Bioinformatic Advancements in GWAS Analysis: Addressing Causality, Multiple Testing, and Regulatory Roles of Non-coding SNPs

Rutger Ballieux, 6919294
Supervisors:
Michiel Thiecke
Jeroen de Ridder

Abstract

Genome-wide association studies (GWAS) have transformed genetic research by uncovering thousands of genetic variants linked to complex traits and diseases, offering opportunities for personalized medicine. Yet, three major obstacles remain: (1) statistical associations alone do not confirm causation, (2) expanding datasets exacerbate the multiple testing burden, and (3) most identified variants fall in non-coding regions whose regulatory mechanisms are difficult to decipher.
In response, advanced bioinformatic tools, such as Fine-mapping, colocalization, Mendelian Randomization, and transcriptome-wide association studies (TWAS), have emerged to localize probable causal variants, integrate molecular QTL data, and test pathways underlying gene regulation. Two key trends now drive these innovations. First, integrative strategies increasingly combine multi-omic and tissue-specific datasets to reveal how non-coding SNPs influence gene expression and disease pathways. Second, methodological convergence merges complementary techniques in multi-step workflows, boosting causal inference and highlighting the most functionally relevant genetic factors.
Despite progress, challenges persist, from the need for higher-resolution tissue data to the computational demands of integrating large-scale datasets. Nevertheless, as reference resources expand and analytical methods mature, these integrative, convergent approaches promise deeper insights into disease mechanisms. Ultimately, such advances stand to accelerate clinically meaningful applications of GWAS, paving the way toward more precise diagnostics, interventions, and truly personalized healthcare.

Plain Language Summary

Genome-wide association studies (GWAS) scan the human genome to identify small genetic differences, called single-nucleotide polymorphisms (SNPs), linked to various traits or diseases. While GWAS have uncovered many potential genetic signals, three main obstacles limit the full understanding of these findings.

First, the correlations observed do not always imply causation, leaving uncertainty about which SNPs truly drive a trait. Second, the large number of SNPs tested raises the risk of statistical errors, both false positives and missed genuine associations. Finally, tying SNPs to actual biological mechanisms is challenging, particularly when most associated SNPs lie in non-coding regions that do not directly encode proteins but instead influence gene regulation.

To overcome these hurdles, researchers have introduced several specialized techniques. Fine-mapping pinpoints the most likely causal SNPs within a genome region flagged by GWAS, filtering out nearby, less relevant variants. Colocalization checks whether one SNP might explain both a GWAS association and a molecular signal (for instance, changes in gene activity), helping confirm that the observed link is not just a coincidence. Mendelian Randomization leverages certain SNPs as “natural experiments” to determine whether a factor, such as the amount of a particular molecule in the blood, possibly causes a disease, rather than merely correlating with it. Finally, transcriptome-wide association studies (TWAS) connect genetic variation to gene expression, clarifying how specific SNPs may drive health outcomes by altering RNA levels.

Across these methods, two major trends have emerged. The first is the rise of integrative approaches, in which data from different layers of biology (e.g., DNA variants, gene expression profiles, and epigenetic features) are combined to provide a clearer picture of how non-coding variants might regulate nearby or distant genes. Some studies now incorporate single-cell technologies, offering even finer resolution by revealing how certain variants operate in specific cell types rather than in entire tissues. This helps uncover hidden regulatory networks that standard analyses might miss.

The second trend is methodological convergence, where several analytical techniques are brought together in a multi-step workflow. For example, after fine-mapping narrows down candidate SNPs, colocalization can verify if these variants also explain related molecular signals. Mendelian Randomization can then test whether those signals reflect a genuine cause-and-effect relationship, while TWAS pinpoints which genes are dysregulated by the implicated SNPs. When used in concert, these approaches could drastically sharpen our view of which genetic variations are truly important, lowering the chance of chasing false signals.

Still, challenges remain. The complexity of non-coding regions requires innovative statistical models and rigorous replication across different populations. Assumptions underlying Mendelian Randomization or colocalization analyses must be carefully evaluated, as hidden factors can distort results. Nonetheless, the combination of high-powered data from large cohorts and increasingly refined analytical methods has substantially advanced our ability to interpret GWAS findings. By clarifying which SNPs exert real effects and how they do so, researchers move closer to using genetic insights for disease prediction, tailored treatments, and other applications in precision medicine.

Introduction

Genome-Wide Association Studies (GWAS): A Cornerstone in Genetic Research

Genome-wide association studies (GWAS) have become a cornerstone of modern genetics, revealing over 625,000 lead associations across more than 15,000 traits[1]. GWAS systematically scan single-nucleotide polymorphisms (SNPs) in large populations, ranging from a few hundred to millions of individuals, to identify genetic variants correlated with diverse phenotypes. Such phenotypes encompass a broad spectrum, including height, smoking initiation, educational attainment, blood pressure, obesity, autoimmune diseases, psychiatric disorders, heart disease, neurodegenerative diseases, and cancer. The sheer breadth of traits studied underscores the widespread applicability of GWAS to human health[2][3].
Rapid expansions in biobanks and large-scale collaborations, including consumer genetics initiatives like 23andMe, Inc., have dramatically increased cohort sizes[3]. This growth enhances statistical power to detect associations for common and rare traits alike[3].
To investigate continuous traits such as height or blood pressure, linear regression models are typically used. These models can be summarized as:

$$Y \sim W\alpha + X_s\beta_s + g + e \text{ (eq 1)}$$

where, for each individual, Y represents the phenotype, W is the covariate matrix (e.g., age, sex), α denotes effect sizes, X_s indicates genotype values for SNP s, β_s is the fixed effect size of SNP s, g is a random effect capturing polygenic contributions, and e represents residual errors. Binary traits, such as disease status, are analyzed using logistic regression, which models the probability of an outcome[2].

Although these regression frameworks offer straightforward methods for linking genotype to phenotype, two major issues limit their ability to illuminate the underlying biology of complex traits:

1. Multiple Testing Burden (MTB): As cohort sizes and SNP counts expand, the sheer number of tests inflates the risk of false positives and negatives. This tradeoff can be particularly detrimental for detecting variants underlying rare phenotypes, which may require even larger or more specialized cohorts[3].
2. Limited Mechanistic Insights: Regression-based GWAS findings are primarily correlational and thereby fail to definitively pinpoint how associated SNPs influence biological processes that underlie specific phenotypes.

Consequently, current developments in the field of GWAS largely focus on mitigating the multiple testing burden while improving mechanistic variant-to-phenotype understanding. I will first describe the predominance of non-coding SNPs and their regulatory significance, then, I will outline the central challenges and tradeoffs in GWAS, and provide a concise overview of emerging bioinformatic approaches that integrate diverse datasets to address these obstacles.

Non-coding SNPs: Predominance and Biological Significance

Compounding these statistical challenges is the remarkable finding that roughly 93% of disease-associated variants reside in the “dark genome”, a term used to describe all DNA that does not contain protein-coding regions, including both intergenic and intronic areas[4]. Initially deemed non-functional, these regions are now actively being investigated for their critical roles in gene regulation. Non-coding SNPs can influence gene expression by modulating regulatory elements such as transcription factor binding sites, enhancers, and silencers (Figure 1A-D)[5][6][7]. These regulatory elements can act over vast genomic distances, controlling the transcriptional activity of distant genes through mechanisms like chromatin looping and three-dimensional genomic interactions (Figure 1A)[8]. Moreover, Chromatin accessibility analyses reveal that many GWAS SNPs cluster in open chromatin regions marked by DNase I hypersensitivity sites (DHSs), indicating active regulatory elements[4]. Moreover, some non-coding SNPs intersect with long noncoding RNAs (lncRNAs) (Figure 1E), which can modulate transcriptional and post-transcriptional activity[9][7]. SNPs within lncRNAs can alter their function, potentially contributing to disease susceptibility. Although gene regulatory effects are a primary focus of SNP investigations, non-coding SNPs may also exert effects through entirely different and potentially unknown mechanisms. Understanding these non-coding regulatory processes is crucial for bridging the gap between SNP association and functional insight.

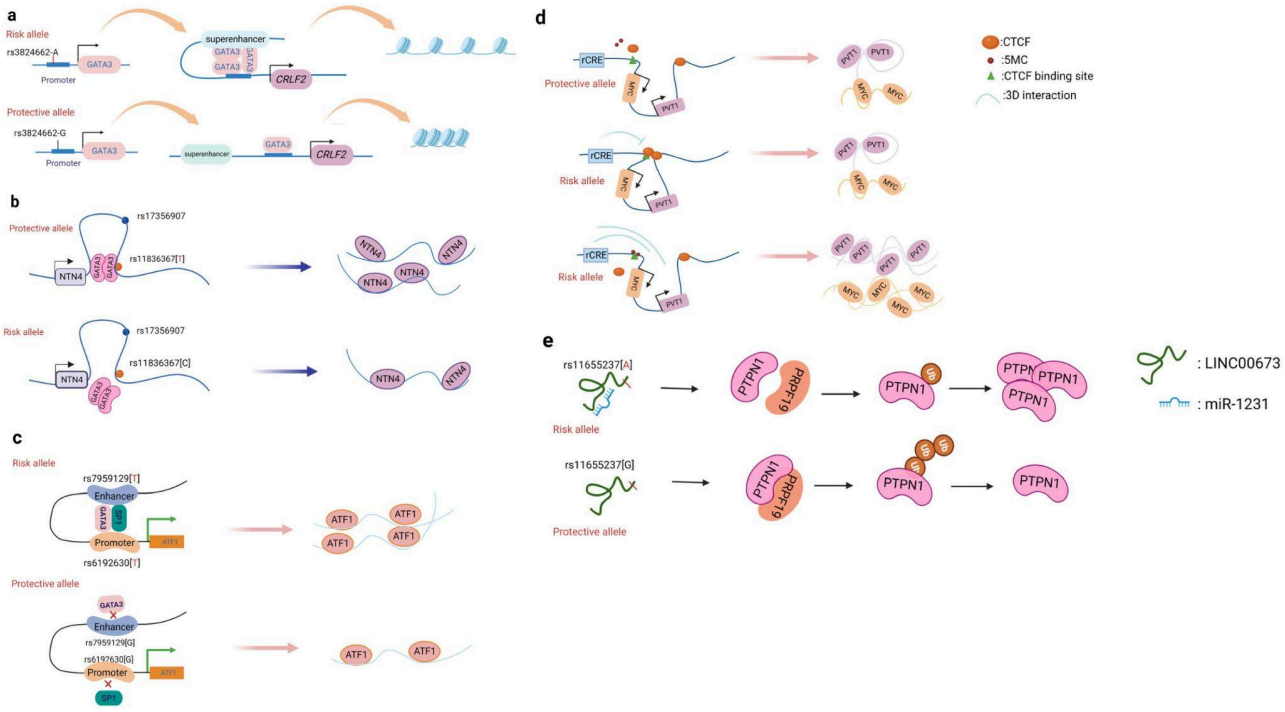

Figure 1: Illustrative mechanisms employed by non-coding SNPs in gene regulation
This figure demonstrates examples of mechanisms by which non-coding SNPs can influence gene regulation. While specific SNPs, genes, and promotors are annotated for illustrative purposes, the focus is on the broader regulatory principles.
(A): The SNP rs3824662 allele A increases chromatin accessibility by inducing *GATA3* expression, promoting the binding of GATA3 to the CRLF promoter and facilitating the formation of a chromatin loop.
(B): The *NTN4* enhancer risk variant rs11836367 binds to the transcription factor GATA3, regulating

