Decoding disease: from genomes to networks to phenotypes

Aaron K. Wong[1,4], *Rachel S. G. Sealfon*[1,4], *Chandra L. Theesfeld*[2,4] *and Olga G. Troyanskaya*[1,2,3]✉

Abstract | Interpreting the effects of genetic variants is key to understanding individual susceptibility to disease and designing personalized therapeutic approaches. Modern experimental technologies are enabling the generation of massive compendia of human genome sequence data and associated molecular and phenotypic traits, together with genome-scale expression, epigenomics and other functional genomic data. Integrative computational models can leverage these data to understand variant impact, elucidate the effect of dysregulated genes on biological pathways in specific disease and tissue contexts, and interpret disease risk beyond what is feasible with experiments alone. In this Review, we discuss recent developments in machine learning algorithms for genome interpretation and for integrative molecular-level modelling of cells, tissues and organs relevant to disease. More specifically, we highlight existing methods and key challenges and opportunities in identifying specific disease-causing genetic variants and linking them to molecular pathways and, ultimately, to disease phenotypes.

Understanding human disease requires comprehensive interpretation of the genome, including characterization of the impact of any variant on gene function and regulation. Broadly, this means that for any letter change in DNA, we must precisely identify its effects on biochemical properties such as protein structure, splicing and levels of expression, and then interpret these effects in terms of their phenotypic consequences. In the past decade, genomic sequencing has generated enormous amounts of data, profiling both normal genetic variation and disease-related mutations[1–3]. Concurrently, functional experiments profiling the epigenomic landscape of various cells and tissues[4,5] have provided a window into the regulatory signals controlling where and when genes are expressed.

These experimental advances have fuelled the development of machine learning algorithms that can predict the functional impact of any DNA variant. Whereas interpretation of exonic mutations that disrupt protein coding and function is guided by the genetic code, for missense mutations of uncertain impact, computational approaches evaluating conservation[6–11], secondary and tertiary protein structures[12–16] and additional biophysical properties (for example, protein stability[17–19], preservation of interactions with proteins[20–22] and other small molecules[14,23,24]) have made great strides in predicting the molecular impact of any amino acid substitution on gene function. However, the majority of variants associated with disease lie outside exons in the non-coding portion of the genome that makes up more than 98% of human DNA and there is no uniform way to decode their impacts (FIG. 1a). Recently, computational approaches opened the door to interpretation of the molecular and phenotypic effects of regulatory mutations, including those never before observed in populations but that might appear in sequencing additional genomes[25–28]. In particular, a powerful kind of machine learning, deep learning, is opening frontiers in modelling complex biology and linking it to genomic sequence changes to give biomedical researchers and clinical geneticists insight into regulatory changes that contribute to human health and disease. Deep learning approaches, which typically require large amounts of training data, have become increasingly useful for diverse biological applications[29–33] as high-throughput data sets, particularly consortium efforts providing systematically generated data across tissues, conditions and diseases, and greater computing power have become available[34,35]. Furthermore, work to provide easy to use frameworks to specify and share deep learning models for genomics[36,37], improve interpretability of deep learning models[38,39] and allow automated optimization of network architecture[40] has improved the accessibility and power of deep learning approaches and made deep learning models the preferred method for many genomic applications[41].

Quantitative genetics studies have clearly shown that most complex human diseases involve contributions from multiple genetic variants, but there remains a ‘missing heritability’ problem: only a small fraction of

Deep learning
A family of machine learning approaches involving multilayer models composed of interconnected nodes, where the output of each node in the model is a function of its inputs.

[1] *Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, NY, USA.*

[2] *Lewis–Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA.*

[3] *Department of Computer Science, Princeton University, Princeton, NJ, USA.*

[4] *These authors contributed equally: Aaron K. Wong, Rachel S. G. Sealfon, Chandra L. Theesfeld.*

✉ *e-mail: ogt@cs.princeton.edu*

https://doi.org/10.1038/s41576-021-00389-x

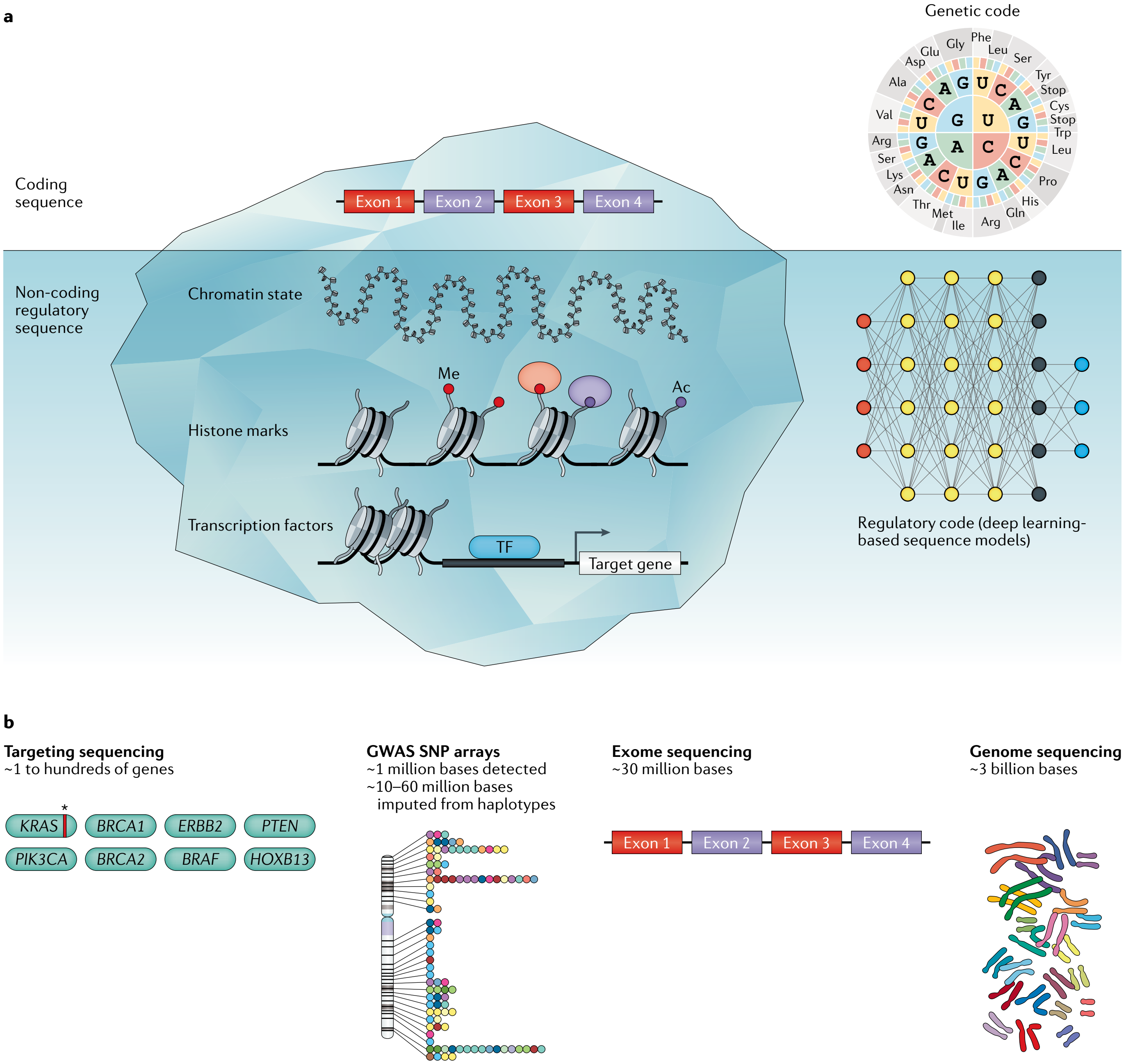


Fig. 1 | **Interpreting the other 99% of the genome. a** | The genetic code guides our interpretation of disease-associated mutations in coding regions of exon sequences; these mutations can disrupt protein coding and function. However, the majority of disease-associated variants lie outside exons in the non-coding portion of the human genome that makes up 99% of DNA and there is no uniform way to decode their impacts. Recent developments in deep learning-based sequence models address this challenge by modelling the relationship between non-coding sequences and properties influencing gene regulation, such as chromatin modifications, DNA accessibility and transcription factor (TF) binding. These models can then be used to predict the effects of variants in the non-coding portion of the genome. **b** | Detection of common and rare human genome variation. Variants are discovered by targeted sequencing of genes (for example, panels of cancer-related genes), or sequencing of whole exomes or whole genomes. Population-wide sampling in genome-wide association studies (GWAS) is done using single-nucleotide polymorphism (SNP) genotyping arrays, which are chips with probes for ~1 million locations of common variation in the genome. With this information, sequences for many more positions can be imputed using population reference sequence haplotypes as guides.

total heritability can be explained by disease-associated variants identified by genome-wide association studies (GWAS), even as these studies gain more power to detect associations[42]. Many studies have suggested that much of the heritability could be explained by the cumulative weak effects of many variants, many of which fall below statistical significance for association[43–46]. The question of how so many genomic regions with weak effects can influence disease susceptibility has prompted theories such as the 'omnigenic' model, where any gene expressed in a disease-relevant tissue can affect core disease genes, and thus disease risk, through interactions in a complex, interconnected network[47,48]. Thus, understanding how diverse protein products interact in the cell under specific contexts, such as in specific organs or tissues, is crucial to pinpointing genes related to disease and the 'network effects' of their dysregulation.

In this Review, we discuss the challenges and advances in using omics data to interpret genetic variants associated with disease. We cover the major sources of genetic variation and recent methodological advances that predict the regulatory effects of non-coding mutations — including biochemical, gene expression and pathogenic impact. Finally, we discuss methods that integrate omics data into tissue-specific systems-level models, which can then be used to identify the genes and dysregulated biological processes associated with specific diseases.

Characterizing genetic variation

Rapid technological improvements and associated reductions in costs have now made it possible to sequence the genomes of hundreds of thousands of people, and projects have launched that aim to sequence the complete genomes of millions of participants[49,50]. This unprecedented growth in available genomic data has resulted in enormous progress in deciphering the genetic underpinnings of human phenotypic variation and disease traits. At the same time, understanding the functional impact of specific genetic changes remains challenging, especially for alterations in non-coding regions of the genome.

Sequence variants can be profiled at several different levels of resolution (FIG. 1b). Single-nucleotide polymorphism (SNP) arrays can be used to genotype polymorphic positions throughout the genome, with more than a million SNPs frequently considered. For example, the landmark UK Biobank data set includes array-based genetic profiling for nearly 500,000 individuals, which can then be investigated in combination with phenotypic information and medical histories to identify disease-associated genetic signals[2]. To pinpoint genetic signatures for particular traits, GWAS can compare the variants observed in individuals who have a particular phenotype (such as a disease) with control individuals to identify loci that are associated with the trait[51]. Genotyping arrays incorporate common SNPs to tile along the genome backbone. To increase SNP density and add statistical power, millions of additional SNPs can be inferred by 'imputation' based on haplotypes of populations under consideration[52,53]. However, the associated SNPs may have indirect relationships to the trait. For instance, although the common variants on the array may be statistically associated with a disease, the identity of the actual allele that is causally linked to the trait can be a rare allele that is not assayed in either the array-based or imputation-based genotyping steps. Furthermore, GWAS may identify multiple positions in linkage disequilibrium that are associated with the trait, in which case further investigation of the associated SNPs through fine mapping is necessary to pinpoint trait-associated alleles[54].

