nature methods

SUPPLEMENTARY INFORMATION

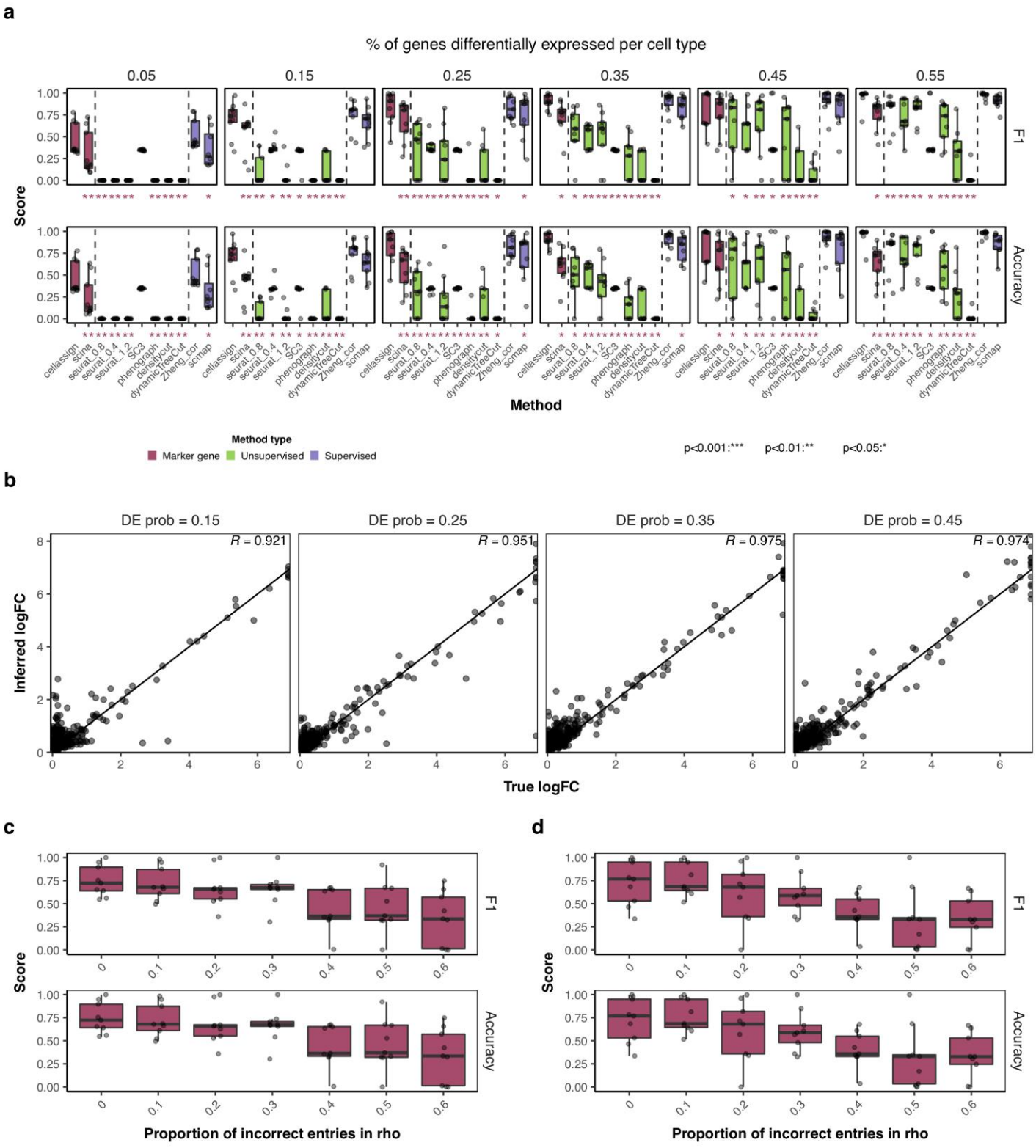
https://doi.org/10.1038/s41592-019-0529-1

In the format provided by the authors and unedited.

Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling

Allen W. Zhang[1,2,3], Ciara O'Flanagan[1], Elizabeth A. Chavez[4], Jamie L. P. Lim[1,2], Nicholas Ceglia[2], Andrew McPherson[1], Matt Wiens[1], Pascale Walters[1], Tim Chan[1], Brittany Hewitson[1], Daniel Lai[1], Anja Mottok[4,5], Clementine Sarkozy[4], Lauren Chong[4], Tomohiro Aoki[4,6], Xuehai Wang[7], Andrew P Weng[7], Jessica N. McAlpine[8], Samuel Aparicio[1,6], Christian Steidl[4], Kieran R. Campbell[1,9,10]* and Sohrab P. Shah[1,2,6]*

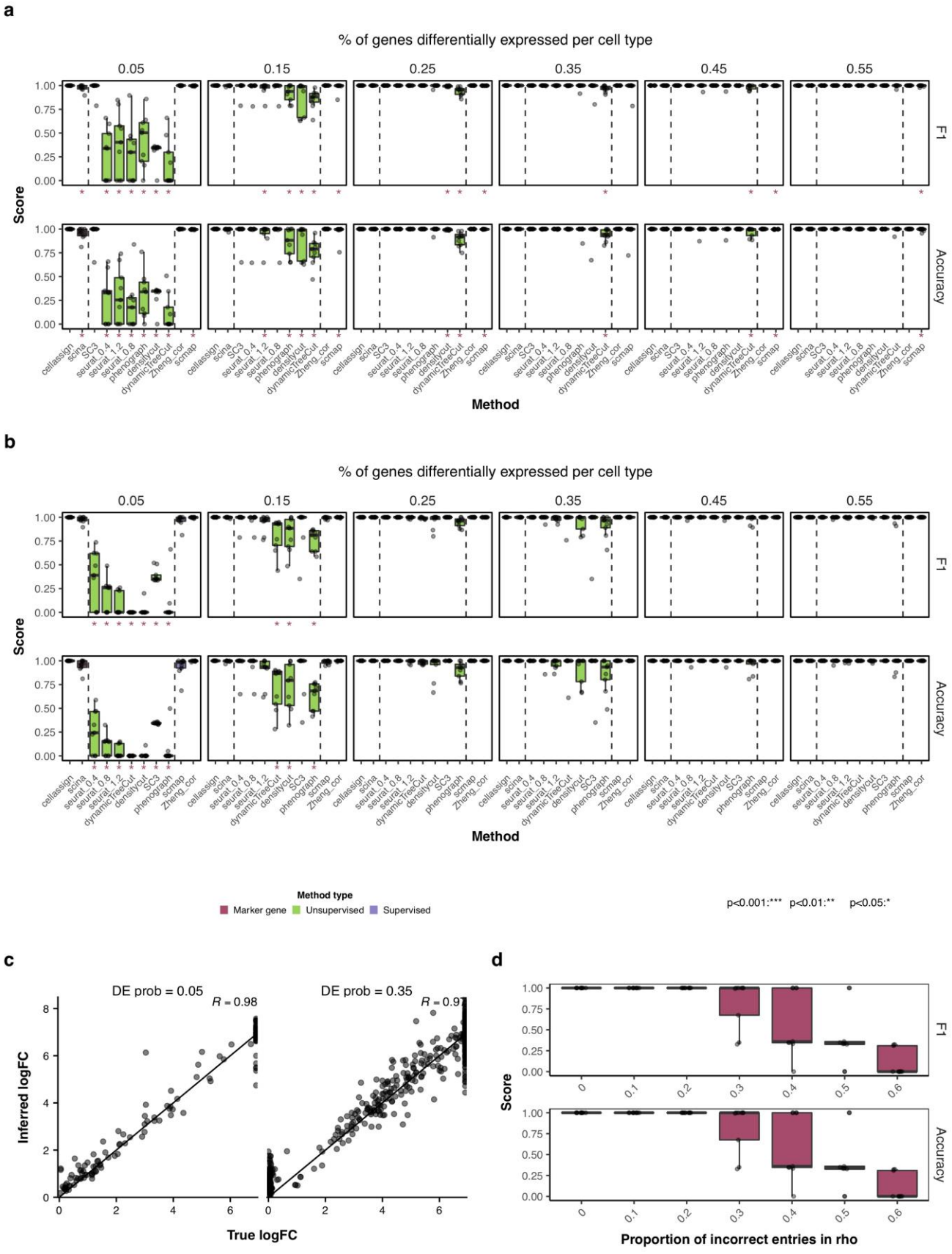
[1]Department of Molecular Oncology, British Columbia Cancer Research Centre, Vancouver, British Columbia, Canada. [2]Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [3]BC Children's Hospital Research, Vancouver, British Columbia, Canada. [4]Centre for Lymphoid Cancer, British Columbia Cancer Research Centre, Vancouver, British Columbia, Canada. [5]Institute of Human Genetics, Ulm University and Ulm University Medical Center, Ulm, Germany. [6]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada. [7]Terry Fox Laboratory, British Columbia Cancer Research Centre, Vancouver, British Columbia, Canada. [8]Department of Obstetrics & Gynaecology, University of British Columbia, Vancouver, British Columbia, Canada. [9]Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada. [10]UBC Data Science Institute, University of British Columbia, Vancouver, British Columbia, Canada. *e-mail: kieran.campbell@stat.ubc.ca; shahs3@mskcc.org



Supplementary Figure 1

Simulation performance across a range of proportions of differ- entially expressed genes, using differential expression parameters derived from comparing naïve CD8+ and naïve CD4+ T cells.

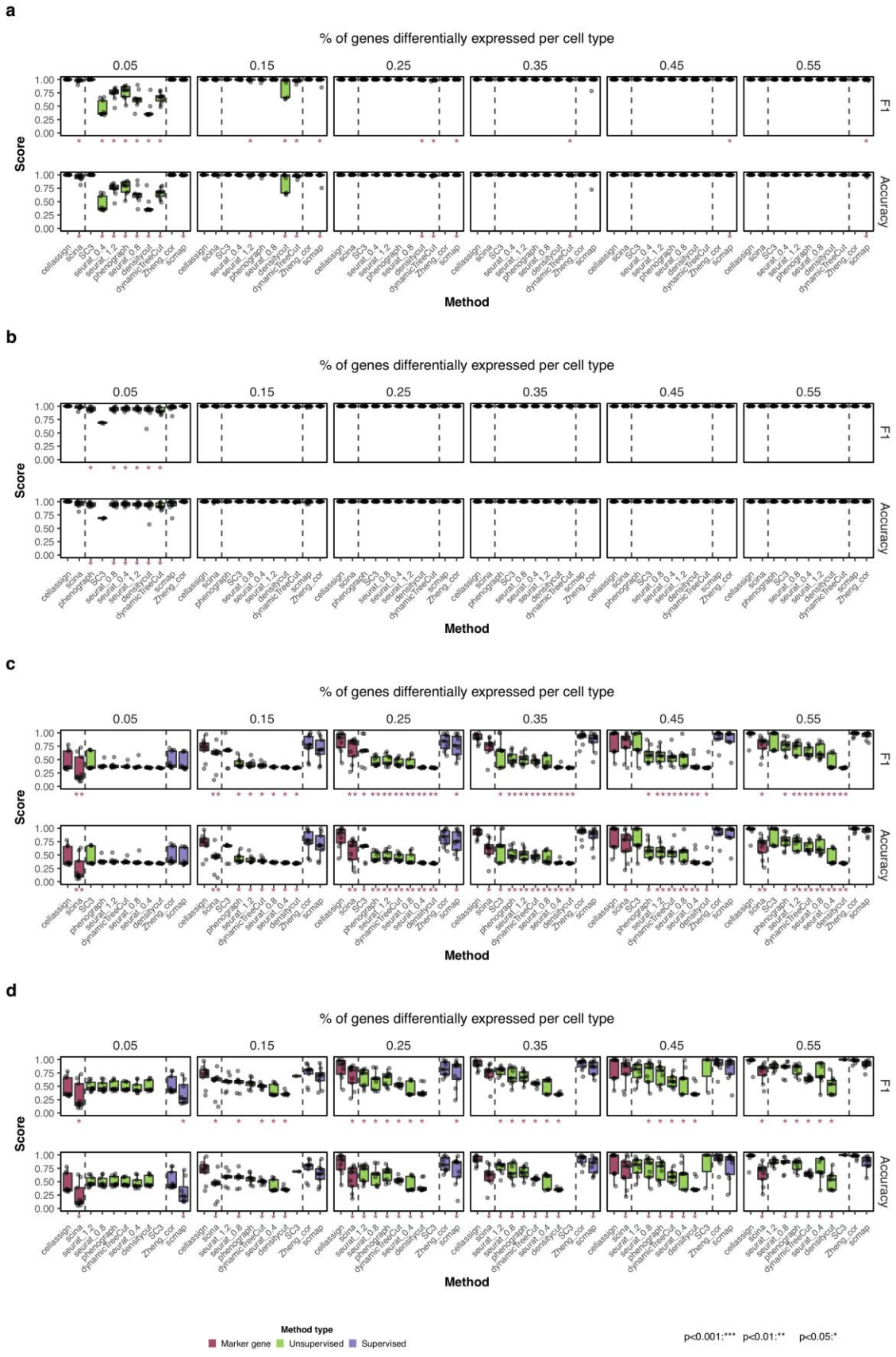
(**a**) Accuracy and cell-level F1 score (**Methods**) for varying proportions of differentially expressed genes per cell type. All methods were pro- vided with expression data for the same set of marker genes. *, **, *** denote FDR-adjusted p-vaues (Wilcoxon signed-rank test) for pairwise comparisons between CellAssign and other methods < 0.05, 0.01, 0.001 respectively. Nine simulated datasets were generated for each parameter setting. Dotted lines separate marker-based, unsupervised, and super- vised methods. (**b**) Correspondence between true simulated log fold change values and log fold change (δ) values inferred by CellAssign. R refers to the Pearson correlation be- tween true and inferred logFC values for CellAssign. (**c** and **d**) Performance of CellAssign where a certain proportion of entries in the marker gene matrix are flipped at random, us- ing (**c)** 5 and (**d)** 20 marker genes per cell type. Nine simulated datasets were generated for each parameter setting. In all boxplots, lower and upper hinges denote the 1st and 3rd quartiles, with whiskers extending to the largest value less than 1.5 × the inter-quartile range.


a
% of genes differentially expressed per cell type
0.05
0.15
0.25
0.35
0.45
0.55
F1
Accuracy
Score
Method
b
% of genes differentially expressed per cell type
Method type
Marker gene
Unsupervised
Supervised
p<0.001:*** p<0.01:** p<0.05:*
c
DE prob = 0.05
R = 0.98
DE prob = 0.35
R = 0.97
Inferred logFC
True logFC
d
Score
Proportion of incorrect entries in rho

Supplementary Figure 2

Simulation performance across a range of proportions of differentially expressed genes, using differential expression parameters derived from com- paring B and CD8+ T cells.

