Imputation of cell-type specific expression using RNA-seq data from mixed cell populations

Wei-Yu Lin[1], Melissa Kartawinata[2,3], Bethany R Jebson[2,3], Restuadi Restuadi[2,3], CLUSTER Consortium, Lucy R Wedderburn[2,3,4], Chris Wallace[1,5]

[1]MRC Biostatistics Unit, East Forvie Building, Forvie Site Robinson Way, Cambridge Biomedical Campus, Cambridge, CB2 0SR, UK

[2]Infection, Immunity and Inflammation Research and Teaching Department, UCL Great Ormond Street Institute of Child Health, University College London (UCL), London, UK

[3]Centre for Adolescent Rheumatology Versus Arthritis at UCL University College London Hospital (UCLH) and Great Ormond Street Hospital (GOSH), University College London (UCL), London, UK

[4]National Institute for Health Research (NIHR) GOSH Biomedical Research Centre, London, UK

[5]Cambridge Institute of Therapeutic Immunology and Infectious Disease (CITIID), Jeffrey Cheah Biomedical Centre, Cambridge Biomedical Campus, University of Cambridge, Cambridge, UK

Abstract:

RNA-seq studies to infer differential gene expression (DGE) in disease or other conditions often use bulk sequencing of mixed cell populations because single cell or sorted cell sequencing may be prohibitively expensive. However, differential gene expression may be restricted to specific cell populations, meaning that mixed cell studies may lose power. Computational deconvolution can be used to estimate cell fractions from bulk expression data and may also infer average cell-type expression in a set of samples (eg cases or controls), but imputing sample-level cell-type expression is less commonly addressed.

In this study, we first assessed the accuracy of cell fractions estimated by three domain-specific methods-CIBERSORTx, bMIND and debCAM/swCAM using a real dataset where mixed peripheral blood mononuclear cells (PBMC) and sorted (CD4, CD8, CD14, CD19) RNA sequencing data as well as flow cytometry data were generated from the same subjects (N=158). Furthermore, we investigated the potential of cross-domain machine learning methods, multiple response LASSO and RIDGE, for imputing sample-level cell-type expression and compared their performance to the three deconvolution methods aforementioned.

Estimates of cell fractions by deconvolution methods for CD8, CD14 and CD19 had modest/high correlations r (0.62-0.77) with those from flow data but were less accurate for CD4 (0.43-0.54). All methods appeared to impute sample-level cell-type gene expression well across cell types (median r > 0.85); however, there was a high variation of correlations across subjects per gene. Nevertheless, LASSO/RIDGE exhibited marginally better accuracy than the deconvolution approaches. We considered an alternative measure of accuracy, differential gene expression (DGE) recovery, based on simulating case/control status and comparing detection of DGE between imputed and observed cell type data. We observed higher sensitivity but lower specificity of DGE recovery in LASSO than in deconvolution methods, although overall receiver operating characteristic (ROC) analysis revealed that LASSO/RIDGE had higher area under curves (AUC, median=0.84-0.87 across cell types) than CIBERSORTx (0.62-0.77), bMIND (0.69-0.76) and swCAM (0.64-0.72).

We conclude that machine learning methods have the potential to outperform domain-specific methods when suitable training data are available and suggest that further research in this area may optimise machine learning approaches to this problem.

Introduction

Tissues are a heterogeneous environment, comprised of various types of cell populations. In immune-mediated diseases, gene expression profiling of immune cells identifies subsets of genes characterising disease prognosis [1,2]. This approach enables better discrimination of disease pathogenesis than at mixed cell level [3] motivating study of immune cell transcriptomes in these diseases. Studying cell-type-specific expression has revealed gene expression signatures, e.g. CD8 T cell exhaustion, that predicted disease course [4]. However, flow sorting of target cells followed by RNA extraction for expression profiling of cell types in parallel, is labour- and resource-intensive. Single-cell RNA sequencing (scRNAseq) is more robust to many of these factors, but is expensive, especially in a large-scale study of many subjects [5,6]. These bottlenecks mean many studies of immune cells use mixed cell populations, such as peripheral blood mononuclear cells (PBMC), which might hinder the discovery of genes that exert their roles in a cell-type-specific manner.

Computational deconvolution of cell specific transcriptomes from mixed cell RNA-seq data provides an alternative to address this challenge. It is generally hypothesised that expression at a given gene in a mixed cell sample is the summation of its cell-type-specific expression weighted by corresponding cell fractions [7,8]

$$m = H \times f$$

where *m* is the vector of observed gene expression profiles of n genes, *H* is a latent *n* × *c* matrix representing gene expression profiles in each of *c* cell types and *f* is a vector of cell fractions [7]. The initial aim of most deconvolution approaches is to estimate *f*, and many deconvolution methods have been developed to solve this equation [9–23], divided into supervised and unsupervised types depending on whether *H* or *f* is used to guide deconvolution [7]. Fraction deconvolution methods rely on pre-computed cell type reference/signature gene expression (*H*) profiles, and differ in regression models/optimising strategies employed to minimise the sum of the squares between fitted expression and *m* [7], as well as their data preprocessing strategies and whether they allow for unknown cell types in the mixture. For example, both EPIC [12] and quanTIseq [15] employ constrained least square regression, FARDEEP implements adaptive least trimmed squares to automatically

detect and remove expression outliers that might lead to inaccurate f estimates [14] and CIBERSORT utilises linear nu-support vector regression that is robust to noise, unknown cell types in mixtures, and collinearity among closely related cell types in *H* [10].

CIBERSORTx extends the functionalities of CIBERSORT to estimate the average cell-type specific gene expression in a set of samples, *H1* ($n \times c$), using non-negative least squares (NNLS) regression and to estimate *f* [13]. Jaakkola and Elo [21] introduced Rodeo, which assumes the cell fraction is known, and then estimates H1 by fitting robust linear regression for each gene iteratively so that cell types with negative expression are excluded until regression coefficients for the remaining cell types are positive. csSAM estimates *H1* using standard least square regression, setting negative regression coefficients to zero [9]. Furthermore, if we write *M* and *F* for the matrix analogues of *m* and *f* for *k* samples, unsupervised deconvolution methods directly decompose *M* into *F* and *H1* simultaneously but require prior knowledge of the numbers of cell types and additional scRNAseq expression data for annotating the results [16,18,19,22]. Differentially expressed genes can then be identified by estimating *H1* separately by a condition of interest, e.g. disease status, with variance estimated via bootstrapping [6] or repeated *H1* deconvolution with permutation processes [9], but covariates or other conditions can not be taken into account in this way. In this case, sample-level cell type gene expression is needed but additional constraints or assumptions are needed to estimate the sample-level expression given only *M* and *F* [13,23].

We identified five existing methods that have developed strategies to impute sample-level cell-type gene expression: CIBERSORTx [13], CellR [22], MIND [17], bMIND [20] and swCAM [23]. CIBERSORTx assumes each gene can be analysed independently and some evidence of cell-type specific differential expression is detectable in bulk tissue [13]. For each gene, if significantly expressed in at least one cell type, CIBX applies bootstrapped NNLS iteratively to two subsets of mixture samples grouped by the similar expression for sample-level cell-type gene expression [13]. CellR models sample-level cell-type gene expression as a function of RNAseq read counts follows negative binomial distribution and infer sample-level cell-type expression using simulated annealing processes [22]. MIND implements an expectation-maximization

algorithm for sample-level cell-type gene expression by leveraging multiple transcriptomes from the same subjects [17]. bMIND [20], developed by the same authors of MIND, overcomes the limitation of multiple measurements per subject and uses a Bayesian mixed-effects model to estimate cell-type expression in each sample via Markov chain Monte Carlo sampling. swCAM [23] is built on debCAM [18], which does not require a signature matrix for fraction and average cell-type gene expression deconvolution in the convex analysis of mixtures framework. swCAM infers impute cell-type gene expression for each sample using low-rank matrix factorisation, assuming cell-type expression variations across samples resulted from a small number of cell-type specific functional modules, such as transcription factor regulatory network [23].

Prediction accuracy of cell fractions among deconvolution methods and factors affecting the performance have been extensively investigated based on synthetic data from scRNAseq data sampled with designated cell proportions [24–26]. However, less is known about the accuracy of deconvolution-based approaches in imputing cell-type gene expression at the sample level. Here, we use RNA-seq data from mixed and sorted cell populations from the same individuals to examine the accuracy of CIBERSORTx, bMIND and swCAM and compare to alternative machine learning methods, multivariate LASSO and RIDGE, that are not dependent on the deconvolution equation. We excluded CellR, which relies on predefined cell-type clusters in scRNAseq data, and MIND, developed by the same authors of bMIND, which requires multiple bulk expression measurements per subject.

Results

The CLUSTER Consortium aims to use immune cell RNA-seq data to find transcriptional signatures which predict treatment response in childhood arthritis. In order to balance the competing goals of maximising both the number of patients studied and the number of cell specific assays from each patient within a fixed budget, we designed an RNA-sequencing experiment with PBMC samples included from all available subjects and specific immune cells from a subset of subjects, with the aim to use the subjects with coverage of both to learn rules to impute cell specific gene expression into the complete dataset. This design also allowed us to split our

data into training (80 samples) and testing (between 52 and 71 samples depending on cell type) sets (Figure 1a) to compare the performance of potential imputation approaches.

Estimation accuracy of sample-level cell frequencies

CIBERSORTx, equipped with an inbuilt leukocyte gene signature matrix LM22, also allows the creation of custom gene signatures. We created a custom signature using our sorted cell expression in the training set (Figure 1b). We deconvoluted CD4, CD8, CD14, and CD19 cell fractions from PBMC mixed cells based on different scenarios: CIBERSORTx with inbuilt (CIBX-inbuilt) and custom (CIBX-custom) matrices, bMIND using the custom matrix (bMIND-custom), debCAM using cell-type specific markers derived from expression in our sorted cell populations (debCAM-custom) and compared estimates of cell fractions to measures of ground truth derived from flow cytometry (Figure 1b).

Correlations were highest in CD14, followed by CD19, CD8, and CD4 regardless of methods and signature matrices (Figure 2). Generally speaking, CIBX-custom performed less well across all four cell types than the other three approaches, while CIBX-inbuilt and debCAM performed the best, although the exact ordering did vary between cell types. This difference between CIBX-custom and CIBX-inbuilt emphasises the importance of a well trained gene expression signature matrix. We note that CD14 predicted fractions were over-estimated regardless of approach and CD4 generally under-estimated (Figure 2).

Estimation accuracy of sample-level cell type gene expression

Our main goal was to compare accuracy in imputing sample-level cell-type gene expression from PBMCs for the four sorted cell populations: CD4 T cells, CD8 T cells, CD14 monocytes and CD19 B cells. In addition to predicted expression by CIBERSORTx using inbuilt and custom signature matrices, we derived cell-type expression profiles using true cell fractions (estimated by flow cytometry) with bMIND and swCAM algorithms (Figure 1b). Finally, we trained regularised multivariate LASSO and RIDGE models to predict cell-type expression using all genes. While CIBERSORTx predicts only a subset of the most confident genes (5881-11795

depending on cell type), all other methods were able to predict expression for all or almost all 18871 genes across cell types (Supplementary Figure 3).

Despite the differing number of imputed genes among methods (Supplementary Figure 3), the initial analysis comparing observed and predicted expression of genes in the same subjects suggested that all methods could predict cell-specific expression well, as judged by high correlations (median r > 0.85) in the test data, although correlations were generally higher and root mean square error (RMSE) lower in LASSO and RIDGE than the other approaches (Figure 3a,b). To better interpret the correlations, we also calculated a “baseline” correlation between observed expression in one individual and estimated expression in the same cell type from a different individual. These were also high, and only marginally lower than correlations between observed and estimated expression within the same individuals, limiting the utility of this measure to discriminate among methods (Supplementary Figure 8).

We therefore complemented this with a comparison of the observed and predicted expression across subjects for each gene. Correlation varied considerably between genes, irrespective of approaches (Figure 3c). All methods had comparable correlation per gene for each cell type, except for swCAM, which exhibited suboptimal performance for CD8, CD14, and CD19 (Figure 3c). Similar RMSE per gene was seen for each cell type across methods, although LASSO and RIDGE had slightly lower median values than other approaches (Figure 3d).

Despite correlation and RMSE being commonly used to assess predictive accuracy, they do not necessarily capture performance of predictions in intended downstream analyses. We therefore defined a new measure of performance, differential gene expression (DGE) recovery, which used simulated phenotypes that deliberately correlated with observed expression across a subset of genes and conducted DGE analysis in parallel using predicted and observed cell-type data. DGE recovery measured the degree to which significant and non-significant signals in the observed data could be correctly identified in the imputed data (Figure 4a,4b). According to this measure, LASSO and RIDGE exhibited similar median values for the area under the receiver operating characteristic curve (AUCs), ranging from 0.84-0.87 across cell types, which were higher than the AUCs achieved by CIBERSORTx with inbuilt

and custom, bMIND and swCAM, which had AUCs ranging from 0.62-0.72, 0.70-0.77, 0.69-0.76 and 0.64-0.72, respectively (Figure 4c). The results suggest that regularised multivariate models performed better than the other three deconvolution-based methods. More detailed examination showed that, generally, LASSO and RIDGE had higher sensitivity than CIBERSORTx, bMIND and swCAM, but also lower specificity (Supplementary Figure 5a). In all cases, imputed estimates of $\log_2$ fold changes were attenuated in the imputed data, with average slopes of 0.60-0.70 in CIBERSORTx inbuilt, 0.64-0.76 in CIBERSORTx custom, 0.66-0.84 in bMIND, 0.55-0.90 in swCAM, 0.69-0.76 in LASSO and 0.68-0.76 in RIDGE (Supplementary Figure 5c).

