deCS: A tool for systematic cell type annotations of single-cell RNA sequencing data among human tissues

Guangsheng Pei[1], Fangfang Yan[1], Lukas M. Simon[2], Yulin Dai[1], Peilin Jia[1*], Zhongming Zhao[1,3,4,5*]

[1]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[2]Therapeutic Innovation Center, Baylor College of Medicine, Houston, TX 77030, USA

[3]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[4]MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, USA

[5]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, USA

*Corresponding authors:

Email: zhongming.zhao@uth.tmc.edu (Zhao Z), peilin.jia@gmail.com (Jia P).

Running title: *Pei G et al / Decoding Cell Type-specificity*

Total number of words: 6277.

Total number of words for article title: 16.

Total number of words for running title: 4.

Total number of words for Abstract: 188.

Total number of Keywords: 5.

Total number of figures: 7.

Total number of tables: 1.

Total number of supplementary figures: 12.

Total number of supplementary tables: 7.

Total number of references: 74.

Abstract

Single-cell RNA sequencing (scRNA-seq) is revolutionizing the study of complex and dynamic cellular mechanisms. However, cell-type annotation remains a main challenge as it largely relies on manual curation, which is cumbersome and less accurate. The increasing number of scRNA-seq data sets, as well as numerous published genetic studies, motivated us to build a comprehensive human cell type reference atlas. Here, we present *deCS* (decoding Cell type-Specificity), an automatic cell type annotation method based on a comprehensive collection of human cell type expression profiles or a list of marker genes. We applied *deCS* to single-cell data sets from various tissue types, and systematically evaluated the annotation accuracy under various conditions, including reference panels, sequencing depth and feature selection. Our results demonstrated that expanding the references is critical for improving annotation accuracy. Compared to the existing state-of-the-art annotation tools, *deCS* significantly reduced computation time while increased accuracy. *deCS* can be integrated into the standard scRNA-seq analytical pipeline to enhance cell type annotation. Finally, we demonstrated the broad utility of *deCS* to identify trait-cell type associations in 51 human complex traits, providing deeper insight into the cellular mechanisms of disease pathogenesis.

Keywords: Cell type-specific enrichment analysis, scRNA-seq, cell type annotation, trait-cell type association, cell type marker gene.

Introduction

Recent breakthroughs of single-cell RNA sequencing (scRNA-seq) have greatly boosted our ability to characterize the heterogeneity of cell types in various tissues [1, 2], leading to deeper understanding of disease pathogenesis and development [3, 4]. The typical scRNA-seq data analysis involves unsupervised clustering of the cells, followed by annotation of each cluster [5, 6]. Accurate identification of cell types is critical for downstream analysis (9). However, this process relies heavily on prior knowledge of cell type marker genes, which is subjective and time-consuming.

In contrast to manual annotation, many automatic annotation tools were recently developed, including *SingleR* [7], *CHETAH* [8], *CellAssign* [9], *scMatch* [10], and *scCATCH* [11], *scPred* [12], *scibet* [13], *scVi* [14] and *Cell BLAST* [15]. However, the benchmarking analysis by Abdelaal et al. showed large difference in method performance, which strongly depended upon the reference database [16]. Thus, there is a pressing need to curate high-quality and comprehensive reference panels of human cell type expression profiles, which include the annotated scRNA-seq data from various tissues or sequencing protocols. So far, most of these methods have been based on machine learning models, such as support vector machine (SVM) [16] and random forest classifier [17], which are computation-intensive and cannot be directly applied to huge datasets, such as human cell landscape (HCL) [18], human cell atlas [19], and human cell atlas of fetal (HCAF) gene expression [20]. Moreover, the annotation largely relies on a single reference, which may be inaccurate when cell types in “query” and “reference” dataset are not well matched. Therefore, a computational method that can efficiently integrate the annotation results among multiple references is urgently needed. Furthermore, the cell type inference is mostly conducted at single cell level rather than cluster level. The annotation accuracy will significantly decrease, especially for low sequencing depth cells, due to the dropout effect in scRNA-seq

data. Even though imputation methods [21, 22] can be applied to recover missing gene expression and improve the annotation accuracy, they are time-consuming [22].

Here, we present *deCS* (decoding Cell type-Specificity), an automatic scRNA-seq cell type annotation method by decoding cell type-specificity. As an improved version of the "*deTS*" algorithm [23], deCS runs fast, allows the integration of cell annotation from multiple references, and defines cell types on cell cluster level rather than single cell level. Specifically, we first created a high-quality and comprehensive human cell type reference panel by including all publicly available large-scale scRNA-seq data sets from various tissues or sequencing protocols, such as human cell landscape, human cell atlas and human cell atlas of fetal gene expression. Then, we calculated the *t*-statistic or z-score based measurements to "decode cell type-specificity" of gene sets. As an R package, *deCS* can be implemented with different statistical methods. It supports input for either gene expression profiles (correlation analysis) or list of genes (Fisher's exact test). Benchmarking results showed that *deCS* reduced computation time and improved annotation accuracy in most tissues compared to other state-of-the-art methods, especially when the cell-type composition of query and reference are not conserved. Therefore, we anticipate *deCS* will become a scRNA-seq routine annotation tool especially when users do not have enough prior knowledge about the cell type markers. Lastly, the curated cell types and their signature genes can be used to explore the cell type-specificity of disease-risk gene lists. By defining cell type-specific genes (CTGenes) and decoding cell type-specificity, we demonstrated the utility of *deCS* to characterize the relationships between human complex diseases and cell types. Accordingly, these results can provide novel insight into the cellular mechanism of the poorly understood traits and diseases.

Results

Overview of *deCS* workflow

Our goal is to develop a tool to perform automatic cell type annotation across datasets of different sequencing protocols and levels of complexity. To this end, we first collected various public cell type expression profiles. As shown in **Figure 1**, we included several public human RNA-seq references from *SingleR* [7], including its BlueprintEncode [24, 25], the Database of Immune Cell Expression (DICE) [26], and MonacoImmune [27]. In addition, we downloaded the most comprehensive human single-cell expression data resource from the human cell landscape (HCL) [18] and human cell atlas of fetal (HCAF) projects [20]. For each data set, we applied different analytical strategies (see Materials and Methods for more details) to define cell type-specific genes (CTGenes). Finally, we integrated the manually curated CTGenes from the CellMatch database [11]. Based on genes *t*-statistic/*z*-score and CTGenes, we implemented two statistical tests for cell type annotation (**Figure 1**). Depending on the type of "query data", *deCS* identifies whether a "gene expression profile" of an unknown cell type is significantly similar to a cell type in the reference by using correlation analysis. In addition, *deCS* determines whether a set of "candidate genes" of interest overlap with a cell type in the reference by using Fisher's exact test (FET).

Cross-reference validation and biological pathway enrichment of CTGenes

To validate the reliability and consistency of CTGenes, we first calculated internal cell type correlation and then utilized Uniform Manifold Approximation and Projection (UMAP) dimension reduction analysis [28] to visualize the global landscape of expression profiles (**Figure S1**). We further validated the consistency of CTGenes across different references, including BlueprintEncode, DICE, MonacoImmune, and HCL. Accordingly, we systematically compared the CTGenes (by default using the top 5% highest genes by *t*-statistic) between two different references. We observed a large variation between references, with matching rate ranging from 20% to 70%, although more than 80% of matched cell types shared the highest cell type-specific marker genes (**Figure S2**). For example, when comparing the BlueprintEncode reference with

the HCL reference, only 36 out of 102 (35.3%) cell types in HCL shared at least 200 CTGenes with the BlueprintEncode reference, while most cell types in HCL reference [e.g., alveolar type 2 (AT2) cells in lung, mast cells in blood] were missed in BlueprintEncode.

To further assess the validity of CTGenes identified by *t*-statistic, we performed pathway enrichment analysis for each cell type using the top 200 CTGenes (**Table S1**). We constructed a network containing all significantly enriched pathways (FDR < 0.05) and corresponding cell types. As shown in **Figure 2**, more than 90% CTGenes were correctly enriched in biologically relevant pathways: adipocytes were enriched in "*Regulation of lipolysis in adipocytes*"; B-cells were enriched in "*B-cell receptor signaling pathway*"; CD4+ and CD8+ T-cells were enriched in "*T-cell receptor signaling pathway*"; HSCs were enriched in "*Spliceosome*"; melanocytes were enriched in "*Melanogenesis*"; and neuron cells were enriched in "*GABAergic synapse*" and "*Glutamatergic synapse*". In addition, the relationship of CTGenes could be observed based on a force-directed layout network. For example, several cell types were well clustered: myocytes and skeletal muscles; CD4+, CD8+ T-cells and NK cells; and monocytes, dendritic cells (DC), and macrophages.

scRNA-seq cell type annotation

Peripheral blood mononuclear cells

Human peripheral blood mononuclear cells (PBMCs) are routinely studied in the biomedical fields including immunology and the development of diagnostics and therapeutics for human diseases [29]. To explore the utility of *deCS* for direct annotation of scRNA-seq data, we first analyzed a PBMC data set, which is available from 10x Genomics using the MonacoImmune reference. The basic workflow of two alternative annotation pipelines is depicted in **Figure S3**. We followed the standard pre-processing workflow for PBMC scRNA-seq data in Seurat [5]. We extracted the union of cell cluster-specific marker genes, and then calculated the average

expression profile for each cluster, followed by z-score normalization. We applied two statistical approaches: correlation analysis and Fisher's exact test would match the cell label with the highest PCC or intersection ratio (IR) to each query cluster (**Figure 3**). Overall, the main labels predicted by all 9 clusters were the same between these two methods on the PBMCs data set. The similarity scores of most clusters mapped to a single main label and were significantly higher compared to the remaining cell types, indicating high specificity on main cell label annotation. When mapping to fine labels, many cell clusters were mapped to multiple fine labels. Nevertheless, most of them were mapped to closely related cell types. For example, native CD4 T-cells were predicted as one subtype CD4 T-helper 1 cells (Th1) [30].

Solid tissues on adult and fetal samples

We next evaluated the performance of *deCS* in six additional scRNA-seq data sets from different solid tissues. The first data set were bronchoalveolar immune cells (BIC) from COVID-19 infected patients [31]. The results indicated that *deCS* found the best-matched cell type for the majority of cell clusters (8 out of 10) on BlueprintEncode panel (**Figure 4A, B**). Two exceptions were mast cells and plasmacytoid dendritic cells (pDC), which were matched to erythrocytes (PCC = 0.5, IR = 0.25) and B cells (PCC = 0.28, IR = 0.2), respectively. Here, IR refers to the intersection ratio. One major concern of the misclassification is that cell type is not represented in the used BlueprintEncode reference [8]. One possible solution is to annotate them as undetermined cells if they are too dissimilar to any references (e.g., PCC < 0.3) [32]. Alternatively, users can map their query data to other available references. As shown in **Figure S4**, mast cells and pDC were mapped to their counterpart when using the HCL (PCC = 0.82, IR = 0.77) and MonacoImmune (PCC = 0.74, IR = 0.9) references, respectively.

The second data set was derived from fetal cardiac cells (FCC) [33]. As shown in **Figure S5**, almost half cell clusters (5/9) failed to find any matched cell types (using PCC > 0.3 as the threshold) when mapping to BlueprintEncode. However, when using the HCL or HCAF reference,

there were 8 out of 9 cell clusters (except for C5 valval cell) mapped to the annotated cell type. Similarly, for the four human scRNA-seq datasets of liver, lung, spleen and esophagus tissues, only 50~70% of cell clusters were mapped to their counterpart when using BlueprintEncode, HCL or HCAF reference individually (**Table S2**). When combining the highest-scoring labels across multiple references, mapping rate was improved to 87.8% or above. This demonstrated that integrating results from multiple high-quality curated reference panels could significantly improve annotation accuracy and utility.

Annotation in single cell resolution

In the above analysis, we used the averaged expression level in each cluster to evaluate the performance of *deCS* [34]. These traditional routines, however, overlook an important characteristic of scRNA-seq data [35]: cellular heterogeneity. Considering the true hierarchical cluster structure for a cell population (e.g., both CD4+ and CD8+ are T cells) [35], cellular heterogeneity is widely observed when applying unsupervised clustering [36]. The ideal scenario is each cell can be annotated to one specific branch of the hierarchical tree [8]. Therefore, we further evaluated *deCS* in bronchoalveolar immune cells (BIC) data set [31]. As shown in **Figure 4C**, the majority of cells within 10 clusters can be matched to the best cell type through the integrative analyses across multiple references. There may exist misclassifications only in closely related cells. For example, for myeloid cells, 1,514 and 2,076 cells in macrophages cluster were erroneously annotated as monocytes and neutrophils, respectively. Some neutrophils were annotated as monocytes or macrophages (**Figures S6**, **S7**).

Previously the scMatch method [10] showed that the cells with more reads were more likely to be correctly classified than those with lower read depth. Therefore, based on the sequencing depth, we divided 63,103 cells from BIC data into three groups: low (1k ≤ UMI < 3k), median (3k ≤ UMI < 6k), and high (UMI ≥ 6k) depth groups. As **Figure 5A** showed, although low sequencing depth has a slight impact on annotation accuracy (e.g., 10 B cells were annotated to

macrophages), *deCS* approach had a good performance for other cell types. As a negative control, we further compared the annotation recall among healthy controls and moderate or severe COVID-19 infected patients. As shown in **Figure 5B**, the cell type frequencies among different groups demonstrated significant differences [Kruskal–Wallis test ($p < 0.05$)], especially for epithelial and neutrophil cell types [37].

