CausalCellInfer: Resolving cell-type-specific disease mechanisms from biobank-scale GWAS

Liangying Yin[1,2#], Yujia Shi[1#], Ruoyu Zhang[1], Yong Xiang[1], Jinghong Qiu[1], Pak-Chung Sham[3], Hon-Cheong So[1,4-9,*]

[1]School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong
[2]Eric and Wendy Schmidt Center, Broad Institute of MIT and Harvard
[3]Department of Psychiatry, The University of Hong Kong, Hong Kong
[4]KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research of Common Diseases, Kunming Institute of Zoology and The Chinese University of Hong Kong, China
[5]Department of Psychiatry, The Chinese University of Hong Kong, Hong Kong
[6]CUHK Shenzhen Research Institute, Shenzhen, China
[7]Margaret K.L. Cheung Research Centre for Management of Parkinsonism, The Chinese University of Hong Kong, Shatin, Hong Kong
[8]Brain and Mind Institute, The Chinese University of Hong Kong, Hong Kong SAR, China
[9]Hong Kong Branch of the Chinese Academy of Sciences Center for Excellence in Animal Evolution and Genetics, The Chinese University of Hong Kong, Hong Kong SAR, China

[#]Both authors contribute equally
*Corresponding author Email: hcso@cuhk.edu.hk

Abstract

Integrating the cellular resolution of single-cell RNA sequencing (scRNA-seq) with the phenotypic depth of population-scale biobanks is essential for elucidating the cellular basis of complex diseases. However, this integration is often hindered by the limited sample sizes of scRNA-seq cohorts and the lack of cell-type resolution in massive biobank datasets.

We present CausalCellInfer, a scalable computational framework designed to bring cellular resolution to bulk and genotype-imputed transcriptomes. CausalCellInfer utilizes an invariant causal prediction-inspired procedure (scI-GCM) to identify environment-stable marker genes, employs a parsimonious deep neural network for robust cell-fraction deconvolution, and leverages regularized matrix completion to reconstruct cell-type-specific (CTS) expression profiles. This architecture is specifically optimized for biobank-scale data, where technical heterogeneity and limited gene overlap are prevalent.

Validated across simulated data, pseudo-bulk mixtures, and real PBMC datasets, CausalCellInfer demonstrated superior accuracy and computational efficiency compared to existing methods. Applied to ~500,000 UK Biobank participants, the framework enabled cell-resolved analyses for 29 traits, identifying known pathological shifts, such as reduced pancreatic β-cell proportions in Type 2 Diabetes, and uncovering novel biological signals, including disrupted excitatory neuron and oligodendrocyte interactions in depression. Furthermore, inferred CTS differential expression patterns showed significant concordance with independent single-cell studies and were enriched for OpenTargets disease genes. Overall, CausalCellInfer bridges the gap between single-cell insights and population-scale genomics, providing a powerful tool for systematic discovery of disease mechanisms at cellular resolution.

Keywords:
Causal principle informed marker (gene) finding; Cell deconvolution; Cell type specific expression inference; Deep learning; Regularized matrix completion

Introduction

Single-cell RNA-sequencing (scRNA-seq) has transformed our understanding of cellular diversity and disease biology by enabling gene expression profiling at single-cell resolution[1–4]. However, routine application of scRNA-seq in large cohorts remains limited by cost, technical complexity, and often sparse clinical data on symptoms, prognosis, and comorbidities. Most single-cell studies are designed for a specific disease or question, and therefore often include limited subject-level clinical information (e.g., symptoms, prognosis, comorbidities, medications) and small numbers of individuals. These constraints make it difficult to connect cellular mechanisms to diverse clinical outcomes, to systematically assess effect modifiers, or to distinguish causes from consequences of disease.

In contrast, bulk RNA-seq and genotype-imputed transcriptome data are available at population scale. Resources such as TCGA and GTEx offer bulk expression across tissues, and large biobanks like the UK Biobank (UKBB) provide extensive genotype data together with deep clinical and biological phenotypes. Genotype data further enable the prediction of individual-level gene expression, creating transcriptome-like measurements at biobank scale[5–7]. However, both measured bulk expression and genetically predicted expression lack cellular resolution. Bulk expression profiles aggregate signals across heterogeneous cell types, obscuring the cell-type-specific (CTS) drivers of complex traits. Bridging the mechanistic resolution of scRNA-seq with the statistical power and phenotypic breadth of biobanks would enable cell-type–resolved discovery in large cohorts and across a variety of clinical outcomes.

We address this gap with **CausalCellInfer**, a framework inspired by the concept of invariant causal prediction (ICP), as proposed by Peters et al.[8]. CausalCellInfer uses scRNA-seq reference data to identify robust, cell-defining features and then transfers this information to bulk transcriptomes, including genotype-imputed expression in large cohorts, to infer *(i) cell-type proportions and (ii) CTS expression*.

The design goal is biobank readiness: the proposed framework is scalable to very large sample sizes and reliable under distribution shifts. By anchoring inference at the cell-type level, CausalCellInfer enables biobank-scale analyses of how cellular composition and CTS expression relate to risk factors, disease onset / progression, prognosis, and treatment response. Although the framework can be used with any bulk transcriptome data, one of its most powerful applications arise in biobanks where broad phenotyping, large sample size, and individual-level covariates enable rigorous, cell-type-resolved association analyses.

Using inferred cell-type proportions and CTS expression, we support an application suite that is difficult to implement with conventional single-cell study designs because of limited subject-level phenotyping and small sample sizes:

- **Cell-type specific TWAS (CTS-TWAS)**: We test whether genetically predicted CTS expression is associated with complex traits/diseases , localizing gene–trait associations to the most relevant cell types.
- **CTS-TWAS with causal gene prioritization (CTS-TWAS-Causal):** By integrating causal feature selection, we prioritize genes whose CTS expression is most consistent

with a direct effect on traits, helping to distinguish core drivers from correlated signals.

- **Cell-type proportion association studies:** We assess whether genetically inferred cell-type abundances relate to disease risk or clinical phenotypes. This highlights cellular contributors to pathophysiology (e.g., expansion or depletion of specific immune or stromal populations) and nominates candidate cell types for therapeutic targeting.
- **Cell-type proportion interaction analyses**: We evaluate interactions within and across tissues (e.g., immune–parenchymal crosstalk) to identify synergistic or antagonistic relationships that affect disease expression, prognosis, or treatment response.
- **Conditional TWAS:** We adjust bulk or imputed expression analyses for inferred cell-type proportions to reduce confounding by cellular composition when identifying differentially expressed genes. This improves interpretability by separating compositional shifts from CTS expression changes.

These analyses operate on individual-level genotype and imputed expression, and also scale readily to thousands of participants and thousands of phenotypes, typical in biobanks. This scale yields several practical advantages. First, comprehensive covariates available in biobanks (e.g., age, sex, BMI, smoking, medications, socioeconomic factors) can be included to reduce confounding and to test effect modification. Second, large samples naturally support interaction models, stratified analyses, and prediction frameworks that are underpowered in typical single-cell cohorts. Third, when expression and cell-type measures are genetically predicted (from germline variations), associations are less susceptible to reverse causation and time-varying confounders (e.g., disease severity, treatment), strengthening causal interpretation relative to observational single-cell measurements. While not a definitive proof of causality, this approach helps infer whether cellular features are more likely antecedents rather than consequences of a disease.

Applying CausalCellInfer to UKBB dataset, we connect cell-type proportions and CTS expression to a wide spectrum of clinical phenotypes, from cardiometabolic traits and immune disorders to neurological diseases. This biobank-scale setting enables us to: (i) map cell types most strongly linked to specific outcomes; (ii) nominate CTS genes whose dysregulation associates with disease liability; and (iii) reveal cellular interactions that may underlie comorbidities or variable treatment response. We further assess biological plausibility by demonstrating significant overlaps between CTS differentially expressed genes inferred from UKBB and those reported in independent scRNA-seq studies, providing external validation across diseases and tissues.

Here we highlight key differences and strengths of our proposed approach relative to several existing deconvolution methods. Signature-based deconvolution methods (e.g. CIBERSORTx, DWLS) rely on reference gene expression profiles for marker genes derived from scRNA-seq[9–14]. However, applying signature-based approaches to genotype-imputed transcriptomes is often challenging because reliable imputed expression is available for only a *subset* of genes (typically on the order of a few thousand, depending on tissue and prediction models), which can remove or weaken key markers and reduce identifiability of cell types. Also, while signature-based methods can perform well when the reference and target data are well matched, in practice their accuracy can degrade under technical and biological mismatches

(e.g., batch effects, platform differences, disease states, subject heterogeneity etc.) because the marker set and reference profiles may not transfer very well across settings.

Signature-free, deep learning–based methods (e.g., Scaden, TAPE) improve robustness by learning directly from single-cell-derived training data[15–17], but they typically require high-dimensional features as input and substantial computational resources. Also, their performance may be sensitive to how well the training distribution matches the target cohort, creating risks of overfitting or reduced generalization under distribution shifts[18–21].

Our framework is designed for the biobank use case: it focuses on identifying a parsimonious set of cell-defining features that are stable across environments (motivated by invariance ideas[8,22]), and it explicitly targets transfer to bulk and genotype-imputed expression where feature availability may be constrained. By prioritizing transferable, compact feature sets and emphasizing out-of-cohort reliability, our framework is intended to remain accurate across cohorts with different characteristics (e.g. demographics, technologies, disease burdens) while scaling to very large samples and thousands of phenotypes. These design choices are driven not by methodological novelty per se, but by the practical need to execute cell-type–resolved analyses across very large samples and many phenotypes.

In summary, the major objective of CausalCellInfer is to bring cellular resolution to biobank-scale discovery. By inferring cell-type proportions and CTS expression from bulk and genotype-imputed transcriptomes, we enable:

- Systematic **CTS-TWAS** to localize gene–trait associations to the most relevant cell types;
- Prioritization of putatively **causal CTS genes** for mechanistic studies and therapeutic target nomination;
- Association and interaction analyses of **cell-type abundances** within and across tissues; and
- Confounding-aware differential expression through **conditional TWAS**.

Together, these capabilities allow investigators to interrogate disease mechanisms at the cellular level while retaining the statistical power, phenotypic breadth, and covariate control afforded by large cohorts.

Methods

Overview of CausalCellInfer

In this study, we introduce an integrated framework, CausalCellInfer, that combines invariant causal prediction and DNNs with scRNA data to estimate cell proportions and expression profiles from bulk or genotype-imputed expression data across large cohorts.

Given (1) a reference scRNA-seq dataset with cell-type labels and (optionally) 'environment' annotations and (2) a target cohort with bulk or genetically-predicted expression, CausalCellInfer produces two outputs for each target individual: **estimated cell-type proportions** and **cell-type-specific (CTS) expression profiles.**

The proposed workflow consists of three main steps (Fig. 1):

1. **Invariant cell-marker identification** from the scRNA-seq reference using an invariant causal prediction (ICP)-inspired feature selection procedure (scI-GCM).
2. **Cell-type deconvolution** in the target cohort using a deep neural network (DNN) trained on pseudo-bulk mixtures generated from the scRNA-seq reference, restricted to the selected marker genes.

3. **CTS expression reconstruction** using a regularized matrix completion approach (ENIGMA) informed by the estimated cell proportions and scRNA-seq reference profiles.

Subsequent sections will provide a detailed description of each step.

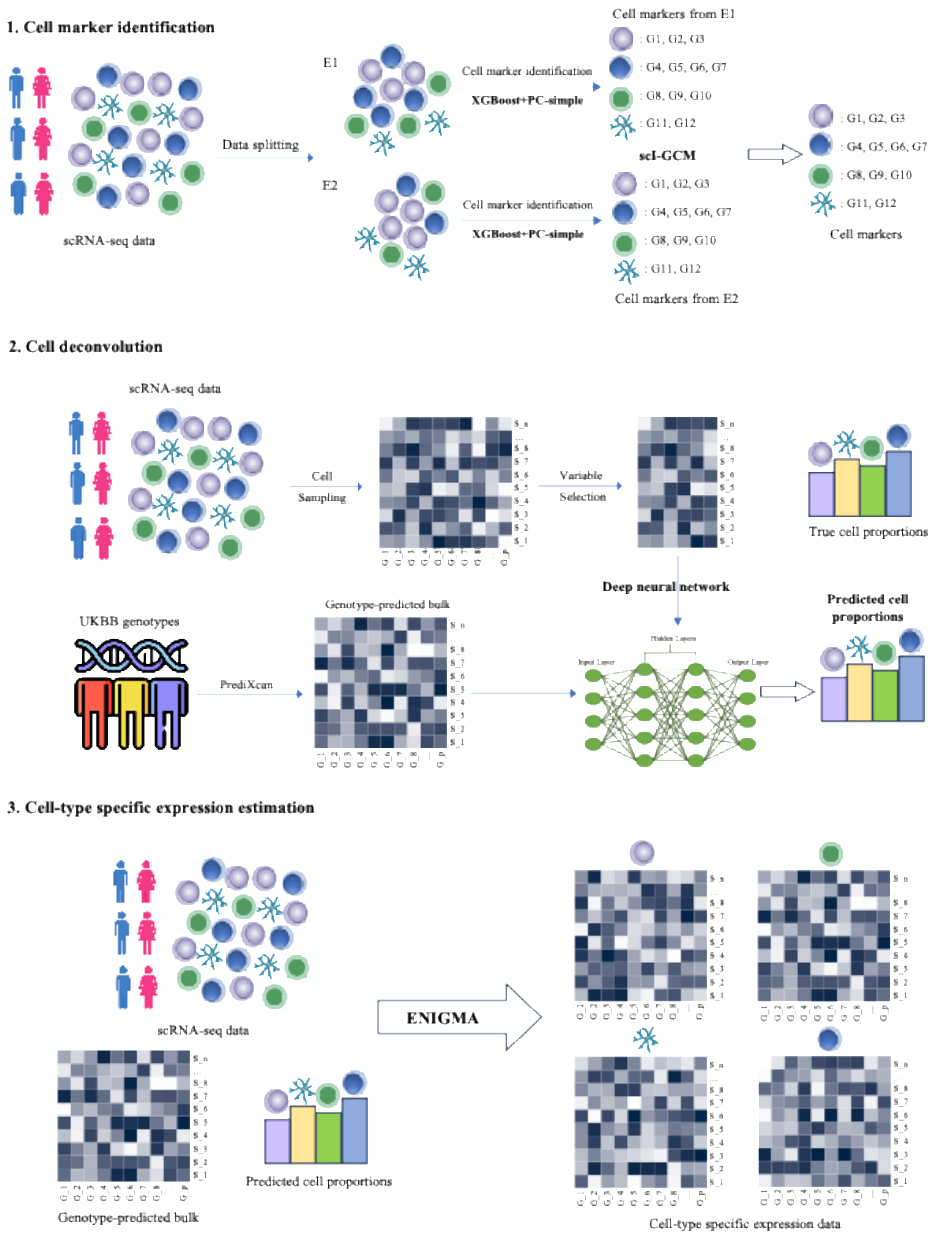

Figure 1 Workflow of our proposed framework for cell decomposition and cell type-specific expression estimation. 1. **Cell marker identification**: We proposed invariant generalized covariance measure (scI-GCM) to identify cell markers for each cell type from the reference scRNA-seq data; 2. **Cell deconvolution**: We modified Scaden, a DNN-based deconvolutional method, to estimate cell proportions from bulk expression data using a small yet informative gene set; 3. **Cell type-specific expression estimation**: We employed a regularized matrix completion method (ENIGMA) to reconstruct cell type-specific expression data by leveraging information from the reference RNA-seq data and the estimated cell proportions from step 2.

