A deconvolution framework that uses single-cell sequencing plus a small benchmark dataset for accurate analysis of cell type ratios in complex tissue samples

Shuai Guo, Xiaoqian Liu, Xuesen Cheng, et al.

Genome Res. published online November 25, 2024
Access the most recent version at doi:10.1101/gr.278822.123

A deconvolution framework that uses single-cell sequencing plus a small benchmark dataset for accurate analysis of cell type ratios in complex tissue samples

Shuai Guo[1,14], Xiaoqian Liu[1,13,14], Xuesen Cheng[2,14], 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], Sanghoon Lee[9], Anil K. Sood[9], Scott Kopetz[10], John Paul Shen[10], John N. Weinstein[1,11], Margaret M. DeAngelis[4,5,6,12], Rui Chen[2,15], Wenyi Wang[1,15*]

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 Gynecologic Oncology and Reproductive Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
10. Department of Gastrointestinal Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.
11. Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
12. VA Western New York Healthcare System, Buffalo, NY, USA.
13. Current affiliation: Department of Statistics, University of California at Riverside, Riverside, CA, USA.
14. Authors contributed equally.
15. Authors contributed equally.

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

Running title: DeMixSC enhances deconvolution with benchmark data

Abstract

Bulk deconvolution with single-cell/nucleus RNA-seq data is critical for understanding heterogeneity in complex biological samples, yet the technological discrepancy across sequencing platforms limits deconvolution accuracy. To address this, we utilize an experimental design to match inter-platform biological signals, hence revealing the technological discrepancy, and then develop a deconvolution framework called DeMixSC using this well-matched, i.e., benchmark data. Built upon a novel weighted nonnegative least-squares framework, DeMixSC identifies and adjusts genes with high technological discrepancy and aligns the benchmark data with large patient cohorts of matched-tissue-type for large-scale deconvolution. Our results using two benchmark datasets of healthy retinas and ovarian cancer tissues suggest much-improved deconvolution accuracy. Leveraging tissue-specific benchmark datasets, we applied DeMixSC to a large cohort of 453 age-related macular degeneration patients and a cohort of 30 ovarian cancer patients with various responses to neoadjuvant chemotherapy. Only DeMixSC successfully unveiled biologically meaningful differences across patient groups, demonstrating its broad applicability in diverse real-world clinical scenarios. Our findings reveal the impact of technological discrepancy on deconvolution performance and underscore the importance of a well-matched dataset to resolve this challenge. The developed DeMixSC framework is generally applicable for accurately deconvolving large cohorts of disease tissues, including cancers, when a well-matched benchmark dataset is available.

Keyword:

Transcriptomic deconvolution, technological discrepancy, single-cell/nucleus RNA sequencing, bulk RNA sequencing, retina, age-related macular degeneration, high-grade serous ovarian cancer.

Introduction

Although recent advances in single-cell/nucleus RNA sequencing (sc/snRNA-seq) offer valuable insights into cell types and states in healthy (Haniffa et al., 2021) and diseased tissues (Gohil et al., 2020; Zeng et al., 2023), high expense and complex sample preparation procedures have restricted its widespread adoption in clinical settings (Li and Wang, 2021). Bulk RNA-seq, on the other hand, retains its essential role, especially in large disease-based cohort studies, for which 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 (Cao et al., 2022; Ratnapriya et al., 2019; Stark et al., 2019). Nevertheless, bulk RNA-seq comes with a significant drawback: it captures averaged gene expression across heterogeneous cell types, thus confounding downstream analysis (Li and Wang, 2021). 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 (Anghel et al., 2015; Wang et al., 2018) employ bulk RNA-seq data from normal tissues or cell lines as the reference. They were 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 (Aliee and Theis, 2021; Chu et al., 2022; Cobos et al., 2023; Dong et al., 2021; Erdmann-Pham et al., 2021; Fan et al., 2022; Newman et al., 2019; Tsoucas et al., 2019; Wang et al., 2019), 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 precision and accuracy. The limitations arise from inconsistencies in gene expression profiles between bulk and sc/snRNA-seq data. Those inconsistencies are attributable to technique variations in sample acquisition, preparation, and sequencing (Aird et al., 2011; Denisenko et al., 2020; Hippen et al., 2023; Stoler and Nekrutenko, 2021; Wery et al., 2013). 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 for deconvolving publicly available large bulk cohorts (Fan et al., 2022; Jin and Liu, 2021; Sturm et al., 2019). Most existing benchmarking designs (Cobos et al., 2020; Jin and Liu, 2021; Sturm et al., 2019) often employ 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 discrepancy. Researchers have become aware of these discrepancies (Cobos et al., 2023; Dietrich et al., 2022; Hippen et al., 2023; Sutton et al., 2022). A recent study (Hippen et al., 2023) generates matched bulk and scRNA-seq data from seven high-grade serous ovarian cancer (HGSC) samples as a benchmark and discusses the impact of technological discrepancy on deconvolution analysis. However, current attempts to address these issues have achieved only limited success (Cobos et al., 2023; Dong et al., 2021; Newman et al., 2019). CIBERSORTx (Newman et al., 2019) implements a batch effect correction step but offers limited improvements in deconvolving complex bulk tissues. The ensemble approach of SCDC (Dong et al., 2021) uses matched bulk and scRNA-seq data from two normal tissue samples (e.g., mouse breast) but lacks generalizability to patient cohorts. The most recent SQUID (Cobos et al.,

2023) builds on top of DWLS (Tsoucas et al., 2019) with a Bisque-based linear transformation step(Jew et al., 2020) to align matched bulk and scRNA-seq data; it can distort gene expression profiles, risking overcorrection. Therefore, there is still need for methods to effectively mitigate such discrepancy by taking full advantage of the well-matched benchmark dataset.

In this paper, we offer a new solution to improve deconvolution performance. To accomplish this, we generate a specialized benchmark dataset of 24 healthy retinal samples, ensuring technological discrepancy as the main confounding factor. Using this dataset, we demonstrate that technological discrepancy significantly affects the expression profiles of bulk and single-nucleus data and thus reduces 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 (Ruppert and Wand, 1994) to identify and adjust for genes consistently affected by technological discrepancy. DeMixSC is generalizable to any tissue type, given a small representative benchmark dataset, to effectively deconvolve a large tissue-type-matched bulk cohort. We validated the improved deconvolution performance of DeMixSC by comparing it on our benchmark dataset with eight existing deconvolution methods. When applied to 453 peripheral retinal samples from patients with age-related macular degeneration (AMD) (Ratnapriya et al., 2019), DeMixSC achieved more realistic cell type estimates that reflect subtle changes in cell type proportions among AMD grades, suggesting its reliability and generalizability in real-world settings. Notably, DeMixSC exhibited superior deconvolution performance on an HGSC cohort (Lee et al., 2020) by employing the available HGSC benchmark dataset (Hippen et al., 2023). DeMixSC accurately captured the proportional differences associated with different treatment responses and identified a trend of increased macrophage infiltration linked to poorer treatment outcomes. This further highlights the capability and generalizability of DeMixSC in deconvolving highly heterogeneous tumor samples, an intractable challenge for existing deconvolution methods. In summary, DeMixSC fills the gap in resolving the technological discrepancy in bulk deconvolution and serves as an accurate and adaptable tool for estimating cell type proportions.

Results

Use benchmark data to assess technological discrepancy

We designed and generated a specialized benchmark dataset to assess the technological discrepancy between bulk and sc/sn sequencing platforms (Fig. 1A and Supplemental Fig. S1). This dataset comprises 24 healthy retinal samples from donors' eyes collected within six hours postmortem (ages of death between 53 and 91, Supplemental Table S1), for two batches of sequencing experiments. Both bulk and snRNA-seq profiling were performed on each sample from the same single-nucleus suspension aliquot using a template-switching method to generate full-length cDNA libraries (see Methods). Because single-cell protocols can be biased toward retaining certain cell types (Mereu et al., 2020), hence changing the cell type proportions, this special approach maximizes our chance that the matched sequencing data shares approximately the same cell type proportions. We performed cell type annotation for snRNA-seq data with known markers (see Methods, Supplemental Table S2). The resulting snRNA-seq data was summed to create matched pseudo-bulk RNA-seq data (see Methods). We hypothesized that any major differences in gene expression profile between the matched pseudo-bulk and real bulk RNA-seq would be technological, rather than due to biological discrepancies.

We observed much larger batch differences between real-bulk and pseudo-bulk data, than the small differences in cell type distributions across samples in snRNA-seq or differences between the two experimental batches (Supplemental Fig. S2A-E). Total read counts from bulk RNA-seq data were significantly lower than total UMI counts from matched pseudo-bulk data (Supplemental Fig. S2F). Assuming that the difference in read depth does not impact the relative expression of each gene, we expected gene expression correlation to be a better metric for identifying technological discrepancy. We observed a low-to-moderate correlation of gene expression, consistent across samples, between the paired bulk datasets (Fig. 1B, mean Spearman's correlation coefficient = 0.31 for batch-1 and 0.41 for batch-2). Further differential expression (DE) analysis between the paired bulk and pseudo-bulk samples identified more than 5,000 DE genes in each experimental batch (Fig. 1C, adjusted *P*-values < 0.05), with more than 60% of those genes overlapping across the experiments (Fig. 1D and Supplemental Fig. S3). We next converted the retina bulk data to transcripts per million (TPM) to account for gene length effects, considering the incomplete gene coverage from the 10x single-cell platforms, and observed even lower correlations with paired pseudo-bulk data (mean Spearman's correlation coefficient = 0.18 for batch-1 and 0.25 for batch-2), indicating that TPM normalization did not ameliorate these discrepancies. We further analyzed a benchmark dataset from seven primary high-grade serous ovarian cancer (HGSC) samples (Hippen et al., 2023) with matched single-cell and three types of bulk data: dissociation with poly(A) enrichment (Disso&poly(A)+), dissociation with rRNA depletion (Disso&rRNA-), and tissue chunk with rRNA depletion (Chunk&rRNA-) (see Methods, Supplemental Fig. S4A). The HGSC benchmark dataset exhibited significant technological discrepancy (more than 5,000 DE genes) between bulk and pseudo-bulk RNA-seq data, with consistent DE patterns across seven samples (Supplemental Fig. S4B, C).

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 technological discrepancy (Denisenko et al., 2020; Stoler and Nekrutenko, 2021; Tung et al., 2017). We, therefore, expect that the reference matrices derived from sc/snRNA-seq data will

not fully represent cell type-specific expression profiles in bulk samples (Cobos et al., 2023; Hippen et al., 2023). Given such discrepancies, the performance of existing deconvolution methods are compromised, as their key assumption about the representative reference is violated.

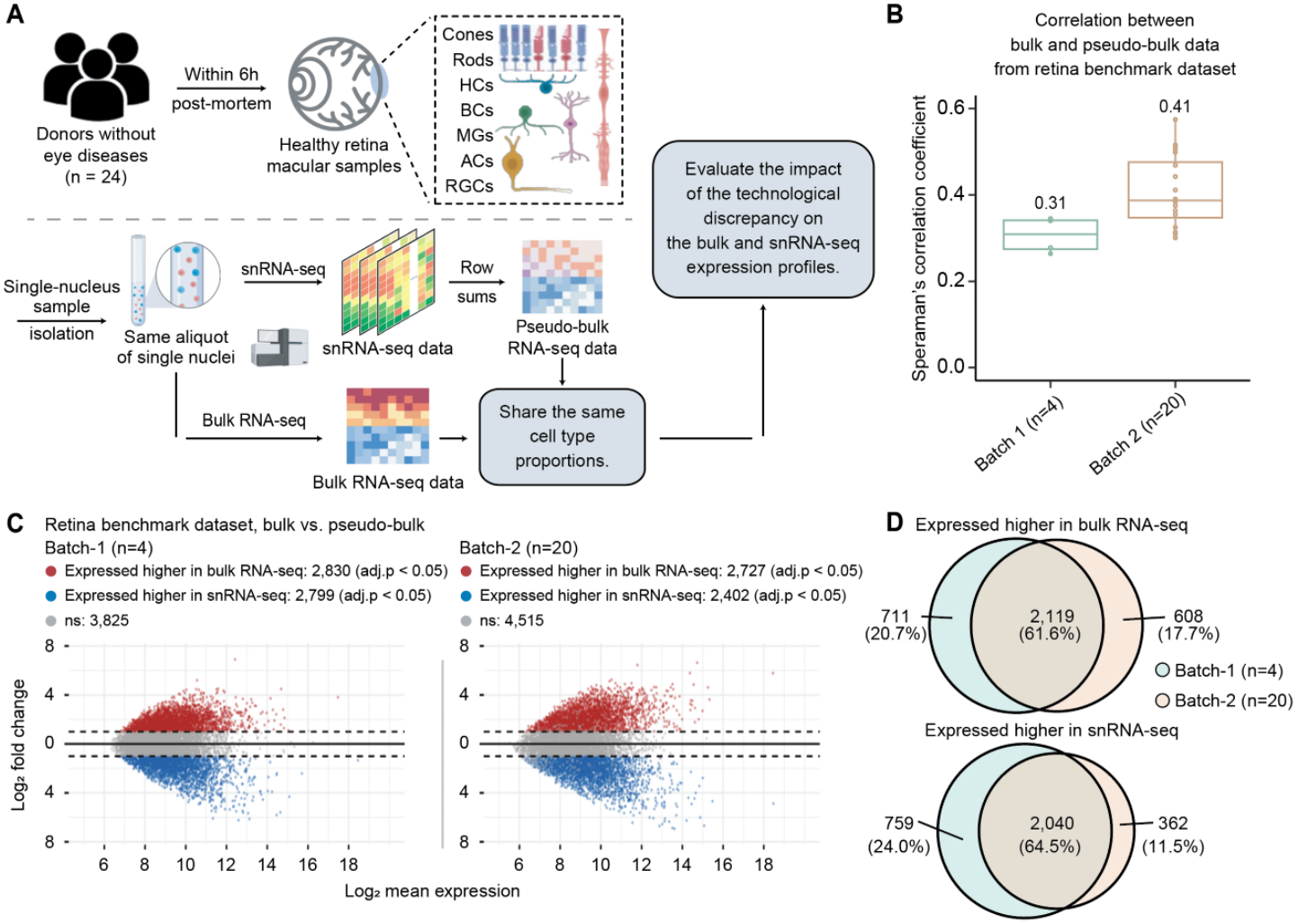

Figure 1. Assessing technological discrepancy between bulk and single-cell sequencing platforms using matched single-nucleus aliquots.

