Supervised Tumor Cell Subtype Identification via SCAN

Russell Kunes [1 2] **Siyu He** [3] **Yang Xiao** [3] **Simon Tavaré** [1 2] **David Knowles** [4]

Abstract

Clustering tumor cells based on single cell RNA sequencing data collected from a cohort of multiple patients is challenging because inter- patient variation dominates other sources of variation. To address this, we introduce *simultaneous clustering and normalization* (SCAN): a Bayesian clustering approach that removes bias in situations where cluster proportions vary across individuals. Normalization prior to clustering removes meaningful signal and creates artifacts. Our approach is novel in two respects. First, by *jointly* modeling the presence of clusters and inter-patient heterogeneity we are able to discover clusters that are present across individuals while taking into account the possibility that their cluster proportions may vary. Second, we introduce a natural method of incorporating quantitative data collected in parallel with scRNA-seq data (termed sSCAN), allowing us to encourage clusters that separate along a certain response variable.

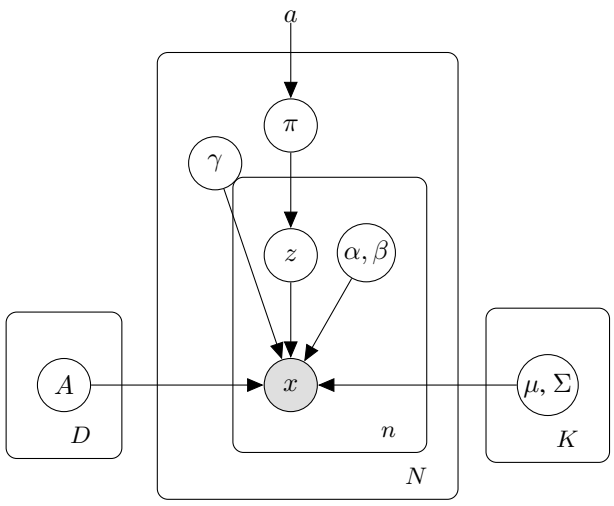
1. Introduction

Single cell RNA sequencing (scRNA-seq) allows for measuring counts of mRNA transcripts from thousands of single cells, permitting characterization of gene expression variation within cell populations and identification of clusters. A large body of literature focuses on extracting meaningful signal regarding cell type or activity state from scRNA-seq data in an unsupervised manner, generally via clustering methods or matrix factorization [(Kotliar et al., 2019) (Sun et al., 2019) (Kharchenko et al., 2014) (Lopez et al., 2018) (Farahbod & Pavlidis, 2019)]. There are a number of challenges with analysis of scRNA data such as dropouts and variation in library size. (Prabhakaran et al., 2016) demonstrates that ad hoc normalization techniques can lead to biases in downstream analysis.

In this work, we address the problem of cell subtype identification from scRNAseq data from the brain tumor stem-like cells (BTSCs) derived from 10 Glioblastoma Multiforme (GBM) patients, studied in an ex vivo model of the perivascular niche (PVN)(Xiao et al., 2019). The PVN is thought to play a crucial role in tumor cell migration, where microvessels might serve as tracks for cell movement. (Xiao et al., 2019) develop a microvasculature on-a-chip system as a model for the perivascular niche. The authors correlate colocalization to the PVN with previously defined (based on expression profile) tumor cell subtypes [(Patel et al., 2014) (Verhaak et al., 2010) (Brennan et al., 2013)]. The fundamental difficulty in identifying tumor cell subtypes via a clustering approach is illustrated in Figure 2 (left), a UMAP plot of single cell gene expression. The expression variation is dominated by individual heterogeneity and cells from the same individuals are completely separable in gene expression space. Reasonable clustering algorithms will tend to categorize cells by the individual they came from rather than identifying meaningful intra-patient variation and subtypes present across individuals. Prior work that addresses the problem of normalization across multiple experiments and data modalities [(Butler et al., 2018), (Welch et al., 2019)] is on the surface well suited to correcting this issue. However, none of these methods take into account the fact that some of this inter-patient heterogeneity may arise from differences in the proportions of cell subtypes between individuals and is thus relevant for downstream clustering. These approaches may remove variation due to differences in cell subtype proportion, thus confounding downstream clustering analysis. In order to ameliorate this issue, we present a modeling approach: a matrix factorization and mixture model (SCAN). The second problem is that we would like to leverage the colocalization coefficients (measured from the experimental assay, and defined roughly as average distance to microvasculature) to inform cluster centers. For this, we introduce a supervised version of the model (sSCAN) that incorporates supervisory signal from the colocalization coefficients.

