Trait Causal Models: from GWAS, perturbations, and omics to predictive models of human biology

Emma Dann[*,1,2], Jeffrey P. Spence[*,3,4], Mineto Ota[‡,5], Jesse M. Engreitz[‡,1,6-9], Alexander Marson[‡,2,3,10-15], and Jonathan K. Pritchard[‡,1,16]

[1]Department of Genetics, Stanford University, [2]Gladstone-UCSF Institute of Genomic Immunology, [3]Institute for Human Genetics, University of California, San Francisco, [4]Department of Epidemiology & Biostatistics, University of California, San Francisco, [5]Department of Allergy and Rheumatology, Graduate School of Medicine, The University of Tokyo, [6]Basic Science and Engineering Initiative, Stanford Children's Health, Betty Irene Moore Children's Heart Center, Stanford, CA, [7]Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Broad Institute of MIT and Harvard, [8]Gene Regulation Observatory, Broad Institute of MIT and Harvard, [9]Stanford Cardiovascular Institute, Stanford University, Stanford, CA, [10]Weill Foundation West Coast Cancer Hub, Stanford, CA, [11]Department of Medicine, University of California, San Francisco, [12]University of California, San Francisco Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, [13]Parker Institute for Cancer Immunotherapy, San Francisco, [14]Innovative Genomics Institute, University of California, Berkeley, [15]Department of Microbiology and Immunology, University of California, San Francisco, [16]Department of Biology, Stanford University

October 15, 2026

Abstract. *A central goal of human genetics is to understand the molecular pathways that link genes to functions in cells and tissues, and ultimately to their roles in health and disease. Genome-wide association studies (GWAS) provide a unique — and powerful — tool for finding causal relationships between variants or genes and human traits, yet it remains challenging to gain mechanistic insights from these associations at genome scale. Here, we argue that recent advances in high-throughput perturbation methods and observational "omics" studies provide valuable complementary data that will be crucial for building the next generation of models in human genetics: formal models linking genes to pathways to cell and tissue biology to traits. These models should be mechanistic, causal, and predictive. We refer to such models as Trait Causal Models.*

Introduction

In 1996, Neil Risch and Kathleen Merikangas wrote a provocative and prescient article, "The Future of Genetic Studies of Complex Human Diseases" [1]. They proposed that in order to understand the biology of complex traits we would need to use what is now known as GWAS. When

* equal contribution
‡ senior author
Correspondence to: emmadann@stanford.edu, jeff.spence@ucsf.edu, pritch@stanford.edu

they wrote that article, it was still a major effort to genotype a few tens of variants in a few hundred individuals, but their vision galvanized the field to build the platforms needed to implement modern GWAS [2–4]. Thirty years later, GWAS — including both traditional genotyping and newer sequencing-based approaches — are routinely used to study complex traits. With new tools available to perturb and measure molecular phenotypes at scale, now is the time to evaluate what GWAS have taught us about the molecular mechanisms underlying the biology of complex traits and what data sources beyond GWAS will drive the next thirty years of discovery.

By many measures, GWAS have been remarkably successful at shaping our understanding of complex traits [5,6]. With genome-wide genotype or sequence data for millions of individuals, we have found hundreds of thousands of replicable associations across thousands of traits [7–9]. For example, in 2022 a landmark study of 5 million people reported a "fully saturated" map of the genetic basis of height, encompassing about 12,000 distinct significant loci spanning approximately 20% of the genome that together explain essentially all of the genetic variance of height [10].

GWAS have led directly to new life-saving drugs [11–13] and gene-therapies [14,15], delivered new insights into disease mechanisms [16–19], and helped clarify the cell types driving certain diseases [20–23]. GWAS have enabled polygenic prediction tools that can impact decisions about medication and medical screening [24–26], and genetic associations are a valuable predictor for which drug development programs will succeed [27].

And yet, even with these important successes, one might still feel just a little disquiet when considering what insights we gain from a well-powered GWAS. In particular, it remains difficult to move from GWAS signals to a systems-level understanding of trait biology. What specific gene regulatory cascades connect variant effects to genes, cell and tissue functions, and ultimately to disease onset [28, 29]? Can we *predict* the *quantitative* effects of variants: for example, does decreasing expression of a particular gene increase or decrease blood pressure and by how much? If we observe hundreds, or even thousands, of independently associated variants for a trait, surely it should be possible to develop a unified model that organizes these into coherent effects on specific pathways in specific cell types.

Instead, despite the wealth of information in a well-powered GWAS, we are often left with mainly qualitative results in genome-scale analyses, such as enrichment of association in broadly-defined gene sets, with mechanistic understanding for only a small fraction of the loci, and with systems-level models of disease risk out of reach.

Thus, we currently sit at a point where GWAS have been remarkably successful at finding associations and for important downstream applications. And yet, there still remains a huge untapped potential for the discovery of mechanistic biology. Fundamentally, GWAS identify the causal effects of variants or genes on traits, but do not directly tell us how, or why, each gene matters. In this article we argue that the emergence of new types of experimental data — focusing particularly on genome-scale CRISPR perturbations and single-cell omics — can be combined with GWAS data to provide a new interpretative toolkit for the next generation of human biology.

Defining "success" in human genetics: A vision for *Trait Causal Models*

Human genetics seeks to understand the genetic basis of complex traits, but what, concretely, does that mean? We argue that this is a question of building *causal* models that can *predict the effects of perturbations* at the molecular, biochemical, cellular and physiological scales.

To understand the genetic basis of a disease, it is not enough to simply identify the genes, pathways, or tissues involved. While finding the key players is an important first step, associations alone are insufficient to make even the most basic predictions: would increasing the expression of some key gene in a particular cell type increase or decrease risk? Would perturbing that gene result in similar effects to perturbing another disease-associated gene? To claim we "understand" the biology of a trait we must be able to make *directional predictions*.

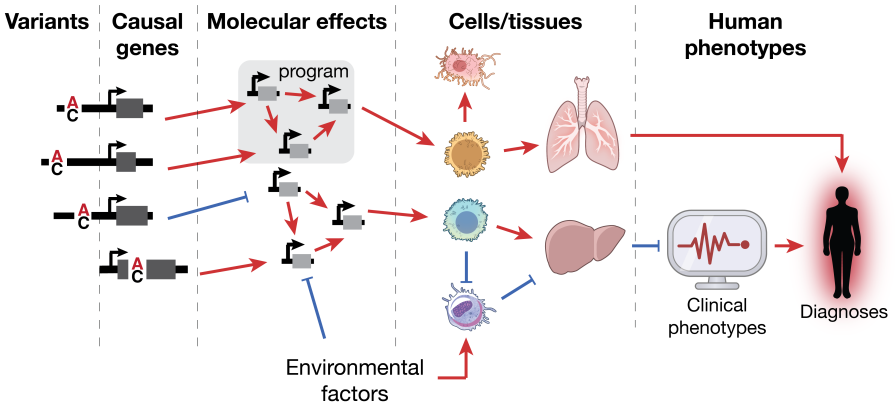


Figure 1: ***Trait Causal Models:*** *Conceptualizing the flow of information through biological systems, from genetic variation to human-level phenotypes.*

Moreover, we need *causal* models that can predict how disease risk will respond to *interventions*. This is in contrast to predictive, but non-causal models such as polygenic scores (PGS) [30]. While PGS can accurately predict individual-level disease risk and can be clinically useful [31], these models do not inherently capture the mechanisms by which variants collectively affect risk, and would not accurately predict the outcome of experimentally altering genotypes. Mechanistic models generalize to new loci, genes, and cells: a model capturing the pathways by which variants affect risk can predict the effect of previously unmeasured variants based on their effects on intermediate biological pathways.

Ultimately, we could claim to fully understand the genetic basis of complex traits if we had predictive causal models that describe how effects flow from genes, through gene regulatory networks, to cell and tissue biology to traits (Figure 1). These "Trait Causal Models" (TCMs) would solve many important downstream problems: identifying drug targets as genes that can be perturbed to reduce disease risk or mitigate disease severity with minimal effects on other traits; guiding pharmacogenomics with models of the pathways a genotype affects; and discovering disease biomarkers by modelling shared pathways affecting distinct traits.

We propose that future genetic studies of complex human diseases should work toward this goal of building TCMs. Progress on both the experimental and computational fronts is making

this goal increasingly tangible. In the remainder of this article we outline what is necessary to make progress toward this vision.

Data Pillars

Causal effects in science are often estimated by experimentally perturbing a system [32], but this is infeasible or unethical in humans outside of expensive, low-throughput clinical trials. Instead, we must combine different data sources, each with their own strengths and weaknesses, to build TCMs (Figure 2).

While there are many potentially useful types of data, for concreteness we focus here on three key *Data Pillars*: 1) biobank-scale whole-genome sequencing data from deeply-phenotyped individuals (for brevity we refer to this as "GWAS data" but include all forms of genomic association studies including conventional SNP-based associations, gene-level burden tests, and molecular QTLs); 2) genome-scale CRISPR perturbations in model systems; and 3) human subjects molecular profiling data measuring molecular correlations with human-level phenotypes, which we refer to as "observational omics". We illustrate potential examples of data pillars for two disease conditions in Box 1.

Pillar 1: GWAS. Because genotypes are randomized at conception, they are causally upstream from phenotypes (assuming that population- and family-level confounders have been appropriately controlled [33]). This makes GWAS data unique in providing genome-wide *causal* information about how variants and genes affect *human-level* traits.

However, GWAS have two main shortcomings. First, while a GWAS association tells us that a variant or gene causally affects a human-level trait, without further data it does not tell us which biological processes lie between the gene and the trait. Second, natural selection on trait variation strongly biases what we discover in GWAS, often making it harder to discover the variants with the largest effects and variants that affect many traits [34–37].

Pillar 2: Experimental perturbations. Modern perturbation experiments provide a powerful tool for filling the interpretative gaps in GWAS. Recent technical advances now enable CRISPR-based perturbation of thousands of genes in pooled screens with high-content phenotypic readouts such as gene expression [38–48] or cell imaging [49–51]; or in bespoke functional assays [52–56]. Crucially, perturbations allow us to identify which other genes or cellular processes lie causally downstream of any gene of interest. These data are often remarkably useful for delineating gene regulatory networks and cellular pathways [42–44, 57], and we see genome-scale perturbations as an essential building block of TCMs.

In principle, there is also complementary information available in trans-eQTLs or trans-pQTLs for estimating gene regulatory effects [58–61]. While perturbation experiments are necessarily limited to *model systems* (e.g., cells, model organisms, organoids), QTL-based approaches may better capture trait-relevant conditions. However, these recent successes notwithstanding, trans-QTL mapping is often limited by low power and discovery biases due to natural selection [36, 37].