Discussion

Using PBMC RNA-seq, sorted-cell RNAseq, and flow cytometry data from the same individuals, our study investigated the accuracy of estimates of cell type fractions by the state-of-the-art domain-specific tool. The composition of cell types in the signature matrix is the key to deconvolution performance [7,26]. CIBERSORTx's better estimates of cell fractions using the inbuilt signature than our custom one (Figure 2), CD19 in particular, might be due to a lack of crucial cell types in our custom matrix. It has been shown that CD19 B cell prediction accuracy was affected by missing monocytes/macrophages [27], which expressed CD14 presented in the custom signature. We found comparable correlations of inbuilt B cell naive (r=0.55) and plasma cell (r=0.47) with ground-truth CD19 fractions (Supplementary Figure 4). Our sorted CD19 expression is a weighted expression of various immune cells, including B and plasma cells. Custom CD19 signature genes derived from this sorted expression differentiated CD19 cells from three other cell types (Supplementary Figure 2e). However, they lacked CD19 sub-lineage-specific markers that could distinguish CD19 sub-cell lineages. These suggested plasma cells might be a crucial cell type absent from our custom signature. Notably, bMIND and debCAM, which utilised signature and cell-type specific genes from our four sorted cell populations, yielded results comparable to CIBERSORTx using its inbuilt signature matrix, suggesting that different deconvolution algorithms can overcome the absence of certain cell types in training data. All methods performed least well for CD4. For CIBERSORTx, closer inspection of fraction correlations between inbuilt CD4 cell

lineages with ground-truth CD4 found the highest r of 0.37 for T cells regulatory (Tregs) followed by T cells CD4 naive of 0.33 and low correlations (|r| < 0.1) for other 3 T cell subsets (Supplementary Figure 4), suggesting that merging CD4 cell lineage as CD4 was not the driver of the moderate correlation observed. Moderate correlations of CD4 by CIBERSORTx were consistent with the observations from Sturm et al. (2019), in which CD4+ correlation accuracy by CIBERSORT ranged from 0.36 to 0.44 in the simulated and actual RNA-seq validation data [25]. In addition, CIBERSORTx provided the most accurate estimates for CD8, despite not performing as well for CD4 compared to other methods. On the other hand, debCAM provided the best CD4 estimates but was less accurate for CD8. This suggests that accurately estimating fractions of these two related cell types remains challenging, possibly due to shared signature genes or a limited number of specific cell-type genes (Supplementary Figures 2a and 2c).

We provide a real data comparison of sample-level cell type specific expression imputation, including off-the-shelf machine learning methods, multivariate LASSO and RIDGE, as comparators. Correlation has been used to evaluate the accuracy of predicted cell-type expression, and good correlations per subject have been reported [13,20,23], consistent with our observations. However, we also found high correlations in between-subject comparisons (Supplementary Figure 8), which presumably reflects that cell type explains the greatest proportion of variability in gene expression. Good correlations at the sample level might not necessarily reflect accuracy at the gene level, as evidenced by low to moderate correlation per gene observed in our study, which was consistent with the findings of bMIND [20] and swCAM [23]. These suggest correlation is not an optimal measure of performance. In contrast, our proposed DGE recovery measure, which mimics DGE analysis and measures the capability to reconstruct DGE signals, could be more indicative than correlation. We observed better accuracy using LASSO and RIDGE than in three deconvolution-based approaches (CIBERSORTx, bMIND and swCAM), particularly based on DGE recovery that reflected the reconstruction of the DGE signals.

To impute cell-type gene expression for samples, deconvolution methods first need an estimate of cell type fractions. CIBERSORTx estimates these from the RNA-seq data, and had good accuracy albeit with notable underestimation of the fraction of CD4+ T cells, while bMIND and swCAM utilised our flow cytometry measured cell type fractions, which we expect to be more accurate. We might expect that using this additional data would allow bMIND and swCAM to be more accurate, but DGE recovery was comparable across deconvolution methods, although the number of genes with estimated expression does vary between methods (lower for CIBERSORTx) . In contrast, LASSO and RIDGE, forms of penalised linear regression, use a one-step approach that does not rely on estimated cell fractions. Instead, it learns directly from a training set of PMBC and cell-type expression data. Rather than treat each gene as an independent problem, we used multivariate LASSO/RIDGE, batching genes with correlated expression in the target cell type to enable the solution for each gene in a batch to share information about which PMBC genes were important predictors.

Nonetheless, there are limitations in LASSO/RIDGE. Most obviously, LASSO/RIDGE requires a training dataset, consisting of bulk and cell-type gene expression data from the same subjects to train the model, unlike deconvolution-based methods that do not need such data. Moreover, LASSO/RIDGE demands high computational resources. Regarding CPU running time, LASSO and RIDGE take 17 or 93 times as long as the fastest method CIBERSORTx. While bMIND is only 3 times slower than CIBERSORTx, swCAM is 258 times slower (Supplementary Table 1). Furthermore, LASSO and RIDGE require 88 or 306 times more memory usage than CIBERSORTx, while bMIND and swCAM only need 18% or 22% of CIBERSORTx's memory usage (Supplementary Table 1). Full details of the running time and RAM usage are shown in the online execution report, https://b8307038.gitlab.io/deconvimpvexpr/nextflow_report.html. We also note that predicted fold changes using imputed expression systematically shrunk the fold changes estimated from observed data across all methods, so that DGE analysis using imputed data can only be supported for detection of differentially expressed genes and direction of differential expression, not for unbiased estimation of fold changes.

LASSO and RIDGE are the machine learning methods we considered, and we have not attempted to optimise their performance. There is presumably potential to improve performance further, with consideration of how genes are batched for prediction, or by considering other approaches. Their better performance should motivate further exploration of non-domain specific methods in this space.

Methods & Materials

Study subjects

Data utilised in the study came from 158 subjects recruited in the CLUSTER consortium. Around 80% of subjects (N=126) have juvenile idiopathic arthritis (JIA); the rest are healthy controls from adults and children, and about 58% are female (N=91). Peripheral blood samples were obtained in accordance with the ethics approved by the London-Bloomsbury Research Ethics Committee (REC 05/Q0508/95, 95RU04, and 11/LO/0330) with full informed consent and age-appropriate assent. The diagnosis for JIA followed the internationally agreed classification as described in [28].

Blood was collected in a heparinised tube and peripheral blood mononuclear cells (PBMC) were isolated by density gradient centrifugation with Lymphoprep™ (Stem Cell Technologies). The blood samples were collected from the JIA patients at different time points of treatment.

Cell sorting

Isolated PBMC were sorted by cell sorter (BD FACSAria™ III, BD Biosciences) into different cell populations with CD4-BV711 (clone OKT4, Biolegend 317440), CD8-APC (clone SK1, Biolegend 344722), CD14- FITC (clone 61D3, eBioscience 11-0149-42), and CD19-PE-Cy7 (clone HIB19, Biolegend 302216). CD3-BV605 (clone OKT3, Biolegend 317322) was used to differentiate between T cell and non-T cell populations. Dead cells were excluded before sorting using 4,6-diamidino-2-phenylindole (DAPI; Sigma). Sorted cell purity was accessed and on

average was>90%. For each subject, we divided CD4, CD8, CD14, and CD19 cell counts by the sum of CD4, CD8, CD14 and CD19 cells to obtain cell-type fractions.

RNA sequencing & data processing

Unsorted PBMC and sorted immune cells were extracted with PicoPure™ RNA Isolation Kit (Applied Biosystems™, KIT0204). The extracted RNA samples were sent to UCL Genomics for library preparation and sequencing.

RNA sequencing was carried out in four batches using Illumina NovaSeq6000. PBMC and sorted cell RNAseq data were processed using the RSSnextflow (Resources), an RNAseq pipeline customised for unique molecular identifiers (UMIs) tagged RNAseq data built under the Nextflow framework [29]. Briefly, sequencing reads (2x100bp) were mapped to the reference genome GRCh38 using STAR aligner [30]. Two-passing mapping mode was used, with gene annotated features (Homo_sapiens.GRCh38.103.gtf) and the options of --twopassMode Basic and --sjdbOverhang 99. Default parameters were used unless otherwise specified. Read PCR duplicates were identified based on alignment coordinates and up to 1 mismatched UMI sequence using the je suite tool [31]. After deduplication, aligned reads were summarised over the gene features using the featureCounts programme [32], and a read count table was generated for each batch.

We selected RNA samples that have RIN ≥ 5.0 and library concentration ≥ 4.5nM (~0.7ng/μL) for RNA sequencing. Illumina TruSeq mRNA stranded v2 and Roche Kapa mRNA HyperPrep were used to create libraries. Samples from the same subjects were sequenced in the same batch, and we employed Combat-seq [33] to minimise batch effects (Supplementary Figure 6). Read counts from across the four batches were analysed together. A total of 723 RNAseq samples from 158 subjects with data on PBMC and at least one cell type were used in the downstream analysis. We filtered out genes with counts-per-million expression <0.836, equivalent to 10 read counts in our median library size of 11.95 million reads, in less than 96 samples. Also, genes were excluded if their total read counts across samples were less than 15. These filtering steps were conducted using the edgeR filterByExpr function [34,35], with cell type information as the group argument. After filtering out low expressed genes, transcripts per million (TPM), as recommended by the authors of CIBERSORT [36], was estimated and utilised as observed expression.

Training/testing set

Of 158 subjects with PBMC gene expression data, 80 had complete RNAseq data on all four sorted cells and PBMCs and formed the training set. The remaining 78 had data on PBMCs and partially complete gene expression data in sorted cells and were used as test samples, with numbers for each cell type of CD4: 71, CD8: 65, CD14: 52, CD19: 57 (Figure 1a). Sequencing batch did not differ significantly between training and testing sets (Chi-squared test $p > 0.05$, Supplementary Figure 1).

Signature gene matrices & cell-type specific genes

Two signature gene matrixes were utilised. CIBERSORTx in-built LM22 was derived from microarray gene expression in 22 purified leukocyte subsets [10]. We constructed a custom signature based on CD4, CD8, CD14, and CD19 TPM in the training subjects using the CIBERSORTxFractions module, with the default settings of "--G.min 300 --G.max 500 --q.value 0.01 --QN FALSE --single_cell FALSE" for sorted RNAseq. There were fewer signature genes in the inbuilt matrix (N=547) than the custom one (N=1589), which we attributed to the dynamic ranges of gene expression measured by two different platforms, microarray for inbuilt and RNAseq for custom (Supplementary Figures 2a and 2b).Sorted cell expression in training samples was also used in debCAM to identify cell-type specific genes for fraction deconvolution. debCAM OVE.FC (one versus everyone - fold change) criteria of 1, 2, 5 and 10 were used (Supplementary Figure 2c), and we selected 1247 cell-type specific genes from OVE.FC of 10 because this number was comparable to custom signature genes. Despite differences inbuilt/custom signature and debCAM cell-type genes separated testing samples well based on their cell types (Supplementary Figures 2d, 2e and 2f).

CIBERSORTx analysis

We ran CIBERSORTx locally with a token requested from the CIBERSORTx website (Resources). We ran the CIBERSORTxHiRes module for deconvolution, with RNAseq default settings and "--variableonly TRUE" for only outputting genes with variation in expression across subjects. When the LM22 signature was applied, two

additional arguments of “--classes” and “--rmbatchBmode” were specified to aggregate cell type expression for 11 major leukocytes based on the shared lineage of 22 leukocytes [10] and minimise measuring variations in gene expression introduced by platforms, respectively.

CIBERSORTx fractions of CD4, CD8, CD14, and CD19 for LM22 were derived from the sum of proportions in their shared-lineage leukocytes (Supplementary Figures 2a and 4). CD4 was the sum of the proportions of T cells CD4 naive, T cells CD4 memory resting, T cells CD4 memory activated, T cells follicular helper, and T cells regulatory (Tregs). T cells CD8 was used as CD8 fraction. CD14 fraction was the sum of Monocytes, Macrophages M0, Macrophages M1, and Macrophages M2, and B cells naive and B cells memory was added together for CD19 (Supplementary Figures 2a and 4). We then scaled the resultant proportions to the sum of 1. Imputed cell-type expression was log2 transformed for downstream evaluation.

bMIND and debCAM/swCAM prediction

bMIND cell fraction deconvolution was carried out using the authors’ bMIND function with our custom signature. Cell-type specific genes that were 10-fold over-expressed in one cell type compared to others, as previously described, were specified in debCAM AfromMarkers function for estimating cell fractions from PBMC mixture.

We followed the authors’ instructions to run bMIND and swCAM in a supervised mode on true cell fractions as measured by flow cytometry in the same samples to predict cell type gene expression for samples. More specifically, bMIND predicted expression profiles were obtained from the bMIND function of the MIND package [20], which took log2(TPM+1) transformed PBMC expression and cell fractions as inputs.

swCAM consisted of two steps. In the fine-tuning step, we conducted 10-fold cross-validation, randomly removing one-tenth of gene expressions from the sample and gene expression matrix followed by imputing back expression missingness, to determine the optimal lambda with the minimum of RMSE between missing and imputed expression. The R script, script-swCAM-cv.R, was used. In the predicting step, lambda of 800, PBMC TPM expression, true cell fractions, and grouped cell expressions (cell types * gene matrix) derived from NNLS were used in the sCAMfastNonNeg function for imputing sample-wise cell type expression. All the R

scripts and functions related to swCAM were obtained from the authors' GitHub repository https://github.com/Lululuella/swCAM [23]. We then log2 transformed the estimates of cell type expression for downstream analysis.