Although *deCS* demonstrated good performance at the single-cell level, there remained a small fraction of misclassified cells. To address this issue, we further investigated top marker genes for each cluster to better understand the underlying reason. As shown in **Figure S8**, the top marker gene of naive CD4 T cell, "*CCR7*", which only expressed in almost 40% naive CD4 T cells due to "dropout" effect in scRNA-seq. We also observed that gene expression imputation approaches [21, 38, 39] could improve single cell resolution annotation performance. Thus, we recommend cell type annotation on clusters instead of individual cell after imputation, which is consistent with other publications [5, 40].

Benchmarking analysis on different tissues

To demonstrate the superior performance of *deCS*, we compared it with other methods, including *SingleR* [7], *CHETAH* [8], *scPred* [12], *scibet* [13], *scVi* [14] and *Cell BLAST* [15] using the above eight datasets. We first applied the methods to two PBMC data sets. As summarized in **Table 1**, the overall performances of *deCS*, *SingleR*, *scPred* and *sciBet* ranged from 87.55% to 89.41%. Among them, *scPred* was the best-performing classifier (88.78%), which is consistent with previous benchmark work, demonstrating that SVM classifiers have overall the best performance [16]. However, as classical genes are not expressed in every single cell, *deCS* showed better performance (89.41%) than *scPred* (**Table S3**) after imputation [21]. In addition, we noticed that *CHETAH*, *Cell BLAST* and *scSorter* had weaker robustness on the annotation accuracy. One possible reason is that the rejection threshold is too strict, such that most cells were annotated to be ambiguous or an intermediated branch of cell types. Another possible reason is potential overcorrection when some query cell types are non-overlapping with

reference (details are shown in **Table S2** and **Figure S9**).

Since most methods showed comparable performance on PBMC, we further benchmarked them on six extra datasets, including bronchoalveolar, heart, liver, lung, oesophagus and spleen. As described in **Table 1**, the *deCS* correlation analysis showed an average of 93.6% accuracy. When *deCS* Fisher's exact test failed to consider the "weight" feature, it only obtained an average of 89.1% accuracy. *sciBet* (86.0%) *and SingleR* (76.8%) also achieved good performance on most tissues, as both of them incorporated a comprehensive reference panel with middle size (>30 cell types). However, the accuracy for most other methods was lower than 50% due to inappropriate reference panels, even when counting those undetermined cells (when query cell types are non-overlapping with reference). For example, as most cell types from spleen were identical with PBMC, *scPred* showed 79.6% accuracy on spleen tissue. However, on liver tissue, where there was a large proportion of hepatocyte (48.2%) (non-overlapping with reference), the accuracy was below 50%, as 41.3% of the hepatocytes were annotated to monocytes (rather than unassigned cell type). In most studies, researchers typically do not have enough prior knowledge about the cell type composition in the investigated samples. Therefore, the expanded reference in *deCS* can be widely used on various tissues.

***deCS* reduced the feature collinearity and impact of different sequencing depth**

The superiority of *deCS* is not only due to the expanded cell type. Even after excluding those query-specific cell types (e.g. ventricle cardiomyocytes in heart, hepatocytes in liver, alveolar cells in lung), the *deCS* showed better annotation accuracy in distinguishing biologically related cell types (e.g. CD4+, CD8+ T-cells and NK, monocyte and macrophage). We speculated feature selection (comparing to *SingleR*) is still an essential step before correlation analysis. These identified anchor genes (cluster marker genes) can effectively reduce the collinearity compared to the case when using more genes (top variable genes). As showed in **Figure 6A**, the query cluster is naive CD4 T cells and the PCC was 0.87 when using "union of cluster-specific genes", but dropped to 0.45 when using top 2000 variable genes. In contrast, the PCC between query

cluster and CD8+ T cells was still high (0.32) (**Figure 6B**). Although there are lots of other methods for feature selection and model prediction (e.g. SVM, random forest) [12, 13], it is not realistic to conduct such time-consuming process on huge HCL or HCAF raw data (>4 million cells) [20]. In consideration of factors of both speed and scalability when combining the highest-scoring labels across multiple references, we decided to provide the cell cluster aggregated *t*-statistic and *z*-scores. On the other hand, *deCS* drastically reduced the required memory and runtime, which is helpful for annotation of large data set using a personal computer.

Due to the stochasticity and inefficient mRNA capture in scRNA-seq, varying sequencing depth across batches is a major driver of batch effects [7, 41]. Thus, we evaluated the level of sequencing depth (different sized pools, n from 1 to 50) of single cell transcriptomes to the pseudobulk level (cluster averaged). As shown in **Figure 6C**, we observed a gradually diminishing improvement of inter-correlation with increasing sequencing depth. Take T cells as an example, the inter correlation between single cell and pseudobulk level in PBMC and BIC data sets was 0.707 and 0.898, respectively. As the median number of detected genes in BIC was almost 4-fold more than PBMC data set, it only needed to cumulate 5 cells to approximate 95% bulk level in BIC, while it needed at least cumulated 10 cells in PBMC data (**Figure 6C**). To put it another way, when a gene is observed at a low or moderate expression level in BIC cells, it is probably not detected in PBMC cells. This phenomenon is also known as “dropout” [42]. For better demonstration, we randomly selected 5 bulk RNA-seq samples and down-sampled each sample to 10-100k reads followed by quantification normalization (**Table S4**). Interestingly, in the low sequencing depth condition, we noticed a pronounced effect of “sequencing depth” compared to the “identity” level in the PCA plot. In contrast, z-score normalization effectively reduced the “batch effect” (**Figure S10**). Furthermore, standard UMAP analysis was applied to two different data batches (PBMC and BIC). As shown in **Figure 6D**, we observed a strong batch effect between PBMC and BIC log normalized expression matrices (e.g., NK and T cells were clustered by two samples). In contrast, after z-score normalization (**Figure 6E**), most cell types

form distinct clusters separated by cell types (e.g., B cells, NK cells, T cells and DC cells).

Decoding the cell type-specificity on other applications

In addition to scRNA-seq cell type annotation, *deCS* reference panels can be used to detect cell-type specificity and to provide insightful understanding of many other biological problems. For example, we can apply *deCS* to genome-wide association studies (GWAS) data for trait-cell type associations, and also bulk RNA-seq data from induced pluripotent stem cell (iPSC) derived cells. We demonstrated the utility of *deCS* with two applications below.

Understanding the underlying context of human complex diseases is an important step to unveil the etiology of disease origin [43], yet tissues are complex milieus consisting of various cell types [44]. Therefore, tissue level association may fail to elucidate cell type contributions in disease [45]. To uncover novel associations between cell types and diseases, *deCS* was applied to 51 GWAS data (**Table S5**) and the association was evaluated by Fisher's exact test. As shown in **Figure 7**, we observed (1) neuropsychiatric disease, college, and education associated genes being mainly enriched in neurons; (2) immune-related traits enriched in B-cells, T-cells, and NK-cells (lymphoid cells); (3) blood red cell related traits enriched in erythroid cells (**Figure S11**); (4) bone mineral density enriched in chondrocytes; and (5) high-density lipoproteins enriched in adipocytes. Of note, Alzheimer's disease, a neurodegenerative disease of the brain, was found to be associated with myeloid lineage cells, rather than neurons. We believe these results collectively shed light on the casual cell type underlying complex traits for both related and unrelated traits. Taken together, the accurate detection of the cell types underlying genetic variants will not only improve our understanding of the molecular mechanisms of complex diseases at the cell type level [46], but also can serve as an important instrumental role in cell type-specific transcriptome-wide association studies (TWAS) [47] and colocalization test [48], among other integrative genetic analyses.

The second application of *deCS* was applied to bulk RNA expression profiles. Human iPSCs have revolutionized the study of the biological mechanisms underlying psychiatric disorders by

establishing cellular models that account for a patient's genetic background [49]. In most iPSC studies, differentiation quality is routinely assessed [49]. After we normalized the bulk RNA-seq data iPSC-derived neuronal precursor cells (NPCs) and neurons [50], the top matched cell types in 12 of 13 iPSC-NPCs were most enriched in "fetal progenitor cells" or "human embryonic stem cells (hESC)" (**Figure S12**) while the top matched cell types for iPSC-neurons were most enriched in "fetal neuron". Therefore, *deCS* can also be applied for decoding the cell type-specificity on bulk RNA-seq.

Discussion

Single-cell sequencing is a fast-evolving technique that provides deeper insight into the complexity of cellular heterogeneity within the same tissue. One critical step in the analysis is cell type annotation. However, this step is largely done manually, which is time-consuming and subjective. A list of methods was recently developed for automatic cell type annotation [16]. However, these methods are not capable of discovering novel cell types, as they only incorporate a rejection option. In addition, due to the minute amount of starting tissue material, scRNA-seq data is prone to batch effects [51]. To overcome these limitations, one strategy is to conduct cell type annotation on each reference respectively, and then combine the highest-scoring labels across multiple references [7, 10]. In this study, we collected multiple human cell type expression profiles from different platforms. Our benchmarking results showed that *deCS* achieved the best performance by using the combined highest-scoring labels across multiple reference panels.

deCS relies on the predefined cell clusters. There are multiple algorithms for unsupervised clustering, including k-means clustering [52], hierarchical clustering, and graph-based clustering [5], which may generate different clusters and affect *deCS*'s performance. We performed the test across various clustering methods and resolutions, which showed greater than 90% annotation consistency (**Table S6**). This suggests that deCS is robust to predefined cell clusters. Of note,

deCS has several advantages over other models. First, the union of cluster-specific genes combined with *t*-statistics and z-scores in *deCS* drastically reduces the collinearity and the required memory and runtime for cell annotation while achieving comparable accuracy on biologically related cell types (e.g. CD4+, CD8+ T-cells and NK, monocyte and macrophage). Second, with the expanded references, *deCS* exhibits superior performance to other methods. *deCS* can combine the highest-scoring labels across multiple references and dramatically reduce the incorrect cell type annotation when real cell type counterparts are missing in the reference. Although the parameter selection, such as marker genes and correlation measure, may slightly affect the annotation accuracy, the collected comprehensive human cell type profiles in *deCS* provide wider applicability ranging from blood cells to adult and fetal solid tissues. Third, *deCS* supports scRNA-seq cell type annotation for either gene expression profiles or list of genes without prior knowledge. *deCS* is a new tool that it not only benefits the studies of scRNA-seq data, but also yields novel insights for better understanding the molecular basis of thousands of human complex diseases or traits.

We applied *deCS* algorithm to the GWAS data in our recently developed database, Cell type-Specific Enrichment Analysis DataBase (CSEA-DB), which covers 55 unique tissues [53]. Among a total of 10,250,480 trait-cell type associations, we observed significant cell type association in 598 (11.68%) traits. Some human complex diseases or traits were associated with multiple cell types. For example, asthma was found to be associated with both immune and epithelial cells (AT2) [53]. Moreover, comparing to previous studies [23] that most TAGs enriched in “blood” tissue, *deCS* showed immune TAGs were enriched in lymphoid lineage cells, while red blood cells related traits were enriched in erythroid cells. It has been frequently recognized that pleiotropic effects are ubiquitous in human complex traits [54]. For example, individuals carrying schizophrenia risk alleles tended to be also associated with high risk of Crohn’s Disease [55]. We expect to identify more explicit associations with the increasing high-quality curation of tissue-cell types.

There are several limitations in our study. First, *deCS* does not take tissue information (except for HCAF) into account which may bias the annotation results. For example, 93.2% cells from fetal thymus were annotated as proliferating T-cells (cluster 52), and 85% cells from fetal brain were annotated as fetal neurons (cluster 11) (**Table S7**). We strongly encourage users to combine prior biological knowledge and apply tissue-matched cell type reference to mitigate potential bias (e.g., almost no neurons or astrocytes in peripheral blood). Second, as a mapping-based annotation toolkit, *deCS* fails to characterize novel cell types. Although we included more than one hundred cell types and most of them come from healthy tissues. It is expected that *deCS* might be challenging for cancer cell annotations due to the expression profile difference between healthy and malignant cells [56]. We believe some “semi-supervised” cell type assignment methods like *scSorter* [57] can address it by providing the marker gene in corresponding cell type. Owing to the efforts by the human cell atlas [19] as well as time-dependent change of cell states in gene expression profiles [58], the available large numbers of single cell data sets can be fed in as reference data sets to improve the annotation of future experiments. Despite these limitations, we demonstrated the feasibility and availability of *deCS* for broad applications. Importantly, *deCS* method is not limited to disease gene lists. As our cell type reference panels are only built upon expression profiles, including those unstudied or poorly annotated genes [23, 43], users can highlight the cell type-specificity from any analysis. Taken together, *deCS* is a new tool that not only benefits studies on scRNA-seq data of complex tissues, but also yields novel insights for better understanding the molecular basis of various human diseases and traits.

Materials and methods

Data collection for cell type reference panels

The BlueprintEncode data

We downloaded the BlueprintEncode RNA-seq data from Aran et al. (2019) [7]. The raw data comprised of 259 bulk RNA-seq samples in total. All cell types were aggregated into 24 broad classes ("main cell types") with 43 cell types ("fine cell types"). Raw gene expression data was normalized by Transcripts Per Million (TPM), followed by log2 transformation.

The Database of Immune Cell Expression data

The Database of Immune Cell Expression (DICE) reference contained 1,561 bulk RNA-seq samples from pure populations of human immune cells. We downloaded the TPM normalized values for 5 main (15 fine) immune cell types or subtypes from https://dice-database.org/downloads [26].

The MonacoImmune data

The MonacoImmune reference comprised 114 PBMCs bulk RNA-seq samples from 4 Singaporean-Chinese individuals. We downloaded the TPM-normalized values for 10 main (29 fine) immune cell types (GSE107011) from Monaco G et al. study [27].