NTN4 expression and contributing to the initiation and progression of breast carcinoma.
(C): The enhancer SNP rs7959129 risk allele G interacts with promoter SNP rs6192603 risk allele G, enhancing *ATF1* expression through the binding of transcription factors GATA3 and SP1.
(D): The risk allele rs11986220, along with higher methylation at –10 Kb, synergistically increases tumor risk. In contrast, hypomethylation at –20 Kb inhibits the function of the risk SNP via an enhancer-blocking insulator loop mediated by CTCF.
(E): The risk variant rs11655237 in *LINC00673* creates a binding site for miR-1231, interfering with *LINC00673* expression and contributing to pancreatic cancer susceptibility.
Note: Adapted from (Yang et al., 2022)[7]

Central Challenges and Tradeoffs in GWAS

Many complex diseases exhibit heritability, which stems from the combined effects of large numbers of non-coding genetic variants[10][11]. To illustrate: a large-scale study involving over five million participants found that ~12,000 SNPs were required to explain 40% of the observed phenotypic variance in human height[3][12]. Studies of this magnitude underscore both the power and the complexity inherent in GWAS.
Increasingly large cohort sizes offer several advantages. (1) They enhance statistical power, allowing the detection of more subtle SNP effects with greater confidence. (2) As sample sizes grow, precision in estimating the effect sizes of genetic variants improves, reducing variance and yielding more reliable associations. (3) Larger consortia also capture a broader range of ancestries and environmental backgrounds, enabling subgroup analyses and mitigating population-specific biases. (4) The expanded scale allows researchers to identify variants of low frequency or rarity that smaller studies would lack the power to detect[13][3][2].

At the same time, the expansion of GWAS cohorts could also heighten certain challenges:

1. Correlation vs. Causation
   Although GWAS excels at revealing correlations between genetic variants and traits, it does not inherently distinguish whether a variant plays a causal role. Larger sample sizes refine statistical associations and boost confidence in the observed signals. However, the fundamental challenge of differentiating correlation from causation remains, especially for complex traits influenced by multiple interacting genetic and environmental factors.

2. Multiple Testing Burden
   With millions of SNPs tested across increasingly large cohorts, the risk of false positives inflates dramatically. Researchers typically impose more stringent significance thresholds to control Type I errors (false positives). This stringency can however also inflate Type II errors (false negatives), causing some true associations to go undetected. Although higher statistical power in large samples helps mitigate these tradeoffs, striking the right balance between sensitivity and specificity remains a core concern[13].

3. Bridging SNPs to Biological Mechanisms
   Uncovering the biological mechanisms by which SNPs influence disease is arguably the biggest hurdle for translating GWAS findings into clinical or therapeutic insights.

Most GWAS-identified variants are located in non-coding regions, making it difficult to infer which genes or regulatory elements they affect.

Complicating the difficulty of pinpointing causal variants and bridging SNPs to biological mechanisms further is linkage disequilibrium (LD), the non-random association of alleles at different loci within a population. LD arises from the co-inheritance of alleles that are physically close on a chromosome, diminishing with increasing genomic distance and influenced by recombination events such as crossing over[2][14]. The most basic measure of LD is defined as:

$$D = pAB - pApB \text{ (eq 2)}$$

where pAB is the observed frequency of the AB haplotype, and pA and pB are the frequencies of alleles A and B, respectively. If no association exists between the two loci, $D = 0$. A more commonly used measure is r^2, which normalizes D by the product of all four allele frequencies:

$$r^2 = \frac{D^2}{pA(1-pA)\,pB(1-pB)} \text{ (eq 3)}$$

The r^2 measure can be interpreted as the squared correlation coefficient between the presence of allele A at the first locus and allele B at the second locus[14]. High r^2 values indicate strong LD, which complicates the identification of the causal variants among multiple associated SNPs. When SNPs are in LD, several may be equally associated with the disease, making it challenging to determine which SNP has the most significant effect[2].
Additionally, pleiotropy, where a single genetic variant influences multiple traits, further complicates the interpretation of GWAS results. Pleiotropic effects can obscure the attribution of SNPs to specific disease mechanisms, as the same variant may be involved in different biological pathways or phenotypic outcomes[15][2].

Interpreting GWAS results: advances in Bioinformatic Approaches

To overcome these hurdles, researchers have developed a suite of computational strategies that increasingly incorporate multi-omics data. By combining genomics, transcriptomics, epigenomics, and other molecular profiles, these approaches aim to refine association signals, and try to bridge the gap between variants and their biological function. Four key methodologies are particularly relevant:

1. Fine-Mapping aims to pinpoint the most likely causal variants within a genomic region identified by GWAS by combining statistical methods and functional annotations[16][17].
2. Colocalization Analysis evaluates whether a single SNP underlies associations seen in both GWAS and molecular QTL datasets[18].
3. Mendelian Randomization (MR) uses genetic variants as instrumental variables to infer causal relationships between risk factors and diseases[19].
4. Transcriptome-Wide Association Studies (TWAS) link GWAS to gene expression profiles, aiming to identify gene targets influenced by SNPs[20].

A unifying theme among these tools is the integration of multiple analytical approaches and diverse omics data to enhance GWAS interpretation. By combining various molecular layers, researchers can more accurately identify and functionally characterize associated variants[21]. Recent advances also emphasize the integration of single-cell transcriptomics, providing higher-resolution insights into gene expression and regulation at the cellular level.

These advances are the latest developments in a challenging and ongoing endeavor to address the core GWAS challenges and capture the full impact of non-coding variants. Gradually providing a more comprehensive understanding of biomechanistic disease etiology and moving us closer to the promise of personalized medicine.

Scope and Objectives of This Review

This review examines recent advances in bioinformatic methodologies for analyzing GWAS data, focusing on how these methodologies address the three core GWAS challenges:

1. Correlation vs. Causation
2. Multiple Testing Burden
3. Bridging SNPs to Biological Mechanisms

By reviewing the latest developments in Fine-mapping, Colocalization analysis, Mendelian Randomization, and TWAS, this review highlights how these tools improve the identification and prioritization of candidate variants. Additionally, the integration of various omics datasets, such as QTLs, chromatin accessibility profiles, epigenomic markers, and single-cell transcriptomic data, is explored to demonstrate their role in enhancing biomechanistic understanding of phenotype-associated variants. Special attention is given to TWAS due to its more recent emergence and active development. Ultimately, this review aims to provide a comprehensive overview of how these bioinformatic approaches advance the interpretation of non-coding variants in GWAS, expanding fundamental biological knowledge and opening up new and personalized diagnostic and therapeutic avenues.

Bioinformatic Approaches to GWAS Functional Analysis

Fine-Mapping

Fine-mapping is a critical analytical step following GWAS that aims to identify causal genetic variants within associated loci[22][2]. Although GWAS can highlight genomic regions linked to complex traits and diseases, pinpointing the specific causal variants remains challenging due to LD and the indirect associations of tag SNPs with phenotypes. This challenge is especially pronounced for non-coding SNPs, whose functional impacts are not immediately apparent. Fine-mapping refines the list of candidate variants by analyzing association signals and LD patterns, thereby disentangling causal variants from those merely associated due to LD.

Early fine-mapping efforts relied on heuristic methods based on LD patterns to prioritize candidate variants[22]. For example, LD thresholding would retain variants showing high pairwise LD (e.g., r^2 above a certain threshold) with the lead SNP, assuming these were more likely causal. Methods like hierarchical clustering grouped SNPs into clusters based on LD, identifying haplotype blocks that might harbor causal variants. While these approaches offered initial insights, they had notable drawbacks. The use of arbitrary thresholds risked excluding true causal variants, and their inability to account for the joint effects of multiple SNPs limited their effectiveness, especially given the polygenic nature of most complex traits.
To address these limitations, penalized regression models were introduced, enabling simultaneous estimation of SNP effect sizes and variable selection in high-dimensional genomic data[22]. Methods like LASSO and Elastic Net apply penalties to regression coefficients, shrinking smaller effects toward zero and selecting relevant SNPs for inclusion in the model. These approaches improved upon heuristic methods but still grappled with calibration issues and the challenge of highly correlated variants that can obscure true causal signals, an issue especially acute for non-coding SNPs lacking straightforward functional validation.

Bayesian fine-mapping methods emerged as a response to these shortcomings, providing a probabilistic framework that accounts for uncertainty and allows incorporation of prior information about variants and models. Popular tools include CAVIAR[16] and FINEMAP[23], which compute posterior probabilities (PIPs) for each variant to reflect the probability of causality given the observed data[22].
Bayesian approaches offer several advantages. They enable direct probabilistic interpretations, allow incorporation of functional annotations, and jointly model multiple variants[16][23][22]. These attributes are especially helpful in complex LD structures commonly encountered in non-coding regions, where many correlated variants might exist but only a subset truly influences gene regulation or disease risk.
A major advancement, functional fine-mapping, integrates additional genomic data, such as epigenetic marks, eQTLs, and chromatin accessibility, to highlight variants with likely regulatory roles (Figure 2)[24]. By tying non-coding variants to their potential impacts on gene regulation, functional fine-mapping narrows down the candidate list to those variants with meaningful biological impacts, bridging the gap between GWAS association and mechanism.