LASSO/RIDGE training and prediction

A penalised multi-response linear regression (LASSO/RIDGE) was performed on log2(TPM+1) in the training set for each cell type, with PBMC and cell-type expression as predictors and dependent variables, respectively. To ease the computational burden, we clustered genes into chunks on the basis of their expression profiles (in log2 TPM), with a size of up to 500 genes. Specifically, we performed a hierarchical clustering analysis on Euclidean distances between genes for each cell type in the training data. We grouped clustering dendrograms into the initial number of chunks, the minimum multiple of 500 to include all genes, using the R cutree function. For each chunk exceeding the size of 500 genes, we repeated clustering analysis on genes in a given chunk, followed by dendrogram grouping using the cutreeDynamic function [37] specified with "minClusterSize" of 250, if necessary, reducing by steps of 5, until all the resultant chunks met the desired size of < 500.

The numbers of chunks (sizes) were 74 (49-481) for CD4, 71 (71-493) for CD8, 76 (76-496) for CD14, and 83 (41-486) for CD19, respectively. For each chunk, the fine-tuned LASSO/RIDGE model from a 5-fold cross-validation based on the mean squared error (MSE) criterion was used to impute cell-type expression in the testing samples. Penalised modelling was carried out using the R glmnet package [38], for which"family=mgaussian", "type.measure =mse", and alpha=1 for LASSO /alpha=0 for RIDGE were specified in the cv.glmnet function for model training, and predict function was used for imputation.

Performance measurement

Predicted cell fractions were compared to the flow cytometry estimates described above, and Pearson r correlations and root mean square errors (RMSE) were calculated for measuring performance. We calculated correlations and RMSE per gene between imputed and observed expression to evaluate and benchmark the

predicted accuracy of expression among CIBERSORTx using custom and inbuilt, bMIND using custom, swCAM with cell-type specific markers, LASSO and RIDGE.

We simulated ten phenotypes that correlated with gene expression in the observed cell type data, including testing samples, using principal component analysis (PCA). Ten simulated phenotypes corresponded to the first ten principal components (PC); for each PC, samples with PC >0 were designated as 1 for pseudo-cases; otherwise, 0 for pseudo-controls. DGE analysis was carried out twice in parallel under the limma framework [39]:

1. in all observed data (test + training)
2. In observed data from the training samples + imputed data from the test samples which had observed data for comparison

We treated (1) as the ground truth, and (2) as the likely analysis adopted in any real world study. Dimensions of expression matrix used in (1) were the same as (2) in limma DGE analysis. A false discovery rate (FDR) of 0.05 was set as the significance level for true signals in the observed data. We varied FDRs in the imputed data from 0 to 1 by steps of 0.05. Receiver operating characteristic analysis was carried out by cell type and combinations of method and scenario using the pROC package [40].

All analyses were performed with R-4.1.2 [41] unless otherwise stated.

Data availability

The Metadata and processed data that support the findings of this study are available in Zenodo (https://doi.org/10.5281/zenodo.10000430). Source data are provided with this paper.

Author contributions

W-Y.L. analysed the data, prepared figures and wrote the manuscript. M.K. performed experiments and wrote the manuscript. B.R.J performed experiments. R.R. contributed to data QC. L.R.W conceived and supervised the project. C.W designed and supervised the project and drafted the manuscript.

Resources

CLUSTER consortium-Childhood arthritis and its associated uveitis: stratification through endotypes and mechanism to deliver benefit, https://www.clusterconsortium.org.uk/

RSSnextflow workflow for processing RNAseq FASTQ files to generate analysis-ready read counts https://gitlab.com/b8307038/rssnextflow

CIBERSORTx, https://cibersortx.stanford.edu/

bMIND, https://github.com/randel/MIND

swCAM, https://github.com/Lululuella/swCAM

Code availability

Nextflow workflow and R markdown file to run the analyses, to generate and to summarise results in this work presented here,

https://gitlab.com/b8307038/deconvimpvexpr

References

1. Lee, J. C. *et al.* Gene expression profiling of $CD8^+$ T cells predicts prognosis in patients with Crohn disease and ulcerative colitis. *J. Clin. Invest.* **121**, 4170–4179 (2011).
2. McKinney, E. F. *et al.* A CD8+ T cell transcription signature predicts prognosis in autoimmune disease. *Nat. Med.* **16**, 586–591 (2010).
3. Lyons, P. A. *et al.* Novel expression signatures identified by transcriptional analysis of separated leucocyte subsets in systemic lupus erythematosus and vasculitis. *Ann. Rheum. Dis.* **69**, 1208–1213 (2010).
4. McKinney, E. F., Lee, J. C., Jayne, D. R. W., Lyons, P. A. & Smith, K. G. C. T-cell exhaustion, co-stimulation and clinical outcome in autoimmunity and infection. *NATURE* vol. 523 612+ (2015).

5. Sturm, G., Finotello, F. & List, M. In Silico Cell-Type Deconvolution Methods in Cancer Immunotherapy. *BIOINFORMATICS FOR CANCER IMMUNOTHERAPY: Methods and Protocols* vol. 2120 213–222 (2020).
6. Steen, C. B., Liu, C. L., Alizadeh, A. A. & Newman, A. M. Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx. *STEM CELL TRANSCRIPTIONAL NETWORKS: METHODS AND PROTOCOLS, 2ND EDITION* vol. 2117 135–157 (2020).
7. Cobos, F. A., Vandesompele, J., Mestdagh, P. & De Preter, K. Computational deconvolution of transcriptomics data from mixed cell populations. *BIOINFORMATICS* vol. 34 1969–1979 (2018).
8. Finotello, F. & Trajanoski, Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. *Cancer Immunol. Immunother.* **67**, 1031–1040 (2018).
9. Shen-Orr, S. S. *et al.* Cell type–specific gene expression differences in complex tissues. *Nat. Methods* **7**, 287–289 (2010).
10. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *NATURE METHODS* vol. 12 453+ (2015).
11. Li, B. *et al.* Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* **17**, 174 (2016).
12. Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* **6**, e26476 (2017).
13. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *NATURE BIOTECHNOLOGY* vol. 37 773+ (2019).
14. Hao, Y., Yan, M., Heath, B. R., Lei, Y. L. & Xie, Y. Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed

squares. *PLOS Comput. Biol.* **15**, e1006976 (2019).

15. Finotello, F. *et al.* Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Med.* **11**, 34 (2019).
16. Kang, K. *et al.* CDSeq: A novel complete deconvolution method for dissecting heterogeneous samples using gene expression data. *PLOS Comput. Biol.* **15**, e1007510 (2019).
17. Wang, J., Devlin, B. & Roeder, K. Using multiple measurements of tissue to estimate subject- and cell-type-specific gene expression. *Bioinformatics* **36**, 782–788 (2020).
18. Chen, L. *et al.* debCAM: a bioconductor R package for fully unsupervised deconvolution of complex tissues. *Bioinformatics* **36**, 3927–3929 (2020).
19. Kang, K., Huang, C., Li, Y., Umbach, D. M. & Li, L. CDSeqR: fast complete deconvolution for gene expression data from bulk tissues. *BMC Bioinformatics* **22**, 262 (2021).
20. Wang, J., Roeder, K. & Devlin, B. Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data. *Genome Res.* gr.268722.120 (2021) doi:10.1101/gr.268722.120.
21. Jaakkola, M. K. & Elo, L. L. Computational deconvolution to estimate cell type-specific gene expression from bulk data. *NAR Genomics Bioinforma.* **3**, lqaa110 (2021).
22. Doostparast Torshizi, A., Duan, J. & Wang, K. A computational method for direct imputation of cell type-specific expression profiles and cellular compositions from bulk-tissue RNA-Seq in brain disorders. *NAR Genomics Bioinforma.* **3**, lqab056 (2021).

23. Chen, L. *et al.* swCAM: estimation of subtype-specific expressions in individual samples with unsupervised sample-wise deconvolution. *Bioinformatics* **38**, 1403–1410 (2022).
24. Jin, H. & Liu, Z. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *GENOME BIOLOGY* vol. 22 (2021).
25. Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *BIOINFORMATICS* vol. 35 I436–I445 (2019).
26. Cobos, F. A., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *NATURE COMMUNICATIONS* vol. 11 (2020).
27. Schelker, M. *et al.* Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat. Commun.* **8**, 2032 (2017).
28. Petty, R. E. *et al.* International League of Associations for Rheumatology classification of juvenile idiopathic arthritis: second revision, Edmonton, 2001. *J. Rheumatol.* **31**, 390–392 (2004).
29. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
30. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
31. Girardot, C., Scholtalbers, J., Sauer, S., Su, S.-Y. & Furlong, E. E. M. Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. *BMC Bioinformatics* **17**, 419 (2016).
32. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

33. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics Bioinforma.* **2**, lqaa078 (2020).

34. Chen, Y., Lun, A. & Smyth, G. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline [version 2; peer review: 5 approved]. *F1000Research* **5**, (2016).

35. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

36. Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M. & Alizadeh, A. A. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. *CANCER SYSTEMS BIOLOGY: METHODS AND PROTOCOLS* vol. 1711 243–259 (2018).

37. Langfelder, P., Zhang, B. & Horvath, with contributions from S. *dynamicTreeCut: Methods for Detection of Clusters in Hierarchical Clustering Dendrograms*. (2016).

38. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).

39. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

40. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).

41. R Core Team. R: A Language and Environment for Statistical Computing. (2021).

Acknowledgements

This research was funded by the MRC (MR/R013926, MC_UU_00002/4), Wellcome Trust (WT220788) and the NIHR Cambridge Biomedical Research Centre

(BRC-1215-20014). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

CLUSTER is supported by grants from the Medical Research Council (MRC) [MR/R013926/1] and Versus Arthritis [Grant: 22084], Great Ormond Street Hospital Children's Charity [VS0518], and Olivia's Vision. This work is supported by the NIHR GOSH Biomedical Research Centre, the NIHR Manchester Biomedical Research Centre, and the British Society for Rheumatology (BSR), and the "UK's Experimental Arthritis Treatment Centre for Children, supported by Versus Arthritis (grant: 20621)". The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. Wedderburn is additionally supported by Versus Arthritis (grant: 21593) at the Centre for Adolescent Rheumatology Versus Arthritis. Hyrich and Thomson are additionally supported by the Centre for Epidemiology Versus Arthritis (grant: 21755) and the Centre for Genetics and Genomics Versus Arthritis (grant: 21754) at the University of Manchester, UK. This study acknowledges the use of the following UK JIA cohort collections: The Biologics for Children with Rheumatic Diseases (BCRD) study (funded by Arthritis Research UK grant: 20747); The British Society for Paediatric and Adolescent Rheumatology Etanercept Cohort Study (BSPAR-ETN) (funded by a research grant from the British Society for Rheumatology (BSR); BSR has previously also received restricted income from Pfizer to fund this project; Childhood Arthritis Prospective Study (CAPS) (funded by Versus Arthritis UK, grant: 20542); Childhood Arthritis Response to Medication Study (CHARMS) (funded by Sparks UK, reference 08ICH09; and the Medical Research Council, reference MR/M004600/1), United Kingdom Juvenile Idiopathic Arthritis Genetics Consortium (UKJIAGC). This study also acknowledges the use of the following two UK-wide JIA-associated uveitis clinical trials: the SYCAMORE Trial (funded by Arthritis Research UK, grant: 19612 and the National Institute of Health Research Health Technology Assessment, grant: 09/51/01); and the APTITUDE Trial (funded by Arthritis Research UK, grant: 20659).

This research was funded in whole, or in part, by the Wellcome Trust [WT220788]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Members of the CLUSTER Consortium are as follows:

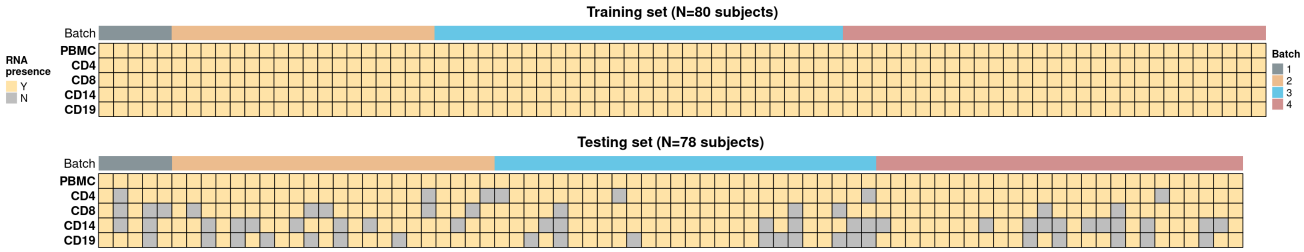
Prof Lucy R. Wedderburn, Dr Melissa Kartawinata, Ms Zoe Wanstall, Ms Bethany R Jebson, Ms Freya Luling Feilding, Ms Alyssia McNeece, Ms Elizabeth Ralph, Ms Vasiliki Alexiou, Mr Fatjon Dekaj, Ms Aline Kimonyo, Ms Fatema Merali, Ms Emma Sumner, Ms Emily Robinson (UCL GOS Institute of Child Health, London); Prof Andrew Dick, (UCL Institute of Ophthalmology, London); Prof Michael W. Beresford, Dr Emil Carlsson, Dr Joanna Fairlie, Dr Jenna F. Gritzfeld (University of Liverpool), Ms Karen Rafferty, Ms Laura Whitty, Ms Jessica Fitzgerald; Prof Athimalaipet Ramanan, Ms Teresa Duerr (University Hospitals Bristol); Prof Michael Barnes, Ms Sandra Ng, (Queen Mary University, London); Prof Kimme Hyrich, Prof Stephen Eyre, Prof Soumya Raychaudhuri, Prof Andrew Morris, Dr Annie Yarwood, Dr Samantha Smith, Dr Stevie Shoop-Worrall, Ms Saskia Lawson-Tovey, Dr John Bowes, Dr Paul Martin, Dr Melissa Tordoff, Ms Jeronee Jennycloss, Mr Michael Stadler, Prof Wendy Thomson, Dr Damian Tarasek (University of Manchester); Dr Chris Wallace, Dr Wei-Yu Lin (University of Cambridge); Prof Nophar Geifman (University of Surrey); Dr Sarah Clarke (School of Population Health sciences and MRC Integrative Epidemiology Unit, University of Bristol); Dr Victoria J Burton, Dr Thierry Sornasse (AbbVie Inc.); Daniela Dastros-Pitei MD, PhD, Sumanta Mukherjee, PhD (GlaxoSmithKline Research and Development Limited.); Dr Michael McLean, Dr Anna Barkaway, Dr Victoria Basey (Pfizer); Dr Peyman Adjamian (Swedish Orphan Biovitrum AB (publ) (Sobi)); Helen Neale (UCB Biopharma SRL.); The CLUSTER Champions.

