNING THE REGULATORY CODE

Before quantifying the impact of regulatory variants, one needs to understand where the functional sequence elements are and what function they impart. Going back to the days of the Human Genome Project, models for gene regulation have been built at levels that range from interactions of binding sites with the DNA to the function encoded in a single enhancer, to the expression of the affected gene, and ultimately to the organismal phenotype.

### 4.1. Biophysical Models of Binding and Position Weight Matrices

Since the early days of computational biology, researchers have been interested in models that adequately describe the target sequences of proteins binding to nucleic acids. Few features are

identical across functional instances (restriction enzymes are an exception), and therefore flexible representations of short sequences of the same functionality (sequence elements), so-called motifs, have been needed. Looking at a typical TF binding site bound by the same TF, some positions will be flexible as to their nucleotide preferences, and others will be strict. Early attempts to capture this took the form of consensus sequences (regular expressions that include characters in addition to A, C, G, and T, such as W) to describe the presence of a nucleotide involved in a weak hydrogen bond (A or T). Such representations have the advantage of allowing for very fast string matching algorithms, but they are obviously limited and do not allow for more nuanced representations (e.g., that the presence of A or T may not be equally probable).

The most widespread representations of DNA/RNA sequence motifs have long been as matrices of size $4 \times N$. Given a set of aligned short sequences of length $N$, in our case target sequences of a specific TF, a position frequency matrix (PFM) first tabulates how many instances of each nucleotide are observed at each position and divides the entries by $N$. Each of the $N$ columns then represents a multinomial distribution over the alphabet (here, of the 4 nucleotides), meaning that all N positions are considered independently. To avoid zero probabilities, one typically modifies the initial counts by including prior information (e.g., in the form of pseudocounts).

Given a longer sequence, the probability of each subsequence of length $N$ being a target for the TF can then be computed by looking up the probabilities of each nucleotide of the subsequence at the corresponding position in the matrix and multiplying them. This operation of moving a short pattern (motif) along a larger signal (a DNA sequence) and recording the resulting scores is called a convolution. The matrix representation gained traction not least because of its clear connection to statistical mechanics and binding affinity (78). In practice, the PFM is converted into a position weight matrix (PWM) (or position-specific scoring matrix), which normalizes the observed frequencies by a background of, for example, the nucleotide composition of the genome (79).

Since their introduction, several extensions of PWMs have been proposed, particularly to address issues of fixed length (targets of some TFs contain a spacer region of variable length) and of the independence between positions of a binding site. More intricate probabilistic models such as hidden Markov models or Bayesian networks have been developed and have shown clear improvements for at least some TFs (80). For some time, small amounts of available data limited the practical applicability of models with a larger number of parameters, a situation that changed with the availability of ChIP-seq experiments and high-throughput assays for determining in vitro binding affinities with typically hundreds to thousands of putative TF binding sites (81, 82). Resources such as JASPAR provide precomputed models for large numbers of TFs for several eukaryotic species (83).

## 4.2. Deep Learning–Based Sequence Models

Deep neural networks are scalable, flexible, and automatically learn predictive tasks without the need for feature engineering (3) (i.e., the manual design or selection of informative input variables). These properties, and state-of-the-art performance, have led to their widespread adaptation in the field of genomics and elsewhere. The word "deep" in deep learning stems from the many layers present in these models, which consist of interconnected computational units called neurons. By stacking layers whose neurons perform linear operations followed by nonlinear activation functions, these models learn highly nonlinear functions of the input data (3).

Soon after deep neural networks began outperforming other methods on image classification tasks (84), they started being applied to predict gene-regulatory functions from sequence (85–87). The majority of sequence models are based on convolutional neural networks (CNNs).

