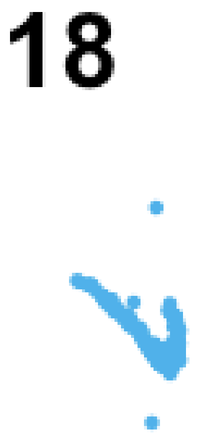
BarcodeBender for Mapping Nuclei in Slide-Tags

Jake Shapiro

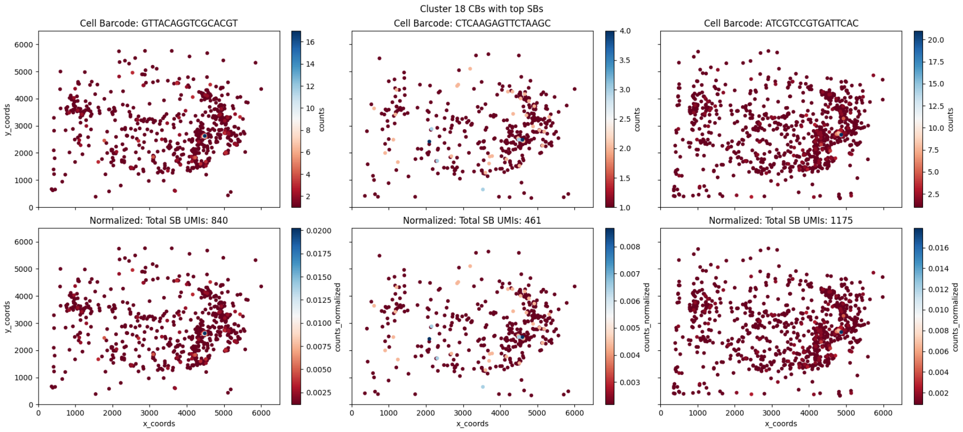
November 2023

1 Barcode Bender V0 - Inference on Cluster 18 Nuclei

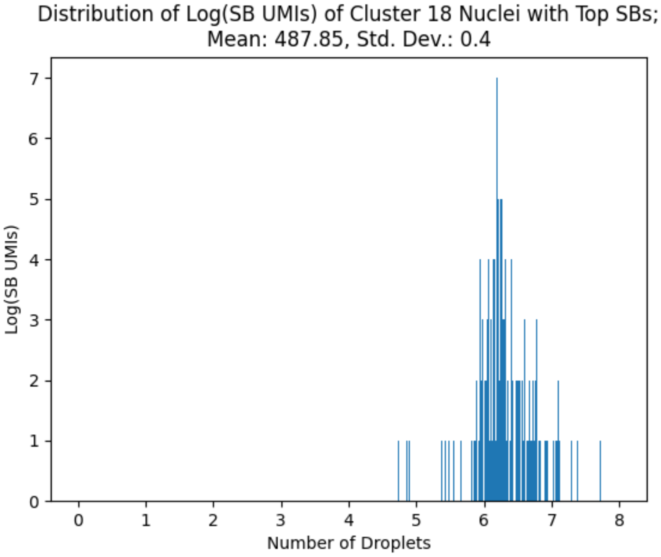
As a first run, I want to develop a model to infer spatial positions of nuclei that can be confidently mapped, which are of the same cell type, and with as few parameters as possible. For this, I will run inference on droplets from the deeply-sequenced gel-2 Slide-Tags run that were able to be mapped by DBSCAN, which were assigned to the same cell type cluster, and only use the top 10k SBs by total UMIs in those droplets. The DBSCAN mapping of these nuclei is shown below:



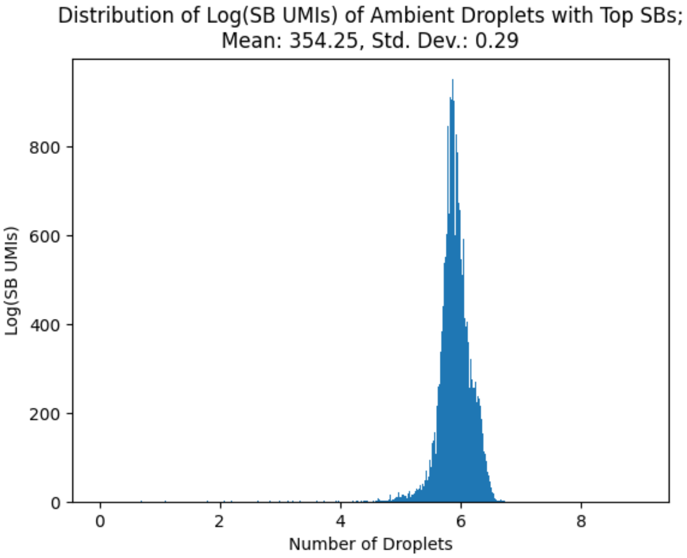
This nuclei-SB set provides several advantages. First, these are all droplets that were called by CellRanger, since the DBSCAN pipeline is run only on CellRanger-called droplets. Second, it is likely that if DBSCAN was able to map these nuclei, then there is a strong spatial signal that BarcodeBender can pick up on and (hopefully) use to improve upon DBSCAN. Third, by having the nuclei be of the same cell type, we remove cell type as a confounding variable. There also seems to be a defined boundary for the mapping of these nuclei. Finally, by only retaining the top 10k SBs, we reduce the number of parameters and the complexity of the model. A sampling of these CBs with raw, normalized, and background-removed plots is shown below and there seems to be a decently-strong spatial signal in each (darker blue spots):



Specifically, this dataset contains 261 nuclei and 9,497 SBs. The log-normalized distribution of SB UMIs in these nuclei is shown below:



For ambient droplets, I will use droplets of rank 40,000 - 100,000 by GEX UMIs, whose log-normalized SB UMI distribution is shown below:



These droplets almost certainly do not contain nuclei, based on GEX UMIs, but still have lots of SB UMIs. Thus, they likely faithfully capture the profile of ambient SBs in solution and can be used to estimate a prior for the ambient SB distribution.

The model I will use is as follows:

$c_{ji}^{\text{obs}} = c_{ji}^{\text{nuc}} + c_{ji}^{\text{noise}}$

$c_{ji}^{\text{nuc}} \sim \text{Poisson}[\epsilon_j^{\text{capt}} \epsilon_j^{\text{perm}} k_{ji}]$

$\epsilon_j^{\text{capt}} \sim \text{Beta}(\epsilon_\alpha^{\text{capt}}, \epsilon_\beta^{\text{capt}})$

$\epsilon_j^{\text{perm}} \sim \text{Beta}(\epsilon_\alpha^{\text{perm}}, \epsilon_\beta^{\text{perm}})$

$k_{ji} = \rho_i k(|\mathbf{r}_i - \mathbf{R}_j|; \sigma_i)$

$\rho_i \sim \text{LogNormal}(\rho_{loc}, \rho_{scale})$

$\sigma_i \sim \text{LogNormal}(\sigma_{loc}, \sigma_{scale})$

$c_{ji}^{\text{noise}} \sim \text{Poisson}[\epsilon_j^{capt} d_j^{\text{drop}} \chi_i^{\text{amb}}]$

$d_j^{\text{drop}} \sim \text{LogNormal}(d_\mu^{\text{drop}}, d_\sigma^{\text{drop}})$

where:

- $c_{ji}^{\text{obs/nuc/noise}}$ are the counts of SB i in droplet j due to observation, capture by the nucleus, or from noise, respectively. The observed counts due to nucleus capture and noise are Poisson-distributed because each SB has a small probability of being capture by the nucleus or droplet.

- $\epsilon_j^{\text{capt}} \in [0, 1]$ is the mRNA capture efficiency during RT of droplet j. It is distributed according to a Beta distribution with parameters $\epsilon_\alpha^{\text{capt}}, \epsilon_\beta^{\text{capt}}$.

- $\epsilon_j^{\text{perm}} \in [0, 1]$ is the permeability of nucleus j. It is distributed according to a Beta distribution with parameters $\epsilon_\alpha^{\text{perm}}, \epsilon_\beta^{\text{perm}}$.