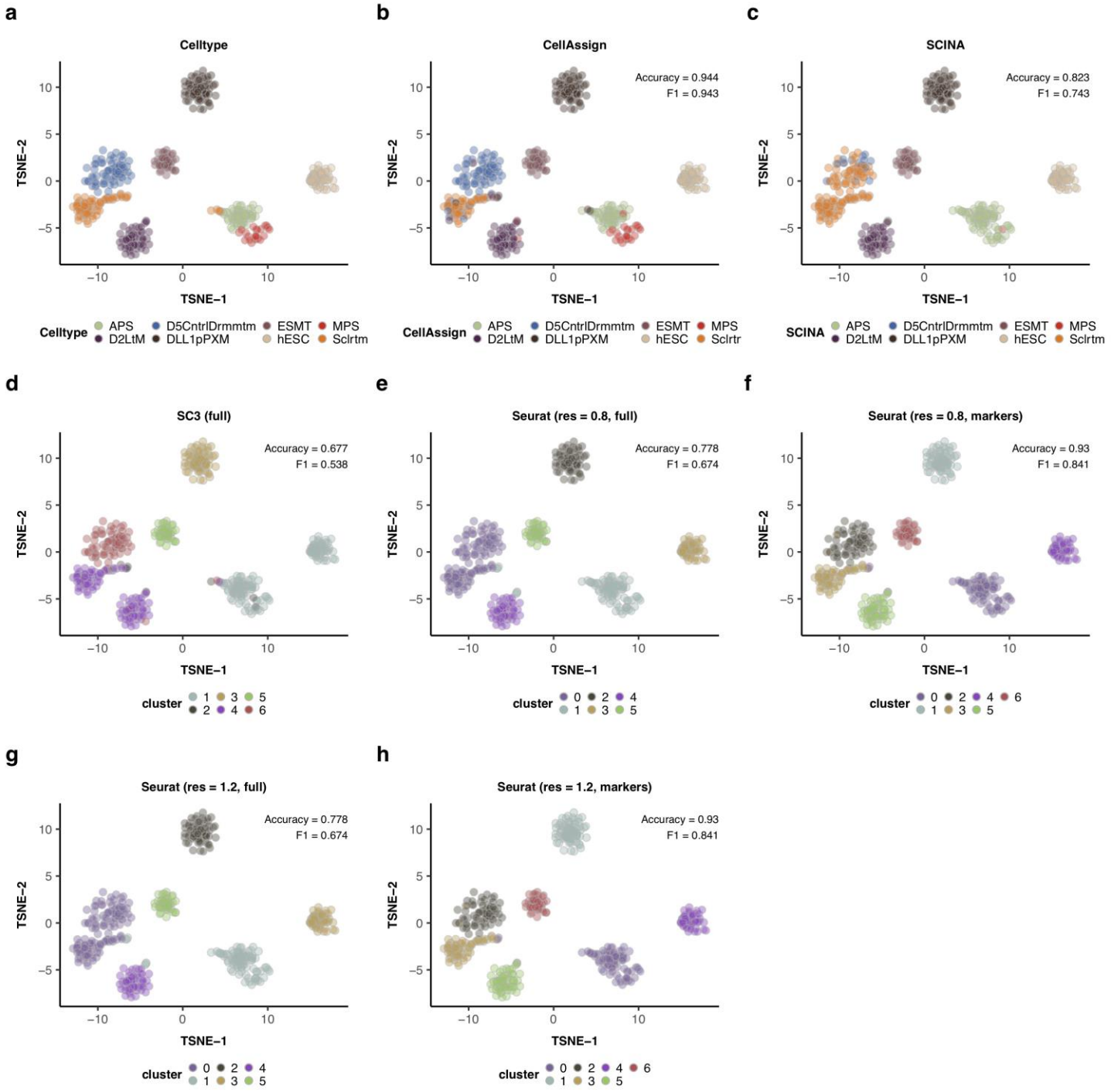
(**a**) Accuracy and cell-level F1 score (**Methods**) for varying proportions of differentially expressed genes per cell type. CellAssign was provided with a set of marker genes (**Methods**); all other methods were provided with all genes. Asterisks indicate FDR-adjusted statistical significance (Wilcoxon signed-rank test) for pairwise comparisons between CellAssign and other methods. Nine simulated datasets were generated for each parameter setting. (**b**) Accuracy and cell-level F1 score for varying proportions of differentially expressed genes per cell type. All methods were provided with the same set of marker genes. Nine simulated datasets were generated for each parameter setting. *, **, *** denote FDR-adjusted p-vaues (Wilcoxon signed-rank test) for pairwise comparisons between CellAssign and other methods < 0.05, 0.01, 0.001 respectively. (**c**) Correspondence between true simulated log fold change values and log fold change (δ) values inferred by CellAssign. R refers to the Pearson correlation between true and inferred logFC values for CellAssign. n = 1000 single-cells were simulated. (**d**) Performance of CellAssign where a certain proportion of entries in the marker gene matrix are flipped at random. Nine simulated datasets were generated for each parameter setting. Dotted lines separate marker-based, unsupervised, and supervised methods. In all boxplots, lower and upper hinges denote the 1st and 3rd quartiles, with whiskers extending to the largest value less than 1.5 × the inter-quartile range.


a
b
c
d
% of genes differentially expressed per cell type
0.05
0.15
0.25
0.35
0.45
0.55
F1
Accuracy
Score
Method
cellassign
scina
SC3
seurat_0.4
seurat_1.2
phenograph
seurat_0.8
densitycut
dynamicTreeCut
Zheng_cor
scmap
Method type
Marker gene
Unsupervised
Supervised
p<0.001:*** p<0.01:** p<0.05:*

Supplementary Figure 3

Simulation performance (accuracy and cell-level F1 score) across a range of proportions of differentially expressed genes, where clusters from "unsupervised" algorithms are assigned to respective cell types through correlation to the transcriptomes of purified cell types.

Dotted lines separate marker-based, unsupervised, and supervised methods. (**a**) & (**b**) Simulating data based on parameters for B and CD8+ T cells, and running algorithms for whole-transcriptome and marker gene data only, respectively. (**c**) & (**d**) Simulating data based on parameters for naïve CD4+ and CD8+ T cells, and running algorithms for whole-transcriptome and marker gene data only, respectively. In all boxplots, lower and upper hinges denote the 1st and 3rd quartiles, with whiskers extending to the largest value less than 1.5 × the inter-quartile range. *, **, *** denote FDR-adjusted p-values (two-sided Wilcoxon signed-rank test) for pairwise comparisons between CellAssign and other methods < 0.05, 0.01, 0.001 respectively. In all cases, n = 1000 single-cells were simulated.

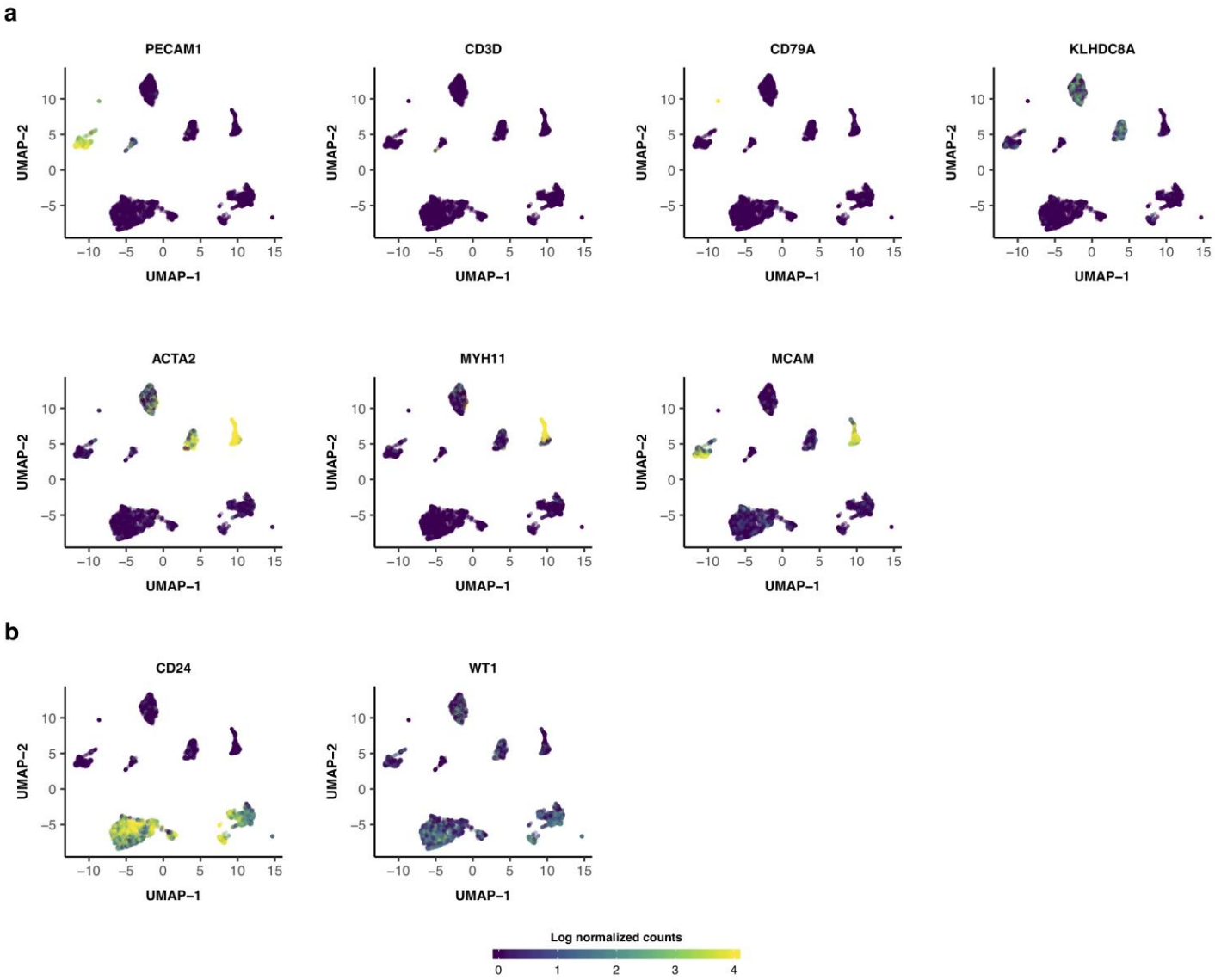


Supplementary Figure 4

Performance (accuracy and cell type-level F1 score, Methods**) of CellAssign and the best-performing clustering methods evaluated by [6] on FACS-purified H7 human embryonic stem cells in various stages of differentiation.**

t-SNE plots of (**a**) ground-truth FACS annotations; (**b**) CellAssign-derived annotations; (**c**) SCINA-derived annotations; (**d**) SC3 clusters (using all genes); (**e**) Seurat clusters (resolution = 0.8, using all genes); (**f**) Seurat

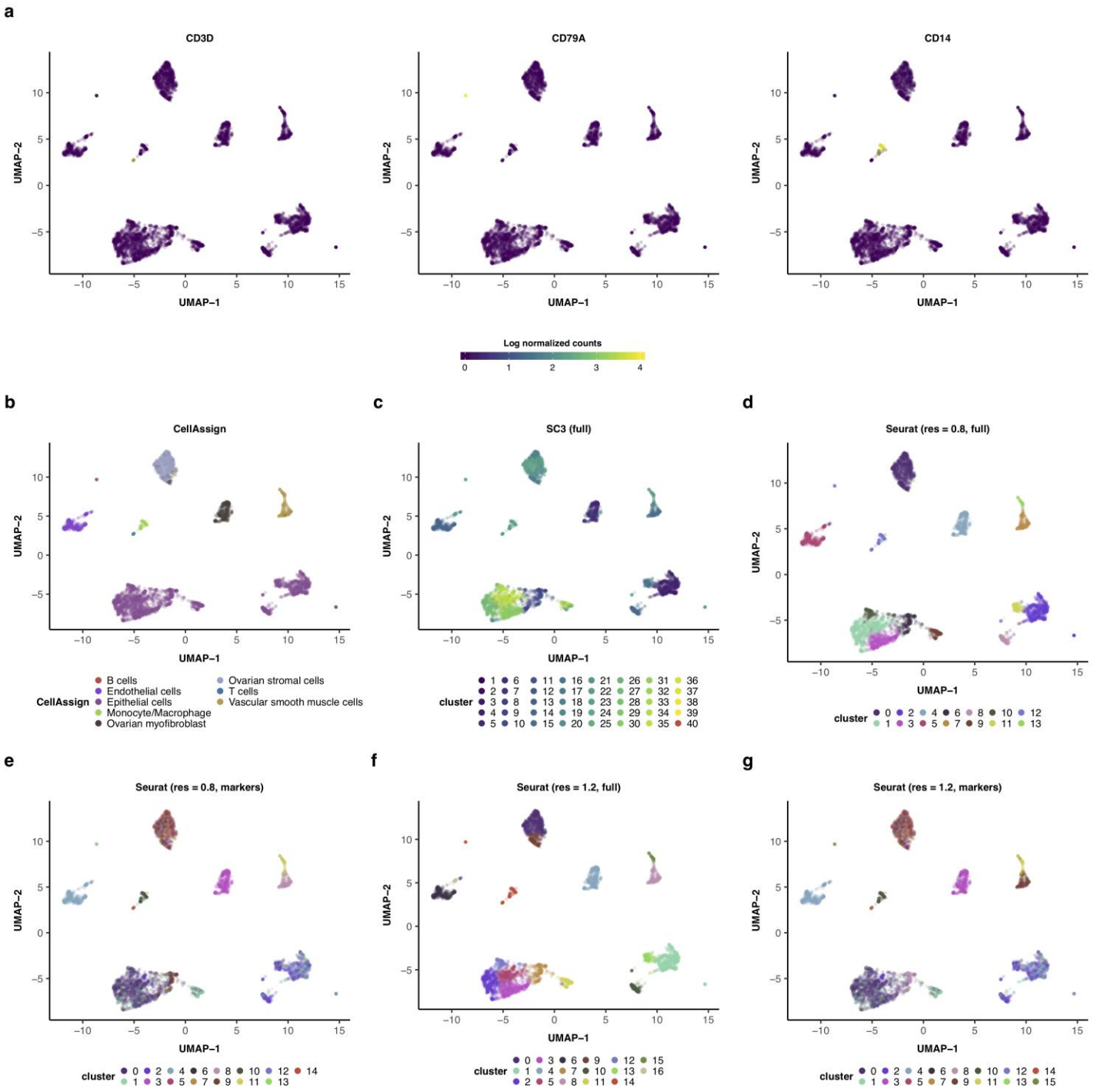
clusters (resolution = 0.8, using the same marker gene set used by CellAssign); (**g**) Seurat clusters (resolution = 1.2, using all genes); (**h**) Seurat clusters (resolution = 1.2, using the same marker gene set used by CellAssign).