[1]Irving Institute for Cancer Dynamics, Columbia University [2]Department of Statistics, Columbia University [3]Department of Biomedical Engineering, Columbia University [4]Department of Computer Science, Columbia University. Correspondence to: Russell Kunes <rk3064@columbia.edu>.

Presented at the ICML 2020 Workshop on Computational Biology (WCB).



for each individual **do**
  Draw cell state distribution $\pi \sim \text{Dirichlet}(a)$
  Draw baseline factors $\gamma \sim p(\gamma)$
  for each cell **do**
    Draw state $z_{ij} \sim \text{Categorical}(\pi)$
    Draw normalization parameters α, β
    Draw expression vector $\mathbf{x}_{ij}$

Figure 1. Left: SCAN graphical model. *Right:* SCAN generative process.

1.1. Notation

Let D denote the total number of genes in the study, N the number of individuals and n_i the number of cells from individual i. The logged mRNA counts for cell j of individual i is denoted $\mathbf{x}_{ij} \in \mathbb{R}^D$. As in (Prabhakaran et al., 2016), we define the log library size as $\sum_{d=1}^{D} \mathbf{x}_{ijd}$. The analysis is based on BISCUIT, which simultaneously models cell type and technical variation. Let $\mu_k \in \mathbb{R}^D$, $\alpha_{ij} \in \mathbb{R}$, $\Sigma_k \in \mathbb{R}^{D\times D}$, $\beta_{ij} \in \mathbb{R}$:

$$\begin{aligned}\pi &\sim \text{Dirichlet}(a)\\ z_{ij} &\sim \text{Categorical}(\pi)\\ \mathbf{x}_{ij} &\sim \mathcal{N}_D(\alpha_{ij}\mu_{z_{ij}}, \beta_{ij}\Sigma_{z_{ij}})\end{aligned}$$

BISCUIT is fundamentally a Gaussian mixture model with scalar cell level heterogeneity parameters to account for technical variation.

2. Simultaneous clustering and normalization

2.1. Introduction and Setup

We present a combined matrix factorization and mixture model (SCAN) for the purpose of normalizing out individual level heterogeneity. SCAN is a generative mixture model for scRNAseq data that allows for three types of shared variance components in datasets measured over multiple individuals and cells: that is, variation from individuals in the study, cell types, and each individual cell. One primary assumption of SCAN is that variation among individuals is restricted to a L dimensional linear subspace of $\mathbb{R}^D$. Variance parameters are sampled from an Inverse Gamma or Log-Normal distribution while parameters with support on the real line are sampled from a Normal distribution.

$$\begin{aligned}\pi_i &\sim \text{Dirichlet}(a)\\ \mu_{kd}, A_{dl}, \gamma_{il} &\sim \mathcal{N}(u, v^2)\\ \alpha_{ij} &\sim \log\mathcal{N}(t, s^2)\end{aligned}$$

$$\begin{aligned}\beta_{ij} &\sim \text{InverseGamma}(k, l)\\ z_{ij} &\sim \text{Categorical}(\pi_i)\\ \mathbf{x}_{ij} &\sim \mathcal{N}_D(\alpha_{ij}(\mu_{z_{ij}} + A\gamma_i), \beta_{ij}\Sigma_{z_i})\end{aligned}$$

with $A \in \mathbb{R}^{D\times L}$ and $\gamma_i \in \mathbb{R}^L$. The generative process for SCAN is summarized in Figure 1. Simulated data from SCAN is presented in Figure 6, showing that the model contains the desired structure for the problem of normalizing out interpatient heterogeneity. In our experiments, we take Σ_{z_i} to be diagonal.

