https://doi.org/10.26599/CAI.2024.9390003 Cybernetics and Intelligence

# Modeling the Causal Mechanism between Genotypes and Phenotypes using Large-Scale Biobank Data and Context-Specific Regulatory Networks

Wenran Li[1], Wanwen Zeng[2], and Wing Hung Wong[2,3] ✉

## ABSTRACT

The relationship between genetic variation and human phenotypes is crucial for developing effective treatments and personalized medicine. However, our understanding of the regulatory mechanisms by which variants influence human traits and diseases is far from complete. Context-specific regulatory network is a typical tool that provides detailed understanding of gene regulation in specific biological contexts, allowing us to identify key regulators and pathways that are important for a particular phenotype. In this review, we summarize the large international biobanks and reference omics data that provide diverse datasets for the genotype-phenotype analysis and the construction of context-specific regulatory networks, and discuss the importance of context-specific regulatory networks in explaining the underlying causal mechanism between genotypes and phenotypes. We emphasize the significance of QTL studies in explaining the correlation between genotypes and omics features, and present various computational approaches for the construction of context-specific regulatory networks. With continued advancements in biobanking, genomics, and computational biology, the context-specific regulatory networks may serve as an increasingly powerful tool for modeling the causal mechanisms that underlie the relationship between genotypes and phenotypes.

## KEYWORDS

genotype; phenotype; biobanks; multi-omics; reference data; context-specific regulatory networks

There are millions of genetic variants in the human genome, including single nucleotide polymorphisms (SNPs), insertions, deletions, and copy number variations[1, 2, 3]. While many of these variants are benign and do not have any significant impact on phenotypes, some variants can have functional effects on genes, proteins, and other molecular factors, which can in turn influence phenotypes and disease risk[4, 5, 6, 7]. Exploring the causal mechanism of the phenotypic variants is crucial for the interpretation of the molecular basis of complex diseases and traits[8], as well as for the development of effective therapies and personalized medicine[9, 10, 11]. However, our understanding of the regulatory mechanisms by which variants influence human traits and diseases is far from complete.

With the development of sequencing technologies and the accumulation of various omics data[12, 13, 14], regulatory networks built based on multiple omics data have become an increasingly important tool for understanding the molecular mechanisms between genotypes and phenotypes[15, 16, 17]. General regulatory networks describe interactions in a generic way across all contexts, however, the mechanisms underlying the biological processes are highly dynamic and can vary across different biological contexts, such as different tissues, developmental stages, or disease states[18]. Thus, there is a need for context-specific regulatory networks that describe the interactions between molecules involved in a particular tissue, cell line, or cellular state. Context-specific regulatory networks provide a more accurate and detailed understanding of gene regulation in specific biological contexts, and can help to identify key regulators and pathways that are specifically important in that context[19].

Context-specific regulatory networks interpret the underlying mechanism between genotypes and phenotypes by identifying the specific regulatory interactions that link genetic variation to phenotypic variation[20]. For example, a context-specific regulatory network can identify key regulatory nodes or hubs that are responsible for regulating the expression of multiple downstream genes and proteins[21, 22]. Then, by analyzing the connectivity patterns within the regulatory network, we can identify genetic variants that affect the activity of these regulatory nodes and ultimately lead to changes in gene expression and the phenotype of interest[23]. Besides, context-specific regulatory networks can be used to identify signaling pathways and other functional modules that are responsible for a particular biological response or phenotype[24, 25].

One of the main challenges and limitations of using context-specific regulatory networks to interpret the underlying mechanism between genotype and phenotype is the requirement for large and diverse omics datasets[12]. The accuracy and reliability of context-specific regulatory networks depend on the availability of high-quality data, including genome-wide association studies (GWAS) data[26], gene expression data[27], chromatin accessibility

1 CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China
2 Department of Statistics, Stanford University, Stanford, CA 94305, USA
3 Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 94305, USA
Wenran Li and Wanwen Zeng contribute equally to this work.
Address correspondence to Wing Hung Wong, whwong@stanford.edu

data[28, 29], protein profiling[30], and other relevant omics data[31]. The most widely used context-specific regulatory network is the gene co-expression network, where the edges are interactions between co-expressed genes predicted using gene expression data[32, 33]. In addition, other omics data, such as chromatin accessibility, DNA methylation, histone modification data, and 3D chromatin interactions can also be used to identify regulatory relationships[34, 35, 36, 37]. For example, ChIP-seq can be used to identify the genomic binding sites of TFs and other chromatin-associated proteins. By integrating ChIP-seq data with gene expression data, it is possible to infer the regulatory relationships between TFs and their target genes[38].

There have been lots of genomics and omics data released for research use, including a series of biobanks providing genotypic and phenotypic data[39], thousands of summary statistics of GWASs[40], various omics data13, and other abundant biological resources[41, 42, 43]. Fig. 1 shows the biological process from genotype to phenotype involving different layers of omics data, which are typically generated in a context-specific manner. In Fig. 2, we emphasize the difference between i) personal data such as genome sequence and clinical information, which are expected to be generated in routine healthcare or in population-level biobanks, and ii) reference data such gene expression and chromatin accessibility profiles. These reference data are typically context-dependent and not expected to be available on the same scale as the personal data, but they can be used to construct reference models to support the interpretation of the personal data. By integrating different types of omics data and simulating the effects of genetic variants, we can gain a more comprehensive understanding of the underlying regulatory mechanisms between genotypes and phenotypes and can have the opportunity to develop new strategies for drug development and genetic engineering[44].

This review discusses the importance of context-specific regulatory networks in explaining the underlying mechanism from genotypes to phenotypes. First, we review the state-of-the-art biobanks that provides abundant genotype and phenotype data for the development of GWAS analysis to capture the links between genotypes and phenotypes. Then, we highlight the value of reference omics data in the construction of context-specific regulatory networks and emphasize the significance of QTL (Quantitative Trait Locus) studies in identifying the correlation between genotypes and omics features. Moreover, we present various computational approaches for the construction of context-specific regulatory networks and show that the networks can be applied in modeling the causal mechanism between genotypes and phenotypes, providing insight into disease, and suggesting potential targets for therapeutic interventions. We also outline the challenges and opportunities that lie ahead, including the demand for more comprehensive and diverse datasets, the need of more reliable approaches for data integration and network construction, and the difficulty of network interpretation. Overall, we believe that the continued advancements in biobanking, genomics, and computational biology will lead to a better application of context-specific regulatory networks in modeling the causal mechanism between genotypes and phenotypes and pave the way for personalized medicine.

## 1 Population-level biobanks provide genotype and phenotype data

Population-level biobanks are large research resources of biological samples and associated data that typically include genotype data from thousands or even millions of individuals, as well as their various phenotypes[13, 45]. The genotype data provided by biobanks typically consist of information about an individual's genetic makeup, including variants in specific genetic markers or throughout the whole genome. The phenotype data include information about an individual's observable traits, such as height, weight, blood pressure, and medical histories, as well as information about their living environment, such as exposure to pollutants or lifestyle factors.

There are two common techniques to generate genotype data, one is microarray genotyping[46] and the other is whole-genome sequencing (WGS)[47]. Microarray uses a chip with thousands or even millions of DNA probes that bind to specific regions of the genome and determines the genotype at each location by measuring the intensity of the signal from the probes[48]. Microarray genotyping is a high-throughput and cost-effective method widely used to generate genotypes for large numbers of genetic markers simultaneously. However, the coverage and resolution of microarray is limited by the number and density of probes on the chip, which may miss important genetic variants and make it difficult to detect large structural variations, such as

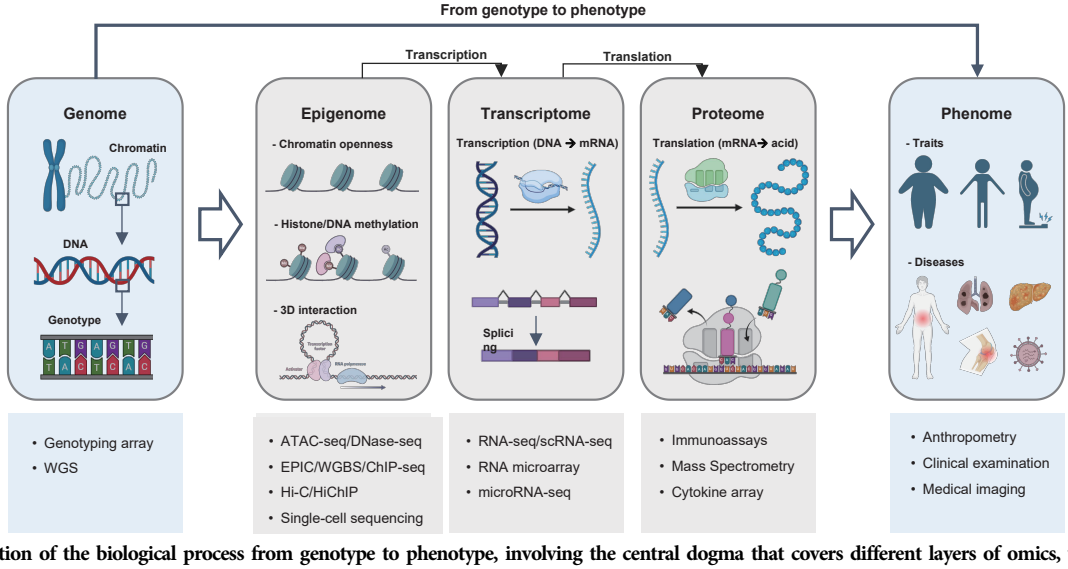


**Fig. 1 Illustration of the biological process from genotype to phenotype, involving the central dogma that covers different layers of omics, which typically happens within a cell.**

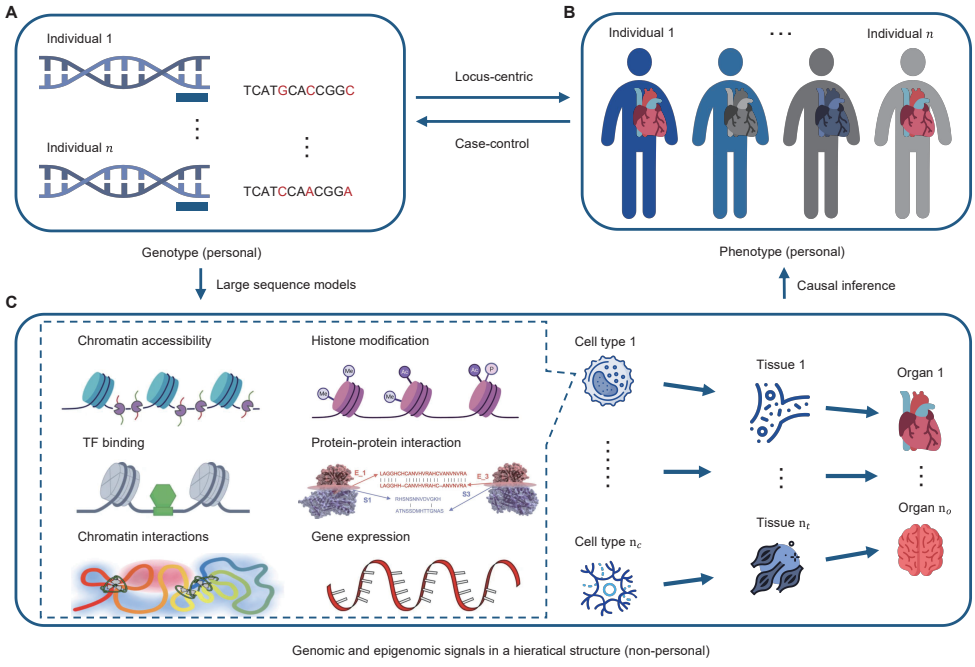


**Fig. 2 A simplified diagram that shows the understanding of how personal genotypes (A) affect personal phenotypes (B) will require the modeling of the relationship between different layers of omics based on non-personal reference data, which are usually generated in a context-specific manner (C).**

copy number variants (CNVs) and translocations[49].

