Improved estimation of cell type-specific gene expression through deconvolution of bulk tissues with matrix completion

Weixu Wang[1,#], Jun Yao[1,#], Yi Wang[2], Chao Zhang[3], Wei Tao[3], Jiahua Zou[4], Ting Ni[1,*]

[1] State Key Laboratory of Genetic Engineering, Collaborative Innovation Center of Genetics and Development, Human Phenome Institute, School of Life Sciences and Huashan Hospital, Fudan University, Shanghai, 200438, P.R. China

[2] Ministry of Education (MOE) Key Laboratory of Contemporary Anthropology, Human Phenome Institute, School of Life Sciences, Fudan University, Shanghai, 200438, P.R. China

[3] MOE Key Laboratory of Cell Proliferation and Differentiation, School of Life Sciences, Peking University, Beijing, 100871 China

[4] Guangdong Provincial Key Laboratory of Bioengineering Medicine, National Engineering Research Center of Genetic Medicine, Institute of Biomedicine, College of Life Science and Technology, Jinan University, Guangzhou 510632, P.R. China

[#] These authors contributed equally

* Correspondence: tingni@fudan.edu.cn

Abstract

Cell type-specific gene expression (CSE) brings novel biological insights into both physiological and pathological processes compared with bulk tissue gene expression. Although fluorescence-activated cell sorting (FACS) and single-cell RNA sequencing (scRNA-seq) are two widely used techniques to detect gene expression in a cell type-specific manner, the constraints of cost and labor force make it impractical as a routine on large patient cohorts. Here, we present ENIGMA, an algorithm that deconvolutes bulk RNA-seq into cell type-specific expression matrices and cell type fraction matrices without the need of physical sorting or sequencing of single cells. ENIGMA used cell type signature matrix generated from either FACS RNA-seq or scRNA-seq as reference, and applied matrix completion technique to achieve fast and accurate deconvolution. We demonstrated the superior performance of ENIGMA to previously published algorithms (TCA, bMIND and CIBERSORTx) while requiring much less running time on both simulated and realistic datasets. To prove its value in biological discovery, we applied ENIGMA to bulk RNA-seq from arthritis patients and revealed a pseudo-differentiation trajectory that could reflect monocyte to macrophage transition. We also applied ENIGMA to bulk RNA-seq data of pancreatic islet tissue from type 2 diabetes (T2D) patients and discovered a beta cell-specific gene co-expression module related to senescence and apoptosis that possibly contributed to the pathogenesis of T2D. Together, ENIGMA provides a new framework to improve the CSE estimation by integrating FACS RNA-seq and scRNA-seq with tissue bulk RNA-seq data, and will extend our understandings about cell heterogeneity on population level with no need for experimental tissue disaggregation.

Introduction

Each cell type has unique role in the ecosystem of tissues. In recently 20 years, from microarray to Next Generation Sequencing (NGS), gene expression quantification has been a useful resource for understanding the complex regulation relate to different diseases or phenotypes[1, 2]. There are a lot of large cohort study, like TCGA[3] or GTEx[4], to build the connection of the expression molecular signature with the various phenotypic data. The sequenced samples are the collection of various cell types (bulk tissue), therefore the expression profile of a given sample is the convolution of gene expression from heterogenous cell types. It has been shown that cell type-specific gene expression alteration can contribute to our understanding of the principles of biological regulation. Therefore, it is necessary to isolate cell type-specific gene expression profile from mixed bulk tissue.

To overcome the challenge, researchers have pursued several advanced techniques such as fluorescence-activated cell sorting (FACS)[5] and single-cell RNA-sequencing (scRNA-seq)[6] to quantify cell type-specific expression (CSE) profile on sample level. These technologies bring new insights to us about the landscape within the tissues, but such data still has some limitations. For FACS dataset, it relies on small combinations of preselected marker genes, limiting the cell types that could be studied. For scRNA-seq dataset, although it could collect a large number of cells from tissues, and characterize a much

richer heterogenous cell state information compared with FACS, its relatively high cost property does not allow the assessment of tissues from hundreds of samples, thus limits its application on large cohort cell-type-specific phenotype-gene expression association analysis[7].

Previous researches have pursued to integrate FACS and single-cell RNA-seq datasets to help us deconvolute the bulk RNA-seq dataset. The methods like CIBERSORT[8], TIMER[9], BSEQ-sc[10] and MuSiC[11], used several regression models to deconvolute the bulk RNA-seq dataset into cell type fraction matrix, and helped us to infer the proportion information of each cell type. Although these methods have been widely applied in previous researches to help find cell type fractional variation, they could not provide the cell type-specific expression profile on sample level. Besides, there were methods that had been developed for deconvolving cell type-specific gene expression on sample level (e.g. bMIND[12] and CIBERSORTx[8]). bMIND uses full Bayesian model, bases on cell type reference matrix derived from single cell RNA-seq as the prior information, and applies Markov chain Monte Carlo (MCMC) to estimate parameters[12]. As bMIND estimates the cell type-specific expression for each gene independently, it ignores the nature of latent correlation among genes, which is important for gene co-expression network analysis[13]. CIBERSORTx requires case and control bulk RNA-seq datasets to estimate CSE profile based on non-negative least squares algorithm[7], therefore it cannot be broadly applied to unpaired bulk RNA-seq data. In addition, these two methods are all time-consuming, and the expensive time consumption leads the parameter tuning and optimization impracticable.

To address these deficiencies, we developed a deconvolution method based on regularized matrix completion (ENIGMA) to refine the estimation of CSE profile for each bulk sample in an extremely short time. To make the inferred CSE reliable, we minimize the distance between each inferred CSE average expression level and the reference expression profile. We validated its inference accuracy on both simulated and real datasets through showing its superior performance compared with other state-of-the-art methods. We proved that ENIGMA could give better CSE estimation as reflected by its inferred gene-level or sample-level correlation with simulated ground truth CSE under different expression noise and reference noise setting. Also, it showed better cell type-specific differential expression gene (DEG) identification compared with other algorithms. Moreover, we applied ENIGMA on three different scenarios, and proved that it could identify unique cell type-specific expression pattern and co-expression network. We identified the differentiation trajectory of monocyte in arthritis patients. Meanwhile, we found a senescence-associated gene co-expression module in pancreas islet beta cells, which is reported relevant to type 2 diabetes (T2D) patients.

Result

The overview of ENIGMA

We illustrated the general strategy of ENIGMA about performing bulk gene expression deconvolution to infer CSE on sample-level. We regarded this task as a matrix completion problem, with bulk gene expression matrix could be represented as the linear combination of several cell type-specific expression matrix. The linear combination coefficient is cell

type fraction (θ_i), which could be estimated through using a cell type fraction estimation algorithm or experimentally measurements through cell sorting. Our goal is to minimize the distance between observed bulk gene expression matrix and reconstituted bulk expression matrix through linear combination of inferred CSE (Fig. 1a).

Besides the minimization of the main object function, we also added two additional constraint terms to make the inference more robust. First, we considered each aggregated CSE should be as close as to the corresponding reference, which could also be regarded as the prior expression information derived from reference matrix. Second, we used regularization term to constraint the rank of inferred CSE. In matrix completion problem, low-rank property of inferred signal is ubiquitous in real-world application. Also, previous analysis of the spectrum of bulk DNA methylation profile enlighten that the mixing profile has the latent variation/eigenvector that associates with the cell type fraction variation among samples[14]. Therefore, assuming the relatively low-rank property in CSE is consistent with our known knowledge.

ENIGMA has three main steps (Fig. 1b). First, ENIGMA requires cell type reference expression matrix (signature matrix), which could be derived from either FACS RNA-seq or scRNA-seq datasets through calculating the average expression value of each gene from each cell type. Previous researches have widely used reference matrix curated from different platforms, for instance, Newman et al. used LM22 immune signature matrix which derived from microarray platform to deconvolute bulk RNA-seq dataset[8]. However, we have to note that use references from different platforms would introduce unwanted batch effect between reference and bulk RNA-seq matrix, especially for the reference matrix derived from low coverage scRNA-seq dataset. To overcome this challenge, we used previously presented method that is specifically designed for correcting batch effect among bulk RNA-seq matrix and reference matrix (Method)[7]. Second, ENIGMA applied robust linear regression model[15] to estimate each cell type fractions among samples based on reference matrix derived from the first step. Third, based on reference matrix and cell type fraction matrix estimated from step 1 and step 2, ENIGMA applied constrained matrix completion algorithm to deconvolute bulk RNA-seq matrix into CSE on sample-level. In order to constraint the model complexity to prevent overfitting, we proposed to use two different norm penalty functions to regularize resulted CSE. Finally, the returned CSE could be used to identify cell type-specific DEG, visualize each gene's expression pattern on the cell type-specific manifold space (e.g. t-SNE[16], UMAP[17]), and build the cell type-specific co-expression network to identify modules that relevant to phenotypes of interest.

ENIGMA gives better estimation of CSE profile on sample-level

We benchmarked the performance of ENIGMA with other CSE inference tools with simulated datasets and realistic datasets. First, we tested whether ENIGMA could yield a better gene expression inference compared with other algorithms. For comparison, we employed two other methods: 1) TCA, a frequentist approach, designed for bulk DNA methylation data, but also applicable for CSE estimation[18]; 2) bMIND, a full Bayesian model, modeling the gene expression as Gaussian distribution with the prior information from reference matrix, and using MCMC to learn the parameters[12]. We compared the performance of two versions of ENIGMA, which correspond to their norm function for

regularization (see Method for details), to these two algorithms. An ideal CSE inference algorithm should recover the cell type-specific sample variation and gene variation. We tested this by generating simulated bulk RNA-seq dataset through linear combination of the simulated cell type-specific expression profiles and simulated observation noise with four different levels (Method). After reconstructing CSE based on each algorithm, we calculated sample-level correlation and gene-level correlation between inferred CSE and ground truth CSE in each simulated cell type.

The correlation per sample and per gene between ENIGMA inferred CSE and ground truth CSE was comparable to those from TCA and bMIND at low-noise levels and was higher than those from TCA and bMIND at high-noise levels (Fig. 2a,b). It also showed that the performance of L2-max norm base regularization (default) ENIGMA is better than trace norm base regularization ENIGMA at different noise levels (Fig. 2a,b).

To further support the superiority of our algorithm, we benchmarked the CSE inference on pseudo bulk RNA-seq datasets generating from scRNA-seq of head and neck squamous cell carcinoma (HNSCC)[19] and melanoma scRNA-seq dataset[20] (see Methods for details). The results still indicated that ENIGMA performed either comparable or better than TCA and bMIND on seven cell types (Fig. 2c, Supplementary Fig. 1a, b).

We then extended our analysis to the real bulk RNA-seq dataset with corresponding FACS RNA-seq as the ground truth. Previous study used surgically resected primary non-small cell lung cancer (NSCLC) tumor biopsies (n = 20 patients) to generate RNA-seq libraries of four major subpopulations purified by FACS: epithelial/cancer, immune, endothelial and fibroblast cells[21]. After deriving a signature matrix from one patient, we applied ENIGMA to perform in-silico expression purification on other 19 patients, and compared ENIGMA with TCA, bMIND and additional algorithms, which merely deconvolute limited genes or cell types (CIBERSORTx[7], ISOpure[22] and Demix[23]). ENIGMA has higher Spearman correlation between inferred CSE and ground truth (FACS-purified tumor cells) CSE per sample than all other methods (Supplementary Fig. 1c, d).