Competing interests

The CLUSTER consortium has been provided with generous grants from AbbVie and Sobi. CW receives funding from MSD and GSK and is a part-time employee of GSK. These companies had no involvement in the work presented here.

Figures

a


Training set (N=80 subjects)
Batch
PBMC
CD4
CD8
CD14
CD19
RNA presence
Y
N
Batch
1
2
3
4
Testing set (N=78 subjects)
Batch
PBMC
CD4
CD8
CD14
CD19

b

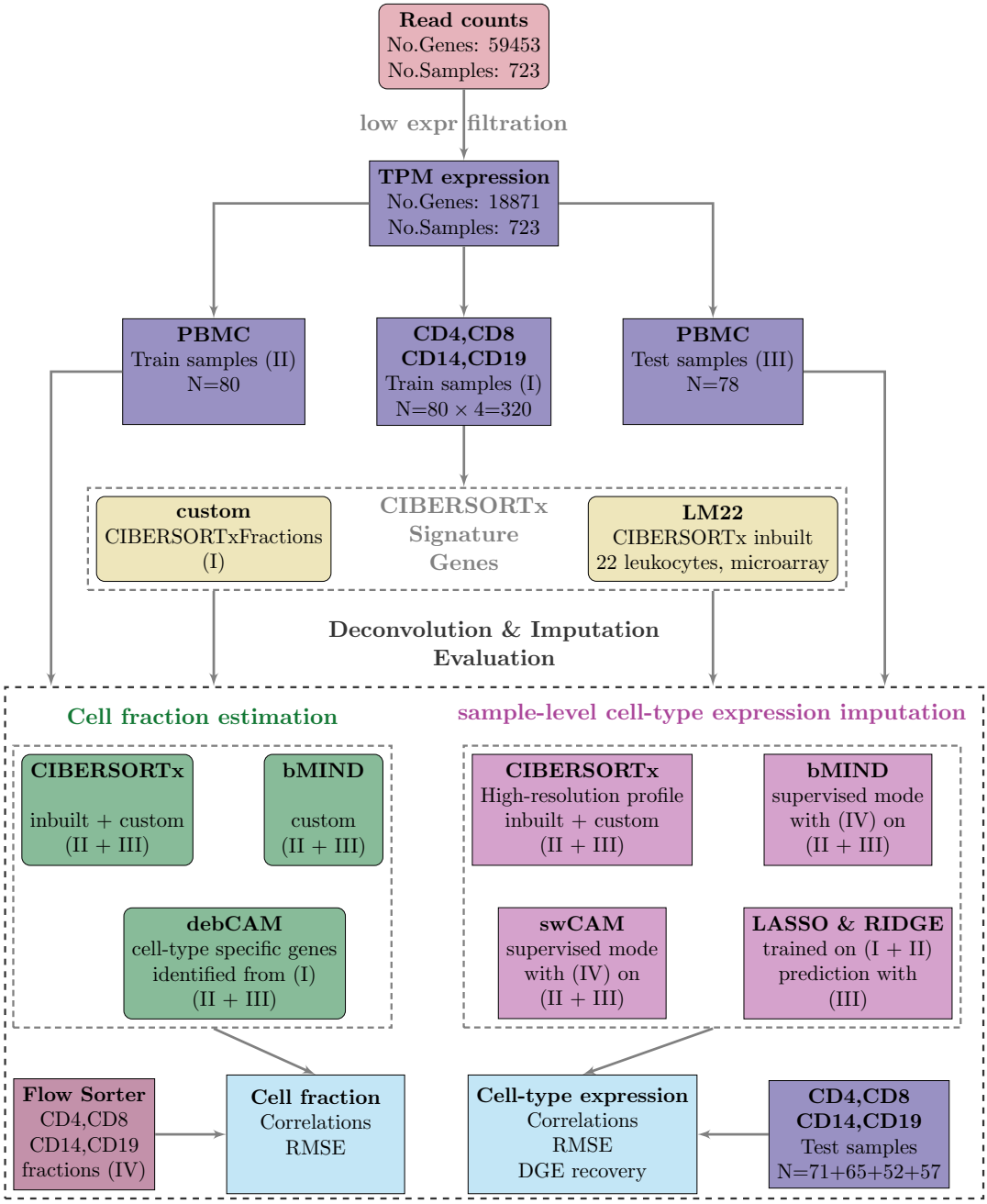

Read counts
No.Genes: 59453
No.Samples: 723
low expr filtration
TPM expression
No.Genes: 18871
No.Samples: 723
PBMC
Train samples (II)
N=80
CD4,CD8
CD14,CD19
Train samples (I)
N=80 × 4=320
PBMC
Test samples (III)
N=78
custom
CIBERSORTxFractions
(I)
CIBERSORTx
Signature
Genes
LM22
CIBERSORTx inbuilt
22 leukocytes, microarray
Deconvolution & Imputation
Evaluation
Cell fraction estimation
sample-level cell-type expression imputation
CIBERSORTx
inbuilt + custom
(II + III)
bMIND
custom
(II + III)
debCAM
cell-type specific genes
identified from (I)
(II + III)
CIBERSORTx
High-resolution profile
inbuilt + custom
(II + III)
bMIND
supervised mode
with (IV) on
(II + III)
swCAM
supervised mode
with (IV) on
(II + III)
LASSO & RIDGE
trained on (I + II)
prediction with
(III)
Flow Sorter
CD4,CD8
CD14,CD19
fractions (IV)
Cell fraction
Correlations
RMSE
Cell-type expression
Correlations
RMSE
DGE recovery
CD4,CD8
CD14,CD19
Test samples
N=71+65+52+57

Figure 1: Data and study design. (a) CLUSTER samples by cell type (row) and subject (column). Cells are coloured based on the availability of RNA (Y for yes, N for no), and the top panel annotations indicate the RNA sequencing batch (Batch) (b) Data analysis workflow. Transcript per million (TPM) was calculated after excluding low-expressed genes. CD4, CD8, CD14, and CD19 TPMs from 80 training samples were used to generate the custom signature genes using the CIBERSORTxFractions module. We deconvoluted the cell fractions from PBMC based on inbuilt and custom signatures using CIBERSORTx. We also estimated cell fractions based on custom signature genes using bMIND and cell-type specific genes using debCAM. Estimates of cell fractions were compared to the ground-truth cell fractions from the flow cytometry, and we assessed fraction accuracy using Pearson correlation and RMSE (root mean square error). Next, we estimated sample-level cell-type gene expression based on inbuilt and custom signature matrices using the CIBERSORTx high resolution module. In parallel, we ran bMIND and swCAM, with the actual cell fractions, in a supervised mode for estimating cell-type expression. Also, for each cell type, a LASSO/RIDGE model was trained on PBMC and cell-type gene expression with 5-fold cross-validation and was then used to predict cell-type gene expression. We compared imputed cell-type expressions with the observed ones and evaluated and benchmarked the performance using Pearson correlation, RMSE and differential gene expression (DGE) recovery.

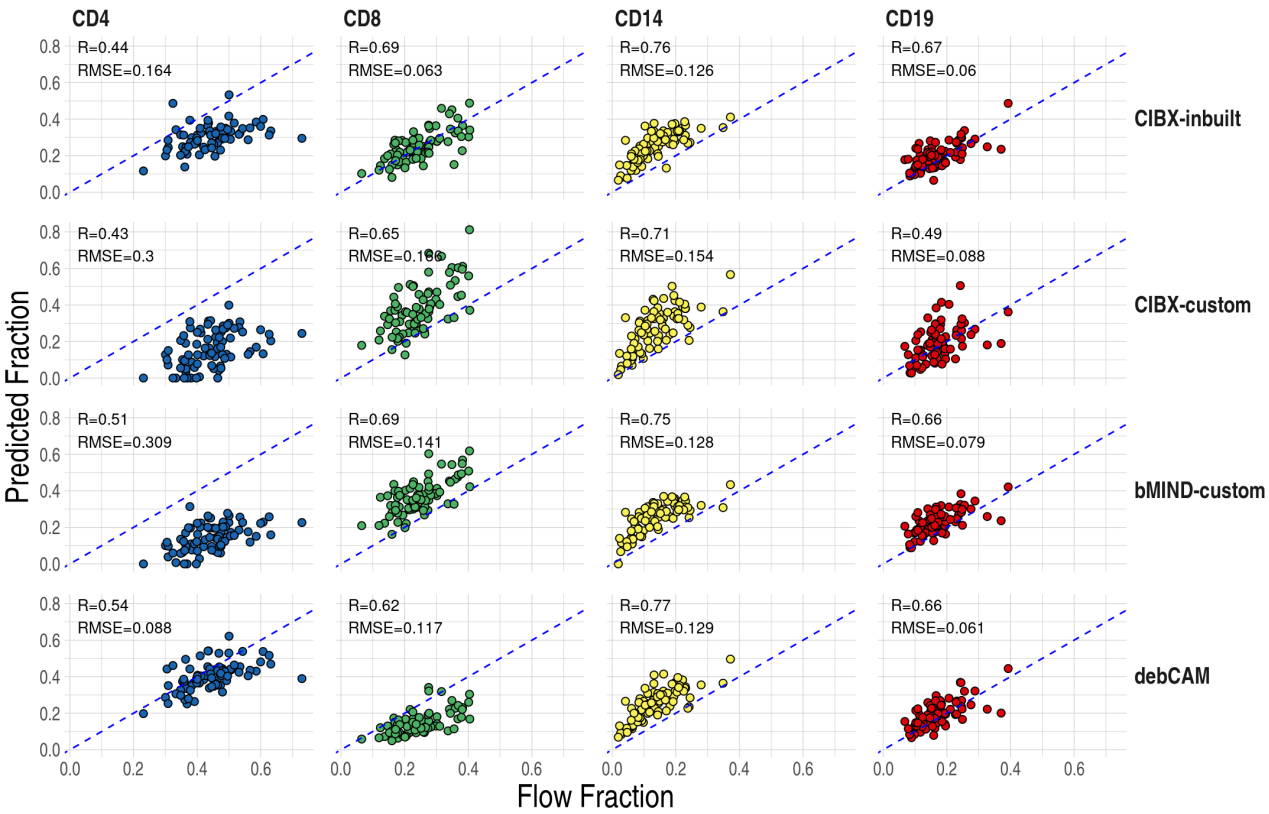
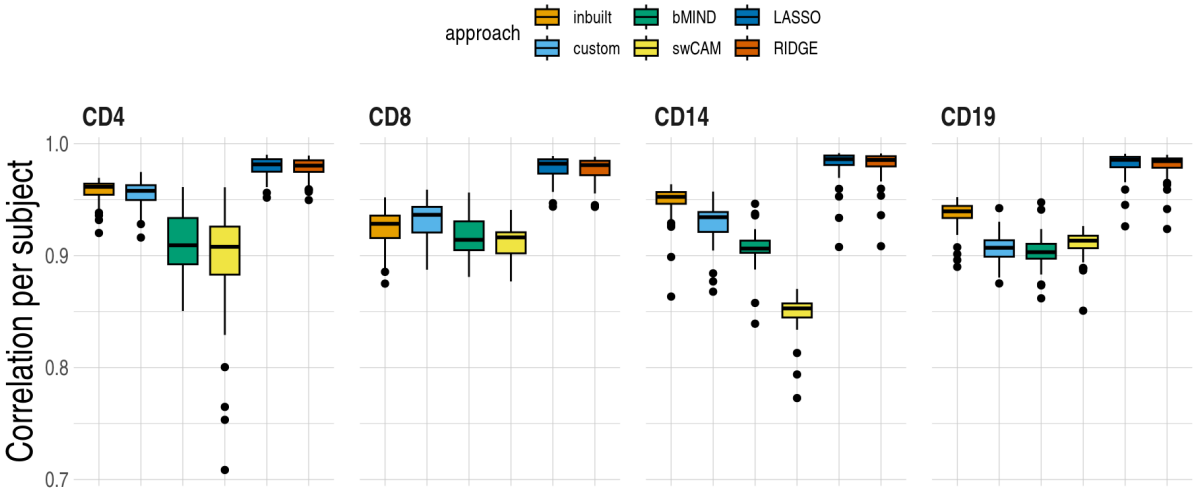
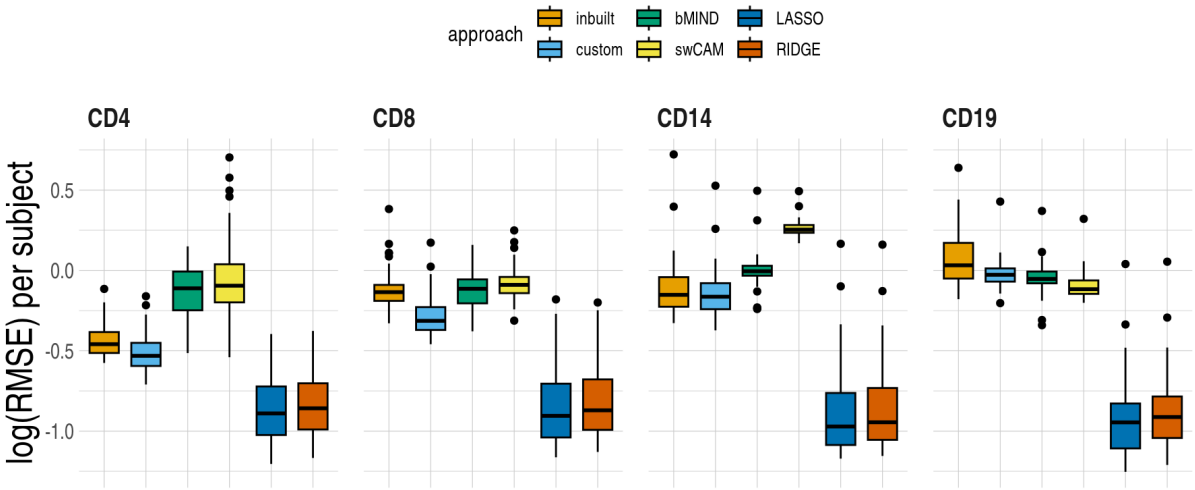