Step 1 - Cell marker identification

Genes used as "markers" in scRNA-seq are usually cell-type enriched, but they can still vary substantially with disease states, age, sex, technical batch, or other sample-specific factors. For deconvolution and transfer to external cohorts, we seek a *parsimonious* set of cell-defining genes whose relationship to cell identity is stable across environments.

Robust cell markers tend to exhibit stability across diverse environments. In this context, we define "environments" (E) as subject-related or experimental background features that may shift expression distributions. These environments may include variations in disease pathology, demographics, spatiotemporal factors, and batch effects, alongside experimental perturbations like gene knockdown or overexpression. However, the "environment" should not include the outcome variable and its descendants[8]. According to the theory of Invariant Causal Prediction (ICP), the conditional distribution of a target variable Y is invariant across environments when conditioned on its direct causes.

Operationally, we seek to identify a gene set X^S such that the association between cell-type Y and E disappears after conditioning on X^S, satisfying the invariance criterion:

$$E \perp Y | X^S$$

Under standard ICP assumptions, this independence ensures that X^S captures the fundamental, stable signals defining cell identity.

To identify such markers within large-scale scRNA-seq data, we introduce an *adapted* and improved Invariant Generalized Covariance Measure (I-GCM) framework for **s**ingle-**c**ell analysis, which we also refer to as "scI-GCM" below. The presented approach refines our previously proposed I-GCM framework[23], originally developed to identify clinical and genetic risk factors in epidemiological cohorts[23]. Here we optimized the I-GCM pipeline to tackle the high-dimensional challenges inherent in single-cell transcriptomic data. The resulting framework is grounded in causal graphical models and invariant prediction principles, operating under standard causal assumptions such as partial faithfulness, unconfoundedness, and stable causal mechanisms. Crucially, the method incorporates a scalable, hierarchical feature selection process to ensure computational efficiency.

Below we summarize the key steps and then describe adaptations specific to cell marker selection:

1. **Candidate marker identification via two-stage hierarchical screening** To address the computational intensity of standard causal discovery, we implemented a two-stage hierarchical screening process:

   (i) ***Initial Screening (XGBoost)***: The scRNA-seq data was first partitioned into training and testing sets. The training set was then subdivided into subsets based on selected environmental variables, specifically disease status and gender for this study. Within each environment stratum, we trained an XGBoost model (max tree depth 2, learning rate 1) to predict the binary cell type (Y) from gene expression data (X). This step performs a "coarse" reduction, retaining genes with positive importance scores.

   (ii) ***Causal Refinement (PC-simple)***: To eliminate spurious associations captured by XGBoost, we refine the retained genes using the PC-simple algorithm within each environment stratum. We set the maximum conditioning order to 3 and significant level

(α) to 0.05. The union of genes across all environments formed the initial candidate set to be analyzed in the next stage.

Regarding the principles of PC-simple, it is a causal inference algorithm that generalizes ordered correlation screening to select direct causal variables. A candidate gene X^j is retained if the null hypothesis of conditional independence between Y and X^j given a conditioning set X^S is rejected using Fisher's Z-transformed partial correlation:

$$(n - |s| - 3)^{1/2} \left| Z(Y, X^j | X^S) \right| > \phi^{-1}(1 - \alpha/2) \quad (1)$$

where $Z\left(Y, X^j \middle| X^s\right) = \frac{1}{2}\log\{\frac{1+\hat{\rho}\left(Y, X^j \middle| X^s\right)}{1-\hat{\rho}\left(Y, X^j \middle| X^s\right)}\}$ denotes Fisher's Z-transform of the partial correlation, n denotes the sample size, $|s|$ is the conditioning set size, ϕ^{-1}indicates the is the standard normal quantile function.

2. **Invariant marker identification via generalized covariance measure (GCM)**

The final stage identifies markers that define cell identity amidst environmental noise. We employed the generalized covariance measure (GCM) (T^n) [24] to test whether the conditional association between the environment variable(E) and cell type(Y) remains stable when conditioning on a specific gene set ($E \perp Y | X^S$). The GCM statistic(T^n) tests independence based on residuals(R):

$$T^n = \frac{\sqrt{n}.\frac{1}{n}\sum_{i=1}^{n} R_i}{(\frac{1}{n}\sum_{i=1}^{n} R_i^2 - (\frac{1}{n}\sum_{r=1}^{n} R_r)^2)^{1/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n} R_i}{[(\frac{1}{n}\sum_{i=1}^{n} R_i^2 - (\frac{1}{n}\sum_{r=1}^{n} R_r)^2)^{1/2}]/\sqrt{n}} \quad (2)$$

Here $R_i = (e_i^j - f(x_i^S))(y_i^j - f(x_i^S))$ represents the product between residuals from the prediction functions of E and Y using X^S for each observation. We utilize XGBoost as the prediction model to calculate this conditional independence; genes are iteratively removed if their inclusion does not contribute to improving They would have gone the invariance. Other machine learning methods can also be used as substitute for XGBoost, offering high flexibility. Significant deviation from the null hypothesis(indicating a change in GCM distance between consecutive conditional gene sets) is rejected if:

$$\frac{\Delta T_{S_{j-1},S_j}^n}{sqrt\left(\mathrm{Var}\left(\Delta T_{S_{j-1},S_j}^n\right)\right)} > \phi^{-1}(1 - \alpha) \quad (3)$$

$\Delta T_{S_{j-1},S_j}^n$ defines the change in the distance of GCM from zero between two consecutive conditional sets, with $\alpha = 0.05$. For more technical details, please refer to supplementary text.

Step 2: Cell-type deconvolution with ensemble DNN

Training data generation (pseudo-bulk mixtures)

In our framework, we employ a deep neural network (DNN) model to estimate cell proportions. It uses simulated bulk gene counts for the selected marker genes derived from reference scRNA-seq data to estimate cell proportions in bulk expression samples.

We train a DNN to map bulk expression of the selected marker genes to cell-type fractions. Because the target cohort may be bulk RNA-seq or genotype-imputed expression, we train using pseudo-bulk mixtures generated from the scRNA-seq reference, following established pseudo-bulk simulation procedures.

To train the DNN models, we simulated 5,000 pseudo-bulk samples with random fractions of different cell types ($f_c \epsilon [0,1] \, for \, all \, c \, \epsilon \{1,2,..C\}, where \, \sum_1^C f_c = 1$) by subsampling cells from the reference RNA-seq data, following the pseudo-bulk data generation approach by Chen et al.[16]. Only genes selected by our proposed scI-GCM approach were included as model features. Each simulated bulk sample was constructed from 500 randomly sampled cells. The number of cells for each cell type in each bulk sample was $r_c = f_c \cdot 500$, where r_c refers to the cell number of cell type c and f_c refers to the fraction of cell type c. We generated the final pseudo-bulk expression profile by summing up the single-cell expression levels for each sample.

Model architecture and training

We used the MinMaxScaler() class from Sklearn to scale the input data to the range [0,1]. We trained an ensemble of 3 DNN models with distinct architectures and degrees of dropout regularization, as the Ensemble DNN model has been shown to outperform individual models[15]. Each model outputs a C-dimensional fraction vector via a Softmax output layer.

We optimize the mean absolute error (L1 loss):

$$L1(y_i, \widehat{y_i}) = \frac{1}{n} |y_i - \widehat{y_i}| \tag{4}$$

where y_i is the vector of ground-truth cell fractions of the pseudo-bulk sample i , $\widehat{y_i}$ is the vector of predicted fractions of the sample i, and n is the sample size. The final estimated proportions are the ensemble average:

$$\widehat{y_c} = \frac{\widehat{y_c^1} + \widehat{y_c^2} + \widehat{y_c^3}}{3} \tag{5}$$

Models were implemented in PyTorch. Trained models are then applied to target bulk (or calibrated imputed) expression to estimate cell-type proportions per individual.

Step 3: CTS expression estimation

We employed a regularized matrix completion-based approach called ENIGMA[25] to estimate cell-type specific expression profiles based on predicted cell proportions and genotype-imputed bulk expression profiles. ENIGMA is developed based on two assumptions. The first assumption is that the expectation of each cell type-specific (CTS) expression profile aligns with the reference profile matrix from the external scRNA-seq data[26]. The second assumption is that observation noise is consistent across different samples. We estimate the CTS expression by the trace norm model, following[25]:

$$(\widehat{X_i})_{i=1,2,..C} = argmin(\frac{1}{2}\alpha \| O - \sum_{i=1}^{C} X_i diag(\theta_i) \|_F^2 + \frac{1}{2}(1-\alpha) \sum_{i=1}^{C} \|X_i a - R_i\|_2^2 + \beta \sum_{i=1}^{C} \widehat{\theta_i} \|X_i\|_*) \tag{6}$$

where

i denotes the cell type index,
$O \in R^{m \times n}$ denotes the bulk count expression matrix for m genes and n subjects,
X_i denotes the true CTS expression for the ith cell type,
θ_i denotes the cell proportion of the ith cell type,
R_i denotes the reference expression vector for the ith cell type,
$\alpha \epsilon (0,1)$ denotes the weights of the aggregated CTS expression profiles (note that $1-\alpha$ represents the 'agreement weight' between the aggregated CTS expression profiles and its corresponding reference expression vector),
$\beta \in (0, +\infty)$ denotes the regularization strength,
$a = (\frac{1}{n}, \dots, \frac{1}{n})^T \epsilon R^n$ with n indicating bulk sample size.

To reduce the computational complexity of the optimization problem, we employed the maximum L_2 norm model to deconvolute bulk samples [27]:

$$(\widehat{X_i})_{i=1,2,..C} = (\frac{1}{2}\alpha \| O - \sum_{i=1}^{C} X_i diag(\theta_i) \|_F^2 + \frac{1}{2}(1-\alpha) \sum_{i=1}^{C} \|X_i a - R_i\|_2^2 + \beta \sum_{i=1}^{C} \hat{\theta}_i \|t(X_i)\|_{2,max}^2) \quad (7)$$

where $\|.\|_{2,max}^2$ indicates the square of the maximum L_2 row norm of a matrix. The optimization procedure was described in[27].

Calibrating PrediXcan-imputed expression

PrediXcan-imputed expression values are typically standardized per gene across individuals (often reported as Z-like scores). This per-gene standardization preserves between-individual variation for a given gene, but obscures cross-gene magnitudes *within* an individual. In other words, PrediXcan outputs are suitable for comparing a single gene across individuals (e.g., a value of 1 for a gene in person B indicates higher expression than a value of 0 for the same gene in person A), but they do not reflect actual expression differences between genes within the same individual (e.g., a value of 0 for gene A and -0.3 for gene B in the same person cannot be directly compared). Consequently, the scale of the PrediXcan data is incompatible with scRNA-seq reference data, preventing direct inference of cell composition and cell-type-specific expression.

To mitigate this limitation, we mapped the PrediXcan-imputed values back into the expression space of the scRNA-seq reference, using a heuristic pseudo-bulk calibration. Specifically, we generated 1,000 pseudo-bulk RNA-seq samples from the control-group scRNA-seq data using a procedure similar to that used in the TAPE approach[16]:

1. We took the reference cell-type proportions as a baseline. For each pseudo-bulk sample, we perturbed these proportions by adding Gaussian noise to each cell type, clipped negatives to zero, and renormalized to sum to one.
2. We then sampled 1,000 cells from the scRNA-seq reference with probabilities given by the perturbed proportions and summed their profiles to form one pseudo-bulk sample.

From these 1,000 pseudo-bulk samples, we computed, for each gene g, the mean μ_g and standard deviation σ_g . For each subject s, let $Z_{g,s}$ denote the PrediXcan-imputed standardized value for gene g. We define the calibrated pseudo-bulk-like value:

$$\tilde{O}_{g,s} = Z_{g,s}\,\sigma_g + \mu_g.$$

We use $\tilde{O}$ as the bulk input O for both DNN deconvolution and ENIGMA reconstruction.