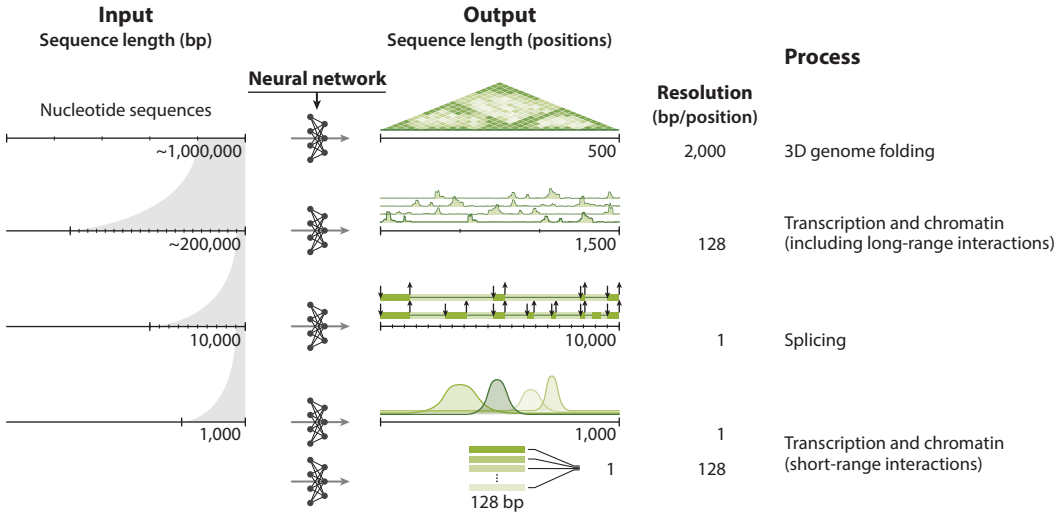


**Figure 3**

Sequence models of noncoding regulatory processes. Deep neural networks predict genomics data from the underlying nucleotide sequences at varying scales and resolutions. Long-range processes are modeled with larger input sequences. Because some models contain pooling operations, the output length can be shorter than the input length. Output data vary in their level of resolution, measured in base pairs per position. Splicing and highly resolved data (e.g., ChIP-nexus) can be modelled at base pair resolution. The model in the bottom row predicts coverage of the center 128-bp sequence from the 1,000-bp surrounding sequence. Abbreviations: bp, base pairs; ChIP-nexus, chromatin immunoprecipitation with nucleotide resolution through exonuclease, unique barcode, and single ligation.

Convolutional layers (i.e., the building blocks of CNNs) contain many fixed-size filters, which are scanned across sequences (4). In the first layer of a network, filters act as short motif finders (typically, ~3–25 bp) and can learn sequence patterns capturing known binding preferences of regulatory proteins (87, 88). However, the deep model structure allows for the representation of motif variants and combinations, positional dependencies, and multiple pattern occurrences within a larger sequence. Recent reviews have covered model architectures and applications in genomics (4, 89), as well as the interpretation of neural networks for genomics using so-called explainable AI methods (90). Here, we focus on core concepts for the analysis of noncoding regions using sequence models and highlight recent advances.

Sequence models predict functional annotations like those presented in **Figure 1** from the underlying nucleotide sequence and, optionally, inputs that capture sequence context (e.g., gene annotations) (91) (**Figure 3**). Sequences are typically represented by one-hot encoding on the nucleotide level (4), and models are trained to predict the presence of either binary features like peaks (classification) (85, 86) or continuous readouts such as signal strength (regression) (49). Software packages facilitate transforming genomics file formats into the data types required for deep learning (92, 93).

Models that predict features from DNA sequence can learn a variety of tasks at different levels of gene regulation, including binding of individual TFs (85, 86) or RNA-binding proteins (85, 91), accessibility of DNA (87), histone modifications (86, 94), transcription (95, 96), splicing (97), and 3D genome folding (98, 99) (**Figure 4**). Models are often trained to predict many outputs for the same input sequence (called multitarget or multilabel)—for example, readouts from many functional genomics experiments across time points and tissues. Because the regulatory code is largely conserved, models can be trained on data from multiple species at once (100). Specialized

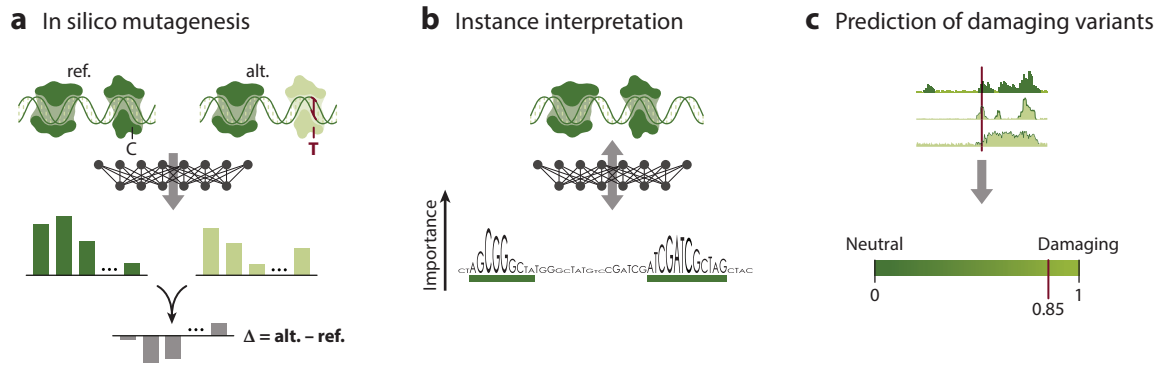


**Figure 4**

Sequence interpretation with neural networks, and prediction of damaging variants. (*a*) The variant C → T disrupts binding of a transcription factor. A deep neural network is presented with both the reference sequence (ref.) and alternative sequence (alt.) containing G (two forward passes). The model predicts lower probabilities of observing DNase accessibility peaks for the alternative sequence. The difference Δ between reference and alternative shows the direction of effects. (*b*) A gradient-based interpretation method is applied to the neural network using the reference sequence as input (forward and backward pass). The method highlights transcription factor binding sites. (*c*) The prediction of noncoding damaging variants typically relies on functional genomics data.

architectures confer biophysical interpretations to model parameters and components (101), and prior knowledge (e.g., TF binding preferences) can be used to initialize model parameters, which can increase performance (94).