Genetic variants can also be identified by genome sequencing, which can cover either limited regions or nearly the entire genome. Specific gene panels may be sequenced to test patients for exonic mutations that are known or suspected to be involved in disease. Gene panel testing is now routinely used in clinical practice, for example in prediction of breast cancer risk[55]. In whole-exome sequencing (WES), coding regions of essentially all genes in the genome are sequenced. This approach has the advantage of dramatically decreasing the fraction of the genome that must be sequenced relative to whole-genome sequencing (WGS), while still profiling the most readily interpretable genetic variants. Exome sequencing can also be used to detect somatic copy number variation, which is important for characterizing tumour genomes[56]. However, many disease-causing mutations are located in non-coding portions of the genome, and WES will not typically provide information on these variants, although targeted non-coding regions (for example, non-coding regions of known clinical importance) can be sequenced in addition to exomes. Moreover, WES sequencing may fail to capture some exonic portions of the genome, in particular regions that are GC-rich[57]. By contrast, in WGS, nearly the entire genome, including both coding and non-coding portions, is sequenced. The inclusion of non-coding portions of the genome provides crucial information for understanding the genetic architecture of disease traits, as an estimated ~90% of disease-related genomic intervals are in non-coding space[58]. WGS is increasingly seeing use in direct clinical assessments, identifying genetic contributors to both acute illness and disease risk[59–61].

Numerous key consortia have aimed to collect the genomic sequences of large numbers of individuals. A pioneering effort in this area was the 1000 Genomes Project, active from 2008 to 2015, which studied healthy individuals and aimed to catalogue most genetic variants at 1% frequency or higher in the sampled populations[1]. The recently launched All of Us consortium aims to sequence the genomes of at least a million American participants with diverse ethnic backgrounds and medical histories[62]. In parallel, several databases have been developed to make large collections of sequence data accessible to researchers. For example, the Genome Aggregation Database (gnomAD) contains more than 125,000 exome sequences and 70,000 whole-genome sequences as of its version 3 release[63]. Efforts to perform WGS on the UK Biobank cohort are in progress[2].

Discovery of the functional impact of mutations and attributing these effects to disease causation, is a major challenge. Large-scale sequencing data have been used to identify genomic regions that are likely to be disease-associated based on their patterns of observed versus expected variation, regardless of coding or non-coding space[64–66] (reviewed by Eilbeck et al.[67]). For population-based studies, although large sample sizes associate increasing numbers of loci with diseases, they still do not explain most of the estimated heritability of complex traits. For complex, highly polygenic diseases, enormous sample sizes (tens of millions of individuals) are needed to unravel causality of most related loci, each of which may individually have a small impact on susceptibility[68]. Furthermore, many possible mutations are never observed, either because they do not occur in a given sample or because they would be lethal before birth. Thus, methods that do not depend on observing an alteration in a population are needed to complement observation-based studies in order to disentangle the relationship between genomic sequences and disease traits.

Interpretation of coding mutations

Computational frameworks for assessing the impact of genetic alterations in the coding portion of the genome, although still an area of rapid development, are comparatively mature relative to approaches for interpreting non-coding variants[12,15,69–73]. Multiple types of evidence, such as the type of alteration to the protein sequence (missense, nonsense or frameshift), the degree of similarity between a reference and a substituted amino acid, the evolutionary conservation of the altered position and the predicted biophysical effect on protein structure can contribute to understanding the likely effect of a change in the coding portion of the genome. A large number of methods have been developed to leverage these factors to predict the impact of coding mutations[12,15,69–73]. We describe here a few commonly used approaches that illustrate different classes of methods for coding variant effect prediction.

One class of methods relies primarily on sequence conservation for predicting variant effects. For example, SIFT (Sorting Intolerant From Tolerant)[74,75] takes a conservation-based approach to variant effect prediction, examining conservation across related proteins in a protein family. The premise is that if an amino acid is highly conserved among members of a protein family, it is likely to be essential, whereas if more variation is tolerated at that position, it is less likely that an alteration will have a large effect.

Another key class of features for predicting the impact of a coding variant involve protein structural predictions. One example is the VIPUR (Variant Interpretation and Prediction Using Rosetta) framework, which incorporates protein structural models to predict variant effects[12]. Rather than simply determining how deleterious a mutation is likely to be, VIPUR uses the protein structural models to predict the precise effect of the variant (for example, whether the active site of the protein is disrupted, whether the stability of the protein is altered or whether the protein folding is likely to be disrupted). PertInInt explicitly integrates structural information with small-molecule binding information to predict structure-based impacts of protein variants on binding to RNA, peptides, ions and drugs[13,14].

Finally, there are many frameworks that integrate combinations of feature types to generate an aggregate prediction reflecting the impact of the mutation on various sequence and protein properties. An example of this type of method is PolyPhen-2 (Polymorphism Phenotyping v2), which uses a combination of sequence-based and structure-based features to predict the impact of a coding variant[16]. A naive Bayesian classifier is then used to predict the overall impact of the change from the individual features. Combined Annotation-Dependent Depletion (CADD)[69,76] and Eigen[77] are two other widely used machine learning frameworks that integrate a large number of feature types in order to predict the effects of both coding and non-coding variants.

Models for understanding the impact of coding genetic variants have played a key role in identifying alleles that contribute to human disease[78–81]. A strength of these methods is that their output is often readily interpretable. However, most approaches used to understand the effect of coding variants are highly tailored to leverage properties of protein-coding portions of the genome. Fundamentally different computational approaches are needed to elucidate the impact of non-coding variants.

Modelling transcriptional effects

As only a small fraction of the human genome is protein-coding and most variants are situated in the non-coding portion of the genome, developing approaches that can address the problem of understanding the effects of non-coding variants is a critical challenge. Methods that can model the relationship between non-coding sequences and properties influencing gene regulation, such as chromatin modifications, DNA accessibility and transcription factor binding, can be used to predict the effects of variants in the non-coding portion of the genome[25,27,28,82–84] (TABLE 1). Key to these efforts are large collections of genome-wide profiles for chromatin regulators in a range of cell types and tissues[4,5], enabling methods to predict the impact of variants even in primary tissues[25,27,28,83,85].

A strength of models that can predict variant effects based only on sequences is that they enable the prediction of molecular and disease impact of rare as well as commonly observed genetic alterations, and even mutations that have never (or not yet) been observed. This is an important feature because unique and rare variants are rapidly being discovered: a recent paper by the Trans-Omics for Precision Medicine (TOPMed) consortium identified 400 million varying bases in >53,000 newly sequenced whole genomes. Almost all variants were present at low allele frequency (<1%) and 46% were present in only one individual[86], making any statistical association studies infeasible for linking SNPs to disease due to lack of power.

Several frameworks use traditional supervised machine learning algorithms or probabilistic modelling to predict the impact of non-coding alterations. For example, gkm-SVM uses a support vector machine (SVM) with gapped *k*-mers as features to predict protein binding to sequences[28,87]. gkm-SVM significantly outperformed an earlier implementation using ungapped *k*-mer features, illustrating the utility of incorporating a longer sequence context for predicting regulatory function.

Success at such prediction problems has been seen recently using deep learning-based frameworks, which leverage genome-scale data on these attributes to generate sequence-based predictive models[25–27,82,84,88–90]. A network architecture that is particularly suited for extracting complex patterns from sequence data is the convolutional neural network (CNN), which was

Variant effect
The biochemical or phenotypic impact of a genetic variant relative to a reference allele.

Deleterious
The attribute of an allele as it relates to phenotypic impact; this can be through decreased organismal fitness that is often associated with increased disease risk.

Support vector machine
(SVM). A standard supervised machine learning approach that identifies the hyperplane (dividing line in high-dimensional space) that optimally separates positive examples from negative examples.

k-mers
Short lengths of nucleic acid used in computational algorithms, oligomers of '*k*' length, in bases.

Convolutional neural network
(CNN). A class of deep learning models that use structure in the input data (for example, adjacencies of pixels in an image or of base pairs in a sequence) to inform connections between nodes of the model. Successive outputs often model features at increasing spatial scales (for example, for sequence models: sequence → motifs → larger sequence contexts).

Table 1 | **Methods for transcriptional/biochemical impact**

Model overview			Input data	Resources	
Model	**Method**	**Prediction task**		**Public interactive Web interface**	**Access**
DeepSEA[25,26]	Deep learning, CNN	Chromatin, TF binding	TF, HM, DHS	https://hb.flatironinstitute.org/deepsea	https://github.com/FunctionLab/selene
gkm-SVM/ delta-SVM[28]	SVM	Chromatin, TF binding	TF, HM, DHS	–	http://www.beerlab.org/deltasvm/
DanQ[82]	Deep learning, CNN, BLSTM	Chromatin, TF binding	TF, DHS, HM	–	http://github.com/uci-cbcl/DanQ
Basset[27]	Deep learning, CNN	Chromatin accessibility	DHS	–	http://www.github.com/davek44/Basset
DeepCpG[89]	Deep learning, CNN, GRU	CpG state	Bisulfite sequencing	–	https://github.com/PMBio/deepcpg
ExPecto[95]	Deep learning, CNN, linear regression	Expression prediction	TF, HM, DHS, RNA-seq	https://hb.flatironinstitute.org/expecto	https://github.com/FunctionLab/ExPecto
Basenji[85]	Deep learning, CNN	Expression prediction	TF, DHS, HM, CAGE peaks	–	https://www.github.com/calico/basenji
BPNet, DeepLIFT, TFModisco[84]	Deep learning, CNN	TF binding	TF	–	https://github.com/kundajelab/bpnet/
ChromDragoNN[83]	Deep learning, CNN, ResNet	Chromatin accessibility	DHS, RNA-seq	–	https://github.com/kundajelab/ChromDragoNN
Xpresso[83,94]	Deep learning, CNN	Expression prediction	CAGE peaks, gene annotations	https://xpresso.gs.washington.edu/	–
AMBER[40]	Auto machine learning, RNN, deep learning, CNN	Chromatin, TF binding	TF, HM, DHS	–	https://github.com/zj-zhang/AMBER

All methods have the sequence as input and single-nucleotide resolution for variant effect predictions. All models can indicate cell type-specific predictions, depending on training data. BLSTM, bidirectional long short-term memory; CAGE, cap analysis gene expression; CNN, convolutional neural network; DHS, DNase hypersensitive site (DNA accessibility); GRU, gated recurrent network; HM, histone marks; ResNet, residual neural network; RNA-seq, RNA sequencing; RNN, recurrent neural network; SVM, support vector machine; TF, transcription factor.

initially widely applied to image analysis problems and uses spatial relationships across features to guide network connectivity. For sequence models, CNNs take raw DNA sequences as input and model regulatory and epigenomic elements at increasing genomic scales without the need for manual selection of features. Intuitively, nodes in successive layers can model individual motif elements, groups of motifs, regulatory regions and relationships across regulatory regions.

An early advance in modelling large-scale data was the DeepSEA framework, a CNN-based approach trained on a compendium of genome-wide transcription factor binding, DNase hypersensitivity and chromatin modification data, modelling >2,000 chromatin features that impact transcription in >200 cell types[25,26] (FIG. 2). The model then predicts the molecular effect of a variant with single base resolution given only sequence data. The DeepSEA framework can be tailored to specific diseases by training on disease-relevant data[91]. In another method, DeepBind[88], individual CNNs more accurately modelled protein-binding data compared with 26 other traditional non-deep learning and *k*-mer methods.

