The DeMixSC deconvolution framework uses single-cell sequencing plus a small benchmark dataset for improved analysis of cell-type ratios in complex tissue samples

Shuai Guo[1,11], Xiaoqian Liu[1,11], Xuesen Cheng[2,11], Yujie Jiang[1, 3], Shuangxi Ji[1], Qingnan Liang[2], Andrew Koval[1, 3], Yumei Li[2], Leah A. Owen[4,5,6], Ivana K. Kim[7], Ana Aparicio[8], John N. Weinstein[1], Scott Kopetz[9], John Paul Shen[9], Margaret M. DeAngelis[4,5,6,10], Rui Chen[2,12], Wenyi Wang[1,12*]

1. Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.
2. Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA.
3. Department of Statistics, Rice University, Houston, TX, USA.
4. Department of Ophthalmology, Jacobs School of Medicine and Biomedical Engineering, SUNY University at Buffalo, Buffalo, NY, USA.
5. Department of Population Health Sciences, University of Utah School of Medicine, Salt Lake City, UT, USA.
6. Department of Ophthalmology and Visual Sciences, University of Utah School of Medicine, Salt Lake City, UT, USA.
7. USA Retina Service, Harvard Medical School, Massachusetts Eye and Ear, Boston, MA, USA.
8. Department of Genitourinary Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.
9. Department of Gastrointestinal Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.
10. VA Western New York Healthcare System, Buffalo, NY, USA.
11. Authors contributed equally.
12. Authors contributed equally.

* Correspondence:
wwang7@mdanderson.org. Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center. 7007 Bertner Ave., Houston, TX 77030.

Abstract

We introduce a novel deconvolution framework, DeMixSC, to resolve technological discrepancies between bulk and single-cell/nucleus RNA-seq data, a critical issue unaddressed by existing single-cell-based deconvolution methods. Built upon the weighted non-negative least squares framework, DeMixSC introduces two key improvements: it leverages a small benchmark dataset to identify and rescale genes affected by technological discrepancies; it employs a novel weight function to account for variations across subjects and cells. The advanced utility of DeMixSC is demonstrated by its superior deconvolution accuracy on a benchmark dataset of healthy retinas and its broad applicability to a large aged-macular degeneration (AMD) cohort. Our work is the first to systematically evaluate the impact of technological discrepancies on deconvolution performance and underscores the importance of using a benchmark dataset to counteract these discrepancies. Our study positions DeMixSC as a transferable tool for accurate deconvolution of large bulk RNA-seq cohorts, necessitating only a tissue-type match between the benchmark and targeted datasets.

Keyword:

Transcriptomic deconvolution, technological discrepancy, single-cell/nucleus RNA sequencing, bulk RNA sequencing, age-related macular degeneration.

Introduction

While recent advances in single-cell/nucleus RNA sequencing (sc/snRNA-seq) offer valuable insights into cell types and states in healthy[1] and diseased tissues[2,3], the high expenses and complex sample preparation procedures have restricted its widespread adoption in clinical settings[4]. Bulk RNA-seq, on the other hand, retains its essential role, especially in large disease-oriented cohort studies, where its cost-efficiency, streamlined sample processing, and high-throughput analytic capabilities establish it as the method of choice for both preliminary screenings and exhaustive population-level analyses[5–7]. Nevertheless, bulk RNA-seq comes with a significant drawback: it captures averaged gene expression across heterogeneous cell types, thus confounding downstream analysis[4]. To mitigate this drawback, deconvolution methods have been developed to delineate the cell-type-specific signals from bulk RNA-seq data. Traditional bulk-based deconvolution methods[8,9] employ bulk RNA-seq data from normal tissues or cell lines as the reference. They are typically constrained by low-resolution estimates, limited to identifying only two or three cellular components within the bulk samples. The progress of sc/snRNA-seq techniques opens the door to the emergence of single-cell-based deconvolution methods[10–18], which tap into the granularity of even a modest set of single-cell data to provide far superior resolution in estimating cell-type proportions in complex tissues, thereby offering a cost-effective alternative.

Single-cell-based deconvolution methods are not without their disadvantages, however. Though affording remarkable resolution, they encounter a substantial challenge in achieving precise accuracy. This limitation arises from inconsistencies in gene expression profiles between bulk and sc/snRNA-seq data, which are attributable to technique variations in sample acquisition, preparation, and sequencing[19,20]. Such inconsistencies, which we refer to as "technological discrepancies", have caused prior deconvolution studies to produce suboptimal estimates of cell-type proportions, particularly when unpaired sc/snRNA-seq data serve as the reference matrix for deconvolving publicly available large bulk cohorts[16,21,22]. Several studies have been aware of and attempted to address these issues but ended up with limited success[10,14,18]. CIBERSORTx[18] simply implements a batch effect correction step but offers limited improvements in deconvolving complex bulk tissues. The ensemble approach of SCDC[14] uses matched bulk and scRNA-seq data from two normal mice breast tissues, which lacks generalizability to patient cohorts. The most recent SQUID[10] builds on top of DWLS[15] with a Bisque-based linear transformation step[23] to align matched bulk and scRNA-seq data, which can distort gene expression profiles, risking overcorrection. Meanwhile, existing benchmarking designs[10,21,22,24] often employ unfavorable datasets such as simulated pseudo-bulk data, cell line mixtures, or publicly available data, none of which are tailored to reveal the negative effect of technological discrepancies. Therefore, there is an unmet demand for a comprehensive benchmark study that rigorously illustrates and assesses the impact of technological discrepancies and a robust deconvolution framework that effectively mitigates these discrepancies to enhance analytical accuracy.

In this study, we thoroughly explore technological discrepancies and address their impact on deconvolution performance to make adjustments correspondingly. To accomplish this, we generate a specialized benchmark dataset of 24 healthy retinal samples, ensuring technological discrepancies as the main confounding factor. Using this dataset, we demonstrate that technological discrepancies significantly affect the expression profiles of bulk and single-nucleus data and thus reduce the accuracy of existing single-cell-based deconvolution methods. Against this backdrop, we introduce a novel deconvolution method called DeMixSC, which employs a benchmark dataset and an improved weighted nonnegative least-squares (wNNLS) framework[25] to identify and adjust for genes consistently affected by technological discrepancies. DeMixSC is generalizable to any major tissue types, by necessitating a small representative benchmark dataset to effectively deconvolve a large tissue-type-matched bulk cohort. We validate the improved deconvolution performance of DeMixSC by comparing it to eight existing methods on our benchmark dataset. When applied to 453 peripheral retinal samples from patients with age-related macular degeneration (AMD)[7], DeMixSC achieves more realistic cell-type estimates which reflect subtle changes in cell-type proportions between AMD conditions, confirming its reliability and generalizability in real-world settings. DeMixSC fills the gap in resolving technological discrepancies in bulk deconvolution and serves as an accurate and adaptable instrument for estimating cell-type proportions.

Results

Use benchmark data to assess technological discrepancy

We design a specialized benchmark dataset to assess the technological discrepancy between bulk and sc/sn sequencing platforms (Fig. 1A). This dataset comprises 24 healthy retinal samples from donors' eyes collected within six hours postmortem (ages of death between 53-91, Supplementary Table 1), from two batches of sequencing experiments. Both bulk and snRNA-seq profiling are performed on each sample from the same single-nucleus suspension aliquot using a template-switching method to generate full-length cDNA libraries (see Methods). As single-cell protocols can be biased toward retaining certain cell types[26], hence changing the cell-type proportions, this special approach maximizes our chance that the matched sequencing data share approximately the same cell-type proportions. The corresponding snRNA-seq data is summed to create matched pseudo-bulk RNA-seq data. We hypothesize that any major differences in the gene expression profiles between the matched pseudo-bulk and real bulk RNA-seq would be technological as compared to biological discrepancies.

We indeed observe much larger batch differences between real-bulk and pseudo-bulk data, in contrast to small differences observed in cell-type distributions across samples in snRNA-seq or differences between the two experimental batches (Extended Data Figs. 1A-E). Total read counts from bulk RNA-seq data are significantly lower than total UMI counts from matched pseudo-bulk data (Extended Data Fig. 1F). Batch-1 exhibits mean read depths of $3.7\text{x}10^7$ and $7.9\text{x}10^7$ for bulk and pseudo-bulk RNA-seq data, respectively (t-test P-value < 0.001), while the mean read depths in batch-2 are $4.6\text{x}10^6$ and $4.7\text{x}10^7$, respectively (t-test P-value < 0.001). Assuming the difference in read depth does not impact the relative expression for each gene, we expect gene expression correlation to be a better metric for identifying technological discrepancy. We observe a low-to-moderate correlation of gene expression between the paired bulk data, which is consistent across samples (Fig. 1B, mean Pearson correlation coefficient = 0.12 for batch-1, and 0.47 for batch-2). Further differential expression (DE) analysis between the two bulks identifies more than 5,000 DE genes in each experimental batch (Fig. 1C, adjusted P-values < 0.05), with more than 60% of these genes overlapping across the experiments (Fig. 1D). Our observations suggest a consistent technological effect across experiments. In broader contexts, factors such as library preparation, RNA capture efficiency, reverse transcription protocol, and sequencing depth could serve as potential sources of the technological discrepancy[19,20,27]. Thus, we expect the reference matrix derived from sc/snRNA-seq data does not fully represent cell-type-specific expression profiles in bulk samples[10,22]. Given these discrepancies, performances of existing deconvolution methods will be compromised, as their key assumption about the representative reference is violated.

Overview of DeMixSC

Here, we present our novel deconvolution framework, DeMixSC, and illustrate how it addresses the observed consistent technological discrepancy in order to enhance the estimation accuracy of cell-type proportions. The DeMixSC framework, as depicted in Fig. 2, is built upon the commonly used wNNLS approach[15,17,25] with several essential improvements (see Methods and Supplementary Note 1). Concretely, for a subject *j*, DeMixSC estimates its cell-type proportions, denoted by $\hat{p}_j$, by minimizing a composite of two weighted squared error terms, as given below in (1):

$$\hat{p}_j = \underset{p_j \geqslant 0}{\operatorname{argmin}} \left(\sum_{g \in G_1} w_{jg} \left(y_{jg} - n_j \sum_{k \in K} p_j^k \widehat{R}_g^k \right)^2 + \sum_{g \in G_2} w_{jg} \left(\frac{y_{jg}}{a_g} - n_j \sum_{k \in K} p_j^k \frac{\widehat{R}_g^k}{a_g} \right)^2 \right). \quad (1)$$

Here, y_{jg} is the observed expression value of gene *g* from subject *j* in bulk RNA-seq data, n_j is a normalization constant, $\widehat{R}_g^k$ is the estimated cell-type-specific expression value of cell type *k* in the reference matrix derived from sc/snRNA-seq data, and w_{jg} is the weight. The gene sets G_1 and G_2 comprise genes with minimal and substantial impact by technological discrepancy, respectively. The first innovation of DeMixSC is the identification of and adjustment to genes presenting consistently high technological discrepancy (G_2), using a small representative benchmark dataset such as our special matched RNA-seq data from 24 retinal samples (Fig. 2A). DeMixSC uses a DE analysis between bulk and matched pseudo-bulk RNA-seq data to segregate genes with low inter-platform discrepancy (non-DE genes, G_1) from those highly affected by technological discrepancy (DE genes, G_2). It then employs a partitioned loss function and adjusts genes from G_2 by rescaling their expressions by a_g (Figs. 2B, C) to mitigate the influence of technological discrepancy. Here we assign a_g as the $\log_2$ transformed mean expression of the matched bulk and pseudo-bulk RNA-seq data.

The second innovation of DeMixSC comes from our proposed function w_{jg}^*, which is given by

$$w_{jg}^{*-1} = (\hat{y}_{jg})^2 + (y_{jg} - \hat{y}_{jg})^2 + 2, \quad (2)$$

where $\hat{y}_{jg}$ denotes the fitted expression value. This weight function comprises three terms: the squared fitted expression, the squared residual, and a baseline constant (Fig. 2C). The fitted part addresses genes with high expression levels. The squared residual accounts for the remaining variance after fitting. The baseline constant constrains the weight range. Previous weight functions contain either the first or the second part only[10,14–17]. We reason that both parts are needed to comprehensively account for variations across cells and across samples, hence introducing this new weight function. In a test experiment using the benchmark dataset (see Methods, Extended Data Fig. 2), we observe that the previous method assigns higher weights to genes with low technological discrepancy rather than cell-type-specific genes, yielding high cell-type-wise collinearities. DeMixSC instead selects more cell-type-specifically expressed genes and reduces the cell-type-wise collinearity.

DeMixSC operates as a two-tier model in application. First, DeMixSC utilizes a specifically designed benchmark dataset to identify and adjust genes with high inter-platform discrepancies (Fig. 2A). Second, to deconvolve a large unmatched bulk RNA-seq dataset, DeMixSC aligns the large bulk cohort with the bulk RNA-seq data in the small benchmark dataset[28] (Fig. 2B) to generalize the detected technological discrepancy and then runs the refined wNNLS framework for deconvolution (Fig. 2C). Our main prerequisite is a matched tissue type between the small benchmark dataset and the large targeted cohort.