Supplementary Figure 5

Expression of select marker genes in HGSC single cell RNA-seq data.

(**a**) Expression (log normalized counts) of *PECAM1* (for endothelial cells), *CD3D* (for T cells), *CD79A* (for B cells), *KLHDC8A* (for ovary-derived cells), *ACTA2* (for myofibroblasts and smooth muscle), *MYH11* (for smooth muscle), and *MCAM* (for vascular cell types including endothelial cells, vascular smooth muscle, and pericytes). (**b**) Expression (log normalized counts) of marker genes expressed in epithelial ovarian cancers but not in normal ovarian tissue. Expression values were winsorized between 0 and 4.

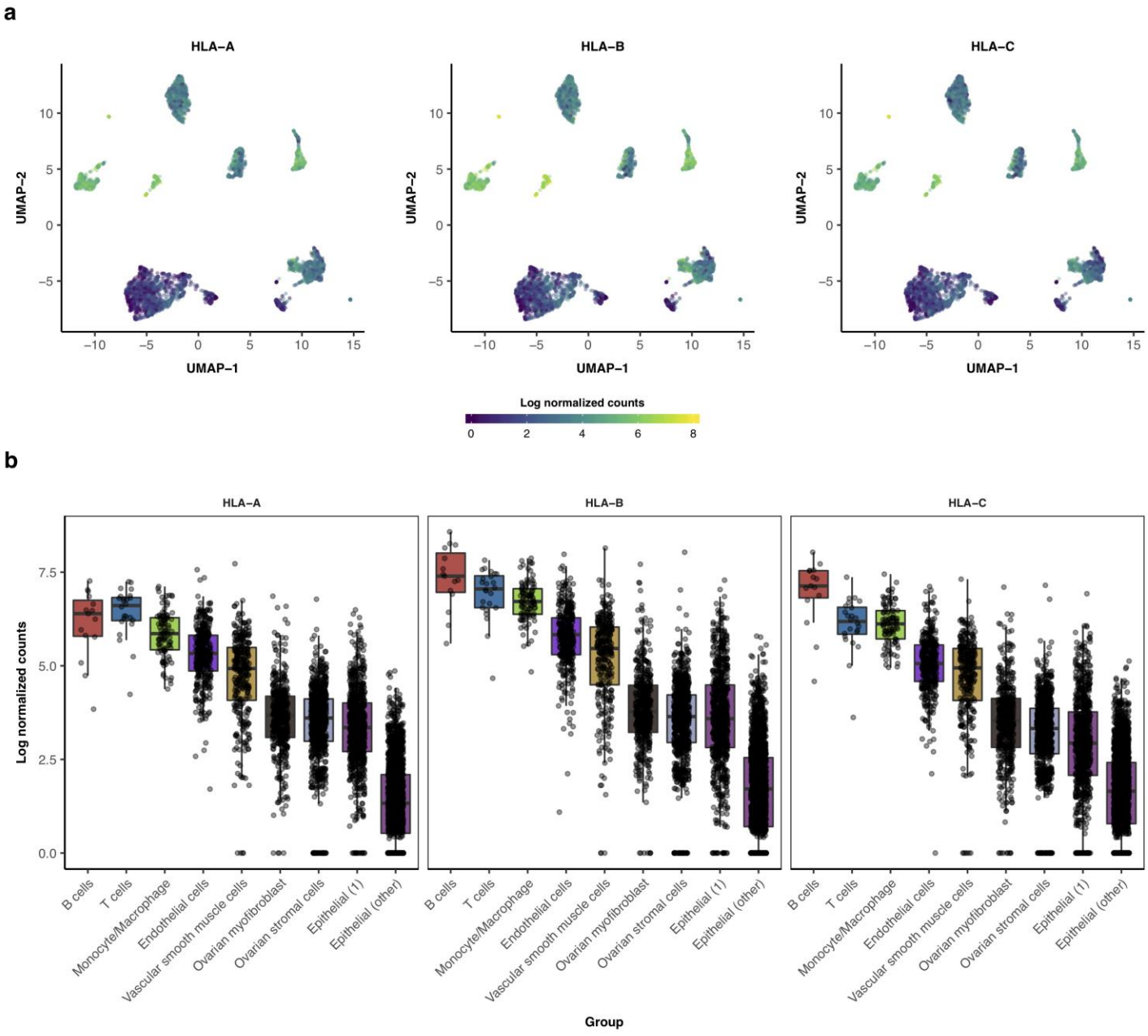


Supplementary Figure 6

Comparison of clusters from CellAssign and state-of-the-art un- supervised clustering approaches [6] on HGSC single cell RNA-seq data.

(**a**) Expression (log normalized counts) of key marker genes of hematopoietic subpopulations *CD3D* (for T cells), *CD79A* (for B cells), and *CD14* (for monocytes/macrophages). Expression values were winsorized between 0 and 4. UMAP plots of (**b**) CellAssign-derived annotations; (**c**) SC3 clusters (using all genes); (**d**) Seurat clusters (resolution

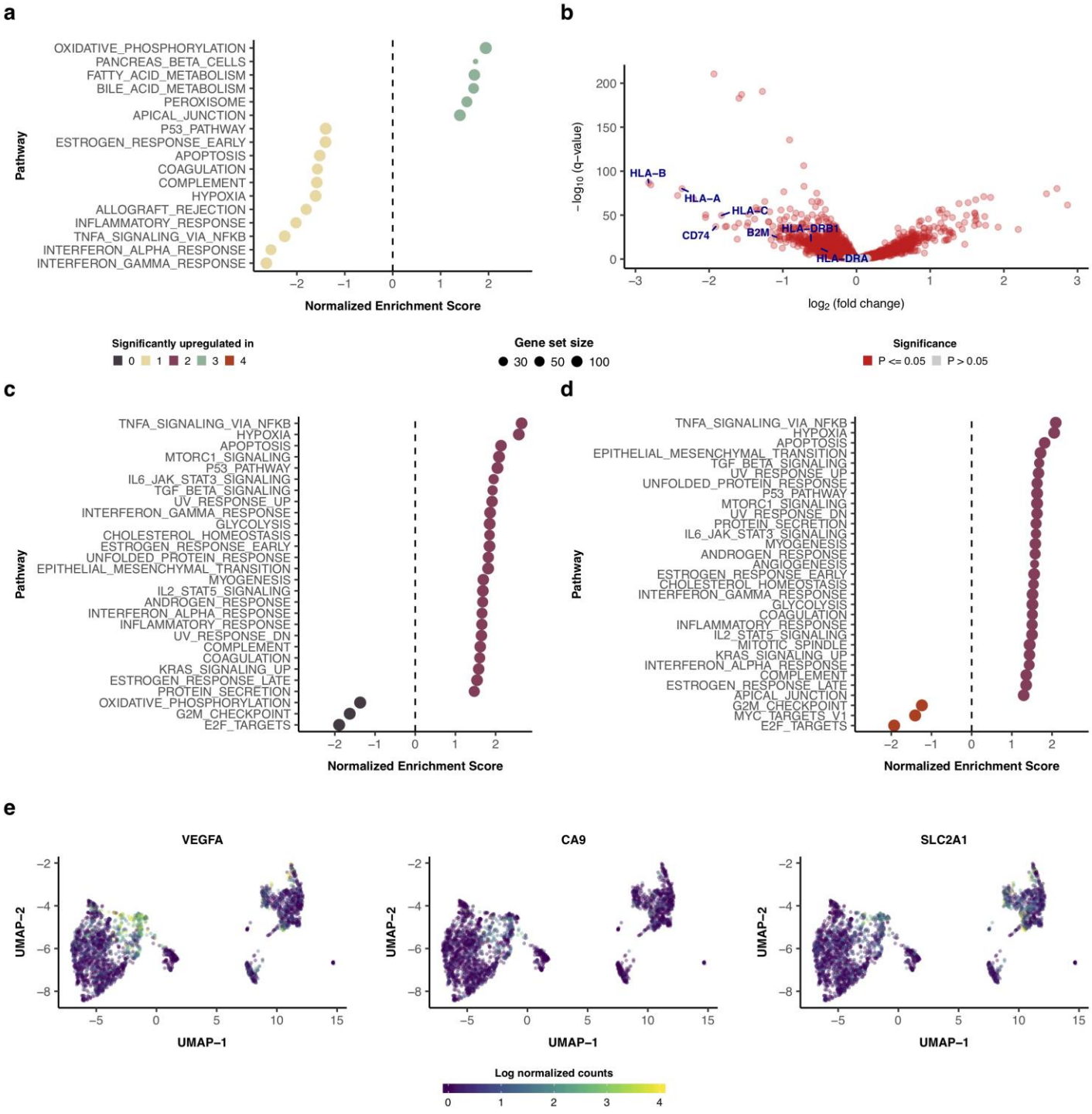
= 0.8, using all genes); (**e**) Seurat clusters (resolution = 0.8, using the same marker gene set used by CellAssign); (**f**) Seurat clusters (resolution = 1.2, using all genes); (**g**) Seurat clusters (resolution = 1.2, using the same marker gene set used by CellAssign).



Supplementary Figure 7

Cluster-specific HLA expression in HGSC epithelial cells.

(**a**) Expression (log normalized counts) of HLA class I genes in all HGSC cells. Expression values clipped from 0 to 8. (**b**) Expression of HLA class I genes across cell types in all HGSC cells. Epithelial (1): epithelial cells from cluster 1. Epithelial (other): epithelial cells from all other clusters. Lower and upper hinges denote the 1st and 3rd quartiles, with whiskers extending to the largest value less than 1.5 × the inter-quartile range. n = 4847 single-cells total (15 B cells, 24 T cells, 99 Monocyte/Macrophage, 361 endothelial, 300 vascular smooth muscle cells, 438 ovarian myofibroblast, 809 ovarian stromal 750 epithelial (1), 2051 epithelial (other)).

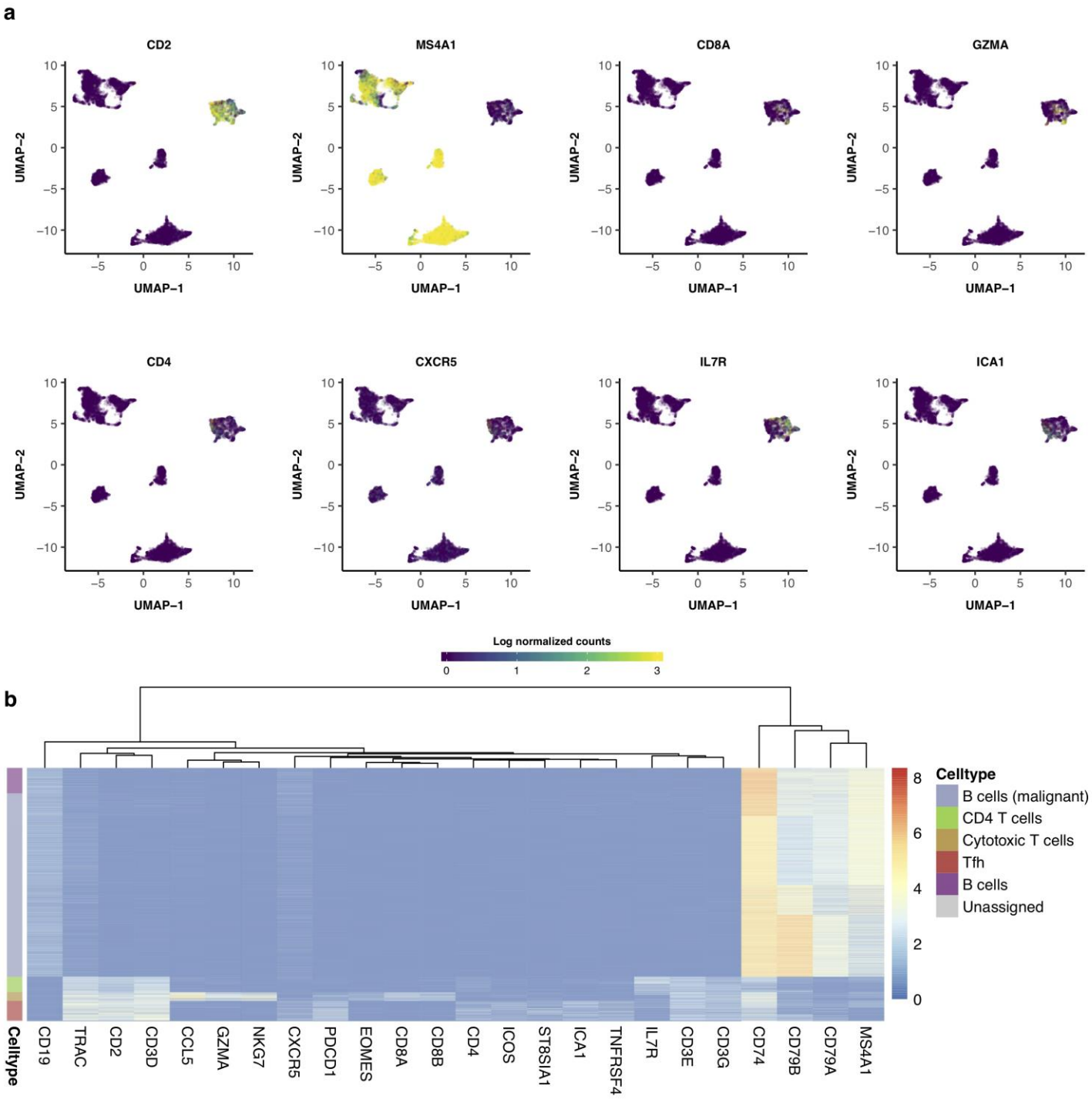


Supplementary Figure 8

Cluster-specific phenotypes in HGSC epithelial cells.

