Cell Genomics

CellPress
OPEN ACCESS

Perspective

From function to translation: Decoding genetic susceptibility to human diseases via artificial intelligence

Erping Long,[1,2,5,*] Peixing Wan,[3,5] Qingyu Chen,[4] Zhiyong Lu,[4] and Jiyeon Choi[2,*]
[1]Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
[2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[3]Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[4]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
[5]These authors contributed equally
*Correspondence: erping.long@ibms.pumc.edu.cn (E.L.), jiyeon.choi2@nih.gov (J.C.)
https://doi.org/10.1016/j.xgen.2023.100320

SUMMARY

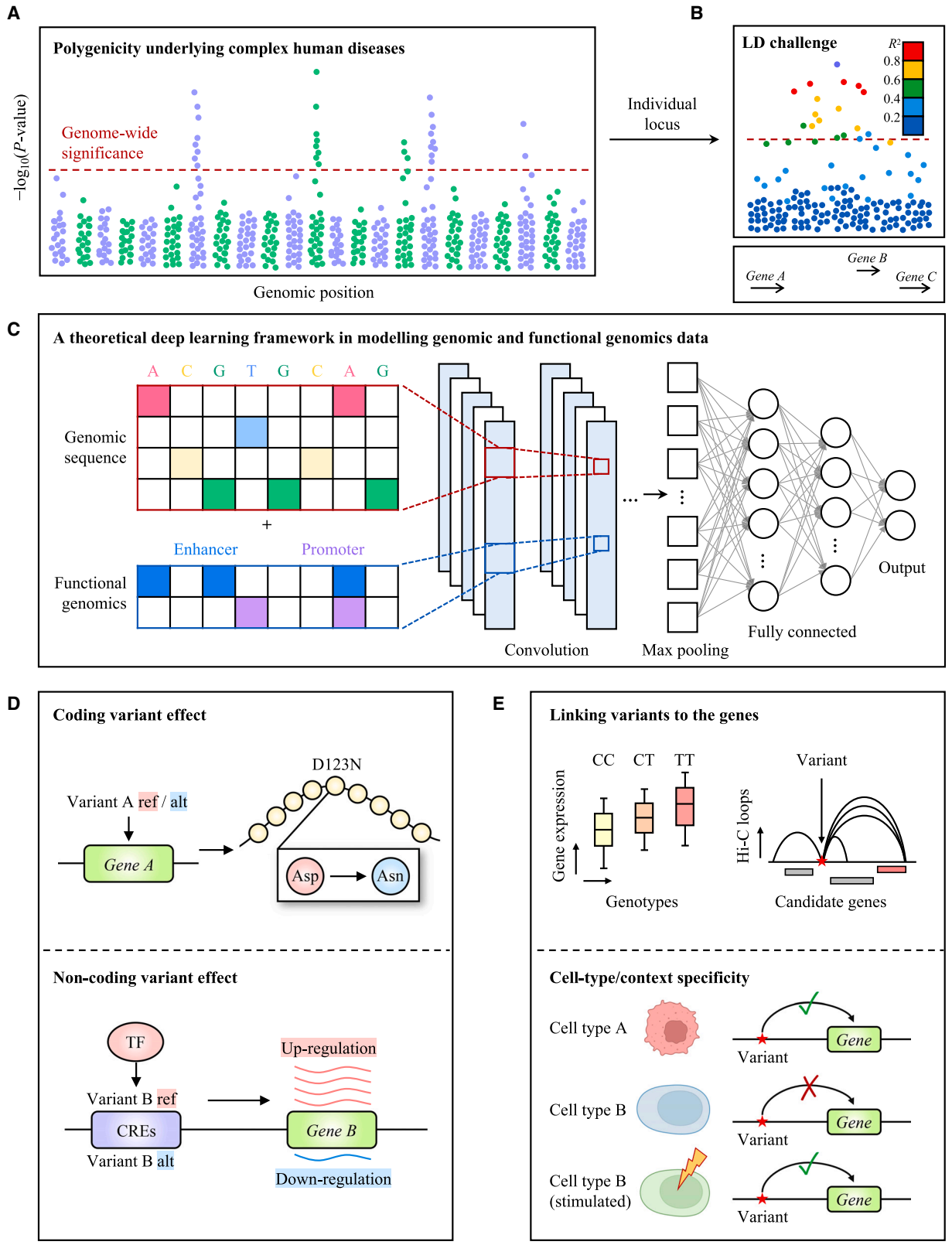
While genome-wide association studies (GWAS) have discovered thousands of disease-associated loci, molecular mechanisms for a considerable fraction of the loci remain to be explored. The logical next steps for post-GWAS are interpreting these genetic associations to understand disease etiology (GWAS functional studies) and translating this knowledge into clinical benefits for the patients (GWAS translational studies). Although various datasets and approaches using functional genomics have been developed to facilitate these studies, significant challenges remain due to data heterogeneity, multiplicity, and high dimensionality. To address these challenges, artificial intelligence (AI) technology has demonstrated considerable promise in decoding complex functional datasets and providing novel biological insights into GWAS findings. This perspective first describes the landmark progress driven by AI in interpreting and translating GWAS findings and then outlines specific challenges followed by actionable recommendations related to data availability, model optimization, and interpretation, as well as ethical concerns.

INTRODUCTION

Decoding disease susceptibility is a central question in human genetics and precision medicine. Genome-wide association studies (GWAS) have discovered thousands of loci across the genome that are significantly associated with human diseases risk, indicating the genetic component and polygenicity underlying most of the common diseases[1,2] (Figure 1A). However, it is challenging to determine the functional mechanisms for GWAS-identified loci for several reasons. First, the association of a locus does not specify the "causal variants" because most variants with statistical significance are not biologically causal but, rather, are correlated with the causal one(s) due to linkage disequilibrium (LD)[3] (Figure 1B). Second, it is unclear which "target genes" are affected since most of these risk variants are non-protein coding and could regulate the transcription, mRNA splicing, mRNA stability, and translation of nearby or distant genes often in a tissue-/cell-type-specific manner[4] (Figure 1B). Because of these challenges in deciphering the functional mechanisms, it is even more difficult to translate GWAS findings into clinical benefits for the patients.

