nature methods

ARTICLES

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

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]*

Single-cell RNA sequencing has enabled the decomposition of complex tissues into functionally distinct cell types. Often, investigators wish to assign cells to cell types through unsupervised clustering followed by manual annotation or via 'mapping' to existing data. However, manual interpretation scales poorly to large datasets, mapping approaches require purified or pre-annotated data and both are prone to batch effects. To overcome these issues, we present CellAssign, a probabilistic model that leverages prior knowledge of cell-type marker genes to annotate single-cell RNA sequencing data into predefined or de novo cell types. CellAssign automates the process of assigning cells in a highly scalable manner across large datasets while controlling for batch and sample effects. We demonstrate the advantages of CellAssign through extensive simulations and analysis of tumor microenvironment composition in high-grade serous ovarian cancer and follicular lymphoma.

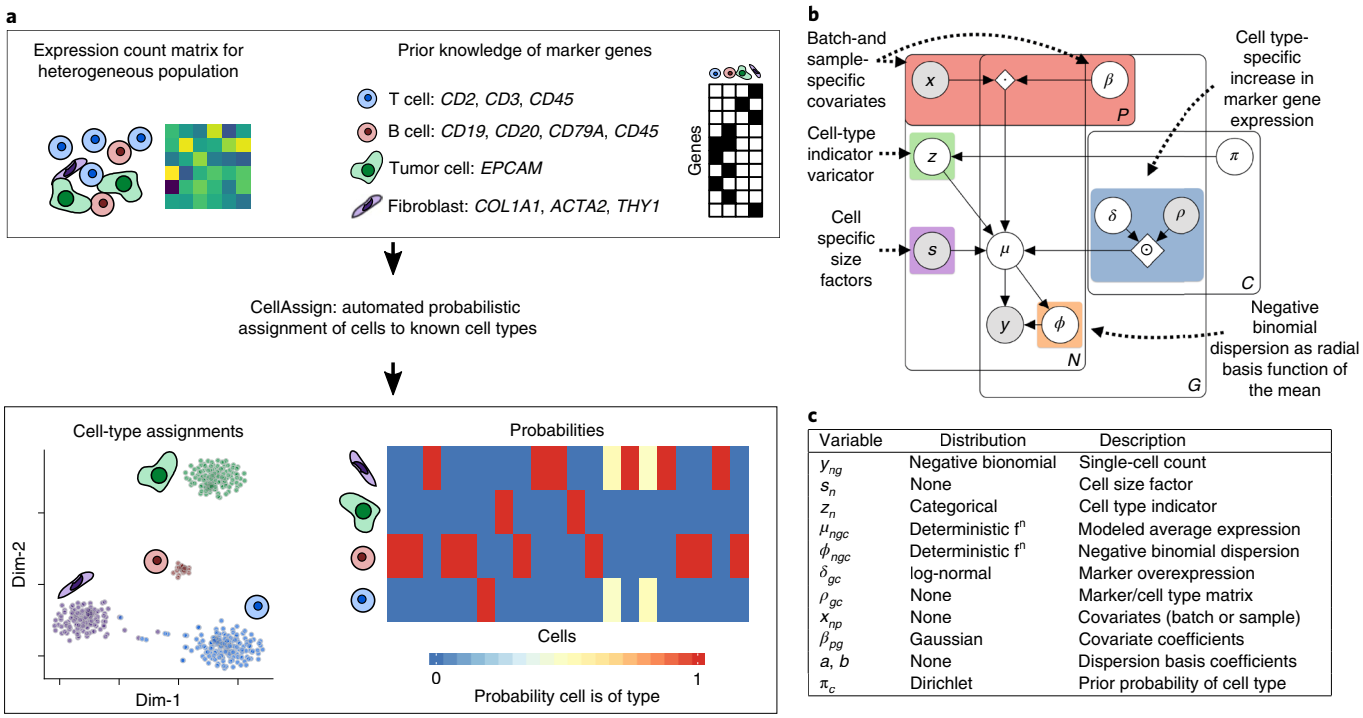
Gene expression observed at the single-cell resolution in human tissues enables the study of cell-type composition and dynamics of mixed cell populations in a variety of biological contexts. Cell types inferred from single-cell RNA sequencing (scRNA-seq) data are typically annotated in a two-step process, whereby cells are clustered using unsupervised algorithms and clusters are then assigned to cell types according to aggregated cluster-level expression profiles[1]. A myriad of methods for unsupervised clustering of scRNA-seq have been proposed, such as SC3 (ref. [2]), Seurat[3], pcaReduce[4] and PhenoGraph[5], along with studies evaluating their performance[6,7]. However, clustering of low-dimensional projections may limit biological interpretability due to low-dimensional projections not encoding variation present in high-dimensional inputs[8] and overclustering of populations that are not sufficiently variable.

In the context of robust clustering that recapitulates biological cell states or classes, few principled methods for annotating clusters of cells into known cell types exist. Typical workflows employ differential expression analysis between clusters to manually classify cells according to differentially expressed markers, aided by recent databases linking cell types to canonical gene-based markers[9]. In situations where investigators wish to identify and quantify specific cell types of interest across multiple samples or replicates, such workflows can be cumbersome and differences in clustering strategies can affect downstream interpretation[6]. Alternatively, cell types may be assigned by gating on marker gene expression, but this strategy is difficult to implement in practice since it relies on knowledge of marker gene expression levels; cells that fall outside these gates will not be assigned to any type rather than being probabilistically assigned to the most likely cell type.

Another approach to cell-type annotation is to leverage single-cell transcriptomic data from pre-annotated and purified cell types to establish robust profiles to which new data can be mapped. For example, scmap-cluster[10] calculates the medioid expression profile for each cell type in the known transcriptomic data and then assigns input cells based on maximal correlation to those profiles. However, such approaches require existing purified or pre-annotated scRNA-seq data for all populations of interest. Given the technical effects associated with differences in experimental design and processing, expression profiles for reference populations may not be directly comparable to those for other scRNA-seq experiments[11].

To address the challenges inherent in existing approaches, we developed CellAssign (https://github.com/irrationone/cellassign), a statistical framework that assigns cells to both known and de novo cell types in scRNA-seq data. CellAssign automates the process of annotation by computing a probabilistic assignment for each cell to a cell type—defined by a set of marker genes—or to an 'unassigned' class. Such panels of markers that uniquely identify cell types may be established through expert knowledge based on the literature, databases such as CellMarker, or derived directly from resources such as PanglaoDB (Supplementary Note 3). CellAssign allows for flexible expression of marker genes, assuming that marker genes are more highly expressed in the cell types they define relative to others. Implemented in Google's TensorFlow framework, CellAssign is

[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



Variable	Distribution	Description
y_{ng}	Negative bionomial	Single-cell count
s_n	None	Cell size factor
z_n	Categorical	Cell type indicator
μ_{ngc}	Deterministic f^n	Modeled average expression
ϕ_{ngc}	Deterministic f^n	Negative binomial dispersion
δ_{gc}	log-normal	Marker overexpression
ρ_{gc}	None	Marker/cell type matrix
x_{np}	None	Covariates (batch or sample)
β_{pg}	Gaussian	Covariate coefficients
a, b	None	Dispersion basis coefficients
π_c	Dirichlet	Prior probability of cell type

Fig. 1 | Overview of CellAssign. a, CellAssign takes raw count data from a heterogeneous scRNA-seq population, along with a set of known marker genes, for various cell types under study. Using CellAssign for inference, each cell is probabilistically assigned to a given cell type without any need for manual annotation or intervention, accounting for any batch- or sample-specific effects. **b**, An overview of the CellAssign probabilistic graphical model. The random variables and data that form the model, along with the distributional assumptions, are shown. **c**, Descriptions of the random variables used in the CellAssign probabilistic model, along with their prior distributions.

highly scalable and capable of annotating thousands of cells in seconds while controlling for inter-batch, patient and site variability.

We evaluated CellAssign across a range of simulations, on ground-truth fluorescence-activated cell sorting (FACS)-purified human embryonic stem cell (hESC) data[12], pre-annotated data and cell line data for multiple scRNA-seq platforms[13]. CellAssign outperforms both clustering and mapping and is robust to errors in marker gene specification. Additionally, we generated two datasets to exemplify the ability of CellAssign to delineate the composition of the tumor microenvironment (TME) across anatomical space and temporal sampling. Overall, CellAssign provides a robust statistical approach through which varying compositions in tissues comprised of mixed cell populations can be quantified and interpreted.

Results

CellAssign: probabilistic and automated cell-type assignment. The CellAssign statistical framework (Fig. 1) models observed gene expression for a heterogeneous cell population as a composite of multiple factors including cell type, library size, and batch. The inputs consist of raw scRNA-seq read counts and a marker gene set for each cell type of interest. Marker genes are assumed to be overexpressed in cell types where they are markers—not necessarily at similar levels—compared to those where they are not. Other experimental and biological covariates such as batch and patient-of-origin are optionally encoded in a standard design matrix. Using this information, CellAssign employs a hierarchical statistical framework to compute the probability that each cell belongs to the modeled cell types, while jointly estimating all model parameters using an expectation–maximization inference algorithm. To prevent misassignment when unknown cell types (unspecified in the marker matrix) are present, CellAssign designates cells that do not belong to any provided cell type as 'unassigned'. Detailed model specification, implementation and runtime performance are described in the Methods.

Performance of CellAssign relative to alternative approaches. We benchmarked CellAssign's performance relative to standard workflows, including unsupervised clustering followed by manual annotation and methods that map cells to existing data from purified populations. Using an adapted version of the splatter model[14] fitted to data for peripheral blood naïve CD8$^+$ and CD4$^+$ T cells, we simulated scRNA-seq data for multiple cell populations (Methods) across a wide range of values for differentially expressed gene fractions (0.05–0.45). We evaluated the performance of unsupervised methods (Seurat[3], SC3 (ref. [2]), PhenoGraph, densityCut[15], dynamicTreeCut[16]), supervised methods (scmap-cluster[10], correlation-based[17]) (Methods), and another marker gene-based approach (SCINA[18]). Half of the simulated cells (n = 1,000 training, n = 1,000 evaluation) were reserved exclusively for training the supervised methods. Marker genes for CellAssign were selected based on simulated log fold change values and mean expression (Methods). For all values of differentially expressed gene fractions, CellAssign performed better than alternative workflows in both accuracy and F_1 score (Fig. 2a and Supplementary Table 1). CellAssign's assignments remained more accurate than the other methods when the analysis was repeated providing other methods with marker genes only (Supplementary Fig. 1a), on data simulated from parameter estimates fitted to B cells and CD8$^+$ T cells (Supplementary Fig. 2a,b and Supplementary Table 1) and when clusters were mapped to existing purified cell types based on maximum correlation (Methods and Supplementary Fig. 3).