In CSE inference, recover the correlation structure among samples is also a very important aspect that could help us to identify the latent higher-order data structure in sequenced cohort on cell type level. However, this important aspect is ignored by previous CSE inference strategies[12, 18]. To benchmark this aspect, we simulated the CSE which has latent sample-wise correlation structure that could be organized as a transition trajectory on the principal component space, and we also attached each sample a pseudo-time label to show the relative order information[24]. Then, we mixed this CSE with other three simulated CSE matrices with specific proportion to build a pseudo-bulk expression profile. After deconvolution, the result showed that both versions (L2-max norm and trace norm) of ENIGMA could recover the trajectory structure of the specific CSE, while TCA and bMIND totally lost this information (Fig. 2d). Taken together, these results demonstrate that ENIGMA could infer CSE with better estimation on both sample-level and gene-level concordance, and could recover the correlation structure of CSE better than previous methods.

The robustness and scalability of ENIGMA

As reference matrix is extremely important for CSE inference and different sources of

reference matrices (library type, sequence depth, batches, etc.) may introduce variability of the reference, we therefore evaluated the influence of the quality of reference matrix by adding variable noises at different levels in the simulated data. We found that although sample-level correlation between inferred and ground truth CSE was not distinctly affected, the gene-level correlation was clearly influenced by the noise of reference matrix (Supplementary Fig. S2a), especially for TCA and bMIND. The maximum gene-level correlation coefficient decrease could reach 0.26 in TCA and 0.32 in bMIND, while the maximum correlation coefficient decreases just reach 0.2 and 0.06 in ENIGMA and ENIGMA (trace norm), respectively (Supplementary Fig. S2b). We noticed that the quality of reference matrix also seriously influenced the prediction accuracy of cell type fraction matrix, with the average correlation coefficient drops from 0.94 to 0.4 (Supplementary Fig. S2c). We also compared the all-gene level correlation coefficient decrease among four methods, and found that both ENIGMA and ENIGMA (trace norm) showed significantly better robustness compared with TCA and bMIND (*P* value < 2.2e-16, Supplementary Fig. S2d).

We also tested the influence of the number of latent cell types in bulk gene expression matrix to the CSE inference through admixing a range of cell type numbers to build bulk gene expression matrices. The results indicated that both versions of ENIGMA performed consistently better than TCA and bMIND regardless the number of cell types we simulated (Supplementary Fig. S3).

Efficiency of computing is very important for CSE inference since faster runtime will greatly facilitate parameter tuning and optimization especially analyzing large cohort datasets. To evaluate ENIGMA's computational performance, we measured total runtime usage of ENIGMA versus other methods. We created five benchmark datasets with 100, 500, 1000, 5000, 10000 genes with fixed sample size (100 samples), and reported the runtime for each condition. The results showed that ENIGMA's runtime scaled well for all datasets (ranging from 4 to 24 seconds), while runtime of the other two methods (TCA and bMIND) showed exponential increase when gene numbers rise from 100 to 10000 (Supplementary Fig. S4). These results demonstrate that ENIGMA is computationally efficient and capable of analyzing all genes with large sample size on servers or even personal computers.

ENIGMA reconstructed CSE benefits cell type fraction estimation

Previous researches about cell type fraction estimation use a common assumption that the CSE of each sample is the same. Therefore, the cell type fraction could be estimated through performing regression model on each sample with a single cell type-specific reference matrix[8, 25]. This assumption obviously ignores the cell type-specific variation across the samples, as in realistic scenario, each sample's CSE shall be different. Using ENIGMA, we could estimate CSE on sample level. To test the ENIGMA estimated CSE could benefit the cell type fraction estimation, we started with scRNA-seq dataset generated from NSCLC[26], and generated pseudo-bulk matrix and reference matrix based on independent samples (see Methods for details). After running ENIGMA on pseudo-bulk matrix, we independently performed robust linear regression on each sample using their corresponding sample-level CSE as the reference matrix. We could find that using

ENIGMA inferred CSE, almost all cell types showed improved cell type fractions estimation, as reflected by their increased Pearson correlation with the ground truth cell type fractions (Supplementary Fig. S5a,b; paired Wilcoxon test, *P* value = 0.001953).

We also validated the improvement on realistic dataset with known cell type fractions determined by flow cytometry and fluorescence immunophenotyping experiment[8]. We used previously generated leukocyte signature matrix (LM22)[8] to distinguish six cell type (B cells, CD4 T cells, CD8 T cells, NK cells, Neutrophils and Monocyte) and regarded it as the reference matrix on a validation cohort of bulk RNA-seq profiles of blood obtained from 12 healthy adults[8]. We found that following application of ENIGMA, deconvolution results substantially improved and compared favorably with ground truth cell proportions (Supplementary Fig. S5c-e; paired Wilcoxon test, *P* value < 0.0004).

ENIGMA benefits cell type-specific-DEG (CTS-DEG) identification

A well inferred CSE also could help us to identify cell type-specific differentially expressed genes (CTS-DEGs) associated with cohort phenotypes[7, 12]. To evaluate ENIGMA's ability to discover the CTS-DEG, we generated simulated dataset with known DEG for each cell type, and calculated the sensitivity and specificity for the identified CTS-DEGs. Besides TCA and bMIND, we also ran differential expression analysis on raw simulated bulk dataset as baseline to make comparison.

We applied similar benchmark evaluation as Rahmani et al[18], and assessed sensitivity and specificity as the function of effect size (the magnitude of expression differences). We found that ENIGMA showed better sensitivity at low effect size compared with TCA and bMIND (Fig. 3a,b). Meanwhile, ENIGMA (trace norm) showed consistently higher sensitivity compared with other algorithms in all scenarios (Fig. 3a,b), indicating that ENIGMA tends to identify more true CTS-DEG compared with other algorithms. To reveal that ENIGMA could refine the phenotype-associated expression difference, we compared the alteration (expression difference) we estimated from mixed expression profile and ENIGMA inferred CSE with the ground truth in each cell type. We found that the correlation between gene expression alteration estimated from ENIGMA inferred CSE and ground truth was better than the correlation between gene expression alteration estimated from mixed expression profile and ground truth (Fig. 3c,d), supporting that ENIGMA could find the cell type-specific phenotype associated alterations that diminished or enwombed in mixed expression profile by the sample-wise mixing proportions of different cell types. We also performed power analysis through assessing the influence of sample size to the sensitivity and specificity of CTS-DEG identification, and reflected both of them using the Area Under Curve (AUC) value. The results showed that ENIGMA (trace norm) and bMIND tended to perform consistently better than other algorithms across different sizes of samples (Supplementary Fig. S6). Moreover, ENIGMA (trace norm) showed better performance (higher AUC) than other methods when effect size increases (Supplementary Fig. S6).

While the simulation study could give us a golden standard of CTS-DEG that could help us to benchmark the algorithms, in realistic application, the CTS-DEG inference would be influenced by many different factors (e.g. the number of latent cell types, the quality of reference expression profile). To show that ENIGMA has potential in identifying CTS-DEG

genes in realistic dataset, we used RNA-seq dataset of flow-sorted malignant cells, endothelial cells, immune cells, fibroblasts, and bulk samples from freshly resected human primary non-small-cell lung tumors to benchmark CTS-DEG identification (n = 26 patients)[21]. To construct reference expression profile, we used previously developed algorithm[7] to construct a batch effect corrected reference profile based on single cell NSCLC RNA-seq dataset from six donors[26] (Fig. 3e). Previous research indicated that both malignant cells and fibroblast cells of lung squamous cell carcinoma (LUSC) and lung adenocarcinomas (LUAD) had distinct expression pattern differences[7, 21]. Therefore, we benchmarked our CTS-DEG identification in these two cell types. We used concordance index based on previous work[7] to reflect the concordance between our predicted CTS-DEG expression trend and ground truth and then generated its statistics distribution using a bootstrap method to evaluate the performance in identifying DEGs in a cell type-specific manner (Method). The result showed that expression concordance between ENIGMA-inferred CTS-DEGs' CSE profile and flow-sorted expression profile was significantly higher than the concordance between bulk RNA-seq derived DEGs' profile and flow-sorted expression profile (Fig. 3f, Supplementary Fig. 7a,b). Such results were further corroborated by scRNA-seq dataset of NSCLC tumors (Supplementary Fig. 7c,d). To make the analysis more comprehensive, we also defined the pseudo-ground truth CTS-DEGs based on flow-sorted gene expression profile and scRNA-seq with very strict criteria (|FoldChange| > 4 & FDR < 0.01). Based on this, we analyzed the CTS-DEG identification sensitivity, and plotted its function as the number of top DEG for each method. We found that both two versions of ENIGMA showed consistently higher sensitivity compared with all other algorithms (Fig. 3g). To further support the sensitivity of ENIGMA, we inspected the significances (*P* value) of those pseudo-ground truth DEGs of tumor cells in ENIGMA-inferred CSE and bulk RNA expression profile. The results showed that these genes were more significance in ENIGMA-sorted tumor expression profile compared with bulk RNA-seq dataset regardless up-regulated or down-regulated (Fig. 3i). These above results support that ENIGMA improves the identification of CTS-DEG.

ENIGMA reveals monocyte pseudo-differentiation trajectory in arthritis patients

We then explored the potential applications of ENIGMA for characterizing cellular heterogeneity in synovial biopsies from patients with arthritis. Rheumatoid arthritis (RA) is an autoimmune disease with chronic inflammation in the synovium of the joint tissue[27]. This inflammation leads to joint destruction, disability and shortened life span[27]. Studying the histological change from osteoarthritis (OA) to rheumatoid arthritis (RA), and defining key cellular subsets with their activation states in the inflamed tissue are critical to define new therapeutic targets for RA[28]. Using flow-sorted RNA-seq profiles from RA to build a reference matrix[29], we inferred CSE on sample level to dissect four major cell types (includes Fibroblast, Monocyte, B cell, T cell) from the transcriptomes of 174 bulk RA & OA profiled by an independent study[30] (Fig. 4a). We benchmarked our *in-silico* purification through evaluating the CTS-DEG expression change concordance with the external single cell RNA-seq datasets generated from synovial biopsies of patients with RA and OA[29]. In all four cell types, ENIGMA exhibited superior performance compared with other algorithms in relation to the DEG expression concordance with the single cell RNA-seq dataset (Fig.

4b, Supplementary Fig. S8b). Meanwhile, we found that using batch effect correction strategy could also improve the DEG concordance, suggesting the importance to remove the confounder effect between reference and bulk expression matrix. Notably, we repeated our analysis through using reference matrix derived from single cell RNA-seq data from RA and observed similar results (Supplementary Fig. S8a).

Within *in-silico* purified cell subsets, we discovered many significant CTS-DEGs in monocytes and fibroblasts that distinguish two disease states of arthritis (Fig. 4c). For instance, in fibroblasts, we found that a lot of interferon response associated genes were up-regulated in RA patients. In monocytes, we found that some chemokine and cytokine associated genes (e.g. *IL1B*, *IL1RN* and *CCL3*) were up-regulated in RA patients (Fig. 4c). They play a crucial role in homeostasis, generation of cellular and humoral immune responses, as well as pathologic immune contribution in rheumatoid arthritis[29, 31-33]. The abundance and activation of macrophages in the inflamed synovial membrane significantly correlates with the severity of RA[34]. In the RA synovial membrane, it is known that a pivot differentiation step is from immigrated monocytes to mature macrophages[35, 36]. Therefore, studying the sample-wise co-expression structure of purified CSE of monocyte may give us a deep understanding of pathogenetic stage of arthritis. Using t-SNE to visualize the CSE of monocyte cell and cluster the samples, we observed a pseudo-differentiation trajectory related to pathogenetic progression of RA, which includes Development state, Median state and Inflammatory state (Fig. 4e). Besides, we found that OA patients were enriched in Developmental and Median states (Supplementary Fig. S9b). Using Gene Ontology (GO) analysis, highly expressed genes in each state were enriched with a number of biologically significant pathways related to the progression (Fig. 4f). The first state overexpressed genes related to histone modification, covalent chromatin modification and dendrite development (Fig. 4f). Highly expressed gene in the median stage related to RNA splicing and processing, while the final state overexpressed genes related to inflammatory pathways like neutrophil activation involved in immune response and protein localization to endoplasmic reticulum (Fig. 4f). The inflammatory state was also supported by classical macrophage markers (*CD86* and *CD68*)[37] (Fig. 4g). Also, the primary state was supported by some genes related to primary monocyte, like *PPM1A*, which plays the role in inhibiting monocyte differentiate to macrophage[38] (Fig. 4g). In contrast, raw bulk RNA-seq data could not identify this pseudo-trajectory structure and related states (Supplementary Fig. S9a). In conclusion, ENIGMA reveals pseudo-differentiation trajectory that could reflect monocyte to macrophage transition in arthritis patients.

