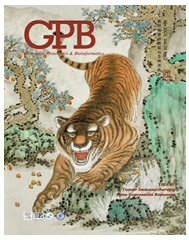
Genomics Proteomics Bioinformatics 21 (2023) 370–384

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com



METHOD

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*

Received 19 September 2021; revised 25 March 2022; accepted 7 April 2022
Available online 22 April 2022

Handled by Jin Gu

KEYWORDS

Cell type-specific enrichment analysis;
scRNA-seq;
Cell type annotation;
Trait–cell type association;
Cell type marker gene

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 *a priori* knowledge and manual curation, which is cumbersome and subjective. The increasing number of scRNA-seq datasets, as well as numerous published genetic studies, has motivated us to build a comprehensive human cell type reference atlas. Here, we present decoding Cell type Specificity (*deCS*), an automatic **cell type annotation** method augmented by a comprehensive collection of human cell type expression profiles and marker genes. We used *deCS* to annotate scRNA-seq data from various tissue types and systematically evaluated the annotation accuracy under different conditions, including reference panels, sequencing depth, and feature selection strategies. Our results demonstrate that expanding the references is critical for improving annotation accuracy. Compared to many existing state-of-the-art annotation tools, *deCS* significantly reduced computation time and 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

* Corresponding authors.
E-mail: zhongming.zhao@uth.tmc.edu (Zhao Z), pjia@big.ac.cn (Jia P).
§ Current address: CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2022.04.001

identify **trait–cell type associations** in 51 human complex traits, providing deep insights into the cellular mechanisms underlying disease pathogenesis. All documents for *deCS*, including source code, user manual, demo data, and tutorials, are freely available at https://github.com/bsml320/deCS.

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 deep 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. 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 have recently been developed, including SingleR [7], CHETAH [8], CellAssign [9], scMatch [10], scCATCH [11], *scPred* [12], SciBet [13], scVI [14], and Cell BLAST [15]. A recent comprehensive benchmarking evaluation by Abdelaal et al. revealed a large difference in performance among these methods, and such performance was strongly dependent on the reference database used [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 are 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 the 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" datasets 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 the single-cell level rather than cluster level. The annotation accuracy will significantly decrease, especially for cells with low sequencing depth due to the dropout effect in scRNA-seq data. Even though imputation methods [21,22] can be applied to recover missing gene expression data and improve the annotation accuracy, they are time-consuming [22].

Here, we present decoding Cell type Specificity (*deCS*), 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 at the cell cluster level rather than the 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 datasets from various tissues or sequencing protocols, such as HCL and HCAF gene expression. Then, we calculated the *t*-statistic- or *z*-score-based measurements to "decode cell type specificity" of gene sets. The *deCS* algorithm was implemented in an R package with different statistical methods. It supports input for either gene expression profiles (*e.g.*, a query scRNA-seq dataset with clusters to be annotated) or list of genes (*e.g.*, a query list of genes). Benchmark 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 datasets are not conserved. Therefore, we anticipate that *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 genes. 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 insights into the cellular mechanisms underlying the poorly understood traits and diseases.

Method

Data collection for cell type reference panels

The BlueprintEncode data

We downloaded the BlueprintEncode RNA-seq data from Aran and his colleagues [7]. The raw data comprised 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 were normalized by transcripts per million (TPM), followed by $\log_2$ transformation.

The database of immune cell expression data

The database of immune cell expression (DICE) reference contained 1561 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 [24].

The MonacoImmune data

The MonacoImmune reference comprised 114 peripheral blood mononuclear cell (PBMC) bulk RNA-seq samples from 4 Singaporean-Chinese individuals. We downloaded the TPM-normalized values for 10 main (29 fine) immune cell types (GEO: GSE107011) from the study by Monaco and his colleagues [25].

The HCL data

The 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 based on known marker genes. These data were derived from 18 fetal tissues, 35 adult tissues, and several intermediate tissues (*e.g.*, cord blood and placenta). We downloaded the gene expression profiles from http://bis.zju.edu.cn/HCL/landscape.html.

The HCAF data

The 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. The raw data are available at https://descartes.brotmanbaty.org/bbi/human-gene-expression-during-development/.

CellMatch

The CellMatch reference [11] was curated from various resources, including CellMarker [26], MCA [27], CancerSEA [28], 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 dataset is available at https://github.com/ZJUFanLab/scCATCH.

Measurement of cell type specificity of genes

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 [23] 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 $\log_2$-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 belongs to the cell type in the examination, and $x_i = 0$ if the sample belongs 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:

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

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

$$MSE = \frac{1}{N}(Y - X(X^T X)^{-1} X^T Y)^T (Y - X(X^T X)^{-1} X^T Y) \quad (2)$$

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 datasets 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 $\log_2$-transformed expression of the gene in the *i*-th 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 dataset, 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 were conducted using *Pvclust* package [29].

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, and also to consider variation among datasets, it is better to select features that co-occur in the reference and query data. Although we have already pre-calculated the cell type specificity score in the reference, we still 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 query 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 [30] 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 or trait-associated genes), we implemented Fisher's exact test to examine if they are significantly enriched in CTGenes. We allow users to define the threshold, *e.g.*, the top 5% genes ranked by *t*-statistics or *z*-scores as CTGenes. Specifically, for the query gene set and CTGenes in a given cell type, we built a dichotomous 2 × 2 contingency table as follows:

$$\begin{array}{c} \text{CTGenes } (r) \\ \hline \text{Query genes } (q) \end{array} \begin{pmatrix} |Q \cap R| & |\bar{Q} \cap R| \\ |Q \cap \bar{R}| & |\bar{Q} \cap \bar{R}| \end{pmatrix} \quad (3)$$

Here, q denotes the set of query genes and r denotes the set of CTGenes in a given cell type. $|Q \cap R|$ denotes the intersect gene number between Q and R, $|\bar{Q} \cap R|$ denotes the number of genes only in R, $|Q \cap \bar{R}|$ denotes the number of genes only in Q, and $|\bar{Q} \cap \bar{R}|$ represents the number of genes that neither in Q or R. In addition, the intersection ratio (IR) $|Q \cap R|/|Q|$ will also be calculated. An IR of 1 indicates that 100% of the query genes overlap 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 (v1.19.0) [31] for pathway enrichment analysis. Both Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations were used. Benjamini and Hochberg's approach [32] was used for multiple test correction. Significant pathways were defined as those with a false discovery rate (FDR) < 0.05. The networks of CTGenes enriched with KEGG pathways were presented by Cytoscape [33].