We then evaluated the performance of CellAssign on real scRNA-seq data from experimentally sorted populations. For FACS-purified H7 hESCs in various stages of differentiation (8 cell types)[12], we used bulk RNA-seq data from the same cell types to define a set of 84 marker genes for CellAssign based on differential expression results (Supplementary Table 2 and Methods). CellAssign outperformed SCINA and the most competitive unsupervised methods from systematic analysis (SC3, Seurat)[6] according to accuracy and

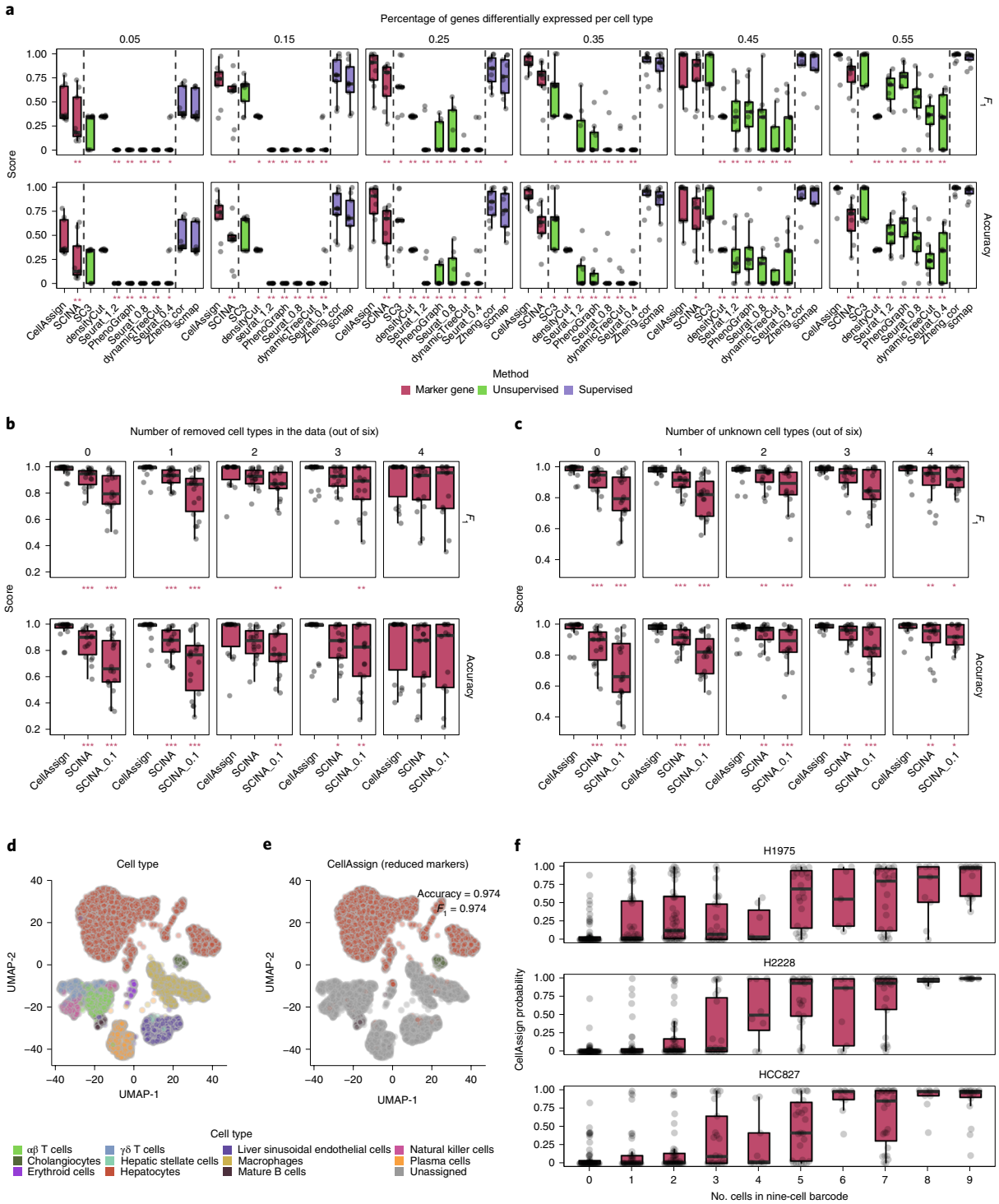


Fig. 2 | Performance of CellAssign on simulated data. a, Accuracy and cell-level F_1 score (Methods) for varying proportions of differentially expressed genes per cell type, with other differential expression parameters set to MAP estimates determined from comparing naïve CD8$^+$ and CD4$^+$ T cells (Methods). CellAssign was provided with a set of marker genes (Methods); all other methods were provided with all genes. The asterisks denote false discovery rate-adjusted *P* values (Wilcoxon signed-rank test) for pairwise comparisons between CellAssign and other methods. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, respectively. The dotted lines separate marker-based, unsupervised and supervised methods. **b**, Accuracy and cell-level F_1 scores for CellAssign, SCINA (default parameters) and SCINA (sensitivity cutoff of 0.1) for simulated data from six cell types, where zero to four cell types were removed from the data (but kept in the marker gene list). **c**, Accuracy and cell-level F_1 score for CellAssign, SCINA (default sensitivity cutoff) and SCINA (sensitivity cutoff of 0.1) for simulated data from six cell types, where zero to four cell types were removed from the marker gene list. Marker genes were inferred without knowledge of the removed cell types. **d**, Cell-type labels for human liver data from MacParland et al.[19]. **e**, CellAssign MAP assignments for human liver data, where marker genes for only hepatocytes, cholangiocytes and mature B cells from MacParland et al.[19] were specified. **f**, CellAssign probabilities for cell line mixture data from Tian et al.[13], where known proportions of three lung adenocarcinoma cell lines (H1975, H2228, HCC827) were mixed in nine-cell combinations. Thirty bulk RNA-derived marker genes for each cell line were used (Supplementary Note 2.7). In the boxplots, the lower and upper hinges denote the first and third quartiles, with the whiskers extending to the largest value less than 1.5× the interquartile range.

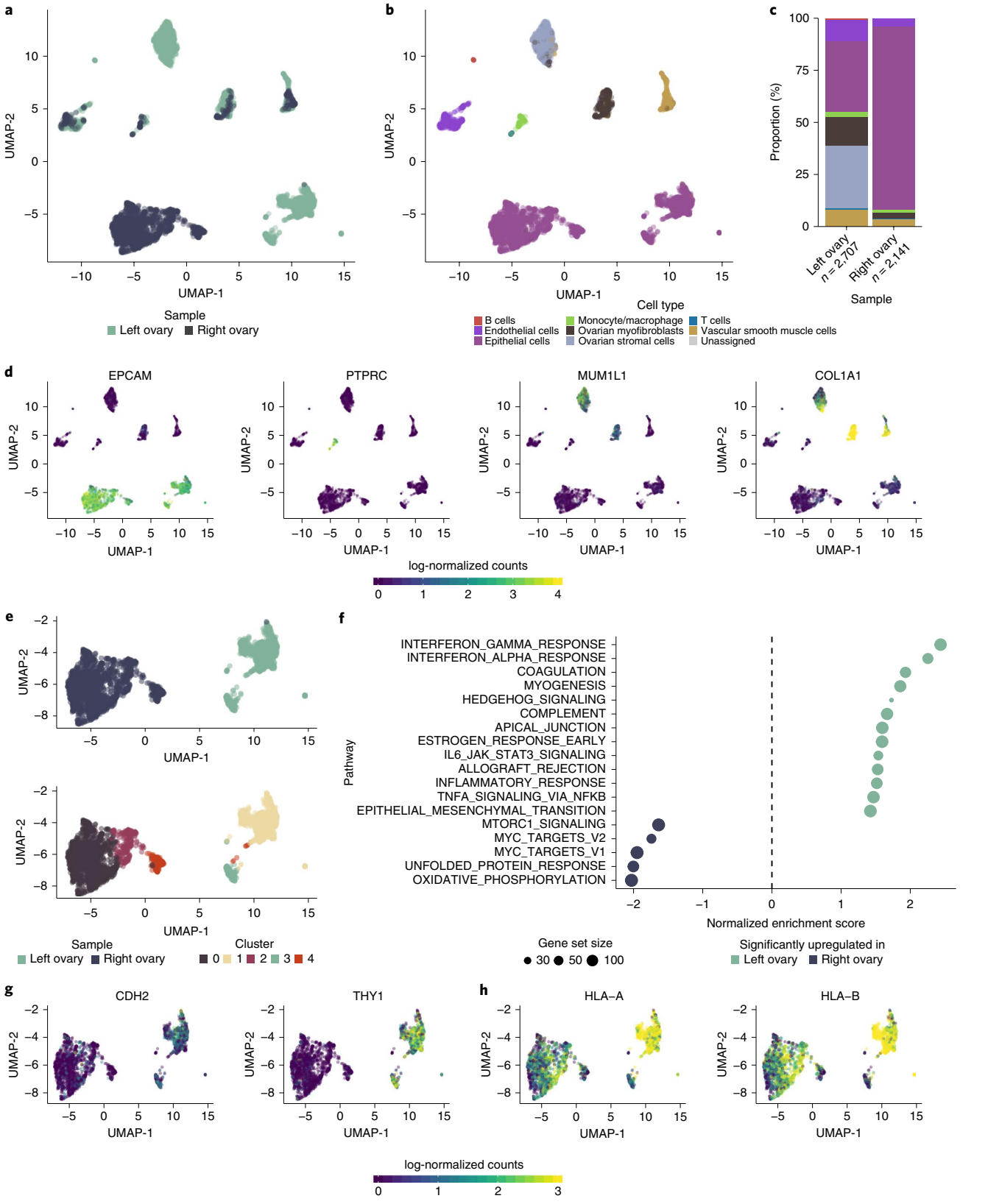


Fig. 3 | CellAssign infers the composition of the HGSC microenvironment. a, UMAP plot of HGSC single-cell expression data, labeled by sample. **b**, UMAP plot of HGSC single-cell expression data, labeled by maximum probability assignments from CellAssign. **c**, Proportions of CellAssign cell types in each sample, with total cell counts indicated. **d**, Expression (log-normalized counts) of *EPCAM* (for epithelial cells), *CD45/PTPRC* (for hematopoietic cells), *MUM1L1* (for ovary-derived cells) and *COL1A1* (for collagen-producing fibroblasts and smooth muscle cells). Expression values were winsorized between 0 and 4. **e**, Hallmark pathway enrichment results for left versus right ovary epithelial cells (Methods). **f**, Unsupervised clustering of epithelial cells (Methods). **g**, Expression (log-normalized counts) of epithelial–mesenchymal transition associated markers, N-cadherin (*CDH2*) and *CD90/THY1* in epithelial cells. **h**, Expression (log-normalized counts) of select HLA class I genes in epithelial cells.

