Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

*Nat Methods*. Author manuscript; available in PMC 2020 September 11.

Published in final edited form as:

*Nat Methods*. 2019 October ; 16(10): 1007–1015. doi: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 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], **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

## Abstract

Single-cell RNA sequencing (scRNA-seq) has enabled decomposition of complex tissues into functionally distinct cell types. Often, investigators wish to assign cells to cell types, performed through unsupervised clustering followed by manual annotation, or via "mapping" procedures to

Correspondence: kieran.campbell@stat.ubc.ca, shahs3@mskcc.org.

Author Contributions

Study design: A.W.Z., K.R.C., S.P.S.; Writing: A.W.Z., K.R.C., S.P.S.; Manuscript review: A.W.Z., C.O., 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., S.P.S.; Data interpretation: A.W.Z., S.A., C. Steidl, K.R.C., S.P.S.; Data curation: B.H., D.L., L.C., C. Sarkozy; Data analysis: A.W.Z., K.R.C., N.C., M.W., P.W., T.C., X.W.; Model development: A.W.Z., K.R.C., S.P.S.; Single cell processing: C.O., E.A.C., J.L.P.L.; Case identification: A.M., J.N.M., C. Steidl, C. Sarkozy.; Supervision: K.R.C., S.P.S., C. Steidl, S.A.

Data availability

Raw sequencing data for all experiments in this paper is available from the European Genome-Phenome Archive (EGA) under access ID EGAD00001004585.

Competing Interests

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 (www.github.com/irrationone/cellassign), a probabilistic model that leverages prior knowledge of cell type marker genes to annotate scRNA-seq data into pre-defined 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.

## Editorial Summary:

CellAssign uses a probabilistic model to assign single cells measured with RNA-seq to a given cell type defined by known marker genes, enabling automated annotation of cell types present in the tumor microenvironment.

## 1. Introduction

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-seq (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 [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 over-clustering of populations that are not sufficiently variable.

In the context of robust clustering which 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 as it relies on knowledge of marker gene expression levels and 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-

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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, a statistical framework that assigns cells to both known and de novo cell types in scRNAseq 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 which uniquely identify cell types may be established through expert knowledge based on the literature, databases such as CellMarker [12], or derived directly from data from resources such as PanglaoDB (Supplementary Notes 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 highly scalable, 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 FACS-purified human embryonic stem cell data [13], pre-annotated data, and cell line data for multiple scRNA-seq platforms [14]. CellAssign outperforms both clustering and "mapping" and is robust to errors in marker gene specification. Additionally, we generated two novel datasets to exemplify the ability of CellAssign to delineate the composition of the tumour microenvironment (TME) across anatomic 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.

## 2 Results

### 2.1 CellAssign: probabilistic and automated cell type assignment

The CellAssign statistical framework (Figure 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 single cell RNA-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 mis-assignment 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 Methods.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

### 2.2 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 [15] 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 fraction (0.05 to 0.45).We evaluated the performance of unsupervised methods (Seurat [3], SC3 [2], phenograph [16], densitycut [17], dynamicTreeCut [18]), supervised methods (scmap-cluster [10], correlation-based [19]) (Methods), and another marker gene-based approach (SCINA [20]). Half of the simulated cells (n=1000 training, n=1000 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 fraction, CellAssign performed better than alternative workflows in both accuracy and F1 score (Figure 2A, Supplemental Table 1). CellAssign's assignments remained more accurate than the other methods when the analysis was repeated providing other methods with marker genes only (Supplemental Figure 1A), on data simulated from parameter estimates fitted to B cells and CD8+ T cells (Supplemental Figure 2A,B, Supplemental Table 1), and when clusters were mapped to existing purified cell types based on maximum correlation (Methods, Supplemental Figure 3).

We then evaluated the performance of CellAssign on real scRNA-seq data from experimentally sorted populations. For FACS-purified H7 human embryonic stem cells in various stages of differentiation (8 cell types) [13], 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 (Supplemental Table 2; Methods). CellAssign outperformed SCINA and the most competitive unsupervised methods from systematic analysis (SC3, Seurat) [6] according to accuracy and cell type-level F1 score (Supplemental Figure 4A–E,G; Methods), with similar results obtained using only marker gene expression data (Supplemental Figure 4F,H) (CellAssign F1: 0.943, accuracy: 0.944; best F1 of other methods: 0.841, accuracy: 0.93). As an example of CellAssign's ability to discriminate highly related cell types, anterior primitive streak (APS) and mid primitive streak (MPS) cells were accurately classified (83/84 correct), while no other method could reliably do so, assigning APS and MPS cells to the same cluster (Supplemental Figure 4).

We next tested the robustness of CellAssign across a range of mis-specified 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 (Supplemental Figure 1C–D, Supplemental Figure 2D, Supplementary Notes 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 on a recent real scRNA-seq dataset of the human liver [21], CellAssign maintained high accuracy of assignment in these situations, with superior performance to SCINA when too many cell types are specified (CellAssign F1: 0.985, accuracy: 98.5%, SCINA F1: 0.910, accuracy: 91.0%; Figure 2B–E,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 3 specific marker genes at modest expression levels, mature B cells could be differentiated from biologically similar cell types present in the data (Supplementary Notes 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 Notes 2.6). Finally, using a recent study of mixed "pseudo-cells" [14], 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 (Figure 2F, Supplementary Notes 2.7 and 2.8).

### 2.3 Delineating the tumour microenvironment composition of spatially sampled HGSC

We next exemplified CellAssign by decomposing cancer tissues from patients into constituent microenvironmental components and profiled variation across anatomic space and between malignant clones. We generated scRNA-seq data for 5233 cells from 2 spatial sites from an untreated high-grade serous ovarian cancer patient at the time of primary debulking surgery. Dimensionality reduction with uniform manifold approximation and projection (UMAP [22]) revealed 4 major site-specific populations and 4 mixed populations with representation from both samples (Figure 3A). Using a panel of literature-derived marker genes (Supplemental Table 2, Methods), CellAssign identified 8 major epithelial, stromal, and immune cell types (Figure 3B,C), which were consistent with well-known marker gene expression (Figure 3D, Supplemental Figure 5A, Methods). Unlike other non-epithelial 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 was available, demonstrating CellAssign can annotate TME cell types for which marker genes have been orthogonally derived in the literature but 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%; Figure 3C) and dominated by myeloid populations (67% and 87.5% of hematopoietic cells in 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 non-hematopoietic cell types (Supplemental Figure 6). Thus for TME decomposition and profiling, subtle differences between constituent cell types may be better distinguished by CellAssign over standard approaches [23].

We next 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 [24, 25] (Supplemental Figure 5B). Within epithelial cells we identified 5 clusters using Seurat (Figure 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 [NES] = 1.42, Q = 0.039), including N-cadherin (CDH2) and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

CD90 (THY1) (Figure 3E–G), and downregulation of E-cadherin (CDH1; log-fold change = −0.32, Q = 1.1e-19). Immune-associated pathways were also significantly upregulated, primarily due to cluster 1, one of the two clusters from the left ovary (Figure 3E,F,H, Supplemental Figure 7A, Supplemental Figure 8A–B, Methods). HLA class I genes were among the most differentially expressed genes associated with these pathways (Supplemental Figure 8B). While HLA expression in cluster 1 was comparable to levels in stromal cells and myofibroblasts, expression levels in other clusters were lowest across all cell types (Supplemental Figure 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 NES > 2.05, Q < 0.0012; Supplemental Figure 8C–E). Accordingly, apoptosis and glycolysis pathways were also upregulated while cell cycle and oxidative phosphorylation-associated pathways were downregulated, consistent with hypoxiainduced cell cycle arrest and metabolic dependence on glycolysis (Supplemental Figure 8C,D). Together, profiling of multi-site HGSC samples demonstrate 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.