The human cell landscape data

The human cell landscape (HCL) reference contained more than 700,000 scRNA-seq expression profiles from more than 50 human different tissues [18]. These cells were grouped into 102 major clusters, and have been well-annotated base on known marker genes. These data were derived from 18 fetal tissues, 35 adult tissues, and several intermediate tissues (e.g., cord blood, placenta). We downloaded the gene expression profiles from http://bis.zju.edu.cn/HCL/landscape.html.

The human cell atlas of fetal data

The human cell atlas of fetal (HCAF) gene expression contained ~4,000,000 single cells from 15 human fetal organs ranging from 72 to 129 days in estimated post-conceptual age [20]. These cells were grouped into 172 major cell types and raw data is available at https://descartes.brotmanbaty.org/bbi/human-gene-expression-during-development/

CellMatch

The CellMatch reference [11] was curated from various resources, including CellMarker [59], MCA [60], CancerSEA [61], and the CD Marker Handbook. In this study, we only obtained 183 cell types for humans with an average of 163 marker genes per cell type. The raw data set is available from https://github.com/ZJUFanLab/scCATCH/blob/master/R/sysdata.rda.

Measurement of cell type-specificity of gene

All references derived from bulk RNA-seq were classified into two tiers: tier 1 for the broad (main) cell type category and tier 2 for a more granular (fine) cell type category [7]. For both tiers, we implemented our previous method [23] by fitting a regression model for each gene and computed *t*-statistics to measure the cell type-specificity. Briefly, for cell types in tier 1, we followed our previous work and fitted a regression model for each cell type independently as $Y \sim X$, where X is the cell group status (0 or 1) and Y is the log2 transformed gene expression. Specifically, for a cell type in the examination, we defined $X = (x_i)$, $i = 1,\ldots, N$, where N is the total number of samples; $x_i = 1$ if the sample belonged to the cell type in the examination, and $x_i = 0$ if the sample belonged to any cell types that are not in the same group. We then selected the *t*-statistic for the explanatory variable X in the standard way:

$$t = \frac{(X^T X)^{-1} X^T Y}{\sqrt{MSE \cdot (X^T X)^{-1}}}$$

where *MSE* is the mean squared error of the fitted model, i.e.,

$$MSE = \frac{1}{N}(Y - X(X^T X)^{-1} X^T Y)^T (Y - X(X^T X)^{-1} X^T Y).$$

Of note, most cell types in tier 2 were biologically related, such as "*memory B-cells*", "*naive B-cells*" and "*plasma cells*". An inclusion of all these instinctively related cell types in one regression model would underestimate their cell type-specificity due to the potential collinearity. Therefore, for the cell types in tier 2, the regression models were fitted after excluding the

samples from the same group (main cell type) but keeping the other cell types [23].

The predefined cell number per tissue in the HCL and HCAF data sets is large: 102 (HCL) and 172 (HCAF) major clusters. Accordingly, we utilized *z*-score to measure the cell type-specificity of genes. For each gene, a *z*-score is calculated as $z_i = (e_i - \text{mean}(E))/sd(E)$, where e_i is the average expression of the gene in the ith cluster, E represents the collection of its average expression in all clusters, and *sd* denotes the standard deviation of E.

For each cell type, those genes with the highest *t*-statistic or *z*-score are considered as CTGenes in the cell type in examination. The user can define the cutoff values (e.g., the top 5% genes as CTGenes). Because the CellMatch database did not include gene expression information, we directly utilized the predefined CTGenes [11].

To create cell type classification tree, for each reference data set, we conducted cell type hierarchical clustering based on Euclidian distance and ward.D2 linkage aggregation. The bootstrap resampling and probability values (p-values) for each cluster was conducted using *Pvclust* package [62].

Algorithm of cell type-specific enrichment analysis

Feature selection is an important step to determine the cell type. To deal with feature gene collinearity and dropout rate differences, while facing all kinds of variations between datasets, we recommend users to pre-process their query dataset, and to use the union of cell cluster-specific genes as input of *deCS*. These recommendations can improve the accuracy and efficiency.

Depending on the query data type, we implemented two test approaches. If the query is an expression profile, we provided two methods to minimize the batch effects between two different datasets, which is a particular concern in scRNA-seq: (1) The z-score strategy normalizes the query expression data by $e_n = (e_q - u_s)/sd_s$, where e_q and e_n are the queries and normalized expression, and u_s and sd_s are the mean and standard deviation of a gene from the query

expression data. (2) The abundance correction approach [63] normalizes the query RNA-seq data by $e_n = log_2(e_q+1)/(log_2(u_s+1)+1)$. After normalization, we calculated the *Pearson correlation coefficient* (PCC) of cell type-specificity of genes between the query and each of the reference cell types, respectively. The most relevant cell type(s), measured by the highest PCC score(s), are annotated to query profiles, possibly with further fine-tuning to resolve closely related cell type(s). On the other hand, we also incorporated a rejection option (e.g. minimum correlation coefficient threshold), which allows the detection of potentially novel cell populations.

When the query is a list of genes (e.g., marker genes of a cell cluster, traits associated genes), we implemented Fisher's exact test to examine if they are significantly enriched in cell type-specific genes (CTGenes). We allow users to define the threshold, e.g., the top 5% genes as CTGenes. Specifically, for the query gene set and CTGenes in a given cell type, we build a dichotomous 2×2 contingency table as follows:

	Cell type-specific genes (*r*)	
Query genes (*q*)	$\lvert Q \cap R \rvert$	$\lvert \bar{Q} \cap R \rvert$
	$\lvert Q \cap \bar{R} \rvert$	$\lvert \bar{Q} \cap \bar{R} \rvert$

Here, *q* denotes the set of query genes and *r* denotes the set of CTGenes in a given cell type. $\lvert Q \cap R \rvert$ denotes the intersect gene number between *Q* and *R*, $\lvert \bar{Q} \cap R \rvert$ and $\lvert Q \cap \bar{R} \rvert$ denote the number of genes only in *R* or *Q*, respectively, and $\lvert \bar{Q} \cap \bar{R} \rvert$ represents the number of genes that neither in *Q* or *R*. In addition, the intersection ratio (IR) $\lvert Q \cap R \rvert / \lvert Q \rvert$ will also be calculated. An intersection ratio (IR) of 1 indicates that the 100% query genes are overlapped with CTGenes (default, the top 5% genes) of a given cell type, while 0 indicates no overlap. The most relevant cell type(s), measured by the highest IR(s), are annotated to query profiles.

Pathway enrichment analysis of cell type-specific genes

For the top 200 CTGenes (genes with the highest *t*-statistic or *z*-score) of each cell type, we used RDAVIDWebService (v 1.19.0) [64] for pathway enrichment analysis. Both Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations are used. Benjamini and Hochberg's approach [65] was used for multiple test correction. Significant pathways were defined as those with a false positive rate (FDR) less than 0.05. The network of CTGenes enriched KEGG pathways was presented by Cytoscape [66].

Data collection and preprocessing

Single cell RNA-seq data

(1) Two 3k peripheral blood mononuclear cells (PBMC) were collected from 10x Genomics website (https://support.10xgenomics.com/single-cell-gene-expression/data sets).

(2) The test data set of bronchoalveolar immune cells (BIC) was collected from healthy controls (HC, n = 3) and patients with moderate (M, n = 3) and severe (S, n = 6) COVID-19 infection [31, 67]. The BIC data included 63,103 single cells. The raw data and predefined labels are available from https://github.com/zhangzlab/covid_balf.

(3) The fetal cardiac cells (FCC) scRNA-seq data were collected from 18 human embryos, ranging from 5 weeks (5W) to 25W of gestation [33]. The UMI count of 3,842 cells and predefined labels were downloaded from the GEO website (GSE106118).

(4) The human liver scRNA-seq data were collected from five fresh hepatic tissues [68]. The raw dataset of 8,444 cells and annotation are available through the GEO accession (GSE115469).

(5) The scRNA-seq data of human lung, spleen and esophagus tissues were collected from 12 organ donors [69]. The raw data of 239,224 single cells and annotation are available from https://www.tissuestabilitycellatlas.org/

Traits associated gene list from GWAS summary statistics

Previous studies have suggested that most traits associated genes (TAGs) showed strong tissue-specific associations [23, 70]. Therefore, we further expanded 51 publicly available genome-wide association studies (GWAS) spanning a wide range of phenotype measurements to investigate the potential association between human traits and cell types.

Considering the statistical noise in GWAS data and more than 90% of genetic variants from GWAS are located in the non-coding regions, for each GWAS trait, we used moderately significant associated SNPs with chi-squared *p*-value < 10^{-3}. This strategy allowed us to have a list of genes with weak to strong association signals. Additionally, to predict their potential regulatory effects, we applied our DeepFun model [71, 72] based on deep convolutional neural networks (CNN) on a comprehensive epigenomic feature. We defined variants with a maximum SAD score greater than 0.1 (or less than -0.1) as regulatory loci (potential causal variants). Then we employed Pascal software [73] to map them to gene level if these SNPs were located within a range of 50 kb upstream or downstream of corresponding gene transcription start sites by taking into account linkage disequilibrium (LD) and gene length information. Any genes with at least one regulatory locus (max SAD > 0.1) were regarded as trait-associated genes (TAGs).

Bulk RNA-seq data

The test bulk RNA-seq data were generated from schizophrenia (SCZ) associated human iPSC-derived cell lines from the population isolate of the Central Valley of Costa Rica (CVCR) [74]. One clone from each subject was differentiated into neural progenitor cells (NPCs) and neurons, respectively. In total, we collected RNA-seq data from iPSC-NPCs (n=13) and iPSC-neurons (n=11).

Single cell permutation analysis

Single-cell permutation analysis was performed to assess the cell cluster internal correlation

and detected gene richness. For a given number (n from 1 to 50) of cells belong to the same cell type, we performed random sampling 100 times without replacement, then calculated the PCC of the averaged gene expression level from the single cells compared to the pseudo-bulk level (cell cluster averaged). In addition, we estimated the total number of detected genes that could be identified (at least one UMI in at least one cell), along with the cumulated number of cells.

Statistical analysis

Uniform Manifold Approximation and Projection (UMAP) analysis [28] was used to visualize scRNA-seq batch effect. For comparison of the cell annotation accuracy among healthy controls, moderate, and severe COVID-19 infected patients, we used *kruskal.test* function in *R* software. The performance of *deCS* was evaluated by recall, defined as TP/(TP+FN) where TP and FN denote the number of true positives and false negatives.

Evaluation and software implementation

The evaluation was conducted on a desktop equipped with i7-7700HQ CPU and 16 GB of RAM. We ran *deCS* using two approaches: correlation analysis and Fisher's exact test. *deCS* runs fast when applying correlation analysis. It took only ~7 seconds for a gene expression matrix with BIC data set (63,103 cells) by *deCS.correlation* function. When applying Fisher's exact test, it took only ~50 seconds for a list of 43,514 genes across the 51 traits.

Data and code availability

The source codes and results are implemented in R and are freely available at GitHub: https://github.com/GuangshengPei/deCS.

CRediT author statement

Guangsheng Pei: Conceptualization, Methodology, Resources, Data Curation, Visualization, Software, Writing - original draft. **Fangfang Yan**: Validation, Writing - review & editing. **Lukas M. Simon**: Validation, Writing - review & editing. **Yulin Dai**: Validation, Writing - review & editing.

Peilin Jia: Conceptualization, Methodology, Writing - review & editing. **Zhongming Zhao**: Conceptualization, Methodology, Project administration, Supervision, Writing - review & editing, Funding acquisition. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Acknowledgments

The authors thank all the members the Bioinformatics and Systems Medicine Laboratory (BSML). We also thank those investigators who generated the reference data. This study was partially supported by National Institutes of Health grants (R01LM012806, R01DE030122, and R01DE029818). We thank the resource support from Cancer Prevention and Research Institute of Texas (CPRIT RP180734 and RP210045). Funding for open access charge: CPRIT (RP180734) and NIH (R01LM012806).

ORCID

0000-0003-1804-7598 (Guangsheng Pei)
0000-0001-6231-5225 (Fangfang Yan)
0000-0001-6148-8861 (Lukas M Simon)
0000-0003-1874-7893 (Yulin Dai)
0000-0003-4523-4153 (Peilin Jia)
0000-0002-3477-0914 (Zhongming Zhao)

References

[1] Islam S, Kjällquist U, Moliner A, Zajac P, Fan J-B, Lönnerberg P, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res 2011;21:1160-7.

[2] Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med 2018;24:1277-89.

[3] Nomura S. Single-cell genomics to understand disease pathogenesis. J Hum Genet 2021;66:75-84.

[4] Angelidis I, Simon LM, Fernandez IE, Strunz M, Mayr CH, Greiffo FR, et al. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. Nat Commun 2019;10:963.

[5] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single-cell data. Cell 2019;177:1888-902.

[6] Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nat Methods 2017;14:309-15.

[7] Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol 2019;20:163-72.

[8] de Kanter JK, Lijnzaad P, Candelli T, Margaritis T, Holstege FCP. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. Nucleic Acids Res 2019;47:e95.

[9] Zhang AW, O'Flanagan C, Chavez EA, Lim JLP, Ceglia N, McPherson A, et al. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. Nat Methods 2019;16:1007-15.

[10] Hou R, Denisenko E, Forrest ARR. scMatch: a single-cell gene expression profile annotation tool using reference datasets. Bioinformatics 2019;35:4688-95.