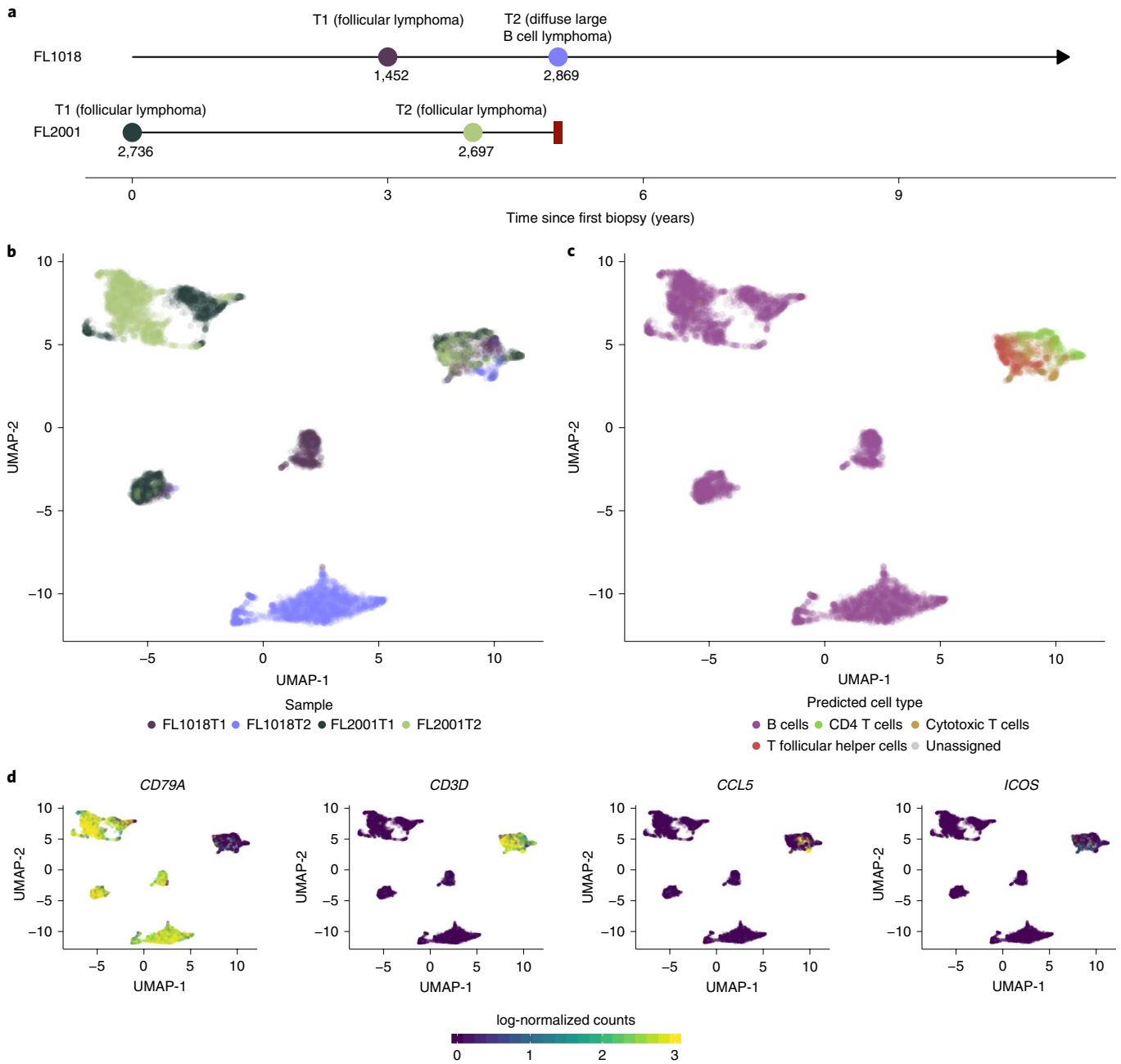


Fig. 4 | CellAssign infers the composition of the follicular lymphoma microenvironment. a, Sample collection times for FL1018 (transformed follicular lymphoma) and FL2001 (progressed follicular lymphoma). FL1018 is alive while FL2001 was lost to follow-up (indicated by the red rectangle). The number of cells collected for each sample is indicated. **b**, UMAP plot of follicular lymphoma single-cell expression data, labeled by sample. **c**, UMAP plot of follicular lymphoma single-cell expression data, labeled by maximum probability assignments from CellAssign. **d**, Expression (log-normalized counts) of select marker genes *CD79A* (for B cells), *CD3D* (for T cells), *CCL5* (for *CD8*$^+$ T cells) and *ICOS* (for T follicular helper cells). Expression values were winsorized between 0 and 3.

cell type-level F_1 score (Supplementary Fig. 4a–e,g and Methods), with similar results obtained using only marker gene expression data (Supplementary Fig. 4f,h; CellAssign $F_1 = 0.943$, accuracy = 0.944; best F_1 of other methods = 0.841, accuracy = 0.93). As an example of CellAssign's ability to discriminate highly related cell types, anterior primitive streak and mid primitive streak cells were accurately classified (83 out of 84 correct), while no other method could reliably do so, assigning anterior primitive streak and mid primitive streak cells to the same cluster (Supplementary Fig. 4).

We next tested the robustness of CellAssign across a range of misspecified inputs that reflect real-world scenarios. We found CellAssign was robust to erroneous specification of the specified marker genes. High assignment accuracy was maintained in scenarios where even 30% of marker gene entries were incorrect (Supplementary Fig. 1c,d, Supplementary Fig. 2d and Supplementary Note 2.1). We also tested the ability of CellAssign to accurately assign cell types when too many or too few cell types are specified compared to those that actually exist in the data. On both simulated data (Methods) and a recent real scRNA-seq dataset of the human liver[19], CellAssign maintained high accuracy of assignment in these situations, with superior performance to SCINA when too many cell types were specified (CellAssign $F_1 = 0.985$, accuracy = 98.5%;

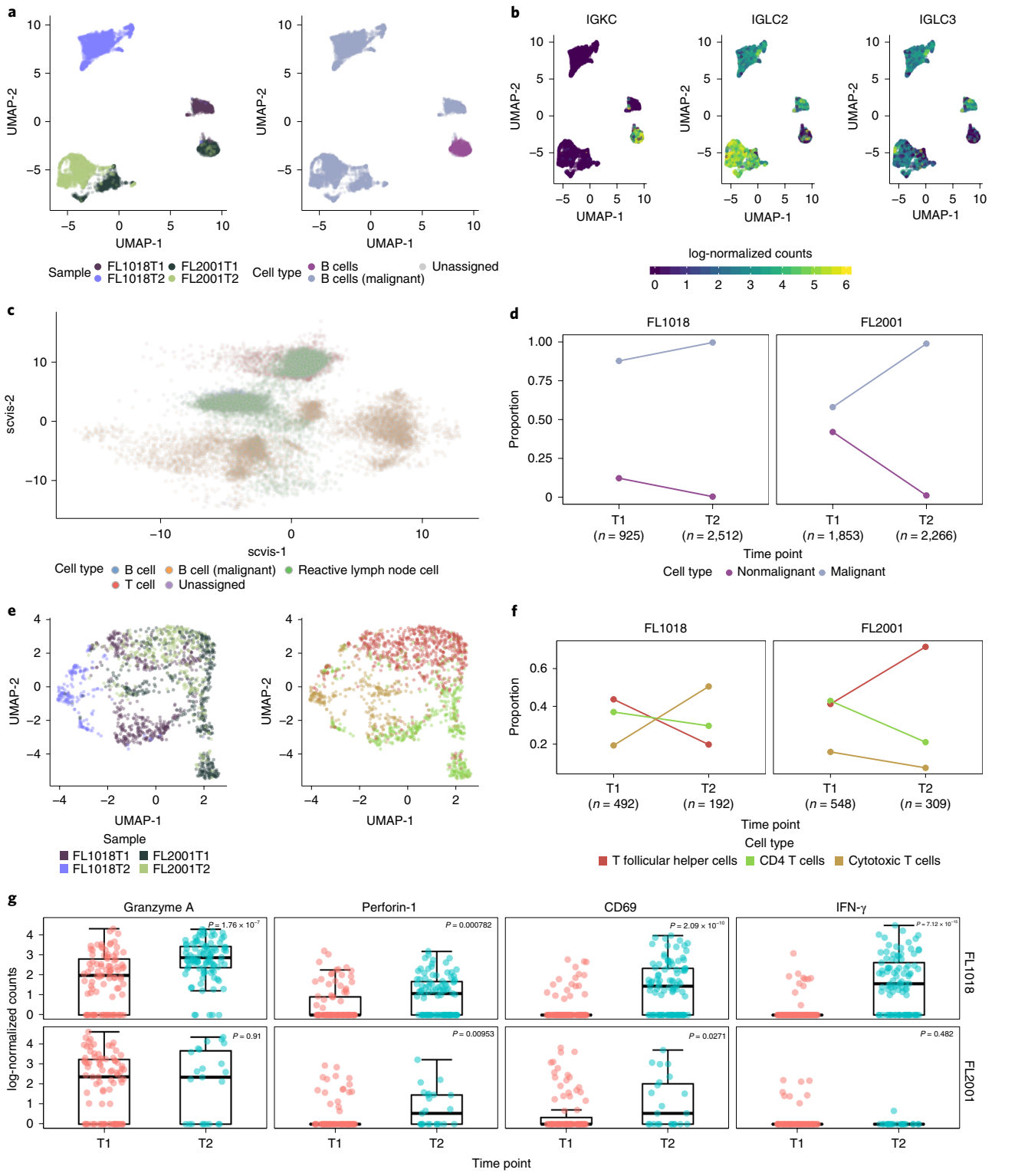


Fig. 5 | Temporal changes in nonmalignant cells in the follicular lymphoma microenvironment. a, Left: UMAP plot of CellAssign-inferred B cells, labeled by sample. Right: UMAP plot of CellAssign-inferred B cells, labeled by putative malignant/nonmalignant status. **b**, Expression (log-normalized counts) of κ (*IGKC*) and λ (*IGLC2* and *IGLC3*) light chain constant region genes. Expression values were winsorized between 0 and 6. **c**, scvis plot of follicular lymphoma data and scRNA-seq data of lymphocytes from reactive lymph nodes from healthy patients. The follicular lymphoma data were used to train the variational autoencoder and produce the two-dimensional embedding. Indicated cell types are B cell (nonmalignant B cell from follicular lymphoma), B cell (malignant; malignant B cell from follicular lymphoma), T cell (T cell from follicular lymphoma), reactive lymph node cell. **d**, Relative proportion of B cell subpopulations over time, with total B cell counts indicated. **e**, UMAP plots of follicular lymphoma T cells, labeled by sample and CellAssign-inferred cell type. **f**, Relative proportion of T cell subpopulations over time, with total T cell counts indicated. **g**, Normalized expression of *CD8*$^+$ T cell activation markers over time. *P* values computed with a two-sided Wilcoxon rank-sum test and adjusted with the Benjamini–Hochberg method. $n = 95$, 96, 90 and 23 single cells identified as *CD8*$^+$ T cells in FL1018T1, FL1018T2, FL2001T1 and FL2001T2, respectively. In the boxplots, the lower and upper hinges denote the first and third quartiles, with the whiskers extending to the largest value less than 1.5× the interquartile range.