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 cells (RPEs). Samples are isolated into single-nucleus suspensions. The same aliquot of single-nucleus 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 discrepancy (i.e., the shot-gun sequencing procedure) on the bulk and snRNA-seq expression profiles. **B** and **C** show the influence of technological discrepancy at the sample and gene level, respectively. **B**, Spearman's correlation coefficient across genes between the matched real-bulk and pseudo-bulk RNA-seq data for one sample at a time for both batches. The correlations were calculated using quantile-normalized expression data (relative abundances). **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 2-fold change 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.

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 (Ruppert and Wand, 1994; Tsoucas et al., 2019; Wang et al., 2019) with several essential improvements (see Methods and Supplementary Note). 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,

$$\hat{\boldsymbol{p}}_j=\underset{\boldsymbol{p}_j\geq 0}{\operatorname{argmin}}\left(\sum_{g\in G_1} w_{jg}\left(y_{jg}-n_j\sum_{k\in K}p_j^k\hat{r}_{jg}^k\right)^2+\sum_{g\in G_2} w_{jg}\left(\frac{y_{jg}}{a}-n_j\sum_{k\in K}p_j^k\frac{\hat{r}_{jg}^k}{a}\right)^2\right).$$

Here, y_{jg} is the observed expression value of gene g from subject j in bulk RNA-seq data, n_j is the number of all cells, $\hat{r}_{jg}^k$ 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 an associated 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 a gene partitioning approach, which identifies and adjusts the expression levels of genes that exhibit consistently high technological discrepancy (G_2). We do so with 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 (G_1) from those highly affected by technological discrepancy (G_2) (see Methods). It then employs a partitioned loss function and adjusts genes from G_2 by rescaling their expressions by a positive constant adjustment factor a to mitigate the influence of technological discrepancy (see Methods and Supplementary Note).

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

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

where $\hat{y}_{jg}$ denotes the fitted expression value of gene g in subject j. This weight function comprises three terms: the squared fitted expression, the squared residual, and a baseline constant, which is distinct from previously proposed weights (Cobos et al., 2023; Dong et al., 2021; Fan et al., 2022; Tsoucas et al., 2019; Wang et al., 2019). The fitted term addresses genes with high expression levels, the squared residual accounts for the remaining variance after fitting, and the baseline constant c adds a reasonable upper bound on the weight (see Methods and Supplementary Note). These two innovations enable DeMixSC to more effectively address the technological discrepancy compared to non-differential weighting approaches, e.g., test statistics (see Methods).

DeMixSC runs as a three-tier model in application. First, DeMixSC uses a specifically designed benchmark dataset to identify and adjust genes with high inter-platform discrepancy (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 (Zhang et al., 2020) (Fig. 2B) to generalize the technological discrepancy detected. Last, DeMixSC runs the refined wNNLS framework iteratively for deconvolution (Fig. 2C), allowing for dynamic updates as the model fit improves and progressively enhancing estimation accuracy. A diagram (Supplemental Fig. S5) complementary to Fig. 2 visualizes the complete workflow of DeMixSC with

more technical details. Our main prerequisite is a matched tissue type between the small benchmark dataset and the large, targeted cohort.

DeMixSC includes a quantile normalization step and a batch effect correction step, both of which operate under specific assumptions. Quantile normalization assumes symmetric differential expression between conditions and similar gene expression distributions across samples. Batch effect correction requires the bulk benchmark data to share similar tissue microenvironment with the large cohort. We note that DeMixSC is compatible with any batch effect correction method, yet effective batch effect correction is essential for its success.

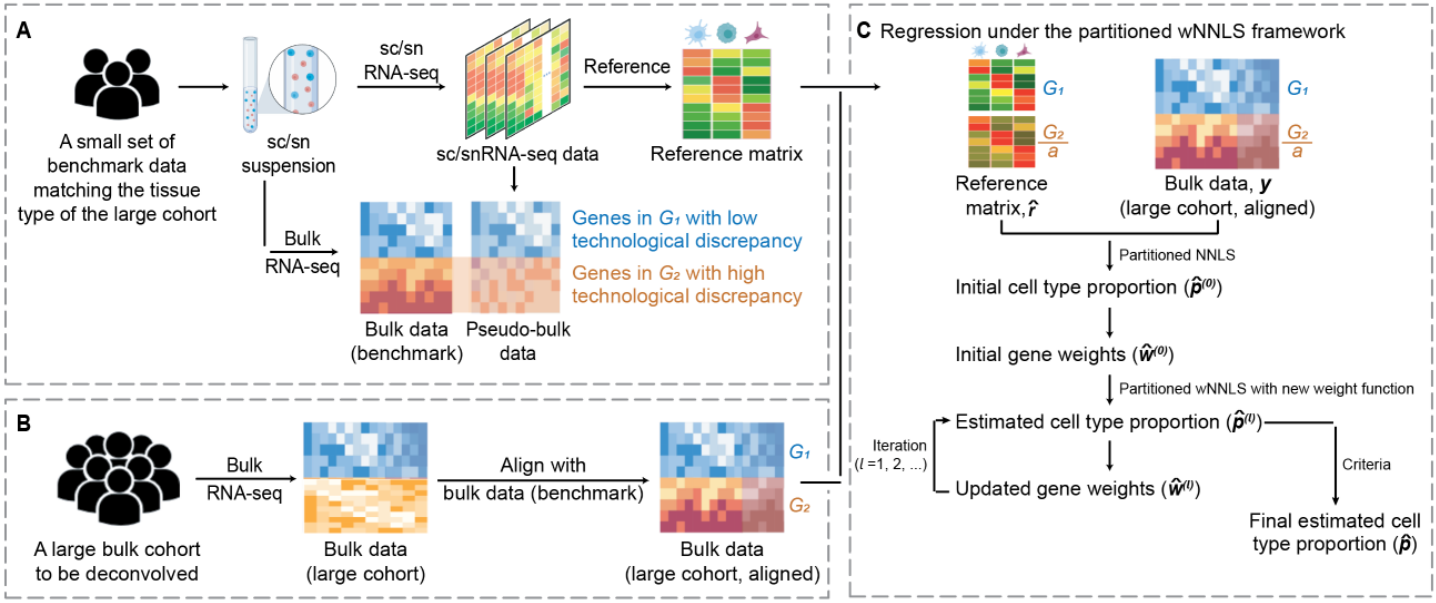

Figure 2. Overview of DeMixSC.
The DeMixSC framework for deconvolution analysis of bulk RNA-seq data using sc/sn RNA-seq data as a reference. **A,** The framework starts with a benchmark dataset of matched bulk and sc/snRNA-seq data with the same cell type proportions. Pseudo-bulk mixtures are generated from the sc/sn data. DeMixSC identifies genes in G_1 and G_2 with 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 more impacted by the technological discrepancy (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 under a weighted nonnegative least squares (wNNLS) framework with two improvements: 1) partitioning and adjusting genes with high technological discrepancy, and 2) a new weight function. The final estimates are obtained when either the algorithm converges or reaches the prespecified maximum number of iterations. Here, G_1 is genes with low technological discrepancy, G_2 is genes with high technological discrepancy, a is a user-defined positive constant that serves as an adjustment factor, $\hat{\boldsymbol{r}}$ is the reference matrix derived from the sc/snRNA-seq data, $\boldsymbol{y}$ is the observed expression in bulk RNA-seq data, $\hat{\boldsymbol{p}}$ is the vector of estimated cell type proportions, and $\hat{\boldsymbol{w}}$ is the estimated gene weights.

Comparing the estimation accuracy of DeMixSC with that of other existing deconvolution methods

Using our retina benchmark data, we compared the performance of DeMixSC with that of eight existing deconvolution methods (Aliee and Theis, 2021; Chu et al., 2022; Cobos et al., 2023; Dong et al., 2021; Erdmann-Pham et al., 2021; Newman et al., 2019; Tsoucas et al., 2019; Wang et al., 2019): AutoGeneS, BayesPrism, CIBERSORTx, DWLS, MuSiC, RNAseive, SCDC, and SQUID (see Methods, Fig. 3A). The retinal tissue samples in our benchmark dataset comprised ten distinct cell types. We focused 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 accounted for 98% of the total cell population (Liang et al., 2023).

Overall, DeMixSC achieved the lowest root mean squared error (RMSE) and the highest Spearman's correlation coefficient in deconvolving bulk RNA-seq data, with mean values of 0.03 and 0.86 (Fig. 3B, C). Moreover, DeMixSC produced comparable RMSEs and Spearman's correlation coefficients for deconvolving bulk and pseudo-bulk RNA-seq data (mean RMSE: bulk 0.03, pseudo-bulk 0.03; and mean Spearman's correlation: bulk 0.86, pseudo-bulk 0.92). Those results suggested that DeMixSC adjusts well to undesired technological discrepancies. In contrast, existing methods performed reasonably well for pseudo-bulk but much poorly for bulk data. In that sense, technological discrepancies that compromise deconvolution accuracy remained unaddressed by other existing approaches (Fig. 3B, C). Specifically, AutoGeneS showed higher RMSE for pseudo-bulk data, likely due to its inability to distinguish between Rod and Cone cells, which share largely similar expression profiles (Fig. 3D). DWLS excelled in deconvolving pseudo-bulk samples but falls short for bulk RNA-seq data, possibly due to overfitting. Using the tree-based deconvolution in MuSiC or the ensemble option in SCDC did not improve their accuracy (Supplemental Fig. S6). CIBERSORTx presented 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 observed systematic biases across other methods. Most methods underestimate the proportions of ACs, BCs, and Cones while overestimating HCs and Rods (Fig. 3D and Supplemental Fig. S7). DeMixSC accurately estimated the proportions of all seven major cell types and improves the deconvolution results for ACs, BCs, Cones, HCs, and MGs (Fig. 3D, E; mean RMSE: 0.01, 0.04, 0.03, 0.02, 0.03, respectively). DeMixSC also performed better in correlations of the estimated versus the true cell proportions, particularly for the top three prevalent cell types (Fig. 3F; Rods, MGs, and BCs, Spearman's correlation coefficients of 0.78, 0.73, and 0.58, respectively).

In addition, we tested the robustness of these methods under varied data formats (Dillies et al., 2013), including RPM, RPKM, and TPM (see Methods) and found DeMixSC to be robust to data normalizations (Supplemental Fig. S8). In line with previous benchmarking studies (Cobos et al., 2020), we found using raw counts as input is sufficient to obtain good results. Finally, SQUID delivered the least desirable results in this benchmarking study (mean RMSE and Spearman's correlation in bulk data: 0.25 and 0.31). The issue with SQUID possibly lies in its data transformation step (Jew et al., 2020), which has the potential to misrepresent gene expression profiles. In summary, our DeMixSC framework has achieved the most accurate deconvolution among the compared methods by successfully addressing the key issues with the technological discrepancy between pairs of sequencing platforms. Regarding the required sample size for the benchmark dataset, we found that DeMixSC exhibited satisfying deconvolution performance with a sample size of four, and its performance becomes stable when the sample size is over seven in the retina data (Supplemental Fig. S9).

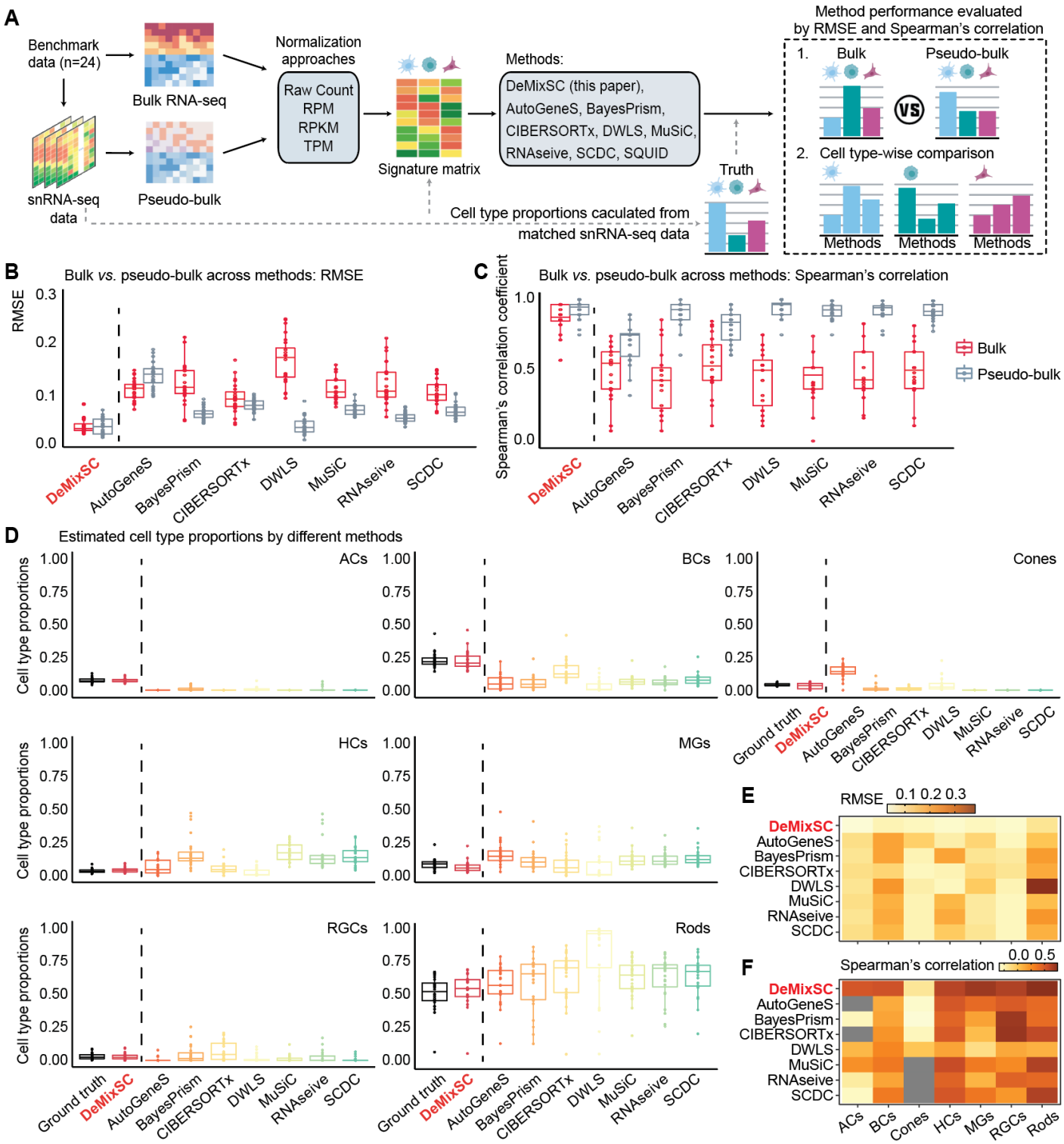

Figure 3. Compare the estimation accuracy of DeMixSC to existing deconvolution methods.
A, Workflow for the deconvolution benchmarking design. We use benchmark data from retinal samples. The cell count proportions for each cell type are used as ground truth for the corresponding tissue samples. We assess 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 Spearman's correlation coefficient. **B** and **C**, Boxplots showing the deconvolution performance of eight deconvolution methods for the bulk and pseudo-bulk data. RMSE and Spearman's correlation coefficient values are calculated across seven major cell types for each sample, with gray denoting pseudo-bulk and red denoting real-bulk. Smaller RMSEs or larger Spearman's correlations 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 Spearman's correlation coefficient, respectively. Lighter colors correspond to lower RMSE or Spearman's correlation coefficient values. Gray means NA.

