Estimating cell compositions and cell-type-specific expressions from GWAS data using invariant causal prediction, deep learning and regularized matrix completion

Liangying Yin[1#], Yujia Shi[1#], Ruoyu Zhang[1], Yong Xiang[1], Jinghong Qiu[1], Pak-Chung Sham[2], Hon-Cheong So[1,3-8,*]

[1]School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong
[2]Department of Psychiatry, The University of Hong Kong, Hong Kong
[3]KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research of Common Diseases, Kunming Institute of Zoology and The Chinese University of Hong Kong, China
[4]Department of Psychiatry, The Chinese University of Hong Kong, Hong Kong
[5]CUHK Shenzhen Research Institute, Shenzhen, China
[6]Margaret K.L. Cheung Research Centre for Management of Parkinsonism, The Chinese University of Hong Kong, Shatin, Hong Kong
[7]Brain and Mind Institute, The Chinese University of Hong Kong, Hong Kong SAR, China
[8]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

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

Abstract

Dissecting large bulk RNA-seq data into cell proportions and cell type specific expression profiles could significantly enhance our understanding of disease mechanisms at cell level and facilitate the identification of novel drug targets and development of efficient intervention strategies.

In this study, we presented a scRNA-seq marker (gene) informed cell deconvolution and expression inference (CausalCellInfer) framework. CausalCellInfer employed causal inference principles to automatically identify a small set of critical marker genes from the reference scRNA-seq dataset. It then integrates deep neural networks with regularized matrix completion algorithms to deconvolute cell proportions and estimate cell type specific (CTS) expression profiles. Most importantly, we pioneered the application of the proposed framework to imputed expression data from large-scale genome-wide association studies (GWAS).

We verified the efficacy of our proposed method by comparing it against existing state-of-the-art cell deconvolution methods, including CIBERSORTx, DWLS, Scaden, and

MuSic across various real and pseudo-bulk samples. Furthermore, we proposed the use of a wide range of enrichment analyses to demonstrate the reliability of CausalCellInfer in estimating CTS profiles. Our proposed framework consistently outperformed existing methods with significantly higher concordance correlation coefficient (CCC), lower mean absolute error (MAE) and root mean square error (RMSE) across all testing real and pseudo bulk samples. Importantly, it also demonstrated superior computational efficiency compared to all benchmarked methods except MuSic. We also applied our trained models to deconvolute 4 tissue-specific expression cell proportions and estimate the corresponding CTS expression profiles, leveraging UK Biobank data (UKBB). We conducted a series of cellular-level analyses, including cell proportion association analysis, causal gene detection, transcriptome wide association analysis (TWAS), for 24 phenotypes in UKBB based on estimated cell composition and CTS expression profiles. Of note, the estimated proportions of various cell types were indicative of disease onset. For example, T2DM patients demonstrated a significant decrease in the proportions of alpha and beta cells in comparison to controls. We also tested their associations. Our method exhibited satisfactory positive predictive values (PPV) in uncovering differentially expressed (DE) genes for the majority of cell types. Encouragingly, most identified CTS causally relevant genes were found to be significantly enriched in target diseases or related pathophysiology.

In conclusion, we presented a novel framework for inferring cell-type proportions and CTS expression, with novel applications to GWAS-imputed expression data from large-scale biobank program. Our work also shed light on how differential cell-type proportion and CTS expression may be associated with susceptibility to different diseases and their prognoses, bridging scRNA-seq and clinical phenotypes in large-scale biobank studies.

Introduction

Last decade has witnessed the great success of single-cell RNA-sequencing (scRNA-seq) in revealing diverse biological mechanisms and gene interactions across various conditions or diseases [1-4]. scRNA-seq allows us to analyze gene expression at the single-cell level. Despite its potential to offer valuable insights into disease mechanisms, the large-scale application and extensive study of scRNA-seq are generally not feasible due to its technical complexity and high cost. Moreover, scRNA-seq data are typically collected for a specific disease or purpose; additional relevant outcomes or features, such as detailed clinical data on symptoms or prognosis, and other comorbidities, are often unavailable. The sample size at the subject level is usually very limited for scRNA-seq studies, which greatly restricts our ability to fully harness the potential of scRNA-seq data to gain insights into disease prevention and intervention.

Bulk RNA-seq data, unlike scRNA-seq data, is usually collected from larger cohorts, and repositories such as TCGA and GTEx offer a wealth of bulk resources derived from large cohorts. It is important to note that bulk expression data is often gathered with specific objectives in mind, resulting in the availability of only a limited set of clinical variables for

the individuals providing the data. Concurrently, there's a growing interest in predicting bulk expression data from GWAS data [5-7]. The advent of large-scale biobanks, including the UK Biobank (UKBB), provides easy access to comprehensive GWAS data, together with a wide array of clinical variables.

Can we combine the best of the two worlds, namely scRNA-seq and large biobanks (or other large-scale GWAS datasets)? If we can leverage the latter to estimate cell type proportions and infer cell type-specific expression profiles, we have the potential to unlock novel insights into the pathophysiology underlying complex traits and diseases at the resolution of specific cell types, while simultaneously enjoying the advantages of large samples and comprehensive clinical or biological phenotypes available in biobanks. Bridging the gap between cell composition, cell-specific expression profiles, and clinical phenotypes could significantly enhance our understanding of disease mechanisms at the cellular level. This may facilitate the identification of new drug targets and development of novel therapeutic strategies.

A framework for cell deconvolution and expression inference (CausalCellInfer) based on the invariance principle

In this study, we present an innovative scRNA-seq marker (gene) informed cell deconvolution and expression inference (CausalCellInfer) framework. This approach is capable of estimating cellular compositions and their corresponding cell type-specific expression profiles from bulk expression data. We propose a novel causality-inspired data-driven marker-gene-informed cell deconvolution method. The proposed method is widely applicable to any bulk transcriptome data, however in this study we focused on the application to genotype-predicted expression datasets in large cohorts.

The method draws inspiration from the concept of invariant causal prediction (ICP), as proposed by Peters et al. [8]. Briefly, ICP is a method used in causal inference that exploits the invariance of causal relations across different environments or conditions. It operates under the invariant assumption of causal relations to uncover true causal predictors.

There are several advantages of the proposed framework. Firstly, as our proposed method does not require prior knowledge of cell markers (the markers are derived from the data with machine learning (ML) and causal-inference based algorithms), it offers considerable flexibility and can potentially be applied to deconvolute any GWAS-predicted bulk samples with available referencing scRNA-seq data. This greatly improved the flexibility and applicability of our proposed method. Secondly, the proposed framework is highly computationally efficient, particularly for large-scale datasets, as it only uses a small set of target-defining markers for cell deconvolution. Thirdly, the use of the invariance principle in defining cell-type markers for deconvolution makes our methods less susceptible to the influence by confounders. Importantly, it also enhances the generalizability of our method in deconvolution of new or external large-scale bulk samples. Briefly, the causal genes markers that differentiate various cell-types are likely invariant across different environments, conditions or interventions. On the other hand, the expression of some marker

genes might be *spuriously* associated with cell-type, due to confounding only. These genes may still be regarded as 'true' cell-types markers by other methods, yet since they are not causal, their ability to deconvolve cell-types in a new sample (i.e., the generalizability) will be limited.

Application to GWAS-imputed expression data and associated advantages

The proposed framework integrates causal inference principles with deep neural networks and regularized matrix completion. This combination allows us to efficiently estimate cell proportions and cell-type specific (CTS) expression profiles from bulk transcriptome data, including genotype-imputed bulk data in large biobanks. We thereby propose a variety of novel analyses, including cell-type specific transcriptome-wide association study (CTS-TWAS), CTS-TWAS with causal gene identification, association and interaction studies of cell-type proportions, as well as conditional TWAS accounting for cell-type proportions which is more immune to risk of confounding. In essence, the proposed framework opens up almost unlimited opportunities that enables cell-type level analyses to be performed on individual-level genotype data. For example, essentially any regression models, machine learning approaches, interaction models and conditional analysis on cell-type proportions and CTS expression may be performed.

There are several other advantages of applying the method to imputed transcriptome data. Firstly, since the cell proportions are genetically imputed, when compared to direct single-cell studies, the association analysis results are unlikely to be affected by other confounders (e.g., medication usage) and reverse causality (significant findings suggest cell proportion/CTS expression changes as a *cause* of the disease in our analyses). On the other hand, in conventional single-cell studies, it is often unclear whether differences in cell-type proportion or CTS expression represent a cause or a consequence of the disease. Secondly, for large biobanks such as UKBB, a broad range of clinical phenotypes are available with large sample sizes. Our framework enables us to harness the potential of these variables to gain insights into disease prevention and intervention. Moreover, the use of large biobanks and the proposed framework allow for the control of additional covariates/confounders when conducting single-cell level analyses.

Comparison to previous works

In this section, we review relevant literature on methods for deconvoluting bulk RNA-seq data, which can be broadly categorized into two types: signature-based and signature-free approaches. Traditional signature-based approaches, such as CIBERSORT[9], CIBERSORTx[10], DeconRNASeq[11], and dampened weighted least squares (DWLS)[12], are designed to deconvolute bulk data based on pre-defined reference gene expression profiles (GEPs) derived from a reference scRNA-seq dataset. These GEPs encompass the expression profiles of marker genes for all cell types involved in a specific tissue. Marker genes can be derived either from domain knowledge (e.g., CIBERSORT) or existing marker gene detection methods (e.g., DESeq2[13], Limma[14]). Detecting marker genes and their corresponding GEPs from reference scRNA-seq datasets is typically computationally intensive. Consider the DWLS method as an example. Initially, it employs the hurdle model in the MAST R package

to pinpoint the marker genes. Subsequently, a signature matrix is constructed using the genes identified by MAST [12]. These two steps are notably time-intensive, especially when dealing with large scRNA-seq data sets. Despite the small size of the reference GEPs, the application of signature-based methods in deconvoluting large-scale bulk data, such as the UKBB, can be time-consuming and resource-demanding.

Also, it is important to note that each marker gene is assumed to have a constant expression in each cell type. This assumption is often violated in actual cell expression profiles, where subject-level heterogeneity and cell-level stochasticity are common in both reference and target samples. As a result, signature-based cell deconvolution methods are susceptible to the influence of confounders such as batch effects and individual-level heterogeneity. The presence of batch effects may also lead to the identification of inaccurate marker genes [15]. In light of this, extra caution is necessary when inferring reference GEPs from large-scale scRNA-seq data, particularly those collected from highly heterogeneous individuals or different sequencing technologies. Another issue is that only a subset of all genes (maximally ~7,000) can be reliably predicted from genotypes, using established tools such as PrediXcan. The information loss is more severe in some cases, in which only around 20% of genes will be preserved by PrediXcan. This feature makes it extremely challenging for existing signature-based methods to find valid reference GEPs for the target bulk samples.

Our method alleviates the above issues by using a causality-inspired, data-driven marker (gene)-informed and deep learning-based cell deconvolution method. It can automatically and efficiently identify a small set of target-defining markers from the reference scRNA-seq data that are also present in genotype-predicted bulk data. Furthermore, our method is highly computationally efficient, particularly for large-scale datasets, as it only uses a small set of target-defining markers for cell deconvolution. The signature-free nature makes our method immune to the influence of batch effects and individual level heterogeneity.

Signature-free approaches, such as Scaden [16] and TAPE [17], do not depend on reference GEPs for cell deconvolution. Instead, they directly make use of the reference scRNA-seq data to estimate the cell composition of target bulk samples. More specifically, they identify important genes in a 'data-driven' manner and then utilize the cell expression profiles for these genes to construct a cell deconvolution model. With the rapid development of deep learning techniques, there is an increasing interest in exploiting deep learning frameworks for cell composition estimation. Deep learning methods do not rely on the constant marker gene expression assumption to deliver unbiased estimation. Notably, they generally achieve better performance than the traditional signature-based methods [18].

However, there are several limitations to the deep learning approach. First, training a deep learning-based cell deconvolution model is highly time-consuming and memory-demanding. The predicament exacerbates when the sample and feature size of the training set are considerably large, a scenario frequently encountered in biomedical research. Even though many deep learning methods performed feature selection prior to model training, they

tend to retain a substantial number of features (~10,000 features were preserved for both Scaden and TAPE after feature selection). Secondly, high-dimensional features may also introduce unwanted noise to the model, potentially leading to reduced deconvolution accuracy[19,20]. Efficient feature selection for large-scale datasets has the potential to dramatically improve both the accuracy and computational efficiency of deep learning-based prediction models [21]. Thirdly, deep learning prediction models built from a small set of target-defining features may be more generalizable than those built from a larger or full set of features, due to lower risk of overfitting [22].