Data collection and pre-processing

scRNA-seq data

Two PBMC datasets were collected from 10X Genomics website (https://www.10xgenomics.com/resources/datasets). The test dataset of bronchoalveolar immune cells (BICs) was collected from healthy controls (HC, $n = 3$), as well as patients with moderate (M, $n = 3$) and severe (S, $n = 6$) COVID-19 infection [34,35]. The BIC data included 63,103 single cells. The raw data and predefined labels are available at https://github.com/zhangzlab/covid_balf. The fetal cardiac cell (FCC) scRNA-seq data were collected from 18 human embryos, ranging from 5 weeks (5 W) to 25 W of gestation [36]. The unique molecular identifier (UMI) count of 3842 cells and predefined labels were downloaded from the GEO website (GSE106118). The human liver scRNA-seq data were collected from five fresh hepatic tissues [37]. The raw dataset of 8444 cells and annotation are available through the GEO accession (GSE115469). The scRNA-seq data of human lung, spleen, and esophagus tissues were collected from 12 organ donors [38]. The raw data of 239,224 single cells and annotation are available at https://www.tissuestabilitycellatlas.org/.

Trait-associated gene lists from genome-wide association study summary statistics

Previous studies have suggested that most trait-associated genes (TAGs) show strong tissue-specific associations [23,39]. Therefore, we further collected 51 publicly available genome-wide association study (GWAS) datasets 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 being located in the non-coding regions, for each GWAS trait, we used moderately significant associated single nucleotide polymorphisms (SNPs) with chi-squared P value $< 1 \times 10^{-3}$. This strategy allowed us to have a list of genes with weak-to-strong association signals. Additionally, we applied our DeepFun model [40,41] to predict their potential regulatory effects. We defined variants with an absolute maximum SNP activity difference (SAD) score > 0.1 as regulatory loci (potential causal variants). Then we employed Pascal software [42] 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 and gene length information. Any gene with at least one regulatory locus was regarded as a TAG.

Bulk RNA-seq data

The test bulk RNA-seq data were generated from schizophrenia-associated human induced pluripotent stem cell (iPSC)-derived cell lines from the population isolate of the Central Valley of Costa Rica [43]. One clone from each subject was differentiated into neuronal precursor 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 detect gene richness. For a given number (n from 1 to 50) of cells belonging to the same cell type, we performed random sampling 100 times without replacement, and 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 [44] 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 s for a gene expression matrix with BIC dataset (63,103 cells) by *deCS.correlation* function. When applying Fisher's exact test, it took only ~ 50 s for a list of 43,514 genes across the 51 traits.

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 bulk RNA-seq data such as BlueprintEncode [45,46], DICE [24], and MonacoImmune [25]. Although these were bulk data, they were generated using cell lines and have already been used as reference datasets in

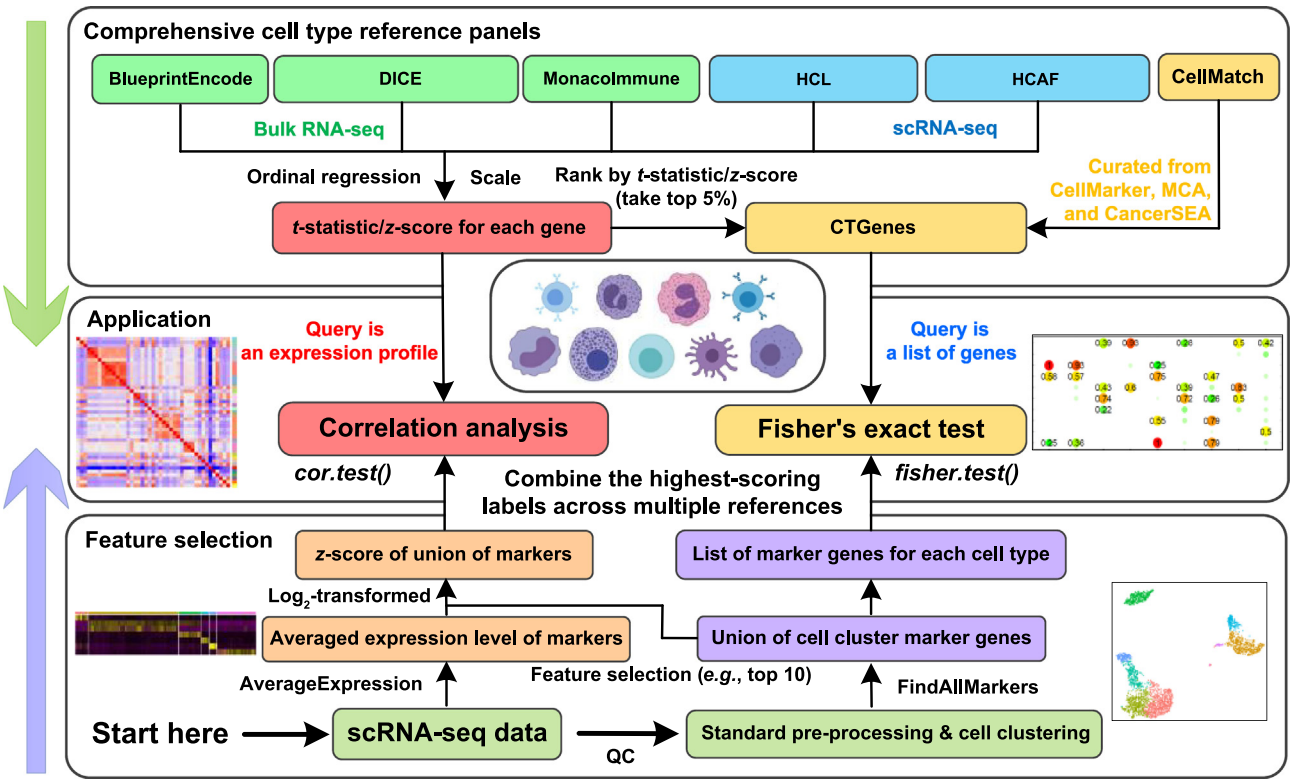


Figure 1 Overview of *deCS* flowchart

For each cell type, we compute *t*-statistic and *z*-score 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 CellMatch database. Depending on the type of "query data", when the query input is a gene expression profile, *deCS* calculates PCC or SCC between query scaled expression profiles and *t*-statistics (or *z*-scores) of each cell type in the reference, and then assigns the label with the highest score to the query profile. When the query input is a list of genes, *deCS* is analogous to existing tools for identifying candidate genes that are overrepresented in specific GO terms or KEGG pathways [31]. Finally, the top enriched cell type is annotated to query data. *deCS*, decoding Cell type Specificity; RNA-seq, RNA sequencing; scRNA-seq, single-cell RNA sequencing; CTGene, cell type-specific gene; DICE, database of immune cell expression; HCL, human cell landscape; HCAF, human cell atlas of fetal; PCC, Pearson correlation coefficient; SCC, Spearman's correlation coefficient; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; QC, quality control.

SingleR [7]. In addition, we downloaded the most comprehensive human single-cell expression data resource from the HCL [18] and HCAF [20] projects. For each dataset, we applied different analytical strategies (see Method for more details) to define CTGenes. Finally, we integrated the manually curated marker genes (used as CTGenes in *deCS*) from the CellMatch database [11]. Based on genes' *t*-statistics/*z*-scores and CTGenes, we implemented two statistical tests for cell type annotation (Figure 1). If the query data are a "gene expression profile" with unknown cell types, *deCS* applies the correlation analysis to determine which cell types in the reference are significantly similar to those in the query data. If the query data are a set of "candidate genes" of interest, *deCS* determines if they overlap significantly with a cell type in the reference dataset by using Fisher's exact test.