### 2.4 Temporal immune microenvironment dynamics accompanying follicular lymphoma progression and transformation

We next applied CellAssign to delineate temporal microenvironmental changes in follicular lymphoma (FL) through scRNA-seq of 9754 cells from temporally collected lymph node biopsies of 2 FL patients at two time points each. Histopathological transformation to diffuse large B cell lymphoma (DLBCL) occurred in one patient (FL1018), while progression occurred in the other (FL2001) 2 years after rituximab treatment (Figure 4A). We first computed a UMAP representation, yielding three major patient-specific and two mixed populations comprised of cells from both patients (Figure 4B). Leveraging literature-derived marker gene information (Supplemental Table 2), we applied CellAssign to identify 4 major T and B cell types (Figure 4C,D, Supplemental Figure 9, Methods). In comparison, most unsupervised approaches were unable to cleanly resolve T cell subpopulations in the microenvironment (Supplemental Figure 10). Hypothesizing the mixed B cell population likely contained nonmalignant B cells (Figure 5A), we examined immunoglobulin light chain constant domain expression using CellAssign (Figure 5B) to confirm heterogeneous light chain expression (κ/IGKC or λ/IGLC) in the polyclonal nonmalignant B cell population and homogeneous light chain restriction in the clonally identical malignant B cell population [26]. The three patient-specific B cell populations were largely IGLC positive, consistent with malignant expansion of λ-chain expressing cells. Applying CellAssign to the mixed population (Supplemental Table 2) showed that 576/907 cells (63.5%) were IGKC+ (FL1018: 76/118 (64.4%), FL2001: 500/789 (63.4%)), consistent with the expected polyclonal 60:40 ratio in normal lymphoid organs [27] (Supplemental Figure 11). In addition, scRNA-seq data of reactive lymph node (RLN) B cells from four healthy donors mapped onto the mixed B cell population [28], (Figure 5C, Supplemental Figure 12). This population also expressed significantly lower levels of follicular lymphoma markers BCL2 and BCL6 [26,29–31] than the other B cells (all log-fold change values < −0.34, Q < 5.4e-07; Supplemental Figure 13, Supplemental Table 3). Together these results demonstrate

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the ability of CellAssign to distinguish malignant from nonmalignant B cells, thereby enhancing cell decomposition capacity and cell type interpretation for lymphoid cancers.

We next 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% to 0.8%; FL2001: 42.5% to 1.4%) (Figure 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 diagnostic samples (Figure 5E,F). In FL1018, these compositional changes were accompanied by significant upregulation of immune-associated pathways such as cytokine signalling [32] 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, IFNG, GZMA, and PRF1 in cytotoxic T cells [33]; Supplemental Figure 13, Figure 5G, Supplemental Table 3). Together, these results illustrate how CellAssign can be applied to study compositional and phenotypic changes in the tumour microenvironment at the level of individual cell types.

## 3 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, though these could be incorporated by extending the model. Nevertheless, we suggest 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 consequently intrinsically mitigates issues common to existing unsupervised clustering approaches, including batch effects on clustering and the need of post-hoc 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 [34]. CellAssign is therefore 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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

# 9 Online Methods

## 9.1 Ethics

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

## 9.2 The CellAssign model

**9.2.1 Model description—**Let $Y$ be a cell-by-gene expression matrix of raw counts for $N$ cells and $G$ genes. Suppose among those cells we have $C$ total cell types, each of which is defined by high expression of several "marker" genes. We encode the relationship between cells and marker genes through a binary matrix $\rho$, 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 an indicator vector $z = \{z_n\}$ that encodes to which of the C cell types each cell belongs:

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

In order to assign cells to cell types we perform statistical inference of the probability that each cell is of a given cell type for which we must compute the quantity $p\left(z_n = c | Y, \hat{\Theta}\right)$ where $\hat{\Theta}$ are the 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

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

Where

$$\log \mu_{ngc} = \log s_n + \delta_{gc}\rho_{gc} + \beta_{g0} + \sum_{p=1}^{P} \beta_{gp} x_{pn}$$

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

The intuition here 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 $\delta_{gc}$ is inferred. In this way we put no restriction that marker genes can't 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 $\delta_{gc}$ corresponds to the average log fold change that gene g is over-expressed in cell c, which only occurs for marker genes for cell types since $\rho_{gc}$ must equal 1 for this to contribute to the likelihood. In simulations we found that CellAssign was able to accurately estimate these parameters (Supplemental Figure 1b, Supplemental Figure 2c). By default we impose a lower bound such that $\delta > \log 2$, making the interpretation that a marker gene must be over-expressed 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 control for technical or sample effects through the matrix $X$.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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

Remaining model parameters are initialized as follows:

- $\beta_{gp}$ is drawn from a $\mathcal{N}(0, 1)$ distribution
- $\log \delta_{gc}$ is drawn from a $\mathcal{N}(0, 1)$ 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

$$y_{ng} \mid z_n = c \sim \mathcal{NB}(\mu_{ngc}, \tilde{\phi}_{ngc})$$

where NB is the negative binomial distribution parametrized by a mean $\mu$ and a $\mu$-specific dispersion $\tilde{\phi}_{ngc}$. We define $\tilde{\phi}_{ngc}$ as a sum of radial basis functions dependent on the modelled mean $\mu_{\text{ngc}}$ as proposed by a recent publication [35]:

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

where $a_{\text{i}}$ and $b_{\text{i}}$ represent RBF parameters to be inferred, $B$ is the total number of *centers* of the RBF, and x$i$ is center $i$. The centers are set to be equally spaced apart from 0 to the maximum number of counts max $y_{ng}$.

**9.2.2 Inference—**Using expectation-maximization (EM) for inference, the latent variables are $z \equiv \{z_{\text{n}}\}$ while the model parameters to be maximized are $\boldsymbol{\delta} = \{\delta_{\text{gc}}\}$, $\boldsymbol{\beta} = \{\beta_{\text{g0}}, \beta_{gp}\}$, $\boldsymbol{a} = \{a_{\text{i}}\}$, and $\boldsymbol{b} = \{b_{\text{i}}\}$.

**E-step:** Compute

$$\gamma_{nc} := p\left(z_n = c \mid \boldsymbol{y}_n, \boldsymbol{\delta}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{a}^{(t-1)}, \boldsymbol{b}^{(t-1)}\right) = \frac{\prod_g \mathcal{NB}(\mu_{ngc}, \tilde{\phi}_{ngc})}{\sum_{c'} \prod_{g'} \mathcal{NB}(\mu_{ng'c'}, \tilde{\phi}_{ng'c'})}$$

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

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**M-step:** During the M-step we optimize the above Q-function using the ADAM optimizer [36] as implemented in Google's Tensorflow [37]. By default, we use a learning rate of 0.1, allow a maximum of $10^5$ ADAM iterations per M-step, and consider the M-step converged when the relative change in the Q function value falls below $10^{-4}$. By default we consider the EM algorithm converged when the relative change in the marginal log likelihood falls below $10^{-4}$.

**9.2.3 Code availability—**CellAssign is available as an R package at www.github.com/irrationone/cellassign.

## 9.3 Simulation

**9.3.1 Model description and rationale—**Initially, we attempted to simulate multi-group data from the splatter model. We employed 10x Chromium data for peripheral blood mononuclear cells (PBMC) [19] with cell type labels derived from [38] to determine realistic parameter estimates for the differential expression component of the model (see below). In order to do so, group-specific log fold-change (logFC) values were drawn from a mixture distribution of a central, narrow Gaussian-Laplace mixture (representing non-differentially expressed genes) and two flanking, absolute value-transformed Gaussians (representing downregulated/upregulated genes). This mixture distribution was fitted to logFC values derived from differential expression analysis (see below).

However, inspection of posterior predictive samples for multiple fits, using labeled single-cell RNA-seq data from [19] and FACS-purified data from Koh et al. [13] (Supplemental Figure 14A–D), revealed that this model systematically underestimates extreme logFC values (Supplemental Figure 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 logFC distributions from this modified model better fit the observed data (Supplemental Figure 14D, Supplemental Figure 14H). Consequently, we used this model to perform simulation analysis.