Deep learning methods performed extremely well in extracting complex prediction rules from data. However biomedical data is frequently subject to the influence of selection bias and confounders. Yet, spurious correlations stemming from the training data are irrelevant to the causal explanation of the prediction target, making them hard to generalize to new test data distributions. We alleviate these issues by employing causal principles to inform the marker gene identification. By identifying the true defining variables for the target of interest (invariant correlations), we can reduce risks of confounding[23]. Also, it could significantly improve the computational efficiency of deep learning-based cell deconvolution method while exhibiting superior performance and greater generalizability.

To the best of our knowledge, we are pioneers in incorporating causal inference principles into a deep neural network-based cell deconvolution framework to estimate cell composition and expression profiles from large-scale genotype-imputed bulk data. A key feature of our framework is its ability to *preselect a small set of target-defining genes based on invariant causal prediction principles for model training.* This feature enhances the efficiency and generalizability of our method in estimating cell compositions and cell type specific profiles from external bulk samples. Additionally, it provides us with the capacity to investigate the associations between the genotype-predicted cell compositions and disease onset and prognoses. It also facilitates the identification of cell-type-specific genes relevant to disease susceptibility and outcomes. In essence, our analysis establishes connections among the inherited genomic information, cell-type composition, cell-type specific expression profiles and disease outcomes. This extrapolates the utility of large-scale Biobank data in comprehending disease mechanisms and prognoses.

Summary of contributions

To summarize, our contributions include the following:

1) Proposal of a novel framework for cell-type deconvolution using deep learning, leveraging the principles of invariant causal prediction (ICP). The method performed favorably when compared to other competing approaches. Briefly, the proposed framework has better computational efficiency, showed better predictive performance, and has better generalization ability (especially when there is a shift in environment) as the ICP principle was employed. The framework is generally applicable to any bulk expression data.

2) We pioneered cell-type deconvolution and cell-type specific expression analysis in GWAS-imputed expression data, and propose a variety of novel analytic approaches:
   - Cell-type specific transcriptome-wide association study (CTS-TWAS) to detect *cell-type specific expression changes* related to traits/diseases of interest
   - CTS-TWAS integrated with the PC-simple algorithm (CTS-TWAS-causal) to prioritize *direct causal* genes
   - Cell-type proportion association study to identify which cell-types may contribute to traits/diseases
   - Cell-type proportion interaction analysis to decipher interactions between cell-types in affecting traits/diseases, with cross-tissue cell-type interactions also explored
   - Conditional TWAS controlling for cell-type proportions, hence reducing risk of confounding when identifying differentially expressed genes (DEGs)

The approaches above can leverage large-scale biobanks which enable the study of comprehensive clinical phenotypes in very large samples, and allow other covariates to be adjusted for, as well as interaction or ML analysis that requires large individual-level data. In essence, our work bridges the gap between GWAS and single-cell studies.

3) We applied our framework to the UK-Biobank on a wide range of clinical phenotypes and outcomes, showcasing the power of our approach to gain cell-type-level biological insights in large datasets with rich clinical data. We also provide further validation of our approach by showing significant overlap of cell-type specific DEGs inferred from UKBB with those identified from real single-cell transcriptome studies.

Method

In this study, we introduce an integrated framework, CausalCellInfer, that synergistically combines XGBoost and scRNA-assisted deep neural networks to estimate cell proportions and expression profiles from bulk expression data across large cohorts. Our proposed framework is a three-step process encompassing cell marker identification, cell decomposition, and cell expression estimation. Subsequent sections will provide a detailed description of each step.

Cell marker identification

Cell markers, by definition, serve as a means to characterize distinct cell types, irrespective of their specific conditions such as disease status, age group, gender, and so forth. This leads us to hypothesize that ('causal') cell markers exhibit stability and robustness across various environments. In this context, 'environments' refer to any features related to the subjects or the background of the experiment/study. For example, the study could be performed across different disease status, gender, age, or across different locations, time-points, batches, etc. Sometimes the study may be performed with different interventions, such as knockdown/over-expression of different genes. All the above may serve as indicators of 'environments'.

Inspired by the principles of invariant causal prediction (ICP) [8,24], we propose the use of ***invariant*** XGBoost (**I**-XGBoost) for the identification of cell markers in large-scale single-cell RNA sequencing (scRNA-seq) data. XGBoost [25], a supervised machine learning method, is composed of a sequence of simpler, weaker decision trees. It operates by iteratively adding new trees, each one designed to correct the errors made by its predecessors, until the model achieves an optimal fit to the data. The complexity of the trees is penalized to prevent overfitting and enhance the model's generalizability. By integrating the invariance principle of causal inference into XGBoost, we aim to improve the validity and robustness of the identification of cell markers. We wish to highlight that other machine learning methods can also be used.

Initially, we partitioned the scRNA-seq data into 2 subsets (i.e., training and testing set). Then, we partitioned the scRNA-seq data in the training set into distinct subsets based on the selected environmental variable. In this study, we used disease status and gender as environment variable for different scenarios. For each subset, we independently applied XGBoost to identify cell markers unique to each cell type using all available features (genes) in the dataset. In this study, we employed the xbg.train() and xbg.importance() functions in R package "xgboost" to train the prediction model function and to extract features that contribute to the trained model(binary cell type label as outcome variable). The maximum depth of each tree and the learning rate were respectively set to 2 and 1. We extracted the model-contributing features (i.e., xgb.importance>0) from each trained environment-specific model. The invariant contributing genes across all environments (prediction models) were regarded as the definitive cell markers for each cell type.

In some cases, there may be too few invariant cell markers identified for cell-type deconvolution; in such scenarios, we suggest the use of Combined XGBoost (C-XGBoost) to integrate cell markers from all environments by taking the union of cell markers. Empirically, we advise assessing the overlap rate of cell markers between cell types prior to determining the strategy for selecting definitive cell markers. When the overlap rate is high, in general more genes may be needed to differentiate the potentially closely related cell types[26]. Conversely, if the overlap is low, we suggest using the 'invariant' cell markers across all environments.

Cell deconvolution

In our framework, we adaptively modified Scaden to estimate cell proportions. Scaden is a deconvolutional deep neural network (DNN) built by simulated gene counts bulk from reference scRNA-seq to estimate the cell proportions of bulk expression samples. It selects ~ 10,000 informative genes (expression variance >=0.1) to train the deconvolutional DNN model. We replaced this feature selection step with our proposed I-XGBoost approach, leading to a much smaller but more informative (and more likely causal) gene-set for model training and inference. The details of our proposed framework are described below.

First, we used the MinMaxScaler() class from the Sklearn preprocessing module to scale the input data to the value 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 perform better than the individual model[16]. Each DNN model attaches a SoftMax layer as the output layer to predict cell fractions. The $L1$ loss was selected as the optimization objective:

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

where y_i is the vector of ground-truth fractions of the sample i , $\widehat{y_i}$ is the vector of predicted fractions of the sample i . We used the average of the three ensemble models' prediction as the ultimate cell-type composition, denoted as

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

To efficiently and robustly train the DNN models, we simulated 5,000 pseudo-bulk samples with random fractions of different cell types $(f_c \epsilon [0,1]\ for\ \forall c \epsilon \{1,2,..C\},\ \sum_1^C f_c = 1)$ by subsampling cells from the reference RNA-seq data to train our deconvolutional DNN following the pseudo-bulk data generation approach proposed by Chen et al. [17]. Only genes selected by our proposed I-XGBoost approach were used during the training stage. 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 the cell type c We generated the final pseudo-bulk expression profile by summing up the expression value of the randomly selected scGEPs for each sample. The deconvolutional DNN model was built using the Pytorch library. Prediction models were constructed for cell proportions based on target bulk expression data using the trained ensemble DNN.

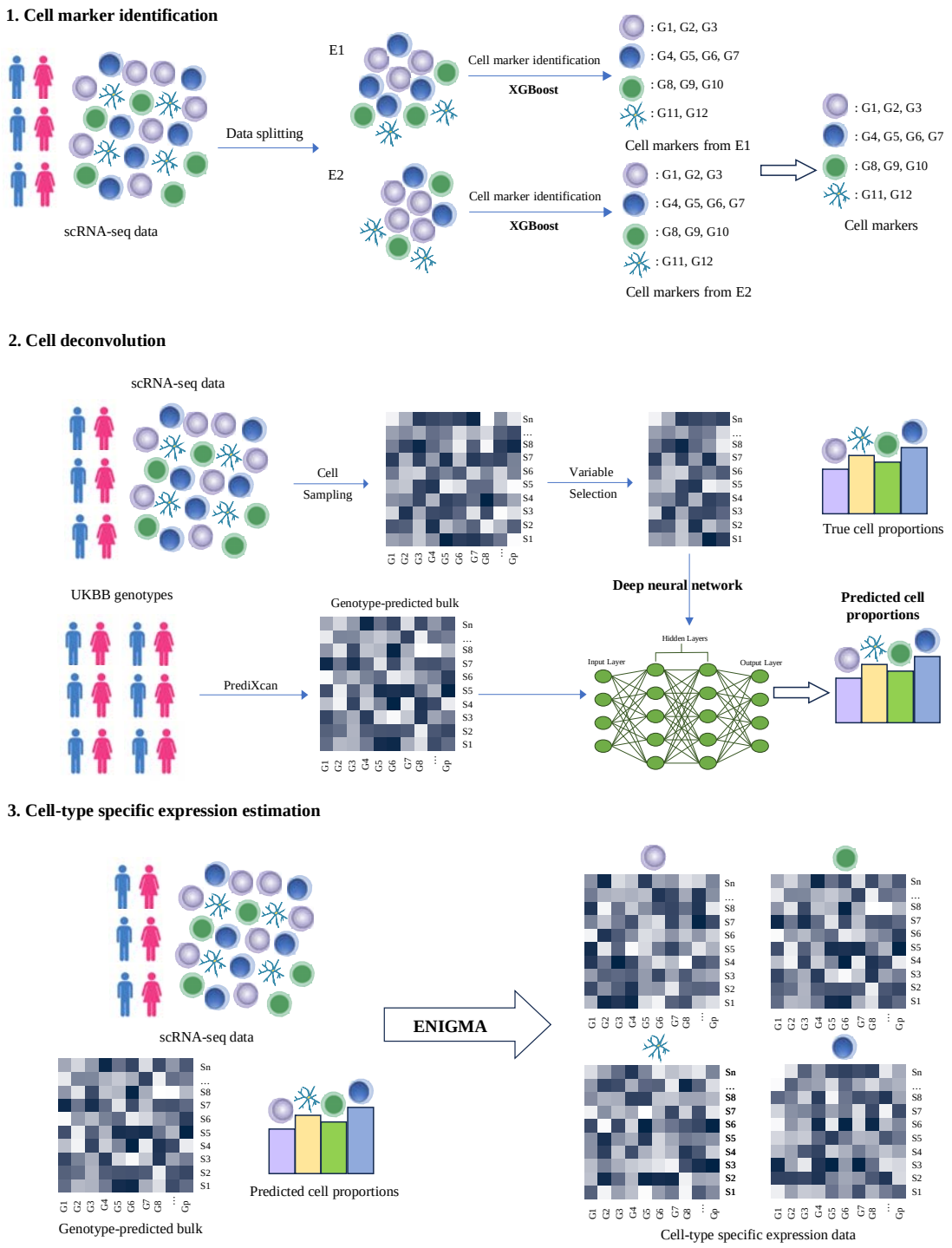

Figure 1 Workflow of our proposed framework for cell decomposition and cell type-specific expression estimation. 1. Cell marker identification: We proposed using invariant XGBoost(I-XGBoost) to find cell markers for each cell type from the reference RNA-seq data; 2. Cell deconvolution: We modified Scaden, 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.