CTGenes among different reference datasets share strong consistency

To validate the reliability and consistency of CTGenes, we first calculated internal cell type correlation, and then performed UMAP dimension reduction analysis [44] 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% genes ranked by *t*-statistics) 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 and mast cells in blood] were missed in BlueprintEncode.

To further assess the validity of CTGenes identified by *t*-statistics, we performed KEGG 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

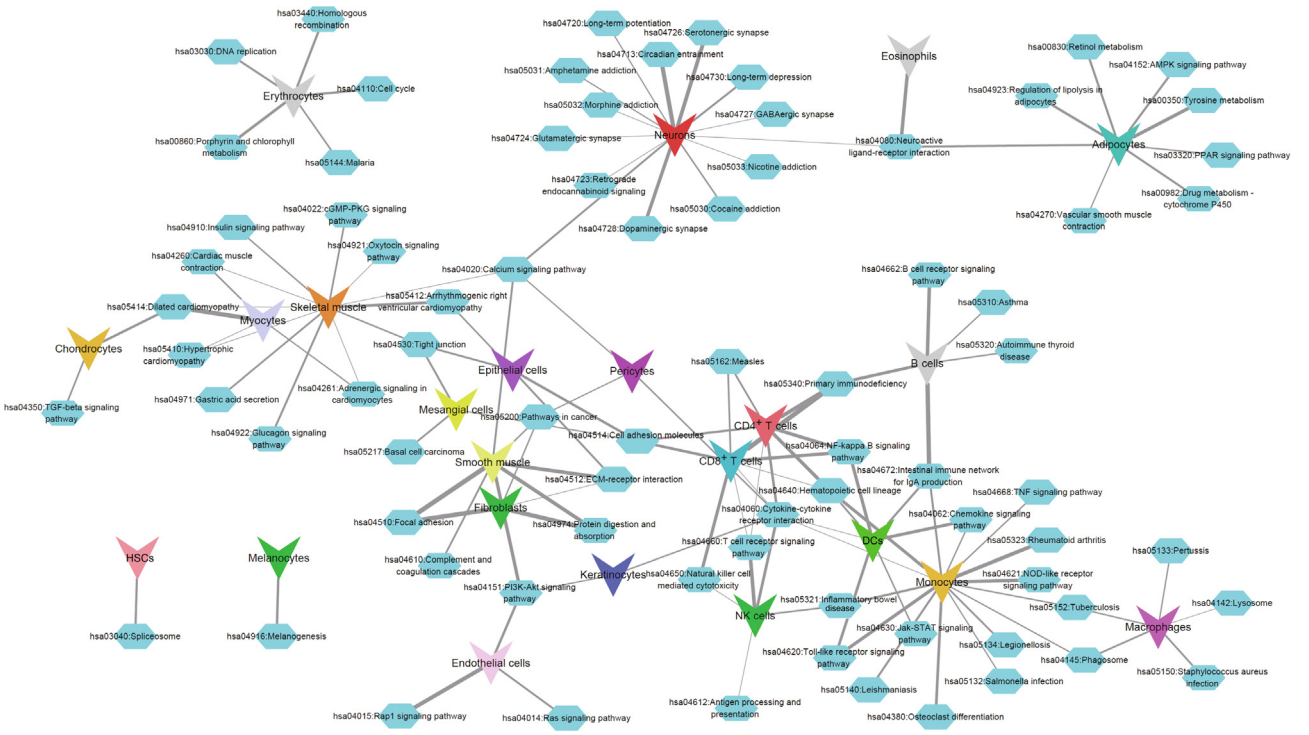


Figure 2 KEGG pathway enrichment of cell type-specific genes

The network shows the relationship between CTGenes and KEGG pathways with significant association (FDR < 0.05). Each arrow head node in different color represents a cell type, and each hexagon node in light blue represents a enriched pathway. Edge width between two nodes is proportional to $-\log_{10}$-transformed FDR from pathway enrichment analysis. The network layout is based on a force-directed graph. FDR, false discovery rate; DC, dendritic cell; NK, natural killer; HSC, hematopoietic stem cell.

were enriched in "B cell receptor signaling pathway"; $CD4^+$ and $CD8^+$ T cells were enriched in "T cell receptor signaling pathway"; hematopoietic stem cells (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 muscle; $CD4^+$ T cells, $CD8^+$ T cells, and natural killer (NK) cells; and monocytes, dendritic cells (DCs), and macrophages.

scRNA-seq cell type annotation by *deCS* demonstrates its advantage when combining multiple references

PBMCs

Human PBMCs are routinely studied in the biomedical fields including immunology and the development of diagnostics and therapeutics for human diseases [47]. To explore the utility of *deCS* for direct annotation of scRNA-seq data, we first analyzed a PBMC dataset, 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 [5]. We extracted the union of top 10 cell cluster-specific marker genes for each cluster, and then calculated the average expression profile followed by *z*-score normalization (Figure 1). We applied two statistical approaches, correlation analysis and Fisher's exact test, which would match the cell label with the highest PCC or 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 PBMC dataset. 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 (Th1) cells [48].

Solid tissues on adult and fetal samples

We next evaluated the performance of *deCS* in six additional scRNA-seq datasets from different solid tissues. We first considered the BIC dataset from COVID-19 infected patients [34]. 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 4**A). Two exceptions were mast cells and plasmacytoid dendritic cells (pDCs), which were matched to erythrocytes (PCC = 0.5, IR = 0.25) and neurons (PCC = 0.27) or B cells (IR = 0.3), respectively. 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) [49]. Alternatively, users can map their query data to other available references to identify novel cell types. As shown in Figure S4, mast cells and pDCs were mapped to their counterpart when