ENIGMA identifies a senescence-associated gene co-expression module in beta cells from T2D patients

ENIGMA could also help us to reveal cell type-specific co-expression modules, and measure their association with phenotypes in large patient cohort. To further demonstrate and evaluate ENIGMA, we extended application with a well-studied tissue, pancreatic islet, which plays important roles in metabolic hemostasis and whose dysfunction leads to diabetes[39]. We applied ENIGMA to bulk pancreatic islet RNA-seq samples from 89 donors to estimate proportion and CSE for each cell type[40]. Next, we characterized their associations with hemoglobin A1c (HbA1c) level, an important biomarker for type 2

diabetes (T2D)[41]. After constructing cell type reference matrix from published pancreatic islet scRNA-seq dataset from normal donors, we dissected bulk RNA-seq dataset into five CSE profiles at sample level (Fig. 5a). We validated our *in-silico* purification through evaluating all gene expression concordance with independently published pancreatic islet scRNA-seq dataset from both T2D and normal donors[42], and ENIGMA showed superior expression concordance with independent realistic dataset compared with other algorithms (Supplementary Fig. S10).

Pancreatic beta cells, which secrete insulin, are gradually lost during T2D[43, 44]. As our expectation, ENIGMA detected the significant negative association of beta cell proportion with HbA1c level (Fig. 5b, P value < 0.05). Recent works also suggested that insulin resistance induced the expression of aging markers, and beta cell aging (or senescence) could accelerate the progression toward diabetes[43, 45, 46]. We further asked whether ENIGMA could recover the latent senescence signal of beta cell in T2D patients. We constructed the expression-based senescence score depends on the senescence-associated genes downloaded from CellAge[47] to predict the degree of senescence for each sample (Method), and applied the score on the beta cell CSE. We found that senescence score of beta cells showed a significantly positive correlation with HbA1c level and Body mass index (BMI) (Fig. 5b, P value < 0.05), consistent with the previous finding that beta cell senescence increases with T2D and BMI[46]. To test whether senescence specifically happens in insulin secreting cell type, we applied senescence score on *in-silico* purified CSE and bulk RNA-seq samples (Supplementary Fig. S11). Notably, although we found that the senescence score in bulk RNA-seq sample showed positive correlation with HbA1c level and BMI (Supplementary Fig. S11), in CSE, only beta cells out of five cell types show significantly positive correlation with both HbA1c level and BMI (Supplementary Fig. S11). To show this result could be reproduced, we constructed the second senescence score based on senescence-associated secretory phenotype (SASP) signatures[48], and the new senescence score of beta cell showed the strongest positive correlation with both HbA1c level and BMI than that of other cell types (Supplementary Fig. S12).

Gene co-expression networks capture biologically important patterns in gene expression data, enabling functional analyses of genes, discovery of biomarkers, and interpretation of phenotype associated biological variation. Previous network analyses were limited to the total gene expression levels in a whole tissue, while using ENIGMA could help us to deconvolute a mixing expression signal into a set of cell type-specific signals, further enable us to build cell type-specific gene co-expression network. We applied ENIGMA and weighted gene co-expression network analysis (WGCNA)[13] method to find representative gene co-expression modules for each cell type. By performing network module discovery on all CSE, we identified several cell type-specific modules (Fig. 5c, Supplementary Fig. 13). For instance, in acinar cell, we identified the modules enriched with pathways like digestion and vitamin metabolic process, in line with the fact that pancreatic acinar cells are highly specialized exocrine factories that produce copious amounts of digestive enzymes for intestinal digestion[49]. We also identified synapse organization and synapse assembly pathways in delta cells, consistent with the idea that delta cells highly express the marker gene *GHSR* (ghrelin receptor), which is frequently involved in the process of synaptic organization and transmission[42]. These results support that ENIGMA could

identify cell type-specific co-expression modules, making it possible to identify cell specific modules that related to the pathogenesis of complex diseases.

Next, we built the gene co-expression network in beta cells to study their role in the pathogenesis of T2D (Fig. 5d). We identified six modules in beta cells that correlated with one of the phenotypes (HbA1c, BMI or Age) and four of them (skyblue, darkgreen, blue and darkolivergreen) enriched with GO pathways (Fig. 5e). In blue and darkolivegreen modules, we found that vesicle-mediated transportation and cycling associated pathways were significantly enriched (Fig. 5f), and these two modules were negatively correlated with HbA1c and BMI (Fig. 5e), suggesting that both HbA1c and BMI factors may link to the deficiency of the secreting ability of beta cells. Meanwhile, we identified that the darkgreen module, which was positively associated with HbA1c and BMI, enriched a lot of pathways that associated with the regulation of apoptotic signaling pathways and wound healing (Fig. 5g), suggesting the dysfunction of beta cell in T2D patients. We also calculated the darkgreen module activity on all cell type-specific expression profile, and found that only in beta cell that this module was positively associated with HbA1c across all five cell types (Fig. 5h, Supplementary Fig. S14). We also performed hypergeometric test and found that this module was significantly overlapped with senescence-associated gene signatures (Fig. 5i, P value = 0.001). Performing regression analysis between gene significance (which measured as the correlation significance with the HbA1c level) and gene connectivity (importance measurement of each gene in the module) showed a significant positive correlation, suggesting the hub genes of the module strongly connected to the HbA1c levels (Fig. 5j). The top 10 hub genes in the darkgreen module also recovered previously validated genes that play a role in T2D pathogenesis (Fig. 5k). For example, *TNFSF10*, which is part of TRAIL (tumor necrosis factor related apoptosis-inducing ligand), a type II transmembrane protein and member of TNF-ligand family, has been extensively studied for its preferential ability to induce apoptosis of cancer cell[50, 51]. Growing lines of evidence support its involvement in the development of obesity and diabetes[52], and *TNFSF10* corporates with Caspase family protein to activate intrinsic mitochondrial pathway and cell apoptosis[52]. The hub genes also included some apoptosis and senescence-related cysteine peptidase (*CAPS7*, *CAPS3*)[53-55], further support the role of extrinsic signaling induced cell senescence and apoptosis in beta cell dysfunctions.

Discussion

We developed a matrix completion-based algorithm, ENIGMA, for *in-silico* dissecting tissue bulk RNA-seq dataset through borrowing prior reference information derived from either FACS RNA-seq or scRNA-seq datasets. Comparing with previously proposed methods, ENIGMA used matrix completion method that jointly optimize all genes/samples' reconstructed CSE profile, gave its capability to preserve or recover sample/gene correlation structure. Meanwhile, its great computational efficiency compared with previous algorithms make it possible to be applied on large cohort study. We conducted extensive benchmark on simulation datasets and application on realistic datasets to compare ENIGMA with state-of-the-art (SOTA) methods, demonstrating that ENIGMA improved the CSE estimation accuracy and recovered specific sample-wise correlation structure. Also,

we validated that ENIGMA is relatively robust to the number of factors that may influence the inferred CSE quality compared with other SOTA methods. What's more, the required running time of ENIGMA is dozens of times faster than other SOTA methods.

We also showed that using ENIGMA could identify CTS-DEGs from bulk RNA expression profile, and showed its advanced specificity and sensitivity compared with other algorithms. Nonetheless, ENIGMA also had limitations in CTS-DEG predictions. For instance, for those cell type which has rare abundance in cohort, both L2-max norm-based and trace-norm based ENIGMA would not conduct accurate estimation, thus influence the DEG detections. Such issue could not be addressed by other methods either. Besides, ENIGMA was sensitive to the phenotype associated variations, therefore it tended to call more CTS-DEGs compared with other methods, which may suggest that it has relatively higher false discovery rate compared with some SOTA methods specifically design for cell type-specific variation detection (e.g. CellDMC)[56]. But this drawback could be improved through controlling the gene effect size. Also, in the realistic application on NSCLC dataset analysis, after controlling the number of predicted CTS-DEGs according to the effect size (expression difference), ENIGMA showed improved precision compared with other CSE inference methods (Fig. 3g).

ENIGMA also enabled the identification of meaningful cell type-specific pseudo-trajectory that may help to explain the progression of complex diseases. In both our benchmark simulation and application on arthritis patients, we showed that ENIGMA could reconstruct the pseudo-trajectory of CSE. However, we found that previously developed CSE inference methods could not capture the trajectory structure. The power of preserving correlation structure also gave ENIGMA benefits to identify gene co-expression modules, and perform association study with the phenotype data. With these features, we anticipated that ENIGMA would improve our understanding of the underlying mechanism of pathogenicity of complex diseases relying on specific cell types.

ENIGMA currently relied on flow-sorted RNA-seq or scRNA-seq as the reference expression profile for deconvolution. The cross-platform between reference profile and bulk RNA-seq expression measurements may bring unwanted technique effect that distort the estimation. To prevent the technique effect, we applied S-mode and B-mode methods to correct batch effect[7]. The applications on real datasets also supported that these correction methods could benefit the CSE inference (Supplementary Fig. S8). Another key factor that may influence the performance of ENIGMA was the choose of parameters. Fortunately, the short runtime of ENIGMA made the tuning of parameters more feasible. We used grid-search algorithm or genetic algorithm to help users to tune specific parameters on their own dataset.

In summary, ENIGMA represented a broadly applicable algorithm that could be used to deconvolute bulk RNA-seq dataset into cell type-specific expression profile, helping researchers to study the underlying heterogeneity in tissues, and systemically determine correlation between CSE and a variety of phenotypes. Therefore, ENIGMA could be utilized as an efficient tool for biological discovery, and provide a framework to perform integrative analysis on scRNA-seq and bulk RNA-seq dataset. We believe that this approach has great potential to enhance the analysis on multicellular organisms, and facilitate the applications of mixture RNA-seq in the era of cell type-specific expression

study.

Methods

The mathematics model of ENIGMA

The input of ENIGMA includes the bulk RNA-seq data and reference expression data. A matrix, O, represents expression profile from bulk RNA-seq data with columns representing n samples and rows representing m genes. The reference matrix, R, represents the expression levels of all genes across different cell types of interest. Our goal is to completely deconvolute bulk expression data into series of cell type-specific expression matrix, $(X_i)_{i=1\ldots ct} \in R^{m\times n}$, in which ct represents the number of cell types.