Cell expression estimation

We employed a regularized matrix completion-based approach called ENIGMA [27] to estimate cell-type specific expression profiles based on predicted cell proportions and genotype-imputed bulk expression profiles. ENIGMA is developed based on two key assumptions. The first assumption is that the expectation of each cell type-specific (CTS) expression profile aligns with the reference profile matrix learned from the external scRNA-seq data[28]. The second assumption is that observation noise is consistent across different samples. Suppose these two assumptions are satisfied, then we can estimate the CSE by the trace norm model, following [27]:

$$(\widehat{X_i})_{i=1,2,..C} = argmin(\frac{1}{2}\alpha\left\|O - \sum_{i=1}^{C} X_i diag(\theta_i)\right\|_F^2 + \frac{1}{2}(1-\alpha)\sum_{i=1}^{C}\left\|X_i a - R_i\right\|_2^2 + \beta\sum_{i=1}^{C}\widehat{\theta_i}\left\|X_i\right\|_*) \quad (3)$$

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 ith cell type, θ_i indicates the cell proportion of ith cell type, R_i indicates the reference expression vector for the ith cell type, $\alpha\epsilon(0,1)$ denotes the weights of the aggregated CTS expression profiles. $1-\alpha$ denotes the agreement weight between aggregated CTS expression profiles and its corresponding reference expression vector, $\beta \in (0,+\infty)$ indicates 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:

$$(\widehat{X_i})_{i=1,2,..C} = (\frac{1}{2}\alpha\left\|O - \sum_{i=1}^{C} X_i diag(\theta_i)\right\|_F^2 + \frac{1}{2}(1-\alpha)\sum_{i=1}^{C}\left\|X_i a - R_i\right\|_2^2 + \beta\sum_{i=1}^{C}\widehat{\theta_i}\left\|t(X_i)\right\|_{2,max}^2) \quad (4)$$

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

As our bulk RNA-seq data was imputed using PrediXcan, we standardized our reference scRNA-seq data to ensure that both datasets were on the same scale. In this way, we can derive the reference profile matrix for each cell type from the reference scRNA-seq data by averaging the expression profile of corresponding cells. The inferred cell type proportion matrix(θ), reference profile matrix(R), and the bulk RNA-seq data(O) were further plugged into formula 2 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 advised to utilize those inferred from the cell type in the reference scRNA-seq dataset with a total cell count exceeding 2000.

Evaluation of the proposed I-XGBoost approach

To validate the advantages of the I-XGBoost method which employs the invariance principle, we initially established a baseline simulation by comparing the I-XGBoost with analogous algorithms (C-XGBoost, XGBoost) using artificially generated scRNA-seq datasets confounded by certain environment variables. We wish to test whether I-XGBoost is able to identify cell-type specific markers more accurately than other methods, and predict cell proportions better, especially when there is a change in the environment.

Specifically, we utilized scDesign3 [29] to create scRNA-seq data based on a publicly available 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 for the manipulation of interpretable parameters, enabling the introduction of environmental effects (leading to confounding) and the subsequent simulation of new scRNA-seq and bulk data impacted by these effects. We carried out simulations on bulk samples consisting of 4 cell types with varied confounding effects respectively. For more details, please refer to the supplementary text.

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

To verify the accuracy and reliability of our proposed framework, we initially compared the performance of our method with existing deconvolution approaches across various evaluation metrics, including root mean square error (RMSE), mean absolute error (MAE), and Lin's Concordance Correlation Coefficient (CCC) on pseudo-bulk expression profiles generated from reference scRNA-seq data of four different tissues, i.e., the frontal cortex, pancreas islet, and subcutaneous and visceral adipose tissues. The pseudo-bulk expression profiles were simulated in silico from scRNA-seq data with randomly generated cell type proportions, essentially making them a summation of numerous single-cell expression profiles [17]. We added noise to each generated pseudo-bulk expression profile to mimic the measurement errors in real RNA-seq data [17]. The noise was randomly generated from a Gaussian distribution, which was determined by the cells used to generate the corresponding pseudo-bulk profile. In this study, we only included genes that could be reliably predicted from genotypes by PrediXcan for model training.

Additionally, we conducted benchmarking on real datasets with available bulk PBMC RNA-seq and their corresponding cell proportions, assessed with flow cytometry from the study by Monaco et al. [30]. To generate pseudo-bulk datasets for data training and to perform I-XGBoost to identify cell markers, we merged three peripheral blood mononuclear cells (PBMC) scRNA-seq datasets obtained from different individuals. These datasets, namely PBMC6K, PBMC8k, and PBMC10k, 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. Initially, we trained our deep neural network model using the training set. Subsequently, we tested our trained model using the pseudo-bulk samples generated from the testing set. The training set was also utilized to obtain references for MuSic, DWLS, and CIBERSORTx.

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-specific expression profiles were imputed from genotypes.

Specifically, we identified differentially expressed (DE) genes from both actual scRNA-seq data and genotype-predicted cell-specific 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 that are differentially expressed between two cell groups. For our genotype-predicted cell expression profiles, we applied the same test to detect DE genes from the estimated CTS expression profiles. 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 'gold standard'.

Identifying cell types associated with disease susceptibility

We conducted additional analysis to investigate how different cell types may affect the risks of developing different diseases. More precisely, we analyzed the associations between estimated cell proportions and disease status by regression. To adjust for multiple testing, we employed the Benjamini-Hochberg procedure to control the false discovery rate (FDR) at 0.05. Besides, we examined whether interactions between cell types, including cell types *across* different tissues, could affect disease phenotypes using regression models.

Distinguishing cell type-specific genes involved in disease susceptibility

Upon acquiring the CTS expression profiles for each target disease/trait from UKBB, we identified the genes that contribute to disease susceptibility for each cell type. This may be considered as a cell type-specific TWAS (CTS-TWAS). To mitigate potential confounding effects arising from population stratification, we adjusted the estimated cell type-specific expression profiles and target trait/disease using the top 10 principal components (PCs) derived from the corresponding genotype dataset.

Subsequently, we utilized the PC-simple algorithm to identify the potential direct causal genes for the outcome variable. In essence, the PC-simple algorithm uses ordered correlation screening (repeated conditional association tests) to deduce the causal relationships between covariates and a target of interest. 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 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 CSE 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). As an exploratory analysis, we concentrated on key prognostic variables of clinical significance for these diseases.

Specifically, we focused on treatment resistance status and depression-related hospitalization frequency for patients with depression (tissue studied: frontal cortex); T2DM mortality and T2DM-related hospitalization frequency for T2DM patients (tissue studied: pancreas); coronary artery disease (CAD) mortality and CAD-related hospitalization frequency for obesity patients (tissue studied: subcutaneous and visceral adipose). Please refer to the supplementary text for more details.

Moreover, we conducted exploratory analyses for 12 other related disorders/phenotypes, i.e., 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). Here we assumed the models built on a certain population/disease group also (approximately) applies to other related disorders. Again, we adjusted the CSE profiles and target prognosis variables by the top 10 PCs to mitigate potential bias from population structure. We then employed PC-simple to identify potentially direct causal genes contributing to specific disease outcomes.

OpenTargets enrichment analysis

We further validated our methodology by assessing whether our identified gene set (associated with disease susceptibility) was significantly enriched among disease genes listed in an independent, established database. We leveraged the OpenTargets database [33] in this study. This database provides an overall association score for each 'target' gene, aggregating seven different types of evidence: genetic associations, somatic mutations, known drugs, affected pathways, RNA expression, literature support, and animal models. Our enrichment analysis was conducted at the disease level; we combined all identified cell type-specific genes together for each disease.

Further enrichment analysis

We conducted a pathway enrichment analysis on the gene sets deemed causal, utilizing the "ConsensusPathDB" database [34]. More specifically, we performed an over-representation analysis on the causal gene set identified, employing the hypergeometric test.

Conditional TWAS (controlling for cell type proportions) and cell type-specific TWAS (CTS-TWAS)

We firstly 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. Note that cell proportion is a potential confounder that may lead to spurious associations between imputed expression and outcomes.

We performed two set of univariate tests for each gene using linear regression by adjusting for different variables. Specifically, we respectively adjusted for the top 10 PCs and the 10 PCs plus the estimated cell proportions. Then, we compared the coefficients derived from these two models using the method recommended by Clogg et al.[35].

Secondly, with the inferred CTS file, we performed CTS-TWAS analysis for each target trait. We used the Benjamin-Hochberg procedure to control the false discovery rate (FDR; FDR<0.05 considered significant) in each TWAS analysis.

Results

Overview

In this study, we introduced a framework, CausalCellInfer, which integrates I-XGBoost with a deep neural network to deconvolute cell proportions. This framework also employs a regularized matrix completion algorithm (ENIGMA) to estimate the corresponding cell type-specific expression (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. Firstly, CausalCellInfer mainly employs I-XGBoost/C-XGBoost to identify markers (genes) for each cell type. It then uses these identified markers to train the deep neural network-based cell deconvolution model. Subsequently, it employs a regularized matrix completion algorithm, ENIGMA, to infer the corresponding CTS 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 CIBERSORTx, DWLS, Scaden, and MuSic 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). With the deconvoluted the cell compositions corresponding CTS profiles, we conducted a wide range of analyses for 29 clinically-relevant traits (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 evaluation of the proposed I-XGBoost approach

We evaluated the performance of our proposed I-XGBoost in deconvoluting cell compositions, with C-XGBoost and XGBoost using the simulated scRNA-seq and bulk data. Even though I-XGBoost utilizes the smallest marker gene sets, it outperformed all competitors in terms of Concordance Correlation Coefficient (CCC), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) in scenarios involving 4 cell types (Figure 2).

As the extent of confounding effects increased, the performance improvement of our method became increasingly pronounced. Of note, I-XGBoost demonstrates superior performance in identifying highly-ranked differentially expressed genes (DEGs) than its competitors. These DEGs were ranked based on the mean difference across cell types in decreasing order.

In essence, marker gene set identified from I-XGBoost exhibited a higher PPV in uncovering DEGs compared to those identified by C-XGBoost and XGBoost (Table 1).

Table 1 Comparison of our method and the competitors in identifying DEG genes

Scenario	**Positive predictive value(PPV)**		
	Our method	C-XGBoost	XGBoost
4 cell types with modest confounding effects	0.816	0.736	0.791
4 cell types with comparably significant confounding effects	0.9	0.803	0.842

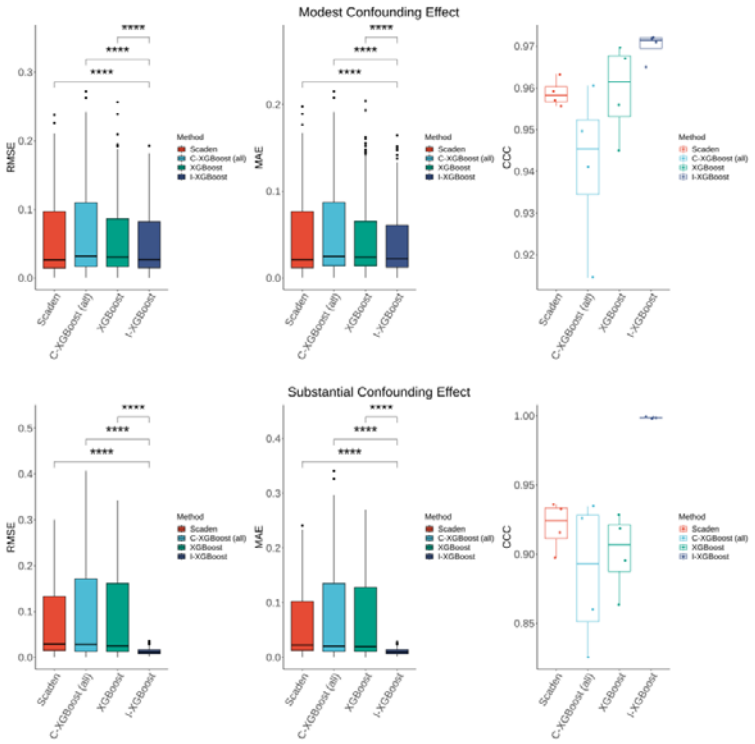

Figure 2 Benchmarking of cell deconvolution accuracy on simulated bulk samples comprising 4 cell types with varied confounding effects.

