Trends in Genetics

CellPress OPEN ACCESS

Feature Review

From genetic associations to genes: methods, applications, and challenges

Ting Qi[1,2,3,*], Liyang Song[1,2,3], Yazhou Guo[1,2,3], Chang Chen[1,2,3], and Jian Yang[1,2,*]

Genome-wide association studies (GWASs) have identified numerous genetic loci associated with human traits and diseases. However, pinpointing the causal genes remains a challenge, which impedes the translation of GWAS findings into biological insights and medical applications. In this review, we provide an in-depth overview of the methods and technologies used for prioritizing genes from GWAS loci, including gene-based association tests, integrative analysis of GWAS and molecular quantitative trait loci (xQTL) data, linking GWAS variants to target genes through enhancer–gene connection maps, and network-based prioritization. We also outline strategies for generating context-dependent xQTL data and their applications in gene prioritization. We further highlight the potential of gene prioritization in drug repurposing. Lastly, we discuss future challenges and opportunities in this field.

Highlights

Integrating genome-wide association studies (GWASs) with molecular quantitative trait loci (xQTLs), including context-dependent (cd)-xQTLs, across multiple 'omics levels helps unveil the putative causal genes underlying GWAS signals, the relevant cell types, and the likely genetic regulation mechanisms of the prioritized genes.

High-resolution enhancer–gene connection maps can be utilized for gene prioritization by linking the fine-mapped GWAS variants in regulatory elements to their target genes.

Incorporating GWAS data with biological networks aids in identifying disease-associated genes, even those with weak GWAS signals.

Integrative approaches that leverage GWAS findings, perturbation-induced transcriptomic profiles, and biological networks hold immense potential for drug repurposing.

Deciphering genetic associations: the hurdles and complexities

GWAS is an experimental design that has led to tremendous success in uncovering genetic variants associated with human traits, including diseases [1,2]. Over the past 15 years, the sample sizes of GWASs and the number of investigated traits have increased rapidly [3–5], leading to the identification of numerous genetic variants associated with over 5000 human traits, as reported in the GWAS catalog [6,7]. Most of these variants are common, because conventional GWASs are primarily designed to detect such variants. Despite the success of GWASs in discovering genetic associations, the mechanisms underlying most GWAS loci remain elusive due to the difficulty in determining the causal variants and genes responsible for these associations (Figure 1). One major hurdle is the vast number of variants in linkage disequilibrium (LD) with the lead variants, making it challenging to discern the causal variants driving the observed associations [8]. Moreover, the fact that most GWAS signals are in non-coding regions of the genome [9] adds another layer of complexity to establishing a definitive link between the genetic association signals and genes. Additionally, the complex nature of gene regulation, along with the potential involvement of multiple genes within a single locus, further complicates pinpointing the causal genes.

These challenges have stimulated the development of methods and analytical paradigms geared toward identifying the putative causal genes in GWAS loci. This process, often referred to as gene prioritization, holds immense potential for deepening our understanding of disease etiology, guiding the development of new therapeutics, and discovering biomarkers for early disease detection [10]. While experimental approaches, such as gene knockout and knockdown, are essential for understanding the relevance of genes to specific traits, in this review, we focus primarily on the statistical methods and related technologies that can be utilized to prioritize genes underlying common variant association signals identified through GWAS. We outline the key features of the methods, as well as some of their applications, and discuss the persistent challenges and emerging opportunities in this field.

[1]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou 310024, China
[2]School of Life Sciences, Westlake University, Hangzhou 310024, China
[3]These authors contributed equally to this work.

*Correspondence:
qiting@westlake.edu.cn (T. Qi) and
jian.yang@westlake.edu.cn (J. Yang).

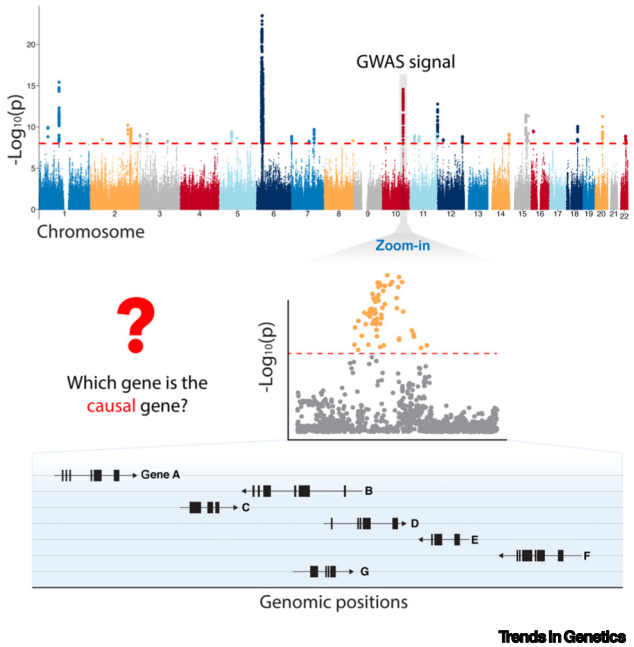

Figure 1. The challenge in prioritizing genes underlying genome-wide association study (GWAS) signals.

Gene-based association tests

Recent evidence suggests that most complex traits, including diseases, are polygenic and that trait-associated genetic variants are enriched near genes [11]. Thus, the power of gene discovery can be enhanced by examining the aggregated effect of a set of variants within and around a gene. The current common practice for conducting gene-based association tests is to use GWAS summary statistics, supplemented with LD information from a reference sample (Figure 2A). This approach is favored due to its adaptability to various study designs and the relative ease of obtaining GWAS summary statistics, and, thus, has been adopted by several gene-based test methods, such as Versatile Gene-based Association Study (VEGAS) [12], Pascal [13], fastBAT [14], Multi-marker Analysis of GenoMic Annotation (MAGMA) [15], and mBAT-combo [16]. These methods share a common feature of testing the aggregated effect of variants within a gene locus by summing up their chi-squared statistics (i.e., sum of squared z-statistics), with or without weighting, but differ in how they assess the statistical significance of the test.

To assess the significance of a gene-based test statistic, VEGAS utilizes reference LD data to simulate z-statistics of variants within a gene locus under the null hypothesis, while preserving the correlations among them, and then compares the observed gene-level sum of squared z-statistics to those obtained from simulations [12]. Under this method framework, the precision of the *P* value is capped by the number of simulations, and obtaining high-precision *P* values requires considerable computational resources. This has led to the development of methods that use numerical approaches to evaluate the distribution of a quadratic form of multivariate normal variables, denoted as $\boldsymbol{T} = \boldsymbol{z}^T \mathbf{I} \boldsymbol{z}$, with $\boldsymbol{z}$ being a vector of GWAS z-statistics for a set of variants and $\mathbf{I}$ being an identity matrix. These methods allow for the computation of a gene-level *P* value without the need for running simulations or permutations. Examples of these methods include Pascal [13], fastBAT [14], and MAGMA [15]. The recently developed mBAT-combo [16] method further improves power in effectively handling masking effects (i.e., situations where the product of the effects of two variant alleles is in the opposite direction to their LD correlation).

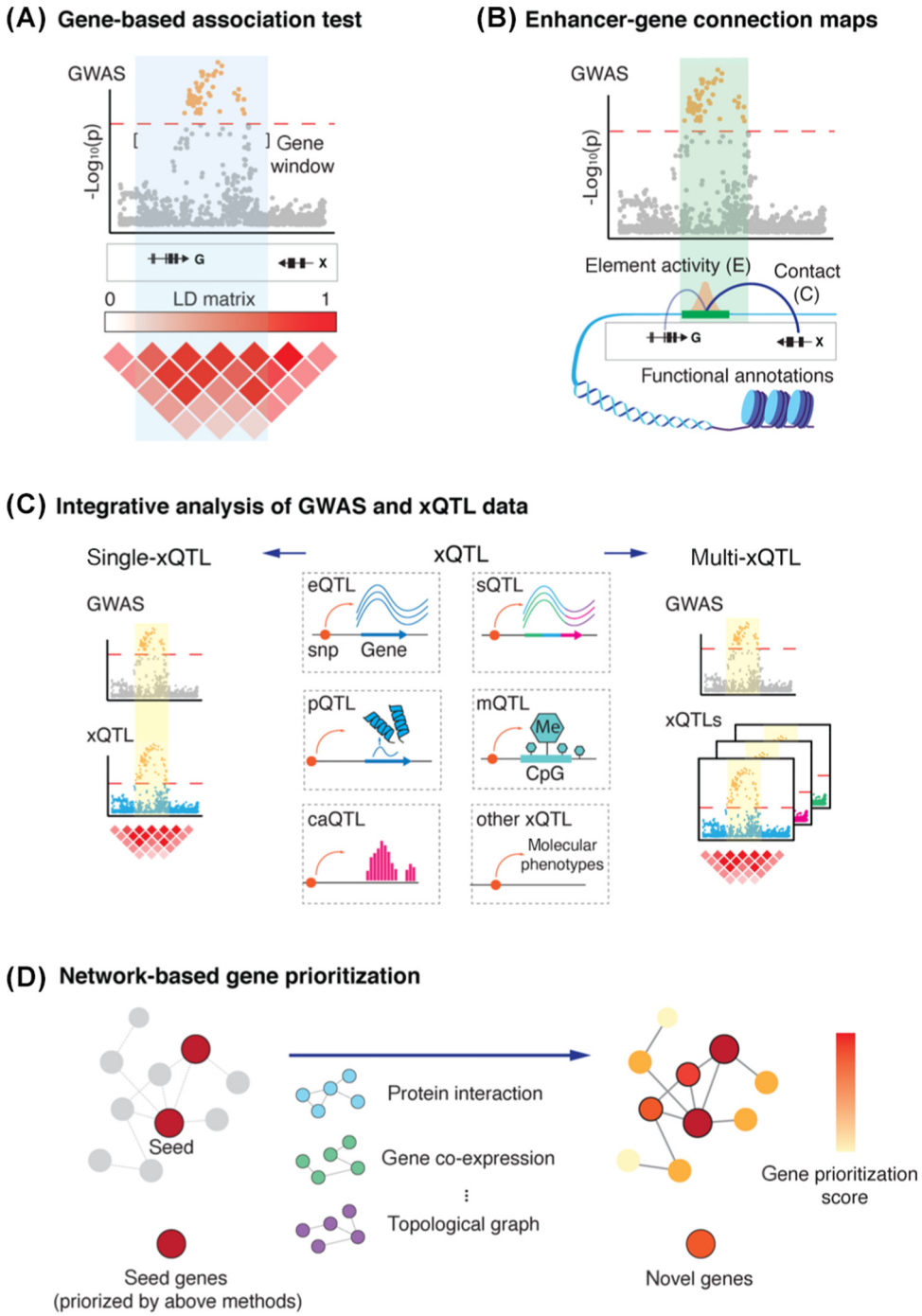

Figure 2. Overview of gene prioritization methods. Abbreviations: caQTL, chromatin accessibility quantitative trait loci; eQTL, expression quantitative trait loci; GWAS, genome-wide association study; LD, linkage disequilibrium; mQTLs, DNA methylation quantitative trait loci; pQTLs, protein abundance quantitative trait loci; sQTLs, splicing quantitative trait loci; xQTLs, molecular quantitative trait loci.

Genes identified by gene-based association tests can serve as valuable resources for subsequent analysis, such as inferring tissues and cell types relevant to a trait of interest [17,18]. Despite the successful applications of gene-based association tests, several challenges remain. First, these tests aggregate variant-level association signals into gene-level scores without explicitly considering the biological mechanisms underlying the associations, which may limit the interpretability of the findings. Second, gene-based association tests may not fully capture the effects of distal regulatory elements (e.g., those located far from their target genes, possibly millions of base pairs away or even on different chromosomes), because these tests typically focus on variants within or close to the gene boundaries, potentially resulting in a loss of power [19].