In previous research, Jiebiao Wang et al proposed to use a full Bayesian model to model the relationship between bulk gene expression and cell type-specific gene expression[12].

$$O_{gs} = W_{s\cdot}X_{gs\cdot} + \sigma_{gs}, \quad X_{gs\cdot} \sim N(\alpha_g, \Sigma_g), \sigma_{gs} \sim N(0, \sigma_g I),$$

in which $W \in R^{m\times ct}$ represents cell type fraction matrix, $X_{gs\cdot}$ represents sample s level of gene g expression value across different cell types, and the parameter Σ_g represents the expression co-variance of gene g across different cell types, which could model its distribution through inverse Wishart distribution. Wang et al used Markov chain Monte Carl to estimate each parameter in this linear model, which is a sampling technique with high time consumption. To improve the computational speed so that to make estimation of the cell type-specific expression (CSE) on large cohort more feasible, we assume the expression value of each cell type shall be very distinct and the gene g expresses in a specific cell-type shall not influence its expression level in other cell type. Therefore, co-variance of gene g across different cell types approximates to diagonal matrix. Also, we assume the observation noise across different samples is the same. Based on the above assumption, and inspired by previous imputation algorithm in scRNA-seq[57], we proposed to optimize the following loss function to estimate $(X_i)_{i=1\ldots ct}$

$$(\hat{X}_i)_{i=1\ldots ct} = argmin(\tfrac{1}{2}\alpha\left\|O - \textstyle\sum_{i=1}^{ct} X_i diag(\theta_i)\right\|_F^2 + \tfrac{1}{2}(1-\alpha)\textstyle\sum_{i=1}^{ct}\|X_i a - R_i\|_2^2 + \beta\textstyle\sum_{i=1}^{ct}\hat{\theta}_i rank(X_i)),$$

in which θ_i represents $W_{\cdot i}$ and a is a column vector with the size 1 by n and each element in a is 1/n. Rank(X) is the rank of the matrix X. The corresponding parameter beta controlling the latent dimensionality of each deconvoluted expression matrix. Considering that the spectrum radial of each CSE matrix is different, constraint them with the same value may lead to wrong estimation, therefore we also multiply beta with the expectation of theta $\hat{\theta}_i = \frac{\sum_{j=1}^{n} W_{ji}}{n}$, which means we assign a weak constraint on latent dimensionality of the cell types that didn't appear frequently in bulk sample because the cell type expression information shall not informative in input bulk expression matrix. 1-alpha is the weight for the agreement between the vector of aggregated CSE matrix and its corresponding reference expression vector.

Trace Norm model

We employed to compute the optimal solution as follows:

$$\left(\hat{X}_i\right)_{i=1\ldots ct} = argmin(\tfrac{1}{2}\alpha\left\|O - \textstyle\sum_{i=1}^{ct} X_i diag(\theta_i)\right\|_F^2 + \tfrac{1}{2}(1-\alpha)\textstyle\sum_{i=1}^{ct}\|X_i a - R_i\|_2^2 + \beta\textstyle\sum_{i=1}^{ct}\hat{\theta}_i\|X_i\|_*),$$

in which a represent [1/n…1/n] vector with length n. We used nuclear norm to regularize each decomposed expression profile, which is the convex envelope of rank function, calculated as the sum of the singular value of matrix. We used the following steps to calculate X in the above formula.

We transformed above constraint by introducing the auxiliary variable Y, and formulated optimization problem as augmented Lagrangian method as follow:

$$\begin{aligned}\left(\hat{X}_i, \hat{Y}_i\right)_{i=1\ldots ct} = \operatorname{argmin}(&\frac{1}{2}\alpha\left\|O - \sum_{i=1}^{ct} X_i diag(\theta_i)\right\|_F^2 + \beta\sum_{i=1}^{ct}\hat{\theta}_i\|Y_i\|_* \\ &+\frac{1}{2}(1-\alpha)\sum_{i=1}^{ct}\|X_i a - R_i\|_2^2 + \sum_{i=1}^{ct} < A_i, X_i - Y_i >_F \\ &+\sum_{i=1}^{ct}\frac{\gamma}{2}\|X_i - Y_i\|_F^2)\end{aligned}$$

The Alternating Direction Method of Multipliers (ADMM)[58] iteration scheme can be written as follows:

$$\begin{aligned}{X_j}^{k+1} = \operatorname{argmin}(&\frac{1}{2}\alpha\left\|O - \sum_{i=1}^{ct} X_i diag(\theta_i)\right\|_F^2 + \frac{1}{2}(1-\alpha)\left\|X_j a - R_j\right\|_2^2 \\ &+< {A_j}^k, X_j - {Y_j}^k >_F + \frac{\gamma}{2}\left\|X_j - {Y_j}^k\right\|_F^2)\end{aligned}$$

$${Y_j}^{k+1} = \operatorname{argmin}(\beta\hat{\theta}_i\left\|Y_j\right\|_* + < {A_j}^k, {X_j}^{k+1} - Y_j >_F + \frac{\gamma}{2}\left\|{X_j}^{k+1} - Y_j\right\|_F^2)$$

$${A_j}^{k+1} = {A_j}^k + \gamma\left({X_j}^{k+1} - {Y_j}^{k+1}\right)$$

We took the derivative with respect to X_i and made it equal to 0 to get the equation as follow:

$$\begin{aligned}\alpha\left(\sum_{i\neq j} X_i diag(\theta_i) - 0\right) diag\left(\theta_j\right) + \alpha X_j diag\left({\theta_j}^2\right) + (1-\alpha)X_j a a^T - (1-\alpha)R_j a^T + A_j^k \\ + \gamma\left(X_j - {Y_j}^k\right) = 0\end{aligned}$$

Let $\alpha\left(\sum_{i\neq j} X_i diag(\theta_i) - 0\right) diag\left(\theta_j\right) = T_j$, the above equation could be reorganized as follow:

$$X_j\left(\alpha \cdot diag\left({\theta_j}^2\right) + (1-\alpha)aa^T + \gamma I\right) = (1-\alpha)R_j a^T - A_j^k + \gamma {Y_j}^k - T_j$$

Therefore, the updated function of X_j could be written as follow:

$${X_j}^{k+1} = \left((1-\alpha)R_j a^T - A_j^k + \gamma {Y_j}^k - T_j\right)\left(\alpha \cdot diag\left({\theta_j}^2\right) + (1-\alpha)aa^T + \gamma I\right)^{-1}$$

For the optimization of Y_j, we could rewrite the equation as:

$$Y_j^{k+1} = \operatorname{argmin}\left(\beta\hat{\theta}_j\|Y_j\|_* + < A_j^k, X_j^{k+1} - Y_j >_F + \frac{\gamma}{2}\|X_j^{k+1} - Y_j\|_F^2\right)$$

$$= \operatorname{argmin}\left(\frac{\beta\hat{\theta}_j}{\gamma}\|Y_j\|_* + < \frac{A_j^k}{\gamma}, X_j^{k+1} - Y_j >_F + \frac{1}{2}\|X_j^{k+1} - Y_j\|_F^2 + \frac{1}{2}\left\|\frac{A_j^k}{\gamma}\right\|_F^2\right)$$

$$= \operatorname{argmin}\left(\frac{\beta\hat{\theta}_j}{\gamma}\|Y_j\|_* + \frac{1}{2}\left\|\frac{A_j^k}{\gamma} + X_j^{k+1} - Y_j\right\|_F^2\right)$$

Further, we could use singular value thresholding[59] to calculate the Y_j^{k+1}

$$Y_j^{k+1} = D_{\frac{\beta\hat{\theta}_j}{\gamma}}\left(\frac{A_j^k}{\gamma} + X_j^{k+1}\right)$$

in which

$$D_\gamma(Z) = U\begin{bmatrix}\sigma_1 - \gamma & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_n - \gamma\end{bmatrix}V^T, \mathrm{SVD(Z)} = U\begin{bmatrix}\sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_n\end{bmatrix}V^T$$

Finally, the updating of A_j could write as follow:

$$A_j^{k+1} = A_j^k + \gamma\left(X_j^{k+1} - Y_j^{k+1}\right)$$

Putting all pieces together, we present an iterative way to address trace norm based regularization problem (Algorithm 1).

Algorithm 1. D**E**co**n**volut**i**on based on Re**g**ularized **M**atrix Completion algorithm (trace norm version)
Input: the bulk expression profile $\mathrm{O} \in R^{m\times n}$, the reference matrix $\mathrm{R} \in R^{m\times ct}$, cell type proportion matrix $\theta \in R^{n\times ct}$, parameter α, β, γ, **Output:** predict cell type-specific gene expression profile $[X_j]_{j=1\dots ct}$ Initialize each expression profile X, auxiliary variable Y and A as matrix O, $\mathrm{k} \leftarrow 1$; Initialize vector a as $\left(\frac{1}{n}, \dots, \frac{1}{n}\right)^T \in R^n$ Initialize matrix $\left(\alpha \cdot diag\left(\theta_j{}^2\right) + (1-\alpha)aa^T + \gamma I\right)^{-1}$(represented as F_j) repeat for j in 1…ct (the number of cell type) $\alpha\left(\sum_{i\neq j} X_i{}^k diag(\theta_i) - \mathrm{O}\right)diag\left(\theta_j\right) = T_j{}^k$ $X_j{}^{k+1} = \left((1-\alpha)R_j a^T - A_j^k + \gamma Y_j{}^k - T_j{}^k\right)F_j$ $Y_j^{k+1} = D_{\frac{\beta\hat{\theta}_j}{\gamma}}\left(\frac{A_j{}^k}{\gamma} + X_j{}^{k+1}\right)$ $A_j^{k+1} = A_j^k + \gamma\left(X_j{}^{k+1} - Y_j^{k+1}\right)$ end $\mathrm{k} \leftarrow \mathrm{k} + 1$ until converge return $[X_j{}^k]_{j=1\dots ct}$

Maximum L2 norm model

For deconvoluting a bulk sample, we could use following loss function to deconvolute it:

$$\left(\hat{X}_i\right)_{i=1\ldots ct} = argmin(\tfrac{1}{2}\alpha\left\|O - \textstyle\sum_{i=1}^{ct} X_i\theta_i\right\|_F^2 + \tfrac{1}{2}(1-\alpha)\textstyle\sum_{i=1}^{ct}\left\|X_i - R_j\right\|_2^2 + \beta\textstyle\sum_{i=1}^{ct}\|X_i\|_2^2),$$

in which we regularized the deconvoluted expression vector through L2 norm, which penalizes the weight (deconvoluted expression) matrices from being too large. It was called 'shrinkage'[60], which was frequently used in ridger regression or elastic net regression model to prevent overfitting. We could extend this formula to address multi-sample deconvoluton as follow:

$$\left(\hat{X}_i\right)_{i=1\ldots ct} = argmin(\frac{1}{2}\alpha\left\|O - \sum_{i=1}^{ct} X_i diag(\theta_i)\right\|_F^2 + \frac{1}{2}(1-\alpha)\sum_{i=1}^{ct}\left\|X_i a - R_j\right\|_2^2 + \beta\sum_{i=1}^{ct}\|X_i\|_F^2)$$

However, the F-norm regularization will constraint the overall expression value. We could release the constraint through considering the maximum L2 norm across each column (sample) of deconvoluted cell type-specific expression matrices as follow:

$$\left(\hat{X}_i\right)_{i=1\ldots ct} = argmin(\tfrac{1}{2}\alpha\left\|O - \textstyle\sum_{i=1}^{ct} X_i diag(\theta_i)\right\|_F^2 + \tfrac{1}{2}(1-\alpha)\textstyle\sum_{i=1}^{ct}\left\|X_i a - R_j\right\|_2^2 + \beta\textstyle\sum_{i=1}^{ct}\|t(X_i)\|_{2,max}^2),$$

where $\|\cdot\|_{2,max}^2$ denotes the square of maximum L2 row norm of a matrix:

$$\|\cdot\|_{2,max}^2 \coloneqq max_j \sum_k A_{jk}^2$$

We used proximal-point algorithm to address above loss function. Replace the decision variable X with $\mathrm{A} = \mathrm{X}'$. With this substitution, the above regularization problem reduced to

$$minimize_{(\mathrm{A}_i)_{i=1\ldots ct}} f((\mathrm{A}_i)_{i=1\ldots ct}) + \beta\sum_{i=1}^{ct}\|A_i\|_{2,max}^2$$

While it is hard to calculate this loss function gradient directly, we could approximate it. When perform gradient decent, we could use linearized proximal-point method[61] to replace the cost function f with a quadratic approximation localized at the previous iterate $\left(A_i^k\right)_{i=1\ldots ct}$.

$$A_i^{k+1} = argmin < \nabla f\left(A_i^k\right), A > + \frac{1}{2}\tau^{-1}\left\|A - A_i^k\right\|_F^2 + \beta\left\|A_i^k\right\|_{2,max}^2$$

We complete the square to rewrite this formula to:

$$\begin{aligned} A_i^{k+1} &= argmin < \nabla f\left(A_i^k\right), A > + \frac{1}{2}\tau^{-1}\left\|A - A_i^k\right\|_F^2 + \beta\left\|A_i^k\right\|_{2,max}^2 \\ &= argmin \frac{\tau^2}{2}\nabla^2 f\left(A_i^k\right) + \tau < \nabla f\left(A_i^k\right), \left(A - A_i^k\right) > + \frac{1}{2}\left\|A - A_i^k\right\|_F^2 + \frac{\beta\tau}{2}\left\|A_i^k\right\|_{2,max}^2 \\ &= argmin \left\|A - (A_i^k - \tau\nabla f\left(A_i^k\right))\right\|_F^2 + \beta\tau\left\|A_i^k\right\|_{2,max}^2 \end{aligned}$$

The new formula could be solved through previously proposed *squash* algorithm[61]. Which could be denoted as:

$$A_i^{k+1} = squash(A_i^k - \tau\nabla f\left(A_i^k\right), \beta\tau)$$

Putting all pieces together, we present an iterative way to address maximum L2-norm

based regularization problem (Algorithm 2).