Recently, increasing resources and datasets have been available to facilitate GWAS functional and translational studies based on approaches such as expression quantitative trait locus (eQTL)[4] profiling, chromatin annotations (e.g., as cataloged in ENCODE[5]), and genome perturbation (e.g., CRISPR-based screening[6]). These datasets provided fruitful features to infer the "causal variants" and target genes underlying the GWAS loci but are accompanied by substantial multiplicity and heterogeneity. For example, multiple datasets using different approaches (e.g., molecular QTL and chromatin interaction studies) are valuable in linking variants to genes. However, they often provide complementary or discordant results, making it difficult to accurately weigh these complexities to achieve the consensus. Moreover, even functional datasets using the same experimental approach can be generated under specific contexts (tissue/cell types) with different statistical models or parameters, making the integration of such heterogeneous information challenging.

Artificial intelligence (AI) technologies can be used to address these challenges of dealing with diverse and heterogeneous data. AI is an umbrella term for any computer program that has the touch of human intelligence, encompassing machine learning and deep learning.[7] Here, we used supervised deep learning, a certain type of AI algorithm,[8] as an example to demonstrate the strengths of AI. Specifically, supervised deep learning (a classic framework showcased in Figure 1C) enables (1) feature extraction and pattern recognition based on the data itself, which negates the need for labeling complex features; (2) iteration and gradient strategy during training, which allows automatic optimization of



(legend on next page)

parametric weights for accurate extraction and summarization of complex features; and (3) multi-layer architecture consisting of millions of neurons, which enables modeling of massive datasets with data heterogeneity. For readers who are not familiar with supervised deep learning, we have provided brief descriptions of the aforementioned terms in Box 1. Please note that, in contrast to supervised deep learning, there are other types of AI algorithms with varied architectures and learning processes (e.g., semi-supervised or unsupervised).[9] While deep-learning approaches have been applied to genetics and genomics in general (reviewed in Eraslan et al.,[9] Novakovsky et al.,[10] Zou et al.,[11] and Angermueller et al.[12]), their promise and utility in GWAS follow-up studies have not yet been thoroughly discussed. In the first part of this perspective, we illustrate a few GWAS functional studies that have benefited from AI approaches, including the prediction of variant effects[13] and target gene assignment.[14] We then introduce ongoing AI efforts toward translating GWAS findings into clinical practice, including genetics-supported drug repurposing and polygenic risk score (PRS) prediction. Finally, we discuss the prospects and challenges of applying AI approaches to GWAS functional and translational studies, including issues such as the scarcity of ground-truth data, challenges in interpretation due to the "black-box" nature of AI, and ethical concerns. We also provide several actionable recommendations for addressing these challenges.

GWAS FUNCTIONAL STUDIES DRIVEN BY AI TECHNOLOGY

Decoding the effects of non-coding variants

Although disease-associated protein-coding variants often have comprehensible functional consequences in protein products (reviewed in Shameer et al.[15]), most GWAS risk variants (~90%) are non-protein coding and have unknown functional consequences. Previous studies indicated that non-coding variants can mainly function by *cis*-regulation of gene expression levels (Figure 1D). Furthermore, the regulatory activities of non-coding variants could be highly tissue-/cell-type-specific and could involve diverse and complex mechanisms, such as binding of transcriptional factors (TFs), RNA-binding proteins, or microRNAs (miRNAs) to affect transcription initiation, alternative splicing, and mRNA stability. Given such complexity of potential mechanisms and biological contexts, it is challenging to identify the functional variants among numerous candidate non-coding variants that are linked by LD.

AI technologies were successfully applied to decoding non-coding variant effects for different aspects of regulatory activities. One example is predicting the binding of regulatory proteins (e.g., TFs) to their target sequences, which is the basis of the allelic effects in mediating gene transcription. The TF-binding event is not easily recognizable by conventional methods because it involves not only sequence information but also the competition or synergistic effects between proteins. One of the earlier AI efforts include a deep-learning approach, DeepBind, which predicts the sequence specificities of DNA- and RNA-binding proteins.[16] DeepBind includes a convolution module that extracts the features from local sequences and a prediction module that synthesizes local features into higher-level signals. The performance of DeepBind was evaluated in nearly 1,000 publicly available datasets, including DNA-binding (chromatin immunoprecipitation) and protein-binding microarrays. DeepBind performed better than the previous methods developed with extensive biological knowledge, which could discover regulatory motifs and interpret the effects of genetic variants.[16] A more recent AI-based development enabled the mapping of DNA sequences to interpretable regulatory classes across the whole genome. This deep-learning model, named Sei,[17] was trained to predict >20,000 features including TF binding and chromatin modification/accessibility peaks from >1,300 cell-line/tissue datasets for each of the 100 bp DNA sequences tiled across the genome. By clustering and defining the genome-wide predictions into 40 regulatory classes (e.g., promoter, cell-type-specific enhancers), Sei could predict the effects of any sequence or variant of interest, including those not previously investigated in GWAS. For example, predicted promoter and enhancer classes from Sei were strongly enriched for GWAS heritability beyond that explained by the baseline functional annotations alone, suggesting extra regulatory elements defined by Sei underlying GWAS signals.[17] These examples highlight the ability of AI-based methods to use complex and multi-dimensional datasets to predict the effects of non-coding GWAS variants in a more interpretable and scalable manner.

Figure 1. AI accelerates GWAS functional studies

(A) A schematic diagram of GWAS Manhattan plot is shown to illustrate the polygenicity underlying complex human diseases. Each dot represents a variant. The x axis refers to genomic location. The y axis refers to -$\log_{10}$(GWAS p value). The red horizontal line indicates the genome-wide significance cutoff.

(B) A schematic diagram presents the challenges in nominating causal variants and target genes due to linkage disequilibrium (LD) and the prevalence of non-protein-coding variants. Gene diagrams are shown with arrows (directionality of transcription). Each dot represents a variant. LD (R^2) is color coded.

(C) A schematic diagram showcases a classical deep-learning framework in modeling genomic and functional genomics data. The left part illustrates a process of formatting the base-resolution information of genomic sequences (four nucleotides A/T/C/G, shown in four colors) and functional annotation (such as enhancer or promoter region, colored when enhancer and promoter overlap) into input matrices. The right parts display a conventional deep-learning architecture, including convolutional layer (feature extraction process by applying filters to input data to generate feature maps), max pooling layer (feature summarization process from clusters of neurons to reduce the data dimensions), fully connected layer (integration process by connecting every neuron in one layer to every neuron in another layer), and classification output. These layers are responsible for feature extraction/processing, pattern recognition, summarizing, and outputting.