[11] Shao X, Liao J, Lu X, Xue R, Ai N, Fan X. scCATCH: automatic annotation on cell types of clusters from single-cell RNA sequencing data. iScience 2020;23:100882.

[12] Alquicira-Hernandez J, Sathe A, Ji HP, Nguyen Q, Powell JE. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. Genome Biol 2019;20:264.

[13] Li C, Liu B, Kang B, Liu Z, Liu Y, Chen C, et al. SciBet as a portable and fast single cell type identifier. Nat Commun 2020;11:1818.

[14] Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods 2018;15:1053-8.

[15] Cao ZJ, Wei L, Lu S, Yang DC, Gao G. Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST. Nat Commun 2020;11:3458.

[16] Abdelaal T, Michielsen L, Cats D, Hoogduin D, Mei H, Reinders MJT, et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol 2019;20:194.

[17] Johnson TS, Wang T, Huang Z, Yu CY, Wu Y, Han Y, et al. LAmbDA: label ambiguous domain adaptation dataset integration reduces batch effects and improves subtype detection. Bioinformatics 2019;35:4696-706.

[18] Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, et al. Construction of a human cell landscape at single-cell level. Nature 2020;581:303-9.

[19] Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The human cell atlas. Elife 2017;6:e27041.

[20] Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, et al. A human cell atlas of fetal gene expression. Science 2020;370:eaba7721.

[21] Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, et al. SAVER: gene expression recovery for single-cell RNA sequencing. Nat Methods 2018;15:539-42.
[22] Hou W, Ji Z, Ji H, Hicks SC. A systematic evaluation of single-cell RNA-sequencing imputation methods. Genome Biol 2020;21:218.
[23] Pei G, Dai Y, Zhao Z, Jia P. deTS: tissue-specific enrichment analysis to decode tissue specificity. Bioinformatics 2019;35:3842-5.
[24] Martens JH, Stunnenberg HG. BLUEPRINT: mapping human blood cell epigenomes. Haematologica 2013;98:1487-9.
[25] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57-74.
[26] Schmiedel BJ, Singh D, Madrigal A, Valdovino-Gonzalez AG, White BM, Zapardiel-Gonzalo J, et al. Impact of genetic polymorphisms on human immune cell gene expression. Cell 2018;175:1701-15.
[27] Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carré C, et al. RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. Cell Rep 2019;26:1627-40.
[28] McInnes L, Healy J, Saul N, Grossberger L. UMAP: uniform manifold approximation and projection. Journal of Open Source Software 2018;3:861.
[29] Navas A, Giraldo-Parra L, Prieto MD, Cabrera J, Gómez MA. Phenotypic and functional stability of leukocytes from human peripheral blood samples: considerations for the design of immunological studies. BMC Immunology 2019;20:5.
[30] Luckheeram RV, Zhou R, Verma AD, Xia B. $CD4^{+}T$ cells: differentiation and functions. Clin Dev Immunol 2012;2012:925135.
[31] Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. Nat Med 2020;26:842-4.
[32] Kiselev VY, Yiu A, Hemberg M. scmap: projection of single-cell RNA-seq data across data sets. Nat Methods 2018;15:359-62.
[33] Cui Y, Zheng Y, Liu X, Yan L, Fan X, Yong J, et al. Single-cell transcriptome analysis maps the developmental track of the human heart. Cell Rep 2019;26:1934-50.
[34] Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet 2019;20:273-82.
[35] Wu Z, Wu H. Accounting for cell type hierarchy in evaluating single cell RNA-seq clustering. Genome Biol 2020;21:123.
[36] Hoffman JA, Papas BN, Trotter KW, Archer TK. Single-cell RNA sequencing reveals a heterogeneous response to Glucocorticoids in breast cancer cells. Commun Biol 2020;3:126.
[37] Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. Journal of the American statistical Association 1952;47:583-621.
[38] Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun 2018;9:997.
[39] Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nat Commun 2019;10:390.
[40] Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 2018;19:15.
[41] Tung PY, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK, et al. Batch effects and the effective design of single-cell gene expression studies. Sci Rep 2017;7:39921.
[42] Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods 2014;11:740-2.
[43] Wells A, Kopp N, Xu X, O'Brien DR, Yang W, Nehorai A, et al. The anatomical distribution of genetic associations. Nucleic Acids Res 2015;43:10804-20.
[44] Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol 2017;18:220.

[45] Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun 2019;10:380.
[46] Pei G, Wang YY, Simon LM, Dai Y, Zhao Z, Jia P. Gene expression imputation and cell-type deconvolution in human brain with spatiotemporal precision and its implications for brain-related disorders. Genome Res 2021;31:146-58.
[47] Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 2015;47:1091-8.
[48] Liu B, Gloudemans MJ, Rao AS, Ingelsson E, Montgomery SB. Abundant associations with gene expression complicate GWAS follow-up. Nat Genet 2019;51:768-9.
[49] Hoffman GE, Hartley BJ, Flaherty E, Sklar P, Brennand KJ. Transcriptional signatures of schizophrenia in hiPSC-derived NPCs and neurons are concordant with post-mortem adult brains. Nat Commun 2017;8:2225.
[50] Walss-Bass C, Liu W, Lew DF, Villegas R, Montero P, Dassori A, et al. A novel missense mutation in the transmembrane domain of neuregulin 1 is associated with schizophrenia. Biol Psychiatry 2006;60:548-53.
[51] Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 2018;36:421-7.
[52] Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods 2017;14:483-6.
[53] Dai Y, Hu R, Manuel AM, Liu A, Jia P, Zhao Z. CSEA-DB: an omnibus for human complex trait and cell type associations. Nucleic Acids Res 2021;49:D862-D70.
[54] Pei G, Sun H, Dai Y, Jia P, Zhao Z. Investigation of multi-trait associations using pathway-based analysis of GWAS summary statistics. BMC Genomics 2018;20:79.
[55] Pickrell JK, Berisa T, Liu JZ, Ségurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. Nat Genet 2016;48:709-17.
[56] Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 2016;352:189-96.
[57] Guo H, Li J. scSorter: assigning cells to known cell types according to marker genes. Genome Biol 2021;22:69.
[58] Park JE, Botting RA, Domínguez Conde C, Popescu DM, Lavaert M, Kunz DJ, et al. A cell atlas of human thymic development defines T cell repertoire formation. Science 2020;367:eaay3224.
[59] Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, et al. CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res 2019;47:D721-D8.
[60] Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, et al. Mapping the mouse cell atlas by microwell-Seq. Cell 2018;172:1091-107.
[61] Yuan H, Yan M, Zhang G, Liu W, Deng C, Liao G, et al. CancerSEA: a cancer single-cell state atlas. Nucleic Acids Res 2019;47:D900-D8.
[62] Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 2006;22:1540-2.
[63] Skene NG, Bryois J, Bakken TE, Breen G, Crowley JJ, Gaspar HA, et al. Genetic identification of brain cell types underlying schizophrenia. Nat Genet 2018;50:825-33.
[64] Fresno C, Fernandez EA. RDAVIDWebService: a versatile R interface to DAVID. Bioinformatics 2013;29:2810-1.
[65] Benjamini Y, Hochberg Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. J R Statist Soc B 1995;57:289-300.
[66] Shannon P, Markiel A, Ozier O, Baliga N, Wang J, Ramage D, et al. Cytoscape: a software enviroment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498-504.

[67] Liu T, Jia P, Fang B, Zhao Z. Differential expression of viral transcripts from single-cell RNA sequencing of moderate and severe COVID-19 patients and Its implications for case severity. Front Microbiol 2020;11:603509.
[68] MacParland SA, Liu JC, Ma XZ, Innes BT, Bartczak AM, Gage BK, et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. Nat Commun 2018;9:4383.
[69] Madissoon E, Wilbrey-Clark A, Miragaia RJ, Saeb-Parsy K, Mahbubani KT, Georgakopoulos N, et al. scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. Genome Biol 2019;21:1.
[70] Jia P, Dai Y, Hu R, Pei G, Manuel AM, Zhao Z. TSEA-DB: a trait-tissue association map for human complex traits and diseases. Nucleic Acids Res 2020;48:D1022-D30.
[71] Pei G, Hu R, Dai Y, Manuel AM, Zhao Z, Jia P. Predicting regulatory variants using a dense epigenomic mapped CNN model elucidated the molecular basis of trait-tissue associations. Nucleic Acids Res 2021;49:53-66.
[72] Pei G, Hu R, Jia P, Zhao Z. DeepFun: a deep learning sequence-based model to decipher non-coding variant effect in a tissue- and cell type-specific manner. Nucleic Acids Res 2021;49:W131-W9.
[73] Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. PLoS Comput Biol 2016;12:e1004714.
[74] Stertz L, Di Re J, Pei G, Fries GR, Mendez E, Li S, et al. Convergent genomic and pharmacological evidence of PI3K/GSK3 signaling alterations in neurons from schizophrenia patients. Neuropsychopharmacology 2020;46:673-82.

Figure legends

Figure 1. Overview of *deCS* flowchart. For each cell type, we compute *t*-statistics and *z*-scores for each gene in the bulk RNA-seq and scRNA-seq derived references, respectively. Then, we define genes with the highest *t*-statistics or *z*-scores (top 5%) as CTGenes. We further integrate cell signature gene sets from one curated CellMatch database. Depending on the type of "query data", when the query input is gene expression profile, it will calculate the *Pearson correlation coefficient* (PCC) or *Spearman's rank correlation coefficient* between query scaled expression profiles and *t*-statistics (or z-scores) of each cell type in the reference, then assign the label with the highest score to the query profile. When the query input is list of genes, it is analogous to existing tools for identifying candidate genes that are overrepresented in specific Gene Ontology (GO) terms or KEGG pathways [64]. Finally, the top enriched cell type is annotated to query data.

Figure 2. KEGG pathway enrichment of cell type-specific genes. The network shows the relationship between CTGenes and KEGG pathways with significant association (FDR < 0.05). The network layout is based on a force-directed graph. Edge width is proportional to $-\log_{10}$(transformed FDR).

Figure 3. Comparison of the two methods in *deCS* using PBMC scRNA-seq data. (A) Correlation analysis by PCC. **(B)** Fisher's exact test. *X*-axis: 9 cell type clusters (query). *Y*-axis: the 10 main (up panel) and 29 fine (bottom panel) immune cell types from the MonacoImmune reference [27]. The colors represent the PCC or genes intersection (IR) ratio, and the sizes represent the $-\log_{10}$(transformed *p*-value). Non-significant associations (*p*-value > 0.05) were labeled by white color.

Figure 4. *deCS* annotation of bronchoalveolar immune cells from COVID-19 infected patients. (A) Correlation analysis by PCC. **(B)** Fisher's exact test between 10 cell major clusters marker genes and CTGenes in BlueprintEncode reference. *X*-axis: 10 cell type clusters (query). *Y*-axis: the 24 main cell types from BlueprintEncode. The colors represent the PCC or genes intersection ratio, and the sizes represent the -$\log_{10}$(transformed *p*-value). **(C)** Individual cell level annotation of 63,103 single cells profiled on the 10x Genomics platform. We annotated each cell by using the top matched mode (cell type with the highest PCC). Using integrated references (BlueprintEncode: green; MonacoImmune: blue; HCL: red), *deCS* classifies all cells that are nearly identical to that by manual classification. The colors and the numbers representing the percentages and the total cell counts were annotated to a given cell type.

Figure 5. Comparison of *deCS* annotation on cells with different sequencing depth and individuals. (A) Annotation comparison of cells sequenced among low (1k ≤ UMI < 3k), median (3k ≤ UMI < 6k) and high (UMI ≥ 6k) depth groups. **(B)** Annotation comparison of cells derived from healthy controls (HC) and patients with moderate (M) and severe (S) COVID-19 infection groups. We only use the union of top 20 marker genes of each cluster. The colors and the numbers represent the percentages and the total cell counts that were annotated to a given cell type. Only cell types are annotated with at least 5% cells from 10 predefined cell clusters were labeled with the number of cell counts.

Figure 6. *deCS* pipeline reduces the collinearity and impact of sequencing depth difference. Correlation comparison by using the union of cell cluster-specific gene and top 2000 most variable genes between query (CD4 T) cells and reference on **(A)** CD4 T (positive) or **(B)** CD8 T cells (negative). **(C)** Rarefaction curve of T cells in PBMC and BIC data sets. *X*-axis: the

sized pools (n from 1 to 50). *Y*-axis: the inter-correlation between the gene expression of pooled cells and bulk level (cluster averaged). Plotted are interquartile ranges (IQRs; boxes). The red, green and blue lines indicate the number of single cells that hypothetically enabled the averaged expression levels of pooled cells are approximate to bulk populations (with PCC > 0.9, 0.95 and 0.975). **(D, E)** Uniform manifold approximation and projection (UMAP) integrative analysis of PBMC and BIC data sets by **(D)** log normalized expression matrix or **(E)** further z-score scaled. Each circle and triangle represents a single cell derived from BIC or PBMC data sets.

Figure 7. Association between trait-associated genes and cell type-specific genes. 51 traits were analyzed. Heat map shows the significant trait-tissue associations (*p*-value < 0.05). The numbers on the cells indicate the shared gene counts between TAGs and CTGenes in BlueprintEncode. The colors and sizes represent the shared gene counts and the $-\log_{10}$(transformed *p*-value). Since more significant SNPs are likely found in immune-related diseases and blood-related traits than other traits, we recommend users to focus on the top 3 most relevant cell types for those traits with higher TAGs number.

Table

Table 1. Performance comparison of *deCS* with six other models among eight data sets.