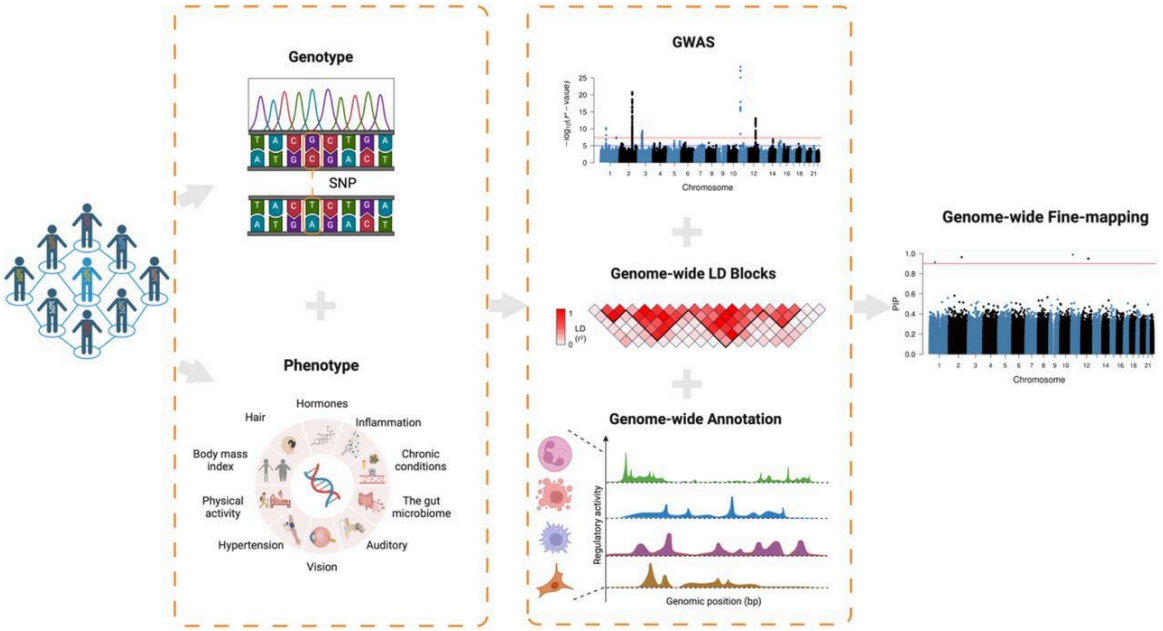

Figure 2: Schematic overview of a Fine-Mapping method utilizing LD reference panels and functional annotations

This figure illustrates a fine-mapping workflow that uses GWAS summary statistics, genome-wide LD reference panels, and functional annotations to identify likely causal variants. LD patterns are used to refine candidate variants, while functional annotations help prioritize those with potential biological relevance. The schematic outlines the process, from GWAS signal identification to variant prioritization.

Note: Figure from (Wu et al., 2024)[25]

In recent years, novel fine-mapping efforts have also expanded from variant-level resolution to gene-level resolution through transcriptomic data integration. Methods like FOCUS[17] and FOGS[26] incorporate TWAS frameworks to map non-coding variants onto the genes they influence, offering potentially more interpretabe findings. However, challenges remain around relying on predicted gene expression from reference panels or proxy tissues, which may not capture all regulatory contexts.

Further developments, such as GIFT[27] and FABIO[28], extend TWAS-based fine-mapping across entire chromosomes. These models account for correlations among genetically regulated expression across multiple genes and accommodate binary disease traits. These methods promise more comprehensive modeling of genetic architecture, reducing credible gene sets, and aiding discovery of both known and novel causal genes. However, their increased computational demands and complexity raise concerns about whether they consistently outperform simpler methods or produce biologically meaningful conclusions for non-coding SNPs. Such unresolved issues call for further validation and benchmarking.

On the variant-centric side, Bayesian calibration refinements seek to address the overconfidence in PIPs, which often occur in highly polygenic contexts. Some models[29] now incorporate infinitesimal effect assumptions alongside sparse, large-effect variants, aiming to reduce replication failures and yield more reliable PIPs. Although promising, the practical advantages of these refined models remain to be firmly established.

Recent advancements in fine-mapping highlight the integration of functional data within statistical frameworks, providing more context for understanding the possible role of non-coding variants.

Colocalization

After fine-mapping refines the list of candidate variants in a GWAS locus, colocalization analyses take the next step by examining whether two association signals, typically one from a GWAS and another from a molecular QTL study, stem from the same causal variant[18]. By testing for shared causality, colocalization aims to distinguish genuine pleiotropy from coincidental overlaps due to LD[15]. This approach can be particularly informative when non-coding SNPs are suspected to influence regulatory processes, enabling researchers to explore whether a trait-associated variant may also affect gene expression or other molecular phenotypes[21][30].

One of the foundational and most widely used tools for colocalization is Coloc[18]. Coloc uses a Bayesian framework to evaluate five hypotheses, ranging from no association to complete colocalization (Figure 3). By calculating approximate Bayes factors and posterior probabilities, Coloc quantifies the likelihood that a GWAS and QTL signal are driven by the same causal variant. Its simplicity, efficiency, and modest computational demands have established Coloc as the standard method in the field, making it the de facto benchmark for new methods[31]. However, Coloc assumes a single causal variant per locus, which can limit its effectiveness in regions with allelic heterogeneity or multiple functional SNPs.

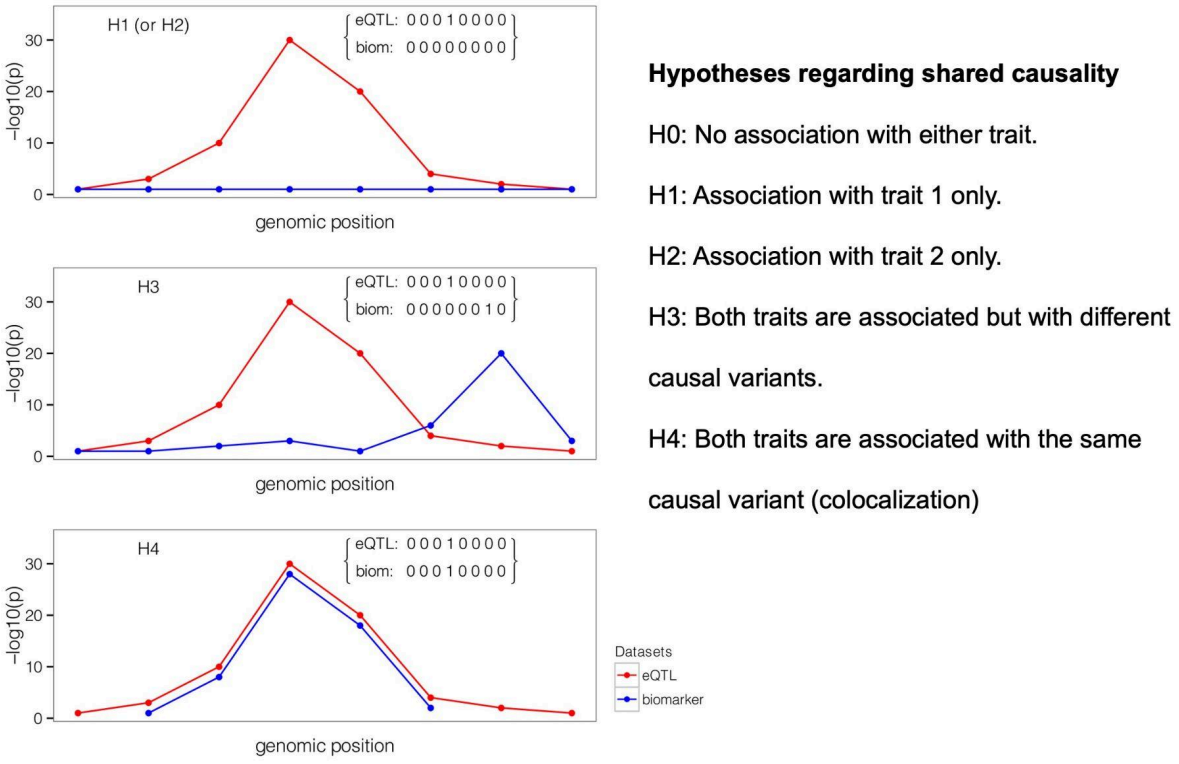

Figure 3: Colocalization hypotheses in a single genomic region

Colocalization analyses evaluate five distinct hypotheses regarding the shared causality between two association signals, such as those from a GWAS and a molecular QTL study, within a single genomic region. Each hypothesis represents a different relationship between the traits and their causal variants. The figure illustrates panels showing -log10(p) values of association signals across eight shared variants within the region. Red lines represent the QTL dataset and blue lines represent the GWAS dataset. Binary vectors indicate whether a variant is causally involved (1) or not (0) for each trait. The specific hypotheses (H0 to H5) are outlined on the right side of the figure.

Top panel (H1 or H2): Association present in only one dataset.
Middle panel (H3): Associations present in both datasets but driven by distinct causal variants.
Bottom panel (H4): Associations in both datasets driven by the same causal variant, indicating colocalization.
Note: Adapted from (Giambartolomei et al., 2014)[18]

To overcome this limitation, newer methods emerged. eCAVIAR[32] extends the Coloc framework by accommodating multiple causal variants within a locus. By integrating fine-mapping outputs in colocalization analysis, eCAVIAR can identify sets of SNPs that jointly drive both GWAS and QTL signals, thereby offering a more nuanced view of complex loci. Similarly, enloc[33] takes a Bayesian hierarchical approach that infers QTL enrichment directly from the data and also supports multiple causal variants per locus. These enhancements are crucial for capturing the polygenic and regulatory landscapes that define many complex traits.
While coloc and its immediate successors largely centered on two-trait scenarios, moloc[34] broadens the Bayesian framework to simultaneously handle multiple datasets, such as eQTLs, methylation QTLs, and other molecular phenotypes. By evaluating more than two traits at once, moloc aims to provide a broader perspective on how a single variant might influence a network of molecular outcomes. However, this added complexity requires greater computational resources.

The methods discussed above, along with other newer approaches, build upon Coloc's principles and address its key limitations by accommodating multiple causal variants, estimating QTL enrichment, and handling diverse data types. However, as noted by Zang et al. (2024)[31], the adoption of these advanced methods remains limited. Researchers often use multiple methods in tandem, with Coloc serving as a baseline for comparison. Therefore, further validation is necessary to determine whether the added complexity of these newer methods reliably enhances the interpretability and robustness of colocalization results[31].

Recent advancements in colocalization analysis build on foundational methods like Coloc, addressing its limitations. However, adoption of these newer approaches remains limited.

Mendelian Randomization (MR)

Mendelian Randomization (MR) extends the insights gained from colocalization by testing whether a trait-associated SNP possibly exerts a causal effect on a particular exposure, such as a biomarker or gene expression level, and subsequently on disease phenotype[19]. This framework treats genetic variants as instrumental variables, leveraging the random assortment of alleles during meiosis to mimic the design of a randomized controlled trial (Figure 4A). This approach also helps reduce confounding and reverse causation (Figure 4B-C), making it especially valuable for clarifying how non-coding SNPs might functionally influence disease risk[19][2][21][30].

MR operates through two main steps. First, genetic variants (SNPs) that are strongly associated with the exposure of interest are selected as instrumental variables. Second, the causal effect estimation involves assessing the association between these genetic instruments and the phenotype or outcome of interest. By analyzing this relationship, MR aims to infer whether the exposure potentially exerts a causal influence on the outcome[19][2][21][30].