The predictive tasks are reflected in the key properties of sequence models, including the length of the input sequence in nucleotides, the choice of output coding and resolution, and the receptive field (102) of neurons in the output layers (**Figure 3**). The receptive field measures the number of nucleotides in the input sequence that can influence the predictions at every position in the output layer; that is, it captures the ability of the model to integrate long-range interactions. CNNs use spatial pooling and fully connected layers before the output layer to grow the receptive field. Fully convolutional CNNs rely on pooling and exponentially dilated convolutions in order to grow the receptive field to tens of thousands of base pairs (49, 97). Models for processes involving the folding of DNA or RNA (e.g., transcription or splicing) benefit from larger receptive fields (96, 97).

Recently, transformers (103) have been used to increase the theoretical receptive field to hundreds of thousands of base pairs in order to improve the performance of transcription prediction (96). Models with larger receptive fields more accurately predict accessibility, ChIP-seq, and transcription (measured by CAGE), showing that the underlying data reflect the results of both local (~100 bp) and long-range (>10,000 bp) interactions (49). However, most signals tend to be captured at short scales (49, 96, 97), and predictive performance may not be the best measure to determine the utility of a model for downstream tasks (104).

## 4.3. In Silico Analysis of Regulatory Elements with Sequence Models

Given a trained sequence model, in silico mutagenesis predicts the functional effects of genetic variants (86, 87) (**Figure 4*a***). Because most models are trained only on the reference sequence, ignoring genetic variants, this task can be seen as a form of zero-shot transfer learning (i.e., repurposing for a different task without adjusting the model's parameters). First, a model predicts functional annotations for the reference sequence and the alternative sequence containing the variant. The difference between reference and alternative prediction is referred to as the allelic functional variant effect prediction.

Saturated in silico mutagenesis investigates all possible single-nucleotide substitutions of a sequence and can highlight important nucleotides. However, this approach is computationally expensive, as it relies on many predictions (forward passes through the network), and it might not capture all important parts of the sequence if there is redundancy (105, 106).

Methods for sequence instance interpretation rely on the forward propagation of activation differences (e.g., DeepLIFT or DeepSHAP) (105, 107) or backward propagation of gradients (108) in order to highlight important nucleotides (**Figure 4*b***). While computationally more efficient than mutagenesis, one drawback of these methods is that they have to compare against a background, the choice of which can considerably affect the results. Occlusion-based methods such as the recently proposed Scrambler (106) learn to predict occlusion masks that either completely destroy or preserve the prediction of a trained model by reshuffling. The occlusion masks (or their inverses) are used to highlight important subsequences.

Once important short subsequences are identified, downstream algorithms identify common patterns in these subsequences across instances in order to distill the global knowledge encoded in the model (91). This approach resembles the search for PWMs, but it is highly nonlinear. For example, TF-MoDISco (transcription factor motif discovery from importance scores) uses a series of clustering and other steps to generate consolidated motifs (109).

Learned attention mechanisms allow neural networks to dynamically weigh (attend to) and combine different parts of a sequence for prediction (110). This allows models to highlight important parts of sequences from a single forward pass. Co-attention to subsequences has been used to identify cooperative binding between TFs (111). Self-attention constructs 2D attention matrices that capture pairwise interactions between even very distant parts of sequences (103), and these 2D attention matrices can be useful for enhancer–promoter contact prediction (96).

Finally, sequence models allow complex in silico experiments to be performed that are difficult to perform in vivo. For example, by reversing all short sequences predicted to attract the TF CTCF, Fudenberg et al. (98) confirmed that their model learned the previously described influence of CTCF-binding motif directionality on loop formation (113). Avsec et al. (114) utilized a base-pair-resolution model to perform computational experiments to answer questions about spacing and cooperativity between TF binding sites.

## 4.4. Evaluation of Sequence Models

The evaluation of sequence models has focused on both the predictive performance of data held out for the target task (the task optimized during training) and models' ability to correctly predict the effects of genetic variants (transfer task). Holdout data comprise data on genomic regions not used during model training. Because functional genomics data are noisy, perfect predictions are never expected. The concordance between experimental replicates serves as a baseline for achievable model performance (49). Evaluating a model's performance on the same sequences that were used for training (data leakage) should be avoided, even for holdout experimental replicates, as it inflates performance estimates (115). This has been seen as less of a problem, and, in fact, almost unavoidable, for the transfer task of predicting genetic variant effects. Predicted effects on gene expression can be evaluated directly using data from reporter assays or fine-mapped eQTL data from matched cell types (96).

Functional predictions can be linked to association signals from GWAS or QTL studies through methods such as stratified LD score regression (116) or overlap/enrichment-based approaches, reviewed by Cano-Gamez & Trynka (74). The results of these analyses should always be contrasted against direct use of the training data or other strong baselines (104). Finally, the utility of models can be estimated on transfer tasks such as the prediction of damaging variants (86), as outlined in the next section.

## 5. PREDICTION OF DAMAGING VARIANTS

Damaging variants are functional variants that negatively impact biomolecular function (117). The majority of algorithms for the prediction of damaging variants have focused on coding variation. Protein-truncating variants and splice donor/acceptor variants are generally considered damaging (118), while variants that change the amino acid sequence are further contextualized using sequence conservation and molecular modeling (9, 119, 120). Sequence conservation can be considered either across species (121) or within the human population (64). The absence of genetic variation between species or within populations is seen as evidence for variation intolerance, and variants in variant-depleted regions are considered potentially damaging (64).