A related approach is used in the Basset framework, which learns a CNN model to predict cell type-specific DNase hypersensitive sites from sequence data[27]. Basenji, an extended and modified version of Bassett, can predict the quantitative impact of sequence alterations on regulatory modifications, as well as considering the impact of sequences in distal genomic regions[85].

An important challenge is predicting chromatin effects of human variation for any cell type. However, it is experimentally intractable to generate genome-wide chromatin immunoprecipitation followed by sequencing (ChIP–seq) profiles for all binding factors in all cell types and tissues. Thus, generalizable methods that can be used to impute missing data (for example, binding in cell types unobserved in model training) are crucial[92,93]. FactorNet, developed from DanQ, addressed this through combining recurrent neural network (RNN) architectures that use information learned previously to inform new training examples[82,90]. FactorNet trains on transcription factor binding in reference cell types and tests on distinct cell types. ChromDragoNN addresses prediction across cell contexts through coupling the learning of *cis*-regulatory sequences with gene expression of *trans* regulators using multimodal, residual neural network architectures that have the flexibility to integrate distinct data types[83]. FactorNet and ChromDragoNN approaches demonstrate the promise that modelling the accessibility of a genomic region in available data can be a powerful baseline predictor for biological contexts with no available data. Deep neural networks have also been applied to epigenomics where sequence modifications are difficult to assay. For example, DeepCpG predicts missing methylation at CpG sites in single cells using a modular arrangement of neural networks to learn sequence patterns around measured CpG sites (using a CNN) and capture the

Sequence models
A deep learning framework that models the relationship between genetic sequences and properties influencing gene regulation.

Recurrent neural network
(RNN). A type of neural network in which learning of inputs is influenced by past instances of input examples, and so output varies depending on the sequence of inputs. For example, in speech recognition, applying the context of prior words is useful for determining the meaning of a new word. (Hidden layers pass weights to input information based on previously learned examples.)

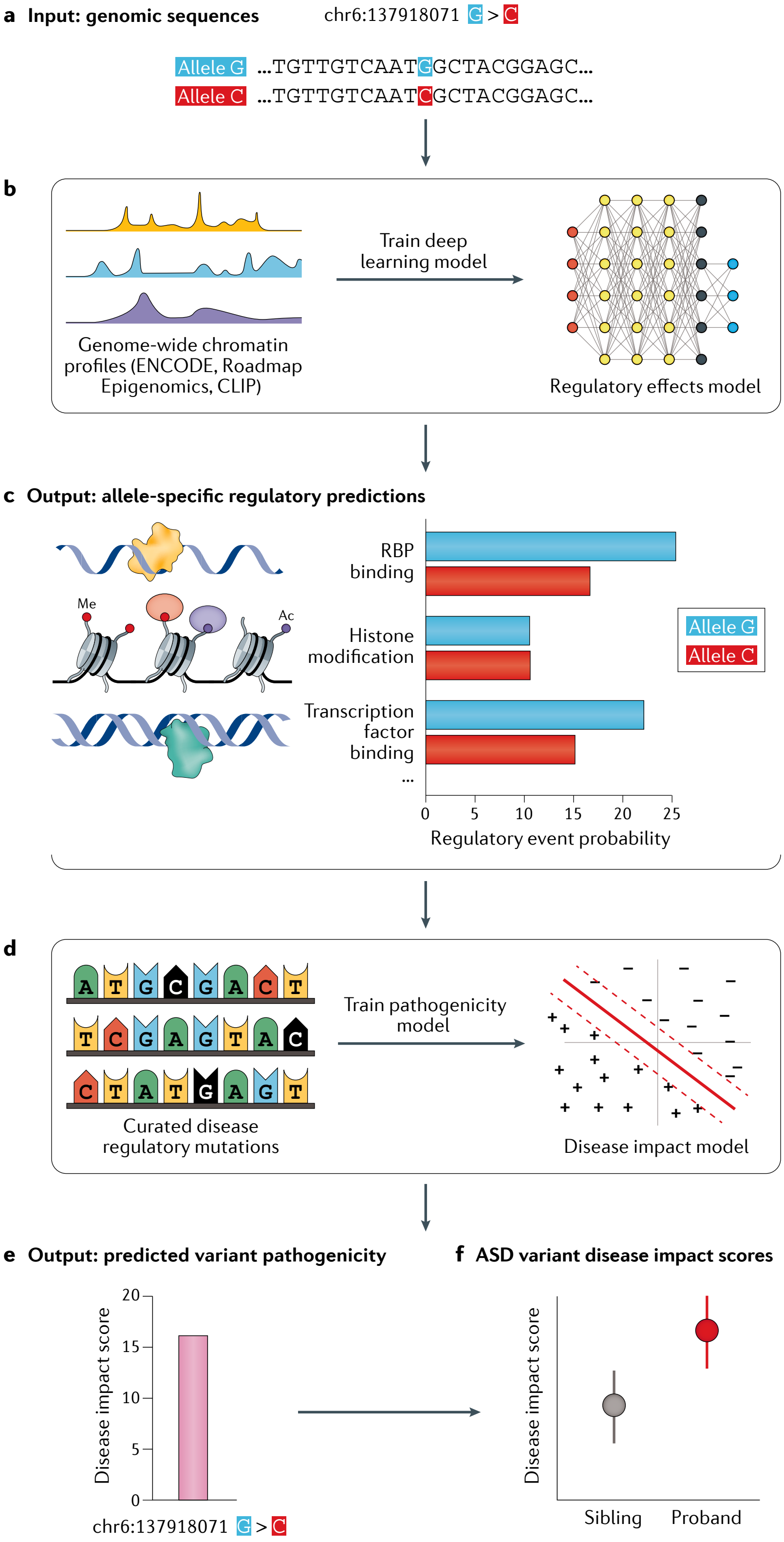

a Input: genomic sequences
chr6:137918071 G > C
Allele G ...TGTTGTCAATGGCTACGGAGC...
Allele C ...TGTTGTCAATCGCTACGGAGC...
b
Genome-wide chromatin profiles (ENCODE, Roadmap Epigenomics, CLIP)
Train deep learning model
Regulatory effects model
c Output: allele-specific regulatory predictions
Me
Ac
RBP binding
Histone modification
Transcription factor binding
...
Allele G
Allele C
0
5
10
15
20
25
Regulatory event probability
d
A T G C G A C T
T C G A G T A C
C T A T G A G T
Curated disease regulatory mutations
Train pathogenicity model
Disease impact model
e Output: predicted variant pathogenicity
f ASD variant disease impact scores
Disease impact score
20
15
10
5
0
chr6:137918071 G > C
Disease impact score
Sibling
Proband

Fig. 2 | **Learning the regulatory code to predict variant effects. a** | Sequence-based deep learning models predict the biochemical impact of alterations in non-coding sequences by comparing predicted effects between two alleles. **b** | Models are trained on genome-wide profiles of attributes such as chromatin modifications, transcription factor binding sites and RNA–protein binding sites. They can then predict the profiles of the novel input sequences, pinpointing the effect of sequence variants. As the models require only sequences as input, and are not trained on variant information, they can predict the effect of mutations that are rarely or never seen in sequence databases. **c** | The model predicts the specific impact of a given variant on attributes such as binding of specific RNA-binding proteins (RBPs), histone modifications and transcription factor binding. **d**,**e** | These biochemical effects can be used for further analyses. For example, to assess how biochemical effects relate to disease impact, a pathogenicity model was trained on predicted biochemical effects of known regulatory mutations that contribute to human disease (part **d**), enabling prediction of pathogenicity for any variant (part **e**). **f** | This framework was applied to predict the impact of de novo mutations in probands with autism spectrum disorder (ASD) and unaffected siblings. Mutations in probands had significantly higher predicted disease impact than mutations in unaffected siblings[26]. CLIP, cross-linking immunoprecipitation; ENCODE, Encyclopedia of DNA Elements.

neighbourhood of CpG site across cells (using a gated recurrent network), enabling single-cell genome-wide analysis of methylation states[89].

The development of interpretable models is crucial because they can lead to biological insight, improve user confidence in predictions and reveal potential technical biases. As deep learning frameworks are composed of many layers of transformations and, in effect, subject input data to a complex mathematical transformation, it is often difficult to trace how input features contribute to the final prediction. One approach to interpreting sequence models is to systematically alter bases in the input data and observe how the output predictions change (in silico mutagenesis)[25,88]. Another approach, exemplified by the DeepLIFT framework, uses backpropagation through the neural network to elucidate which input features contribute to which aspects of the output[38]. For sequence models, the ability of DeepLIFT to define motifs for transcription factor binding was applied to understand regulatory grammar syntax[84], and *trans*-regulatory mechanisms for the usage of transcription start sites[94].

Output from models that predict proximal molecular effects of variants can be incorporated into frameworks to predict the wider functional impact of these changes. Information on the chromatin modification and transcription factor binding profile of sequences can be used directly to predict effects on gene expression. ExPecto uses a modified version of the DeepSEA model and adds a tissue-specific, regularized linear model of gene expression based on the chromatin mark, DNase hypersensitivity and transcription factor binding profile of its upstream sequence[95]. By comparing the predicted expression of the gene in the presence of a variant versus a reference allele, ExPecto can infer the likely effect of a sequence mutation on gene expression in a specific tissue. A recent approach integrated SNP effect predictions from multiple methods, used multiple approaches to link genes to SNPs and predicted expression from sequence features in order to infer disease impact of variants[96]. By leveraging predictions of the direct effects of non-coding changes (for example, alterations in transcription factor binding or epigenetic profile or expression in specific cell types and tissues) to understand the impact on gene expression, these frameworks facilitate the understanding of the disease impact of non-coding changes at the biochemical, regulatory and phenotypic levels.

Making deep learning models and resources available to the broader biomedical community is key to accelerating development and adoption by researchers. This includes user-friendly dynamic web interfaces where researchers can directly generate predictions and model repositories that collate available models[36]. Resources that empower scientists who are not deep learning experts to develop algorithms and train and test deep learning models include code libraries such as Selene[37].

Modelling post-transcriptional effects

Variants that alter post-transcriptional properties of genes, such as interaction with RNA-binding proteins or splicing, can also contribute to disease risk[97–100]. Sequence-based deep learning models can be used to predict the precise post-transcriptional effect of a specific variant, including regulatory effects of synonymous mutations[26,101–106] (TABLE 2).

One key task is to predict the impact of variants that affect splicing, and various innovative methods have taken advantage of the molecular mechanisms of splicing to detect splicing events. For example, the DeepSplice and SpliceRover frameworks use CNN models to predict splice junctions from sequences[101,104]. The Deep Splicing Code framework trained multiple deep learning models for distinct prediction tasks, including identifying whether an exon has alternative 5′ or 3′ splice sites and determining whether an exon is constitutive or alternatively spliced[102]. Deep learning models have also been used to infer intronic RNA splicing branch points[107,108]. Despite the extensive progress in identifying alternative splicing events, largely through data from short-read next-generation sequencing technologies and genome annotations, determining full-length and alternative RNA isoforms, and how variants impact their formation, is still an open challenge[109,110]. Sequencing technologies (for example, long-read, direct RNA sequencing and targeted deep sequencing) and modelling methods are rapidly rising to find end to end full-length RNA species[111–113], reviewed recently in REF.[114].