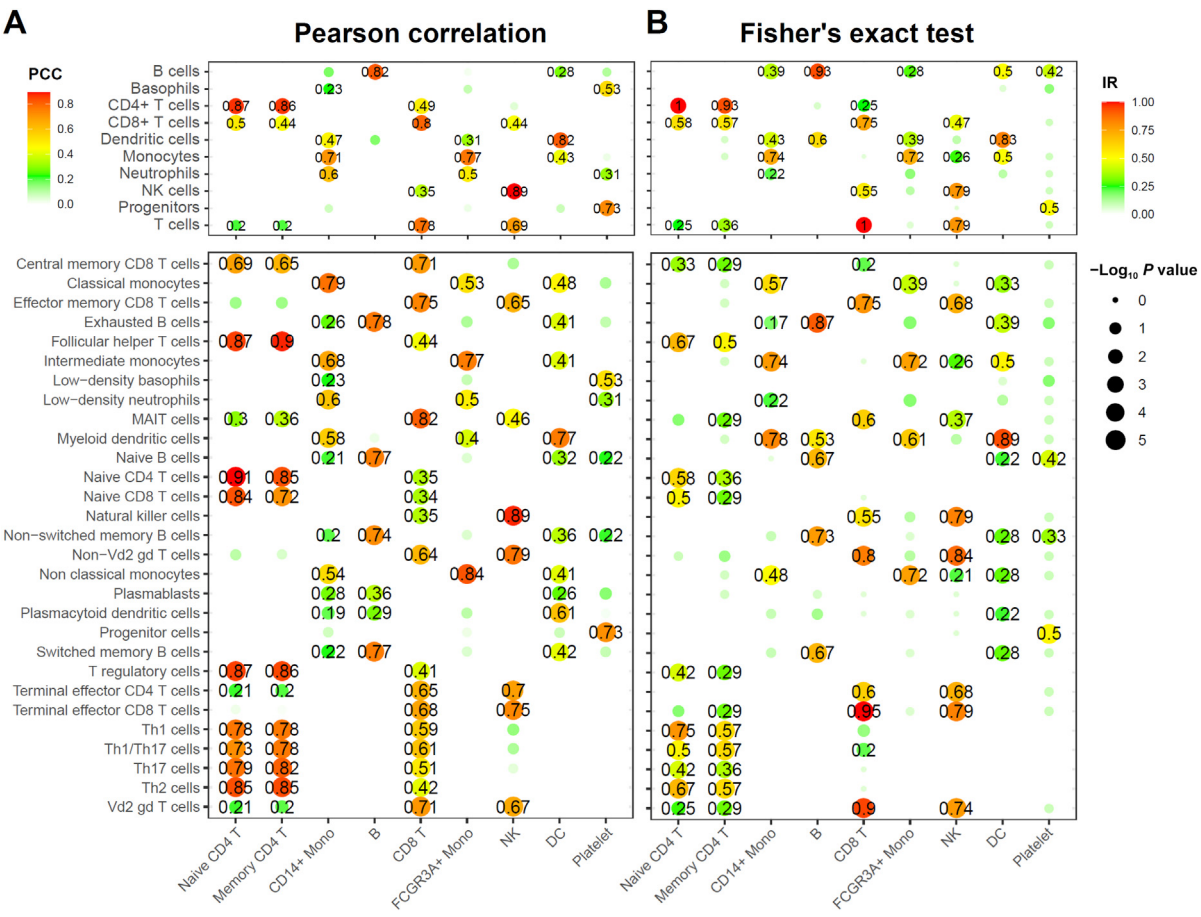


Figure 3 Comparison of the two methods in *deCS* using PBMC scRNA-seq data

A. Correlation analysis by PCC. **B.** Fisher's exact test. X-axis shows the 9 cell type clusters (query). Y-axis shows the 10 main (up panel) and 29 fine (bottom panel) immune cell types from the MonacoImmune reference [25]. The colors represent the PCC or IR, and the sizes of circles represent the $\log_{10}$-transformed P values. Non-significant associations ($P > 0.05$) were labeled by white color. Names of cell type clusters on X-axis and immune cell types on Y-axis are derived from original references. IR, intersection ratio.

using the HCL (PCC = 0.82, IR = 0.7) and MonacoImmune (PCC = 0.74, IR = 0.9) references, respectively.

The second dataset was derived from FCC [36]. 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 demonstrates that integrating results from multiple high-quality curated reference panels could effectively identify novel cell types that cannot be detected by other methods, thereby significantly improving annotation accuracy and utility.

Annotation at single-cell resolution

In the aforementioned analysis, we used the averaged expression level in each cluster to evaluate the performance of *deCS* [50]. These traditional routines, however, overlook an important characteristic of scRNA-seq data [51]: cellular heterogeneity. Considering the true hierarchical cluster structure for a cell population (*e.g.*, both $CD4^+$ and $CD8^+$ are T cells) [51], cellular heterogeneity is widely observed when applying unsupervised clustering [52]. The ideal scenario is that each cell can be annotated to one specific branch of the hierarchical tree [8]. Therefore, we further evaluated *deCS* in BIC dataset [34]. As shown in Figure 4B, 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, 1514 and 2076 cells in macrophage cluster were erroneously annotated as monocytes and neutrophils, respectively. Some neutrophils were annotated as monocytes or macrophages (Figures S6 and 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-depth (1000 ≤ UMI count < 3000), median-depth (3000 ≤ UMI count < 6000), and high-depth (UMI count ≥ 6000) groups. As Figure 5A showed, although