Compare the estimation accuracy of DeMixSC to existing deconvolution methods

Using our benchmark data, we compare the performance of DeMixSC to eight existing deconvolution methods[10–15,17,18], including AutoGeneS, BayesPrism, CIBERSORTx, DWLS, MuSiC, RNAseive, SCDC, and SQUID (see Methods, Fig. 3A). The retinal tissue samples in our benchmark dataset encompass ten distinct cell types. We focus our evaluation of different deconvolution methods on seven major cell types (Fig. 1A; amacrine cells, ACs; bipolar cells, BCs; Cone cells; horizontal cells, HCs; Müller glial cells, MGs; retina ganglion cells, RGCs; Rod cells), which on average account for 98% of the total cell population[30].

Overall, DeMixSC achieves the lowest root mean squared error (RMSE) and mean absolute error (MAE) scores in deconvolving bulk RNA-seq data, with mean values of 0.06 and 0.04 (Figs. 3B, C). Moreover, DeMixSC produces similar RMSE and MAE scores for deconvolving bulk and pseudo-bulk RNA-seq data (mean RMSE: bulk 0.06, pseudo-bulk 0.04; and mean MAE: bulk 0.04, pseudo-bulk 0.03). This suggests that DeMixSC properly adjusts for undesired technological discrepancies. In contrast, existing methods perform poorly for bulk as compared to perform reasonably for pseudo-bulk, reflecting that technological discrepancies compromise deconvolution accuracy and remain unaddressed by current approaches (Figs. 3B, C). Specifically, AutoGeneS shows higher RMSE and MAE scores for pseudo-bulk data, likely due to an inability to distinguish between Rod and Cone cells, which share largely similar expression profiles (Fig. 3D). DWLS excels in deconvolving pseudo-bulk samples but fails short in bulk RNA-seq data, possibly due to overfitting. Using the tree-based deconvolution in MuSiC or the ensemble option in SCDC does not improve their accuracy (Extended Data Fig. 3). CIBERSORTx presents overall reasonable performances in both bulk and pseudo-bulk data, likely because of its batch effect correction step. Looking further at the cell-type level, we observe systematic biases across methods. Most methods underestimate the proportions of ACs and BCs while overestimating HCs and MGs (Fig. 3D). DeMixSC accurately estimates the proportions of all seven major cell types and improves the deconvolution results for ACs, BCs, HCs, and MGs (Figs. 3D-F; RMSE: 0.04, 0.06, 0.03, 0.03 and MAE: 0.03, 0.05, 0.03, 0.02 for four cell types, respectively). Similar to most other methods, DeMixSC also underestimates the proportion of Cone cells (true mean proportion at 0.04). Finally, we test the performance of each method using various data transformation procedures (see Methods). We test the robustness of these methods under varied data formats[29], including RPM, RPKM, and TPM, and find DeMixSC to be robust to data transformations (Extended Data Fig. 4). In line with previous benchmarking studies[24], we find using raw counts as input is sufficient to obtain good results. Finally, we find SQUID delivers the least desirable results in this benchmarking study (mean RMSE and MAE in bulk data: 0.25 and 0.17). The pitfall with SQUID possibly lies in its data transformation step[23], which has the potential to misrepresent gene expression profiles. In summary, our DeMixSC framework achieves the most accurate deconvolution among the compared methods by successfully addressing key issues with technological discrepancy between two sequencing platforms.

Apply DeMixSC to human peripheral retina bulk RNA-seq data

Age-related macular degeneration (AMD) is characterized as a deterioration of retina and choroid that leads to a substantial loss in visual acuity, with the loss of Rod cells being a major manifestation. It is the leading cause of blindness among the global elderly population[31]. However, the molecular and cellular events that underlie AMD remain poorly understood, impeding the development of effective treatments[32]. Understanding the molecular and cellular dynamics is essential for targeting the progression of AMD. We aim to examine cell-type proportion changes during AMD progression using bulk RNA-seq samples from 453 human peripheral retina[7] (see Methods). Among these retina samples, 105 are in the Minnesota Grading System 1 (MGS1), with 97 in MGS2, 88 in MGS3, and 61 in MGS4). An MGS1 rating suggests healthy retina, and an MGS4 rating suggests AMD, while MGS2 and MGS3 represent intermediate stages[33].

We run the two-tier DeMixSC to first align the AMD cohort with the bulk data from our specialized benchmark dataset of retina samples, and then run wNNLS to estimate cell-type proportions in the AMD cohort (see Methods, Fig. 4A). As we have more than one single-nucleus reference sample, DeMixSC produces robust deconvolution estimates among the consensus and each individual single-nucleus reference at both cell-type and sample levels (see Methods, Figs. 4B, C). The top-weighted genes selected by DeMixSC present low cell-type-wise collinearity across samples (see Methods, Extended Data Fig. 5). DeMixSC achieves cell-type proportions (Fig. 4D) that are closer to experimental measures for non-AMD samples[34] , with an MAE of 0.03 (Supplementary Table 2). In contrast, three other methods, MuSiC2, CIBERSORTx, and SQUID, overestimate the proportion of Rod cells and underestimate all other cell types (Extended Data Fig. 6). DeMixSC identifies consistent changes in proportions in three major cell types over the axis of MGS1-4 (Extended Data Fig. 7), including a statistically significant decrease in proportions of Rod cells (non-AMD: 0.57 vs. AMD: 0.53; t-test *P*-value = 0.02), as well as statistically significant increases in proportions of BCs (non-AMD: 0.16 vs. AMD: 0.17, t-test *P*-value = 0.05) and MGs (non-AMD: 0.12 vs. AMD: 0.15, t-test *P*-value < 0.001). It is known that the adult retina is not able to generate new cells[31,32,35], so we hypothesize that it is the loss of Rod cells that results in a decreased total number of cells, hence inflating the Rod cell proportions as well as proportions of other cells in the AMD condition. Indeed, we find losing 15% of all Rod cells can result in the observed subtle drop from 0.57 to 0.53 for the proportions of Rod cells, as well as the subtle increases in proportions of the other two cell types (see Methods).

One major utility of obtaining accurate cell-type proportions is to enable the downstream accurate evaluation of cell-type-specific gene expressions between disease conditions. We test the utility of DeMixSC by running the kernel of integrated single cells (KERIS), a method that estimates cell-type specific gene expression levels between conditions, on the proportion estimates of non-AMD and AMD samples provided by DeMixSC and MuSiC2 (see Methods, Supplementary Tables 4-7). Pathway enrichment analysis using the top 100 expressed genes in Rod cells, MGs, and BCs, respectively, shows that with the majority cell type where a similar trend in proportions is presented by both methods, e.g., in Rod cells (Fig. 4D, Extended Data Fig. 6A), their cell-type specific enriched pathways also overlap (Fig. 4E); when the proportions are underestimated by MuSiC2, e.g., in MGs and BCs (Fig. 4D, Extended Data Fig. 6A), DeMixSC-based estimates generates pathways that are biologically meaningful, such as extracellular matrix remodeling and hypoxia response in MGs[36,37] and synapse regulation and membrane potential in BCs[38] (Fig. 4E). In conclusion, DeMixSC demonstrates high accuracy and generalizability in deconvolving unmatched large bulk RNA-seq datasets using a small tissue-type-matched benchmark dataset.

Discussion

This work addresses technological discrepancies between bulk and sc/sn RNA-seq data in order to improve the deconvolution accuracy of bulk RNA-seq data. We construct a specialized benchmark dataset of healthy retina samples and thoroughly evaluate the impact of technological discrepancies on existing single-cell-based deconvolution methods[10–18]. Utilizing this benchmark dataset, we introduce the DeMixSC deconvolution method that makes innovative improvements on the wNNLS framework to address the consistently observed technological discrepancy at a gene level. The distinct advantage of DeMixSC lies in its superior deconvolution accuracy and broad generalizability. As demonstrated in our benchmark study, DeMixSC achieves more accurate estimates of cell-type proportions compared to existing methods. More importantly, DeMixSC is generalizable to deconvolving unmatched large bulk datasets, only requiring a matched tissue type between a small benchmark dataset and the targeted bulk cohorts. In the application to complex retina samples from patients with AMD, DeMixSC identifies subtle yet critical cell-type proportion changes, highlighting its ability to reflect cellular dynamics during disease progressions and facilitate the cell-type-specific gene expression analysis. Most existing single-cell-based deconvolution methods[10–18] are adept at discerning between 7 to 13 cell types from bulk RNA-seq data, DeMixSC aligns with these capabilities as demonstrated in our study. DeMixSC is computationally efficient, completing the analysis of 453 AMD samples in under five minutes, and presents robust convergence against different starting values (see Methods, Extended Data Figs. 8). The implementation of the DeMixSC framework is freely available at https://github.com/wwylab/DeMixSC.

The generation of the benchmark dataset in DeMixSC is crucial for accurate and reliable estimates of cell-type proportions. Our study employs a specifically tailored cDNA library preparation procedure to generate the benchmark dataset of retinal samples. A critical step in the data generation process is to ensure the 'matchness' of paired bulk and snRNA-seq data. In our setup, the cDNA library for bulk RNA-seq is generated using the Smartseq v4 ultralow input RNA kit, a protocol that is similar to that of the snRNA-seq. We advocate for the preparation of ample tissue samples when executing this pipeline to ensure a reliable benchmark dataset.

The stride of DeMixSC is noteworthy, yet there is potential room for improvements which we leave for future work. The key of DeMixSC rests on effectively discerning and down-weighting genes with technological discrepancies. A potential challenge arises in gene identification when applying DeMixSC to tissue types (e.g., tumors) with high cellular plasticity. In this scenario, a stratified categorization of genes into three distinct groups can be beneficial: technologically stable genes, biologically stable genes (e.g., global tumor signature genes[6]), and the remaining unstable genes. Moreover, DeMixSC can gain from principled statistical methods to identify and adjust genes simultaneously. Additionally, alternative methods to ComBat[28] for aligning the large cohort with the benchmark dataset can be considered when dealing with tumor samples, which often are extremely heterogeneous with complex batch structures. Considering these potential adaptations, we anticipate the promising utility of DeMixSC in cancer research. By utilizing a concise benchmark dataset derived from matched tissue specimens, DeMixSC can be leveraged to accurately deconvolve large bulk cohorts acquired through either surgery or needle biopsies, which accelerates the discovery of cell subtypes and cell-type-specific markers among diverse patient groups.

Methods

Human retina sample collection. These samples were obtained from 24 individuals between the ages of 73 and 91 who had passed away due to respiratory or heart failure or from a myocardial infarction (Supplementary Table 1). Human donor eyes were obtained through the Utah Lions Eye Bank. For this study, we included samples collected within six hours postmortem. Dissections of donor eyes were performed immediately following a previously published protocol[39]. Macular retinal tissue was collected using a six mm disposable biopsy punch (Integra, Cat # 33-37), flash-frozen, and stored at -80°C. Only one eye was used per donor, and donors with any history of retinal degeneration, diabetes, macular degeneration, or drusen were excluded from the study. Additionally, each donor underwent an ophthalmology check to ensure that the eye was in a healthy condition. Institutional approval for patients consent to donate their eyes was obtained from the University of Utah, and the study adhered to the principles of the Declaration of Helsinki. All retinal tissues were deidentified in accordance with HIPAA Privacy Rules.

Generation of benchmark data from 24 human retinal samples.

Single-nucleus mRNA sequencing. Nuclei were isolated with prechilled fresh-made RNase-free lysis buffer (10 mM Tris-HCl, 10 mM NaCl, 3 mM MgCl2, 0.02% NP40). The frozen tissue was resuspended and triturated to break the tissue structure in lysis buffer and homogenized with a Wheaton™ Dounce Tissue Grinder. Isolated nuclei were filtered with 40 µm Flow Cell Strainer and stained with DAPI (4′,6-diamidino-2-phenylindole, 10 µg/ml) before fluorescent cytometry sorting (FACS) on an FACSAria III Cell Sorter (BD, San Jose, CA, USA) in the Cytometry and Cell Sorting Core at Baylor College of Medicine (BCM). All single-nucleus RNA sequencing was performed at the Single Cell Genomics Core (SCGC) at BCM. Single-nucleus cDNA library preparation and sequencing were performed following the manufacturer's protocols (https://www.10xgenomics.com). A single-nucleus suspension was loaded on a Chromium controller to obtain a single-cell GEMS (gel beads-in-emulsions) for the reaction. The snRNA-seq library was prepared with chromium single cell 3′ reagent kit v3 (10x Genomics). The product was then sequenced on an Illumina NovaSeq 6000 (https://www.illumina.com).