**9.3.2 Model fitting—**The models described above were fit to logFC values derived from real data. Using the labeled 10x Chromium data for 68k PBMCs [19], differential expression was performed with the findMarkers function from the R package scran [39]. To generate corresponding null distributions of logFC values for non-differentially expressed genes, we split data for each cell type into equally sized halves 10 times, running findMarkers to compare the resulting halves. A central Gaussian-Laplace mixture ($\mu = 0$) was first fit to the null logFC values. The distribution of posterior predictive logFC values appeared to be consistent with observed logFC values for this null component (Supplemental Figure 14D). Following this, the entire mixture distribution was fitted to logFC values for pairs of distinct cell types, using maximum a posteriori (MAP) estimates of parameters for the central Gaussian-Laplace component. Posterior distributions of model parameters were inferred using the no U-turn sampler (NUTS) in pymc3, using 4 independent chains, 1000 tuning iterations, and 2500 additional iterations per chain. Trace plots and the Gelman-Rubin diagnostic were used to assess convergence.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**9.3.3 Simulating multi-group data—**Expression count matrices were simulated using a modified version of the splatter package. 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) logFC values between naïve CD4+ and naïve CD8+ T cells (Supplemental Figure 14A); and (2) logFC values between B cells and CD8+ T cells (section 9.3.1) for the differential expression component. The proportion of downregulated genes out of differentially expressed genes was set to 0.5 (i.e. equally probable for a differentially expressed gene to be downregulated vs. 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+ vs. naïve CD8+ T cells, and 0.153 for B vs. CD8+ T cells.) The number of simulated cells, n, was set to 2000, and 1000 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 $\rho$, $p_d$ was set to 0.25 and the number of simulated cells n to 1500.

Simulations were run 9 times with unique random seeds for each combination of parameter settings.

**9.3.4 Clustering multi-group data—**Count matrices were normalized with scater normalize and the top 50 principal components were computed from the top 1000 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 PCs. For SC3, the entire normalized SingleCellExperiment object was passed as input instead. For supervised methods (scmap- cluster [10] and correlation-based [19]), expression data for both training and evaluation sets was 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 logcounts. 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 5th percentile of differentially expressed genes according to logFC and the top 10th 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 $\rho$ were flipped from 0 to 1 (or vice versa). All other parameters were set to the defaults.

**9.3.5 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Spearman correlation between these logFC values and the simulated (true) logFC values for each simulated group. Each inferred cluster was mapped to 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 cluster 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.

**9.3.6 Evaluation—**Accuracy and cell-level F1 score were computed to evaluate clustering performance. The cell-level F1 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.

**9.3.7 Marker gene overspecification/underspecification analysis on simulated data—**For analyses of to test the robustness of CellAssign to overspecification of marker gene matrices, we simulated 6 groups at equal proportions (n = 1500, $p_{d = 0.15)}$, following the methods described insection 9.3.3 above. Following this, marker gene matrices were generated for the simulated cell types (see section 9.3.4). Zero to 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 (1000). These are analogous settings for both tools that consider all marker genes and allow for inference of novel cell types. Cell type assignments from both methods were evaluated as described in section 9.3.6. While CellAssign can automatically discount cell types that don't exist in the data, SCINA was run with various values of sensitivity_cutoff, which facilitates "removal" of those cell types [20].

Similarly, we simulated 6 groups at equal proportions as described above to test robustness to underspecification of the marker gene matrix (for novel cell type discovery). Zero to 4 cell types were then removed prior to marker gene selection (but retained in the data), to ensure that marker genes were being selected with no knowledge of "non-existent" 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.

**9.3.8 Benchmarking—**We generated synthetic datasets for benchmarking from the modified splatter model (section 9.3.1) with 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 $N \in \{1000,2000,4000,8000,10000,20000,40000,80000\}$ 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 5000 cells. On simulated data for 80000 cells from 2 cell types, CellAssign completed in under 2 minutes, appearing to scale at worst linearly in the number of cell types and marker genes used per cell type (Supplemental Figure 15). Five separate CellAssign runs were timed for each combination of parameters.

## 9.4 Koh et al. dataset

The Koh et al. [13] dataset consists of scRNA-seq data for 531 cells of human embryonic stem cells at various stages of differentiation.

### 9.4.1 Preprocessing and normalization of single-cell RNA-seq data—

Preprocessed data was obtained from the R package DuoClustering2018 [6, 13]. Cell- types with both single-cell RNA-seq data and bulk RNA-seq data were used: hESC (day 0 human embryonic stem cell), APS (day 1 anterior primitive streak), MPS (day 1 mid primitive streak), DLL1pPXM (day 2 DLL1+ paraxial mesoderm), ESMT (day 3 somite), Sclrtm (day 6 sclerotome), D5CntrlDrmmtm (day 5 dermomyotome), D2LtM (day 2 lateral mesoderm). Normalization and dimensionality reduction was performed with scater normalize, runPCA, runTSNE, and runUMAP. The top 500 most variable genes were used to compute the top 50 principal components, and the top 50 PCs were used as input for t-SNE.

### 9.4.2 Identification of marker genes from bulk RNA-seq data—

Differential expression analysis results for 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 supplemental materials of [13] were used to compute log-scale relative gene expression levels. Next, we identified gene-specific thresholds for defining the cell types in which each gene is a marker. For each gene, relative expression levels across cell types were sorted in ascending order, denoted as $E_1$, …, $E_C$, where C is the total number of cell types. The maximum difference between sorted expression levels, $\max_{1\leq i<C(Ei+1\,-\,Ei),}$ was then computed. Denote the index i for gene g in which this difference is maximal $i_{g.}$ For gene g, cell types in which relative expression values were equal to or greater than $E_{ig+1}$ were considered cell types with gene g as a marker. Genes with a maximum difference value in the the top 20th percentile were used as marker genes.

### 9.4.3 CellAssign—

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

### 9.4.4 Unsupervised clustering—

Unsupervised clustering was performed on the top 50 PCs with Seurat [3] (resolution $\in \{0.8,1.2\}$; these represent low-moderate and high levels within the recommended range) and on the SingleCellExperiment object of raw and normalized counts with SC3 [2]. We also provided [3] with only the marker genes used by CellAssign (SC3 failed to run when provided with this number of genes). Inferred clusters

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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.

**9.4.5 SCINA—**SCINA was run on normalized logcounts using the same marker gene matrix used for CellAssign. As above, SCINA was run with rm_overlap = 0 and allow_unknown = 1, with 10 times the default number of maximum iterations (1000).

**9.4.6 Evaluation—**Accuracy and cell type-level F1 score were computed to evaluate clustering performance. The cell type-level F1 score is defined as the arithmetic mean of F1 scores computed for each cell type separately.

### 9.5 High-grade serous ovarian cancer

**9.5.1 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 VGH/UBC Anatomical Pathology specimen handling procedures. Tissues were dissociated at low temperature [40] using a modified protocol (O'Flanagan et al., manuscript in preparation). Briefly, after finely chopping and weighing in a cell culture dish, tissue was transferred into a gentleMACS C tube, and 1mL of 10 mg/mL Bacillus lichenformis protease (Creative Enzymes NATE-0633) was added to each 25 mg of tissue. The resulting solution was incubated and mechanically disrupted at 6○C using the Miltenyi Biotec MACS Separator (programs h_tumour_01, h_tumour_02, h_tumour_03) for 1 hour. Following dissociation, cells were assessed for viability using the cell counter (5μL cells + 5μL trypan blue) under a microscope.

Samples were then diluted with cold HFN 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 (5μL cells + 5μL trypan blue) under a microscope, and kept on ice. Cells were spun down and the pellet resuspended in 100μL of Miltenyi Dead Cell Removal MicroBeads and incubated at room temperature for 15 minutes. Viable cell enrichment was performed using the positive selection column type MS with a MACS Separator.

**9.5.2 Library preparation and sequencing—**Single-cell RNA-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 (75bp paired end reads) using a modified 58bp R2 at the UBC Biomedical Research Centre.