After obtaining the re-scaled bulk RNA-seq data (O), we can derive the reference profile matrix R by averaging the expression profiles of cells within each cell type in the scRNA-seq reference. Subsequently, the inferred cell type proportion matrix(θ), reference profile matrix (R), and the re-scaled bulk RNA-seq data (O) were plugged into the above formula to infer the final CTS expression profile. The quality of the inferred CTS expression profiles can be significantly influenced by the reference CTS expression profiles. To ensure the reliability of these estimated profiles for subsequent analyses, we recommend using those inferred from the cell types in the reference scRNA-seq dataset with a total cell count exceeding 2,000. Additionally, we aimed to balance the reconstructed CTS so that it resembles the reference scRNA-seq dataset, while ensuring the reconstructed bulk RNA-seq data approximates the observed bulk data. Therefore, we set α to 0.5 in our analysis and kept β at the default setting.

Evaluation of the proposed framework

Simulations to evaluate scI-GCM for single-cell analysis

To validate the advantages of the scI-GCM method, we first established a baseline simulation using simulated scRNA-seq datasets confounded by environmental variables. We compared scI-GCM's performance against analogous algorithms (PC-simple) and state-of-the-art deconvolution methods (DWLS, Scaden). Notably, the analogous PC-simple method refers to our approach but only replacing scI-GCM in step 1 with PC-simple. Our objective was to test whether scI-GCM more accurately identifies cell-type-specific gene markers in the presence of confounding factors, thereby leading to more robust predictions of cell type proportions.

We utilized scDesign3[27] to create scRNA-seq data based on a publicly available peripheral blood mononuclear cells (PBMC) dataset. scDesign3 is a statistical simulator that generates realistic single-cell data using a unified probabilistic model, the parameters of which are derived from actual data. This model allows us to manipulate interpretable parameters, introducing environmental effects that lead to confounding, and subsequently simulating new scRNA-seq and bulk data impacted by these effects. We conducted simulations on bulk samples comprising four cell types with varying confounding effects. For more details, please refer to the supplementary text.

Evaluation of the proposed approach on real and pseudo-bulk data

To verify the reliability of our proposed framework, we compared the performance of our method with several existing deconvolution approaches across various evaluation metrics, including root mean square error (RMSE), mean absolute error (MAE), and Lin's Concordance Correlation Coefficient (CCC).

We used pseudo-bulk expression profiles generated from reference scRNA-seq data of four different tissues: the frontal cortex, pancreas islet, and subcutaneous and visceral adipose tissues. These pseudo-bulk expression profiles were simulated in silico from scRNA-seq data with randomly generated cell type proportions, essentially a summation of single-cell expression profiles[16]. We added noise to each generated pseudo-bulk expression profile to

mimic the measurement errors in real RNA-seq data[16]. The noise was randomly generated from a Gaussian distribution based on the cells used to create the corresponding pseudo-bulk profile. We only included genes that could be reliably predicted from genotypes by PrediXcan for model training.

Additionally, we conducted benchmarking on real datasets, using available bulk PBMC RNA-seq data and their corresponding cell proportions assessed by flow cytometry from Monaco et al.[28]. To generate pseudo-bulk datasets for data training and to perform scI-GCM to identify cell markers, we merged five PBMC scRNA-seq datasets obtained from different individuals, and these datasets were all downloaded from the 10x Genomics database.

To mitigate overfitting, we partitioned each reference scRNA-seq dataset into two subsets based on subject ID: a training set and a testing set. We initially trained our deep neural network model using the training set and subsequently tested it with pseudo-bulk samples generated from the testing set. The training set was also used to obtain references for MuSiC[29], DWLS[30], and Scaden[15].

Validating estimated CTS expression profiles via differential expression analysis

To further validate the robustness of our proposed framework, we performed a comparative analysis on its ability to detect differentially expressed genes across different cell types, especially when cell-type specific (CTS) expression profiles were imputed from genotypes.

Specifically, we identified differentially expressed (DE) genes from both actual scRNA-seq data and genotype-predicted CTS expression profiles. For raw scRNA-seq data, we utilized the FindMarkers() function from the Seurat[31] package to detect DE genes using the default settings. This function by default employs the Wilcoxon Rank Sum test to identify genes differentially expressed between two cell groups. For our genotype-predicted CTS expression profiles, we applied the same test to detect DE genes. We used positive predictive value (PPV) for the evaluation, considering the DE genes (p-value<=0.05) identified from the actual scRNA-seq dataset as the reference set.

UK Biobank analyses

Cell-type proportion association and interaction analyses with clinical phenotypes

We tested associations between inferred cell-type proportions and disease status/phenotypes using regression models, adjusting for relevant covariates and controlling false discovery rate (FDR) at 0.05 via the Benjamini–Hochberg procedure. Besides, we examined whether interactions between cell types, including cell types *across* different tissues, could affect disease phenotypes using regression models.

Identifying cell type-specific genes involved in disease susceptibility (CTS-TWAS) and causal gene prioritization

Upon acquiring the CTS expression profiles for each target disease/trait from UKBB, we identified genes contributing to disease susceptibility for each cell type. This may be considered as a cell type-specific TWAS (CTS-TWAS). To mitigate potential confounding effects from population stratification, we adjusted the estimated cell type-specific expression

profiles and phenotypes using the top 10 principal components (PCs) derived from the corresponding genotype dataset.

Subsequently, we utilized the PC-simple algorithm to identify potential direct causal genes for the outcome variable. For binary and count data, we applied PC-simple using linear regression-based partial correlations as an approximation, treating the outcome as a continuous variable. This follows common practice in high-dimensional causal discovery[5,32]. Following ref[32], we considered an alpha threshold of 0.05 and an order of 3 for the correlation tests. For a more comprehensive description of the algorithm, please refer to ref[5,32] and the supplementary text.

Uncovering cell type-specific genes involved in disease prognosis

UKBB is a comprehensive database that provides extensive phenotypic and prognostic information on various diseases. With the CTS profiles, we could probe the involvement of cell-specific genes in disease *prognoses*, such as mortality and hospitalization frequency. Our application to UKBB mainly focused on 4 diseases/outcomes with imputed expression in relevant tissues (i.e., depression in frontal cortex, T2DM in pancreas, and obesity in subcutaneous and visceral adipose).

We concentrated on key prognostic variables of clinical significance for these diseases. Specifically, we focused on treatment resistance and psychiatric hospitalization frequency for depression (tissue studied: frontal cortex); mortality and related hospitalization frequency for T2DM (tissue studied: pancreas); coronary artery disease (CAD) mortality and related hospitalization frequency for obesity (tissue studied: subcutaneous and visceral adipose). Please refer to the supplementary text for more details.

Moreover, we conducted analyses for 12 other related disorders/phenotypes, including psychosis, bipolar disorder, anxiety (tissue studied: frontal cortex); type 1 diabetes mellitus (T1DM), prediabetes, T2DM-related hospitalization, T2DM requiring insulin (tissue studied: pancreas); CAD, T2DM, heart failure, hypertension, stroke (tissue studied: subcutaneous and visceral adipose). We adjusted the CTS profiles and target prognosis variables by the top 10 PCs to mitigate population stratification. We then employed PC-simple to identify potentially direct causal genes contributing to disease outcomes.

Enrichment analysis

We further validated our methodology by assessing whether the identified gene set was significantly enriched among disease genes listed in an independent, established database. We leveraged the OpenTargets database[33], which provides an overall association score for each gene, aggregating seven types of evidence: genetic associations, somatic mutations, known drugs, affected pathways, RNA expression, literature support, and animal models. Enrichment was performed at the disease level by aggregating CTS genes across cell types.

For further functional interpretation of prioritized (putatively causal) genes, we performed pathway over-representation analysis using ConsensusPathDB[34] with hypergeometric testing.

Conditional TWAS controlling for cell-type proportions

We first performed a standard TWAS analysis on the tissue-specific expression data for each target trait. Then, we examined whether controlling for the estimated cell proportions would significantly affect the effect size of trait-associated genes. Cell proportion is a potential confounder that may lead to spurious associations between imputed expression and outcomes.

For each gene, we fit two univariate regression models: one adjusting for the top 10 genotype PCs and one adjusting for PCs plus estimated cell fractions. We compared coefficients using the method of Clogg et al.[35].

Results

Overview of the proposed framework

We proposed CausalCellInfer, a framework which integrates scI-GCM with a deep neural network to deconvolute cell proportions and employs a regularized matrix completion algorithm (ENIGMA) to estimate the corresponding CTS profiles. CausalCellInfer can reliably and efficiently deconvolute cell compositions and estimate the corresponding expression profiles from large-scale bulk expression data, including GWAS-predicted expression.

CausalCellInfer mainly employs scI-GCM to identify markers (genes) for each cell type. scI-GCM is based on the principle of invariant causal prediction, designed to identify the "causal" genes for each cell-type. It then uses these identified markers to train the deep neural network-based cell deconvolution model. Subsequently, it employs ENIGMA to infer the corresponding CTS expression profiles based on the estimated cell proportions.

We verified the efficacy of our proposed method by comparing it against existing state-of-the-art cell deconvolution methods, including NNLS, DWLS, MuSiC, Scaden, and DISSECT across various real and pseudo-bulk samples. Finally, we applied our proposed framework to deconvolute four GWAS-predicted tissue-specific expression profiles in the UKBB, i.e., frontal cortex, pancreas, subcutaneous adipose tissue (SAT) and visceral adipose tissue (VAT), using disease status as environment variable. With the deconvoluted cell compositions and corresponding CTS expression profiles, we conducted a wide range of analyses for 29 clinically relevant phenotypes (4 'main' disorders, 8 prognostic traits pertaining to the 4 main disorders, 17 other related disorders) in the UKBB, including cell type association analysis, cell type-specific TWAS and causal gene detection, and (tissue-level) TWAS controlled for cell proportions.

Performance of scI-GCM for marker selection and cell-type deconvolution under confounding

To evaluate marker selection under environment-induced confounding, we compared scI-GCM against another causal-screening approach (PC-simple) and three representative state-of-the-art methods (DWLS, MuSiC, and Scaden) using simulated scRNA-seq and bulk transcriptomic data. We considered two scenarios with different confounding effect sizes (modest and substantial) to mimic realistic settings in which factors such as disease status impose strong confounding effects on cell-specific expression profiles.

Despite using the smallest marker gene set, scI-GCM outperformed all competitors in terms of concordance correlation coefficient (CCC), root mean square error (RMSE), and mean absolute error (MAE) in simulations with four cell types (Figure 2). Notably, as the magnitude of confounding increased, the performance advantage of scI-GCM became more pronounced.

In contrast, DWLS struggled to deconvolute samples simulated from diseased conditions (Figure 2, the median of CCC, MAE and RMSE is comparable to other methods), although it achieved good performance on control samples. This discrepancy likely arises because DWLS constructs its deconvolution signature matrix from control samples, which limits its generalizability to samples affected by confounding factors such as disease.

We also assessed cell marker quality by different algorithms. scI-GCM also demonstrated superior performance in identifying highly ranked differentially expressed genes (DEGs) relative to PC-simple. Genes were ranked by the mean inter-cell-type difference in decreasing order; the top 100 were considered DEGs. Marker gene sets identified by scI-GCM achieved higher PPV for DEGs than those derived from PC-simple (Table 1). Moreover, as confounding effects intensified, scI-GCM more reliably identified trustworthy DEGs for downstream deconvolution.

Table 1 Comparison of scI-GCM and PC-simple in identifying DEG genes

Scenario	**Positive predictive value (PPV)**	
	scI-GCM	PC-simple
4 cell types with modest confounding effects	0.476	0.207
4 cell types with substantial confounding effects	0.627	0.208

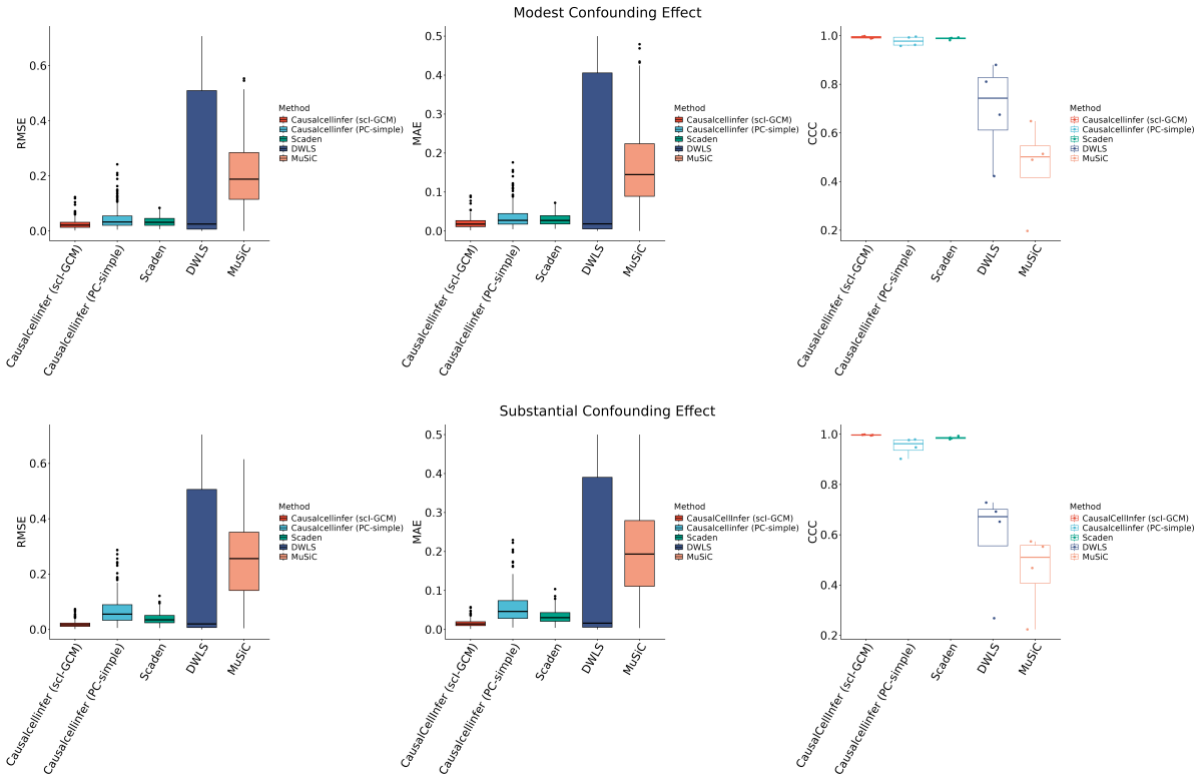

Figure 2 Benchmarking of cell deconvolution accuracy on simulated bulk samples comprising 4 cell types with varied confounding effects.