A broader goal is predicting diverse post-transcriptional regulatory consequences, including not only splicing but also RNA structure, stability and translation efficiency through interactions with RNA-binding proteins[115–119], including tRNAs, and microRNAs[120,121]. Notably, sequence-based models of RNA–protein interactions using CNN frameworks have been developed for de novo non-coding variant effect prediction[26,88,122]. For example, Seqweaver was used to predict the post-transcriptional regulatory impact of variants from a whole-genome analysis of simplex autism families, demonstrating significant RNA regulation disruption in de novo variants of children with autism spectrum disorder[26]. In an even broader application of this model, Park et al. found that the dysregulation of wide-ranging categories of post-transcriptional regulation is a primary contributor to the inherited risk of psychiatric disorders[122].

Table 2 | **Methods for post-transcriptional impact**

Model overview			Input data			Resources	
Model	**Method**	**Prediction task**	**RBP profiles**	**Chromatin profiles**	**Genome annotations**	**Public interactive Web interface**	**Access**
DeepBind[88]	Deep learning, CNN	Chromatin and RBP binding	PBM, CLIP	SELEX,TFs	–	http://tools.genes.toronto.edu/deepbind	–
DeepSplice[188]	Deep learning, CNN	Splice junctions	–	–	Splice sites	–	https://github.com/zhangyimc/DeepSplice
LaBranchoR[107]	Deep learning, BLSTM	Intronic branch point sites	–	–	Annotated branch points	http://bejerano.stanford.edu/labranchor/	–
SpliceRover[101]	Deep learning, CNN	Splice donor and acceptor sites	–	–	Annotated donor and acceptor sites	http://bioit2.irc.ugent.be/rover/splicerover/	–
Seqweaver[26,122]	Deep learning, CNN	RBP binding impact	CLIP	–	–	https://hb.flatironinstitute.org/seqweaver	https://hb.flatironinstitute.org/seqweaver/about
Deep Splicing Code[102]	Deep learning, CNN	Alternative splice site usage	–	–	ESTs, HEXEvents annotated splice events	–	https://github.com/louadi/DSC
SpliceAI[109]	Deep learning, CNN	Splice donor and acceptor sites	–	–	GENCODE RNA annotations	–	https://github.com/Illumina/SpliceAI

All methods have the sequence as input and have single-nucleotide resolution for variant effect predictions. All models can indicate cell type-specific predictions, depending on training data. BLSTM, bidirectional long short-term memory; CLIP, cross-linking immunoprecipitation; CNN, convolutional neural network; EST, expressed sequence tag; PBM, protein-binding microarrays; RBP, RNA-binding protein; SELEX, systematic evolution of ligands by exponential enrichment; TF, transcription factor.

Deciphering pathogenic variant effects

In order to decipher the importance of a variant in contributing to disease processes, it is crucial to go beyond biochemical effect and understand the disease impact and, ultimately, clinical consequence. Several data resources aggregate current knowledge on variant impact. For example, the ClinVar database is a public resource from the US National Institutes of Health (NIH) that has so far compiled information on >800,000 genetic variants annotated to multiple diseases and with various levels of clinical significance using literature to document evidence and panel review to indicate confidence in each annotation (ClinVar statistics, 8 December 2020)[123–126]. The commercial Human Gene Mutation Database (HGMD) is a curated collection of more than 275,000 variants (HGMD 2019.4) sourced from published associations between genetic variants and human disease[127]. The American College of Medical Genetics and Genomics currently maintains a list of 59 genes with mutations that lead to well-defined and highly penetrant clinical phenotypes[128] that can be systematically profiled in clinical genetic testing and reported. MaveDB[129] and BioGRID Open Repository of CRISPR Screens (ORCS)[130] collect experimentally determined molecular impacts for thousands of sequence variants systematically profiled as part of biomedical research[131]. Through synthesizing and making accessible the current state of knowledge on disease-causing variants, such databases provide an invaluable resource for interpreting genetic data.

However, such clinical information is available for a very small and likely biased fraction of variants, especially so in the non-coding space. Further systematic characterization of variants at the phenotypic and endophenotypic level, such as through GWAS, clinical databases and experimental studies, for example in organoids, is essential (see ‘Integrative network models’ below for discussion of systematic network-based methods deployed for this purpose). That said, full characterization of the genome must rely on computational approaches, especially for the vast non-coding genome (approximately >2.94 billion bp) where most disease-associated mutations map. Given the regulatory consequences of variants predicted by the deep learning models (described above and in FIG. 2 and TABLES 1,2), machine learning methods can be trained to predict how deleterious each variant might be on the phenotypic level by training on example variants with associated phenotypic consequences (FIG. 2d; TABLE 3).

The resulting pathogenicity scores predict which variants of molecular impact will also have potential clinical impact. Pathogenicity refers to the impact on organism fitness, which is often closely correlated with the appearance of diseases, especially those that manifest before or during reproductive years. For example, CADD scores[76,132] reflect the deleteriousness of variants (the scores differentiate variants based on the likelihood they are evolutionarily constrained or neutral) through integrating dozens of features including genome annotation (such as intron or untranslated region (UTR)), evolutionary (GERP++[7,10] and SIFT[73]), and functional information on variants in both coding (PolyPhen-2 (REF.[16])) and non-coding (epigenomic profiles from the Encyclopedia of DNA Elements (ENCODE)[133] and Roadmap Epigenomics[5]) portions of the genome. The predicted regulatory features from DeepSEA and Seqweaver for both coding and non-coding variants can be leveraged to predict the disease impact score (DIS) for any mutation. The DIS is calculated by training a logistic regression model with predicted regulatory effects for verified disease-causing mutations in HGMD (positives) and low-frequency non-disease variants (negatives). This score was used in whole-genome analysis to discover increased mutational burden in probands with autism compared with unaffected matched siblings, improving upon traditional count-based analyses by enabling assessment of the functional impact of de novo mutations[26]. Constraint violation scores predict, on a per-gene and per-tissue basis, disease allele versus non-disease allele status for variants without knowledge of disease association. These scores integrate predicted expression effects from ExPecto with inferred evolutionary constraints in a given tissue[95]. For example, if a variant (a SNP or small insertion or deletion (indel)) causes a specific highly expressed, positively constrained gene to have a significant decrease in expression, constraint is violated and the variant is likely to be deleterious and pathogenic. This method accurately prioritized putative causal SNPs in the GWAS catalogue over SNPs in linkage disequilibrium. Another approach, LINSIGHT, combines modelling the effects of natural selection (INSIGHT) with functional molecular data to predict the deleteriousness of non-coding variants[134,135].

There is great potential in developing more accurate pathogenicity scores, both through integrating additional information and by leveraging newly characterized disease mutations as such data become available. Benchmarking of variant prediction methods against well-annotated sets of disease mutations and functional assay data (for example, in the extensively characterized *BRCA1* and *BRCA2* genes) indicates that variant predictions are significantly correlated with, but do not fully recapitulate, experimental data, suggesting that although such methods can aid in clinical applications they cannot yet be used in isolation to confidently determine the clinical consequences of a variant[136–138]. Although the variant databases provide samples of disease mutations, the comparatively small number and lack of strong evidence for disease association makes for noisy representation of true disease mutations, their associated functional genomic profiles and evolutionary signatures. Thus, a limitation in developing any sequence model is the lack of unbiased experimentally and clinically confirmed variant effects that can be used as gold standards in evaluations. Despite this, an orthogonal analysis demonstrated that the predicted variant effects from sequence models are significantly enriched in disease heritability, providing additional evidence that they are informative for discovering disease-influencing variants[139].

With regard to evaluations, the ClinVar variants reviewed using community-agreed standards, girded with multiple lines of evidence and reviewed by an expert panel will become an important benchmark in developing and evaluating variant impact scores. Methods

Endophenotypic
An aspect of a complex trait that may be more experimentally measurable than the entire complex trait and may be closer to an underlying biological process. For example, educational attainment is an endophenotype examined in the study of autism genetics because it is a readily measurable trait associated with autism, and expression levels of insulin receptors are endophenotypes contributing to type 2 diabetes.

Non-coding genome
The portion of the genome that does not encode proteins, which comprises more than 98% of the total human genome length.

Probands
In a genetic study, individuals with the disease (typically, the particular affected individuals within families who are the starting points for genetic analyses).

Table 3 | **Methods for pathogenicity**

Model overview				**Input data**							**Resources**	
Model	**Method**	**Prediction task**	**Tissue specificity**	**RNA expression**	**Protein structure**	**Amino acid biochemistry/ biophysics**	**Genome annotation (CDS, UTR)**	**Conservation**	**Chromatin profiles**	**RBP profiles**	**Public interactive Web interface**	**Access**
PhastCons[6,9]	MSA, phylo-HMM	Within conserved elements, per-base estimate of negative selection	–	–	–	–	–	Yes	–	–	http://compgen.cshl.edu/phast/	https://github.com/CshlSiepelLab/phast
GERP++[7,10]	MSA, MLE	Per-base and per-element score of constraint	–	–	–	–	–	Yes	–	–	–	http://mendel.stanford.edu/SidowLab/downloads/gerp/
PhyloP[9,11]	MSA statistical analysis by clade	*P* value, conservation/ acceleration scores, per base (divergence from neutral rate, allowing clade-specific selection pressures)	–	–	–	–	–	Yes	–	–	http://compgen.cshl.edu/phast/	https://github.com/CshlSiepelLab/phast
CADD[69,76]	Multi-data integration classifier, SVM	Coding and non-coding variant impact: deleteriousness of SNVs and small indels based on functional and evolutionary information	–	–	–	Yes	Yes	Yes	Yes	–	https://cadd.gs.washington.edu/	–
LINSIGHT[8,135]	Multi-data integration score, GLM, probabilistic model of constraints	Non-coding mutation fitness impact based on functional and evolutionary information	–	RNA-seq	–	–	Yes	Yes	Yes	–	–	https://github.com/CshlSiepelLab/LINSIGHT
DANN[132]	Deep learning version of CADD, CNN	Coding and non-coding variant fitness impact for SNVs and small indels based on functional and evolutionary information	–	–	–	Yes	Yes	Yes	Yes	–	–	https://cbcl.ics.uci.edu/public_data/DANN/
Eigen[77]	Unsupervised meta-score	Non-coding mutation impact	–	–	Yes	Yes	Yes	Yes	TF, DHS, HM	–	–	http://www.columbia.edu/~ii2135/eigen.html
Constraint violation score (ExPecto)[95]	Multi-data integration score	Non-coding variant impact: disease risk associated with SNV and small indels based on variant impact violation of cumulative regulatory impacts across genomic interval	By gene and tissue	RNA-seq	–	–	Yes	–	TF, DHS, HM	–	–	Supp Data 2 has directionality scores for calculating CVS

Table 3 (cont.) | **Methods for pathogenicity**

Model overview				**Input data**							**Resources**	
Model	**Method**	**Prediction task**	**Tissue specificity**	**RNA expression**	**Protein structure**	**Amino acid biochemistry/ biophysics**	**Genome annotation (CDS, UTR)**	**Conservation**	**Chromatin profiles**	**RBP profiles**	**Public interactive Web interface**	**Access**
Disease impact score (DeepSEA)[26]	Multi-data integration classifier, SVM	Coding and non-coding variant impact: disease association of SNVs and small indels based on transcriptional regulatory impact in a genome sequence model	–	–	–	–	–	–	TF, DHS, HM	Yes	https://hb.flatironinstitute.org/deepsea/, https://hb.flatironinstitute.org/asdbrowser	–

CADD, Combined Annotation-Dependent Depletion; CDS, coding sequence; CNN, convolutional neural network; DHS, DNase hypersensitive site (DNA accessibility); GLM, generalized linear model; HM, histone marks; indel, insertion or deletion; MLE, maximum likelihood estimation; MSA, multiple sequence alignment; phylo-HMM, phylogenetic hidden Markov model; RBP, RNA-binding protein; RNA-seq, RNA sequencing; SNV, single-nucleotide variant; SVM, support vector machine; TF, transcription factor; UTR, untranslated region.