Bulk mRNA sequencing of retina single-nucleus suspension. To ensure the ‘matchness’ of paired bulk and snRNA-seq data, the mRNA library for bulk RNA-seq followed the same pipeline as for snRNA-seq. Specifically, matched samples with snRNA-seq were used for RNA isolation by applying TRIzol (Invitrogen) to the separated single-nucleus resuspension. cDNA was prepared from ~1 ng of total RNA by using the Smartseq v4 ultralow input RNA Kit according to the manufacturer’s directions (Takara). The libraries were made using

Nextera XT library prep (Illumina). Full-length RNA-seq was performed on NovaSeq 6000 sequencers according to the manufacturer's directions (Illumina).

Preprocessing of snRNA-seq and bulk RNA-seq data. Retina snRNA-seq UMI (unique molecular identifier) count matrices were obtained using CellRanger (version 3.1.0) following the official guide to estimate absolute counts and were then processed using the Seurat[40] package (version 3.6.0). Specifically, for each snRNA-seq dataset, we first removed genes expressed in fewer than 5% of cells; then filtered out cells with either fewer than 500 total UMIs or 200 expressed genes, or more than 50% total UMI counts derived from mitochondrial genes. The total numbers of transcripts of each cell were then normalized to 10,000, followed by a natural log transformation. Highly variable genes were detected and used for principal component analysis (PCA). Cells were then clustered using the Seurat package at a resolution of 0.5.

For bulk RNA-seq data, the quality of raw sequencing data was first evaluated by FastQC (version 0.11.9), and low-quality reads and adapters were then trimmed by Trimmomatic (version 0.4.0). Next, reads that passed quality control were aligned to GRCh38 using the 2-pass mode of STAR (version 2.7.7b), and read counts were obtained by featureCount function from the Subread package (version 1.22.2) following the standard pipeline.

Cell type annotation for snRNA-seq data. Seven major cell types, including Cone cells, Rod cells, horizontal cells (HCs), bipolar cells (BCs), amacrine cells (ACs), retinal ganglion cells (RGCs), and Müller glia cells (MGs), were annotated using known marker genes[34,35] (Extended Data Figs. 1A, B; Supplementary Table 3). For the deconvolution analysis of bulk AMD retinal samples[7], we included additional 3 minor cell types, including astrocytes, microglia cells, and retina pigmental epithelium (RPE).

Generation of ground truth proportion, and pseudo-bulk mixtures. With each annotated snRNA-seq data, the true proportion of each cell type was calculated as the number of cells in the cell type divided by the total number of cells. Pseudo-bulk mixtures corresponding to each bulk were calculated by adding up the UMI counts from all the annotated cells per gene from the matched snRNA-seq data.

Statistical analysis. We used paired Student's t-tests to identify the differentially expressed (DE) genes between matched bulk and pseudo-bulk RNA-seq data. The *P*-values for DE analysis were adjusted for multiple testing by the Benjamini-Hochberg (BH) method[41]. We used Student's t-tests to compare the estimated cell-type proportions between non-AMD and AMD groups from different deconvolution methods. We used Wilcoxon rank-sum tests to compare the sequencing read depth between bulk and pseudo-bulk data. For all *P*-values in this study, significance levels were denoted as follows: no significant (ns), *P*-value >0.05; **P*-value ≤0.05; ***P*-value ≤0.01; and ****P*-value ≤0.001.

DeMixSC deconvolution framework. DeMixSC is a reference-based model built upon the wNNLS deconvolution framework with several improvements. Our model explicitly requires a benchmark dataset for training. To begin with, we revisit the core equation of existing deconvolution methods[10,11,14–17,25], which is

$$\hat{p}_j=\underset{p_j\geq 0}{\mathrm{argmin}}\sum_{g\in G} w_{jg}\left(y_{jg}-n_j\sum_{k\in K}p_j^k\widehat{R}_g^k\right)^2 \quad (1),$$

where y_{jg} is the observed expression value of gene g from subject j in the bulk RNA-seq data, $\widehat{R}_g^k$ is the estimated expression value of gene g and cell type k in the reference matrix derived from sc/snRNA-seq data, w_{jg} is the weight of each gene g from subject j, $\hat{p}_j$ is the estimated vector of cell-type proportions, and n_j is a normalization constant. The main drawback of model (1) is that it does not address technological discrepancies observed in our benchmark data. To explain this, we split the squared term in Eq(1) into two components, and rewrite the model as

$$\hat{p}_j=\underset{p_j\geq 0}{\mathrm{argmin}}\sum_{g\in G} w_{jg}\left(\left(\tilde{y}_{jg}-n_j\sum_{k\in K}p_j^kR_g^k\right)+\left(\epsilon_{jg}-n_j\sum_{k\in K}p_j^kb_g^k\right)\right)^2 \quad (2),$$

where $\tilde{y}_{jg}$ is the true expression value in the bulk data, $\tilde{y}_{jg}+\epsilon_{jg}=y_{jg}$, and R_g^k is the true cell-type-specific reference matrix, $n_j\sum_{k\in K}p_j^kR_g^k+n_j\sum_{k\in K}p_j^kb_g^k=n_j\sum_{k\in K}p_j^k\widehat{R}_g^k$. The left component of Eq(2), $\left(\tilde{y}_{jg}-n_j\sum_{k\in K}p_j^kR_g^k\right)$, consists of the true bulk-level expression $\tilde{y}_{jg}$ and the fitted value based on true cell-type-specific mean expression $n_j\sum_{k\in K}p_j^kR_g^k$. This component reflects the true estimation errors that we aim to minimize. The right component in Eq(2), $\left(\epsilon_{jg}-n_j\sum_{k\in K}p_j^kb_g^k\right)$, defines the difference in noise introduced by the bulk (ϵ_{jg}) and the sc/snRNA-seq data (denoted by $p_j^kb_g^k$ for cell type k). Therefore, the right component in Eq(2) represents the measurable technological discrepancy between sequencing platforms. Genes with higher levels of technological discrepancy contribute more to the right component (see Supplementary Note 1). Thus, when the technological discrepancy overtakes the true signal, instead of minimizing estimation errors (the left component), this model is geared towards minimizing technological discrepancies (the right component) and is no longer fitting the expression profiles of individual bulk samples.

To address the issue of overcorrection for technological discrepancies in model (1), we introduce DeMixSC, which estimates cell-type proportions by minimizing a partitioned loss function, as shown below:

$$\hat{p}_j=\underset{p_j\geq 0}{\mathrm{argmin}}\left(\sum_{g\in G_1} w_{jg}\left(y_{jg}-n_j\sum_{k\in K}p_j^k\widehat{R}_g^k\right)^2+\sum_{g\in G_2} w_{jg}\left(\frac{y_{jg}}{a_g}-n_j\sum_{k\in K}p_j^k\frac{\widehat{R}_g^k}{a_g}\right)^2\right) \quad (3),$$

where G_1 is a set of genes hardly affected by technological discrepancies, i.e., the non-DE genes between matched bulk and pseudo-bulk data, G_2 contains genes highly affected by technological discrepancies, i.e., the DE genes between matched bulk and pseudo-bulk (see Supplementary Note 1), and a_g is the $\log_2$ transformed mean expression of the matched bulk and pseudo-bulk RNA-seq data. We use a_g to rescale the expression levels of DE genes to reduce the impact of the large differences in their error terms, as shown in Eq(2). Instead of directly filtering them out, our approach preserves those DE genes in our model, as they could still have biological significance and contribute to mixed expression levels. In addition, we introduce a new weight function (w_{jg}^*) to reduce the influence of highly expressed genes and assign lower rankings for genes with large variances:

$$w^{*-1}_{jg}=(\hat{y}_{jg})^2+(y_{jg}-\hat{y}_{jg})^2+2 \quad (4).$$

The current literature uses either the squared fitted value[10,15] $(\hat{y}_{jg})^2$ or the variance[14,16,17] $(y_{jg}-\hat{y}_{jg})^2$ for weights, but never both. The constant 2 in Eq(4) is introduced for controlling the range of weights. Using the summation of the three terms as our new weight function improves model fit, accounts for variability, and enhances the statistical robustness of our framework. The detailed mathematical derivation is in Supplementary Note 1.

Collinearity in top-weighted genes. To test the cell-type-wise collinearity among samples, we selected the top 1000 weighted genes from each deconvolved sample and used them to compute the cell-type-wise collinearity matrix. The cell-type-wise collinearity was quantified using the Pearson correlation coefficient. We then took the average of cell-type-wise collinearity matrices across all samples to achieve the final collinearity heatmap (Extended Data Fig. 2 and Extended Data Fig. 4).

Convergence property of the DeMixSC algorithm. To evaluate how robust DeMixSC is against different initial values, we randomly selected a sample from the AMD cohort[7] as a case study. To create different initial values, we set three different scale factors $n = \{100, 380, 1000\}$. For each scale factor, we chose 10 extreme starting values for the proportions $\hat{p}$, with the proportion of one out of 10 cell types being one and the rest being zero (Extended Data Figs. 8A). Finally, we used the 30 values of $n\times\hat{p}$ to initialize the wNNLS framework and then compared the estimates of DeMixSC (Extended Data Figs. 8B).

Data normalization of bulk mixtures. We applied the following data normalizations to the bulk raw count matrices[29]: (i) reads per million mapped reads (RPM); (ii) reads per kilobase of transcript, per million mapped reads (RPKM); and (iii) transcripts per million (TPM). Both RPKM and TPM include an additional step that uses the gene length to obtain normalized counts per million.

Computational deconvolution. Nine deconvolution methods that use scRNA-seq as a reference were tested in our study[10–18]. We first used the default settings of each method as described in the GitHub repository or the websites (AutoGeneS: https://github.com/theislab/AutoGeneS, BayesPrism: https://github.com/Danko-Lab/BayesPrism, CIBERSORTx: https://cibersortx.stanford.edu/, DWLS: https://github.com/dtsoucas/DWLS, MuSiC: https://github.com/xuranw/MuSiC, MuSiC2: https://github.com/Jiaxin-Fan/MuSiC2/tree/master/R, RNAseive: https://github.com/songlab-cal/rna-sieve, SCDC: https://github.com/meichendong/SCDC, and SQUID: https://github.com/favilaco/deconv_matching_bulk_scnRNA). For CIBERSORTx, we followed the recommended built-in batch correction method for the deconvolution analysis of bulk samples (batch mode = S).

Additionally, we evaluated the performance of the tree-guided deconvolution of MuSiC[17] and the ensemble option of SCDC[14]. For tree-guided MuSiC, we first performed hierarchical clustering on the single-cell reference dataset; based on the hierarchical clustering results, we grouped Cone and Rod cells to form a mega cell cluster (Extended Data Fig. 3A), and each of the remaining cell types also formed a cluster. Cell-type-specific marker genes of cones and rods were obtained using FindAllMarkers function from Seurat[40] package under the bimod likelihood ratio test. We ran MuSiC deconvolution first at the cell cluster level and then again within the Rod and Cone clusters. For the SCDC ensemble option, we ran deconvolution on SCDC with 3 different sc references; then, we ran the SCDC_ENSEMBLE function to obtain the ensemble deconvolution results.

Evaluation metrics for the deconvolution performance. We evaluated the estimated cell proportions of each method using two metrics, root-mean-square error (RMSE) and mean absolute error (MAE), at both sample and cell-type levels: $\mathrm{RMSE}^j=\sqrt{\frac{\sum_{k=1}^{K}(\hat{p}_j^k-p_j^k)^2}{K}}$, $\mathrm{MAE}^j=\frac{1}{K}\sum_{k=1}^{K}|\hat{p}_j^k-p_j^k|$, $\mathrm{RMSE}^k=\sqrt{\frac{\sum_{j=1}^{J}(\hat{p}_j^k-p_j^k)^2}{J}}$, $\mathrm{MAE}^k=\frac{1}{J}\sum_{j=1}^{J}|\hat{p}_j^k-p_j^k|$, where $\hat{p}_j^k$ denotes the estimated cell proportion by the investigated method for cell type k and sample j, and p_j^k is the corresponding ground truth. We use J to denote the total number of samples and K to denote the total number of cell types. Smaller RMSE and MAE values indicate a better deconvolution performance.

Quality control for the AMD cohort. The large AMD bulk cohort[7] comprised 523 samples. We conducted quality control following the pipeline described in the original study. Briefly, samples were filtered out due to ambiguous clinical features (n=26), poor sequencing results (n=16), inconsistent genotyping results (n=14), and divergent ancestry (n=6). A total of 70 samples were removed, with a total of 453 samples remaining to be used to perform the deconvolution analysis.

Reference matrices for deconvolving the AMD cohort. For our deconvolution analysis of AMD retinal samples[7], we generated a consensus single-nucleus reference by integrating seven samples from Batch-2 (Sample 5, 10, 12, 18, 19, 21 and 23, Supplementary Tables 1). We selected these samples as they adequately represented these three minor cell types: astrocytes, microglia cells, and RPE. For each sample, we randomly selected up to 500 cells per cell type, using all available cells for types with fewer than 500 cells. Relative abundance and cell size for every cell type were calculated for each sample. A consensus reference matrix was subsequently derived by averaging relative abundance and cell size across the selected samples.