Performance evaluation of the proposed approach on real and pseudo-bulk data

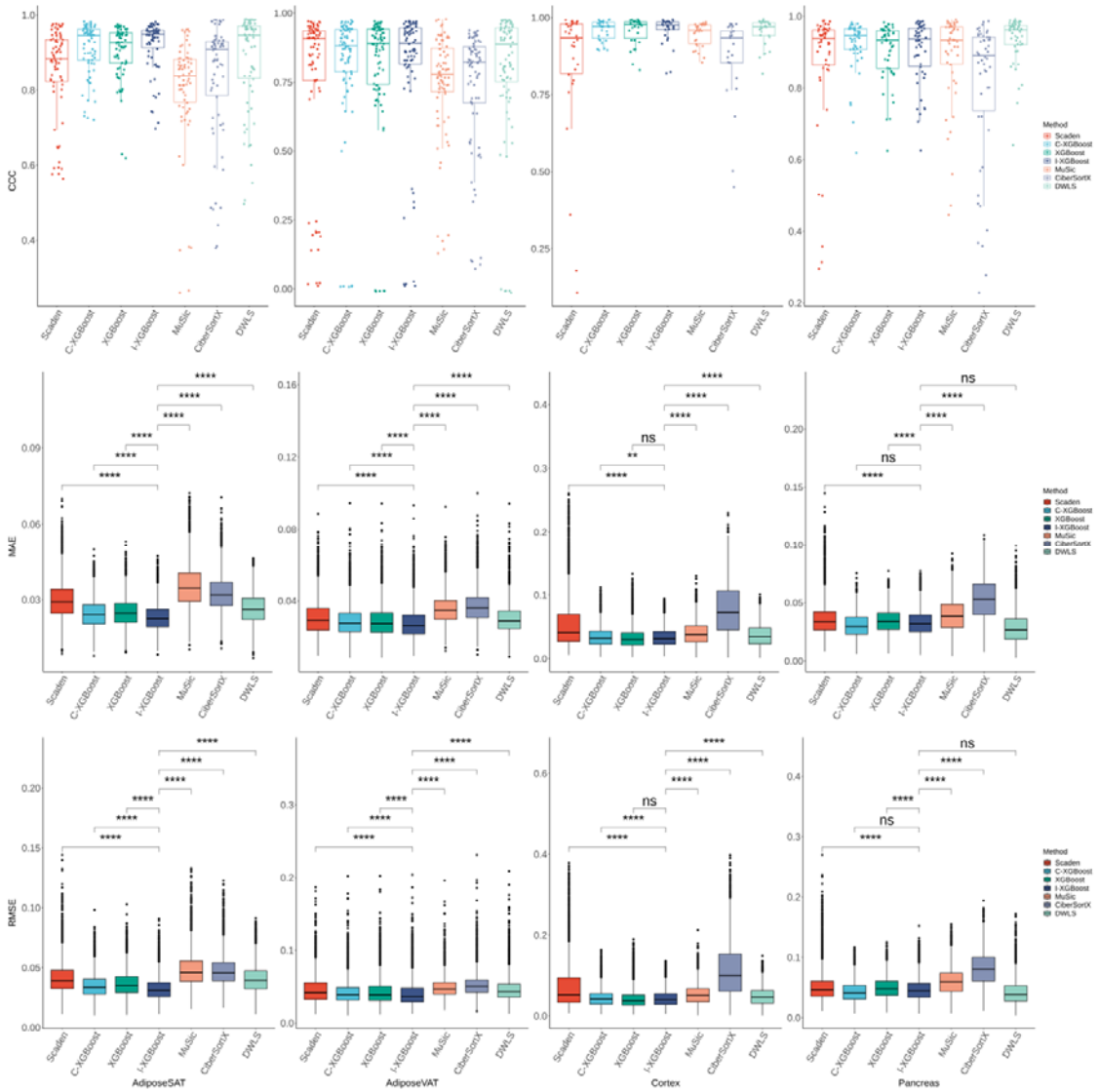

Figure 3 Benchmarking of cell deconvolution accuracy on pseudo-bulk data constructed by combining reference scRNA-seq examples.

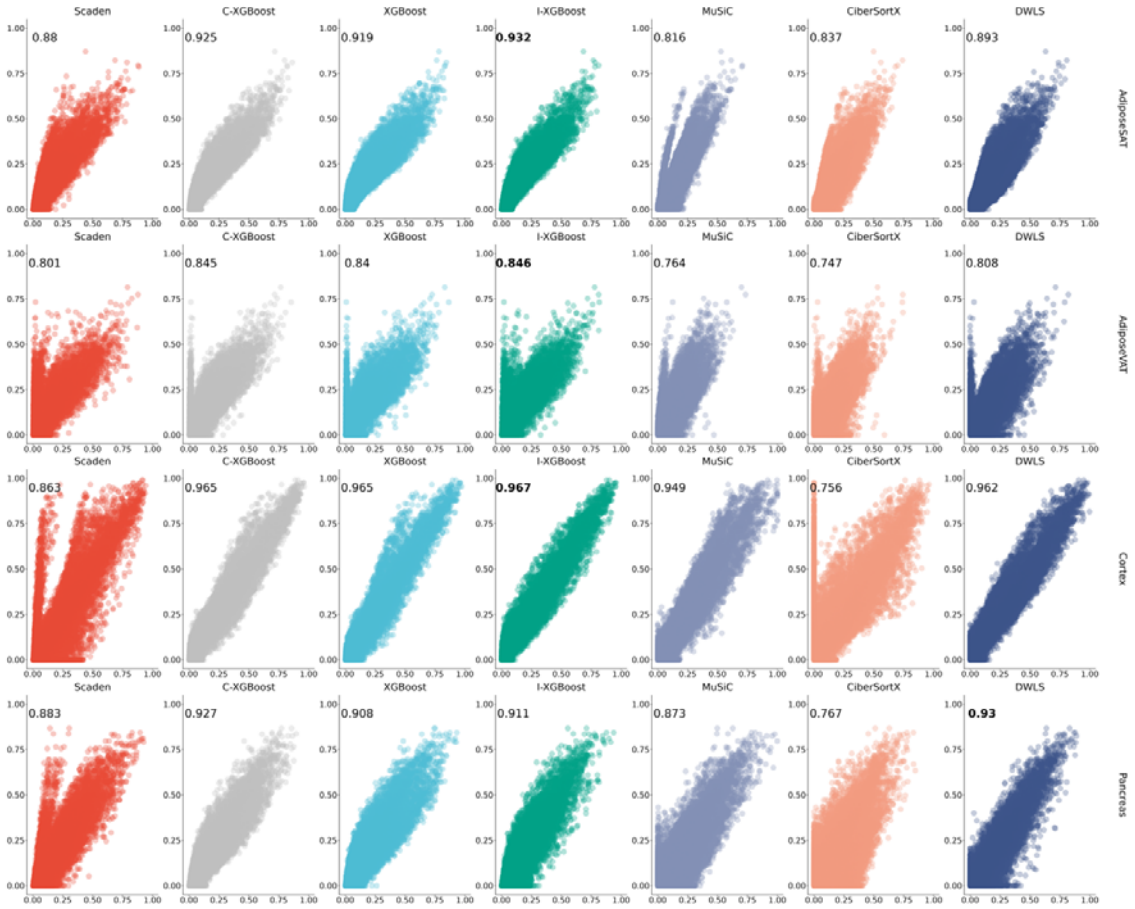

Figure 4 Benchmarking of cell deconvolution accuracy on pseudo-bulk data constructed by combining reference scRNA-seq examples with Lin's Concordance Correlation Coefficient (CCC).

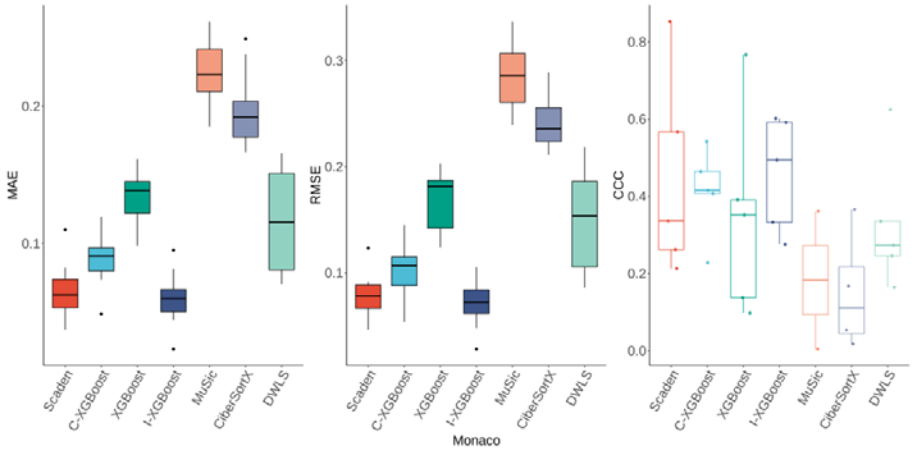

Figure 5 Benchmarking of cell deconvolution accuracy on real bulk data

Table 2 Benchmark of time complexity across different methods

Method	Sample size/running time(s)							
	100	200	300	400	500	600	700	800
Scaden	1158.77	1163.64	1151.67	1063.66	1117.92	1067.01	1072.44	1078.65
CausalCellInfer (C-XGBoost)	822.54	826.97	823.73	820.30	830.87	809.33	822.64	826.45
CausalCellInfer (I-XGBoost)	818.58	823.94	824.90	817.52	821.66	823.40	820.35	821.27
DWLS	14077.33	14396.41	14781.02	15160.65	15604.56	15919.06	16357.16	16970.25
MuSiC	75.69	79.18	86.98	86.29	105.25	99.25	112.82	126.94

We utilized reference datasets to train four deconvolutional Deep Neural Networks (DNNs) on the four tissues delineated in the methods section. The comparative results with existing methodologies are illustrated in Figures 3-4, Table S1. Our proposed method, CausalCellInfer (utilizing both I-XGBoost and C-XGBoost), was benchmarked against four state-of-the-art cell deconvolution techniques: CIBERSORTx[10], DWLS[12], Multi-subject Single Cell deconvolution (MuSiC) [36], and Scaden [16]. CIBERSORTx, 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, both Scaden 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 the 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[16].

In conclusion, our proposed method, CausalCellInfer (CausalCellInfer by I-XGBoost or C-XGBoost), demonstrated consistent superiority over existing methods such as CIBERSORTx, Scaden, MuSiC, and DWLS in deconvoluting pseudo-bulk RNA-seq data. Notably, it achieved the highest Concordance Correlation Coefficient (CCC) and the lowest Mean Absolute Error (MAE) across all four pseudo-bulk scenarios when compared to other cell deconvolution methods. A similar trend was observed for the Root Mean Square Error (RMSE), with our method achieving the lowest RMSE in nearly all tested scenarios. Although DWLS performed comparably to our method in the pancreas scenario, our method delivered significantly superior results in the remaining three scenarios. These findings underscore the reliability and feasibility of our proposed method, CausalCellInfer, in deconvoluting cell compositions from large-scale bulk samples.

Both Scaden and our proposed method employ deep neural networks for cell deconvolution. However, our method consistently outperforms Scaden, as evidenced by significantly CCCs and lower MAE and RMSE across all test scenarios. Moreover, our method exhibits a markedly smaller variance in CCCs across all cell types (Fig. 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 I-XGBoost/C-XGBoost, as opposed to all highly variable genes. Our results indicate that focusing solely on informative genes that contribute to cell-type identification can significantly enhance performance within a deep learning framework, compared to the use of highly variable genes.

This approach also yields an additional advantage: it leads to a less complex, yet more robust deconvolution model. We hypothesize that the inferior performance of Scaden may be due to the additional noise present in the training data. This observation warrants further investigations.

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 dataset, our proposed method (CausalCellInfer by I-XGBoost) excelled, surpassing all other methods in terms of both MAE and RMSE metrics. Furthermore, it achieved the highest CCC among all evaluated methods, as depicted in Figure 5. Interestingly, our comparative results align with the findings of Chen et al. [17], who conducted a benchmarking study of various cell deconvolution methods using identical datasets.

We subsequently compared the runtime of DWLS, Scaden, and MuSiC on a local server with the runtime of the CIBERSORTx web application. Our observations echoed the findings reported by Chen et al. [17] and Menden et al. [16], wherein deep-learning approaches exhibited a lower time complexity compared to signature-based methods such as DWLS and CIBERSORTx. CIBERSORTx is expected to have similar time complexity with DWLS. However, since CIBERSORTx is a web application, the ultimate time cost for CIBERSORTx may be much longer than DWLS, as the web tool is frequently occupied by other submitted tasks. Notably, our method demonstrated a significantly improved computational complexity compared to Scaden, with Scaden's running time being ~30-40% longer than our approach (Table 2). Furthermore, our approach does not necessitate training on the full gene set, resulting in a model that is considerably less complex than the original Scaden or its competitors. A significant benefit of this streamlined model is its substantially lower memory consumption. Although MuSiC displayed the lowest time complexity among all the methods benchmarked, it was found to be the least accurate in most simulation scenarios, as indicated by previous results.