ID	Tissue name	deCS(Cor)	deCS(FET)*	SingleR	scPred	sciBet	CHEATH	Cell blast	scSorter**
1	PBMC1	88.74%	-	87.81%	88.78%	87.55%	54.34%	77.38%	75.81%
2	PBMC2	88.52%	-	87.81%	88.59%	87.79%	54.42%	78.40%	71.86%
3	Bronchoalveolar	97.34%	95.79%	71.34%	<50%	91.54%	<50%	<50%	-
4	Heart	80.97%	80.97%	63.33%	<50%	<50%	<50%	<50%	-
5	Liver	99.56%	92.37%	<50%	<50%	83.99%	<50%	<50%	-
6	Lung	98.59%	87.78%	70.72%	<50%	67.77%	<50%	<50%	-
7	Oesophagus	87.64%	87.07%	86.82%	<50%	96.34%	<50%	<50%	-
8	Spleen	97.34%	90.54%	91.90%	79.63%	90.35%	<50%	<50%	-

* We benchmarked PBMC data sets at the single cell level. Because deCS(FET) model needs cell-cluster level information, which is not available from these PBMC data sets.

** scSorter is a “semi-supervised” cell type assignment method. Because the prior knowledge of markers would strongly impact the annotation, we did not benchmark scSorter on 6 solid tissues.

Supplementary materials

Figure S1. Summary of three bulk-RNA derived cell type references. (A-C) *Spearman's rank* correlation analysis for each cell type by using the average expression for each gene. **(D-F)** Uniform Manifold Approximation and Projection (UMAP) dimension reduction analysis investigated the global landscape of cell type expression profiles. **(G-I)** Variance comparison of different gene features. Each dot represents one gene feature. *X*-axis: type of gene feature (e.g. protein coding, lincRNA). *Y*-axis: the variance of gene feature, scaled by absolute abundance.

Figure S2. Shared cell type-specific genes comparison between bulk-RNA and scRNA-seq derived references. (A-C) Main CTGenes comparison between BlueprintEncode and ImmuneCellExpression, BlueprintEncode and MonacoImmune, ImmuneCellExpression and MonacoImmune. **(D-F)** Fine CTGenes comparison between BlueprintEncode and ImmuneCellExpression, BlueprintEncode and MonacoImmune, ImmuneCellExpression and MonacoImmune. **(G)** CTGenes comparison between BlueprintEncode and human cell landscape. The number indicates the shared CTGenes number between two cell types.

Figure S3. Overview of *deCS* flowchart for scRNA-seq annotation. We recommend users apply *Seurat* standard pre-processing workflow, including transcription noise cells cleaning, highly variable genes identification, dimension reduction and cell clustering analysis. Then, *Seurat* FindAllMarkers() function can use to identify cluster-specific marker genes, and AverageExpression() function can use to calculate the gene average expression level for each cell cluster. Base on provided *t*-statistics and *z*-score for each gene, we implemented two test approaches for scRNA-seq cell type annotation. Strategy A: user can define genes with the highest *t*-statistics or *z*-score (top 5%) as CTGenes, then investigate whether a set of

"cluster-specific marker genes" of interest are disproportionately overlapped with CTGenes by using Fisher's exact test. Strategy B: user can calculate the *Pearson correlation coefficient* (PCC) or *Spearman's rank correlation coefficient* between query scaled expression profiles and *t*-statistics (or *z*-score) of each cell type in reference, then assigned the label with the highest score to query profile.

Figure S4. Cell type annotation of bronchoalveolar immune cells data set by using different references. (A) BlueprintEncode. **(B)** Human cell landscape. **(C)** MonacoImmune. *X*-axis: 10 cell type clusters (query). *Y*-axis: the reference cell types. The colors representing the PCC or genes intersection ratio, and the sizes representing the -log10 transformed *p*-value. Non-significant associations (*p*-value > 0.05) were replaced by blank color.

Figure S5. Cell type annotation of cardiac cells from human fetal cardiac cells by using different reference or correlation measurement. (A-C) *Pearson correlation coefficient* on BlueprintEncode, human cell landscape and human cell atlas of fetal. **(D-F)** *Spearman's correlation coefficient* on BlueprintEncode, human cell landscape and human cell atlas of fetal. *X*-axis: 9 cell type clusters (query). *Y*-axis: the reference cell types. The colors representing the PCC or *Spearman's correlation coefficient*, and the sizes representing the -log10 transformed *p*-value. Non-significant associations (*p*-value > 0.05) were replaced by blank color.

Figure S6 Overview of cell type trajectory in human immune system.

Figure S7 Hierarchical clustering analysis of different cell type in each reference. Numbers at the nodes indicate bootstrap values for 100 replicates.

Figure S8. Violin plot shows the expression probability distributions across each cluster's canonical marker genes (top 1) in PBMC data set. Each dot represents a single cell.

Figure S9. Single cell annotation results by CHETAH model. (A) Top 30 marker genes for each cell cluster. **(B)** Top 2000 most variable genes. **(C)** All detected genes.

Figure S10. Principal component analysis (PCA) of down-sampled bulk RNA-seq data set. (A) Sequencing depth identity. **(B)** Sample identity based on raw TPM value. Z-score normalization by **(C)** sequencing depth and **(D)** sample identity. S1 to S5 denote five samples NSC353, NSC412, NSC413, NSC416 and NSC419, respectively.

Figure S11. Association between 51 trait-associated genes and cell type-specific genes. Heat map shows the significant trait-tissue associations (*p*-value < 0.05). The numbers on the cells indicate the shared gene counts between trait-associated genes and CTGenes in human cell landscape. The colors and sizes represent the shared gene counts and the $-\log_{10}$(transformed *p*-value). Since more significant SNPs are likely found in immune-related diseases and blood-related traits than other traits, we recommend users to focus on the top 3 most relevant cell types for those traits with higher TAGs number.

Figure S12. *deCS* annotation on bulk RNA-seq of iPSC-derived cell lines. iPSC-NPCs (n=13) and iPSC-neurons (n=11).

Table S1. GO and KEGG pathway enrichment analysis of cell type-specific genes.

Table S2. *deCS* annotation results on 6 different solid tissues.

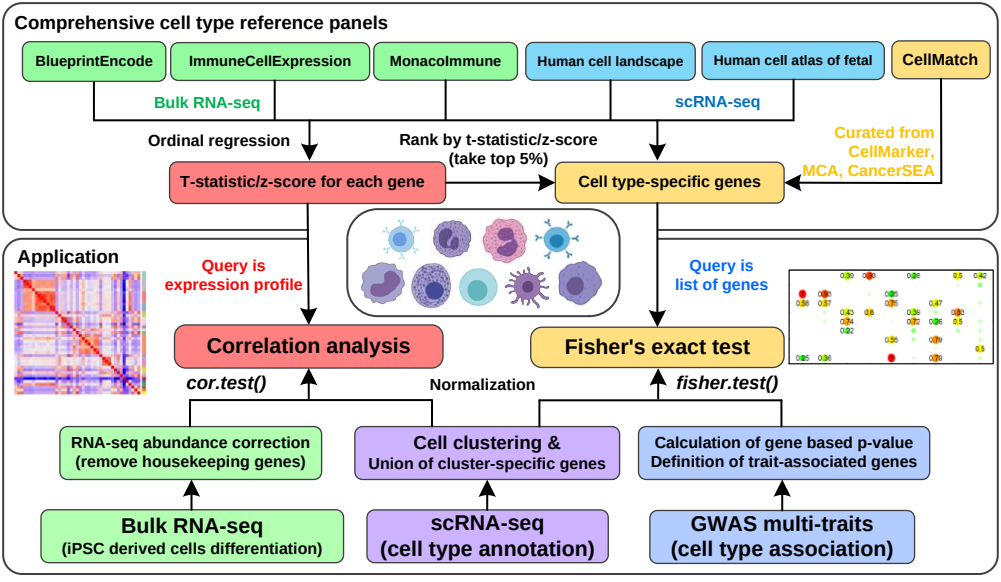
Table S3. Performance comparison of *deCS* and other software on PBMC data set.

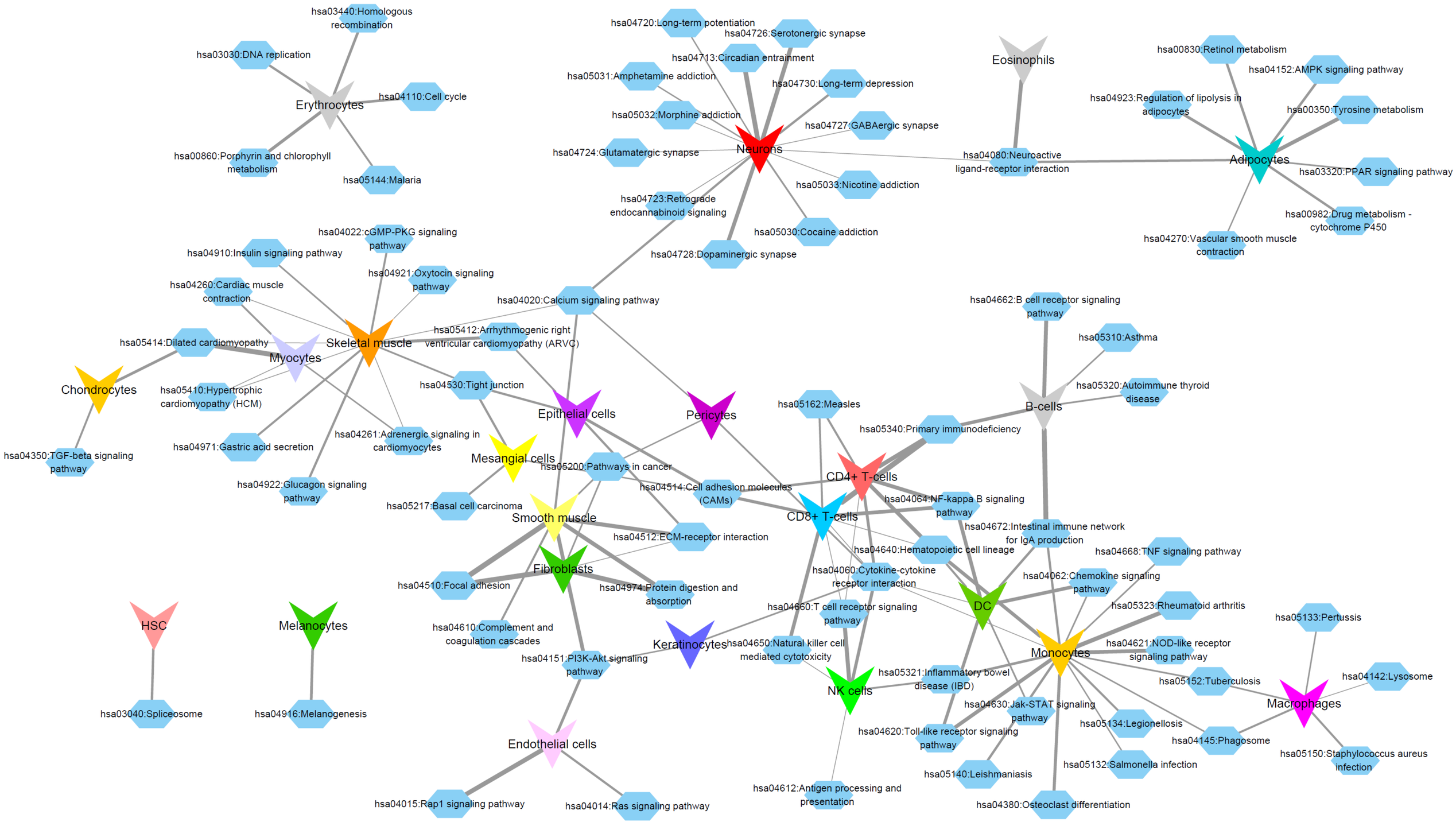
Table S4. Simulated TPM expression profiles of 5 down-sampled bulk RNA-seq samples.