Pillar 3: Observational omics. GWAS identifies causal effects of variants or genes on human-level traits; experimental perturbations measure how genes affect low-level cellular processes, but always in model systems. This leaves a gap that can be filled by observational omics data: How

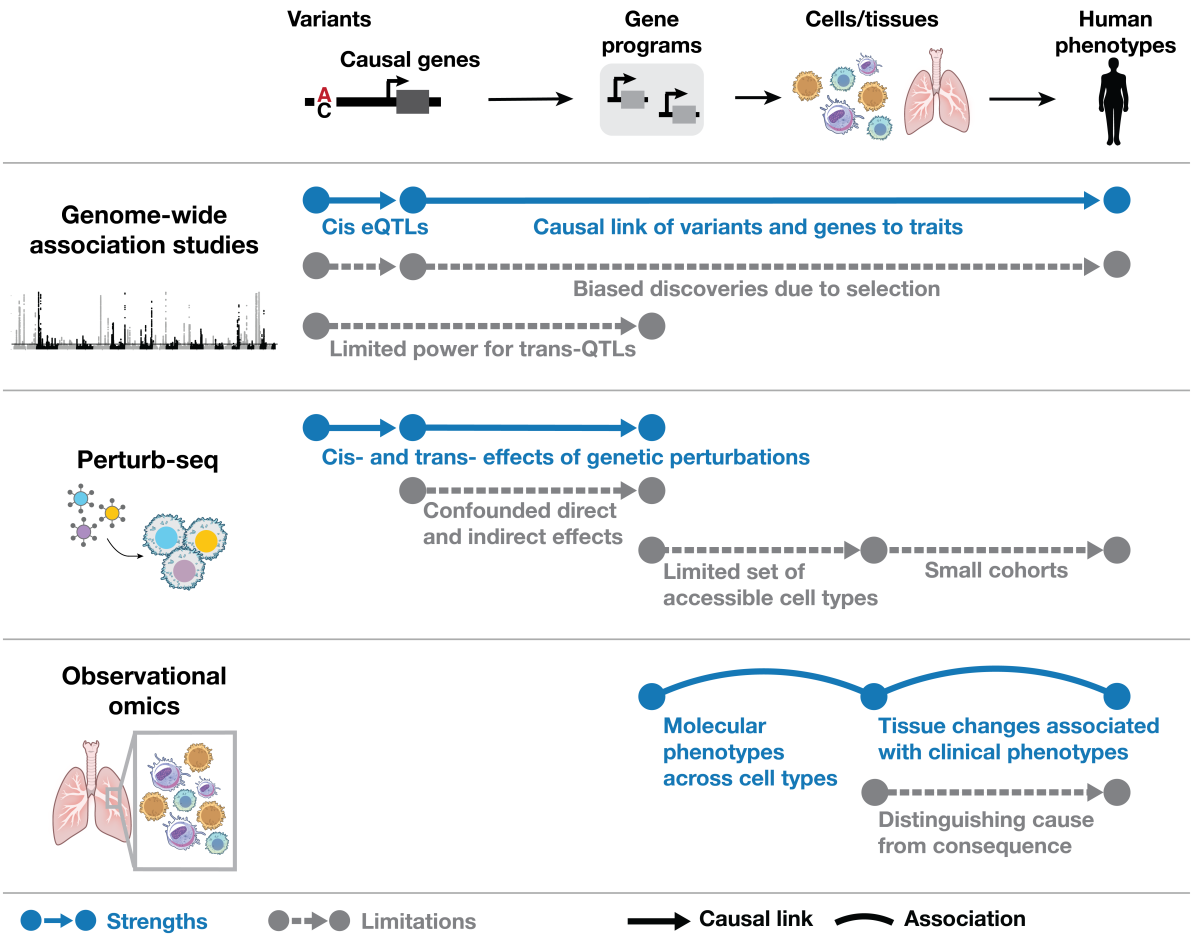


Figure 2: ***Data pillars for Trait Causal Models.*** *Each pillar provides complementary causal or correlative information on connections between biological layers.*

do human subjects — especially disease cases and controls — differ in transcriptional, cellular, and clinical phenotypes, and in which cell types and states might causal variants act?

We define observational omics expansively as the application of any of a range of technologies — for example, RNA-seq, ATAC-seq, spatial transcriptomics; metabolite or proteome measurements in blood or urine; or cellular, tissue and organ imaging — to biobank cohorts or case-control studies to identify which phenotypes are altered in disease states *in humans*. In cohort studies, several phenotypes can be measured in parallel on the same population to test associations between molecular, cellular, and clinical phenotypes.

Unlike GWAS and perturbations, most omics data are fundamentally *observational*. There is no guarantee that detected signatures are causal for disease; they may be co-regulated with disease-relevant pathways, or even be downstream consequences of disease.

Yet, in practice, omics studies have been impactful in nominating disease-associated pathways and the tissues, cell types, and patient subpopulations in which they are active, even with modest sample sizes [78–81]. For example, tissue profiling studies mapped Alzheimer's Disease-

Box 1. Opportunities and gaps: data pillars for *Coronary Artery Disease (CAD)* and *Crohn's Disease*.

To make the discussion more concrete, we consider two complex disorders with different opportunities and challenges: CAD and Crohn's Disease.

Both disorders are impacted by multiple processes acting through different cell types and organ systems. CAD is driven by atherosclerosis, in which damage to the inner lining of the arteries and the buildup of LDL cholesterol in the artery wall trigger the formation of plaques [62]. Crohn's disease is an inflammatory bowel disease driven by dysregulated innate and adaptive immune responses, involving immune cells such as CD4+ T cells (especially Th17), macrophages, and intestinal epithelial cells [63].

Here we provide an overview of key features and data resources currently available for building TCMs for both traits, while acknowledging that important data gaps remain, including a critical need for genome-scale perturb-seq in the relevant cell types:

	CAD	**Crohn's Disease**
Key features:		
Major cell types	Smooth muscle cells, endothelial cells, macrophages, hepatocytes	Macrophages, CD4+ T cells (Th1/Th17), epithelial cells
Access to primary human cells	Challenging (explanted transplant-recipients hearts, *post-mortem*)	Tractable (blood, routine intestinal biopsies)
Example experimental conditions/stimuli	flow models, oxLDL treatment	LPS/TCR stimulation, inflammatory cytokine treatment
Disease-relevant clinical phenotypes	Blood pressure, LDL, triglycerides	autoantibodies, inflammatory markers (C-reactive protein, fecal calprotectin)
Current therapeutic targets	LDL/lipids; vessel-wall inflammation; blood pressure	myeloid/Th17 activation, immune trafficking to gut
Representative Data Pillar Resources:		
GWAS/xQTL	Disease risk [18, 64], BP, LDL [65, 66]; limited cell-type resolved eQTL data [67]	Disease risk [68–70]; immune and intestinal sc-eQTL [71]
Perturb-seq	Endothelial cell model (teloHAEC) [43]; Liver cell line (HepG2) [72], mouse liver *in situ* [73]	CD4+ T cells (genome-scale) [44], mouse macrophages [51], macrophage cell line (THP-1) [74]
Omics	scRNA-seq of cardiovascular tissues [75] and liver [76]; lipids [66], other metabolites [77]	PBMC and intestinal scRNA-seq from disease cohorts [71]

associated genes to microglial function [82], and early transcriptomics work identified mutually exclusive Th2 and Th17 inflammatory pathways in asthma patients [83, 84], informing clinical trials on cytokine receptor blockade. Moreover, unlike GWAS and perturbations, omics data — especially when collected across timepoints — can be informative about the *dynamics* of disease progression through time [85, 86].

Building Trait Causal Models

Figure 1 showed a conceptualization of a trait causal model, with genetic effects flowing through levels of biological structures, ultimately impacting human-level traits. If we have measured the relevant data pillars for a trait, how can we infer this graph in practice?

For inspiration, we turn to a growing literature that jointly analyzes pairs of data pillars to gain biological insights (Figure 3A). Among these, several studies have identified GWAS hit enrichment within "programs" of co-expressed genes that are either identified from perturb-seq [43, 44, 55, 87], or that are differentially expressed in observational data from cases and controls [23, 88–90]. In both analyses, the enrichment of GWAS hits within gene programs suggests that these programs, or related cellular functions, play a causal role in disease. Meanwhile, recent work has used perturb-seq data to identify the upstream regulators of observed gene programs. For example, one can combine differential expression signatures from purely observational data with perturb-seq to nominate master regulators that may explain the signatures [44, 91].

While important biological insights have been found by these approaches, the focus is mainly on using enrichment tests to discover trait-relevant regulators or programs. Instead of finding enrichment, can we adapt and extend these methods to build *predictive* models?

A recent paper from our group developed a prototype TCM (Figure 3B), pairing GWAS data from red blood cell traits with perturb-seq data from the erythroleukemic cell line K562 (Ota et al. 2026 [57]). Following [43], Ota et al. used the perturb-seq data to infer a simplified regulatory network in erythroid precursors consisting of two layers, with effects flowing from perturbed *regulators* to *programs* of measured genes. The assumption is that causal effects flow from *regulators* $\rightarrow$ *programs* $\rightarrow$ *traits*. In this model, we can think of core genes (as defined in [29]) as residing in programs, and controlled by upstream regulators. Genetic variation can impact this network at either the regulator or program levels.

At this stage, instead of identifying programs enriched for GWAS hits, we can obtain a much richer description if we measure the *direction of effects* of genes on traits by using variants with a known direction of effect on gene expression. For example, Ota et al. used loss-of-function mutations for this purpose. In a correct model we should expect that the directional effects of regulators on traits can be explained by their directional effects on programs and the effects of the genes within the program on traits. We can leverage this intuition along with estimates of the *regulator* $\rightarrow$ *program* effects from perturb-seq to decompose *gene* $\rightarrow$ *trait* estimates from GWAS data into a *regulator* $\rightarrow$ *program* $\rightarrow$ *trait* TCM. Importantly, this model also enables quantitative predictions: when Ota et al. held some genes out of the model training, they could use the perturb-seq data to predict the signed effects of the held-out genes on traits with reasonable accuracy. Although the ability to predict unobserved values is an important criterion for evaluating models, this has historically been challenging in human genetics.