Validity of inferred cell-type specific expression profile

We conducted a comparison analysis on the cell-type specific differentially expressed (DE) genes identified from our estimated CTS profiles (from GWAS data) and the reference single-cell RNA sequencing (scRNA-seq) datasets for similar phenotypes. If the inferred CTS expression profiles are reliable, there should be a significant overlap between the DEG identified from the two datasets. Here we regard the DEGs from the actual scRNA-seq data as the 'gold standard'.

Our observations reveal that the DE genes, identified via the CTS expression profiles based on our method and GWAS data, exhibit satisfactory positive predictive values (PPV) for the majority of cell types (here the DE genes identified from real scRNA-seq data were considered as the 'standard' for comparison). For instance, the DE genes 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 S2). Remarkably, 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. In conclusion, our findings validate the reliability of our proposed method for inferring CTS expression profiles.

Cell proportion and expression estimation for various traits in UKBB

To evaluate the effectiveness of our approach using real bulk expression data, we deconvoluted the tissue-specific imputed expression from PrediXcan for approximately 500,000 subjects from the UKBB. Specifically, we estimated the proportions of cells and the profiles of cell-specific expression for depression in the frontal cortex, T2DM in the pancreas, and obesity in both SAT and VAT. These estimations were made using our four trained deconvolutional DNNs. All binary traits excluding obesity were defined using a combination of self-reported diagnoses and International Classification of Diseases (ICD) codes. For obesity, we classified individuals with a Body Mass Index (BMI) exceeding 30 as obese.

Distinguishing cell types associated with various diseases

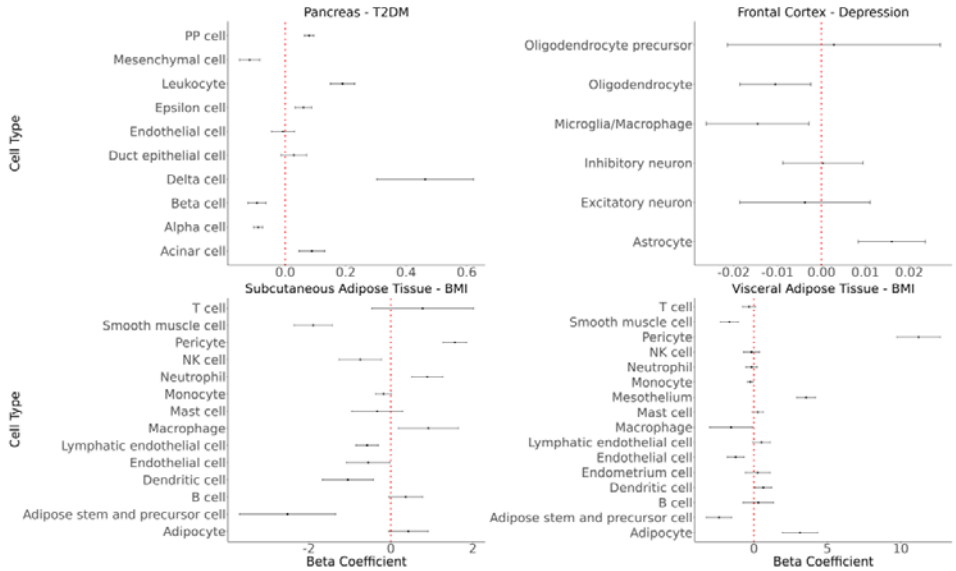

Figure 6 Associations between estimated cell proportions and disease onset

To elucidate the relationship between the estimated cell proportions and diseases/traits, we conducted a univariate test for each cell type via logistic/linear regression. As expected, 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 S3). Notably, a lower proportion of micro/macro cells and oligodendrocytes were associated with depression, while the proportion of astrocytes was elevated. These observations are consistent with existing literature; for instance, Rajkowska et al.[37] observed a decrease in both the density and ultrastructure of oligodendrocytes in the prefrontal cortex of patients diagnosed with depression.

Type 2 Diabetes

In our analysis of ten distinct cell types within the pancreas, we observed that the proportions of almost all cell types, excluding duct epithelial and endothelial cells, exhibited a significant association with the risk of developing T2DM. Notably, significant lower proportions of alpha and beta cells were observed with T2DM, with the reduction in alpha cells being more pronounced. Conversely, we detected associations of elevated proportions of acinar, delta, PP, and epsilon cells with T2DM, with the increase in PP cell proportions being especially prominent (Figure 6). Numerous studies have underscored the link between T2DM and a decrease in beta-cell mass and function [38,39]. Moreover, previous GWAS have pinpointed *IRX3* as a susceptibility gene for T2DM. The knockdown of the *IRX3* orthologue, irx3a, in zebrafish led to an increase in the number of ghrelin-producing epsilon cells and a decrease in the number of insulin-producing beta cells and glucagon-producing alpha cells [40]. These findings are in alignment with our observations on these cell types.

Obesity

Similar patterns were observed for BMI in both subcutaneous and visceral adipose tissues. Utilizing reference scRNA-seq datasets, we deconvoluted these tissues into 14 and 16 distinct cell types, respectively. More than half of these cell types, including adipose-derived stem and progenitor cell (ASPC), pericyte, smooth muscle cells (SMC), and adipocyte, were found to be significantly or nominally 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 monocytes and macrophages, were associated with obesity in both tissues (SAT and VAT), while others, like neutrophils and natural killer cells (NK cells), were associated with obesity in only one tissue. Compared to normal subjects, a significant decrease in the estimated proportions of most obesity-associated cell types, including ASPC, lymphatic endothelial cell (LEC), SMC, endothelial cells, and monocytes, was observed in the subcutaneous adipose tissue. Conversely, a significant increase in adipocytes was noted.

This inverse relationship between ASPC and adipocytes may be attributed to the role of adipogenesis in obesity, which involves the proliferation and differentiation of adipocyte precursor cells into mature adipocytes [41]. Our results suggest that abnormal adipogenesis occurs in the obese population, where a high level of adipocyte precursor cells undergoes the adipogenesis process and differentiates into adipocytes, subsequently leading to obesity [42]. Notably, a significant increase in pericyte proportions was also observed in both adipose

tissues, indicating its extensive involvement in the development of obesity. Supporting our findings, Tang et al.[43] demonstrated that pericytes are one of the precursor cells of adipocytes, underscoring their crucial role in obesity development. Watanabe et al.[44] showed that mice fed with a high-fat diet (HFD) exhibit a higher level of pericyte count than those on a chow diet, reinforcing the association between elevated pericytes and obesity.

Interactions between cell types and their associations with disease

We tested the effects of cell type interaction on 4 tissue-specific outcome variables using multivariate regression analyses, i.e., depression in the frontal cortex, T2DM in the pancreas, BMI in the SAT, and BMI in the VAT. Only cell types with significant main effects (i.e., univariate association with outcome) were considered for interaction analysis. Our results revealed that interactions between certain cell types were significantly associated with disease risks (refer to Table 3 and Table S4). For example, we discovered that astrocyte*microglia/macrophage interaction was associated with depression (based on expression data in the frontal cortex). A recent study by Matejuk et al.[45] suggested that the interaction pattern between astrocyte and microglia may contribute to the development of a broad spectrum of neurodegenerative disorders.

In addition, we also found that interaction of different cell types *across different tissues* may contribute to disease risks. For example, we found that the interaction between B cells in SAT and dendritic cells in VAT was significantly associated with obesity. Both cell types play crucial roles in maintaining immune homeostasis, and previous studies [46,47] has revealed the intensive 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 disorder status

Tissue	Disorder	Cell type interaction	P-value
Frontal cortex	Depression	Astrocyte*Microglia/Macrophage	3.79E-06
Frontal cortex	Depression	Microglia/Macrophage*Oligodendrocyte precursor	2.32E-02
Pancreas	Diabetes	Acinar cell*Leukocyte	4.26E-08
Pancreas	Diabetes	Acinar cell *PP cell	5.61E-03
Pancreas	Diabetes	Alpha cell*PP cell	2.67E-03
Pancreas	Diabetes	Beta cell*Leukocyte	4.29E-03
Pancreas	Diabetes	Beta cell*PP cell	4.53E-02
Pancreas	Diabetes	Epsilon cell*Leukocyte	3.93E-02
Pancreas	Diabetes	Leukocyte*Mesenchymal cell	2.04E-08
Pancreas	Diabetes	Leukocyte*PP cell	6.16E-03
Pancreas	Diabetes	Mesenchymal cell*PP cell	1.38E-05
SAT	Obesity	Adipose stem and precursor cell*Dendritic cell	6.75E-04
SAT	Obesity	Adipose stem and precursor cell *Pericyte	9.54E-07
SAT	Obesity	Dendritic cell*Lymphatic endothelial cell	2.75E-04
SAT	Obesity	Lymphatic endothelial cell *Pericyte	1.07E-02

SAT	Obesity	Lymphatic endothelial cell *Endothelial cell	1.24E-02
SAT	Obesity	Macrophage*Pericyte	3.68E-02
SAT	Obesity	Neutrophil*Pericyte	3.52E-10
SAT	Obesity	Pericyte*Smooth muscle cell	6.08E-07
VAT	Obesity	Adipose stem and precursor cell *Pericyte	6.94E-10
VAT	Obesity	Adipose stem and precursor cell * Smooth muscle cell	2.64E-03
VAT	Obesity	Dendritic cell * Lymphatic endothelial cell	4.95E-02
VAT	Obesity	Dendritic cell*Endothelial cell	2.35E-02
VAT	Obesity	Macrophage*Pericyte	4.05E-03
VAT	Obesity	Adipocyte*Pericyte	2.87E-02
VAT	Obesity	Monocyte* Smooth muscle cell	4.80E-02
VAT	Obesity	Pericyte* Smooth muscle cell	3.84E-06
VAT	Obesity	Pericyte*Endothelial cell	5.09E-03
VAT	Obesity	Pericyte*Mesothelium	9.04E-12
Frontal cortex*Pancreas	Depression	Astrocyte*Alpha cell	3.91E-02
Frontal cortex*SAT	Depression	Astrocyte* Smooth muscle cell	2.79E-02
Frontal cortex*SAT	Depression	Astrocyte*Endothelial	2.13E-02
Frontal cortex*SAT	Depression	Astrocyte*Mast cell	4.69E-02
Frontal cortex*SAT	Depression	Astrocyte*Monocyte	2.53E-02
Frontal cortex*SAT	Depression	Astrocyte*Neutrophil	3.22E-02
Frontal cortex*SAT	Depression	Astrocyte*NK cell	3.16E-02
Frontal cortex*SAT	Depression	Astrocyte*Pericyte	4.73E-02
Frontal cortex*SAT	Depression	Oligodendrocyte precursor*Adipose stem and precursor cell	2.14E-02
Frontal cortex*SAT	Depression	Oligodendrocyte precursor*Dendritic cell	4.54E-02
Frontal cortex*SAT	Depression	Oligodendrocyte precursor*Neutrophil	4.54E-02
Frontal cortex*SAT	Depression	Oligodendrocyte precursors*NK cell	6.45E-03
Frontal cortex*SAT	Depression	Oligodendrocyte precursor*Pericyte	4.09E-02
Frontal cortex*VAT	Depression	Astrocyte*Dendritic cell	4.56E-02
Frontal cortex*VAT	Depression	Astrocyte*Macrophage	4.31E-02
Frontal cortex*VAT	Depression	Excitatory neuron*Adipose stem and precursor cell	4.04E-02
Frontal cortex*VAT	Depression	Oligodendrocyte precursor*Smooth muscle cell	2.62E-02
Frontal cortex*VAT	Depression	Oligodendrocyte precursor*Adipocyte	3.38E-02
Frontal cortex*VAT	Depression	Oligodendrocyte precursor*Mast cell	1.07E-02
Frontal cortex*VAT	Depression	Oligodendrocyte precursor*Mesothelium	2.11E-02
Frontal cortex*VAT	Depression	Oligodendrocyte precursor*Monocyte	1.03E-02
Frontal cortex*VAT	Depression	Oligodendrocyte precursor*Neutrophil	2.42E-02
SAT*VAT	Obesity	Lymphatic endothelial cell*Dendritic cell	3.50E-02
SAT*VAT	Obesity	Adipocyte*Mesothelium	4.95E-02
SAT*VAT	Obesity	B cell*Dendritic cell	2.13E-02