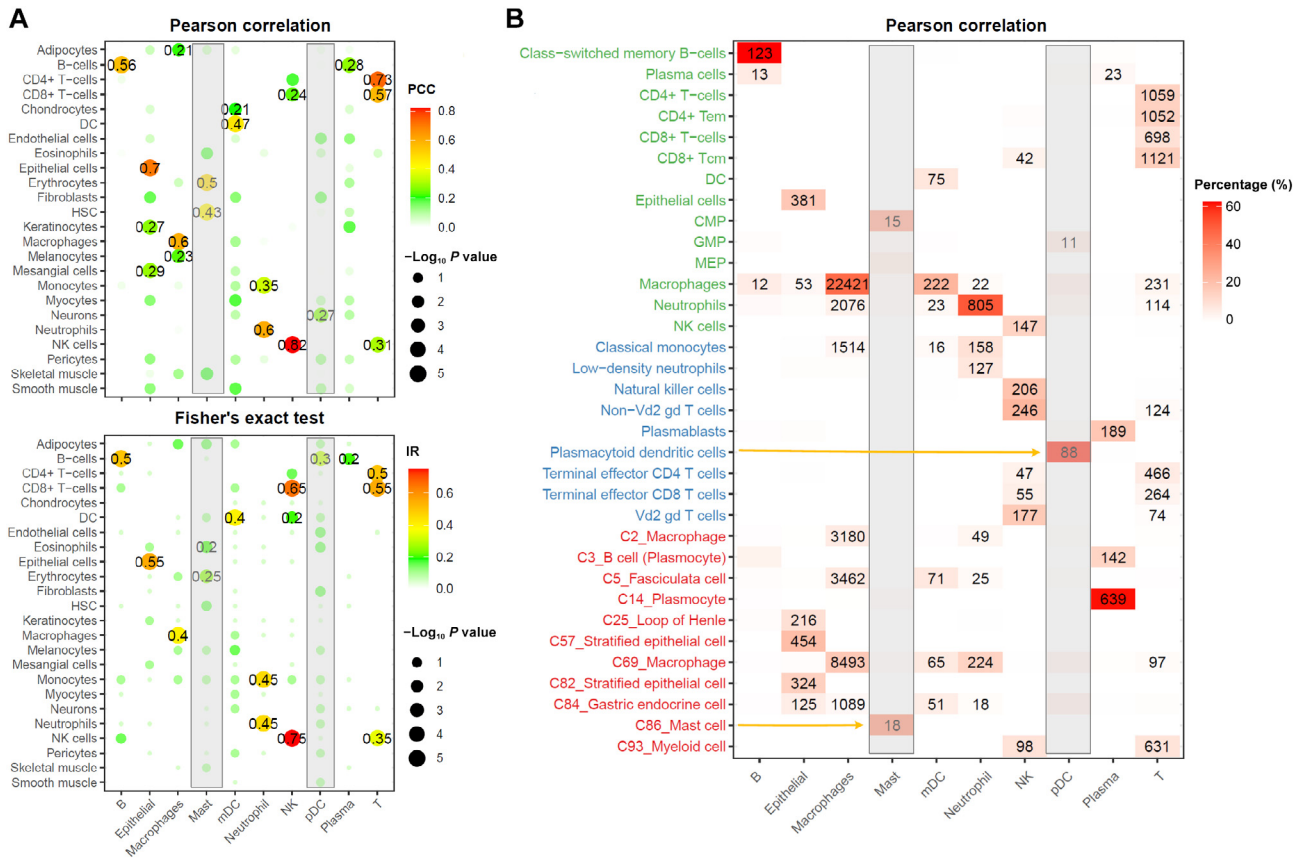


Figure 4 ***deCS* annotation of BICs from COVID-19 infected patients**

A. Correlation analysis by PCC and Fisher's exact test between 10 major cell type clusters (query) and 24 main cell types from the BlueprintEncode reference. The colors represent the PCC or IR, and the sizes of circles represent the $\log_{10}$-transformed *P* values. **B.** 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 represent the percentages and the total cell counts that were annotated to a given cell type. Names of cell type clusters on X-axis and cell types on Y-axis are derived from original references. BIC, bronchoalveolar immune cell; pDC, plasmacytoid dendritic cell; mDC, myeloid dendritic cell; CMP, common myeloid progenitor; MEP, megakaryocyte-erythrocyte progenitor; GMP, granulocyte-monocyte progenitor.

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 [53].

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, *CCR7*, the top marker gene of naive $CD4^+$ T cell, was expressed in approximately 40% of the naive $CD4^+$ T cells due to the dropout effect in scRNA-seq. The annotation accuracy was decreased when cell type inference was conducted at single-cell level, while utilizing gene expression imputation approaches [21,54,55] could improve annotation performance at single-cell resolution. Thus, we recommend cell type annotation on clusters instead of individual cell after imputation, which is consistent with other publications [5,56].

Benchmarking analysis on different tissues

To demonstrate the superior performance of *deCS*, we compared it with other six methods, including SingleR [7], CHETAH [8], *scPred* [12], SciBet [13], Cell BLAST [15], and scSorter [57] using the aforementioned eight datasets. We first applied the methods to two PBMC datasets. As summarized in **Table 1**, the overall performances of *deCS*, SingleR, *scPred*, and SciBet ranged from 87.55% to 88.78%. 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

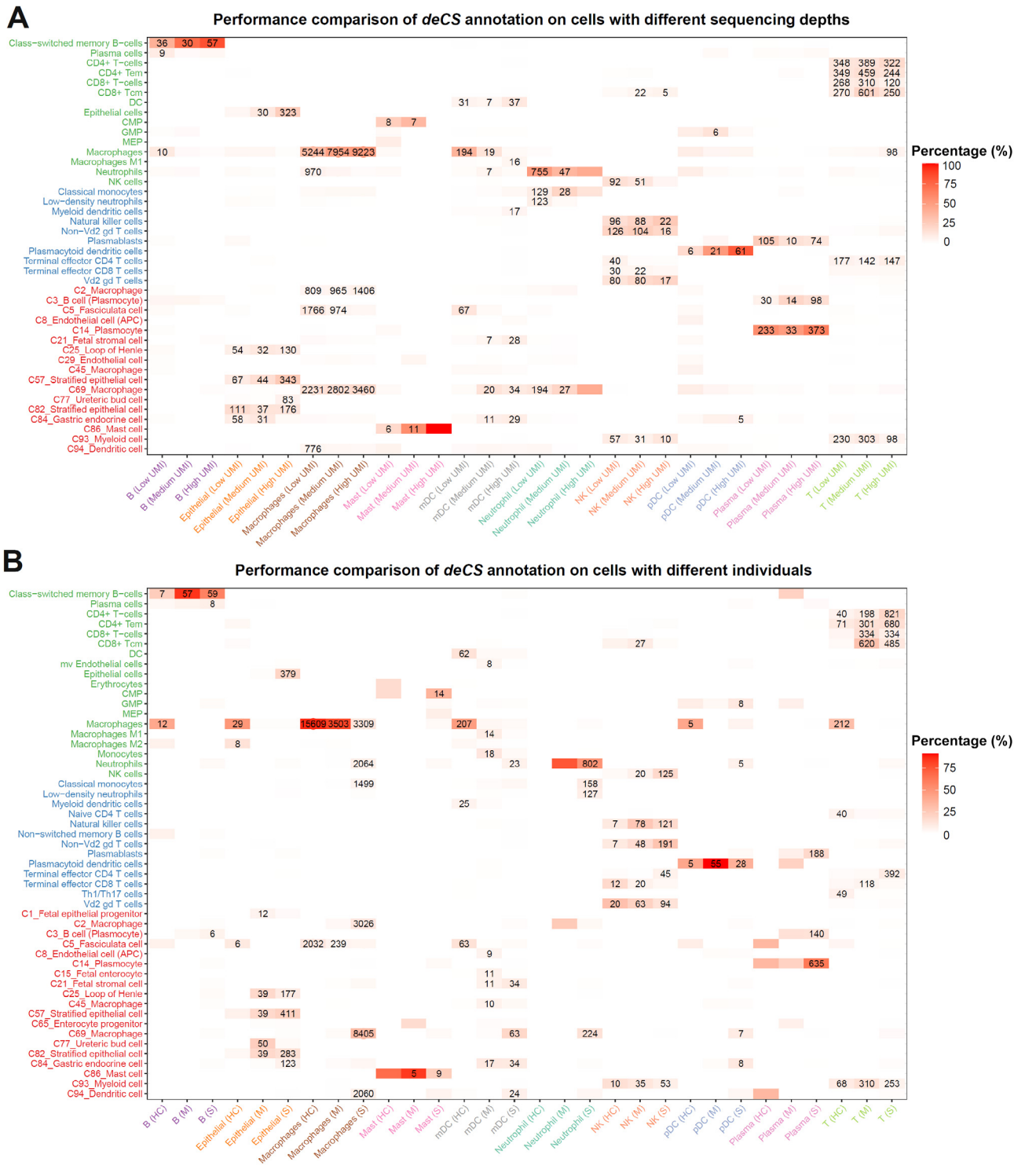


Figure 5 Comparison of *deCS* annotation on cells with different sequencing depth and individuals

A. Annotation comparison of cells sequenced among low (1000 ≤ UMI count < 3000), median (3000 ≤ UMI count < 6000), and high (UMI count ≥ 6000) depth groups. **B.** Annotation comparison of cells derived from HC, M, and S groups of COVID-19 infected patients. 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 annotated with at least 5% cells from 10 predefined cell clusters were labeled with cell counts. Names of cell type clusters on X-axis and cell types on Y-axis are derived from original references. UMI, unique molecular identifier; HC, healthy control; M, moderate; S, severe.