WGS is a more comprehensive method than microarray for generating genotype information, as it sequences the entire genome of an individual, rather than targeting specific regions with probes[64]. WGS technique has become increasingly important in the field of genomics because it 1) enables the identification of both common and rare variants, 2) can provide important insights into the biological mechanisms that underlying the development of diseases, and 3) can be used to identify individuals who may be at increased risk of developing certain diseases and to develop personalized treatment plans for individuals, as well as help explore the most effective drugs or therapies for specific diseases[65]. By analyzing genetic variants and their impact on gene expression and protein function, we can gain a better understanding of how diseases develop and progress. Overall, whole-genome sequencing data provides a wealth of information that can help advance our understanding of human genetics and biology, and has the potential to revolutionize personalized medicine and disease prevention[66].

While genomics focuses on the study of genetic variation at the DNA level, phenomics examines the development, physiology, and behavior of an organism. Phenotype data can be broadly classified into several categories based on their characteristics and properties. Some of the main classes of phenotypes include morphological phenotypes, biochemical phenotypes, physiological phenotypes, behavioral phenotypes, and clinical phenotypes[67]. Morphological phenotypes are physical traits that can be observed and measured, such as height, weight, and body mass index (BMI). Biochemical phenotypes are phenotypes that are related to the chemical composition and metabolic processes of the body, such as blood glucose levels and cholesterol levels. Physiological phenotypes usually reflect the function of different physiological systems in the body, such as blood pressure and heart rate. Behavioral phenotypes are related to an individual's behavior, such as sleep patterns, diet, and exercise habits. Clinical phenotypes refer to disease or other medical conditions, such as diagnosis codes, medication use, and hospitalization records.

There are several techniques that can be used to generate phenotype data. One of the most typical techniques is the clinical examination, which involves physical and diagnostic examinations conducted to assess an individual's health status. The digital records of an individual's medical history, including diagnoses, medication use, and other clinical data, can be collected into electronic health records (EHRs) to serve as standardized phenotypes[68]. Another common way to generate phenotype data is through questionnaires and surveys, which are self-reported measures that capture information on an individual's behavior, lifestyle, and other exposures. In recent years, wearable devices have become a new technique to collect detailed phenotypes of individuals by monitoring the physiological or behavioral parameters, such as heart rate, activity levels, and sleep patterns.

Owing to the widespread adoption of EHRs and advancements in experimental and computational platforms for cost-effective population-scale sequencing and analysis, large-scale biobanks have emerged as a crucial resource for accelerating biomedical researches. Biobank data typically include genomic and phenotypic data from thousands to millions of individuals, and often includes data from individuals with diverse genetic and environmental backgrounds, allowing for analysis of genetic and environmental factors that contribute to the development of disease and other phenotypes. Table 1 provides a brief overview of the state-of-the-art biobanks developed by different countries and organizations. One of the most advanced biobanks is the UK Biobank (UKBB)[69], which is a prospective cohort study that recruited half a million individuals aged 40-69 years old across the United Kingdom between 2006 and 2010. UKBB is a large-scale biomedical resource that integrates genome-wide genetic data with extensive phenotype data, including data from lifestyle questionnaires, physical measures, biomarkers in blood and urine, accelerometry, multimodal imaging and other sources (Table 2).

The UKBB cohort is unprecedented in size, and the extensive phenotyping and genome-wide genotype data, supplemented with high-density imputation, have enhanced power for genetic discovery and enable well-powered GWASs of hundreds of quantitative traits, including anthropometric traits, blood traits, cognitive traits, and numerous blood and urine biomarkers. The

**Table 1 Summary of state-of-the-art biobanks proposed by organizations from different countries.**

| Biobank Name | Country | Year | Number of Individuals | Research Focus | Homepage |
|---|---|---|---|---|---|
| UK Biobank[12] | UK | 2006 | >500, 000 | Wide range of health conditions | https://www.ukbiobank.ac.uk/ |
| FinnGen[45] | Finland | 2017 | ~500, 000 | Genetic factors of diseases | https://www.finngen.fi/en |
| China Kadoorie Biobank[50] | China | 2004 | >512, 000 | Chronic diseases | https://www.ckbiobank.org |
| BioBank Japan[51] | Japan | 2003 | >200, 000 | Precision medicine | https://biobankjp.org/ |
| Biobank Graz[52] | Austria | 2008 | >1, 200, 000 | Metabolic diseases | https://biobank.medunigraz.at/ |
| LifeGene[53] | Sweden | 2007 | >50, 000 | Environmental and genetic factors on health | https://www.lifegene.se/en/ |
| Estonian Biobank[54] | Estonia | 2000 | >200, 000 | Genetic factors of diseases | https://genomics.ut.ee/en/content/estonian-biobank |
| Qatar Biobank[55] | Qatar | 2016 | >30, 000 | Health conditions prevalent in Qatar | https://www.qatarbiobank.org.qa/ |
| The Cancer Genome Atlas[56] | USA | 2006 | >11, 000 | Cancer genomics | https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga |
| The National Cancer Institute's Genomic Data Commons[57] | USA | 2016 | >86, 500 | Cancer genomics | https://gdc.cancer.gov/ |
| CanPath[58] | Canada | 2008 | >330, 000 | Environmental and lifestyle factors on health | https://www.partnershipfortomorrow.ca/ |
| Saudi Human Genome Program[59] | Saudi Arabia | 2013 | >100, 000 | Genetic factors of diseases | https://shgp.kacst.edu.sa/en/Pages/default.aspx |
| deCODE Genetics[60] | Iceland | 1996 | >250, 000 | Genetic risk factors | https://www.decode.com/ |
| Estonian Biobank[61] | Estonia | 2000 | >200, 000 | Medical science | https://genomics.ut.ee/en/content/estonian-biobank |
| Korean Genome and Epidemiology Study[62] | South Korea | 2001 | >10, 000 | Genetic and environmental factors of diseases | http://www.koGES.re.kr/eng/ |
| The International Agency for Research on Cancer (IARC) Biobank (IBB)[63] | France | 1972 | >562, 000 | Disease biomarkers | https://ibb.iarc.fr/ |

access infrastructure provided with the UKBB study has made it one of the most valuable human genetics bioresources ever generated.

## 2 GWAS captures the links between genotypes and phenotypes

GWAS captures the links between genotypes and phenotypes by examining the statistical correlation between the phenotype of interest and millions of SNPs across the entire genome of a large number of individuals[70]. GWAS has been widely used to identify phenotypic variants that are associated with a broad range of human phenotypes, including complex diseases, traits, and drug response[71]. The basic steps of GWAS include selecting a large sample of individuals with and without a particular phenotype, genotyping these individuals using high-throughput genotyping technologies, and then analyzing the genotype data to identify SNPs that are significantly associated with the given phenotype[72]. The statistical significance of the associations is usually assessed using a genome-wide significance threshold to account for multiple testing[73]. This process is illustrated in Fig. 2: "case-control".

GWAS has led to the identification of thousands of genetic variants that are associated with various human phenotypes, providing insights into the biological mechanisms underlying complex diseases and traits[26]. The availability of genome-wide genotype data collected from all UKBB participants, together with the biobank's vast amount of phenotype data, have generated a singular resource of considerable size that provides opportunities for the discovery of new genetic associations and the genetic basis of complex traits and diseases.

Although GWAS can capture the statistical correlation between genotypes and phenotypes, it cannot provide much insight into the mechanisms by which these SNPs affect the phenotypes. To unravel the mechanisms underlying these associations, integrating GWAS results with omics data such as epigenomics, transcriptomics, and proteomics is necessary[74]. Epigenetic modifications, such as DNA methylation and histone modifications, can affect gene expression and potentially mediate the effects of genetic variants on phenotype[75]. Transcriptomics involves the study of gene expression, which can provide information about which genes are differentially expressed in individuals with the phenotype of interest[76]. Proteomic data can complement GWAS by providing additional information on the functional consequences of genetic variation[77]. In general, the integration of various omics data with GWAS data can help identify the regulatory elements that are affected by genetic variants and make it possible for us to explore the functioning pathways of genetic variants that modulate complex traits and diseases.

## 3 Omics data for the construction of context-specific regulatory networks

With the advancement of next-generation sequencing technologies, it has become increasingly feasible to generate large-scale genomic data from individuals. Germline genotypes and phenotypes information in health records are expected to become available for most individuals with access to good healthcare systems[78]. However, to clarify their relationship, we will need to construct models to connect the different omics layers in Fig. 1. The construction of these models is typically based on reference data, which refers to genomic data and omics data that has been generated from large population cohorts and serves as a reference for comparison with personal genomic data. Reference omics data can be used to identify genetic variants that are common or rare in the population, and to annotate functional elements in the genome such as regulatory regions, protein-coding genes, and non-

**Table 2 Data overview of UK Biobank.**

| Data Type | Detail | Number of phenotypes | Number of Participants |
|---|---|---|---|
| Questionnaire and interview | | | |
| Sociodemographic data | Includes ethnicity, education, employment, household information, Townsend deprivation index | 29 | ~500, 000 |
| Family history and early life | Includes illnesses of fathermothersiblings, age of parents, age parents died, number of siblings, birthplace, birth weight, breastfed, childhood body size and height, maternal smoking, handedness, adopted, and part of multiple birth | 28 | ~500, 000 |
| Psychosocial factors | Includes social support, bipolarmajor depression, anxiety, nerves, psychological traits, and mood | 48 | ~500, 000 |
| Lifestyle | Includes information of smoking, alcohol consumption, physical activity, diet, sleep, electronic device use, sun exposure, and sexual factors | 155 | ~500, 000 |
| Medical history | Includes medical conditions, medications, operations, cancer screening, pain, oral health, eyesight, hearing, and general health | 102 | ~500, 000 |
| Cognitive function | Includes prospective memory, pairs matching, fluid intelligence, reaction time, and numeric memory | 121 | ~500, 000 |
| Physical measures | | | |
| Blood pressure | Includes two blood pressure measures taken 1 min apart using a digital blood pressure monitor | 10 | ~500, 000 |
| Hand grip strength | Includes right and left hand isometric grip strength | 5 | ~500, 000 |
| Anthropometrics | Includes standing/sitting height, waist/hip circumference, weight body mass index, and whole body bio-impedance measures | 59 | ~500, 000 |
| Spirometry | Includes two to three blows measurement within a 6 min period | 37 | ~500, 000 |
| Heel bone density | Includes ultrasound measurement of the heel | 41 | ~500, 000 |
| Arterial stiffness | Includes pulse wave velocity using infra-red sensor at the finger | 14 | ~200, 000 |
| Hearing test | Includes reaction on speech-in-noise | 31 | ~200, 000 |
| Eye measures | Includes eye surgery complications, visual acuity, autorefraction, intraocular pressure, and retinal coherence tomography | 333 | ~100, 000 |
| Cardiorespiratory fitness plus ECG | Includes heart rate monitoring results using a four-lead electrocardiograph during cycle ergometry on a stationary bike | 45 | ~100, 000 |
| Web-based questionnaires | | | |
| Diet | Includes information on consumption of over 200 food and drink items over the last 24 hour | 473 | ~210, 000 |
| Cognitive function | Includes a series of cognitive tests, of which four were repeated from the baseline assessment (fluid intelligence, reaction time, numeric memory, pairs test) in addition to two further tests (trail making, symbol digit substitution) | 56 | ~120, 000 |
| Occupational history | Included information on lifetime employment history, occupational exposures and related medical information | 100 | ~120, 000 |
| Mental health | Included information on lifetime mental health events (including depression, bipolar affective disorder, and generalized anxiety disorder), alcohol and cannabis use, unusual and psychotic experiences, traumatic events, self-harm behaviours and subjective wellbeing | 142 | ~150, 000 |
| Enhancement | | | |
| Physical activity monitor | Includes results from Axivity AX3 tri-axial wrist accelerometer for a 7-day period | 210 | ~100, 000 |
| Biochemical measures | Includes thirty-four biomarkers using the plasma, serum, red blood cells, and urine samples. Biomarkers are selected because they are established risk factors for disease (e.g. sex hormones for cancer), diagnostic measures (e.g. HbA1C for diabetes) or they are used to characterize phenotypes (e.g. cystatin C and creatinine for renal function). | 978 | ~500, 000 |
| Genotyping | Includes SNP array results covers ☒800 000 SNPs and indel markers covering markers of specific interest, rare coding variants and genome-wide coverage. Seventy-three million SNPs, short indels, and large structural variants have been imputed. WGS are still ongoing but will be released soon | 271 | ~500, 000 |
| Multi-modal imaging | Includes MRI of brain, heart and body, carotid ultrasound and whole body DXA scan of bones and joints | 2 691 | ~100, 000 |
| Electronic medical records | | | |
| Death registry | ICD-10 coded national death registry data obtained from the Health and Social Care Information Centre (now NHS Digital) for England and Wales and the Information Services Department (ISD) for Scotland. Contains information on source of death report, date, age and cause(s) of death | 8 | ~14, 000 |
| Cancer registry | ICD-9 and -10 coded national cancer registry data obtained from HSCIC for England and Wales and the ISD for Scotland. Contains information on source of cancer report, date and age at diagnosis, site, histology, and behaviour of the cancer. | 9 | ~79, 000 |
| Hospital inpatient data | ICD-9 and -10 coded hospital inpatient episodes obtained from the Hospital Episode Statistics provider for England, the Patient Episode Data for Wales and the Scottish Morbidity Records for Scotland. Contains information on admission and discharge, operations, diagnoses, maternity care, and psychiatric care. Main and secondary diagnoses/operations as well as date of diagnosis/operation are included. | 80 | ~400, 000 |
| Primary care data | Contain coded data from primary care records, including diagnoses, prescriptions, referrals etc. | 3 | pending |

coding RNAs.