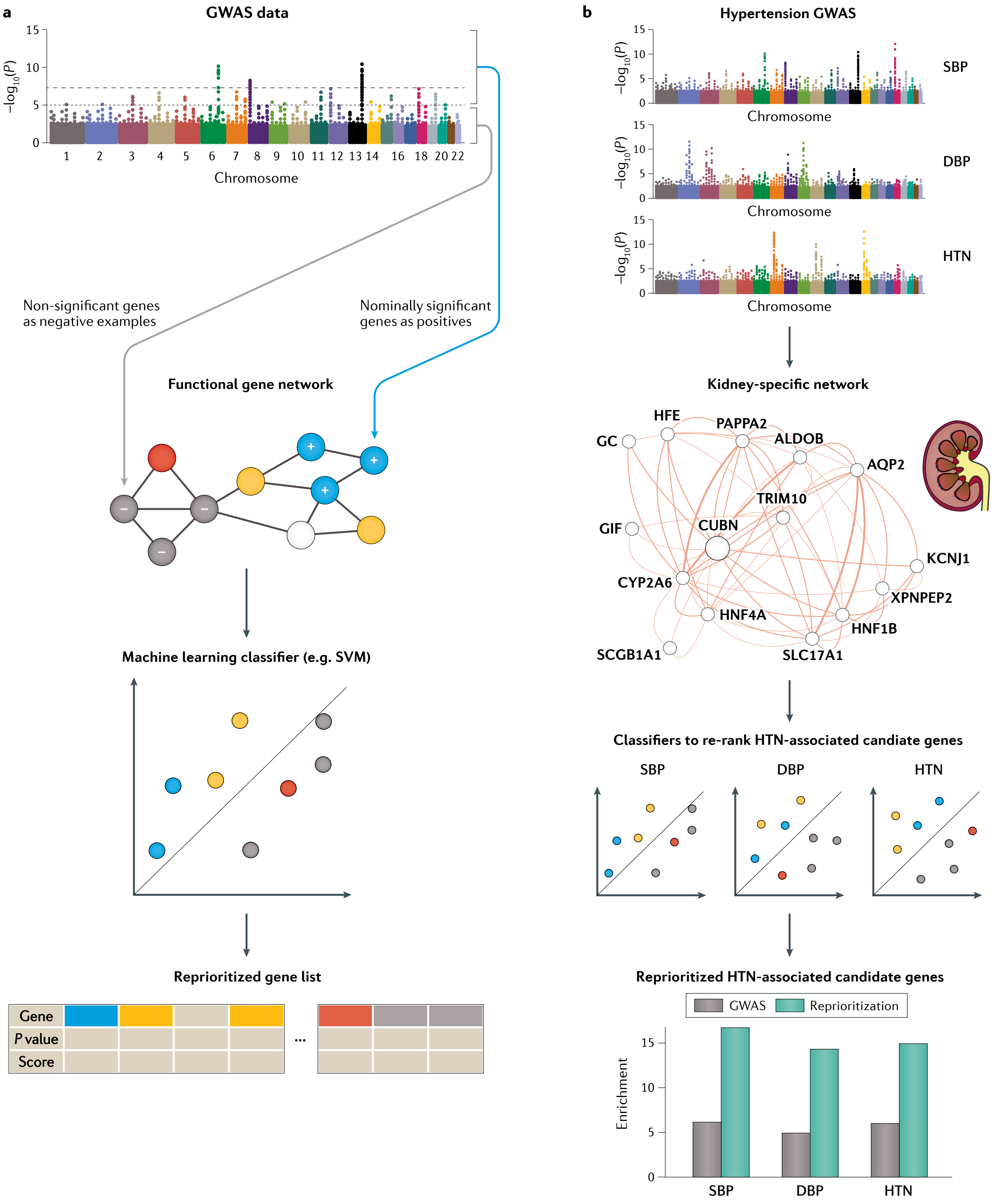
that use conservation scores or allele frequencies are of less value for sets of SNPs that contribute to disease but are not directly linked to fitness. For example, cumulatively, common variants contribute significantly to risk for complex diseases but may be under weak selection, and disease variants in linkage disequilibrium with beneficial alleles can be present in the population at higher frequency than expected. Such conservation-based methods will underscore these variant sets. Scores that accurately predict functional clinical impact that operate at a wide range of variant frequencies, in coding and non-coding regions of the genome and regardless of genetic architecture (for example, common inherited variants and de novo rare variants) will continue to have increasing utility for biomedical researchers studying the aetiology of diseases and clinicians choosing treatment modalities to meet the promise of personalized medicine.

Integrative network models

Although sequence models can predict the molecular effects of mutations, including on tissue-specific gene expression, explaining how these alterations give rise to disease phenotypes requires an understanding of the dysregulated pathways and processes. Recent work attempting to explain how even common variants with weak effects can still increase disease susceptibility theorizes an ‘omnigenic model’ wherein any gene expressed in a disease-relevant tissue can affect core disease genes through interactions between genes that function together in processes, pathways and larger networks. Thus, understanding how the molecular effects of genomic mutations influence disease requires models of cellular networks and pathways that are active in various tissues in healthy or disease states. Diverse functional genomic data, including gene expression[4,5,140,141] and proteomics[142,143], provide a window into the genome-scale interactions between biomolecules in various contexts. Inferring accurate models from these data is challenging because these data are produced on heterogeneous data modalities, experimental designs and technological platforms. Furthermore, the models must capture both shared and unique processes across cell types and physiological conditions. Sophisticated computational techniques are needed to extract context-specific biological relationships from the raw data sets.

Integrative approaches that combine diverse data sources can identify genome-scale functional maps of cells across biological and environmental contexts[144–149]. Whereas individual experiments vary in biological relevance and signal, by analysing data collections as a whole, faint but recurrent signals can be amplified. These methods deploy various machine learning techniques to summarize massive collections of human omics data (for example, transcriptional profiles, protein–protein interactions and sequence motifs) into gene interaction networks specific to a biological context. In these genome-wide networks, nodes represent genes, and edges reflect a predicted confidence that two genes participate in similar biological processes. These networks can provide both a systems-level view and specific experimentally testable hypotheses of gene function and interactions.

Context-specific networks can be constructed directly from correlations of gene expression profiles[150–152] or by combining global protein–protein interaction networks with disease or tissue-specific gene expression[153]. However, these methods are limited by the availability and quality of tissue expression profiling data, which for many human tissues and cell types remains challenging or even infeasible to assay experimentally. More recent methods address these limitations by applying machine learning to predict tissue-specific functional interactions from large compendia of genomic data[148,149,154–157]. For example, GIANT (Genome-Wide Integrated Analysis of gene Networks in Tissues)[148] (now available at humanbase.io) uses a regularized Bayesian classifier to predict networks for more than 100 tissues and cell types by assessing and weighting each data set by relevant tissue signal. The method can predict gene–gene interactions in tissues for which few tissue-specific genome-scale data exist. A systematic study by Huang et al.[158] evaluated 21 human interaction networks on their ability to predict disease genes and found ConsensusPathDB[159], GIANT[148] and STRING[147] to have the best performance



Tissue-specific network
A network that captures relationships between genes that participate in similar biological processes for a particular tissue or cell type.

overall, although the evaluation was limited to global (that is, tissue-unaware) networks.

Network analysis of quantitative genetics data

Network-based machine learning approaches can be used to leverage prior experimental knowledge and improve the interpretation of large-scale quantitative genetics studies or individual targeted, disease-specific studies. Intuitively, these methods use the functional genomic information about pathways encoded in these networks to increase the signal to noise ratio in genetics studies. Non-coding regulatory variants can be linked to putative target genes, and these target genes can then be subjected to network-based analyses.

One approach is to reprioritize GWAS hits based on information in networks that reflect relationships among genes encapsulated in thousands of biological experiments[160]. Broadly, these methods identify disease-specific connectivity patterns based on the behaviour of GWAS-prioritized genes, and reprioritize all genes based on these patterns (FIG. 3). Importantly, these methods are discovery-driven and do not depend on prior literature-based knowledge in their reprioritization; instead, they use the genome-wide information from disease-focused GWAS projects. For example, the NetWAS (Network-Wide Association Study) framework is a machine learning method that takes as input marginally significant GWAS hits as positives and lower-ranked genes as negatives, and uses tissue-specific network edge weights (for a disease-relevant tissue) as the feature set in a support vector classifier to reprioritize candidate genes[148]. A recent NetWAS 2.0 approach also includes a subsampling procedure to weight negative examples proportional to their GWAS *P* value[155]. Another method, Camoco (Coanalysis of molecular components), uses co-expression networks to identify GWAS hits that share similar expression patterns across experimental assays, finding that networks constructed using tissue-relevant co-expression data sets showed the best performance[161].

Network-based methods can also directly use mutational data from sequencing patient genomes to identify genes and pathways significantly associated with disease. For example, a network-based method was used to evaluate potential disease genes mutated in patients with hereditary plastic paraplegias and to predict additional candidate disease genes[162]. Another approach, network-neighbourhood differential enrichment analysis (NDEA), uses functional networks to boost power in enrichment analysis; in application to autism spectrum disorder, NDEA identified pathways significantly associated with autism based on the proband-specific mutational burden in brain-specific network neighbourhoods[26].

Another group of methods aims to identify cancer genes and pathways by overlaying mutational data onto molecular networks[163–167] (reviewed in REF.[168]). These methods, such as NetSig[169], are based on the observation that although many genes carry genetic alterations in a particular tumour profile, only a small number of these genes are cancer drivers[3]. Such driver genes are likely to function in shared biological processes and pathways across multiple tumours[170]. By examining the connectivity of mutated genes in gene networks, it becomes possible to identify subnetworks with greater mutational burden, thus increasing power to detect cancer drivers as well as elucidating their functional impact.

Molecular architecture of disease

More broadly, a large number of methods aim to identify candidate disease genes based on the molecular networks that summarize large omics data compendia[146,171–173] (reviewed in REF.[174]). These methods analyse the connectivity patterns of known disease genes (from literature, GWAS and so on) in a network and identify novel candidates with similar patterns. Importantly, the molecular network used for disease–gene prediction is crucial for accurate and relevant predictions, particularly because the dysregulated genes underlying diseases are frequently involved in tissue-specific and context-specific processes[175,176]. For example, in a recent study, a brain-specific network significantly outperformed the global, tissue-unaware network in predicting autism genes, and the top predicted genes from the brain network showed significant enrichment in autism probands relative to unaffected siblings[177]. Networks can further serve to systematically integrate information and data across organisms, facilitating the effective use of model organisms in the study of human disease[154,178–180]. In the diseaseQuest method, human quantitative genetics data are coupled with tissue-specific networks from *Caenorhabditis elegans* to predict candidate Parkinson disease genes and then to experimentally verify them in the worm[154].

