Modelling Single Cell RNA-Sequencing Data

Modelling A Single Homogenous Cell Population

The current view that transcription itself is stochastic and the probe used to sample transcripts is imprecise poses a problem for inference. The implications are that:

1. Even for a homogenous cell population the true transcript counts will vary from cell to cell despite an unchanged transcriptional process.
2. Differences in sample processing can introduce systematic variation where there is none.

The first implication has been addressed by empirical observations that the distribution of transcript counts is well-explained by a negative binomial, where— within a homogenous cell population —for gene g and cell c the observed number of counts is represented by:

$$Y_{g,c} \sim \text{NegBinom}(\mu_g, \phi_g)$$

with probability mass function:

$$P_{\text{NB}}(n \mid \mu, \phi) = \binom{n+\phi-1}{n} \left(\frac{\mu}{\mu+\phi}\right)^n \left(\frac{\phi}{\mu+\phi}\right)^\phi$$

and moments:

$$\mathbb{E}(Y) = \mu$$
$$\text{Var}(Y) = \mu + \frac{\mu^2}{\phi}$$

The second implication highlights that due to inefficiencies in RNA capture, sequencing, and mapping; it is unlikely that directly estimating μ_g will be equivalent to estimating an absolute measure of expression. Instead, we assume the true expression parameter η_g(i.e, the one we would have been able to estimate had we measured all transcripts within a cell) is scaled by a batch-specific "size factor" ζ_b— representing how well the sample was processed:

$$\begin{aligned} Y_{b,g,c} &\sim \text{NegBinom}(\mu_{b,g}, \phi_g) \\ \mu_{b,g} &= \zeta_b \cdot \eta_g \end{aligned}$$

With multiple batches $n_b > 1$, one might assume this would allow us to estimate all parameters as η_g is constant across batches(fixing g), while ζ_b is constant across genes(fixing b).

Unfortunately, without additional information, the model above is unidentifiable.

A Tangent on Models and Identifiability

- A **statistical model** $\mathcal{M}$ is a set(or *family*) of probability distributions on a given sample space $\mathcal{S}$. A parameterized statistical model is a set of parameterized probability distributions $\mathcal{M} := \{\pi_\theta : \theta \in \Theta\}$ where there exists a known mapping from the parameter space Θ to the model $\mathcal{M}$.
- Each element in $\mathcal{M}$ is also referred to as a **model configuration**.
- A parameterized statistical model $\mathcal{M}$ is **identifiable** if two equal model configurations will always have the same parameters, i.e, $\pi_\theta = \pi_{\theta'} \Longrightarrow \theta = \theta'$ for all $\pi_\theta, \pi_{\theta'} \in \mathcal{M}$.

To see this in an example, let $\mathcal{M} := \{\pi_\theta : \theta \in \Theta\}$ be a parameterized model where $\theta := (\zeta, \eta, \phi)$ and $\pi_\theta := P_{\text{NB}}(n \mid \zeta \cdot \eta, \phi)$.

Notice that if $\theta = (\zeta, \eta, \phi)$ and $\theta' = (\zeta', \eta', \phi')$ are two distinct parameters such that for an arbitrary constant $k \in \mathbb{R}^+$:

- $\zeta = k\zeta'$
- $\eta = \frac{1}{k}\eta'$,
- $\phi = \phi'$

Then the two model configurations(probability distributions) π_θ and $\pi_{\theta'}$ are equal despite distinct parameters:

$$\begin{aligned}
\pi_\theta &= P_{\text{NB}}(n \mid \zeta \cdot \eta, \phi) \\
&= P_{\text{NB}}\left(n \mid \left(\frac{k}{k}\right)\zeta \cdot \eta, \phi\right) \\
&= P_{\text{NB}}\left(n \mid k\zeta \cdot \frac{1}{k}\eta, \phi\right) \\
&= P_{\text{NB}}(n \mid \zeta' \cdot \eta', \phi') \\
&= \pi_{\theta'}
\end{aligned}$$

Implying the model is unidentifiable.