In addition to conservation, many algorithms for the prediction of damaging variants in noncoding regions take into account functional sequence annotations (8) (**Figure 4*c***). These annotations include the location relative to transcripts (e.g., promoter, UTR, intron, splice region) or transcript-agnostic annotations (e.g., CTCF binding site, accessible region, chromatin states). Annotations can also include functional variant effect predictions derived from sequence models. For example, the popular variant effect prediction tool CADD (combined annotation-dependent depletion) has been updated to include predictions from SpliceAI, a deep learning model that predicts splicing (97, 122). However, most tools for noncoding variants do not yet incorporate allelic functional effect predictions and therefore lack the ability to distinguish variants with opposing predicted functional effects at the same location.

Methods that predict damaging variants are optimized or evaluated using verified variants from databases [e.g., ClinVar (123) or HMGD (Human Gene Mutation Database) (124)] or allele frequency data. The latter approach distinguishes between common variants, which are depleted of damaging variants by natural (purifying) selection (59, 117), and rare variants like singletons, which have not been depleted. Generally, constructing appropriate training sets for these methods is challenging, and independent evaluations have exposed a lack of generalizability (125, 126). Therefore, algorithms that do not rely on predefined sets of variants have been proposed (9).

## 6. FUNCTIONAL PREDICTIONS IN RARE-VARIANT ASSOCIATION STUDIES

We have previously introduced GWAS that test variants individually. For the many rare variants observed by sequencing-based genotyping in seqGWAS, this approach lacks statistical power: First, the effect size for any single rare variant would need to be very large in order to reach statistical significance. Second, LD is low for rare variants, and therefore noncausal rare variants cannot effectively tag nearby causal variants. Third, the many very rare variants and singletons would drastically increase the burden of multiple rounds of testing (besides being statistically inappropriate).

To address these problems, rare-variant association tests increase the weights of potentially causal variants and aggregate variants in groups (127) (**Figure 5**). In addition, variant inclusion criteria based on functional annotations or effect predictions are used to increase the fraction of potentially causal variants. Filtering and weighting increase power if the causal mechanisms at a locus are correctly identified and noncausal variants are excluded from the test. As the true biology of each locus and its influence on the trait are unknown, it is common to vary the inclusion criteria for variants and perform multiple tests per locus (10, 128, 129). The *p*-values arising from different variant groups tested at the same locus can be aggregated, for example, using omnibus tests like the Cauchy combination test (130).

The two main types of association tests that aggregate variants are variant collapsing tests (also called burden tests) and kernel-based tests (also called variance-component tests). Variant

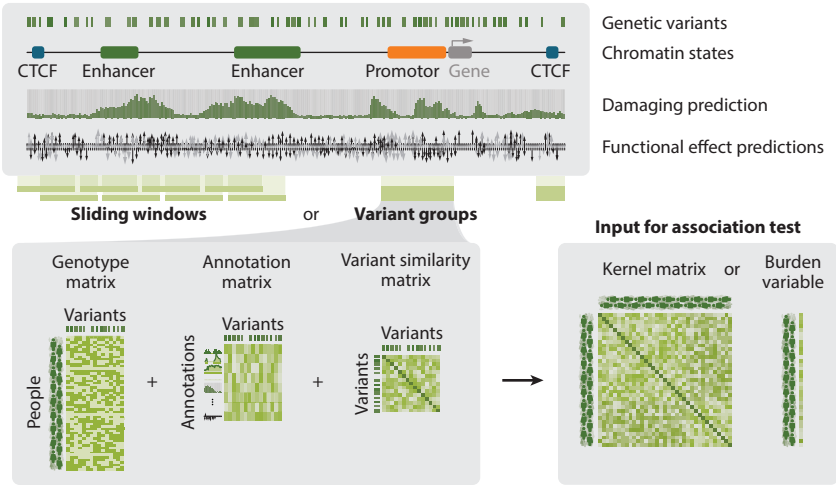


**Figure 5**

Sequence annotations in rare-variant association studies. Models predict effects for all observed genetic variants, and variants are grouped by the type of variant effect prediction, their locations (sliding windows), or their overlaps with predicted functional elements. Each group of variants is tested separately. Functionally informed rare-variant association tests integrate groups' genotypes with variant annotations and variant–variant similarities. Noncollapsing tests construct kernel matrices that capture genetic similarities between individuals, whereas burden tests (collapsing tests) test single variables. Abbreviation: CCTF, CCCTC-binding factor.

collapsing tests aggregate all variants within a group into a single variable before testing. They have high statistical power if the variants affect the phenotype in the same direction (e.g., increasing the probability of disease) and most included variants are causal. Kernel-based tests are advantageous if variants have opposing directions of effect or if there are fewer causal variants (131). Tests that combine collapsing and kernel-based tests in order to universally increase statistical power have been proposed (132, 133).