Algorithm 2. D**E**co**n**volut**i**on based on Re**g**ularized **M**atrix Completion algorithm (maximum L2-norm version)
Input: the bulk expression profile $\mathrm{O} \in R^{m\times n}$, the reference matrix $\mathrm{R} \in R^{m\times ct}$, cell type proportion matrix $\theta \in R^{n\times ct}$, parameter α, β, τ, **Output:** predict cell type-specific gene expression profile $[X_j]_{j=1\ldots ct}$ Initialize each expression profile X, Y and A as matrix O, $\mathrm{k} \leftarrow 1$; Initialize vector a as $\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)^T \in R^n$ repeat for j in 1…ct (the number of cell type) $\nabla f\left(X_j^k\right) = \alpha\left(\sum_i X_i^k diag(\theta_i) - \mathrm{O}\right) diag\left(\theta_j\right) + (1-\alpha) X_j^k a a^T - (1-\alpha) R_j a^T$ $X_j^{k+1} = t(squash(t(X_j^k - \tau \nabla f\left(X_j^k\right)), \beta\tau))$ end $\mathrm{k} \leftarrow \mathrm{k} + 1$ until converge return $[X_j^k]_{j=1\ldots ct}$

Compare with trace-norm based regularization, this method would not promote rank reduced cell type-specific gene expression matrix, but the smoothed gene expression matrix with each entry of the gene expression matrix is shrinkage. Also, training this model shall be faster than trace-norm based model, as its optimization procedure doesn't involve the singular value decomposition or inversion of any matrix.

Generation of simulated data for comparing the CSE inference accuracy

We simulated the cell type-specific expression (CSE) data of four cell types using the Bioconductor package *Splatter*[24]. We used the *splatSimulateGroup* function to generate the simulation data with 180 samples and 1000 genes with default parameters setting. Each sample is simulated to express 99.3% genes. Further, we simulated each cell type fractions through sampling from a dirichlet distribution. Then multipled the simulated cell type fraction matrix with CSE matrix to generate *in silico* mixture matrix (M). We also considered the measurement error, and simulated the additional measurements noise through combining it with mixture matrix M as follow:

$$\mathrm{B} = \mathrm{M} + \theta * \mathrm{noise}, \mathrm{noise} \sim \mathrm{N}(0,1),$$

in which theta refers as the noise level. We simulated 4 different noise levels (0.1, 0.5, 1, 1.5), and benchmarked each method under varied noise level through measuring the gene-gene Spearman correlation, sample-sample Spearman correlation across each cell type. To benchmark ENIGMA with other algorithms on real datasets, we used single cell RNA-seq to generate pseudo-bulk expression data through aggregating the expression profile of the cells from the same donor. For each scRNA-seq dataset for benchmarking, we randomly selected the cells from two donors to construct reference profile, in which we calculated each cell type average expression value. The rest cells from other batches were used to generate pseudo-bulk expression data. The cell type fractions of each donor were

calculated through measuring the proportion of the cells from the specific cell type.

Generation of simulated data for benchmarking sample correlation structure reconstruction

Samples in CSE may have latent biological correlation relationship, such as the cell differential trajectory, or cell sub-population. To make the deconvolution has better biological meanings, it is necessary to inspect whether the inferred CSE could capture the sample-wise correlation structure. We simulated samples which have internal differential trajectory structure through R package *Splatter* with function *splatSimulatePath.* Then mixed it with the simulated two CSE matrices which have no differential trajectory structures. We produced *in silico* mixture and bulk expression matrix based on above described workflow.

Generation of simulated data for benchmarking detection of differential expression genes (DEGs)

One advantage of inferring CSE is to help us to detect cell type-specific DEGs. We simulated gene expression datasets with DEG based on previously works[62]. We firstly simulated cell type expression matrix with each column represents an expression vector for a cell type, which denoted as $\mathrm{H1} \in R^{g \times ct}$, then we randomly added or delated value on each column of H1 to simulate DEG, wherein the value represents the signal strength of the DEG. Next, we denoted the processed matrix as $\mathrm{H2} \in R^{g \times ct}$. Further, we sampled cell type fractions (R1 and R2) on dirichlet distribution for H1 and H2, and calculated the *in silico* mixture through multiplying them, which represents as $\mathrm{M1} = \mathrm{H1} \times \mathrm{R1}^T$, $\mathrm{M2} = \mathrm{H2} \times \mathrm{R2}^T$. Finally, we generated simulated bulk expression matrix through adding white noise on each matrix. The signal to noise ratio (SNR) was calculated as the ratio of the DEG signal strength to the noise level. We simulated 1000 genes, 5 cell types and 100 samples. We also simulated 50 up-regulated genes and 50 down-regulated genes in H2. Meanwhile, we set the noise level equal to one, and simulated DEGs with varied SNR (1.8, 2.4,3, 3.6, 4.2, 4.8).

Benchmark DEGs detection requires cautious choose of criterion, otherwise it would lead ones to biased comparison and conclusion. Therefore, we used sensitivity, specificity and precision to benchmark different methods. The sensitivity and specificity reflect whether the methods are sensitive to differential and non-differential expression, respectively. And the precision reflects the degree of false discovery rate. To calculate these three indices, we required the predicted positive DEGs shall have the same differential expression trends (up- or down-) with the ground truth and passed the threshold of significance (adjusted *P* value < 0.05), otherwise it's a negative prediction. We used linear regression model to perform cell type-specific DEG analysis for bMIND, TCA and ENIGMA. Also, we noted that bMIND and TCA also provided their own statistical model to perform DEG analysis, therefore, we also included them to benchmark. We used Benjamini–Hochberg to correct multiple testing when calculating *P* value.

Expression concordance estimation between ground truth expression profile and inferred CSE

Validate the DEG prediction performance on real dataset is difficult because of the following considerations: 1) We don't know the ground truth cell type-specific DEG; 2) Different DEG selection criterion (*P* value, FoldChange) would influence the performance evaluation. The key concern in expression deconvolution is that we want to know the gene expression variate or trend across case and control samples. Inspired by previous research[7], we calculated 'DEG concordance' as the fraction of genes within predicted DEG genes with the same direction of differential expression between inferred CSE and the validation data (higher in both or lower in both) after mean-aggregating the expression data by known phenotypic groupings. To make the comparison of different methods more faithful, we randomly changed the predicted DEG selection criterion as follow: First, we set a fixed threshold for *P* value or adjusted *P* value (< 0.05) to filter the genes, which was considered as tested significantly. Then we sampled the Fold Change (FC) cutoff on the FC distribution, and filtered the genes that > FC cutoff or < -FC cutoff, where these genes were considered as the predict DEG genes, and calculated their DEG concordance value. We repeated the selection of FC cutoff for 1000 times so that we could get a DEG concordance distribution. Using this method, we also could calculate the expression concordance on all genes, which measure the overall predicted trends similarity with the ground truth dataset.

Construct reference cell type-specific gene expression profile

Building cell type-specific reference profile is necessary for cell type deconvolution. While the technical variation between bulk sample expression profiles and reference matrix will bring distortions in cell abundance and cell type-specific expression estimation, especially when we used scRNA-seq with technical dropout and noise. We applied previously developed methods, named as B-mode and S-mode normalization[7], to remove cross-platform batch effect.

B-mode normalization was used for the reference matrix derived from sorted bulk RNA-seq or microarray base expression matrix. It also could be applied on the reference matrix derived from full length scRNA-seq with sufficient sequence depth platform, e.g. SMART-Seq2[63]. S-mode normalization was used for the reference matrix derived from scRNA-seq platform with high technical dropout, e.g. 10X Genomics and droplet-based sequencing. S-mode normalization aimed to generate pseudo-bulk matrix through admixing the reference profile, and used ComBat[64] to correct the pseudo-bulk matrix with bulk expression matrix (which need to be deconvoluted), and restored the reference matrix through non-negative least square (NNLS) regression. We implemented both methods based on R and the source code could be found at https://github.com/WWXkenmo/ENIGMA .

Running of other deconvolution algorithms

We also benchmarked cell type omic profile inference method TCA and bMIND packages. For TCA, we set the maximum iteration time as maxiter = 100 when running tca() function to fit the data, so that it could converge. One thing we need to note that is this algorithm is originally designed for DNA methylation, therefore when applying this algorithm on gene expression profile one would encounter some errors because gene expression range and distribution are distinctly different with DNA methylation data. For bMIND, we followed the

running tutorial https://github.com/randel/MIND/blob/master/MIND-manual.pdf. When running bMIND, we also used the reference gene expression profile of ENIGMA. All three methods need the cell fraction matrix (CTM), we therefore used robust linear regression with ad-hoc adjustment method[65] to estimate CTM.

We also compared ENIGMA with other methods when analyzing NSCLC datasets. These methods have intrinsic shortcomings thus are not suitable for performing extensive simulation study and benchmark. The first category is limited component method, ISOpure and DemixT, which only deconvolute the bulk RNA-seq dataset from tumor into two or three components (tumor cell or non-tumor cell), while the TCA, bMIND and ENIGMA could deconvolute multiple cell types. The second category is non-negative factorization method, CIBERSORTx, which is a web-server based tool, and time-consuming when applying it on large cohort. Also, it does not provide open source code for implementation. Thus, we used these three methods when performing NSCLC deconvolution, which is a bulk RNA-seq dataset sampling from NSCLC tumor with small sample size. We followed the benchmark workflow provided by CIBERSORTx[7], and applied Spearman correlation to measure the deconvoluted expression concordance with ground truth gene expression profile curated from corresponding FACS RNA-seq datasets.

Data resource and analysis

Analysis of NSCLC

We used the Gentles et al. NSCLC bulk RNA-seq dataset[21] and Lambrechts et al. NSCLC scRNA-seq (10X Genomics) dataset[26] as the training datasets for predicting sample-wise cell type-specific expression profile. Then validate through Newman et al. NSCLC FACS cell type-specific RNA-seq. For NSCLC scRNA-seq dataset, we first normalized the row count data as the count per million (CPM). This step was implemented through *Seurat* R-package[66] with R function NormalizeData(NSCLC_seurat, normalization.method = "RC", scale.factor = 10000). Further, we constructed cell type reference expression profile based on NSCLC scRNA-seq dataset from Patient #3[26] through “S-mode” method described in the section “Construct reference cell type-specific gene expression profile” and set n_pseudo_bulk parameter as 500. When estimating cell type fraction matrix, we used robust linear regression model based on our inferred reference expression profile and non-log transformed bulk expression profile. When performing bulk RNA-seq deconvolution to infer CSE, we used log transformed bulk expression and reference expression profile.