Accounting for the total cell loss in the AMD cohort. The decline in overall cell count induced by Rod cell loss likely amplifies the cell-type proportions in the AMD samples[31,35,37]. To quantify the true proportion of Rod cell loss, we use x to represent the mean percentage of lost Rod cells in the AMD group. With the mean estimated fraction of Rod cells is 0.57 in non-AMD and 0.53 in AMD (Fig. 4D), we derived the relation: 0.57(1-x)/(1-0.57x)=0.53. Solving for x gives 0.15. This 15% reduction in Rod cells in the peripheral AMD retina aligns well with biological evidence indicating the primary impact of AMD on the macular region[31].

Then, we demonstrated the observed subtle increases in BCs and MGs are driven by the death of Rod cells. The mean estimated fractions of BCs and MGs in the non-AMD group were 0.16 and 0.12 (Fig. 4D). Based on our Rod cell loss metric, the expected fractions in the AMD group would be 0.16/(1-0.57*0.15)=0.17 for BCs and 0.12/(1-0.57x0.15)=0.13 for MGs, which were close to the DeMixSC's estimates of BCs (0.17) and MGs (0.15) in the AMD group (Fig. 4D).

Cell-type-specific functional analysis. we applied KERIS[42] with default settings as described in the GitHub repository (https://github.com/YosefLab/keris). Briefly, KERIS takes bulk gene expression, cell-type proportion estimates, and a vector of sample conditions (e.g., non-AMD vs AMD) as input, and gives the mean expression of each cell type as output. For each cell type, we selected the top 100 genes ranked by the estimated mean expression levels (Supplementary Tables 4 and 5) and performed a Gene Ontology (GO) enrichment analysis on three major cell types Rod, BC, and MG using clusterProfiler4.0[43] following the standard pipeline (Supplementary Tables 6 and 7) (http://yulab-smu.top/biomedical-knowledge-mining-book/GOSemSim.html).

Declarations

Ethics approval and consent to participate

Institutional approval for patient consent to donate their eyes was obtained from the University of Utah, and the study adhered to the principles of the Declaration of Helsinki. All retinal tissues were deidentified in accordance with HIPAA Privacy Rules.

Consent for publication

All authors have approved the manuscript for submission.

Availability of data and materials

The snRNA-seq data from the human retina tissue will be deposited and released at the Human Cell Atlas Data Portal. The raw data and count matrix of bulk RNA-seq data from human retinal tissue will be available at GEO under accession number: GSE175937.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

S.G. is supported by Human Cell Atlas Seed Network - Breast, Chan Zuckerberg Institute, MD Anderson Colorectal Cancer Moon Shot Program, DoD PC210079. X.L. is supported by R01CA239342. A.K. is supported by 5T32CA096520-15. J.P.S. is supported by the Cancer Prevention and Research Institute of Texas as a CPRIT Scholar in Cancer Research and by National Institutes of Health (K22CA234406). R.C. is supported by CZI (CZF2019-02425), National Eye Institute (R01EY022356, R01EY018571), and Retinal Research Foundation. W.W. is supported by Human Cell Atlas Seed Network - Retina, Chan Zuckerberg Institute, NIH R01CA268380, DoD PC210079, P30CA016672. This work is also supported by CPRIT Single Core grant RP180684 and the Cytometry and Cell Sorting Core at Baylor College of Medicine with funding from the CPRIT Core Facility Support Award (CPRIT-RP180672), the NIH (CA125123 and RR024574) and the assistance of Joel M. Sederstrom. Single-nucleus RNA sequencing was performed at the Single Cell Genomics Core at BCM partially supported by NIH shared instrument grants (S10OD018033, S10OD023469, S10OD025240), P30CA125123, P30EY002520, and CPRIT Comprehensive Cancer Epigenomics Core Facility RP200504.

Authors' contributions

W.W. and R.C. supervised the research. W.W. and R.C. conceived the ideas and designed the study. S.G. and X.L. developed the method, implemented the R package, and conducted all the analysis. X.C. conducted the sequencing experiments and generated the paired bulk and single-nuclei RNA-seq data for benchmarking purposes. Y.J. and S.J. help with the initial development of the method and evaluation. Q.L. and Y.L. help with the sequencing experiments. L.A.O., I.K.K., A.A., S.K., J.P.S., M.M.D., and R.C. provided samples, advised on the study design, and assisted with the interpretation of results. A.K. performed the initial benchmarking analysis. J.N.W. and R.C. contributed technical suggestions. W.W., S.G., and X.L. wrote the paper, with input from all authors. All authors reviewed and approved the final version of the manuscript.

References

1. Haniffa, M. *et al.* A roadmap for the Human Developmental Cell Atlas. *Nature 2021 597:7875* **597**, 196–205 (2021).

2. Zeng, Q. *et al.* Understanding tumour endothelial cell heterogeneity and function from single-cell omics. *Nature Reviews Cancer 2023 23:8* **23**, 544–564 (2023).

3. Gohil, S. H., Iorgulescu, J. B., Braun, D. A., Keskin, D. B. & Livak, K. J. Applying high-dimensional single-cell technologies to the analysis of cancer immunotherapy. *Nature Reviews Clinical Oncology 2020 18:4* **18**, 244–256 (2020).

4. Li, X. & Wang, C. Y. From bulk, single-cell to spatial RNA sequencing. *International Journal of Oral Science 2021 13:1* **13**, 1–6 (2021).

5. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat Rev Genet* **20**, 631–656 (2019).

6. Cao, S. *et al.* Estimation of tumor cell total mRNA expression in 15 cancer types predicts disease progression. *Nature Biotechnology 2022 40:11* **40**, 1624–1633 (2022).

7. Ratnapriya, R. *et al.* Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nature Genetics 2019 51:4* **51**, 606–610 (2019).

8. Anghel, C. V. *et al.* ISOpureR: An R implementation of a computational purification algorithm of mixed tumour profiles. *BMC Bioinformatics* **16**, 1–11 (2015).

9. Wang, Z. *et al.* Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration. *iScience* **9**, 451–460 (2018).

10. Cobos, F. A. *et al.* Effective methods for bulk RNA-seq deconvolution using scnRNA-seq transcriptomes. *Genome Biology 2023 24:1* **24**, 1–22 (2023).

11. Aliee, H. & Theis, F. J. AutoGeneS: Automatic gene selection using multi-objective optimization for RNA-seq deconvolution. *Cell Syst* **12**, 706-715.e4 (2021).

12. Erdmann-Pham, D. D., Fischer, J., Hong, J. & Song, Y. S. Likelihood-based deconvolution of bulk gene expression data using single-cell references. (2021)

13. Chu, T., Wang, Z., Pe'er, D. & Danko, C. G. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nature Cancer 2022 3:4* **3**, 505–517 (2022).

14. Dong, M. *et al.* SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief Bioinform* **22**, 416–427 (2021).

15. Tsoucas, D. *et al.* Accurate estimation of cell-type composition from gene expression data. *Nature Communications 2019 10:1* **10**, 1–9 (2019).

16. Fan, J. *et al.* MuSiC2: cell-type deconvolution for multi-condition bulk RNA-seq data. *Brief Bioinform* **23**, (2022).

17. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications 2019 10:1* **10**, 1–9 (2019).

18. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology 2019 37:7* **37**, 773–782 (2019).

19. Denisenko, E. *et al.* Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol* **21**, 1–25 (2020).

20. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom Bioinform* **3**, (2021).

21. Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).

22. Jin, H. & Liu, Z. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *Genome Biol* **22**, 1–23 (2021).

23. Jew, B. *et al.* Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature Communications 2020 11:1* **11**, 1–11 (2020).

24. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature Communications 2020 11:1* **11**, 1–14 (2020).

25. Ruppert, David, and Matthew P. Wand. Multivariate locally weighted least squares regression. *The annals of statistics* 1994: 1346-1370 (1994).

26. Mereu, E. *et al.* Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nature Biotechnology 2020 38:6* **38**, 747–755 (2020).

27. Tung, P. Y. *et al.* Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports 2017 7:1* **7**, 1–15 (2017).

28. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

29. Dillies, M. A. *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* **14**, 671–683 (2013).

30. Liang, Q. *et al.* A multi-omics atlas of the human retina at single-cell resolution. *Cell genomics* **3**, (2023).

31. Fleckenstein, M. *et al.* Age-related macular degeneration. *Nature Reviews Disease Primers 2021 7:1* **7**, 1–25 (2021).

32. Khanani, A. M. *et al.* Review of gene therapies for age-related macular degeneration. *Eye 2021 36:2* **36**, 303–311 (2022).

33. Olsen, T. W. & Feng, X. The Minnesota Grading System of eye bank eyes for age-related macular degeneration. *Invest Ophthalmol Vis Sci* **45**, 4484–4490 (2004).

34. Liang, Q. *et al.* Single-nuclei RNA-seq on human retinal tissue provides improved transcriptome profiling. *Nature Communications 2019 10:1* **10**, 1–12 (2019).

35. Menon, M. *et al.* Single-cell transcriptomic atlas of the human retina identifies cell types associated with age-related macular degeneration. *Nature Communications 2019 10:1* **10**, 1–9 (2019).

36. Xin, X. *et al.* Hypoxic retinal Muller cells promote vascular permeability by HIF-1-dependent up-regulation of angiopoietin-like 4. *Proc Natl Acad Sci U S A* **110**, (2013).

37. Ambati, J., Atkinson, J. P. & Gelfand, B. D. Immunology of age-related macular degeneration. *Nature Reviews Immunology 2013 13:6* **13**, 438–451 (2013).

38. Hoon, M. *et al.* Neurotransmission plays contrasting roles in the maturation of inhibitory synapses on axons and dendrites of retinal bipolar cells. *Proc Natl Acad Sci U S A* **112**, 12840–12845 (2015).

39. Owen, L. A. *et al.* The Utah Protocol for Postmortem Eye Phenotyping and Molecular Biochemical Analysis. *Invest Ophthalmol Vis Sci* **60**, 1204 (2019).

40. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

41. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

42. Rahmani, E., Jordan, M. I. & Yosef, N. Identifying systematic variation at the single-cell level by leveraging low-resolution population-level data. *bioRxiv* 2022.01.27.478115 (2022)

43. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).

Figures:

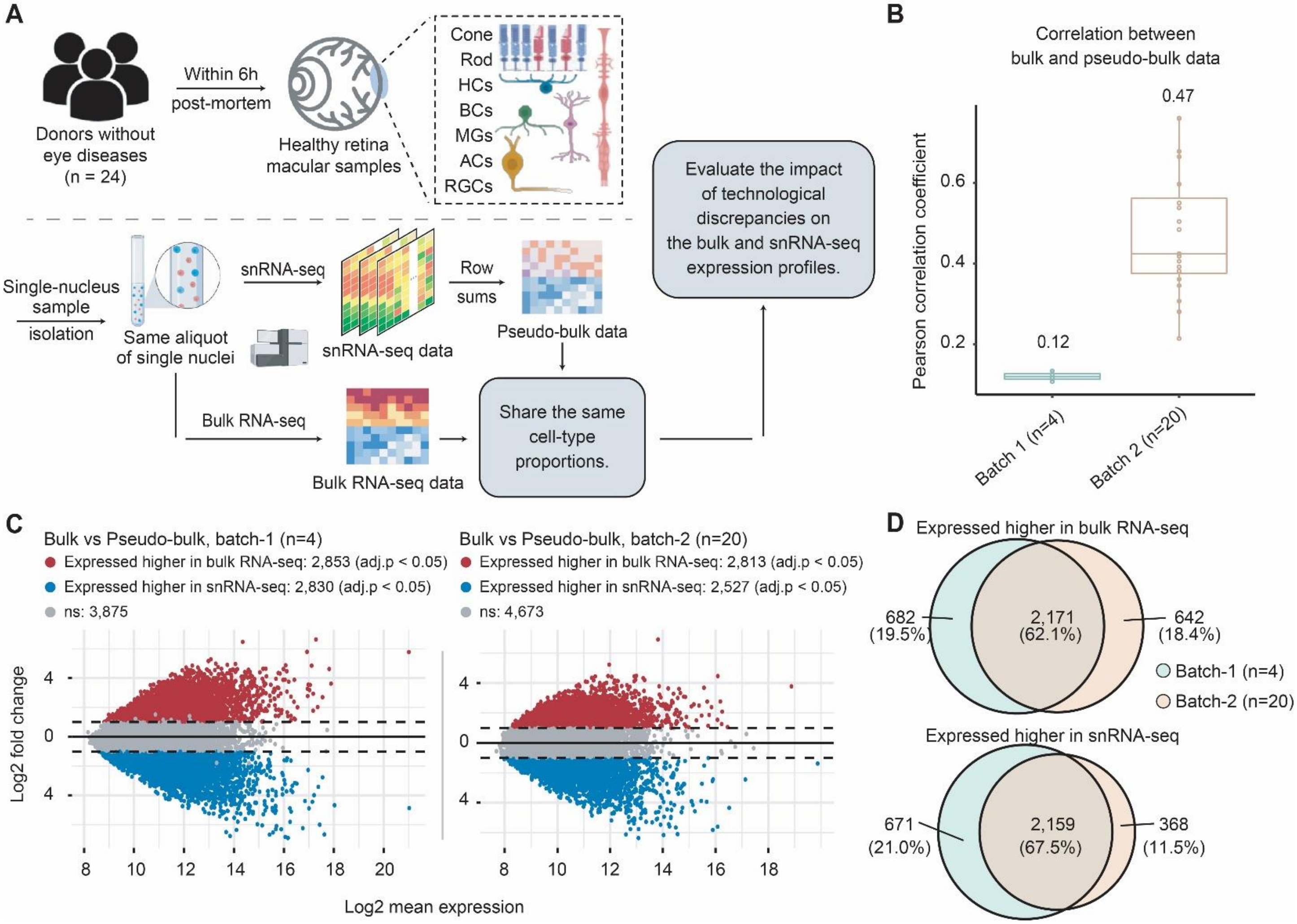


Figure 1 | Assessing technological discrepancies between bulk and single-cell sequencing platforms using matched single-nuclei aliquots.