Networks also provide a powerful approach to analysing and visualizing relationships among genes or gene sets, for example to identify groupings (that is, modules) that elucidate functional themes among disease-implicated genes (FIG. 4a,b). Various approaches have been developed for network-based module discovery[177,181,182] with publicly available implementations in tools such as Cytoscape[183] (cytoscape.org) and HumanBase (humanbase.io). For example, applying a Louvain community detection algorithm to genes associated with autism spectrum disorder[177] in a brain-specific functional network identifies modules enriched in

◀ Fig. 3 | **Informing quantitative genetics data with network models.** Molecular networks can be used to reprioritize the statistical associations from genome-wide association studies (GWAS) using functional information. **a** | Methods such as NetWAS (Network-Wide Association Study) first convert GWAS single-nucleotide polymorphism (SNP) *P* values to gene-level *P* values (using various methods[189]) and then train a machine learning classifier using nominally significant GWAS genes as positive examples (blue nodes), non-significant genes as negative examples (grey nodes) and features as weights from a gene network. The classifier outputs effectively re-rank all genes by their connectivity to the top GWAS genes, identifying candidates most likely to be associated with the studied disease/phenotype. **b** | An example of NetWAS in reprioritizing hypertension (HTN), systolic blood pressure (SBP) and diastolic blood pressure (DBP) GWAS genes[148] (Manhattan plots and the kidney-specific network shown are schematic). Gene functional relationships from a kidney-specific network were used as features in a support vector machine (SVM). For each GWAS, positive examples were GWAS hits with nominal gene $P < 0.01$ and non-significant genes as negative examples. The SVM effectively re-ranked all genes by their connectivity to the top GWAS genes, enriching for disease-relevant signals including, for example, hypertension-associated genes, processes and pathways.

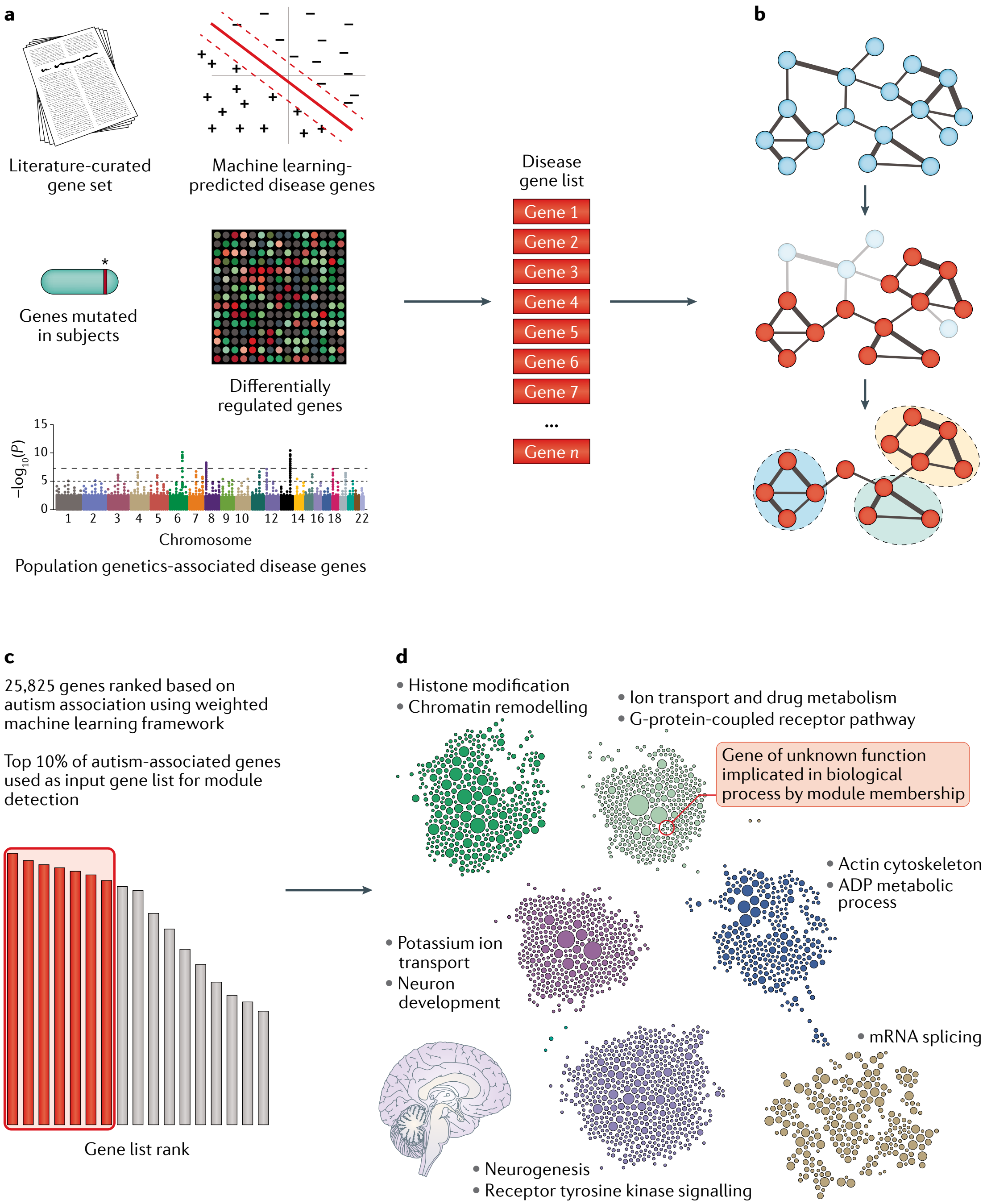

a
Literature-curated gene set
Machine learning-predicted disease genes
Genes mutated in subjects
Differentially regulated genes
15
10
5
0
−log10(P)
1 2 3 4 5 6 7 8 9 10 12 14 16 18 22
Chromosome
Population genetics-associated disease genes
Disease gene list
Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Gene 6
Gene 7
...
Gene n
b
c
25,825 genes ranked based on autism association using weighted machine learning framework
Top 10% of autism-associated genes used as input gene list for module detection
Gene list rank
d
• Histone modification
• Chromatin remodelling
• Ion transport and drug metabolism
• G-protein-coupled receptor pathway
Gene of unknown function implicated in biological process by module membership
• Actin cytoskeleton
• ADP metabolic process
• Potassium ion transport
• Neuron development
• mRNA splicing
• Neurogenesis
• Receptor tyrosine kinase signalling

Fig. 4 | **Systems-level functional interpretation of disease molecular architecture.** Functional module discovery identifies groups of genes that are tightly connected in a molecular network and likely to participate in similar biological processes. This can distil a large group of genes into biologically coherent sets, and can be used to elucidate disease-specific processes and infer the function of unknown genes. **a** | Gene lists for module clustering may be any group of genes of interest, and may be drawn from literature curation, sets of genes mutated in subjects with a specific disease, genes differentially regulated in a condition of interest, genes computationally implicated in a disease (for example, based on machine learning predictions) or genes associated with a disease in population genetics-based studies. **b** | The molecular network is first filtered by the input gene list (red nodes) and then partitioned into modules using a community clustering algorithm based on the connectivity of the genes in a functional network. The discovered modules can be biologically characterized based on Gene Ontology enrichment of member genes. **c** | In an application of module clustering to autism biology, 25,825 genes were ranked using a weighted machine learning framework based on their association with autism, and the top 10% of genes were used for module clustering. **d** | Functional modules of 2,500 predicted autism-associated genes were identified using a brain-specific functional network[177]. The resulting modules captured distinct aspects of autism biology. Each node represents a gene and the size of the node reflects connectivity of the gene (that is, the average edge weight of edges to the gene) in the network. Edges not shown for clarity. The function of unknown genes can be inferred based on their module context. An intuitive interface for researchers to perform functional module discovery is available at HumanBase (hb.flatironinstitute.org/module).

synaptic transmission and plasticity, sensory perception and chromatin remodelling (FIG. 4c,d). More recently, there have been community-led efforts to benchmark disease module discovery methods on various network configurations[184]. The prospects of these methods that predict disease association are strongly dependent on the quality and coverage of available molecular networks. Importantly, integrated network models can accommodate the growing breadth and scale of new omics data (for example, metabolomics, lipidomics and proteomics) and sequencing in more contexts (single cell types within tissues and across diseases).

Conclusions and perspectives

Increasingly rapid and inexpensive sequencing technologies are making it possible to initiate projects aiming to sequence ever-larger numbers of participants. Coupled with rich phenotypic information and improving computational approaches, this wealth of genetic information will lead to dramatic advances in the understanding of the genetic architecture of complex diseases and traits.

The goal of human genetics is linking genotype to phenotype, and this is especially challenging for variants in the non-coding genome. An emerging class of approaches use deep neural networks trained on large compendia of transcription factor binding and chromatin modification data to predict the biochemical impact of non-coding variants. Such predictions, which, in effect, function as in silico ChIP–seq assays, can be further linked to properties such as gene expression. Crucially, sequence-based deep learning methods can predict the impact of rare or never-seen variants that may be discovered with more sequencing, providing information that cannot be readily learned from phenotypic association studies of large genomic databases. Improvements in deep learning methods, including the development of more interpretable models as well as increasing the quality and coverage of training epigenetic data, are likely to result in further advances in the understanding of the non-coding genome. Broader availability of WGS studies, especially with careful controls, and improved representation of more diverse populations are also critical to advancement of this field.

Another key direction for understanding human disease processes involves developing network and pathway models in disease-relevant biological contexts, including specific cell types[155,157], developmental stages and environmental conditions. Identifying genes, modules and pathways most relevant to specific diseases, as well as elucidating how network perturbations and dysregulations contribute to disease, represents an important direction for study. The accuracy and disease relevance of such models can be improved by accounting for tissue differences in network wiring, which can include key differences for understanding disease processes in specific cell types and organ systems. Large-scale, cross-tissue data sets for training such models are increasingly available through consortium efforts[5,133,140,185–187], but additional data, especially at single-cell resolution, will be highly valuable for improving the resolution and accuracy of these models.

The availability of accurate, context-specific models of molecular architecture of disease will, in turn, be key to interpreting large numbers of variants affecting different genes and processes. An understanding of how to combine each variant's direct biochemical impact and its effect on downstream regulatory, interaction and functional networks will be critical to quantifying its contribution to phenotypic alterations and disease properties across diverse populations. Such an understanding can be facilitated by the ability to deeply profile individuals at the molecular level, including their genomic sequence, cell type-specific expression patterns and epigenetic profiles, and longitudinal phenotypic information. Advances in understanding variant impact, taken together with multimodal molecular and phenotypic data, will bring the field closer to broad implementation of precision medicine, in which genetically or molecularly targeted therapeutic approaches can be designed in a way that is responsive to the disease processes and drug sensitivities in specific patients.