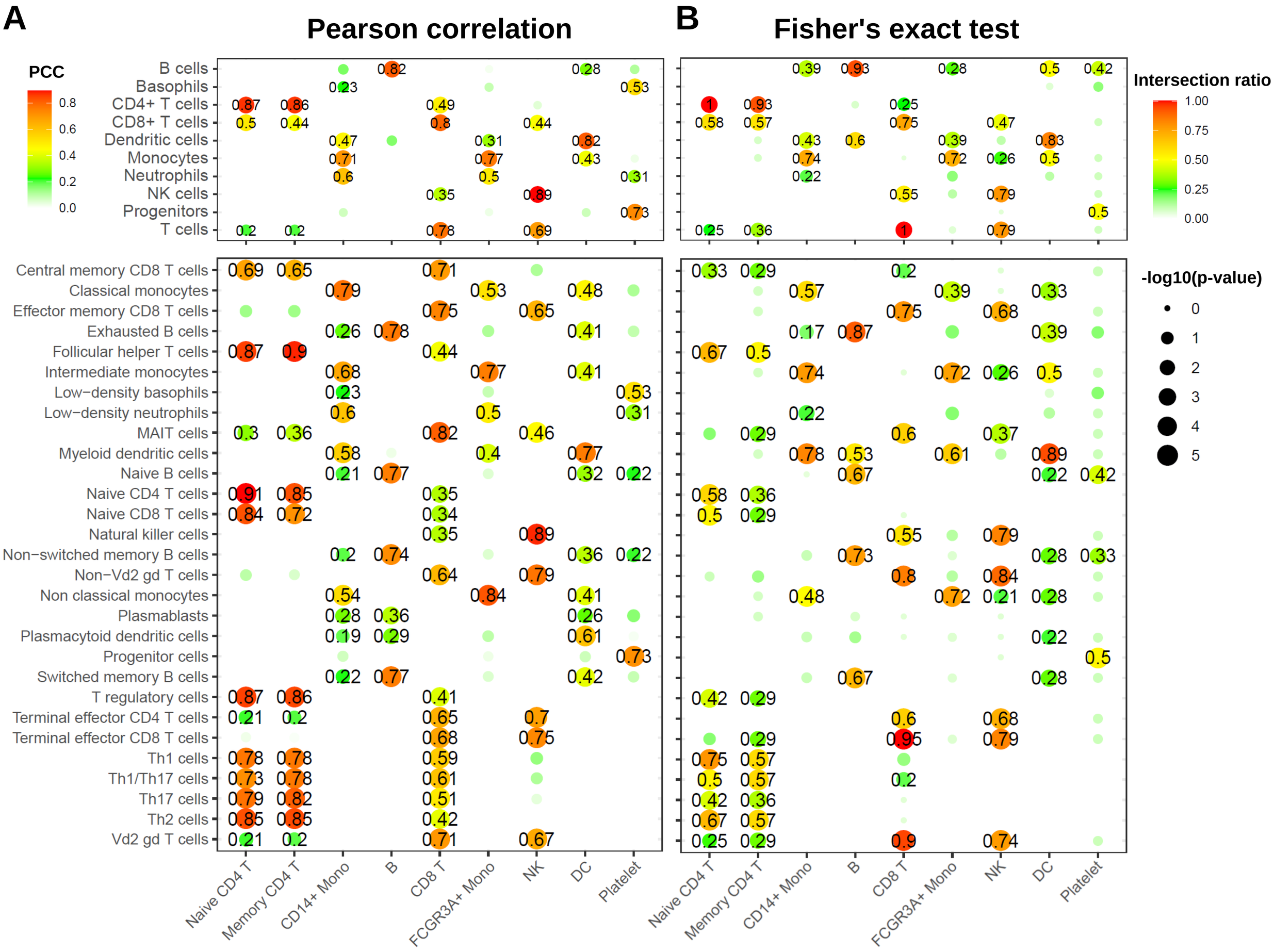
Table S5. GWAS summary and regulatory loci associated genes statistics for 51 human complex traits.

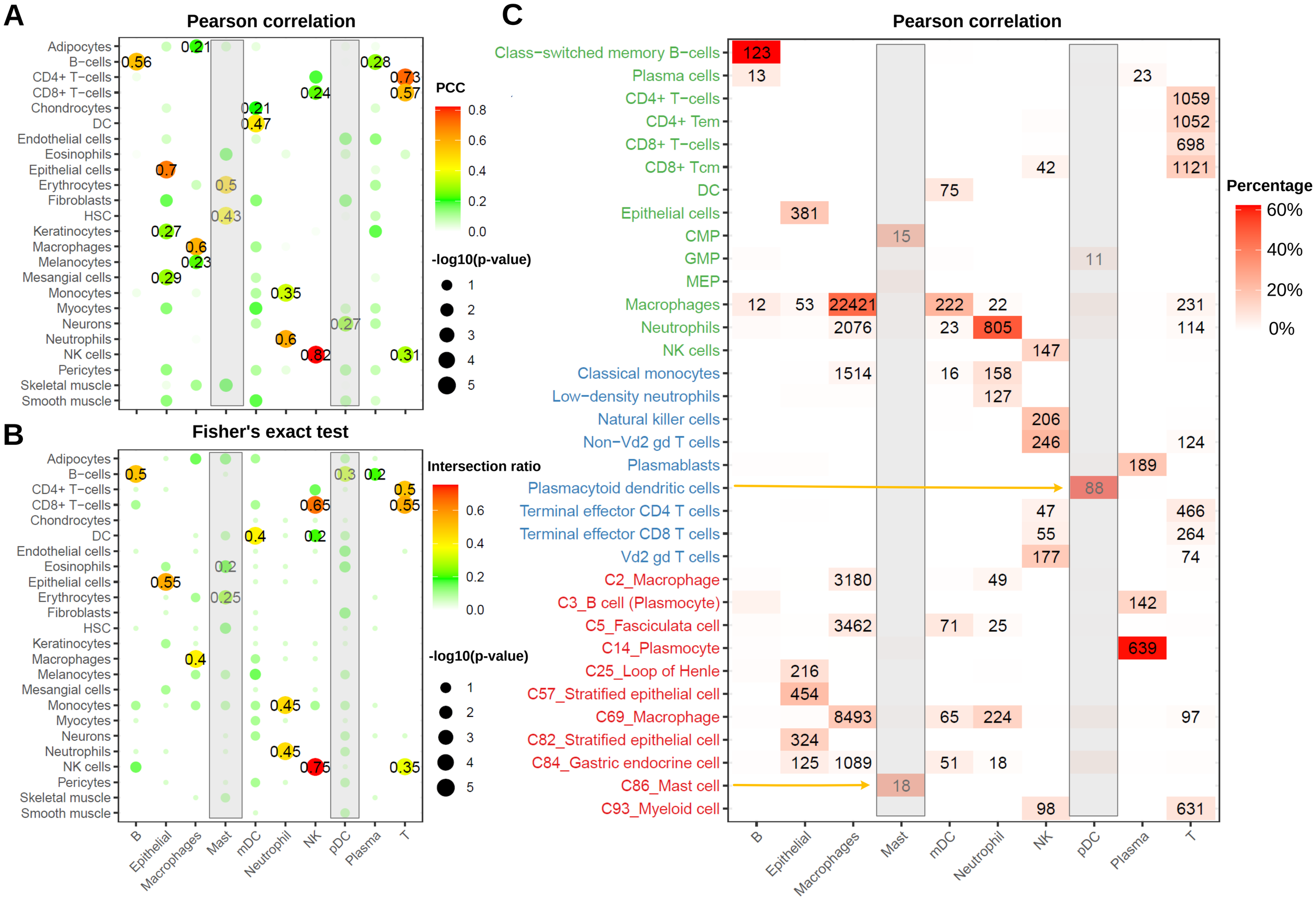
Table S6. *deCS* annotation results consistency comparison between different clustering methods or under different clustering resolution.

Table S7. Summary of cell type composition of major human organs from human cell landscape reference.


Comprehensive cell type reference panels
BlueprintEncode
ImmuneCellExpression
MonacoImmune
Human cell landscape
Human cell atlas of fetal
CellMatch
Bulk RNA-seq
scRNA-seq
Ordinal regression
Rank by t-statistic/z-score
(take top 5%)
Curated from
CellMarker,
MCA, CancerSEA
T-statistic/z-score for each gene
Cell type-specific genes
Application
Query is
expression profile
Query is
list of genes
Correlation analysis
Fisher's exact test
cor.test()
Normalization
fisher.test()
RNA-seq abundance correction
(remove housekeeping genes)
Cell clustering &
Union of cluster-specific genes
Calculation of gene based p-value
Definition of trait-associated genes
Bulk RNA-seq
(iPSC derived cells differentiation)
scRNA-seq
(cell type annotation)
GWAS multi-traits
(cell type association)

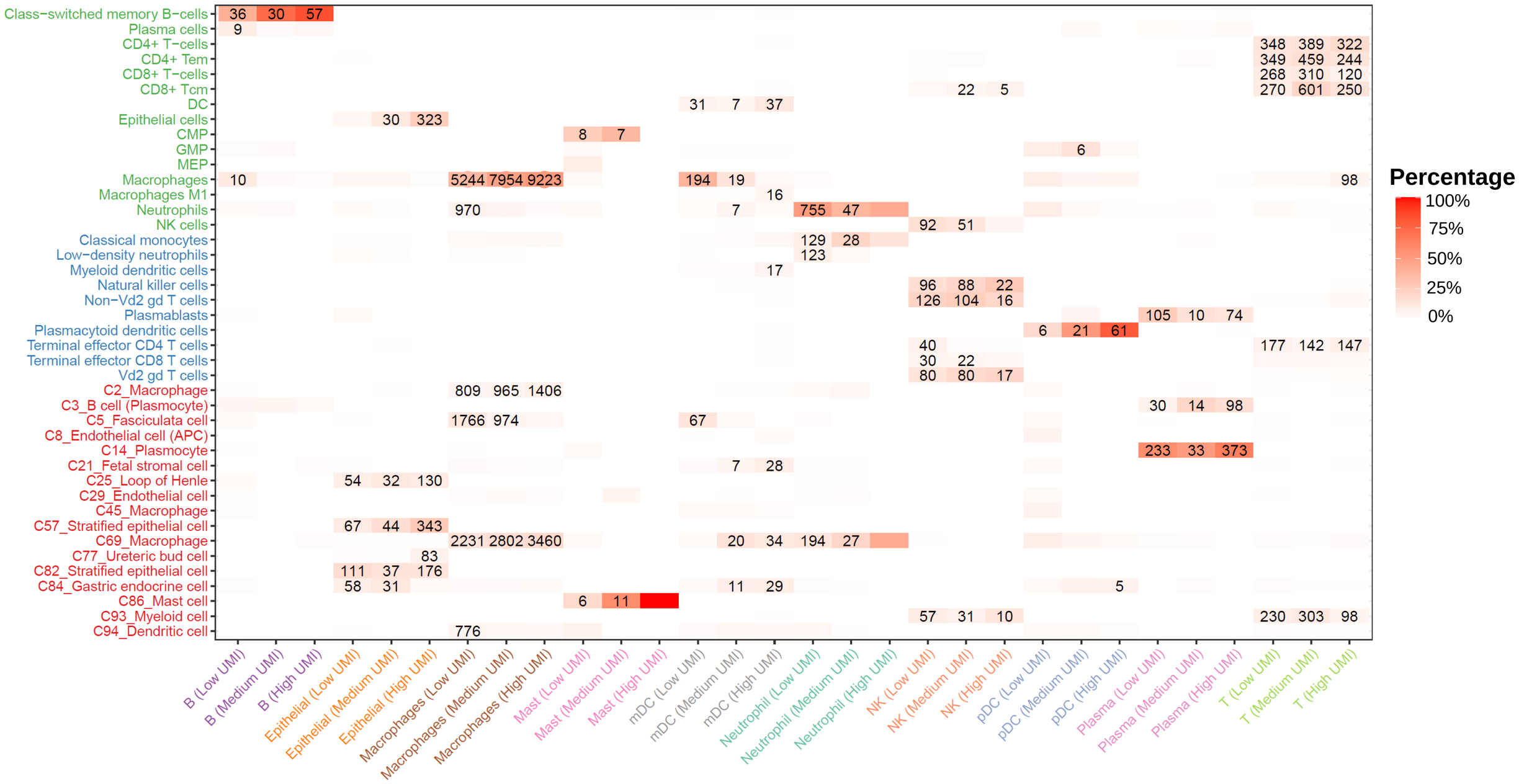

hsa03440:Homologous recombination
hsa03030:DNA replication
Erythrocytes
hsa04110:Cell cycle
hsa00860:Porphyrin and chlorophyll metabolism
hsa05144:Malaria
hsa04720:Long-term potentiation
hsa04726:Serotonergic synapse
hsa04713:Circadian entrainment
hsa05031:Amphetamine addiction
hsa04730:Long-term depression
hsa05032:Morphine addiction
hsa04727:GABAergic synapse
hsa04724:Glutamatergic synapse
Neurons
hsa05033:Nicotine addiction
hsa04723:Retrograde endocannabinoid signaling
hsa05030:Cocaine addiction
hsa04728:Dopaminergic synapse
Eosinophils
hsa04080:Neuroactive ligand-receptor interaction
hsa00830:Retinol metabolism
hsa04152:AMPK signaling pathway
hsa04923:Regulation of lipolysis in adipocytes
hsa00350:Tyrosine metabolism
Adipocytes
hsa03320:PPAR signaling pathway
hsa00982:Drug metabolism - cytochrome P450
hsa04270:Vascular smooth muscle contraction
hsa04022:cGMP-PKG signaling pathway
hsa04910:Insulin signaling pathway
hsa04921:Oxytocin signaling pathway
hsa04260:Cardiac muscle contraction
hsa04020:Calcium signaling pathway
hsa05414:Dilated cardiomyopathy
Skeletal muscle
hsa05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC)
Myocytes
Chondrocytes
hsa05410:Hypertrophic cardiomyopathy (HCM)
hsa04530:Tight junction
Epithelial cells
Pericytes
hsa04350:TGF-beta signaling pathway
hsa04971:Gastric acid secretion
hsa04261:Adrenergic signaling in cardiomyocytes
Mesangial cells
hsa05200:Pathways in cancer
hsa04922:Glucagon signaling pathway
hsa05217:Basal cell carcinoma
hsa04514:Cell adhesion molecules (CAMs)
Smooth muscle
hsa04512:ECM-receptor interaction
Fibroblasts
hsa04510:Focal adhesion
hsa04974:Protein digestion and absorption
hsa04610:Complement and coagulation cascades
hsa04151:PI3K-Akt signaling pathway
HSC
hsa03040:Spliceosome
Melanocytes
hsa04916:Melanogenesis
Keratinocytes
Endothelial cells
hsa04015:Rap1 signaling pathway
hsa04014:Ras signaling pathway
hsa04662:B cell receptor signaling pathway
hsa05310:Asthma
hsa05320:Autoimmune thyroid disease
B-cells
hsa05162:Measles
hsa05340:Primary immunodeficiency
CD4+ T-cells
hsa04064:NF-kappa B signaling pathway
CD8+ T-cells
hsa04672:Intestinal immune network for IgA production
hsa04640:Hematopoietic cell lineage
hsa04668:TNF signaling pathway
hsa04060:Cytokine-cytokine receptor interaction
hsa04062:Chemokine signaling pathway
DC
hsa04660:T cell receptor signaling pathway
hsa05323:Rheumatoid arthritis
hsa05133:Pertussis
hsa04650:Natural killer cell mediated cytotoxicity
Monocytes
hsa04621:NOD-like receptor signaling pathway
hsa05321:Inflammatory bowel disease (IBD)
hsa05152:Tuberculosis
hsa04142:Lysosome
NK cells
hsa04630:Jak-STAT signaling pathway
Macrophages
hsa05134:Legionellosis
hsa04620:Toll-like receptor signaling pathway
hsa04145:Phagosome
hsa05150:Staphylococcus aureus infection
hsa05132:Salmonella infection
hsa05140:Leishmaniasis
hsa04612:Antigen processing and presentation
hsa04380:Osteoclast differentiation