(D) The two representative mechanisms of GWAS variants (coding/non-coding) effects. The top part showcases a variant located in a coding region, which results in corresponding amino acid changes. The bottom part showcases a variant located in a *cis*-regulatory element (CRE) in a non-coding region, which has allelic effects on binding a transcriptional factor (TF) and regulates the expression level of a target gene.

(E) Two main challenges in characterizing the functional consequences of GWAS variants, including variant-gene linkage and cell-type/context specificity. The top part presents two mainstream approaches linking variants to the genes, expression quantitative trait locus (eQTL; top left) and chromatin interaction (top right). The bottom part depicts a context-dependent gene regulation, which is active in cell type A but conditionally active in cell type B (only with a stimulus).

Box 1. Brief description of terms mentioned in the introduction

Deep learning: deep learning consists of a range of AI algorithms, which commonly involves a neural network with multiple layers and numerous neurons in each layer. These neural networks are biologically inspired by human brains—albeit far from matching their ability—allowing it to learn from massive amounts of data.

Label: a label is used to explain a piece of data information, which could provide ground-truth guidance (e.g., disease/normal) that AI algorithms can learn from (called supervised learning). Notably, unsupervised/self-supervised learning do not need ground-truth labels.

Training: the AI training process is to feed data into the algorithm with labels (e.g., disease/normal). Over iterations, the algorithm extracts the features and recognizes patterns from the data, achieving optimized parametric weights to distinguish the disease from normal (called trained model).

Layer: a layer in deep-learning algorithms is a structure or network topology, which takes information from the previous layer and then passes it to the next layer. There are several commonly used architectures (e.g., convolutional neural network) and associated layers (e.g., convolution, pooling, and fully connected) in deep learning, underlying different mathematics and functions.

Neuron: neurons in deep-learning algorithms are nodes through which data and computations flow. Neurons usually receive one or more input signals. These input signals can come from either the raw dataset or from neurons positioned at a previous layer of the neural network.

Linking variants to genes

Given that "causal variants" in GWAS loci are predominantly regulatory, their target genes are not necessarily the closest ones to "causal variants." For example, a recent functional follow-up study of kidney disease GWAS showed that 66% of candidate variants were not assigned to the closest gene as their target.[18] Moreover, it has been reported that GWAS variants could impact long-range chromatin interaction and thus regulate genes far away.[19] Molecular QTL or chromatin interaction approaches have been used to link the variants to the target genes they regulate (Figure 1E), but they often provide complementary or discordant results, and therefore accurately weighting these complexities to achieve the consensus is challenging.[20] To address this issue, researchers from Open Targets developed a systematic framework of integrating GWAS and functional data and developing a machine-learning model to identify target genes across GWAS loci. Specifically, they established a unified pipeline in performing fine mapping and colocalization analysis and integrated them with GWAS and other functional data, which resulted in four main feature categories at the locus level: *in silico* pathogenicity prediction, colocalization of molecular QTLs, chromatin interaction, and gene distance to credible set variants weighted by fine-mapping probabilities.[14] For each GWAS locus, locus-to-gene scores were derived from these features both by aggregating variant-to-gene scores from multiple variants (e.g., *in silico* prediction, chromatin interaction) and summarizing the gene-level evidence (e.g., conditional and tissue-dependent colocalization). A machine-learning-based model (XGBoost gradient-boosting classifier) was used to incorporate all the locus-level features and was trained on a manually curated set of 445 gold-standard-positive genes, for which the target gene assignment is deemed credible. The model output is the target gene ranking based on the predicted locus-to-gene score. The full machine-learning model (area under the curve [AUC] = 0.93) outperformed the classical model only considering the variant-to-gene distance (AUC = 0.76–0.79) across over 100,000 published GWAS loci associated with various human traits.[14] Although this is a remarkable advancement, future efforts are warranted to improve the target gene assignment given the limited sample size and potentially biased sources of gold-standard labels for training (see discussion in challenges and recommendations).

Decoding cell-type specificity and predicting downstream pathways

Given that the genetic regulation is often cell-type specific (Figure 1E), deciphering the relevant cell types/states is important for interpreting GWAS findings at both the locus level and the downstream pathway level.[21] Single-cell-based sequencing technologies provide an opportunity to address this challenge, which enables high-throughput profiling of transcriptome and other modalities (e.g., chromatin accessibility) at a single-cell resolution. Typical outputs from single-cell sequencing are large scale and contain high dimensionality information of thousands of cells, which is challenging to handle using conventional algorithms. AI-driven efforts have been made to process the single-cell datasets and to decode the cell-type specificity for GWAS findings. For example, a deep-learning model was trained by a single-cell assay for transposase-accessible chromatin sequencing (ATAC-seq) profile of over 50,000 cells representing 13 human retinal cell types.[22] The trained model could predict the per-base differences in chromatin accessibility between reference and alternate alleles specific to each cell type. The authors identified 23 GWAS risk variants of eye diseases showing a "high effect" on allelic chromatin accessibility, including rs1532278 (associated with myopia risk) in Müller glia and rs1874459 (associated with glaucoma risk) in bipolar and amacrine cells.[22]

Moving beyond the single locus-level gene identification from GWAS data, gene regulatory network (GRN) inference can provide pathway-level insights for understanding disease susceptibility. Although initially developed with bulk expression data, various machine-learning-based algorithms have been applied to single-cell expression data to predict the cell-type-relevant GRNs, especially in the context of the differentiation and cell-state transition.[23] Machine-learning-based pipelines (e.g., elastic net regression) were shown to identify cell-type-specific GRNs and cell-type-specific disease genes by integrating single-cell multi-modal data with the GWAS variants, which improved clinical phenotype prediction.[24] A deep convolutional neural network[25] was also used to infer gene-gene relationship and disease causality. This approach used gene expression levels from complex single-cell data converted into an image of a 2D histogram as an input for the deep-learning process, which outperforms previous methods in predicting TF target genes and the causality (direction) within the pathways. These studies have showcased the merit of machine-learning algorithms in handling single-cell datasets and characterizing GWAS risk loci as well as their interaction and downstream pathways.

UPCOMING BREAKTHROUGHS IN GWAS TRANSLATIONAL STUDIES

Candidate targets for drug repurposing

Drug repurposing, one of the notable directions in GWAS translational studies, refers to the identification of new indications for approved or investigational (including clinically failed) drugs that have not been approved. It is reported that only 0.02% of drug candidates from preclinical testing make it to market.[26] Given the high attrition rates, substantial cost, and slow pace of *de novo* drug discovery, exploiting known drugs can help improve their efficacy while minimizing side effects in clinical trials.