2.2. Variational Inference

The posterior distribution of SCAN is intractable and approximate inference is necessary. Recall that in variational inference, we optimize a lower bound to the marginal data log likelihood $\log p(\mathbf{x})$ given by:

$$\mathcal{L}(\nu) = \mathbb{E}_{q_\nu}\Big\{\log p(\mathbf{x}, \mathbf{z})\Big\} + H(q_\nu)$$

where (in an abuse of notation) $\mathbf{z}$ denotes the set of all parameters of the model and expectations are taken with respect to $q_\nu(\mathbf{z})$. In this case q_ν represents an approximating family of distributions indexed by parameters ν (termed the variational parameters), and $H(q_\nu)$ is the entropy of this variational distribution. This lower bound to the marginal log likelihood is referred to as the evidence lower bound (ELBO). Remark that the difference $\log p(x) - \mathcal{L}(\nu)$ is given by

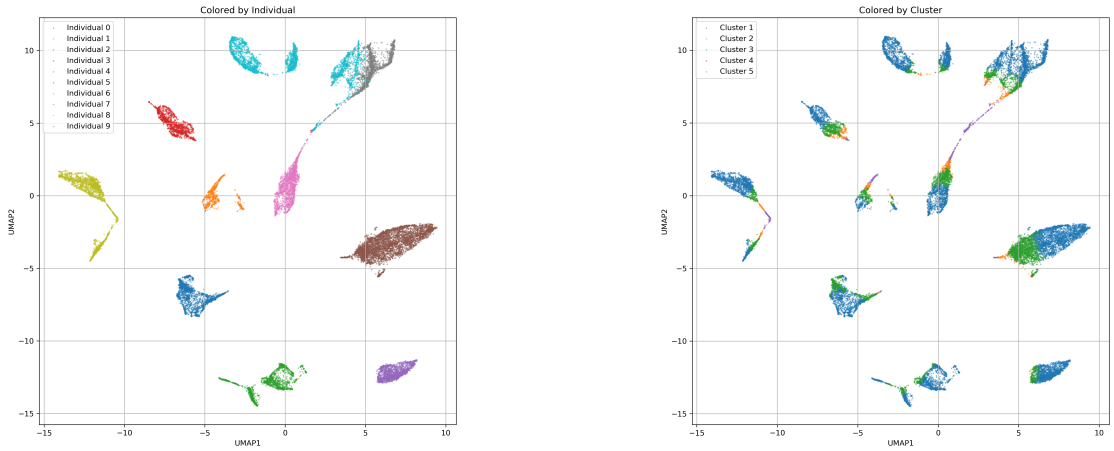


Figure 2. Left: UMAP plot of log expression counts, $\mathbf{x}$, labeled by individual. Remark that the variation is dominated by variation across individuals and the cells naturally form clusters based on the individuals. *Right:* The same data colored by clusters estimated by the SCAN procedure. Note that SCAN successfully finds clusters that are represented across individuals.

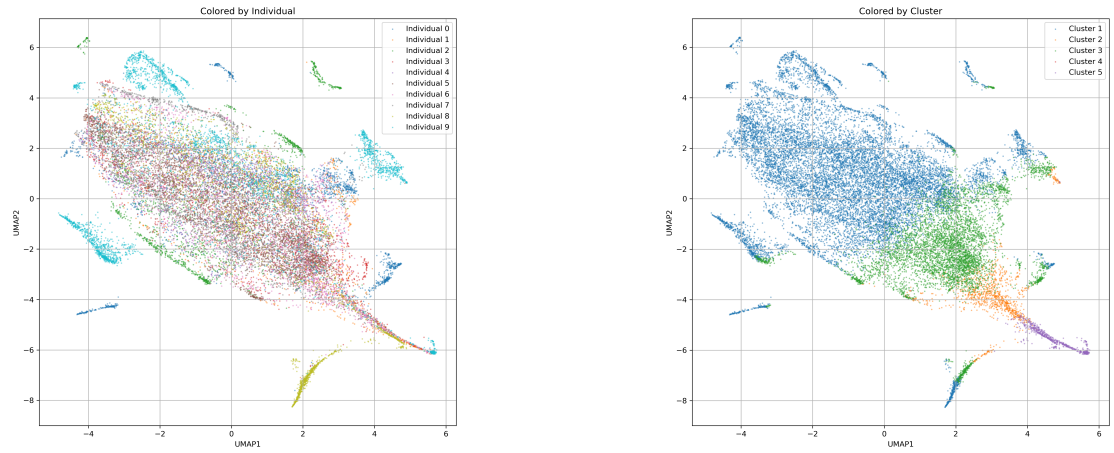


Figure 3. Left: UMAP plot of log expression counts, after normalization via the SCAN procedure, labeled by individual. *Right:* The same normalized data colored by clusters estimated by the SCAN procedure.