Applying DeMixSC to human peripheral retina bulk RNA-seq data

Age-related macular degeneration (AMD) is characterized by deterioration of retina and choroid that leads to substantial decreased visual acuity, with loss of Cone and Rod cells as a major manifestation. It is the leading cause of blindness among the elderly population globally (Fleckenstein et al., 2021). However, the molecular and cellular events that underlie AMD remain poorly understood, impeding the development of effective treatments (Khanani et al., 2022). 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 retinas (Ratnapriya et al., 2019) (see Methods). Among these retina samples, 105 have been scored in the Minnesota Grading System as grade 1 (MGS1), 175 as MGS2, 112 as MGS3, and 61 as MGS4. An MGS1 rating indicates non-AMD healthy retina, and an MGS4 rating indicates AMD. MGS2 and MGS3 represent intermediate stages (Olsen and Feng, 2004).

We ran DeMixSC to first align the AMD cohort with the bulk data from our specialized benchmark dataset of retina samples, and then estimate cell type proportions in the AMD cohort (see Methods, Fig. 4A). For the reference matrix in wNNLS, we constructed a consensus reference by integrating expression profiles from seven single-nucleus samples (see Methods) to achieve reliable deconvolution. DeMixSC produced overall robust deconvolution estimates among the consensus and each individual single-nucleus references at both cell type and sample levels, only with low-to-moderate correlations observed in some conditions due to variations in the ranking of low-abundance cell types across samples (Fig. 4B, C). DeMixSC achieved cell type proportions that are closer to experimental measures for non-AMD samples (Liang et al., 2019), with mean RMSE of 0.04 and mean Spearman's correlation coefficient of 0.75 (see Methods, Supplemental Table S3). DeMixSC revealed changes in cell type proportions between non-AMD and AMD samples (Fig. 4D). We observed statistically significant decreases in photoreceptors, including Rod cells (*P*-value = 0.047) and Cone cells (*P*-value = 0.035), and HCs (*P*-value = 0.005). Besides, DeMixSC identified increases in glial cells, specifically Astrocytes (*P*-value = 0.006) and MGs (*P*-value = 0.002). The increase of BCs is of marginal significance (*P*-value = 0.068). These changes in cell type proportions showed consistent patterns across the progression of AMD severity from MGS1 to MGS4 (Supplemental Fig. S10), reflecting the progressive nature of the disease. For comparison, we deconvolved the same cohort with MuSiC2 (Fan et al., 2022), CIBERSORTx (Newman et al., 2019), and SQUID (Cobos et al., 2023), where MuSiC2 was chosen for its added ability to leverage conditionally stable genes from healthy references in analyzing diseased tissue. Among the four methods compared, DeMixSC exhibited the least bias for three out of seven major cell types (ACs, BCs, and Rod cells; Supplemental Table S3 and Supplemental Fig. S11). CIBERSORTx is the least biased for Cones and MGs, MuSiC2 showed the least bias for HCs and RGCs, and SQUID demonstrated the most biased estimates across all cell types. Additionally, while DeMixSC, CIBERSORTx, and MuSiC2 all detected a decrease in Rod cells in AMD, only DeMixSC identified a statistically significant reduction in Cone cells, consistent with AMD pathology affecting both photoreceptors (Curcio et al., 1996). We further evaluated DeMixSC framework on the AMD cohort under different benchmark alignment conditions (see Methods, Supplemental Table S3, and Supplemental Fig. S12). Limma (Ritchie et al., 2015) also effectively corrected batch effects and yielded comparable deconvolution performance (mean RMSE: 0.05, mean Spearman's correlation: 0.68), while both no batch correction and VSN (Huber et al., 2002) showed

inferior results (mean RMSE: 0.12 and 0.20; mean Spearman's correlation: 0.57 and -0.61), highlighting the importance of effective batch correction.

It is known that adult retinal photoreceptors (Cone and Rod cells) cannot regenerate after injury (Fleckenstein et al., 2021; Khanani et al., 2022; Menon et al., 2019). We hypothesized that photoreceptor loss reduced the total cell count, hence inflating the cell proportions in AMD. Indeed, we found that losing 12.53% of total photoreceptors resulted in the observed subtle drop for Rod cells (2.72%) and Cone cells (0.61%) (see Methods). The increased proportion of BCs primarily resulted from photoreceptor loss, while MGs and Astrocytes showed actual increases beyond this effect (see Methods). These results aligned well with the current understanding that photoreceptor degeneration is accompanied by reactive gliosis (Pfeiffer et al., 2020; Tomita et al., 2021), characterized by glial cell activation and proliferation. In summary, our findings demonstrated DeMixSC's ability to capture subtle yet biologically relevant changes in retinal cell composition in AMD.

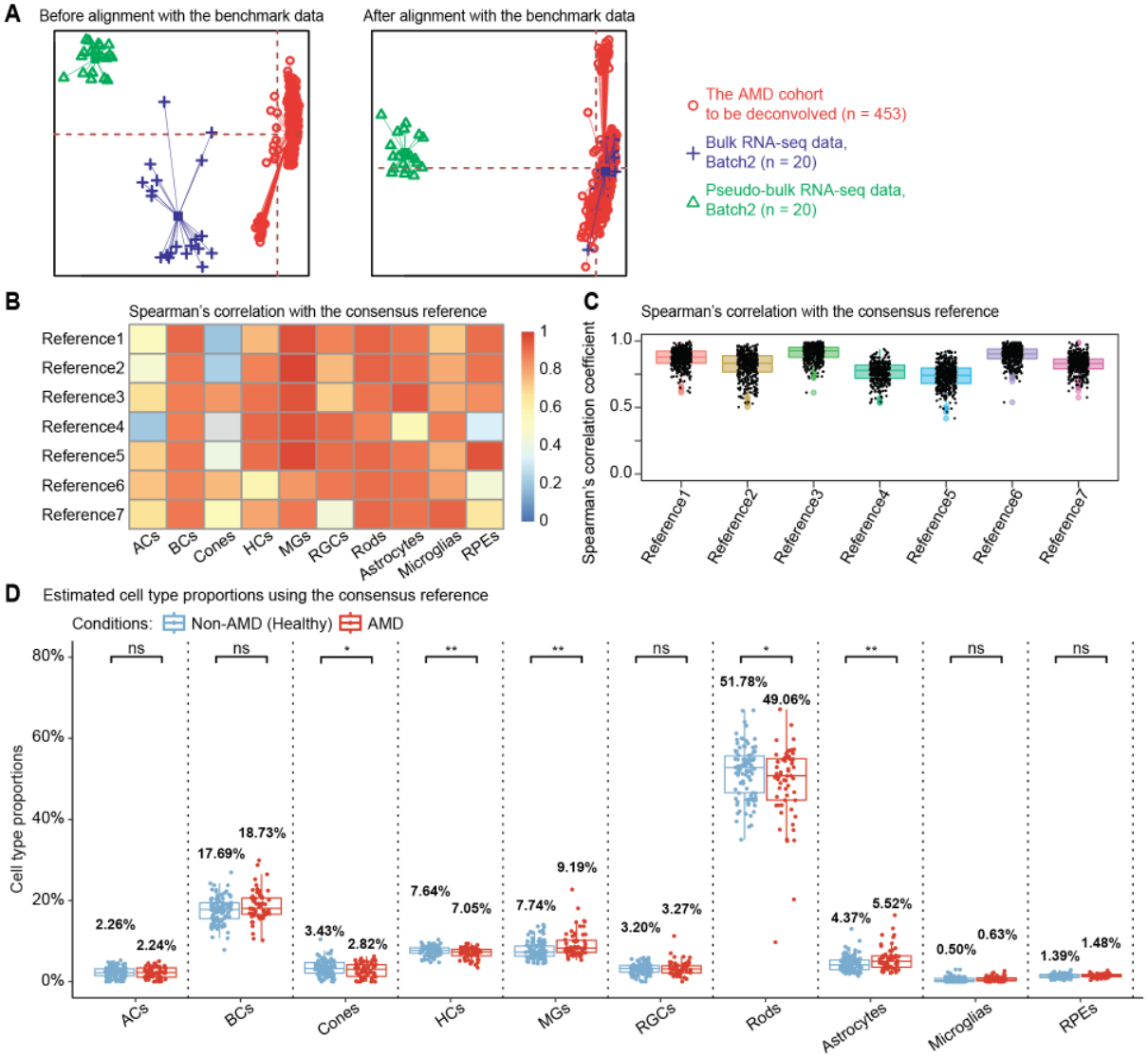

Figure 4. Using DeMixSC to deconvolve a large cohort of human peripheral retinal samples.
A, PCA plots of both the retina cohort data and the benchmark data. Red denotes the bulk data to be deconvolved, blue denotes the benchmark bulk data, and green denotes the benchmark pseudo-bulk data. **B** and **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 (healthy) 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.

Applying DeMixSC to high-grade serous ovarian cancer data

High-grade serous ovarian cancer (HGSC) is the most common and lethal subtype of epithelial ovarian cancer, yet it remains poorly understood (Veneziani et al., 2023). A major challenge in comprehending and treating HGSC lies in its extensive tumor heterogeneity, characterized by the presence of diverse cell populations. This heterogeneity contributes to differential responses to therapy and various clinical outcomes. Accurate deconvolution analysis of HGSC bulk cohorts with well-documented clinical follow-ups is crucial for dissecting cellular interactions underlying disease progression and treatment response.

We compared the deconvolution performance of DeMixSC to seven existing methods using a HGSC benchmark dataset (Hippen et al., 2023) (see Methods, Supplemental Fig. S4A). DeMixSC notably outperformed other methods when deconvolving Disso&poly(A)+ samples, achieving the lowest RMSE (mean: 0.09) and the highest Spearman's correlation coefficient (mean: 0.72) (Fig. 5A). At the cell type level, DeMixSC accurately estimated the proportions of all 13 cell types in the Disso&poly(A)+ samples, with marked improvements in the deconvolution of epithelial cells, endothelial cells, and T cells (Supplemental Fig. S13A). In comparison, the next best-performing method, CIBERSORTx, had a mean RMSE of 0.13 and a mean Spearman's correlation of 0.49 for the Disso&poly(A)+ data type. Additionally, we evaluated DeMixSC on two other data types (Disso&rRNA- and Chunk&rRNA-) from the HGSC dataset, which were made with lower level of technical matchness with the scRNA-seq data. DeMixSC did not outperform other methods on these two data types (Fig. 5A and Supplemental Fig. S13B, C; mean RMSE: 0.14 and 0.14, mean Spearman's correlation: 0.27 and 0.30, for Disso&rRNA- and Chunk&rRNA-, respectively), suggesting that the benchmark dataset needs to be specifically designed for optimal performance.

To demonstrate the generalizability of DeMixSC, we utilized the Disso&poly(A)+ data as the benchmark dataset to deconvolve an unmatched HGSC cohort (Lee et al., 2020) with detailed clinical annotations (see Methods). This cohort contains 30 primary treatment-naïve tumor samples, categorized into three groups based on their responses to treatment: those who underwent complete gross resection (R0, n=10), those who received neoadjuvant chemotherapy with an excellent response (ER, n=10), and those with a poor response (PR, n=10). As in the AMD deconvolution analysis, we applied DeMixSC, MuSiC, CIBERSORTx, and SQUID for deconvolution (see Methods). DeMixSC achieved the most biologically realistic estimations of cell type proportions among all compared methods (Fig. 5B and Supplemental Fig. S14). DeMixSC was the only method successfully captured proportion differences in epithelial cells (R0 vs ER, *P*-value = 0.013; R0 vs PR, *P*-value = 0.007) and macrophages (R0 vs ER, *P*-value = 0.085; R0 vs PR, *P*-value = 0.044) across three distinct response groups. Notably, DeMixSC revealed a decrease in epithelial cells from patients with no need for chemotherapy (R0) to only showing partial response to chemo-treatment (PR) (Supplemental Fig. S14), aligning with previous clinical observations (Caudle et al., 2010). Additionally, DeMixSC identified a consistent trend of increased proportion of macrophages from R0 to PR groups, suggesting that higher pre-treatment macrophage infiltrations may be associated with decreased treatment response. The deconvolution-based estimates were further validated by immunofluorescent staining (Lee et al., 2020) (see Methods, Spearman's correlation coefficient = 0.82, Fig. 5C). In contrast, the other three methods produced lower Spearman's correlations with the staining

results (MuSiC: 0.60, SQUID: 0, and CIBERSORTx: 0.56) and failed to discern biological differences across response groups (Supplemental Fig. S14).