A, Workflow for generating a benchmark dataset. We collect 24 healthy human retinal samples within six hours of postmortem. An illustration shows the layer and cell compositions of the human retina. Seven major cell types include photoreceptors (Rod and Cone cells), bipolar cells (BCs), retinal ganglion cells (RGCs), horizontal cells (HCs), amacrine cells (ACs), and Müller glia cells (MGs). Three minor cell types are not depicted in the illustration: astrocytes, microglia cells, and retinal pigment epithelial (RPE) cells. Samples are isolated into single-nucleus suspensions. The same aliquot of single-nuclei is used for both bulk and snRNA-seq profiling. The matched pseudo-bulk mixtures are generated as conventionally done by summing UMI counts across cells from all cell types in each sample. This data generation pipeline guarantees the matched bulk and snRNA-seq data share the same cell-type proportions, which enables us to evaluate the impact of technological discrepancies (i.e., the shot-gun sequencing procedure) on the bulk and snRNA-seq expression profiles. **B** and **C** show the influence of technological discrepancies at sample- and gene-level, respectively. **B**, Pearson correlation coefficient across genes between the matched real-bulk and pseudo-bulk RNA-seq data for one sample at a time for both batches. **C**, MA-plots displaying the mean expression levels of all genes between matched real-bulk and pseudo-bulk data. Differentially expressed (DE) genes are identified using the paired t-test with Benjamini-Hochberg (BH) adjustment. Red represents genes expressed higher in the real-bulk, and blue represents genes expressed higher in the pseudo-bulk. The horizontal dotted lines denote a log2 fold change of one between matched real-bulk and pseudo-bulk data. adj.p: adjusted *P*-values. **D**, Venn diagrams showing genes consistently expressed higher in the bulk (upper) or the pseudo-bulk (bottom) between the two batches, which were generated using different tissue samples and at a different time.

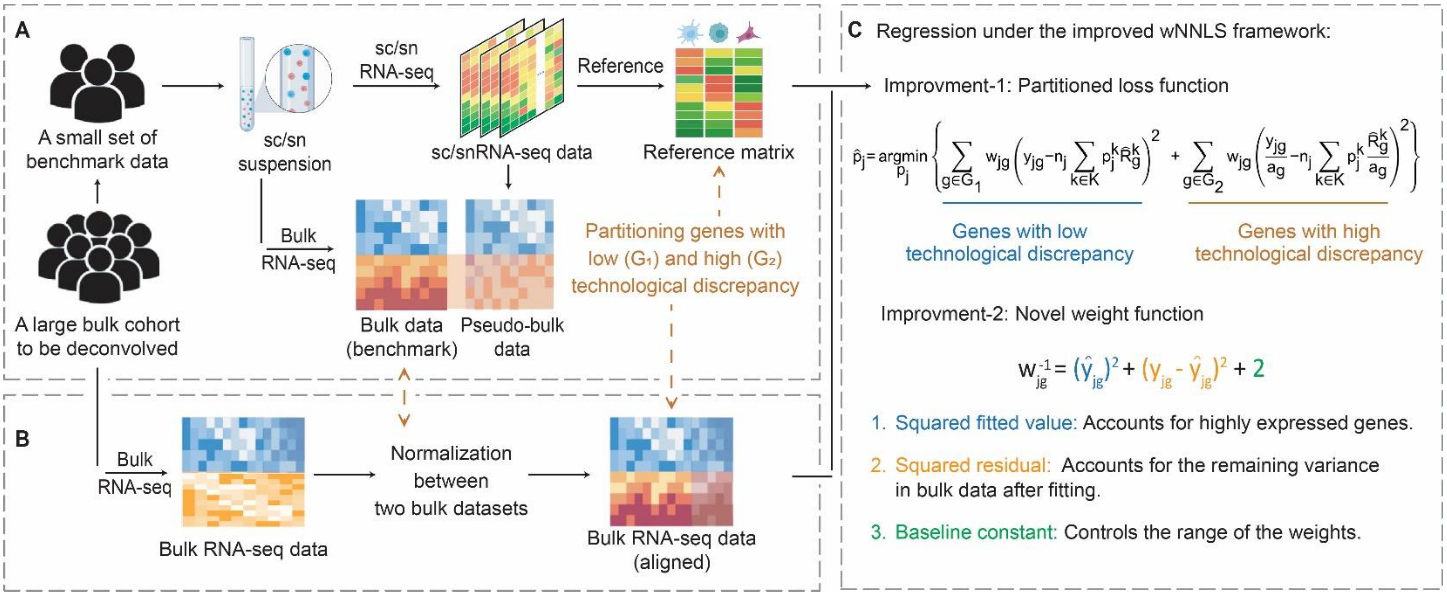

A
A small set of benchmark data
sc/sn suspension
sc/sn RNA-seq
sc/snRNA-seq data
Reference
Reference matrix
Bulk RNA-seq
Bulk data (benchmark)
Pseudo-bulk data
Partitioning genes with low (G₁) and high (G₂) technological discrepancy
A large bulk cohort to be deconvolved
B
Bulk RNA-seq
Bulk RNA-seq data
Normalization between two bulk datasets
Bulk RNA-seq data (aligned)
C Regression under the improved wNNLS framework:
Improvment-1: Partitioned loss function
$\hat{p}_j = \underset{p_j}{\operatorname{argmin}} \left\{ \sum_{g \in G_1} w_{jg} \left(y_{jg} - n_j \sum_{k \in K} p_j^k \hat{R}_g^k \right)^2 + \sum_{g \in G_2} w_{jg} \left(\frac{y_{jg}}{a_g} - n_j \sum_{k \in K} p_j^k \frac{\hat{R}_g^k}{a_g} \right)^2 \right\}$
Genes with low technological discrepancy
Genes with high technological discrepancy
Improvment-2: Novel weight function
$w_{jg}^{-1} = (\hat{y}_{jg})^2 + (y_{jg} - \hat{y}_{jg})^2 + 2$
1. Squared fitted value: Accounts for highly expressed genes.
2. Squared residual: Accounts for the remaining variance in bulk data after fitting.
3. Baseline constant: Controls the range of the weights.

Figure 2 | Overview of DeMixSC.

The DeMixSC framework for deconvolution analysis of bulk RNA-seq data using sc/sn RNA-seq data as reference. **A,** The framework starts with a benchmark dataset of matched bulk and sc/snRNA-seq data with the same cell-type proportions. Pseudo-bulk mixtures are generated from the sc/sn data. DeMixSC identifies DE genes and non-DE genes between the matched real-bulk and pseudo-bulk data. The non-DE genes are considered stably captured by both sequencing platforms (blue), while the DE genes are highly affected by technological discrepancies (orange). **B,** DeMixSC then employs a normalization procedure to perform the alignment between two bulk RNA-seq datasets (e.g., with ComBat). **C,** DeMixSC estimates cell-type proportions by regression under a weighted non-negative least square (wNNLS) framework with two improvements: 1) partitioning and adjusting genes with high technological discrepancies, and 2) a new weight function. Here, *g* is the index of gene, *j* is the index of subject, *k* is the index of cell type, $\hat{p}_j$ is the estimated cell-type proportions, w_{jg} is the weight, n_j is the normalization constant, $\hat{R}_g^k$ is the reference expression value derived from the sc/snRNA-seq data, a_g is the $\log_2$ transformed mean expression of the matched bulk and pseudo-bulk RNA-seq data, y_{jg} is the observed expression value in the bulk RNA-seq data, and $\hat{y}_{jg}$ is the corresponding fitted value.

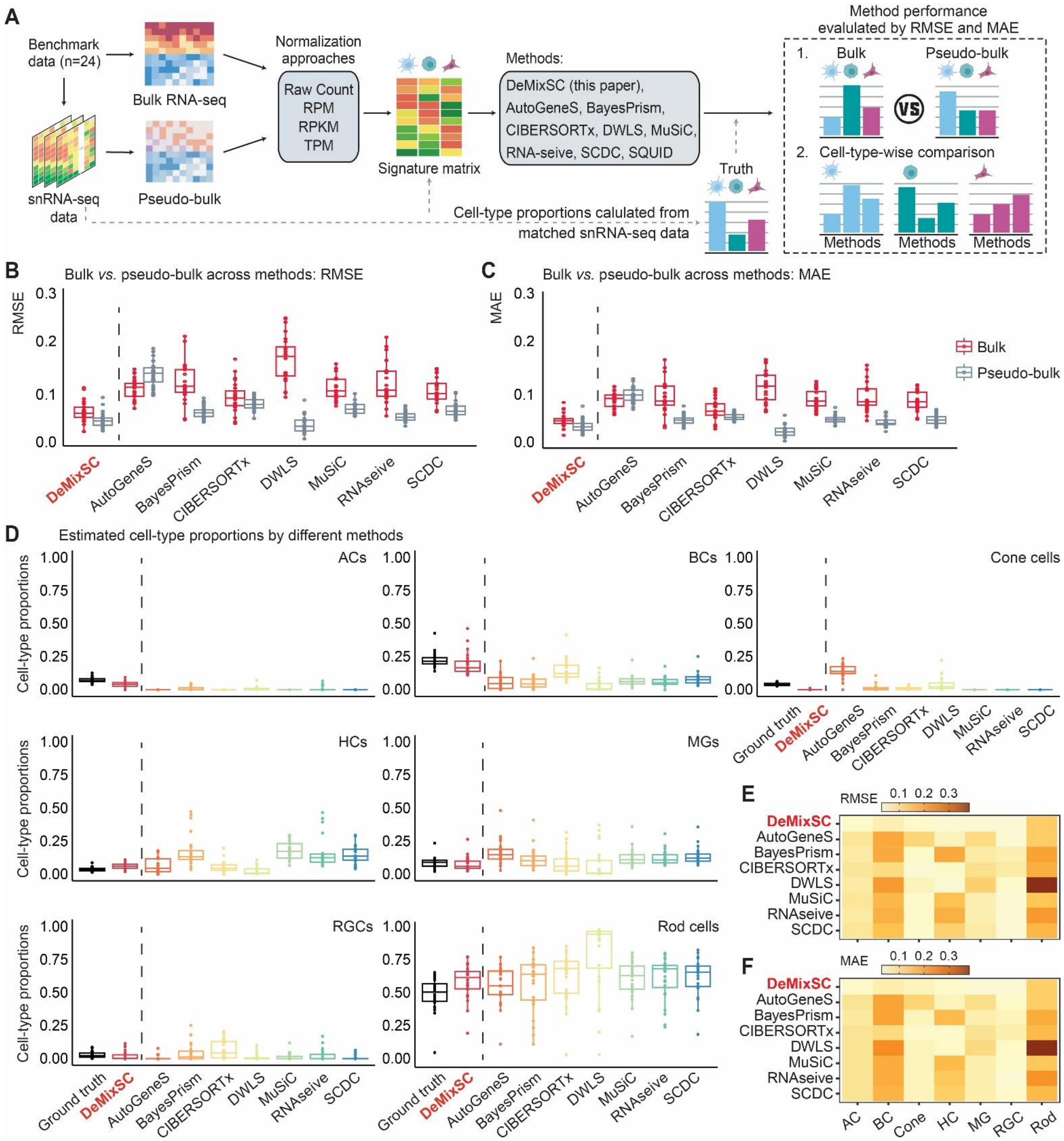


Figure 3 | Compare the estimation accuracy of DeMixSC to existing deconvolution methods.