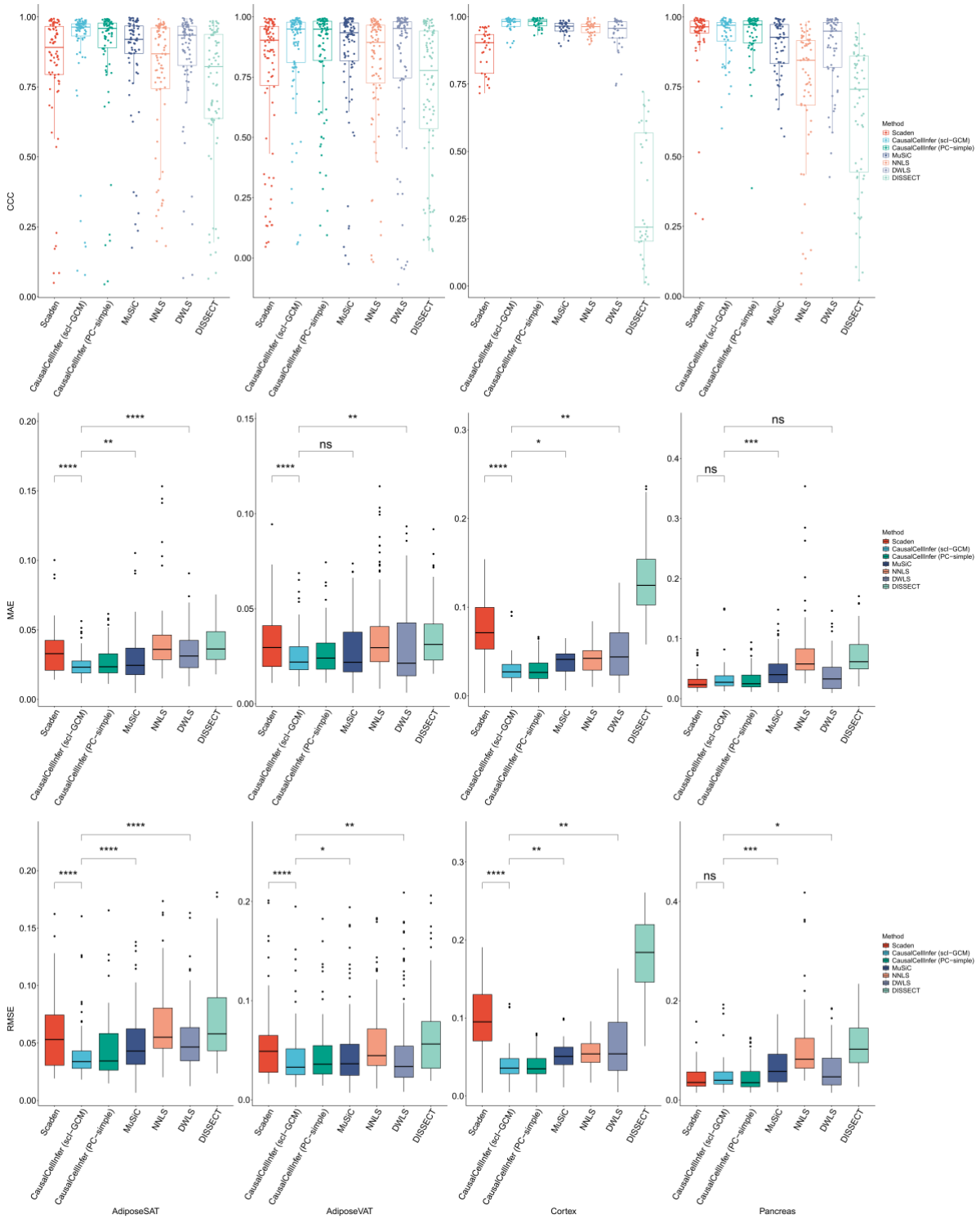

Figure 3 Benchmarking of cell deconvolution accuracy on pseudo-bulk data constructed by combining reference scRNA-seq examples.

Table 2 Comparison of cell deconvolution accuracy on pseudo-bulk data constructed by combining reference scRNA-seq examples

Methods	Cortex			Adipose SAT			Adipose VAT			Pancreas		
	CCC	RMSE	MAE	CCC	RMSE	MAE	CCC	RMSE	MAE	CCC	RMSE	MAE
Scaden	0.758	0.0964	0.0743	0.894	0.0578	0.0347	0.870	0.0555	0.0331	**0.954**	**0.0448**	**0.0295**

CausalCellInfer (scI-GCM)	**0.930**	*0.0422*	*0.0321*	**0.944**	**0.0409**	**0.0248**	**0.925**	**0.0437**	**0.0259**	*0.944*	0.0508	0.0357
CausalCellInfer (PC-simple)	*0.907*	**0.0387**	**0.0295**	*0.932*	*0.0454*	*0.0279*	*0.920*	*0.0459*	*0.0271*	0.938	*0.0464*	*0.0314*
MuSiC	0.874	0.0508	0.0371	0.911	0.0506	0.0299	0.890	0.0485	0.0279	0.909	0.0664	0.0476
NNLS	0.887	0.0535	0.0415	0.863	0.0665	0.0429	0.861	0.0587	0.0376	0.705	0.1104	0.0781
DWLS	0.893	0.0633	0.0499	0.905	0.0541	0.0337	0.865	0.0520	0.0306	0.925	0.0633	0.0423
DISSECT	0.401	0.1775	0.1287	0.811	0.0705	0.0389	0.795	0.0656	0.0357	0.713	0.1128	0.0736

The top 2 performing methods are underlined. The best performing method is also bolded, while the 2nd best methods are in italics.
Note: CCC is the abbreviation for Lin's Concordance Correlation Coefficient, RMSE is the abbreviation for Root Mean Square Error, MAE is the abbreviation for Mean Absolute Error.

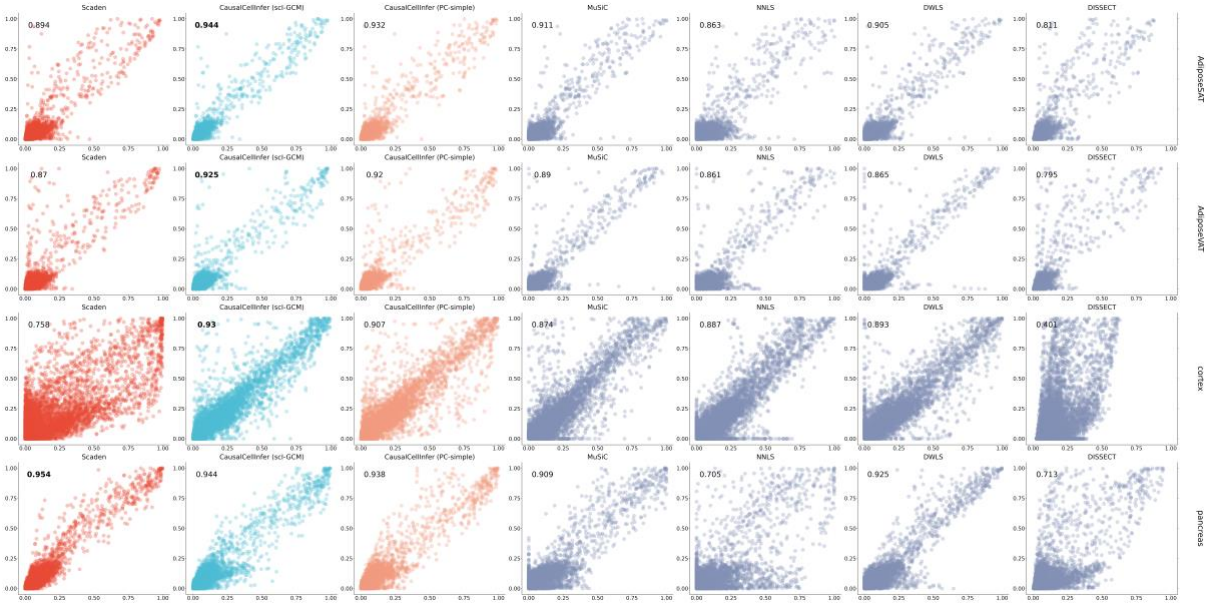

Figure 4 Benchmarking of cell deconvolution accuracy on pseudo-bulk data constructed by combining reference scRNA-seq examples with Lin's Concordance Correlation Coefficient (CCC).

Deconvolution accuracy on pseudo-bulk mixtures

We next benchmarked deconvolution accuracy on pseudo-bulk mixtures derived from scRNA-seq references from four tissues (cortex, pancreas, SAT, VAT). The comparative results with existing methodologies are illustrated in Figures 3-4 and Table 2. Our proposed method, CausalCellInfer (utilizing either scI-GCM or PC-simple alone for marker selection), was benchmarked against five state-of-the-art cell deconvolution techniques: NNLS, DWLS, MuSiC, DISSECT and Scaden. NNLS, DWLS, and MuSiC are all signature-based methods that leverage reference scRNA-seq data to generate the gene expression profiles (GEPs) necessary for cell deconvolution. In contrast, Scaden, DISSECT and our method are deep-learning based methods that employ the reference scRNA-seq data for pseudo-bulk training set construction, which is subsequently used to train cell deconvolution models. To mitigate the risk of overfitting, we terminated the training of deep-learning models after 5000 steps, adhering to the Scaden pipeline.

In summary, our proposed method, CausalCellInfer demonstrated consistent superiority over existing methods such as NNLS, MuSiC, DWLS, DISSECT and Scaden, in deconvoluting pseudo-bulk RNA-seq data. Notably, our approach achieved satisfactory Concordance Correlation Coefficient (CCC), Root Mean Square Error (RMSE) and Mean Absolute Error

(MAE) across the four pseudo-bulk scenarios, when compared to other cell deconvolution methods.

Considering RMSE and MAE, CausalCellInfer(scI-GCM) was the best performing method in two out of the four tissues tested (SAT, VAT); CausalCellInfer (PC-simple) performed the best for cortex, yet CausalCellInfer(scI-GCM) ranked second with very close performance. For pancreas, Scaden performed the best, however CausalCellInfer (both scI-GCM and PC-simple) ranked 2nd and 3rd with close performance. Considering the performance according to CCC, CausalCellInfer(scI-GCM) was the best for three tissues (cortex/SAT/VAT). For pancreas, Scaden again led based on CCC, followed closely by CausalCellInfer(scI-GCM) and CausalCellInfer(PC-simple). Taken together, our proposed framework achieves top performance in cell type deconvolution.

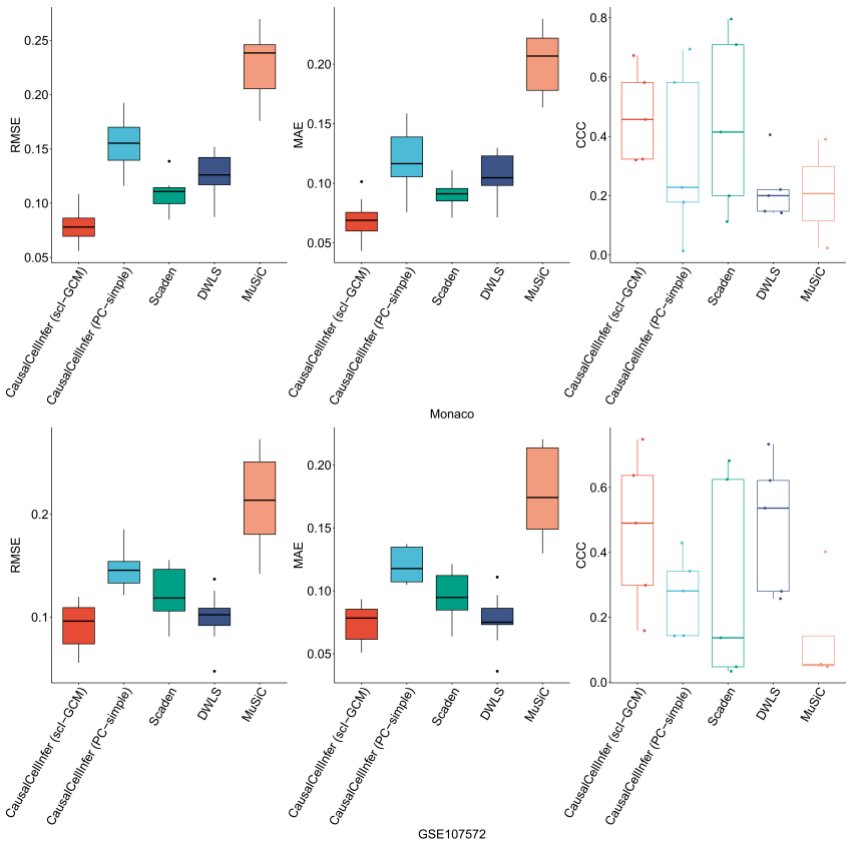

Figure 5 Benchmarking of cell deconvolution accuracy on real bulk data

Both Scaden and our proposed method employ DNNs for cell deconvolution. However, our method consistently outperforms Scaden, as evidenced by significantly CCCs and lower MAE and RMSE across most test scenarios. Moreover, our method exhibits a markedly smaller variance in CCCs across all cell types (Figure 3), underscoring its robustness in deconvoluting cell compositions. During the training phase of the deconvolution model, we opted to use selected marker genes identified by scI-GCM/PC-simple, as opposed to all highly variable genes. Our results indicate that focusing on a subset of informative marker genes that are stable across environments can significantly enhance performance within a deep learning framework. This approach likely leads to a less complex, more robust and transferable deconvolution model less affected by noise in the data.