SCINA $F_1 = 0.910$, accuracy = 91.0%; Fig. 2b–e and Supplementary Notes 2.2 and 2.3). We next tested the ability of CellAssign to resolve cell types when cells had few distinguishing marker genes. On the same real dataset of human liver cells, we found that even with as few as three specific marker genes at modest expression levels, mature B cells could be differentiated from biologically similar cell types present in the data (Supplementary Note 2.4). Furthermore, when analyzing cell types related through hierarchical differentiation, assigned cell types were consistent regardless of whether CellAssign was run on all cell types upfront or in a nested manner on each level of the hierarchy (Supplementary Note 2.6). Finally, using a recent study of mixed 'pseudo-cells'[13], we demonstrated that CellAssign is robust across scRNA-seq platforms (10x Chromium, CEL-Seq2, Drop-Seq; all accuracy ≥99.9%) and that the assignment probabilities from CellAssign correspond to cell-type purity (Fig. 2f and Supplementary Note 2.7 and 2.8).

Delineating the TME composition of spatially sampled high-grade serous carcinoma (HGSC). We next exemplified CellAssign by decomposing cancer tissues from patients into constituent microenvironmental components and profiled variation across anatomical space and between malignant clones. We generated scRNA-seq data for 5,233 cells from 2 spatial sites from an untreated HGSC patient at the time of primary debulking surgery. Dimensionality reduction with uniform manifold approximation and projection (UMAP[20]) revealed four major site-specific populations and four mixed populations with representation from both samples (Fig. 3a). Using a panel of literature-derived marker genes (Supplementary Table 2 and Methods), CellAssign identified eight major epithelial, stromal and immune cell types (Fig. 3b,c), which were consistent with well-known marker gene expression (Fig. 3d, Supplementary Fig. 5a and Methods). Unlike other nonepithelial cell types, ovarian stromal cells were largely restricted to the left ovary. For cell types such as ovarian stromal cells, no scRNA-seq data from purified populations were available, demonstrating that CellAssign can annotate TME cell types for which marker genes have been orthogonally derived in the literature but where scRNA-seq data for purified populations is unavailable. Hematopoietic cells (B cells, T cells and myeloid cells) were rare in both samples (left ovary: 3.9%; right ovary: 1.5%; Fig. 3c) and dominated by myeloid populations (67 and 87.5% of hematopoietic cells in the left and right ovary, respectively). While CellAssign resolved hematopoietic cell types in a manner consistent with the expression patterns of canonical marker genes, most unsupervised approaches did not resolve some of these cell types, such as B cells, from other hematopoietic or nonhematopoietic cell types (Supplementary Fig. 6). Thus, for TME decomposition and profiling, subtle differences between constituent cell types may be better distinguished by CellAssign over standard approaches[21].

Next, we characterized variation within the epithelial cells identified by CellAssign, all of which were determined to be malignant based on ubiquitous expression of epithelial ovarian cancer markers[22,23] (Supplementary Fig. 5b). Within epithelial cells we identified 5 clusters using Seurat (Fig. 3e) with three (0, 2, 4) derived from the right ovary and two (1, 3) from the left ovary. Differential expression between clusters revealed significant upregulation of genes associated with epithelial–mesenchymal transition in the left ovary (normalized enrichment score = 1.42, $Q = 0.039$), including N-cadherin (*CDH2*) and *CD90/THY1* (Fig. 3e–g), and downregulation of E-cadherin/*CDH1* (log fold change = −0.32, $Q = 1.1 \times 10^{-19}$). Immune-associated pathways were also significantly upregulated, primarily due to cluster 1, one of the two clusters from the left ovary (Fig. 3e,f,h, Supplementary Figs. 7a and 8a,b and Methods). Human leukocyte antigen (HLA) class I genes were among the most differentially expressed genes associated with these pathways (Supplementary Fig. 8b). While HLA expression in cluster 1 was comparable to levels in stromal cells and myofibroblasts, expression levels in other clusters were the lowest across all cell types (Supplementary Fig. 7b), suggestive of subclonal HLA downregulation. Examining cluster-specific gene expression among epithelial cells in the right ovary, hypoxia response was significantly upregulated in cluster 2 relative to the other right ovary clusters (all normalized enrichment scores > 2.05, $Q < 0.0012$; Supplementary Fig. 8c–e). Accordingly, apoptosis and glycolysis pathways were also upregulated while cell cycle-and oxidative phosphorylation-associated pathways were downregulated, consistent with hypoxia-induced cell cycle arrest and metabolic dependence on glycolysis (Supplementary Fig. 8c,d). The profiling of multisite HGSC samples demonstrates how CellAssign can be leveraged within analytical workflows, superseding standard clustering approaches to decompose the TME without compromising the ability to characterize variation within major cell types.

Temporal immune microenvironment dynamics accompanying follicular lymphoma progression and transformation. We next applied CellAssign to delineate temporal microenvironmental changes in follicular lymphoma through scRNA-seq of 9,754 cells from temporally collected lymph node biopsies of two follicular lymphoma patients at two time points each. Histopathological transformation to diffuse large B cell lymphoma occurred in one patient (FL1018), while progression occurred in the other (FL2001) 2 years after rituximab treatment (Fig. 4a). We first computed a UMAP representation, yielding three major patient-specific and two mixed populations comprised of cells from both patients (Fig. 4b). Leveraging literature-derived marker gene information (Supplementary Table 2), we applied CellAssign to identify four major T and B cell types (Fig. 4c,d, Supplementary Fig. 9 and Methods). In comparison, most unsupervised approaches were unable to cleanly resolve T cell subpopulations in the microenvironment (Supplementary Fig. 10). Hypothesizing that the mixed B cell population probably contained nonmalignant B cells (Fig. 5a), we examined immunoglobulin light chain constant domain expression using CellAssign (Fig. 5b) to confirm heterogeneous light chain expression (κ/immunoglobulin kappa constant (*IGKC*) or λ/immunoglobulin lambda constant (*IGLC*)) in the polyclonal nonmalignant B cell population and homogeneous light chain restriction in the clonally identical malignant B cell population[24]. The three patient-specific B cell populations were largely *IGLC*$^+$, consistent with malignant expansion of lambda chain-expressing cells. Applying CellAssign to the mixed population (Supplementary Table 2) showed that 576 out of 907 cells (63.5%) were *IGKC*$^+$ (FL1018: 76 out of 118 (64.4%); FL2001: 500 out of 789 (63.4%)), consistent with the expected polyclonal 60:40 ratio in normal lymphoid organs[25] (Supplementary Fig. 11). In addition, scRNA-seq data of reactive lymph node B cells from four healthy donors mapped onto the mixed B cell population[26] (Fig. 5c and Supplementary Fig. 12). This population also expressed significantly lower levels of the follicular lymphoma markers BCL2 and BCL6 (refs. [24,27–29]) than the other B cells (all log fold change values < −0.34, $Q < 5.4 \times 10^{-7}$; Supplementary Fig. 13 and Supplementary Table 3). Together these results demonstrate the ability of CellAssign to distinguish malignant from nonmalignant B cells, thereby enhancing cell decomposition capacity and cell-type interpretation for lymphoid cancers.

Next, we investigated the temporal dynamics of these cell types in the two patients. The relative proportion of nonmalignant B cells decreased dramatically over time in both cases (FL1018: 12.4–0.8%; FL2001: 42.5–1.4%; Fig. 5d), consistent with clonal expansion of malignant cells during disease progression. Among T cells, the relative proportions of each cell type were comparable between patients in the diagnostic samples (Fig. 5e,f). In FL1018, these compositional changes were accompanied by significant upregulation of immune-associated pathways such as cytokine signaling[30] and T cell activation

and effector molecules among cytotoxic T cells, T follicular helper cells and *CD4*$^+$ T cells after transformation (*CD69* in all T cells, interferon-γ (*IFNG*), granzyme A and perforin-1 in cytotoxic T cells[31]; Supplementary Fig. 13, Fig. 5g and Supplementary Table 3). Together, these results illustrate how CellAssign can be applied to study compositional and phenotypic changes in the tumor microenvironment at the level of individual cell types.

Discussion

CellAssign is intended for scenarios where well-understood marker genes exist, meaning poorly characterized cell types (or unknown cell types or cell states) may be invisible. Furthermore, we make no a priori distinction between medium or high expression of the same marker in two different cell types, although these could be incorporated by extending the model. Nevertheless, we suggest that a large proportion of clinical applications profiling complex tissues start with hypotheses relating the composition of known cell types to disease states. As such, CellAssign fills an important role in the scRNA-seq analysis toolbox, providing interpretable output from biologically motivated prior knowledge. It intrinsically mitigates issues common to existing unsupervised clustering approaches, including batch effects on clustering and the need of post hoc or ad hoc interpretation of clusters in terms of known cell types[8].

The volume of scRNA-seq data will increase over time in that both the number of cell types profiled will increase—thereby expanding databases of known marker genes—and it will become more widely available in research and clinical settings[32]. Therefore, CellAssign is poised to provide scalable, systematic and automated assignment of cells based on known parameters of interest, such as cell type, clone-specific markers or genes associated with drug response. By appropriately modifying the observation model, CellAssign can be extended to annotate cell types in data generated by other single-cell measurement technologies such as mass cytometry. We anticipate the CellAssign approach will help unlock the potential for large-scale population-wide studies of cell composition of human disease and other complex tissues through encoding biological prior knowledge in a robust probabilistic framework.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0529-1.