It has been consistently reported that drug targets with genetic support (e.g., GWAS associations) are more likely to be successful in clinical trials and drug development.[27,28] Pleiotropic and genetic correlation analyses can link target genes identified from one disease to multiple correlated diseases and thereby suggest new indications for known drugs. For example, raloxifene is a selective estrogen receptor modulator that was initially developed for osteoporosis and then successfully repositioned for breast cancer.[29] Consistently, GWAS-based pleiotropic analyses have indicated that *ESR1*, a gene that encodes an estrogen receptor, contributed to the shared genetic basis between bone mineral density and breast cancer risk.[30] It should be noted that pleiotropy can also result in unintended side effects when targeting certain genes, and therefore both potential beneficial and harmful effects should be considered in drug development processes. Moreover, multiple GWAS have identified signals on or near the genes encoding the targets of known pharmacological agents, such as *HMGCR* from low-density lipoprotein GWAS to statins[31] and *IL23R* from psoriasis GWAS to ustekinumab[32] (reviewed in Reay and Cairns[33]). However, it is still controversial how much weight should be given to GWAS-driven susceptibility genes in prioritizing drug targets to pursue.

To enable a more robust connection between GWAS signals and target genes as potential drug targets, multiple existing approaches can be applied, including eQTL, chromatin interaction (see linking variants to genes section), transcriptome-wide association studies (TWASs) for gene-level association tests, and Mendelian randomization (MR) for genetically informed causal inference[34] (Figure 2A). Additional information sources such as literature, drug-gene interaction databases (e.g., DGIdb), and clinical trial records (Figure 2A) can be valuable for assessing the feasibility of these GWAS-driven susceptibility genes as potential targets of drug development. To integrate all this information, AI-empowered platforms have been initiated. For example, PandaOmics is a platform that uses deep-learning models to predict the potential of drug targets in the form of the likelihood of a drug candidate entering phase 1 of clinical trials within 5 years. Their models included "omics scores" and "text scores." Omics scores consider all available genetic datasets, such as GWAS findings and gene-disease associations identified by MR or TWASs. Text scores are based on literature, grants, and patents searches. These models have been applied to identify candidate drug targets for amyotrophic lateral sclerosis.[35] Another platform developed by Open Targets integrated tissue specificity from open-access expression profiles, biological knowledge from Gene Ontology, and protein-protein interaction networks into a machine-learning model, which achieved improved power in identifying the drug target-disease pairs.[36] These studies have shown the value of AI in translating multidimensional knowledge into GWAS-informed drug prioritization, but further efforts are required to prioritize robust targets with better efficacy and specificity to ultimately achieve improved therapeutics for complex diseases.

PRS

Another important direction of GWAS translational studies is PRS, which calculates the cumulative effects of many genetic variants and provides a quantitative measure of an individual's genetic predisposition to diseases (Figure 2B). PRS has potential in a variety of medical scenarios, including disease risk prediction, diagnostic refinement, and therapeutic response management.[37] However, the clinical implementation of PRS is still rare, which could be attributed to two main obstacles. First, the discriminative ability of PRS is compromised by the multifactorial contributors to complex diseases. To address this issue, a neural-network-based model integrating polygenic and clinical predictors has been developed for cardiovascular disease risk.[38] This model is designed with three linear layers followed by deep survival machines,[39] a mixture of three linear layers to parameterize a mixture of Weibull distributions. The neural-network-based model was trained using features of 29 cardiovascular risk factors and 6 PRSs from different cohorts as input. The model was then validated using spatially separated samples from individual assessment centers of participants in the UK Biobank cohort to predict the occurrence of a major adverse cardiac event within 10 years. Compared with an existing model and a Cox proportional hazards model trained on the same data, the neural network model achieved better integration of polygenic and clinical predictors and improved predictive performance.[38] However, it should be noted that this model was developed and validated in the UK Biobank cohort and has not been evaluated in an entirely independent cohort.

The second challenge associated with PRS is poor transferability, a problem that is also encountered in many other statistical models. Transferability is defined as the accuracy of a model's predictions for an independent dataset, and in the context of PRS, poor transferability refers to the situation where a PRS generated from GWASs in one population does not perform well in other populations. The accuracy of PRS depends on the adequate estimation of allelic effect sizes and genetic similarity between training and target sets. Therefore, PRS is more predictive in European populations compared with underrepresented populations due to the larger sample sizes available for training in European populations. One solution to improve PRS transferability is to introduce biological priors and select "candidate causal" variants using functional genomics data. This approach is valid under the assumption that causal variants are largely shared across populations. A recent study has shown that cell-type-specific regulatory annotations can improve the transferability of PRS from European to East Asian populations across 21 human phenotypes.[40] Another study introduced polygenic transcription risk scores (PTRS) based on the predicted transcript levels, which presented higher portability from a European to an African population than variant-based PRSs.[41] Future