This tells us that estimating an absolute measure of expression is practically impossible with data from a typical single-cell experiment. Instead, we settle for the next best thing by substituting the original model with one that constrains the size factors(now denoted by the variable s) to sum to 1:

$$\vec{1} \cdot \vec{s} = \sum_{b=1}^{n_b} s_b = 1$$

Where n_b denotes the number of batches.

This provides with a new model that is 1) identifiable and 2) retains relative differences in magnitude between relative expression quantities(now denoted by the variable q):

$$\boxed{\begin{aligned}
Y_{b,g,c} &\sim \text{NegBinom}\left(\mu_{b,g}, \phi_g\right) \\
\mu_{b,g} &= s_b \cdot q_g \\
\vec{1} \cdot \vec{s} &= 1
\end{aligned}}$$

We prove both statements below:

The Constrained Model Is Identifiable

Let $\mathcal{M} := \{\pi_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ be a parameterized model where:

- $\boldsymbol{\theta} := \left(\vec{s}, \vec{q}, \vec{\phi}\right)$ with the constraint $\vec{1} \cdot \vec{s} = 1$.
- $\theta_{b,g} = \left(s_b, q_g, \phi_g\right)$.
- $\pi_{\theta_{b,g}} := P_{\mathrm{NB}}\left(n \mid s_b \cdot q_g, \phi_g\right)$.

Let $\boldsymbol{\theta} = \left(\vec{s}, \vec{q}, \vec{\phi}\right)$ and $\boldsymbol{\theta}' = \left(\vec{s}', \vec{q}', \vec{\phi}'\right)$ with equality $\pi_{\boldsymbol{\theta}} = \pi_{\boldsymbol{\theta}'}$ if and only if $\pi_{\theta_{b,g}} = \pi_{\theta'_{b,g}}$ for all b, g.

Note that if two negative binomial distributions are equal, then both their realized location parameters μ and their inverse overdispersion factors ϕ are equal. Since we are given $\pi_{\boldsymbol{\theta}} = \pi_{\boldsymbol{\theta}'}$, it is immediate that $\phi_g = \phi'_g$ for all g. Thus, we must prove that equality of the location parameters $\mu_{b,g} = \mu'_{b,g} = s_b \cdot q_g = s'_b \cdot q'_g$ implies $\left(s_b, q_g\right) = \left(s'_b, q'_g\right)$ for all b, g.

The realization of the location parameter for all b, g can be rewritten as the following outerproduct $\mathbf{M} = \vec{s} \cdot \vec{q}^{\top}$ with the b-th row and g-th column storing $\mu_{b,g}$. Thus, equality of location parameters is equivalent to:

$$\begin{aligned} \mathbf{M} &= \mathbf{M}' \\ \vec{s} \cdot \vec{q}^{\top} &= \vec{s}' \cdot \vec{q}'^{\top} \\ \vec{1} \cdot \vec{s} \cdot \vec{q}^{\top} &= \vec{1} \cdot \vec{s}' \cdot \vec{q}'^{\top} \\ \vec{q}^{\top} &= \vec{q}'^{\top} \end{aligned}$$

Immediately implying $(\vec{s}, \vec{q}) = \left(\vec{s}', \vec{q}'\right)$.