**9.5.3 Processing and normalization of single-cell RNA-seq data—**Raw sequence files were processed with CellRanger v2.1.0. The resulting filtered count matrices were read into SingleCellExperiment objects. Outlier cells according to quality control parameters (≥ 3 median absolute deviations from the median) were filtered out using the scater R package. Additionally, cells with ≥ 20% mitochondrial UMIs or ≥ 50% ribosomal UMIs were removed (ovarian cancer cells can have higher mitochondrial percentages than

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

other cell types, as in [41]). Size factors were computed using quickCluster and computeSumFactors from the scran R package. Following this, data normalization was performed using scater normalize. Principal components analysis was performed on the resultant normalized logcounts for the top 1000 most variable genes. The first 50 PCs were used as input for UMAP.

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

Cell cycle scores were computed with cyclone from the scran package [39, 42].

#### 9.5.4 CellAssign

- B cells: VIM[c], MS4A1[c], CD79A[c], PTPRC[c], CD19[c], BANK1 [43]
- 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 [44], WFDC2 [44]
- Ovarian stromal cells: VIM[c], MUM1L1 [45], FOXL2 [45], ARX [45], DCN [44], TPT1 [50], RBP1 [50]
- Ovarian myofibroblast: VIM[c], MUM1L1 [45], FOXL2 [45], ARX [45], ACTA2[c], COL1A1[c], COL3A1[c], SERPINH1 [44]
- Vascular smooth muscle cells: VIM[c], ACTA2[c], MYH11[c], PLN [46], MYLK [c], MCAM [47], COL1A1[c], COL3A1[c], SERPINH1 [48]
- Endothelial cells: VIM[c], EMCN[c], CLEC14A [49], CDH5[c], PECAM1[c], VWF [c], MCAM [47], SERPINH1 [44]

c: canonical marker

The marker gene list described above and in Supplemental Table 2 was used for CellAssign [43–45]. 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 [44] (these were the top 3 genes upregulated in normal fibroblasts by log fold change where $Q < 0.05$). Ovarian stromal cells and myofibroblasts were identified based on expression of MUM1L1 and ARX, ovary-specific markers known to be expressed in stroma from bulk RNA-seq and immunohistochemistry [45] (Figure 3D, Supplemental Figure 5A), with myofibroblasts distinguished by higher expression of α-smooth muscle actin and various collagen genes [44] (Figure 3D, Supplemental Figure 5A). A group of cells expressing vascular smooth muscle markers α-smooth muscle actin, MYH11, and MCAM [47] was also identified with CellAssign (Supplemental Figure 5A). CellAssign was run with default parameters, and 5 random initializations.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**9.5.5 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 [2], using default parameters. For Seurat, resolutions of 0.8 and 1.2 were used (these represent lowmoderate 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).

**9.5.6 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, using default parameters with 10,000 permutations, and the hallmark pathway gene set [50]. Annotations for cell cycle-associated pathways (E2F targets, G2M checkpoint, and mitotic spindle), and immune-associated pathways (including interferon gamma response, interferon alpha response, coagulation, complement, IL6- JAK/STAT signalling, and allograft rejection) were taken from [50]. All reported Q values refer to Benjamini-Hochberg corrected P values for two-sided tests.

### 9.6 Follicular lymphoma

**9.6.1 Sample preparation—**Leftovers from clinical flowed samples were collected and frozen in fetal calf serum containing 10% DMSO. Cells were thawed and washed according to the steps outlined in the 10X Genomics Sample Preparation Protocol. Cells were stained with PI for viability and sorted in a BD FACSAria Fusion using a 85um nozzle. Sorted cells were collected in 0.5 ml of medium, centrifuged and diluted in 1X PBS with 0.04% bovine serum albumin.

**9.6.2 Library preparation and sequencing—**Cell concentration was determined by using a Countess II Automated Cell Counter and approximately 3,500 cells were loaded per well in the Single Cell 3’ Chip. Single cell libraries were prepared according to the Chromium Single Cell 3’Reagent Kits V2 User Guide. Single cell libraries from two samples were pooled and sequenced on one HiSeq 2500 125 base PET lane.

**9.6.3 Preprocessing and normalization of single-cell RNA-seq data—**Preprocessing steps for the follicular lymphoma data were identical to those for HGSC single-cell RNA-seq data, described in section 9.5.3, with the exception of different mitochondrial and ribosomal thresholds (cells with ≥ 10% mitochondrial UMIs or ≥ 60% ribosomal UMIs were removed).

**9.6.4 scvis analysis—**scvis train (v0.1.0) [28] was run with default settings on the top 50 PCs to produce a 2dimensional embedding of the follicular lymphoma data. Early stopping was added to scvis, so that the model would terminate after 3 successive iterations of no improvement (relative improvement in ELBO < $10^{-5}$). The resultant model was saved and used for mapping in section 9.7.4.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

#### 9.6.5 CellAssign

- 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 helper T cells: CD2[c], CD3D[c], CD3E[c], CD3G[c], TRAC[c], CD4[c], CXCR5[c], PDCD1[c], TNFRSF4 [43], ST8SIA1 [43], ICA1 [43], ICOS [43]
- Other CD4+ T cells: CD2[c], CD3D[c], CD3E[c], CD3G[c], TRAC[c], CD4[c], IL7R [43]

c: canonical marker

The marker gene list described above and in Supplemental Table 2 was used for CellAssign [43, 52]. CellAssign was run with default parameters, and 5 random initializations.

Patient was added as an additional covariate into the design matrix $\boldsymbol{X}$ (section 9.2.1). The best result according to marginal log-likelihood at convergence was kept. Optimization was considered converged after 3 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), NK cells (NCAM1 expression), and myeloid cells (CD14/CD16 and LYZ expression) was detected.

**9.6.6 Unsupervised clustering—**Unsupervised clustering of all cells was performed with Seurat and SC3 [2], using default parameters. For Seurat, resolutions of 0.8 and 1.2 were used (these represent lowmoderate 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).

**9.6.7 Classifying B cells—**B cells from CellAssign were further subclassified into ‘malignant’ or ‘nonmalignant’ groups according to expression of the constant region of the immunoglobulin light chain (kappa or lambda type) and the results of PCA. Seurat [3] (resolution = 0.8) was used to separate B cells into clusters, based on the top 50 PCs. Following this, the sole cluster associated with IGKC (immunoglobulin light chain kappa-type 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 the B cells.

**9.6.8 Differential expression between timepoints—**Differential expression analysis between timepoints for a given celltype and patient was performed using voom from the limma package for each patient and cell type separately, with timepoint 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 timepoint and any interactions.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**9.6.9 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). Overrepresentation of Reactome [32] pathways was assessed using the R package ReactomePA. Pathways were considered significantly overrepresented if the adjusted P-value $\leq 0.05$ (two-sided) and at least 2 genes from the pathway were present.

## 9.7 Reactive lymph node data

**9.7.1 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 DAPI and viable cells (DAPI negative) were sorted on a FACS ARIAIII or FACS Fusion (BD Biosciences) instrument.

**9.7.2 Library preparation and sequencing—**Approximately 8,700 cells per sample were loaded into a Chromium Single Cell 3’ Chip kit v2 (PN-120236) and processed according to the Chromium Single Cell 3’Reagent kit v2 User Guide. Libraries were constructed using the Single 3’ Library and Gel Bead Kit v2 (PN-120237) and Chromium i7 Mulitiplex Kit v2 (PN-120236). Single cell libraries from two samples were pooled and sequenced on one HiSeq 2500 125 base PET lane.

**9.7.3 Preprocessing and normalization of single cell RNA-seq data—**Preprocessing steps for the reactive lymph node data were identical to those for follicular single-cell RNA-seq data, described in section 9.6.3.

**9.7.4 scvis analysis—**The identities of the top 1000 most variable genes and PCA loadings from follicular lymphoma data analysis were used to compute a 50-dimensional embedding for the reactive lymph node data. Following this, the resultant 50 PCs were provided as input to scvis map [28], using the model trained in section 9.6.4 and default settings.