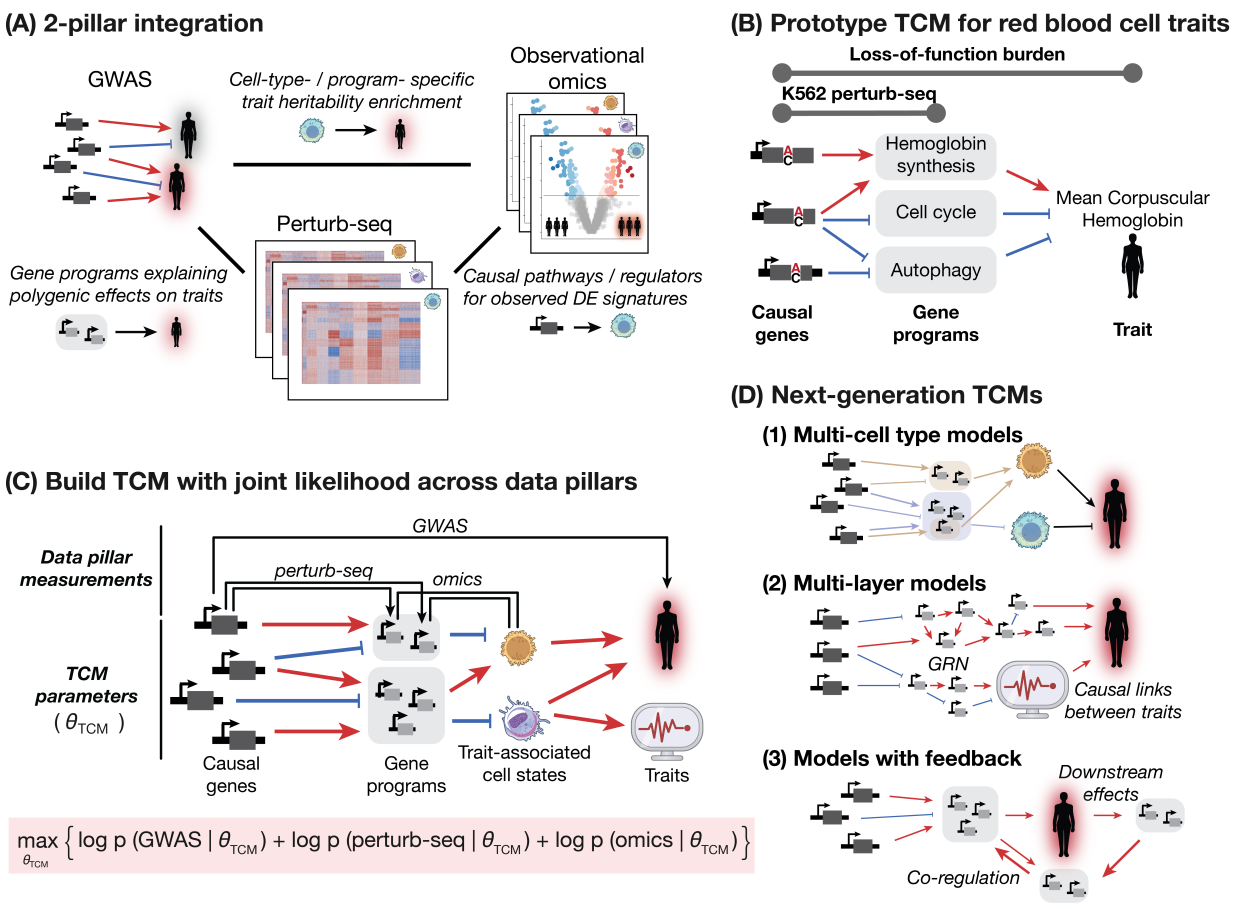


Figure 3: ***Building Trait Causal Models.*** ***(A.)*** *Many existing approaches in functional genomics focus on integration between pairs of data pillars to connect biological layers.* ***(B.)*** *A first generation TCM for red blood cell traits [57] explicitly considered directions of effects within perturbation graphs and for genetic associations;* ***(C.)*** *Example TCM framework incorporating quantitative measurements across 3 data pillars.* ***(D.)*** *Future extensions of the TCM framework could incorporate further aspects of trait biology including:* ***(1.)*** *jointly modeling genetic effects in multiple cell types;* ***(2.)*** *additional layers of regulation between biological entities;* ***(3.)*** *feedback both upstream and downstream of traits.*

We anticipate an expanding opportunity to develop more detailed TCMs for a wide variety of traits as the available perturbation and omics data expand. We expect that future TCMs will aim to provide joint, quantitative modeling of multiple pillars. While this modeling could go in many directions, we suggest one hypothetical framework in Figure 3C. Here, the TCM links genes across layers of a regulatory network between regulators and programs, with activity in multiple trait-relevant cell types. Future inference methods would aim to explain data collected in all three pillars within a single coherent model; one could potentially fit this by minimizing a joint loss function across all three data pillars.

Moreover, this basic framework will likely need to be extended in several key dimensions:

- *Multi-cell type models.* Most traits are controlled by variation in multiple cell types. Thus we

can imagine TCMs with shared or cell type-specific nodes and edges (Figure 3D);

- *Multi-layer graphs.* Regulator-program models are clearly a simplification of both gene regulatory networks, and the layers of complexity at higher levels of structure, including tissues, organs and clinical phenotypes (Figure 3D). Existing techniques such as Mendelian Randomization [92] and matrix factorization of GWAS signals [17, 93, 94] may be valuable for modeling relationships among traits;
- *Models with feedback and co-regulation.* Some molecular processes affect human-level traits, while others are causally downstream of human-level traits. We anticipate TCMs that model this sort of feedback between layers of biological complexity (Figure 3D).

Looking forward, one question is what level of granularity to use for modeling connections within and across biological scales. Gene regulatory networks offer an instructive example. Current approaches for perturb-seq analysis mainly use latent factor models to build simplified representations of the gene regulatory network (GRN) based on programs [38, 43, 95]. These simplified models can be easier to interpret and may require less data to estimate. New approaches come closer to estimating the full GRN [96–98], which may trade interpretability for increased fidelity to the underlying biology.

These considerations raise an important issue: advances in machine learning and AI mean that we will be able to perform both inference and prediction in highly complex graphs. These extremely granular TCMs may prove best for prediction problems like selecting drug targets, but we also view human interpretation as an essential output, especially for design of new experiments and therapeutic translation. This implies a need for either human-interpretable models, or AI tools for presenting interpretable summaries of complex models.

Lastly, we note that to achieve quantitative TCMs, we need to be able to connect very different kinds of measurements, in different biological units, from each pillar. Instead of focusing on significance testing, experiments and analyses should be designed to learn biologically meaningful *parameters* representing connections between biological components, corresponding to edges or groups of edges in the TCM.

In the next section we outline the key challenges that we see for the practical implementation of TCMs.

From Data Pillars to Trait Causal Models: open questions

Analysis methods for *variant* $\rightarrow$ *gene* $\rightarrow$ *trait*. Most signals in GWAS come from noncoding variants that act through cis-regulation of nearby genes. There has been important work on identifying causal variants [99], which genes they regulate [100–102], and the cell types in which causal variants may impact gene expression [21, 23, 90, 103–105].

Much less attention has been paid to estimating the quantitative effects (or even directions of effects) of *genes* on traits. This is in contrast with most perturbation and omics datasets, which measure and perturb gene expression. To build joint, quantitative models of all three pillars, we need to be able to extract corresponding information from GWAS. In practice, this requires

estimating how a variant affects gene expression, β, and then decomposing *variant* $\rightarrow$ *trait* into *variant* $\rightarrow$ *gene* (i.e., β) and *gene* $\rightarrow$ *trait* (which we denote by $\gamma(\beta)$, Figure 4A).

With current data and approaches, estimating β for individual variants is difficult, with loss-of-function (LoF) mutations in protein-coding genes being an important exception. LoFs generally reduce the amount of viable gene product by half, allowing us to interpret their estimated effects on traits as being the effect of halving gene expression on the trait [106], assuming no compensatory upregulation of the functional allele. As such, LoF burden tests have been widely used to quantify the effect of changes in gene expression on traits [11, 57, 107, 108].

These successes notwithstanding, LoFs are generally extremely rare, resulting in LoF burden tests being underpowered even at current sample sizes. Increasing power will require increasing the sizes of sequenced cohorts and combining results via collaborative meta-analyses [109], as well as developing analysis tools that increase statistical efficiency for a given sample size [57,110–112].

Even with these advances, building TCMs will require approaches that use other kinds of variants. First, some genes are under such strong natural selection that there simply are not enough LoF carriers in the world to ever obtain meaningful burden signals [37,113]. Second, there is growing evidence that the downstream effects of changes in gene expression are non-linear [106, 114]. While LoFs are informative about the effects of *decreasing* gene expression on a trait, this may not be informative about the effect of *increasing* gene expression.

For these reasons, it is essential to develop methods to extract more quantitative information from the common, non-coding variants measured in standard GWAS, but new approaches will be needed to overcome the noise in estimates of their effect on gene expression, β. One approach is to pool information across variants, by imagining the relationship between effects on expression and effects on the trait as a *gene dosage response curve* (GDRC; Figure 4B). In principle, β determines where a variant falls along the x-axis of the GDRC and can be estimated using eQTL data, while its y-axis position can be estimated from GWAS. The GDRC can then be estimated by fitting a curve through these points. This is exactly the intuition behind approaches like TWAS [115, 116] and Mendelian Randomization [117] (both assume a linear GDRC), but these methods currently have limited power and high rates of sign-errors [118]. However, we are optimistic that it will soon be possible to resolve these challenges by combining complementary approaches: (1) increasing sample sizes of eQTL studies [71, 119]; (2) measuring β experimentally using experimental tools including massively parallel reporter assays (MPRAs) [120–122], enhancer-targeted CRISPRi [45, 123–125] or base-editing [126–128]; and (3) machine learning-based predictions of variant effects [129–131]. Machine learning approaches currently show mixed results [132–136], but we expect that accuracy will improve with more data and algorithmic advances.

Non-coding variants also offer another key advantage over LoFs: they provide information about which cell types are responsible for mediating observed gene-trait associations. LoFs generally reduce expression across all cell types, making it difficult to pinpoint the cell types in which the reduction in expression matters for the trait. In contrast, non-coding variants can be highly context-specific (Figure 4C) [105, 137–139], and new methods are needed to leverage that information to enable building multi-cell type TCMs.

Perturbations: experimental needs and opportunities. There have been tremendous technological advances in our ability to perturb and measure molecular readouts in single cells, enabling an

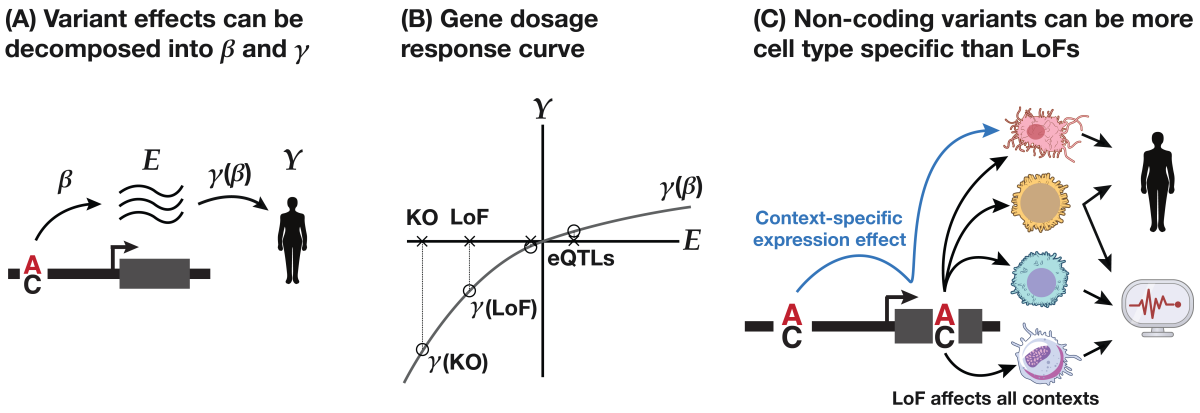


Figure 4: ***Recasting GWAS as parameter estimation.*** ***(A.)*** *The effect of a variant on a trait can be conceptualized as being the result of its effect on a gene, β, and the effect of that gene on the trait, γ.* ***(B.)*** *The gene dosage response curve (GDRC) conceptualizes the relationship between gene expression and expected phenotype.* ***(C.)*** *LoFs generally affect gene function in all cellular contexts while regulatory variants can be more specific.*

incredible array of experimental designs. While many of these designs will surely provide exciting insights into cellular biology, here we ask what perturbation datasets will be most useful for building TCMs.

Which perturbations? For the purpose of building TCMs, the first focus should be on measuring how gene-level perturbations affect the global transcriptome, proteome, and other cellular phenotypes. Gene-level perturbations give the most flexibility to study responses to many types of gene dosage changes, via approaches including CRISPR-KO, CRISPRi, and CRISPRa, as well as stable repression and activation with epigenetic editing in both gene promoters and enhancers [140,141], enabling comparison to GWAS data. Ultimately, we imagine combining multiple perturbation modes to estimate GDRCs relating the expression of pairs of genes [114]. While combinatorial perturbations are useful for elucidating molecular pathways [142,143], single gene perturbations can be scaled genome wide and genetic effects appear to be largely additive in human genetics [144]. Therefore, we see individual gene perturbations as a tractable and powerful starting point for building TCMs.

Which readouts? Single cell RNA-seq is emerging as a versatile and convenient high-dimensional readout (i.e., perturb-seq). The last decade has shown that scRNA-seq is a remarkably useful indicator of cell states, and readouts from perturb-seq will power the inference of trans-regulatory interactions between genes. RNA-seq data has the additional advantages of being able to be analyzed by established quantitative methods [145–148] and providing a natural common coordinate framework (measurements on the same genes) for comparisons to gene-level omics datasets across conditions and cell types.