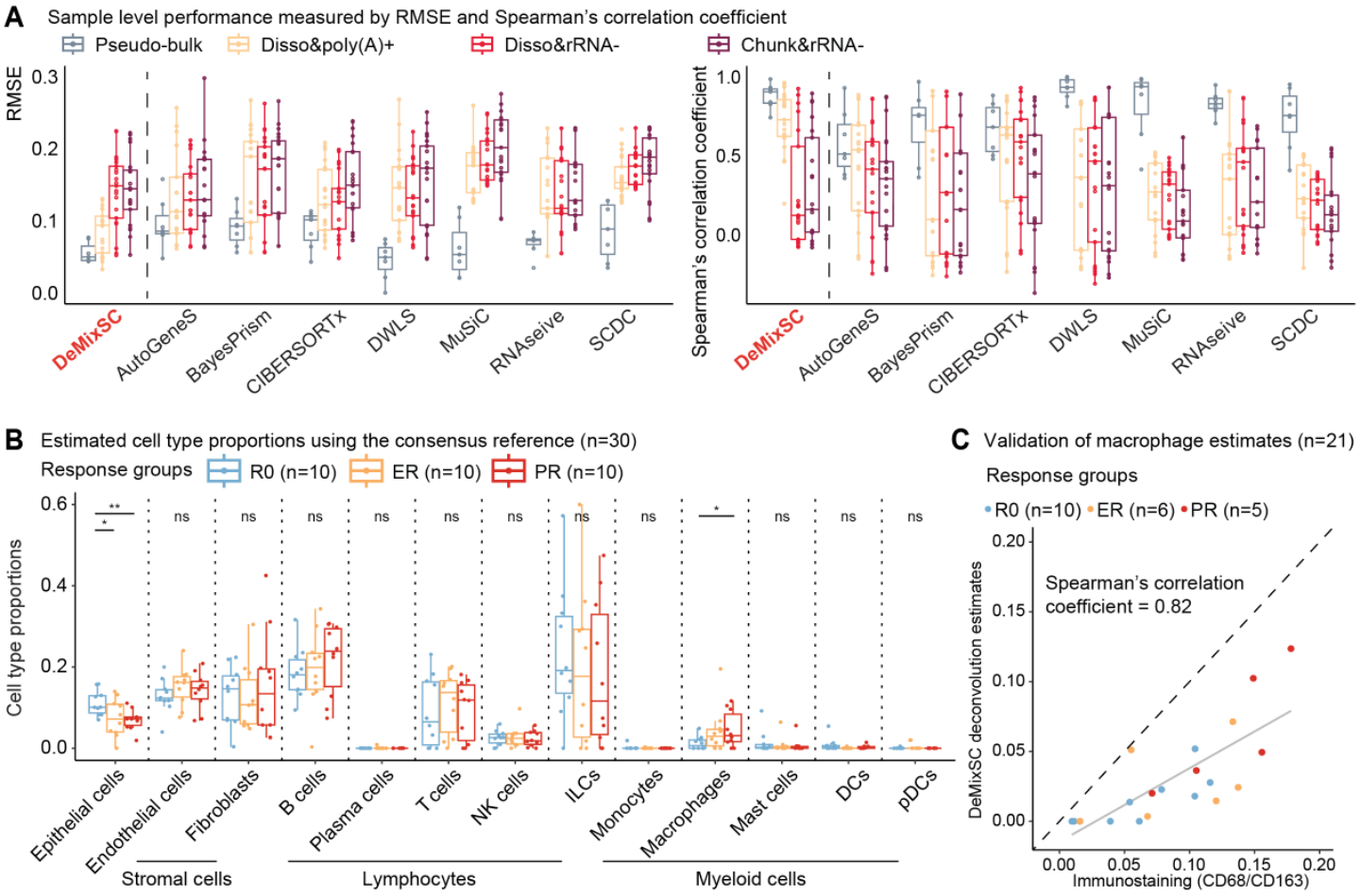

Figure 5. Using DeMixSC to deconvolve HGSC samples.
A, Boxplots showing the deconvolution performance of eight deconvolution methods for the pseudo-bulk and three types of bulk data in the HGSC benchmark dataset. RMSE values and Spearman's correlation coefficients are calculated across 13 cell types for each sample. Smaller RMSEs or larger Spearman's correlations indicate higher accuracy in proportion estimation. **B,** Distributions of DeMixSC estimated cell type proportions of *Lee et al.* data using consensus references. Each panel corresponds to a given cell type. NK cells: natural killer (NK) cells; ILC: innate lymphoid cells; DC: dendritic cells macrophages; pDC: plasmacytoid dendritic cells. The *P*-values for Student's *t*-tests comparing the estimated cell type proportions across R0, ER, and PR 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. **C,** Scatter plot comparing DeMixSC estimates of macrophages with experimental measurements from immunofluorescent (CD68/CD163) in 21 HGSC samples. The black dashed line represents the diagonal, while the grey solid line indicates the linear fit across the data points.

Discussion

This study addresses the technological discrepancy between bulk and sc/sn RNA-seq data in order to improve the deconvolution accuracy of bulk RNA-seq data. We constructed a specialized benchmark dataset of healthy retina samples and demonstrated the impact of technological discrepancy on existing single-cell-based deconvolution methods (Aliee and Theis, 2021; Chu et al., 2022; Cobos et al., 2023; Dong et al., 2021; Erdmann-Pham et al., 2021; Fan et al., 2022; Newman et al., 2019; Tsoucas et al., 2019; Wang et al., 2019). Using this benchmark dataset, we introduce the DeMixSC deconvolution method that makes innovative improvements to the wNNLS framework to address the consistently observed technological discrepancy at the gene level. The distinct advantage of DeMixSC lies in its superior deconvolution accuracy and broad generalizability. As demonstrated in the benchmarking study, DeMixSC achieves more accurate estimates of cell type proportions than other existing deconvolution methods. In our application to complex retina samples from patients with AMD, DeMixSC was able to accurately delineate 7 to 10 cell types and identify subtle yet critical changes in cell type proportions. Furthermore, DeMixSC succeeded in deconvolving ovarian cancer data by utilizing a publicly available HGSC benchmark dataset, where it achieved considerably more accurate deconvolution performance and discovered proportional differences associated with different treatment responses. Our studies support the capability and generalizability of DeMixSC in deconvolving large heterogeneous bulk cohorts, only requiring a small set of tissue-type-matched benchmark data. DeMixSC is computationally efficient, completing the analysis of 453 AMD samples within five minutes, and exhibits robust convergence against different starting values (see Methods, Supplemental Fig. S15).

Generation of the benchmark dataset in DeMixSC is crucial for accurate and reliable estimation of cell type proportions. Our study employed 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 procedure, the cDNA library for bulk RNA-seq was generated using the Smart-seq v4 ultralow input RNA kit procedure, a protocol similar to that used in snRNA-seq. The improved performance of DeMixSC in large cohort bulk data demonstrates the benchmark data generation is a worthwhile one-time investment. One available benchmark dataset can be utilized for unlimited times to deconvolve any large cohort of the same tissue type. Second, the required sample size for the benchmark dataset is small. Eight samples were sufficient to ensure accurate deconvolution in the retina benchmark dataset. Single-cell data for the tissue of interest are already being generated and are needed to apply existing single-cell-based deconvolution methods. Saving the remainder dissociated cell/nucleus suspension for a minimum of eight bulk RNA-seq experiments, an additional step that typically costs less than $2000, can provide valuable benchmark data for enhanced deconvolution accuracy. In addition, given its importance to the success of DeMixSC, we expect the specialized benchmark data can improve other deconvolution methods, such as the deep learning-based Scaden (Menden et al., 2020) and the guided topic modeling-based GTM-decon (Swapna et al., 2023), by providing insights into cross-platform technical discrepancies.

The advance represented by DeMixSC is noteworthy, but there is potential room for improvements in future work. The key to DeMixSC rests on effectively identifying and down-weighting genes with high technological discrepancy. A potential challenge arises in gene identification when applying DeMixSC to tissue

types (e.g., tumors) with high cellular plasticity. In that 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 (Cao et al., 2022)), and the remaining unstable genes. Moreover, DeMixSC can be expected to gain from machine learning models to simultaneously identify and adjust genes. Additionally, alternative methods to ComBat (Zhang et al., 2020) for aligning the large cohort with the benchmark dataset can be considered when dealing with tumor samples, which often are highly heterogeneous with complex batch structures. DeMixSC also holds the potential to address the challenge of missing cell types in single-cell reference samples by analyzing the residual information from the deconvolution process (Ivich et al., 2024).

Considering such potential adaptations, we anticipate that DeMixSC will prove useful in cancer research. By using a concise benchmark dataset derived from matched tissue specimens, DeMixSC can be leveraged to accurately deconvolve large bulk cohorts acquired through either surgical or biopsy samples. DeMixSC's enhanced deconvolution accuracy can improve the reliability of downstream cell type-specific differential expression analysis with any methods that rely on estimated cell type proportions (Luca et al., 2021; Wang et al., 2021). This capability can be expected to accelerate the discovery of cell subtypes and cell type-specific markers among diverse patient groups with a variety of different types of cancer.

Methods

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.

Human retina sample collection. These samples were obtained from 24 individuals between age of 73 to 91 who had passed away due to respiratory or heart failure or from a myocardial infarction (Supplemental Table S1). 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 published protocol (Owen et al., 2019). 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.

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 $MgCl_2$, 0.02% NP-40). 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 Smart-seq 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).

Benchmark design for matched single-cell/nucleus and bulk RNA-seq data. The workflow for the benchmark design was summarized in Supplemental Fig. S1. Two steps were essential for the 'matchness' of the paired bulk and sc/snRNA-seq in the benchmark dataset. First, tissue chunks needed to be dissociated into cell or nucleus suspensions, and the paired bulk and sc/snRNA-seq profiling was carried out using the same aliquot (Supplemental Fig. S1A). This process guaranteed the two sequencing datasets share approximately equal cell type proportions. Second, it was necessary to employ the same cDNA library preparation protocol for both sequencing data (Supplemental Fig. S1B). In our study, both bulk and single-nucleus cDNA libraries were generated using the ploy(A) enrichment method. These two critical steps together ensured that any technological discrepancies stem solely from the sequencing platforms.

Preprocessing of snRNA-seq and bulk RNA-seq data. Retina snRNA-seq UMI (unique molecular identifier) count matrices were obtained using CellRanger (Zheng et al., 2017) (version 3.1.0) following the official guide to estimate absolute counts and were then processed using the Seurat (Hao et al., 2021) 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 (Babraham Bioinformatics, 2019) (version 0.11.9), and low-quality reads and adapters were then trimmed by Trimmomatic (Bolger et al., 2014) (version 0.4.0). Next, reads that passed quality control were aligned to GRCh38 using the 2-pass mode of STAR (Dobin et al., 2013) (version 2.7.7b), and read counts were obtained by featureCount (Liao et al., 2014) 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 (Liang et al., 2019; Menon et al., 2019) (Supplemental Fig. S2A, B

and Supplemental Table S2). For the deconvolution analysis of bulk AMD retinal samples (Ratnapriya et al., 2019), we included additional three 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 (Benjamini and Hochberg, 1995). We used Student's *t*-tests to compare the estimated cell type proportions between non-AMD and AMD conditions 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: not 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 (Aliee and Theis, 2021; Cobos et al., 2023; Dong et al., 2021; Fan et al., 2022; Ruppert and Wand, 1994; Tsoucas et al., 2019; Wang et al., 2019), which is

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

where y_{jg} is the observed expression (relative abundance) of gene g from subject j in the bulk RNA-seq data, $\hat{r}_g^k$ is the estimated expression value of gene g in cell type k in the reference matrix derived from sc/snRNA-seq data, w_{jg} is the weight of each gene g for subject j , n_j is the total number of cells in subject j, and $\hat{\boldsymbol{p}}_j$ is the estimated vector of cell type proportions. 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

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

where $\tilde{y}_{jg}$ is the true expression value of gene g in the bulk data with $\tilde{y}_{jg}+\epsilon_{jg}=y_{jg}$, ϵ_{jg} is the measurement noise for gene g in subject j at the bulk level, γ_{jg} is the accumulated measurement noise at the single-cell level, and r_{jg}^k is the true cell type-specific reference matrix (see Supplemental Note). The component $\left(\tilde{y}_{jg}-n_j\sum_{k\in K} p_j^k r_{jg}^k\right)$

consists of the true bulk level expression $\tilde{y}_{jg}$ and the true expression value derived based on the true cell type-specific mean expression $n\sum_{k\in K}p_j^k r_{jg}^k$. This component reflects the true estimation error that we aim to minimize. The component $(\epsilon_{jg}\text{-}\gamma_{jg})$ defines the difference in noises introduced by the bulk (ϵ_{jg}) and the sc/snRNA-seq (γ_{jg}) data, which represents the measurable technological discrepancy between sequencing platforms. Genes with highly inconsistent expressions, or equivalently inconsistent noises, across different platforms suffer from high technological discrepancy (see Supplemental Note). Thus, when the technological discrepancy overtakes the true signal, instead of minimizing estimation errors, this model is geared towards minimizing the technological discrepancy and is no longer fitting the expression profiles of individual bulk samples.