## 9.8 General statistical methods

On all boxplots, whiskers denote data within 1.5 times the interquartile range of the upper and lower quartiles. Where plotted over boxplots, points were horizontally, but not vertically jittered. Correlations were calculated using the cor function in the R statistical language (version 3.5.0).

## 9.9 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. Until the data is uploaded, it will be available from the authors upon request. Count matrices will be made

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

available on Zenodo before publication, and until then are also available from the authors upon request.

### 9.10 Materials availability

Further information and requests for resources and reagents should be directed to and will be fulfilled by Sohrab P. Shah (shahs3@mskcc.org). Limited quantities of the HGSC and follicular lymphoma patient tissue and cell suspensions used to generate single cell RNA-seq data are available.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

We thank V Svensson for his feedback on this manuscript. We also thank WW Wasserman, BH Nelson, PT 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), the Canadian Statistical Sciences Institute (CANSSI), and the UBC Data Science Institute. S.P.S. is a Susan G. Komen scholar. We acknowledge generous funding support provided by the BC Cancer Foundation. In addition, S.P.S. receives operating funds from the CIHR (grant FDN-143246), Terry Fox Research Institute (grants 1021 and 1061) and the Canadian Cancer Society (grant 705636). This work was supported by Cancer Research UK grant C31893/A25050 (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. CS is an Allen Distinguished Investigator supported by the Allen Frontiers Group.

## References

1. Consortium TM et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature (2018).
2. Kiselev VY et al. SC3: consensus clustering of single-cell RNA-seq data. Nature methods 14, 483 (2017). [PubMed: 28346451]
3. Butler A, Hoffman P, Smibert P, Papalexi E & Satija R Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology. ISSN: 1087–0156. doi:10.1038/nbt.4096. https://www.nature.com/articles/nbt.4096 (2018).
4. Zurauskiene J & Yau C pcaReduce: hierarchical clustering of single cell transcriptional profiles. BMC bioinformatics 17, 140 (2016). [PubMed: 27005807]
5. Levine JH et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell 162, 184–197 (2015). [PubMed: 26095251]
6. Duò A, Robinson MD & Soneson C A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F1000Research 7 (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. F1000Research 7, 1297 ISSN: 2046–1402 (Aug. 2018). [PubMed: 30228881]
8. Kiselev VY, Andrews TS & Hemberg M Challenges in unsupervised clustering of single-cell RNA-seq data. Nature Reviews Genetics, 1 ISSN: 1471–0056 (Jan. 2019).
9. Zhang X et al. CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Research. ISSN: 0305–1048. doi:10.1093/nar/gky900. http://www.ncbi.nlm.nih.gov/pubmed/30289549%20https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gky900/5115823 (Oct. 2018).
10. Kiselev VY, Yiu A & Hemberg M scmap: projection of single-cell RNA-seq data across data sets. Nature methods 15, 359 (2018). [PubMed: 29608555]

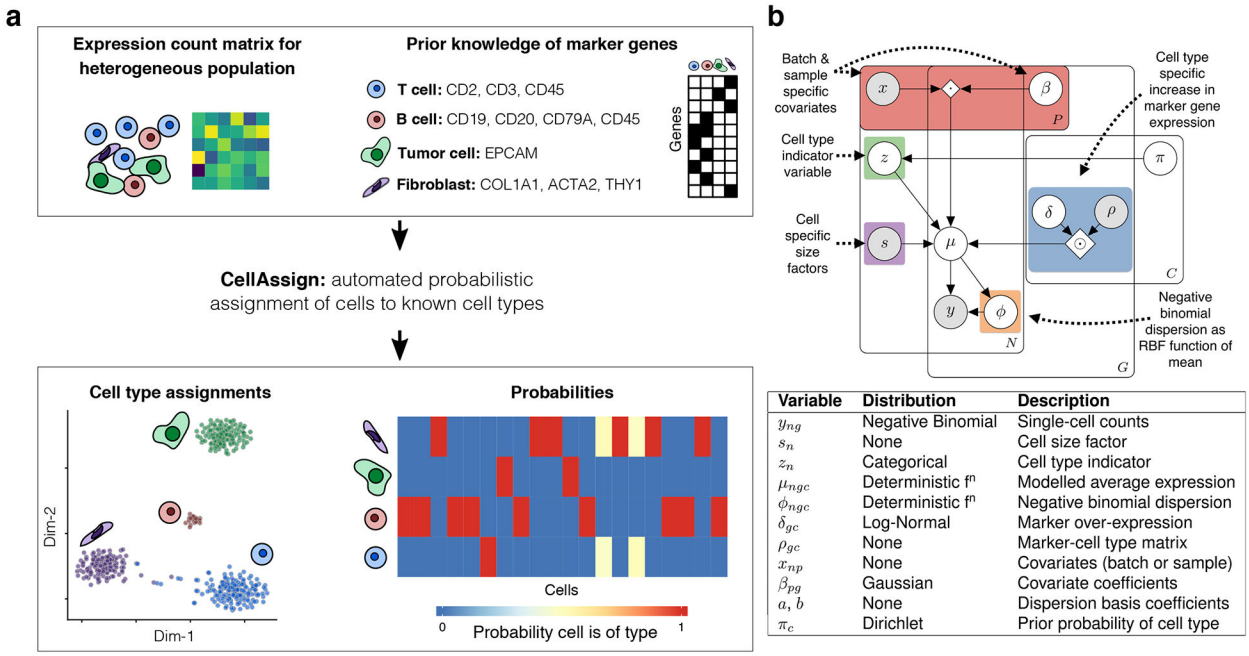
Author Manuscript Author Manuscript Author Manuscript Author Manuscript


11. Hicks SC, Townes FW, Teng M & Irizarry RA Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics (2017).
12. Zhang X et al. CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic acids research 47, D721–D728 (2018).
13. Koh PW 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). [PubMed: 27996962]
14. Tian L et al. scRNA-seq mixology: towards better benchmarking of single cell RNA- seq protocols and analysis methods. BioRxiv, 433102 (2018).
15. Zappia L, Phipson B & Oshlack A Splatter: simulation of single-cell RNA sequencing data. Genome Biology 18, 174 ISSN: 1474–760X (Dec. 2017). [PubMed: 28899397]
16. Levine JH et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell 162, 184–197. ISSN: 00928674 (7 2015). [PubMed: 26095251]
17. Ding J, Shah S & Condon A densityCut: an efficient and versatile topological approach for automatic clustering of biological data. Bioinformatics 32, 2567–2576. ISSN: 1367–4803 (Sept. 2016). [PubMed: 27153661]
18. Langfelder P, Zhang B & Horvath S Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 24, 719–720. ISSN: 1460– 2059 (Mar. 2008). [PubMed: 18024473]
19. Zheng GXY et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications 8, 14049 ISSN: 2041–1723 (Jan. 2017).
20. Zhang Z et al. SCINA: Semi-Supervised Analysis of Single Cells in silico BioRxiv (2019).
21. MacParland SA et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. Nature communications 9, 4383 (2018).
22. McInnes L & Healy J Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018).
23. Zhang AW et al. Interfaces of Malignant and Immunologic Clonal Dynamics in Ovarian Cancer. Cell 173, 1755–1769.e22. ISSN: 00928674 (6 2018). [PubMed: 29754820]
24. Kristiansen G et al. CD24 Is Expressed in Ovarian Cancer and Is a New Independent Prognostic Marker of Patient Survival. The American Journal of Pathology 161, 1215–1221. ISSN: 00029440 (Oct. 2002). [PubMed: 12368195]
25. Hylander B et al. Expression of Wilms tumor gene (WT1) in epithelial ovarian cancer. Gynecologic Oncology 101, 12–17. ISSN: 0090–8258 (Apr. 2006). [PubMed: 16263157]
26. Andor N et al. Single-cell RNA-Seq of lymphoma cancers reveals malignant B cell types and co-expression of T cell immune checkpoints. Blood, blood–2018–08–862292. ISSN: 1528–0020 (Dec. 2018).
27. Jefferis R & Lefranc M-P Human immunoglobulin allotypes: possible implications for immunogenicity. mAbs 1, 332–8. ISSN: 1942–0870 (2009). [PubMed: 20073133]
28. Ding J, Condon A & Shah SP Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. Nature Communications 9, 2002 ISSN: 2041–1723 (Dec. 2018).
29. 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, 265–72. ISSN: 0006–4971 (Jan. 1996). [PubMed: 8547651]
30. Gu K et al. t(14;18)-negative follicular lymphomas are associated with a high frequency of BCL6 rearrangement at the alternative breakpoint region. Modern Pathology 22, 1251–1257. ISSN: 0893–3952 (Sept. 2009). [PubMed: 19465899]
31. Hatzi K & Melnick A Breaking bad in the germinal center: how deregulation of BCL6 contributes to lymphomagenesis. Trends in molecular medicine 20, 343–52. ISSN: 1471–499X (6 2014). [PubMed: 24698494]
32. Fabregat A et al. The Reactome Pathway Knowledgebase. Nucleic Acids Research 46, D649–D655. ISSN: 13624962 (Jan. 2018). [PubMed: 29145629]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript


33. Freeman BE, Hammarlund E, Raué H-P & Slifka MK Regulation of innate CD8 + T-cell activation mediated by cytokines. doi:10.1073/pnas.1203543109. www.pnas.org/cgi/doi/10.1073/pnas.1203543109.
34. Hwang B, Lee JH & Bang D Single-cell RNA sequencing technologies and bioinformatics pipelines. Experimental & Molecular Medicine 50, 96 ISSN: 2092–6413 (Aug. 2018).
35. Eling Nils Richard Arianne C., R. S M. JC V. CA Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data. Cell Systems 7 (2018).
36. Kingma DP & Ba J Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).
37. Abadi Martin et al. {TensorFlow}: Large-Scale Machine Learning on Heteroge neous Systems 2015 https://www.tensorflow.org/.
38. Sinha D, Kumar A, Kumar H, Bandyopadhyay S & Sengupta D dropClust: efficient clustering of ultra-large scRNA-seq data. Nucleic Acids Research 46, e36–e36. ISSN: 0305–1048 (Apr. 2018). [PubMed: 29361178]
39. Lun AT, McCarthy DJ & Marioni JC A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research 5, 2122 ISSN: 2046–1402 (Oct. 2016). [PubMed: 27909575]
40. Adam M, Potter AS & Potter SS Psychrophilic proteases dramatically reduce single-cell RNA-seq artifacts: a molecular atlas of kidney development. Development (Cambridge, England) 144, 3625–3632. ISSN: 1477–9129 (2017).
41. Schelker M et al. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. Nature Communications 8, 2032 ISSN: 2041–1723 (Dec. 2017).
42. Scialdone A et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. Methods 85, 54–61 (2015). [PubMed: 26142758]
43. Newman AM et al. Robust enumeration of cell subsets from tissue expression profiles. Nature Methods 12, 453–457. ISSN: 1548–7091 (5 2015). [PubMed: 25822800]
44. Shih AJ et al. Identification of grade and origin specific cell populations in serous epithelial ovarian cancer by single cell RNA-seq. PLOS ONE 13 (ed Orsulic S) e0206785 ISSN: 1932–6203 (Nov. 2018). [PubMed: 30383866]
45. Uhlen M et al. A pathology atlas of the human cancer transcriptome. Science 357, eaan2507 ISSN: 0036–8075 (Aug. 2017). [PubMed: 28818916]
46. Perisic Matic L et al. Phenotypic Modulation of Smooth Muscle Cells in Atherosclerosis Is Associated With Downregulation of LMOD1, SYNPO2, PDLIM7, PLN, and SYNMArteriosclerosis, Thrombosis, and Vascular Biology 36, 1947–1961. ISSN: 1079–5642 (Sept. 2016).
47. Espagnolle N et al. CD146 expression on mesenchymal stem cells is associated with their vascular smooth muscle commitment. Journal of cellular and molecular medicine 18, 104–14. ISSN: 1582–4934 (Jan. 2014). [PubMed: 24188055]
48. Rocnik E, Saward L & Pickering JG HSP47 expression by smooth muscle cells is increased during arterial development and lesion formation and is inhibited by fibrillar collagen. Arteriosclerosis, thrombosis, and vascular biology 21, 40–6. ISSN: 1524– 4636 (Jan. 2001).
49. Mura M et al. Identification and angiogenic role of the novel tumor endothelial marker CLEC14A. Oncogene 31, 293–305. ISSN: 0950–9232 (Jan. 2012). [PubMed: 21706054]
50. Liberzon A et al. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Systems 1, 417–425. ISSN: 24054712 (Dec. 2015). [PubMed: 26771021]
51. Payne D, Drinkwater S, Baretto R, Duddridge M & Browning MJ Expression of chemokine receptors CXCR4, CXCR5 and CCR7 on B and T lymphocytes from patients with primary antibody deficiency. Clinical and experimental immunology 156, 254–62. ISSN: 1365–2249 (5 2009). [PubMed: 19250276]
52. Deenick EK & Ma CS The regulation and role of T follicular helper cells in immunity. Immunology 134, 361–7. ISSN: 1365–2567 (Dec. 2011). [PubMed: 22043829]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript




| Variable | Distribution | Description |
|---|---|---|
| $y_{ng}$ | Negative Binomial | Single-cell counts |
| $s_n$ | None | Cell size factor |
| $z_n$ | Categorical | Cell type indicator |
| $\mu_{ngc}$ | Deterministic f$^n$ | Modelled average expression |
| $\phi_{ngc}$ | Deterministic f$^n$ | Negative binomial dispersion |
| $\delta_{gc}$ | Log-Normal | Marker over-expression |
| $\rho_{gc}$ | None | Marker-cell type matrix |
| $x_{np}$ | None | Covariates (batch or sample) |
| $\beta_{pg}$ | Gaussian | Covariate coefficients |
| $a, b$ | None | Dispersion basis coefficients |
| $\pi_c$ | Dirichlet | Prior probability of cell type |

**Figure 1.**

(**a**) Overview of CellAssign. CellAssign takes raw count data from a heterogeneous single-cell RNA-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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