Published online 2 August 2021

1. 1000 Genomes Project Consortium, et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
2. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
3. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
4. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
5. Bernstein, B. E. et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
6. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
This paper uses the PhastCons method to give per-base estimates of negative selection within conserved elements using multiple sequence alignments and hidden Markov models.
7. Davydov, E. V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).
This paper uses the pathogenicity GERP++ method to give nucleotide and element-level constraint scores from profiling substitution rates in multiple sequence alignments.
8. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* **47**, 276–283 (2015).
9. Ramani, R., Krumholz, K., Huang, Y.-F. & Siepel, A. PhastWeb: a web interface for evolutionary conservation scoring of multiple sequence alignments using phastCons and phyloP. *Bioinformatics* **35**, 2320–2322 (2019).
10. Cooper, G. M. et al. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
11. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
This paper discusses the pathogenicity scoring method PhyloP using multiple sequence alignments and gives the per-base *P* value for conservation/acceleration scores per clade that reflect divergence from the neutral rate.
12. Baugh, E. H. et al. Robust classification of protein variation using structural modelling and large-scale data integration. *Nucleic Acids Res.* **44**, 2501–2513 (2016).
13. Kobren, S. N., Chazelle, B. & Singh, M. PertInInt: an integrative, analytical approach to rapidly uncover cancer driver genes with perturbed interactions and functionalities. *Cell Syst.* **11**, 63–74.e7 (2020).
This paper uses PertInInt to assess protein variants for cancer relevance based on predicting the functional impact on physical interactions between proteins and other proteins, nucleic acids, ions, drugs and other small molecules.
14. Kobren, S. N. & Singh, M. Systematic domain-based aggregation of protein structures highlights DNA-, RNA- and other ligand-binding positions. *Nucleic Acids Res.* **47**, 582–593 (2019).
15. Ancien, F., Pucci, F., Godfroid, M. & Rooman, M. Prediction and interpretation of deleterious coding variants in terms of protein structural stability. *Sci. Rep.* **8**, 4480 (2018).
16. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
17. Gerasimavicius, L., Liu, X. & Marsh, J. A. Identification of pathogenic missense mutations using protein stability predictors. *Sci. Rep.* **10**, 15387 (2020).
18. Rodrigues, C. H., Pires, D. E. & Ascher, D. B. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res.* **46**, W350–W355 (2018).

19. Pires, D. E. V., Ascher, D. B. & Blundell, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.* **42**, W314–W319 (2014).
20. Rodrigues, C. H. M., Myung, Y., Pires, D. E. V. & Ascher, D. B. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic Acids Res.* **47**, W338–W344 (2019).
21. Li, M., Simonetti, F. L., Goncearenco, A. & Panchenko, A. R. MutaBind estimates and interprets the effects of sequence variants on protein–protein interactions. *Nucleic Acids Res.* **44**, W494–W501 (2016).
22. Dehouck, Y., Kwasigroch, J. M., Rooman, M. & Gilis, D. BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations. *Nucleic Acids Res.* **41**, W333–W339 (2013).
23. Pires, D. E. V., Blundell, T. L. & Ascher, D. B. mCSM-lig: quantifying the effects of mutations on protein–small molecule affinity in genetic disease and emergence of drug resistance. *Sci. Rep.* **6**, 29575 (2016).
24. Ghersi, D. & Singh, M. Interaction-based discovery of functionally important genes in cancers. *Nucleic Acids Res.* **42**, e18 (2014).
25. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015). **This paper presents DeepSEA, a multitask deep learning model that trains and predicts cell type-specific regulatory factor binding to genomic sequence for >900 features and cell types.**
26. Zhou, J. et al. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat. Genet.* **51**, 973–980 (2019). **This paper, for the first time, finds a significant contribution of non-coding mutations to complex disease risk by demonstrating higher functional impact of de novo mutations from probands with autism compared with siblings, using mutational impacts inferred from deep learning sequence models of transcriptional and post-transcriptional effects.**
27. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).
28. Lee, D. et al. A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015). **This paper presents, among the first approaches to predict the tissue-specific impact of changes in the non-coding genome without information on evolution or genome annotations, gkm-svm implementing an SVM classifier that uses only sequence *k*-mers as input.**
29. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
30. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
31. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
32. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* **6**, 26094 (2016).
33. Zhavoronkov, A. et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
34. Ching, T. et al. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**, 20170387 (2018).
35. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
36. Avsec, Ž. et al. The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.* **37**, 592–600 (2019).
37. Chen, K. M., Cofer, E. M., Zhou, J. & Troyanskaya, O. G. Selene: a PyTorch-based deep learning library for sequence data. *Nat. Methods* **16**, 315–318 (2019).
38. Shrikumar, A., Greenside, P. & Kundaje, A. in *ICML'17 Proc. 34th Int. Conf. Machine Learning* (eds Precup, D. & Teh, Y. W.) 3145–3153 (PMLR, 2017).
39. Binder, A. et al. Morphological and molecular breast cancer profiling through explainable machine learning. *Nat. Mach. Intell.* **3**, 355–366 (2021).
40. Zhang, Z., Park, C. Y., Theesfeld, C. L. & Troyanskaya, O. G. An automated framework for efficiently designing deep convolutional neural networks in genomics. *Nat. Mach. Intell.* **3**, 392–400 (2021).
41. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* **20**, 389–403 (2019).
42. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
43. Wainschtein, P. et al. Recovery of trait heritability from whole genome sequence data. Preprint at *bioRxiv* https://doi.org/10.1101/588020 (2019).
44. Wood, A. R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
45. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
46. Yang, J. et al. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525 (2011).
47. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
48. Liu, X., Li, Y. I. & Pritchard, J. K. Trans effects on gene expression can drive omnigenic inheritance. *Cell* **177**, 1022–1034.e6 (2019).
49. Lappalainen, T., Scott, A. J., Brandt, M. & Hall, I. M. Genomic analysis in the age of human genome sequencing. *Cell* **177**, 70–84 (2019).
50. Shendure, J., Findlay, G. M. & Snyder, M. W. Genomic medicine—progress, pitfalls, and promise. *Cell* **177**, 45–57 (2019).
51. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
52. Pasaniuc, B. et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**, 2906–2914 (2014).
53. Schurz, H. et al. Evaluating the accuracy of imputation methods in a five-way admixed population. *Front. Genet.* **10**, 34 (2019).
54. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
55. Easton, D. F. et al. Gene-panel sequencing and the prediction of breast-cancer risk. *N. Engl. J. Med.* **372**, 2243–2257 (2015).
56. Robbins, C. M. et al. Copy number and targeted mutational analysis reveals novel somatic events in metastatic prostate tumors. *Genome Res.* **21**, 47–55 (2011).
57. Meienberg, J., Bruggmann, R., Oexle, K. & Matyas, G. Clinical sequencing: is WGS the better WES? *Hum. Genet.* **135**, 359–362 (2016).
58. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
59. French, C. E. et al. Whole genome sequencing reveals that genetic conditions are frequent in intensively ill children. *Intensive Care Med.* **45**, 627–636 (2019).
60. Hou, Y.-C. C. et al. Precision medicine integrating whole-genome sequencing, comprehensive metabolomics, and advanced imaging. *Proc. Natl Acad. Sci. USA* **117**, 3053–3062 (2020).
61. Cassini, T. A. et al. Whole genome sequencing reveals novel IGHMBP2 variant leading to unique cryptic splice-site and Charcot–Marie–Tooth phenotype with early onset symptoms. *Mol. Genet. Genom. Med.* **7**, e00676 (2019).
62. All of Us Research Program Investigators, et al. The 'All of Us' research program. *N. Engl. J. Med.* **381**, 668–676 (2019).
63. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
64. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013).
65. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
66. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat. Genet.* **51**, 88–95 (2019).
67. Eilbeck, K., Quinlan, A. & Yandell, M. Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.* **18**, 599–612 (2017).
68. Watanabe, K. et al. A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
69. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
70. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14**, S3 (2013).
71. Shihab, H. A. et al. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum. Genomics* **8**, 11 (2014).
72. Ioannidis, N. M. et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
73. Sim, N.-L. et al. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452–W457 (2012).
74. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
75. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).
76. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019). **This paper uses the CADD pathogenicity score as an SVM classifier that integrates multiple functional genomic and evolutionary data to predict coding and non-coding variant impacts.**
77. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016). **This paper presents the Eigen pathogenicity score as an unsupervised meta-score of non-coding variant fitness impact.**
78. Park, J. S. et al. Brain somatic mutations observed in Alzheimer's disease associated with aging and dysregulation of tau phosphorylation. *Nat. Commun.* **10**, 1–12 (2019).
79. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e18 (2018).
80. Buja, A. et al. Damaging de novo mutations diminish motor skills in children on the autism spectrum. *Proc. Natl Acad. Sci. USA* **115**, E1859–E1866 (2018).
81. Wright, C. F. et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**, 1305–1314 (2015).
82. Quang, D. & Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* **44**, e107–e107 (2016).
83. Nair, S., Kim, D. S., Perricone, J. & Kundaje, A. Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. *Bioinformatics* **35**, i108–i116 (2019).
84. Avsec, Ž. et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (2021). **This paper describes a novel deep learning framework for functional genomics sequence modelling, which combines neural network models with model interpretation tools to discover high-resolution motif syntax.**
85. Kelley, D. R. et al. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* **28**, 739–750 (2018). **This paper uses Basenji as a CNN sequence model that predicts regulatory factor binding and expression based on cap analysis gene expression (CAGE) peak data.**
86. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
87. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped *k*-mer features. *PLoS Comput. Biol.* **10**, e1003711 (2014).
88. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and

RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
This paper uses DeepBind as a framework of individual CNNs that train on and predict regulatory factor binding to DNA and RNA.