SAT*VAT	Obesity	Endothelial*Dendritic cell	4.47E-02
SAT*VAT	Obesity	Endothelial*Monocyte	4.67E-02
SAT*VAT	Obesity	Mast cell*Adipocyte	4.78E-02
SAT*VAT	Obesity	Mast cell*Mast cell	2.91E-02
SAT*VAT	Obesity	Monocyte*Dendritic cell	4.56E-02
SAT*VAT	Obesity	Neutrophil*Dendritic cell	3.62E-02
SAT*VAT	Obesity	NK cell*Dendritic cell	3.58E-02
Pancreas*VAT	T2DM	Acinar cell*Dendritic cell	2.76E-02
Pancreas*VAT	T2DM	Delta cell*Dendritic cell	3.67E-02
Pancreas*VAT	T2DM	Delta cell*T cell	4.19E-02

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

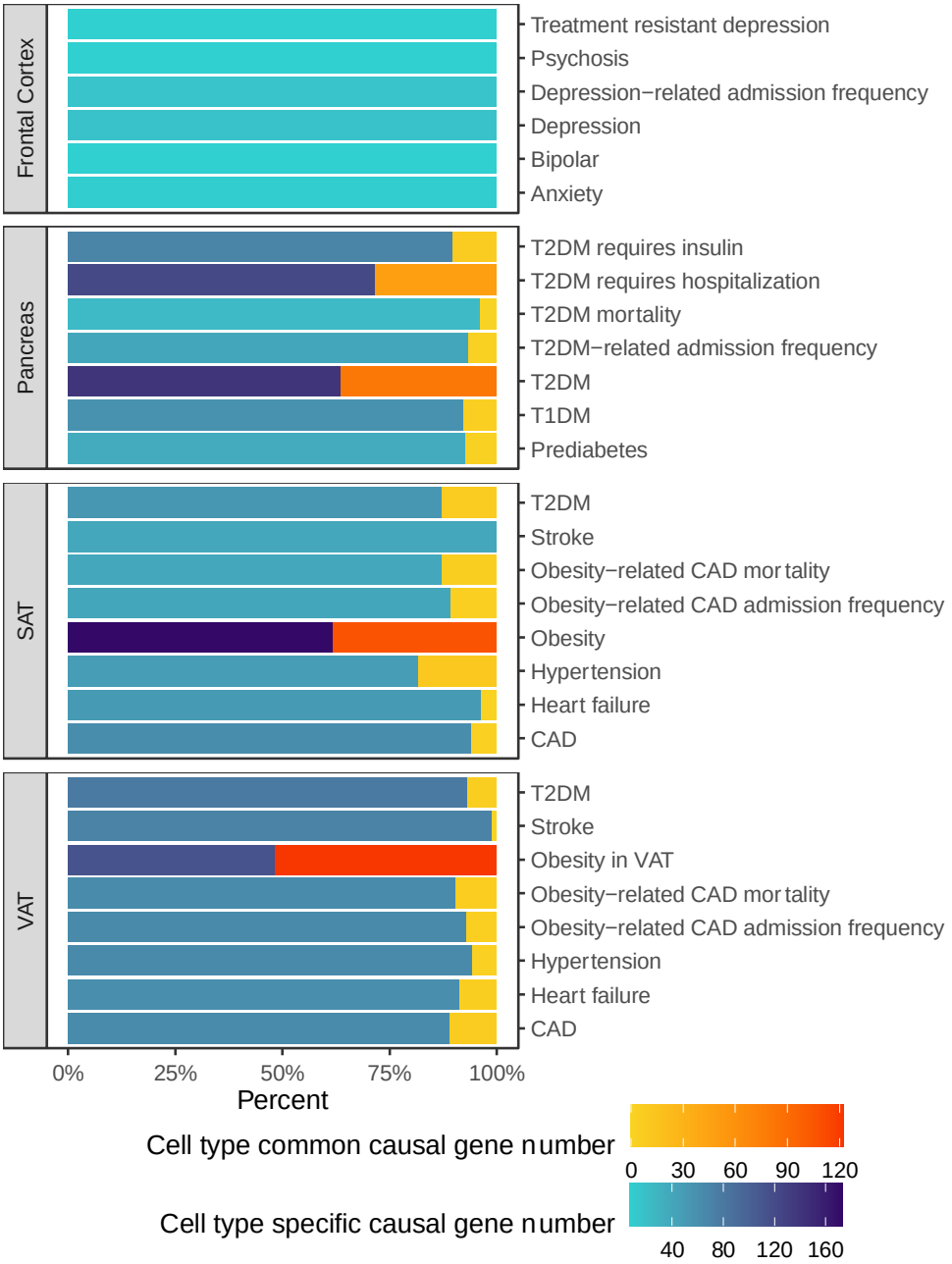

Figure 7 Cell specificity comparison of the identified causal genes for disease onset and prognoses

Causal genes for disease susceptibility were found to exhibit variation across different cell types (Figure 7, Table S5). We define genes that were identified as causal (by the PC-simple algorithm) in only one single cell type as "cell-type specific" causal genes, while those identified in >= 2 cell types were labelled as "cell-type common". All 18 direct causal genes associated with increased risks of depression were specific to a particular cell type. For instance, C10orf82 was identified as a causal gene for depression susceptibility exclusively in exhibitory cells (refer to Table 4 and supplementary Table S5 for details). An integrated

analysis of methylomic and transcriptomic analysis conducted by Xie et al.[48] revealed that C10orf82 is among the top 3 up-regulated genes, based on methylation levels associated with major depressive disorder.

Interestingly, we observed both altered cell proportions and gene expression changes in certain cell types associated with disease status. In contrast, other cell types exhibited only changes in gene expression. For example, while the proportion of inhibitory neurons was not significantly associated with depression susceptibility, this cell type had the highest number of identified direct causal genes.

Table 4 Top causal genes (ranked by P-value) from inferred CTS expression profiles

Tissue	Disorder	Genes
Frontal cortex	Depression	*RPS20, MCM3, C10orf82, CHDHD5, TMEM106C, GRP155*
Pancreas	T2DM	*ALS2CL,CWC22,LTF,TTC4,GPRC5A,MBLAC1,RTCA,USP8,PAX9,RGL3*
SAT	Obesity	*PIK3R4,NME4,PI15,KCNN2,TMEM206,MBLAC1,PAXIP1.AS2,AGRN, CAMK2A*
VAT	Obesity	*NGEF,ACSM2A,ECT2L,NAA11,MAK16,ADCK1,SGSH,GDPD1,MARVELD1,PAOX,PEPD,EPCAM,PLCD 3,FAM111A,RGCC*

For T2DM, we identified 222 direct causal genes associated with disease risk, with varying involvement across different cell types. Unlike depression, a larger proportion of these direct causal genes were common across multiple cell types (Figure 7, Table S5). Specifically, 81 of these genes were identified as causal in more than two cell types. For instance, *DOK3* was identified as a direct causal gene in two distinct cell types, Beta and Epsilon cells (detailed information can be found in the Table S5). A study showed that *DOK3* was among the most significantly upregulated genes in patients with diabetes, and its expression can be reverted by the antidiabetic agent, metformin [49].

We identified a substantial number of direct causal genes for obesity in both subcutaneous and visceral adipose tissues. While the proportion of 'cell-type common' causal genes in the subcutaneous adipose tissue was comparable to that for T2DM, the proportion of common genes was much higher in the visceral adipose tissue (~51.9%). Specifically, we identified 107(38.2%) and 122(51.9%) genes in subcutaneous and visceral adipose tissues, respectively, that contribute to obesity in at least two cell types.

In other words, most of the causal genes for obesity in subcutaneous adipose tissue exhibited cell type specificity. For example, *PIK3R4* was identified as a primary causal gene for subcutaneous adipose tissues in adipocytes. This is consistent with a previous study that demonstrated the involvement of PIK3R4 in adipocyte autophagy[50]; notably, diminished adipocyte autophagy may promote obesity[51]. These findings substantiate the potential causal role of *PIK3R4* in the development of obesity.

Cell type-specific genes involved in disease prognosis (direct causal genes as prioritized by PC-simple)

Identifying specific genes that contribute to disease outcomes or prognosis is crucial for developing personalized treatment or interventions. Detailed prognosis data are often

missing in single-cell studies but are readily available in the UKBB. Here we conducted an exploratory analysis to identify the direct causal genes associated with disease-specific mortality and hospitalizations(i.e., number of admissions).

For depression, we revealed a number of genes associated with the frequency of hospitalizations and some with treatment resistance. All the direct causal genes identified for treatment resistance were specific to a single cell type(Table S5.a). Similar pattern was observed for the identified direct causal genes associated with the frequency of hospital admissions(Table S5.a).

As for T2DM, we noted a modest proportion (6.8%) of direct causal genes influencing T2DM-related hospitalization frequencies across multiple cell types(Table S5.b). In contrast, the causal genes identified for T2DM mortality exhibited a distinct pattern. The majority of the identified causal genes for T2DM-related mortality were cell-type specific(Table S5.b). For example, we identified the gene *LIPC* as a determinant of mortality in T2DM within leukocytes. This finding aligns with previous research indicating a relationship between genetic polymorphism of *LIPC* and T2DM [52] and insulin resistance[53].

Obesity is a well-documented risk factor for coronary artery disease (CAD), with a number of studies confirming its association with an increased incidence of CAD [54]. Our findings reveal that a multitude of genes play a significant role in CAD-related mortality and hospitalization among obese patients. Notably, the majority of these genes exert their influence on CAD mortality and admission in specific cell types(Table S5.c and Table S5.d).

Similar patterns of causal gene distribution were observed for other related disorders studied. Most of the identified causal genes, typically exceeding 80%, for the 17 studied disorders (anxiety, bipolar disorder, psychosis [tissue studied: frontal cortex]; prediabetes, T1DM, T2DM requiring hospitalization(binary outcome), T2DM requiring insulin [tissue studied: pancreas], CAD, T2DM, hypertension, heart failure, stroke related to obesity [tissues studied: SAT and VAT]), were found to be cell-type specific. The proportion of cell-type common causal genes for depression-related disorders were in general lower than that for T2DM and obesity-related disorders, possibly due to the inferior performance of PrediXcan in estimating expression profiles for the frontal cortex, owing to a limited sample size of the reference expression dataset.

OpenTargets enrichment results

We conducted enrichment analyses for depression, T2DM and obesity. Notably, we performed four sets of enrichment analyses for each pair of tissues associated with obesity. Specifically, we investigated whether the genes we identified were more likely to be listed as targets for BMI/obesity/body weight/WHR (waist-hip-ratio) in the OpenTargets database than expected.

Encouragingly, the sets of causal genes we identified for all the studied diseases, with the exception of depression, were significantly/nominally enriched among the disease targets catalogued in the OpenTargets database (Table 5). This observation further validates the accuracy 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	**3.72E-04**
SAT	Obesity	***5.76E-02***
SAT	Body weight	**1.75E-03**
SAT	BMI	**3.42E-04**
SAT	WHR	**4.09E-02**
VAT	Obesity	2.92E-01
VAT	Body weight	**2.42E-02**
VAT	BMI	4.32E-01
VAT	WHR	***5.72E-02***
Frontal cortex	Depression	6.84E-01

Results of further pathway enrichment analyses

We combined the identified cell-type specific causal genes (by PC-simple algorithm) to perform pathway enrichment analysis. The enriched pathways were shown in Table S6. For instance, "spliceosome assembly" was significantly enriched for depression. In a relevant study, Li et al.[55] reported that the differentially expressed genes for depressive disorders were also significantly enriched in this pathway. For T2DM, pathways like "Endocytosis", "Aspartate and asparagine metabolism", and "Regulation of signaling by NODAL" were significantly enriched. Interestingly, a study[56] discovered that an increase in NODAL expression could lead to a rise in apoptosis in beta cells. It has been established that dysregulation of beta cells plays a pivotal role in the onset of T2DM [57].

We identified similar patterns among the enriched pathways based on gene sets associated with disease prognosis (Table S6). For instance, the "Sucrose degradation"

pathway was significantly enriched in treatment-resistant depression. Kim et al.[58] reported that sucrose withdrawal could result in depression-like behaviors through overexpression of Kir2.1 in the nucleus accumbens. Further studies indicated an association between hyper-active sucrose metabolism and major depressive disorder[59]. The "Glial cell differentiation" pathway was notably enriched in relation to the frequency of depression-related admissions. Rajkoska et al.[60] implicated that glia cells are involved the neuropathology of depression and suggested their potential as therapeutic targets. Zhou et al.[61] further suggested that specialized glial cells, i.e., oligodendrocyte lineage (OL) cells, were implicated in the pathogenesis of depression.