Evaluation on real-world datasets

In the evaluation of real-world datasets, our method exhibited comparable performance to Scaden, as indicated by analogous RMSE and MAE values. Importantly, our methodology outperformed alternative approaches by attaining a superior CCC across most cell types. Upon application to the Monaco and GSE107572 dataset, our proposed method (CausalCellInfer(scI-GCM)) excelled, surpassing all other methods in terms of both MAE and RMSE. Furthermore, it achieved the highest average CCC among all evaluated methods (Figure 5). Interestingly, This pattern is consistent with prior benchmarking on the same dataset[17], where MuSiC underperformed in real-world PBMC dataset.

Validity of inferred cell-type specific expression profile

We conducted a comparison analysis on the cell-type specific DEGs identified from our estimated CTS profiles (from GWAS data) and the reference scRNA-seq datasets for similar phenotypes. If the inferred CTS expression profiles are reliable, we expect a significant overlap between the DEG identified from the two types of datasets. Here we regard the DEGs from the actual scRNA-seq data as the reference for comparison.

We reveal that the DEGs identified via the CTS expression profiles based on our method and GWAS data exhibit substantial overlap with satisfactory PPV for the majority of cell types. For instance, the DEGs identified for primary cell types in the pancreas, namely acinar cells, alpha cells, beta cells, and delta cells, achieved a PPV exceeding or around 0.5, significantly higher than expectation (Table S1). In adipose tissue, the two predominant cell types, adipocytes and adipose-derived stem and progenitor Cells (ASPCs), demonstrated the highest PPV across all cell types, with values greater than 0.8 in SAT and 0.6 in VAT. These findings validate the reliability of our proposed method for inferring CTS expression profiles.

UKBB application: cell-type proportion and cell-type specific expression estimation at biobank scale

We applied CausalCellInfer to PrediXcan-imputed tissue-specific expression for ~500,000 UKBB participants, estimating cell-type proportions and CTS expression profiles for depression (frontal cortex), type 2 diabetes (pancreas), and obesity (SAT and VAT). Binary traits (except obesity) were defined using self-reported diagnoses and ICD codes. Obesity was defined as BMI > 30.

Distinguishing cell types associated with various diseases

To elucidate the relationship between the estimated cell proportions and diseases/traits, we conducted univariate association tests for each cell type via logistic/linear regression, with the trait/disease of interest as outcome. We revealed that the estimated proportions of various cell types were associated with different diseases, as detailed below.

Depression

We deconvoluted the frontal cortex into six distinct cell types and discovered that the estimated cell proportions for astrocytes, micro/macro cells, and oligodendrocytes were significantly associated with risk of depression (Figure 6, Table S2). Notably, a lower inferred proportion of microglial/macrophage cells and oligodendrocytes were associated with depression, while the proportion of neurons was elevated. These observations are consistent with existing literature; for instance, Rajkowska et al.[36] observed a decrease in both the density and ultrastructure of oligodendrocytes in the prefrontal cortex in depression.

Type 2 Diabetes

In our analysis of ten distinct cell types within the pancreas, we observed that the inferred proportions of six cell types, including beta cells, exhibited a significant association with T2DM risks. Notably, significantly lower inferred proportions of beta cells were observed in T2DM. Conversely, we detected associations of elevated proportions of acinar, delta, duct epithelial, and mesenchymal cells with T2DM, with the increase in mesenchymal cell proportions being especially prominent (Figure 6). We also observed significant alpha beta ratio increase in T2DM patients. Numerous studies have underscored the link between T2DM and a decrease in beta-cell mass and function[37,38]. Also, previous GWAS have pinpointed *IRX3* as a susceptibility gene for T2DM, and knockdown of the *IRX3* orthologue in zebrafish led to a decrease in insulin-producing beta cells[39].

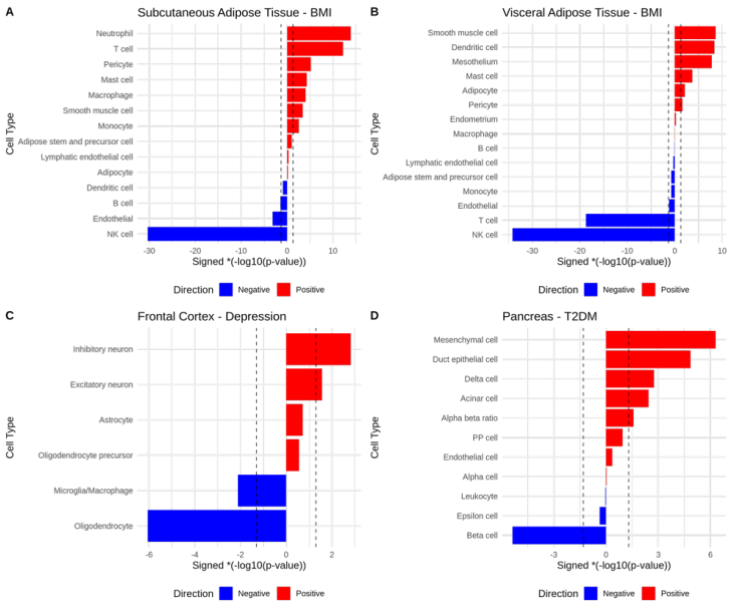

Figure 6 Associations between estimated cell proportions and disease onset

Obesity

In subcutaneous and visceral adipose tissues (SAT/VAT), we deconvoluted 14 and 15 distinct cell types respectively, and tested associations with obesity status. More than half of these cell types, including smooth muscle cells (SMC), and adipocyte, were found to be associated with obesity. Interestingly, the involvement of immune-related cell types in obesity was observed to vary between adipose tissues. Some cell types, such as natural killer cells (NK

cells) and T cells, were associated with obesity in both tissues (SAT and VAT), while others, like neutrophils and macrophage, were associated with obesity in only one tissue.

Compared to normal subjects, there was a significant increase in the estimated proportions of various cell types, including SMCs, macrophage, mast, monocyte, neutrophil, pericyte and T cells in SAT. Conversely, a significant decrease in NK, endothelial and B cell proportion was noted.

Notably, a significant increase in pericyte proportions was also observed in both adipose tissues. Supporting our findings, Tang et al.[40] demonstrated that pericytes are one of the precursor cells of adipocytes, underscoring their crucial role in obesity development. Watanabe et al.[41] also showed that mice fed with a high-fat diet (HFD) exhibit a higher level of pericyte count than those on a chow diet.

Interactions between cell types and their associations with disease

We tested the effects of cell type interactions on 4 tissue-specific outcome variables using multivariate regression analyses, including depression in the frontal cortex, T2DM in the pancreas, BMI in the SAT, and BMI in the VAT.

Our results revealed that interactions between certain cell types were significantly associated with disease risks (Table 3 and S3). For example, the interaction between astrocytes and microglia/macrophages was linked to depression (based on expression from the frontal cortex). This is consistent with the literature suggesting a disrupted structural communication between excitatory neuron and oligodendrocyte in the frontal cortex, which contributes to both cognitive retardation and emotional instability commonly seen in depression [42].

In addition, we found that interaction of cell types *across different tissues* may contribute to disease risks. For example, the interaction between B cells on SAT and dendritic cells in VAT was significantly associated with obesity. Both cell types play crucial roles in maintaining immune homeostasis, and previous studies[43,44] has revealed significant involvement of inflammation in obesity. To conclude, interactions between certain cell types may play an important role in disease pathogenesis.

Table 3 Cell type interactions significantly associated with diseases

Tissue	Trait	Cell type interaction	P-values
Frontal cortex	Depression	Excitatory * Oligodendrocyte	1.11E-02
Pancreas	T2DM	Duct epithelial cell * Mesenchymal cell	3.22E-03
		Leukocyte * Beta cell	3.19E-03
		Mesenchymal cell * Alpha cell	3.06E-02
		Mesenchymal cell * PP cell	5.65E-03
		Alpha cell * PP cell	3.28E-03
SAT	Obesity	Adipose stem and precursor cell * Lymphatic endothelial cell	4.73E-03
		Adipose stem and precursor cell * Mast cell	1.78E-03
		Adipose stem and precursor cell * NK cell	3.17E-04
		Adipose stem and precursor cell * Pericyte	1.43E-03
		Adipose stem and precursor cell * T cell	4.13E-07
		Lymphatic endothelial cell * Smooth muscle cell	1.81E-02
		Lymphatic endothelial cell * Adipocyte	1.88E-02
		Lymphatic endothelial cell * B cell	2.52E-04
		Lymphatic endothelial cell * Dendritic cell	1.76E-03

		Lymphatic endothelial cell * Mast cell	9.74E-03
		Lymphatic endothelial cell * NK cell	9.25E-06
		Lymphatic endothelial cell * Pericyte	4.09E-02
		Lymphatic endothelial cell * T cell	4.59E-03
		Smooth muscle cell * Adipocyte	4.13E-03
		Smooth muscle cell * Macrophage	9.21E-03
		Smooth muscle cell * Mast cell	2.89E-02
		Smooth muscle cell * Neutrophil	2.40E-02
		Smooth muscle cell * NK cell	4.25E-05
		Smooth muscle cell * Pericyte	1.31E-06
		Smooth muscle cell * T cell	3.46E-03
		Adipocyte * Mast cell	4.24E-03
		Adipocyte * T cell	6.89E-03
		B cell * Mast cell	4.21E-03
		B cell * Neutrophil	3.84E-02
		B cell * NK cell	1.05E-02
		B cell * Pericyte	7.65E-03
		B cell * T cell	1.97E-04
		Dendritic cell * Neutrophil	2.39E-03
		Macrophage * NK cell	4.25E-03
		Macrophage * Pericyte	1.24E-02
		Macrophage * T cell	2.48E-04
		Mast cell * NK cell	1.32E-03
		Mast cell * Pericyte	2.82E-02
		Neutrophil * NK cell	7.85E-07
		Neutrophil * Pericyte	1.37E-03
		NK cell * Pericyte	2.46E-02
		Pericyte * T cell	2.23E-02
VAT	Obesity	Adipose stem and precursor cell * Adipocyte	7.95E-03
		Adipose stem and precursor cell * B cell	1.36E-02
		Adipose stem and precursor cell * Endometrium	5.41E-03
		Adipose stem and precursor cell * Mesothelium	3.07E-02
		Lymphatic endothelial cell * Adipocyte	9.66E-03
		Lymphatic endothelial cell * Endometrium	5.57E-08
		Lymphatic endothelial cell * Macrophage	4.76E-02
		Lymphatic endothelial cell * Mesothelium	5.68E-03
		Lymphatic endothelial cell * NK cell	3.75E-03
		Lymphatic endothelial cell * Pericyte	1.67E-04
		Smooth muscle cell * Endometrium	1.03E-04
		Smooth muscle cell * NK cell	1.28E-04
		Smooth muscle cell * Pericyte	3.45E-05
		Adipocyte * Endometrium	2.78E-09
		Adipocyte * Mesothelium	2.23E-02
		Adipocyte * NK cell	3.19E-05
		Adipocyte * Pericyte	2.73E-05
		B cell * Endometrium	2.91E-08
		B cell * Mesothelium	1.70E-04
		B cell * NK cell	1.46E-06
		B cell * Pericyte	1.22E-07
		Dendritic cell * Endometrium	1.53E-02
		Dendritic cell * Macrophage	4.10E-02
		Dendritic cell * Mesothelium	7.23E-03
		Dendritic cell * Pericyte	1.41E-03
		Endometrium * Macrophage	5.64E-04

Endometrium * NK cell	2.82E-07
Macrophage * Pericyte	1.90E-02
Mast cell * Monocyte	2.42E-02
Mesothelium * NK cell	1.27E-04
NK cell * Pericyte	9.47E-04

Cell type-specific genes linked to disease susceptibility ***(focusing on putative direct casual genes as prioritized by PC-simple)***

Putative direct causal effect genes for disease susceptibility were found to exhibit variation across different cell types (Figure 7, Table S4). We refer to genes selected by PC-simple as putative direct causal genes. We define genes detected in exactly one cell type as cell-type-specific and genes detected in ≥2 cell types as cell-type-common.

Depression

For depression, approximately one-third of the 94 prioritized genes were cell-type-specific. A notable example is *KLC3*, identified as a causal gene for depression susceptibility exclusively within excitatory neurons (Table 4; Table S4). This finding aligns with a previous study[45] reporting neuronal hypomethylated changes of *KLC3* in the prefrontal cortex of bipolar patients, suggesting a potential epigenetic link to mood disorders.