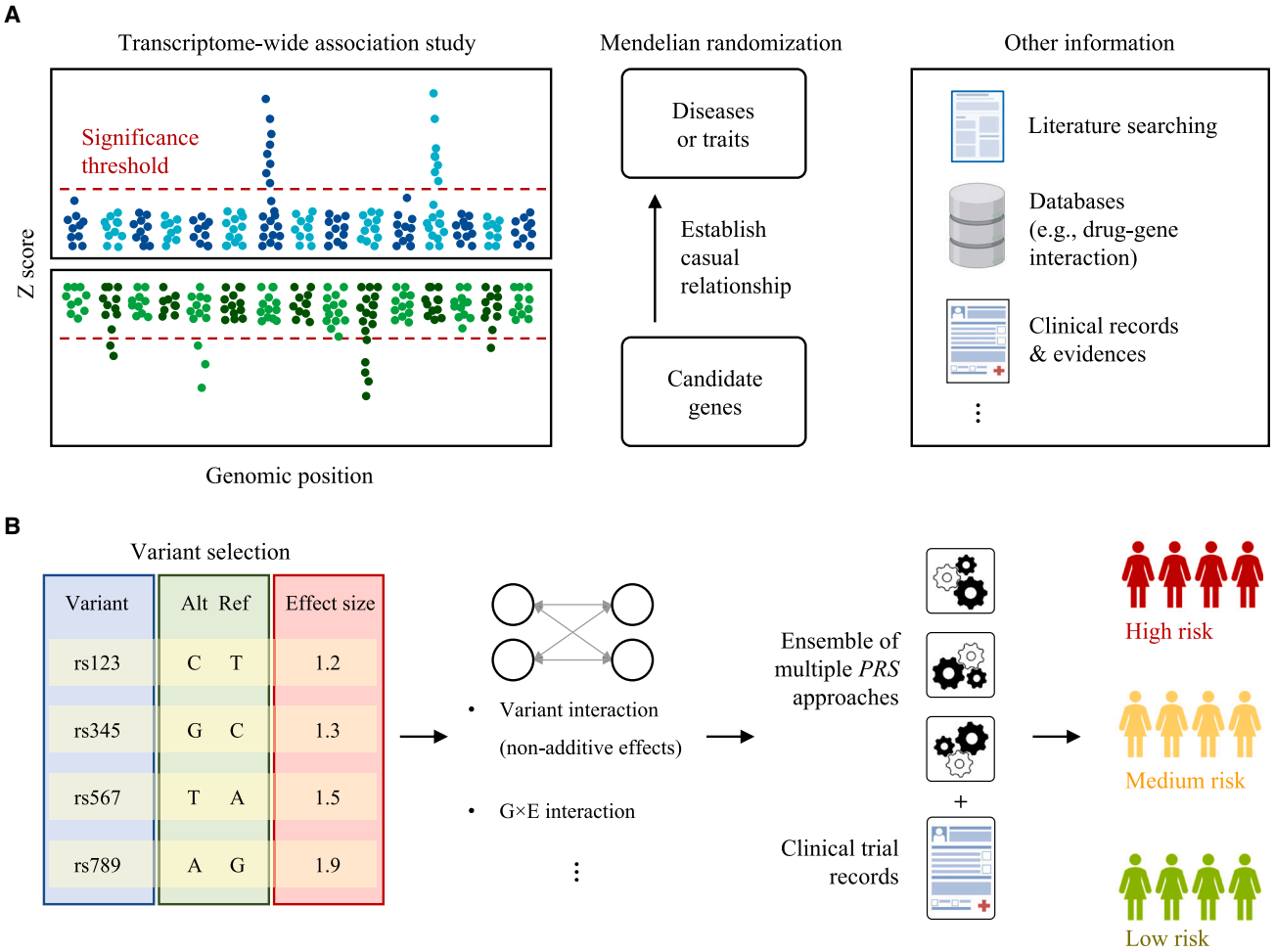


Figure 2. AI potentials in GWAS translational studies

(A) Multiple approaches/data are useful in translating GWAS findings into repurposed drug targets. Transcriptome-wide association studies (TWASs) are a gene-based association approach that can integrate GWAS and eQTL data to identify transcriptionally regulated genes associated with diseases. The top and bottom panels show positive and negative TWAS *Z* scores, respectively. Mendelian Randomization (MR) is a causal-inference approach to estimate the causal effects of target genes on risk of diseases or traits. Other information includes literature on target gene biology, established databases of drug-gene interaction, and clinical trial records relevant to the drug efficacy in patients.

(B) The polygenic risk score (PRS) is typically constructed as the weighted sum of a set of genetic variants, usually from GWASs (e.g., risk allele, effect size). Different PRS approaches can be used by considering non-additive variant effects or interactions between genetic and environmental factors (G × E). An ensemble of PRS approaches and a combination of clinical indicators are valuable in improving the predictive performance, resulting in relative risk estimation in a given population.

studies using AI and a combination of other functional data are anticipated to improve PRS transferability across populations.

CHALLENGES AND RECOMMENDATIONS

Scarcity of ground-truth training data

Most of the state-of-the-art approaches in GWAS functional and translational studies have relied on supervised AI models, but these attempts have been hindered by the lack of adequately sized ground-truth datasets for training and validation. Genetic susceptibility involves many potential target genes and biological pathways and interplay between them, which is often context dependent. Consequently, it is challenging for AI to make reliable predictions and unambiguously evaluate the predictive performance because confidently labeled data, such as causal variants and target genes underlying most of the GWAS loci, are rare.[14] It is promising that fast-evolving high-throughput assays (e.g., CRISPR and single-cell approaches) evaluating variant and gene function in diverse biological contexts as well as those involving animal models will help improve ground-truth causal variant/gene assignment in the coming years. As these data accumulate, it will be important to establish guidelines on what qualifies as a ground-truth gene/variant assignment, especially for those GWAS loci with multiple non-coding variants in high LD.

Recommendations

Beside ongoing efforts to improve the ground-truth datasets, a few AI-based analytic approaches could be considered to

alleviate the problem. One way to improve the data size and quality is to search and curate the training set in a more precise and efficient manner. AI-based tools have shown their potentials in handling the growing volume of literature with unstructured information. Regarding data searching, a tool enabled by algorithms of natural language processing can automatically recognize and extract genetic variant information with related entities (e.g., allele) from literature. This tool presents a state-of-the-art performance with over 90% in F-measure for variant recognition and is now available for the entire PubMed and PMC datasets.[42] Moreover, AI-based tools have been developed to improve the data curation. For example, a deep-learning-based approach achieves a precision 2.99 times higher than current query-based approach in curating the most comprehensive GWAS database, the NHGRI-EBI GWAS Catalog.[43] This deep-learning approach can efficiently identify relevant literature and thus significantly reduced the number of papers that required review in the manual curation process.

Another approach to address the challenge of insufficient ground-truth training datasets is to use label-free strategies. One example is using human-in-the-loop AI approaches,[44] which could train AI models using a relatively small size of existing ground-truth labels to generate large-scale new labels. The trained model could be further calibrated by a correction process of their newly generated labels by humans, which could ultimately generate the labels with high confidence. Such a human-in-the-loop strategy has been primarily applied to generating labels for the data from hematoxylin and eosin staining in immunohistochemistry.[45] Alternatively, generative models can be used to synthesize new data with labels based on prior knowledge from relevant resources. Considering the sparsity and heterogeneity of existing labels of coding variant pathogenicity, a deep generative model was developed without training on any of the existing labels.[46] Assuming that the evolutionary constraints from natural sequences reflected the propensity of variant pathogenicity, the generative model could learn the distribution of sequence variation across species and thus approximate the likelihood of each variant to be pathogenic by assigning them to clusters (benign, uncertain, or pathogenic).[46] This model outperforms conventional variant-effect-prediction approaches that rely on existing labels. It is expected that the label-free generative strategies could be applied to GWAS functional and translational studies to address the issue with limited ground-truth labels.