The biological process illustrated in Fig. 1 typically happens within a single cell, which may belong to a specific cell line, tissue, or organ (Fig. 2C). Thus, the different layers of omics data processed in the cell are context-specific. The context-specific regulatory network is a common bridge used to link genotypes to phenotypes, which is typically constructed based on public reference data to provide a wealth of information on the omics molecules and their interactions. Advancements in high-throughput omics technologies have led to an explosion of reference data[79, 80, 81]. In the past decade, multiple levels of omics data—including whole-genome DNA sequencing data, DNA methylation, chromatin accessibility, histone modifications, the binding of transcription factors, chromatin interactions, RNA expression levels, and proteomics—have been generated to explore biological regulatory process and model the mechanism between genotypes and phenotypes[82, 83]. The resources of these omics data have been collected by organized projects such as the Encylopedia of DNA Elements (ENCODE) project[13], the Genotype-Tissue Expression (GTEx) project[14], ROADMAP epigenomics project[84], and so on. Here is a brief introduction for several of the most commonly used projects.

ENCODE project[13]: The ENCODE project is a collaborative effort involving hundreds of studies from around the world to identify and annotate all functional elements in the human genome. ENCODE delivers 9, 239 experiments (7, 495 in human and 1, 744 in mouse) in more than 500 cell types and tissues[40], including mapping of transcribed regions and transcript isoforms, regions with transcription factor binding or histone modifications, open chromatin elements, 3D chromatin interactions and other functional annotation. These data are publicly available at the ENCODE portal (http://www.encodeproject.org) and have been widely used to study the function and regulation of the human genome.

GTEx project[14]: GTEx studies the relationship between genetic variation and gene expression across multiple human tissues. GTEx has generated gene expression data for 54 non-diseased tissue sites across nearly 1, 000 individuals, as well as genomic data on more than 840, 000 genetic variants, primarily for molecular assays including WGS, WES, and RNA-Seq. The latest version of GTEx provides the genotypes of 838 donors and the expression levels of 17, 382 samples in 52 tissues and two cell lines. The project has also developed a number of tools and resources for analyzing and visualizing the data, including an online portal that allows us to explore gene expression patterns across different tissues and genetic backgrounds. All resources can be found at https://gtexportal.org/home/.

ROADMAP epigenomics project[84]: The NIH Roadmap Epigenomics Mapping Consortium produces a public resource of human epigenomic data to catalyze basic biology and disease-oriented research. The project has generated high-quality, genome-wide maps of several key histone modifications, chromatin accessibility, DNA methylation and mRNA expression across over 100 human cell types and tissues, providing uniformly processed datasets, integrative analysis products and interactive genome browser sessions. The processed data are available at https://egg2.wustl.edu/roadmap/web_portal/processed_data.html.

Besides the above projects that focus on bulk-level data, recent advancements in single-cell technologies have revolutionized our ability to dissect the complex tissues with single-cell resolution. The Human Cell Atlas (HCA)[85] is an international collaborative consortium that charts the cell types in the healthy body. It aims to create comprehensive reference maps of all human cells as a basis for identifying the common cell types in tissues from the major human organs and understanding human health and diseases. So far, HCA scientists have identified more than 39 million cells from 15 major organ systems, such as 11.1 million nervous system cells, 5.8 million embryonic and fetal cells, 3.4 million lung cells, and 7.2 million immune cells. These atlases also include important human diseases, such as nearly 4.8 million cells derived from individuals infected with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The processed data are available at https://www.humancellatlas.org. The Human BioMolecular Atlas Program (HuBMAP)[86] is another public database that provides a comprehensive map of the human body at the cellular and molecular level. The database contains data from various sources, including imaging data, genomic data, and clinical data. The goal of the database is to provide a comprehensive view of the human body, which can be used to study the complex regulation mechanism and develop new treatment for various diseases.

In addition to conventional single cell data, spatial transcriptomics technologies[87] have emerged to allow simultaneous profiling of transcriptomes and spatial locations of cells. This type of data allows us to study transcriptomes of cells in relation to their cellular organization. Many studies indicate that spatially variable genes are the potential novel markers or essential regulators for tissue pattern formation and homeostasis[88]. Thus, these spatial transcriptomics data have great potential to provide detailed molecular maps for investigating the complex context-specific regulatory networks.

## 4 Computational approaches for the construction of regulatory networks

Approaches for regulatory network construction typically involve the integration of various types of experimental and computational data[89, 90]. Wang et al. discuss the efforts of integrating the DNA accessibility data, transcriptional data, and functional genomic regions together to enable the accurate interpretation of regulatory landscape[91]. Duren et al. propose a statistical approach, named PECA, to build gene regulatory networks based on paired expression and chromatin accessibility data across diverse cellular contexts[92]. Xin et al. develop a variant interpretation methodology (vPECA) to identify active selected regulatory elements and associated regulatory network based on temporal data of paired ATAC-seq and RNA-seq data[93]. In general, network inference approaches use statistical models to identify regulatory interactions between genes and other molecules, then the regulatory networks can be spontaneously constructed using the predicted interactions[94]. Up to now, there have been lots of computational methods developed to construct context-specific regulatory networks from omics data, including regression-based methods, Bayesian networks, machine learning models, and single-cell based methods[95, 96, 97, 98, 99].

Regression-based methods: Regression-based methods assume a linear or non-linear relationship between the abundance or expression of the target molecules and its potential regulators in a particular tissue or a specific biological context. The inference methods can mine the essential rules on partial omics data, discover interactions reflected by the molecular level and finally present complex regulatory relationships in the form of network[100]. Compared with simple correlation-based methods, regression models can be combined with regularization approaches, such as linear regression[101], elastic net[102], or support vector regression (SVR)[103], to improve the accuracy of the

regulatory network and reduce overfitting.

One of the typical applications of the regression-based methods in the construction of regulatory networks is the calculation of the correlation between genotypes and omics features, which is also known as QTLs. QTL studies are widely used to explore how genetic variation functions and affects the quantitative molecules in a specific cellular context[104] (Fig. 3). In a QTL study, researchers typically genotype a large number of individuals and measure the levels of certain molecules (e.g. CpGs, genes, or proteins). Then by analyzing the correlation between the genotypes and molecular levels, regions of the genome that are likely to harbor genetic variants that influence the molecules can be identified[105].

mQTLs (methylation quantitative trait loci) are genetic variants associated with changes in DNA methylation levels, where DNA methylation is an epigenetic modification that can affect gene expression and not alter the DNA sequence itself[106]. eQTLs (expression quantitative trait loci) are genetic variants associated with changes in gene expression levels. These variants can be located within a gene or in regions that regulate gene expression. eQTL studies statistically link SNPs to genes and can help to identify genes and pathways that participate in the mechanism of disease or other complex traits104. pQTL (protein quantitative trait loci) are genetic variants that are associated with changes in protein levels or protein function. pQTL studies can help to identify genes involved in the regulation of protein function, which is important for the understanding of many biological processes[107]. We introduced several studies of mQTLs, eQTLs, and pQTLs in Table 3, all of whose QTLs are publicly available. Overall, these different types of QTLs can be applied as part of the context-specific regulatory networks and provide a complementary understanding of the genetic regulation of complex traits.

Bayesian networks: Bayesian network model has become a powerful tool for constructing gene regulatory networks with its solid theoretical foundation, natural representation of knowledge structure, and flexible reasoning ability[121]. In a Bayesian framework, the probability of a regulatory interaction between two molecules is calculated based on the available data and prior knowledge. One common application of Bayesian networks is to construct dynamic regulatory networks using the longitudinal data processed during biological development or in response to a perturbation such as drug treatments or genetic manipulations[122]. For example, the dynamic Bayesian network can be used to model the time-varying relationships within molecules and captures interactions that drive changes over time[123].

Machine learning models: To use machine learning methods for regulatory network construction, we first need gene expression or other relevant genomics data, as well as a set of known regulatory interactions as a training set. The resulting trained model can then be used to predict new regulatory interactions between molecules[124]. The most popular machine learning models used to construct regulatory networks include decision trees, random forests, support vector machines (SVMs), and deep learning-based neural networks[125]. For example, Zhou et al. developed a deep learning-based framework, DeepSEA, that directly learns a regulatory sequence code from large-scale chromatin-profiling data and enables the prediction of chromatin effects of sequence alterations with single-nucleotide sensitivity[126], and the underlying mechanism of this model was interpretated by NeuronMotif[127]. Avsec et al. proposed a deep learning architecture, Enformer, which substantially improves gene expression prediction accuracy from DNA sequences, yielding more accurate variant effect predictions on gene expression and providing predicted enhancer-promoter interactions[128]. Both DeepSEA and Enformer can be used to predict the regulatory effects of non-coding variants on context-specific gene expression, which links the SNPs to genes and provides functional interactions for the construction of the context-specific regulatory networks.

Single cell-based methods: Single-cell expression data are especially promising for computing gene regulatory networks (GRNs) because they do not obscure biological signals by averaging over all the cells in a sample. However, these data have features including substantial cellular heterogeneity, cell-to-cell variation in sequencing depth, high sparsity caused by dropouts and cell-cycle-related effects, that pose significant difficulties.