(**a**) Hall- mark pathway enrichment results for epithelial clusters 3 (n = 161 single cells) vs. 1 (n = 750 single cells) from the left ovary sample. (**b**) Gene-level differential expression for epithelial clusters 3 vs. 1, with statistical testing performed using the findMarkers func- tion in the scran R package. (**c** and **d**) Hallmark pathway enrichment

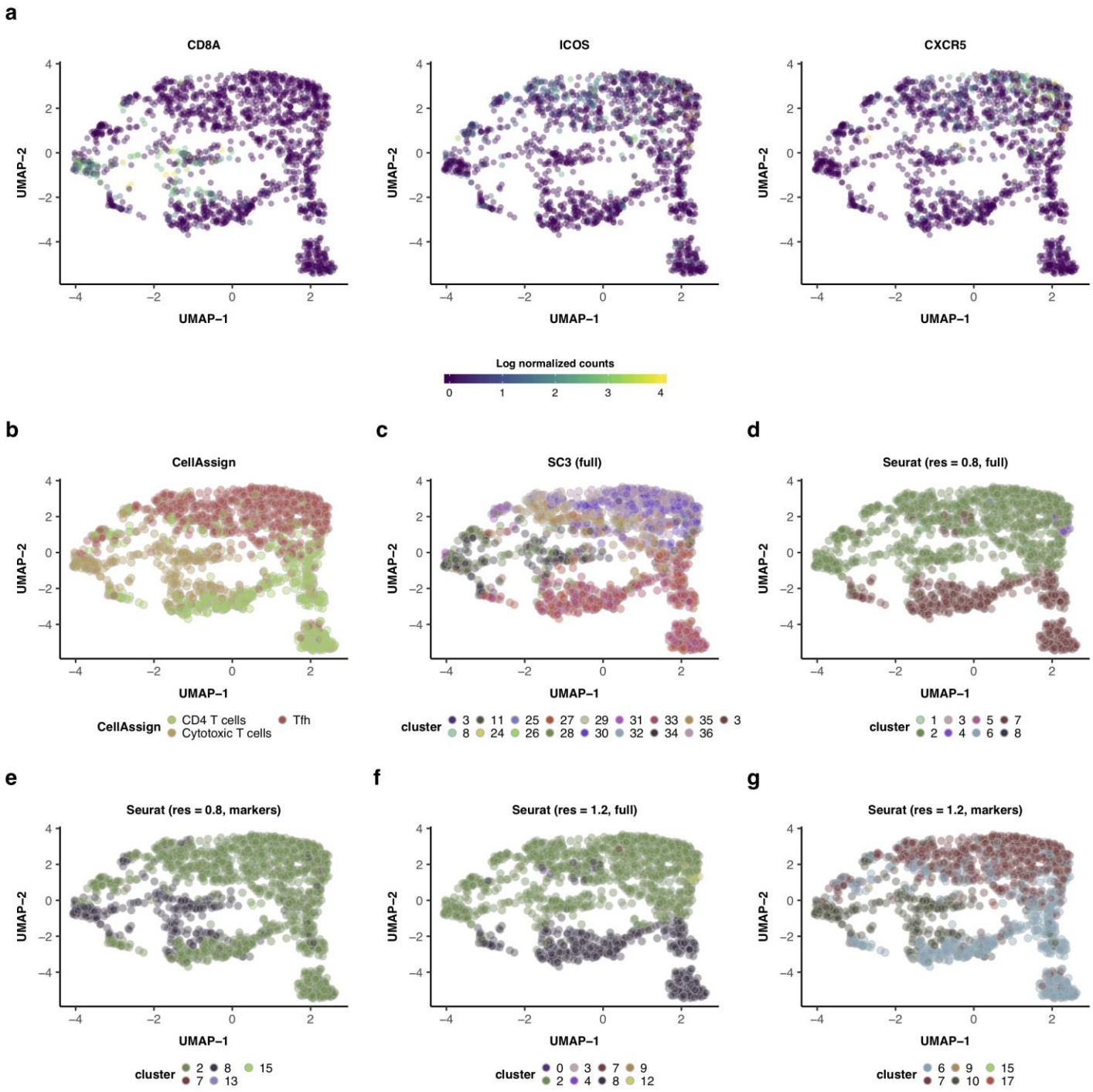
results for epithelial clusters (**c**) 2 vs. 0; and (**d**) 2 vs. 4. (**e**) Expression (log normalized counts) of select hypoxia-associated markers in HGSC epithelial cells. Expression values were winsorized between 0 and 4.



Supplementary Figure 9

Cell-type specific expression in follicular lymphoma.

(**a**) Expression (log normalized counts) of select marker genes *CD2* (for T cells), *MS4A1* (for B cells), *CD8A* and *GZMA* (for CD8+ T cells), *CD4* (for T follicular helper cells and other CD4+ T cells) and *CXCR5* and *ICA1* (for T follicular helper cells). Expression values were winsorized between 0 and 3. (**b**) Heatmap of marker gene expression, labeled by maximum probability CellAssign-inferred cell types.

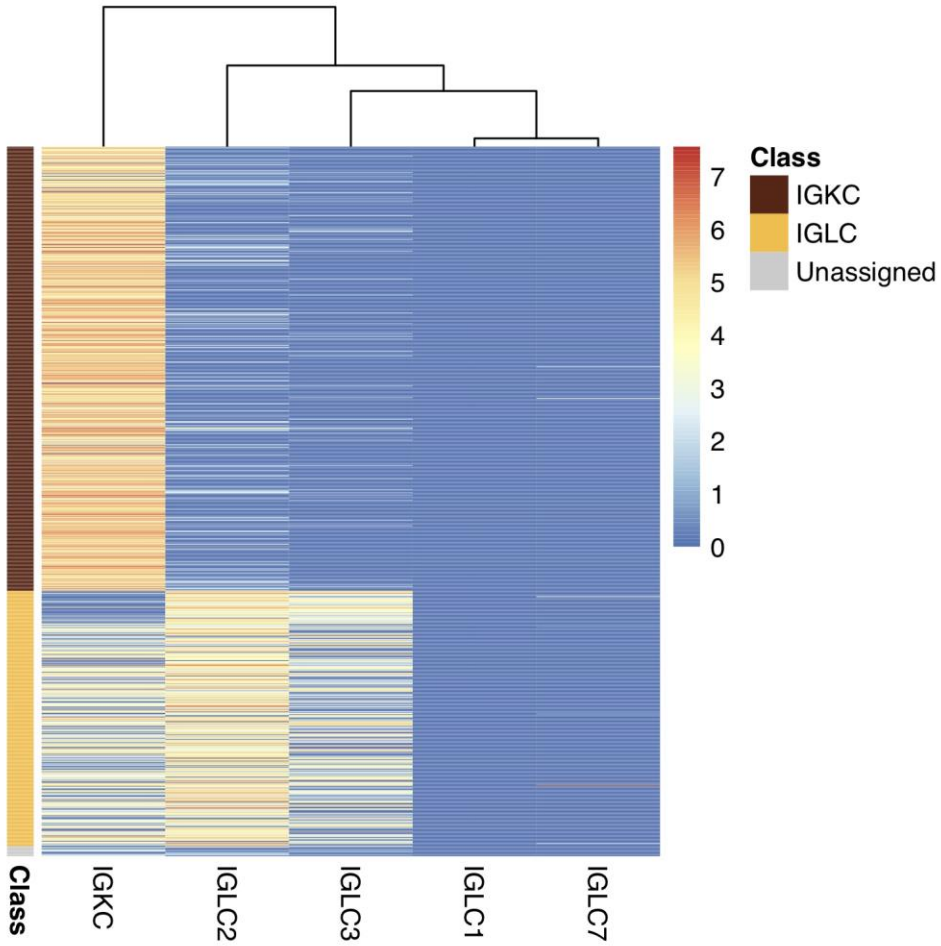


Supplementary Figure 10

Comparison of clusters from CellAssign and state-of-the-art unsupervised clustering approaches [6] on follicular lymphoma single cell RNA-seq data (showing only T cell subtypes).

(**a**) Expression (log normalized counts) of key T cell sub- population marker genes *CD8A* (for cytotoxic T cells), *ICOS* and *CXCR5* (for T follicular helper cells). Expression values were winsorized between 0 and 4. UMAP plots of (**b**) CellAssign-derived annotations; (**c**) SC3 clusters (using all genes); (**d**) Seurat clusters (resolution = 0.8, using all

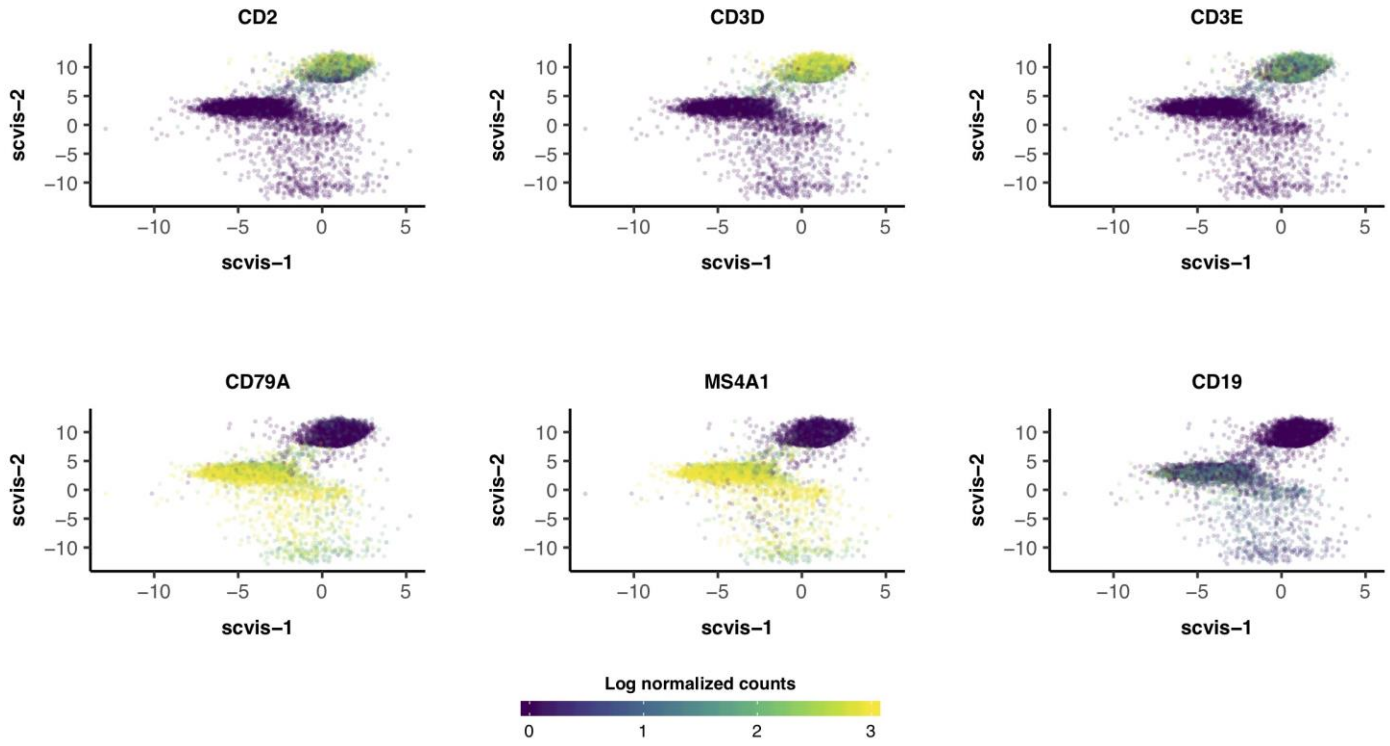
genes); (**e**) Seurat clusters (resolution = 0.8, using the same marker gene set used by CellAssign); (**f**) Seurat clusters (resolution = 1.2, using all genes); (**g**) Seurat clusters (resolution = 1.2, using the same marker gene set used by CellAssign).



Supplementary Figure 11

Expression (log normalized counts) of κ and λ light chain constant region genes in nonmalignant B cells.

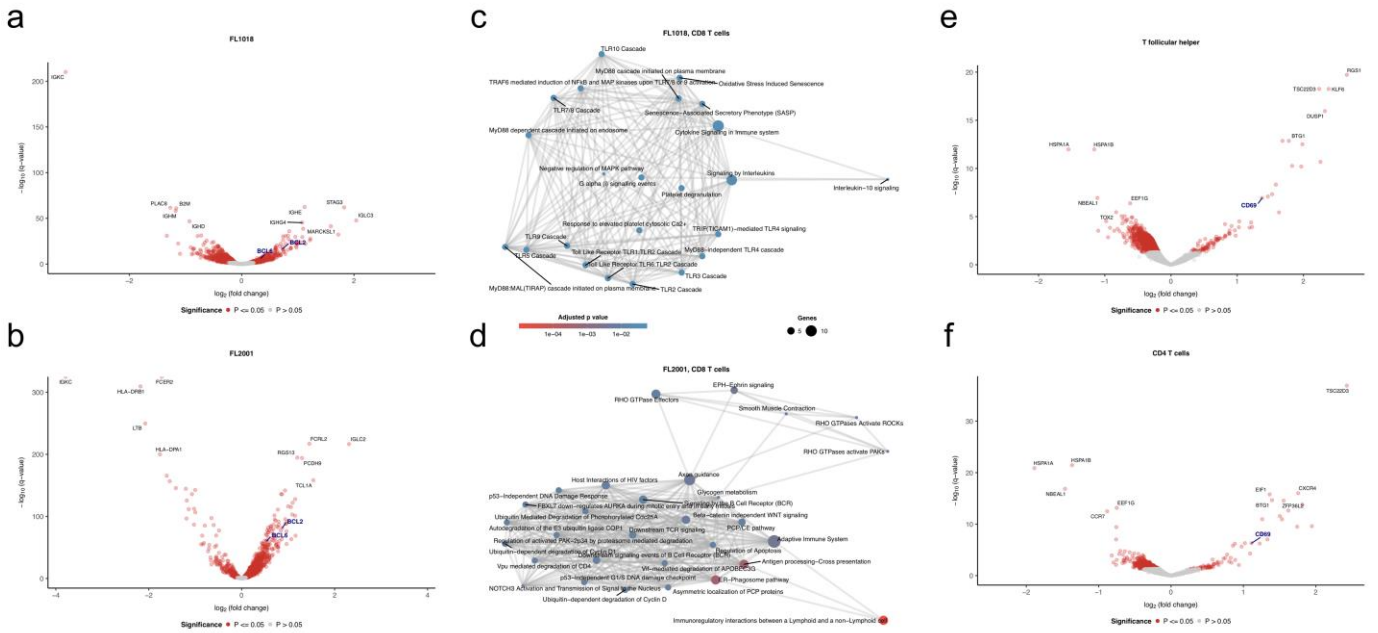
Class assignments were determined by CellAssign (**Methods**).



Supplementary Figure 12

Expression (log normalized counts) of selected marker genes (*CD2*, *CD3D*, and *CD3E* for T cells; *CD79A*, *MS4A1*, and *CD19* for B cells) in scvis embedding of reactive lymph node data.