Analysis of rheumatoid arthritis

We used the Guo et al. synovial tissue bulk RNA-seq dataset sampled from rheumatoid arthritis and osteoarthritis patients as the discovery datasets[30]. We also used Zhang et al. FACS sorted RNA-seq dataset from four purified cell types (T cell, B cell, Monocyte and Fibroblast) as reference[29], and used its scRNA-seq (CEL-seq2) dataset as the validation dataset.

We processed the data as follow. To correct the batch effect while preserving biological variation between bulk RNA-seq dataset from whole synovial tissue and FACS RNA-seq dataset from four cell types, we used “B-mode” correction to construct the reference gene expression profile (as illustrated in ‘Construct reference cell type-specific gene expression profile’). We also transformed the scRNA-seq count matrix into CPM. To calculate

differential expression genes (DEGs), we used Wilcoxon test to calculate DEGs across different cell types and used Benjamini–Hochberg method to correct multiple testing when calculate *P* value.

Analysis of pancreatic islet

We used Fadista et al. pancreatic islet sampled from type 2 diabetes and normal donors bulk RNA-seq dataset and corresponding patient metadata as the discovery dataset[40]. We also used Baron et al. pancreatic islet scRNA-seq dataset as the reference dataset[10], in which it only sampled the pancreatic islet from normal donors. Considering that the Baron et al. scRNA-seq dataset contains some rare cell types with no sufficient number of cells to generate pseudo bulk samples, we kept the cell type with its fraction in reference dataset more than 5%. We used Segerstolpe et al. pancreatic islet scRNA-seq dataset as the validation dataset[42].

We first normalized both Baron et al. and Segerstolpe et al. datasets into CPM, and used "S-mode" method combined with Baron dataset to build cell type reference expression matrix. After inferring CSE, we calculated DEGs between T2D and normal controls on both inferred CSE and Segerstolpe dataset, and compared the DEG expression concordance and all gene expression concordance of our prediction with that of Segerstolpe dataset. Noteworthy, bMIND only predicted two DEGs (P value < 0.1) in alpha cells and it may bias our conclusion of the DEG expression concordance, therefore we discarded this part of result, and only compared the all gene expression concordance.

We also performed weighted gene co-expression network analysis (WGCNA) on purified CSE. We first analyzed CSE by concatenating all cell types. A signed weighted correlation network was constructed by creating a matrix of pairwise correlations between all pairs of genes across the predict CSE. Further, the adjacency matrix was constructed by raising the co-expression measure, $0.5 + 0.5 \times$ correlation matrix, to the power $\beta = 10$. The power value, which is interpreted as a soft-threshold of the correlation matrix, was selected through scale free fitness optimization routine[13]. Based on the resulting adjacency matrix, we calculated the topological overlap matrix (TOM), which is a robust and biologically meaningful measure of network interconnectedness[13]. Genes with highly similar co-expression relationships were grouped together by performing average linkage hierarchical clustering on the topological overlap. We used the Dynamic Hybrid Tree Cut algorithm to cut the hierarchal clustering tree, and defined modules as branches from the tree cutting. We summarized the expression profile of each module by representing it as the first principal component (referred to as module eigengene). Modules whose eigengenes were highly correlated (correlation coefficient above 0.85) were merged. When constructing beta cell-specific weighted gene co-expression network, we first kept the top 8000 variable genes in beta cell CSE, then followed the above described way to construct the gene co-expression network step by step. We chose the power value 24 according to its scale free fitness result.

In senescence score assessment, we used two different ways to assess senescence. We curated two gene sets associated with cell senescence. One is SASP gene list[48] and the other is CellAge senescence associated genes list[47]. We then quantified expression variation of these two gene sets across samples through GSVA (Gene set variation analysis)[67].

Data availability

ENIGMA is freely available at https://github.com/WWXkenmo/ENIGMA.

Acknowledgements

This work was supported by National Key Research and Development Program of China [2018YFC1003500], National Natural Science Foundation of China [91949107, 31771336, 31521003] and Shanghai Municipal Science and Technology Major Project [2017SHZDZX01].

Author contributions

The manuscript was written by W.W., T.N. and polished by Y.W., C.Z., W.T. and J.Z.. The method was conceived by W.W. and J.Y. and the algorithm is implemented by W.W. and J.Y. Computational analyses and algorithm evaluations was conducted by W.W. and J.Y. This work was supervised by T.N.

Competing interests

The authors declare no competing interests

Reference

1. Costa, V., Aprile, M., Esposito, R. & Ciccodicola, A. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *European Journal of Human Genetics* **21**, 134-142 (2013).
2. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics* **10**, 57-63 (2009).
3. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology* **19**, A68 (2015).
4. Lonsdale, J. et al. The genotype-tissue expression (GTEx) project. *Nature genetics* **45**, 580-585 (2013).
5. Basu, S., Campbell, H.M., Dittel, B.N. & Ray, A. Purification of specific cell population by fluorescence activated cell sorting (FACS). *Journal of visualized experiments: JoVE* (2010).
6. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods* **6**, 377-382 (2009).
7. Newman, A.M. et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology* **37**, 773-782 (2019).
8. Newman, A.M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods* **12**, 453-457 (2015).
9. Li, B. et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome biology* **17**, 1-16 (2016).
10. Baron, M. et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems* **3**, 346-360. e344 (2016).
11. Wang, X., Park, J., Susztak, K., Zhang, N.R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications* **10**, 1-9 (2019).
12. Wang, J., Roeder, K. & Devlin, B. Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data. *Genome Research*, gr. 268722.268120 (2021).
13. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology* **4** (2005).
14. Rahmani, E. et al. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nature methods* **13**, 443 (2016).
15. Yu, C. & Yao, W. Robust linear regression: A review and comparison. *Communications in Statistics-Simulation and Computation* **46**, 6261-6282 (2017).
16. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9** (2008).
17. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
18. Rahmani, E. et al. Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nature communications* **10**, 1-11 (2019).
19. Puram, S.V. et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**, 1611-1624. e1624 (2017).
20. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196 (2016).
21. Gentles, A.J. et al. A human lung tumor microenvironment interactome identifies clinically relevant cell-type cross-talk. *Genome biology* **21**, 1-22 (2020).

22. Quon, G. et al. Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome medicine* **5**, 1-20 (2013).
23. Ahn, J. et al. DeMix: deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics* **29**, 1865-1871 (2013).
24. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology* **18**, 174 (2017).
25. Avila Cobos, F., Vandesompele, J., Mestdagh, P. & De Preter, K. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* **34**, 1969-1979 (2018).
26. Lambrechts, D. et al. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nature Medicine* **24**, 1277-1289 (2018).
27. Guo, Q. et al. Rheumatoid arthritis: pathological mechanisms and modern pharmacologic therapies. *Bone Res* **6**, 15-15 (2018).
28. Huang, W. et al. Parallel comparison of fibroblast-like synoviocytes from the surgically removed hyperplastic synovial tissues of rheumatoid arthritis and osteoarthritis patients. *BMC Musculoskelet Disord* **20**, 591-591 (2019).
29. Zhang, F. et al. Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. *Nat Immunol* **20**, 928-942 (2019).
30. Guo, Y. et al. CD40L-Dependent Pathway Is Active at Various Stages of Rheumatoid Arthritis Disease Progression. *The Journal of Immunology* **198**, 4490 (2017).
31. Pharoah, D.S. et al. Expression of the inflammatory chemokines CCL5, CCL3 and CXCL10 in juvenile idiopathic arthritis, and demonstration of CCL5 production by an atypical subset of CD8+ T cells. *Arthritis Research & Therapy* **8**, R50 (2006).
32. Jahid, M., Rehan Ul, H., Chawla, D., Avasthi, R. & Ahmed, R.S. Association of polymorphic variants in IL1B gene with secretion of IL-1β protein and inflammatory markers in north Indian rheumatoid arthritis patients. *Gene* **641**, 63-67 (2018).
33. Ramírez-Pérez, S. et al. High expression of interleukine-1 receptor antagonist in rheumatoid arthritis: Association with IL1RN*2/2 genotype. *Autoimmunity* **50**, 468-475 (2017).
34. Kinne, R.W., Bräuer, R., Stuhlmüller, B., Palombo-Kinne, E. & Burmester, G.-R. Macrophages in rheumatoid arthritis. *Arthritis Research & Therapy* **2**, 189 (2000).
35. Koch, A.E. et al. Monoclonal antibodies detect monocyte/macrophage activation and differentiation antigens and identify functionally distinct subpopulations of human rheumatoid synovial tissue macrophages. *Am J Pathol* **138**, 165-173 (1991).
36. Paoletti, A. et al. Monocyte/Macrophage Abnormalities Specific to Rheumatoid Arthritis Are Linked to miR-155 and Are Differentially Modulated by Different TNF Inhibitors. *J Immunol* **203**, 1766-1775 (2019).
37. Dong, P. et al. $CD86^+/CD206^+$, Diametrically Polarized Tumor-Associated Macrophages, Predict Hepatocellular Carcinoma Patient Prognosis. *Int J Mol Sci* **17**, 320-320 (2016).
38. Smith, S.R. et al. The phosphatase PPM1A controls monocyte-to-macrophage differentiation. *Sci Rep* **8**, 902-902 (2018).
39. Röder, P.V., Wu, B., Liu, Y. & Han, W. Pancreatic regulation of glucose homeostasis. *Exp*

Mol Med **48**, e219-e219 (2016).
40. Fadista, J. et al. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc Natl Acad Sci U S A* **111**, 13924-13929 (2014).
41. Leong, A. & Meigs, J.B. Type 2 Diabetes Prevention: Implications of Hemoglobin A1c Genetics. *Rev Diabet Stud* **12**, 351-362 (2015).
42. Segerstolpe, Å. et al. Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell metabolism* **24**, 593-607 (2016).
43. Sone, H. & Kagawa, Y. Pancreatic beta cell senescence contributes to the pathogenesis of type 2 diabetes in high-fat diet-induced diabetic mice. *Diabetologia* **48**, 58-67 (2005).
44. Maedler, K. & Donath, M.Y. β-Cells in Type 2 Diabetes: A Loss of Function and Mass. *Hormone Research in Paediatrics* **62(suppl 3)**, 67-73 (2004).
45. Tomita, T. Apoptosis in pancreatic β-islet cells in Type 2 diabetes. *Bosn J Basic Med Sci* **16**, 162-179 (2016).
46. Aguayo-Mazzucato, C. et al. Acceleration of β Cell Aging Determines Diabetes and Senolysis Improves Disease Outcomes. *Cell Metabolism* **30**, 129-142.e124 (2019).
47. Tacutu, R. et al. Human Ageing Genomic Resources: new and updated databases. *Nucleic Acids Res* **46**, D1083-D1090 (2018).
48. Coppé, J.-P., Desprez, P.-Y., Krtolica, A. & Campisi, J. The senescence-associated secretory phenotype: the dark side of tumor suppression. *Annu Rev Pathol* **5**, 99-118 (2010).
49. MacDonald, R.J., Swift, G.H. & Real, F.X. in Progress in Molecular Biology and Translational Science, Vol. 97. (ed. K.H. Kaestner) 1-40 (Academic Press, 2010).
50. Gonzalvez, F. & Ashkenazi, A. New insights into apoptosis signaling by Apo2L/TRAIL. *Oncogene* **29**, 4752-4765 (2010).
51. Thorburn, A. Tumor Necrosis Factor-Related Apoptosis-Inducing Ligand (TRAIL) Pathway Signaling. *Journal of Thoracic Oncology* **2**, 461-465 (2007).
52. Harith, H.H., Morris, M.J. & Kavurma, M.M. On the TRAIL of obesity and diabetes. *Trends in Endocrinology & Metabolism* **24**, 578-587 (2013).
53. Hattangadi, D.K. et al. Influence of p53 and caspase 3 activity on cell death and senescence in response to methotrexate in the breast tumor cell. *Biochemical Pharmacology* **68**, 1699-1708 (2004).
54. Marcotte, R., Lacelle, C. & Wang, E. Senescent fibroblasts resist apoptosis by downregulating caspase-3. *Mechanisms of Ageing and Development* **125**, 777-783 (2004).
55. Yang, C. et al. Transcriptional activation of caspase-6 and -7 genes by cisplatin-induced p53 and its functional significance in cisplatin nephrotoxicity. *Cell Death & Differentiation* **15**, 530-544 (2008).
56. Zheng, S.C., Breeze, C.E., Beck, S. & Teschendorff, A.E. Identification of differentially methylated cell types in epigenome-wide association studies. *Nature Methods* **15**, 1059-1066 (2018).
57. Peng, T., Zhu, Q., Yin, P. & Tan, K. SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data. *Genome biology* **20**, 88-88 (2019).
58. Boyd, S., Parikh, N. & Chu, E. Distributed optimization and statistical learning via the alternating direction method of multipliers. (Now Publishers Inc, 2011).