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

$$\hat{\boldsymbol{p}}_j=\underset{\boldsymbol{p}_j\geq 0}{\operatorname{argmin}}\left(\sum_{g\in G_1} w_{jg}\left(y_{jg}\text{-}n_j\sum_{k\in K}p_j^k r_{jg}^k\right)^2+\sum_{g\in G_2} w_{jg}\left(\frac{y_{jg}}{a}\text{-}n_j\sum_{k\in K}p_j^k\frac{r_{jg}^k}{a}\right)^2\right) \quad (3),$$

where G_1 is a set of genes hardly affected by technological discrepancy and G_2 contains genes highly affected by technological discrepancy. DeMixSC employs a DE analysis to identify genes affected by technological discrepancy (G_2). This process begins with a paired *t*-test on matched bulk and pseudo-bulk data. Genes with BH adjusted *P*-values less than 0.05 are selected and ranked from high to low based on mean expression across both data types. The top-ranking genes are most impacted by technological discrepancy and designated to G_2. To mitigate this technological discrepancy, DeMixSC introduces a positive adjustment factor (a) to rescale the gene expression and thereby reduce the contribution of squared residuals from genes in G_2. Rather than excluding them, DeMixSC preserves the high discrepancy genes in G_2 in the wNNLS model, acknowledging their potential biological significance and contribution to mixed expression levels. The size of G_2 and the value of a are user-definable parameters in DeMixSC, allowing for flexibility in different analytical contexts. We have tested the model's performance with various G_2 gene selections and values of a (Supplemental Fig. S16), and we set the size of G_2 to be 5,000 and a to be 1,000 by default.

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:

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

The current literature uses either the squared fitted value $(\hat{y}_{jg})^2$ (Cobos et al., 2023; Tsoucas et al., 2019) or the variance $(y_{jg}\text{-}\hat{y}_{jg})^2$ (Dong et al., 2021; Fan et al., 2022; Wang et al., 2019) in the weight, but never both. The constant term c is introduced for controlling the range of the weight. A range of positive values can be appropriate for the constant c (Supplemental Fig. S17). We treat c as a tuning parameter in the DeMixSC software and set it to 2 by default for users' convenience. Using the summation of these three terms in Eq(4) as our new weight function improves model fit, accounts for variability, and enhances the numerical stability of the DeMixSC framework. Detailed mathematical derivation is provided in the Supplemental Note. Implementation of the DeMixSC framework is available as an R package (R Core Team, 2023) at https://github.com/wwylab/DeMixSC.

Evaluation of gene partitioning and weight function. To validate the efficacy of our proposed gene partitioning and weight function, we tested an alternative method with the retina bulk benchmark data. We used the technological discrepancy (i.e., the test statistics from the paired *t*-test) as inverse weights and did not rescale the gene expressions using adjustment factor (i.e., no gene partition). The tested inverse weights were calculated as the *t*-statistics: $w_g^{-1} = \frac{\bar{d}_g}{s_{d_g}/\sqrt{n}}$, where $\bar{d}_g$ is mean of the differences between paired bulk and pseudo-bulk samples (i.e., $\bar{d}_g = y_{jg}^{bulk} - y_{jg}^{pseudo\text{-}bulk}$) of gene g, s_{d_g} is the standard deviation of the differences, and n is the number of sample pairs. This approach yielded decreased deconvolution accuracy, with mean RMSE increasing from 0.03 to 0.13 and mean Spearman's correlation decreasing from 0.86 to 0.73.

Evaluation of batch correction methods. We tested DeMixSC framework on the AMD retina cohort with four batch correction approaches: no correction, ComBat (Zhang et al., 2020) (implemented in DeMixSC), Limma (Ritchie et al., 2015), and VSN (Huber et al., 2002). Each method was applied following its standard procedure.

Data normalization of bulk mixtures. We applied the following data normalizations to the bulk raw count matrices (Dillies et al., 2013): (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.

Convergence property of the DeMixSC algorithm. To evaluate how robust DeMixSC is against different initial values, we randomly selected a sample from the AMD retina cohort 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. Finally, we used the 30 pairs of $n \times \hat{p}$ to initialize the wNNLS framework and then compared the estimates of DeMixSC.

Computational deconvolution with existing methods. Eight deconvolution methods that use the same scRNA-seq data as the reference were tested in our benchmarking study (Aliee and Theis, 2021; Chu et al., 2022; Cobos et al., 2023; Dong et al., 2021; Erdmann-Pham et al., 2021; Newman et al., 2019; Tsoucas et al., 2019; Wang et al., 2019). 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, 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 (Wang et al., 2019) and the ensemble

option of SCDC (Dong et al., 2021). 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 (Supplementary Fig. 6A), 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 (Hao et al., 2021) 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. For deconvolving the AMD cohort, we used MuSiC2 (Fan et al., 2022) following the tutorial provided with default settings (https://github.com/Jiaxin-Fan/MuSiC2).

Evaluation metrics for the deconvolution performance. We evaluated the performance of each method using: (1) Root-mean-square error (RMSE) at both sample ($RMSE^j$) and cell type ($RMSE^k$) levels: $RMSE^j=\sqrt{\frac{\sum_{k=1}^{K}(\hat{p}_j^k-p_j^k)^2}{K}}$ and $RMSE^k=\sqrt{\frac{\sum_{j=1}^{J}(\hat{p}_j^k-p_j^k)^2}{J}}$, 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. A smaller RMSE value indicates a better deconvolution performance.

(2) Spearman's correlation coefficient (ρ) at both sample (ρ^j) and cell type (ρ^k) levels: $\rho^j=1-\frac{6\sum d_k^2}{K(K^2-1)}$ and $\rho^k=1-\frac{6\sum d_j^2}{J(J^2-1)}$, where d_j denotes the difference between the ranks of $\hat{p}_j^k$ and p_j^k. A higher Spearman's correlation coefficient indicates a better deconvolution performance.

(3) Deviation from the ground truth: To show the variability across samples and deviation from the ground truth, we calculate the differences $\hat{p}_j^k-p_j^k$. A score of 0 indicates perfect concordance, greater than 0 means over-estimation, and less than 0 suggests under-estimation.

Deconvolution analysis on the human healthy retina benchmark dataset. The retina benchmark dataset comprised two batches, batch-1 (n=4) and batch-2 (n=20). To account for potential batch effects and ensure optimal deconvolution performance for each method, the analysis was performed separately for each batch. All deconvolution methods were performed using the same single-nucleus reference derived from the retinal benchmark dataset.

Deconvolution analysis on the human diseased retina cohort (aged-macular degeneration, AMD).

Data acquisition and quality control. The expression matrix of the AMD cohort comprised 523 samples was obtained from Ratnapriya et al., 2019's study under the accession GSE115828. We conducted quality control following the pipeline described in their original study. Samples were filtered out due to ambiguous clinical features (n=26), poor sequencing results (n=24), 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. We further evaluated DeMixSC's performance on genetically diverse populations by including six samples with ancestry diverging from Caucasian, which were previously excluded based on QC criteria (Fan et al., 2022; Ratnapriya et al., 2019). DeMixSC achieved comparable results (mean RMSE: 0.04 and mean Spearman's correlation: 0.75) to our original analysis, demonstrating DeMixSC's potential applicability across diverse genetic backgrounds.

Computational deconvolution with consensus reference. We performed deconvolution using four methods: DeMixSC, MuSiC2, CIBERSORTx , and SQUID. The three existing methods were selected for the following reasons: (1) MuSiC2 leveraged conditionally stable genes from healthy references to analyze diseased tissue; (2) CIBERSORTx was ranked second in our benchmarking study; and (3) SQUID also deconvolved unmatched large bulk cohorts using benchmark data. To run each method, we generated a consensus reference by integrating seven samples from batch-2 (Sample 5, 10, 12, 18, 19, 21 and 23, Supplemental Table S1). 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 θ_{jg}^{k} and cell size s_j^k for each cell type k were calculated for each sample j. A consensus reference matrix $\boldsymbol{r}$ was subsequently derived by multiplying the averaged relative abundance and the averaged cell size across the selected samples. Mathematically, the consensus reference is defined as $r_g^k=\bar{\theta}_g^k\,\bar{s}^k$, where $\bar{\theta}_g^k=\frac{\sum_j \theta_{jg}^k}{7}$ and $\bar{s}^k=\frac{\sum_j s_j^k}{7}$ are the averaged abundance and averaged cell size over the seven samples, respectively.

Validating deconvolution performance using reference proportions from healthy human peripheral retina. To validate methods performance in deconvolving the AMD cohort, we compared the estimated cell type proportions of non-AMD samples (n=105) with the previously reported proportions in healthy human peripheral retina tissues (Liang et al., 2019). The non-AMD samples were peripheral retina tissues from donors aged between 55 and 94, with a mean age of 80 ± 9.95 years. Reference proportions were calculated based on snRNA-seq profiling of the peripheral retina from three human donors aged 60 to 80 years, matching the non-AMD samples by age. Seven major cell types were identified with the following proportions: ACs, 7.74%; BCs, 15.6%; Cones, 4.61%; HCs, 3.61%; MGs, 14.08%; RGCs, 1.07%; and Rods, 53.29% (Supplemental Table S3). To account for the additional three minor cell types (RPE, astrocytes, and microglia) estimated in the non-AMD samples but not measured in the reference, we rescaled the proportions of the seven major cell types so that they sum to 1 by dividing their total.

Accounting for the total cell loss in the AMD cohort. The decrease in overall cell count induced by photoreceptors (Cone and Rod cells) loss likely amplifies the cell type proportions in the AMD samples (Ambati et al., 2013; Fleckenstein et al., 2021; Menon et al., 2019). The mean estimated fraction of photoreceptors is 55.21% (3.43% Cone +51.78% Rod) in non-AMD and 51.88% (2.82% Cone + 49.06% Rod) in AMD (Fig. 4D). We use "x" to represent the mean percentage of lost photoreceptors in the AMD condition and derive the relation: $0.5521(1-x)/(1-0.5521x)=0.5188$. Solving for x shows a 12.53% reduction in photoreceptors in the peripheral AMD retina, which aligns well with biological evidence that the peripheral retina experiences a more modest

photoreceptors loss (10% to 20%) when compared to the macular region (>30%) (Curcio et al., 1996). Next, we investigated whether the observed increases in BCs, MGs, and Astrocytes were driven by the death of photoreceptors or cell proliferation. Using our photoreceptor loss metric, we estimated the expected cell fractions in AMD: (estimated non-AMD proportion) / (1 - 0.5521 * 0.1253). The expected fractions were 19.00% for BCs (from 17.69% in non-AMD), 8.31% for MGs (from 7.74%), and 4.69% for Astrocytes (from 4.37%). The expected fraction of BCs closely matches DeMixSC's estimate (18.73%), suggesting the increase of BCs was due to photoreceptor loss. For glial cells, DeMixSC's estimates (9.19% for MGs and 5.52% for Astrocytes) exceeded the expected fractions, suggesting an increase of these cells in the AMD condition.

Deconvolution analysis on the human primary HGSC benchmark dataset. The HGSC benchmark dataset, obtained from GSE217517 (Hippen et al., 2023), comprised seven primary HGSC samples. For each of these samples, three types of bulk RNA-seq data were generated, with three technical replicates for each data type: (1) dissociation with poly(A) enrichment (Disso&poly(A)+, n=21); (2) dissociation with rRNA depletion (Disso&rRNA-, n=21); and (3) tissue chunk with rRNA depletion (Chunk&rRNA-, n=21). For the matched single-cell data, we followed the cell type annotation as described in the original paper. Thirteen cell types were identified: epithelial cells, endothelial cells, fibroblasts, B cells, plasma cells, natural killer (NK) cells, innate lymphoid cells (ILCs), monocytes, macrophages, mast cells, dendritic cells (DCs), plasmacytoid dendritic cells (pDCs), and T cells. All deconvolution methods were performed using the same single-nucleus reference derived from the HGSC benchmark dataset.

Deconvolution analysis on the unmatched human primary HGSC cohort.

Data acquisition and quality control. The unmatched human primary HGSC cohort was obtained from Lee et al., 2020's study under the accession EGAD00001005238 (Lee et al., 2020). This cohort contains 30 primary HGSC samples categorized into three treatment response groups: complete gross resection (R0), received neoadjuvant chemotherapy with excellent (ER) or poor (PR) response.

Computational deconvolution with consensus reference. We performed deconvolution with four methods, including DeMixSC, MuSiC, CIBERSORTx, and SQUID. Unlike the AMD study, we used MuSiC rather than MuSiC2 because MuSiC and MuSiC2 shared the same computational framework, but MuSiC2 was specifically designed to use normal references for deconvolving disease samples, which was not applicable in the HGSC study. To run each method, we generated a consensus reference by integrating all seven scRNA-seq samples from the HGSC benchmark dataset. For each sample, we randomly selected up to 1,000 cells per cell type, or all available cells if fewer than 1,000 were present. The methodology for generating the consensus reference matrix followed the same approach described in the "*Deconvolution analysis on the human diseased retina cohort (aged-macular degeneration, AMD)*" section.

Comparing with the immunostaining results. Out of the 30 primary HGSC samples, 21 had both RNA-seq and immunostaining data for macrophage, as measured using CD68 and CD163 antibodies. The detailed data description and immunostaining results were obtained from the original study (Lee et al., 2020).

Software availability

DeMixSC is freely available as an R package and can be downloaded from our GitHub repository: https://github.com/wwylab/DeMixSC. A tutorial for DeMixSC is available at https://wwylab.github.io/DeMixSC/. The DeMixSC source code is also available as Supplemental Code.

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE175937.

The human retina tissue snRNA-seq data in this study have been submitted to the Human Cell Atlas Data Portal (https://explore.data.humancellatlas.org/projects/9c20a245-f2c0-43ae-82c9-2232ec6b594f).

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 NIH (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 NIH (K22CA234406). R.C. is supported by Human Cell Atlas Seed Network - Retina, Chan Zuckerberg Institute (CZF2019-02425), National Eye Institute (R01EY022356 and 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), the Cytometry and Cell Sorting Core at Baylor College of Medicine with funding from the CPRIT Core Facility Support Award (CPRIT-RP180672), 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, and 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-nucleus 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. S.L and A.K.S. helped with ovarian

769 cancer data analysis, W.W., S.G., and X.L. wrote the paper, with input from all authors. All authors reviewed and
770 approved the final version of the manuscript.