A, Workflow for the deconvolution benchmarking design. We use benchmark data from retinal samples. The cell count proportions for each cell type are used as ground truth for the corresponding tissue samples. We assess the deconvolution performance of DeMixSC and seven existing methods for both bulk and pseudo-bulk mixtures. In addition to the raw counts, we also test RPM, RPKM, and TPM. The deconvolution performance is assessed by RMSE and MAE. **B** and **C**, Boxplots showing the deconvolution performance of eight deconvolution methods for the bulk and pseudo-bulk data. RMSE and MAE values are calculated across seven major cell types for each sample, with gray denoting pseudo-bulk and red denoting real-bulk. Smaller values indicate higher accuracy in proportion estimation. **D,** Boxplots showing the distributions of deconvolution estimates at the cell-type level for all 24 retinal samples. Each color corresponds to a given deconvolution method, with black denoting the ground truth, and each panel corresponds to a given cell type. **E** and **F**, An overview of deconvolution performance at the cell-type level across the eight methods using RMSE and MAE, respectively. Lighter colors correspond to lower RMSE or MAE values.

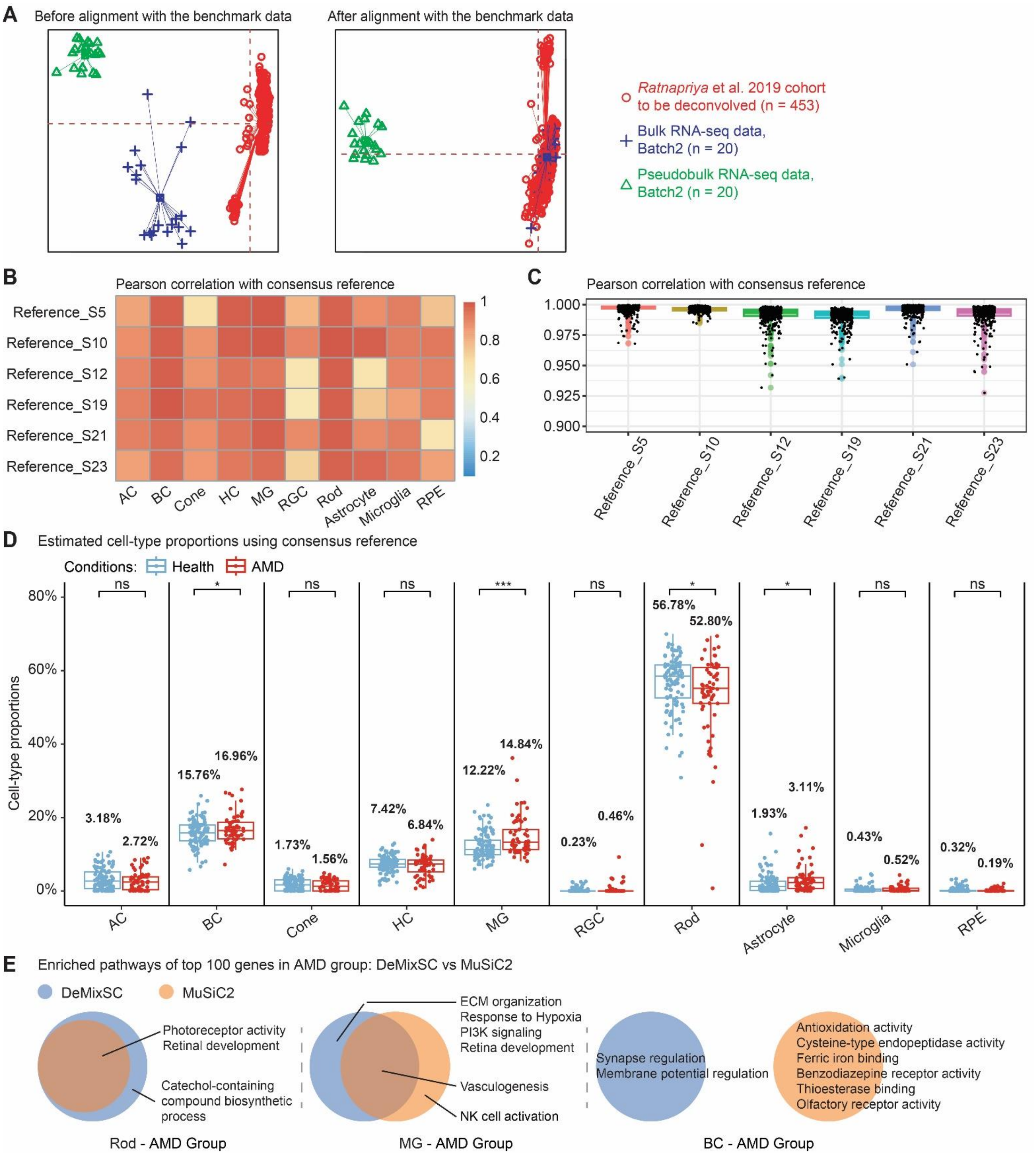
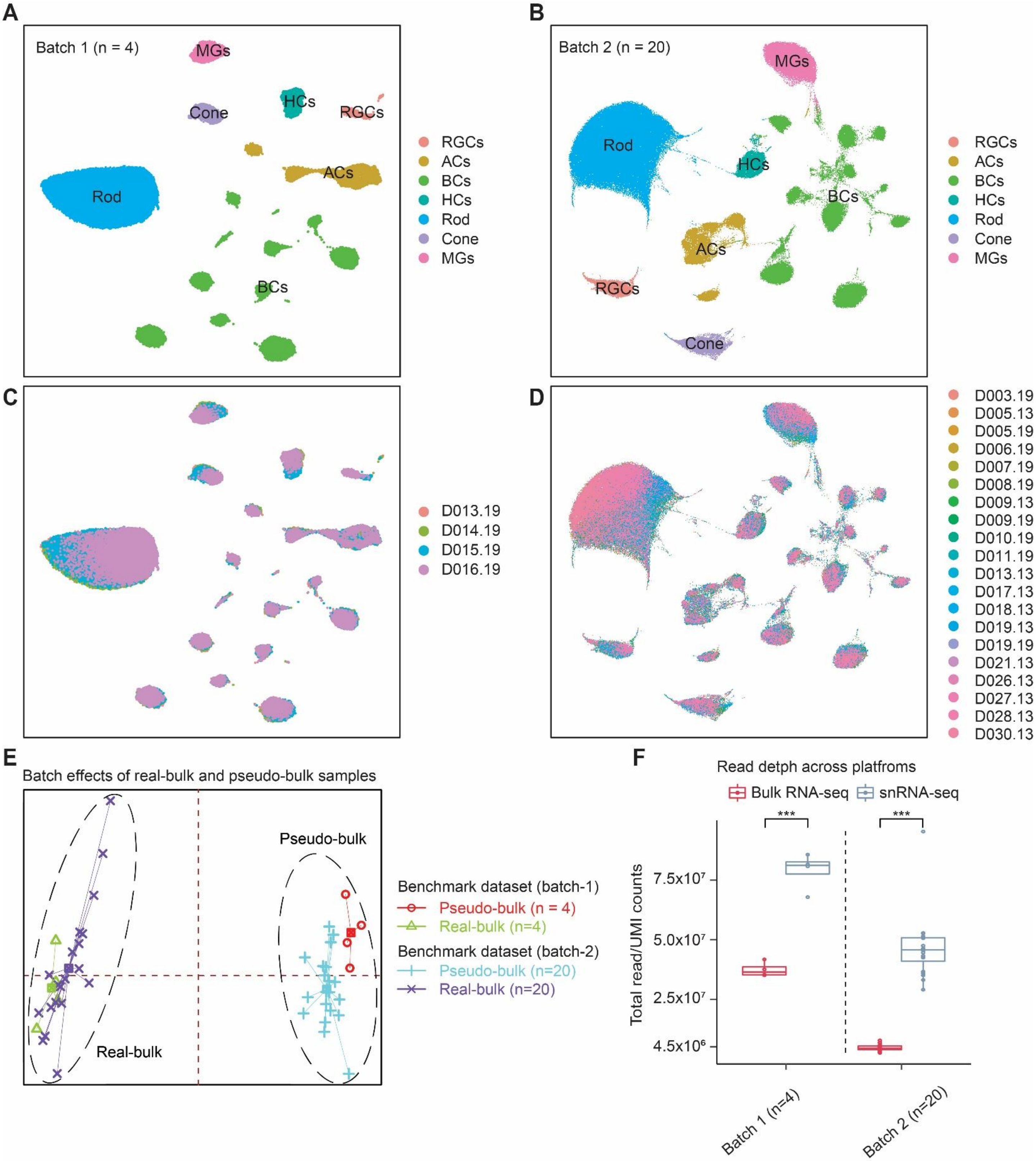
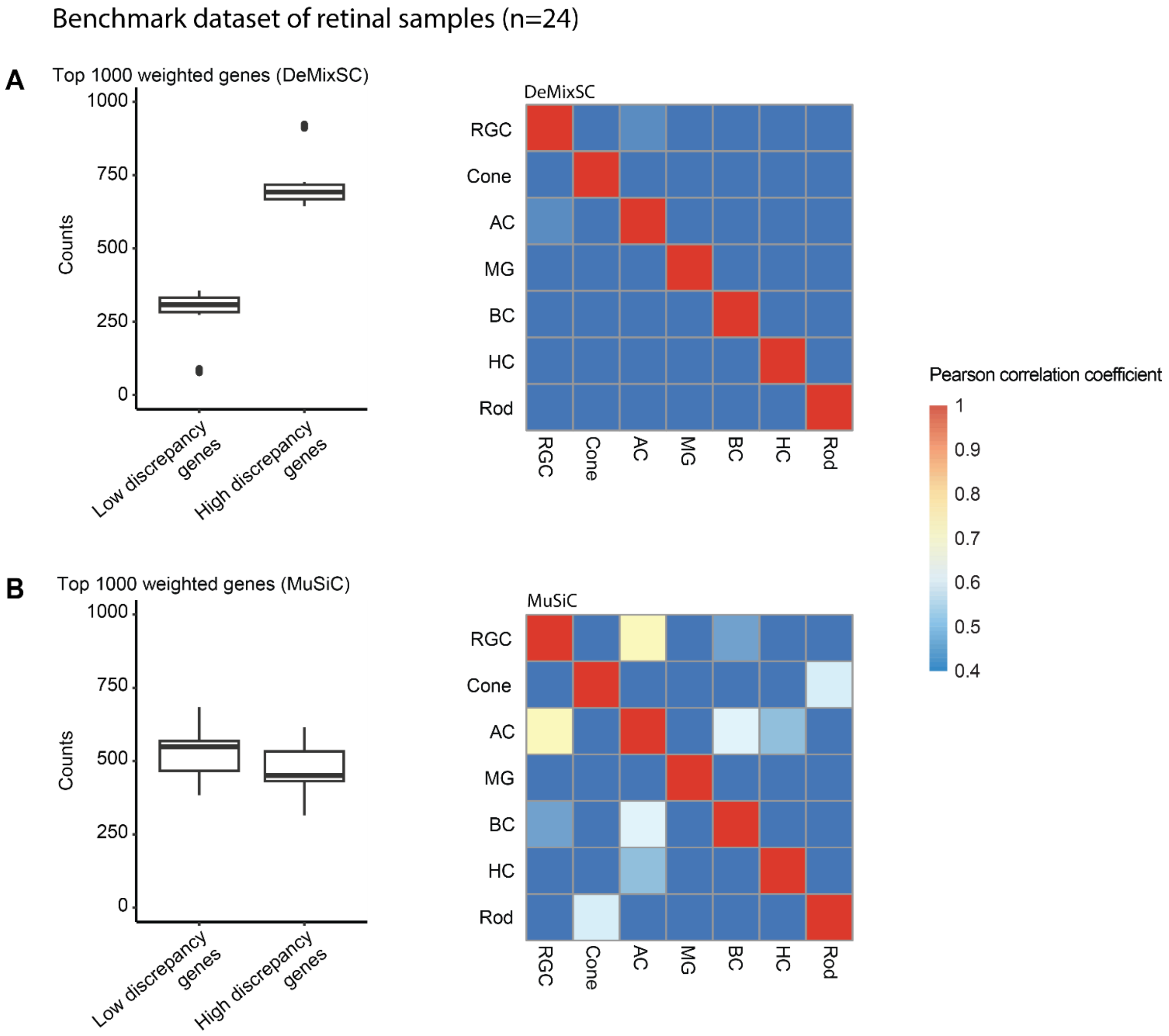


Figure 4 | Using DeMixSC to deconvolve a large cohort of human peripheral retinal samples.