A
Pearson correlation
B
Fisher's exact test
PCC
0.8
0.6
0.4
0.2
0.0
Intersection ratio
1.00
0.75
0.50
0.25
0.00
-log10(p-value)
0
1
2
3
4
5
B cells
Basophils
CD4+ T cells
CD8+ T cells
Dendritic cells
Monocytes
Neutrophils
NK cells
Progenitors
T cells
Central memory CD8 T cells
Classical monocytes
Effector memory CD8 T cells
Exhausted B cells
Follicular helper T cells
Intermediate monocytes
Low−density basophils
Low−density neutrophils
MAIT cells
Myeloid dendritic cells
Naive B cells
Naive CD4 T cells
Naive CD8 T cells
Natural killer cells
Non−switched memory B cells
Non−Vd2 gd T cells
Non classical monocytes
Plasmablasts
Plasmacytoid dendritic cells
Progenitor cells
Switched memory B cells
T regulatory cells
Terminal effector CD4 T cells
Terminal effector CD8 T cells
Th1 cells
Th1/Th17 cells
Th17 cells
Th2 cells
Vd2 gd T cells
Naive CD4 T
Memory CD4 T
CD14+ Mono
B
CD8 T
FCGR3A+ Mono
NK
DC
Platelet


A
Pearson correlation
Adipocytes
B−cells
CD4+ T−cells
CD8+ T−cells
Chondrocytes
DC
Endothelial cells
Eosinophils
Epithelial cells
Erythrocytes
Fibroblasts
HSC
Keratinocytes
Macrophages
Melanocytes
Mesangial cells
Monocytes
Myocytes
Neurons
Neutrophils
NK cells
Pericytes
Skeletal muscle
Smooth muscle
PCC
0.8
0.6
0.4
0.2
0.0
-log10(p-value)
1
2
3
4
5
B
Fisher's exact test
Intersection ratio
0.6
0.4
0.2
0.0
-log10(p-value)
1
2
3
4
5
B
Epithelial
Macrophages
Mast
mDC
Neutrophil
NK
pDC
Plasma
T
C
Pearson correlation
Class−switched memory B−cells
Plasma cells
CD4+ T−cells
CD4+ Tem
CD8+ T−cells
CD8+ Tcm
DC
Epithelial cells
CMP
GMP
MEP
Macrophages
Neutrophils
NK cells
Classical monocytes
Low−density neutrophils
Natural killer cells
Non−Vd2 gd T cells
Plasmablasts
Plasmacytoid dendritic cells
Terminal effector CD4 T cells
Terminal effector CD8 T cells
Vd2 gd T cells
C2_Macrophage
C3_B cell (Plasmocyte)
C5_Fasciculata cell
C14_Plasmocyte
C25_Loop of Henle
C57_Stratified epithelial cell
C69_Macrophage
C82_Stratified epithelial cell
C84_Gastric endocrine cell
C86_Mast cell
C93_Myeloid cell
B
Epithelial
Macrophages
Mast
mDC
Neutrophil
NK
pDC
Plasma
T
Percentage
60%
40%
20%
0%

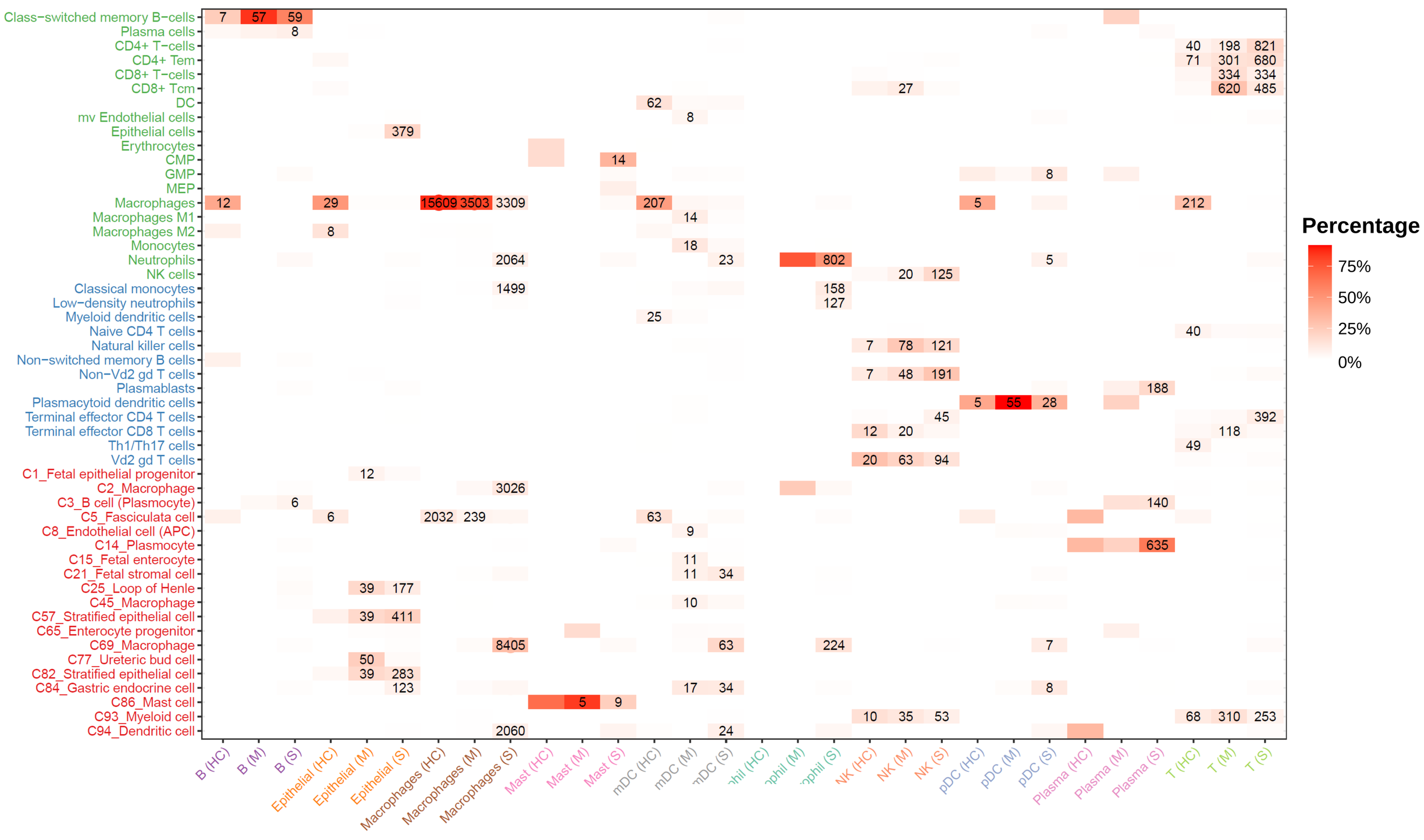
A

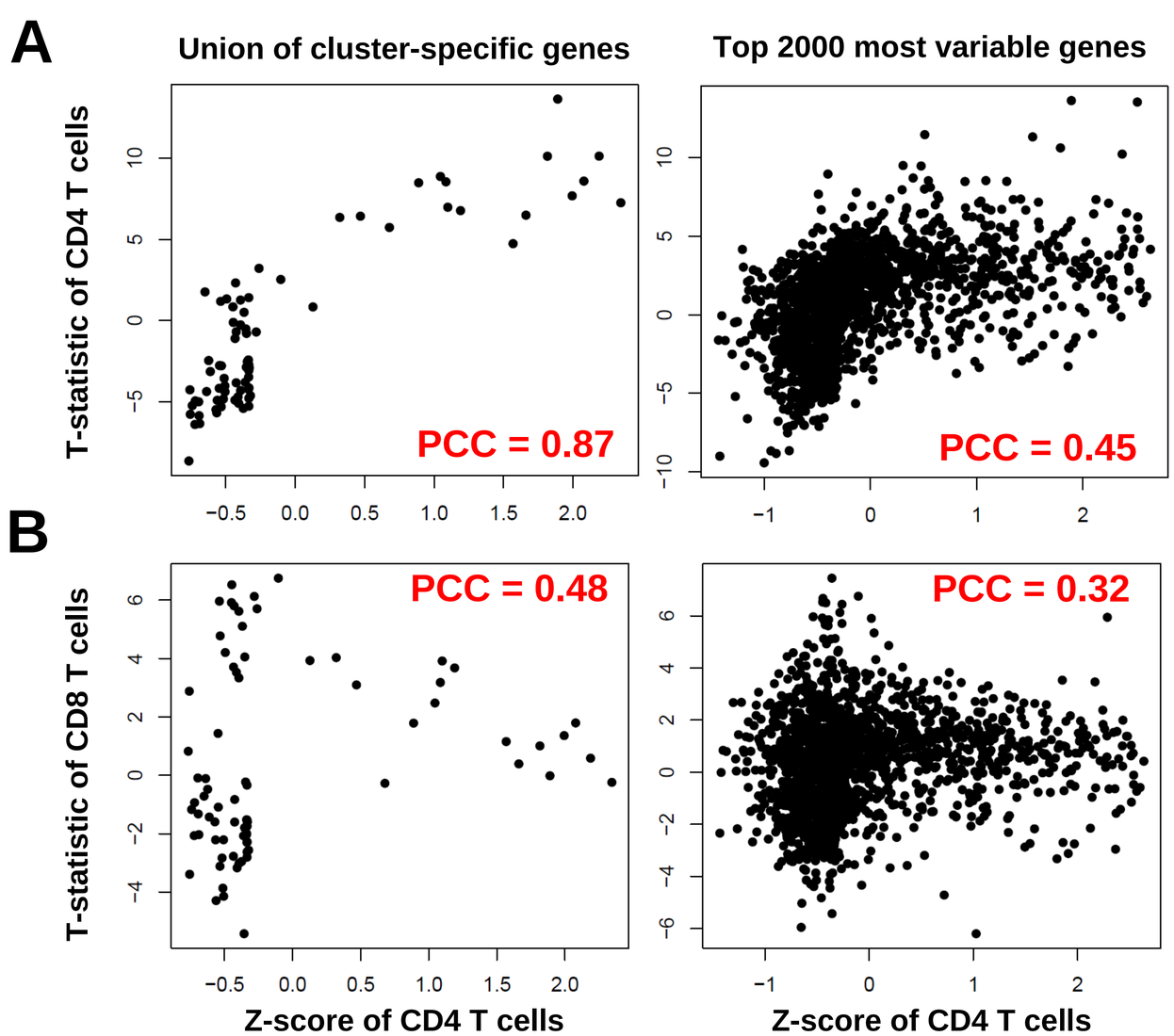
Performance comparison of *deCS* annotation on cells with different sequencing depth