Integration of GWAS and molecular QTLs to identify putative causal genes

The observation that most GWAS signals are in noncoding regions of the genome implies that genetic variants may influence traits through gene regulation. Advances in high-throughput technologies have enabled the generation of extensive genome-wide molecular phenotype data, greatly facilitating the identification of genetic variants associated with molecular phenotypes (i.e., xQTLs), such as expression QTLs (eQTLs), DNA methylation QTLs (mQTLs), splicing QTLs (sQTLs), chromatin accessibility QTLs (caQTLs), and protein abundance QTLs (pQTLs). Several extensive xQTL resources have been generated, and their summary statistics have been made available to the research community (Table 1). The integration of GWAS and xQTL data can enhance our understanding of whether genetic variants influence complex traits by regulating molecular phenotypes, which not only facilitates the identification of putative causal genes, but

Table 1. Summary of commonly used xQTL data sets[a]

Study	Sample size	Tissue or cell type	Refs
caQTL			
Kumasaka *et al.*	100	Lymphoblastoid cell line	[185]
Bryois *et al.*	272	Brain	[186]
hQTL			
BLUEPRINT	200	Three immune cell types	[187]
ROSMAP	561	Brain	[188]
mQTL			
GTEx	424	Nine tissues	[189]
Brain-mMeta	1160	Brain	[56]
LBC+BSGS	1980	Blood	[40]
GoDMC	27 750	Blood	[190]
eQTL and sQTL			
eQTLGen (eQTL)	32 000	Blood	[30]
GTEx (eQTL and sQTL)	832	49 tissues	[24]
OneK1K (eQTL)	982	14 blood cell types	[100]
BrainMeta (eQTL and sQTL)	2865	Brain	[41]
pQTL			
INTERVAL	3301	Plasma	[191]
SCALLOP	30 931	Plasma	[192]
deCODE	35 559	Plasma	[193]
UKB_PPP	35 571	Plasma	[194]

[a]Abbreviation: hQTL, histone modification QTL.

also provides valuable insights into the molecular mechanisms underpinning the genetic associations. Such information can also guide researchers toward potential targets for functional validation and the development of therapeutic strategies. Moreover, the integrative analysis may uncover genes where nearby genetic variants do not meet the genome-wide significance threshold with the current sample size but attain genome-wide significance when examined with a larger sample size [20]. Recognizing the potential and benefits provided, a variety of methods have been developed for integrating GWAS and xQTL data, which can be broadly classified into three categories: transcriptome-wide association studies (TWASs), colocalization analysis, and Mendelian randomization (MR).

Transcriptome-wide association studies

The concept of TWAS typically refers to the analysis that associates the expression levels of genes across the transcriptome with a trait of interest. In the context of gene prioritization, it refers specifically to the analysis that links genetically predicted gene expression levels with a trait. Various TWAS methods have been developed, including PrediXcan [21] and FUSION-TWAS [22] among others, each with its distinct methodological design. PrediXcan uses an elastic net regression model [23] to estimate variant weights from individual-level genotype and gene expression data. These weights are subsequently used to predict gene expression levels in a GWAS cohort, which are then assessed for their associations with a trait. PrediXcan requires individual-level genotype and phenotype data in the GWAS cohort, which are often unavailable, particularly for GWAS meta-analyses. FUSION-TWAS circumvents this problem by performing predictions based on GWAS summary statistics. Specifically, it estimates the z-score of association between a genetically predicted gene expression level and a trait as a linear combination of variant-trait z-scores multiplied by variant–gene association weights. FUSION-TWAS uses individual-level genotype and gene expression data, which can be obtained from a reference panel, such as the Genotype-Tissue Expression (GTEx) [24], to estimate these weights by multiple strategies, such as the Bayesian sparse linear mixed model (BSLMM) [25].

Given that genetic variants used to predict gene expression in TWAS are often enriched in regulatory elements, methods such as EpiXcan [26] and MOSTWAS [27] have been developed to improve the prediction accuracy of gene expression by incorporating functional annotation data, thereby boosting the power of detecting gene–trait associations. EpiXcan integrates epigenomic annotation data (e.g., DNA methylation and histone modifications) with *cis*-eQTL data to obtain priors that reflect the likelihood of genetic variants being involved in gene regulation. It then uses an adaptive mapping approach to rescale these priors and leverages a weighted elastic net model for gene expression prediction. When applied to 58 traits and 14 eQTL data sets, EpiXcan exhibited an increase of over 18% in the number of gene–trait associations, compared with PrediXcan. MOSTWAS improves the prediction accuracy by incorporating genetic predictors of mediating biomarkers (e.g., DNA methylation and miRNAs) of gene expression as well as distal eQTLs mediated by local biomarkers into the prediction model.

While TWAS methods have shown promising results in prioritizing genes for complex traits [28,29], they do have several challenges. First, the accuracy of the gene expression prediction model, which is fundamental to the performance of the TWAS methods, depends on various factors, including training sample size, heritability of each gene, and homogeneity between the transcriptome reference sample and testing sample. Most existing TWAS methods utilize genetic effects on gene expression estimated from relatively small reference panels, limiting the number of genes that can be accurately predicted. Second, some TWAS methods, such as FUSION-TWAS and PrediXcan, require individual-level genotype and gene expression data, making them inapplicable to eQTL summary data from large-scale meta-analyses, such as eQTLGen

[30]. One solution is to train the gene expression prediction model using summary-level eQTL data, as implemented in the recently developed method, OTTERS [31]. Third, significant TWAS signals can arise when two distinct causal variants, one affecting gene expression and the other influencing the trait, are in LD with each other, resulting in the prioritization of noncausal genes [32]. Moreover, the correlation of predicted expression among genes, which could be due to shared or distinct eQTLs in LD, may introduce the risk of identifying irrelevant genes [32].

Colocalization

Colocalization analysis commonly refers to a statistical analysis used to ascertain whether the genetic association signals from two traits overlap within a certain genomic region due to shared causal variant(s). This approach has been extensively utilized to prioritize genes from GWAS loci by scrutinizing the colocalization of GWAS and eQTL signals. Among the colocalization methods is the Regulatory Trait Concordance (RTC) method, which is an empirical approach based on the assumption that, if a GWAS signal and an eQTL signal are driven by the same causal variant, the eQTL signal would be markedly reduced or even eliminated after correcting the expression phenotype for the GWAS variant [33]. QTLMatch uses a likelihood ratio test to identify genes for which the lead GWAS and eQTL variants are colocalized due to a shared causal variant [34]. However, both methods require individual-level data, which restricts their widespread application. To overcome this problem, COLOC [35] was developed, which only requires GWAS and eQTL summary statistics, even without the need for reference LD. COLOC uses an approximate Bayes factor to calculate the posterior probabilities of a variant being causal for two traits. Due to its flexibility with input data, COLOC has been widely utilized to test the colocalization of GWAS and eQTL signals.

COLOC assumes that a single causal variant underlies the colocalized GWAS and eQTL signals. While this assumption is convenient, it is often unrealistic. When this assumption does not hold, COLOC tends to underestimate the true posterior probability of colocalization, leading to a loss of power [36]. To account for the possibility of multiple causal variants within one locus, eQTL and GWAS Causal Variant Identification in Associated Regions (eCAVIAR) utilizes a fine-mapping approach to identify putative causal variants for both GWAS and eQTL signals [37]. It then uses a probabilistic model to estimate the colocalization posterior probability for each fine-mapped variant, calculated as the product of the probabilities of the variant being causal for both the GWAS and eQTL signals. However, this approach leads to an exponentially increasing computational load as the number of assumed causal variants increases. Of note, both eCAVIAR and COLOC require users to specify priors of the colocalization models. If the priors are severely mis-specified relative to the true models, this could result in an inflated type I error rate. One solution to this issue is to estimate priors from data, which has been shown to improve the robustness of the colocalization analysis [38]. Moreover, colocalization methods do not provide the direction or magnitude of the effect of a prioritized gene on the trait, which could be instrumental in inferring the molecular mechanisms underlying the genetic association signals and advancing the prioritized genes toward drug development or repurposing. Additionally, the colocalization methods do not distinguish between horizontal pleiotropy (i.e., when a genetic variant regulates a trait and a gene independently) and causality (i.e., when a genetic variant affects a trait through gene regulation).

Mendelian randomization

MR [39] is a statistical method that leverages one or multiple genetic variants as instrumental variables (IVs) to examine a causal relationship between an exposure and an outcome. This methodological framework has been applied or repurposed to test for causal or pleiotropic associations between molecular phenotypes and a trait of interest, or even between molecular phenotypes.

MR was initially developed based on a single sample where both exposure and outcome are measured. It has since been extended to accommodate the scenario where exposure and outcome are measured on two independent samples, a concept known as two-sample MR. This extension has remarkably broadened the application of MR, including its use for gene prioritization. Summary-data-based Mendelian Randomization (SMR) [20] is a variant of the two-sample MR method that integrates GWAS and eQTL summary data to identify genes, the expression levels of which are associated with a trait through shared genetic effects. Although originally developed for eQTL data, SMR has been utilized with other types of xQTL data [40,41]. The original version of SMR only uses the lead *cis*-xQTL variant as IV, and it has since been extended to SMR-multi to accommodate the potential presence of multiple *cis*-xQTL causal variants [40]. SMR-multi uses multiple, typically correlated variants as IVs and a set-based test, akin to the gene-based association test described in the preceding text, to combine the SMR test statistics across multiple IVs, with their correlations accounted for [40].

When only one genetic variant or multiple correlated variants in a locus are used as IVs, MR often cannot distinguish between the causality model (where the effect of a causal variant on the outcome is mediated through the exposure), the pleiotropy model (where the causal variant influences the exposure and outcome through separate paths), or even the linkage model (where two different causal variants independently affect the exposure and outcome). This situation is common, particularly when only *cis*-xQTL data are available. This issue is not exclusive to MR methods but is a common problem for most methods that integrate GWAS and xQTL data to establish associations between molecular phenotypes and traits, including TWAS. Compared with the causality and pleiotropy models, the linkage model has the least biological relevance. To reject the linkage model, the HEterogeneity In Dependent Instruments (HEIDI) method was developed and included as part of an SMR analysis to test whether the overlapping GWAS and xQTL signals, as detected by SMR or SMR-multi, are driven by the same set (i.e., the pleiotropy or causality model) or distinct (i.e., the linkage model) sets of causal variants [20]. In practice, when HEIDI analysis is not feasible for reasons such as the lack of signed xQTL effects, COLOC can be used in combination with SMR or SMR-multi to reject the linkage model.

Even if a GWAS signal is known to be colocalized with an xQTL signal due to the same set of causal variants, it remains challenging to further distinguish between the causality and pleiotropy models. One possible solution is to perform an MR analysis with multiple independent genetic variants, such as those located on different chromosomes. A comprehensive review of these multi-IV MR methods can be found elsewhere [42]. These include MR-inverse variance weighted (IVW) [43], weighted median [44], weighted mode-based estimate [45], heterogeneity test [46], GSMR [47], and MR-Egger [48]. The use of multiple independent variants can help mitigate potential confounding from pleiotropy, thereby improving the robustness of causal inference, although the degree of this robustness varies across methods [49]. Nevertheless, these methods have been used to identify putative causal genes associated with complex traits. For instance, Zheng *et al.* utilized the MR-IVW method to estimate the causal effects of 66 proteins, which have both *cis*- and *trans*-pQTLs, on a range of traits [50]. Additionally, during the coronavirus disease 2019 (COVID-19) pandemic, multiple studies applied multi-IV MR approaches to integrate eQTL and pQTL data into GWAS to identify potential therapeutic targets for COVID-19, leading to the discovery of several promising targets, including *OAS1* and *IFNAR2* [51,52].

Collectively, the MR approaches provide a unique toolkit for gene discovery and causality assessment. However, for a valid causal inference, genetic variants used as IVs must satisfy three conditions: (i) they are strongly associated with the exposure; (ii) they are not associated with confounders; and (iii) they are not associated with the outcome other than through the exposure.

Violation of any of these assumptions can lead to a biased result. Thus, we recommend adhering to a stringent threshold (e.g., 5e-8) for IV selection and resisting the temptation to lower this threshold to include more IVs, thereby ensuring that the first assumption is always valid. Moreover, it has been demonstrated that all MR methods may exhibit bias in the presence of strong pleiotropic effects [49], violating the third assumption. As such, adopting a triangulation strategy, such as by utilizing multiple MR methods, is advisable. Significant discrepancies in MR estimates across different methods could indicate biases.

Despite the widespread use of the integrative methods discussed in the preceding text [53], these methods still pose several challenges. One such challenge is co-regulation, where multiple genes or molecular phenotypes are correlated because they are regulated by the same, or distinct but correlated, causal variants. This can lead to correlated effects of genes on traits, making it difficult to pinpoint the putative causal gene(s). While methods such as COLOC and HEIDI can, to some extent, identify scenarios where two genes have distinct but linked causal variants, they cannot identify cases where genes share the same causal variant(s). Fine-mapping of Causal Gene Sets (FOCUS) [54] and Transcriptome-wide MR (TWMR) [55] have attempted to address this issue by jointly analyzing multiple genes in a region, but they may not fully account for LD between eQTLs of a noncausal gene and GWAS causal variants. Additionally, many xQTL studies are based on samples from a single or limited number of tissues, which may not be relevant to the trait of interest. Thus, it is important to consider factors, such as tissue specificity, developmental stage, and environmental context, when validating findings derived from these methods.