Despite its many advantages, RNA-seq is an indirect readout of cellular biology. Cellular functions are largely carried out by *proteins*, and changes in protein activity may not be fully reflected in the transcriptome. Cell-cell communication is crucial for many aspects of cell functions but not measured in current perturbation experiments. Transcriptomic responses can also be challeng-

ing to relate to disease-relevant cellular phenotypes, such as exhaustion in T cells or scarring in fibroblasts [149]. To address these issues, transcriptome readouts could be complemented with screens on bespoke functional assays, or imaging-based readouts with Optical Pooled Screening (OPS) [49,50,73,150]. Joint analysis of such experiments with perturb-seq would identify expression signatures linked to functional endophenotypes or non-cell-autonomous effects on tissues.

Which biological contexts? The last critical question is how to select the most relevant model systems. Perturbation studies have employed cellular contexts including cancer cell lines, iPSC-derived lines, organoids, *ex vivo* primary cells, and *in vivo* mouse tissues [42–45,47,73,149,151–153]. These systems vary in terms of convenience, accessibility, and biological fidelity to human tissue. Recent work shows that perturb-seq results are highly dependent on cell types, experimental conditions and even measurement timepoints [42,44,72,154]. Therefore, it is clear that distinct human traits will require distinct experimental systems, each covering all the major cell types involved in the disease and including disease-relevant experimental stimuli. For a given disease, there will be a need for measurements on thousands of perturbations in tractable cells as well as smaller perturbation panels in more biologically faithful systems. Of the axes along which perturb-seq experiments can scale (perturbation type, number of perturbations, readout modality, and biological context), we believe expanding the space of biological contexts is most urgent for building the first generation of TCMs.

Perturbations: prediction for unseen contexts. It is clearly impractical to measure every perturbation with every readout in every context relevant to a given trait. Many disease-relevant cell types are challenging or impossible to culture or edit *ex vivo*, such as non-proliferating cells. Even with remarkable experimental advances, we will realistically be left with a patchwork: dense observations in some contexts, sparse or absent coverage in others. Could we impute, *in silico*, the perturbation effects we cannot measure?

This task has become a popular machine learning challenge in biology in recent years [155,156], proposed as central to building "AI Virtual Cells" [157], and will be key for building TCMs for traits where relevant cell types are experimentally inaccessible. To realize this vision, these models must be trained and evaluated in real-world use-cases. A realistic setting is having access to thousands of perturbations in cell lines, but needing to predict the effect of perturbations in a transcriptionally distant primary cell type. In the best cases there may be a small number of perturbations available in the target cell type, but more often methods will need to rely solely on single-cell RNA-seq, ATAC-seq, and other molecular profiles from the unperturbed baseline in the target cell type.

For now, we remain far from the inflection point in training data needed for accurate perturbation effect prediction across disease-relevant cell types: the joint space of cell contexts, perturbation types, and readouts remains largely unexplored.

Incorporating observational omics into TCMs. Observational omics will be key to building TCMs, confirming that causal pathways predicted from GWAS and perturbation data are active in patients in physiological contexts (Figure 5A). Given the vast toolkit of possible molecular and clinical measurements, which omics data will be most useful for TCMs?

Both *platform omics* and *bespoke assays* will play important roles. Platform omics, such as single-

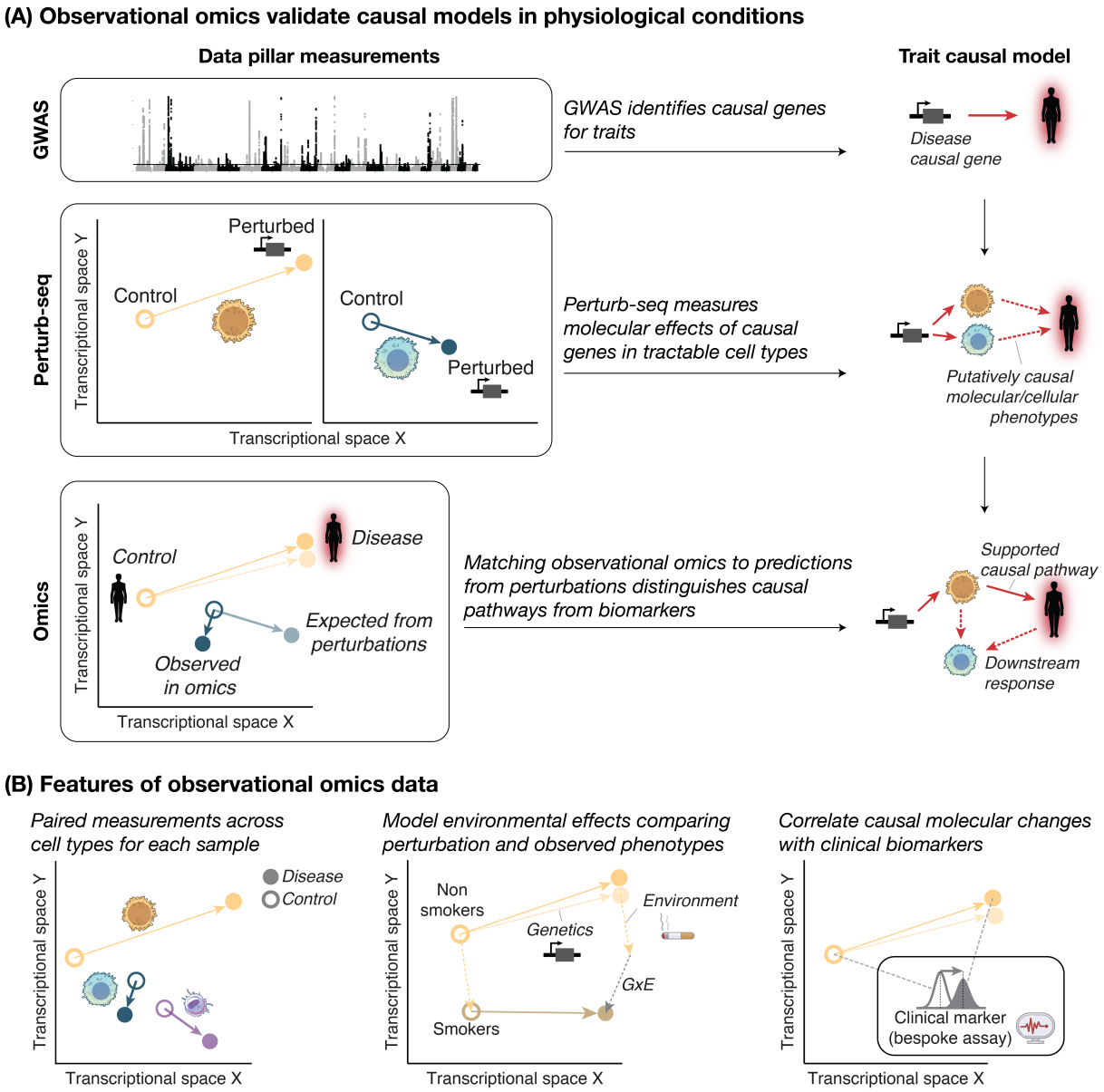


Figure 5: ***Building up TCMs with perturbation and observational omics data.*** ***(A.)*** *Perturb-seq data across perturbation modes quantifies molecular responses to causal gene dosage across biological contexts. This generates hypotheses on the cell types and molecular mechanisms by which causal genes affect traits.* ***(B.)*** *Features of observational omics data that inform TCMs.*

cell RNA-seq and ATAC-seq, work robustly across a range of tissues, and can be matched to the readouts of perturbation experiments, enabling the joint modeling of perturbation effects on gene expression and expression profiles of disease-relevant cell types observed in patients. These data will also provide the substrate to infer quantitative effects of variants on gene expression in disease-relevant cell types and states [101, 131, 158]. At the same time, bespoke assays for any given disease will also be needed to connect molecular changes to functional phenotype. Performing both platform omics and these bespoke assays in the same cohort will enable connecting

gene-level measurements to other layers of disease biology. For example, in an Alzheimer's disease cohort scRNA-seq might be paired with measurements of amyloid plaque deposition, while for autoimmune conditions we might measure inflammatory protein abundance in blood with targeted proteomics panels.

On top of serving as a validation layer, observational omics present several key features to advance TCMs (Figure 5B):

- **Measurements across cell types.** With observational omics we can routinely measure multiple cell types in the same sample. This enables joint modeling of perturb-seq across cell types and provides the measurements needed to predict effect sizes of non-coding variants on gene expression. Multi-cell type data also captures cell-cell interactions and shared responses to the effects of the environment and genetic background.
- **Modeling of environmental effects.** Omics cohorts can be used to understand the effects of environmental exposures on tissue phenotypes and their interaction with cell-intrinsic gene regulatory network effects. Omics studies should maximize the amount of donor-level data collected alongside molecular measurements to facilitate these discoveries.
- **Biomarkers of causal pathways.** Not all phenotypes associated with a trait of interest capture core processes affected by genetic variation: these might include phenotypes that reflect secondary effects contributing to symptoms, protective responses to pathogenesis or response to therapy. While insufficient to nominate causal core processes in isolation, deeply phenotyped omics cohorts can be used to identify proxies of core disease processes that are detectable in patient samples, such as blood biomarker correlates of activity of a causal pathway in a hard to access tissue.

With these goals in mind, what should we estimate from patient omics data to incorporate it into TCMs? The richness and complexity of omics datasets has led to a proliferation of analytical methods, often making it difficult to relate biological findings across diseases and cohorts. We suggest focusing on key parameters that can be estimated and interpreted across cohorts. For example, from scRNA-seq data we can estimate (a) gene expression signatures for each cell type, (b) disease-associated changes in cell type fractions, and (c) disease-associated changes in gene expression profiles for each cell type. These summary statistics are directly comparable to perturb-seq perturbation effect estimates and eQTL effects, and can be meta-analysed across studies and compared across traits.

We believe robust estimates of cell type-specific parameters will power the first generation of trait causal models. We are excited to see how analytical frameworks will evolve to accommodate more complex, multi-modal or time-series datasets, incorporating trait effects on many molecular readouts or modeling longitudinal measurements [159].

Discussion

GWAS have been remarkably successful during the past 20 years, but it still remains difficult to uncover new biological mechanisms at genome scale from GWAS data alone. We argue that modern genome-scale perturbation experiments are a critical tool for revealing the cellular pathways

through which genetic effects act on human traits, and that observational omics can provide direct readouts of these pathways in disease states.

We propose that these different data modalities and views of biology can be unified into Trait Causal Models. TCMs capture quantitative parameters connecting biological layers — from variants to genes to pathways in specific cell types and tissues — and they can predict the effects of perturbations of these entities.

The first generation of TCMs does not need to be built at full resolution to start being useful. We can begin with coarse graphs, then progressively refine them by resolving the intermediate pathways of cellular function, but much work remains to fully realize this promise. We view this as an opportunity for new experimental and computational methods development tackling grand challenges for the field:

1. **Choosing and developing tractable experimental models for genome-scale perturbation studies.** Building TCMs will require new experimental models of disease-relevant human tissues. New methods are needed for determining which models and experimental contexts are trustworthy for which traits and which aspects of TCMs.

2. **Expanded readouts of perturbation responses.** Not all biological processes leave clear patterns in transcriptomic data measured at a single timepoint. To more deeply characterize the functional consequences of perturbations will require additional readouts, including image-based readouts at subcellular and multi-cell/tissue scales, protein measurements, and multiple-timepoint readouts to study dynamic processes.