59. Cai, J.-F., Candès, E.J. & Shen, Z. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization* **20**, 1956-1982 (2010).
60. Copas, J.B. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society: Series B (Methodological)* **45**, 311-335 (1983).
61. Lee, J., Recht, B., Salakhutdinov, R.R., Srebro, N. & Tropp, J.A. (Neural Information Processing Systems, 2010).
62. Shen-Orr, S.S. et al. Cell type–specific gene expression differences in complex tissues. *Nature methods* **7**, 287-289 (2010).
63. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods* **10**, 1096-1098 (2013).
64. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E. & Storey, J.D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882-883 (2012).
65. Teschendorff, A.E., Breeze, C.E., Zheng, S.C. & Beck, S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC bioinformatics* **18**, 1-14 (2017).
66. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* **36**, 411-420 (2018).
67. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics* **14**, 1-15 (2013).

Main Figure

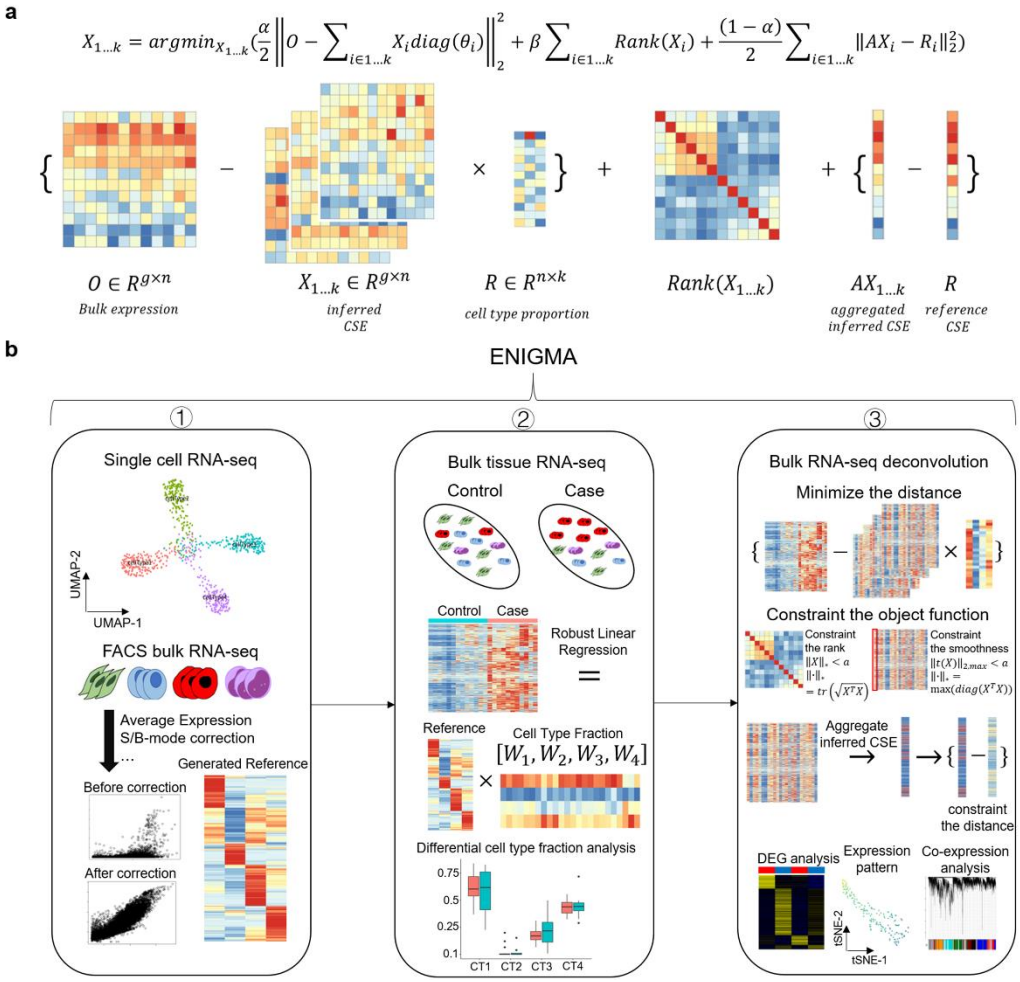

Fig. 1: Cell type-specific expression (CSE) inference using ENIGMA

a. Schematic overview of the ENIGMA algorithm. The loss (object) function is shown on the top, which has three terms. The first term means the difference between the observed bulk gene expression matrix and the linear combination of the CSE. The mixing coefficient for each CSE is quantified as the cell fraction matrix, which could be estimated from classical cell type deconvolution algorithms (e.g. CIBERSORT). The second term is regularization term, which can be used to control the rank of CSE or shrink the CSE (extreme values in a CSE are “shrunk” towards a central value). The third term represents the difference between aggregated inferred CSE and reference profile across each cell type.

b. Workflow overview of the ENIGMA algorithm. The main steps of ENIGMA involve: 1) Transcriptome profiling of flow-sorted cell subpopulations or single cells to define a reference matrix consisting of features (genes) that can discriminate each cell subset of interest in a given tissue type. To prevent batch effect, ENIGMA applied pre-developed methods to correct the technique artifacts induced by cross-platform; 2) The bulk gene expression profile was deconvoluted through using robust linear regression and reference matrix to estimate each cell type fractions in each sample. The inferred cell fraction information could be used to perform differential cell fractions analysis; 3) Using matrix completion based algorithm to deconvolute bulk RNA profile. There have two versions of ENIGMA, ENIGMA (L2 max norm, default) and ENIGMA (trace norm). ENIGMA constrained that each average expression value inferred by CSE need to be close to the reference expression profile. After minimizing the object function to estimate the CSE, we could use the inferred CSE to perform differential gene expression (DEG) analysis, gene expression pattern analysis in each CSE, and gene co-expression analysis, etc.

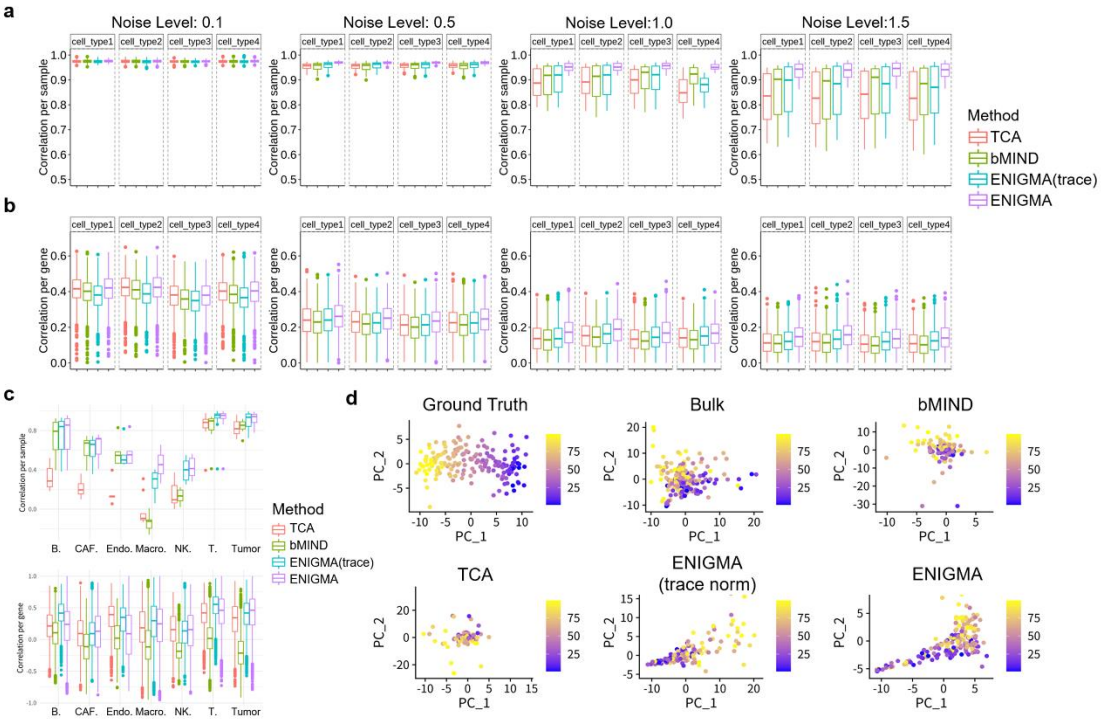

Fig. 2: Validation of ENIGMA on simulation data.

a.b. Each imputed cell type-specific expression (CSE) profile was compared against its corresponding simulated ground truth CSE by Spearman correlation with different observation noise level (0.1, 0.5, 1.0, 1.5). For each cell type, we computed correlation across genes for each sample (a) and correlation across samples for each gene (b). The CSEs were inferred through four methods, TCA, bMIND, ENIGMA, and ENIGMA (trace norm). The ground truth CSE and bulk expression profile for deconvolution were generated through simulation (Method).

c. Each imputed CSE profile was compared against its corresponding ground truth CSE by Spearman correlation. For each cell type, we computed correlation across genes for each sample (upper panel) and correlation across samples for each gene (lower panel). The ground truth CSE and bulk expression profile for deconvolution was generated from scRNA-seq dataset of head and neck squamous cell carcinoma.

d. The principal component analysis (PCA) plots of gene expression profile. Each dot represents a sample and colored according to the pseudo-time label on the right. The simulated ground truth CSE profile was posed with a “trajectory” like structure (left of the upper panel), labeled as “Ground Truth”. The middle of the upper panel denotes ground truth profile mixed with other three simulated CSE profiles to generate a simulated bulk RNA expression profile, labeled as “Bulk”. We further used bMIND, TCA, ENIGMA and ENIGMA (trace norm) to deconvolute the admixture sample to infer the ground truth profile and inspect whether the algorithms recover the trajectory structure through visualization on PCA plot.