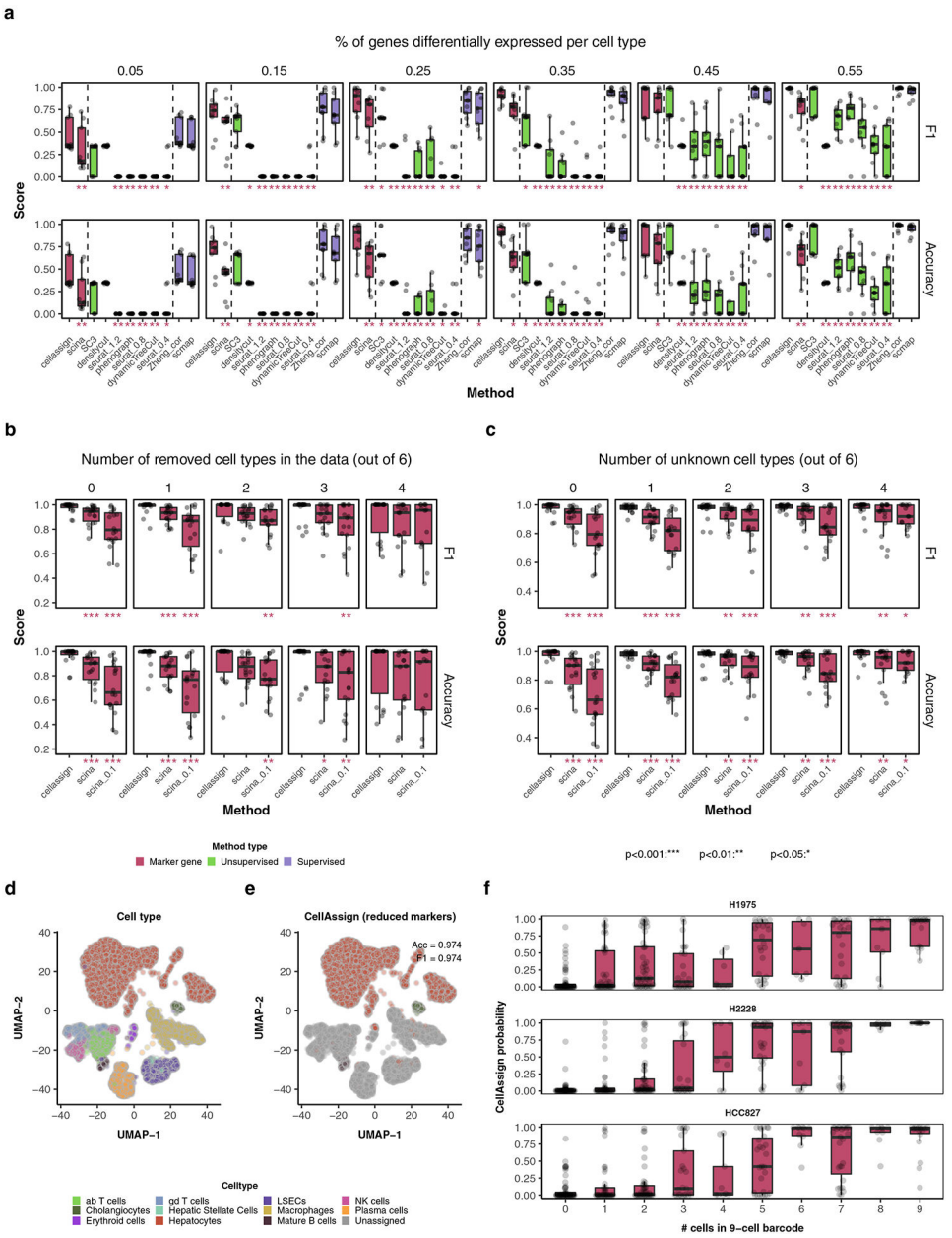



**Figure 2.**

Performance of CellAssign on simulated data. (**a**) Accuracy and cell-level F1 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 naïve CD4+ T cells (Methods). CellAssign was provided with a set of marker genes (Methods); all other methods were provided with all 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. Dotted lines separate marker-based, unsupervised, and supervised methods. (**b**) Accuracy and cell-level F1 score for CellAssign, SCINA (default parameters) and SCINA (sensitivity cutoff of 0.1) for simulated data from 6 cell types, where zero to 4 cell types were removed from the data (but kept in the marker

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

gene list). (**c**) Accuracy and cell-level F1 score for CellAssign, SCINA (default sensitivity cutoff) and SCINA (sensitivity cutoff of 0.1) for simulated data from 6 cell types, where zero to 4 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 [21]. (**e**) CellAssign MAP assignments for human liver data, where marker genes for only hepatocytes, cholangiocytes, and mature B cells from [21] were specified. (**f**) CellAssign probabilities for cell line mixture data from [14], where known proportions of 3 lung adenocarcinoma cell lines (H1975, H2228, HCC827) were mixed in 9-cell combinations. 30 bulk RNA-derived marker genes for each cell line were used (Supplementary Notes 2.7). Lower and upper hinges denote the 1st and 3rd quartiles on boxplots, with whiskers extending to the largest value less than 1.5 × the inter-quartile range.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

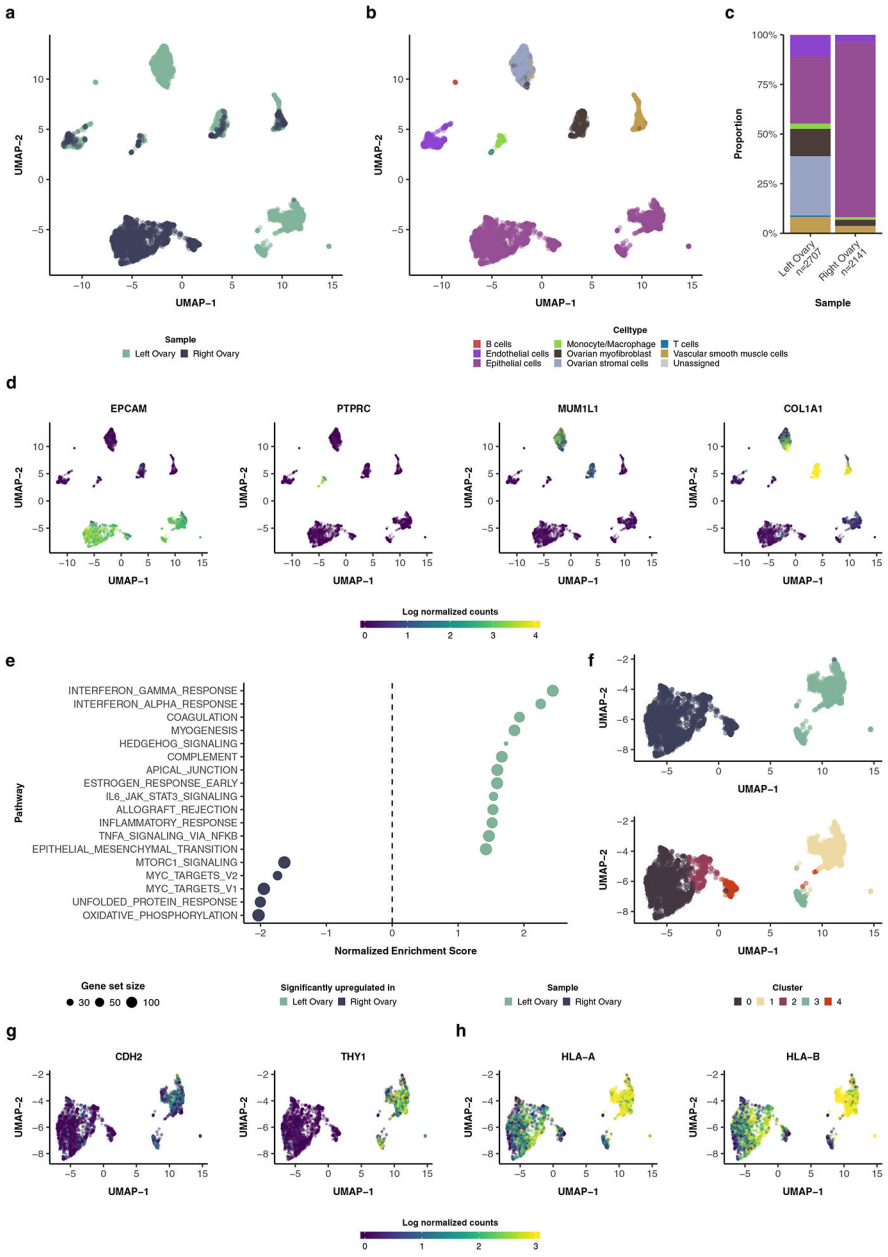



**Figure 3.**

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 collagenproducing fibroblasts and smooth muscle cells). Expression values were winsorized between 0 and 4. (**e**) Hallmark pathway enrichment results for left ovary vs. right ovary epithelial cells (Methods). (**f**) Unsupervised clustering of epithelial cells (Methods). (**g**) Expression (log normalized counts) of epithelial-mesenchymal transition (EMT) 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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