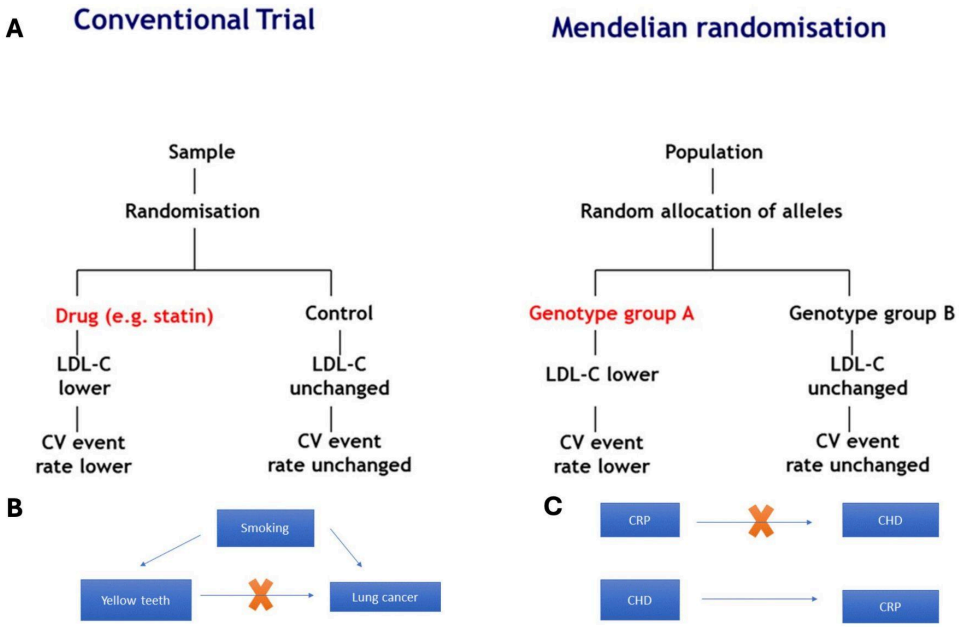

Figure 4: Design of Mendelian Randomization study and common challenges in observational studies

(A): Comparison of a randomized controlled trial (RCT) and a Mendelian Randomization (MR) study. This panel illustrates how MR mimics the structure of a RCT by using the random assortment of alleles during meiosis as a natural experiment, assigning individuals genetic variants akin to treatment and control in a RCT. Both approaches aim to identify causal relationships.

(B): Visual representation of confounding in observational studies. The diagram demonstrates how a spurious association (e.g., between yellow teeth and lung cancer) can arise due to a shared confounder (e.g., smoking). The arrow represent proposed causal directions, and the cross indicates that the direct link between yellow teeth and lung cancer is false.

(C): Visual representation of reverse causation in observational studies. This panel shows how an association (e.g., between C-reactive protein [CRP] levels and coronary heart disease [CHD]) might be misinterpreted. The arrows represent the direction of proposed causality, with the cross highlighting that the assumed direct link between CRP and CHD is incorrect. Current evidence suggests the reverse, CHD increases CRP levels.

Note: Figure adapted from (Bennet & Holmes, 2017)[35]

For MR to produce reliable causal inferences, three assumptions must hold:

1. Relevance: The genetic instrument is strongly associated with the exposure.
2. Independence: The instrument is independent of any confounders that affect both the exposure and the outcome.
3. Exclusion Restriction: The instrument influences the outcome exclusively through the exposure and not through any alternative pathways.

Violations of these assumptions can lead to false causal claims. For instance, horizontal pleiotropy, where a SNP affects multiple traits, breaks the exclusion restriction by introducing unintended causal routes[15]. Similarly, LD blocks can confound MR analyses if correlated SNPs influence different exposures or outcomes, obscuring the true causal mechanism. As a result, MR findings must be interpreted with caution, especially for complex traits with overlapping genetic influences[19][2][21][30].

One strategy to mitigate these pitfalls is to pair MR with colocalization. Colocalization complements MR by assessing whether the exposure and outcome signals arise from the same causal variant[30]. If the association signals fail to colocalize, the SNPs driving the exposure and outcome may be distinct, undermining the fundamental premise of MR. Thus, a combined MR-colocalization approach could reinforce causal claims by ensuring that the inferred evidence is not distorted by multiple linked variants or pleiotropic effects.

Several MR frameworks have been developed to handle diverse research questions[2][21][30]:

- Two-sample MR: Uses summary statistics from two independent GWAS datasets, one for the exposure, one for the outcome.
- eQTL-based MR: Uses eQTL data as exposures to test whether genetically regulated gene expression causally influences disease risk.
- Cis-MR: Restricts the instrumental SNPs to those near the target gene (cis-eQTLs), tightening specificity for local regulatory effects.
- Polygenic MR: Aggregates multiple SNPs associated with the exposure to estimate the combined causal effect, accounting for polygenic architectures in complex diseases.

More advanced methods, like Graph-MRcML[36], extend MR beyond pairwise trait comparisons to infer causal networks among multiple traits. This method employs a two-step process: constructing a total causal network and applying network deconvolution to distinguish direct from indirect causal effects. This framework aims to provide a more detailed understanding of how genetic variants exert effects across multiple traits simultaneously. The efficacy of this approach awaits validation, but it has the possibility of broadening MR's potential in elucidating the genetic architecture of complex diseases.

Mendelian Randomization aims to link genetic associations with potential causal mechanisms by using genetic variants as instrumental variables. Its effectiveness relies on the validation of core assumptions, which can be difficult to satisfy in complex diseases.

Transcriptome-Wide Association Studies (TWAS)

Transcriptome-wide association studies (TWAS) offer a gene-centered approach to interpreting GWAS results by linking genetic variation to gene expression changes[20][21]. This framework is particularly valuable for non-coding SNPs, because rather than testing millions of SNPs individually, TWAS aggregates the effects of multiple eQTLs to infer genetically regulated expression (GReX) for each gene. Thereby potentially uncovering mechanistic pathways that underlie complex traits and diseases.

TWAS operate under the hypothesis that genetic variants, especially eQTLs, collectively regulate gene trascription, and that these genetically altered gene expression levels influence disease risk or trait variability. TWAS proceeds in two main steps (Figure 5). First, for each gene, GReX is imputed by combining the transcriptional regulatory effects of eQTLs under an additive genetic model. A vector of gene expression levels is predicted using a genotype matrix and eQTL effect sizes. In this initial stage, an independent reference panel with both genotype and expression data is used to estimate eQTL effects, effectively learning how genetic variation shapes gene expression. Second, these estimated weights are then applied to a GWAS cohort that lacks direct expression measurements. By multiplying the cohort's genotype data by the inferred eQTL weights, TWAS aims to reconstruct the genetically regulated component of expression for each gene. This approach can be tailored to tissue-specific reference data, ensuring that the predicted expression captures the most biologically relevant regulatory context[20][21].
By aggregating the effects of multiple eQTLs on gene expression, TWAS provide a gene-level association testing framework that directly links genetic variation to gene expression changes associated with disease risk[21]. This gene-centric focus facilitates the functional characterization of non-coding SNPs and elucidates how they may influence gene regulation. Moreover, by aggregating the regulatory contributions of multiple variants, TWAS enhance statistical power, enabling the detection of associations that might be missed when examining individual SNPs with small effect sizes. Additionally, conducting association tests at the gene level significantly reduces the multiple testing burden, thereby lowering the risk of false positives and increasing the robustness and reliability of identified gene-trait associations. Consequently, TWAS serves as an efficient and effective tool for uncovering the genetic architecture of complex traits.

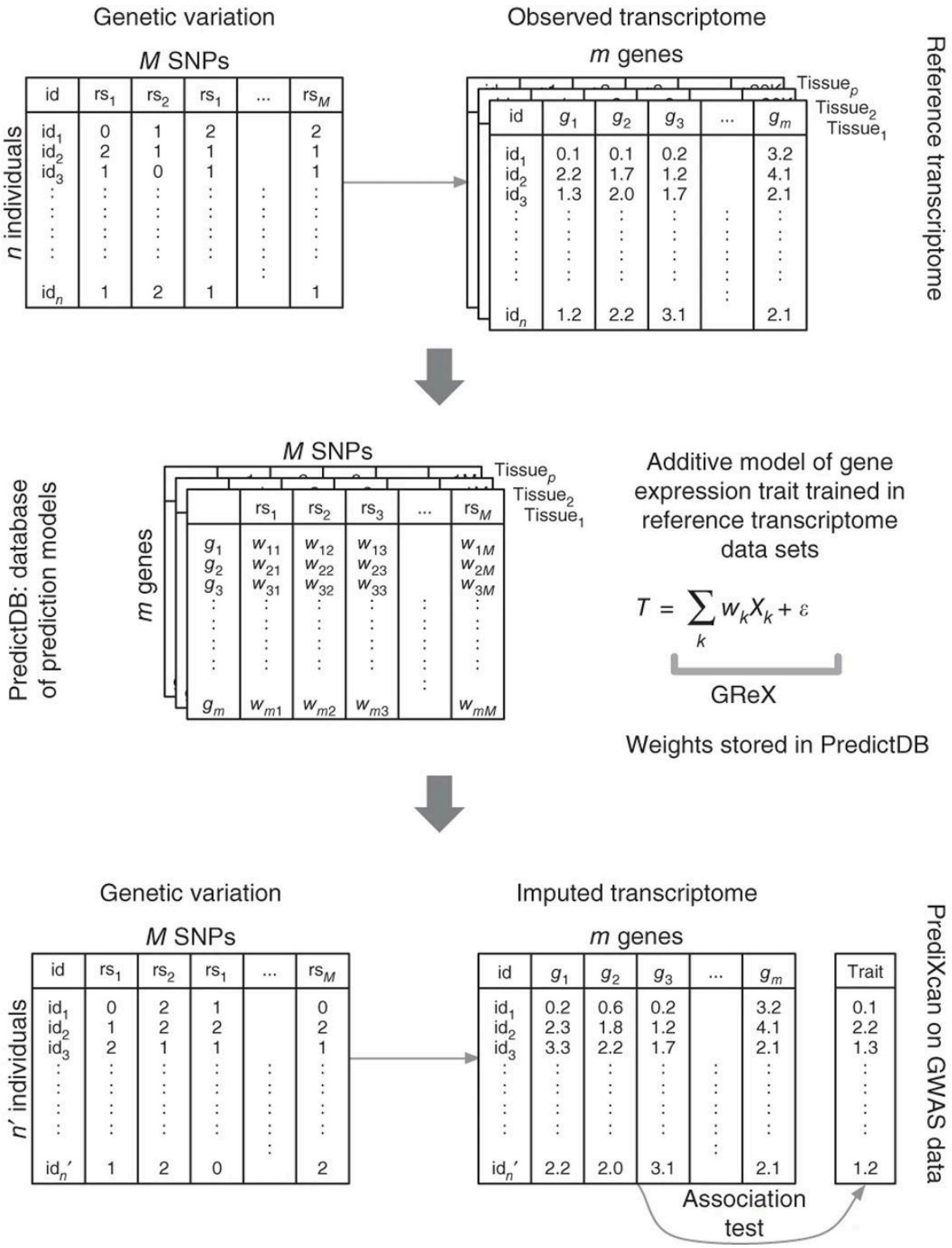

Figure 5: Workflow of Transcriptome-Wide Association Studies (TWAS) using PrediXcan
This schematic illustrates the PrediXcan TWAS workflow. Initially, genotype and gene expression data from reference transcriptome studies are integrated to develop predictive models of gene expression based on genetic variants. These models are then applied to genotype data from a GWAS cohort lacking direct expression measurements to impute genetically regulated expression (GReX) for each gene. Finally, the imputed expression levels are statistically associated with the trait of interest, enabling the identification of gene-trait associations by linking genetic variation to expression changes underlying complex traits and diseases.
Note: Figure from (Gamazon et al., 2015)[20]

Initial TWAS methods, such as PrediXcan[20] (Figure 5), used individual-level genotype and phenotype data to estimate the association between imputed gene expression and complex traits. PrediXcan employed elastic net regression to estimate eQTL effect sizes and impute GReX levels, demonstrating the utility of integrating genomic and transcriptomic data. Recognizing the difficulty of obtaining large individual-level datasets, S-PrediXcan[37] extended this approach to GWAS summary statistics, broadening its applicability.
Concurrent with these developments, FUSION[38], a TWAS framework that also utilizes GWAS summary statistics arose. FUSION introduced a Bayesian sparse linear mixed model to impute GReX, accounting for the polygenic nature of gene expression regulation and complex LD structures. This method improved upon previous approaches by providing more accurate estimates of gene expression and enhancing the detection of gene-trait associations.
Subsequent tools sought to improve expression imputation further. TIGAR[39] used a Bayesian nonparametric latent Dirichlet process to capture both sparse and polygenic regulatory architectures. TIGAR demonstrated improved imputation performance over earlier methods, particularly for complex diseases where gene expression regulation may involve intricate genetic architectures. TIGAR-V2[40] further enhanced this approach by making the model more modular, faster, and convenient, facilitating its application to large-scale datasets.