Our analysis revealed a nuanced relationship between cell-type alterations and disease status. Some cell types exhibited both altered proportions and changes in gene expression, while others showed only gene expression alterations without significant proportional changes. For instance, despite no significant association between oligodendrocytes precursor cells (OPCs) proportion and depression susceptibility, this cell type harbored 16 putative direct-effect genes.

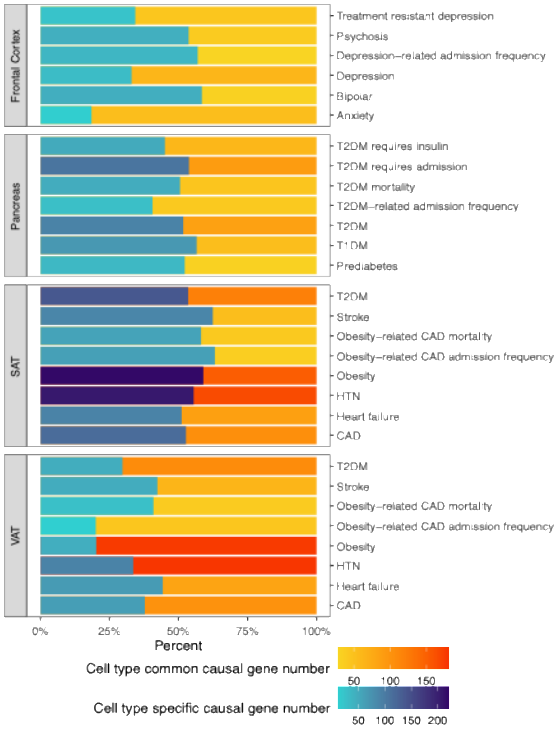

Figure 7 Cell specificity comparison of the identified putative direct causal genes by PC-simple for disease onset and prognoses

Table 4 Top putative direct causal genes by PC-simple (ranked by P-value) from inferred cell-type specific (CTS) expression profiles

Tissue	Disorder	Genes
Frontal cortex	Depression	*KLC3, ZSCAN9, RP9, CCDC64B, ZSCAN21, SPATC1L*
Pancreas	T2DM	*JAZF1, IRX3, CLEC18A, HLA-DQA2, CEP68, NOTCH2, UBA7, CILP2, MAN2C1, OPRL1*
SAT	Obesity	*LMOD1, HSD17B12, C1QTNF4, ITIH4, TRIM73, CLEC18A, ATP8B3, UHRF1BP1, PRSS16, EIF3C, ADORA2B, SERPINC1, SLC5A11, SULT1A2*
VAT	Obesity	*SULT1A2, CENPO, CLEC18A, TRIM73, RNF123, CTSW, DNAH10, ATP8B3, SLC5A11, SERPINC1, C1QTNF4, RBM6, HSD17B12, TTLL6, ADORA2B*

T2DM

For T2DM, we identified 178 putative direct-effect genes associated with disease risk, with varying involvement across different cell types. In contrast to depression, where cell-type specificity was more prominent, nearly half (86 genes, ~48.3%) of the T2DM-associated genes were common, detected in two or more cell types (Figure 7, Table S4). For instance, *JAZF1* was identified as a direct causal gene in 6 distinct cell types, such as acinar, alpha, beta, and delta cells (Table S4). The association of *JAZF1* with increased T2DM risk is well-established[46,47].

Obesity

Our analysis of obesity revealed a substantial number of putative direct-effect genes in both subcutaneous and visceral adipose tissues, but with a large difference in the distribution of cell-type-specific versus common genes. In SAT, a high degree of cell-type specificity was observed, with approximately 69.7% of causal genes being specific to a single cell type (only 77 genes, or 30.3% of ~254 identified genes, were cell-type-common). Conversely, VAT displayed a much higher proportion of common genes, with 178 genes (79.8% of ~223 identified genes) contributing to obesity across multiple cell types. This suggests a broader, multi-cellular involvement of causal mechanisms in VAT.

A notable cell-type-specific example in SAT is *PRSS53* (also known as *POL3S*), identified as a causal gene exclusively in adipocytes. This finding is supported by a previous study[48] showing *PRSS53* downregulation in obesity. Given *PRSS53*'s critical role in maintaining mitochondrial health and function for metabolism[49], its potential causal role in obesity development within SAT is further substantiated.

Cell type-specific genes involved in disease prognosis (focusing on putative direct casual genes as prioritized by PC-simple)

Identifying specific genes that contribute to disease outcomes or prognosis is crucial for developing personalized treatments. Leveraging detailed prognosis data from the UKBB, which are often absent in single-cell studies, we conducted analyses to pinpoint putative direct-effect genes associated with disease-specific mortality and hospitalization frequencies.

Depression

For depression, our analysis revealed distinct patterns for treatment resistance and hospitalization frequency. A substantial proportion (~65.7%) of putative direct-effect genes linked to treatment resistance were common, identified in at least two cell types (Table S4.a). Conversely, genes associated with the frequency of hospital admissions were predominantly cell-type specific (~56.9%) (Table S4.a).

Type 2 Diabetes Mellitus

In T2DM, we observed that a majority (~59.4%) of putative direct-effect genes influencing hospitalization frequencies were common across multiple cell types (Table S4.b). However, for T2DM-related mortality, a slightly different distribution emerged, with approximately half (~49.5%) of the causal genes being cell-type common (Table S4.b). For example, we identified the gene *LIMK1* as a determinant of mortality in T2DM within beta cells. This finding aligns with a recent research indicating a relationship between genetic polymorphisms of *LIMK1* and glucose dysregulation and insulin exocytosis[50].

Obesity

Given obesity's established role as a risk factor for coronary artery disease (CAD)[51], we investigated genes influencing CAD-related mortality and hospitalization frequency in obese patients. A notable divergence was observed between adipose tissues: SAT showed a higher proportion of cell-type-specific genes, while VAT exhibited a higher proportion of common genes influencing these outcomes (Table S4.c and Table S4.d).

Other Disorders

The distribution of identified causal genes varied considerably across tissues. In SAT, for most obesity-related traits (CAD, T2DM, hypertension, heart failure, stroke), cell-type-specific causal genes predominated. The inverse pattern was consistently found for these same traits in VAT.

Similarly, for traits associated with the frontal cortex (anxiety, bipolar disorder, psychosis) and pancreas (prediabetes, T1DM, T2DM requires admission, T2DM requires insulin), the majority of identified causal genes were found to be cell-type specific.

OpenTargets enrichment results

We conducted enrichment analyses for depression, T2DM and obesity. For obesity, we performed four sets of enrichment analyses for each pair of tissues associated with obesity. Specifically, we investigated whether the genes we identified were significantly overrepresented as targets for BMI/obesity/weight/WHR (waist-hip-ratio) in the OpenTargets database.

Encouragingly, the sets of causal genes identified for all the studied diseases were significantly enriched among the disease targets catalogued in OpenTargets (Table 5). This observation provides further validation for the usefulness of our proposed framework in estimating cell proportions and CTS expression profiles.

Table 5 OpenTargets enrichment results based on gene-sets identified for disease susceptibility

Tissue	Trait	Enrichment P-value
Pancreas	T2DM	1.034E-14
	T2DM requires admission	1.092E-13
	T2DM requires insulin	1.535E-04
SAT	Obesity	1.207E-02
	Body weight	4.568E-21
	BMI	2.699E-35
	WHR	1.566E-10
VAT	Obesity	3.965E-04
	Body weight	4.370E-18
	BMI	6.224E-32
	WHR	2.377E-11
Frontal cortex	Depression	4.654E-02

Results of further pathway enrichment analyses

We combined the identified cell-type specific causal genes ((prioritized by the PC-simple algorithm) for pathway enrichment analysis. The enriched pathways were shown in Table S5. For instance, "Pyrophosphate hydrolysis" was significantly enriched for depression, consistent with Sha et al.[52] work suggesting its protective role in astrocytes against depression via lysosomal function.

For T2DM, pathways like "Vitamin B3 (nicotinate and nicotinamide) metabolism", "Pre-NOTCH Processing in the Endoplasmic Reticulum", and "Pre-NOTCH Processing in Golgi T2DM" were significantly enriched. This aligns with Lu et al.[53] findings on dysregulated Notch signaling worsening T2DM through hepatic fat deposition and glucogenesis.

We also identified enriched pathways based on gene sets associated with disease prognosis (Table S5). For instance, the "Apoptotic cleavage of cellular proteins" pathway was significantly enriched in treatment-resistant depression (TRD). This aligns with Kosten et al.[54] who showed that antidepressant treatment can differentially regulate apoptotic genes in the brain structures implicated in the etiology and treatment of depressive disorder. Fries et al.[55] also showed that apoptotic events may be responsible for the chronic low-grade inflammation in depressive patients. The "Ceramide signaling pathway" pathway was enriched in relation to the frequency of depression-related admissions. Gulbins et al.[56] reported that increased ceramide concentration can lead to the reversal of stress-induced MDD via activation of autophagy.

"Role of mitochondria in apoptotic signaling" emerged as a significantly enriched pathway in relation to T2DM-related mortality, supporting existing literature that apoptotic signaling drives cell death in diabetic complications like cardiovascular diseases (CVD) [57] and beta cell failure[58], increasing T2DM mortality.

Results from conditional TWAS (controlled for cell-type proportions)

We compared trait-associated genes identified from two different linear models using bulk expression data, one with estimated cell proportions as covariates (conditional TWAS) and one without. Genes from the adjusted TWAS are shown in Table 6 and S6. Adjusting for estimated cell proportions when performing TWAS analysis led to significant coefficient changes for numerous genes (Table 7 and S7-8), consistent with cell composition acting as a confounder or mediator of bulk-level associations. For example, adjusting for cell proportions resulted in significant effect size changes for ~59.7% of T1DM-associated genes. Some genes were significant only in the standard TWAS (e.g., *EFEMP2* in SAT obesity), illustrating that bulk-level associations can be sensitive to compositional variation.

Results from CTS-TWAS (univariate association tests)

Here we report the results from CTS-TWAS based on univariate association tests (distinct from PC-simple, which involves repeated conditional tests). Genes identified in at least two cell types were classified as 'cell-type common', while those in only one were 'cell-type specific'.

Similar to direct causal gene analysis by PC-simple, the majority of genes identified from CTS-TWAS were cell-type specific (Table 8, S9-10). The number of significant genes for each trait are listed in Table S9, with FDR<0.05 considered as 'significant' genes. Among traits with significant genes, T2DM-related traits typically showed higher cell type-specificity, whereas most obesity-related traits showed broader sharing across cell types.

OpenTargets enrichment analysis of CTS-TWAS genes for 4 tissue-disorder pair with >=5 genes (including T2DM and hospitalization frequency (pancreas), obesity (SAT/VAT)) confirmed significant or nominal enrichment among known disease targets listed in OpenTargets (Table 9).

Furthermore, we compared the genes identified from a causal inference approach (PC-simple, which prioritizes putative direct-effect genes) and univariate association analysis (TWAS) using inferred CTS expression profiles. For most traits, we identified considerably more associated genes from univariate TWAS (Table S9). As PC-simple relies on repeated conditional independence testing, it is more stringent than univariate testing. Interestingly, PC-simple also uncovered novel associations not found in standard association tests (Table S11). We hypothesize that some TWAS-identified genes may be "peripheral" (indirect causal) genes acting through the direct causal genes identified by PC-simple. As expected, if we relax the multiple testing correction criteria, there is an increased overlap between PC-simple and standard association analyses (Table S11).

Table 6 Examples of top (ranked by P value) trait-associated genes identified from TWAS after adjustment for estimated cell-type proportions

Tissue	Disorder	Genes
Frontal cortex	Depression	*ZSCAN9, BTN3A2, C4A*
Pancreas	T2DM	*IRX3, NOTCH4, CD84*
SAT	Obesity	*RTN4, TPCN2, SEMA3F*
VAT	Obesity	*TBL2, NOVA1, SULT1A2*

Table 7 Trait-associated genes from a standard TWAS analysis with significant coefficient changes after adjusting for estimated cell-type proportions