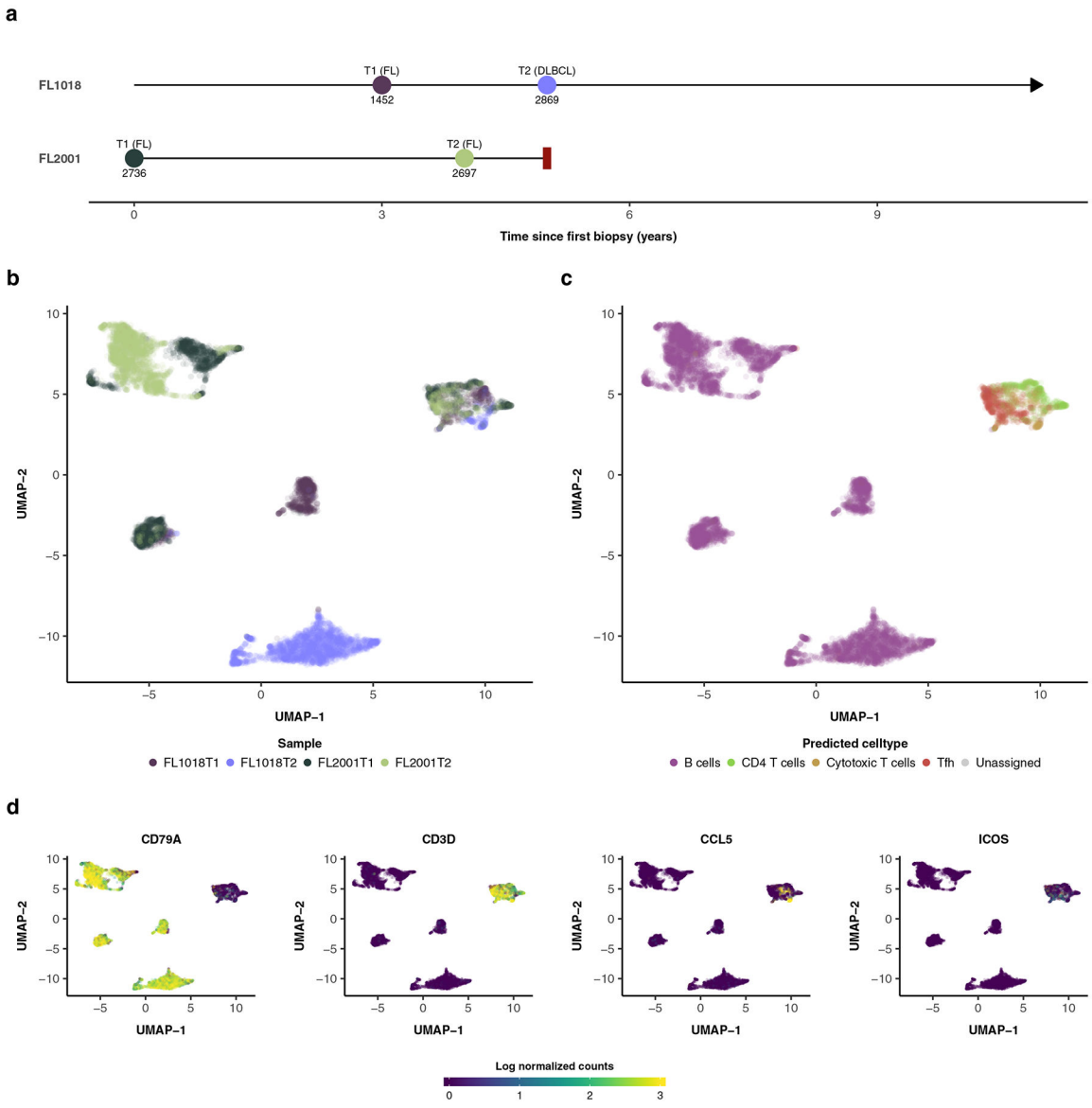



**Figure 4.**

CellAssign infers the composition of the follicular lymphoma microenvironment. (**a**) Sample collection times for FL1018 (transformed FL) and FL2001 (progressed FL). FL1018 is alive while FL2001 was lost to followup (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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

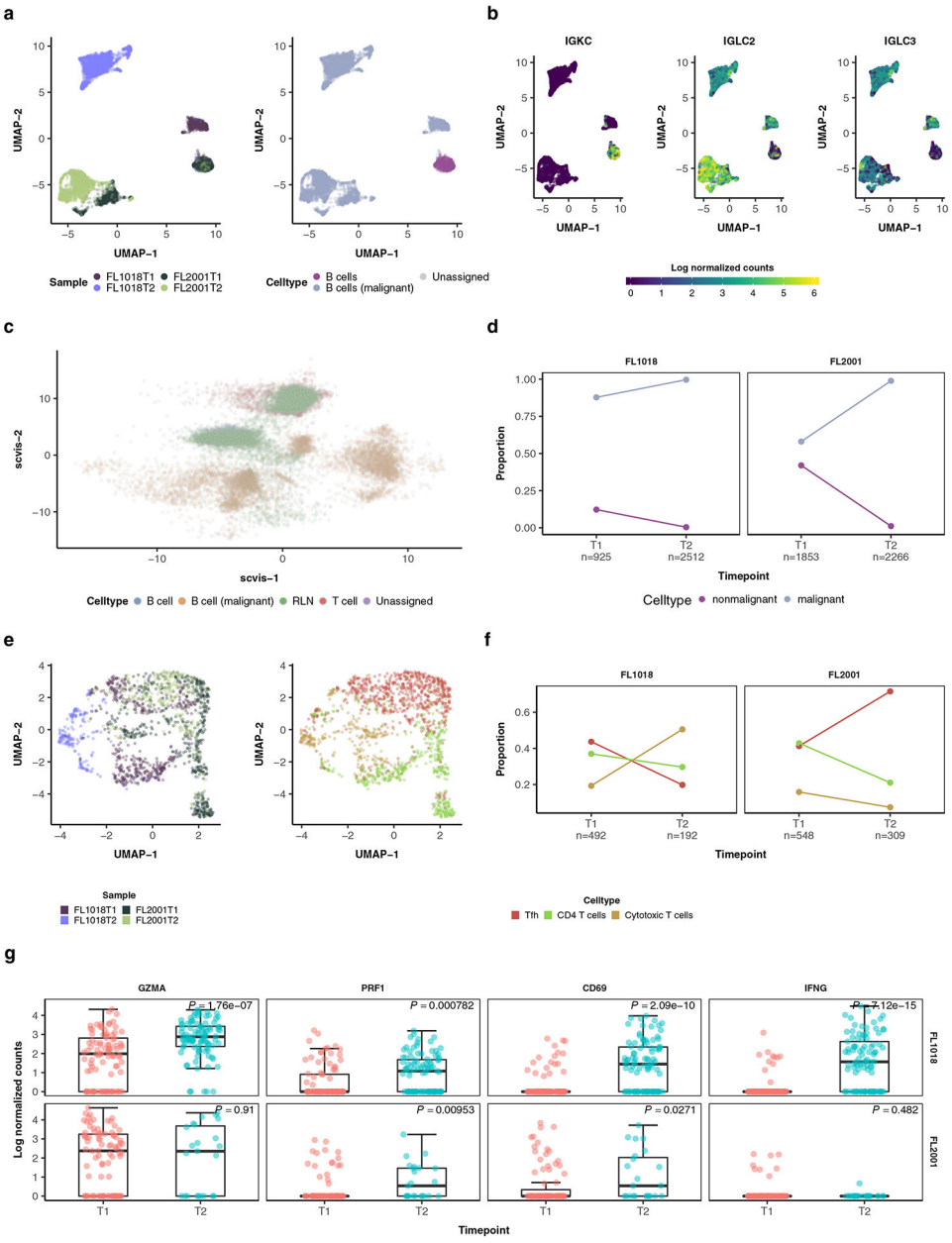



**Figure 5.**

Temporal changes in nonmalignant cells in the follicular lymphoma microenvironment. (**a**) Left: UMAP plot of CellAssign-inferred B cells, labeled by sample. Right: UMAP pot 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 single cell RNA-seq data of lymphocytes from reactive lymph nodes from healthy patients. The follicular lymphoma data was used to train the variational autoencoder and produce the two-dimensional embedding. Indicated cell types are B cell (nonmalignant B cell from FL), B cell (malignant) (malignant B cell from FL), T cell (T cell from FL), RLN (reactive lymph node cell). (**d**) Relative proportion of B cell subpopulations

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

over time, with total B cell counts indicated. (**e**) UMAP plots of FL T cells, labeled by sample and CellAssign-inferred celltype. (**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 the 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. Lower and upper hinges denote the 1st and 3rd quartiles on boxplots, with whiskers extending to the largest value less than 1.5 × the inter-quartile range.