Integration of GWAS and xQTL data from multiple tissues or 'omics layers

Integration of GWAS and eQTL data from multiple tissues

While gene expression levels are known to vary in different tissues, genetic factors affecting gene expression in *cis* are largely shared across tissues [24]. For instance, the correlation of *cis*-eQTL effects between the brain and blood is estimated to be >0.70 [56]. When eQTL data from multiple tissues are available, one common practice is to either analyze the tissue most relevant to the trait or analyze each tissue separately. However, this approach could limit the power of the analysis due to the small to modest eQTL sample sizes available for most tissues, especially those that are difficult to acquire.

Recognizing the shared eQTL effects across tissues, one strategy to enhance the power of detecting gene–trait associations is to perform an integrative analysis of GWAS data with eQTL data from multiple tissues, as implemented in MultiXcan and S-MultiXcan [57]. While MultiXcan requires individual-level GWAS data and S-MultiXcan only needs summary-level GWAS data, both methods still necessitate individual-level genotype and gene expression data of the expression training set. The multi-tissue analysis is achieved by regressing the principal components (PCs) of predicted expression across tissues on the trait. This strategy is computationally efficient and improves power over PrediXcan, but complicates the interpretation of the effect size and direction for each PC association. Unified Test for Molecular Signatures (UTMOST) applies a meta-analysis framework to combine single-tissue TWAS results from multiple tissues, which improves the power to detect trait-associated genes and also allows the identification of gene–tissue pairs that exhibit the strongest associations with the trait of interest [58]. Additionally, Joint-tissue Imputation (JTI) is an extension of PrediXcan. It uses a Bayesian hierarchical model to integrate multi-tissue transcriptomic data and atlases of regulatory elements, which improves the accuracy in predicting gene expression, thereby enhancing the power of identifying gene–trait associations in each tissue [59]. Colocalization and Fine-mapping in the Presence of Allelic Heterogeneity (CAFEH) is another Bayesian approach that integrates genetic association data from multiple traits (including tissues) and LD information to improve the identification of putative causal variants

shared by multiple traits [60]. By leveraging the widespread allelic heterogeneity in genetic regulation and accounting for the tissue specificity of *cis*-eQTLs, CAFEH enables the prioritization of both the putative causal genes and the relevant tissues underlying GWAS loci.

Integration of GWAS and xQTL data from multiple 'omics layers

Despite extensive efforts in eQTL mapping, the fraction of GWAS signals explained by eQTLs remains modest [61,62], and the proportion of trait heritability mediated by eQTLs has been estimated to be as low as ~10% [63]. Gene regulation is a complex process involving multiple layers of control, including but not limited to, chromatin accessibility, epigenetic modifications, RNA splicing, and translation. The increasing availability of summary statistics for various types of xQTL, such as caQTL, mQTL, and pQTL, offers an opportunity to gain deeper insights into the molecular mechanisms underlying genetic associations. Even when xQTL data from multiple 'omics levels are available, one common practice is to analyze each type of xQTL separately with the GWAS data to identify molecular phenotypes associated with the trait and then to analyze pairwise xQTL data sets to identify associations between molecular phenotypes (e.g., associations between gene expression and DNA methylation) [40]. However, the number of tests required for such an analysis increases quadratically with the increase in the number of 'omics layers. When considering all potential combinations of 'omics layers, this results in an exponential increase, not only elevating computing costs, but, more importantly, also exacerbating the multiple testing burden. One solution is to fit the GWAS and multi-xQTL data in a joint model, as in the methods described in the following sections.

Multiple-Trait COLOC (MOLOC) is an extension of COLOC that can integrate GWAS summary data with xQTL data from multiple 'omics layers simultaneously to uncover loci where the GWAS signal is colocalized with one or more xQTL signals due to a shared causal variant [64]. Analyses with schizophrenia GWAS and brain eQTL and mQTL data demonstrated that MOLOC using both xQTL data sets resulted in a 1.5-fold increase in gene discovery, compared with using only eQTL data [64]. However, when analyzing more than four phenotypes, MOLOC becomes computationally impractical due to the exponential growth in the number of causal configurations with each additional phenotype. This computational hurdle can be circumvented by computing the posterior probability of colocalization using an approximate approach, as in Hypothesis Prioritization for Multi-Trait Colocalization (HyPrColoc) [65], which has enabled simultaneous colocalization analysis of a vast number of complex traits. However, when applying HyPrColoc to multi-omics xQTL data, it is necessary to consider more complicated situations, such as diverse coverages of molecular phenotypes across the genome and multiple sites per molecule within a locus [66]. Primo is another method of this kind, but it accommodates scenarios where multiple causal variants exist in a locus [67]. By re-estimating the effects of GWAS variants conditional on other lead xQTL variants, Primo mitigates spurious associations due to LD, albeit at the cost of reduced statistical power. Omics PIEiotRopic Association (OPERA) is an extension of SMR under the Bayesian framework, which enables joint analysis of GWAS summary statistics and xQTL summary statistics from multiple 'omics layers [66]. It can effectively control the false discovery rate (FDR) while maintaining a high detection power for various patterns of associations, both between molecular phenotypes and the trait of interest, and among molecular phenotypes. As with the HEIDI test in the SMR analysis, a multi-exposure HEIDI test has been developed and included as part of the OPERA analysis to filter out associations due to the linkage model. Applying OPERA to summary-level GWAS data for 50 complex traits and xQTL data from seven 'omics layers revealed that 51% of the GWAS signals were shared with at least one xQTL, approximately half of which were not eQTLs.

Beyond the integrative analysis approaches based on conventional statistical frameworks, machine-learning approaches, such as Locus to Gene (L2G) [68], have also been developed

for gene prioritization. The L2G model was trained using 445 gold-standard genes, along with a large collection of fine-mapped genetics and functional genomics data, including transcriptomic, proteomic, and epigenomic data, as well as disease-molecular trait colocalization analysis results across 92 cell types and tissues. The L2G score generated by the model represents the likelihood of a gene being causal for a specific trait. However, a primary limitation of this approach is the requirement for a sizable set of high-quality gold-standard genes for effective model training, and each gold-standard source could introduce biases.

Linking fine-mapped GWAS variants to their target genes

Recent advancements in statistical fine-mapping techniques have substantially improved our ability to identify putative causal variants [8,69–73]. However, it remains challenging to establish direct links between fine-mapped variants and their target genes, because most variants identified through GWAS are noncoding and do not necessarily regulate the nearest gene [9]. For instance, variants in enhancer regions, which regulate gene expression by interacting with the transcription machinery at the promoter, can be found at various distances from the target gene [74,75]. Approximately 77% of blood trait GWAS loci contain at least one fine-mapped variant that overlaps with an enhancer region [76]. Hence, developing high-precision enhancer–promoter interaction maps can facilitate gene prioritization [74] (Figure 2B).

Chromosome conformation capture techniques

Mapping spatial contacts between chromatins has the potential to reveal the target genes of candidate *cis*-regulatory elements (cCREs). This is possible because chromatin fibers fold into higher-order structures, allowing distant DNA fragments to come into proximity in 3D space. High-throughput detection of chromatin interactions has been achieved by chromosome conformation capture (3C) techniques, such as Hi-C [77] and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) [78]. Promoter capture Hi-C (PCHi-C) [79] is a specialized Hi-C method designed for identifying interactions involving promoters throughout the genome. It has been used to map cCREs to their target genes in various tissue types and cell lines. Notable examples include studies conducted by Jung *et al.* [80] and Javierre *et al.* [81], which utilized PCHi-C to map the interactomes involving ~18 000 promoters in 27 different human tissues and cell types, and ~30 000 promoters in 17 hematopoietic cell types, respectively. Jung *et al.* further used the chromatin interaction data to infer target genes for 27 325 noncoding variants associated with 2117 physiological traits and diseases [80]. However, all the 3C-based methods operate on the assumption that spatial proximity reflects functional interaction, an assumption that may not always hold. Additionally, the relatively low resolution of 3C techniques, typically several kilobases, might limit their precision in associating individual cCREs with specific target genes or in distinguishing associations between proximal cCREs and gene promoters [80]. Moreover, these techniques have only been applied to a limited range of human tissues and cell types, partly due to the high costs associated with them.

Inferring chromatin interactions through associations between epigenomic features

Epigenomic features, such as chromatin accessibility and histone modifications, can be harnessed to identify cCRE–gene links by assessing the correlation of chromatin accessibility or activity between pairs of cCREs. One such approach, Cicero, uses a graphical Least Absolute Shrinkage and Selection Operator (LASSO) model to identify co-accessible pairs of cCREs across bins of related cells within a cell type [82], and has been widely used to associate cCREs with putative target genes in specific cell types [83,84]. For example, in brain cell types, Cicero identified 2.82 million co-accessible links between cCREs using single cell chromatin accessibility data, with ~20% confirmed as physical interactions [84].

While promoter accessibility is often used as a proxy for gene activity in co-accessibility analysis, it may not always accurately reflect gene activity due to the complexity of gene regulation. To address this issue, a new set of methods has been developed to identify 'co-activity' between the activity of cCREs and gene expression [85,86]. One notable effort is the Epigenome Integration across Multiple Annotation Projects (EpiMap) [75], which incorporates data generated by multiple assays, including RNA-sequencing (RNA-seq), ChIP-sequencing (ChIP-seq), and DNase I hypersensitivity sequencing (DNase-seq), in various tissues and cell types to predict enhancer–gene links. The prediction is based on the Pearson correlation between gene expression and the histone mark activity of nearby enhancers within 1 Mb. An XGBoost classifier was trained on the positive set of valid links against their paired negative links, using precomputed correlations and the distance to the transcription start site (TSS) as features. EpiMap has predicted 3.3 million tissue-specific enhancer–gene links, with each gene associated with an average of 13 enhancers and each enhancer linked to ~1.5 genes, typically at a median distance of 42 359 bp. However, the correlation analysis approaches have limitations. They can identify associations that do not necessarily reflect direct *cis*-regulatory interactions, but instead indicate associations driven by confounding factors. Furthermore, these analyses often rely on individual-level data, which are typically available only in small samples.

The aforementioned integrative methods can also be harnessed to detect associations between epigenomic features due to shared genetic factors, using only xQTL summary data. One such example is the application of the SMR & HEIDI method [20] to mQTL data for predicting chromatin interactions [87], where CpG sites in a gene promoter are used as hooks to detect their pleiotropic associations with other CpG sites. Integrating the predicted chromatin interactions with GWAS summary data highlights links among CpG sites and genes associated with complex traits. This approach offers several advantages. Unlike experimental assays, such as Hi-C and PCHi-C, this approach is cost-effective, because it reuses data from experiments not originally designed for this purpose. Furthermore, in contrast to correlation analysis methods, it utilizes a genetic model to perform an MR analysis, meaning that the detected associations are unlikely to be confounded by nongenetic factors.

Activity-by-contact model

Even when an enhancer is inactive, it can still maintain close physical contact with the promoter of a gene [88], indicating that such contact, while critical, is only part of the mechanism for understanding their functional interaction. This led to the development of the activity-by-contact (ABC) model to link enhancers to gene promoters [89], assuming that the regulatory effect of an enhancer on a gene depends on both the activity of the enhancer and its physical contact with the promoter of the gene. The ABC score for an enhancer–gene pair is calculated by the product of the activity of the enhancer (measured using DNase-seq and H3K27ac ChIP-seq data) and its contact frequency with the promoter of a gene (measured using chromatin interaction assays, such as PCHi-C), which is then normalized by the sum of such products for all elements within a specified genomic distance from the gene, typically within 5 Mb. This method was shown to be effective in predicting enhancer–gene relationships in a cell type-specific manner, offering a valuable tool for understanding the functional implications of genetic variants. For instance, in an analysis involving 72 complex traits, the ABC model linked 5036 fine-mapped GWAS signals to 2249 unique genes, including 577 genes that appear to influence multiple traits through variants in enhancers that function in different cell types [90]. The ABC model requires high-quality chromatin interaction data to quantify contact frequency. When such data are unavailable, the physical distance between the enhancer and promoter can serve as a proxy [90].