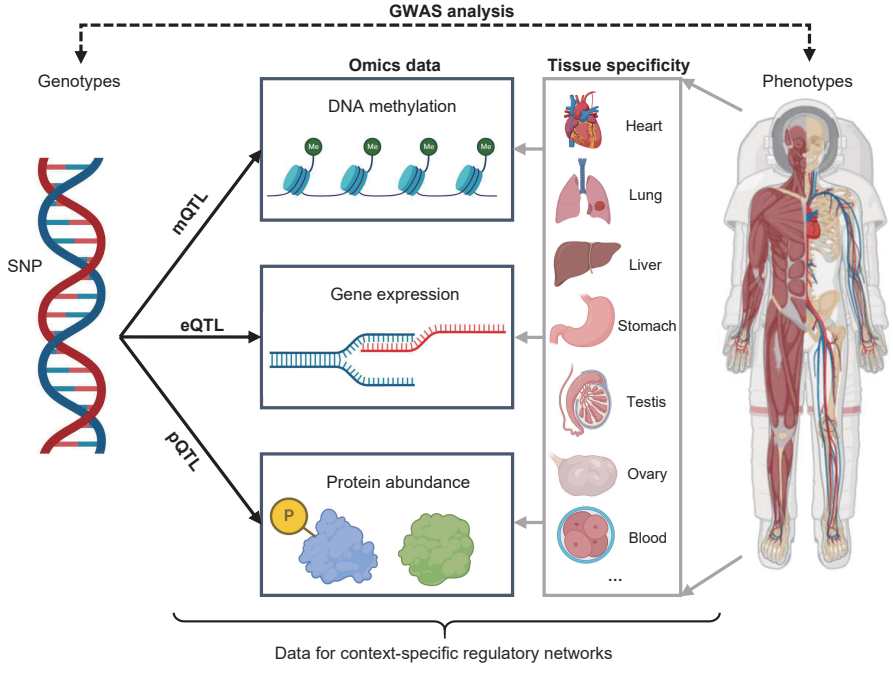


**Fig. 3 Scheme of QTLs and its role in mechanism modeling.**

**Table 3 Summary of the resources for different types of QTLs.**

| Type | Database/Paper | Sample Size | Number of QTLs | Year | Homepage |
|---|---|---|---|---|---|
| mQTL | Methylation QTL Database[106] | 5, 533 | 175, 091 | 2016 | http://www.mqtldb.org/ |
| mQTL | BIOS QTL Browser[108] | 3, 841 | 272, 037 cis; 18, 764 trans | 2017 | http://www.genenetwork.nl/biosqtlbrowser |
| mQTL | Min et al.[109] | 27, 750 | >270, 000 | 2021 | http://mqtldb.godmc.org.uk |
| mQTL | Hawe ea al.[110] | ~7, 000 | 11, 165, 559 | 2022 | https://zenodo.org/record/5 196 216#.YRZ3TfJxeUk |
| mQTL | FHS_meQTLs[111] | 4 170 | 4, 700, 000 cis; 630, 000 trans | 2019 | https://ftp.ncbi.nlm.nih.gov/eqtl/original_submissions/FHS_meQTLs/ |
| mQTL | Pancan-meQTL[112] | 7, 242 | 8, 028, 964 cis; 965, 050 trans | 2018 | http://gong_lab.hzau.edu.cn/Pancan-meQTL/ |
| eQTL | GTExPortal[14] | ~10,000 | ~30,000,000 | 2020 | https://gtexportal.org/home/ |
| eQTL | GTExPortal[14] | ~10,000 | ~30,000,000 | 2020 | https://gtexportal.org/home/ |
| eQTL | eQTLGen Consortium eQTL[104] | 31, 684 | 10, 507, 665 cis; 59, 787 trans | 2021 | https://www.eqtlgen.org/ |
| eQTL | GEUVADIS[113] | 462 | 18, 366 | 2013 | https://www.ebi.ac.uk/ |
| eQTL | MRCA eQTLs[114] | 1, 350 | 7, 302 | 2013 | https://www.hsph.harvard.edu/liming-liang/software/eqtl/ |
| eQTL | Brain eQTL[115] | 412 | 1, 815, 172 | 2018 | https://eqtl.brainseq.org/ |
| eQTL | PancanQTL[116] | 9, 196 | 5, 606, 570 cis; 231, 210 trans | 2017 | http://gong_lab.hzau.edu.cn/PancanQTL/ |
| pQTL | Sun et al.[117] | 3, 301 | 1, 927 | 2018 | http://www.phpc.cam.ac.uk/ceu/proteins/ |
| pQTL | Ferkingstad et al.[118] | 35, 559 | 18, 084 | 2021 | https://www.nature.com/articles/s41588-021-00978-w#MOESM1 |
| pQTL | Yao et al.[119] | 6, 861 | >16, 000 | 2018 | https://preview.ncbi.nlm.nih.gov/gap/eqtl/studies/ |
| pQTL | Zhang et al.[103] | ~9, 000 | 4, 069 | 2022 | http://nilanjanchatterjeelab.org/pwas |
| pQTL | Gudjonsson et al.[120] | 5, 368 | 4, 035 | 2022 | https://doi.org/10.528 1/zenodo.5711426 |

Despite these challenges, over a dozen methods have been developed or used to infer GRNs from single-cell data. We can categorize these methods into three groups based on how the network is constructed: differential equation, gene correlation, and correlation ensemble over pseudo-time. For example, PPCOR is a differential equation-based R package that computes the partial and semi-partial correlation coefficients for every pair of genes, with respect to all the other genes[129]. SCENIC computes the regulatory network for each gene independently using tree-based ensemble methods[130]. LEAP utilizes pseudo time-ordered data and calculates the Pearson's correlation of normalized mapped-read counts over temporal windows of a fixed size with different lags to construct the GRN[131].

The above methods can be incorporated with the prior knowledge about the biology of the system, such as known interactions between genes and proteins provided by GO[132], KEGG[133], STRING[134] and other public databases, to construct a more comprehensive regulatory network.

## 5 Applications of the regulatory networks in modeling the causal mechanism

Context-specific regulatory networks have been used in a wide range of applications to demonstrate and understand the complex relationships between genes, proteins, and other biological molecules in a particular cellular context[135]. The networks can be constructed based on the complex interactions within epigenomics, transcriptomics and proteomics, where the nodes typically represent biological entities such as genes, proteins, or other molecules, and the edges represent the interactions between these entities[136]. The typical context-specific regulatory networks include gene regulatory network[34], gene co-expression network[137], protein-protein interaction networks[134], promoter-enhancer networks[138], as well as the integration of multi-omics networks[139, 140, 141, 142, 143, 144].

These networks can help to explain how genetic variations or perturbations affect cellular behavior and lead to changes in phenotype, contributing to the explanation of the relationship between genotypes and phenotypes[145, 146]. Here are several examples of how the context-specific regulatory networks have been used in different applications (Fig. 4).

Interpretation of GWAS: GWAS typically identify many genetic variations associated with a complex trait, but can hardly recognize which variations are causally linked to the phenotype[147]. Context-specific regulatory networks can be used to prioritize candidate genes by identifying which genes are functionally related to the phenotype of interest and are likely to be affected by the genetic variation[148]. Integrating GWAS results with regulatory networks can identify the genes and pathways that are dysregulated in the disease state and the genetic variations that contribute to this dysregulation[149]. For example, Finucane et al. introduce a method, stratified linkage disequilibrium (LD) score regression, for partitioning heritability from GWAS summary statistics while accounting for linked functional elements[150]. Zhu et al. develop a Bayesian framework that integrates GWAS summary statistics with context-specific regulatory networks to infer genetic enrichments and associations simultaneously[151].

Cancer genomics deciphering: Cancer cells are heterogeneous and can differ in their genomic, transcriptomic, and epigenomic profiles[152]. The context specificity of the regulatory network allows us to model the context-specific molecular interactions that contribute to cancer heterogeneity and helps to identify mutations that play an important role in cancer development and progression[153, 154]. Besides, with the context-specific regulatory network, we can also detect genes and pathways that are dysregulated in cancer cells[155, 156], providing insights into the molecular mechanisms underlying cancer genomics.

Personalized medicine: The context-specific regulatory networks can help with the identification of personalized treatment options based on an individual's genetic profile[157]. By

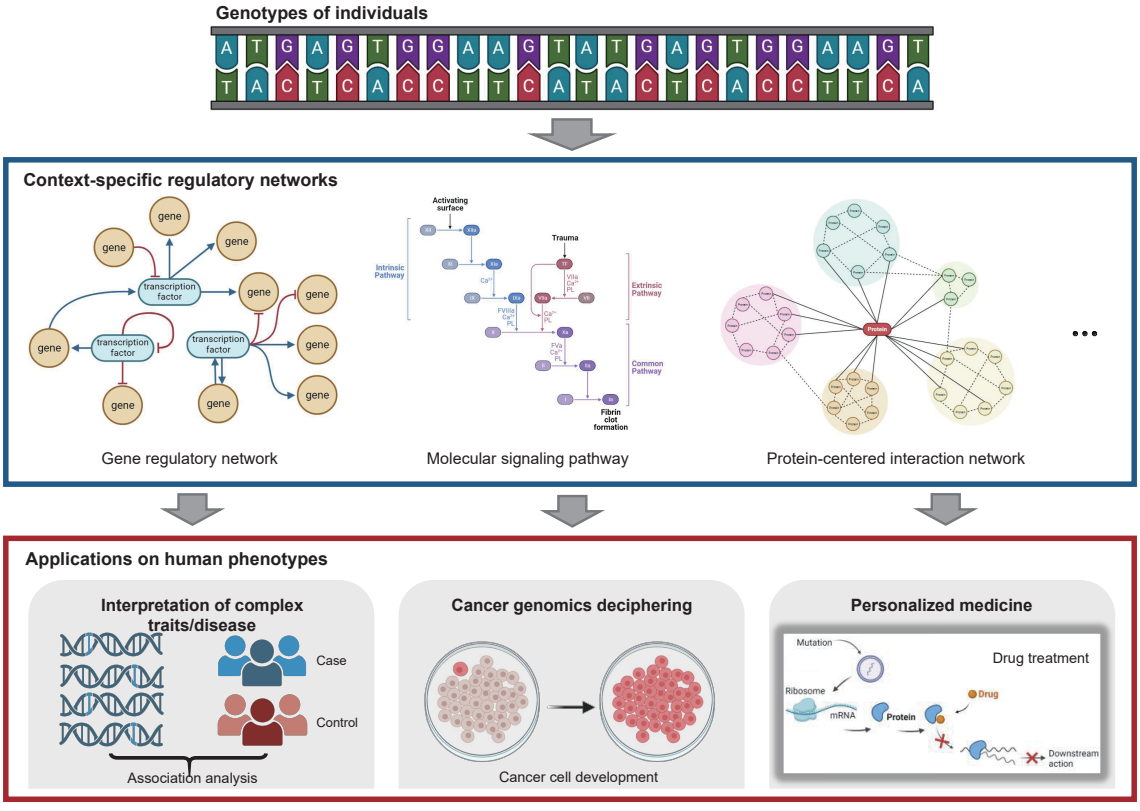


**Fig. 4 Applications of the context-specific regulatory network in modeling the causal mechanism of phenotypes.**

modeling the interactions between genes, proteins, and other molecules in an individual's cells, we can identify the optimal treatment options for a particular disease for the individual. The context-specific regulatory network can also be used to predict the response to specific mutations by modeling the interactions between the mutation and the phenotype of interest[158]. By identifying the genes and pathways that are affected by the mutation and their downstream effects on cellular behavior, we can predict the influence of the individual's mutation on the given phenotype, providing a solution for the development of personalized medicine[159].

Collectively, the context-specific regulatory network is a powerful tool for deciphering the relationships between genotypes and phenotypes, providing insights into the molecular mechanisms underlying complex traits and diseases and contributing to the identification of potential targets for therapeutic interventions.

## Future goals and challenges

In this paper, we review diverse genomics and omics data provided by biobanks and other organized projects, discuss the approaches commonly used to construct context-specific regulatory networks with omics data, and elaborate the importance of context-specific regulatory networks as part of the causal model linking genotypes to phenotypes. To sum up, modeling the causal mechanism between genotypes and phenotypes via context-specific regulatory networks helps in improving our understanding of biological systems underlying various phenotypes. However, some technological and analytical improvements will still be needed for the construction and application of reliable context-specific regulatory networks.

First, context-specific regulatory networks provide a static snapshot of the regulatory landscape, which may not capture dynamic changes in regulatory interactions over time or in response to different stimuli or conditions[160]. To address this limitation, we may need to generate additional datasets under different conditions or perturbations to obtain a more comprehensive and dynamic view of the regulatory network and apply causal inference method for analysis[161]. Perturbing a biological system, such as a cell or a tissue, can provide valuable insights into the underlying regulatory mechanisms that govern the system's behavior[162]. Generating reference data after perturbation provides important information on how the regulatory network changes in response to the perturbation[163]. For example, if a gene is knocked out or silenced, its downstream targets may also be affected, resulting in changes to the regulatory network. By comparing the regulatory network before and after perturbation, we can identify the specific regulatory interactions that are affected by the perturbation and gain insights into the underlying mechanism.