In exome-sequencing studies, which currently are still the majority of seqGWAS, variants are often grouped by gene, and potentially damaging coding variants are readily identified. Collapsing tests are effective in coding regions because the majority of damaging coding variants lead to a loss of function, and therefore have aligned effects on the phenotype within the same gene. We have recently shown that kernel-based tests have advantages if gain-of-function variants are present, which are typically limited to only few amino acid positions in a gene (129).

For noncoding variants, the grouping of variants is less straightforward, and sliding windows and groups defined by specific chromatin states have been used (e.g., enhancer-like regions, CTCF binding sites) (10). Methods that aggregate signal at the gene level by incorporating genome folding data have also been proposed (134, 135).

Rare-variant tests typically allow for the incorporation of variant weights, which can be derived using variant annotations such as allele frequencies or functional variant effect predictions (129, 132, 136–138). The directionality of functional variant effect predictions can also be taken into account, which could enable, for example, the separate collapsing of variants predicted to increase or decrease expression at a specific locus. The different components (i.e., matrices) used to construct a functionally informed rare-variant test are depicted in **Figure 5**. Both variant annotations (e.g., functional effect predictions) and variant similarities (e.g., physical proximity between variants)

can be taken into account. By incorporating functional variant effect predictions, functionally informed tests directly provide hypotheses on the implicated *cis*-regulatory mechanisms.

Reporting standards for aggregated tests are still being established (139). When reporting the results of grouped tests, it has been recommended to report all the variants that contributed to the combined test and, ideally, the type of variant effect and algorithm that was used to identify these variants (if functional predictions were performed). This allows novel variants that appear in the same sequence context to be contextualized.

## 7. DISCUSSION

As functional genomics and genotype data keep expanding, driven by advances in experimental and sequencing technologies, the analysis of noncoding regions is becoming increasingly complex. In this review, we have introduced topics relevant to the analysis of noncoding regions, with a focus on functional genomics data for gene expression, variant effect prediction with sequence models, and the integration of these predictions with (seq)GWAS. We focused on single short genetic variants; however, phased genotype data could allow variant interactions to be investigated.

We showed how deep learning has become a valuable tool to perform in silico experiments and predict the functional effects of variants through transfer learning. We see great promise in improvements of model interpretability (90), as well as in architectures that allow for the integration of large sequence contexts (103).

In order to make use of these advances in the context of rare variants, software for rare-variant association tests needs to be flexibly designed to accommodate the many types of variant annotations and effect predictions available. Reporting standards for associations found by such methods need to be established (139). This is critical for their application in personalized healthcare or for the inclusion of variants in polygenic scores.

While genomics deep learning has greatly profited from advances in natural language processing and image analysis, algorithms tailored to the specific properties of genomics data could further increase performance and interpretability. There is a need for models that more accurately predict cell type–specific effects by incorporating new types of data or combining existing datasets. Single-cell data analysis, reporter assays using CRISPR (19), and the analysis of complex structural variants provide promising avenues for future research.

In general, the linkage of functional variant effect predictions for noncoding variants to association signals from GWAS, their incorporation into rare-variant association tests, and their application in clinical settings require further research. We observe a lack of consensus in evaluation strategies and independent benchmarks, which makes it difficult to assess the utility of predictions for downstream applications (125, 126).

When designing algorithms for tasks like variant effect prediction, ethical considerations increasingly need to be taken into account. For example, if data used to train models come only from individuals of a specific ancestry, models may not generalize well to other ancestries, which could increase disparities in care. The reporting of results from such algorithms on the personal level also needs to be scrutinized (140).

Finally, self-supervised learning could help create powerful zero-shot models for noncoding variants, as has been shown for large corpora of text in natural language processing (141), and recently applied to protein variation (9).

## 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

U.O. acknowledges support from the German Federal Ministry of Education and Research (BMBF grant 01IS18053B MechML). We thank Philipp Jordan for his excellent contribution to the figures.