To overcome the limitations of traditional TWAS methods, which primarily focus on (proximal) cis-eQTLs and may miss significant regulatory effects mediated by (distal) trans-eQTLs, recent methodological advancements have sought to incorporate a broader range of regulatory variants. Tools such as BGW-TWAS[41] integrate both cis- and trans-eQTLs, leveraging Bayesian variable selection to efficiently manage large numbers of SNPs. Utilizing a scalable expectation-maximization Markov Chain Monte Carlo (EM-MCMC) algorithm, BGW-TWAS reduces computational burden, enabling genome-wide TWAS that includes millions of trans-SNPs.
Incorporating trans-eQTLs is particularly beneficial for non-coding SNPs, as it captures distal regulatory effects that cis-focused TWAS methods might overlook[41]. For example, when applied to breast and ovarian cancer GWAS data[42], BGW-TWAS integrated both cis- and trans-eQTLs to identify over 100 susceptibility genes, including novel candidates predominantly driven by trans-eQTLs, such as *ACAP3*. This study underscores the importance of including trans-eQTLs to more accurately capture the regulatory landscape influencing complex traits.

Other recent methods emphasize multi-omic data integration to refine gene expression imputation. For example, ETWAS[43] incorporates epigenetic features, such as chromatin states, DHSs, and TFBS, into the TWAS framework. By considering these regulatory elements, ETWAS aims to improve the identification of genes influenced by epigenetic modifications. Similarly, MOSTWAS[44] includes trans-eQTLs and defines regulatory elements, such as TFs, microRNAs, CpG methylation sites, and chromatin-binding factors, as potential mediators in gene expression prediction. These approaches could enhance the interpretation of non-coding SNPs by linking them not just to gene expression but also to the regulatory elements and epigenetic modifications that shape expression profiles.

To overcome the limitations of individual-level eQTL data, such as limited sample sizes and privacy concerns, recent methodologies have leveraged large-scale summary-level eQTL

data to expand the applications of TWAS. For instance, SUMMIT[45] uses penalized regression with various penalty types to model gene expression using extensive consortia datasets, including eQTLGen[46]. Similarly, OTTERS[47] integrates multiple polygenic risk score (PRS) methods to estimate eQTL weights by combining individual p-values via the Aggregated Cauchy Association Test (ACAT-O). By avoiding the need to pre-specify the optimal method, OTTERS aims to enhance TWAS power and accommodate diverse genetic architectures, thereby improving the robustness and accuracy of eQTL weight estimation.

Advances in statistical methodologies have further enhanced TWAS analyses by addressing challenges related to false discovery rates and complex LD structures. For instance, TWAS-GKF[48] introduces a GhostKnockoff-based[49] TWAS method that achieves finite-sample false discovery rate (FDR) control. By utilizing GhostKnockoff variables derived from summary statistics as robust negative controls, TWAS-GKF promises to enhance the precision of gene-trait association identification, particularly in regions with complex LD.
Moreover, methods like SR-TWAS[50] leverage ensemble machine learning to pool predictions from multiple imputation models trained on diverse reference panels, regression methods, and tissue types. By employing stacked regression to optimally combine base models, SR-TWAS claims to enhance prediction accuracy and discover a wider range of causal genes.

While these ensemble and knockoff-based solutions expand on linear assumptions, a key frontier lies in deep learning. Ramprasad et al. (2024)[51] propose a convolutional neural network (CNN) to capture intricate SNP-gene expression interactions that linear or polynomial models often overlook. By incorporating over fifty functional annotations, such as histone marks, chromatin states, and enhancers, directly into the training process, their network learns biologically informed constraints without succumbing to severe overfitting. Empirical results show modest yet notable improvements in predictive accuracy compared to elastic net regression, with some genes showing substantial gains[51]. These improvements could potentially translate into more reliable genotype-phenotype associations for highly complex loci. However, these networks demand considerable computational resources and are currently only evaluated on GTEx data, leaving questions of scalability and generalizability to other populations and tissues.

An alternative deep learning paradigm appears in the scPrediXcan preprint[52], which diverges from both linear TWAS methods and the architecture employed by Ramprasad et al. (2024)[51]. This method builds on a pre-trained genome-wide model known as Enformer[53]. Enformer's key contribution is its ability to predict gene expression from DNA sequence alone by considering regulatory elements and enhancer–promoter interactions spanning up to 100 kb. Rather than deeply embedding functional annotations, scPrediXcan uses a transfer learning step, applying Enformer's learned representations within a simpler multilayer perceptron (MLP). What further sets scPrediXcan apart methodologically is its translation of the deep learning predictions back into a linear, SNP-based elastic net model. This hybrid workflow preserves the predictive abilities of a deep learning model while remaining compatible with standard TWAS pipelines, thereby offering a pragmatic compromise that balances accuracy, interpretability, and computational requirements. And herein lies scPrediXcan's core significance, bridging the gap between computationally intensive advanced deep learning models and the scalable, linear frameworks familiar in TWAS pipelines. This balancing act is promising but awaits further validation.

TWAS aims to refine GWAS interpretation by linking genetic variants to gene-level regulatory insights. Among GWAS methodologies, it represents one of the most rapidly evolving areas, driven by innovations in imputation algorithms and data integration.

Single-Cell Developments in GWAS Interpretation

Single-cell genomics offers a transformative lens for understanding how non-coding variants drive complex traits by uncovering cell-type-specific regulatory mechanisms. Whereas traditional bulk-tissue analyses can mask critical cellular heterogeneity, single-cell RNA sequencing (scRNA-seq) and related technologies allow researchers to map genetic variants to distinct cell populations or states (Figure 6). This higher resolution is particularly beneficial for interpreting non-coding SNPs, as it facilitates the investigation of their potential roles in regulating specific cellular processes and regulatory elements[54][55][56][57][58].

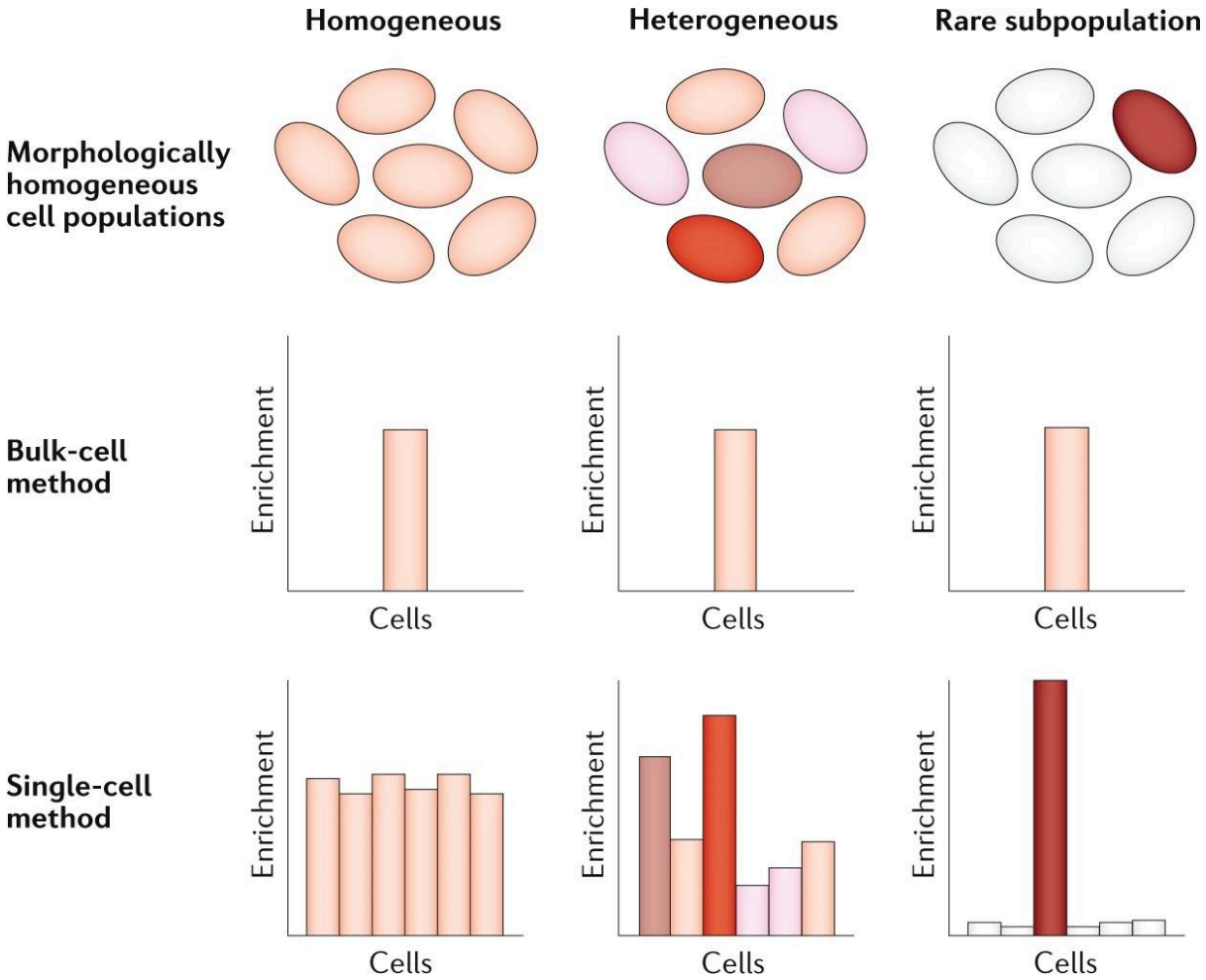

Figure 6: Single-cell techniques provide greater resolution than bulk-cell methods
The figure illustrates three morphologically homogeneous cell populations exhibiting different patterns of cellular heterogeneity, with darker shading representing higher levels of variability. Bulk-cell analysis produces an average value for the entire population, obscuring the distinct heterogeneity patterns. In contrast, single-cell approaches can clearly differentiate between varying degrees of heterogeneity (middle), identify rare subpopulations (right), and recognize uniformly homogeneous populations (left). This demonstrates the superior ability of single-cell methods to capture and resolve the diverse cellular states that bulk analyses cannot discern.
Note: Figure from (Carter & Zhao, 2021)[59]