Transparency and AI "black box"

Another important challenge is that AI algorithms, such as deep neural networks, are considered black boxes that predict outputs from inputs without regard to the internal rationale (i.e., "end-to-end" strategy) and thus provide limited mechanistic insights for GWAS functional and translational studies. The parameters within the neural network are subject to extensive mathematical optimization during training, leading to a dense web of neural connections neither tied to an actual system nor based on human reasoning.

Recommendations

One way to address the "black-box issue" is to use models with fewer parameters and select a minimal set of features for prediction. For example, the study by the Open Targets group trained their variant-to-gene predictive model using an XGBoost gradient-boosting classifier with a binary logistic learning objective function.[14] Leave-one-in/-out analyses were performed, by leaving one feature in/out of model training at a time, to determine the individual feature's contribution to the output, which identified several key features in predicting target genes. It should be noted that such models with simple architecture may not capture the full complexity of genetic susceptibility to human diseases and may result in a loss of information. Another idea is to use models with a hierarchical resolution, whose internal logic fits naturally with biological systems and deep neural networks.[10] A notable example is DCell, a visible neural network with a hierarchical structure to predict the cellular growth of a eukaryotic cell (budding yeast) based on the genotypes.[47] Specifically, the neurons inside this neural network are organized into banks, each of which maps to a known cellular component. A predicted change in cellular growth (output) caused by combinatorial gene disruptions (input) can then be interpreted by examining the functional states (active or inactive) of underlying cellular components.[47] This group applied a similar strategy to predicting the drug response of cancer cells, where an interpretable deep-learning model was used to couple the inner workings of the model to the known hierarchical structure of human cell biology.[48] Nguyen et al. introduced an interpretable deep-learning tool, Varmole, which takes the genotypes and gene expression data as inputs to predict the disease phenotypes.[49] Varmole embeds prior biological knowledge of QTLs and GRNs into the deep-learning network, which enables the prioritization of the genetic variants and genes underlying the disease phenotypic prediction.[49] Wang et al. developed the deep structured phenotype network (DSPN), which adds a series of intermediate layers between the prediction of genotype and phenotype.[50] These intermediate layers could be associated with specific genes (e.g., expression level or chromatin status) or gene groups (e.g., coexpression modules) for mechanistic interpretations from genotypes to traits. It should be noted that the development of both Varmole and the DSPN benefited from the large-scale and comprehensive functional genomics resources of the PsychENCODE Consortium, including uniformly processed bulk transcriptome, chromatin, genotype, Hi-C, and single-cell transcriptomic data of brain tissues.[50] Moreover, the models with hierarchical resolution may require significant computational resources and expertise and may not be applicable to all types of genetic data. Despite these cautions, these studies present promising examples of interpreting AI models to inform mechanistic insights into genetic susceptibility to human diseases.

Ethical concerns and biases

While the application of AI may enable higher accuracy and better performance in conducting GWAS follow-up studies, it is also accompanied by a series of ethical concerns and biases that need to be addressed as the field moves forward. One notable concern is that AI system could exacerbate injustice and discrimination by amplifying the existing social inequity. For example, the predictive performance of AI is affected by the biases that are already presented in the training dataset, which

could create inequity in AI-driven health benefits for underrepresented racial and gender populations. Specifically, current functional genomics resources such as molecular QTL- and single-cell-based datasets lack diversity in terms of the represented ancestries, tissue/cell types, and cellular contexts.[21] It is notable that many of the current advances in GWAS functional studies using AI are largely limited to European populations and often involving blood traits and immune cells, reflecting the composition of current GWAS and functional datasets. Similarly, biased representation of socioeconomic, racial, and gender groups in the GWAS, clinical trials, and other medical studies could affect the predictive performance and generalizability of PRS and AI pipelines for drug repurposing, which could exacerbate the existing health disparities. Furthermore, the black-box nature of the AI process combined with embedded bias in the training data could potentially create algorithmic biases with misleading causal information, such as misinterpreting social and environmental effects as genetic effects in the prediction outcome.[51] Another ethical concern in AI-driven research is the privacy and data protection issue, given that large-scale and diverse genomic and clinical data are typically collected in generating AI models. For example, PRS development and evaluation using deep-learning-based models are prone to involving more diverse patient phenotype data and health records in addition to individual-level genotype data, which could create more complex data protection issues. It is still controversial regarding the extent to which anonymization can be ensured with the large amounts of data used for AI studies. In addition to these fairness and privacy issues in AI ethics, other concerns also exist, including regulatory uncertainties.[52]

Recommendations

To reduce bias and injustice in AI-based GWAS follow-up studies, conscious efforts should be made to include functional genomics and clinical datasets representing diverse populations for training deep-learning models. This process should certainly be accompanied by systematic efforts by the biomedical research community in generating more diverse genomic and medical databases/studies that accurately represent the whole populations that the medicine should serve. To this end, technology-driven solutions such as remote digital clinical trials using wearable devices without clinical sites could potentially help reduce the barriers in participation to previously underrepresented populations.[53] At the algorithm level, fairness-enhancing AI approaches have been developed. For example, multi-objective learning was proposed to mitigate the fairness via simultaneous optimization and automatic balance among accuracy and multiple fairness measures.[54] With regard to the privacy and data protection issues, some technical solutions could be considered in addition to implementing AI-specific health data protection policies. Federated learning, for example, could allow model training without sharing raw data in the multi-center collaboration setting, where separate training is performed by each center and model updates are shared and aggregated through a trusted central server.[55] Other approaches such as differential privacy, which involves randomly disrupting individual-level data while keeping global patterns of the dataset, and homomorphic encryption, which uses encrypted input data, could also be considered and combined with other solutions.[53] Next-generation privacy preservation techniques such as privacy-preserving federated learning are thus highly anticipated.