CRISPRi-based enhancer screening

Enhancer–gene links can be experimentally assessed through clustered regularly interspaced short palindromic repeats (CRISPR) screening. Using CRISPR interference (CRISPRi) techniques, it is possible to simultaneously perturb the activity of numerous regulatory elements, and examine the subsequent impact on gene expression in a high-throughput manner. One typical example is the multiplex, eQTL-inspired framework introduced by Gasperini *et al.* [91]. This approach uses random combinations of CRISPRi/dCas9-mediated perturbations in a multitude of cells, followed by single-cell (sc)RNA-seq. By utilizing dCas9-KRAB, 5920 candidate enhancers in K562 cell lines were perturbed, and the effects of these perturbations were measured by profiling 254 974 single cell transcriptomes, resulting in the identification of 470 high-confidence enhancer–gene pairs. Similarly, Morris *et al.* focused on characterizing the functional effects of GWAS variants [76]. They used CRE-silencing CRISPRi perturbations to inhibit cCREs, derived from fine-mapped variants for blood traits, in the human erythroid progenitor cell line K562. By targeting 543 variants across 254 loci and generating comprehensive scRNA-seq data, they identified *cis*-target genes for 134 cCREs, most of which are the closest gene. The integration of CRISPRi-based enhancer screening with scRNA-seq facilitates accurate identification of putative causal genes, even though these experiments are technically complex, involving multiple steps, such as guide RNA design, cloning, transfection or transduction, cell sorting, and scRNA-seq. Additionally, while these methods are feasible for immortalized cell lines, it is crucial to extend them to other cell lines and primary cells for the next stage of target gene identification and characterization for diverse complex traits.

Network-based gene prioritization

Genes or proteins that interact with each other often participate in similar cellular functions and contribute to related organismal traits. By leveraging the principle of guilt-by-association, molecular networks have proven useful in predicting the function or trait relevance of human genes. In the context of GWAS, these networks can enhance gene discovery by using genes identified from post-GWAS analyses as seeds to identify additional trait-associated genes. Various algorithms have been developed for this purpose, including similarity-based and propagation-based algorithms. These algorithms can be integrated with diverse types of network, such as protein–protein interaction (PPI), pathway and gene co-expression networks.

Similarity-based methods

Similarity-based algorithms are commonly used to nominate genes with similar functions or those within the same pathways. One such method is NetWAS 2.0 [92,93]. In this approach, significant genes identified from gene-based association analysis (i.e., VEGAS) are considered positive cases, while randomly sampled nonsignificant genes serve as negative cases. The weights of labeled genes to all genes in the functional networks across various human tissues and cell types are used as features. NetWAS 2.0 then trains support vector machine (SVM) classifiers to enhance the identification of gene–trait associations. When applied to Alzheimer's disease (AD), NetWAS 2.0 successfully identified genes associated with axon plasticity that connect amyloid-beta (Aβ), aging, and tau protein in susceptible neurons, by leveraging functional networks specific to seven types of human neuron. Data-driven Expression Prioritized Integration for Complex Traits (DEPICT) is another versatile method for gene prioritization, gene set enrichment, and tissue enrichment [94]. It utilizes 14 462 reconstituted gene sets derived from 77 840 publicly available expression microarrays. Each gene set contains z-scores representing the membership strength of each gene within the set. The gene prioritization algorithm in DEPICT involves two steps. First, it identifies genes (*g*) in the trait-associated loci, defined as genes overlapping variants with LD r^2 >0.5 to a lead variant. Next, it assesses the correlation of each gene with the remaining genes in *g* across the reconstituted gene sets. Genes with stronger overall correlations within *g* are prioritized more

highly. Another recently developed similarity-based method, Polygenic Priority Score (PoPS), utilizes a ridge regression framework to prioritize genes from GWAS [95]. It utilizes gene-level associations computed from GWAS summary statistics using a gene-based test tool (i.e., MAGMA) to learn joint polygenic enrichments of gene features derived from cell type-specific gene expression, pathways, and PPIs. To nominate putative causal genes, PoPS assigns a priority score to every protein-coding gene according to these enrichments.

Propagation-based methods

Gene networks can be represented and analyzed as graphs, denoted as $G = (V, E)$, where nodes V represent genes or proteins and edges E represent interactions between genes or proteins. Prioritization methods based on network propagation utilize the network topology to capture the flow of influence or information through the network when ranking genes. Initially, a set of known trait-associated genes obtained through post-GWAS analyses serves as the seed genes to discover additional genes that are closely connected to these seed genes (Figure 3D). One widely used propagation-based algorithm is the Random Walk with Restart (RWR) [96]. It calculates the steady probabilities of each gene or protein being visited based on the network proximity, with a higher probability indicating a closer similarity to the seed gene and a stronger trait association. For example, Priority index (Pi) is a scoring system for drug target discovery that uses the RWR algorithm [97]. Specifically, Pi begins by identifying seed genes prioritized from GWAS based on genomic predictors, such as physical distance, chromatin conformation, and eQTLs. The RWR algorithm is then used to identify non-seed genes based on their PPI network connectivity to seed genes. This approach has identified both existing therapeutic targets and unexplored potential targets for 30 immune-related traits. Another method, the integrative risk gene selector (iRIGS), integrates multi-omics data and RWR in a Bayesian framework [98] to identify plausible trait-associated genes. As an extension of the RWR, Personalized PageRank (PPR) allows for the incorporation of personalized probabilities to a specific set of nodes. This increased flexibility has led to the widespread application of PPR in network-based gene prioritization. For instance, Barrio-Hernandez *et al.* utilized PPR to augment a pool of risk genes prioritized from GWAS for 1002 traits. This algorithm effectively recapitulated known disease genes or drug targets and identified groups of genes connected to the trait-associated genes but exhibiting weak GWAS signals [99].

In summary, network-based gene prioritization methods facilitate the discovery of trait-associated genes that may be masked due to weak GWAS signals. Despite their advantages, these methods have some limitations. First, network-based methods rely on the 'guilt-by-association' principle, which means they may fail to uncover trait-associated genes that lie outside currently known functional networks [81]. Second, algorithms such as RWR tend to prioritize genes that are closely connected to the seed genes, often resulting in the prioritization of hub genes involved in multiple networks, even though they may not be causal genes. Third, existing gene networks are primarily derived from limited tissues and cell types, necessitating the development of context-specific networks. Lastly, certain approaches solely focus on the network topology, neglecting the integration of other modalities.

Integration of GWAS with cellular xQTL data

While we have discussed the integration of xQTLs from one or multiple tissues for gene prioritization, most currently available xQTL data primarily reflect the genetic control of a molecular phenotype in the entire tissue. If an xQTL effect is highly dependent on a specific cell type or cell state (i.e., context-dependent xQTL or cd-xQTL), it could remain undetected when using bulk tissue data, especially when the sample size is not sufficiently large. Therefore, the extent to which cd-xQTLs can resolve the GWAS loci where no gene is prioritized using bulk xQTL data remains

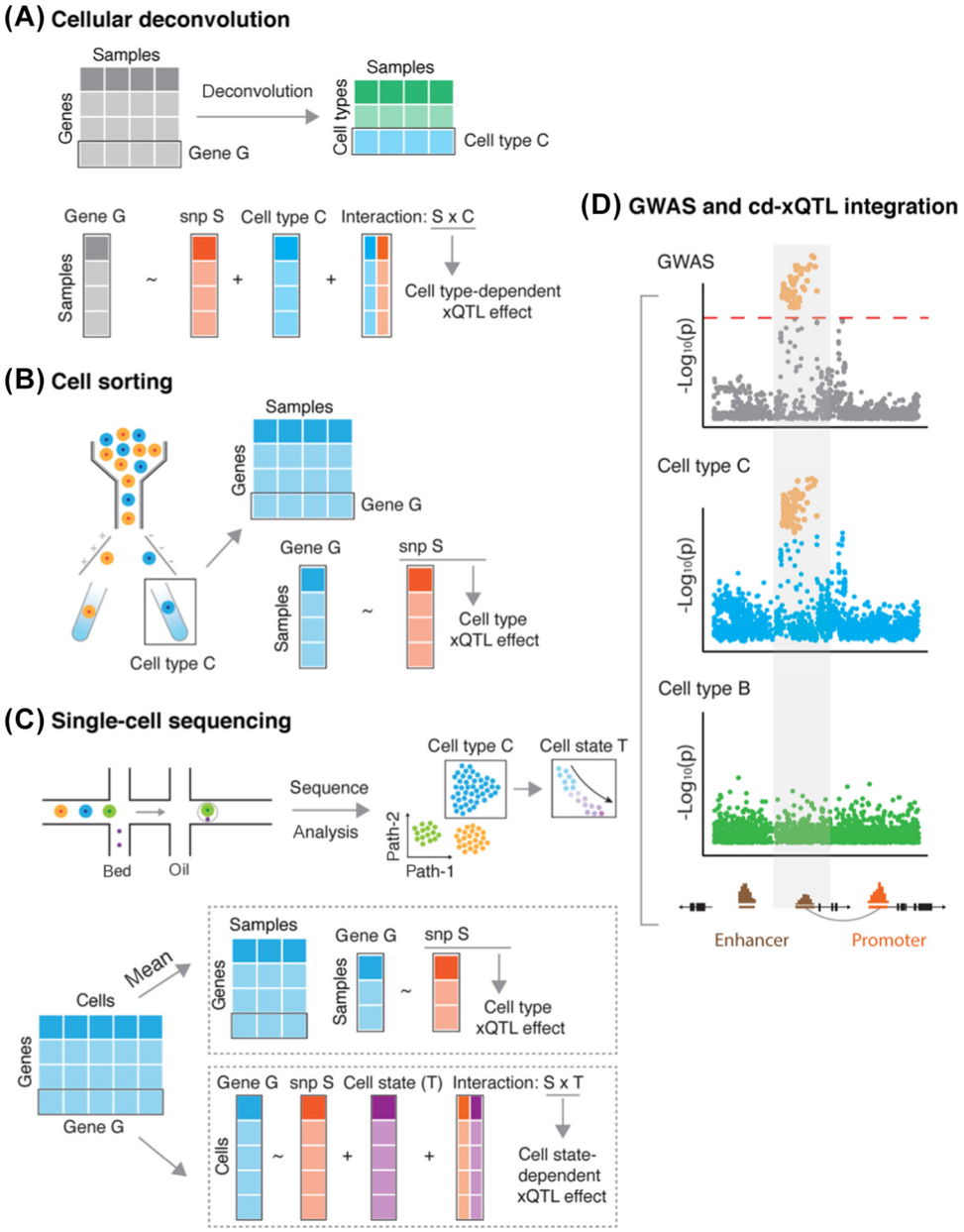

Figure 3. Overview of cellular molecular quantitative trait loci (xQTL) mapping strategies. Abbreviations: cd, context dependent; GWAS, genome-wide association study; xQTL, molecular quantitative trait loci.

unclear. The rapid progress in single cell technologies has catalyzed extensive efforts to decipher the genetic regulations governing molecular phenotypes within diverse cellular contexts, spanning specific cell types to distinct states. This evolving landscape has given rise to a new avenue of investigation, focusing on uncovering xQTLs at the cellular level. By integrating these cellular-level xQTLs with GWAS data, an opportunity arises to unveil gene–trait associations that may remain concealed in analyses conducted using bulk tissue data [100–103]. In this section, we provide an overview of the cellular xQTL mapping strategies accompanied by currently available cellular xQTL studies and

summarize the main findings in gene prioritization when integrating GWAS with cellular xQTL data (Figure 3).