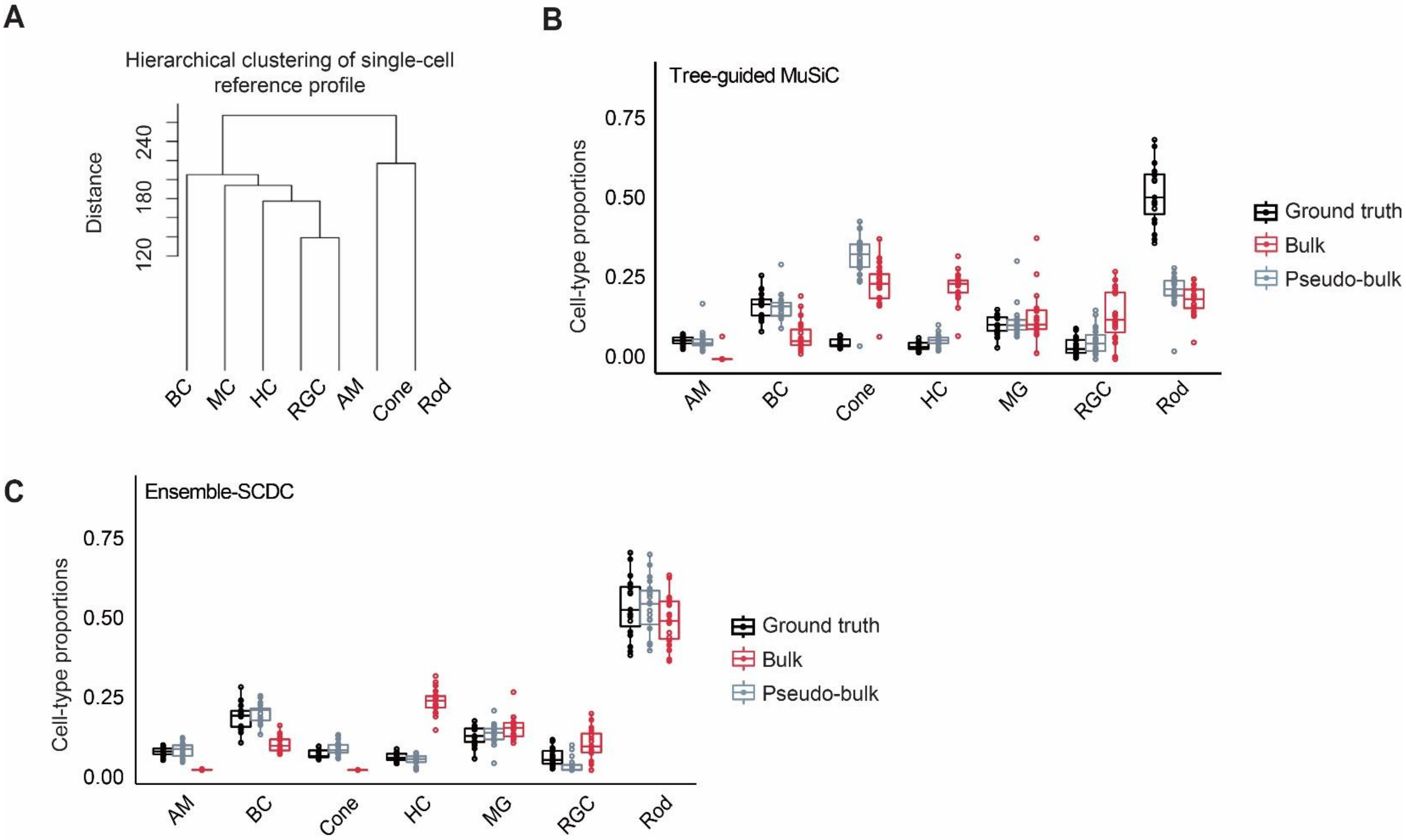
A, PCA plots of both the retina cohort data and the benchmark data. Red denotes the bulk data to be deconvolved, blue denotes the benchmark bulk data, and green denotes the benchmark pseudo-bulk data. **B** and **C** demonstrate the robustness of DeMixSC to different reference matrices at both cell-type and sample levels. Higher correlation coefficients indicate better performance. **D,** Distributions of DeMixSC estimated cell-type proportions of *Ratnapriya et al.* data using consensus references. Each panel corresponds to a given cell type. The *P*-values for Student's t-tests comparing the estimated cell-type proportions between non-AMD and AMD groups are denoted as follows: not significant (ns), *P*-value >0.05; **P*-value ≤0.05; ***P*-value ≤0.01; and ****P*-value ≤0.001. **E**, Venn diagram showing the GO annotation pathways identified using KERIS output based on DeMixSC or MuSiC2, for Rod, MGs, and BCs in the AMD samples.



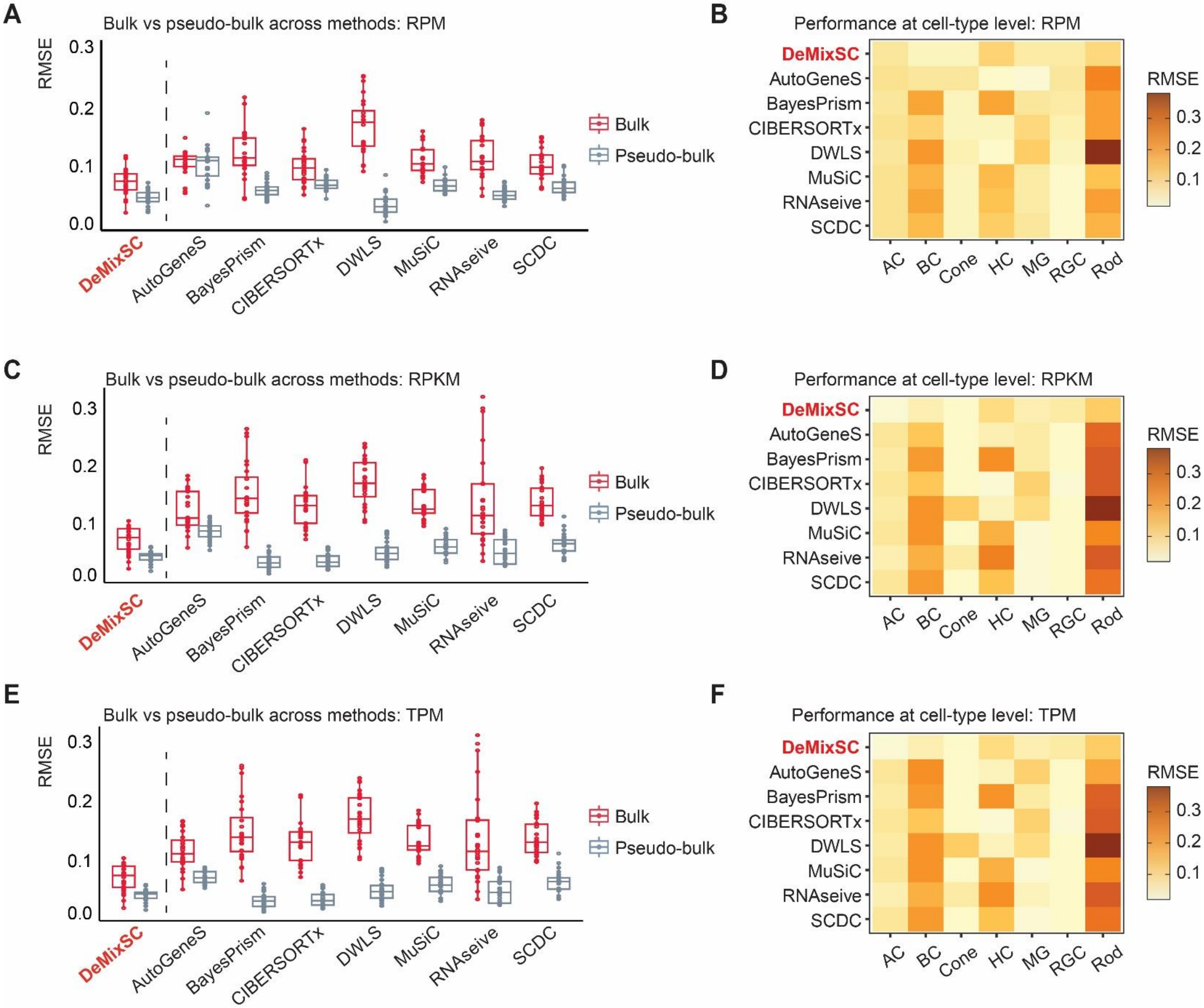
Extended Data Figure 1 | Overview of the matched bulk and snRNA-seq data.
A and **B**, UMAP projection of snRNA-seq data from 4 healthy retinal samples in batch-1 and 20 healthy retinal samples in batch-2, annotated by cell types. **C** and **D,** UMAP projection of snRNA-seq data from 4 healthy retinal samples in batch-1 and 20 healthy retinal samples in batch-2, annotated by sample IDs. Cells were clustered by their biological annotations instead of sample origins, suggesting weak-to-none batch effects. **E**, Distribution of the first two principal components for the matched real-bulk and pseudo-bulk RNA-seq data in the benchmark dataset. **F**, Boxplot showing the raw read depth between bulk and pseudo-bulk RNA-seq data from batch-1 and batch-2. The *P* values for Wilcoxon rank-sum tests comparing sequencing read depth between bulk and pseudo-bulk data are denoted as follows: **P*-value ≤0.05; ***P*-value ≤0.01; and ****P*-value ≤0.001.



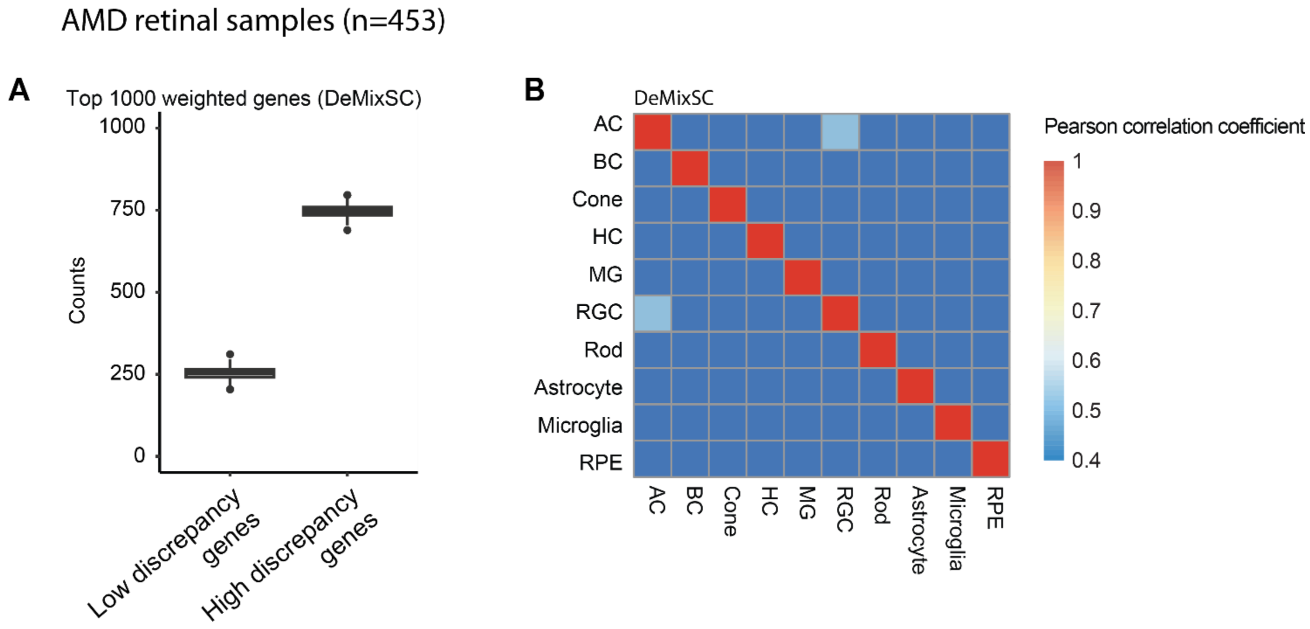
Extended Data Figure 2 | DeMixSC reduces the collinearity among top-weighted genes in the benchmark dataset. **A** and **B**, Boxplots showing numbers of low and high discrepancy genes within the top 1000 weighted genes for each sample, and heatmaps showing the averaged cell-type-wise collinearity across samples: DeMixSC **(A)** and MuSiC **(B)**.