## LITERATURE CITED

1. Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, et al. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* 583:699–710
2. Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB. 2010. Annotating non-coding regions of the genome. *Nat. Rev. Genet.* 11:559–71
3. Goodfellow I, Bengio Y, Courville A. 2016. *Deep Learning*. Cambridge, MA: MIT
4. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. 2019. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* 20:389–403
5. Barshai M, Tripto E, Orenstein Y. 2020. Identifying regulatory elements via deep learning. *Annu. Rev. Biomed. Data Sci.* 3:315–38
6. Halldorsson BV, Eggertsson HP, Moore KH, Hauswedell H, Eiriksson O, et al. 2022. The sequences of 150,119 genomes in the UK Biobank. *Nature* 607:732–40
7. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, et al. 2021. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 590:290–99
8. Zhu C, Miller M, Zeng Z, Wang Y, Mahlich Y, et al. 2020. Computational approaches for unraveling the effects of variation in the human genome and microbiome. *Annu. Rev. Biomed. Data Sci.* 3:411–32
9. Frazer J, Notin P, Dias M, Gomez A, Min JK, et al. 2021. Disease variant prediction with deep generative models of evolutionary data. *Nature* 599:91–95
10. Hu Y, Stilp AM, McHugh CP, Rao S, Jain D, et al. 2021. Whole-genome sequencing association analysis of quantitative red blood cell phenotypes: The NHLBI TOPMed program. *Am. J. Hum. Genet.* 108:874–93
11. DiCorpo D, Gaynor SM, Russell EM, Westerman KE, Raffield LM, et al. 2022. Whole genome sequence association analysis of fasting glucose and fasting insulin levels in diverse cohorts from the NHLBI TOPMed program. *Commun. Biol.* 5:756
12. Ellingford JM, Ahn JW, Bagnall RD, Baralle D, Barton S, et al. 2022. Recommendations for clinical interpretation of variants found in non-coding regions of the genome. *Genome Med.* 14:73
13. Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* 15:272–86
14. Andersson R, Sandelin A. 2020. Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.* 21:71–87
15. Gebauer F, Schwarzl T, Valcárcel J, Hentze MW. 2021. RNA-binding proteins in human genetic disease. *Nat. Rev. Genet.* 22:185–98
16. Vierstra J, Lazar J, Sandstrom R, Halow J, Lee K, et al. 2020. Global reference mapping of human transcription factor footprints. *Nature* 583:729–36
17. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444:499–502
18. Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P. 2018. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat. Ecol. Evol.* 2:152–63
19. 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.* 51:1664–69
20. Zuin J, Roth G, Zhan Y, Cramard J, Redolfi J, et al. 2022. Nonlinear control of transcription through enhancer–promoter interactions. *Nature* 604:571–77
21. de Wit E, Vos ES, Holwerda SJ, Valdes-Quezada C, Verstegen MJ, et al. 2015. CTCF binding polarity determines chromatin looping. *Mol. Cell* 60:676–84
22. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, et al. 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 153:307–19

23. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, et al. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485:376–80
24. Sun F, Chronis C, Kronenberg M, Chen XF, Su T, et al. 2019. Promoter-enhancer communication occurs primarily within insulated neighborhoods. *Mol. Cell* 73:250–63
25. Winick-Ng W, Kukalev A, Harabula I, Zea-Redondo L, Szabó D, et al. 2021. Cell-type specialization is encoded by specific chromatin topologies. *Nature* 599:684–91
26. Kornberg RD. 1974. Chromatin structure: a repeating unit of histones and DNA: Chromatin structure is based on a repeating unit of eight histone molecules and about 200 DNA base pairs. *Science* 184:868–71
27. Cotney J, Leng J, Oh S, DeMare LE, Reilly SK, et al. 2012. Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. *Genome Res.* 22:1069–80
28. Bernstein BE, Humphrey EL, Erlich RL, Schneider R, Bouman P, et al. 2002. Methylation of histone H3 Lys 4 in coding regions of active genes. *PNAS* 99:8695–700
29. Schübeler D, MacAlpine DM, Scalzo D, Wirbelauer C, Kooperberg C, et al. 2004. The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev.* 18:1263–71
30. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *PNAS* 107:21931–36
31. Kouzarides T. 2007. Chromatin modifications and their function. *Cell* 128:693–705
32. Boros J, Arnoult N, Stroobant V, Collet JF, Decottignies A. 2014. Polycomb repressive complex 2 and H3K27me3 cooperate with H3K9 methylation to maintain heterochromatin protein 1$\alpha$ at chromatin. *Mol. Cell. Biol.* 34:3662–74
33. Holliday R, Pugh JE. 1975. DNA modification mechanisms and gene activity during development: Developmental clocks may depend on the enzymic modification of specific bases in repeated DNA sequences. *Science* 187:226–32
34. Gardiner-Garden M, Frommer M. 1987. CpG islands in vertebrate genomes. *J. Mol. Biol.* 196:261–82
35. Wiener D, Schwartz S. 2021. The epitranscriptome beyond $m^6a$. *Nat. Rev. Genet.* 22:119–31
36. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing *Nat. Mathods* 4:651–57
37. Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316(5830):1497–503
38. 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
39. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5:621–28
40. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326:289–93
41. Darnell RB. 2010. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *WIREs RNA* 1:266–86
42. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. 2012. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* 58:268–76
43. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, et al. 2017. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* 65:631–43.e4
44. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, et al. 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523:486–90
45. 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* 10:1213–18
46. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, et al. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *PNAS* 100:15776–81
47. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* 129:823–37