Figure 2: Prediction accuracy of cell fractions by cell type (column) and approaches (row). Pearson correlation (R) and root mean square errors (RMSE) between fraction estimates (y-axis) and flow sorter (x-axis) were calculated. Each point is a testing sample and dashed blue lines indicate $y = x$. CIBX-inbuilt: CIBERSORTx fraction deconvolution using the inbuilt signature matrix. We derived CD4, CD8, CD14, and CD19 fractions from corresponding collapsed members (Supplementary Fig. 2a; Supplementary Fig. 4) by summing members' fractions and then scaled the resultant CD4, CD8, CD14 and CD19 fractions to the sum of 1; CIBX-custom: CIBERSORTx fraction deconvolution using the custom signature matrix derived from our sorted cell (CD4, CD8, CD14, CD19) gene expression by CIBERSORTx; bMIND-custom: bMIND fraction estimation using the custom signature matrix; debCAM-custom: debCAM fraction estimation using cell-type specific genes identified by debCAM from sorted cell expression in 80 training samples.

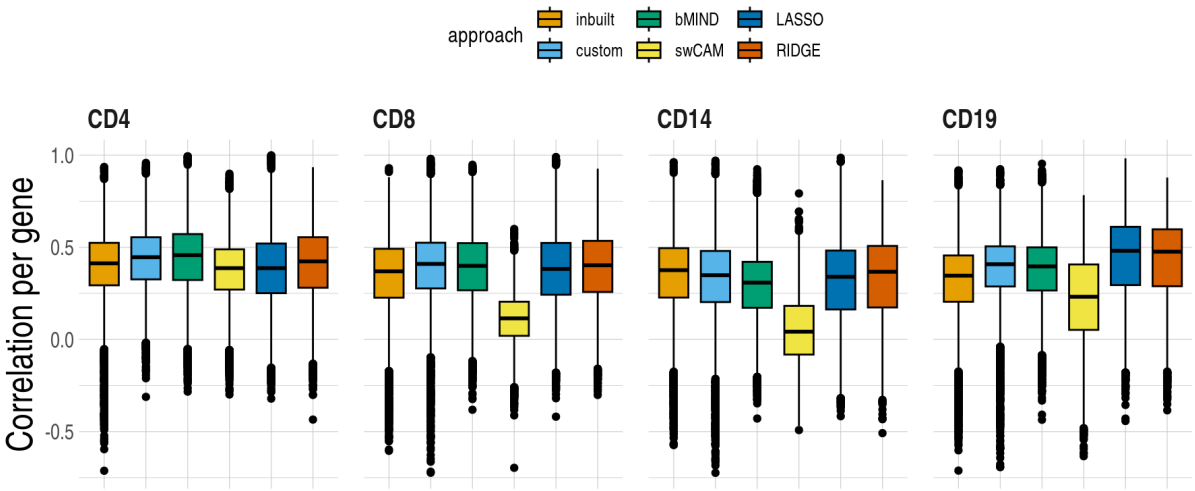
a


approach
inbuilt
custom
bMIND
swCAM
LASSO
RIDGE
CD4
CD8
CD14
CD19
Correlation per subject
1.0
0.9
0.8
0.7

b


approach
inbuilt
custom
bMIND
swCAM
LASSO
RIDGE
CD4
CD8
CD14
CD19
log(RMSE) per subject
0.5
0.0
-0.5
-1.0

c


approach
inbuilt
custom
bMIND
swCAM
LASSO
RIDGE
CD4
CD8
CD14
CD19
Correlation per gene
1.0
0.5
0.0
-0.5

d

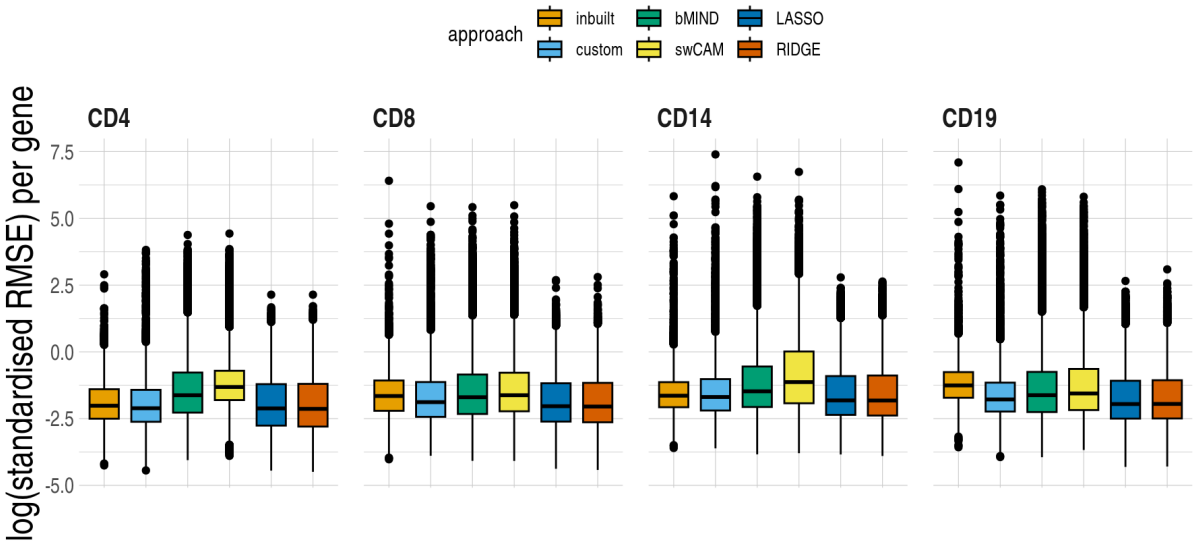


Figure 3: Prediction accuracy of sample-level cell-type expression by approach. (a) Pearson correlation and (b) root mean square error (RMSE) by comparing observed to predicted cell-type expression of genes from the same subjects. One estimate per subject. (c) Pearson correlation and (d) RMSE between observed and predicted cell-type expression across testing samples for each gene. One estimate per gene. RMSE per gene was standardised by the average observed expression. RMSE per subject/gene (b,d) is shown in the log scale. inbuilt: CIBERSORTx expression deconvolution with the inbuilt signature matrix; custom: CIBERSORTx expression deconvolution with a custom signature matrix derived from sorted cell-type expression in training samples; bMIND: bMIND expression deconvolution with flow fractions; swCAM: swCAM deconvolution with flow fractions; LASSO/RIDGE: expression predicted from regularised multi-response Gaussian models.