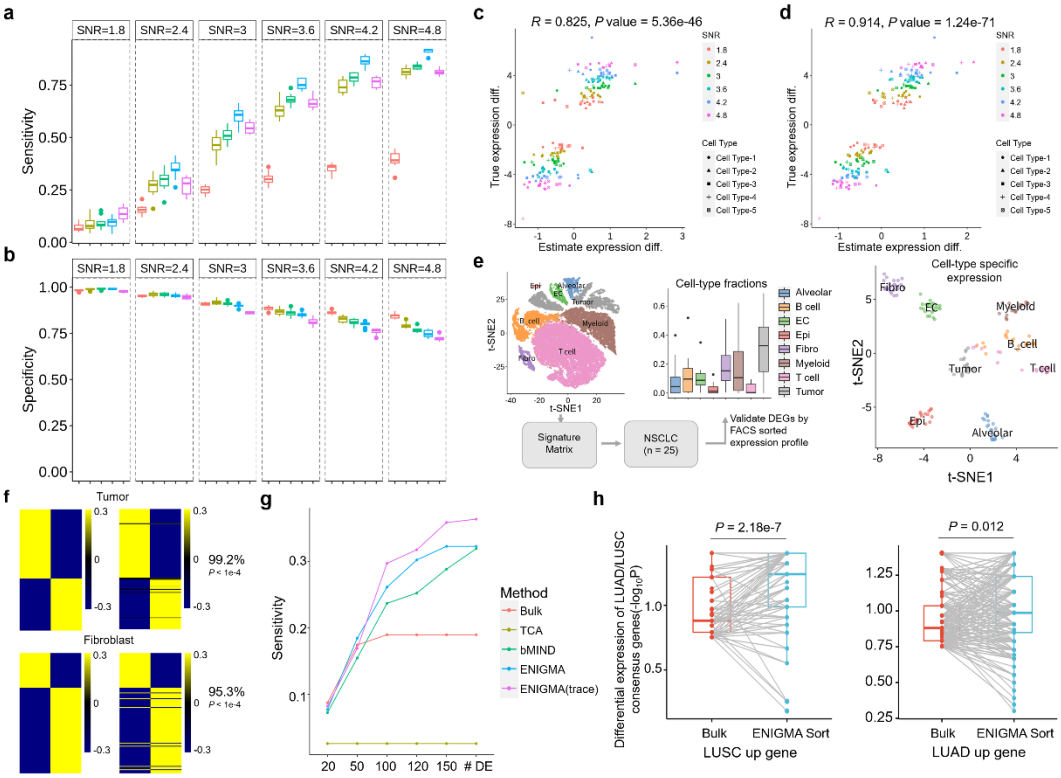

Fig. 3: Benchmark of ENIGMA on CTS-DEG detection.

a-b. Boxplots of sensitivity (a) and specificity (b) for the evaluation of detecting cell type-specific differentially expressed genes (CTS-DEG) on the simulation dataset. The rate of true positives (sensitivity) and the rate of true negatives (specificity) were measured under a range of signal to noise ratio (SNR). All of the simulations were evaluated under the setting of five constituting cell types, 1000 genes, 100 samples and Gaussian observation noise with unit variance. Each cell type has 100 DEGs with randomly down- or up-regulated.

c,d. Scatterplots of the true gene expression difference in affected cell types (y-axis) versus predicted gene expression difference (x-axis) from raw bulk RNA profile (c) or CSE profile imputed from trace-norm based ENIGMA (d). Data points are shown for 500 DEGs, five affected cell types and six SNR levels. Each point is colored according to SNR level and shaped by the cell type it comes from.

e. Schema of deconvoluting NSCLC tumor tissue bulk RNA profile using ENIGMA. The reference matrix was constructed from previous scRNA-seq dataset, and it was applied on the bulk RNA profile from 26 patients to deconvolute the bulk RNA profile into cell type fraction matrix and cell type-specific expression profile. Left: A t-SNE plot showing eight major tumor subpopulations profiled from six donors by scRNA-seq. Each dot represents a cell and is colored according to the cell type. Middle: A barplot showing eight major tumor subpopulation fractions in 26 patient samples[21]. Right: A t-SNE plot showing eigth major tumor subpopulations deconvoluted from bulk RNA profile of 26 patients. Each dot represents a patient and is colored according to the cell type it belongs to.

f. Analysis of the concordance between ENIGMA-imputed (left) and flow-sorted (right) CTS-DEG expression profiles in LUAD and LUSC. The heatmaps show the averaged mean-centered log2-adjusted expression values for each DEG identified by ENIGMA. The DEGs were defined as the genes with |log2FC| > 2 and adjusted *P* value < 0.01. 'DEG concordance' denotes the fraction of genes with the same direction of differential expression between ENIGMA and flow-sorted profile. The *P* value is calculated through Monte Carlo strategy (Method).

g. The curves showing the sensitivity vs. the number of selected top DEGs. The sensitivity is calculated based on the number of top DEGs estimated from each method. The top DEGs (adjusted *P* value < 0.05) are ranked through |log2FC|. For each method, when the number of top DEGs is exceed the total number of the DEGs, the sensitivity is calculated according to the total DEGs. The pseudo ground truth DEGs were defined from flow-sorted profile and scRNA-seq profile (Method).

h. Significance comparison between bulk RNA profile and ENIGMA-imputed profile for pseudo ground truth DEGs from panel g. The significance (-log10(*P* value)) was calculated by Wilcoxon rank-sum test.

Fig 4. ENIGMA reveals monocyte pseudo-differentiation trajectory in arthritis patients

a. Schema outlining the application of ENIGMA to identify cellular signature in arthritis patient dataset. The t-SNE plot represents the external single cell RNA-seq data of arthritis patients for validation. We used one sample of flow-sorted RNA-seq (n = 105) to build signature (reference) matrix, and applied the signature matrix to deconvolute bulk RNA expression matrix (n = 174) by ENIGMA. To validate the reliability of *in-silico* purification by ENIGMA, we used scRNA-seq dataset and flow-sorted RNA-seq datasets (n = 104) to perform CTS-DEG external validation.

b. Analysis of the concordance between ENIGMA-imputed (left) and scRNA-seq (right) CTS-DEG expression profiles in OA and RA. The heatmaps show the averaged mean-centered log2-adjusted expression values for each DEG identified by ENIGMA. The DEGs were defined as the genes with |log2FC| > 2 and adjusted *P* value < 0.01. 'DEG concordance' denotes the fraction of genes with the same direction of differential expression between ENIGMA and scRNA-seq profile. The P value is calculated through performing Wilcoxon rank-sum test of concordance index (Method).

c. The heatmap of averaged expression value of pathway marker genes from each group (OA & RA) across ENIGMA predicted CSE of B cell, Fibroblast and Monocyte.

d. The cell type fraction boxplot of four cell types in arthritis patients. Each box is colored according to the phenotype information.

e. A t-SNE plot of monocyte CSE imputed by ENIGMA for both OA and RA. The sample is clustered using Louvian algorithm on the KNN graph built based on monocyte CSE. Each dot represents a

deconvoluted sample belongs to the monocyte and colored according to the cell clustering results.

f. The top five significant GO pathways of the marker genes derived from the differential gene expression analysis corresponding to the cluster label in e. The pathways are ordered according to its statistics significance. The GO analysis is conducted using R package *clusterProfiler.*

g. The gene expression pattern of three genes is visualized on t-SNE plot. Each dot represents a deconvoluted sample and was colored according to gene expression value.

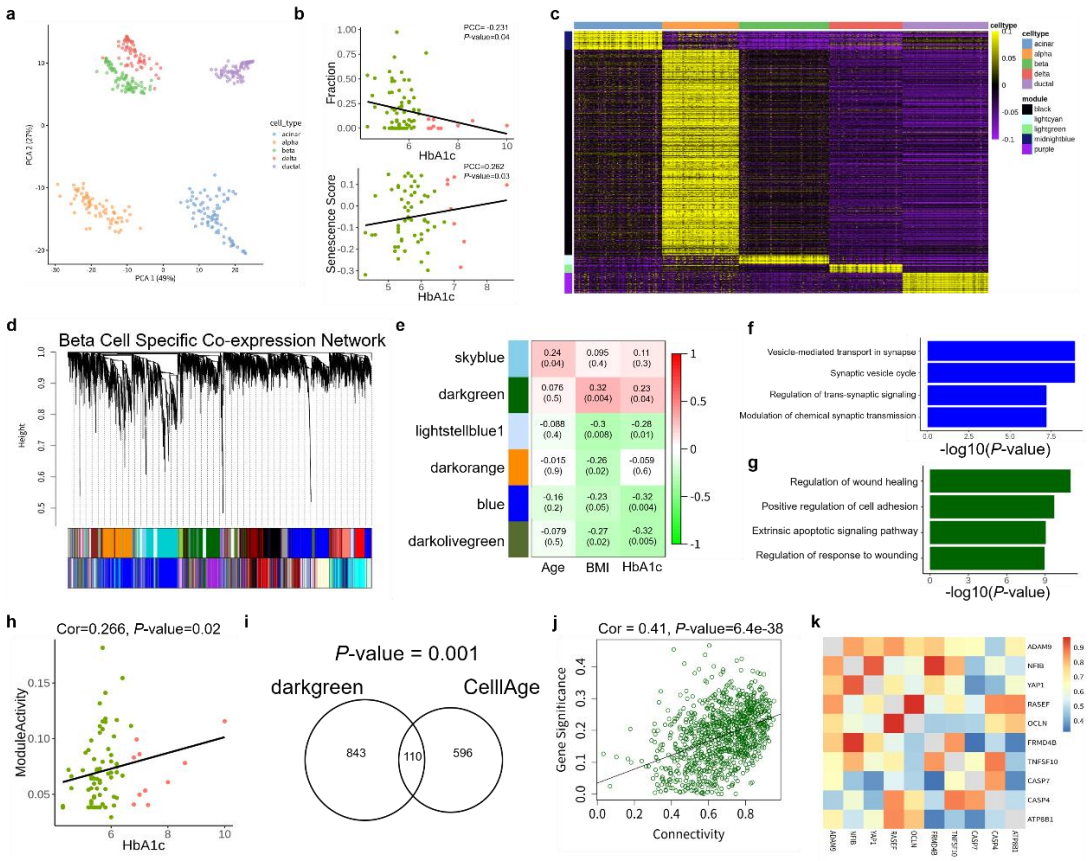

Fig 5. ENIGMA identifies cellular co-expression network in pancreas islet from T2D patients.

a. A PCA plot of ENIGMA derived CSE of pancreatic islet bulk RNA profile sampled from T2D and normal patients. Each dot represents a deconvoluted sample and was colored according to the cell type its belongs to.
b. The scatter plots of beta cell fractions (top) and senescence score (bottom) with the HbA1c level. The correlation is calculated using Pearson correlation coefficients. Each dot represents a deconvoluted sample.
c. The heatmap of CSE of five co-expression modules that are associated with each cell type. The row represents a gene and the column represents a sample.
d. Hierarchical cluster tree showing co-expression modules identified using WGCNA in beta cells. Modules correspond to branches and are labelled by colors as indicated by the first color band underneath the tree. The modules identified by bulk RNA-seq are labelled by colors as indicated by the second color band underneath the tree.
e. The heatmap of the association matrix between identified modules and three phenotypes (Age, BMI, HbA1c). The correlation coefficients are calculated using the Pearson correlation of each pair of the module eigengene expression value and phenotype. The module eigengene expression is calculated as the first left singular vector of the module gene expression matrix.
f. The top four significant GO pathways enriched in blue module in panel e. The pathways are ordered according to their significance. GO analysis is conducted through R-package *clusterProfiler.*
g. The top four significant GO pathways enriched in darkgreen module in panel e. The pathways are ordered according to their significance. GO analysis is conducted through R-package *clusterProfiler.*
h. The scatter plot of darkgreen module (in panel e) activity with the HbA1c level. The correlation is calculated using Pearson correlation coefficient (Cor). Each dot represents a deconvoluted sample.
i. Venn diagram showing the overlap of genes in darkgreen modules (in panel e) and genes in cell senescence signature database (CellAge)[47]. The significance of overlap is calculated using hypergeometric test.
j. The scatter plot of gene significance (defined as the correlation coefficients between gene expression and HbA1c level, y-axis) and the gene connectivity (defined as correlation coefficients between gene expression and module eigengene expression, x-axis). Each dot represents a gene in darkgreen module in panel e.
k. The heatmap visualization of topological overlap correlation network for top 10 hub genes from darkgreen module in panel e.