Second, biobank data provide a wealth of genomic and phenotypic data from large and diverse populations and include a variety of data types, such as epigenetic data, gene expression data, and clinical data, enabling the development of context-specific regulatory networks that capture the complexity of the biological system. However, there are also challenges for regulatory networks to interpret biobank data. On the one hand, biobank data can be highly heterogeneous, with variation in sample size, data quality, and data types. This heterogeneity can make it challenging to integrate different types of data into a coherent model. On the other hand, biological systems are highly complex, and the interactions between genes, proteins, and other molecules can be difficult to model. This complexity would make it difficult to identify causal relationships between genotypes and phenotypes.

Third, the accuracy of regulatory networks is highly dependent on the computational methods used to analyze the omics data and construct networks. Therefore, developing new statistical and

machine learning approaches that can handle large and diverse datasets and are less prone to overfitting or misinterpretation is of significance. Besides, the integration of multiple types of omics data can provide a more comprehensive view of the underlying regulatory mechanisms that govern a biological system. Computational approaches for the integration of genomic, epigenomic, transcriptomic, proteomic, and other relevant data to build a more complete picture of the regulatory landscape should also be developed. In addition, context-specific regulatory networks can also be integrated with other types of biological networks, such as metabolic networks and molecular signaling networks, to better understand the complex interplay between different biological processes and their relationship to phenotype.

Fourth, biological systems are highly complex and involve numerous interacting components, such as genes, proteins, and regulatory elements, that operate at multiple levels and are subject to a wide range of internal and external factors. Since context-specific regulatory networks are based on statistical and machine learning models, they can be sensitive to noise and biases in the data, leading to incorrect or misleading interpretations of the underlying regulatory mechanisms. Besides, the interpretation of context-specific regulatory networks requires a deep understanding of the biological processes and pathways involved, as well as the technical details of the data generation and analysis. This can be a significant challenge for researchers with limited experience in computational biology and bioinformatics.

Overall, the goal of modeling the causal mechanism between genotypes and phenotypes via context-specific regulatory networks is to provide a more comprehensive and accurate understanding of the underlying biological mechanisms between genotypes and phenotypes. However, this requires overcoming a number of challenges, including the generation of more reliable datasets, the development of more accurate and robust computational methods, the integration of diverse and complex datasets, and the translation of this knowledge into clinical practice.

## Acknowledgments

W.L. is supported by the National Natural Science Foundation of China (NSFC) (Grant No. 32200472), China Postdoctoral Science Foundation (Grant No. 2021M693274 and BX2021336). W.W. and W.Z. are supported by NIH (Grant No. HG007735 and HG010359).

## Article History

Received: 29 March 2023; Revised: 15 May 2023; Accepted: 7 July 2023

## References

[1] Richard A Gibbs, The human genome project changed everything, *Nature Reviews Genetics*, vol. 21, no. 10, pp. 575–576, 2020.

[2] Ting Wang, Lucinda Antonacci-Fulton, Kerstin Howe, Heather A Lawson, Julian K Lucas, Adam M Phillippy, Alice B Popejoy, Mobin Asri, Caryn Carson, Mark JP Chaisson, et al, The human pangenome project: a global resource to map genomic diversity, *Nature*, vol. 604, no. 7906, pp. 437–446, 2022.

[3] Rachel M Sherman and Steven L Salzberg, Pan-genomics in the human genome era, *Nature Reviews Genetics*, vol. 21, no. 4, pp. 243–254, 2020.

[4] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al, The mutational constraint spectrum quantified from variation in 141, 456 humans, *Nature*, vol. 581, no. 7809, pp. 434–443, 2020.

[5] Gundula Povysil, Slavé Petrovski, Joseph Hostyk, Vimla Aggarwal, Andrew S Allen, and David B Goldstein, Rare-variant collapsing analyses for complex traits: guidelines and applications, *Nature Reviews Genetics*, vol. 20, no. 12, pp. 747–759, 2019.

[6] Yukihide Momozawa and Keijiro Mizukami, Unique roles of rare variants in the genetics of complex diseases in humans, *Journal of human genetics*, vol. 66, no. 1, pp. 11–23, 2021.

[7] Wenran Li, Zhana Duren, Rui Jiang, and Wing Hung Wong, A method for scoring the cell type-specific impacts of noncoding variants in personal genomes, *Proceedings of the National Academy of Sciences*, vol. 117, no. 35, pp. 21364–21372, 2020.

[8] Quanli Wang, Ryan S Dhindsa, Keren Carss, Andrew R Harper, Abhishek Nag, Ioanna Tachmazidou, Dimitrios Vitsios, Sri VV Deevi, Alex Mackay, Daniel Muthas, et al, Rare variant contribution to human disease in 281, 104 uk biobank exomes, *Nature*, vol. 597, no. 7877, pp. 527–532, 2021.

[9] Eric Vallabh Minikel, Konrad J Karczewski, Hilary C Martin, Beryl B Cummings, Nicola Whiffin, Daniel Rhodes, Jessica Alföldi, Richard C Trembath, David A van Heel, Mark J Daly, et al, Evaluating drug targets through human loss-of-function genetic variation, *Nature*, vol. 581, no. 7809, pp. 459–464, 2020.

[10] Ryan L Collins, Harrison Brand, Konrad J Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C Francioli, Amit V Khera, Chelsea Lowther, Laura D Gauthier, Harold Wang, et al, A structural variation reference for medical and population genetics, *Nature*, vol. 581, no. 7809, pp. 444–451, 2020.

[11] Keren J Carss, Aimee M Deaton, Alberto Del Rio-Espinola, Dorothée Diogo, Mark Fielden, Diptee A Kulkarni, Jonathan Moggs, Peter Newham, Matthew R Nelson, Frank D Sistare, et al, Using human genetics to improve safety assessment of therapeutics, *Nature Reviews Drug Discovery*, vol. 22, no. 2, pp. 145–162, 2023.

[12] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al, The uk biobank resource with deep phenotyping and genomic data, *Nature*, vol. 562, no. 7726, pp. 203–209, 2018.

[13] EA Feingold, PJ Good, MS Guyer, S Kamholz, L Liefer, K Wetterstrand, FS Collins, TR Gingeras, D Kampa, EA Sekinger, et al, The encode (encyclopedia of dna elements) project, *Science*, vol. 306, no. 5696, pp. 636–640, 2004.

[14] GTEx Consortium, The gtex consortium atlas of genetic regulatory effects across human tissues, *Science*, vol. 369, no. 6509, pp. 1318–1330, 2020.

[15] A Marieke Oudelaar and Douglas R Higgs, The relationship between genome structure and function, *Nature Reviews Genetics*, vol. 22, no. 3, pp. 154–168, 2021.

[16] Inigo Barrio-Hernandez, Jeremy Schwartzentruber, Anjali Shrivastava, Noemi Del-Toro, Asier Gonzalez, Qian Zhang, Edward Mountjoy, Daniel Suveges, David Ochoa, Maya Ghoussaini, et al, Network expansion of genetic associations defines a pleiotropy map of human cell biology, *Nature genetics*, vol. 55, no. 3, pp. 389–398, 2023.

[17] Pisanu Buphamalai, Tomislav Kokotovic, Vanja Nagy, and Jörg Menche, Network analysis reveals rare disease signatures across multiple levels of biological organization, *Nature communications*, vol. 12, no. 1, pp. 6306, 2021.

[18] William Villiers, Audrey Kelly, Xiaohan He, James Kaufman-Cook, Abdurrahman Elbasir, Halima Bensmail, Paul Lavender, Richard Dillon, Borbála Mifsud, and Cameron S Osborne, Multi-omics and machine learning reveal context-specific gene regulatory activities of pml:: Rara in acute promyelocytic leukemia, *Nature Communications*, vol. 14, no. 1, pp. 724, 2023.

[19] Sascha Jung and Antonio Del Sol, Multiomics data integration unveils core transcriptional regulatory networks governing cell-type identity, *NPJ systems biology and applications*, vol. 6, no. 1,

pp. 26, 2020.

[20] Andrea Califano, Atul Butte, Stephen Friend, Trey Ideker, and Eric Schadt, Integrative network-based association studies: Leveraging cell regulatory models in the post-gwas era, *Nature Precedings*, pp. 1–1, 2011.

[21] Sarvenaz Choobdar, Mehmet E Ahsen, Jake Crawford, Mattia Tomasoni, Tao Fang, David Lamparter, Junyuan Lin, Benjamin Hescott, Xiaozhe Hu, Johnathan Mercer, et al, Assessment of network module identification across complex diseases, *Nature methods*, vol. 16, no. 9, pp. 843–852, 2019.

[22] Sara Brin Rosenthal, Sarah N Wright, Sophie Liu, Christopher Churas, Daisy Chilin-Fuentes, Chi-Hua Chen, Kathleen M Fisch, Dexter Pratt, Jason F Kreisberg, and Trey Ideker, Mapping the common gene networks that underlie related diseases, *Nature protocols*, pp. 1–15, 2023.

[23] Mark B Gerstein, Anshul Kundaje, Manoj Hariharan, Stephen G Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, Ekta Khurana, Joel Rozowsky, Roger Alexander, et al, Architecture of the human regulatory network derived from encode data, *Nature*, vol. 489, no. 7414, pp. 91–100, 2012.

[24] Megha Padi and John Quackenbush, Detecting phenotype-driven transitions in regulatory network structure, *NPJ systems biology and applications*, vol. 4, no. 1, pp. 16, 2018.

[25] Phillip L Davidson, Haobing Guo, Jane S Swart, Abdull J Massri, Allison Edgar, Lingyu Wang, Alejandro Berrio, Hannah R Devens, Demian Koop, Paula Cisternas, et al, Recent reconfiguration of an ancient developmental gene regulatory network in heliocidaris sea urchins, *Nature Ecology & Evolution*, vol. 6, no. 12, pp. 1907–1920, 2022.

[26] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina De Vries, Yukinori Okada, Alicia R Martin, Hilary C Martin, Tuuli Lappalainen, and Danielle Posthuma, Genome-wide association studies, *Nature Reviews Methods Primers*, vol. 1, no. 1, pp. 59, 2021.

[27] Emily Clough and Tanya Barrett, The gene expression omnibus database, *Statistical Genomics: Methods and Protocols*, pp. 93–110, 2016.

[28] Kai Zhang, James D Hocker, Michael Miller, Xiaomeng Hou, Joshua Chiou, Olivier B Poirion, Yunjiang Qiu, Yang E Li, Kyle J Gaulton, Allen Wang, et al, A single-cell atlas of chromatin accessibility in the human genome, *Cell*, vol. 184, no. 24, pp. 5985–6001, 2021.

[29] Silvia Domcke, Andrew J Hill, Riza M Daza, Junyue Cao, Diana R O'Day, Hannah A Pliner, Kimberly A Aldinger, Dmitry Pokholok, Fan Zhang, Jennifer H Milbank, et al, A human cell atlas of fetal chromatin accessibility, *Science*, vol. 370, no. 6518, pp. eaba7612, 2020.

[30] Andreas Digre and Cecilia Lindskog, The human protein atlas—spatial localization of the human proteome in health and disease, *Protein Science*, vol. 30, no. 1, pp. 218–233, 2021.

[31] Wanwen Zeng, Qiao Liu, Qijin Yin, Rui Jiang, and Wing Hung Wong, Hichipdb: a comprehensive database of hichip regulatory interactions, *Nucleic Acids Research*, vol. 51, no. D1, pp. D159–D166, 2023.

[32] Jérémie Breda, Mihaela Zavolan, and Erik van Nimwegen, Bayesian inference of gene expression states from single-cell rna-seq data, *Nature Biotechnology*, vol. 39, no. 8, pp. 1008–1016, 2021.

[33] Weiguang Mao, Elena Zaslavsky, Boris M Hartmann, Stuart C Sealfon, and Maria Chikina, Pathway-level information extractor (plier) for gene expression data, *Nature methods*, vol. 16, no. 7, pp. 607–610, 2019.