"Glycerolipid metabolism" emerged as a significantly enriched pathway in relation to T2DM mortality. Prentki et al.[62] implicated the involvement of Glycerolipid cycle in the regulation of energy homeostasis, insulin secretion and appetite control. They also suggested that pharmacological intervention on this pathway may result in antidiabetic effects.

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

We compared the 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. The trait-associated genes from adjusted TWAS are shown in Table 6 and Table S7. As anticipated, cell type could introduce confounding factors into the TWAS results. Adjusting for estimated cell proportions when performing TWAS analysis led to significant coefficient changes for many genes (Table 7; Table S8-9). For example, adjusting for cell proportions resulted in significant effect size changes for approximately 63% of the identified T1DM-associated genes. While some trait-associated genes identified from standard TWAS remained associated with the target trait in conditional TWAS, others did not. This highlights the influence of cell type proportions on TWAS results.

Results from CTS-TWAS (univariate association tests)

Here we report the results from CTS-TWAS based on univariate association tests (instead of PC-simple which involves repeated conditional tests). Similar to PC-simple analysis, we defined trait-associated genes identified from at least two cell-types as 'cell-type common' TWAS genes, and those identified from only one cell-type as 'cell-type specific'.

Similar to direct causal gene analysis by PC-simple, the majority of genes identified from CTS-TWAS were cell-type specific TWAS genes (Table S10-11). The number of significant genes for each trait are listed in Table S10. Here we consider FDR<0.05 as 'significant' genes. Of note, not all studied traits that have trait-associated genes can pass multiple testing correction. Under the above criteria for gene inclusion, for traits that indeed have 'significant' genes, the proportion of cell-type specific TWAS genes mostly exceeds 80%. It also worth noting that most top-ranked (by absolute effect size from linear regression) genes from CTS-TWAS were cell-type specific (Table 8). For example, among the top one

trait-associated genes for all 10 cell-types in the pancreas, only 1 gene was cell-type common(Table S11.b).

We also conducted OpenTargets enrichment analysis on the identified trait-associated genes from CTS-TWAS for 4 tissue-specific disorders (i.e., T2DM in pancreas, T2DM related hospitalization in pancreas, Obesity in subcutaneous adipose, Obesity in visceral adipose) with >=5 identified genes. The trait-associated gene sets we identified for these disorders, were significantly or at least nominally enriched among the disease targets listed in the OpenTargets database (Table 9).

Furthermore, we compared the genes identified from a causal inference approach (PC-simple, which prioritizes direct causal genes) and univariate association analysis (TWAS) using the inferred CTS expression profiles. For most traits with both causal genes and "significant" trait associated genes, we identified considerably more trait-associated genes from TWAS association analysis than from PC-simple(Table S10). As PC-simple relies on repeated conditional independence testing, it is more stringent than univariate association testing. We also found that PC-simple can uncover new associations compared to standard association tests (Table S12). We hypothesized that some identified associated genes from TWAS may be peripheral (indirect causal) genes that influence the target traits through the identified (direct) causal genes. As anticipated, relaxing the FDR cutoff lead to an increased gene overlap rate between results from the PC-simple and standard association analyses(Table S12).

Table 6 Examples of trait-associated genes identified from TWAS after adjustment for estimated cell-type proportions

Tissue	**Disorder**	**Genes**
Frontal cortex	Depression	ZSCAN9,BTN3A2,C4A
Pancreas	T2DM	B3GALT1,ZFP91,CRYZP1
SAT	Obesity	SEMA3F,RBM6,UHRF1BP1
VAT	Obesity	SULT1A2,PMS2P3,STAG3L1

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	*GSTZ1*	0.00339	1.42E-08	0.00266	3.64E-05	2.44E-03	1.22E-06	2.53E-03
Pancreas	T1DM	*NOMO3*	0.00146	4.65E-05	0.00100	7.18E-03	4.39E-06	2.65E-03	2.10E-01
Pancreas	T2DM	*GABRR2*	-0.00434	4.29E-04	-0.00276	5.13E-02	2.41E-02	2.01E-02	3.44E-01
Pancreas	T2DM	*CRHR2*	0.01193	2.30E-07	0.01116	1.50E-06	7.81E-04	3.94E-05	2.96E-04
Pancreas	T2DM related admission	*ESRRB*	0.00855	2.10E-04	0.00705	3.19E-03	1.63E-02	8.67E-03	5.82E-02
Pancreas	T2DM related admission	*TMCO6*	0.01602	1.52E-09	0.01523	1.11E-08	2.67E-03	4.87E-07	3.18E-06
SAT	CAD	*PAX8*	0.00939	1.26E-05	0.00929	1.55E-05	3.70E-02	3.46E-03	5.07E-02

SAT	CAD	*TNFSF14*	0.01605	1.71E-03	0.01596	1.81E-03	3.07E-02	7.92E-02	1.00E+00
SAT	HTN	*RBCK1*	-0.01609	2.86E-04	-0.01599	3.11E-04	3.11E-02	5.25E-03	1.00E+00
SAT	HTN	*PMS2P3*	-0.01333	2.84E-07	-0.01325	3.36E-07	5.46E-03	1.70E-05	1.10E-03
SAT	Heart Failure	*FKBP11*	-0.01068	1.85E-04	-0.01072	1.74E-04	2.07E-02	1.79E-02	5.70E-01
SAT	Heart Failure	*NME7*	0.01573	1.73E-05	0.01105	3.35E-03	1.58E-07	2.58E-03	1.00E+00
SAT	T2DM	*MLKL*	-0.01074	1.70E-15	-0.02111	1.81E-17	6.47E-07	9.27E-13	5.95E-14
SAT	T2DM	*LARP4*	0.00788	1.78E-03	0.00795	1.62E-03	6.75E-04	3.96E-02	1.00E+00
VAT	CAD	*ZNF575*	-0.02154	3.82E-07	-0.02012	2.81E-06	3.95E-02	1.05E-03	7.75E-03
VAT	CAD	*PHLDB3*	0.01968	1.41E-04	0.01739	9.70E-04	2.52E-02	3.00E-02	1.00E+00
VAT	HTN	*CEP63*	0.02587	6.59E-09	0.02596	5.87E-09	9.64E-03	8.66E-07	1.62E-05
VAT	HTN	*SMDT1*	0.01517	5.15E-07	0.01509	5.94E-07	2.90E-02	3.23E-05	1.64E-03
VAT	Heart Failure	*WDSUB1*	-0.00377	1.66E-07	-0.00376	1.82E-07	2.99E-02	4.57E-05	5.01E-04
VAT	Heart Failure	*SLFN12L*	-0.00993	1.36E-07	-0.00084	7.14E-01	6.57E-12	4.18E-05	1.00E+00
VAT	T2DM	*NPM1P26*	-0.00354	7.99E-04	-0.00350	9.09E-04	1.30E-04	2.42E-02	1.00E+00
VAT	T2DM	*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

Cell type	Genes	Estimates	Pvalues	Qvalues
acinar cell	*PAX9*	-0.0959	1.97E-04	4.10E-02
acinar cell	*MYRFL*	0.0499	2.22E-04	4.37E-02
acinar cell	*LINGO4*	-0.0329	8.22E-06	5.12E-03
alpha cell	*EXTL1*	0.0600	9.78E-05	1.79E-02
alpha cell	*ESPNL*	-0.0432	4.30E-04	3.50E-02
alpha cell	*ZNF215*	0.0355	8.14E-06	3.78E-03
beta cell	*USP6*	-0.0546	5.26E-04	4.19E-02
beta cell	*HLA-DQB2*	-0.0538	4.24E-04	3.78E-02
beta cell	*IP6K3*	-0.0506	7.12E-04	4.93E-02
delta cell	*RMI2*	0.0339	3.61E-04	4.83E-02
delta cell	*TNNI3K*	-0.0303	4.43E-04	4.89E-02
delta cell	*FAM153A*	0.0290	1.38E-05	6.47E-03
duct epithelial cell	*SLC25A41*	-0.0471	7.26E-05	1.51E-02
duct epithelial cell	*IRX3*	-0.0216	4.86E-04	4.91E-02
duct epithelial cell	*GPRC5A*	0.0213	1.73E-08	6.47E-05
endothelial cell	*GNA15*	-0.0349	3.13E-06	2.86E-03
endothelial cell	*CRHR2*	-0.0328	2.69E-05	1.12E-02
endothelial cell	*RNF122*	0.0291	9.44E-05	1.77E-02
epsilon cell	*PRAP1*	-0.0737	1.12E-04	1.14E-02
epsilon cell	*WNT10B*	-0.0530	3.60E-04	2.47E-02

epsilon cell	*CLNK*	-0.0485	3.43E-04	2.42E-02
leukocyte	*GYS2*	-0.0616	1.93E-04	2.38E-02
leukocyte	*NAALAD2*	0.0579	1.43E-06	8.92E-04
leukocyte	*TDRKH*	-0.0225	2.09E-04	2.38E-02
mesenchymal cell	*PAX9*	0.2298	2.39E-19	8.93E-16
mesenchymal cell	*HCG9*	0.0546	7.89E-06	3.67E-03
mesenchymal cell	*SLC6A19*	-0.0394	4.78E-04	4.97E-02
PP cell	*PCDHA2*	-0.0599	1.06E-04	1.04E-02
PP cell	*C10orf62*	-0.0541	4.79E-05	6.57E-03
PP cell	*CDC25A*	-0.0466	1.02E-05	2.95E-03

Note: Qvalues indicated false discovery rate which are calculated from Pvalues 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	**1.71E-05**
Pancreas	T2DM related admission	***6.09E-02***
SAT	Obesity	**2.39E-05**
SAT	Body weight	**4.35E-05**
SAT	BMI	**2.59E-03**
SAT	WHR	**5.10E-03**
VAT	Obesity	5.44E-01
VAT	Body weight	**1.06E-02**
VAT	BMI	6.49E-01
VAT	WHR	**4.94E-02**

Discussion

In this study, we introduced CausalCellInfer, a comprehensive framework that integrates XGBoost, deep learning, the invariance principle, causal inference techniques, and regularized matrix completion, with applications to GWAS-imputed transcriptome profiles. This framework is designed to deconvolute cell compositions and estimate Cell Type Specific (CTS) profiles from large-scale bulk samples. As a fully data-driven method, CausalCellInfer first identifies marker genes from the reference single-cell RNA sequencing (scRNA-seq) dataset. 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 four state-of-the-art deconvolution methods, namely DWLS, CIBERSORTx, Scaden, 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 Differential Expression (DE) gene enrichment analysis, OpenTargets enrichment analysis, and pathway enrichment analyses. Moreover, we applied CausalCellInfer to four GWAS-predicted expression datasets in the UKBB, and performed cell-type-level analyses for 29 clinical phenotypes.

The estimated cell compositions and CTS expression profiles provide us with a more comprehensive understanding of disease mechanisms at the cellular level. Given that many genes may exert their effect on disease susceptibility/prognoses in specific cell types, our estimated cell-type-specific DEGs or potential direct causal genes may illuminate novel targets for drug development and personalized treatment.

The framework presented in this study exhibits several strengths. A notable advantage is that CausalCellInfer incorporated a novel marker-gene informed and deep learning-based cell deconvolution model. This model automatically detects marker genes from the reference scRNA-seq dataset, eliminating the need for prior knowledge of marker genes. Secondly, our framework does not require a constant marker-gene expression assumption (i.e., a given marker gene is expressed at a constant level across all 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 in the data than signature-based methods. As also reported in [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. It operates much faster than all competing methods under study.

Most importantly, CausalCellInfer is flexible and can be applied to estimate cell compositions and the corresponding CTS expression profiles, including any large biobanks with available GWAS data, which is another novel contribution from this study. We have shown that the proposed framework enables us to extend standard TWAS analyses to *cell type specific* analyses (e.g., cell type specific TWAS (CTS-TWAS), CTS direct causal gene detection via PC-simple), and detect the association of (genetically-predicted) cell-type proportion with diverse phenotypes using GWAS data. Other novel analyses include testing

cell-type interactions on phenotypes, and conditional TWAS adjusted for cell type proportions.