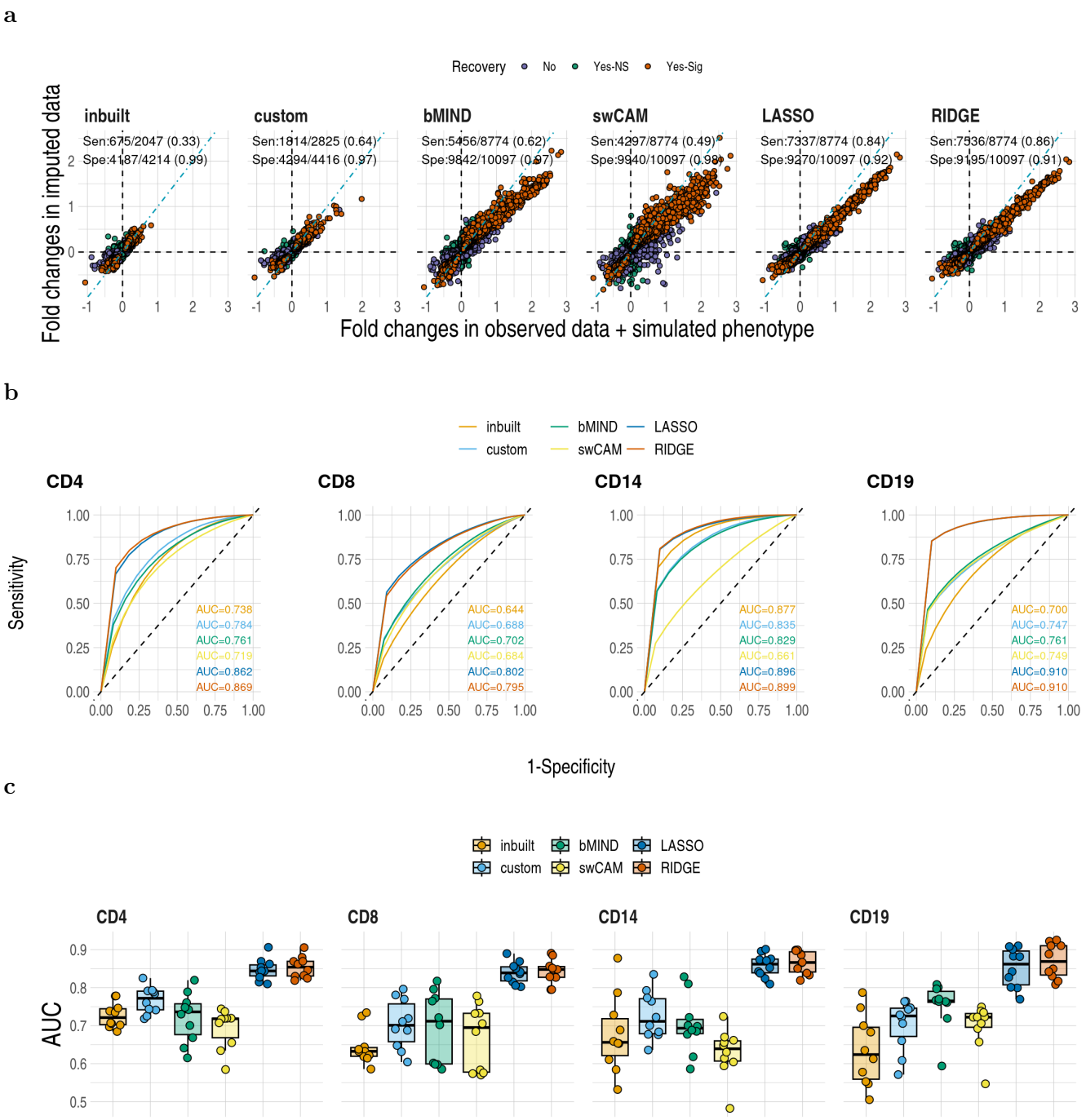
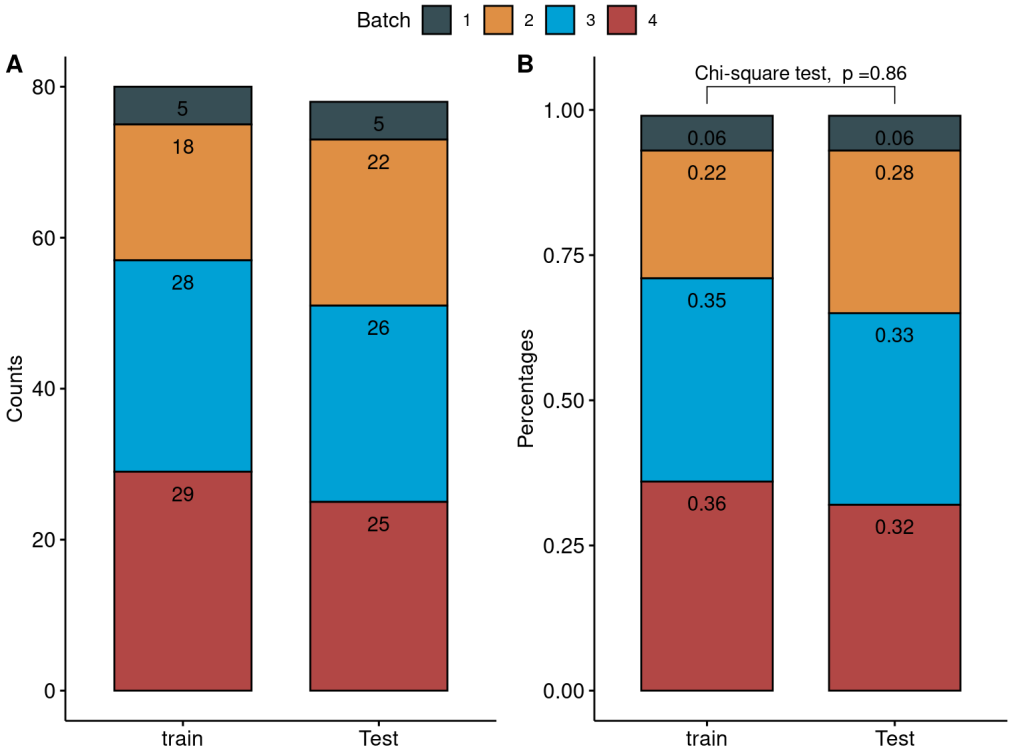
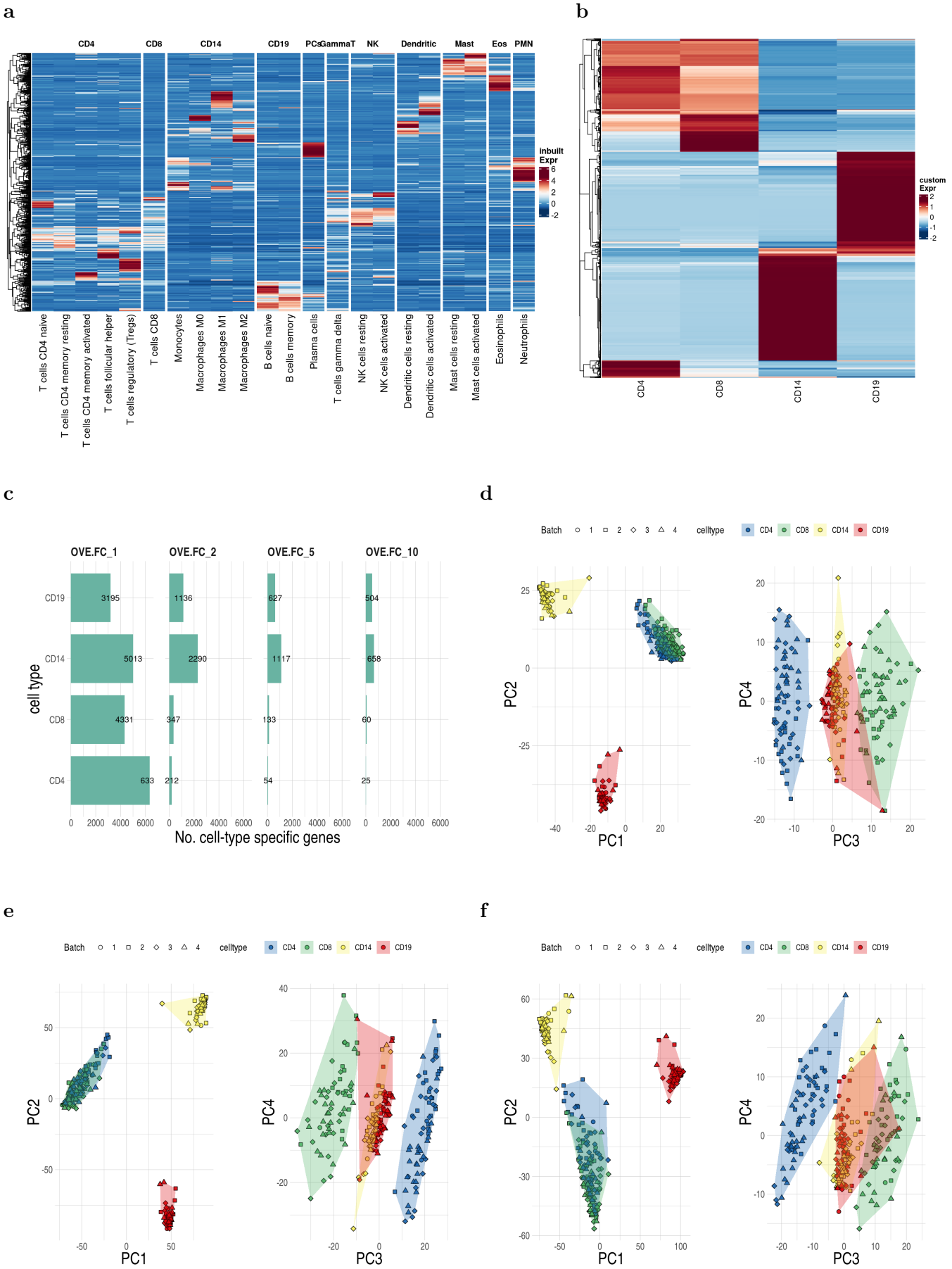


Figure 4: Differential gene expression (DGE) recovery. (a) Comparisons of $\log_2$ fold changes in genes between the observed (x-axis) and imputed (y-axis) data by method (column). DGE analysis was carried out using limma based on one of the simulated phenotypes. Here, an FDR of 0.05 was used in both observed and imputed data, and CD4 DGE recovery is shown. Each point (gene) is coloured by Recovery status: No, Yes-NS, and Yes-Sig. Yes-Sig (sensitivity; Sen): differentially expressed genes in the observed data were also called significance in the imputed data, and the orientations of the effect sizes are the same in both data. Yes-NS (specificity; Spe): genes are called non-significance (NS) in both data. The same results are also in Supplementary Fig. 7, illustrating different recovery statuses by row. (b) Receiver operating characteristic (ROC) curves and estimated area under curve (AUC) by cell type and approach based on one simulated phenotype. FDR was fixed at 0.05 in the observed data and varied from 0 to 1 by 0.05 in the imputed data. (c) AUC distributions by cell type and approach. Each point is a simulated phenotype, and there are ten simulated phenotypes. Box plots showed the AUC distributions, with horizontal lines from the bottom to the top for 25 %, 50 % and 75 % quantiles, respectively. inbuilt: CIBERSORTx with the inbuilt signature matrix; custom: CIBERSORTx with a custom signature matrix derived from sorted cell-type expression in training samples; bMIND: bMIND with flow fractions; swCAM: swCAM with flow fractions; LASSO/RIDGE: regularised multi-response Gaussian models.

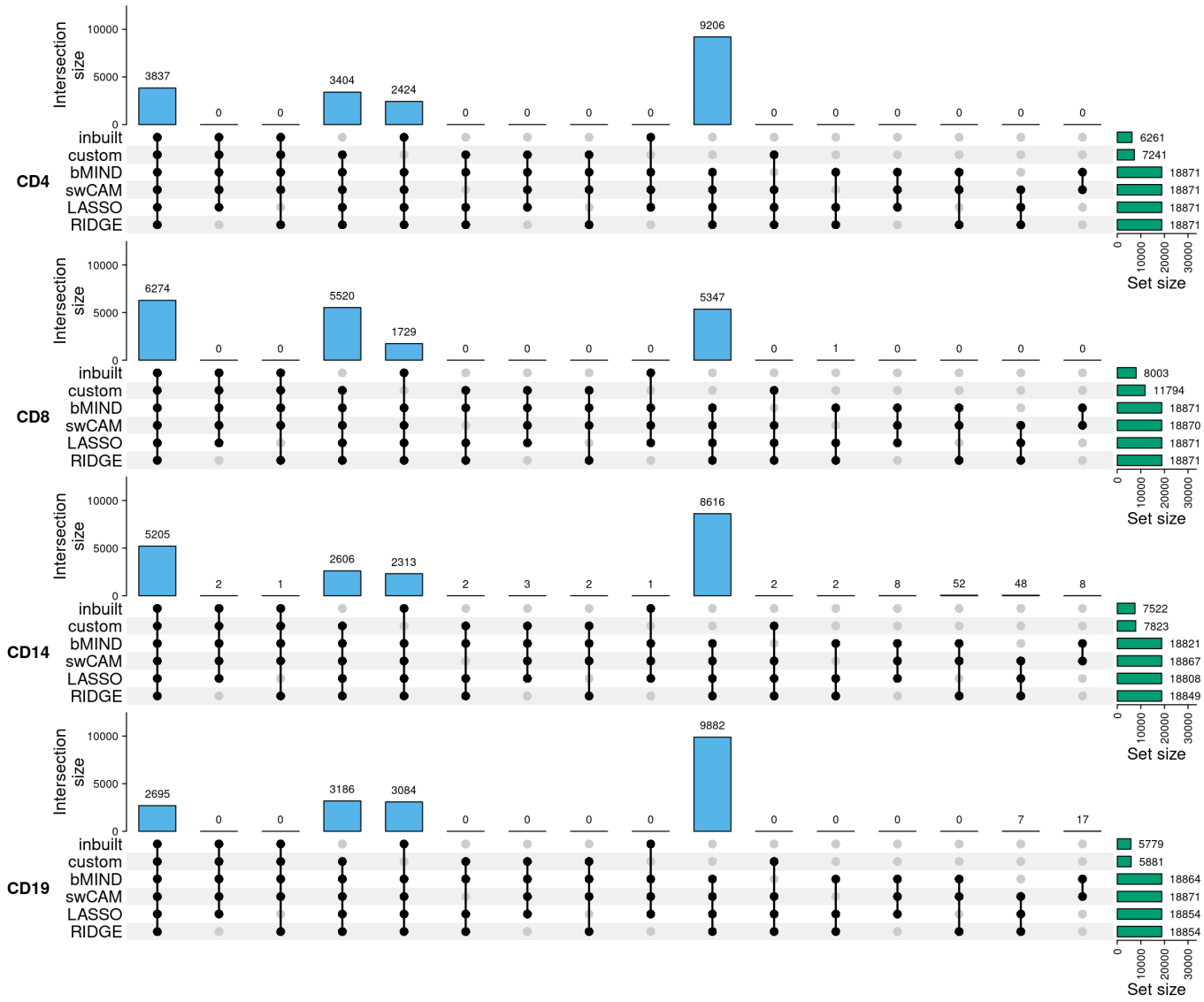
Supplementary Figures and Table



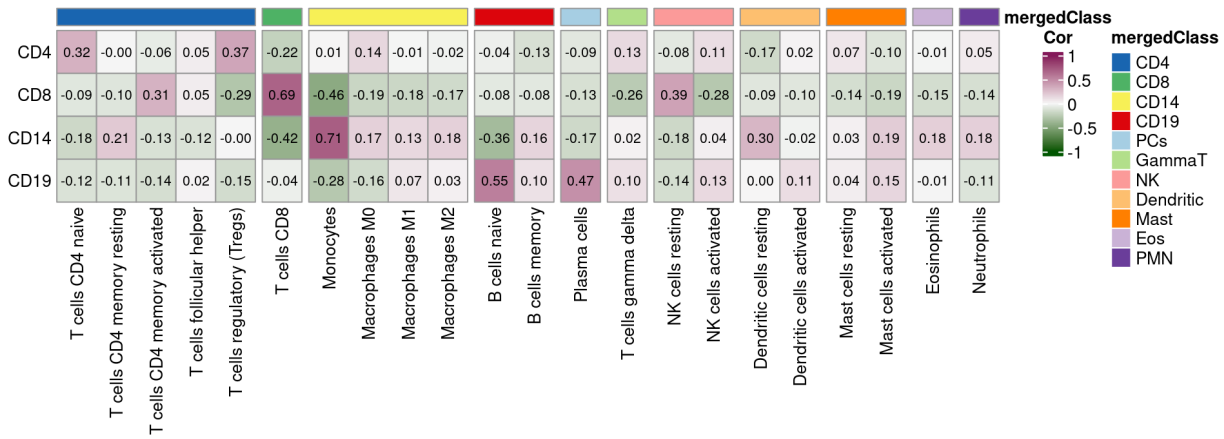
Supplementary Figure 1: Frequencies (A) and percentages (B) of subjects by batch and training/testing set.


a
CD4
CD8
CD14
CD19
PCs
GammaT
NK
Dendritic
Mast
Eos
PMN
inbuilt Expr
T cells CD4 naive
T cells CD4 memory resting
T cells CD4 memory activated
T cells follicular helper
T cells regulatory (Tregs)
T cells CD8
Monocytes
Macrophages M0
Macrophages M1
Macrophages M2
B cells naive
B cells memory
Plasma cells
T cells gamma delta
NK cells resting
NK cells activated
Dendritic cells resting
Dendritic cells activated
Mast cells resting
Mast cells activated
Eosinophils
Neutrophils
b
custom Expr
CD4
CD8
CD14
CD19
c
OVE.FC_1
OVE.FC_2
OVE.FC_5
OVE.FC_10
cell type
CD19
CD14
CD8
CD4
3195
1136
627
504
5013
2290
1117
658
4331
347
133
60
633
212
54
25
No. cell-type specific genes
d
Batch
celltype
PC1
PC2
PC3
PC4
e
f

Supplementary Figure 2: Differences between inbuilt and custom signature genes, and debCAM cell-type specific genes.(a) Expression of inbuilt signature genes (N=547) in 22 leukocyte subsets, curated by the CIBERSORTx team from microarray gene expression. For each gene (row), expression is centred to the mean and scaled by the standard deviation of cell types (column). Columns are split by the collapsed classes. CD4: T cells CD4 naive, T cells CD4 memory resting, T cells CD4 memory activated, T cells follicular helper, and T cells regulatory (Tregs); CD8: T cells CD8; CD14: Monocytes, Macrophages M0, Macrophages M1, and Macrophages M2; CD19: B cells naive and B cells memory; PCs: Plasma cells; GammaT: T cells gamma delta; NK: NK cells resting and NK cells activated; Dendritic: Dendritic cells resting and Dendritic cells activated; Mast: Mast cells resting and Mast cells activated; Eos: EosinophilsPMN: Neutrophils (b) Expression of custom signature genes (N=1589), derived from our sorted-cell RNAseq expression in 80 training subjects using CIBERSORTx. Expression is centred and scaled across cell types (column) by gene (row). (c) Numbers (No.) of debCAM cell-type specific genes by cell type (row) and selection criteria (column). debCAM, which does not have the signature matrix, selects the cell-type-specific genes, that are over-expressed in one cell type verse everyone (OVE). OVE fold change (FC) of 1, 2, 5 and 10 were used in our sorted cell expression of 80 training subjects. The first four principal components from PCA analysis of (d) inbuilt, (e) custom signature and (f) debCAM cell type specific gene expression in test samples. Each dot is a sample, coloured by cell type and shaped by sequencing batch.