48. Feng J, Liu T, Qin B, Zhang Y, Liu XS. 2012. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* 7:1728–40
49. 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.* 28:739–50
50. Ernst J, Kellis M. 2017. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* 12:2478–92
51. Schreiber J, Durham T, Bilmes J, Noble WS. 2020. Avocado: A multi-scale deep tensor factorization method learns a latent representation of the human epigenome. *Genome Biol.* 22:81
52. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* 30:271–77
53. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, et al. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* 30:265–70
54. Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339:1074–77
55. Ernst J, Melnikov A, Zhang X, Wang L, Rogov P, et al. 2016. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat. Biotechnol.* 34:1180–90
56. Monti R, Barozzi I, Osterwalder M, Lee E, Kato M, et al. 2017. Limb-Enhancer Genie: an accessible resource of accurate enhancer predictions in the developing limb. *PLOS Comput. Biol.* 13:e1005720
57. Movva R, Greenside P, Marinov GK, Nair S, Shrikumar A, Kundaje A. 2019. Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. *PLOS ONE* 14:1–20
58. 1000 Genomes Proj. Consort. 2015. A global reference for human genetic variation. *Nature* 526:68–74
59. Karczewski KJ, Martin AR. 2020. Analytic and translational genetics. *Annu. Rev. Biomed. Data Sci.* 3:217–41
60. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, et al. 2021. Genome-wide association studies. *Nat. Rev. Methods Primers* 1:59
61. Ho SS, Urban AE, Mills RE. 2020. Structural variation in the sequencing era. *Nat. Rev. Genet.* 21:171–89
62. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, et al. 2016. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* 70:214–23
63. 100,000 Genomes Proj. Pilot Investig. 2021. 100,000 Genomes Pilot on rare-disease diagnosis in health care—preliminary report. *N. Engl. J. Med.* 385:1868–80
64. 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* 581:434–43
65. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, et al. 2015. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Med.* 12:e1001779
66. Das S, Forer L, Schönherr S, Sidore C, Locke AE, et al. 2016. Next-generation genotype imputation service and methods. *Nat. Genet.* 48:1284–87
67. Flynn E, Lappalainen T. 2022. Functional characterization of genetic variant effects on expression. *Annu. Rev. Biomed. Data Sci.* 5:119–39
68. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, et al. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45:580–85
69. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. 2011. Fast linear mixed models for genome-wide association studies. *Nat. Methods* 8:833–35
70. Loh PR, Kichaev G, Gazal S, Schoech AP, Price AL. 2018. Mixed-model association for biobank-scale datasets. *Nat. Genet.* 50:906–8
71. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, et al. 2018. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* 50:1335–41
72. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, et al. 2010. Locuszoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26:2336–37
73. Schaid DJ, Chen W, Larson NB. 2018. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* 19:491–504

74. Cano-Gamez E, Trynka G. 2020. From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases. *Front. Genet.* 11:424
75. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, et al. 2019. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47(D1):D1005–12
76. Torkamani A, Wineinger NE, Topol EJ. 2018. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* 19:581–90
77. Lambert SA, Abraham G, Inouye M. 2019. Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* 28:R133–42
78. Berg OG, von Hippel PH. 1988. Selection of DNA binding sites by regulatory proteins. *Trends Biochem. Sci.* 13:207–11
79. Stormo GD. 2000. DNA binding sites: representation and discovery. *Bioinformatics* 16:16–23
80. Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, et al. 2005. Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics* 21:2657–66
81. Mathelier A, Wasserman WW. 2013. The next generation of transcription factor binding site prediction. *PLOS Comput. Biol.* 9:e1003214
82. Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, et al. 2013. Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.* 31:126–34
83. Fornes O, Castro-Mondragon JA, Khan A, Van der Lee R, Zhang X, et al. 2020. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 48:D87–92
84. Krizhevsky A, Sutskever I, Hinton GE. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60:84–90
85. Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33:831–38
86. Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* 12:931–34
87. Kelley DR, Snoek J, Rinn JL. 2016. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* 26:990–99
88. Maslova A, Ramirez RN, Ma K, Schmutz H, Wang C, et al. 2020. Deep learning of immune cell differentiation. *PNAS* 117:25655–66
89. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. 2018. A primer on deep learning in genomics. *Nat. Genet.* 51:12–18
90. Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, Mostafavi S. 2022. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nat. Rev. Genet.* 24:125–37
91. Ghanbari M, Ohler U. 2020. Deep neural networks for interpreting RNA binding protein target preferences. *Genome Res.* 30:214–26
92. Chen KM, Cofer EM, Zhou J, Troyanskaya OG. 2019. Selene: a PyTorch-based deep learning library for sequence data. *Nat. Methods* 16:315–18
93. Kopp W, Monti R, Tamburrini A, Ohler U, Akalin A. 2020. Deep learning for genomics using Janggu. *Nat. Commun.* 11:3448
94. Quang D, Xie X. 2016. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* 44:e107
95. Zrimec J, Börlin CS, Buric F, Muhammad AS, Chen R, et al. 2020. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat. Commun.* 11:6141
96. 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.* 53:354–66
97. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, et al. 2019. Predicting splicing from primary sequence with deep learning. *Cell* 176(3):535–48.e24
98. Fudenberg G, Kelley DR, Pollard KS. 2020. Predicting 3D genome folding from DNA sequence with Akita. *Nat. Methods* 17:1111–17
99. Schwessinger R, Gosden M, Downes D, Brown RC, Oudelaar AM, et al. 2020. DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat. Methods* 17:1118–24