$\mathrm{KL}(q_\nu(\mathbf{z})||p(\mathbf{z}|\mathbf{x}))$, so that maximizing the ELBO is equivalent to minimizing the KL-divergence between the approximating family and the true posterior. We use a mean field approximating family (Blei et al., 2003), $q_\nu(\mathbf{z}) = \prod_l q_{\nu_l}(z_l)$, in which possible dependencies among coordinates of the variational posterior are ignored, and each $q_{\nu_l}(z_l)$ is chosen to be from the same family of distributions as it's prior. Under these assumptions, the ELBO for SCAN and mean field approximating family can be computed analytically. Thus, fast approximate inference can be achieved by stochastic gradient descent (SGD) with respect to the variational parameters without any further approximations (Hoffman et al., 2013). The results of SCAN on the dataset of (Xiao et al., 2019) are displayed in Figures 2 and 3.

2.3. Supervised Clustering via sSCAN

In order to encourage the estimated clusters to resemble subtypes that characterize migration properties of the tumor, we also develop a supervised version of this model (sSCAN) that incorporates supervisory signal given by the colocalization coefficient to the PVN. The supervised version of

λ	1.4002	-1.3662	0.0336
Top Genes	TMSB4X	CDKN1A	VDAC1
	NDUFB9	SNRPE	EIF3H
	SNRPE	SOX2	SOX2
	TMEM14C	TMEM14C	HMGB1
	NDUFB10	MYEOV2	TCEB2

Table 1. Top genes (highest value in $E_q[\mu_k]$ after normalizing each row) for a cluster with high λ, a cluster with low λ, and one with intermediate λ. High λ means more predictive of y.

SCAN (sSCAN) can be concisely summarized as follows:

$$\mathbf{x}_{ij} \sim \text{SCAN}$$
$$y_{im} \sim \text{GLM}(\lambda^\top \pi_i; \phi)$$

where y_{im} is the m'th measurement of the PVN colocalization coefficient (in (Xiao et al., 2019)) for individual i, and ϕ is the dispersion parameter of an exponential family. In other words, y is modeled as an exponential dispersion family with mean determined by the dot product of the cell state distribution π_i and a set of global regression coefficients λ. Importantly, the parameters of both parts of the model are jointly estimated, so as to encourage $\{\pi_i\}_{i=1}^N$ to be predictive of the observed colocalization coefficients. In practice, there is a tradeoff between modeling y well and finding meaningful clusters in the data. In order to modulate this tradeoff, the dispersion parameter ϕ is left as a tuning parameter that controls the relative weight of the y and $\mathbf{x}$ terms in the loss function. The results of sSCAN in a Gaussian regression model are presented in Table 1. We show the top genes (subject to normalization) of the cluster most associated with y (high λ), the cluster least associated with y (low λ) and one intermediate cluster. Examining Table 1, the top marker for the negative λ cluster is p21, a CDK inhibitor and major target of tumor suppresor p53. Furthermore, Thymosin Beta 4 (coded for by TMSB4X), an actin sequestering protein that plays a role in cell migration, is differentially expressed by the high λ cluster.

2.4. Simulation Study

In order to validate the model and inference algorithm, we ran a simulation study in order to confirm that model captures desirable properties and that the inference algorithm recovers the true parameters in the well specified case. Since we care about the cluster assignments z_{ij}, we quantify performance of the algorithm by the percentage of points correctly clustered. As a side note, remark that the parameters $\{\mu, A, \alpha, \beta, \Sigma\}$ are recovered up to proportionality since the model parameters are only identifiable up to proportionality, while z and π can only be recovered up to permutation. For a given configuration of model parameters and choice of K, N, n, D, L, we simulate from the model 100 times and run the inference algorithm assuming the correct values of K and L for 3 random restarts (the objective is subject to local optima) [Figure 5]. These experiments confirm a reasonable rate of recovery in the well specified setting.