Table 1 Performance comparison of *deCS* with six other models among eight datasets

ID	Dataset	*deCS* (PCC)	*deCS* (Fisher's exact test)	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%	–

Note: PBMC datasets were benchmarked at the single-cell level, since the cell-cluster level information needed for the *deCS* (Fisher's exact test) model is not available from these PBMC datasets. Considering that scSorter is a "semi-supervised" cell type assignment method and the prior knowledge of markers would strongly affect the annotations, we did not benchmark scSorter on datasets from six solid tissues. "–" means not applicable. PCC, Pearson correlation coefficient; PBMC, peripheral blood mononuclear cell.

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 intermediate 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 datasets, 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 those from PBMC, *scPred* showed 79.6% accuracy on spleen tissue. However, on liver tissue, where there was a large proportion of hepatocytes (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 reduces the feature collinearity and impact of sequencing depth difference

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, and alveolar cells in lung), *deCS* showed better annotation accuracy in distinguishing biologically related cell types (*e.g.*, $CD4^+$ T cells, $CD8^+$ T cells, and NK cells; monocytes and macrophages). We speculated that feature selection (comparing to SingleR) is still an essential step before correlation analysis. The identified anchor genes (cluster marker genes) can effectively reduce the collinearity compared to the case when using more genes (top variable genes). As shown in **Figure 6**A (upper), 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 (PCC = 0.32) when using top 2000 variable genes (Figure 6A, lower). Although there are lots of other methods for feature selection and model prediction (*e.g.*, SVM) [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 both speed and scalability when combining the highest-scoring labels across multiple references, we decided to provide the cell cluster aggregated *t*-statistics and *z*-scores. On the other hand, *deCS* drastically reduced the required memory and runtime, which is helpful for annotation of large dataset 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 [7,58]. Thus, we evaluated the level of sequencing depth (different sized pools, *n* from 1 to 50) of single-cell transcriptomes to the pseudo-bulk level (cluster averaged). As shown in Figure 6B, we observed a gradually diminishing improvement of inter-correlation with increasing sequencing depth. For example, the inter-correlation between single-cell and pseudo-bulk levels for T cells in PBMC and BIC datasets was 0.707 and 0.898, respectively. As the median number of detected genes in BIC dataset was almost 4-fold more than that in PBMC dataset, it only needed to cumulate 5 cells to approximate 95% bulk level in BIC dataset, while it needed at least cumulated 10 cells in PBMC dataset (Figure 6B). To put it another way, when a gene is observed at a low or moderate expression level in BICs, it is probably not detected in PBMCs. This phenomenon is also known as "dropout" [59]. For better demonstration, we randomly selected 5 bulk RNA-seq samples and down-sampled each sample to 10,000–100,000 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 principal component analysis (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 6C, we observed a strong batch effect between PBMC and BIC natural-log-normalized expression matrices (*e.g.*, NK and T cells were clustered by two samples). In contrast, after *z*-score normalization (Figure 6D), most cell types form dis-

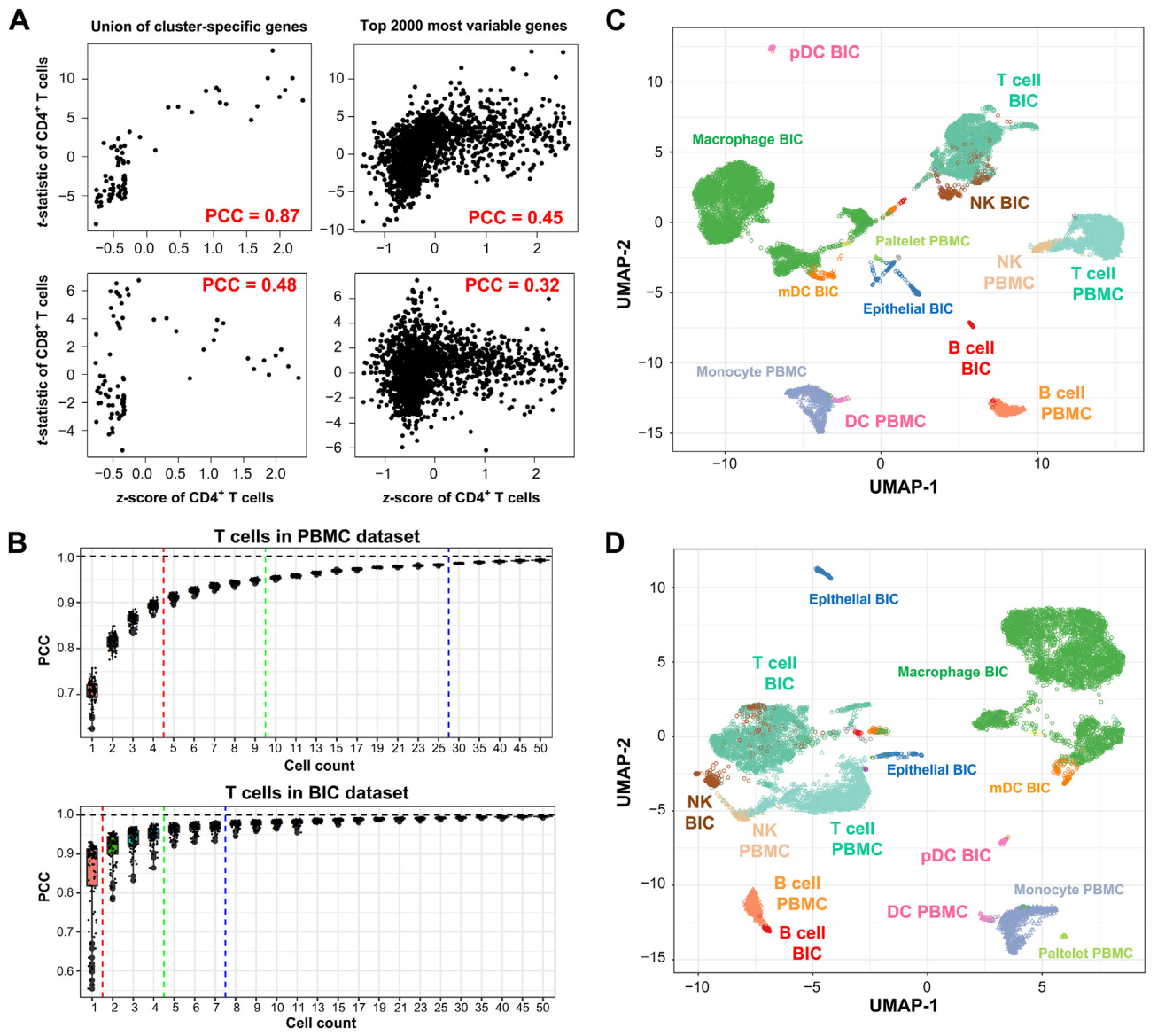


Figure 6 ***deCS* pipeline reduces the collinearity and impact of sequencing depth difference**