Received: 15 January 2019; Accepted: 16 July 2019;
Published online: 9 September 2019

References

1. Schaum, N. et al. Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*. *Nature* **562**, 367–372 (2018).
2. Kiselev, V. Y. et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483–486 (2017).
3. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
4. Žurauskienė, J. & Yau, C. pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics* **17**, 140 (2016).
5. Levine, J. H. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
6. Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res.* **7**, 1141 (2018).
7. Freytag, S., Tian, L., Lönnstedt, I., Ng, M. & Bahlo, M. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Res.* **7**, 1297 (2018).
8. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data.*Nat. Rev. Genet.* **20**, 273–282 (2019).
9. Zhang, X. et al. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* **47**, D721–D728 (2019).
10. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).
11. Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**, 562–578 (2018).
12. Koh, P. W. et al. An atlas of transcriptional, chromatin accessibility, and surface marker changes in human mesoderm development. *Sci. Data* **3**, 160109 (2016).
13. Tian, L. et al. scRNA-seq mixology: towards better benchmarking of single cell RNA-seq protocols and analysis methods. Preprint at *bioRxiv* https://doi.org/10.1101/433102 (2018).
14. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* **18**, 174 (2017).
15. Ding, J., Shah, S. & Condon, A. densityCut: an efficient and versatile topological approach for automatic clustering of biological data. *Bioinformatics* **32**, 2567–2576 (2016).
16. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008).
17. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
18. Zhang, Z. et al. SCINA: A semi-supervised subtyping algorithm of single cells and bulk samples. *Genes* **10**, 531 (2019).
19. MacParland, S. A. et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat. Commun.* **9**, 4383 (2018).
20. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/pdf/1802.03426.pdf (2018).
21. Zhang, A. W. et al. Interfaces of malignant and immunologic clonal dynamics in ovarian cancer. *Cell* **173**, 1755–1769.e22 (2018).
22. Kristiansen, G. et al. CD24 is expressed in ovarian cancer and is a new independent prognostic marker of patient survival. *Am. J. Pathol.* **161**, 1215–1221 (2002).
23. Hylander, B. et al. Expression of Wilms tumor gene (*WT1*) in epithelial ovarian cancer. *Gynecol. Oncol.* **101**, 12–17 (2006).
24. Andor, N. et al. Single-cell RNA-Seq of lymphoma cancers reveals malignant B-cell types and coexpression of T-cell immune checkpoints. *Blood* **133**, 1119–1129 (2019).
25. Jefferis, R. & Lefranc, M.-P. Human immunoglobulin allotypes: possible implications for immunogenicity. *MAbs* **1**, 332–338 (2009).
26. Ding, J., Condon, A. & Shah, S. P. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* **9**, 2002 (2018).
27. Hermine, O. et al. Prognostic significance of bcl-2 protein expression in aggressive non-Hodgkin's lymphoma. Groupe d'Etude des Lymphomes de l'Adulte (GELA). *Blood* **87**, 265–272 (1996).
28. Gu, K. et al. t(14;18)-negative follicular lymphomas are associated with a high frequency of *BCL6* rearrangement at the alternative breakpoint region. *Mod. Pathol.* **22**, 1251–1257 (2009).
29. Hatzi, K. & Melnick, A. Breaking bad in the germinal center: how deregulation of BCL6 contributes to lymphomagenesis. *Trends Mol. Med.* **20**, 343–352 (2014).
30. Fabregat, A. et al. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).
31. Freeman, B. E., Hammarlund, E., Raué, H. P. & Slifka, M. K. Regulation of innate CD8$^+$ T-cell activation mediated by cytokines. *Proc. Natl Acad. Sci. USA* **109**, 9971–9976 (2012).
32. Hwang, B., Lee, J. H., Bang, D. & Single-cell, R. N. A. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 96 (2018).

Acknowledgements

We thank V. Svensson for his feedback on this manuscript. We also thank W. W. Wasserman, B. H. Nelson, P. T. Hamilton and A. Miranda for helpful discussions. A.W.Z. is funded by scholarships from the Canadian Institutes of Health Research (CIHR) (Vanier Canada Graduate Scholarship, Michael Smith Foreign Study Supplement) and a BC Children's Hospital (UBC) MD/PhD studentship. K.R.C. is funded by postdoctoral fellowships from the CIHR (Banting) no. 01353-000, the Canadian Statistical Sciences Institute and the UBC Data Science Institute no. 201803. S.P.S. is a Susan G. Komen scholar. We acknowledge the generous funding support provided by the BC Cancer Foundation. In addition, S.P.S. receives operating funds from the CIHR (grant no. FDN-143246), Terry Fox Research Institute (grant nos. 1021 and 1061) and the Canadian Cancer Society (grant no. 705636). This work was supported by Cancer Research UK (grant no. C31893/A25050 to S.A. and S.P.S.). S.P.S. is supported by the Nicholls-Biondi endowed chair and the Cycle for Survival benefitting Memorial Sloan Kettering Cancer Center. C.S. is an Allen Distinguished Investigator supported by the Allen Frontiers Group no. 12829.

Author contributions

A.W.Z., K.R.C. and S.P.S. designed the study. A.W.Z., K.R.C. and S.P.S. wrote the manuscript. A.W.Z., C.O.F., E.A.C., J.L.P.L., A. McPherson, A. Mottok, N.C., L.C.,

M.W., T.A., A.P.W., J.N.M., S.A., C. Steidl, K.R.C. and S.P.S. reviewed the manuscript. A.W.Z., S.A., C. Steidl, K.R.C. and S.P.S. interpreted the data. B.H., D.L., L.C. and C. Sarkozy curated the data. A.W.Z., K.R.C., N.C., M.W., P.W., T.C. and X.W. analyzed the data. A.W.Z., K.R.C. and S.P.S. developed the model. C.O.F., E.A.C. and J.L.P.L. performed the single-cell processing. A. Mottok, J.N.M., C. Steidl and C. Sarkozy performed the case identification. K.R.C., S.P.S., C. Steidl and S.A. supervised the study.

Competing interests

S.P.S. and S.A. are founders, shareholders and consultants of Contextual Genomics.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0529-1.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to K.R.C. or S.P.S.

Peer review information: Lei Tang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Ethical approval. Ethical approval for this study was obtained from the University of British Columbia (UBC) Research Ethics Board (nos. H08-01411,H14-02304 and H18-01090). Informed consent was obtained from all participants in this study.

The CellAssign model. *Model description.* Let Y be a cell-by-gene expression matrix of raw counts for N cells and G genes. Suppose that among those cells we have C total cell types, each defined by high expression of several 'marker' genes. We encode the relationship between cells and marker genes through a binary matrix ρ, where $\rho_{gc}=1$ if gene g is a marker for cell type c and 0 otherwise. To relate cells to cell types, we introduce a categorical indicator vector $\mathbf{z}=\{z_n\}$ that encodes to which of the C cell types each cell belongs:

$$z_n = c \text{ if cell } n \text{ of type } c$$

To assign cells to cell types, we performed statistical inference of the probability that each cell is of a given cell type for which we must compute the quantity $p(z_n = c|Y, \hat{\Theta})$ where $\hat{\Theta}$ are the maximum a posteriori probability (MAP) estimates of the model parameters.

Let s_n be the size factor for cell n and X be a $P\times N$ matrix of P covariates (such as patient-of-origin). Then our model is:

$$\mathbb{E}\left[y_{ng}|z_n = c\right] = \mu_{ngc}$$

where

$$\overbrace{\log \mu_{ngc}}^{\text{Log mean expression}} = \underbrace{\log s_n}_{\text{Cell size factor}} + \overbrace{\delta_{gc}\rho_{gc}}^{\text{Cell type specific}} + \underbrace{\beta_{g0}}_{\text{Base expression}} + \overbrace{\sum_{p=1}^{P}\beta_{gp}x_{pn}}^{\text{Other covariates (incl.batch)}}$$

with the constraint that $\delta_{gc}>0$.

The intuition is that if gene g is a marker for cell type c, then we expect the expression of g to be multiplied by the factor $e^{\delta_{gc}}$, where δ_{gc} is inferred. In this way, we put no restriction that marker genes cannot be expressed in other cell types and that they must be highly expressed in their cell type, only that they exhibit higher expression in the cells of type for which they are a marker. The quantity δ_{gc} corresponds to the average log(fold change) that gene g is overexpressed in cell c, which only occurs for marker genes for cell types since ρ_{gc} must be equal to 1 for this to contribute to the likelihood. In simulations, we found that CellAssign could accurately estimate these parameters (Supplementary Figs. 1b and 2c). By default, we impose a lower bound such that $\delta>\log(2)$, making the interpretation that a marker gene must be overexpressed by a factor of 2 relative to cells for which it is not a marker, but this is left as an option for the user. We also controlled for technical or sample effects through the matrix X.

We specified a hierarchical shrinkage prior $\delta_{gc}\sim \log-\text{normal}(\bar{\delta},\sigma^2)$ over the cell type-specific overexpression parameters δ_{gc}, where the mean and variance parameters of the log-normal $\bar{\delta}$ and σ^2 are initialized to 0 and 1, respectively. We further specified a hierarchical prior on the cell-type assignments $p(z_n = c) = \pi_c$ and $(\pi_1, \ldots, \pi_C) \approx \text{Dirichlet}(\alpha, \ldots, \alpha)$ with π_c initialized to $1/K$ and $\alpha = 10^{-2}$ by default.

The remaining model parameters are initialized as follows:

- β_{gp} is drawn from an $\mathcal{N}(0, 1)$ distribution;
- $\log \delta_{gc}$ is drawn from an $\mathcal{N}(0, 1)$ distribution truncated at $[\log(\delta_{\min}), 2]$;
- a is initialized to 0;
- b is initialized to twice the square difference between successive spline bases.

The likelihood is given by:

$$y_{ng}|z_n = c \approx \mathcal{NB}\left(\mu_{ngc}, \tilde{\phi}_{ngc}\right)$$

where NB is the negative binomial distribution parametrized by a mean μ and a μ-specific dispersion $\tilde{\phi}_{ngc}$. We define $\tilde{\phi}_{ngc}$ as a sum of radial basis functions dependent on the modeled mean μ_{ngc} as proposed by a recent publication[33]:

$$\tilde{\phi}_{ngc} = \sum_{i=1}^{B} a_i \times \exp(-b_i \times (\mu_{ngc} - x_i)^2)$$

where a_i and b_i represent the radial basis function parameters to be inferred, B is the total number of centers of the radial basis function and x_i is center i. The centers are set to be equally spaced apart from 0 to the maximum number of counts max y_{ng}.

Inference. Using expectation–maximization for inference, the latent variables are $\mathbf{z}\equiv\{z_n\}$ while the model parameters to be maximized are $\boldsymbol{\delta}=\{\delta_{gc}\}$, $\boldsymbol{\beta}=\{\beta_{g0}\beta_{gp}\}$, $\mathbf{a}=\{a_i\}$ and $\mathbf{b}=\{b_i\}$. For the E-step we compute

$$\gamma_{nc} := p\left(z_n = c|y_n, \boldsymbol{\delta}^{(t-1)}, \beta^{(t-1)}, a^{(t-1)}, b^{(t-1)}\right) = \frac{\Pi_g \mathcal{NB}\left(\mu_{ngc}, \tilde{\phi}_{ngc}\right)}{\sum_{c\prime}\Pi_{g\prime}\mathcal{NB}\left(\mu_{ng\prime c\prime}, \tilde{\phi}_{ng\prime c\prime}\right)}$$

where $\theta^{(t)}$ is the value of some parameter θ at iteration t. We then form the Q function:

$$\begin{aligned} &Q\left(\delta^t, \beta^t, a^t, b^t|\delta^{(t-1)}, \beta^{(t-1)}, a^{(t-1)}, b^{(t-1)}\right) \\ &= \mathbb{E}_{z|Y,\delta^{(t-1)},\beta^{(t-1)},a^{(t-1)},b^{(t-1)}}\left[\log p\left(Y|\pi, \delta^t, \beta^t, a^t, b^t\right)\right] \\ &= \sum_{n=1}^{N}\sum_{c=1}^{C}\gamma_{nc}\sum_{g=1}^{G}\log \mathcal{NB}\left(y_{ng}|\mu_{ngc}, \tilde{\phi}_{ngc}\right)\end{aligned}$$

During the M-step, we optimize the Q function using the Adam optimizer[34] as implemented in Google's TensorFlow[35]. By default, we used a learning rate of 0.1, allowed a maximum of 10^5 Adam iterations per M-step and considered that the M-step converges when the relative change in the Q function value falls below 10^{-4}. By default we consider the expectation–maximization algorithm converged when the relative change in the marginal log-likelihood falls below 10^{-4}.