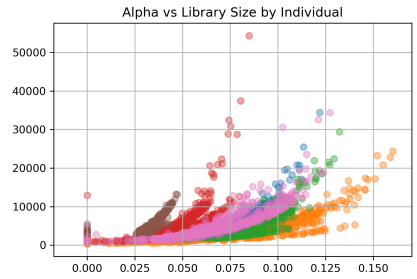


Figure 4. Left: Library size vs. estimated α, colored by individual.. This plot shows that the α parameter serves to normalize out technical variation in the library size of each cell. A separate relationship between α and the library size is estimated for each individual.

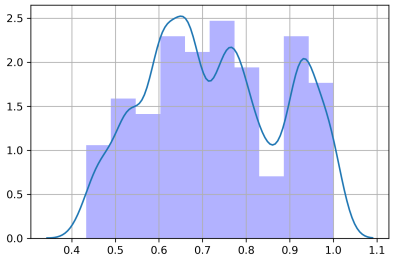


Figure 5. Classification accuracy of clustering assignments over a simulation study of 100 sets of simulated data. The parameters were $L = 1, D = 2, K = 3, N = 5, n = 500, \Sigma = I$

3. Conclusion

We have presented (supervised) *simultaneous clustering and normalization* (SCAN and sSCAN), a novel method for cluster analysis in scRNA-seq datasets where there is substantial heterogeneity across individuals. The model posits that cluster proportions can vary between individuals, allowing us to separate individual baseline expression levels from variation in cluster proportions. sSCAN allows for the incorporation of supervisory signal to inform cluster centers. Future work will include analyzing the results of SCAN on a variety of other scRNA-seq datasets.

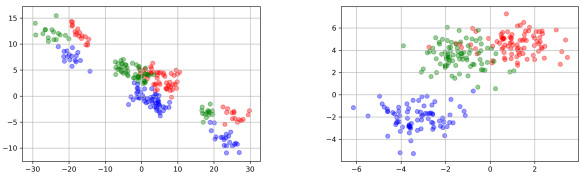


Figure 6. Left: Simulated data from the SCAN model. Note that the axis of variation of the individual means is $1D$ due to the matrix factorization assumption. In this case $D = 2$, $L = 1$, $K = 3$, and $N = 4$. *Right:* Samples from BISCUIT model under the same simulation settings.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

Brennan, C. W., Verhaak, R. G., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., Zheng, S., Chakravarty, D., Sanborn, J. Z., Berman, S. H., et al. The somatic genomic landscape of glioblastoma. *Cell*, 155(2):462–477, 2013.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

Farahbod, M. and Pavlidis, P. Untangling the effects of cellular composition on coexpression analysis. *bioRxiv*, 2019. doi: 10.1101/735951. URL https://www.biorxiv.org/content/early/2019/08/15/735951.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740, 2014.

Kotliar, D., Veres, A., Nagy, M. A., Tabrizi, S., Hodis, E., Melton, D. A., and Sabeti, P. C. Identifying gene expression programs of cell-type identity and cellular activity with single-cell rna-seq. *Elife*, 8, 2019.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., et al. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

Prabhakaran, S., Azizi, E., Carr, A., and Pe'er, D. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In *International Conference on Machine Learning*, pp. 1070–1079, 2016.

Sun, Z., Chen, L., Xin, H., Jiang, Y., Huang, Q., Cillo, A. R., Tabib, T., Kolls, J. K., Bruno, T. C., Lafyatis, R., et al. A bayesian mixture model for clustering droplet-based single-cell transcriptomic data from population studies. *Nature communications*, 10(1):1–10, 2019.

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell*, 17(1):98–110, 2010.

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E. Z. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887, 2019.

Xiao, Y., Kim, D., Dura, B., Zhang, K., Yan, R., Li, H., Han, E., Ip, J., Zou, P., Liu, J., et al. Ex vivo dynamics of human glioblastoma cells in a microvasculature-on-a-chip system correlates with tumor heterogeneity and subtypes. *Advanced Science*, 6(8):1801531, 2019.