Related approaches

Numerous methods have been proposed to deconvolute cell compositions and estimate Cell Type Specific (CTS) profiles from bulk samples. However, these methodologies differ from the current work. For instance, Newman et al. [10] introduced CIBERSORTx, a tool designed to infer cell compositions and CTS expression profiles from bulk tissue transcriptomes. CIBERSORTx is an extension of their previous framework, CIBERSORT, which was initially designed to estimate cell compositions. It utilizes reference gene expression profiles (GEPs) of predetermined marker genes for cell deconvolution. However, it relies on the assumption of constant marker gene expression levels within cell types across different samples, which might not hold in practice.

Another similar line of work is deep learning-based cell deconvolution and CTS expression estimation methods. Menden et al. [16] introduced Scaden, a deep neural network-based cell deconvolution model designed to dissect bulk samples. Rather than relying on GEPs of marker/signature genes, they trained their deconvolution model directly on the reference scRNA-seq using approximately 10,000 highly variable genes. However, this large number of features often leads to unfavorable running times and introduce noise to the model, potentially resulting in poorer performance in estimating cell proportions, as shown in our simulation benchmark. Here we leveraged another machine learning model (XGBoost) and the invariance principle to reduce the number of genes as input, which leads to improvement in computational speed as well as predictive performance. Following a similar approach to Scaden, Chen et al. [17] proposed an autoencoder-based model, TAPE, to infer cell compositions and CTS profiles from bulk samples. Similar to Scaden, TAPE also generates bulk samples from scRNA-seq datasets for training, but TAPE is based on an autoencoder architecture. Following Scaden, TAPE also uses highly variable genes (approximately 10,000 genes) for model training. However, TAPE can only infer the CTS expression profiles for the preselected highly variable genes. But we propose to use the detected marker genes from I-XGBoost (~100 genes) to inform the deconvolution model training, rather than using all highly variable genes. Importantly, the identified invariant marker genes from I-XGBoost were more resistant to data bias and noises. Compared with existing deep learning-based methods, our method is likely more efficient and generalizable, as it depends only on a small set of (invariant) marker genes for cell deconvolution model training. This also makes it particularly suitable for dissecting large-scale bulk RNA-seq data, including those predicted from genotypes.

Another important distinction of our work from all previous studies on cell deconvolution is that we pioneered applications to genotype-imputed expression data, especially genotype data in 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 (estimated) cell proportions and CTS expression, together with diverse clinical phenotypes in a large sample. Such analyses are often not possible in conventional single-cell studies, owing to the small number subjects involved and limited phenotypic information available.

Several methods have been proposed to link scRNA-seq data with GWAS data. In a previous study, 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 diseases/traits. They first use scRNA-seq to infer disease-associated cell types and processes then build connections between them and SNPs. In another study, scGWAS was developed to leverage scRNA-seq to interrogate cell type specific transcriptional changes induced by genetic variants [63]. It identified associations between traits and cell types by examining whether susceptibility genes implicated by GWAS exhibits concordant activation in a particular cell type using proportion tests.

The current work is different from the above-mentioned methods. One of our primary objectives is to estimate the cell proportions and the corresponding CTS expression profiles directly from GWAS data (more specifically, GWAS-predicted bulk expression data), whereas the above methods concentrate on linking susceptibility genes implied by GWAS with those identified from scRNA-seq data. We propose to combine causal inference principles with deep neural networks and regularized matrix completion to estimate cell compositions and their corresponding CTS expression profiles using reference scRNA-seq and GWAS data. A wide range of cell-type-level analyses, such as CTS-TWAS, causal gene discovery, enrichment analysis, are readily accessible based on the inferred cell proportions and CTS expression profiles.

A note on our two marker gene identification methods

In this study, we propose two distinct strategies for marker gene detection. The first strategy focuses on identifying marker genes that remain invariant across various environments, while the second strategy concentrates on the amalgamation of markers from all environments. In the context of causal inference, an environmental variable is typically considered as an “external” variable that can influence the outcome variable of interest. It defines the “experimental” settings for data collection, and can be utilized to categorize subjects into different subgroups. Regarding this, variables such as disease status, gender, and age group can act as environmental variables in our study. Notably, the first strategy tends to identify a significantly smaller number of marker genes compared to the second. This reduced gene pool for modeling expedites the training process of the deconvolution model, and since the genes are selected based on the invariance principle[8], this approach is more robust to confounding and in general has better generalizability to new samples.

However, a high overlap rate between related cell types could compromise the performance of our method in deconvoluting their corresponding cell compositions, and in this case I-XGBoost may return only a small number of genes. The number may be too small to enable reliable model training and accurate prediction. Under this situation, additional information from more genes may help further differentiate closely related cell types. Consequently, in such cases we recommend using marker genes identified from the 2nd strategy to train the deconvolution model, when the overlap rate of marker genes identified from the 1st strategy between cell types is high.

Limitations

The efficacy of machine learning algorithms in identifying target-associated signals is known to be influenced by sample size. Marker genes, essential for the training of deconvolution models, are automatically identified from the reference single-cell RNA sequencing (scRNA-seq) data using either I-XGBoost or C-XGBoost. However, the detection of marker gene may be compromised when the sample size for certain cell types is extremely small. This limitation can be addressed through strategies such as oversampling, data augmentation or data integration. The continual advancements in scRNA-seq technologies have facilitated access to larger scRNA-seq datasets. Consequently, sample size will likely become less of a constraint for our method. The accuracy of ENIGMA (in estimating CTS expression profiles) could be affected if the reference scRNA-seq dataset is biased (e.g. selection bias in sampling subjects). This is because ENIGMA operates on the assumption that the expectation of each CTS expression profile aligns with the reference profile matrix derived from the reference scRNA-seq. This challenge could be mitigated by data augmentation/integration and batch effect correction. In addition to the above, in our application to UKBB, we only focus on how gene expression affects the outcome, following the principles of MR; however, our method does not capture how traits may lead to changes in expression.

While the results from our application are promising, they are preliminary and require further experimental validation, which we have not yet performed. Despite these limitations, our method is highly adaptable and can be extended to address these issues.

In conclusion, the CausalCellInfer framework, integrating an innovative marker gene detection method, offers a robust and efficient strategy for deconstructing large-scale bulk single-cell RNA sequencing (scRNA-seq) data into cellular compositions and corresponding cell type-specific expression profiles. Its distinguishing features include its robustness and accuracy in predictions. Moreover, its ability to handle modest-size datasets and accurately identify relevant genes, even amidst batch effects, establishes it as a promising tool in the field. Importantly, we also demonstrated how the method may be applied to large-scale biobank data with imputed transcriptome profiles, effectively bridging the gap between GWAS, comprehensive clinical data, and single-cell analysis.

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 at Github:

Conflicts of interest
The authors declare no relevant conflicts of interest.

Acknowledgements
This work was supported partially by the 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.

References

1. Wolf FA, Angerer P, Theis FJ. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;19:1–5.

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. *J Hum Genet*. 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. *Nat Genet*. 2022;54(10):1479–1492.

5. Yin L, Feng Y, Lau A, Qiu J, Sham P, So H. A bayesian network-based framework to uncover the causal effects of genes on complex traits based on GWAS data. *medRxiv*. 2022:2022.12. 25.22283943.

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

7. Yin L, Chau CK, Sham P, So H. 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. *Cancer Systems Biology: Methods and Protocols*. 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 G. 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 Biol*. 2014;15(12):1–21.

14. Ritchie ME, Phipson B, Wu DI, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47.

15. Nguyen HC, Baik B, Yoon S, Park T, Nam D. Benchmarking integration of single-cell differential expression. *Nature Communications*. 2023;14(1):1570.

16. Menden K, Marouf M, Oller S, et al. Deep learning–based cell composition analysis from tissue expression profiles. *Science advances*. 2020;6(30):eaba2619.

17. 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.

18. 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.

19. 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.

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

21. 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.

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

23. Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*. 2019.

24. Yin L, Liu M, Shi Y, Qiu J, So H. Direct causal variable discovery leveraging the invariance principle: Application in biomedical studies. .

25. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. . 2016:785–794.

26. Sturm G, Finotello F, Petitprez F, et al. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics*. 2019;35(14):i436–i445.

27. 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.

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

29. 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. *Nat Biotechnol*. 2024;42(2):247–252.

30. 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.

31. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573–3587. e29.

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.

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

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. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications*. 2019;10(1):380.

37. Rajkowska G, Miguel-Hidalgo J. Gliogenesis and glial pathology in depression. *CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders)*. 2007;6(3):219–233.

38. Holter MM, Saikia M, Cummings BP. Alpha-cell paracrine signaling in the regulation of beta-cell insulin secretion. *Frontiers in endocrinology*. 2022;13:934775.

39. Son J, Accili D. Reversing pancreatic β-cell dedifferentiation in the treatment of type 2 diabetes. *Exp Mol Med*. 2023;55(8):1652–1658.

40. 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.

41. Jakab J, Miškić B, Mikšić Š, et al. Adipogenesis as a potential anti-obesity target: A review of pharmacological treatment and natural products. *Diabetes, Metabolic Syndrome and Obesity*. 2021:67–83.

42. Romao JM. Adipogenesis and obesity. In: *MicroRNA in regenerative medicine.* Elsevier; 2015:539–565.

43. Tang W, Zeve D, Suh JM, et al. White fat progenitor cells reside in the adipose vasculature. *Science*. 2008;322(5901):583–586.

44. 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;23:667–684.

45. Matejuk A, Ransohoff RM. Crosstalk between astrocytes and microglia: An overview. *Frontiers in immunology*. 2020;11:1416.

46. Ellulu MS, Patimah I, Khaza'ai H, Rahmat A, Abed Y. Obesity and inflammation: The linking mechanism and the complications. *Archives of medical science*. 2017;13(4):851–863.

47. 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.

48. Xie Y, Xiao L, Chen L, Zheng Y, Zhang C, Wang G. Integrated analysis of methylomic and transcriptomic data to identify potential diagnostic biomarkers for major depressive disorder. *Genes*. 2021;12(2):178.

49. Bhat M, Pasini E, Das A, et al. Diabetogenic effects of immunosuppression: An integrative analysis. *Transplantation*. 2020;104(1):211–221.

50. Zhang Y, Sowers JR, Ren J. Targeting autophagy in obesity: From pathophysiology to management. *Nature Reviews Endocrinology*. 2018;14(6):356–376.

51. Soussi H, Clément K, Dugail I. Adipose tissue autophagy status in obesity: Expression and flux—two faces of the picture. *Autophagy*. 2016;12(3):588–589.

52. Teng M, Wu S, Er L, Hsu L, Chou H, Ko Y. LIPC variants as genetic determinants of adiposity status, visceral adiposity indicators, and triglyceride-glucose (TyG) index-related parameters mediated by serum triglyceride levels. *Diabetology & Metabolic Syndrome*. 2018;10:1–10.

53. Long T, Lu S, Li H, et al. Association of APOB and LIPC polymorphisms with type 2 diabetes in chinese han population. *Gene*. 2018;672:150–155.

54. Jahangir E, De Schutter A, Lavie CJ. The relationship between obesity and coronary artery disease. *Translational Research*. 2014;164(4):336–344.

55. Li J, Ma Q, Ai M. Identification and analyses of crucial genes associated with pathogenesis of major depressive disorder. *Psychiatry and Clinical Psychopharmacology*. 2023;33(4):264.

56. Zhao F, Huang F, Tang M, et al. Nodal induces apoptosis through activation of the ALK7 signaling pathway in pancreatic INS-1 β-cells. *American Journal of Physiology-Endocrinology and Metabolism*. 2012;303(1):E132–E143.

57. Saisho Y. B-cell dysfunction: Its critical role in prevention and management of type 2 diabetes. *World journal of diabetes*. 2015;6(1):109.

58. Kim S, Shou J, Abera S, Ziff EB. Sucrose withdrawal induces depression and anxiety-like behavior by Kir2. 1 upregulation in the nucleus accumbens. *Neuropharmacology*. 2018;130:10–17.

59. Gu X, Ke S, Wang Q, et al. Energy metabolism in major depressive disorder: Recent advances from omics technologies and imaging. *Biomedicine & Pharmacotherapy*. 2021;141:111869.

60. Rajkowska G, Miguel-Hidalgo J. Gliogenesis and glial pathology in depression. *CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders)*. 2007;6(3):219–233.

61. Zhou B, Zhu Z, Ransom BR, Tong X. Oligodendrocyte lineage cells and depression. *Mol Psychiatry*. 2021;26(1):103–117.

62. Prentki M, Madiraju SM. Glycerolipid metabolism and signaling in health and disease. *Endocr Rev*. 2008;29(6):647–676.

63. 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 Biol*. 2022;23(1):220.