3. **Predicting the effects of perturbations in unseen contexts.** We view this challenge as synergistic with the previous challenges, and necessary to fill the gaps from disease-relevant cell states that are inaccessible to experimental perturbations.

4. **Realistic inference of biological networks.** Richer perturbation and omics readouts call for methods that estimate more fine-scaled TCMs, such as moving from the level of gene expression programs to fully resolved GRNs [96–98]. These approaches could be extended to additional molecular layers that model signaling pathways or other biochemical processes, or cell–cell interactions.

5. **Quantitative estimates of gene effects on programs, cell types and human-level traits.** We need tools to estimate gene dosage response curves from common variant GWAS and burden tests. Relatedly, frameworks must be developed for estimating biologically meaningful parameters at the level of genes, programs, and cell types from omics data, that are comparable across observational and perturbation studies.

6. **Making TCMs interpretable to humans.** TCMs will be most useful if they are interpretable. In practice, this will mean connecting pathways or programs in TCMs to named pathways and functions. Here we see considerable promise in the use of agentic AI systems to synthesize prior knowledge and model outputs into interpretable insights.

7. **Therapeutic interventions from causal models.** We will need dedicated approaches to identify interventions that can *reverse* a phenotype from disease TCMs. This is complicated by

the fact that most GWAS conflate genetic drivers of disease onset with drivers of progression, which, when studied separately, often prove distinct [160, 161]. Although expensive and low-throughput, clinical trials could be the ultimate test of a TCM's predictive validity: trials pairing clinical endpoints with omics profiling would provide an invaluable resource to model the causal structure of disease progression and response to intervention.

Reflecting back on the past 30 years since Risch and Merikangas laid out their vision for the genetic study of complex human diseases, it is clear that it has been a period of great discovery, in which we have cataloged a huge number of high-confidence causal relationships between variants and traits. Yet, looking toward the future, it is time to ask new questions and use new approaches. We should no longer be looking to simply discover genes and catalog associations. We should be learning disease biology by building detailed, predictive, and mechanistic models of the pathways from genes to disease. This vision is encapsulated in Trait Causal Models. These goals are ambitious, but the data and computational methods are now attainable with a combination of tools, both current and on the near horizon.

Acknowledgments

We thank Alvina Adimoelja, Howard Chang, Jan Engelmann, Tami Gjorgjieva, Nikhil Milind, Tony Zeng, and Julie Zhu for critical feedback on the manuscript, and all members of the Pritchard and Marson labs for helpful discussions. Cell and organ illustrations in figures are from NIAID NIH BioArt Source (https://bioart.niaid.nih.gov). This work was supported by the National Institutes of Health (grants R01HG008140, R01HG014005, to J.K.P.; R01HL175076 to J.P.S.; UM1HG011972, R01HL164811, P01HL180323, R01HG014216 to J.M.E.). E.D. is supported by a Helen Hay Whitney Foundation fellowship and by an EMBO non-stipendiary fellowship. J.M.E. acknowledges support from the Applebaum Foundation. A.M. is a cofounder of Site Tx, Arsenal Biosciences, and Survey Genomics, serves on the boards of directors at Site Tx and Survey Genomics, is a member of the scientific advisory boards of network.bio, Site Tx, Arsenal Biosciences, Cellanome, Survey Genomics, NewLimit, Amgen, and Tenaya, owns stock in network.bio, Arsenal Biosciences, Site Tx, Cellanome, NewLimit, Survey Genomics, Tenaya and Lightcast and has received fees from network.bio, Site Tx, Arsenal Biosciences, Cellanome, Survey Genomics, Spotlight Therapeutics, NewLimit, Abbvie, Gilead, Pfizer, 23andMe, PACT Pharma, Juno Therapeutics, Tenaya, Lightcast, Trizell, Vertex, Merck, Amgen, Genentech, GLG, ClearView Healthcare, AlphaSights, and Rupert Case Management. A.M. is an investor in and informal advisor to Offline Ventures and a client of EPIQ. The Marson laboratory has received research support from the Parker Institute for Cancer Immunotherapy, the Emerson Collective, Arc Institute, Juno Therapeutics, Epinomics, Sanofi, GlaxoSmithKline, Gilead and Anthem and reagents from 10x, Ultima, Genscript, Illumina and Cellanome.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author(s) used Claude Opus 4.8/5.5 by Anthropic for draft editing and Paperclip by GXL for literature search. The author(s) reviewed and edited the output as needed and take full responsibility for the content of the published article.

References

[1] Risch, N. and Merikangas, K. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, 1996.

[2] Ragoussis, J. Genotyping technologies for genetic research. *Annu. Rev. Genomics Hum. Genet.*, 10(1):117–133, 2009.

[3] International HapMap Consortium,. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.

[4] Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J., et al. Genome-wide

association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[5] Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. 10 years of GWAS discovery: Biology, function, and translation. *American Journal of Human Genetics*, 101(1):5–22, 2017.

[6] Claussnitzer, M., Cho, J. H., Collins, R., Cox, N. J., Dermitzakis, E. T., Hurles, M. E., Kathiresan, S., Kenny, E. E., Lindgren, C. M., MacArthur, D. G., et al. A brief history of human disease genetics. *Nature*, 577(7789):179–189, January 2020.

[7] Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1):D977–D985, 2023.

[8] Kurki, M. I., Karjalainen, J., Palta, P., Sipilä, T. P., Kristiansson, K., Donner, K. M., Reeve, M. P., Laivuori, H., and others,. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*, 613(7944):508–518, 2023.

[9] Verma, A., Huffman, J. E., Rodriguez, A., Conery, M., Liu, M., Ho, Y.-L., Kim, Y., Heise, D. A., and others,. Diversity and scale: Genetic architecture of 2068 traits in the VA Million Veteran Program. *Science*, 385(6706):eadj1182, 2024.

[10] Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., Sakaue, S., Graff, M., Eliasen, A. U., Jiang, Y., Raghavan, S., and others,. A saturated map of common genetic variants associated with human height. *Nature*, 610(7933):704–712, 2022.

[11] Cohen, J. C., Boerwinkle, E., Mosley, T. H., and Hobbs, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *New England Journal of Medicine*, 354(12):1264–1272, March 2006.

[12] Duerr, R. H., Taylor, K. D., Brant, S. R., Rioux, J. D., Silverberg, M. S., Daly, M. J., Steinhart, A. H., Abraham, C., Regueiro, M., Griffiths, A., et al. A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science*, 314(5804):1461–1463, 2006.

[13] Trajanoska, K., Bhérer, C., Taliun, D., Zhou, S., Richards, J. B., and Mooser, V. From target discovery to clinical drug development with human genetics. *Nature*, 620(7975):737–745, August 2023.

[14] Bauer, D. E., Kamran, S. C., Lessard, S., Xu, J., Fujiwara, Y., Lin, C., Shao, Z., Canver, M. C., Smith, E. C., Pinello, L., et al. An erythroid enhancer of *BCL11A* subject to genetic variation determines fetal hemoglobin level. *Science*, 342(6155):253–257, 2013.

[15] Frangoul, H., Altshuler, D., Cappellini, M. D., Chen, Y.-S., Domm, J., Eustace, B. K., Foell, J., de la Fuente, J., Grupp, S., Handgretinger, R., et al. CRISPR-Cas9 gene editing for sickle cell disease and beta-thalassemia. *N. Engl. J. Med.*, 384(3):252–260, January 2021.

[16] Sekar, A., Bialas, A. R., De Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., and others,. Schizophrenia risk from complex variation of complement component 4. *Nature*, 530(7589):177–183, 2016.

[17] Smith, K., Deutsch, A. J., Mcgrail, C., Kim, H., Hsu, S., Huerta-Chagoya, A., Mandla, R., Schroeder, P. H., Westerman, K. E., Szczerbinski, L., and others,. Multi-ancestry polygenic mechanisms of type 2 diabetes. *Nat. Med.*, 30(4):1065–1074, 2024.

[18] Aragam, K. G., Jiang, T., Goel, A., Kanoni, S., Wolford, B. N., Atri, D. S., Weeks, E. M., Wang, M., Hindy, G., Zhou, W., et al. Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *Nat. Genet.*, 54(12):1803–1815, December 2022.

[19] Trubetskoy, V., Pardiñas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T. B., Bryois, J., Chen, C.-Y., Dennison, C. A., Hall, L. S., et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature*, 604(7906):502–508, April 2022.

[20] Guerreiro, R., Wojtas, A., Bras, J., Carrasquillo, M., Rogaeva, E., Majounie, E., Cruchaga, C., Sassi, C., Kauwe, J. S., Younkin, S., and others,. TREM2 variants in Alzheimer's disease. *N. Engl. J. Med.*, 368(2):117–127, 2013.

[21] Finucane, H. K., Reshef, Y. A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.-R., Lareau, C., Shoresh, N., et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.*, 50(4):621–629, April 2018.

[22] Skene, N. G., Bryois, J., Bakken, T. E., Breen, G., Crowley, J. J., Gaspar, H. A., Giusti-Rodriguez, P., Hodge, R. D., Miller, J. A., Muñoz-Manchado, A. B., et al. Genetic identification of brain cell types underlying schizophrenia. *Nature Genetics*, 50(6):825–833, 2018.

[23] Jagadeesh, K. A., Dey, K. K., Montoro, D. T., Mohan, R., Gazal, S., Engreitz, J. M., Xavier, R. J., Price, A. L., and Regev, A. Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-sequencing and human genetics. *Nat. Genet.*, 54(10):1479–1492, October 2022.

[24] Khera, A. V., Emdin, C. A., Drake, I., Natarajan, P., Bick, A. G., Cook, N. R., Chasman, D. I., Baber, U., Mehran, R., Rader, D. J., et al. Genetic risk, adherence to a healthy lifestyle, and coronary disease. *New England Journal of Medicine*, 375(24):2349–2358, 2016.

[25] Natarajan, P., Young, R., Stitziel, N. O., Padmanabhan, S., Baber, U., Mehran, R., Sartori, S., Fuster, V., Reilly, D. F., Butterworth, A., et al. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation*, 135(22):2091–2101, 2017.

[26] Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J. P., Chen, T.-H., and others,. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *American Journal of Human Genetics*, 104(1):21–34, 2019.

[27] Minikel, E. V., Painter, J. L., Dong, C. C., and Nelson, M. R. Refining the impact of genetic evidence on clinical success. *Nature*, 629(8012):624–629, 2024.

[28] Boyle, E. A., Li, Y. I., and Pritchard, J. K. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169(7):1177–1186, June 2017.

[29] Liu, X., Li, Y. I., and Pritchard, J. K. Trans effects on gene expression can drive omnigenic inheritance. *Cell*, 177(4):1022–1034.e6, May 2019.

[30] Lambert, S. A., Gil, L., Jupp, S., Ritchie, S. C., Xu, Y., Buniello, A., McMahon, A., Abraham, G., Chapman, M., Parkinson, H., and others,. The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.*, 53(4):420–425, 2021.

[31] Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., and Kathiresan, S. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219–1224, August 2018.

[32] Pearl, J. *Causality*. Cambridge university press, 2009.

[33] Young, A. I., Benonisdottir, S., Przeworski, M., and Kong, A. Deconstructing the sources of genotype-phenotype associations in humans. *Science*, 365(6460):1396–1400, 2019.

[34] Simons, Y. B., Bullaughey, K., Hudson, R. R., and Sella, G. A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS Biol.*, 16(3):e2002985, 2018.