- k_{ji} is the number of SBs from bead i at the location of nucleus j. It is sampled from the distribution $\rho_i k(|\mathbf{r}_i - \mathbf{R}_j|; \sigma_i)$ where ρ_i is the total number of SBs shed by bead i, $\mathbf{r}_i$ is the location of bead i and is known from *in-situ* sequencing, σ_i is the diffusion radius of SB i, and $\mathbf{R}_j$ is the location of nucleus j.

- $d_j^{\text{drop}} > 0$ is the "size" of droplet j (the larger, the more ambient it scoops). It is distributed according to a log-normal distribution with parameters d_μ^{drop} and d_σ^{drop}.

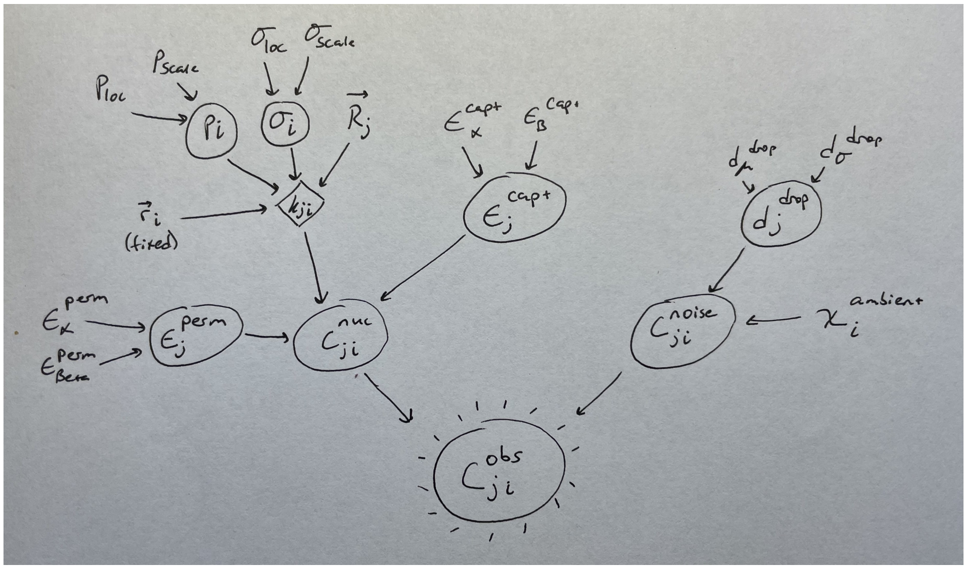
- $\chi_i^{\text{ambient}} \in [0, 1]$ is the relative abundance of SB i in ambient solution. It will be initialized from the normalized distribution of SBs in CBs of rank

40,000 - 100,000 by total GEX UMIs (almost certainly no nuclei, still in the ambient SB regime).

The learned parameters are:

- $\epsilon_{\alpha}^{\text{capt}}, \epsilon_{\beta}^{\text{capt}} > 0$
- $\epsilon_{\alpha}^{\text{perm}}, \epsilon_{\beta}^{\text{perm}} > 0$
- $\mathbf{R}_j \in \mathbb{R}^2$
- $\rho_{loc}, \rho_{scale}, \sigma_{loc}, \sigma_{scale} > 0$
- d_{μ}^{drop} and $d_{\sigma}^{\text{drop}} > 0$
- $\chi^{\text{ambient}}, \chi_i^{\text{ambient}} \in [0, 1], \sum_i \chi_i^{\text{ambient}} = 1$

All parameters except $\mathbf{r}_i$ (again, known from *in-situ* sequencing will be learned using variational inference. χ^{ambient} may also be treated as fixed because we know what the background looks like from the droplets in the ambient regime. A graphical representation of the generative model is below (hand-drawn, sorry):



σ_{loc}
σ_{scale}
ρ_{scale}
ρ_{loc}
ρ_i
σ_i
$\vec{R}_j$
ϵ_{α}^{capt}
ϵ_{β}^{capt}
d_{μ}^{drop}
d_{σ}^{drop}
k_{ji}
$\vec{r}_i$
(fixed)
ϵ_j^{capt}
d_j^{drop}
ϵ_{α}^{perm}
ϵ_{Beta}^{perm}
ϵ_j^{perm}
C_{ji}^{nuc}
C_{ji}^{noise}
$\chi_i^{ambient}$
C_{ji}^{obs}