Simulation. *Model description and rationale.* Initially, we attempted to simulate multigroup data from the splatter model (v1.5.4). We employed 10x Chromium data for peripheral blood mononuclear cells[17] with cell-type labels derived from Sinha et al.[36] to determine realistic parameter estimates for the differential expression component of the model. To do so, group-specific log fold change values were drawn from a mixture distribution of a central, narrow Gaussian–Laplacian mixture (representing nondifferentially expressed genes) and two flanking, absolute, value-transformed Gaussian components (representing downregulated and upregulated genes). This mixture distribution was fitted to log(fold change) values derived from differential expression analysis.

However, inspection of posterior predictive samples for multiple fits, using labeled scRNA-seq data from Zheng et al.[17] and FACS-purified data from Koh et al.[12] (Supplementary Fig. 14a–d), revealed that this model systematically underestimates extreme log(fold change) values (Supplementary Fig. 14c,g). Thus, to accommodate the heavier tails present in observed data, we augmented the splatter model by replacing the flanking, absolute, value-transformed Gaussian components with bounded Student's t-distributions. Posterior predictive log(fold change) distributions from this modified model better fitted the observed data (Supplementary Fig. 14d,h). Consequently, we used this model to perform the simulation analysis.

Model fitting. The models described earlier were fitted to log(fold change) values derived from real data. Using the labeled 10x Chromium data for 68k peripheral blood mononuclear cells[17], differential expression was performed with the FindMarkers function from the R package scran (v1.11.21)[37]. To generate the corresponding null distributions of log(fold change) values for nondifferentially expressed genes, we split the data for each cell type into equally sized halves ten times, running FindMarkers to compare the resulting halves. A central Gaussian–Laplacian mixture ($\mu=0$) was first fitted to the null log(fold change) values. The distribution of posterior predictive log(fold change) values appeared to be consistent with observed log(fold change) values for this null component (Supplementary Fig. 14d). Following this, the entire mixture distribution was fitted to log(fold change) values for pairs of distinct cell types, using MAP estimates of parameters for the central Gaussian–Laplacian component. Posterior distributions of model parameters were inferred using the no-U-turn sampler in PyMC3, using 4 independent chains, 1,000 tuning iterations and 2,500 additional iterations per chain. Trace plots and the Gelman–Rubin diagnostic were used to assess convergence.

Simulating multigroup data. Expression count matrices were simulated using a modified version of the splatter package (http://www.github.com/Irrationone/splatter, commit edb3578d7a5dceaf354022bf4f4b6051583d8ddd). log(fold change) values were simulated according to our model instead of the splatter model. Other settings were kept identical. We used MAP estimates of $\mu+$, $\mu-$, $\sigma+$, $\sigma-$, $\nu+$ and $\nu-$, determined by fitting our simulation model to (1) log(fold change) values between naïve CD4$^+$ and naïve CD8$^+$ T cells (Supplementary Fig. 14a) and (2) log(fold change) values between B cells and CD8$^+$ T cells (see Model description and rationale section) for the differential expression component. The proportion of downregulated genes out of differentially expressed genes was set to 0.5 (that is, equally probable for a differentially expressed gene to be downregulated versus upregulated). Three groups (cell types) were simulated at equal proportions. Other parameters for splatter were fitted from 10x Chromium data for 4,000 T cells available from 10x Genomics.

To assess the performance of CellAssign relative to other clustering methods across a range of P_d values (proportion of genes differentially expressed between each pair of cell types), P_d was chosen from {0.05,0.15,0.25,0.35,0.45,0.55}. (The true MAP estimate of P_d was 0.0746 for naïve CD4$^+$ versus naïve CD8$^+$ T cells and 0.153 for B versus CD8$^+$ T cells.) The number of simulated cells, n, was set to 2,000; 1,000 were randomly set aside for training (for scmap and correlation-based supervised clustering).

To assess the robustness of CellAssign to misspecification of the marker gene matrix ρ, P_d was set to 0.25 and the number of simulated cells n to 1,500.

Simulations were run nine times with unique random seeds for each combination of parameter settings.

Clustering multigroup data. Count matrices were normalized with scater-normalize (scater v1.11.12) and the top 50 principal components were computed from the top 1,000 most variable genes. For PhenoGraph, Seurat (resolution $\in \{0.4,0.8,1.2\}$), densityCut and dynamicTreeCut, unsupervised clustering was performed on the values of these top 50 principal components. For SC3, the entire normalized SingleCellExperiment object was passed as input instead. For the supervised methods (scmap-cluster[10] and correlation-based[17]), expression data for both training and evaluation sets were provided. For CellAssign, the raw count matrix was provided as input, along with a set of marker genes selected based on simulated log(fold change) and mean expression values. For SCINA, the same marker gene matrix used for CellAssign was provided as input, along with normalized log counts. The parameter rm_overlap was set to 0 to ensure that, like CellAssign, SCINA was using all provided marker genes. Specifically, a gene was defined as a marker gene if it was in the top fifth percentile of differentially expressed genes according to log(fold change) and the top tenth percentile of differentially expressed genes according to mean expression. A maximum of 15 marker genes were selected for each group. In simulations of robustness to marker gene misspecification, a proportion of randomly selected entries in the marker gene matrix ρ were flipped from 0 to 1 (or vice versa). All other parameters were set to the defaults.

Mapping clusters to true groups. For assignments derived from unsupervised clustering, clusters were mapped to simulated groups by first performing differential expression between each cluster and the remaining cells. Following this, we computed the Spearman correlation between these log(fold change) values and the simulated (true) log(fold change) values for each simulated group. Each inferred cluster was mapped to the most highly correlated simulated group based on Spearman's ρ where $\rho > 0$ and $P \leq 0.05$. Clusters that could not be mapped based on these criteria were marked as 'unassigned'.

We also implemented a second method for mapping clusters from unsupervised clustering to ground-truth simulated groups. To do so, we computed the mean gene expression vector for each ground-truth group and inferred clustering using calcAverage from the scater R package. Clusters were then mapped onto groups by taking the maximum of pairwise Spearman correlation coefficients (enforcing a fairly lenient minimum of Spearman's $\rho \geq 0.1$) between mean expression vectors with all ground-truth groups.

Evaluation. Accuracy and cell-level F_1 score were computed to evaluate clustering performance. The cell-level F_1 score considers each cell as an individual classification task with a true cell-type assignment (and potentially multiple incorrect cell-type assignments) for the purposes of calculating precision and recall.

Marker gene overspecification/underspecification analysis on simulated data. To test the robustness of CellAssign to overspecification of marker gene matrices, we simulated 6 groups at equal proportions ($n = 1{,}500$, $P_d = 0.15$), following the methods described in the section on Simulating multigroup data. Following this, marker gene matrices were generated for the simulated cell types (see Clustering multigroup data section); 0–4 cell types were then removed from the data to create scenarios where more cell types are specified in the marker gene matrix than those that actually exist in the data. CellAssign (include_other=TRUE) and SCINA (rm_overlap=0, allow_unknown=1) were then run on the resulting data. SCINA was provided with 10 times the default number of maximum iterations (1,000). These are analogous settings for both tools that consider all marker genes and allow for the inference of novel cell types. Cell-type assignments from both methods were evaluated as described in the Evaluation section. While CellAssign can automatically discount cell types that do not exist in the data, SCINA was run with various values of sensitivity_cutoff, which facilitates removal of those cell types[18].

Similarly, we simulated six groups at equal proportions as described earlier to test robustness to underspecification of the marker gene matrix (to discover new cell types); 0–4 cell types were then removed before marker gene selection (but retained in the data) to ensure that marker genes were being selected with no knowledge of 'nonexistent' cell types. CellAssign (include_other=TRUE) and SCINA (rm_overlap=0, allow_unknown=1) were then run on the resulting data. Simulations were run 18 times with unique random seeds for each combination of parameter settings.

Benchmarking. We generated synthetic datasets for benchmarking from the modified splatter model (see Model description and rationale section) with the Student's t-parameters $\mu = 0.1$, $\sigma = 0.1$, $\nu = 1$ and the proportion of differentially expressed genes per cell type set to 20%. Synthetic datasets of various sizes (number of cells $N \in \{1{,}000,2{,}000,4{,}000,8{,}000,10{,}000,20{,}000,40{,}000,80{,}000\}$ and number of cell types $C \in \{2,4,6,8\}$) with a balanced number of cells per type were generated. Markers for CellAssign were selected from genes in the top 20th percentile in terms of log(fold change) among differentially upregulated genes and the top 10th percentile in terms of expression. CellAssign was run with 2, 4, 6 and 8 markers per cell type, with a maximum minibatch size of 5,000 cells. On simulated data for 80,000 cells from 2 cell types, CellAssign completed in under 2 min, appearing to scale at worst linearly in the number of cell types and marker genes used per cell type (Supplementary Fig. 15). Five separate CellAssign runs were timed for each combination of parameters.

Koh et al. dataset. The Koh et al.[12] dataset consists of scRNA-seq data for 531 hESCs at various stages of differentiation.

Preprocessing and normalization of scRNA-seq data. Preprocessed data was obtained from the R package DuoClustering2018 (v1.0.0) (refs. [6,12]). Cell types with both scRNA-seq data and bulk RNA-seq data were used: hESC (day 0), anterior primitive streak (day 1), mid primitive streak (day 1), DLL1-positive paraxial mesoderm (day 2), early somite (day 3), sclerotome (day 6), central dermomyotome (day 5), lateral mesoderm (day 2). Normalization and dimensionality reduction were performed with scater-normalize, RunPCA, RunTSNE and RunUMAP. The top 500 most variable genes were used to compute the top 50 principal components; the top 50 principal components were used as input for t-distributed stochastic neighbor embedding.

Identification of marker genes from bulk RNA-seq data. The differential expression analysis results for the bulk RNA-seq data for the same cell types was used to compute the relative expression of each gene in each cell type. Briefly, bulk RNA-seq log(fold change) values obtained from the supplementary materials of Koh et al.[12] were used to compute the log-scale relative gene expression levels. Next, we identified gene-specific thresholds for defining the cell types where each gene is a marker. For each gene, relative expression levels across cell types were sorted in ascending order and were denoted as $E_1, \ldots, E_C$, where C is the total number of cell types. The maximum difference between sorted expression levels, $\max_{1 \leq i < C}(E_{i+1} - E_i)$, was then computed. Note the index i for gene g where this difference is maximal i_g. For gene g, cell types where relative expression values $\geq E_{ig+1}$ were considered the cell types with gene g as a marker. Genes with a maximum difference value in the top 20th percentile were used as marker genes.