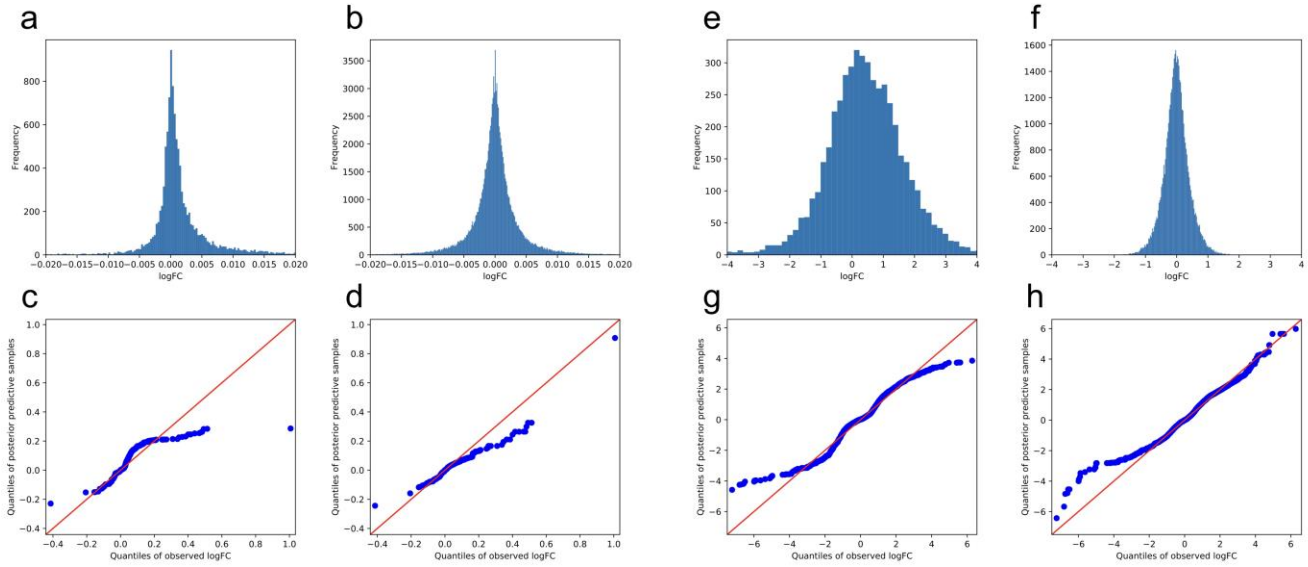
Expression values were winsorized between 0 and 3.



Supplementary Figure 13

Differential gene regulation for FL1018 and FL2001.

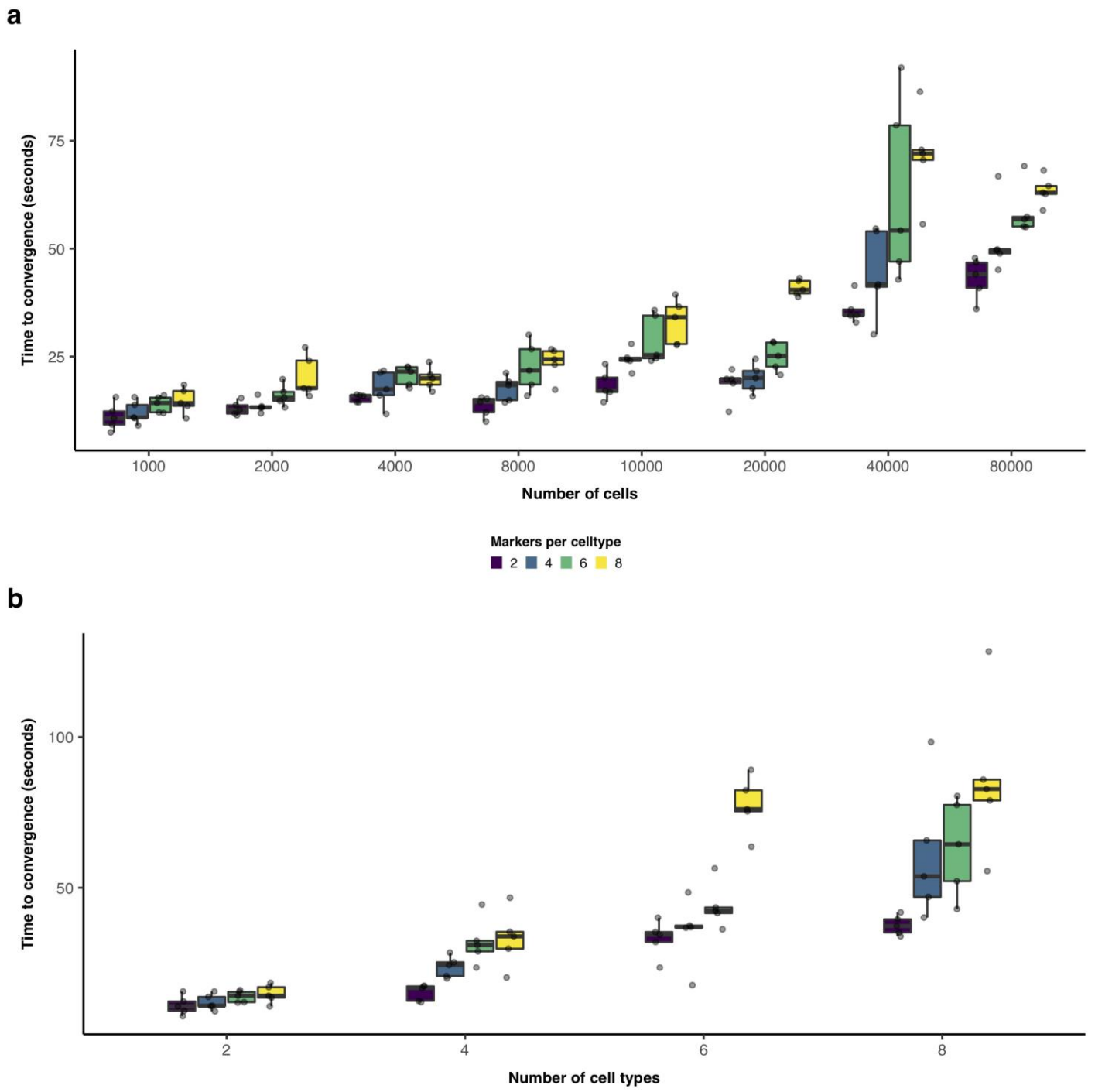
Differ- ential expression results using scran's findMarkers for malignant vs. nonmalignant B cells in (**a**) FL1018 and (**b**) FL2001. Comparisons was performed accounting for timepoint and potential interactions between malignant status and timepoint using a multivariate lin- ear model described in **Methods**. Genes upregulated among malignant cells have logFC values > 0. P-values were adjusted with the Benjamini-Hochberg method. Significantly enriched Reactome pathways (BH-adjusted P -value ≤ 0.05) among the top 50 most highly upregulated genes (ranked by log fold change) in (**c**) FL1018 and (**d**) FL2001. Up to 30 pathways are shown in either plot (**Methods**). Differentially expressed genes (found using scran's findMarkers) for (**e**) T follicular helper and (**f**) other CD4 T cells between T2 vs. T1. Genes upregulated in T2 have log fold change values > 0. The activation marker *CD69* is highlighted. P -values were adjusted with the Benjamini-Hochberg method.



Supplementary Figure 14

Fitting single cell RNA-seq simulation models to the Zheng PBMC 68k dataset, using cell type annotations provided in [51] (n = 66205 single cells), and of FACS prufied data from Koh et al. (n = 369 single cells).

(**a**) Log fold change values computed from differential expression analysis between naïve CD8+ and naïve CD4+ T cells. (**b**) 'Null' log fold change values computed by randomly splitting naïve CD8+ T cells into equally sized halves 10 times. (**c**) Quantile-quantile (QQ) plot comparing ob- served log fold change values between naïve CD8+ and naïve CD4+ T cells and poste- rior predictive samples from the splatter model (**Methods**). (**d**) Quantile-quantile (QQ) plot comparing observed log fold change values between naïve CD8+ and naïve CD4+ T cells and posterior predictive samples from the modified model (**Methods**). (**e**) Log fold change values computed from differential expression analysis between human embry- onic stem cells (hESCs) and day 3 somite cells (ESMT). (**f**) 'Null' log fold change values computed by randomly splitting anterior primitive streak cells into equally sized halves 10 times. (**g**) Quantile-quantile (QQ) plot comparing observed log fold change values be- tween hESC and ESMT cells and posterior predictive samples from the splatter model (**Methods**). (**h**) Quantile-quantile (QQ) plot comparing observed log fold change values between hESC and ESMT cells and posterior predictive samples from the modified model (**Methods**).



Supplementary Figure 15

Benchmarking results for CellAssign across a range of simu- lated data set sizes (number of cells), number of cell types being inferred, and number of marker genes per cell type.

(**a**) Runtime (to convergence, defined as a relative change in log-likelihood < 10^{-3} between successive iterations, as a function of data set size and the number of marker genes used per cell type, on simulated data (**Methods**). Two cell types were used. (**b**) Runtime (to convergence, defined as a relative change in log-likelihood < 10^{-3} between

successive iterations, as a function of the number of cell types and the number of marker genes used per cell type, on simulated data. One thousand cells were used. n = 5 simulated datasets were generated for each parameter setting. In all boxplots, lower and upper hinges denote the 1st and 3rd quartiles, with whiskers extending to the largest value less than 1.5 × the inter-quartile range.

Supplementary Note: Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling

Allen W Zhang[1,2,3], Ciara O'Flanagan[1], Elizabeth A Chavez[4], Jamie LP Lim[1,2], Nicholas Ceglia[2], Andrew McPherson[1], Matt Wiens[1], Pascale Walters[1], Tim Chan[1], Brittany Hewitson[1], Daniel Lai[1], Anja Mottok[4,5], Clementine Sarkozy[4], Lauren Chong[4], Tomohiro Aoki[4,8], Xuehai Wang[6], Andrew P Weng[6], Jessica N McAlpine[7], Samuel Aparicio[1,8], Christian Steidl[4], Kieran R Campbell[1,9,10,*], and Sohrab P Shah[1,2,8,*]

[1] Department of Molecular Oncology, British Columbia Cancer Research Centre, Vancouver, BC, Canada
[2] Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York, USA
[3] BC Children's Hospital Research, Vancouver, BC, Canada
[4] Centre for Lymphoid Cancer, British Columbia Cancer Research Centre, Vancouver, BC, Canada
[5] Institute of Human Genetics, Ulm University and Ulm University Medical Center, Ulm, Germany
[6] Terry Fox Laboratory, British Columbia Cancer Research Centre, Vancouver, BC, Canada
[7] Department of Gynecology and Obstetrics, University of British Columbia, Vancouver, BC, Canada
[8] Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, Canada
[9] Department of Statistics, University of British Columbia, Vancouver, BC, Canada
[10] UBC Data Science Institute, University of British Columbia, Vancouver, BC, Canada
[*]Correspondence: kieran.campbell@stat.ubc.ca, shahs3@mskcc.org

Contents

Cell type	Marker genes
Cholangiocytes	KRT19, EPCAM, FXYD2, CLDN4, CLDN10, SOX9, MMP7, CXCL1, CFTR, TFF2, KRT7, CD24
Mature B cells	MS4A1, IGHD, CD79A[c], PTPRC[c], IGKC[c], CD19[c]
Hepatocytes	ALB, HAMP, ARG1, PCK1, AFP, BCHE
LSECs	CALCRL, VWF, PECAM1[c], CLEC14A[c], EMCN[c]
Hepatic Stellate Cells	ACTA2, COL1A1, TAGLN, COL1A2, COL3A1, SPARC, RBP1, DCN, MYL9
Macrophages	CD68, MARCO, FCGR3A[c], LYZ[c], PTPRC[c], AIF1[c]
ab T cells	CD2, CD3D, TRAC, IL32, CD3E, PTPRC[c]
gd T cells	NKG7, FCGR3A, HOPX, GNLY, CMC1[c], KLRF1[c], CCL3 [3], PTPRC[c]
NK cells	GZMK, KLRF1, CCL3, CMC1, NKG7[c], PTPRC[c]
Plasma cells	CD27, IGHG1, IGHA1[c], IGHM[c], CD79A[c], PTPRC[c], IGKC[c]
Erythroid cells	HBB, SLC25A37, CA1, ALAS2

Supplemental Note Table 1. Revised markers for the MacParland et al. dataset. Rationale for added markers shown. [c]: Canonical marker.

1 Application of CellAssign to MacParland et al. liver dataset

For several robustness tests of CellAssign (**Supplementary Notes** 2 below) we use the McParland et al. [1] dataset of 11 cell types in the liver from 5 patients (with 6872 cells post-QC from the paper).

We use cell type signatures for the 11 cell types from the paper (alpha/beta T cells, gamma/delta T cells, cholangiocytes, erythroid cells, hepatic stellate cells, hepatocytes, liver sinusoidal endothelial cells, macrophages, mature B cells, NK cells, plasma cells), as defined in the authors' GitHub repository from the function `viewHumanLiver()` (`https://github.com/BaderLab/HumanLiver`).

The authors' original cell type classifications are shown in **Supplemental Note Figure** 1A. We applied CellAssign to the entire dataset using the authors' original marker genes for each cell type, as well as using our own revised set of marker genes (**Supplemental Note Table** 1). A UMAP embedding of the dataset coloured by cell assignments in each case can be seen in **Supplemental Note Figure** 1B,C respectively. We further applied SCINA [2] using both marker gene sets, as can be seen in **Supplemental Note Figure** 1D,E respectively. CellAssign achieved higher accuracy and cell-level F1 scores than SCINA in both cases (**Supplemental Note Figure** 1B-E).

2 Robustness of CellAssign

2.1 Erroneously specified marker genes

We next assessed the robustness of CellAssign to misspecification of marker gene information, acknowledging the likely scenario where user-provided marker gene information may be incomplete or incorrect. For example, a shared marker gene may be incorrectly

specified as a cell type-specific marker gene due to incomplete prior information. We randomly flipped a proportion of entries in the binary marker gene matrix to introduce error. When supplied with data for 3 cell types with 5 marker genes per cell type ($n = 1500$ cells total), CellAssign maintained comparable performance in scenarios where up to 30% of matrix entries were misspecified (**Supplemental Figure** 2D, **Supplemental Table** 1). This robustness was maintained even when cells belonged to transcriptionally similar cell types containing fewer highly differentially expressed genes. For example, when cells were simulated based on the degree of dissimilarity between naïve CD4+ and naïve CD8+ T cells, CellAssign prediction accuracy was maintained in scenarios where 30% of marker gene matrix entries were misspecified (**Supplemental Figure** 1C,D, **Supplemental Table** 1).

2.2 Over-specifying the expected number of cell types

We sought to evaluate the robustness of CellAssign to specifying signatures for cell types that are not present in the data. We demonstrate this using both simulations (**Figure** 2B, **Methods**) and the MacParland et al. dataset as described above.

We first simulated data for 6 cell types and then subsetted 2-6 cell types from the result (**Methods**). Marker genes were inferred on data from all 6 cell types and used for CellAssign and SCINA. SCINA was run with `sensitivity_cutoffs` of 1 (the default) and 0.1 (valid values range from 0 to 1), with lower values of this parameter facilitating the automatic removal of non-existent cell types [2]. Across all tested parameter combinations, CellAssign performed superiorly to SCINA (**Figure** 2B).