89. Angermueller, C., Lee, H. J., Reik, W. & Stegle, O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* **18**, 67 (2017).
90. Quang, D. & Xie, X. FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods* **166**, 40–47 (2019).
91. Richter, F. et al. Genomic analyses implicate noncoding de novo variants in congenital heart disease. *Nat. Genet.* **52**, 769–777 (2020).
This paper presentes whole-genome sequence analysis of genetic aetiology of congenital heart disease wherein HeartENN, a deep CNN sequence genomic sequence model, is applied to functional impact prediction of de novo non-coding mutations and an excess burden of high-impact mutations is observed in individuals who are affected compared with controls.
92. Qin, Q. & Feng, J. Imputation for transcription factor binding predictions based on deep learning. *PLoS Comput. Biol.* **13**, e1005403 (2017).
93. Li, H., Quang, D. & Guan, Y. Anchor: *trans*-cell type prediction of transcription factor binding sites. *Genome Res.* **29**, 281–292 (2019).
94. Agarwal, V. & Shendure, J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Rep.* **31**, 107663 (2020).
95. Zhou, J. et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).
This paper presents tissue-specific gene expression prediction from sequence using a deep CNN and a linear model, and application to derive constraint violation score pathogenicity, based on cumulative predicted regulatory impacts in genomic intervals.
96. Dey, K. K. et al. Integrative approaches to improve the informativeness of deep learning models for human complex diseases. Preprint at *bioRxiv* https://doi.org/10.1101/2020.09.08.288563 (2020).
97. Law, A. J., Kleinman, J. E., Weinberger, D. R. & Weickert, C. S. Disease-associated intronic variants in the ErbB4 gene are related to altered ErbB4 splice-variant expression in the brain in schizophrenia. *Hum. Mol. Genet.* **16**, 129–141 (2006).
98. Sangermano, R. et al. ABCA4 midigenes reveal the full splice spectrum of all reported noncanonical splice site variants in Stargardt disease. *Genome Res.* **28**, 100–110 (2018).
99. de Jong, V. M. et al. Post-transcriptional control of candidate risk genes for type 1 diabetes by rare genetic variants. *Genes. Immun.* **14**, 58–61 (2012).
100. Cardo, L. F. et al. A Search for SNCA 3′ UTR variants Identified SNP rs356165 as a determinant of disease risk and onset age in Parkinson's disease. *J. Mol. Neurosci.* **47**, 425–430 (2011).
101. Zuallaert, J. et al. SpliceRover: interpretable convolutional neural networks for improved splice site prediction. *Bioinformatics* **34**, 4180–4188 (2018).
102. Louadi, Z., Oubounyt, M., Tayara, H. & Chong, K. T. Deep splicing code: classifying alternative splicing events using deep learning. *Genes* **10**, 587 (2019).
103. Leung, M. K. K., Xiong, H. Y., Lee, L. J. & Frey, B. J. Deep learning of the tissue-regulated splicing code. *Bioinformatics* **30**, i121–i129 (2014).
104. Zhang, Y., Liu, X., MacLeod, J. & Liu, J. Discerning novel splice junctions derived from RNA-seq alignment: a deep learning approach. *BMC Genomics* **19**, 971 (2018).
105. Zeng, Z. & Bromberg, Y. Predicting functional effects of synonymous variants: a systematic review and perspectives. *Front. Genet.* **10**, 914 (2019).
106. Barash, Y. et al. Deciphering the splicing code. *Nature* **465**, 53–59 (2010).
107. Paggi, J. M. & Bejerano, G. A sequence-based, deep learning model accurately predicts RNA splicing branchpoints. *RNA* **24**, 1647–1658 (2018).
108. Li, Y. I. et al. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
109. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019).
110. Cummings, B. B. et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**, eaal5209 (2017).
111. Ray, T. A. et al. Comprehensive identification of mRNA isoforms reveals the diversity of neural cell-surface molecules with roles in retinal development and disease. *Nat. Commun.* **11**, 3328 (2020).
112. Lagarde, J. et al. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat. Genet.* **49**, 1731–1740 (2017).
113. Gupta, I. et al. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.* **36**, 1197–1202 (2018).
114. Hardwick, S. A., Joglekar, A., Flicek, P., Frankish, A. & Tilgner, H. U. Getting the entire message: progress in isoform sequencing. *Front. Genet.* **10**, 709 (2019).
115. Pan, X. & Shen, H.-B. RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC Bioinforma.* **18**, 1–14 (2017).
116. Pan, X., Rijnbeek, P., Yan, J. & Shen, H.-B. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genomics* **19**, 1–11 (2018).
117. Pan, X. & Shen, H.-B. Predicting RNA–protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics* **34**, 3427–3436 (2018).
118. Yu, H., Wang, J., Sheng, Q., Liu, Q. & Shyr, Y. beRBP: binding estimation for human RNA-binding proteins. *Nucleic Acids Res.* **47**, e26–e26 (2018).
119. Zhang, Z. et al. Deep-learning augmented RNA-seq analysis of transcript splicing. *Nat. Methods* **16**, 307–310 (2019).
120. Wen, M., Cong, P., Zhang, Z., Lu, H. & Li, T. DeepMirTar: a deep-learning approach for predicting human miRNA targets. *Bioinformatics* **34**, 3781–3787 (2018).
121. Kang, Q., Meng, J., Cui, J., Luan, Y. & Chen, M. PmliPred: a method based on hybrid model and fuzzy decision for plant miRNA–lncRNA interaction prediction. *Bioinformatics* **36**, 2986–2992 (2020).
122. Park, C. Y. et al. Genome-wide landscape of RNA-binding protein target site dysregulation reveals a major impact on psychiatric disorder risk. *Nat. Genet.* **53**, 166–173 (2021).
123. Landrum, M. J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).
124. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
125. Rehm, H. L. et al. ClinGen—the clinical genome resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
126. Harrison, S. M. et al. Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar. *Genet. Med.* **19**, 1096–1104 (2017).
127. Stenson, P. D. et al. The human gene mutation database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Hum. Genet.* **139**, 1197–1207 (2020).
128. Kalia, S. S. et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**, 249–255 (2017).
129. Esposito, D. et al. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol.* **20**, 223 (2019).
130. Oughtred, R. et al. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic and chemical interactions. *Protein Sci.* **30**, 187–200 (2021).
131. Gelman, H. et al. Recommendations for the collection and use of multiplexed functional data for clinical variant interpretation. *Genome Med.* **11**, 85 (2019).
132. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763 (2015).
This paper presents a pathogenicity scoring method, which is a deep learning (CNN) version of CADD, for coding and non-coding variant fitness impact.
133. Davis, C. A. et al. The Encyclopedia of DNA Elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
134. Gronau, I., Arbiza, L., Mohammed, J. & Siepel, A. Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. *Mol. Biol. Evol.* **30**, 1159–1171 (2013).
135. Huang, Y.-F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).
This paper uses a pathogenicity scoring method, LINSIGHT, to predict fitness consequences of non-coding human variation using linear modelling of functional genomic data with a probabilistic model of molecular evolution.
136. Ernst, C. et al. Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics. *BMC Med. Genomics* **11**, 35 (2018).
137. Hart, S. N. et al. Comprehensive annotation of BRCA1 and BRCA2 missense variants by functionally validated sequence-based computational prediction models. *Genet. Med.* **21**, 71–80 (2019).
138. Findlay, G. M. et al. Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
139. Kim, S. S. et al. Improving the informativeness of Mendelian disease-derived pathogenicity scores for common disease. *Nat. Commun.* **11**, 6258 (2020).
140. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
141. Consortium, G. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
142. Deutsch, E. W. et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.* **45**, D1100–D1106 (2017).
143. Schwenk, J. M. et al. The human plasma proteome draft of 2017: building on the human plasma peptideatlas from mass spectrometry and complementary assays. *J. Proteome Res.* **16**, 4299–4310 (2017).
144. Huttenhower, C. et al. Exploring the human genome with functional maps. *Genome Res.* **19**, 1093–1106 (2009).
145. Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B. & Botstein, D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl Acad. Sci. USA* **100**, 8348–8353 (2003).
146. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.* **9**, S4 (2008).
147. Snel, B., Lehmann, G., Bork, P. & Huynen, M. A. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* **28**, 3442–3444 (2000).
148. Greene, C. S. et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47**, 569–576 (2015).
149. Wong, A. K., Krishnan, A. & Troyanskaya, O. G. GIANT 2.0: genome-scale integrated analysis of gene networks in tissues. *Nucleic Acids Res.* **46**, W65–W70 (2018).
150. Pierson, E. et al. Sharing and specificity of co-expression networks across 35 human tissues. *PLoS Comput. Biol.* **11**, e1004220 (2015).
151. Keller, M. P. et al. A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. *Genome Res.* **18**, 706–716 (2008).
152. Dobrin, R. et al. Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome Biol.* **10**, R55 (2009).
153. Magger, O., Waldman, Y. Y., Ruppin, E. & Sharan, R. Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS Comput. Biol.* **8**, e1002690 (2012).
154. Yao, V. et al. An integrative tissue-network approach to identify and test human disease genes. *Nat. Biotechnol.* **36**, 1091–1099 (2018).
155. Roussarie, J.-P. et al. Selective neuronal vulnerability in Alzheimer's disease: a network-based analysis. *Neuron* **107**, 821–835.e12 (2020).
156. Goya, J. et al. FNTM: a server for predicting functional networks of tissues in mouse. *Nucleic Acids Res*. **43**, W182–W187 (2015).
157. Ledo, J. H. et al. Lack of a site-specific phosphorylation of Presenilin 1 disrupts microglial gene networks and progenitors during development. *PLoS ONE* **15**, e0237773 (2020).
158. Huang, J. K. et al. Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst.* **6**, 484–495.e5 (2018).
159. Kamburov, A., Wierling, C., Lehrach, H. & Herwig, R. ConsensusPathDB—a database for integrating human functional interaction networks. *Nucleic Acids Res.* **37**, D623–D628 (2009).

160. Califano, A., Butte, A. J., Friend, S., Ideker, T. & Schadt, E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat. Genet.* **44**, 841–847 (2012).
161. Schaefer, R. J. et al. Integrating coexpression networks with GWAS to prioritize causal genes in maize. *Plant. Cell* **30**, 2922–2942 (2018).
162. Novarino, G. et al. Exome sequencing links corticospinal motor neuron disease to common neurodegenerative disorders. *Science* **343**, 506–511 (2014).
163. Leiserson, M. D. M. et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015).
164. Ruffalo, M., Koyutürk, M. & Sharan, R. Network-based integration of disparate omic data to identify 'silent players' in cancer. *PLoS Comput. Biol.* **11**, e1004595 (2015).
165. Vandin, F., Upfal, E. & Raphael, B. J. Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* **18**, 507–522 (2011).
166. Reyna, M. A., Leiserson, M. D. M. & Raphael, B. J. Hierarchical HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics* **34**, i972–i980 (2018).
167. Creixell, P. et al. Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell* **163**, 202–217 (2015).
168. Creixell, P. et al. Pathway and network analysis of cancer genomes. *Nat. Methods* **12**, 615–621 (2015).
169. Horn, H. et al. NetSig: network-based discovery from cancer genomes. *Nat. Methods* **15**, 61–66 (2018).
170. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
171. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. & Sharan, R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.* **6**, e1000641 (2010).
172. Aerts, S. et al. Gene prioritization through genomic data fusion. *Nat. Biotechnol.* **24**, 537–544 (2006).
173. Köhler, S., Bauer, S., Horn, D. & Robinson, P. N. Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* **82**, 949–958 (2008).
174. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).
175. Lage, K. et al. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl Acad. Sci. USA* **105**, 20870–20875 (2008).
176. Winter, E. E., Goodstadt, L. & Ponting, C. P. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res.* **14**, 54–61 (2004).
177. Krishnan, A. et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat. Neurosci.* **19**, 1454–1462 (2016).
178. Chikina, M. D. & Troyanskaya, O. G. Accurate quantification of functional analogy among close homologs. *PLoS Comput. Biol.* **7**, e1001074 (2011).
179. Guan, Y., Ackert-Bicknell, C. L., Kell, B., Troyanskaya, O. G. & Hibbs, M. A. Functional genomics complements quantitative genetics in identifying disease-gene associations. *PLoS Comput. Biol.* **6**, e1000991 (2010).
180. Swarup, V. et al. Identification of evolutionarily conserved gene networks mediating neurodegenerative dementia. *Nat. Med.* **25**, 152–164 (2019).
181. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, 7 (2005).
182. Parikshak, N. N., Gandal, M. J. & Geschwind, D. H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat. Rev. Genet.* **16**, 441–458 (2015).
183. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
184. Choobdar, S. et al. Assessment of network module identification across complex diseases. *Nat. Methods* **16**, 843–852 (2019).
185. Lizio, M. et al. Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. *Nucleic Acids Res.* **47**, D752–D758 (2019).
186. Regev, A. et al. The human cell atlas. *eLife* **6**, e27041 (2017).
187. Lindsay, S. J. et al. HDBR expression: a unique resource for global and individual gene expression studies during early human brain development. *Front. Neuroanat.* **10**, 86 (2016).
188. Zhang, Y. et al. Discerning novel splice junctions derived from RNA-seq alignment: a deep learning approach. *BMC Genomics* **19**, 971 (2018).
189. Mishra, A. & Macgregor, S. VEGAS2: software for more flexible gene-based testing. *Twin Res. Hum. Genet.* **18**, 86–91 (2015).

Acknowledgements

This work is supported by National Institutes of Health (NIH)/National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) grants U24DK100845, UGDK114907 and U2CDK114886 and NIH grant UH3TR002158 to O.G.T. The authors thank C. Park for helpful discussions.

Author contributions

The authors contributed to all aspects of the article.

Competing interests

The authors declare no competing interests.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.