Early efforts to integrate single-cell data with GWAS have focused on direct mapping strategies. For instance, scGWAS[54] employs a network-based framework that leverages cell-type-average gene expression profiles from large scRNA-seq datasets. By identifying gene modules enriched for both GWAS signals and specific cell types, scGWAS highlights candidate cell types and gene networks most relevant to disease. This approach moves beyond simple enrichment analyses, offering a more mechanistic perspective on disease susceptibility genes.
In contrast, scDRS[55] assesses polygenic disease enrichment at the level of individual cells, rather than relying on predefined cell-type groupings, as is the case in scGWAS. It calculates a disease relevance score for each cell based on the expression of GWAS-implicated genes, thereby aiming to uncover cellular heterogeneity and identify rare subpopulations critical to disease mechanisms, which may remain hidden in aggregate tissue data.
A benchmarking study[56] revealed complementary strengths in these two approaches. ScGWAS excelled at pinpointing enriched gene sets and trait-cell type associations, whereas scDRS's cell-level scoring system more effectively captured subtle disease-relevant cellular subpopulations.

Single-cell resources also enhance TWAS by refining gene expression imputation to specific cell types or states. Traditional TWAS often rely on bulk expression measurements, potentially obscuring cell-type-specific regulatory effects. Recent developments like scTWAS Atlas[57] aggregate millions of single-cell TWAS results from multiple studies, traits, and tissues. This resource enables exploration of how certain variants influence expression in distinct cellular populations.
Similarly, scPrediXcan[52] adapts PrediXcan-style methodologies to single-cell expression profiles, aiming to capture the nuanced interplay between non-coding SNPs and cell-type-specific regulatory landscapes.
In tandem, databases dedicated to single-cell eQTLs have emerged as pivotal resources. scQTLbase[58] provides a comprehensive repository of sc-eQTLs across multiple tissues and cell states, and allows users to query how particular variants modulate gene expression in specific cell types and conditions. The ability to align sc-eQTL data with GWAS signals provides a more direct route to linking genetic variants with the cellular environment in which they act.

While single-cell approaches offer unprecedented resolution, additional developments are needed to fully harness their potential for variant interpretation. Building more comprehensive cell-type atlases in disease-relevant tissues and refining computational pipelines for large-scale, multi-omic data remain key priorities[58][52]. Addressing trans-acting effects and cell-state transitions will further enhance our ability to pinpoint how non-coding SNPs modulate gene expression[54][55].

Single-cell transcriptomics is emerging as a critical tool in GWAS pipelines, offering cellular-level resolution for variant interpretation. Its integration with existing GWAS approaches promises to enhance the ability to identify where and how genetic variants exert their effects.

Discussion

Genome-wide association studies (GWAS) comprise an important collection of approaches to explore the genetic factors underlying complex human traits and diseases. Although numerous trait-associated variants have been identified, it remains a challenge to bridge the gap between significant variant associations and functional understanding. This, in turn, impedes the successful translation of GWAS discoveries into clinical applications and personalized medicine.
A central reason for this gap lies in the complex relationship between variants and their associated phenotypes. The vast majority of associated SNPs reside in non-coding regions, obfuscating which genes they affect directly in terms of transcriptional regulation. Moreover, associated variants may affect multiple genes, and in a context dependent manner. For instance, a variant might perturb a regulatory element that is active only during lineage choice within a specific cellular trajectory. Taken together, while associated non-coding variants offer valuable insights into the genetic basis of (disease) phenotypes, fully understanding their biological significance requires the development and application of specialized methods.

This review has discussed a range of bioinformatic tools and methodologies, each addressing specific challenges in moving beyond simple correlative associations.

Fine-Mapping

Fine-mapping aims to identify SNPs within a GWAS locus that are most likely to influence phenotype associations by utilizing linkage disequilibrium (LD) patterns, Bayesian probability frameworks, and functional annotations. This approach effectively reduces the number of candidate variants for further investigation. However, statistical correlations alone do not establish direct causal relationships between SNPs and phenotypes. To address this limitation, recent advancements have focused on integrating biological data, such as epigenomic and transcriptomic information, into Fine-mapping processes. This integration provides functional context that enhances the identification of driver SNPs, particularly in regions with multiple correlated variants. Additionally, other methodological developments, including gene-level analyses and polygenic frameworks, promise a more comprehensive understanding of the genetic architecture of complex traits. Nevertheless, these added layers of complexity bring new assumptions and computational challenges, highlighting the need for rigorous benchmarking and independent validation across diverse genetic backgrounds. Ensuring the robustness of next-generation fine-mapping approaches is essential for their effective application in functional studies and therapeutic development.

Colocalization

Once candidate variants are identified, Colocalization methods are used to determine whether a single genetic variant is responsible for associations observed in both GWAS and molecular QTL datasets. These techniques provide insights into whether a given variant truly exerts pleiotropic effects, or if observed overlaps are due to LD.
Although colocalization analysis has evolved and diversified, with more recent methods accommodating multiple causal variants for instance, adoption of these newer approaches

remains limited. Coloc remains the preferred choice for many standard applications due to its computational efficiency, conceptual clarity, and empirical track record. The ultimate impact of emerging methods will therefore depend on their ability to demonstrate consistent and validated advantages. Nonetheless, continued methodological advancements and better multi-omic integration promise to enhance how colocalization identifies possible functional impacts of non-coding SNPs.

Mendelian Randomization (MR)

Mendelian Randomization leverages genetic variants as instrumental variables to test whether an exposure (such as gene expression levels) possibly contributes causally to a disease outcome. MR aims to differentiate correlation from causation by mimicking an observational study, offering a key bridge between genetic associations and potential biological mechanisms. However, for MR to yield valid inferences, it relies on three assumptions that must not be violated: Relevance, Independence, and Exclusion Restriction. In practice, these assumptions are often challenging to satisfy, especially for complex diseases where multiple genetic variants may influence various biological pathways and traits.
Even so, MR remains a valuable tool, enabling causal hypothesis testing in a way that purely correlation-based methods cannot. Researchers often try to enhance robustness by pairing MR with colocalization to reduce confounding due to LD or pleiotropy. Nevertheless, each assumption still requires careful validation to determine if MR is a suitable approach for the trait of interest.

TWAS

Transcriptome-wide association studies (TWAS) focus on linking genetically regulated gene expression to trait outcomes. By utilizing eQTL data, TWAS aggregates the effects of multiple SNPs within each gene, effectively grouping genetic variants and shifting the analysis from millions of individual SNPs to thousands of genes. This gene-centric approach significantly reduces the multiple-testing burden by limiting the number of statistical tests to the number of genes rather than the vast number of SNPs. Additionally, TWAS provides a more mechanistic understanding of how non-coding variants influence phenotypes by directly associating gene expression levels with traits.
Over time, TWAS methods have expanded from focusing primarily on (proximal) cis-eQTLs and linear regression imputation methods, to incorporating (distal) trans-eQTLs, multi-omic data, and advanced statistical or deep learning-based imputation. These developments are promising but are computationally more demanding and require further validation. However, as these frameworks mature, they promise to deepen our understanding of disease etiology by identifying specific genes or pathways that drive complex traits, thereby guiding targeted experiments and therapeutic interventions.

Single-Cell Developments

Single-cell approaches offer unprecedented resolution for understanding how non-coding variants influence complex traits by uncovering cell-type-specific regulatory mechanisms. Unlike traditional bulk-tissue analyses, which can obscure cellular heterogeneity, single-cell

RNA sequencing and related technologies enable the mapping of genetic variants to distinct cell populations and states.
However, additional developments are needed to fully harness their potential for variant interpretation. Building comprehensive cell-type atlases in disease-relevant tissues and refining computational pipelines for integrating large-scale multi-omic data are essential[58[52]. Additionally, addressing trans-acting regulatory effects and cell-state transitions will further enhance our ability to pinpoint how non-coding SNPs modulate gene expression[54][55].
Nonetheless, the trajectory is clear: single-cell transcriptomics appears poised to become an indispensable component of GWAS pipelines, from preliminary variant discovery to functional follow-up. As single-cell tools integrate with Fine-mapping, Colocalization, MR, and TWAS, researchers could be able to pinpoint not only which variants matter, but where and when they act. This cellular-level resolution promises more comprehensive mechanistic insights, guiding experimental validations and accelerating the translation of GWAS findings into biomedical applications.

Multi-omic Integration

A first major trend in GWAS interpretation is the growing use of multi-omic data, epigenomic, transcriptomic, chromatin accessibility, and, increasingly, single-cell datasets. These layers help contextualize how non-coding SNPs disrupt regulatory elements or transcription factor binding sites, shedding light on the possible mechanisms through which a genetic signal translates into a phenotypic effect. Large-scale biobanks and consortia have been pivotal in enabling such integrative analyses, yet they introduce computational and data-harmonization challenges that remain non-trivial. More recently, single-cell transcriptomics has begun to reveal how risk variants exert effects in specific cell populations or states, improving the mechanistic understanding of pleiotropy and highlighting novel therapeutic targets.

Methodological Integration

A second trend is the increasing tendency to combine aforementioned methods, Fine-mapping, Colocalization, MR, and TWAS, into a more comprehensive pipeline. By leveraging their complementary strengths and compensating for each other's weaknesses, researchers aim to refine variant prioritization, test putative causal pathways, and identify candidate genes for further investigation. For instance, Fine-mapping and Colocalization could narrow down candidate variants most likely to modulate gene function, while MR could test whether such modulation is plausibly causal. TWAS, in turn, could clarify which gene(s) within a locus, if dysregulated, might meaningfully alter disease risk.
Such a multi-method approach could seem more akin to "throwing the kitchen sink" at the data, rather than a carefully thought-out analytical pipeline. Nonetheless, it exemplifies a promising, possibly sophisticated, strategy where different lines of evidence are used to reinforce one another. Thereby guiding the development of biomarker panels or potential therapeutic interventions.

Limitations and Future Directions

Despite these methodological advances, GWAS remains a correlation-based endeavor. Although MR and related techniques aim to approximate causal inferences, each method carries assumptions that might be violated in polygenic diseases with complex biological pathways. Violations of these assumptions (e.g. unaccounted-for pleiotropic pathways) can lead to incorrect predictions of SNP-trait relationships.