References

Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., Gnirke, A., 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 12, 1–14. https://doi.org/10.1186/GB-2011-12-2-R18

Aliee, H., Theis, F.J., 2021. AutoGeneS: automatic gene selection using multi-objective optimization for RNA-seq deconvolution. Cell Syst 12, 706–715. https://doi.org/10.1016/J.CELS.2021.05.006

Ambati, J., Atkinson, J.P., Gelfand, B.D., 2013. Immunology of age-related macular degeneration. Nat Rev Immunol 13, 438–451. https://doi.org/10.1038/nri3459

Anghel, C. V., Quon, G., Haider, S., Nguyen, F., Deshwar, A.G., Morris, Q.D., Boutros, P.C., 2015. ISOpureR: an R implementation of a computational purification algorithm of mixed tumour profiles. BMC Bioinformatics 16, 1–11. https://doi.org/10.1186/S12859-015-0597-X

Babraham Bioinformatics, 2019. FastQC: A Quality Control tool for High Throughput Sequence Data [WWW Document]. URL https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57, 289–300. https://doi.org/10.1111/J.2517-6161.1995.TB02031.X

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114. https://doi.org/10.1093/BIOINFORMATICS/BTU170

Cao, S., Wang, J.R., Ji, S., Yang, P., Dai, Y., Guo, S., Montierth, M.D., Shen, J.P., Zhao, X., Chen, J., Lee, J.J., Guerrero, P.A., Spetsieris, N., Engedal, N., Taavitsainen, S., Yu, K., Livingstone, J., Bhandari, V., Hubert, S.M., Daw, N.C., Futreal, P.A., Efstathiou, E., Lim, B., Viale, A., Zhang, J., Nykter, M., Czerniak, B.A., Brown, P.H., Swanton, C., Msaouel, P., Maitra, A., Kopetz, S., Campbell, P., Speed, T.P., Boutros, P.C., Zhu, H., Urbanucci, A., Demeulemeester, J., Van Loo, P., Wang, W., 2022. Estimation of tumor cell total mRNA expression in 15 cancer types predicts disease progression. Nat Biotechnol 40, 1624–1633. https://doi.org/10.1038/s41587-022-01342-x

Caudle, A.S., Gonzalez-Angulo, A.M., Hunt, K.K., Liu, P., Pusztai, L., Symmans, W.F., Kuerer, H.M., Mittendorf, E.A., Hortobagyi, G.N., Meric-Bernstam, F., 2010. Predictors of tumor progression during neoadjuvant chemotherapy in breast cancer. Journal of Clinical Oncology 28, 1821–1828. https://doi.org/10.1200/JCO.2009.25.3286

Chu, T., Wang, Z., Pe'er, D., Danko, C.G., 2022. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. Nat Cancer 3, 505–517. https://doi.org/10.1038/s43018-022-00356-3

Cobos, F.A., Alquicira-Hernandez, J., Powell, J.E., Mestdagh, P., De Preter, K., 2020. Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nat Commun 11, 1–14. https://doi.org/10.1038/s41467-020-19015-1

Cobos, F.A., Panah, M.J.N., Epps, J., Long, X., Man, T.-K., Chiu, H.-S., Chomsky, E., Kiner, E., Krueger, M.J., di Bernardo, D., Voloch, L., Molenaar, J., van Hooff, S.R., Westermann, F., Jansky, S., Redell, M.L., Mestdagh, P., Sumazin, P., 2023. Effective methods for bulk RNA-seq deconvolution using scnRNA-seq transcriptomes. Genome Biol 24, 1–22. https://doi.org/10.1186/S13059-023-03016-6

Curcio, C.A., Medeiros, N.E., Millican, C.L., 1996. Photoreceptor loss in age-related macular degeneration. Invest Ophthalmol Vis Sci 37, 1236–1249. https://iovs.arvojournals.org/article.aspx?articleid=2180378

Denisenko, E., Guo, B.B., Jones, M., Hou, R., De Kock, L., Lassmann, T., Poppe, D., Poppe, D., Clément, O., Simmons, R.K., Simmons, R.K., Lister, R., Forrest, A.R.R., 2020. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol 21, 1–25. https://doi.org/10.1186/S13059-020-02048-6

Dietrich, A., Sturm, G., Merotto, L., Marini, F., Finotello, F., List, M., 2022. SimBu: bias-aware simulation of bulk RNA-seq data with variable cell-type composition. Bioinformatics 38, 141–147. https://doi.org/10.1093/BIOINFORMATICS/BTAC499

Dillies, M.A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, N.S., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloë, D., Le Gall, C., Schaëffer, B., Le Crom, S., Guedj, M., Jaffrézic, F., 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform 14, 671–683. https://doi.org/10.1093/BIB/BBS046

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/BIOINFORMATICS/BTS635

Dong, M., Thennavan, A., Urrutia, E., Li, Y., Perou, C.M., Zou, F., Jiang, Y., 2021. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. Brief Bioinform 22, 416–427. https://doi.org/10.1093/BIB/BBZ166

Erdmann-Pham, D.D., Fischer, J., Hong, J., Song, Y.S., 2021. Likelihood-based deconvolution of bulk gene expression data using single-cell references. Genome Res 31, 1794–1806. https://doi.org/10.1101/gr.272344.120

Fan, J., Lyu, Y., Zhang, Q., Wang, X., Li, M., Xiao, R., 2022. MuSiC2: cell-type deconvolution for multi-condition bulk RNA-seq data. Brief Bioinform 23, 1–10. https://doi.org/10.1093/BIB/BBAC430

Fleckenstein, M., Keenan, T.D.L., Guymer, R.H., Chakravarthy, U., Schmitz-Valckenberg, S., Klaver, C.C., Wong, W.T., Chew, E.Y., 2021. Age-related macular degeneration. Nat Rev Dis Primers 7, 1–25. https://doi.org/10.1038/s41572-021-00265-2

Gohil, S.H., Iorgulescu, J.B., Braun, D.A., Keskin, D.B., Livak, K.J., 2020. Applying high-dimensional single-cell technologies to the analysis of cancer immunotherapy. Nat Rev Clin Oncol 18, 244–256. https://doi.org/10.1038/s41571-020-00449-x

Haniffa, M., Taylor, D., Linnarsson, S., Aronow, B.J., Bader, G.D., Barker, R.A., Camara, P.G., Camp, J.G., Chédotal, A., Copp, A., Etchevers, H.C., Giacobini, P., Göttgens, B., Guo, G., Hupalowska, A., James, K.R., Kirby, E., Kriegstein, A., Lundeberg, J., Marioni, J.C., Meyer, K.B., Niakan, K.K., Nilsson, M., Olabi, B., Pe'er, D., Regev, A., Rood, J., Rozenblatt-Rosen, O., Satija, R., Teichmann, S.A., Treutlein, B., Vento-Tormo, R., Webb, S., Barbry, P., Bayraktar, O., Behjati, S., Bosio, A., Canque, B., Chalmel, F., Gitton, Y., Henderson, D., Jorgensen, A., Lisgo, S., Liu, J., Lundberg, E., Maitre, J.L., Mazaud-Guittot, S., Robertson, E., Rolland, A., Scharfmann, R., Souyri, M., Sundström, E., Zaffran, S., Zilbauer, M., 2021. A roadmap for the human developmental cell atlas. Nature 597, 196–205. https://doi.org/10.1038/s41586-021-03620-1

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E.P., Jain, J., Srivastava, A., Stuart, T., Fleming, L.M., Yeung, B., Rogers, A.J., McElrath, J.M., Blish, C.A., Gottardo, R., Smibert, P., Satija, R., 2021. Integrated analysis of multimodal single-cell data. Cell 184, 3573–3587. https://doi.org/10.1016/J.CELL.2021.04.048

Hippen, A.A., Omran, D.K., Weber, L.M., Jung, E., Drapkin, R., Doherty, J.A., Hicks, S.C., Greene, C.S., 2023. Performance of computational algorithms to deconvolve heterogeneous bulk ovarian tumor tissue depends on experimental factors. Genome Biol 24, 1–27. https://doi.org/10.1186/S13059-023-03077-7

Huber, W., Von Heydebreck, A., Sültmann, H., Poustka, A., Vingron, M., 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18, S96–S104. https://doi.org/10.1093/BIOINFORMATICS/18.SUPPL_1.S96

Ivich, A., Davidson, N.R., Grieshober, L., Li, W., Hicks, S.C., Doherty, J.A., Greene, C.S., 2024. Missing cell types in single-cell references impact deconvolution of bulk data but are detectable. bioRxiv 2024.04.25.590992. https://doi.org/10.1101/2024.04.25.590992

Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K.M., Sul, J.H., Pietiläinen, K.H., Pajukanta, P., Halperin, E., 2020. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nat Commun 11, 1–11. https://doi.org/10.1038/s41467-020-15816-6

Jin, H., Liu, Z., 2021. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. Genome Biol 22, 1–23. https://doi.org/10.1186/S13059-021-02290-6

Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127. https://doi.org/10.1093/BIOSTATISTICS/KXJ037

Khanani, A.M., Thomas, M.J., Aziz, A.A., Weng, C.Y., Danzig, C.J., Yiu, G., Kiss, S., Waheed, N.K., Kaiser, P.K., 2022. Review of gene therapies for age-related macular degeneration. Eye 36, 303–311. https://doi.org/10.1038/s41433-021-01842-1

Lee, S., Zhao, L., Rojas, C., Bateman, N.W., Yao, H., Lara, O.D., Celestino, J., Morgan, M.B., Nguyen, T. V., Conrads, K.A., Rangel, K.M., Dood, R.L., Hajek, R.A., Fawcett, G.L., Chu, R.A., Wilson, K., Loffredo, J.L., Viollet, C., Jazaeri, A.A., Dalgard, C.L., Mao, X., Song, X., Zhou, M., Hood, B.L., Banskota, N., Wilkerson, M.D., Te, J., Soltis, A.R., Roman, K., Dunn, A., Cordover, D., Eterovic, A.K., Liu, J., Burks, J.K., Baggerly, K.A., Fleming, N.D., Lu, K.H., Westin, S.N., Coleman, R.L., Mills, G.B., Casablanca, Y., Zhang, J., Conrads, T.P., Maxwell, G.L., Futreal, P.A., Sood, A.K., 2020. Molecular Analysis of Clinically Defined Subsets of High-Grade Serous Ovarian Cancer. Cell Rep 31, 107502. https://doi.org/10.1016/J.CELREP.2020.03.066

Li, X., Wang, C.Y., 2021. From bulk, single-cell to spatial RNA sequencing. Int J Oral Sci 13, 1–6. https://doi.org/10.1038/s41368-021-00146-0

Liang, Q., Cheng, X., Wang, J., Owen, L., Shakoor, A., Lillvis, J.L., Zhang, C., Farkas, M., Kim, I.K., Li, Y., DeAngelis, M., Chen, R., 2023. A multi-omics atlas of the human retina at single-cell resolution. Cell Genomics 3, 1–18. https://doi.org/10.1016/J.XGEN.2023.100298

Liang, Q., Dharmat, R., Owen, L., Shakoor, A., Li, Y., Kim, S., Vitale, A., Kim, I., Morgan, D., Liang, S., Wu, N., Chen, K., DeAngelis, M.M., Chen, R., 2019. Single-nuclei RNA-seq on human retinal tissue provides improved transcriptome profiling. Nat Commun 10, 1–12. https://doi.org/10.1038/s41467-019-12917-9

Liao, Y., Smyth, G.K., Shi, W., 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930. https://doi.org/10.1093/BIOINFORMATICS/BTT656

Luca, B.A., Steen, C.B., Matusiak, M., Azizi, A., Varma, S., Zhu, C., Przybyl, J., Espín-Pérez, A., Diehn, M., Alizadeh, A.A., van de Rijn, M., Gentles, A.J., Newman, A.M., 2021. Atlas of clinically distinct cell states and ecosystems across human solid tumors. Cell 184, 5482-5496.e28. https://doi.org/10.1016/J.CELL.2021.09.014

Menden, K., Marouf, M., Oller, S., Dalmia, A., Magruder, D.S., Kloiber, K., Heutink, P., Bonn, S., 2020. Deep learning-based cell composition analysis from tissue expression profiles. Sci Adv 1–12. https://doi.org/DOI: 10.1126/sciadv.aba2619

Menon, M., Mohammadi, S., Davila-Velderrain, J., Goods, B.A., Cadwell, T.D., Xing, Y., Stemmer-Rachamimov, A., Shalek, A.K., Love, J.C., Kellis, M., Hafler, B.P., 2019. Single-cell transcriptomic atlas of the human retina

identifies cell types associated with age-related macular degeneration. Nat Commun 10, 1–9. https://doi.org/10.1038/s41467-019-12780-8

Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D.J., Álvarez-Varela, A., Batlle, E., Sagar, Grün, D., Lau, J.K., Boutet, S.C., Sanada, C., Ooi, A., Jones, R.C., Kaihara, K., Brampton, C., Talaga, Y., Sasagawa, Y., Tanaka, K., Hayashi, T., Braeuning, C., Fischer, C., Sauer, S., Trefzer, T., Conrad, C., Adiconis, X., Nguyen, L.T., Regev, A., Levin, J.Z., Parekh, S., Janjic, A., Wange, L.E., Bagnoli, J.W., Enard, W., Gut, M., Sandberg, R., Nikaido, I., Gut, I., Stegle, O., Heyn, H., 2020. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nat Biotechnol 38, 747–755. https://doi.org/10.1038/s41587-020-0469-4

Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., Diehn, M., Alizadeh, A.A., 2019. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol 37, 773–782. https://doi.org/10.1038/s41587-019-0114-2

Olsen, T.W., Feng, X., 2004. The Minnesota Grading System of eye bank eyes for age-related macular degeneration. Invest Ophthalmol Vis Sci 45, 4484–4490. https://doi.org/10.1167/IOVS.04-0342

Owen, L.A., Shakoor, A., Morgan, D.J., Hejazi, A.A., Wade Mcentire, M., Brown, J.J., Farrer, L.A., Kim, I., Vitale, A., Deangelis, M.M., 2019. The Utah protocol for postmortem eye phenotyping and molecular biochemical analysis. Invest Ophthalmol Vis Sci 60, 1204. https://doi.org/10.1167/IOVS.18-24254