The Constrained Model Preserves Relative Differences

Our assumed "true model" for the data generating process $\mathcal{M} := \{\pi_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ and its constrained counterpart $\mathcal{M}' := \{\pi_{\boldsymbol{\theta}'} : \boldsymbol{\theta}' \in \Theta'\}$ are essentially equivalent. In order for relative differences to be preserved, then for $\pi_{\boldsymbol{\theta}} \in \mathcal{M}, \pi_{\boldsymbol{\theta}'} \in \mathcal{M}'$ and $k \in \mathbb{R}$:

$$\pi_{\boldsymbol{\theta}} = \pi_{\boldsymbol{\theta}'} \Longrightarrow \eta_g = k q'_g \qquad \forall g \in \left\{1, ..., n_g\right\}$$

Let $\pi_{\boldsymbol{\theta}} \in \mathcal{M}, \pi_{\boldsymbol{\theta}'} \in \mathcal{M}'$ be arbitrary with:

- $\boldsymbol{\theta} := \left(\vec{\zeta}, \vec{\eta}, \vec{\phi}\right)$ and $\boldsymbol{\theta}' := \left(\vec{s}', \vec{q}', \vec{\phi}'\right)$ with the constraint $\vec{1} \cdot \vec{s} = 1$.
- $\pi_{\theta_{b,g}} := P_{\mathrm{NB}}\left(n \mid \zeta_b \cdot \eta_g, \phi_g\right)$ and $\pi_{\theta'_{b,g}} := P_{\mathrm{NB}}\left(n \mid s'_b \cdot q'_g, \phi'_g\right)$.
- $b \in \{1, ..., n_b\}$ and $g \in \left\{1, ..., n_g\right\}$.

Then $\pi_{\boldsymbol{\theta}} = \pi_{\boldsymbol{\theta}'}$ implies:

$$\begin{aligned} \pi_{\boldsymbol{\theta}} = \pi_{\boldsymbol{\theta}'} &\Longrightarrow P_{\mathrm{NB}}\left(n \mid \zeta_b \cdot \eta_g, \phi_g\right) = P_{\mathrm{NB}}\left(n \mid s'_b \cdot q'_g, \phi'_g\right) \quad \forall b, g \\ &\Longrightarrow \zeta_b \cdot \eta_g = s'_b \cdot q'_g \quad \forall b, g \\ &\Longrightarrow \vec{\zeta} \cdot \vec{\eta}^{\top} = \vec{s} \cdot \vec{q}^{\top} \\ &\Longrightarrow \vec{1} \cdot \vec{\zeta} \cdot \vec{\eta}^{\top} = \vec{1} \cdot \vec{s} \cdot \vec{q}^{\top} \\ &\Longrightarrow \vec{1} \cdot \vec{\zeta} \cdot \vec{\eta}^{\top} = \vec{q}^{\top} \\ &\Longrightarrow \vec{\eta}^{\top} = \frac{1}{\vec{1} \cdot \vec{\zeta}} \vec{q}^{\top} \end{aligned}$$

Therefore, there exists some constant $k = \frac{1}{\vec{1} \cdot \vec{\zeta}}$ such that $\pi_{\boldsymbol{\theta}} = \pi_{\boldsymbol{\theta}'} \Longrightarrow \eta_g = k q'_g \quad \forall g \in \left\{1, ..., n_g\right\}$.

Modelling Multiple Cell Populations From a Single Tissue

The natural extension to multiple cell populations is introduced by adding another index variable p to group observations by their annotated cell-type:

$$\begin{aligned} Y_{b,p,g,c} &\sim \text{NegBinom}(\mu_{b,p,g}, \phi_g) \\ \mu_{b,p,g} &= s_b \cdot q_{p,g} \\ \vec{1} \cdot \vec{s} &= 1 \end{aligned}$$

Note however that there is potential for this model to be unidentifiable if there exists a batch that by chance contains no cell populations shared with other batches. E.g, if you seek to compare expression for two populations that were processed separately, you cannot infer what differences are due to batch-effects and what differences are due to biology. Fortunately, this scenario is unlikely all samples are derived from the same tissue— it is expected that the same cell populations will repeatedly show up.