A. Correlation comparison by using the union of cell cluster-specific genes and top 2000 most variable genes between query ($CD4^+$ T) cells and reference on $CD4^+$ T (positive; upper panels) or $CD8^+$ T cells (negative; lower panels). **B.** Rarefaction curve of T cells in PBMC and BIC datasets. X-axis indicates the sized pools (*n* from 1 to 50). Y-axis indicates the inter-correlation between the gene expression of pooled cells and pseudo-bulk level (cluster averaged). The red, green, and blue lines indicate the number of single cells that hypothetically enables the averaged expression levels of pooled cells approximate to bulk populations with PCC > 0.90, PCC > 0.95, and PCC > 0.975, respectively. UMAP integrative analysis of PBMC and BIC datasets by natural-log-normalized expression matrix (**C**) or further *z*-score scaled (**D**). Each circle or triangle represents a single cell derived from BIC or PBMC dataset. PBMC, peripheral blood mononuclear cell; UMAP, uniform manifold approximation and projection.

tinct clusters were separated by cell types (*e.g.*, B cells, NK cells, T cells, and DCs).

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 GWAS data for identifying trait–cell type associations, and also to bulk RNA-seq data from 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 [60], yet tissues are complex milieus consisting of various cell types [61]. Therefore, tissue-level association may fail to elucidate cell type contributions in diseases [62]. To uncover novel associations between cell types and diseases, *deCS* was applied to 51 GWAS data (Table S5) and the associations were 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 that 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 [63], but also can serve as an important instrumental role in cell type-specific transcriptome-wide association studies [64] and colocalization test [65], 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 [66]. In most iPSC studies, differentiation quality is routinely assessed [66]. After we normalized the bulk RNA-seq data of iPSC-derived NPCs and neurons [67], the top matched cell types in 12 of 13 iPSC-NPCs were most enriched in "fetal progenitor cells" or "human embryonic stem cells (hESCs)" (Figure S12), while the top matched cell types for iPSC-neurons were most enriched in "fetal neurons". Therefore, *deCS* can also be applied for decoding the cell type specificity on bulk RNA-seq data.

Discussion

Single-cell sequencing is a fast-evolving technique that provides deep insight into the complexity of cellular heterogeneity within the same tissue. One critical step in the scRNA-seq data analysis is cell type annotation. However, this step is largely done manually, which is time-consuming and subjective. A list of methods has been recently developed for automatic cell type annotation [16]. Based on our benchmark results, we demonstrate that most classifiers work well when users provide a reference dataset that is conserved with query data. However, due to the small amount of starting tissue material, scRNA-seq data tend to have batch effects [68]. In most scenarios, researchers do not have enough prior knowledge in the cell type composition of the investigated samples. Thus, finding a suitable reference is a time-consuming and subjective task. If some novel cell types were not labeled in the reference, it would be difficult to control the sensitivity/specificity of most

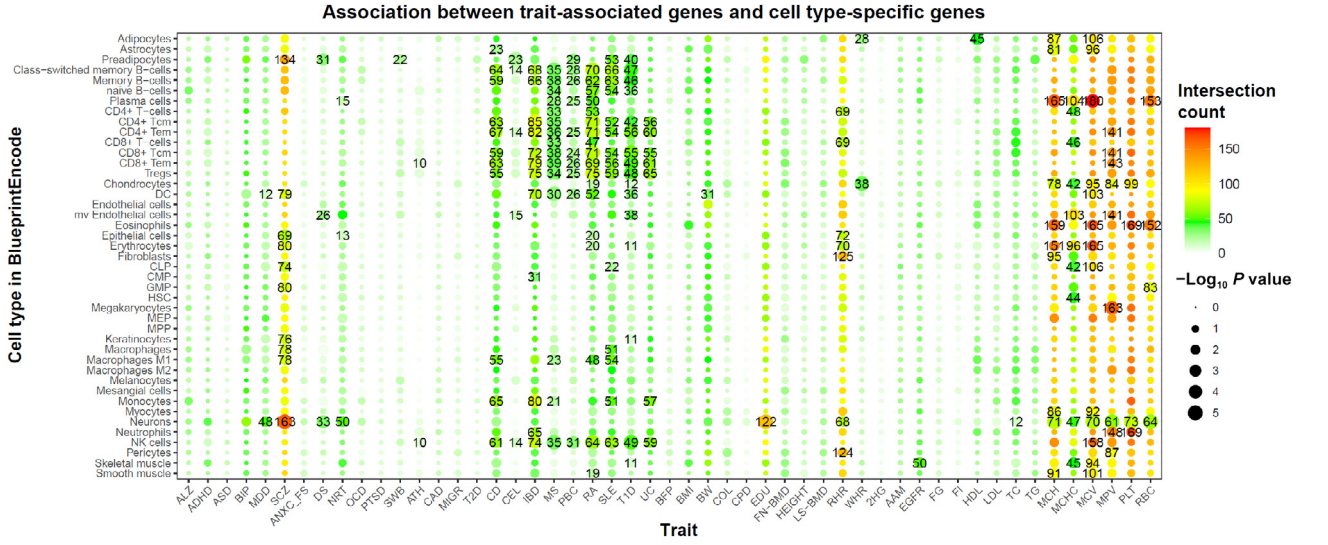


Figure 7 Association between trait-associated genes and cell type-specific genes

Fifty-one traits were analyzed. Heatmap shows the significant trait–tissue associations ($P < 0.05$). 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 values. 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 TAG number. Names of cell types on Y-axis are derived from original references. SNP, single nucleotide polymorphisin; TAG, trait-associated gene; CLP, common lymphoid progenitor; MPP, multipotent progenitor cell; ALZ, Alzheimer's disease; ADHD, attention deficit-hyperactivity disorder; ASD, autism spectrum disorder; BIP, bipolar disorder; MDD, major depressive disorder; SCZ, schizophrenia; ANXC_FS, anxiety angst; DS, depressive symptoms; NRT, neuroticism; OCD, obsessive-compulsive disorder; PTSD, post-traumatic stress disorder; SWB, subjective well-being; ATH, asthma; CAD, coronary artery disease; MIGR, migraine; T2D, type 2 diabetes; CD, Crohn's disease; CEL, celiac disease; IBD, inflammatory bowel disease; MS, multiple sclerosis; PBC, primary biliary cholangitis; RA, rheumatoid arthritis; SLE, systemic lupus erythematosus; T1D, type 1 diabetes; UC, ulcerative colitis; BFP, body fat percentage; BMI, body mass index; BW, birthweight; COL, college; CPD, cigarettes per day; EDU, education; FN-BMD, bone mineral density (femoral neck); HEIGHT, height; LS-BMD, bone mineral density (lumbar spine); RHR, resting heart rate; WHR, waist–hip ratio; 2HG, two-hour glucose; AAM, age at menarche; EGFR, glomerular filtration rate; FG, fasting glucose; FI, fasting insulin; HDL, high-density lipoprotein; LDL, low-density lipoprotein; TC, total cholesterol; TG, triglyceride; MCH, mean cell hemoglobin; MCHC, mean cell hemoglobin concentration; MCV, mean red blood cell volume; MPV, mean platelet volume; PLT, platelet count; RBC, red blood cell count.