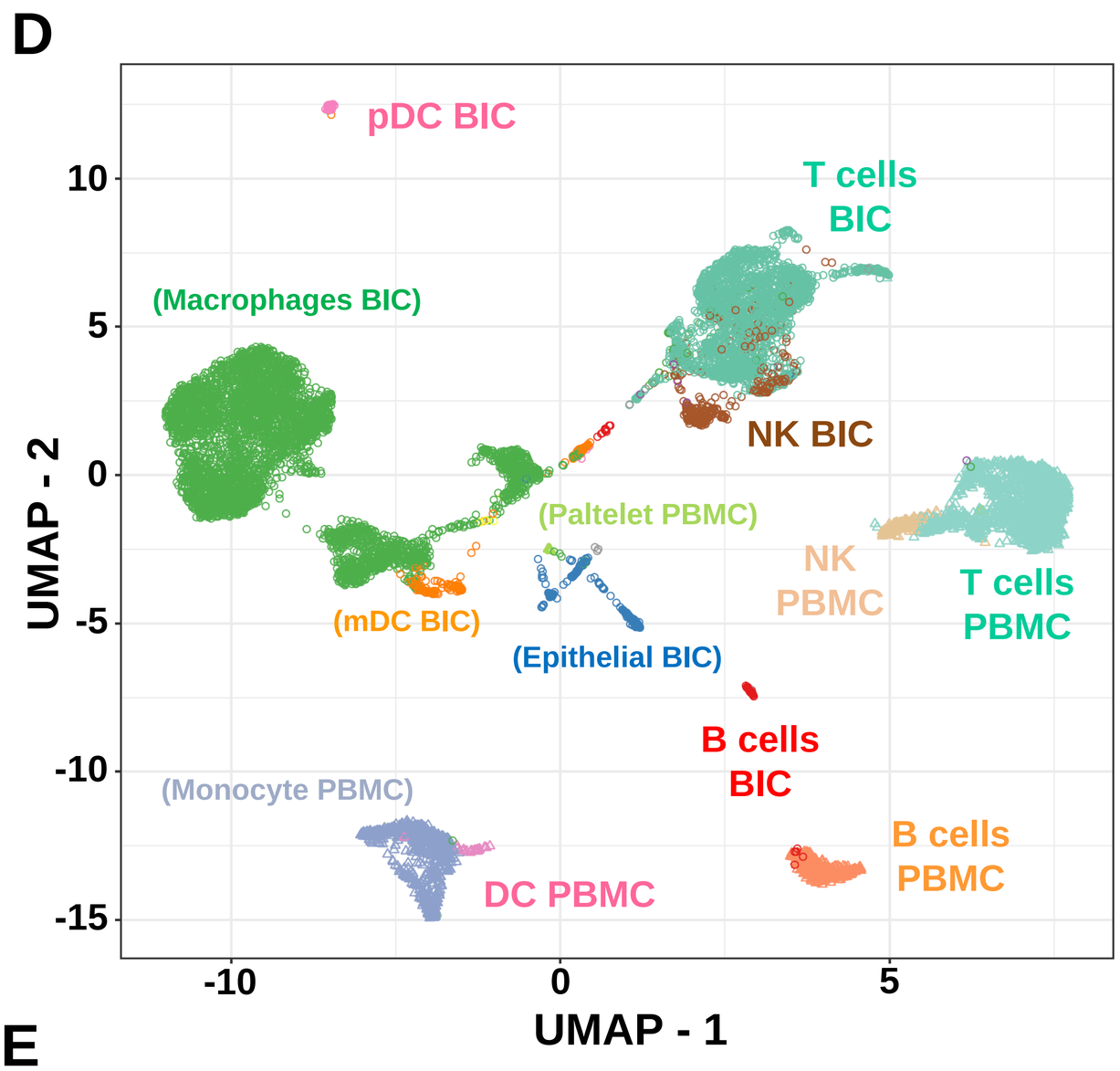


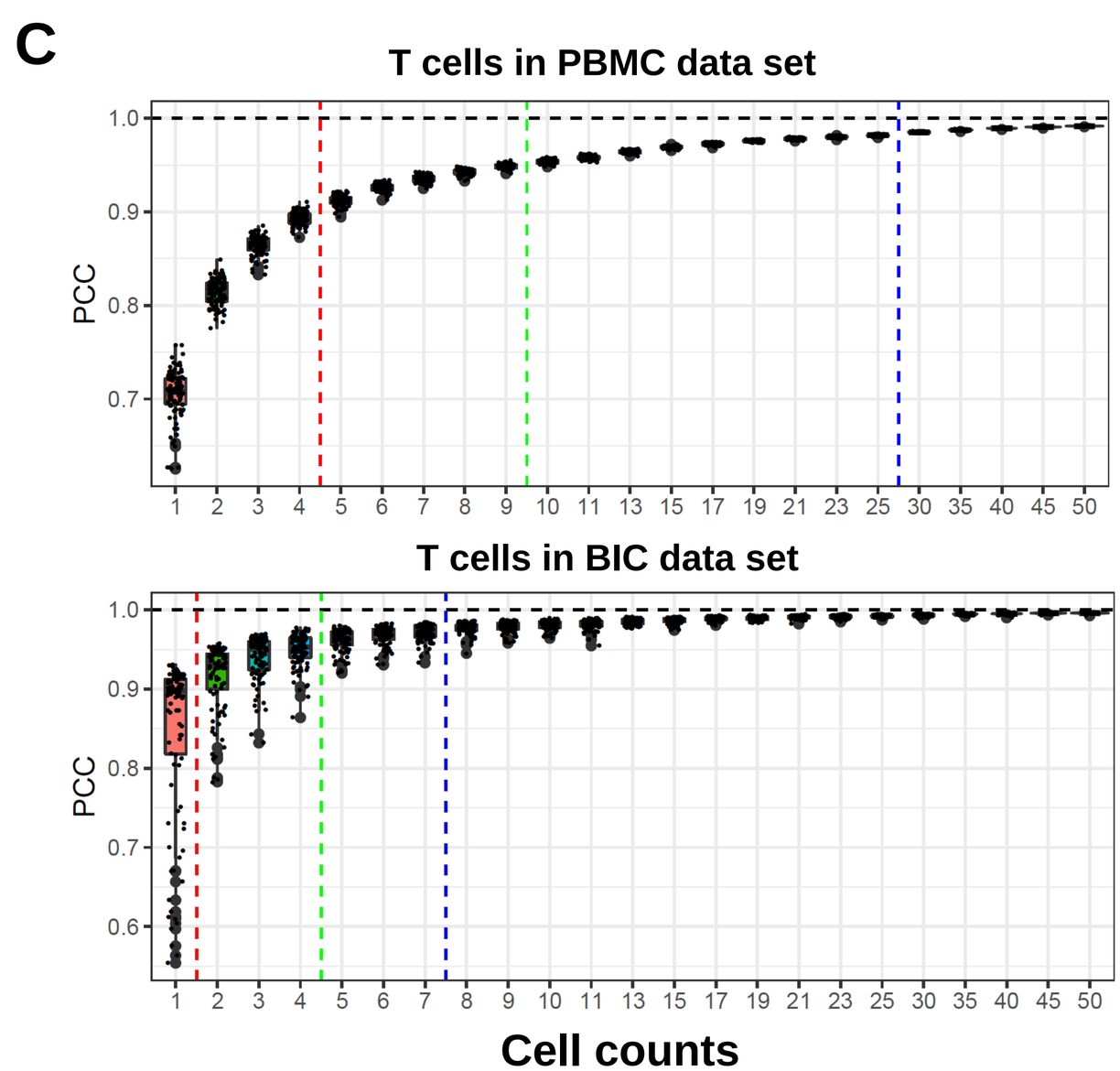
B

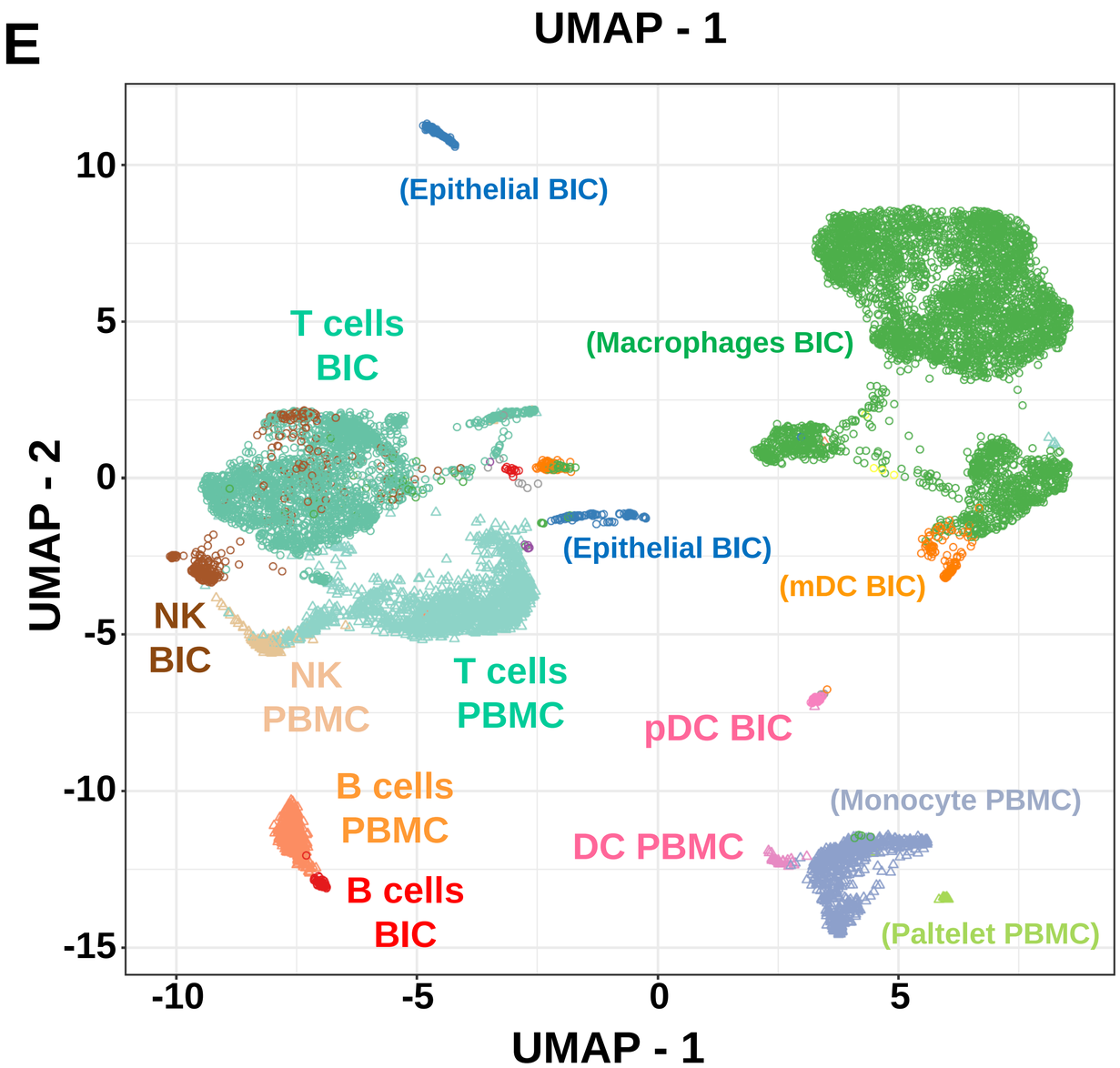
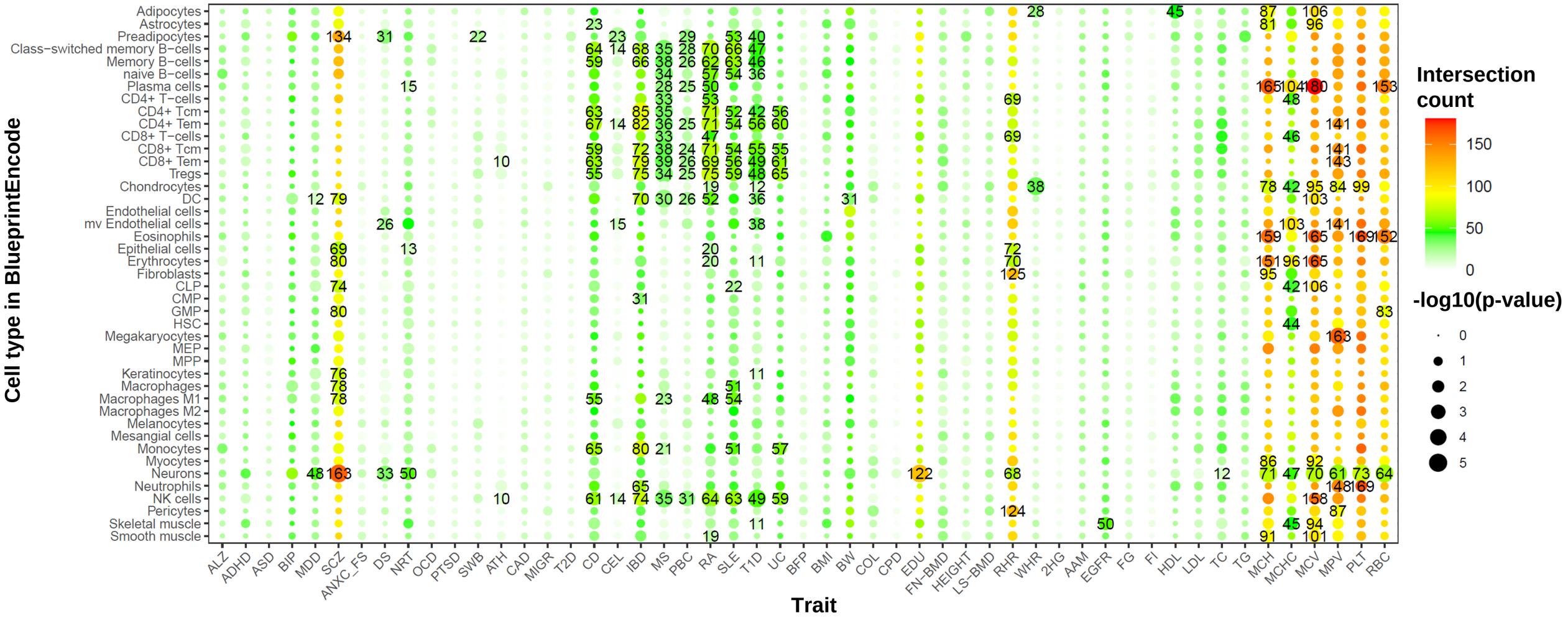
Performance comparison of *deCS* annotation on cells with different individuals




A
Union of cluster-specific genes
Top 2000 most variable genes
T-statistic of CD4 T cells
PCC = 0.87
PCC = 0.45
B
T-statistic of CD8 T cells
PCC = 0.48
PCC = 0.32
Z-score of CD4 T cells
Z-score of CD4 T cells


D
pDC BIC
T cells BIC
(Macrophages BIC)
NK BIC
(Paltelet PBMC)
NK PBMC
T cells PBMC
(mDC BIC)
(Epithelial BIC)
B cells BIC
(Monocyte PBMC)
B cells PBMC
DC PBMC
UMAP - 2
UMAP - 1


C
T cells in PBMC data set
T cells in BIC data set
PCC
Cell counts


E
(Epithelial BIC)
T cells BIC
(Macrophages BIC)
(Epithelial BIC)
(mDC BIC)
NK BIC
NK PBMC
T cells PBMC
pDC BIC
B cells PBMC
(Monocyte PBMC)
DC PBMC
B cells BIC
(Paltelet PBMC)
UMAP - 2
UMAP - 1

Association between trait-associated genes and cell type-specific genes