[35] O'Connor, L. J., Schoech, A. P., Hormozdiari, F., Gazal, S., Patterson, N., and Price, A. L. Extreme polygenicity of complex traits is explained by negative selection. *The American Journal of Human Genetics*, 105(3):456–476, September 2019.

[36] Mostafavi, H., Spence, J. P., Naqvi, S., and Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nature Genetics*, 55(11):1866–1875, October 2023.

[37] Spence, J. P., Mostafavi, H., Ota, M., Milind, N., Gjorgjieva, T., Smith, C. J., Simons, Y. B., Sella, G., and Pritchard, J. K. Specificity, length and luck drive gene rankings in association studies. *Nature*, 649(8098):918–925, 2026.

[38] Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., et al. Perturb-seq: Dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, December 2016.

[39] Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., et al. A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882.e21, December 2016.

[40] Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L. C., Kuchler, A., Alpar, D., and Bock, C. Pooled crispr screening with single-cell transcriptome readout. *Nature Methods*, 14(3):297–301, January 2017.

[41] Xie, S., Duan, J., Li, B., Zhou, P., and Hon, G. C. Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. *Molecular Cell*, 66(2):285–299.e5, 2017.

[42] Replogle, J. M., Saunders, R. A., Pogson, A. N., Hussmann, J. A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E. J., Adelman, K., Lithwick-Yanai, G., and others,. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575, 2022.

[43] Schnitzler, G. R., Kang, H., Fang, S., Angom, R. S., Lee-Kim, V. S., Ma, X. R., Zhou, R., Zeng, T., Guo, K., Taylor, M. S., et al. Convergence of coronary artery disease genes onto endothelial cell programs. *Nature*, 626(8000):799–807, February 2024.

[44] Zhu, R., Dann, E., Yan, J., Reyes Retana, J., Goto, R., Guitche, R. C., Brixi, L., Ota, M., Hartman, A., Roth, T. L., et al. Genome-scale perturb-seq in primary human $CD4^+$ T cells maps context-specific regulators of T cell programs and human immune traits. *Cell*, 2026.

[45] Moonen, D. P. I., Claringbould, A., Gschwind, A. R., Schrod, S., Braunger, J., Feng, C., Rauscher, B., Yi, J., Bi, S. Z., Matthess, Y., et al. Genome-scale mapping of variant, enhancer and gene function in primary human CD4+ T cells. *bioRxiv*, page 2026.03.09.710372, March 2026.

[46] Huang, A. C., Hsieh, T.-H. S., Zhu, J., Michuda, J., Teng, A., Kim, S., Rumsey, E. M., Lam, S. K., Anigbogu, I., Wright, P., et al. X-atlas/orion: Genome-wide perturb-seq datasets via a scalable fix-cryopreserve platform for training dose-dependent biological foundation models. *bioRxiv*, page 2025.06.11.659105, June 2025.

[47] Nourreddine, S., Doctor, Y., Dailamy, A., Forget, A., Lee, Y.-H., Chinn, B., Khaliq, H., Polacco, B., Muralidharan, M., Pan, E., et al. A perturbation cell atlas of human induced pluripotent stem cells. *bioRxiv*, page 2024.11.03.621734, November 2024.

[48] Savage, R. E., McRoberts Amador, C. D., Hock, C. T., Zhang, R., Kuo, H.-C., Gao, A. R., Horlbeck, M. A., Gersbach, C. A., and Buenrostro, J. D. Transcription factor collaboration enables precise T cell state engineering. *bioRxiv*, page 2026.04.20.718569, April 2026.

[49] Feldman, D., Singh, A., Schmid-Burgk, J. L., Carlson, R. J., Mezger, A., Garrity, A. J., Zhang, F., and Blainey, P. C. Optical pooled screens in human cells. *Cell*, 179(3):787–799.e17, October 2019.

[50] Funk, L., Su, K.-C., Ly, J., Feldman, D., Singh, A., Moodie, B., Blainey, P. C., and Cheeseman, I. M. The phenotypic landscape of essential human genes. *Cell*, 185(24):4634–4653.e22, November 2022.

[51] Kudo, T., Lopez, R., Meireles, A. M., Rios, A., Coelho, P., Chandrasekar, V., Lam, D. C., Huetter, J.-C., Ota, M., Rozenblatt-Rosen, O., et al. Scalable multimodal mapping of macrophage regulatory architecture by integrating optical and transcriptomic pooled screens. *bioRxiv*, page 2026.05.27.728345, 2026.

[52] Hamilton, M. C., Fife, J. D., Akinci, E., Yu, T., Khowpinitchai, B., Cha, M., Barkal, S., Thi, T. T., Yeo, G. H. T., Ramos Barroso, J. P., et al. Systematic elucidation of genetic mechanisms underlying cholesterol uptake. *Cell Genom.*, 3(5):100304, May 2023.

[53] Rottner, A. K., Ye, Y., Navarro-Guerrero, E., Rajesh, V., Pollner, A., Bevacqua, R. J., Yang, J., Spigelman, A. F., Baronio, R., Bautista, A., et al. A genome-wide CRISPR screen identifies CALCOCO2 as a regulator of beta cell function influencing type 2 diabetes risk. *Nat. Genet.*, 55(1):54–65, January 2023.

[54] Haney, M. S., Bohlen, C. J., Morgens, D. W., Ousey, J. A., Barkal, A. A., Tsui, C. K., Ego, B. K., Levin, R., Kamber, R. A., Collins, H., et al. Identification of phagocytosis regulators using magnetic genome-wide CRISPR screens. *Nat. Genet.*, 50(12):1716–1727, December 2018.

[55] Freimer, J. W., Shaked, O., Naqvi, S., Sinnott-Armstrong, N., Kathiria, A., Garrido, C. M., Chen, A. F., Cortez, J. T., Greenleaf, W. J., Pritchard, J. K., and Marson, A. Systematic discovery and perturbation of regulatory genes in human t cells reveals the architecture of immune networks. *Nature Genetics*, 54(8):1133–1144, July 2022.

[56] Khan, T. G., Bragazzi Cunha, J., Raut, C., Burroughs, M., Vyas, H. S., Leix, K., Goonewardena, S. N., Smrcka, A. V., Speliotes, E. K., and Emmer, B. T. Functional interrogation of cellular Lp(a) uptake by genome-scale CRISPR screening. *Atherosclerosis*, 403(119174):119174, April 2025.

[57] Ota, M., Spence, J. P., Zeng, T., Dann, E., Milind, N., Marson, A., and Pritchard, J. K. Causal modelling of gene effects from regulators to programs to traits. *Nature*, 650(8101):399–408, 2026.

[58] Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M. W., Fairfax, B. P., and others,. Systematic identification of *trans* eQTLs as putative drivers of known disease associations. *Nature Genetics*, 45(10):1238–1243, 2013.

[59] Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., and others,. 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, 2021.

[60] Ferkingstad, E., Sulem, P., Atlason, B. A., Sveinbjornsson, G., Magnusson, M. I., Styrmisdottir, E. L., Gunnarsdottir, K., Helgason, A., and others,. Large-scale integration of the plasma proteome with genetics and disease. *Nature Genetics*, 53(12):1712–1721, 2021.

[61] Salamone, I. M., Tian, P., Qi, Z., Zhao, J., Zhang, L., Tan, Q., Li, J., Michael, A. N., Thornburg, A. G., Sakabe, N. J., and others,. Trans-regulatory gene mapping prioritizes disease drivers in asthma. *Cell*, 189(17):5250–5265, 2026.

[62] Libby, P., Buring, J. E., Badimon, L., Hansson, G. K., Deanfield, J., Bittencourt, M. S., Tokgözoğlu, L., and Lewis, E. F. Atherosclerosis. *Nat. Rev. Dis. Primers*, 5(1):56, August 2019.

[63] Roda, G., Chien Ng, S., Kotze, P. G., Argollo, M., Panaccione, R., Spinelli, A., Kaser, A., Peyrin-Biroulet, L., and Danese, S. Crohn's disease. *Nat. Rev. Dis. Primers*, 6(1):22, April 2020.

[64] Tcheandjieu, C., Zhu, X., Hilliard, A. T., Clarke, S. L., Napolioni, V., Ma, S., Lee, K. M., Fang, H., Chen, F., Lu, Y., et al. Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. *Nat. Med.*, 28(8):1679–1692, August 2022.

[65] Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S., et al. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.*, 45(11):1274–1283, November 2013.

[66] Graham, S. E., Clarke, S. L., Wu, K.-H. H., Kanoni, S., Zajac, G. J. M., Ramdas, S., Surakka, I., Ntalla, I., Vedantam, S., Winkler, T. W., et al. The power of genetic diversity in genome-wide association studies of lipids. *Nature*, 600(7890):675–679, December 2021.

[67] Franzén, O., Ermel, R., Cohain, A., Akers, N. K., Di Narzo, A., Talukdar, H. A., Foroughi-Asl, H., Giambartolomei, C., Fullard, J. F., Sukhavasi, K., et al. Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. *Science*, 353(6301):827–830, August 2016.

[68] de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S.-G., et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.*, 49(2):256–261, February 2017.

[69] Zhu, R., Zhang, Q., Yuan, K., Zhang, R., Turvey, A. K., Stevens, C. R., Fachal, L., IIBDGC Sequencing Group,, Ahmad, T., Bel Kok, K., et al. Exome sequencing directly implicates 68 genes in inflammatory bowel disease. *medRxiv*, page 2026.05.08.26352648, May 2026.

[70] Liu, Z., Liu, R., Gao, H., Jung, S., Gao, X., Sun, R., Liu, X., Kim, Y., Lee, H.-S., Kawai, Y., et al. Genetic architecture of the inflammatory bowel diseases across East Asian and European ancestries. *Nat. Genet.*, 55(5):796–806, May 2023.

[71] Alegbe, T., Harris, B. T., Fachal, L., Ramirez-Navarro, L., Tutert, M., Krzak, M., Ghouraba, M., Strickland, M., Ozols, M., Cohen, C. E., et al. Cell-type-resolved genetic variation shapes inflammatory bowel disease risk. *Nature*, 656(8126):129–139, August 2026.

[72] Nadig, A., Replogle, J. M., Pogson, A. N., Murthy, M., McCarroll, S. A., Weissman, J. S., Robinson, E. B., and O'Connor, L. J. Transcriptome-wide analysis of differential expression in perturbation atlases. *Nature Genetics*, 57(5):1228–1237, April 2025.

[73] Saunders, R. A., Allen, W. E., Pan, X., Sandhu, J., Lu, J., Lau, T. K., Smolyar, K., Sullivan, Z. A., Dulac, C., Weissman, J. S., and Zhuang, X. Perturb-multimodal: A platform for pooled genetic screens with imaging and sequencing in intact mammalian tissue. *Cell*, 188(17):4790–4809.e22, August 2025.

[74] Papalexi, E., Mimitou, E. P., Butler, A. W., Foster, S., Bracken, B., Mauck, III, W. M., Wessels, H.-H., Hao, Y., Yeung, B. Z., Smibert, P., and Satija, R. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nat. Genet.*, 53(3):322–331, March 2021.

[75] Ma, W. F., Hodonsky, C. J., Turner, A. W., Wong, D., Song, Y., Mosquera, J. V., Ligay, A. V., Slenders, L., Gancayco, C., Pan, H., et al. Enhanced single-cell RNA-seq workflow reveals coronary artery disease cellular cross-talk and candidate drug targets. *Atherosclerosis*, 340:12–22, January 2022.