Mapping cd-xQTLs using cellular deconvolution

Cellular deconvolution methods provide a cost-effective solution for estimating the abundances of specific cell types or states from bulk tissue data [104–106]. These methods enable the mapping of cd-xQTLs by examining the interaction between genotype and cell type abundance [107–109] (Figure 3A). Applying this cd-xQTL mapping approach to cohort-level bulk tissue data led to the identification of thousands of cd-eQTLs and cd-sQTLs [103]. The credibility of these findings was reinforced by functional enrichment analyses, demonstrating that the lead cd-xQTLs are enriched in cell type- or state-specific open chromatin regions [104,110]. The integration of cd-xQTLs into GWAS has led to the discovery of novel colocalized loci. For instance, rs4292, a genetic variant located within a cell type-specific open chromatin region unique to proximal tubule (PT) cells, was identified as an eQTL for the *ACE* gene. The effect size of this cd-eQTL was found to be dependent on the proportion of PT cells and displayed a colocalization pattern with a GWAS signal of systolic blood pressure [110]. In addition to prioritizing genes through colocalizing GWAS signals with cd-eQTLs, alternative approaches, such as MiXcan and CONTENT, directly assess the associations between context-dependent molecular phenotypes (cd-MPs) and complex traits under the TWAS framework [111]. MiXcan utilizes the elastic net model to predict cell type-dependent gene expression levels by treating the estimated cell type abundances as priors. These predictions are then associated with phenotypes, and the association *P* values of genes from various cell types are integrated using the Cauchy combination test. CONTENT [112] decomposes gene expression profiles into context-specific and context-shared components and then constructs genetic predictors for each component. These predictors can be associated with phenotypes individually or in a combined manner. Unlike conventional TWAS, which tests associations averaged across cellular contexts, these context-aware TWAS methods accumulate association signals from different cellular contexts, thereby improving statistical power. Despite the success in prioritizing genes using cd-xQTLs or cd-MPs estimated from bulk tissue data, there remain concerns regarding the biological interpretability of these findings. One prominent concern is the possibility that cd-xQTLs or cd-MPs obtained through the deconvolution (decomposition)-based strategy may not be specific to the focal contexts but rather inflated by other confounding factors.

Mapping cd-xQTLs using sorted cell populations

Compared with the deconvolution-based strategy, mapping cd-xQTLs from sorted cell populations could yield more precise results, thereby enhancing the biological interpretability of the findings (Figure 3B). Several studies have performed cd-xQTL mapping using 'omics data measured from sorted cells [113–119]. For example, Ota *et al.* [115] conducted a large-scale analysis of eQTLs ($N = 416$) across 28 immune cell types, demonstrating that genetic variants can influence disease susceptibility through specific cell types. The study highlighted the variant rs62266700, which is in high LD ($r^2 = 0.85$) with the systemic lupus erythematosus (SLE) GWAS variant rs36059542. This variant only exhibits an eQTL effect on the *ARHGAP3* gene in plasmablasts among the 28 immune cell types, and is located within chromatin regions that are only open in plasmablasts. This example underscores the significance of accounting for cell-type heterogeneity in eQTL mapping, because only the relevant cell type can manifest QTL effects on specific genes, enabling cell type-specific gene prioritization. In addition to eQTLs, Zeng *et al.* performed a large-scale ($N = 616$) caQTL analysis of neurons and microglia across four brain regions [114] and found that only 10.4% of caQTLs are shared between neurons and glia. Beyond the cell type level, cell state-dependent xQTL mapping is also achievable through the use of sorted cells. Strober *et al.* conducted time-series RNA-seq on human cell lines undergoing differentiation

from induced pluripotent stem cells (iPSCs) to cardiomyocytes [113]. They identified hundreds of dynamic eQTLs that only exhibit regulatory effects on gene expression at specific stages during iPSC differentiation. These transient genetic effects provide valuable biological insights into certain GWAS loci. For instance, the variant rs28818910, which manifests an eQTL effect on the gene *C15orf39* exclusively at intermediate stages of differentiation, is associated with body mass index. Although promising, there are several limitations in mapping cd-xQTLs using the cell purification strategy. First, the isolated cell populations may still be contaminated with other types of cell. Second, the selection of cells may be biased toward specific, well-established cell types or states. Finally, due to the high costs and low throughputs associated with this type of technology, sample sizes are often small, which limits the discovery power.

Mapping cd-xQTLs using single cell sequencing

Single cell sequencing (sc-seq) or single nucleus sequencing (sn-seq) provides a powerful and unbiased strategy for profiling cellular phenotypes, which has addressed several limitations in the strategies based on bulk or sorted cell samples mentioned in the preceding text [101]. With the continuous decrease in the cost of sc-seq and the advancement of sample multiplexing techniques, it becomes increasingly feasible to generate sc-seq (or sn-seq) data in population-level cohorts for cd-xQTL mapping [120].

Most single cell-based xQTL studies have relied on methods originally developed for bulk xQTL mapping. These methods assume that the molecular phenotype across all individuals follows a normal distribution and that there is only one observation of each molecular phenotype for each individual. However, these assumptions do not necessarily hold true for sc-seq data due to their sparsity and the presence of multiple cells for each molecular phenotype for each individual. To address this disparity, many studies have utilized the pseudo-bulk approach, which aggregates cells of the same cell type or state within an individual to obtain cell set-specific molecular phenotypes. Due to the high technical noise in sc-seq data, optimizing the pseudo-bulk generation approach is critical for improving the robustness and statistical power. Leveraging sample-matched bulk and scRNA-seq data, Cuomo *et al.* benchmarked different approaches for identifying eQTLs from scRNA-seq data [121] and found that normalizing the data at the cellular level and aggregating cells at the donor-run level (i.e., merging cells not only for each donor, but also for each sequencing run) achieved the highest statistical power and the lowest false positive rate [121]. The pseudo-bulk approach has been applied to map cd-xQTLs across different tissues [100,102,122–125]. For example, Bryois *et al.* performed snRNA-seq in brain tissues and identified eQTLs across eight brain cell types with a sample size of 192 [123]. By integrating these eQTLs with GWAS of brain disorders, they found that rs10792832, which is associated with AD, only functions as an eQTL for *PICALM* in microglia. Yazar *et al.* conducted the largest single cell-based eQTL study to date by performing scRNA-seq in peripheral blood mononuclear cells (PBMCs) from 982 donors [100]. They identified thousands of eQTLs that are active only in specific cell types. By leveraging a colocalization method, they found that 19% of these cell type-specific *cis*-eQTLs might share the same causal variants with GWAS loci of seven autoimmune diseases. In addition to mapping eQTLs in cell types, some studies have also stimulated cells *in vitro* to identify eQTLs that function in specific cell states arising from stimulations [125–132]. For example, Soskic *et al.* mapped eQTLs at different activation stages of CD4+ T cells [126] and identified 127 genes for which eQTLs in the activated T cells colocalized with GWAS loci of immune diseases. These genes were significantly enriched in the gene set exhibiting cell-state dynamic regulations.

Another approach for mapping cd-eQTLs is treating each cell as a unique observation, offering a flexible framework for modeling the continuous state of individual cells [133,134]. The

linear mixed model (LMM) is used to account for correlations among cells from the same donor or sequencing run. Conventional LMM applications often involve log or rank-based inverse normal transformation of phenotypes. The transformed phenotypes are assumed to follow a normal distribution. However, due to the sparsity of single cell data, such transformations may lead to elevated FDRs, as demonstrated by the benchmark analyses from Cuomo *et al.* and Nathan *et al.* [121,134]. To address this issue, Nathan *et al.* proposed the use of the generalized linear mixed model (GLMM) based on the Poisson distribution to fit the count data of cells [134]. Using scRNA-seq data in T cells, Nathan *et al.* utilized canonical correlation analysis (CCA) to obtain the top 20 canonical variates as representations of distinct states associated with T cell cytotoxicity, regulatory functions, and other characteristics. By testing the interaction term between the canonical variates and genotypes of individual cells, they found that approximately one-third of the 6511 eQTLs, identified using the pseudo-bulk approach, exhibited cell state-dependent effects. Furthermore, they discovered that genetic variants associated with autoimmune diseases were significantly enriched in these cell state-dependent eQTLs.

In summary, cellular deconvolution methods offer a cost-effective solution for mapping cd-xQTLs in conventional xQTL mapping cohorts with genotype and bulk molecular phenotype data. By contrast, cell sorting and single cell technologies provide a more precise approach to map xQTLs at higher resolution. While the integration of GWAS with cd-xQTLs opens a promising avenue to uncover trait-associated genes that may be obscured in bulk data due to the heterogeneity of genetic effects across different cell populations [100–103], there are several limitations that need to be addressed. First, small sample sizes of cd-xQTL studies, often due to financial constraints, limit the statistical power and generalizability of findings. Second, cd-xQTLs in solid tissues are less studied because cohort-level scRNA-seq or snRNA-seq data from those tissues are rarely available. Third, there is a need for further exploration of diverse molecular phenotypes beyond gene expression, such as chromatin accessibility, epigenetic modifications, and RNA splicing, to gain a more comprehensive understanding of the genetic regulations under various cellular contexts. Lastly, mapping cd-xQTLs during different developmental stages requires further investigation to capture the dynamics of genetic regulation throughout ontogeny.

Potential of gene prioritization in drug repurposing

Recent advancements in drug development, as exemplified by the development of evolocumab, informed by *PCSK9* gain-of-function mutations in familial hypercholesterolemia [135–137], and romosozumab, inspired by *SOST* loss-of-function mutations in sclerosteosis [138,139], have demonstrated the potential of human genetics to propel the development of new therapeutics. This is particularly true when genetic effects closely mimic pharmacological interventions [140], a notion that is steadily gaining empirical support. For example, Finan *et al.* [141] showed that genomic regions associated with complex traits are enriched with approved drug targets. Moreover, retrospective analyses revealed that drugs targeting genes supported by human genetics are approximately twice as likely to be approved for clinical indications [142,143]. A recent study also indicates that 33 out of 50 drugs approved by the FDA in 2021 were supported by human genetic evidence [144]. These findings enhance the connection between drugs and diseases, opening potential clinical opportunities by uncovering disease-associated genes that could be targeted for therapeutic interventions (Figure 4A). We have discussed in the preceding text the methods for prioritizing genes underlying GWAS signals. However, there remains a challenge regarding how these genes can facilitate drug discovery or repurposing. In this section, we explore various approaches that leverage the insights gained from GWAS to identify potential avenues for drug repurposing.

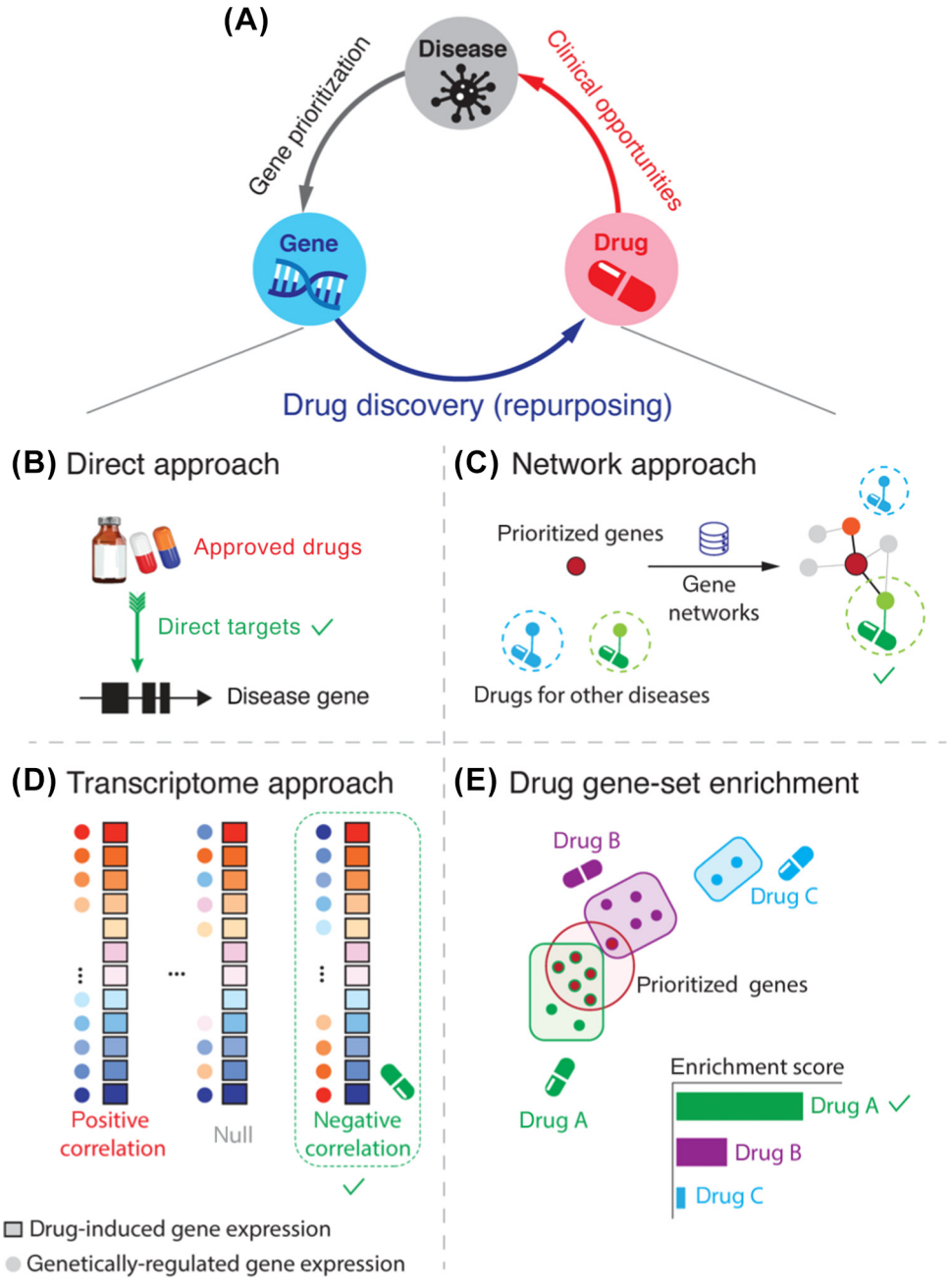

Figure 4. Overview of strategies for drug repurposing utilizing genetic data.