Beyond methodological refinements, collaborative efforts that expand reference panels to more diverse populations, incorporate underrepresented tissues, and standardize large-scale data integration will be pivotal. Such efforts could accelerate personalized medicine, guiding biomarker discovery, targeted therapies, and patient stratification.

No single breakthrough in computational methodology will instantly solve all the challenges inherent in mapping non-coding variants to phenotypes. However, each incremental improvement, whether an advanced Bayesian fine-mapping technique, a next-generation colocalization framework, a refined MR design, or a more powerful TWAS method, brings us closer to realizing the clinical potential of GWAS. Meanwhile, single-cell assays stand poised to reveal additional layers of regulatory complexity, highlighting how cell-state–dependent effects shape gene expression and disease risk.

Conclusion

In summary, the approaches outlined in this review collectively address core obstacles in interpreting GWAS findings, notably the difficulty in identifying causally relevant variants among thousands of correlated non-coding SNPs. By combining multi-omic data, single-cell resolution, and innovative statistical frameworks, these methods help transform broad associations into more refined biological insights. Nonetheless, significant challenges remain, including the inherent constraints of correlation-based approaches, the complexities of integrating heterogeneous datasets, and an evolving understanding of non-coding regulatory elements. As these bioinformatic tools continue to converge and mature, they set the stage for more mechanistically informed, context-specific investigations of complex traits, paving the way toward deeper biological insight and eventual clinical applications.

Addendum: GWAS in the Context of Disease Biology and Clinical Therapies

To broaden the methodological focus of this review, the following section situates GWAS findings within the contexts of disease biology and clinical application. It highlights key successes achieved through genome-wide association studies, along with remaining challenges in translating genetic insights into actionable therapeutic strategies.

Many of GWAS's most notable achievements relate to drug repurposing and rapid therapeutic development. For instance, the association of *IL23R* variants with Crohn's disease led to the repurposing of anti-IL-23 treatments, originally developed for psoriasis, to effectively treat Crohn's disease[3]. Similarly, *TYK2* variants linked to severe COVID-19 informed the successful reuse of baricitinib, reducing mortality by approximately 20%. This example illustrates how GWAS can expedite clinical interventions under urgent circumstances[3]. Beyond drug repurposing, GWAS-based insights also guide gene-editing solutions for conditions like hypercholesterolemia, where pinpointed *PCSK9* variants have led to interventions such as VERVE-101, targeting *PCSK9* for long-term LDL cholesterol reduction[3]. These strategies demonstrate that once strong evidence supports a gene's causal role, often confirmed through larger, more diverse biobanks and functional follow-up, targeted therapies can be developed.
Polygenic risk scores (PRS), derived by aggregating the effects of numerous modest-effect variants, have further expanded GWAS's clinical relevance[3]. GWAS data is used to refine PRS for conditions like cardiovascular disease, diabetes, and certain cancers. Early identification of individuals with elevated polygenic risk offers opportunities for earlier interventions or more targeted surveillance, thereby advancing personalized preventive medicine.

Despite these milestones, several gaps remain in translating GWAS-based insights into robust, clinically actionable outcomes. First, many GWAS associations map to non-coding regions, and even the most sophisticated Fine-Mapping or Colocalization approaches cannot fully elucidate which variants exert their effects, through which pathways, or under what developmental or environmental conditions[60]. Second, GWAS inherently emphasize genomic factors without fully accounting for epigenetic mechanisms, cellular heterogeneity, and environmental influences[60][59]. In complex diseases like cancer or neurodegenerative disorders, epigenetic modifications can reprogram cells, while tissue-specific or cell-state changes create heterogeneity that bulk GWAS measurements often mask[10][11][59].
A growing body of literature highlights that disease progression and patient outcomes are driven by an interplay of genomic, epigenomic, environmental, and immunological factors[60][59][10][11]. In many conditions, factors like the microenvironment, inflammatory cues, or microbiome composition significantly influence gene regulation and disease progression. Furthermore, single-cell analyses reveal that individual cells within the same tissue can differ dramatically in gene expression or regulatory states, leading to variable drug responses or treatment resistance[59]. This heterogeneity is particularly relevant for precision medicine, where subpopulations of tumor or diseased cells may escape therapies designed based on bulk genomic features alone. Incorporating single-cell transcriptomics and multi-omic data into GWAS pipelines can thus refine the interpretation of risk variants, pinpoint the cell types they act upon, and design multi-targeted interventions that address both genetic and non-genetic drivers of disease.

Taken together, recent methodological advances in GWAS should be viewed not as an endpoint but as a foundation for more integrative research programs. By placing GWAS within a broader framework that includes epigenetic factors, environmental influences, and the heterogeneity of human tissues, researchers can move beyond simple variant-to-disease associations. This expanded view could clarify how genotypes interface with diverse cell states, immune responses, and dynamic environmental exposures. As multi-omic studies grow, encompassing epigenomic profiling, proteomic analyses, and single-cell atlases, GWAS findings can be more deeply contextualized within the biological networks that drive disease. This comprehensive perspective could inform targeted therapies that address both genetic pathways and the factors modulating them in specific cellular contexts. Ultimately, merging high-dimensional genomic data with functional, single-cell, and environmental insights holds great promise for unraveling disease complexity and fueling the next generation of personalized interventions.

Note on the use of GenAI

For this writing assignment I used GenAI to assist with generating ideas and structure content, find references, summarize papers, and provide feedback on and improve writing.
Tool references:
OpenAI ChatGPT, version GPT-4o and GPT-o1, May and September 2024.

References

1. Cerezo, M., Sollis, E., Ji, Y., Lewis, E., Abid, A., Bircan, K. O., Hall, P., Hayhurst, J., John, S., Mosaku, A., Ramachandran, S., Foreman, A., Ibrahim, A., McLaughlin, J., Pendlington, Z., Stefancsik, R., Lambert, S. A., McMahon, A., Morales, J., … Harris, L. W. (2024). The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkae1070
2. Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 59. https://doi.org/10.1038/s43586-021-00056-9
3. Abdellaoui, A., Yengo, L., Verweij, K. J. H., & Visscher, P. M. (2023). 15 years of GWAS discovery: Realizing the promise. *The American Journal of Human Genetics*, *110*(2), 179–194. https://doi.org/10.1016/j.ajhg.2022.12.011
4. Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., … Stamatoyannopoulos, J. A. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, *337*(6099), 1190–1195. https://doi.org/10.1126/science.1222794
5. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012). https://doi.org/10.1038/nature11247
6. Javierre, B. M., Burren, O. S., Wilder, S. P., Kreuzhuber, R., Hill, S. M., Sewitz, S., Cairns, J., Wingett, S. W., Várnai, C., Thiecke, M. J., Burden, F., Farrow, S., Cutler, A. J., Rehnström, K., Downes, K., Grassi, L., Kostadima, M., Freire-Pritchett, P., Wang, F., … Fraser, P. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, *167*(5), 1369-1384.e19. https://doi.org/10.1016/j.cell.2016.09.037
7. Yang, W., Zhang, T., Song, X., Dong, G., Xu, L., & Jiang, F. (2022). SNP-Target Genes Interaction Perturbing the Cancer Risk in the Post-GWAS. *Cancers*, *14*(22), 5636. https://doi.org/10.3390/cancers14225636
8. Bonev, B., & Cavalli, G. (2016). Organization and function of the 3D genome. *Nature Reviews Genetics*, *17*(11), 661–678. https://doi.org/10.1038/nrg.2016.112
9. Castellanos-Rubio, A., & Ghosh, S. (2022). Functional Implications of Intergenic GWAS SNPs in Immune-Related LncRNAs. In S. Carpenter (Ed.), *Long Noncoding RNA: Mechanistic Insights and Roles in Inflammation* (pp. 147–160). Springer International Publishing. https://doi.org/10.1007/978-3-030-92034-0_8
10. Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. *Cancer Discovery*, *12*(1), 31–46. https://doi.org/10.1158/2159-8290.CD-21-1059
11. Wilson, D. M., Cookson, M. R., van den Bosch, L., Zetterberg, H., Holtzman, D. M., & Dewachter, I. (2023). Hallmarks of neurodegenerative diseases. *Cell*, *186*(4), 693–714. https://doi.org/10.1016/j.cell.2022.12.032
12. Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., Sakaue, S., Graff, M., Eliasen, A. U., Jiang, Y., Raghavan, S., Miao, J., Arias, J. D., Graham, S. E., Mukamel, R. E., Spracklen, C. N., Yin, X., Chen, S.-H., Ferreira, T., Highland, H. H., … Hirschhorn, J. N. (2022). A saturated map of common genetic variants associated with human height. *Nature*, *610*(7933), 704–712. https://doi.org/10.1038/s41586-022-05275-y
13. de Bakker, P. I. W., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., & Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nature Genetics*, *37*(11), 1217–1223. https://doi.org/10.1038/ng1669
14. Kang, J. T. L., & Rosenberg, N. A. (2020). Mathematical Properties of Linkage Disequilibrium Statistics Defined by Normalization of the Coefficient D = pAB – pApB. *Human Heredity*, *84*(3), 127–143. https://doi.org/10.1159/000504171

15. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., & Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, *14*(7), 483–495. https://doi.org/10.1038/nrg3461
16. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., & Eskin, E. (2014). Identifying Causal Variants at Loci with Multiple Signals of Association. *Genetics*, *198*(2), 497–508. https://doi.org/10.1534/genetics.114.167908
17. Mancuso, N., Freund, M. K., Johnson, R., Shi, H., Kichaev, G., Gusev, A., & Pasaniuc, B. (2019). Probabilistic fine-mapping of transcriptome-wide association studies. *Nature Genetics*, *51*(4), 675–682. https://doi.org/10.1038/s41588-019-0367-1
18. Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genetics*, *10*(5), e1004383. https://doi.org/10.1371/journal.pgen.1004383
19. Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., & Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, *27*(8), 1133–1163.
20. Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. v, Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., Im, H. K., & Consortium, Gte. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, *47*(9), 1091–1098. https://doi.org/10.1038/ng.3367
21. Li, B., & Ritchie, M. D. (2021). From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries. *Frontiers in Genetics*, *12*. https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2021.713230
22. Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, *19*(8), 491–504. https://doi.org/10.1038/s41576-018-0016-z
23. Benner, C., Spencer, C. C. A., Havulinna, A. S., Salomaa, V., Ripatti, S., & Pirinen, M. (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, *32*(10), 1493–1501. https://doi.org/10.1093/bioinformatics/btw018
24. Broekema, R. v., Bakker, O. B., & Jonkers, I. H. (2020). A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biology*, *10*(1), 190221. https://doi.org/10.1098/rsob.190221
25. Wu, Y., Zheng, Z., Thibaut, L., Goddard, M. E., Wray, N. R., Visscher, P. M., & Zeng, J. (2024). Genome-wide fine-mapping improves identification of causal variants. *MedRxiv*, 2024.07.18.24310667. https://doi.org/10.1101/2024.07.18.24310667
26. Wu, C., & Pan, W. (2020). A powerful fine-mapping method for transcriptome-wide association studies. *Human Genetics*, *139*(2), 199–213. https://doi.org/10.1007/s00439-019-02098-2
27. Liu, L., Yan, R., Guo, P., Ji, J., Gong, W., Xue, F., Yuan, Z., & Zhou, X. (2024). Conditional transcriptome-wide association study for fine-mapping candidate causal genes. *Nature Genetics*, *56*(2), 348–356. https://doi.org/10.1038/s41588-023-01645-y
28. Zhang, H., He, K., Li, Z., Tsoi, L. C., & Zhou, X. (2024). FABIO: TWAS fine-mapping to prioritize causal genes for binary traits. *PLOS Genetics*, *20*(12), e1011503. https://doi.org/10.1371/journal.pgen.1011503
29. Cui, R., Elzur, R. A., Kanai, M., Ulirsch, J. C., Weissbrod, O., Daly, M. J., Neale, B. M., Fan, Z., & Finucane, H. K. (2024). Improving fine-mapping by modeling infinitesimal effects. *Nature Genetics*, *56*(1), 162–169. https://doi.org/10.1038/s41588-023-01597-3
30. Zuber, V., Grinberg, N. F., Gill, D., Manipur, I., Slob, E. A. W., Patel, A., Wallace, C., & Burgess, S. (2022). Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *The American Journal of Human Genetics*, *109*(5), 767–782. https://doi.org/10.1016/j.ajhg.2022.04.001