Tissue	Disorder	Genes	Estimates (standard TWAS)	Pvalues (standard TWAS)	Estimates (conditional TWAS)	Pvalues (conditional TWAS)	Coeff_diff_P	Pvalue_BH (standard TWAS)	Pvalue_BH (conditional TWAS)
Pancreas	T1DM	*ZSCAN31*	0.00118	2.76E-07	0.00110	1.96E-06	6.28E-08	2.03E-05	1.44E-04
		TRIM26	-0.00124	2.19E-02	-0.00187	6.77E-04	2.57E-11	3.66E-01	3.48E-02
	T2DM	*BAK1*	0.01193	2.30E-07	0.01102	2.02E-06	1.00E-04	3.94E-05	4.32E-04
		CD84	0.00218	1.17E-01	0.00591	4.36E-04	7.60E-05	5.11E-01	2.02E-02
	T2DM requires admission	*HLA-L*	0.00598	7.37E-07	0.00564	3.27E-06	5.17E-04	8.60E-05	3.51E-04
		AGER	0.01238	3.16E-05	0.01050	5.04E-04	3.66E-04	1.93E-03	1.58E-02
	T2DM requires insulin	*HLA-DMA*	0.00139	2.21E-05	0.00150	6.49E-06	4.34E-02	5.54E-03	2.08E-03
		SKIV2L	-0.00096	2.68E-03	-0.00145	1.79E-05	6.51E-06	1.56E-01	4.61E-03
SAT	CAD	*GRK4*	0.00482	2.06E-04	0.00479	2.22E-04	1.76E-02	2.05E-02	2.10E-02
		C6orf106	0.01231	1.38E-06	0.01241	1.14E-06	1.82E-02	6.75E-04	6.25E-04
	HTN	*BANF1*	0.00542	1.21E-01	0.01031	4.81E-03	5.56E-06	3.70E-01	4.44E-02
		SNX31	-0.02702	2.85E-05	-0.02690	3.09E-05	3.13E-02	8.19E-04	8.44E-04
	Heart failure	*HLA-S*	-0.00242	8.23E-02	-0.00530	2.47E-04	1.35E-13	5.92E-01	2.72E-02
		TSLP	0.01451	7.15E-06	0.01444	7.88E-06	2.41E-03	1.24E-03	2.58E-03
	Obesity	*RTN4*	-0.01358	6.42E-03	-0.01438	3.92E-03	2.52E-05	5.83E-02	4.44E-02
		CENPC	-0.01467	9.49E-06	-0.01476	8.36E-06	7.65E-04	2.76E-04	2.98E-04
	T2DM	*SPTA1*	-0.00482	1.60E-03	-0.00470	2.10E-03	5.85E-03	3.62E-02	5.00E-02
		NPC1	-0.01650	6.72E-06	-0.01643	7.27E-06	1.64E-02	5.01E-04	6.28E-04
VAT	CAD	*ITPR3*	0.01946	6.84E-05	0.02006	4.21E-05	4.92E-02	1.72E-02	1.25E-02
		ADIPOQ	-0.00925	3.12E-03	-0.01174	3.38E-04	9.65E-03	1.08E-01	3.89E-02
	HTN	*MLXIPL*	-0.01251	4.89E-02	-0.01985	2.63E-03	3.90E-05	2.23E-01	3.06E-02
		WASH2P	-0.01919	5.32E-06	-0.01910	5.94E-06	1.42E-02	2.26E-04	2.52E-04
	Heart failure	*MAST2*	0.01151	8.62E-05	0.01158	7.90E-05	7.87E-03	1.13E-02	1.15E-02
		MICB	0.00245	2.31E-03	0.00279	5.78E-04	1.56E-03	1.12E-01	4.58E-02
	Obesity	*TBL2*	0.00999	9.80E-03	0.01204	1.95E-03	1.70E-08	7.27E-02	2.39E-02
		ERAP1	-0.01853	5.76E-06	-0.01861	5.22E-06	1.07E-02	2.01E-04	2.18E-04
	T2DM	*HLA-C*	-0.00229	3.32E-02	-0.00393	7.07E-04	1.58E-04	2.55E-01	2.47E-02
		HLA-L	-0.01872	4.50E-21	-0.01607	2.53E-14	1.68E-04	1.07E-17	6.98E-11

Note: "Coeff_diff_P" indicates the calculated p-value for coefficient change between standard and conditional TWAS, standard TWAS estimates are used as the reference for the calculation of coefficient differences.

Table 8 Top-ranked (by absolute effect size from linear regression) trait-associated genes for T2DM (tissue studied: pancreas) identified from CTS-TWAS(univariate tests)

Cell type	Genes	Estimates	P-values	Q-values
acinar cell	*PAX9*	-0.0959	1.97E-04	4.10E-02
	MYRFL	0.0499	2.22E-04	4.37E-02
	LINGO4	-0.0329	8.22E-06	5.12E-03
alpha cell	*EXTL1*	0.0600	9.78E-05	1.79E-02
	ESPNL	-0.0432	4.30E-04	3.50E-02
	ZNF215	0.0355	8.14E-06	3.78E-03
beta cell	*USP6*	-0.0546	5.26E-04	4.19E-02
	HLA-DQB2	-0.0538	4.24E-04	3.78E-02
	IP6K3	-0.0506	7.12E-04	4.93E-02
delta cell	*RMI2*	0.0339	3.61E-04	4.83E-02
	TNNI3K	-0.0303	4.43E-04	4.89E-02
	FAM153A	0.0290	1.38E-05	6.47E-03
duct epithelial cell	*SLC25A41*	-0.0471	7.26E-05	1.51E-02
	IRX3	-0.0216	4.86E-04	4.91E-02
	GPRC5A	0.0213	1.73E-08	6.47E-05
endothelial cell	*GNA15*	-0.0349	3.13E-06	2.86E-03
	CRHR2	-0.0328	2.69E-05	1.12E-02
	RNF122	0.0291	9.44E-05	1.77E-02
epsilon cell	*PRAP1*	-0.0737	1.12E-04	1.14E-02
	WNT10B	-0.0530	3.60E-04	2.47E-02
	CLNK	-0.0485	3.43E-04	2.42E-02
leukocyte	*GYS2*	-0.0616	1.93E-04	2.38E-02
	NAALAD2	0.0579	1.43E-06	8.92E-04
	TDRKH	-0.0225	2.09E-04	2.38E-02
mesenchymal cell	*PAX9*	0.2298	2.39E-19	8.93E-16
	HCG9	0.0546	7.89E-06	3.67E-03
	SLC6A19	-0.0394	4.78E-04	4.97E-02
PP cell	*PCDHA2*	-0.0599	1.06E-04	1.04E-02
	C10orf62	-0.0541	4.79E-05	6.57E-03
	CDC25A	-0.0466	1.02E-05	2.95E-03

Note: Q-value indicated false discovery rate which are calculated from P-value after correcting for multiple testing by Benjemini-Hochberg method

Table 9 OpenTargets enrichment results based on gene-sets identified from CTS-TWAS using inferred CTS expression profiles

Tissue	Trait	P-value
Pancreas	T2DM	9.405E-19

	T2DM requires admission	6.044E-19
	T2DM requires insulin	7.659E-03
SAT	Obesity	1.514E-64
	Body weight	9.696E-42
	BMI	3.201E-60
	WHR	9.026E-28
VAT	Obesity	1.639E-98
	Body weight	6.682E-54
	BMI	4.919E-63
	WHR	1.559E-38
Cortex	Depression	1.204E-02

Discussion

In this study, we introduced CausalCellInfer, a comprehensive framework that integrates invariant causal prediction, deep learning, and regularized matrix completion, with applications to GWAS-imputed transcriptome profiles. This framework is designed to deconvolute cell compositions and estimate CTS expression profiles from bulk and genotype-imputed transcriptomes at biobank scale (N~500K). As a fully data-driven method, CausalCellInfer first identifies marker genes from the reference scRNA-seq dataset, leveraging the invariant causal prediction principle. It then trains a deep neural network-based deconvolution model using these identified marker genes. Finally, it employs a regularized matrix completion algorithm, ENIGMA, to estimate the corresponding CTS profiles.

We validated the robustness and accuracy of CausalCellInfer in cell deconvolution by comparing it against five state-of-the-art deconvolution methods, namely DWLS, NNLS, Scaden, DISSECT and MuSiC. This comparison was conducted on four simulated and two real peripheral blood mononuclear cell (PBMC) datasets. To further demonstrate the reliability of CausalCellInfer in estimating CTS profiles, we have conducted a wide range of enrichment analyses, including DEGs enrichment analysis, OpenTargets enrichment analysis, and pathway analyses.

Moreover, we applied CausalCellInfer to GWAS-imputed expression datasets from the UKBB (four different tissues) and performed cell-type-level analyses for 29 clinical phenotypes, including susceptibility, prognosis-related outcomes, interaction analyses between cell types, CTS-TWAS, and composition-adjusted (conditional) TWAS. These applications illustrate a central advantage of the approach: it enables cell-type-resolved hypothesis testing in cohorts where the scale and phenotypic depth are far beyond what is typically feasible for single-cell studies.

Strengths of the framework

The framework presented in this study exhibits several strengths. Firstly, a notable advantage is that CausalCellInfer incorporated a novel marker-gene informed and deep learning-based cell deconvolution model. Marker genes are learned from scRNA-seq reference data rather than relying on predefined marker lists.

Secondly, our framework does not require a constant marker-gene expression assumption as required in classical signature-based methods (i.e., a given marker gene is expressed at a constant level across samples within a specific cell type) for accurate estimation, making it more robust to inter-subject heterogeneities or other technical variabilities across samples.

Thirdly, our method is likely more resistant to the influence of confounders and selection bias than signature-based methods. As also reported in ref[16], deep learning methods are able to learn latent representations of cell-types. As such, the network nodes effectively capture higher-order features that are less susceptible to noise and technical bias present in the original input expression data.

Fourth, our incorporation of the invariance principle further enhances the robustness of the method to confounding and improves generalizability to new samples.

Fifth, despite being based on deep neural networks, our method is time- and resource-efficient, as it is trained on a compact set of informative markers. It operates much faster than most competing methods under study.

Sixth, as shown in real data applications, the proposed framework is scalable to large biobanks and genotype-imputed transcriptome data. It is designed to operate when only a subset of genes is available and when expression is derived from genetic prediction, enabling cell-type-resolved analyses at sample sizes that are typically infeasible for scRNA-seq.

Finally, a variety of downstream analyses are enabled by ***individual-level*** inference. Once cell-type proportions and CTS profiles are inferred per person under our framework, many analyses becomes possible, including CTS-TWAS, CTS direct causal gene detection via PC-simple, interaction models, prognosis modeling, and composition-adjusted TWAS.

Related approaches

Numerous methods have been proposed to deconvolute cell compositions and estimate CTS expression profiles from bulk samples. However, these methodologies differ substantially from our current work. For instance, CIBERSORT/CIBERSORTx[9] can infer cell compositions and CTS expression profiles from bulk transcriptomes. It utilizes reference gene expression profiles (GEPs) of predetermined marker genes for cell deconvolution. However, classical signature-matrix methods typically rely on the assumption of constant marker gene expression levels within cell types across samples, which might not hold in practice.

Another similar line of work is deep learning-based cell deconvolution and CTS expression estimation methods, such as Scaden and TAPE. Menden et al.[15] introduced Scaden, a deep neural network-based cell deconvolution model designed to dissect bulk samples. Following a similar approach to Scaden, Chen et al.[16] proposed an autoencoder-based model, TAPE, to infer cell compositions and CTS profiles from bulk samples. Rather than relying on GEPs of marker genes, they trained their deconvolution model directly on the reference scRNA-seq using typically ~10,000 highly variable genes, making them computationally intensive and prone to learning dataset-specific artifacts or noise. CausalCellInfer occupies a strategic middle ground that combines the non-linear flexibility of deep neural networks with the rigorous, parsimonious feature selection of causal inference methods. Leveraging the invariance principle, the identified invariant marker genes from scI-GCM were likely more resistant to data bias and noises, and more generalizable to new datasets.

Another important distinction of our work from all previous studies on cell deconvolution is that we pioneered applications to genotype-imputed expression data, especially from large biobanks. Our work helps to bridge the gap between GWAS/sequencing and scRNA-seq studies and enables a wide variety of individual-level analysis with cell proportions and CTS expression and their relationship with diverse clinical phenotypes in a large sample. Such analyses are generally not possible in conventional single-cell studies due to the limited number of subjects and phenotypic information available.

More broadly, several frameworks link scRNA-seq and GWAS to identify disease-relevant cell types or to interpret GWAS signals in a cellular context. For example, Jagadeesh et al.[4] integrated GWAS summary statistics, epigenomics and scRNA-seq data to detect the underlying cell types and processes by which genetic variants influence the target traits, and utilized stratified LD score regression to calculate heritability enrichment. In another study, scGWAS was developed to leverage scRNA-seq to interrogate CTS transcriptional changes induced by genetic variants[59]. It identified associations between traits and cell types by examining whether genes implicated by GWAS exhibit concordant activation in a particular cell type using proportion tests. Additionally, Zhou et al.[60] proposed a CTS transcriptome-wide association framework (this work was released after our preprint was first posted in Oct 2024[61]). The approach leveraged external single-cell transcriptome and genotype data, and used Enformer-inferred epigenomics features to predict CTS expression; then a linearized model was built and association tests performed using GWAS summary statistics.

The current work is different from the aforementioned methods, which mainly operate with and focus on GWAS summary data. Our objective is fundamentally different: we infer ***individual-level*** cell-type proportions and CTS expression from bulk or genotype-imputed transcriptomes, enabling biobank-scale, cell-type-resolved association analyses across a wide phenotype spectrum. Individual-level data also allows for more diverse types of analysis, such as interaction modeling, cell-type-composition-adjusted TWAS, prioritization of potential direct causal genes via PC-simple, adjustment for other genetic/clinical covariates, flexible definition of outcome phenotypes, and the use of non-linear or machine learning models, which are generally not feasible with GWAS summary statistics alone. Moreover, our approach is methodologically novel and distinct, combining causal inference and invariance principles with deep neural networks and regularized matrix completion to estimate cell compositions and corresponding CTS expression profiles, a comprehensive strategy not previously employed.

Limitations

Several limitations are worth noting but are likely increasingly addressable in practice. First, as with all reference-based methods, performance depends on the coverage and representativeness of the scRNA-seq reference. Rare cell types with few cells can yield less stable markers and less accurate cell fraction estimates. This challenge is expected to diminish as the volume of publicly available single-cell data grows, and it may be further mitigated by integrating multiple reference datasets and employing appropriate harmonization and batch correction methods. Furthermore, like most reference-based approaches, differences in factors such as ancestry, age, disease burden, and technical platforms between the reference and target cohorts (e.g., UK Biobank) can degrade transferability. Our proposed invariance-based marker

selection aims to improve robustness and transferability across diverse environments, though it cannot entirely eliminate the effects of such mismatches.

Second, genotype-imputed expression reflects only the genetically regulated component of expression, and typically has reduced dynamic range relative to measured bulk RNA-seq. Our pseudo-bulk calibration provides a pragmatic way to align scales for deconvolution and CTS reconstruction, but future work could further refine this step or incorporate uncertainty in imputation.