Genetics-informed drug repurposing

An intuitive approach for genetics-informed drug repurposing is to collect disease-associated genes from post-GWAS analyses and evaluate whether existing drugs target these genes (Figure 4B). For example, a putative disruptive missense variant rs11209026 in *IL23R* exhibits a protective effect against Crohn's disease. This insight led to the discovery of IL23-targeted therapies for Crohn's disease, such as risankizumab and ustekinumab, initially developed for

psoriasis treatment [145]. Similarly, prioritizing genes from GWAS loci through multiple sources of evidence, including missense variants, *cis*-eQTLs, and PPIs, for rheumatoid arthritis (RA) revealed that certain drugs, such as flavopiridol and alvocidib, which target the *CDK4/6* genes and were initially approved for cancer treatment, could be repurposed for RA treatment [146].

Randomized controlled trials (RCTs) are considered the gold standard for examining causal effects in medical treatments, and MR is often referred to as a 'natural' RCT that is not limited by ethical, feasibility, and adherence considerations [50,147–149]. MR-based drug repurposing operates on the assumption that the effect of a genetic variant on a disease could mimic the lifelong activation or inhibition of a drug target. For example, MR analysis showed that overexpression of *PCSK9* is a risk factor for hypercholesterolemia, which aligns with the therapeutic action of the approved PCSK9 inhibitor, alirocumab (a monoclonal antibody of PCSK9) [50]. Beyond its potential in identifying drug targets, MR can also assist clinicians in making more informed prescription decisions and uncover potential side effects associated with drug interventions. For example, given the high prevalence of hypertension among patients with psychiatric disorders, Chauquet *et al.* [150] used SMR to demonstrate that reduced *ACE* gene expression level is associated with both decreased systolic blood pressure and increased schizophrenia risk, suggesting that antihypertensive medication (i.e., ACE inhibitors) might increase the risk of schizophrenia. Similarly, MR analyses have associated statins, cholesterol-lowering drugs, with a slightly increased risk of type 2 diabetes mellitus [151]. Compared with traditional drug target discovery, which tends to be lengthy and time-consuming, MR can be used to swiftly and effectively identify potential targets for emerging infectious diseases or pandemics. One prime example is the rapid discovery of a gene target for COVID-19 using SMR. It was found that the expression level of *TYK2* is significantly associated with COVID-19 [152]. Not long after the identification of *TYK2* as a risk gene, the JAK inhibitor baricitinib, which targets the enzyme encoded by *TYK2* and was initially developed for treating RA, received emergency use authorization from the FDA for the treatment of COVID-19 [153].

Transcriptome-based drug repurposing

Transcriptomic data can also be leveraged for drug repurposing based on the signature reversion principle [154–156]. The fundamental assumption is that compounds exhibiting reverse effects on gene expression profiles, characterized by a negative correlation between drug-induced gene expression profile and genetically regulated expression signature of a disease, could serve as prospective drug candidates for disease treatment (Figure 4D). The Connectivity Map (CMap) and the Library of Integrated Network-based Cellular Signatures project (LINCS) L1000 library [157,158] are two widely used databases containing extensive transcriptomic profiles of cell lines treated with numerous genetic and pharmacological perturbagens. CMap comprises ~7000 gene expression profiles from ~1300 compounds on five cancer cell lines, while LINCS L1000 contains ~1 300 000 gene expression profiles from ~42 000 genetic and small-molecular perturbations on ~90 cell lines. These valuable databases have facilitated transcriptome-based drug repurposing. For instance, So *et al.* [154] utilized Spearman and Pearson correlation, as well as Kolmogorov–Smirnov (KS) tests, to compare the transcriptome profiles derived from TWAS analysis against drug-induced gene expression profiles from CMap. Through this analysis, they identified several nonsteroidal anti-inflammatory drugs (NSAIDs; i.e., cyclooxygenase inhibitors), which may have therapeutic potential for AD. Similarly, the Trans-Phar (integration of TWAS and pharmacological database) pipeline performs *in silico* screening of compounds from the LINCS L1000 library through Spearman correlation analysis [156]. This pipeline, when applied to 29 GTEx tissues and 77 LINCS L1000 cell types across 13 distinct categories, identified several promising compounds, including anisomycin for schizophrenia and verapamil for patients hospitalized with COVID-19. Despite the success of

transcriptome-based drug repurposing, one major limitation of this approach is that the identified drug candidates often lack well-defined biological mechanisms of action.

Network-based drug repurposing

We have discussed the value of network-based gene prioritization in the preceding text. A recent study showed that network diffusion is also beneficial in identifying drug-target genes with weak genetic support [159]. Another study suggests that genes identified through GWAS are closely connected to drug-target genes within biological networks [160]. These observations provide possibilities for drug repurposing by examining the proximity of genes identified through GWAS to drug targets in human interactome networks (Figure 4C). One notable example is the Genome-wide Positioning Systems network (GPSnet) [161], which uses RWR to identify disease–gene modules by integrating somatic mutations and transcriptome profiles with human PPI networks. It then tests whether these modules are enriched in drug-induced up- and downregulated gene sets sourced from CMap. Simultaneously, by integrating disease–gene modules with drug–gene networks, network proximity analysis is performed to identify significant associations between drugs and diseases. As a result, GPSnet is capable of not only prioritizing disease-gene modules with high druggable potential, but also identifying new indications for approved drugs. Network information can also be leveraged based on the signature reversion principle, similar to transcriptome-based drug repurposing. For example, Pathway Signatures for Drug Repositioning (PS4DR) [162] is a multimodal and integrative workflow that combines GWAS data, transcriptomic signatures, and pathways to prioritize drugs predicted to reverse disease pathway dysregulations. The primary limitation of network-based drug repurposing is the incompleteness of human interactome network data, particularly in specific tissues and cell types.

Gene set enrichment-based drug repurposing

Databases, such as DrugBank [163], TTD [164], ChEMBL [165], and DGIdb [166], provide information on the relationships between gene sets and drugs, thereby facilitating examination of potential associations between drug sets and disease phenotypes (Figure 4E). One method that enables this examination is Genome for REPositioning drugs (GREP) [167], which uses Fisher's exact test to assess whether genes prioritized from GWAS are enriched in genes targeted by drugs in clinical indication categories, such as Anatomical Therapeutic Chemical (ATC) and International Classification of Diseases 10 (ICD10). Another web platform, Drug Targetor [168], utilizes bipartite drug–gene connections to define gene sets for each drug. It then applies MAGMA gene-set analysis to calculate a genetics-informed drug score, assessing whether a drug–gene set is more associated with a disease compared with other sets. Similarly, Bell *et al.* [169] used this approach, supplementing it with a multiple linear regression model, to determine whether drugs within a specific group exhibit a stronger connection to disease-associated genes compared with others. The Bell *et al.* approach allowed for exploration of both individual drugs and groups of drugs categorized by ATC III code, mechanism of action, and clinical indication. Overall, gene-set enrichment analysis is capable of identifying clinically relevant drugs that target the gene set associated with a specific disease, thereby offering a pool of potential drug candidates for repurposing. However, it primarily uncovers indirect associations inferred from drug–gene interactions, rather than directly linking the pharmacological action of a drug to a disease. Additionally, it does not indicate the direction (i.e., beneficial or adverse) of the drug effect on disease. Finally, it should be acknowledged that the enrichment results might be influenced by the over-representation of drug targets within specific categories, such as G protein-coupled receptors, ion channels, nuclear receptors, enzymes, transporters, and immune checkpoint proteins.

In summary, GWAS and post-GWAS analyses have provided valuable opportunities to identify potential therapeutic targets for drug discovery and repurposing. However, this area has certain

complexities that require careful consideration and attention. First, most analyses discussed in the preceding text heavily rely on historical drugs and their targets, which might bias the prioritization of drug candidates. Another complexity in translating genetics discoveries to drugs is that the magnitude of genetic effect size does not necessarily equal drug effect size [170]. For example, although the genetic effects of the *HMGCR* variants, rs17238484 and rs12916, on low-density lipoprotein (LDL) cholesterol are much smaller than those of the loss-of-function mutations in *PCSK9*, this does not hinder *HMGCR* from being an effective therapeutic target of stains for the treatment of hypercholesterolaemia [136,140,171]. In fact, the drug effects of rosuvastatin (targeting *HMGCR*) and alirocumab (targeting *PCSK9*) on LDL cholesterol are comparable [172–174], suggesting that genetic variants with small effect sizes do not necessarily imply low drug efficacy. In addition, gene associations that link drug target genes to unintended phenotypes indicate the potential risk of adverse events in specific organ systems [175,176]. By using human genetics data to identify potential drug targets and safety liabilities, we can better predict the effects of lifelong modulation of therapeutic targets and anticipate the risk for on-target and off-target adverse events [177].

Concluding remarks

GWASs have propelled the development of various methods for gene prioritization, creating new avenues for the identification of therapeutic targets. Despite the varying performances of different methods and tools, the wealth of available data enables the formulation of hypotheses on disease mechanisms. However, despite the usefulness of the tools and workflows outlined in this review, no universal tool exists for pinpointing causal genes and mechanisms. Therefore, there is a strong need for continued development of user-friendly tools, leveraging new data and novel insights (see Outstanding questions). Moreover, for benchmark studies comparing different methods, the set of genes currently identified as 'causal' is likely to be strongly biased toward those nearest to the GWAS peak for several reasons. For example, genes with high-probability coding variants are more readily identified as causal and generally correspond to the nearest gene. Additionally, genes near the center of the GWAS peak are more likely to be investigated and accumulate evidence for being causal. Thus, a comprehensive set of high-confidence causal genes is in high demand for benchmark studies. Furthermore, functional validation, both *in vitro* and *in vivo*, remains essential for establishing causal links between genes and diseases.

Gene prioritization has greatly benefited from the extensive resource of *cis*-xQTL data. By contrast, *trans*-xQTLs, such as *trans*-eQTLs, which are estimated to account for ~70% of the heritability in mRNA expression [178], hold immense potential in advancing our understanding of distal genetic regulation. Nevertheless, the substantial number of association tests required for genome-wide *trans*-xQTL mapping presents a considerable challenge, necessitating the development of statistical methods to deal with the multiple testing problem and computational strategies to manage the resulting data deluge. In addition, while GWAS have identified numerous loci associated with complex traits in various ancestries, most samples used for large-scale xQTL studies are of European ancestry, resulting in a lack of xQTL data for other ancestries [179]. This limitation could restrict the generalizability of gene prioritization findings, potentially overlooking crucial genetic factors specific to certain ancestry groups. While initiatives such as the Multi-Ethnic Study of Atherosclerosis (MESA) have made progress by characterizing eQTLs in African American (N = 233), Hispanic (N = 352), and European (N = 578) ancestries separately [180], further efforts in this direction are needed.

In addition to integrating GWAS with xQTL data for gene prioritization, enhancer–gene maps can connect genetic variants in enhancers to their target genes. However, creating these maps based on computational methods, such as co-accessibility and co-activity, often

Outstanding questions

How can we improve the precision of gene prioritization methods to distinguish between causal and co-regulated genes?

How can we accurately define positive (causal) and negative (noncausal) genes when benchmarking gene prioritization methods?