[34] Wenran Li, Meng Wang, Jinghao Sun, Yong Wang, and Rui Jiang, Gene co-opening network deciphers gene functional relationships, *Molecular BioSystems*, vol. 13, no. 11, pp. 2428–2439, 2017.

[35] Wanwen Zeng, Yong Wang, and Rui Jiang, Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network, *Bioinformatics*, vol. 36, no. 2, pp. 496–503, 2020.

[36] Wanwen Zeng, Mengmeng Wu, and Rui Jiang, Prediction of enhancer-promoter interactions via natural language processing, *BMC genomics*, vol. 19, pp. 13–22, 2018.

[37] Qiao Liu, Hairong Lv, and Rui Jiang, hicgan infers super resolution hi-c data with generative adversarial networks, *Bioinformatics*, vol. 35, no. 14, pp. i99–i107, 2019.

[38] Yijie Wang, Hangnoh Lee, Justin M Fear, Isabelle Berger, Brian Oliver, and Teresa M Przytycka, Netrex-cf integrates incomplete transcription factor data with gene expression to reconstruct gene regulatory networks, *Communications Biology*, vol. 5, no. 1, pp. 1282, 2022.

[39] Bjarni V Halldorsson, Hannes P Eggertsson, Kristjan HS Moore, Hannes Hauswedell, Ogmundur Eiriksson, Magnus O Ulfarsson, Gunnar Palsson, Marteinn T Hardarson, Asmundur Oddsson, Brynjar O Jensson, et al, The sequences of 150, 119 genomes in the uk biobank, *Nature*, vol. 607, no. 7920, pp. 732–740, 2022.

[40] Federico Abascal, Reyes Acosta, Nicholas J Addleman, Jessika Adrian, Veena Afzal, Bronwen Aken, and Jennifer A Akiyama, Perspectives on encode, *Nature*, vol. 583, no. 7818, pp. 693–699, 2020.

[41] Gene Ontology Consortium, The gene ontology resource: 20 years and still going strong, *Nucleic acids research*, vol. 47, no. D1, pp. D330–D338, 2019.

[42] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, et al, The reactome pathway knowledgebase 2022, *Nucleic acids research*, vol. 50, no. D1, pp. D687–D692, 2022.

[43] Minoru Kanehisa, Miho Furumichi, Yoko Sato, Masayuki Kawashima, and Mari Ishiguro-Watanabe, Kegg for taxonomy-based analysis of pathways and genomes, *Nucleic acids research*, vol. 51, no. D1, pp. D587–D592, 2023.

[44] Tyler Grimes, S Steven Potter, and Somnath Datta, Integrating gene regulatory pathways into differential network analysis of gene expression data, *Scientific reports*, vol. 9, no. 1, pp. 5479, 2019.

[45] Mitja I Kurki, Juha Karjalainen, Priit Palta, Timo P Sipilä, Kati Kristiansson, Kati M Donner, Mary P Reeve, Hannele Laivuori, Mervi Aavikko, Mari A Kaunisto, et al, Finngen provides genetic insights from a well-phenotyped isolated population, *Nature*, vol. 613, no. 7944, pp. 508–518, 2023.

[46] Kevin L Gunderson, Frank J Steemers, Grace Lee, Leo G Mendoza, and Mark S Chee, A genome-wide scalable snp genotyping assay using microarray technology, *Nature genetics*, vol. 37, no. 5, pp. 549–554, 2005.

[47] Kelly E Ormond, Matthew T Wheeler, Louanne Hudgins, Teri E Klein, Atul J Butte, Russ B Altman, Euan A Ashley, and Henry T Greely, Challenges in the clinical application of whole-genome sequencing, *The Lancet*, vol. 375, no. 9727, pp. 1749–1751, 2010.

[48] Ann-Christine Syvänen, Toward genome-wide snp genotyping, *Nature genetics*, vol. 37, no. Suppl 6, pp. S5–S10, 2005.

[49] Philippe Lamy, Jakob Grove, and Carsten Wiuf, A review of software for microarray genotyping, *Human genomics*, vol. 5, no. 4, pp. 304, 2011.

[50] Zhengming Chen, Junshi Chen, Rory Collins, Yu Guo, Richard Peto, Fan Wu, and Liming Li, China kadoorie biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up, *International journal of epidemiology*, vol. 40, no. 6, pp. 1652–1666, 2011.

[51] Akiko Nagai, Makoto Hirata, Yoichiro Kamatani, Kaori Muto, Koichi Matsuda, Yutaka Kiyohara, Toshiharu Ninomiya, Akiko Tamakoshi, Zentaro Yamagata, Taisei Mushiroda, et al, Overview of the biobank japan project: study design and profile, *Journal of epidemiology*, vol. 27, no. Supplement_III, pp. S2–S8, 2017.

[52] Berthold Huppertz, Michaela Bayer, Tanja Macheiner, and Karine Sargsyan, Biobank graz: the hub for innovative biomedical research, *Open journal of bioresources*, vol. 3, no. 1, 2016.

[53] Catarina Almqvist, Hans-Olov Adami, Paul W Franks, Leif Groop,

Erik Ingelsson, Juha Kere, Lauren Lissner, Jan-Eric Litton, Markus Maeurer, Karl Michaëlsson, et al, Lifegene—a large prospective population-based study of global relevance, *European journal of epidemiology*, vol. 26, pp. 67–77, 2011.

[54] Rain Eensaar. Estonia: ups and downs of a biobank project. In Biobanks, pages 56-70. Routledge, 2008.

[55] Asma Al Thani, Eleni Fthenou, Spyridon Paparrodopoulos, Ajayeb Al Marri, Zumin Shi, Fatima Qafoud, and Nahla Afifi, Qatar biobank cohort study: study design and first results, *American journal of epidemiology*, vol. 188, no. 8, pp. 1420–1433, 2019.

[56] Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al, An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics, *Cell*, vol. 173, no. 2, pp. 400–416, 2018.

[57] Mark A Jensen, Vincent Ferretti, Robert L Grossman, and Louis M Staudt, The nci genomic data commons as an engine for precision medicine, *Blood, The Journal of the American Society of Hematology*, vol. 130, no. 4, pp. 453–459, 2017.

[58] Trevor JB Dummer, Philip Awadalla, Catherine Boileau, Camille Craig, Isabel Fortier, Vivek Goel, Jason MT Hicks, Sébastien Jacquemont, Bartha Maria Knoppers, Nhu Le, et al, The canadian partnership for tomorrow project: a pan-canadian platform for research on chronic disease prevention, *Cmaj*, vol. 190, no. 23, pp. E710–E717, 2018.

[59] Saudi Genome Project Team, et al, The saudi human genome program: An oasis in the desert of arab medicine is providing clues to genetic disease, *IEEE pulse*, vol. 6, no. 6, pp. 22–26, 2015.

[60] Gísli Pálsson. The rise and fall of a biobank. In Biobanks: Governance in comparative perspective, pages 41-55. Routledge London, 2008.

[61] Liis Leitsalu, Toomas Haller, Tõnu Esko, Mari-Liis Tammesoo, Helene Alavere, Harold Snieder, Markus Perola, Pauline C Ng, Reedik Mägi, Lili Milani, et al, Cohort profile: Estonian biobank of the estonian genome center, university of tartu, *International journal of epidemiology*, vol. 44, no. 4, pp. 1137–1147, 2015.

[62] Yeonjung Kim, Bok-Ghee Han, and KoGES Group, Cohort profile: the korean genome and epidemiology study (koges) consortium, *International journal of epidemiology*, vol. 46, no. 2, pp. e20–e20, 2017.

[63] Marjan Mohammadi, Biobanking in the developing world; maximum specimens, minimum infrastructure, *Basic & Clinical Cancer Research*, vol. 9, no. 4, pp. 1–3, 2017.

[64] Jianhua Zhao and Struan FA Grant, Advances in whole genome sequencing technology, *Current pharmaceutical biotechnology*, vol. 12, no. 2, pp. 293–305, 2011.

[65] Frederick E Dewey, Megan E Grove, Cuiping Pan, Benjamin A Goldstein, Jonathan A Bernstein, Hassan Chaib, Jason D Merker, Rachel L Goldfeder, Gregory M Enns, Sean P David, et al, Clinical interpretation and implications of whole-genome sequencing, *Jama*, vol. 311, no. 10, pp. 1035–1045, 2014.

[66] Sang Tae Park and Jayoung Kim, Trends in next-generation sequencing and a new era for whole genome sequencing, *International neurourology journal*, vol. 20, no. Suppl 2, pp. S76, 2016.

[67] David Houle, Diddahally R Govindaraju, and Stig Omholt, Phenomics: the next challenge, *Nature reviews genetics*, vol. 11, no. 12, pp. 855–866, 2010.

[68] Peter B Jensen, Lars J Jensen, and Søren Brunak, Mining electronic health records: towards better research applications and clinical care, *Nature Reviews Genetics*, vol. 13, no. 6, pp. 395–405, 2012.

[69] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al, Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age, *PLoS medicine*, vol. 12, no. 3, pp. e1001779, 2015.

[70] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre, Benefits and limitations of genome-wide association studies, *Nature Reviews Genetics*, vol. 20, no. 8, pp. 467–484, 2019.

[71] Barbara E Stranger, Eli A Stahl, and Towfique Raj, Progress and promise of genome-wide association studies for human complex trait genetics, *Genetics*, vol. 187, no. 2, pp. 367–383, 2011.

[72] Ben Hayes, Overview of statistical methods for genome-wide association studies (gwas), *Genome-wide association studies and genomic prediction*, pp. 149–169, 2013.

[73] Melinda C Mills and Charles Rahal, A scientometric review of genome-wide association studies, *Communications biology*, vol. 2, no. 1, pp. 9, 2019.

[74] Masato Akiyama, Multi-omics study for interpretation of genome-wide association study, *Journal of Human Genetics*, vol. 66, no. 1, pp. 3–10, 2021.

[75] Anna Portela and Manel Esteller, Epigenetic modifications and human disease, *Nature biotechnology*, vol. 28, no. 10, pp. 1057–1068, 2010.

[76] Franck J Barrat, Mary K Crow, and Lionel B Ivashkiv, Interferon target-gene expression and epigenomic signatures in health and disease, *Nature immunology*, vol. 20, no. 12, pp. 1574–1583, 2019.

[77] Zachery R Gregorich and Ying Ge, Top-down proteomics in health and disease: Challenges and opportunities, *Proteomics*, vol. 14, no. 10, pp. 1195–1210, 2014.

[78] Noura S Abul-Husn and Eimear E Kenny, Personalized medicine and the power of electronic health records, *Cell*, vol. 177, no. 1, pp. 58–69, 2019.

[79] Jason D Buenrostro, Beijing Wu, Howard Y Chang, and William J Greenleaf, Atac-seq: a method for assaying chromatin accessibility genome-wide, *Current protocols in molecular biology*, vol. 109, no. 1, pp. 21–29, 2015.

[80] Ruth Pidsley, Elena Zotenko, Timothy J Peters, Mitchell G Lawrence, Gail P Risbridger, Peter Molloy, Susan Van Djik, Beverly Muhlhausler, Clare Stirzaker, and Susan J Clark, Critical evaluation of the illumina methylationepic beadchip microarray for whole-genome dna methylation profiling, *Genome biology*, vol. 17, no. 1, pp. 1–17, 2016.

[81] Maxwell R Mumbach, Adam J Rubin, Ryan A Flynn, Chao Dai, Paul A Khavari, William J Greenleaf, and Howard Y Chang, Hichip: efficient and sensitive analysis of protein-directed genome architecture, *Nature methods*, vol. 13, no. 11, pp. 919–922, 2016.

[82] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika, Multi-omics data integration, interpretation, and its application, *Bioinformatics and biology insights*, vol. 14, pp. 1177932219899051, 2020.

[83] Mingon Kang, Euiseong Ko, and Tesfaye B Mersha, A roadmap for multi-omics data integration using deep learning, *Briefings in Bioinformatics*, vol. 23, no. 1, pp. bbab454, 2022.