To test this on a real data example, we consider subsetting 3 cell types from the McParland et al. dataset: hepatocytes, cholangiocytes, and mature B cells. We then add in the marker genes for a fourth cell type — NK cells. A UMAP representation of the data can be seen in **Supplemental Note Figure** 2A. The *maximum a posteriori* (MAP) classifications from CellAssign (using default parameters, and patient as a batch covariate), without any filtering, are shown in **Supplemental Note Figure** 2B. Using marker genes for the 3 cell types present gives an accuracy of 99.8% (3633/3639, F1: 0.998) (**Supplemental Note Figure** 2B) compared to cell type labels from MacParland et al., while using 4 cell types (1 not present) achieves a comparable accuracy of 99.8% (3632/3639, F1: 0.998) (**Supplemental Note Figure** 2C).

To further illustrate this point, we consider an extreme case by providing marker gene lists for all 11 cell types when only 3 are present in the data. CellAssign retains accurate cell type classifications (98.5%, 3583/3639 correct, F1: 0.985) even in the face of extreme cell type overspecification (**Supplemental Note Figure** 2D). We note that the rate of misclassification here could be reduced by, e.g. grouping the data with other data that contains some of the non-existent cell types and specifying a batch variable.

We repeated this analysis with SCINA [2] and found a substantial reduction in accuracy when the number of cell types was overspecified (3311/3639 correct, 91%, F1: 0.91 with default SCINA parameters; 3383/3639 correct, 93% accuracy with a `sensitivity_cutoff` of 0.1 that did not appear to be improved by batch correction with scanorama [4]; **Supplemental Note Figure** 2E-L).

Cell type	Marker genes
Cholangiocytes	KRT19, EPCAM, FXYD2, CLDN4, CLDN10, SOX9, MMP7, CXCL1, CFTR, TFF2, KRT7, CD24
Mature B cells	MS4A1, LTB, CD52, IGHD
Hepatocytes	ALB, HAMP, ARG1, PCK1, AFP, BCHE

Supplemental Note Table 2. Markers for different cell types in the human liver as per MacParland et al. 2018 [1].

2.3 Under-specifying the expected number of cell types

We next sought to test CellAssign's performance in situations where some cell types in the data are not specified in the marker gene list. We first evaluate CellAssign's performance on simulations, where 6 cell types are simulated, but marker genes are only generated for 2-6 of these cell types. Across all tested parameter combinations, CellAssign performed superiorly to SCINA (**Figure** 2C).

We next consider a case where the data contains 4 cell types — hepatocytes, cholangiocytes, mature B cells, and NK cells — but CellAssign only "knows" about the first 3, using the MacParland et al. data as above. **Supplemental Note Figure** 3A shows a UMAP representation of the entire dataset, coloured by original cell type [1]. Applying CellAssign as above accurately recapitulates cholangiocyte, hepatocyte, and mature B cell groups, while classifying NK cells as 'other' (**Supplemental Note Figure** 3B). In total, predictions from 3890/3934 (98.9%, F1: 0.989) cells have the same classes as labeled in the McParland et al. paper, according to their MAP classifications from CellAssign.

2.4 Distinctiveness of novel cell type group

We next consider an extension to the above analysis to address how distinctive the novel cell type group must be to be accurately detected. We applied the above analysis but now with all cell types present in the data and only 3 applied through the marker matrix as input to CellAssign. As we are providing many other cell types, this will allow us to determine how different each cell type has to be in order to be quantified accurately. In this case CellAssign achieves a high overall accuracy (86.4%, F1: 0.864), but incurs some misclassification with the mature B cell set encompassing both mature B cells and alpha-beta T cells. (**Supplemental Note Figure** 3C).

However, this mis-classification appears to be due to ambiguity in the marker genes from McParland et al. (**Supplemental Note Table** 2). While *MS4A1* and *IGHD* are known to be specific to B cells, *LTB* and *CD52* are not. *CD52* is widely known to be expressed on mature lymphocytes and some myeloid populations [5], and *LTB* is likewise known to expressed on activated lymphocyte populations [6]. Plotting the expression of these markers on the MacParland data demonstrates agreement with these observations (**Supplemental Note Figure** 4).

Thus the marker gene specification for mature B cells here is ambiguous relative to other immune populations, particularly alpha-beta T cells, which express high levels of *LTB* and *CD52*, from the plot above. If we instead amend the marker gene list for mature B cells

to be *MS4A1*, *CD19*, *IGHD*, where the well-known canonical B cell marker *CD19* is used in place of *LTB* and *CD52*, we much more accurately recapitulate the reference cell types from McParland et al. (accuracy = 6690/6872 = 97.4% (F1: 0.974), **Figure** 2D,E). SCINA achieves similar performance (**Supplemental Note Figure** 3D-F).

Thus, even with as few as 3 specific marker genes with modest expression levels (maximum logCPM values of approximately 3, from the expression plot shown above), mature B cells can be differentiated from biologically similar cell types in the data (T cells and plasma cells).

2.5 Increasing number of marker genes

To test the robustness of CellAssign to a possibly large number of marker genes and diverse cell types as input, we applied CellAssign to both the HGSC and follicular lymphoma datasets presented in the main text. We concatenated the cell type marker matrices from the main text (**Supplemental Table** 2), eliminating nonsensical overlap (e.g. B cells in the HGSC table with IGKC/IGLC in the follicular lymphoma table), to give **Supplemental Note Table** 3. We take the union of marker genes from both lists where overlap is present (e.g. VIM is also expressed in B cells (kappa) since it is expressed in B cells in the original HGSC marker gene matrix).

The results of applying CellAssign to each cancer separately compared to combining all data can be seen in **Supplemental Note Figure** 5A-C. These assignments demonstrate that follicular lymphoma-specific cell types are restricted to the follicular lymphoma dataset, and likewise for HGSC-specific cell types to the HGSC dataset). We observe high correspondence between the results from the full marker gene list and the 'hierarchical' approach (HGSC and follicular lymphoma separately). The correspondence between the combined HGSC + follicular lymphoma and HGSC classifications for the HGSC dataset is 4641/4848 (95.7%, F1: 0.957). Likewise, the correspondence between the combined HGSC + follicular lymphoma and follicular lymphoma classifications for the follicular lymphoma dataset is 8979/9156 (98.1%, F1: 0.981).

2.6 Hierarchical specification of cell types

We next demonstrate the robustness of CellAssign to the hierarchical specification of cell types. In the follicular lymphoma data presented in the main text, B cells can be further subdivided into IGKC and IGLC subsets. We can then perform cell assignment in 2 ways: (i) classifying B cells first and then separating into IGKC and IGLC subsets, or (ii) classifying as B cells (kappa) or B cells (lambda) from the beginning. Designating (i) as the "hierarchical" approach and (ii) as the "combined" approach, we evaluate correspondence between the probabilities from each method. To evaluate how the outcome of the cell assignment using CellAssign varies in these two settings, we evaluate the consistency of cell type assignments via the correlation of the assignment probabilities to each cell type (**Supplemental Note Figure** 5D). Using default CellAssign parameters with `include_other = FALSE` (as the 'other' group could map to different cell subpopulations in the hierarchical approach), we found high concordance in MAP assignments, both among major cell

Cell type	Marker genes
B cells (kappa)	VIM, CD19, MS4A1, CD79A, CD79B, CD74, CXCR5, IGKC, PTPRC, BANK1
B cells (lambda)	VIM, CD19, MS4A1, CD79A, CD79B, CD74, CXCR5, IGLC2, IGLC3, PTPRC, BANK1
Cytotoxic T cells	VIM, CD2, CD3D, CD3E, CD3G, TRAC, CD8A, CD8B, GZMA, NKG7, CCL5, EOMES, PTPRC
CD4 T cells	VIM, CD2, CD3D, CD3E, CD3G, TRAC, CD4, IL7R, PTPRC
Tfh	VIM, CD2, CD3D, CD3E, CD3G, TRAC, CD4, CXCR5, PDCD1, TNFRSF4, ST8SIA1, ICA1, ICOS, PTPRC
Monocyte/Macrophage	VIM, CD14, FCGR3A, CD33, ITGAX, ITGAM, CD4, PTPRC, LYZ
Epithelial cells	EPCAM, CDH1, KRT8, WFDC2
Ovarian stromal cells	VIM, MUM1L1, FOXL2, ARX, DCN, TPT1, RBP1
Ovarian myofibroblast	VIM, MUM1L1, FOXL2, ARX, ACTA2, COL1A1, COL3A1, SERPINH1
Endothelial cells	VIM, EMCN, CLEC14A, CDH5, PECAM1, VWF, MCAM, SERPINH1

Supplemental Note Table 3. Concatenated marker gene matrix for both ovarian and follicular lymphoma cell types.

types (9113/9156, 99.5%, F1: 0.995), and among lambda/kappa type nonmalignant B cells (861/920, 93.6%, F1: 0.936).

We use the Koh et al. [7] to presented in the main text to further demonstrate this. We assessed the consistency of cell assignment when (i) combining the phenotypically similar anterior primitive streak (APS) and mid-primitive streak (MPS) cells into one group (APS_MPS), and then first performing CellAssign using the APS_MPS group followed by a run on APS_MPS cells to split into separate APS and MPS groups; and (ii) classifying as APS or MPS (among all other cell types) from the beginning. Markers for the APS_MPS group were taken as the intersection of APS and MPS markers. Shared markers between APS and MPS cells were removed when classifying only APS_MPS cells into APS or MPS categories. We show that assignment probabilities are consistent between these two approaches (**Supplemental Note Figure** 5E). We also found high concordance in MAP assignments, both among the major cell types (440/446, 98.7%, F1: 0.987) and APS/MPS cells (86/88, 97.7%, F1: 0.977).

2.7 Robustness across multiple measurement technologies

We additionally show that CellAssign achieves high performance across different modalities of single cell RNA-sequencing. Specifically, we consider single cell RNA-seq data from the CellBench dataset ([8], `github.com/LuyiTian/CellBench_data/`) for the 3 cell lines (H1975, H2228, and HCC827) considered above, for 3 different experimental protocols: 10x Genomics, CEL-Seq2, and Drop-Seq.

We then derive marker genes for these 3 cell types using orthogonal bulk RNA-seq data for these 3 cell lines from the same group [9]. We use limma voom for differential expression, selecting genes that are significantly differentially expressed in at least one of the contrasts (FDR < 0.05) with maximal log fold changes between any pair of cell types ($\geq$99th

percentile). This yielded a marker gene list of 73 genes.

With this input, CellAssign achieves an accuracy of 100% (1332/1332) when jointly considering data for all 3 scRNA-seq methods (**Supplemental Note Figure** 6A-C). Similarly, accuracy was high when analyzing data for each method separately: 99.9% (889/890) for 10x, 100% (232/232) for CEL-Seq2, and 100% (210/210) for Drop-Seq.

2.8 Calibration of output probabilities

The probabilities returned by CellAssign can be used to assess the uncertainty in assignment. To evaluate the concordance of such probabilities to the ground truth, we present analysis of the CellBench data [8] as above. This data contains "pseudo-cells" that have been sequenced using CEL-Seq2. These pseudocells are 9-cell mixtures containing defined mixtures of 3 closely related human lung adenocarcinoma cell lines HCC827, H2228, and H1975. Thus, we can relate the mixture proportions of these cell types to CellAssign probabilities. While the CellAssign probabilities themselves may vary depending on the number and relative strength of marker genes specified for each cell type (i.e. specifying many more marker genes for one cell type than another will increase the confidence of assignments for the former type), we expect to observe an approximately monotonic relationship between mixture proportion and CellAssign probability for each cell type.

To do this comparison, we combine the pure CEL-Seq2 data for these 3 cell lines (`sc_CEL-seq2` on the Github page above) with the "cellmix4" mixture (the mixture considered in their QC file, labeled `sce_SC4_qc`, `https://github.com/LuyiTian/CellBench_data/blob/master/script/data_QC_visualization/all_data_plot.Rmd`). We then randomly subsetted various numbers of marker genes (20, 30, and 50) from the 73 genes described above to demonstrate stability across the number of selected marker genes.

For each cell type, we compare the number of cells in each 9-cell pseudocell (9 corresponds to 100% of that cell type, 5 would correspond to 55.6%, for example), to probabilities for CellAssign (default settings), the results of which can be seen in **Figure** 2F and **Supplemental Note Figure** 6D). Indeed, we observe an approximately monotonic relationship for each cell type and each marker gene matrix. Increasing the number of marker genes appears to increase the confidence of predictions, while relatively even mixtures (e.g. 3-6 cells out of 9) demonstrate uncertainty in classification as expected.

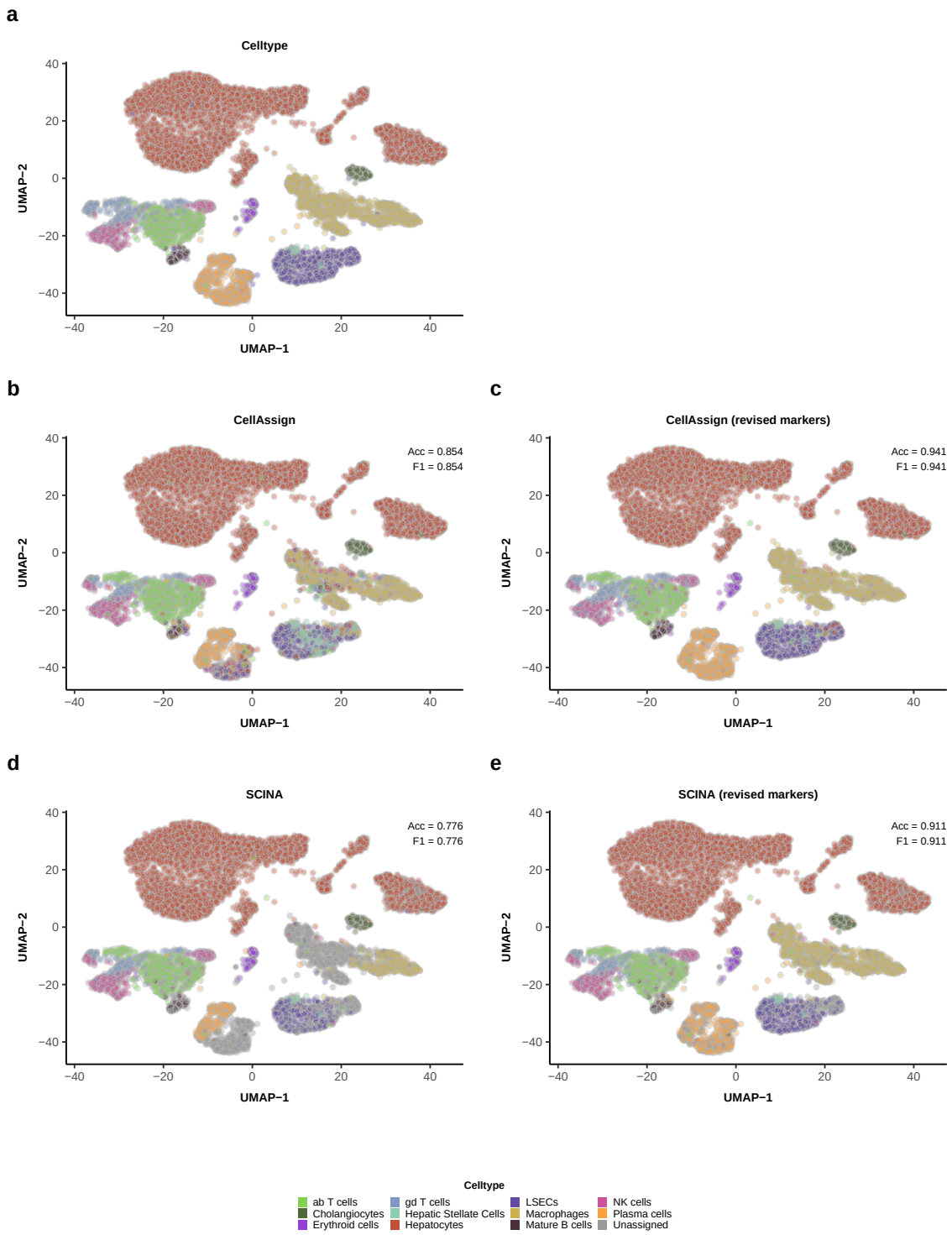
To evaluate the confidence of predictions from CellAssign for each cell, we computed the Shannon entropy of probabilities in the CellAssign probability matrix for each cell. For each selection of the number of marker genes used, pure 'pseudocells' (i.e 9/9 for a particular cell line) had lower entropy values – significantly so for 30 and 50 marker genes. FDR-adjusted Wilcoxon p values are show significant decreases in entropy for both 30 and 50 marker genes (**Supplemental Note Figure** 6E). Thus the deviation in entropy for pure vs. mixed cell types increases for higher numbers of marker genes. Thus the entropy of the assignment probabilities can be used as a diagnostic statistic for how well a cell has been assigned.