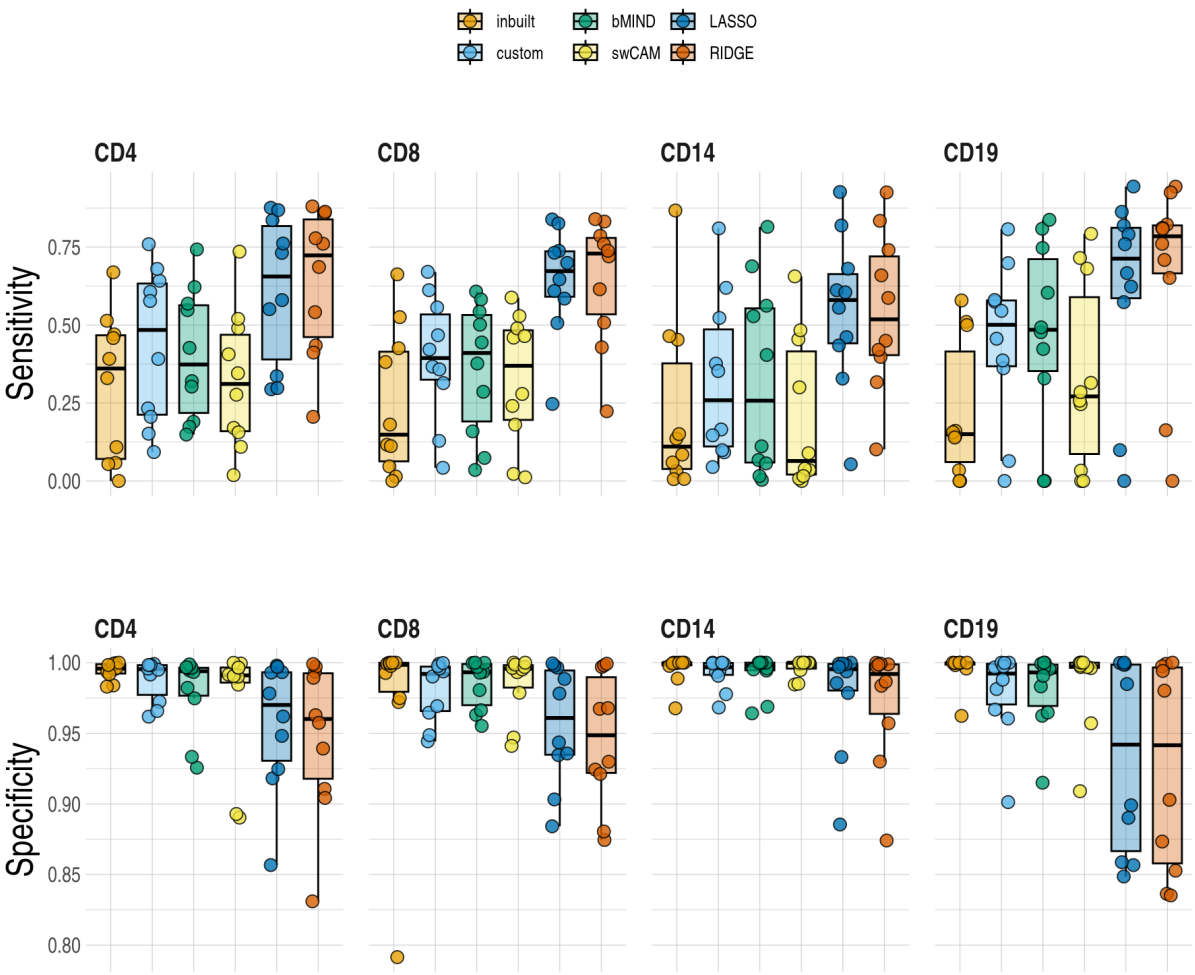


Supplementary Figure 3: Overlap of predicted genes by CIBERSORTx using inbuilt (inbuilt) and our custom signatures (custom), bMIND, swCAM, LASSO and RIDGE. Predicted genes were defined as those with variations in expression across subjects. For each panel (cell type), the right bar plot indicates the numbers of predicted genes (No.Pred.Genes) by approach, and the top bar plot demonstrates No.Pred.Genes common in different combinations of approaches (black dots), but not in the grey-dot approaches, if present.

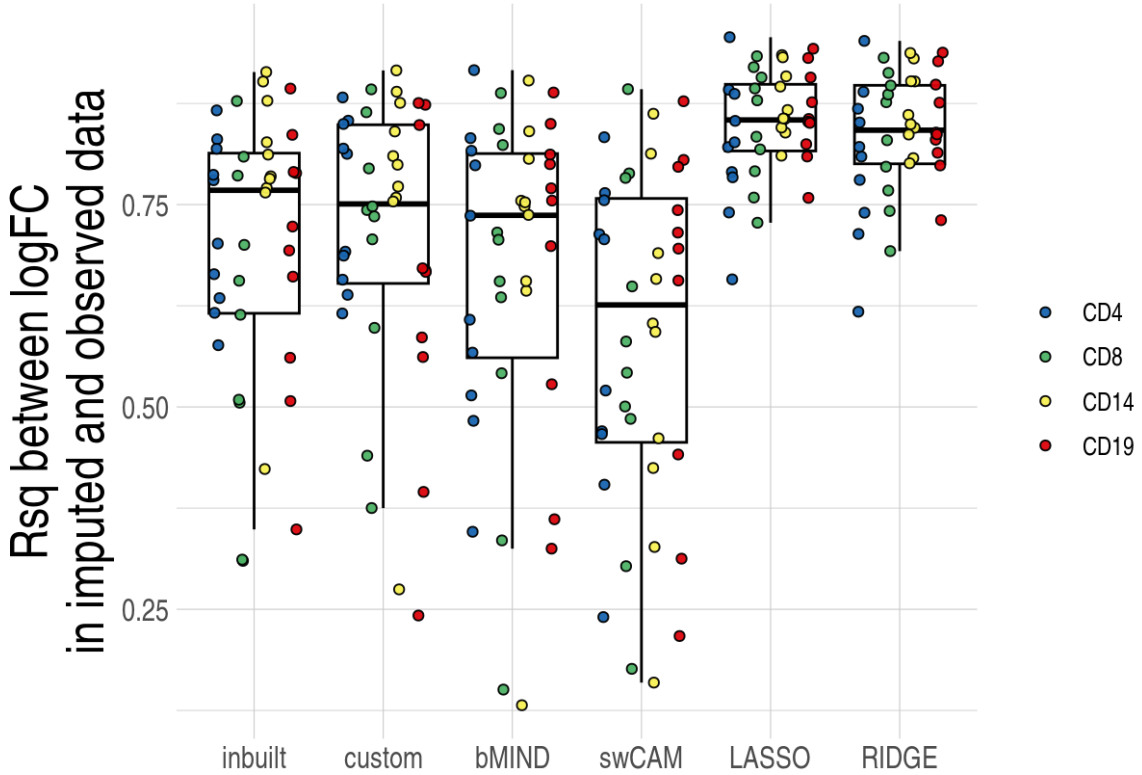


Supplementary Figure 4: Pearson correlations of cell fractions between 22 leucocyte (LM22) and ground-truth flow cell types. Columns are LM22 cell subsets split by their merged classes (top annotation). Rows are ground-truth cell types. Each cell is coloured based on the strength of the correlation.

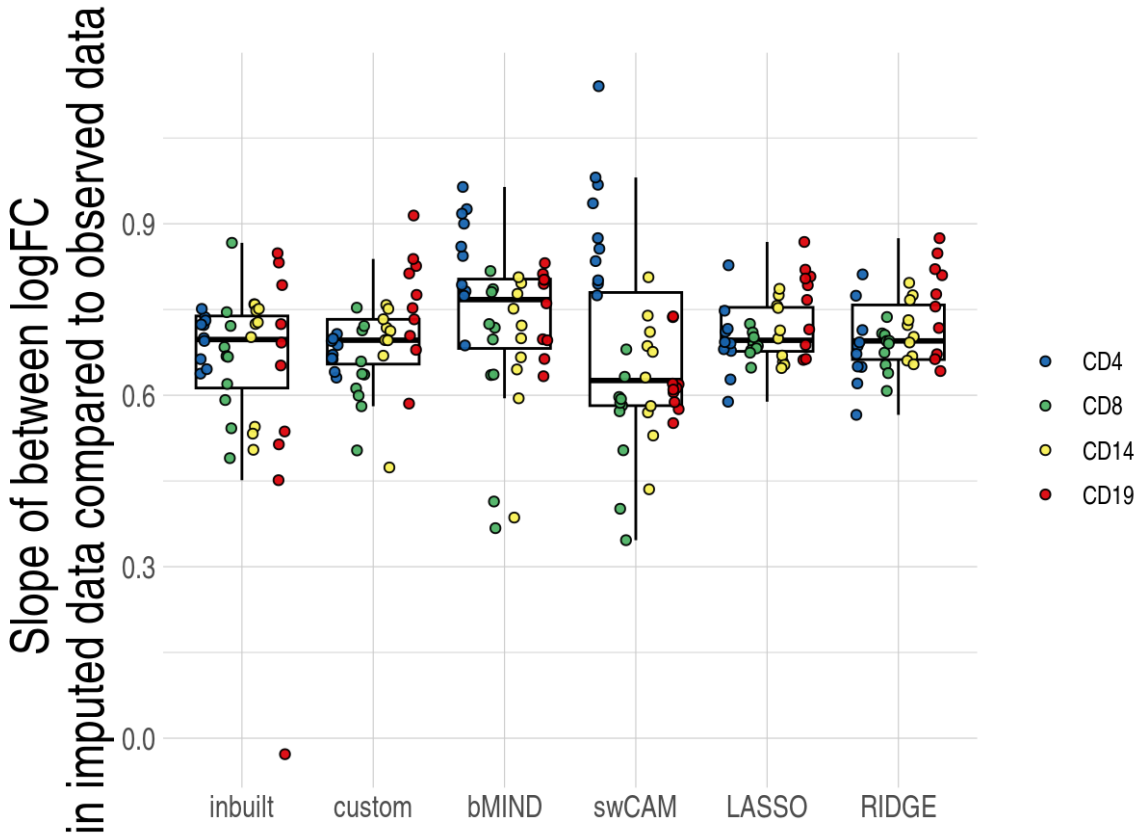
a


inbuilt
custom
bMIND
swCAM
LASSO
RIDGE
CD4
CD8
CD14
CD19
Sensitivity
0.75
0.50
0.25
0.00
Specificity
1.00
0.95
0.90
0.85
0.80

b

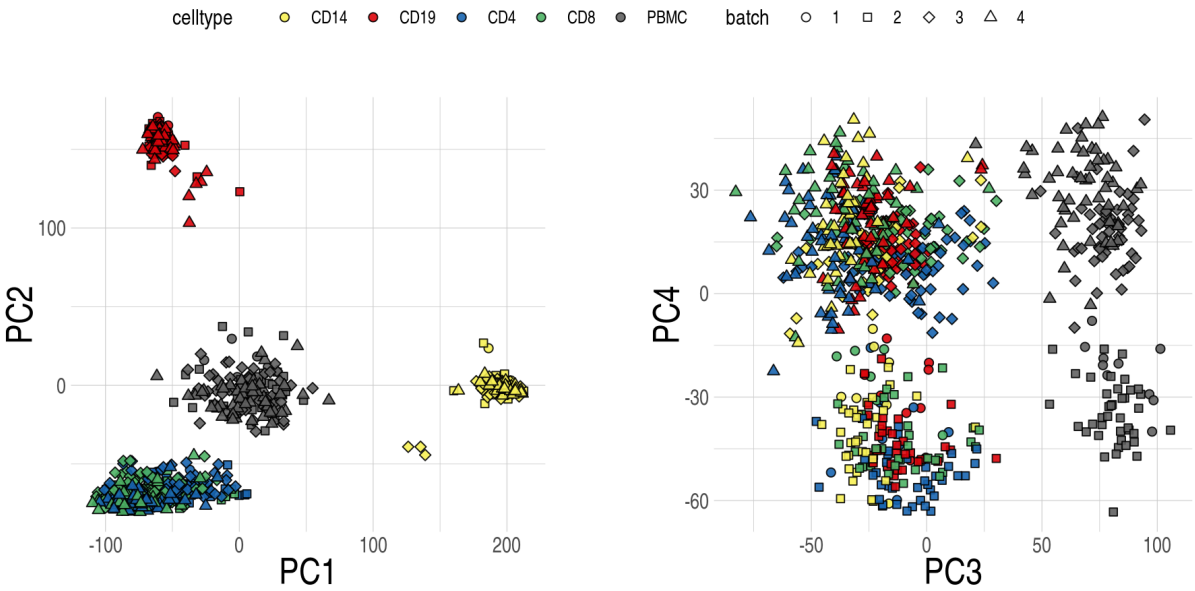

Rsq between logFC
in imputed and observed data
0.75
0.50
0.25
inbuilt
custom
bMIND
swCAM
LASSO
RIDGE
CD4
CD8
CD14
CD19