[84] Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, et al, The nih roadmap epigenomics mapping consortium, *Nature biotechnology*, vol. 28, no. 10, pp. 1045–1048, 2010.

[85] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al, The human cell atlas, *elife*, vol. 6, pp. e27041, 2017.

[86] Caltech-UW TMC Cai Long lcai@ caltech. edu 21 b Shendure Jay 9 Trapnell Cole 9 Lin Shin shinlin@ uw. edu 2 e Jackson Dana 9, UCSD TMC Zhang Kun kzhang@ bioeng. ucsd. edu 15 b Sun Xin 15 Jain Sanjay 24 Hagood James 25 Pryhuber Gloria 26 Kharchenko Peter 8, California Institute of Technology TTD Cai Long lcai@ caltech. edu 21 b Yuan Guo-Cheng 35 Zhu Qian 35 Dries Ruben 35, Harvard TTD Yin Peng peng_yin@hms. harvard. edu 36 37 b Saka Sinem K. 36 37 Kishi Jocelyn Y. 36 37 Wang Yu 36 37 Goldaracena Isabel 36 37, Purdue TTD Laskin Julia jlaskin@

purdue. edu 10 b Ye DongHye 10 38 Burnum-Johnson Kristin E. 39 Piehowski Paul D. 39 Ansong Charles 39 Zhu Ying 39, Stanford TTD Harbury Pehr harbury@ stanford. edu 11 b Desai Tushar 40 Mulye Jay 11 Chou Peter 11 Nagendran Monica 40, et al. The human body at cellular resolution: the nih human biomolecular atlas program. Nature, 574(7777): 187-192, 2019.

[87] Vivien Marx, Method of the year: spatially resolved transcriptomics, *Nature methods*, vol. 18, no. 1, pp. 9–14, 2021.

[88] Franziska Hildebrandt, Alma Andersson, Sami Saarenpää, Ludvig Larsson, Noémi Van Hul, Sachie Kanatani, Jan Masek, Ewa Ellis, Antonio Barragan, Annelie Mollbrink, et al, Spatial transcriptomics to define transcriptional patterns of zonation and structural components in the mouse liver, *Nature communications*, vol. 12, no. 1, pp. 7046, 2021.

[89] Mengyuan Zhao, Wenying He, Jijun Tang, Quan Zou, and Fei Guo, A comprehensive overview and critical evaluation of gene regulatory network inference technologies, *Briefings in Bioinformatics*, vol. 22, no. 5, pp. bbab009, 2021.

[90] Wanwen Zeng, Shengquan Chen, Xuejian Cui, Xiaoyang Chen, Zijing Gao, and Rui Jiang, Silencerdb: a comprehensive database of silencers, *Nucleic acids research*, vol. 49, no. D1, pp. D221–D228, 2021.

[91] Yong Wang, Rui Jiang, and Wing Hung Wong, Modeling the causal regulatory network by integrating chromatin accessibility and transcriptome data, *National science review*, vol. 3, no. 2, pp. 240–251, 2016.

[92] Zhana Duren, Xi Chen, Rui Jiang, Yong Wang, and Wing Hung Wong, Modeling gene regulation from paired expression and chromatin accessibility data, *Proceedings of the National Academy of Sciences*, vol. 114, no. 25, pp. E4914–E4923, 2017.

[93] Jingxue Xin, Hui Zhang, Yaoxi He, Zhana Duren, Caijuan Bai, Lang Chen, Xin Luo, Dong-Sheng Yan, Chaoyu Zhang, Xiang Zhu, et al, Chromatin accessibility landscape and regulatory network of high-altitude hypoxia adaptation, *Nature communications*, vol. 11, no. 1, pp. 4928, 2020.

[94] Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and TM Murali, Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data, *Nature methods*, vol. 17, no. 2, pp. 147–154, 2020.

[95] Chang Lu, Jan Zaucha, Rihab Gam, Hai Fang, Ben Smithers, Matt E Oates, Miguel Bernabe-Rubio, James Williams, Natalie Zelenka, Arun Prasad Pandurangan, et al, Hypothesis-free phenotype prediction within a genetics-first framework, *Nature Communications*, vol. 14, no. 1, pp. 919, 2023.

[96] Vikas Pejaver, Jorge Urresti, Jose Lugo-Martinez, Kymberleigh A Pagel, Guan Ning Lin, Hyun-Jun Nam, Matthew Mort, David N Cooper, Jonathan Sebat, Lilia M Iakoucheva, et al, Inferring the molecular and phenotypic impact of amino acid variants with mutpred2, *Nature communications*, vol. 11, no. 1, pp. 5918, 2020.

[97] Wanwen Zeng, Xi Chen, Zhana Duren, Yong Wang, Rui Jiang, and Wing Hung Wong, Dc3 is a method for deconvolution and coupled clustering from bulk and single-cell genomics data, *Nature communications*, vol. 10, no. 1, pp. 4613, 2019.

[98] Qiao Liu, Mingxin Gan, and Rui Jiang, A sequence-based method to predict the impact of regulatory variants using random forest, *BMC Systems Biology*, vol. 11, no. 2, pp. 1–9, 2017.

[99] Zhana Duren, Fengge Chang, Fnu Naqing, Jingxue Xin, Qiao Liu, and Wing Hung Wong, Regulatory analysis of single cell multiome gene expression and chromatin accessibility data with screg, *Genome biology*, vol. 23, no. 1, pp. 1–19, 2022.

[100] MA Al-Jawary, MI Adwan, and GH Radhi, Three iterative methods for solving second order nonlinear odes arising in physics, *Journal of King Saud University-Science*, vol. 32, no. 1, pp. 312–323, 2020.

[101] Jamshid Pirgazi and Ali Reza Khanteymoori, A robust gene regulatory network inference method base on kalman filter and linear regression, *PloS one*, vol. 13, no. 7, pp. e0200094, 2018.

[102] Artem Sokolov, Daniel E Carlin, Evan O Paull, Robert Baertsch, and Joshua M Stuart, Pathway-based genomics prediction using generalized elastic net, *PLoS computational biology*, vol. 12, no. 3, pp. e1004790, 2016.

[103] Li Chen, Jianhua Xuan, Rebecca B Riggins, Yue Wang, Eric P Hoffman, and Robert Clarke, Multilevel support vector regression analysis to identify condition-specific regulatory networks, *Bioinformatics*, vol. 26, no. 11, pp. 1416–1422, 2010.

[104] Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Seyhan Yazar, et al, Large-scale cis-and trans-eqtl analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression, *Nature genetics*, vol. 53, no. 9, pp. 1300–1310, 2021.

[105] GTEx Consortium, Kristin G Ardlie, David S Deluca, Ayellet V Segrè, Timothy J Sullivan, Taylor R Young, Ellen T Gelfand, Casandra A Trowbridge, Julian B Maller, Taru Tukiainen, et al, The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans, *Science*, vol. 348, no. 6235, pp. 648–660, 2015.

[106] Tom R Gaunt, Hashem A Shihab, Gibran Hemani, Josine L Min, Geoff Woodward, Oliver Lyttleton, Jie Zheng, Aparna Duggirala, Wendy L McArdle, Karen Ho, et al, Systematic identification of genetic influences on methylation across the human life course, *Genome biology*, vol. 17, no. 1, pp. 1–14, 2016.

[107] Jingning Zhang, Diptavo Dutta, Anna Köttgen, Adrienne Tin, Pascal Schlosser, Morgan E Grams, Benjamin Harvey, CKDGen Consortium, Bing Yu, Eric Boerwinkle, et al, Plasma proteome analyses in individuals of european and african ancestry identify cis-pqtls and models for proteome-wide association studies, *Nature Genetics*, vol. 54, no. 5, pp. 593–602, 2022.

[108] Marc Jan Bonder, René Luijk, Daria V Zhernakova, Matthijs Moed, Patrick Deelen, Martijn Vermaat, Maarten Van Iterson, Freerk Van Dijk, Michiel Van Galen, Jan Bot, et al, Disease variants alter transcription factor levels and methylation of their binding sites, *Nature genetics*, vol. 49, no. 1, pp. 131–138, 2017.

[109] Josine L Min, Gibran Hemani, Eilis Hannon, Koen F Dekkers, Juan Castillo-Fernandez, René Luijk, Elena Carnero-Montoro, Daniel J Lawson, Kimberley Burrows, Matthew Suderman, et al, Genomic and phenotypic insights from an atlas of genetic effects on dna methylation, *Nature genetics*, vol. 53, no. 9, pp. 1311–1321, 2021.

[110] Johann S Hawe, Rory Wilson, Katharina T Schmid, Li Zhou, Lakshmi Narayanan Lakshmanan, Benjamin C Lehne, Brigitte Kühnel, William R Scott, Matthias Wielscher, Yik Weng Yew, et al, Genetic variation influencing dna methylation provides insights into molecular mechanisms regulating genomic function, *Nature genetics*, vol. 54, no. 1, pp. 18–29, 2022.

[111] Tianxiao Huan, Roby Joehanes, CI Song, Fen Peng, Yichen Guo, Michael Mendelson, Chen Yao, Chunyu Liu, Jiantao Ma, Melissa Richard, et al, Genome-wide identification of dna methylation qtls in whole blood highlights pathways for cardiovascular disease, *Nature communications*, vol. 10, no. 1, pp. 4267, 2019.

[112] Jing Gong, Hao Wan, Shufang Mei, Hang Ruan, Zhao Zhang, Chunjie Liu, An-Yuan Guo, Lixia Diao, Xiaoping Miao, and Leng Han, Pancan-meqtl: a database to systematically evaluate the effects of genetic variants on methylation in human cancer, *Nucleic acids research*, vol. 47, no. D1, pp. D1066–D1072, 2019.

[113] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter AC 't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzalez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al, Transcriptome and genome sequencing uncovers functional variation in humans, *Nature*, vol. 501, no. 7468, pp. 506–511, 2013.

[114] Liming Liang, Nilesh Morar, Anna L Dixon, G Mark Lathrop, Goncalo R Abecasis, Miriam F Moffatt, and William OC Cookson, A cross-platform analysis of 14, 177 expression quantitative trait loci derived from lymphoblastoid cell lines, *Genome research*, vol. 23, no. 4, pp. 716–726, 2013.

[115] Andrew E Jaffe, Richard E Straub, Joo Heon Shin, Ran Tao, Yuan Gao, Leonardo Collado-Torres, Tony Kam-Thong, Hualin S Xi, Jie

Quan, Qiang Chen, et al, Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis, *Nature neuroscience*, vol. 21, no. 8, pp. 1117–1125, 2018.

[116] Jing Gong, Shufang Mei, Chunjie Liu, Yu Xiang, Youqiong Ye, Zhao Zhang, Jing Feng, Renyan Liu, Lixia Diao, An-Yuan Guo, et al, Pancanqtl: systematic identification of cis-eqtls and trans-eqtls in 33 cancer types, *Nucleic acids research*, vol. 46, no. D1, pp. D971–D976, 2018.

[117] Benjamin B Sun, Joseph C Maranville, James E Peters, David Stacey, James R Staley, James Blackshaw, Stephen Burgess, Tao Jiang, Ellie Paige, Praveen Surendran, et al, Genomic atlas of the human plasma proteome, *Nature*, vol. 558, no. 7708, pp. 73–79, 2018.

[118] Egil Ferkingstad, Patrick Sulem, Bjarni A Atlason, Gardar Sveinbjornsson, Magnus I Magnusson, Edda L Styrmisdottir, Kristbjorg Gunnarsdottir, Agnar Helgason, Asmundur Oddsson, Bjarni V Halldorsson, et al, Large-scale integration of the plasma proteome with genetics and disease, *Nature genetics*, vol. 53, no. 12, pp. 1712–1721, 2021.

[119] Chen Yao, George Chen, Ci Song, Joshua Keefe, Michael Mendelson, Tianxiao Huan, Benjamin B Sun, Annika Laser, Joseph C Maranville, Hongsheng Wu, et al, Genome-wide mapping of plasma protein qtls identifies putatively causal genes and pathways for cardiovascular disease, *Nature communications*, vol. 9, no. 1, pp. 3268, 2018.