Conclusions

AI is proving itself to be a revolutionary technology in GWAS functional and translational studies. These AI-driven efforts have enriched our understanding of genetic susceptibility and empowered the translational potential of drug repurposing and individual risk prediction. However, before AI models can consolidate their role in clinical validity, efforts are required to address several challenges in performance, generalizability, and interpretability, as well as ethical concerns. It should be noted that most of these models were developed recently without accumulating enough time to be applied to diverse scenarios, and their value needs to be tested by the community over times. Novel AI strategies, including generative models and interpretable deep learning, may hold the key to unlocking the full potential of GWAS in providing biological insights and health benefits for complex human diseases.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100320.

ACKNOWLEDGMENTS

This work was supported by the Intramural Research Program (IRP) of the Division of Cancer Epidemiology and Genetics at National Cancer Institute and the National Library of Medicine at the US National Institutes of Health. Q.C. is supported by the K99/R00 Pathway to Independence Award (LM014024-01), National Library of Medicine, National Institutes of Health. E.L. is supported by the National Natural Science Foundation of China (Excellent Youth Scholars Program and 82090011).

AUTHOR CONTRIBUTIONS

E.L. conceived the idea. E.L. and P.W. drafted the manuscript. E.L. and J.C. revised the manuscript. P.W., Q.C., and Z.L. provided valuable comments on manuscript revision.

DECLARATION OF INTERESTS

The author declares no competing interests.

REFERENCES

1. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. Am. J. Hum. Genet. *101*, 5–22.
2. Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An expanded view of complex traits: from polygenic to omnigenic. Cell *169*, 1177–1186.
3. Schaid, D.J., Chen, W., and Larson, N.B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat. Rev. Genet. *19*, 491–504.
4. GTEx Consortium; Laboratory, Data Analysis &Coordinating Center LDACC—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx eGTEx groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

5. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

6. Kampmann, M. (2020). CRISPR-based functional genomics for neurological disease. Nat. Rev. Neurol. *16*, 465–480.

7. Jakhar, D., and Kaur, I. (2020). Artificial intelligence, machine learning and deep learning: definitions and differences. Clin. Exp. Dermatol. *45*, 131–132.

8. Sealfon, R.S.G., Wong, A.K., and Troyanskaya, O.G. (2021). Machine learning methods to model multicellular complexity and tissue specificity. Nat. Rev. Mater. *6*, 717–729.

9. Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F.J. (2019). Deep learning: new computational modelling techniques for genomics. Nat. Rev. Genet. *20*, 389–403.

10. Novakovsky, G., Dexter, N., Libbrecht, M.W., Wasserman, W.W., and Mostafavi, S. (2023). Obtaining genetics insights from deep learning via explainable artificial intelligence. Nat. Rev. Genet. *24*, 125–137.

11. Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. (2019). A primer on deep learning in genomics. Nat. Genet. *51*, 12–18.

12. Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. Mol. Syst. Biol. *12*, 878.

13. Zhou, J., Theesfeld, C.L., Yao, K., Chen, K.M., Wong, A.K., and Troyanskaya, O.G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat. Genet. *50*, 1171–1179.

14. Mountjoy, E., Schmidt, E.M., Carmona, M., Schwartzentruber, J., Peat, G., Miranda, A., Fumis, L., Hayhurst, J., Buniello, A., Karim, M.A., et al. (2021). An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nat. Genet. *53*, 1527–1533.

15. Shameer, K., Tripathi, L.P., Kalari, K.R., Dudley, J.T., and Sowdhamini, R. (2016). Interpreting functional effects of coding variants: challenges in proteome-scale prediction, annotation and assessment. Brief. Bioinformatics *17*, 841–862.

16. Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838.

17. Chen, K.M., Wong, A.K., Troyanskaya, O.G., and Zhou, J. (2022). A sequence-based global map of regulatory activity for deciphering human genetics. Nat. Genet. *54*, 940–949.

18. Liu, H., Doke, T., Guo, D., Sheng, X., Ma, Z., Park, J., Vy, H.M.T., Nadkarni, G.N., Abedini, A., Miao, Z., et al. (2022). Epigenomic and transcriptomic analyses define core cell types, genes and targetable mechanisms for kidney disease. Nat. Genet. *54*, 950–962.

19. Gallagher, M.D., Posavi, M., Huang, P., Unger, T.L., Berlyand, Y., Gruenewald, A.L., Chesi, A., Manduchi, E., Wells, A.D., Grant, S.F.A., et al. (2017). A dementia-associated risk variant near TMEM106B alters chromatin architecture and gene expression. Am. J. Hum. Genet. *101*, 643–663.

20. Gazal, S., Weissbrod, O., Hormozdiari, F., Dey, K.K., Nasser, J., Jagadeesh, K.A., Weiner, D.J., Shi, H., Fulco, C.P., O'Connor, L.J., et al. (2022). Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. Nat. Genet. *54*, 827–836.

21. Long, E., García-Closas, M., Chanock, S.J., Camargo, M.C., Banovich, N.E., and Choi, J. (2022). The case for increasing diversity in tissue-based functional genomics datasets to understand human disease susceptibility. Nat. Commun. *13*, 2907.

22. Wang, S.K., Nair, S., Li, R., Kraft, K., Pampari, A., Patel, A., Kang, J.B., Luong, C., Kundaje, A., and Chang, H.Y. (2022). Single-cell multiome of the human retina and deep learning nominate causal variants in complex eye diseases. Cell Genom. *2*, 100164.

23. Pratapa, A., Jalihal, A.P., Law, J.N., Bharadwaj, A., and Murali, T.M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat. Methods *17*, 147–154.

24. Jin, T., Rehani, P., Ying, M., Huang, J., Liu, S., Roussos, P., and Wang, D. (2021). scGRNom: a computational pipeline of integrative multi-omics analyses for predicting cell-type disease genes and regulatory networks. Genome Med. *13*, 95.

25. Yuan, Y., and Bar-Joseph, Z. (2019). Deep learning for inferring gene relationships from single-cell expression data. Proc. Natl. Acad. Sci. USA *116*, 27151–27158.

26. Kraljevic, S., Stambrook, P.J., and Pavelic, K. (2004). Accelerating drug discovery. EMBO Rep. *5*, 837–842.

27. Nelson, M.R., Tipney, H., Painter, J.L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P.C., Li, M.J., Wang, J., et al. (2015). The support of human genetic evidence for approved drug indications. Nat. Genet. *47*, 856–860.

28. King, E.A., Davis, J.W., and Degner, J.F. (2019). Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. PLoS Genet. *15*, e1008489.

29. Maximov, P.Y., Lee, T.M., and Jordan, V.C. (2013). The discovery and development of selective estrogen receptor modulators (SERMs) for clinical practice. Curr. Clin. Pharmacol. *8*, 135–155.