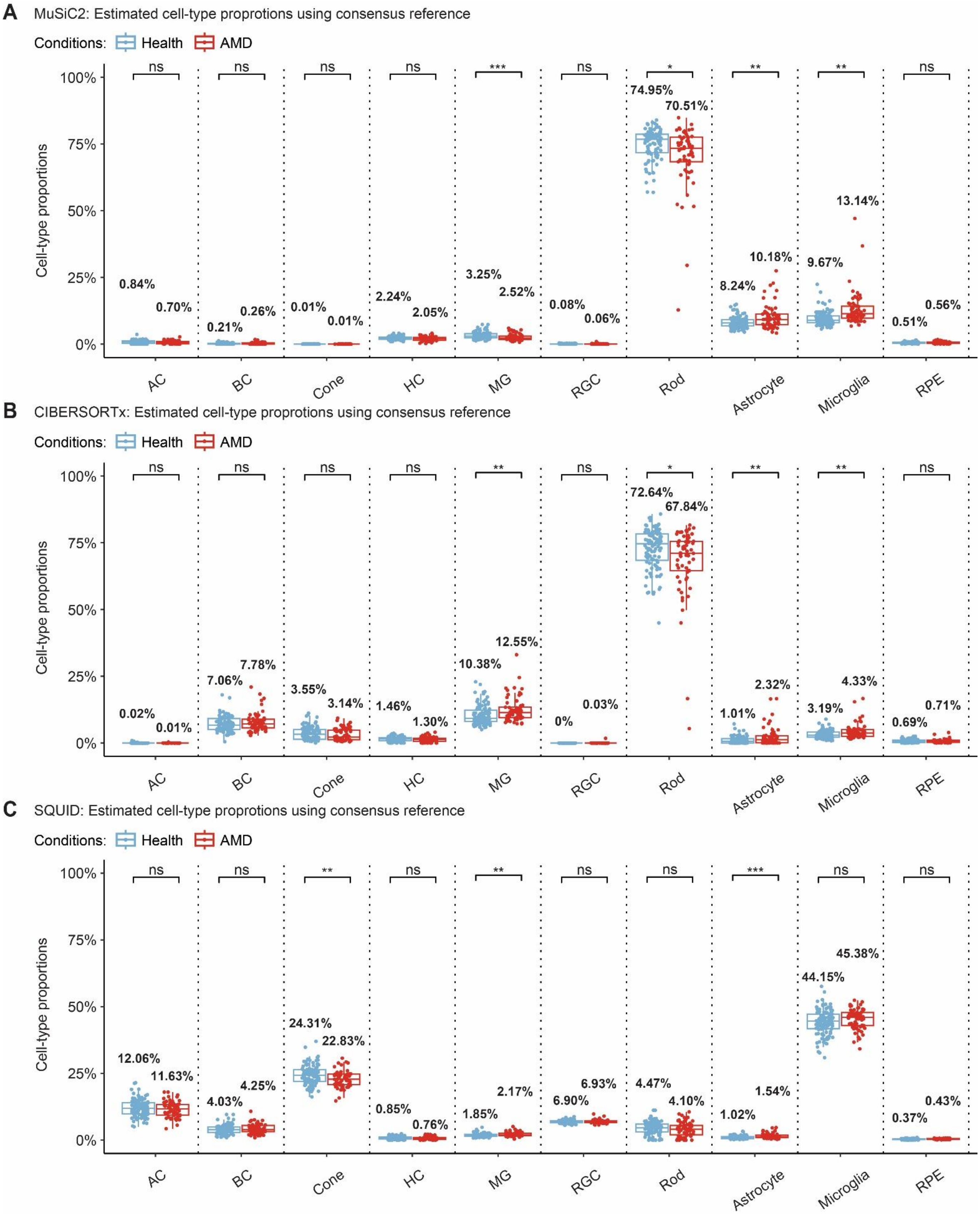
Extended Data Figure 3 | Deconvolution performance of the tree-guided MuSiC and Ensemble-SCDC. **A**, Hierarchical clustering of the cell-type-specific reference matrix. **B**, Boxplot showing the distributions of estimated cell-type proportions from the benchmark data using the tree-guided MuSiC. **C**, Boxplot showing the distributions of estimated cell-type proportions from the benchmark data using the SCDC ensemble mode. Black denotes the ground truth estimated using the snRNA-seq data. Gray denotes estimates from the pseudo-bulk RNA-seq data, and red denotes estimates from the matched bulk RNA-seq data.



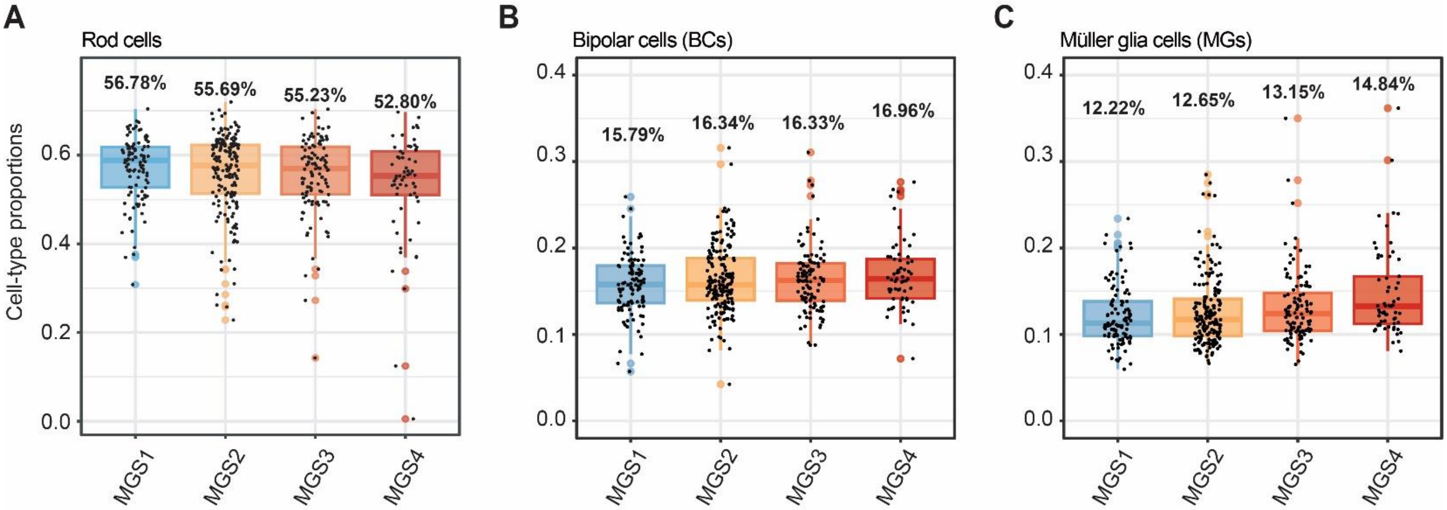
Extended Data Figure 4 | Impact of data normalization on the deconvolution performance.
A, C, and **E**, Boxplots showing the deconvolution performance across DeMixSC and seven current single-cell-based deconvolution methods for bulk and pseudo-bulk mixtures. RMSE values are calculated across seven major cell types for each sample, with gray denoting pseudo-bulk and red denoting real-bulk. Smaller values indicate higher accuracy in proportion estimation. **B, D,** and **F**, Heatmaps showing the deconvolution performance at the cell-type level across the eight methods using RMSE. Lighter colors correspond to lower RMSE values. Each panel corresponds to a normalization strategy: RPM (**B**), RPKM (**D**), and TPM (**F**).



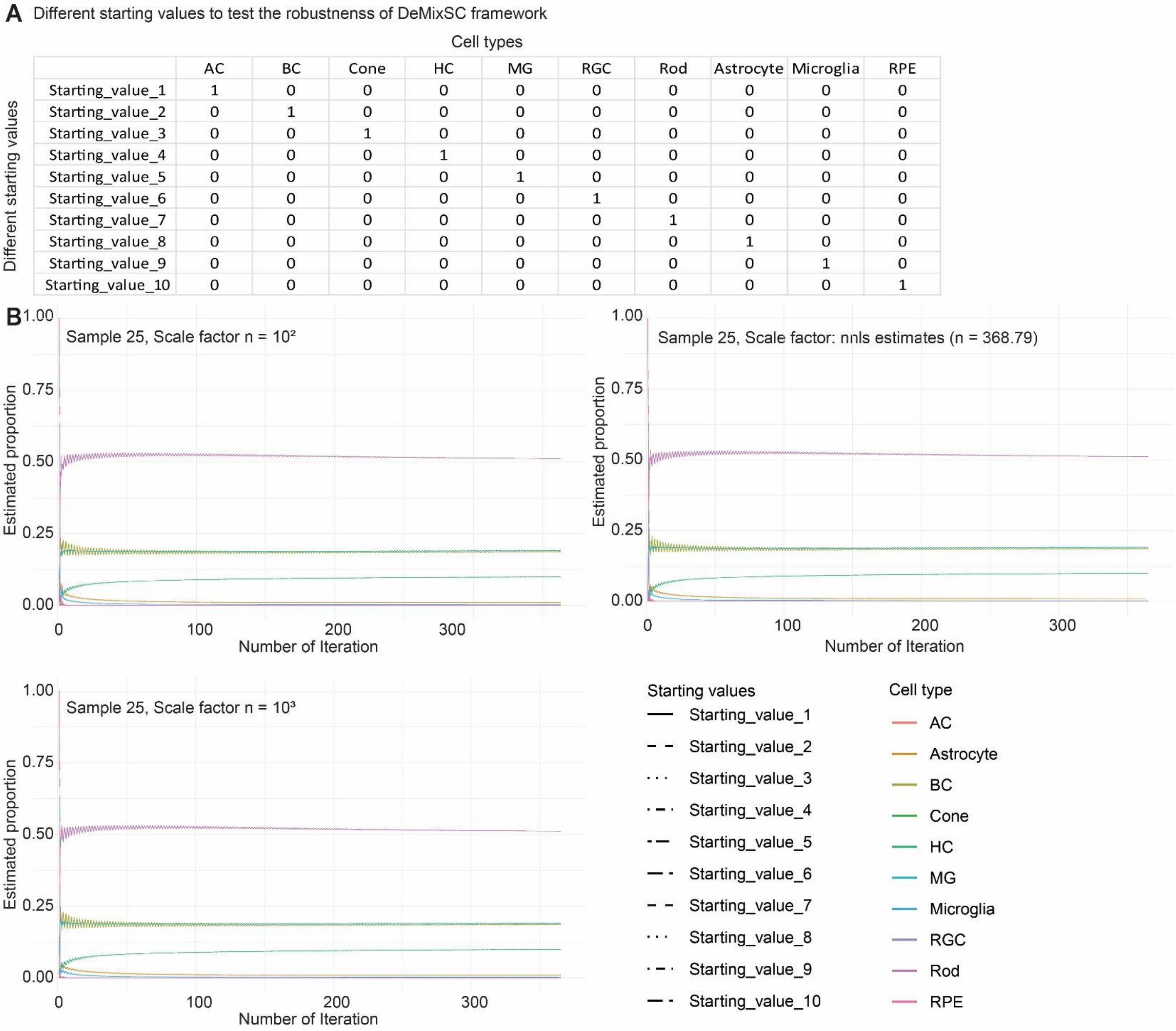
Extended Data Figure 5 | DeMixSC reduces the collinearity among top-weighted genes in the AMD cohort.
A, Boxplot showing numbers of low and high discrepancy genes within top 1000 weighted genes for each sample. **B**, heatmap showing the averaged cell-type-wise collinearity across samples.



Extended Data Figure 6 | Cell-type proportion estimates for the AMD cohort with existing methods. **A**, **B**, and **C**, Boxplots showing the distributions of cell-type proportion estimates for non-AMD retina vs. AMD retina from MuSiC2 (**A**), CIBERSORTx (**B**), SQUID (**C**). The *P* values for Student's t-tests comparing the estimated cell-type proportions between non-AMD and AMD groups are denoted as follows: not significant (ns), *P*-value >0.05; **P*-value ≤0.05; ***P*-value ≤0.01; and ****P*-value ≤0.001.



Extended Data Figure 7 | DeMixSC recovers a dynamic shift in cell-type proportions during the AMD progression. **A**, **B**, and **C**, Boxplots showing the distributions of cell-type proportion estimates across different MGS stages from MGS1 to MGS4. Each panel corresponds to a given cell type: Rod cells (**A**), Bipolar cells (**B**), and Müller glia cells (**C**).



A Different starting values to test the robustnenss of DeMixSC framework

Cell types

Different starting values	AC	BC	Cone	HC	MG	RGC	Rod	Astrocyte	Microglia	RPE
Starting_value_1	1	0	0	0	0	0	0	0	0	0
Starting_value_2	0	1	0	0	0	0	0	0	0	0
Starting_value_3	0	0	1	0	0	0	0	0	0	0
Starting_value_4	0	0	0	1	0	0	0	0	0	0
Starting_value_5	0	0	0	0	1	0	0	0	0	0
Starting_value_6	0	0	0	0	0	1	0	0	0	0
Starting_value_7	0	0	0	0	0	0	1	0	0	0
Starting_value_8	0	0	0	0	0	0	0	1	0	0
Starting_value_9	0	0	0	0	0	0	0	0	1	0
Starting_value_10	0	0	0	0	0	0	0	0	0	1

Extended Data Figure 8 | Convergence of DeMixSC with different starting values.
A, A list of different starting values across ten cell types. **B**, Trace plots of estimated proportions over iterations.