Pfeiffer, R.L., Marc, R.E., Jones, B.W., 2020. Persistent remodeling and neurodegeneration in late-stage retinal degeneration. Prog Retin Eye Res 74, 100771. https://doi.org/10.1016/J.PRETEYERES.2019.07.004

R Core Team, 2023. R: A Language and Environment for Statistical Computing. https://cran.r-project.org/

Ratnapriya, R., Sosina, O.A., Starostik, M.R., Kwicklis, M., Kapphahn, R.J., Fritsche, L.G., Walton, A., Arvanitis, M., Gieser, L., Pietraszkiewicz, A., Montezuma, S.R., Chew, E.Y., Battle, A., Abecasis, G.R., Ferrington, D.A., Chatterjee, N., Swaroop, A., 2019. Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. Nat Genet 51, 606–610. https://doi.org/10.1038/s41588-019-0351-9

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43, e47. https://doi.org/10.1093/nar/gkv007

Ruppert, D., Matthew P. Wand, 1994. Multivariate locally weighted least squares regression. The Annals of Statistics. https://www.jstor.org/stable/2242229

Stark, R., Grzelak, M., Hadfield, J., 2019. RNA sequencing: the teenage years. Nat Rev Genet 20, 631–656. https://doi.org/10.1038/S41576-019-0150-2

Stoler, N., Nekrutenko, A., 2021. Sequencing error profiles of Illumina sequencing instruments. NAR Genom Bioinform 3, 1–9. https://doi.org/10.1093/NARGAB/LQAB019

Sturm, G., Finotello, F., Petitprez, F., Zhang, J.D., Baumbach, J., Fridman, W.H., List, M., Aneichyk, T., 2019. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics 35, 436–445. https://doi.org/10.1093/BIOINFORMATICS/BTZ363

Sutton, G.J., Poppe, D., Simmons, R.K., Walsh, K., Nawaz, U., Lister, R., Gagnon-Bartsch, J.A., Voineagu, I., 2022. Comprehensive evaluation of deconvolution methods for human brain gene expression. Nat Commun 13, 1–18. https://doi.org/10.1038/s41467-022-28655-4

Swapna, L.S., Huang, M., Li, Y., 2023. GTM-decon: guided-topic modeling of single-cell transcriptomes enables sub-cell-type and disease-subtype deconvolution of bulk transcriptomes. Genome Biol 24. https://doi.org/10.1186/S13059-023-03034-4

Tomita, Y., Qiu, C., Bull, E., Allen, W., Kotoda, Y., Talukdar, S., Smith, L.E.H., Fu, Z., 2021. Müller glial responses compensate for degenerating photoreceptors in retinitis pigmentosa. Exp Mol Med 53, 1748. https://doi.org/10.1038/S12276-021-00693-W

Tsoucas, D., Dong, R., Chen, H., Zhu, Q., Guo, G., Yuan, G.C., 2019. Accurate estimation of cell-type composition from gene expression data. Nat Commun 10, 1–9. https://doi.org/10.1038/s41467-019-10802-z

Tung, P.Y., Blischak, J.D., Hsiao, C.J., Knowles, D.A., Burnett, J.E., Pritchard, J.K., Gilad, Y., 2017. Batch effects and the effective design of single-cell gene expression studies. Sci Rep 7, 1–15. https://doi.org/10.1038/srep39921

Veneziani, A.C., Gonzalez-Ochoa, E., Alqaisi, H., Madariaga, A., Bhat, G., Rouzbahman, M., Sneha, S., Oza, A.M., 2023. Heterogeneity and treatment landscape of ovarian carcinoma. Nature Reviews Clinical Oncology 2023 20:12 20, 820–842. https://doi.org/10.1038/s41571-023-00819-1

Wang, J., Roeder, K., Devlin, B., 2021. Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data. Genome Res 31, 1807–1818. https://doi.org/10.1101/GR.268722.120/-/DC1

Wang, X., Park, J., Susztak, K., Zhang, N.R., Li, M., 2019. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun 10, 1–9. https://doi.org/10.1038/s41467-018-08023-x

Wang, Z., Cao, S., Morris, J.S., Ahn, J., Liu, R., Tyekucheva, S., Gao, F., Li, B., Lu, W., Tang, X., Wistuba, I.I., Bowden, M., Mucci, L., Loda, M., Parmigiani, G., Holmes, C.C., Wang, W., 2018. Transcriptome deconvolution of heterogeneous tumor samples with immune infiltration. iScience 9, 451–460. https://doi.org/10.1016/j.isci.2018.10.028

Wery, M., Descrimes, M., Thermes, C., Gautheret, D., Morillon, A., 2013. Zinc-mediated RNA fragmentation allows robust transcript reassembly upon whole transcriptome RNA-Seq. Methods 63, 25–31. https://doi.org/10.1016/J.YMETH.2013.03.009

Zeng, Q., Mousa, M., Nadukkandy, A.S., Franssens, L., Alnaqbi, H., Alshamsi, F.Y., Safar, H. Al, Carmeliet, P., 2023. Understanding tumour endothelial cell heterogeneity and function from single-cell omics. Nat Rev Cancer 23, 544–564. https://doi.org/10.1038/s41568-023-00591-5

Zhang, Y., Parmigiani, G., Johnson, W.E., 2020. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinform 2. https://doi.org/10.1093/NARGAB/LQAA078

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., Gregory, M.T., Shuga, J., Montesclaros, L., Underwood, J.G., Masquelier, D.A., Nishimura, S.Y., Schnall-Levin, M., Wyatt, P.W., Hindson, C.M., Bharadwaj, R., Wong, A., Ness, K.D., Beppu, L.W., Deeg, H.J., McFarland, C., Loeb, K.R., Valente, W.J., Ericson, N.G., Stevens, E.A., Radich, J.P., Mikkelsen, T.S., Hindson, B.J., Bielas, J.H., 2017. Massively parallel digital transcriptional profiling of single cells. Nat Commun 8, 1–12. https://doi.org/10.1038/ncomms14049

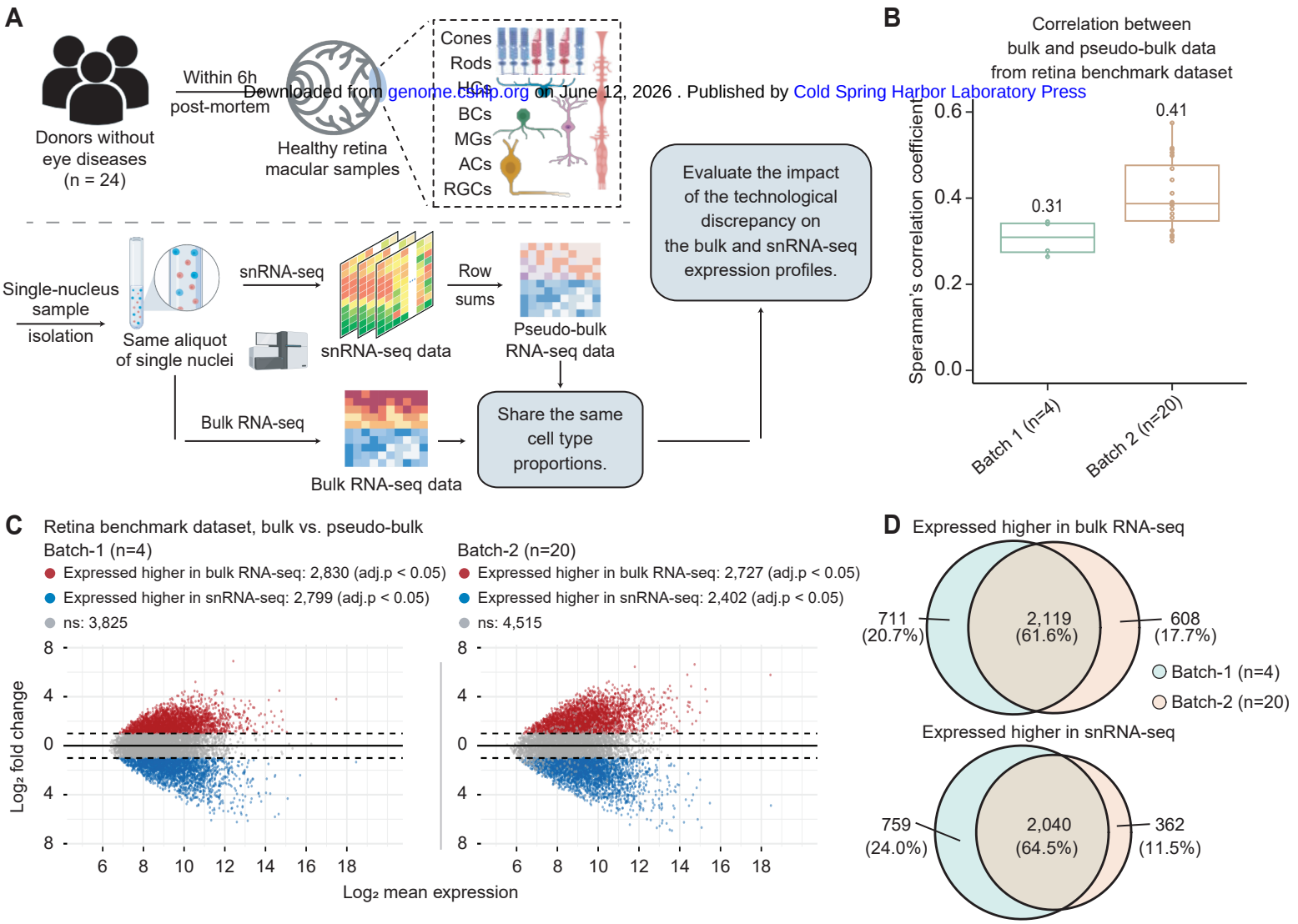

Figure 1. Assessing technological discrepancy between bulk and single-cell sequencing platforms using matched single-nucleus aliquots.

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 cells (RPEs). Samples are isolated into single-nucleus suspensions. The same aliquot of single-nucleus 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 discrepancy (i.e., the shot-gun sequencing procedure) on the bulk and snRNA-seq expression profiles. **B** and **C** show the influence of technological discrepancy at the sample and gene level, respectively. **B**, Spearman's correlation coefficient across genes between the matched real-bulk and pseudo-bulk RNA-seq data for one sample at a time for both batches. The correlations were calculated using quantile-normalized expression data (relative abundances). **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 2-fold change 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.

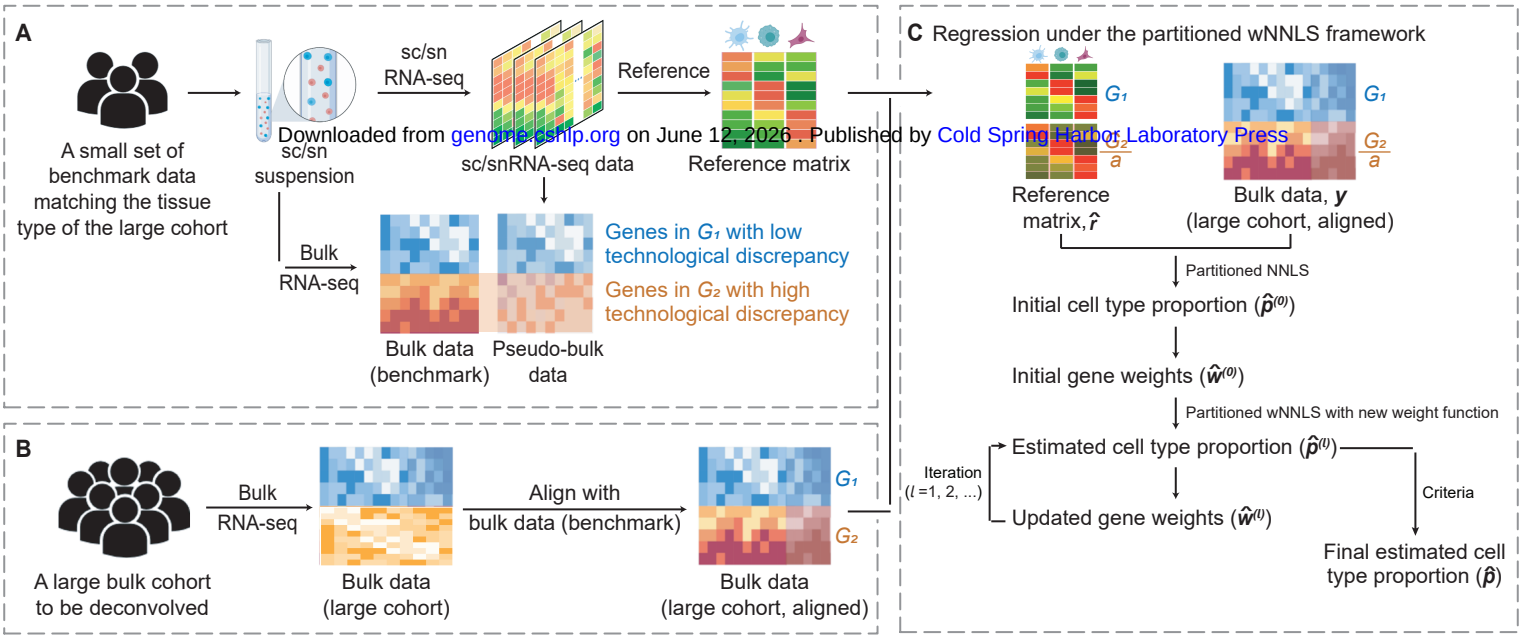

Figure 2. Overview of DeMixSC.
The DeMixSC framework for deconvolution analysis of bulk RNA-seq data using sc/sn RNA-seq data as a reference. **A,** The framework starts with a benchmark dataset of matched bulk and sc/snRNA-seq data with the same cell-type proportions. Pseudo-bulk mixtures are generated from the sc/sn data. DeMixSC identifies genes in G_1 and G_2 with 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 more impacted by the technological discrepancy (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 under a weighted nonnegative least squares (wNNLS) framework with two improvements: 1) partitioning and adjusting genes with high technological discrepancy, and 2) a new weight function. The final estimates are obtained when either the algorithm converges or reaches the prespecified maximum number of iterations. Here, G_1 is genes with low technological discrepancy, G_2 is genes with high technological discrepancy, a is a user-defined positive constant that serves as an adjustment factor, $\hat{\boldsymbol{r}}$ is the reference matrix derived from the sc/snRNA-seq data, $\boldsymbol{y}$ is the observed expression in bulk RNA-seq data, $\hat{\boldsymbol{p}}$ is the vector of estimated cell type proportions, and $\hat{\boldsymbol{w}}$ is the estimated gene weights.