classifiers; in that case, it is good to define them as unknown cell types, rather than annotate them as a wrong cell type. For example, in Figure 4A, if PCC threshold was set 0.25, pDCs would be wrongly annotated as neurons. When PCC threshold was set 0.65, both macrophages and neutrophils would be annotated as unknown cell types. To overcome these limitations, we collected multiple human cell type expression profiles from different platforms. As *deCS* only relies on the pre-calculated cell type specificity score and the correlation analysis without further advanced dimension reduction, it can efficiently conduct cell type annotation on each reference, and then combine the highest-scoring labels across multiple references. When the cell composition between query and reference datasets is not well conserved (*e.g.*, in the cases of bronchoalveolar, heart, liver, and lung in this study), *deCS* achieves 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 [69], 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. In addition, feature selection is also an essential step to obtain the accurate results. In this regard, *deCS* has several advantages over other models. First, *deCS* selects features that co-occur in the reference and query data (discarding useless features in query). Therefore, it drastically reduces the collinearity, the required memory, and runtime for cell annotation, while it achieves comparable accuracy on biologically related cell types (*e.g.*, $CD4^+$ T cells, $CD8^+$ T cells, and NK cells; monocytes and macrophages). Second, with the expanded references, *deCS* exhibits superior performance to other methods. It is worth noting that even for the same cell type, their expression profiles demonstrate strong heterogeneity among different tissues or different individuals. Thus, it will not work well by simply relying on a single reference. *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 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 [70]. 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 are associated with multiple cell types. For example, asthma was found to be associated with both immune and epithelial cells [70]. Moreover, comparing to previous studies [23] that most TAGs enriched in "blood" tissue, *deCS* showed that immune TAGs were enriched in lymphoid lineage cells, while red blood cell-related traits were enriched in erythroid cells. It has been frequently recognized that pleiotropic effects are ubiquitous in human complex traits [71]. For example, individuals carrying schizophrenia risk alleles tend to be also associated with high risk of Crohn's Disease [72]. We expect to identify more explicit associations with the increasing high-quality curation of transcriptome profiles among different 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 have 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 [73]. We believe some "semi-supervised" cell type assignment methods like scSorter [57] can address this problem by providing the marker gene in the 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 [74], the available large numbers of single-cell datasets can be fed in as reference datasets to improve the annotation of future experiments. Despite these limitations, we demonstrate 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 [23,60], 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 complex traits and diseases.

Code availability

The source codes and results are implemented in an R package, and are freely available at GitHub (https://github.com/bsml320/deCS); these data have also been submitted to BioCode at the National Genomics Data Center, Beijing Institute of Gemonics, Chinese Academy of Sciences / China National Center for Bioinformation (BioCode: BT007286), and are publicly accessible at https://ngdc.cncb.ac.cn/biocode/tools/BT007286. We also developed a shiny application as part of the *deCS* package, and it is available at https://gpei.shinyapps.io/decs_cor/ and https://gpei.shinyapps.io/decs_fisher/.

Competing interests

The authors have declared no competing interests.

CRediT authorship contribution statement

Guangsheng Pei: Conceptualization, Methodology, Software, Resources, Data curation, Visualization, Writing – original

draft. **Fangfang Yan:** Data curation, Visualization, 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 have read and approved the final manuscript.

Acknowledgments

We thank all the members the Bioinformatics and Systems Medicine Laboratory (BSML). We also thank the investigators who generated the reference data. Zhongming Zhao was partially supported by National Institutes of Health grants (Grant Nos. R01LM012806, R01DE030122, and R01DE029818). We thank the resource support from Cancer Prevention and Research Institute of Texas (Grant Nos. CPRIT RP180734 and RP210045), United States.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2022.04.001.

ORCID

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

References

[1] Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, 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] 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.
[25] 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.
[26] Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, et al. Cell Marker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res 2019;47:D721–8.
[27] 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.
[28] 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:D900–8.
[29] Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 2006;22:1540–2.
[30] 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.

[31] Fresno C, Fernandez EA. RDAVIDWebService: a versatile R interface to DAVID. Bioinformatics 2013;29:2810–1.
[32] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 1995;57:289–300.
[33] 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.
[34] 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.
[35] 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.
[36] 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.
[37] 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.
[38] 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.
[39] 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:D1022–30.
[40] 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.
[41] 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:W131–9.
[42] 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.
[43] 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 2021;46:673–82.
[44] McInnes L, Healy J, Saul N, Grossberger L. UMAP: uniform manifold approximation and projection. J Open Source Softw 2018;3:861.
[45] Martens JHA, Stunnenberg HG. BLUEPRINT: mapping human blood cell epigenomes. Haematologica 2013;98:1487–9.
[46] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74.
[47] 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 Immunol 2019;20:5.
[48] Luckheeram RV, Zhou R, Verma AD, Xia B. $CD4^+$ T cells: differentiation and functions. Clin Dev Immunol 2012;2012925135.
[49] Kiselev VY, Yiu A, Hemberg M. scmap: projection of single-cell RNA-seq data across data sets. Nat Methods 2018;15:359–62.
[50] Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet 2019;20:273–82.
[51] Wu Z, Wu H. Accounting for cell type hierarchy in evaluating single cell RNA-seq clustering. Genome Biol 2020;21:123.
[52] 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.
[53] Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc 1952;47:583–621.
[54] Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun 2018;9:997.
[55] 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.
[56] Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 2018;19:15.
[57] Guo H, Li J. scSorter: assigning cells to known cell types according to marker genes. Genome Biol 2021;22:69.
[58] 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.
[59] Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods 2014;11:740–2.
[60] 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.
[61] Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol 2017;18:220.
[62] 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.
[63] 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.
[64] 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.
[65] 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.
[66] Hoffman GE, Hartley BJ, Flaherty E, Ladran I, Gochman P, Ruderfer DM, et al. Transcriptional signatures of schizophrenia in hiPSC-derived NPCs and neurons are concordant with post-mortem adult brains. Nat Commun 2017;8:2225.
[67] 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.
[68] 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.
[69] 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.
[70] 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:D862–70.
[71] 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 2019;20:79.
[72] Pickrell JK, Berisa T, Liu JZ, Segurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. Nat Genet 2016;48:709–17.
[73] 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.
[74] 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.