3 Constructing a marker gene list for input to CellAssign

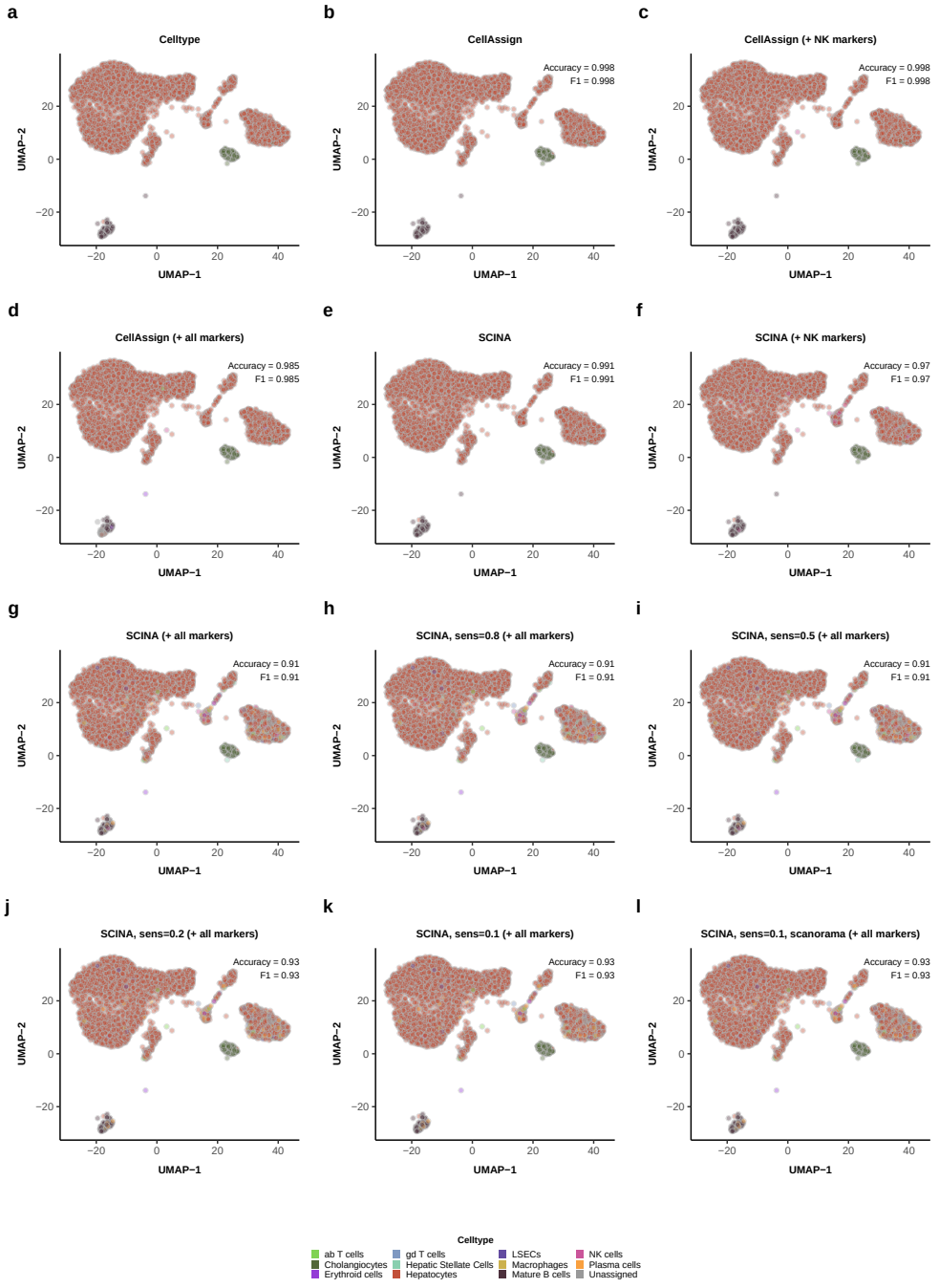
Several such marker panels are now publicly available, including CellMarker [10] (`http://biocc.hrbmu.edu.cn/CellMarker/`), the Human Protein Atlas (which also contains RNA-seq-defined markers, `https://www.proteinatlas.org/`), and PanglaoDB (`https://panglaodb.se/index.html`). PanglaoDB integrates single cell RNA-seq data from over 1000 publicly available samples encompassing 200+ tissue types and 4.5 million+ cells, identifying marker genes (see Datasets -> Marker genes from the home page) under a unified framework. While the list is not complete (e.g. it lacks ovarian cancer-specific markers, such as *WFDC2*, that are used in our manuscript based on RNA-seq derived Human Protein Atlas data), it contains a comprehensive list of marker genes for many tissue types. We do not believe that it would be useful to duplicate their efforts, which are continuously being updated.

We further provide a framework for identifying marker genes from orthogonal bulk RNA-seq data as a vignette in the `cellassign` package (`https://irrationone.github.io/cellassign/constructing-markers-from-purified-data.html`). We have also included an additional example showing how this can be applied to real data (CellBench) from CEL-Seq2 (in addition to the 10x and SMART-Seq-based analyses provided in the paper).

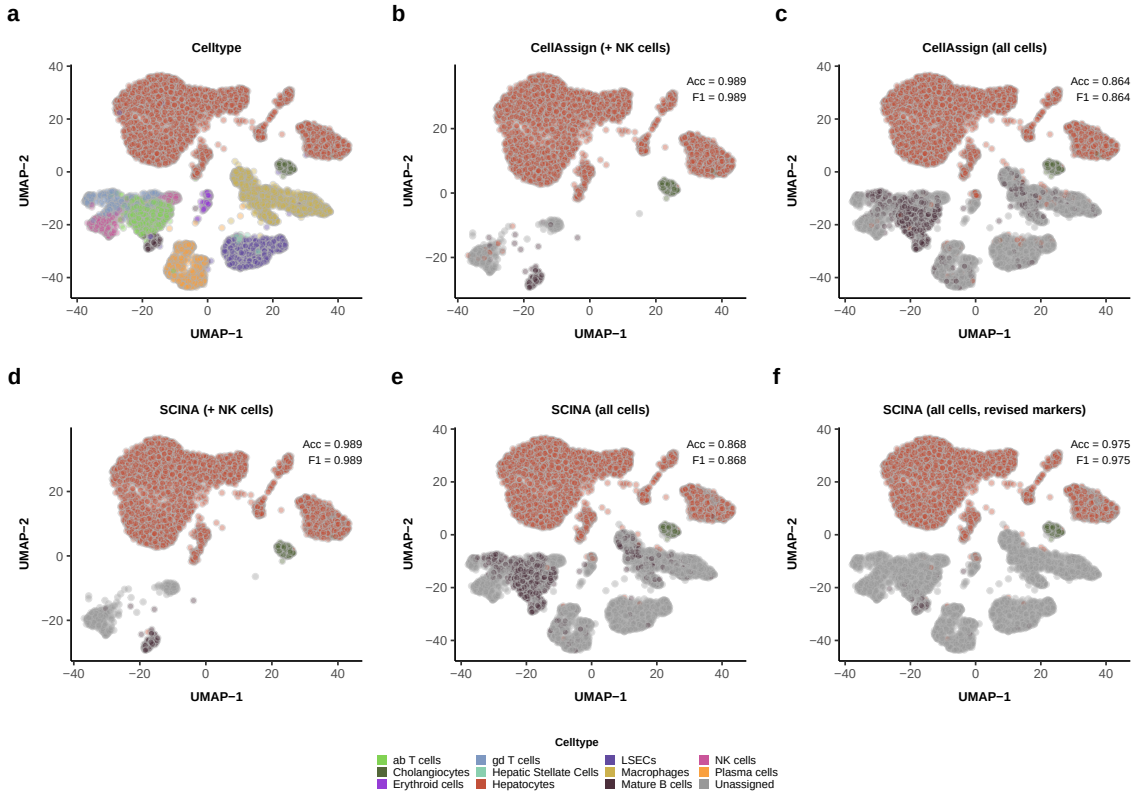
To demonstrate the utility of data from these databases, we provide an example of using the liver cell type signatures in the PanglaoDB database (currently only available for a few cell types) on the MacParland et al. data considered above. Using human marker genes labeled as canonical in PanglaoDB for hepatocytes and cholangiocytes (total of 112 marker genes), we use CellAssign to assign all the data in the MacParland et al. liver dataset. Having run CellAssign with default parameters and patient as a batch variable, a UMAP representation of the data coloured by both original and CellAssign cell type assignment can be seen in **Supplemental Note Figure** 7. We find that using PanglaoDB for marker genes assigns cholangiocytes and hepatocytes in the data with high accuracy (6781/6872, 98.7%, F1: 0.987).



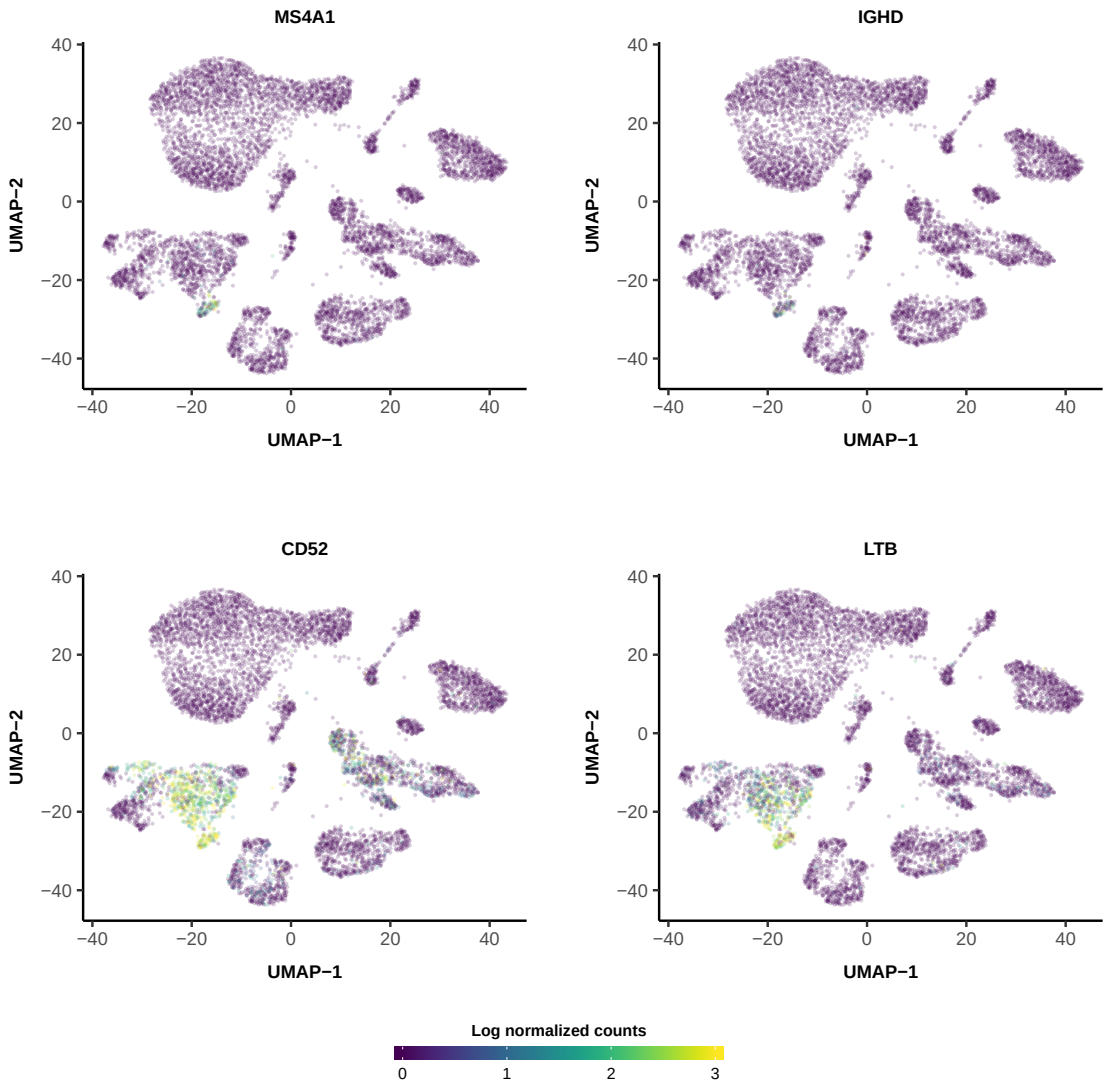
Supplemental Note Figure 1. UMAP representation of the McParland et al. dataset, coloured by (**a**) original cell type assignment [1], (**b**) CellAssign cell type assignment, (**c**) CellAssign cell type assignment using revised marker list, (**d**) SCINA cell type assignment, (**e**) SCINA cell type assignment using revised marker list.