c

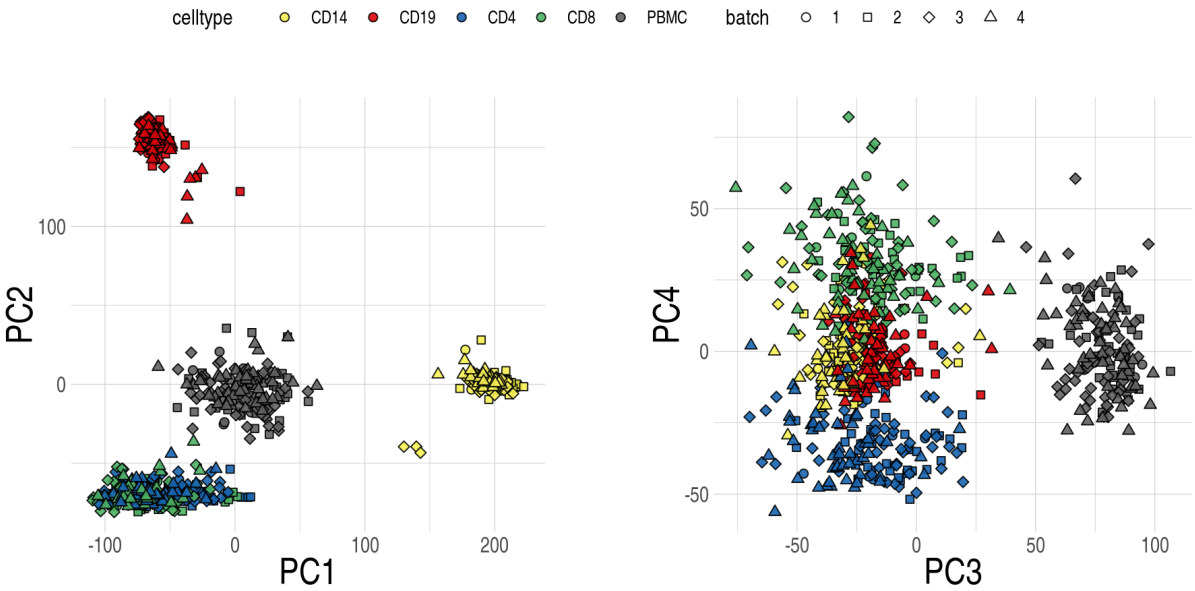


Supplementary Figure 5: Distributions of sensitivity and specificity of DGE recovery, and r-squared and slope of $\log_2$ fold changes between imputed and observed data. An FDR < 0.05 was used in both imputed and observed data. (a) Sensitivity (y-axis, upper panel) and Specificity (y-axis, lower panel) by cell type and approach. Each point is a simulated phenotype. (b) R-squared (Rsq, y-axis) of imputed $\log_2$ fold changes (FC) regression on observed effect sizes by approach. (c) Slopes (y-axis) of imputed $\log_2$ FC regression on observed effect sizes by approach. Each point is a simulated phenotype, coloured by cell type. inbuilt: CIBERSORTx with the inbuilt signature matrix; custom: CIBERSORTx with a custom signature matrix derived from sorted cell-type expression in training samples; bMIND: bMIND with flow fractions; swCAM: swCAM with flow fractions; LASSO/RIDGE: regularised multi-response Gaussian models.

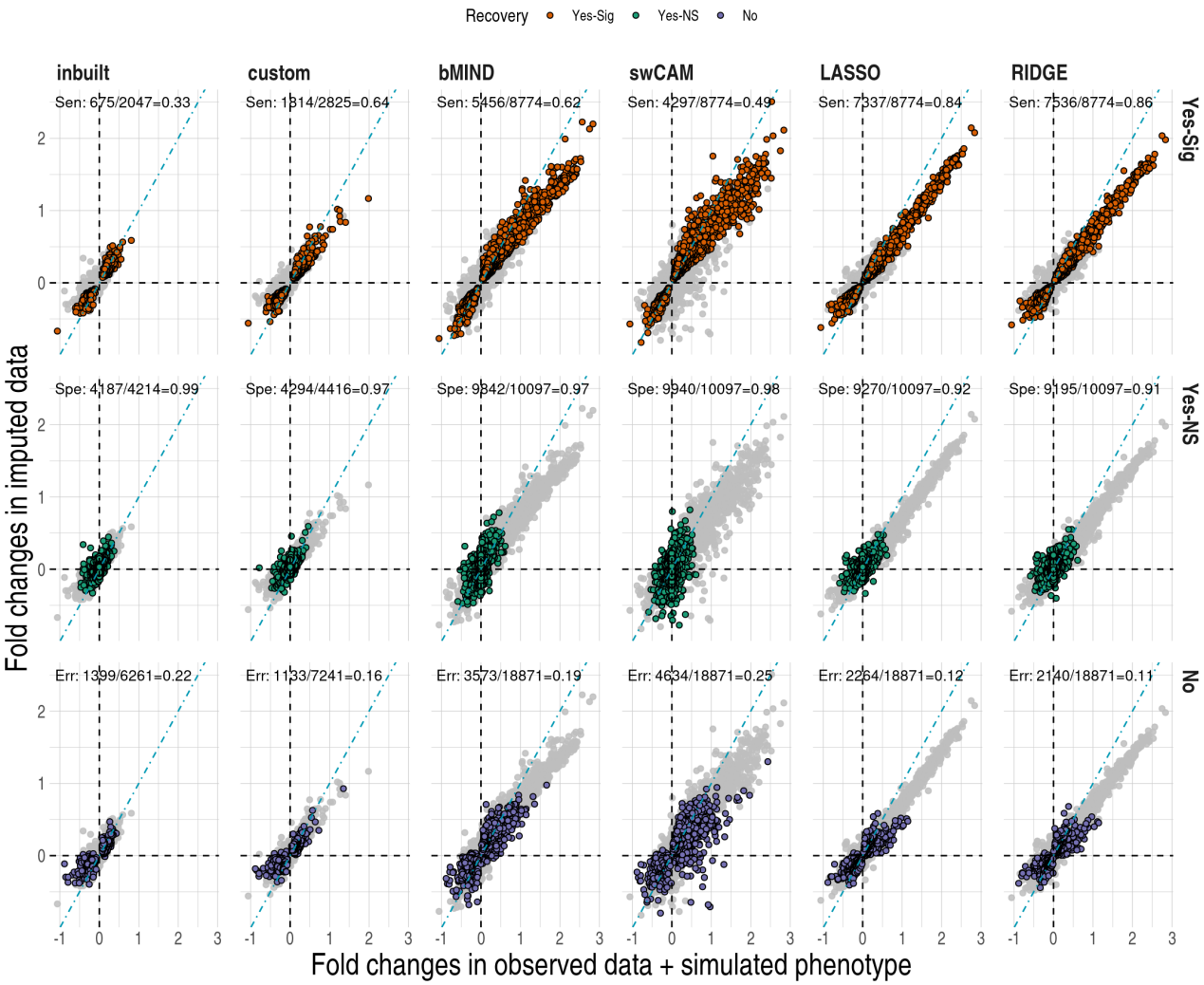
a



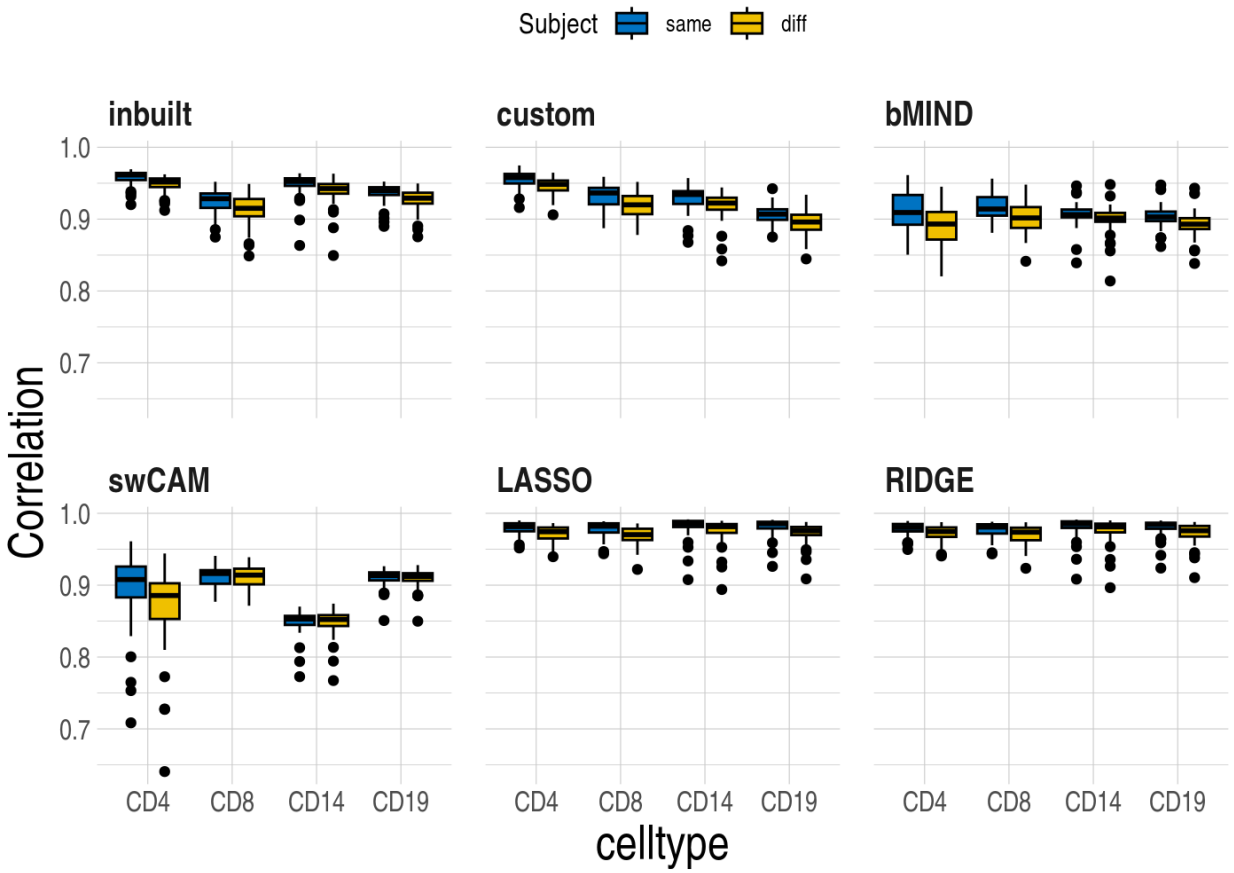
b



Supplementary Figure 6: The first four principal components (PC) from PCA analysis of $\log_2$(count-per-million) expression derived from (a) raw read counts (b) Combat-Seq batch-adjusted read counts in RNAseq samples used in this work.



Supplementary Figure 7: Comparisons of $\log_2$ fold changes in genes between the observed (x-axis) and imputed (y-axis) data by method (column) and by recovery status (row). DGE analysis was carried out using limma based on one of the simulated phenotypes. An FDR of 0.05 was used in both observed and imputed data here, and CD4 DGE recovery is shown. DGE results in each method (column) are the same, with coloured points for genes falling into that category of recovery status (row) and grey points for genes not belonging to the same category. Recovery status: No, Yes-NS, and Yes-Sig. Yes-Sig (sensitivity; Sen): differentially expressed genes in the observed data were also called significance in the imputed data, and the orientations of the effect sizes are the same in both data. Yes-NS (specificity; Spe): genes are called non-significance (NS) in both data. No (error; Err): misclassified genes; Err is calculated as the percentage of misclassified genes to the total number of predicted genes.



Supplementary Figure 8: Distributions of Pearson correlations (y-axis) between observed and imputed expression across genes from the same/different subjects by cell type and approach. One estimate per subject. inbuilt: CIBERSORTx with the inbuilt signature matrix; custom: CIBERSORTx with a custom signature matrix derived from sorted cell-type expression in training samples; bMIND: bMIND with flow fractions; swCAM: swCAM with flow fractions; LASSO/RIDGE: regularised multi-response Gaussian models.

Supplementary Table 1: computational time and memory usage by approach

	Additional step before expression prediction	CPU time (minutes)	RAM usage (Gb)
CIBERSORTx-custom	signature matrix generation	43.7	64.1
bMIND	-	123.2	12.0
swCAM	grid search of optimal Lamda	11296.98	14.3
LASSO	cross-validation model training	765.43	5678.2
RIDGE	cross-validation model training	4067.83	19676.1