How can we integrate GWAS with xQTLs from various 'omics layers, tissues, cell types, and even cell states for a joint analysis?

What strategies can be utilized to investigate cd-xQTLs that have been underexplored for certain molecular phenotypes, such as RNA splicing and protein abundance?

What methodologies can we use to identify the putative causal genes underlying GWAS loci, where the effects of genetic variants on the trait are mediated by spatiotemporal-dependent genetic regulation of molecular phenotypes?

How can we effectively integrate different types of biological network to derive a combined and representative network for network-based gene prioritization?

How can gene prioritization data and related information be utilized to predict the success rate of drug targets in clinical trials?

How can we leverage associations between drug target genes and unintended phenotypes to predict and manage the risk of adverse events?

requires large-scale epigenomic and transcriptomic atlases. While perturbation experiments, such as CRISPRi, can suggest causal links between enhancers and genes, it is challenging to scale these experiments up due to the absence of cell- or tissue-specific experimental protocols and the complexity of the experiments. Given that enhancer–gene connection catalogs are far from complete, there is an urgent need for high-throughput experiments that can generate connection maps under various conditions.

The advent of whole-exome and whole-genome sequencing studies in large cohorts has greatly facilitated the identification of rare variants associated with complex traits [181,182]. Unlike common variants, rare variants generally do not exhibit strong LD with other common or rare variants, suggesting that the most significantly associated rare variants are primarily causal [183]. Moreover, rare variants located in the coding region tend to exert larger effects and can often be directly linked to target genes. However, because the power of detection depends on both effect size and minor allele frequency, large sample sizes are required to detect most of the rare variant associations [184]. With ongoing advances in biobank efforts and growing accessibility of sequencing technology, we expect a continuous increase in the discovery of rare variants and genes, which will further enhance our understanding of the genetic basis of complex traits.

Compared with those from cellular and animal models, findings from human genetics studies are more relevant to human diseases. Approaches based on human genetics are increasingly important for target identification in early drug development and nonclinical safety assessment. However, the effectiveness of genetics-informed drug development has been limited by the polygenic nature and heterogeneity of complex diseases. To bridge the translational gap between human genetics findings and clinical outcomes, a potential direction for future research is the development of methods capable of integrating disease-associated variants with multi-omics data and biological networks, including those at cellular level, as well as clinical trial data to prioritize candidate therapeutic targets.

Acknowledgments

We thank Weiyang Bai and Ruilei Ma for their valuable discussions and assistance in generating the figures. This research was partially funded by the 'Pioneer & Leading Goose' R&D Program (2022SDXHDX0001 and 2024SSYS0032), the Leading Innovative and Entrepreneur Team Introduction Program (2021R01013), the National Natural Science Foundation of China (U23A20165), and the Westlake University Research Center for Industries of the Future (WU2022C002 and WU2023C010).

Declaration of interests

We have no conflicts of interest to declare.

References

1. McCarthy, M.I. *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9, 356–369
2. Uffelmann, E. *et al.* (2021) Genome-wide association studies. *Nat. Rev. Methods Primers* 1, 59
3. Visscher, P.M. *et al.* (2012) Five years of GWAS discovery. *Am. J. Hum. Genet.* 90, 7–24
4. Visscher, P.M. *et al.* (2017) 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22
5. Abdellaoui, A. *et al.* (2023) 15 years of GWAS discovery: realizing the promise. *Am. J. Hum. Genet.* 110, 179–194
6. Buniello, A. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47, D1005–D1012
7. Sollis, E. *et al.* (2023) The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* 51, D977–D985
8. Schaid, D.J. *et al.* (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* 19, 491–504
9. Maurano, M.T. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190–1195
10. Tam, V. *et al.* (2019) Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* 20, 467–484
11. Yang, J. *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–569
12. Liu, J.Z. *et al.* (2010) A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* 87, 139–145
13. Lamparter, D. *et al.* (2016) Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Comput. Biol.* 12, e1004714
14. Bakshi, A. *et al.* (2016) Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. *Sci. Rep.* 6, 32894

15. de Leeuw, C.A. *et al.* (2015) MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* 11, e1004219
16. Li, A. *et al.* (2023) mBAT-combo: a more powerful test to detect gene-trait associations from GWAS data. *Am. J. Hum. Genet.* 110, 30–43
17. Zhang, M.J. *et al.* (2022) Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. *Nat. Genet.* 54, 1572–1580
18. Bryois, J. *et al.* (2020) Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson's disease. *Nat. Genet.* 52, 482–493
19. Ma, S. *et al.* (2021) Powerful gene-based testing by integrating long-range chromatin interactions and knockoff genotypes. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2105191118
20. Zhu, Z. *et al.* (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* 48, 481–487
21. Gamazon, E.R. *et al.* (2015) A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47, 1091–1098
22. Gusev, A. *et al.* (2016) Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* 48, 245–252
23. Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67, 301–320
24. GTEx Consortium (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330
25. Zhou, X. *et al.* (2013) Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* 9, e1003264
26. Zhang, W. *et al.* (2019) Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits. *Nat. Commun.* 10, 3834
27. Bhattacharya, A. *et al.* (2021) MOSTWAS: multi-omic strategies for transcriptome-wide association studies. *PLoS Genet.* 17, e1009398
28. Raj, T. *et al.* (2018) Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nat. Genet.* 50, 1584–1592
29. Wu, L. *et al.* (2018) A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat. Genet.* 50, 968–978
30. Vosa, U. *et al.* (2021) Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* 53, 1300–1310
31. Dai, Q. *et al.* (2023) OTTERS: a powerful TWAS framework leveraging summary-level reference data. *Nat. Commun.* 14, 1271
32. Wainberg, M. *et al.* (2019) Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* 51, 592–599
33. Nica, A.C. *et al.* (2010) Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* 6, e1000895
34. Plagnol, V. *et al.* (2009) Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. *Biostatistics* 10, 327–334
35. Giambartolomei, C. *et al.* (2014) Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* 10, e1004383
36. Hukku, A. *et al.* (2021) Probabilistic colocalization of genetic variants from complex and molecular traits: promise and limitations. *Am. J. Hum. Genet.* 108, 25–35
37. Hormozdiari, F. *et al.* (2016) Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* 99, 1245–1260
38. Wen, X. *et al.* (2017) Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. *PLoS Genet.* 13, e1006646
39. Smith, G.D. and Ebrahim, S. (2003) 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* 32, 1–22
40. Wu, Y. *et al.* (2018) Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat. Commun.* 9, 918
41. Qi, T. *et al.* (2022) Genetic control of RNA splicing and its distinct role in complex trait variation. *Nat. Genet.* 54, 1355–1363
42. Sanderson, E. *et al.* (2022) Mendelian randomization. *Nat. Rev. Methods Primers* 2, 6
43. Burgess, S. *et al.* (2013) Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* 37, 6586-65
44. Bowden, J. *et al.* (2016) Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* 40, 304–314
45. Hartwig, F.P. *et al.* (2017) Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* 46, 1985–1998
46. Bowden, J. *et al.* (2019) Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *Int. J. Epidemiol.* 48, 728–742
47. Zhu, Z. *et al.* (2018) Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* 9, 224
48. Bowden, J. *et al.* (2015) Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* 44, 512–525
49. Xue, A. *et al.* (2024) Unravelling the complex causal effects of substance use behaviours on common diseases. *Commun. Med.* 4, 43
50. Zheng, J. *et al.* (2020) Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* 52, 1122–1131
51. Zhou, S. *et al.* (2021) A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity. *Nat. Med.* 27, 659–667
52. Gaziano, L. *et al.* (2021) Actionable druggable genome-wide Mendelian randomization identifies repurposing opportunities for COVID-19. *Nat. Med.* 27, 668–676
53. Aguet, F. *et al.* (2023) Molecular quantitative trait loci. *Nat Rev Methods Primers* 3, 4
54. Mancuso, N. *et al.* (2019) Probabilistic fine-mapping of transcriptome-wide association studies. *Nat. Genet.* 51, 675–682
55. Porcu, E. *et al.* (2019) Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.* 10, 3300
56. Qi, T. *et al.* (2018) Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat. Commun.* 9, 2282
57. Barbeira, A.N. *et al.* (2019) Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet.* 15, e1007889
58. Hu, Y. *et al.* (2019) A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat. Genet.* 51, 568–576
59. Zhou, D. *et al.* (2020) A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. *Nat. Genet.* 52, 1239–1246
60. Arvanitis, M. *et al.* (2022) Redefining tissue specificity of genetic regulation of gene expression in the presence of allelic heterogeneity. *Am. J. Hum. Genet.* 109, 223–239
61. Chun, S. *et al.* (2017) Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* 49, 600–605
62. Mostafavi, H. *et al.* (2022) Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. *bioRxiv*, Published online May 8, 2022. https://doi.org/10.1101/2022.05.07.491045
63. Yao, D.W. *et al.* (2020) Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* 52, 626–633
64. Giambartolomei, C. *et al.* (2018) A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* 34, 2538–2545
65. Foley, C.N. *et al.* (2021) A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nat. Commun.* 12, 764
66. Wu, Y. *et al.* (2023) Joint analysis of GWAS and multi-omics QTL summary statistics reveals a large fraction of GWAS signals shared with molecular phenotypes. *Cell Genom.* 3, 100344
67. Gleason, K.J. *et al.* (2020) Primo: integration of multiple GWAS and omics QTL summary statistics for elucidation of molecular

mechanisms of trait-associated SNPs and detection of pleiotropy in complex traits. *Genome Biol.* 21, 1–24
68. Mountjoy, E. *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
69. Caliskan, M. *et al.* (2021) A catalog of GWAS fine-mapping efforts in autoimmune disease. *Am. J. Hum. Genet.* 108, 549–563
70. Broekema, R.V. *et al.* (2020) A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biol.* 10, 190221
71. Benner, C. *et al.* (2016) FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* 32, 1493–1501
72. Wang, G. *et al.* (2020) A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 82, 1273–1300
73. Weissbrod, O. *et al.* (2020) Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* 52, 1355–1363
74. Gaulton, K.J. *et al.* (2023) Interpreting non-coding disease-associated human variants using single-cell epigenomics. *Nat. Rev. Genet.* 24, 516–534
75. Boix, C.A. *et al.* (2021) Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* 590, 300–307
76. Morris, J.A. *et al.* (2023) Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens. *Science* 380, eadh7699
77. Lieberman-Aiden, E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293
78. Fullwood, M.J. *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 462, 58–64
79. Mifsud, B. *et al.* (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* 47, 598–606
80. Jung, I. *et al.* (2019) A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat. Genet.* 51, 1442–1449
81. Javierre, B.M. *et al.* (2016) Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* 167, 1369–1384
82. Pliner, H.A. *et al.* (2018) Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* 71, 858–871
83. Chiou, J. *et al.* (2021) Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. *Nature* 594, 398–402
84. Corces, M.R. *et al.* (2020) Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat. Genet.* 52, 1158–1168
85. Granja, J.M. *et al.* (2021) ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* 53, 403–411
86. Fang, R. *et al.* (2021) Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* 12, 1337
87. Wu, Y. *et al.* (2020) Promoter-anchored chromatin interactions predicted from genetic analysis of epigenomic data. *Nat. Commun.* 11, 2061
88. Ghavi-Helm, Y. *et al.* (2014) Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* 512, 96–100
89. Fulco, C.P. *et al.* (2019) Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* 51, 1664–1669
90. Nasser, J. *et al.* (2021) Genome-wide enhancer maps link risk variants to disease genes. *Nature* 593, 238–243
91. Gasperini, M. *et al.* (2019) A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* 176, 377–390
92. Greene, C.S. *et al.* (2015) Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* 47, 569–576
93. Roussarie, J.P. *et al.* (2020) Selective neuronal vulnerability in Alzheimer's disease: a network-based analysis. *Neuron* 107, 821–835
94. Pers, T.H. *et al.* (2015) Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* 6, 5890
95. Weeks, E.M. *et al.* (2023) Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *Nat. Genet.* 55, 1267–1276
96. Kohler, S. *et al.* (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* 82, 949–958
97. Fang, H. *et al.* (2019) A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nat. Genet.* 51, 1082–1091
98. Wang, Q. *et al.* (2019) A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. *Nat. Neurosci.* 22, 691–699
99. Barrio-Hernandez, I. *et al.* (2023) Network expansion of genetic associations defines a pleiotropy map of human cell biology. *Nat. Genet.* 55, 389–398
100. Yazar, S. *et al.* (2022) Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* 376, eabf3041
101. Cuomo, A.S.E. *et al.* (2023) Single-cell genomics meets human genetics. *Nat. Rev. Genet.* 24, 535–549
102. Perez, R.K. *et al.* (2022) Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science* 376, eabf1970
103. Kim-Hellmuth, S. *et al.* (2020) Cell type-specific genetic regulation of gene expression across human tissues. *Science* 369, eaaz8528
104. Song, L. *et al.* (2023) Mixed model-based deconvolution of cell-state abundances (MeDuSA) along a one-dimensional trajectory. *Nat. Comput. Sci.* 3, 630–643
105. Avila Cobos, F. *et al.* (2020) Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* 11, 5650
106. Jin, H. and Liu, Z. (2021) A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *Genome Biol.* 22, 102
107. Westra, H.J. *et al.* (2015) Cell specific eQTL analysis without sorting cells. *PLoS Genet.* 11, e1005223
108. Donovan, M.K.R. *et al.* (2020) Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nat. Commun.* 11, 955
109. de Klein, N. *et al.* (2023) Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. *Nat. Genet.* 55, 377–388
110. Sheng, X. *et al.* (2021) Mapping the genetic architecture of human traits to cell types in the kidney identifies mechanisms of disease and potential treatments. *Nat. Genet.* 53, 1322–1333
111. Song, X. *et al.* (2023) MiXcan: a framework for cell-type-aware transcriptome-wide association studies with an application to breast cancer. *Nat. Commun.* 14, 377
112. Thompson, M. *et al.* (2022) Multi-context genetic modeling of transcriptional regulation resolves novel disease loci. *Nat. Commun.* 13, 5704
113. Strober, B.J. *et al.* (2019) Dynamic genetic regulation of gene expression during cellular differentiation. *Science* 364, 1287–1290
114. Zeng, B. *et al.* (2023) Genetic regulation of cell-type specific chromatin accessibility shapes the etiology of brain diseases. *bioRxiv*, Published online March 2, 2023. https://doi.org/10.1101/2023.03.02.530826
115. Ota, M. *et al.* (2021) Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. *Cell* 184, 3006–3021
116. Lopes, K.P. *et al.* (2022) Genetic analysis of the human microglial transcriptome across brain regions, aging and disease pathologies. *Nat. Genet.* 54, 4–17
117. Liang, D. *et al.* (2021) Cell-type-specific effects of genetic variation on chromatin accessibility during human neuronal differentiation. *Nat. Neurosci.* 24, 941–953
118. Fairfax, B.P. *et al.* (2012) Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* 44, 502–510
119. Kosoy, R. *et al.* (2022) Genetics of the human microglia regulome refines Alzheimer's disease risk loci. *Nat. Genet.* 54, 1145–1154