[76] Alkhawaja, A. A., Currin, K. W., Perrin, H. J., Vadlamudi, S., Etheridge, A. S., Broadaway, K. A., Cannon, G. H., Anderson, C. W., Moxley, A. H., Iuga, A. C., et al. Liver single-nucleus multiome profiling reveals cell-type mechanisms for cardiometabolic traits. *Am. J. Hum. Genet.*, 113(1):133–148, January 2026.

[77] Julkunen, H., Cichońska, A., Tiainen, M., Koskela, H., Nybo, K., Mäkelä, V., Nokso-Koivisto, J., Kristiansson, K., Perola, M., Salomaa, V., et al. Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank. *Nat. Commun.*, 14(1):604, February 2023.

[78] Wirka, R. C., Wagh, D., Paik, D. T., Pjanic, M., Nguyen, T., Miller, C. L., Kundu, R., Nagao, M., Coller, J., Koyano, T. K., et al. Atheroprotective roles of smooth muscle cell phenotypic modulation and the TCF21 disease gene as revealed by single-cell analysis. *Nat. Med.*, 25(8):1280–1289, August 2019.

[79] Corces, M. R., Shcherbina, A., Kundu, S., Gloudemans, M. J., Frésard, L., Granja, J. M., Louie, B. H., Eulalio, T., Shams, S., Bagdatli, S. T., et al. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat. Genet.*, 52(11):1158–1168, November 2020.

[80] Smillie, C. S., Biton, M., Ordovas-Montanes, J., Sullivan, K. M., Burgin, G., Graham, D. B., Herbst, R. H., Rogel, N., Slyper, M., Waldman, J., et al. Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell*, 178(3):714–730.e22, July 2019.

[81] Ma, X. R., Conley, S. D., Kosicki, M., Bredikhin, D., Cui, R., Tran, S., Sheth, M. U., Qiu, W.-L., Chen, S., Kundu, S., et al. Molecular convergence of risk variants for congenital heart defects leveraging a regulatory map of the human fetal heart. *medRxiv*, page 2024.11.20.24317557, November 2024.

[82] Gjoneska, E., Pfenning, A. R., Mathys, H., Quon, G., Kundaje, A., Tsai, L.-H., and Kellis, M. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature*, 518(7539):365–369, February 2015.

[83] Woodruff, P. G., Modrek, B., Choy, D. F., Jia, G., Abbas, A. R., Ellwanger, A., Koth, L. L., Arron, J. R., and Fahy, J. V. T-helper type 2-driven inflammation defines major subphenotypes of asthma. *Am. J. Respir. Crit. Care Med.*, 180(5):388–395, September 2009.

[84] Choy, D. F., Hart, K. M., Borthwick, L. A., Shikotra, A., Nagarkar, D. R., Siddiqui, S., Jia, G., Ohri, C. M., Doran, E., Vannella, K. M., et al. TH2 and TH17 inflammatory pathways are reciprocally regulated in asthma. *Sci. Transl. Med.*, 7(301):301ra129, August 2015.

[85] Lee, D., Koutrouli, M., Masse, N. Y., Hoffman, G. E., Kinrot, S., Wang, X., N M, P., Pjanic, M., Clarence, T., Tsetsos, F., et al. Single-cell atlas of transcriptomic vulnerability across brain disorders. *Nature*, 657(8133):988–1002, September 2026.

[86] Su, Y., Chen, D., Yuan, D., Lausted, C., Choi, J., Dai, C. L., Voillet, V., Duvvuri, V. R., Scherler, K., Troisch, P., et al. Multi-omics resolves a sharp disease-state shift between mild and moderate COVID-19. *Cell*, 183(6):1479–1495.e20, December 2020.

[87] Geiger-Schuller, K., Eraslan, B., Kuksenko, O., Dey, K. K., Jagadeesh, K. A., Thakore, P. I., Karayel, O., Yung, A. R., Rajagopalan, A., Meireles, A. M., et al. Systematically characterizing the roles of e3-ligase family members in inflammatory responses with massively parallel perturb-seq. *bioRxiv*, January 2023.

[88] Ruzicka, W. B., Mohammadi, S., Fullard, J. F., Davila-Velderrain, J., Subburaju, S., Tso, D. R., Hourihan, M., Jiang, S., Lee, H.-C., Bendl, J., et al. Single-cell multi-cohort dissection of the schizophrenia transcriptome. *Science*, 384(6698):eadg5136, May 2024.

[89] Wamsley, B., Bicks, L., Cheng, Y., Kawaguchi, R., Quintero, D., Margolis, M., Grundman, J., Liu, J., Xiao, S., Hawken, N., et al. Molecular cascades and cell type-specific signatures in ASD revealed by single-cell genomics. *Science*, 384(6698):eadh2602, May 2024.

[90] Zhang, M. J., Hou, K., Dey, K. K., Sakaue, S., Jagadeesh, K. A., Weinand, K., Taychameekiatchai, A., Rao, P., Pisco, A. O., Zou, J., et al. Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. *Nat. Genet.*, 54(10):1572–1580, October 2022.

[91] Grabski, I. N., Lee, J., Blair, J., Dalgarno, C., Mascio, I., Bradu, A., Knowles, D. A., and Satija, R. Mapping transcriptional responses to cellular perturbation dictionaries with RNA fingerprinting. *bioRxiv*, page 2025.09.19.676866, September 2025.

[92] Davey Smith, G. and Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.*, 32(1):1–22, 2003.

[93] Suzuki, K., Hatzikotoulas, K., Southam, L., Taylor, H. J., Yin, X., Lorenz, K. M., Mandla, R., Huerta-Chagoya, A., Melloni, G. E. M., Kanoni, S., et al. Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. *Nature*, 627(8003):347–357, 2024.

[94] Morrison, J., Willwerscheid, J., Sylvertooth, D., He, X., and Stephens, M. Genetic factor analysis for characterizing phenome-wide patterns of genetic pleiotropy. *Nature Genetics*, 2026.

[95] Yao, D., Binan, L., Bezney, J., Simonton, B., Freedman, J., Frangieh, C. J., Dey, K., Geiger-Schuller, K., Eraslan, B., Gusev, A., et al. Scalable genetic screening for regulatory circuits using compressed perturb-seq. *Nat. Biotechnol.*, 42(8):1282–1295, August 2024.

[96] Brown, B. C., Tokolyi, A., Morris, J. A., Lappalainen, T., and Knowles, D. A. Large-scale causal discovery using interventional data sheds light on gene network structure in k562 cells. *Nat. Commun.*, 16(1):9628, October 2025.

[97] Cao, Z.-J., Guo, T., Li, C., Zhang, F., Lu, P., Gu, C., Guo, F., and Gao, G. Causality-oriented regulatory inference and rational transcriptomic reprogramming design with CASCADE. *bioRxiv*, 2025. Preprint.

[98] Ge, C. and Li, H. Robust causal gene network estimation for large-scale single-cell perturbation screens using reduced control function. *bioRxiv*, 2026. Preprint, version 2.

[99] Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.*, 82(5):1273–1300, 2020.

[100] Dorans, E., Jagadeesh, K., Dey, K., and Price, A. L. Linking regulatory variants to target genes by integrating single-cell multiome methods and genomic distance. *Nat. Genet.*, 57(7):1649–1658, 2025.

[101] Sheth, M. U., Qiu, W.-L., Ma, X. R., Gschwind, A. R., Jagoda, E., Tan, A. S., Galante, J., Ray, J., Amgalan, D., Einarsson, H., et al. Mapping enhancer-gene regulatory interactions from single-cell data. *Nat. Genet.*, 58(8):1941–1952, August 2026.

[102] Gschwind, A. R., Mualim, K. S., Karbalayghareh, A., Sheth, M. U., Dey, K. K., Jagoda, E., Nurtdinov, R. N., Xi, W., Tan, A. S., Galante, J., et al. An encyclopedia of human enhancer-gene regulatory interactions. *Nature*, 657(8130):179–190, September 2026.

[103] Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11):1228–1235, September 2015.

[104] Ulirsch, J. C., Lareau, C. A., Bao, E. L., Ludwig, L. S., Guo, M. H., Benner, C., Satpathy, A. T., Kartha, V. K., Salem, R. M., Hirschhorn, J. N., et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nature Genetics*, 51(4):683–693, March 2019.

[105] Strober, B. J., Zhang, M. J., Amariuta, T., Rossen, J., and Price, A. L. Fine-mapping causal tissues and genes at disease-associated loci. *Nature genetics*, 57(1):42–52, 2025.

[106] Milind, N., Smith, C. J., Zhu, H., Gjorgjieva, T., Spence, J. P., and Pritchard, J. K. Buffering of gene dosage response curves for human complex traits. *Cell Genom.*, 6(7):101221, 2026.

[107] Backman, J. D., Li, A. H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M. D., Benner, C., Liu, D., Locke, A. E., Balasubramanian, S., et al. Exome sequencing and analysis of 454,787 uk biobank participants. *Nature*, 599(7886):628–634, October 2021.

[108] Hindy, G., Adam, R. C., Sosina, O., Pryce, D., Blair, D., Herman, J., Lee, J., Dornbos, P., Mayerhofer, E., Gilly, A., and others,. FNIP1 variants are associated with favourable metabolism in 1 million humans. *Nature*, pages 1–10, 2026.

[109] Palmer, D. S., Hill, B., Hodgson, S., Jõeloo, M., Kalantzis, G., Kousathanas, A., Koyama, S., Lu, W., Namba, S., Rodriguez, Z. B., and others,. The Biobank Rare Variant consortium powers the discovery of rare genetic associations through global collaboration. *medRxiv*, pages 2026–05, 2026.

[110] Zeng, T., Spence, J. P., Mostafavi, H., and Pritchard, J. K. Bayesian estimation of gene constraint from an evolutionary model with gene features. *Nat. Genet.*, 56(8):1632–1643, 2024.

[111] Clarke, B., Holtkamp, E., Öztürk, H., Mück, M., Wahlberg, M., Meyer, K., Munzlinger, F., Brechtmann, F., Hölzlwimmer, F. R., Lindner, J., and others,. Integration of variant annotations using deep set networks boosts rare variant association testing. *Nat. Genet.*, 56(10):2271–2280, 2024.

[112] Weeks, E. M., Ulirsch, J. C., Cheng, N. Y., Trippe, B. L., Fine, R. S., Miao, J., Patwardhan, T. A., Kanai, M., Nasser, J., Fulco, C. P., and others,. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *Nat. Genet.*, 55(8):1267–1276, 2023.

[113] Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., and others,. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

[114] Domingo, J., Minaeva, M., Morris, J. A., Ghatan, S., Ziosi, M., Sanjana, N. E., and Lappalainen, T. Nonlinear transcriptional responses to gradual modulation of transcription factor dosage. *eLife*, 13:RP100555, 2026.

[115] Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., GTEx Consortium,, Nicolae, D. L., and others,. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.*, 47(9):1091–1098, 2015.

[116] Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., De Geus, E. J., Boomsma, D. I., Wright, F. A., and others,. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.*, 48(3):245–252, 2016.

[117] Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., and others,. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.*, 48(5):481–487, 2016.

[118] Gerlach, P. A., Milind, N., Spence, J. P., and Pritchard, J. K. High false sign rates in transcriptome-wide association studies. *bioRxiv*, page 2025.12.19.695550, 2025.

[119] Chen, M., Wang, X., Krockenberger, L., Tyebally, R., Berg, J. J., Pott, S., Flint, J., Powell, J. E., Balliu, B., Liu, X., and Dahl, A. Cell-type-specific eQTLs underlie the genetic architecture of complex traits. *Nature*, August 2026.

