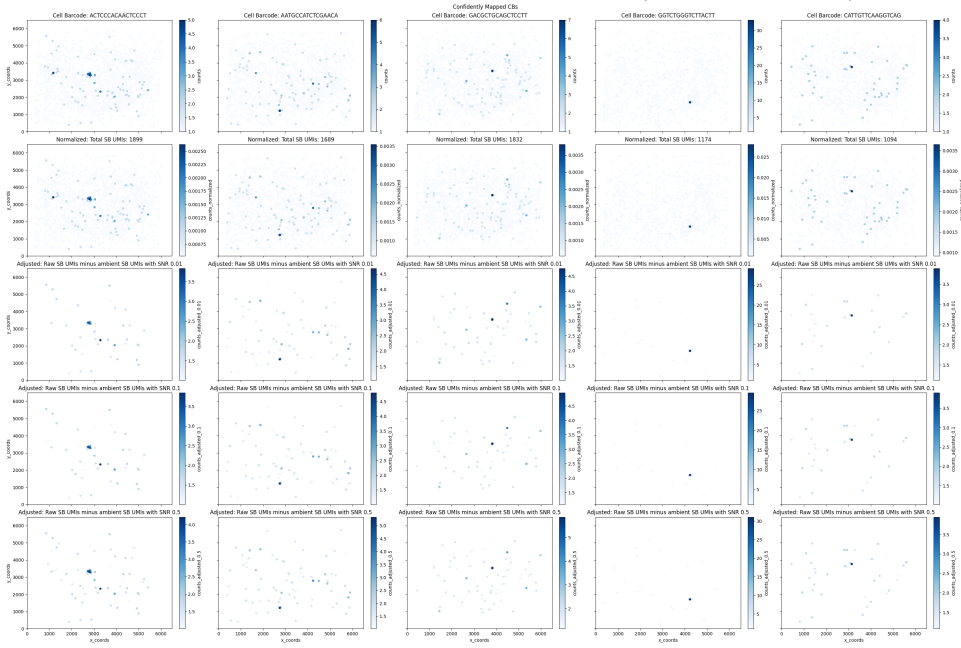
BarcodeBender for Mapping Nuclei in Slide-Tag

Jake Shapiro

November 2023

1 Barcode Bender V0 - Inference on Confidently-Mappable Nuclei

As a first run, I want to develop a model to infer spatial positions of nuclei that can be confidently mapped. 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. This nuclei set provides two 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. 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):



The model I will use is as follows:

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

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

$$\epsilon_j \sim \text{Gamma}(\epsilon_\alpha, \epsilon_\alpha)$$

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

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

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

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

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.
- $\epsilon_j \in [0, 1]$ is the mRNA capture efficiency during RT of droplet j. It is distributed according to a Gamma distribution parameterized by ϵ_α, the same as in CellBender. However, is Gamma(50, 50) $\in [0, 1]$?
- $d_j^{\text{nuc}} > 0$ is the "size" of nucleus j, or how well the nucleus takes up SBs (a proxy for both the size of the nucleus and its permeability). It is parameterized in the same way as in CellBender except that the parameters are global instead of local (all nuclei sizes are sampled from the same distribution and there is no amortization during learning. CellBender parameterized this distribution with a neural network).
- 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 parameterized in the same way as in CellBender.
- $\chi_i^{\text{ambient}} \in [0, 1]$ 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).

All parameters except $\mathbf{r}_i$ (again, known from *in-situ* sequencing will be learned using SVI. A graphical representation of the generative model is below (hand-drawn, sorry):