Third, similar to the principles of Mendelian Randomization (MR), our analysis is based on genetically predicted expression, and focuses on how expression influences outcomes. The approach does not capture reverse causality, where phenotypic traits might induce changes in gene expression. Also, our characterization of identified genes and markers as "causal" is contingent upon satisfying the standard assumptions of causal graphical models and invariant prediction. Consequently, while our findings are promising, these outputs should be interpreted as hypothesis-generating and putative, requiring experimental validation for definitive confirmation. It is worth noting, however, that markers identified as invariant are expected to be less sensitive to context-induced shifts. Therefore, even if they are not strictly causal, they are likely to be more transferable and robust across different cohorts and conditions. Despite the above limitations, our proposed framework is highly flexible and holds potential for extension to address these challenges in future developments.

Conclusions

In summary, CausalCellInfer offers a scalable strategy for converting bulk and genotype-imputed transcriptomes into *cell-type-resolved* readouts. This enables a wide variety of analyses at biobank scale, such as systematic CTS-TWAS, cell type composition-aware inference, interaction analyses, and causal gene prioritization. Through benchmarking and application to the UK Biobank, we demonstrate that cellular hypotheses can be tested with rich phenotypic depth, facilitating the translation of genomic associations into mechanistic and clinically actionable insights.

Data availability: UK biobank data is available to any researchers who formally apply for the data. However, the data is not publicly available due to privacy concerns. Reference scRNA-seq datasets are publicly available at the following sites:
frontal cortex: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE144136
adipose tissues: https://singlecell.broadinstitute.org/single_cell/study/SCP1376/a-single-cell-atlas-of-human-and-mouse-white-adipose-tissue?cluster=Human%20WAT&spatialGroups=--&annotation=fat__type--group--study&subsample=100000#study-summary
pancreas: https://hpap.pmacs.upenn.edu/

Code availability: The source codes and R package to reproduce our experiments for this work is available via GitHub: https://github.com/yujias424/CausalCellInfer.

Conflicts of interest

The authors declare no relevant conflicts of interest.

Acknowledgements
This work was supported partially by a Young Collaborative Research Grant (C4003-23Y), Lo Kwee Seong Biomedical Research Fund from The Chinese University of Hong Kong and the KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research of Common Diseases, Kunming Institute of Zoology and The Chinese University of Hong Kong, China. An earlier version of this manuscript was released as a preprint on medRxiv on 17 Oct 2024. (https://doi.org/10.1101/2024.10.17.24315646)

References

1. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology*. 2018;19(1):15.

2. Zheng Y, Lu P, Deng Y, et al. Single-cell transcriptomics reveal immune mechanisms of the onset and progression of IgA nephropathy. *Cell reports*. 2020;33(12).

3. Nomura S. Single-cell genomics to understand disease pathogenesis. *Journal of Human Genetics*. 2021;66(1):75-84.

4. Jagadeesh KA, Dey KK, Montoro DT, et al. Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-sequencing and human genetics. *Nature genetics*. 2022;54(10):1479-1492.

5. Yin L, Feng Y, Shi Y, et al. Estimation of causal effects of genes on complex traits using a Bayesian-network-based framework applied to GWAS data. *Nature Machine Intelligence*. 2024;6(10):1231-1244.

6. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*. 2015;47(9):1091-1098.

7. Yin L, Chau CK, Sham PC, So HC. Integrating clinical data and imputed transcriptome from GWAS to uncover complex disease subtypes: applications in psychiatry and cardiology. *The American Journal of Human Genetics*. 2019;105(6):1193-1212.

8. Peters J, Bühlmann P, Meinshausen N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 2016;78(5):947-1012.

9. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*. 2015;12(5):453-457.

10. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling tumor infiltrating immune cells with CIBERSORT. In: *Cancer Systems Biology: Methods and Protocols*. Springer; 2018:243-259.

11. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics*. 2013;29(8):1083-1085.

12. Tsoucas D, Dong R, Chen H, Zhu Q, Guo G, Yuan GC. Accurate estimation of cell-type composition from gene expression data. *Nature communications*. 2019;10(1):2975.

13. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*. 2014;15(12):550.

14. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*. 2015;43(7):e47-e47.

15. Menden K, Marouf M, Oller S, et al. Deep learning–based cell composition analysis from tissue expression profiles. *Science Advances*. 2020;6(30):eaba2619. doi:10.1126/sciadv.aba2619

16. Chen Y, Wang Y, Chen Y, et al. Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. *Nature Communications*. 2022;13(1):6735.

17. Tran KA, Addala V, Johnston RL, et al. Performance of tumour microenvironment deconvolution methods in breast cancer using single-cell simulated bulk mixtures. *Nature Communications*. 2023;14(1):5758.

18. Thakkar A, Lohiya R. Fusion of statistical importance for feature selection in deep neural network-based intrusion detection system. *Information Fusion*. 2023;90:353-363.

19. Cheng LH, Hsu TC, Lin C. Integrating ensemble systems biology feature selection and bimodal deep neural network for breast cancer prognosis prediction. *Scientific Reports*. 2021;11(1):14914.

20. Chen Z, Pang M, Zhao Z, et al. Feature selection may improve deep neural networks for the bioinformatics problems. *Bioinformatics*. 2020;36(5):1542-1552.

21. Cui P, Athey S. Stable learning establishes some common ground between causal inference and machine learning. *Nature Machine Intelligence*. 2022;4(2):110-115.

22. Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. Invariant risk minimization. *arXiv preprint arXiv:190702893*. Published online 2019.

23. Yin L, Liu M, Shi Y, Qiu J, So H cheong. Direct causal variable discovery leveraging the invariance principle: application in biomedical studies. *medRxiv*. Published online 2024:2024-08.

24. Shah RD, Peters J. The Hardness of Conditional Independence Testing and the Generalised Covariance Measure. *The Annals of Statistics*. 2020;48(3):1514-1538.

25. Wang W, Zhou X, Wang J, et al. Approximate estimation of cell-type resolution transcriptome in bulk tissue through matrix completion. *Briefings in Bioinformatics*. 2023;24(5):bbad273.

26. Wang J, Roeder K, Devlin B. Bayesian estimation of cell type–specific gene expression with prior derived from single-cell data. *Genome research*. 2021;31(10):1807-1818.

27. Song D, Wang Q, Yan G, Liu T, Sun T, Li JJ. scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nature Biotechnology*. 2024;42(2):247-252.

28. Monaco G, Lee B, Xu W, et al. RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Reports*. 2019;26(6):1627-1640.e7. doi:10.1016/j.celrep.2019.01.041

29. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun*. 2019;10(1):380. doi:10.1038/s41467-018-08023-x

30. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573-3587.e29. doi:10.1016/j.cell.2021.04.048

31. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. 2018;36(5):411-420. doi:10.1038/nbt.4096

32. Bühlmann P, Kalisch M, Maathuis MH. Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm. *Biometrika*. 2010;97(2):261-278.

33. Koscielny G, An P, Carvalho-Silva D, et al. Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res*. 2017;45(D1):D985-D994. doi:10.1093/nar/gkw1055

34. Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res*. 2013;41(D1):D793-D800. doi:10.1093/nar/gks1055

35. Clogg CC, Petkova E, Haritou A. Statistical Methods for Comparing Regression Coefficients Between Models. *American Journal of Sociology*. 1995;100(5):1261-1293.

36. Rajkowska G, Miguel-Hidalgo JJ. Gliogenesis and Glial Pathology in Depression. *CNS & Neurological Disorders - Drug Targets*. 6(3):219-233. doi:10.2174/187152707780619326

37. Holter MM, Saikia M, Cummings BP. Alpha-cell paracrine signaling in the regulation of beta-cell insulin secretion. *Front Endocrinol*. 2022;13. doi:10.3389/fendo.2022.934775

38. Saisho Y. β-cell dysfunction: Its critical role in prevention and management of type 2 diabetes. *World Journal of Diabetes*. 2015;6(1):109-124. doi:10.4239/wjd.v6.i1.109

39. Ragvin A, Moro E, Fredman D, et al. Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3. *Proceedings of the National Academy of Sciences*. 2010;107(2):775-780. doi:10.1073/pnas.0911591107

40. Tang W, Zeve D, Suh JM, et al. White Fat Progenitor Cells Reside in the Adipose Vasculature. *Science*. 2008;322(5901):583-586. doi:10.1126/science.1156232

41. Watanabe E, Wada T, Okekawa A, et al. Stromal cell-derived factor 1 (SDF1) attenuates platelet-derived growth factor-B (PDGF-B)-induced vascular remodeling for adipose tissue expansion in obesity. *Angiogenesis*. 2020;23(4):667-684. doi:10.1007/s10456-020-09738-6

42. Zhou B, Zhu Z, Ransom BR, Tong X. Oligodendrocyte lineage cells and depression. *Mol Psychiatry*. 2021;26(1):103-117. doi:10.1038/s41380-020-00930-0

43. Ellulu MS, Patimah I, Khaza'ai H, Rahmat A, Abed Y. Obesity and inflammation: the linking mechanism and the complications. *Arch Med Sci*. 2017;13(4):851-863. doi:10.5114/aoms.2016.58928

44. Hägglöf T, Vanz C, Kumagai A, et al. T-bet+ B cells accumulate in adipose tissue and exacerbate metabolic disorder during obesity. *Cell Metabolism*. 2022;34(8):1121-1136.e6. doi:10.1016/j.cmet.2022.07.002

45. Bundo M, Ueda J, Nakachi Y, Kasai K, Kato T, Iwamoto K. Decreased DNA methylation at promoters and gene-specific neuronal hypermethylation in the prefrontal cortex of patients with bipolar disorder. *Mol Psychiatry*. 2021;26(7):3407-3418. doi:10.1038/s41380-021-01079-0

46. Liao Z, Wang Y, Qi X, Xiao X. JAZF1, a relevant metabolic regulator in type 2 diabetes. *Diabetes Metabolism Res*. 2019;35(5):e3148. doi:10.1002/dmrr.3148

47. Lee HY, Jang HR, Li H, et al. Deletion of Jazf1 gene causes early growth retardation and insulin resistance in mice. *Proceedings of the National Academy of Sciences*. 2022;119(49):e2213628119. doi:10.1073/pnas.2213628119

48. Hrovatin K, Bastidas-Ponce A, Bakhti M, et al. Delineating mouse β-cell identity during lifetime and in diabetes with a single cell atlas. *Nat Metab*. 2023;5(9):1615-1637. doi:10.1038/s42255-023-00876-x

49. Mizusawa N, Harada N, Iwata T, Ohigashi I, Itakura M, Yoshimoto K. Identification of protease serine S1 family member 53 as a mitochondrial protein in murine islet beta cells. *Islets*. 2022;14(1):1-13. doi:10.1080/19382014.2021.1982325

50. Muffels IJJ, Carter T, Rehmann H, et al. LIMK1 variants are associated with divergent endocrinological phenotypes and altered exocytosis dynamics. *iScience*. 2025;28(6):112585. doi:10.1016/j.isci.2025.112585

51. Jahangir E, De Schutter A, Lavie CJ. The relationship between obesity and coronary artery disease. *Translational Research*. 2014;164(4):336-344. doi:10.1016/j.trsl.2014.03.010

52. Sha L, Li J, Shen H, et al. LHPP-mediated inorganic pyrophosphate hydrolysis-driven lysosomal acidification in astrocytes regulates adult neurogenesis. *Cell Reports*. 2023;42(8):112975. doi:10.1016/j.celrep.2023.112975

53. Lu X, Xie Q, Pan X, et al. Type 2 diabetes mellitus in adults: pathogenesis, prevention and therapy. *Sig Transduct Target Ther*. 2024;9(1):262. doi:10.1038/s41392-024-01951-9

54. Kosten TA, Galloway MP, Duman RS, Russell DS, D'Sa C. Repeated Unpredictable Stress and Antidepressants Differentially Regulate Expression of the Bcl-2 Family of Apoptotic Genes in Rat Cortical, Hippocampal, and Limbic Brain Structures. *Neuropsychopharmacol*. 2008;33(7):1545-1558. doi:10.1038/sj.npp.1301527

55. Fries GR, Saldana VA, Finnstein J, Rein T. Molecular pathways of major depressive disorder converge on the synapse. *Mol Psychiatry*. 2023;28(1):284-297. doi:10.1038/s41380-022-01806-1

56. Gulbins A, Schumacher F, Becker KA, et al. Antidepressants act by inducing autophagy controlled by sphingomyelin–ceramide. *Mol Psychiatry*. 2018;23(12):2324-2346. doi:10.1038/s41380-018-0090-9

57. Volpe CMO, Villar-Delfino PH, dos Anjos PMF, Nogueira-Machado JA. Cellular death, reactive oxygen species (ROS) and diabetic complications. *Cell Death Dis*. 2018;9(2):119. doi:10.1038/s41419-017-0135-z

58. Lee SC, Pervaiz S. Apoptosis in the pathophysiology of diabetes mellitus. *The International Journal of Biochemistry & Cell Biology*. 2007;39(3):497-504. doi:10.1016/j.biocel.2006.09.007

59. Jia P, Hu R, Yan F, Dai Y, Zhao Z. scGWAS: landscape of trait-cell type associations by integrating single-cell transcriptomics-wide and genome-wide association studies. *Genome Biology*. 2022;23(1):220. doi:10.1186/s13059-022-02785-w

60. Zhou Y, Adeluwa T, Zhu L, et al. scPrediXcan integrates deep learning methods and single-cell data into a cell-type-specific transcriptome-wide association study framework. *Cell Genomics*. 2025;5(5):100875. doi:10.1016/j.xgen.2025.100875

61. Yin L, Shi Y, Zhang R, et al. Estimating cell compositions and cell-type-specific expressions from GWAS data using invariant causal prediction, deep learning and regularized matrix completion. *medRxiv*. Published online January 1, 2024:2024.10.17.24315646. doi:10.1101/2024.10.17.24315646