Modelling Multiple Cell Populations of a Single Tissue Across Donors

The extension to multiple donors is also introduced by adding another index variable d to group observations by the donor they were sampled from:

$$\begin{aligned} Y_{b,d,p,g,c} &\sim \text{NegBinom}(\mu_{b,d,p,g}, \phi_g) \\ \mu_{b,d,p,g} &= s_b \cdot q_{d,p,g} \\ \vec{1} \cdot \vec{s} &= 1 \end{aligned}$$

But an issue stems from the fact that individuals often differ in their biology in some way. Differences in genetics and environmental exposures inevitably lead to changes in expression even when looking at the same cell population. I.e, for two donors d_1, d_2 and a cell population p: $\eta_{d_1,p,g}$ is not guarenteed to be equal to $\eta_{d_2,p,g}$ for all genes. This complicates inference as samples from different donors are also often processsed in different batches, so we run into the same issue described above when two populations are processed separately— we cannot distinguish between biology and batch-effect.

We need insight on either the size factors or the gene expression to make our model identifiable again. The ideal method would be to add “molecular spikes”(artificial RNAs with known sequence and fixed concentration) into each batch prior to sequencing(anecdotally the major cause behind batch-effects), and estimate a common intercept α that is fixed across batches for the number of reads S_b mapping to the spike-in sequence:

$$\begin{aligned} Y_{b,d,p,g,c} &\sim \text{NegBinom}(\mu_{b,d,p,g}, \phi_g) \\ \mu_{b,d,p,g} &= s_b \cdot q_{d,p,g} \\ \vec{1} \cdot \vec{s} &= 1 \\ \\ S_b &\sim \text{Poisson}(\mu_b) \\ \mu_b &= s_b \cdot \alpha \end{aligned}$$

The authors behind Tabula Sapiens however did not use molecular spikes. So we must either give up or make assumptions about gene expression. Fortunately, we have good reason to make assumptions about gene expression.

Despite differences in biology, it is plausible that the expression of a gene g is more similar among donors *within* a particular cell population p compared to *between* populations.

We adopt the assumption that for a fixed cell population p and gene g, expression quantities are log-normally distributed across donors with parameters $\left(\alpha_{p,g}, \lambda_{p,g}\right)$. It should be noted that the specific choice of distribution here is somewhat arbitrary; any distribution that provides control over the mean and variance of expression quantities could be used.

$$\begin{aligned}
q_{d,p,g} &\sim \text{Log-Normal}\left(\alpha_{p,g}, \lambda_{p,g}\tau_p\right) \\
\lambda_{p,g} &\sim \text{ Half-Cauchy}^{+}(0, 1)
\end{aligned}$$

The similarity of a cell population's gene expression across donors is represented by $\lambda_{p,g}$— approximately zero for genes whose variation is primarily due to batch-effects. Following [1], we model the shrinkage of this variance term by including a half-Cauchy prior:

Where τ_p represents the global shrinkage parameter.

Bells and Whistles

The bulk of the model has been designed above step-by-step, gradually increasing the complexity and working around issues that pop up. The two final additions are 1) a log-normal prior on the inverse overdispersion factors ϕ_g to partially pool estimates and 2) an exponential prior on the per-population global shrinkage parameter τ_p with $c^* = 1$ by default:

$$\begin{aligned}
Y_{b,d,p,g,c} &\sim \text{NegBinom}\left(\mu_{b,d,p,g}, \phi_g\right) \\
\mu_{b,d,p,g} &= s_b \cdot q_{d,p,g} \\
\vec{1} \cdot \vec{s} &= 1 \\
q_{d,p,g} &\sim \text{Log-Normal}\left(\alpha_{p,g}, \lambda_{p,g}\right) \\
\lambda_{p,g} &\sim \text{ Half-Cauchy}^{+}\left(0, \tau_p\right) \\
\tau_p &\sim \text{Exponential}(c^*) \\
\phi_g &\sim \text{Log-Normal}(\psi, \sigma)
\end{aligned}$$

Validation

Modelling a Single Population Across 2 Donors

Simulation parameters:

- Number of genes : 1000
- Number of genes with nonzero differences : 10
- Number of cells per-donor : 200