[120] Siraj, L., Castro, R. I., Dewey, H. B., Kales, S., Butts, J. C., Nguyen, T. T. L., Kanai, M., Berenzy, D., Mouri, K., Wang, Q. S., et al. Functional dissection of complex trait variants at single-nucleotide resolution. *Nature*, February 2026.

[121] Ho, C.-H., Dippel, M. A., McQuade, M. S., Nguyen, L. P., Mishra, A., Pribitzer, S., Hardy, S., Chandok, H., Chardon, F. M., McDiarmid, T. A., et al. Genetic and epigenetic screens in primary human T cells link candidate causal autoimmune variants to T cell networks. *Nat. Genet.*, 57(10):2536–2545, October 2025.

[122] Ray, J. P., de Boer, C. G., Fulco, C. P., Lareau, C. A., Kanai, M., Ulirsch, J. C., Tewhey, R., Ludwig, L. S., Reilly, S. K., Bergman, D. T., et al. Prioritizing disease and trait causal variants at the TNFAIP3 locus using functional and genomic features. *Nat. Commun.*, 11(1):1237, March 2020.

[123] Morris, J. A., Caragine, C., Daniloski, Z., Domingo, J., Barry, T., Lu, L., Davis, K., Ziosi, M., Glinos, D. A., Hao, S., et al. Discovery of target genes and pathways at gwas loci by pooled single-cell crispr screens. *Science*, 380(6646):eadh7699, May 2023.

[124] Ramste, M., Weldy, C., Kundu, S., Zhao, Q., Li, D., Brand, K., Sharma, D., Ramste, A., Jagoda, E., Ray, J., et al. Enhancer-targeting CRISPR screens at coronary artery disease loci suggest shared mechanisms of disease risk. *medRxiv*, page 2025.08.28.25334684, September 2025.

[125] Nasser, J., Bergman, D. T., Fulco, C. P., Guckelberger, P., Doughty, B. R., Patwardhan, T. A., Jones, T. R., Nguyen, T. H., Ulirsch, J. C., Lekschas, F., and others,. Genome-wide enhancer maps link risk variants to disease genes. *Nature*, 593(7858):238–243, 2021.

[126] Cooper, S. E., Coelho, M. A., Strauss, M. E., Gontarczyk, A. M., Wu, Q., Garnett, M. J., Marioni, J. C., and Bassett, A. R. scSNV-seq: high-throughput phenotyping of single nucleotide

variants by coupled single-cell genotyping and transcriptomics. *Genome Biol.*, 25(1):20, January 2024.

[127] Baglaenko, Y., Mu, Z., Curtis, M., Mire, H. M., Jayanthi, V., Al Suqri, M., Liu, C., Agnew, R., Nathan, A., Mah-Som, A. Y., et al. Precisely defining disease variant effects in CRISPR-edited single cells. *Nature*, 646(8083):117–125, October 2025.

[128] Ryu, J., Barkal, S., Yu, T., Jankowiak, M., Zhou, Y., Francoeur, M., Phan, Q. V., Li, Z., Tognon, M., Brown, L., et al. Joint genotypic and phenotypic outcome modeling improves base editing variant effect quantification. *medRxiv*, page 2023.09.08.23295253, September 2023.

[129] Linder, J., Srivastava, D., Yuan, H., Agarwal, V., and Kelley, D. R. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nat. Genet.*, 57(4):949–961, 2025.

[130] Avsec, Ž., Latysheva, N., Cheng, J., Novati, G., Taylor, K. R., Ward, T., Bycroft, C., Nicolaisen, L., Arvaniti, E., Pan, J., and others,. Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, 649(8099):1206–1218, 2026.

[131] Lal, A., Karollus, A., Gunsalus, L., Garfield, D., Nair, S., Tseng, A. M., Gordon, M. G., Blischak, J., Van De Geijn, B., Bhangale, T., et al. Decoding sequence determinants of gene expression in diverse cellular and disease states. *Nat. Methods*, 23(6):1138–1151, June 2026.

[132] Sasse, A., Ng, B., Spiro, A. E., Tasaki, S., Bennett, D. A., Gaiteri, C., De Jager, P. L., Chikina, M., and Mostafavi, S. Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *bioRxiv*, page 2023.03.16.532969, September 2023.

[133] Huang, C., Shuai, R. W., Baokar, P., Chung, R., Rastogi, R., Kathail, P., and Ioannidis, N. M. Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nat. Genet.*, 55(12):2056–2059, 2023.

[134] Li, S., Luo, R., and Huang, Y. Assessing large-scale genomic language models in predicting personal gene expression: promises and limitations. *bioRxiv*, page 2025.07.09.664024, July 2025.

[135] Drusinsky, S., Whalen, S., and Pollard, K. S. Deep-learning prediction of gene expression from personal genomes. *Genome Biol.*, 27(1):19, 2026.

[136] Rastogi, R., Reddy, A. J., Chung, R., and Ioannidis, N. M. Fine-tuning sequence-to-expression models on personal genome and transcriptome data. *Genome Biol.*, 27(1):234, May 2026.

[137] Aguet, F., Anand, S., Ardlie, K. G., Gabriel, S., Getz, G. A., Graubert, A., Hadley, K., Handsaker, R. E., Huang, K. H., Kashin, S., et al. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, September 2020.

[138] Yazar, S., Alquicira-Hernandez, J., Wing, K., Senabouth, A., Gordon, M. G., Andersen, S., Lu, Q., Rowson, A., Taylor, T. R. P., Clarke, L., et al. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science*, 376(6589):eabf3041, April 2022.

[139] Soskic, B., Cano-Gamez, E., Smyth, D. J., Ambridge, K., Ke, Z., Matte, J. C., Bossini-Castillo, L., Kaplanis, J., Ramirez-Navarro, L., Lorenc, A., et al. Immune disease risk variants regulate gene expression dynamics during CD4+ T cell activation. *Nat. Genet.*, 54(6):817–826, June 2022.

[140] Nuñez, J. K., Chen, J., Pommier, G. C., Cogan, J. Z., Replogle, J. M., Adriaens, C., Ramadoss, G. N., Shi, Q., Hung, K. L., Samelson, A. J., et al. Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing. *Cell*, 184(9):2503–2519.e17, April 2021.

[141] Goudy, L., Ha, A., Borah, A. A., Umhoefer, J. M., Chow, L., Tran, C., Winters, A., Talbot, A., Hernandez, R., Li, Z., et al. Integrated epigenetic and genetic programming of primary human T cells. *Nat. Biotechnol.*, 44:1470–1481, 2026.

[142] Hsiung, C. C.-S., Wilson, C. M., Sambold, N. A., Dai, R., Chen, Q., Teyssier, N., Misiukiewicz, S., Arab, A., O'Loughlin, T., Cofsky, J. C., et al. Engineered CRISPR-Cas12a for higher-order combinatorial chromatin perturbations. *Nat. Biotechnol.*, 43(3):369–383, March 2025.

[143] Wessels, H.-H., Méndez-Mancilla, A., Hao, Y., Papalexi, E., Mauck, III, W. M., Lu, L., Morris, J. A., Mimitou, E. P., Smibert, P., Sanjana, N. E., and Satija, R. Efficient combinatorial targeting of RNA transcripts in single cells with Cas13 RNA perturb-seq. *Nat. Methods*, 20(1):86–94, January 2023.

[144] Sella, G. and Barton, N. H. Thinking about the evolution of complex traits in the era of genome-wide association studies. *Annu. Rev. Genomics Hum. Genet.*, 20(Volume 20, 2019):461–493, 2019.

[145] Robinson, M. D. and Smyth, G. K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

[146] Robinson, M. D. and Smyth, G. K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–332, 2008.

[147] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[148] Love, M. I., Huber, W., and Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12):550, 2014.

[149] Southard, K. M., Ardy, R. C., Tang, A., O'Sullivan, D. D., Metzner, E., Guruvayurappan, K., and Norman, T. M. Comprehensive transcription factor perturbations recapitulate fibroblast transcriptional states. *Nat. Genet.*, 57(9):2323–2334, September 2025.

[150] Laber, S., Strobel, S., Mercader, J. M., Dashti, H., Dos Santos, F. R. C., Kubitz, P., Jackson, M., Ainbinder, A., Honecker, J., Agrawal, S., et al. Discovering cellular programs of intrinsic and extrinsic drivers of metabolic traits using LipocyteProfiler. *Cell Genom.*, 3(7):100346, July 2023.

[151] Fleck, J. S., Jansen, S. M. J., Wollny, D., Zenk, F., Seimiya, M., Jain, A., Okamoto, R., Santel, M., He, Z., Camp, J. G., and Treutlein, B. Inferring and perturbing cell fate regulomes in human brain organoids. *Nature*, 621(7978):365–372, September 2023.

[152] Jin, X., Simmons, S. K., Guo, A., Shetty, A. S., Ko, M., Nguyen, L., Jokhi, V., Robinson, E., Oyler, P., Curry, N., et al. In vivo perturb-seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science*, 370(6520):eaaz6063, November 2020.

[153] Fan, X., Zhou, R., Raftrey, B. C., Coronado, P. E. R., Trimm, E., Clancy, E., Chen, X., Bozeman, J., Chen, M. S., Alimukhamedov, S., et al. A perturb-seq screen guided by species divergence uncovers pathways for collateral artery formation. *bioRxiv*, page 2026.04.29.721711, May 2026.

[154] Jiang, L., Dalgarno, C., Papalexi, E., Mascio, I., Wessels, H.-H., Yun, H., Iremadze, N., Lithwick-Yanai, G., Lipson, D., and Satija, R. Systematic reconstruction of molecular pathway signatures using scalable single-cell perturbation screens. *Nat. Cell Biol.*, 27(3):505–517, March 2025.

[155] Rood, J. E., Hupalowska, A., and Regev, A. Toward a foundation model of causal cell and tissue biology with a perturbation cell and tissue atlas. *Cell*, 187(17):4520–4545, August 2024.

[156] Gavriilidis, G. I., Vasileiou, V., Orfanou, A., Ishaque, N., and Psomopoulos, F. A mini-review on perturbation modelling across single-cell omic modalities. *Comput. Struct. Biotechnol. J.*, 23:1886–1896, December 2024.

[157] Bunne, C., Roohani, Y., Rosen, Y., Gupta, A., Zhang, X., Roed, M., Alexandrov, T., AlQuraishi, M., Brennan, P., Burkhardt, D. B., et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, December 2024.

[158] Pampari, A., Shcherbina, A., Kvon, E., Kosicki, M., Nair, S., Kundu, S., Kathiria, A. S., Risca, V. I., Kuningas, K., Alasoo, K., et al. ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. *bioRxiv*, page 2024.12.25.630221, December 2024.

[159] Cui, H., Tejada-Lapuerta, A., Brbić, M., Saez-Rodriguez, J., Cristea, S., Goodarzi, H., Lotfollahi, M., Theis, F. J., and Wang, B. Towards multimodal foundation models in molecular cell biology. *Nature*, 640(8059):623–633, April 2025.

[160] Lee, J. C., Biasci, D., Roberts, R., Gearry, R. B., Mansfield, J. C., Ahmad, T., Prescott, N. J., Satsangi, J., Wilson, D. C., Jostins, L., et al. Genome-wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease. *Nat. Genet.*, 49(2):262–268, February 2017.

[161] Iwaki, H., Blauwendraat, C., Leonard, H. L., Kim, J. J., Liu, G., Maple-Grødem, J., Corvol, J.-C., Pihlstrøm, L., van Nimwegen, M., Hutten, S. J., et al. Genomewide association study of Parkinson's disease clinical biomarkers in 12 longitudinal patients' cohorts. *Mov. Disord.*, 34(12):1839–1850, December 2019.