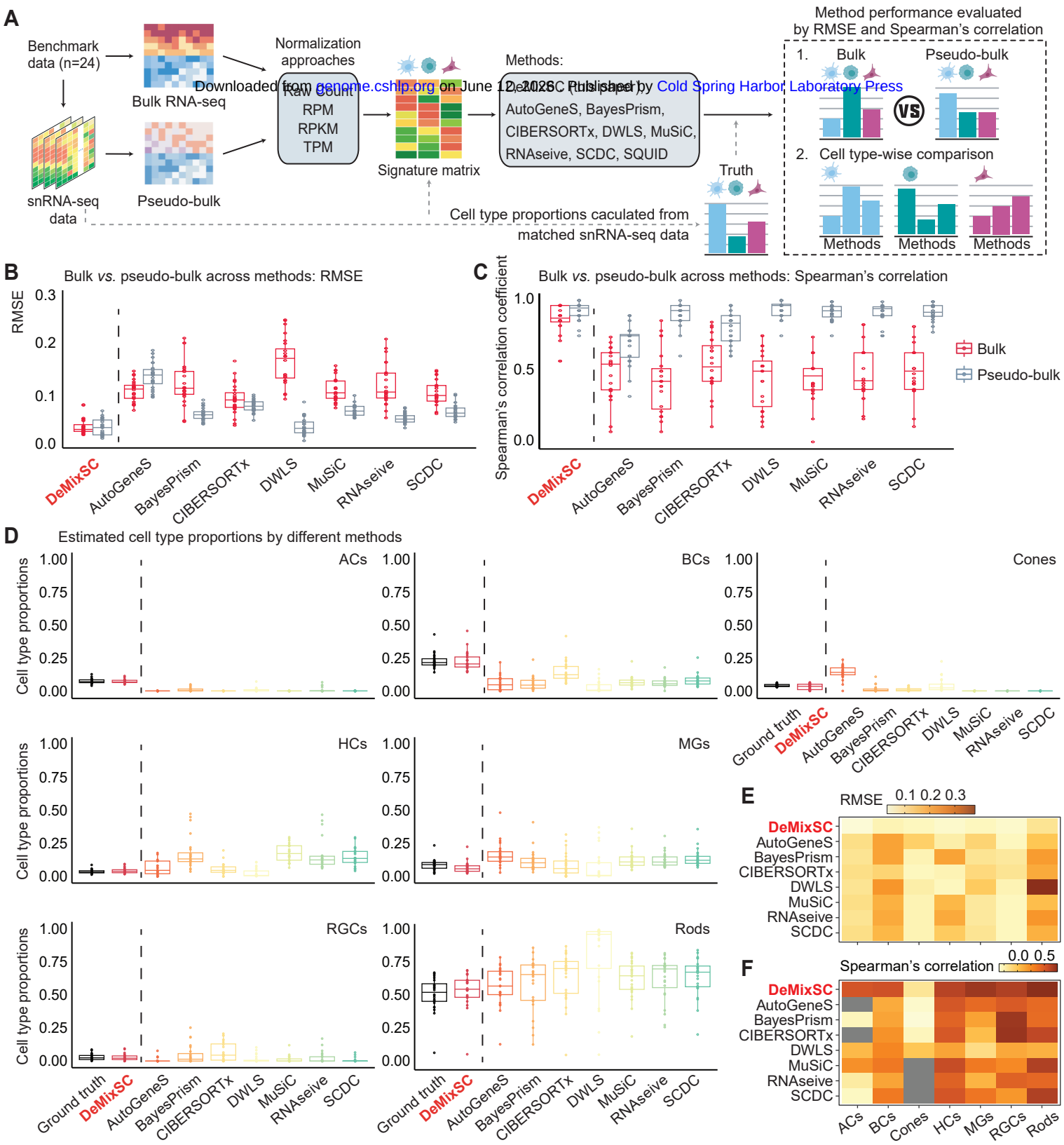

Figure 3. Compare the estimation accuracy of DeMixSC to existing deconvolution methods.
A, Workflow for the deconvolution benchmarking design. We use benchmark data from retinal samples. The cell count proportions for each cell type are used as ground truth for the corresponding tissue samples. We assess 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 Spearman's correlation coefficient. **B** and **C**, Boxplots showing the deconvolution performance of eight deconvolution methods for the bulk and pseudo-bulk data. RMSE and Spearman's correlation coefficient values are calculated across seven major cell types for each sample, with gray denoting pseudo-bulk and red denoting real-bulk. Smaller RMSEs or larger Spearman's correlations 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 Spearman's correlation coefficient, respectively. Lighter colors correspond to lower RMSE or Spearman's correlation coefficient values. Gray means NA.

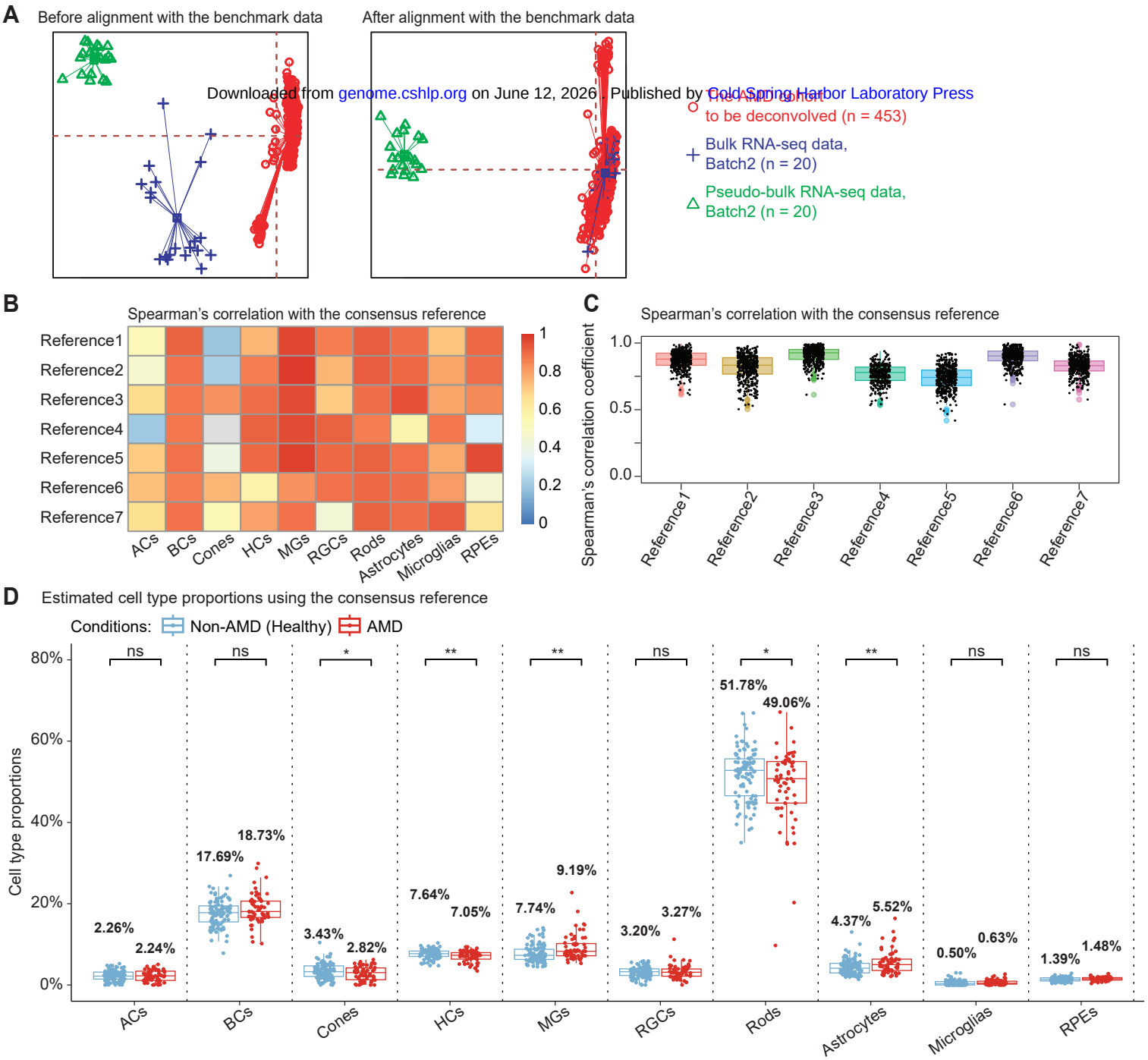

Figure 4. Using DeMixSC to deconvolve a large cohort of human peripheral retinal samples.
A, PCA plots of both the retina cohort data and the benchmark data. Red denotes the bulk data to be deconvolved, blue denotes the benchmark bulk data, and green denotes the benchmark pseudo-bulk data. **B** and **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 (healthy) 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.

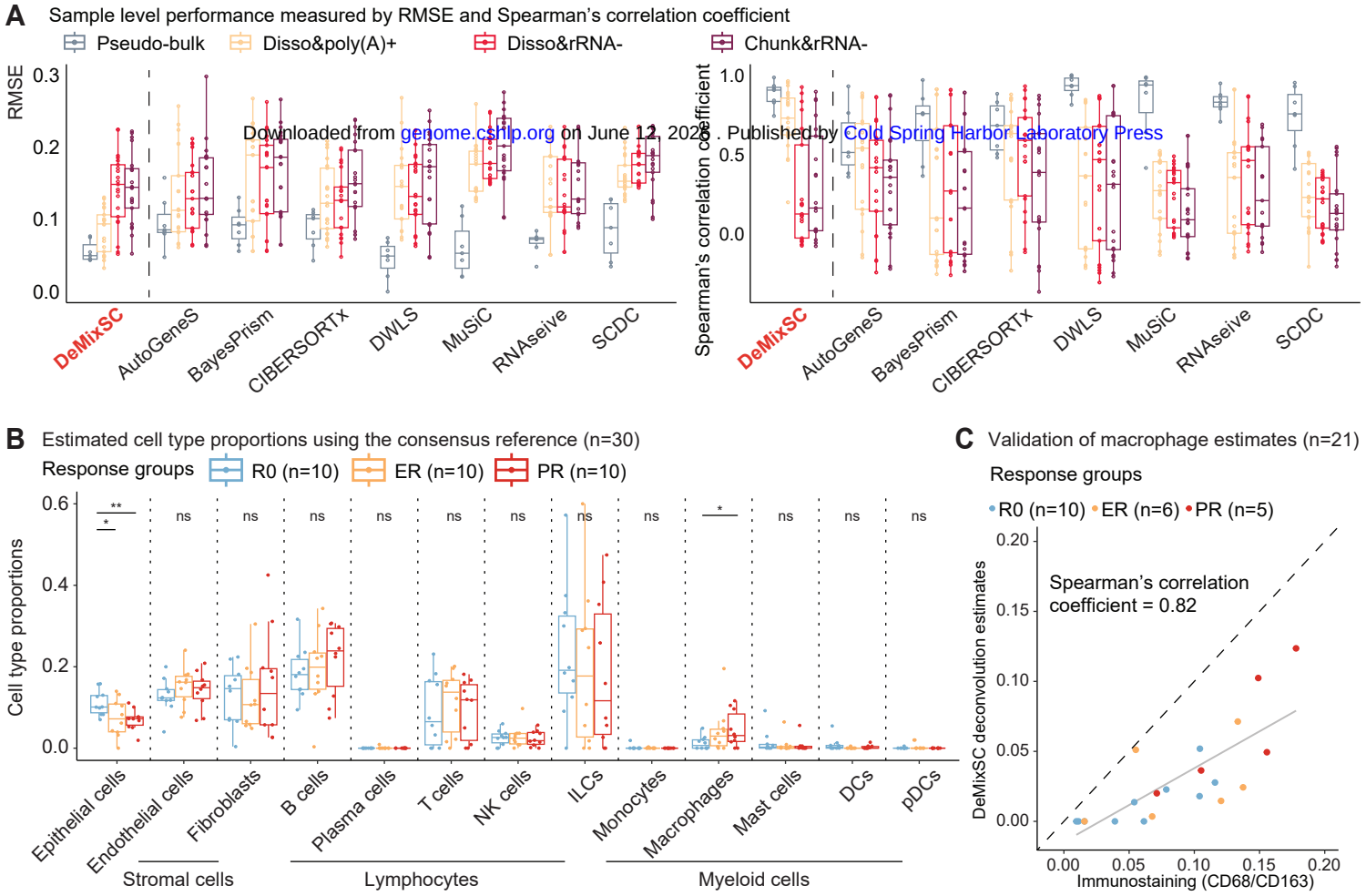

Figure 5. Using DeMixSC to deconvolve HGSC samples.
A, Boxplots showing the deconvolution performance of eight deconvolution methods for the pseudo-bulk and three types of bulk data in the HGSC benchmark dataset. RMSE values and Spearman's correlation coefficients are calculated across 13 cell types for each sample. Smaller RMSEs or larger Spearman's correlations indicate higher accuracy in proportion estimation. **B,** Distributions of DeMixSC estimated cell type proportions of *Lee et al.* data using consensus references. Each panel corresponds to a given cell type. NK cells: natural killer (NK) cells; ILC: innate lymphoid cells; DC: dendritic cells macrophages; pDC: plasmacytoid dendritic cells. The *P*-values for Student's *t*-tests comparing the estimated cell type proportions across R0, ER, and PR 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. **C,** Scatter plot comparing DeMixSC estimates of macrophages with experimental measurements from immunofluorescent (CD68/CD163) in 21 HGSC samples. The black dashed line represents the diagonal, while the grey solid line indicates the linear fit across the data points.