30. Peng, C., Lou, H.-L., Liu, F., Shen, J., Lin, X., Zeng, C.-P., Long, J.-R., Su, K.-J., Zhang, L., Greenbaum, J., et al. (2017). Enhanced identification of potential pleiotropic genetic variants for bone mineral density and breast cancer. Calcif. Tissue Int. *101*, 489–500.

31. Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature *466*, 707–713.

32. Nair, R.P., Duffin, K.C., Helms, C., Ding, J., Stuart, P.E., Goldgar, D., Gudjonsson, J.E., Li, Y., Tejasvi, T., Feng, B.-J., et al. (2009). Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat. Genet. *41*, 199–204.

33. Reay, W.R., and Cairns, M.J. (2021). Advancing the use of genome-wide association studies for drug repurposing. Nat. Rev. Genet. *22*, 658–671.

34. Li, B., and Ritchie, M.D. (2021). From GWAS to gene: transcriptome-wide association studies and other methods to functionally understand GWAS discoveries. Front. Genet. *12*, 713230.

35. Pun, F.W., Liu, B.H.M., Long, X., Leung, H.W., Leung, G.H.D., Mewborne, Q.T., Gao, J., Shneyderman, A., Ozerov, I.V., Wang, J., et al. (2022). Identification of therapeutic targets for amyotrophic lateral sclerosis using PandaOmics - an AI-enabled biological target discovery platform. Front. Aging Neurosci. *14*, 914017.

36. Han, Y., Klinger, K., Rajpal, D.K., Zhu, C., and Teeple, E. (2022). Empowering the discovery of novel target-disease associations via machine learning approaches in the open targets platform. BMC Bioinformatics *23*, 232.

37. Pasaniuc, B., and Price, A.L. (2017). Dissecting the genetics of complex traits using summary association statistics. Nat. Rev. Genet. *18*, 117–127.

38. Steinfeldt, J., Buergel, T., Loock, L., Kittner, P., Ruyoga, G., Zu Belzen, J.U., Sasse, S., Strangalies, H., Christmann, L., Hollmann, N., et al. (2022). Neural network-based integration of polygenic and clinical information: development and validation of a prediction model for 10-year risk of major adverse cardiac events in the UK Biobank cohort. Lancet. Digit. Health *4*, e84–e94.

39. Nagpal, C., Li, X., and Dubrawski, A. (2021). Deep survival machines: fully parametric survival regression and representation learning for censored data with competing risks. IEEE J. Biomed. Health Inform. *25*, 3163–3175.

40. Amariuta, T., Ishigaki, K., Sugishita, H., Ohta, T., Koido, M., Dey, K.K., Matsuda, K., Murakami, Y., Price, A.L., Kawakami, E., et al. (2020). Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. Nat. Genet. *52*, 1346–1354.

41. Liang, Y., Pividori, M., Manichaikul, A., Palmer, A.A., Cox, N.J., Wheeler, H.E., and Im, H.K. (2022). Polygenic transcriptome risk scores (PTRS) can improve portability of polygenic risk scores across ancestries. Genome Biol. *23*, 23.

42. Wei, C.-H., Allot, A., Riehle, K., Milosavljevic, A., and Lu, Z. (2022). tmVar 3.0: an improved variant concept recognition and normalization tool. Bioinformatics *38*, 4449–4451.

43. Lee, K., Famiglietti, M.L., McMahon, A., Wei, C.-H., MacArthur, J.A.L., Poux, S., Breuza, L., Bridge, A., Cunningham, F., Xenarios, I., and Lu, Z. (2018). Scaling up data curation using deep learning: an application to literature triage in genomic variation resources. PLoS Comput. Biol. *14*, e1006390.

44. Ristoski, P., Zubarev, D.Y., Gentile, A.L., Park, N., Sanders, D., Gruhl, D., Kato, L., and Welch, S. (2020). Expert-in-the-loop AI for polymer discovery. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (ACM). https://doi.org/10.1145/3340531.3416020.

45. van der Wal, D., Jhun, I., Laklouk, I., Nirschl, J., Richer, L., Rojansky, R., Theparee, T., Wheeler, J., Sander, J., Feng, F., et al. (2021). Biological data annotation via a human-augmenting AI-based labeling system. NPJ Digit. Med. *4*, 145.

46. Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J.K., Brock, K., Gal, Y., and Marks, D.S. (2021). Disease variant prediction with deep generative models of evolutionary data. Nature *599*, 91–95.

47. Ma, J., Yu, M.K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., and Ideker, T. (2018). Using deep learning to model the hierarchical structure and function of a cell. Nat. Methods *15*, 290–298.

48. Kuenzi, B.M., Park, J., Fong, S.H., Sanchez, K.S., Lee, J., Kreisberg, J.F., Ma, J., and Ideker, T. (2020). Predicting drug response and synergy using a deep learning model of human cancer cells. Cancer Cell *38*, 672–684.e6.

49. Nguyen, N.D., Jin, T., and Wang, D. (2021). Varmole: a biologically drop-connect deep neural network model for prioritizing disease risk variants and genes. Bioinformatics *37*, 1772–1775.

50. Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F.C.P., Clarke, D., Gu, M., Emani, P., Yang, Y.T., et al. (2018). Comprehensive functional genomic resource and integrative model for the human brain. Science *362*, eaat8464.

51. Fritzsche, M.-C., Akyüz, K., Cano Abadía, M., McLennan, S., Marttinen, P., Mayrhofer, M.T., and Buyx, A.M. (2023). Ethical layering in AI-driven polygenic risk scores-New complexities, new challenges. Front. Genet. *14*, 1098439.

52. Nordström, M. (2022). AI under great uncertainty: implications and decision strategies for public policy. AI Soc. *37*, 1703–1714.

53. Acosta, J.N., Falcone, G.J., Rajpurkar, P., and Topol, E.J. (2022). Multimodal biomedical AI. Nat. Med. *28*, 1773–1784.

54. Qingquan, Z., Jialin, L., Zeqi, Z., Junyi, W., Bifei, M., and Xin, Y. (2022). Mitigating unfairness via evolutionary multi-objective ensemble learning. Preprint at arXiv.

55. Kaissis, G.A., Makowski, M.R., Rückert, D., and Braren, R.F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. Nat. Mach. Intell. *2*, 305–311.