120. Mandric, I. *et al.* (2020) Optimized design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. *Nat. Commun.* 11, 5504
121. Cuomo, A.S.E. *et al.* (2021) Optimizing expression quantitative trait locus mapping workflows for single-cell studies. *Genome Biol.* 22, 188
122. Natri, H.M. *et al.* (2023) Cell type-specific and disease-associated eQTL in the human lung. *bioRxiv*, Published online June 29, 2023. https://doi.org/10.1101/2023.03.17.533161
123. Bryois, J. *et al.* (2022) Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nat. Neurosci.* 25, 1104–1112
124. Oelen, R. *et al.* (2022) Single-cell RNA-sequencing of peripheral blood mononuclear cells reveals widespread, context-specific gene expression regulation upon pathogenic exposure. *Nat. Commun.* 13, 3267
125. Daniszewski, M. *et al.* (2022) Retinal ganglion cell-specific genetic regulation in primary open-angle glaucoma. *Cell Genom.* 2, 100142
126. Soskic, B. *et al.* (2022) Immune disease risk variants regulate gene expression dynamics during CD4(+) T cell activation. *Nat. Genet.* 54, 817–826
127. Neavin, D. *et al.* (2021) Single cell eQTL analysis identifies cell type-specific genetic control of gene expression in fibroblasts and reprogrammed induced pluripotent stem cells. *Genome Biol.* 22, 76
128. Jerber, J. *et al.* (2021) Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nat. Genet.* 53, 304–312
129. Cuomo, A.S.E. *et al.* (2020) Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.* 11, 810
130. Schmiedel, B.J. *et al.* (2022) Single-cell eQTL analysis of activated T cell subsets reveals activation and cell type-dependent effects of disease-risk variants. *Sci. Immunol.* 7, eabm2508
131. Sarkar, A.K. *et al.* (2019) Discovery and characterization of variance QTLs in human induced pluripotent stem cells. *PLoS Genet.* 15, e1008045
132. Elorbany, R. *et al.* (2022) Single-cell sequencing reveals lineage-specific dynamic genetic regulation of gene expression during human cardiomyocyte differentiation. *PLoS Genet.* 18, e1009666
133. Kumasaka, N. *et al.* (2023) Mapping interindividual dynamics of innate immune response at single-cell resolution. *Nat. Genet.* 55, 1066–1075
134. Nathan, A. *et al.* (2022) Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature* 606, 120–128
135. Abifadel, M. *et al.* (2003) Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat. Genet.* 34, 154–156
136. Cohen, J.C. *et al.* (2006) Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* 354, 1264–1272
137. Stein, E.A. *et al.* (2012) Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. *N. Engl. J. Med.* 366, 1108–1118
138. Balemans, W. *et al.* (2001) Increased bone density in sclerosteosis is due to the deficiency of a novel secreted protein (SOST). *Hum. Mol. Genet.* 10, 537–544
139. McClung, M.R. *et al.* (2014) Romosozumab in postmenopausal women with low bone mineral density. *N. Engl. J. Med.* 370, 412–420
140. Plenge, R.M. *et al.* (2013) Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* 12, 581–594
141. Finan, C. *et al.* (2017) The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* 9, eaag1166
142. Nelson, M.R. *et al.* (2015) The support of human genetic evidence for approved drug indications. *Nat. Genet.* 47, 856–860
143. King, E.A. *et al.* (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
144. Ochoa, D. *et al.* (2022) Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. *Nat. Rev. Drug Discov.* 21, 551
145. 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
146. Okada, Y. *et al.* (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 506, 376–381
147. Holmes, M.V. *et al.* (2021) Integrating genomics with biomarkers and therapeutic targets to invigorate cardiovascular drug development. *Nat. Rev. Cardiol.* 18, 435–453
148. Schmidt, A.F. *et al.* (2020) Genetic drug target validation using Mendelian randomisation. *Nat. Commun.* 11, 3255
149. Storm, C.S. *et al.* (2021) Finding genetically-supported drug targets for Parkinson's disease using Mendelian randomization of the druggable genome. *Nat. Commun.* 12, 7342
150. Chauquet, S. *et al.* (2021) Association of antihypertensive drug target genes with psychiatric disorders: a Mendelian randomization study. *JAMA Psychiatry* 78, 623–631
151. Walker, V.M. *et al.* (2017) Mendelian randomization: a novel approach for the prediction of adverse drug events and drug repurposing opportunities. *Int. J. Epidemiol.* 46, 2078–2089
152. Pairo-Castineira, E. *et al.* (2021) Genetic mechanisms of critical illness in COVID-19. *Nature* 591, 92–98
153. FDA (2022) *Baricitinib EUA Letter of Authorization*, FDA
154. So, H.C. *et al.* (2017) Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry. *Nat. Neurosci.* 20, 1342–1349
155. Gerring, Z.F. *et al.* (2021) Integrative network-based analysis reveals gene networks and novel drug repositioning candidates for Alzheimer disease. *Neurol. Genet.* 7, e622
156. Konuma, T. *et al.* (2021) Integration of genetically regulated gene expression and pharmacological library provides therapeutic drug candidates. *Hum. Mol. Genet.* 30, 294–304
157. Subramanian, A. *et al.* (2017) A next generation connectivity map: L1000 Platform and the first 1,000,000 profiles. *Cell* 171, 1437–1452
158. Lamb, J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935
159. Sadler, M.C. *et al.* (2023) Multi-layered genetic approaches to identify approved drug targets. *Cell Genom.* 3, 100341
160. Cao, C. and Moult, J. (2014) GWAS and drug targets. *BMC Genomics* 15, 1–14
161. Cheng, F. *et al.* (2019) A genome-wide positioning systems network algorithm for in silico drug repurposing. *Nat. Commun.* 10, 3476
162. Emon, M.A. *et al.* (2020) PS4DR: a multimodal workflow for identification and prioritization of drugs based on pathway signatures. *BMC Bioinformatics* 21, 231
163. Wishart, D.S. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082
164. Zhou, Y. *et al.* (2022) Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. *Nucleic Acids Res.* 50, D1398–D1407
165. Mendez, D. *et al.* (2019) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* 47, D930–D940
166. Freshour, S.L. *et al.* (2021) Integration of the Drug–Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. *Nucleic Acids Res.* 49, D1144–D1151
167. Sakaue, S. and Okada, Y. (2019) GREP: genome for REPositioning drugs. *Bioinformatics* 35, 3821–3823
168. Gaspar, H.A. *et al.* (2019) Drug Targetor: a web interface to investigate the human druggome for over 500 phenotypes. *Bioinformatics* 35, 2515–2517
169. Bell, N. *et al.* (2022) Using genome-wide association results to identify drug repurposing candidates. *medRxiv*, Published online October 17, 2022. https://doi.org/10.1101/2022.09.06.22279660
170. Minikel, E.V. *et al.* (2020) Evaluating drug targets through human loss-of-function genetic variation. *Nature* 581, 459–464
171. Swerdlow, D.I. *et al.* (2015) HMG-coenzyme A reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials. *Lancet* 385, 351–361
172. Schwartz, G.G. *et al.* (2014) Effect of alirocumab, a monoclonal antibody to PCSK9, on long-term cardiovascular outcomes

following acute coronary syndromes: rationale and design of the ODYSSEY outcomes trial. *Am. Heart J.* 168, 682–689
173. Robinson, J.G. *et al.* (2015) Efficacy and safety of alirocumab in reducing lipids and cardiovascular events. *N. Engl. J. Med.* 372, 1489–1499
174. Ridker, P.M. *et al.* (2008) Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *N. Engl. J. Med.* 359, 2195–2207
175. Nguyen, P.A. *et al.* (2019) Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. *Nat. Commun.* 10, 1579
176. Duffy, Á. *et al.* (2020) Tissue-specific genetic features inform prediction of drug side effects in clinical trials. *Sci. Adv.* 6, eabb6242
177. Carss, K.J. *et al.* (2023) Using human genetics to improve safety assessment of therapeutics. *Nat. Rev. Drug Discov.* 22, 145–162
178. Liu, X. *et al.* (2019) Trans effects on gene expression can drive omnigenic inheritance. *Cell* 177, 1022–1034
179. Bhattacharya, A. *et al.* (2022) Best practices for multi-ancestry, meta-analytic transcriptome-wide association studies: lessons from the Global Biobank Meta-analysis Initiative. *Cell Genom.* 2, 100180
180. Mogil, L.S. *et al.* (2018) Genetic architecture of gene expression traits across diverse populations. *PLoS Genet.* 14, e1007586
181. Backman, J.D. *et al.* (2021) Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* 599, 628–634
182. Halldorsson, B.V. *et al.* (2022) The sequences of 150,119 genomes in the UK Biobank. *Nature* 607, 732–740
183. Wu, Y. *et al.* (2017) Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome Biol.* 18, 1–10
184. Yang, J. (2023) Expanding the genetic landscape of obesity. *Cell Genom.* 3, 100400
185. Kumasaka, N. *et al.* (2019) High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nat. Genet.* 51, 128–137
186. Bryois, J. *et al.* (2018) Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. *Nat. Commun.* 9, 3121
187. Kundu, K. *et al.* (2022) Genetic associations at regulatory phenotypes improve fine-mapping of causal variants for 12 immune-mediated diseases. *Nat. Genet.* 54, 251–262
188. Ng, B. *et al.* (2017) An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat. Neurosci.* 20, 1418–1426
189. Oliva, M. *et al.* (2023) DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nat. Genet.* 55, 112–122
190. Min, J.L. *et al.* (2021) Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat. Genet.* 53, 1311–1321
191. Sun, B.B. *et al.* (2018) Genomic atlas of the human plasma proteome. *Nature* 558, 73–79
192. Folkersen, L. *et al.* (2020) Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.* 2, 1135–1148
193. Ferkingstad, E. *et al.* (2021) Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* 53, 1712–1721
194. Sun, B.B. *et al.* (2023) Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* 622, 329–338