[120] Alexander Gudjonsson, Valborg Gudmundsdottir, Gisli T Axelsson, Elias F Gudmundsson, Brynjolfur G Jonsson, Lenore J Launer, John R Lamb, Lori L Jennings, Thor Aspelund, Valur Emilsson, et al, A genome-wide association study of serum proteins reveals shared loci with common diseases, *Nature communications*, vol. 13, no. 1, pp. 480, 2022.

[121] Fei Liu, Shao-Wu Zhang, Wei-Feng Guo, Ze-Gang Wei, and Luonan Chen, Inference of gene regulatory network based on local bayesian networks, *PLoS computational biology*, vol. 12, no. 8, pp. e1005024, 2016.

[122] Sun Yong Kim, Seiya Imoto, and Satoru Miyano, Inferring gene networks from time series microarray data using dynamic bayesian networks, *Briefings in bioinformatics*, vol. 4, no. 3, pp. 228–235, 2003.

[123] Sunyong Kim, Seiya Imoto, and Satoru Miyano, Dynamic bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data, *Biosystems*, vol. 75, no. 1-3, pp. 57–65, 2004.

[124] Ruiqing Zheng, Min Li, Xiang Chen, Fang-Xiang Wu, Yi Pan, and Jianxin Wang, Bixgboost: a scalable, flexible boosting-based method for reconstructing gene regulatory networks, *Bioinformatics*, vol. 35, no. 11, pp. 1893–1900, 2019.

[125] Hantao Shu, Jingtian Zhou, Qiuyu Lian, Han Li, Dan Zhao, Jianyang Zeng, and Jianzhu Ma, Modeling gene regulatory networks using neural network architectures, *Nature Computational Science*, vol. 1, no. 7, pp. 491–501, 2021.

[126] Jian Zhou and Olga G Troyanskaya, Predicting effects of noncoding variants with deep learning-based sequence model, *Nature methods*, vol. 12, no. 10, pp. 931–934, 2015.

[127] Zheng Wei, Kui Hua, Lei Wei, Shining Ma, Rui Jiang, Xuegong Zhang, Yanda Li, Wing H Wong, and Xiaowo Wang, Neuronmotif: Deciphering cis-regulatory codes by layer-wise demixing of deep neural networks, *Proceedings of the National Academy of Sciences*, vol. 120, no. 15, pp. e2216698120, 2023.

[128] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley, Effective gene expression prediction from sequence by integrating long-range interactions, *Nature methods*, vol. 18, no. 10, pp. 1196–1203, 2021.

[129] Seongho Kim, ppcor: an r package for a fast calculation to semi-partial correlation coefficients, *Communications for statistical applications and methods*, vol. 22, no. 6, pp. 665, 2015.

[130] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Van Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, et al, Scenic: single-cell regulatory network inference and clustering, *Nature methods*, vol. 14, no. 11, pp. 1083–1086, 2017.

[131] Alicia T Specht and Jun Li, Leap: constructing gene co-expression networks for single-cell rna-sequencing data using pseudotime ordering, *Bioinformatics*, vol. 33, no. 5, pp. 764–766, 2017.

[132] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al, Gene ontology: tool for the unification of biology, *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[133] Minoru Kanehisa. The kegg database. In 'In silico' simulation of biological processes: Novartis Foundation Symposium 247, volume 247, pages 91-103. Wiley Online Library, 2002.

[134] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al, String v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, *Nucleic acids research*, vol. 47, no. D1, pp. D607–D613, 2019.

[135] Gang Fang, Wen Wang, Vanja Paunic, Hamed Heydari, Michael Costanzo, Xiaoye Liu, Xiaotong Liu, Benjamin VanderSluis, Benjamin Oately, Michael Steinbach, et al, Discovering genetic interactions bridging pathways in genome-wide association studies, *Nature communications*, vol. 10, no. 1, pp. 4274, 2019.

[136] Matti Hoch, Suchi Smita, Konstantin Cesnulevicius, David Lescheid, Myron Schultz, Olaf Wolkenhauer, and Shailendra Gupta, Network-and enrichment-based inference of phenotypes and targets from large-scale disease maps, *NPJ Systems Biology and Applications*, vol. 8, no. 1, pp. 13, 2022.

[137] Paola Paci, Giulia Fiscon, Federica Conte, Rui-Sheng Wang, Lorenzo Farina, and Joseph Loscalzo, Gene co-expression in the interactome: moving from correlation toward causation via an integrated approach to disease module discovery, *NPJ systems biology and applications*, vol. 7, no. 1, pp. 3, 2021.

[138] Wenran Li, Wing Hung Wong, and Rui Jiang, Deeptact: predicting 3d chromatin contacts via bootstrapping deep learning, *Nucleic acids research*, vol. 47, no. 10, pp. e60–e60, 2019.

[139] Zhidong Tu, Li Wang, Michelle N Arbeitman, Ting Chen, and Fengzhu Sun, An integrative approach for causal gene identification and gene regulatory pathway inference, *Bioinformatics*, vol. 22, no. 14, pp. e489–e496, 2006.

[140] Silpa Suthram, Andreas Beyer, Richard M Karp, Yonina Eldar, and Trey Ideker, eqed: an efficient method for interpreting eqtl associations using protein networks, *Molecular systems biology*, vol. 4, no. 1, pp. 162, 2008.

[141] Yoo-Ah Kim, Stefan Wuchty, and Teresa M Przytycka, Identifying causal genes and dysregulated pathways in complex diseases, *PLoS computational biology*, vol. 7, no. 3, pp. e1001095, 2011.

[142] Shining Ma, Xi Chen, Xiang Zhu, Philip S Tsao, and Wing Hung Wong, Leveraging cell-type-specific regulatory networks to interpret genetic variants in abdominal aortic aneurysm, *Proceedings of the National Academy of Sciences*, vol. 119, no. 1, pp. e2115601119, 2022.

[143] Zhanying Feng, Xianwen Ren, Zhana Duren, and Yong Wang, Human genetic variants associated with covid-19 severity are enriched in immune and epithelium regulatory networks, *Phenomics*, vol. 2, no. 6, pp. 389–403, 2022.

[144] Zhanying Feng, Zhana Duren, Ziyi Xiong, Sijia Wang, Fan Liu, Wing Hung Wong, and Yong Wang, hreg-cncc reconstructs a regulatory network in human cranial neural crest cells and annotates variants in a developmental context, *Communications Biology*, vol. 4, no. 1, pp. 442, 2021.

[145] Zhanying Feng, Zhana Duren, Jingxue Xin, Qiuyue Yuan, Yaoxi He, Bing Su, Wing Hung Wong, and Yong Wang, Heritability

enrichment in context-specific regulatory networks improves phenotype-relevant tissue identification, *Elife*, vol. 11, pp. e82535, 2022.

[146] Xi Xi, Haochen Li, Shengquan Chen, Tingting Lv, Tianxing Ma, Rui Jiang, Ping Zhang, Wing Hung Wong, and Xuegong Zhang, Unfolding the genotype-to-phenotype black box of cardiovascular diseases through cross-scale modeling, *Iscience*, vol. 25, no. 8, pp. 104790, 2022.

[147] Farhad Hormozdiari, Steven Gazal, Bryce Van De Geijn, Hilary K Finucane, Chelsea J-T Ju, Po-Ru Loh, Armin Schoech, Yakir Reshef, Xuanyao Liu, Luke O'connor, et al, Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits, *Nature genetics*, vol. 50, no. 7, pp. 1041–1047, 2018.

[148] Andrew P Feinberg and Andre Levchenko, Epigenetics as a mediator of plasticity in cancer, *Science*, vol. 379, no. 6632, pp. eaaw3835, 2023.

[149] Zhonghua Li, Pengcheng Wang, Chunyuan You, Jiwen Yu, Xiangnan Zhang, Feilin Yan, Zhengxiu Ye, Chao Shen, Baoqi Li, Kai Guo, et al, Combined gwas and eqtl analysis uncovers a genetic regulatory network orchestrating the initiation of secondary cell wall development in cotton, *New Phytologist*, vol. 226, no. 6, pp. 1738–1752, 2020.

[150] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, et al, Partitioning heritability by functional annotation using genome-wide association summary statistics, *Nature genetics*, vol. 47, no. 11, pp. 1228–1235, 2015.

[151] Xiang Zhu, Zhana Duren, and Wing Hung Wong, Modeling regulatory network topology improves genome-wide analyses of complex human traits, *Nature communications*, vol. 12, no. 1, pp. 2851, 2021.

[152] Jing Zhang, Donghoon Lee, Vineet Dhiman, Peng Jiang, Jie Xu, Patrick McGillivray, Hongbo Yang, Jason Liu, William Meyerson, Declan Clarke, et al, An integrative encode resource for cancer genomics, *Nature communications*, vol. 11, no. 1, pp. 3696, 2020.

[153] Mahmoud Ghandi, Franklin W Huang, Judit Jané-Valbuena, Gregory V Kryukov, Christopher C Lo, E Robert McDonald III, Jordi Barretina, Ellen T Gelfand, Craig M Bielski, Haoxin Li, et al, Next-generation characterization of the cancer cell line encyclopedia, *Nature*, vol. 569, no. 7757, pp. 503–508, 2019.

[154] M Angela Nieto, Context-specific roles of emt programmes in cancer cell dissemination, *Nature cell biology*, vol. 19, no. 5, pp. 416–418, 2017.

[155] Clare Pacini, Joshua M Dempster, Isabella Boyle, Emanuel Gonçalves, Hanna Najgebauer, Emre Karakoc, Dieudonne van der Meer, Andrew Barthorpe, Howard Lightfoot, Patricia Jaaks, et al, Integrated cross-study datasets of genetic dependencies in cancer, *Nature communications*, vol. 12, no. 1, pp. 1661, 2021.

[156] Salam A Assi, Maria Rosaria Imperato, Daniel JL Coleman, Anna Pickin, Sandeep Potluri, Anetta Ptasinska, Paulynn Suyin Chin, Helen Blair, Pierre Cauchy, Sally R James, et al, Subtype-specific regulatory network rewiring in acute myeloid leukemia, *Nature genetics*, vol. 51, no. 1, pp. 151–162, 2019.

[157] Monique GP Van Der Wijst, Dylan H de Vries, Harm Brugge, Harm-Jan Westra, and Lude Franke, An integrative approach for building personalized gene regulatory networks for precision medicine, *Genome medicine*, vol. 10, no. 1, pp. 1–15, 2018.

[158] Junha Cha and Insuk Lee, Single-cell network biology for resolving cellular heterogeneity in human diseases, *Experimental & molecular medicine*, vol. 52, no. 11, pp. 1798–1808, 2020.

[159] Victoria Yao, Aaron K Wong, and Olga G Troyanskaya, Enabling precision medicine through integrative network models, *Journal of molecular biology*, vol. 430, no. 18, pp. 2913–2923, 2018.

[160] Le Yang, Runpu Chen, Steve Goodison, and Yijun Sun, An efficient and effective method to identify significantly perturbed subnetworks in cancer, *Nature computational science*, vol. 1, no. 1, pp. 79–88, 2021.

[161] Qiao Liu, Zhongren Chen, and Wing Hung Wong. Causalegm: a general causal inference framework by encoding generative modeling. arXiv preprint arXiv: 2212.05925, 2022.

[162] Ruth Stoney, David L Robertson, Goran Nenadic, and Jean-Marc Schwartz, Mapping biological process relationships and disease perturbations within a pathway network, *NPJ systems biology and applications*, vol. 4, no. 1, pp. 22, 2018.

[163] Vahid H Gazestani, Tiziano Pramparo, Srinivasa Nalabolu, Benjamin P Kellman, Sarah Murray, Linda Lopez, Karen Pierce, Eric Courchesne, and Nathan E Lewis, A perturbed gene network containing pi3k-akt, ras-erk and wnt-β-catenin pathways in leukocytes is linked to asd genetics and symptom severity, *Nature neuroscience*, vol. 22, no. 10, pp. 1624–1634, 2019.