CellAssign. CellAssign was run on count data using the marker gene matrix defined from the bulk RNA-seq data described earlier. Three random initializations of expectation–maximization were used. Results from the run reaching the highest marginal log-likelihood at convergence were kept.

Unsupervised clustering. Unsupervised clustering was performed on the top 50 principal components with Seurat[3] (resolution $\in \{0.8,1.2\}$; these represent low-to-moderate and high levels within the recommended range) and on the SingleCellExperiment object of raw and normalized counts with SC3 (ref. [2]). We also provided Seurat[3] with only the marker genes used by CellAssign. (SC3 failed to run when provided with this number of genes.) Inferred clusters were mapped to true (FACS-purified) cell types by computing the pairwise Spearman correlation between mean expression vectors for each cluster and each true cell type. Each cluster was treated as the cell type it was most strongly positively associated with by Spearman's ρ.

SCINA. SCINA was run on normalized log counts using the same marker gene matrix used for CellAssign. As stated earlier, SCINA was run with rm_overlap=0 and allow_unknown=1, with 10 times the default number of maximum iterations (1,000).

Evaluation. Accuracy and cell type-level F_1 score were computed to evaluate clustering performance. The cell type-level F_1 score is defined as the arithmetic mean of F_1 scores computed for each cell type separately.

HGSC. *Sample preparation.* Specimens were placed into cold media in the operating room and brought to the clinical laboratory by messenger porter. Following this, each specimen was assigned a unique research identifier and processed as per Vancouver General Hospital/UBC Anatomical Pathology specimen handling procedures. Tissues were dissociated at low temperature[38] using a modified protocol[39]. Briefly, after finely chopping and weighing in a cell culture dish, tissue was transferred into a gentleMACS C tube, and 1 ml of 10 mg ml^{-1} *Bacillus licheniformis* protease (catalog no. NATE-0633; Creative Enzymes) was added to each 25 mg of tissue. The resulting solution was incubated and mechanically disrupted at 6 °C using the MACS Separator (Miltenyi Biotec; programs h_tumour_01, h_tumour_02, h_tumour_03) for 1 h. Following dissociation, cells were assessed for viability using the cell counter (5 µl cells + 5 µl trypan blue) under a microscope.

Samples were then diluted with cold HFN (Hanks' buffered saline with 2% fetal calf serum and 0.1% sodium azide) and washed with trypsin, dispase and DNase while gently pipetting up and down. Cold ammonium chloride was added to bloody samples. Cells were assessed for viability using the cell counter (5 µl cells + 5 µl trypan blue) under a microscope and kept on ice. Cells were spun down and the pellet resuspended in 100 µl of Dead Cell Removal MicroBeads (Miltenyi Biotec) and incubated at room temperature for 15 min. Viable cell enrichment was performed using the positive selection column type MS with a MACS Separator.

Library preparation and sequencing. scRNA-seq libraries were prepared following the 10x Genomics User Guide for 5′ gene expression library construction.

Single-cell libraries were sequenced on an Illumina NextSeq 500 (75 base pair paired-end reads) using a modified 58 base pair R2 (read 2) at the UBC Biomedical Research Centre.

Processing and normalization of scRNA-seq data. Raw sequence files were processed with Cell Ranger v.2.1.0. The resulting filtered count matrices were read into SingleCellExperiment objects. According to quality control parameters (≥3 median absolute deviations from the median), outlier cells were filtered out using the scater R package. Additionally, cells with ≥20% mitochondrial unique molecular identifiers (UMIs) or ≥50% ribosomal UMIs were removed. (Ovarian cancer cells can have higher mitochondrial percentages than other cell types, as in Schelker et al.[40].) Size factors were computed using quickCluster and computeSumFactors from the scran R package. Following this, data normalization was performed using scater-normalize. Principal component analysis (PCA) was performed on the resultant normalized log counts for the top 1,000 most variable genes. The first 50 principal components were used as input for UMAP.

For the HGSC data, two UMAP parameters were changed from the defaults (umap R package v0.2.0.0) due to the presence of an outlier in UMAP space along the first dimension. The number of neighbors was set to 25 and the minimum distance was set to 0.2.

Cell cycle scores were computed with cyclone from the scran package[37,41].

CellAssign. The following marker gene list (see also Supplementary Table 2) was used for CellAssign: B cells: *VIM*^{c}, *MS4A1*^{c}, *CD79A*^{c}, *PTPRC*^{c}, *CD19*^{c}, *BANK1* (ref. [42]); T cells: *VIM*^{c}, *CD2*^{c}, *CD3D*^{c}, *CD3E*^{c}, *CD3G*^{c}, *CD28*^{c}, *PTPRC*^{c}; monocyte/macrophage: *VIM*^{c}, *CD14*^{c}, *FCGR3A*^{c}, *CD33*^{c}, *ITGAX*^{c}, *ITGAM*^{c}, *CD4*^{c}, *PTPRC*^{c}, *LYZ*^{c}; epithelial cells: *EPCAM*^{c}, *CDH1*^{c}, *KRT8* (ref. [43]), *WFDC2* (ref. [43]); ovarian stromal cells: *VIM*^{c}, *MUM1L1* (ref. [44]), *FOXL2* (ref. [45]), *ARX*[44], *DCN*[43], *TPT1* (ref. [45]), *RBP1* (ref. [44]); ovarian myofibroblast: *VIM*^{c}, *MUM1L1* (ref. [45]), *FOXL2* (ref. [45]), *ARX*[45], *ACTA2*^{c}, *COL1A1*^{c}, *COL33A1*^{c}, SERPINH1 (ref. [43]); vascular smooth muscle cells: *VIM*^{c}, *ACTA2*^{c}, *MYH11*^{c}, *PLN*[46], *MYLK*^{c}, *MCAM*[47], *COL1A1*^{c}, *COL3A1*^{c}, *SERPINH1* (ref. [48]); endothelial cells: *VIM*^{c}, *EMCN*^{c}, *CLEC14A* (ref. [49]), *CDH5*^{c}, *PECAM1*^{c}, *VWF*^{c}, *MCAM*[47], *SERPINH1* (ref. [43]). The superscript letter c indicates a canonical marker.

DCN, *TPT1* and *RBP1* were selected as markers of ovarian stromal cells based on differential expression results comparing normal fibroblasts (ovarian stromal cells) and malignant fibroblasts from Shih et al.[43] (these were the top three genes upregulated in normal fibroblasts by log(fold change) where $Q < 0.05$). Ovarian stromal cells and myofibroblasts were identified based on the expression of *MUM1L1* and *ARX*, ovary-specific markers known to be expressed in stroma from bulk RNA-seq and immunohistochemistry[45] (Fig. 3d and Supplementary Fig. 5a), with myofibroblasts distinguished by higher expression of α-smooth muscle actin and various collagen genes[43] (Fig. 3d and Supplementary Fig. 5a). A group of cells expressing the vascular smooth muscle markers *ACTA2*, *MYH11* and *MCAM*[47] was also identified with CellAssign (Supplementary Fig. 5a). CellAssign was run with default parameters and five random initializations.

Unsupervised clustering. Unsupervised clustering of epithelial cells from CellAssign (probability ≥90%) was performed with Seurat[3] using a resolution parameter of 0.2 (for fairly coarse resolution). Unsupervised clustering of all cells was performed with Seurat and SC3 (ref. [2]) using default parameters. For Seurat, resolutions of 0.8 and 1.2 were used (these represent low-to-moderate and high levels within the recommended range). Additionally, Seurat clustering was also performed using data for the same set of marker genes provided to CellAssign. (SC3 failed to run when provided with this number of genes.)

Differential expression and enrichment analysis. log(fold change) values from the FindMarkers function (filtering out ribosomal and mitochondrial genes) from scran were used as input for gene set enrichment analysis with the fgsea R package (v1.9.4), using default parameters with 10,000 permutations and the hallmark pathway gene set[44]. Annotations for cell cycle-associated pathways (E2F targets, G2-M checkpoint and mitotic spindle) and immune-associated pathways (including IFN-γ and IFN-α response, coagulation, complement, interleukin-6-Janus kinase/signal transducer and activator of transcription signaling and allograft rejection) were taken from Liberzon et al.[44]. All reported *Q* values refer to Benjamini–Hochberg-corrected *P* values for two-sided tests.

Follicular lymphoma. *Sample preparation.* Leftovers from clinical flowed samples were collected and frozen in FCS containing 10% dimethylsulfoxide (DMSO). Cells were thawed and washed according to the steps outlined in the 10x Genomics Sample Preparation Protocol. Cells were stained with propidium iodide for viability and sorted in a BD FACSAria Fusion (BD Biosciences) using an 85-µm nozzle. Sorted cells were collected in 0.5 ml of medium, centrifuged and diluted in 1× PBS with 0.04% BSA.

Library preparation and sequencing. Cell concentration was determined using a Countess II Automated Cell Counter (Thermo Fisher Scientific) and approximately 3,500 cells were loaded per well in the Chromium Single Cell A Chip Kit (3′). Single-cell libraries were prepared according to the Chromium Single Cell 3′ Reagent Kits User Guide (v.2). Single-cell libraries from two samples were pooled and sequenced on one HiSeq 2500 125 base PET lane (Illumina).

Preprocessing and normalization of scRNA-seq data. The preprocessing steps for the follicular lymphoma data were identical to those for the HGSC scRNA-seq data, described in the Processing and normalization of scRNA-seq data section, with the exception of different mitochondrial and ribosomal thresholds (cells with ≥10% mitochondrial UMIs or ≥60% ribosomal UMIs were removed).

Scvis analysis. scvis train v.0.1.0 (ref. [26]) was run with default settings on the top 50 principal components to produce a two-dimensional embedding of the follicular lymphoma data. Early stopping was added to scvis so that the model would terminate after three successive iterations of no improvement (relative improvement in evidence lower bound $<10^{-5}$). The resultant model was saved and used for mapping.

CellAssign. The following marker gene list (see also Supplementary Table 2) was used for CellAssign[42,50]: B cells: *CD19*^{c}, MS4A1^c, *CD79A*^{c}, *CD79B*^{c}, *CD74*^{c}, *CXCR5*[51]; cytotoxic T cells: *CD2*^{c}, *CD3D*^{c}, *CD3E*^{c}, *CD3G*^{c}, *TRAC*^{c}, *CD8A*^{c}, *CD8B*^{c}, *GZMA*^{c}, *NKG7*^{c}, *CCL5*^{c}, *EOMES*^{c}; follicular T helper cells: *CD2*^{c}, *CD3D*^{c}, *CD3E*^{c}, *CD3G*^{c}, *TRAC*^{c}, *CD4*^{c}, *CXCR5*^{c}, *PDCD1*^{c}, *TNFRSF4* (ref. [42]), *ST8SIA1* (ref. [42]), *ICA1* (ref. [42]), *ICOS*[42]; other CD4$^+$ T cells: *CD2*^{c}, *CD3D*^{c}, *CD3E*^{c}, *CD3G*^{c}, *TRAC*^{c}, *CD4*^{c}, *IL7R*[42]. The superscript letter c indicates a canonical marker.