100. Kelley DR. 2020. Cross-species regulatory sequence activity prediction. *PLOS Comput. Biol.* 16:e1008050
101. Tareen A, Kinney JB. 2019. Biophysical models of cis-regulation as interpretable neural networks. arXiv:2001.03560 [q-bio.MN]
102. Luo W, Li Y, Urtasun R, Zemel R. 2016. Understanding the effective receptive field in deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.* 29:4905–13
103. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. 2017. Attention is all you need. *Adv. Neural Inform. Process. Syst.* 30:5999–6009
104. Dey K, Van de Geijn B, Kim SS, Hormozdiari F, Kelley D, Price A. 2019. Evaluating the informativeness of deep learning annotations for human complex diseases. *Nature* 11:4703
105. Shrikumar A, Greenside P, Kundaje A. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*, Vol. 70, pp. 3145–53. New York: Assoc. Comput. Mach.
106. Linder J, La Fleur A, Chen Z, Ljubetič A, Baker D, et al. 2022. Interpreting neural networks for biological sequences by learning stochastic masks. *Nat. Mach. Intell.* 4:41–54
107. Lundberg SM, Lee SI. 2017. A unified approach to interpreting model predictions. *Adv. Neural Inform. Process. Syst.* 30:4765–4774
108. Sundararajan M, Taly A, Yan Q. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*, Vol. 70, pp. 3319–28. New York: Assoc. Comput. Mach.
109. Shrikumar A, Tian K, Avsec Ž, Shcherbina A, Banerjee A, et al. 2018. Technical note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5. arXiv:1811.00416 [cs.LG]
110. Bahdanau D, Cho K, Bengio Y. 2014. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473 [cs.CL]
111. Ullah F, Ben-Hur A. 2021. A self-attention model for inferring cooperativity between regulatory features. *Nucleic Acids Res.* 49(13):e77
112. Deleted in proof
113. Guo Y, Xu Q, Canzio D, Shou J, Li J, et al. 2015. CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* 162:900–10
114. 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.* 53:354–66
115. Schreiber J, Singh R, Bilmes J, Noble WS. 2020. A pitfall for machine learning methods aiming to predict across cell types. *Genome Biol.* 21:282
116. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, et al. 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* 47:1228–35
117. MacArthur D, Manolio T, Dimmock D, Rehm H, Shendure J, et al. 2014. Guidelines for investigating causality of sequence variants in human disease. *Nature* 508:469–76
118. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, et al. 2016. The Ensembl Variant Effect Predictor. *Genome Biol.* 17:122
119. Ng PC, Henikoff S. 2003. Sift: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31:3812–14
120. Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genomics* 76:7.20.1–7.20.41
121. 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.* 15:1034–50
122. 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):886–94
123. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, et al. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 42(D1):980–85
124. Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, 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.* 133:1–9

125. Mahmood K, Jung CH, Philip G, Georgeson P, Chung J, et al. 2017. Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. *Hum. Genom.* 11:10
126. Liu L, Sanderford MD, Patel R, Chandrashekar P, Gibson G, Kumar S. 2019. Biological relevance of computationally predicted pathogenicity of noncoding variants. *Nat. Commun.* 10:330
127. Lee S, Abecasis GR, Boehnke M, Lin X. 2014. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* 95:5–23
128. Wang Q, Dhindsa RS, Carss K, Harper AR, Nag A, et al. 2021. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* 597:527–32
129. Monti R, Rautenstrauch P, Ghanbari M, James AR, Kirchler M, et al. 2022. Identifying interpretable gene-biomarker associations with functionally informed kernel-based tests in 190,000 exomes. *Nat. Commun.* 13:5332
130. Liu Y, Xie J. 2020. Cauchy combination test: a powerful test with analytic *p*-value calculation under arbitrary dependency structures. *J. Am. Stat. Assoc.* 115:393–402
131. Lee S, Abecasis GR, Boehnke M, Lin X. 2014. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* 95:5–23
132. Lee S, Wu MC, Lin X. 2012. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 13:762–75
133. Chen H, Huffman JE, Brody JA, Wang C, Lee S, et al. 2019. Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *Am. J. Hum. Genet.* 104:260–74
134. Sey NY, Hu B, Mah W, Fauni H, McAfee JC, et al. 2020. A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nat. Neurosci.* 23:583–93
135. Ma S, Dalgleish J, Lee J, Wang C, Liu L, et al. 2021. Powerful gene-based testing by integrating long-range chromatin interactions and knockoff genotypes. *PNAS* 118(47):e2105191118
136. Jin B, Capra JA, Benchek P, Wheeler N, Naj AC, et al. 2022. An association test of the spatial distribution of rare missense variants within protein structures identifies Alzheimer's disease–related patterns. *Genome Res.* 32:778–90
137. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. 2015. MAGMA: generalized gene-set analysis of GWAS data. *PLOS Comput. Biol.* 11(4):e1004219
138. Li X, Li Z, Zhou H, Gaynor SM, Liu Y, et al. 2020. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat. Genet.* 52:969–83
139. McMahon A, Lewis E, Buniello A, Cerezo M, Hall P, et al. 2021. Sequencing-based genome-wide association studies reporting standards. *Cell Genom.* 1:100005
140. Lewis AC, Green RC. 2021. Polygenic risk scores in the clinic: new perspectives needed on familiar ethical issues. *Genome Med.* 13:14
141. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, et al. 2020. Language models are few-shot learners. *Adv. Neural Inform. Process. Syst.* 33:1877–901

Annual Review of Biomedical Data Science

Volume 6, 2023

# Contents

**Errata**

An online log of corrections to *Annual Review of Biomedical Data Science* articles may be found at http://www.annualreviews.org/errata/biodatasci