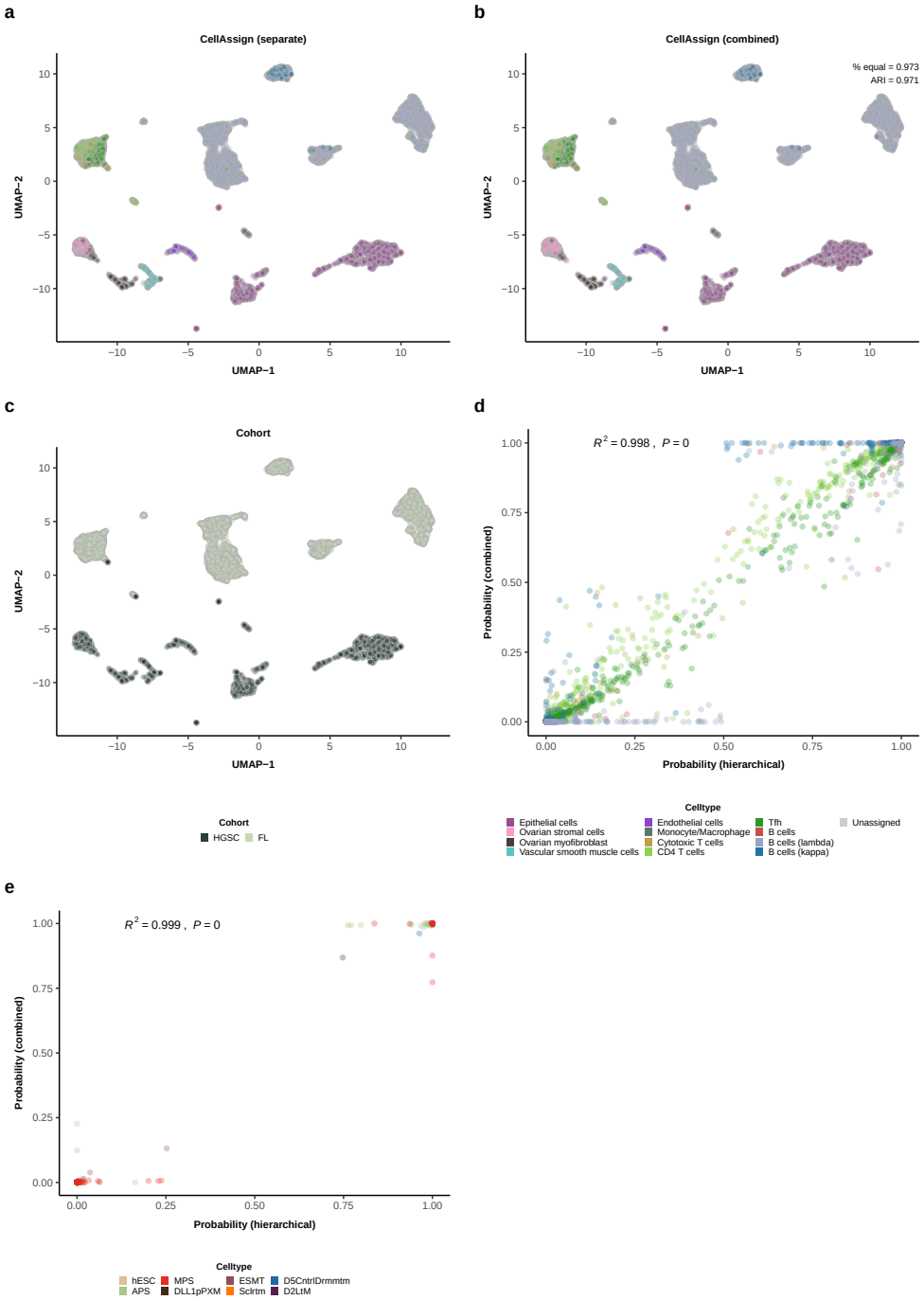
Supplemental Note Figure 2. Robustness of CellAssign when the expected number of cell types is over-specified compared to what exists in the data. UMAP representations of the MacParland dataset subsetted to contain only hepatocytes, cholangiocytes, and mature B cells, coloured by (**a**) Original cell type assignment from MacParland et al., (**b**) CellAssign cell type assignment when supplying markers for only the 3 cell types present, (**c**) CellAssign cell type assignment when supplying markers for the 3 cell types present and 1 cell type not present (NK cells), (**d**) CellAssign cell type assignment when supplying markers for all original cell types, (**e-l**) equivalent assignments using SCINA with sensitivity cutoffs of 1 (default), 0.8, 0.5, 0.2, and 0.1 (with and without batch correction [4]).



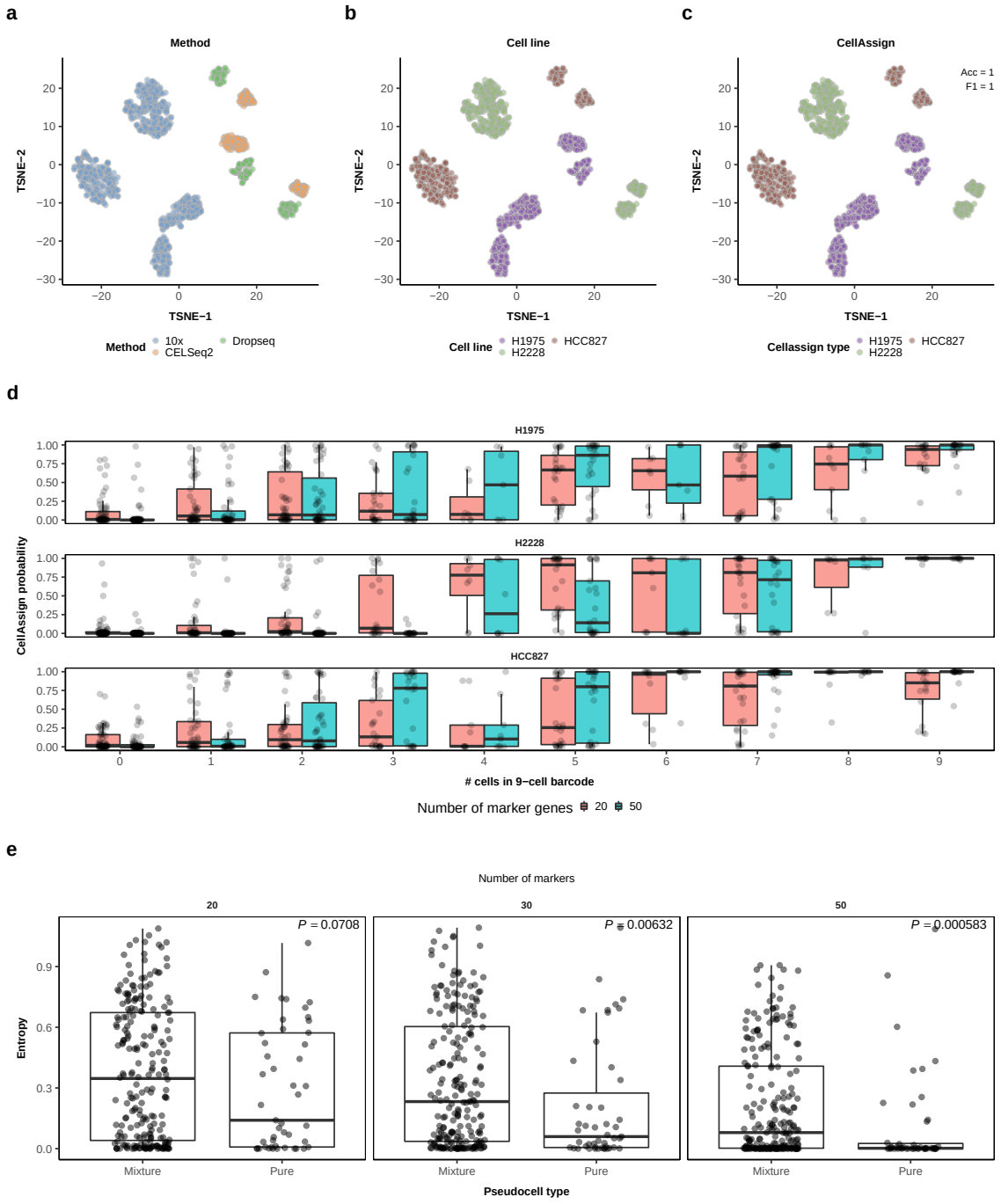
Supplemental Note Figure 3. Robustness of CellAssign when the expected number of cell types is under-specified compared to what exists in the data. UMAP representation of the MacParland et al. data coloured by (**a**) original cell type assignments from MacParland et al., (**b**) CellAssign cell type assignments on the data subsetted to 4 cell types when CellAssign has only 3 as known input, (**c**) CellAssign cell type assignments on all the cell types when CellAssign has only 3 as known input, (**d-e**) equivalent tests for SCINA. (**f**) SCINA, as in (**e**) but run with the revised marker list (analogous to the CellAssign run in **Figure** 2E).



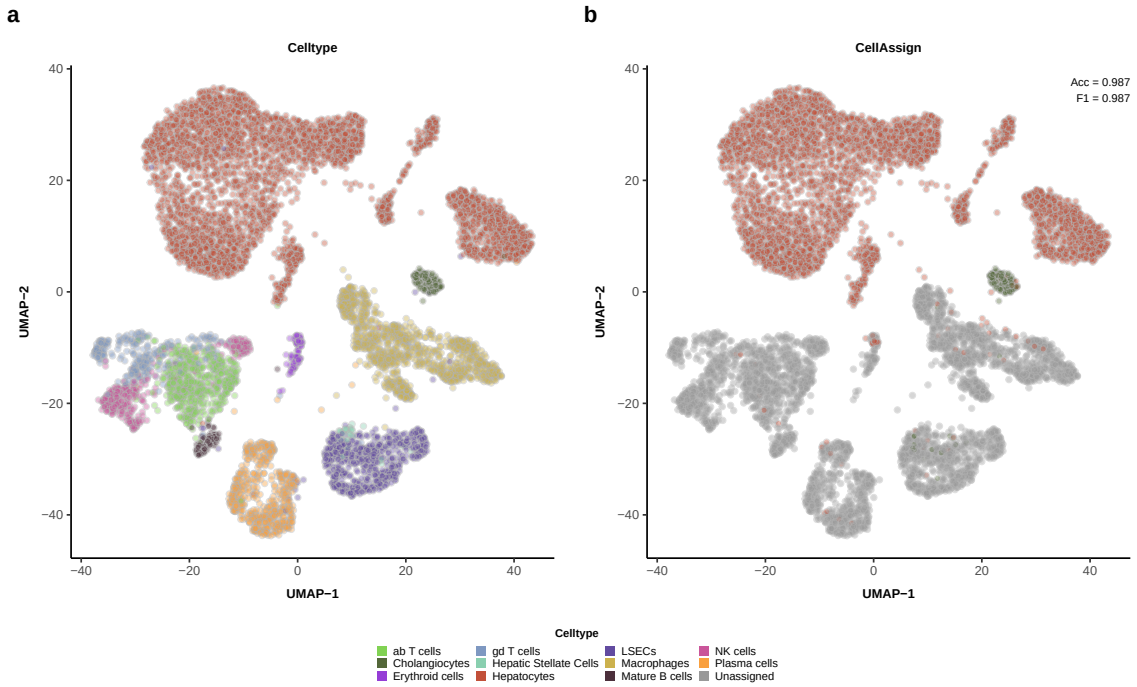
Supplemental Note Figure 4. Expression (log normalized counts) of B cell “marker genes" from McParland et al. [1]. Expression values were winsorized between 0 and 3.



Supplemental Note Figure 5. Consistency of CellAssign cell type assignment to hierarchical specification of the cell types. UMAP representations of both the HGSC and FL datasets coloured by (**a**) CellAssign assignment on each dataset separately, (**b**) CellAssign assignment on the combined dataset with concatenated marker gene matrix, (**c**) dataset of origin. (**d**) Concordance between cell assignment probabilities for follicular lymphoma data ($n = 9156$ single cells) between assignments where the cell types are hierarchically specified (x-axis) or specified all at once (y-axis). (**e**) Concordance between cell assignment probabilities for Koh et al. data ($n = 369$ single cells) between assignments where the cell types are hierarchically specified (x-axis) or specified all at once (y-axis). P-values computed with linear mixed models with random effect intercepts for each cell type.



Supplemental Note Figure 6. CellAssign applied to the CellBench data. (**a**-**c**) T-SNE representation of the data coloured by measurement modality, cell line (ground truth), and CellAssign assigned cell type, respectively. (**d**) Probability of assignment for each $n = 503$ pseudo-cells as the purity of the pseudo-cell increases for each cell line and a variable number of marker genes between cell types. (**e**) The Shannon entropy of cell assignment for $n = 503$ pseudo-cells is lower in pure cell types than mixtures (significance testing performed using FDR-adjusted two-sided Wilcoxon ranked-sum test). In all panels, lower and upper hinges denote the 1st and 3rd quartiles on boxplots, with whiskers extending to the largest value less than 1.5 $\times$ the inter-quartile range.



Supplemental Note Figure 7. CellAssign applied to the MacParland et al. data using markers derived from PanglaoDB for hepatocytes and cholangiocytes. UMAP representation of the data coloured by (**a**) original paper cell assignment and (**b**) by CellAssign cell type assignment.

References

1. MacParland, S. A. *et al.* Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nature communications* **9,** 4383 (2018).
2. Zhang, Z. *et al.* SCINA: Semi-Supervised Analysis of Single Cells in silico (2019).
3. Pont, F. *et al.* The gene expression profile of phosphoantigen-specific human $\gamma\delta$ T lymphocytes is a blend of $\alpha\beta$ T-cell and NK-cell signatures. *European journal of immunology* **42,** 228–240 (2012).
4. Hie, B. L., Bryson, B. & Berger, B. Panoramic stitching of heterogeneous single-cell transcriptomic data. *BioRxiv,* 371179 (2018).
5. Rodig, S. J. *et al.* Heterogeneous CD52 expression among hematologic neoplasms: implications for the use of alemtuzumab (CAMPATH-1H). *Clinical Cancer Research* **12,** 7174–7179 (2006).
6. Browning, J. L. *et al.* Lymphotoxin β, a novel member of the TNF family that forms a heteromeric complex with lymphotoxin on the cell surface. *Cell* **72,** 847–856 (1993).
7. Koh, P. W. *et al.* An atlas of transcriptional, chromatin accessibility, and surface marker changes in human mesoderm development. *Scientific Data* **3,** 160109. ISSN: 2052-4463 (Dec. 2016).
8. Tian, L. *et al.* scRNA-seq mixology: towards better benchmarking of single cell RNA-seq protocols and analysis methods. *BioRxiv,* 433102 (2018).
9. Holik, A. Z. *et al.* RNA-seq mixology: designing realistic control experiments to compare protocols and analysis methods. *Nucleic acids research* **45,** e30–e30 (2016).
10. Zhang, X. *et al.* CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic acids research* **47,** D721–D728 (2018).