31. Zang, K., Brossard, M., Wilson, T., Ali, S. A., & Espin-Garcia, O. (2024). A scoping review of statistical methods to investigate colocalization between genetic associations and microRNA expression in osteoarthritis. *Osteoarthritis and Cartilage Open*, *6*(4), 100540. https://doi.org/https://doi.org/10.1016/j.ocarto.2024.100540
32. Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., & Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *The American Journal of Human Genetics*, *99*(6), 1245–1260. https://doi.org/10.1016/j.ajhg.2016.10.003
33. Wen, X., Pique-Regi, R., & Luca, F. (2017). Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLOS Genetics*, *13*(3), e1006646. https://doi.org/10.1371/journal.pgen.1006646
34. Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., Pickrell, J., Jaffe, A. E., Pasaniuc, B., & Roussos, P. (2018). A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics*, *34*(15), 2538–2545. https://doi.org/10.1093/bioinformatics/bty147
35. Bennett, D. A., & Holmes, M. v. (2017). Mendelian randomisation in cardiovascular research: an introduction for clinicians. *Heart*, *103*(18), 1400. https://doi.org/10.1136/heartjnl-2016-310605
36. Lin, Z., Xue, H., & Pan, W. (2023). Combining Mendelian randomization and network deconvolution for inference of causal networks with GWAS summary data. *PLOS Genetics*, *19*(5), e1010762. https://doi.org/10.1371/journal.pgen.1010762
37. Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., Torres, J. M., Torstenson, E. S., Shah, K. P., Garcia, T., Edwards, T. L., Stahl, E. A., Huckins, L. M., Aguet, F., Ardlie, K. G., Cummings, B. B., Gelfand, E. T., Getz, G., Hadley, K., Handsaker, R. E., … Im, H. K. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications*, *9*(1), 1825. https://doi.org/10.1038/s41467-018-03621-1
38. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., Jansen, R., de Geus, E. J. C., Boomsma, D. I., Wright, F. A., Sullivan, P. F., Nikkola, E., Alvarez, M., Civelek, M., Lusis, A. J., Lehtimäki, T., Raitoharju, E., Kähönen, M., Seppälä, I., … Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, *48*(3), 245–252. https://doi.org/10.1038/ng.3506
39. Nagpal, S., Meng, X., Epstein, M. P., Tsoi, L. C., Patrick, M., Gibson, G., de Jager, P. L., Bennett, D. A., Wingo, A. P., Wingo, T. S., & Yang, J. (2019). TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. *The American Journal of Human Genetics*, *105*(2), 258–266. https://doi.org/10.1016/j.ajhg.2019.05.018
40. Parrish, R. L., Gibson, G. C., Epstein, M. P., & Yang, J. (2022). TIGAR-V2: Efficient TWAS tool with nonparametric Bayesian eQTL weights of 49 tissue types from GTEx V8. *Human Genetics and Genomics Advances*, *3*(1), 100068. https://doi.org/10.1016/j.xhgg.2021.100068
41. Luningham, J. M., Chen, J., Tang, S., de Jager, P. L., Bennett, D. A., Buchman, A. S., & Yang, J. (2020). Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics. *The American Journal of Human Genetics*, *107*(4), 714–726. https://doi.org/10.1016/j.ajhg.2020.08.022
42. Head, S. T., Dezem, F., Todor, A., Yang, J., Plummer, J., Gayther, S., Kar, S., Schildkraut, J., & Epstein, M. P. (2024). Cis- and trans-eQTL TWASs of breast and ovarian cancer identify more than 100 susceptibility genes in the BCAC and OCAC consortia. *The American Journal of Human Genetics*, *111*(6), 1084–1099. https://doi.org/10.1016/j.ajhg.2024.04.012
43. Yao, S., Wu, H., Liu, T.-T., Wang, J.-H., Ding, J.-M., Guo, J., Rong, Y., Ke, X., Hao, R.-H., Dong, S.-S., Yang, T.-L., & Guo, Y. (2021). Epigenetic Element-Based Transcriptome-Wide Association Study Identifies Novel Genes for Bipolar Disorder. *Schizophrenia Bulletin*, *47*(6), 1642–1652. https://doi.org/10.1093/schbul/sbab023

44. Bhattacharya, A., Li, Y., & Love, M. I. (2021). MOSTWAS: Multi-Omic Strategies for Transcriptome-Wide Association Studies. *PLOS Genetics*, *17*(3), e1009398. https://doi.org/10.1371/journal.pgen.1009398
45. Zhang, Z., Bae, Y. E., Bradley, J. R., Wu, L., & Wu, C. (2022). SUMMIT: An integrative approach for better transcriptomic data imputation improves causal gene identification. *Nature Communications*, *13*(1), 6336. https://doi.org/10.1038/s41467-022-34016-y
46. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., Brugge, H., Oelen, R., de Vries, D. H., van der Wijst, M. G. P., Kasela, S., Pervjakova, N., Alves, I., Favé, M.-J., Agbessi, M., … Consortium, i2QTL. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics*, *53*(9), 1300–1310. https://doi.org/10.1038/s41588-021-00913-z
47. Dai, Q., Zhou, G., Zhao, H., Võsa, U., Franke, L., Battle, A., Teumer, A., Lehtimäki, T., Raitakari, O. T., Esko, T., Agbessi, M., Ahsan, H., Alves, I., Andiappan, A. K., Arindrarto, W., Awadalla, P., Battle, A., Beutner, F., Jan Bonder, M., … Yang, J. (2023). OTTERS: a powerful TWAS framework leveraging summary-level reference data. *Nature Communications*, *14*(1), 1271. https://doi.org/10.1038/s41467-023-36862-w
48. Wang, A., Tian, P., & Zhang, Y. D. (2024). TWAS-GKF: a novel method for causal gene identification in transcriptome-wide association studies with knockoff inference. *Bioinformatics*, *40*(8). https://doi.org/10.1093/bioinformatics/btae502
49. He, Z., Liu, L., Belloy, M. E., le Guen, Y., Sossin, A., Liu, X., Qi, X., Ma, S., Gyawali, P. K., Wyss-Coray, T., Tang, H., Sabatti, C., Candès, E., Greicius, M. D., & Ionita-Laza, I. (2022). GhostKnockoff inference empowers identification of putative causal variants in genome-wide association studies. *Nature Communications*, *13*(1), 7209. https://doi.org/10.1038/s41467-022-34932-z
50. Parrish, R. L., Buchman, A. S., Tasaki, S., Wang, Y., Avey, D., Xu, J., de Jager, P. L., Bennett, D. A., Epstein, M. P., & Yang, J. (2024). SR-TWAS: leveraging multiple reference panels to improve transcriptome-wide association study power by ensemble machine learning. *Nature Communications*, *15*(1), 6646. https://doi.org/10.1038/s41467-024-50983-w
51. Ramprasad, P., Ren, J., & Pan, W. (2024). Enhancing Gene Expression Predictions Using Deep Learning and Functional Annotations. *Genetic Epidemiology*. https://doi.org/10.1002/gepi.22595
52. Zhous, Y., Adeluwa, T., Zhu, L., Salazar-Magaña, S., Sumner, S., Kim, H., Gona, S., Nyasimi, F., Kulkarni, R., Powell, J., Madduri, R., Liu, B., Chen, M., & Im, H. K. (2024). scPrediXcan integrates advances in deep learning and single-cell data into a powerful cell-type–specific transcriptome-wide association study framework. *BioRxiv*, 2024.11.11.623049. https://doi.org/10.1101/2024.11.11.623049
53. Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., & Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, *18*(10), 1196–1203. https://doi.org/10.1038/s41592-021-01252-x
54. Jia, P., Hu, R., Yan, F., Dai, Y., & Zhao, Z. (2022). scGWAS: landscape of trait-cell type associations by integrating single-cell transcriptomics-wide and genome-wide association studies. *Genome Biology*, *23*(1), 220. https://doi.org/10.1186/s13059-022-02785-w
55. Zhang, M. J., Hou, K., Dey, K. K., Sakaue, S., Jagadeesh, K. A., Weinand, K., Taychameekiatchai, A., Rao, P., Pisco, A. O., Zou, J., Wang, B., Gandal, M., Raychaudhuri, S., Pasaniuc, B., & Price, A. L. (2022). Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. *Nature Genetics*, *54*(10), 1572–1580. https://doi.org/10.1038/s41588-022-01167-z
56. Townsend, H., Rosenberger, K., Vanderlinden, L., Inamo, J., & Zhang, F. (2024). Single-cell based integrative analysis of transcriptomics and genetics reveals robust associations and

complexities for inflammatory diseases. *BioRxiv*, 2024.06.17.599349. https://doi.org/10.1101/2024.06.17.599349

57. Mai, J., Qian, Q., Gao, H., Fan, Z., Zeng, J., & Xiao, J. (2024). scTWAS Atlas: an integrative knowledgebase of single-cell transcriptome-wide association studies. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkae931
58. Ding, R., Wang, Q., Gong, L., Zhang, T., Zou, X., Xiong, K., Liao, Q., Plass, M., & Li, L. (2024). scQTLbase: an integrated human single-cell eQTL database. *Nucleic Acids Research*, *52*(D1), D1010–D1017. https://doi.org/10.1093/nar/gkad781
59. Carter, B., & Zhao, K. (2021). The epigenetic basis of cellular heterogeneity. *Nature Reviews Genetics*, *22*(4), 235–250. https://doi.org/10.1038/s41576-020-00300-0
60. Bailey, K. R., & Cheng, C. (2010). Conference Scene: The Great Debate: Genome-Wide Association Studies in Pharmacogenetics Research, Good or Bad? *Pharmacogenomics*, *11*(3), 305–308. https://doi.org/10.2217/pgs.10.6