CellAssign was run with default parameters and five random initializations.

'Patient' was added as an additional covariate into the design matrix *X* (see Model description section). The best result according to marginal log-likelihood at convergence was kept. Optimization was considered converged after three consecutive rounds of no improvement (relative change in log-likelihood $<10^{-5}$). MAP assignments from CellAssign were used for downstream analysis.

No evidence of regulatory T cells (*FOXP3* and *IL2RA* expression), natural killer cells (*NCAM1* expression), and myeloid cells (*CD14*/*CD16* and *LYZ* expression) was detected.

Unsupervised clustering. Unsupervised clustering of all cells was performed with Seurat and SC3 (ref. [2]) using default parameters. For Seurat, resolutions of 0.8 and 1.2 were used (these represent low-to-moderate and high levels within the recommended range). Additionally, Seurat clustering was also performed using data for the same set of marker genes used with CellAssign. (SC3 failed to run when provided with this number of genes.)

Classifying B cells. B cells from CellAssign were further subclassified into 'malignant' or 'nonmalignant' groups according to the expression of the constant region of the immunoglobulin light chain (κ or λ type) and the results of the PCA. Seurat[3] (resolution = 0.8) was used to separate B cells into clusters, based on the top 50 principal components. Following this, the sole cluster associated with IGKC (constant region) expression was designated as nonmalignant. We further reasoned this was the case based on the cluster containing a mixture of T1 and T2 cells and constituting only a minor subset of B cells.

Differential expression between time points. Differential expression analysis between time points for a given cell type and patient was performed using voom from the limma package (v3.39.12) for each patient and cell type separately, with time point as the independent variable. Genes with low expression (<500 UMIs in total across all cells) were removed. *P* values were adjusted with the Benjamini–Hochberg method and genes with $Q \leq 0.05$ (two-sided) were considered differentially expressed. Differential expression between malignant and nonmalignant B cells was performed similarly, but using the formula ~malignant_status +timepoint +malignant_status:timepoint to control for time point and any interactions.

Reactome pathway enrichment analysis. Pathway analysis was performed for the top 50 most upregulated and top 50 most downregulated genes (separately) by log(fold change) from limma (where $Q \leq 0.05$, filtering out ribosomal and mitochondrial genes). Over-representation of Reactome[30] pathways was assessed with the R package ReactomePA v1.27.0. Pathways were considered significantly over-represented if the adjusted $P \leq 0.05$ (two-sided) and at least two genes from the pathway were present.

Reactive lymph node data. *Sample preparation.* Cell suspensions from patients with reactive lymphoid hyperplasia but no evidence of malignant disease and collagen disease were used. Leftovers from clinical flowed samples were collected and frozen in FCS containing 10% DMSO. The day of the experiment, cell suspensions were rapidly thawed at 37 °C and washed according to the steps outlined in the 10x Genomics Sample Preparation Protocol. Cells were stained with 4,6-diamidino-2-phenylindole (DAPI) and viable cells (DAPI$^-$) were sorted on a FACSAria III or FACSAria Fusion (BD Biosciences) instrument.

Library preparation and sequencing. Approximately 8,700 cells per sample were loaded into a Chromium Single Cell 3′ Chip Kit v.2 (catalog no. PN-120236; 10x Genomics) and processed according to the user guide. Libraries were constructed using the Chromium Single Cell 3′ Library & Gel Bead Kit v.2 (catalog no. PN-120237) and

Chromium i7 Mulitiplex Kit (catalog no. PN-120262). Single-cell libraries from two samples were pooled and sequenced on one HiSeq 2500 125 base PET lane.

Preprocessing and normalization of scRNA-seq data. The preprocessing steps for the reactive lymph node data were identical to those for the follicular scRNA-seq data.

scvis analysis. The identities of the top 1,000 most variable genes and PCA loadings from the follicular lymphoma data analysis were used to compute a 50-dimensional embedding for the reactive lymph node data. The resultant 50 principal components were provided as input to scvis map[26], using the trained model and default settings.

General statistical methods. In all boxplots, the whiskers denote data within 1.5× the interquartile range of the upper and lower quartiles. Where plotted over the boxplots, points were horizontally, but not vertically jittered. Correlations were calculated using the cor function in the R programming language v.3.5.0.

Raw scRNA-seq read data and count matrices availability. Raw scRNA-seq read data and count matrices for HGSC, follicular lymphoma and reactive lymph node samples have been deposited with the European Genome-phenome Archive (accession no. EGAS00001003452). Raw and normalized count matrices are available from Zenodo (https://doi.org/10.5281/zenodo.3372746).

Materials availability. Further information and requests for resources and reagents should be directed to S.P.S. (shahs3@mskcc.org). Limited quantities of the HGSC and follicular lymphoma patient tissues and cell suspensions used to generate the scRNA-seq data are available.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw sequencing data for all experiments in this paper are available from the European Genome-phenome Archive (accession no. EGAD00001004585).

Code availability

CellAssign is available as an R package at www.github.com/irrationone/cellassign.

References

33. Eling, N., Richard, A. C., Richardson, S., Marioni, J. C. & Vallejos, C. A. Correcting the mean-variance dependency for differential variability testing using single-cell RNA sequencing data. *Cell Syst.* **7**, 284–294.e12 (2018).
34. Kingma, D. P. & Ba, J. L. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/pdf/1412.6980.pdf (2014).
35. Abadi, M. et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. Preprint at https://arxiv.org/pdf/1603.04467.pdf (2015).
36. Sinha, D., Kumar, A., Kumar, H., Bandyopadhyay, S. & Sengupta, D. dropClust: efficient clustering of ultra-large scRNA-seq data. *Nucleic Acids Res.* **46**, e36 (2018).
37. Lun, A. T., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res.* **5**, 2122 (2016).
38. Adam, M., Potter, A. S. & Potter, S. S. Psychrophilic proteases dramatically reduce single-cell RNA-seq artifacts: a molecular atlas of kidney development. *Development* **144**, 3625–3632 (2017).
39. O'Flanagan, C. H. et al. Dissociation of solid tumour tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase associated stress responses. Preprint at *bioRxiv* https://doi.org/10.1101/683227 (2019).
40. Schelker, M. et al. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat. Commun.* **8**, 2032 (2017).
41. Scialdone, A. et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54–61 (2015).
42. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
43. Shih, A. J. et al. Identification of grade and origin specific cell populations in serous epithelial ovarian cancer by single cell RNA-seq. *PLoS ONE* **13**, e0206785 (2018).
44. Liberzon, A. et al. The Molecular Signatures Database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
45. Uhlen, M. et al. A pathology atlas of the human cancer transcriptome. *Science* **357**, eaan2507 (2017).
46. Perisic Matic, L. et al. Phenotypic modulation of smooth muscle cells in atherosclerosis is associated with downregulation of *LMOD1*, *SYNPO2*, *PDLIM7*, *PLN*, and *SYNM*. *Arterioscler. Thromb. Vasc. Biol.* **36**, 1947–1961 (2016).
47. Espagnolle, N. et al. CD146 expression on mesenchymal stem cells is associated with their vascular smooth muscle commitment. *J. Cell. Mol. Med.* **18**, 104–114 (2014).
48. Rocnik, E., Saward, L. & Pickering, J. G. HSP47 expression by smooth muscle cells is increased during arterial development and lesion formation and is inhibited by fibrillar collagen. *Arterioscler. Thromb. Vasc. Biol.* **21**, 40–46 (2001).
49. Mura, M. et al. Identification and angiogenic role of the novel tumor endothelial marker CLEC14A. *Oncogene* **31**, 293–305 (2012).
50. Deenick, E. K. & Ma, C. S. The regulation and role of T follicular helper cells in immunity. *Immunology* **134**, 361–367 (2011).
51. Payne, D., Drinkwater, S., Baretto, R., Duddridge, M. & Browning, M. J. Expression of chemokine receptors CXCR4, CXCR5 and CCR7 on B and T lymphocytes from patients with primary antibody deficiency. *Clin. Exp. Immunol.* **156**, 254–262 (2009).

nature research

Corresponding author(s): Sohrab P Shah, Kieran R Campbell

Last updated by author(s): July 2nd 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	Open source code: CellAssign, commit code 01676a42888e4c2cb393393f3dc8cda3589f2e78 (https://github.com/Irrationone/cellassign) 10x Genomics CellRanger v2.1.0 R 3.5.0 python 3.6.6 Docker version 18.06.1-ce, build e68fc7a conda 4.5.11 tidyverse 1.2.1 R package/suite data.table 1.11.8 R package scater 1.10.0 Bioconductor package scran 1.10.1 Bioconductor package Seurat 2.3.3 Bioconductor package fgsea 1.8.0 Bioconductor package ReactomePA 1.26.0 Bioconductor package limma 3.38.3 Bioconductor package pymc3 3.5 python package

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw single cell RNA-sequencing read data and count matrices for HGSC, follicular lymphoma, and reactive lymph node samples are being deposited in the European Genome-Phenome Archive (EGA) under accession number EGAS00001003452.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. The number of samples were chosen for this exploratory study based on the availability of materials at study time.
Data exclusions	No data was excluded from the analysis. Filtering and quality control of single cell RNA-seq data is described in the Methods.
Replication	All experimental steps are detailed in the manuscript to ensure replication. All data analyses are available online in reproducible format. We have not attempted study replication.
Randomization	The study is exploratory and descriptive to demonstrate the utility of a computational method, and no case control comparisons were performed, so no randomization was considered.
Blinding	No blinding as no case-control comparisons are made.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	characteristics Follicular lymphoma (FL) patients: the patient with progressed FL (male, 38 yo at diagnosis) was treated with rituximab 2 years after diagnosis; the patient with transformed FL (male, 61 yo at diagnosis) was previously treated with radiotherapy for an FL malignancy in the palate. The HGSC patient considered in this study was 67 years old female treatment naive at diagnosis, stage IIIC (representative of advanced HGSC, see Zhang et al. 2018 "Interfaces of malignant and immunologic clonal dynamics", Cell).
Recruitment	Consented patients seen at BC Cancer where sufficient tumour/cell suspension material was collected to perform single-cell RNA-sequencing were recruited.

FL samples: Representative samples were selected based on the availability of fresh frozen cell suspensions (diagnostic and relapse pairs) of FL patients. Samples were centrally reviewed by pathological criteria based on the WHO classification (Swerdlow, WHO 2017).

HGSC: Patients were recruited and screened by Dr. Jessica McAlpine at BC Cancer agency

The authors are aware of no biases in patient recruitment that would impact results.

Ethics oversight

UBC Human Research Ethics Board

Note that full information on the approval of the study protocol must also be provided in the manuscript.