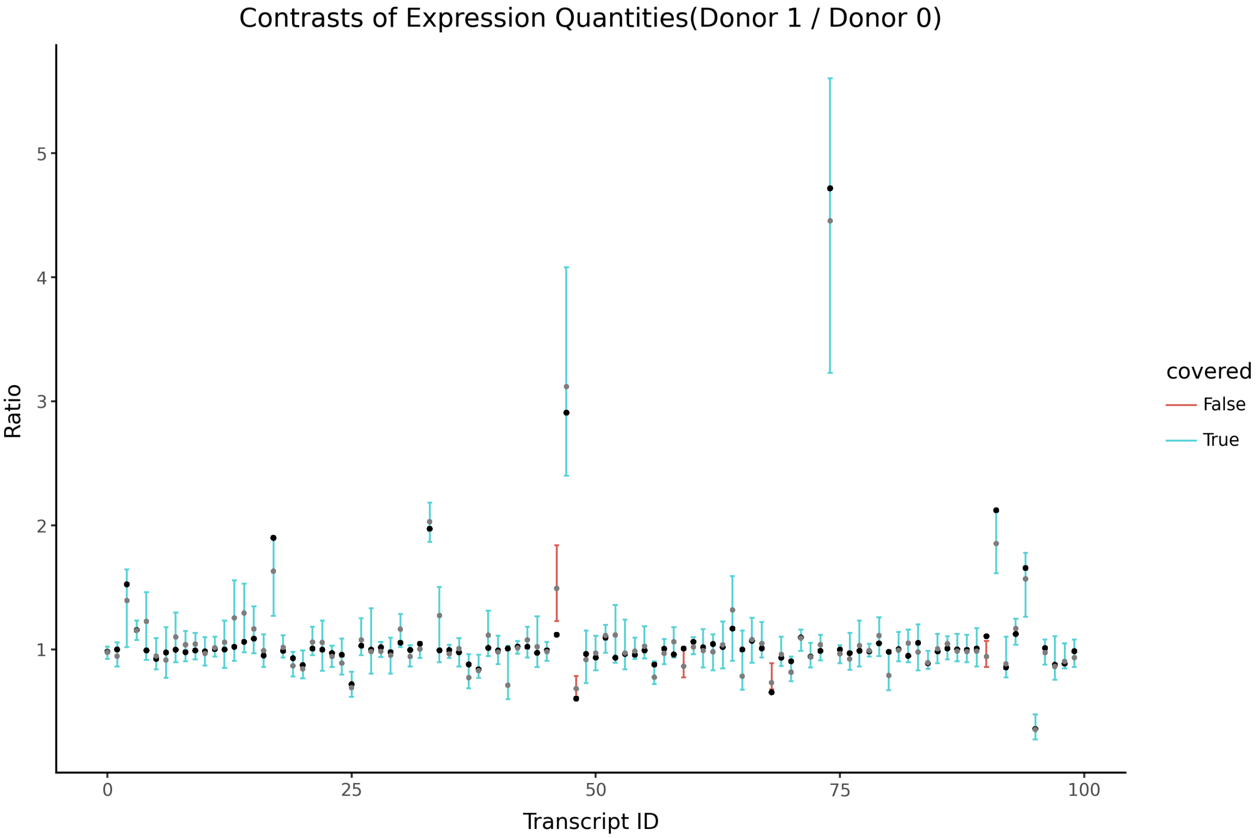

Figure 1: 95% credibility intervals for the ratio of relative expression quantities between donors `1` and `0`. Coloured dots represent the ground truth ratios(black) and size factor-adjusted sample mean ratios(grey). Only the first 100 transcripts are shown here.

Modelling Multiple Populations Across 2 Donors

Simulation parameters:

- Number of populations: 3
- Number of genes : 500
- Number of genes with nonzero differences : 10
- Number of cells per-population : 200

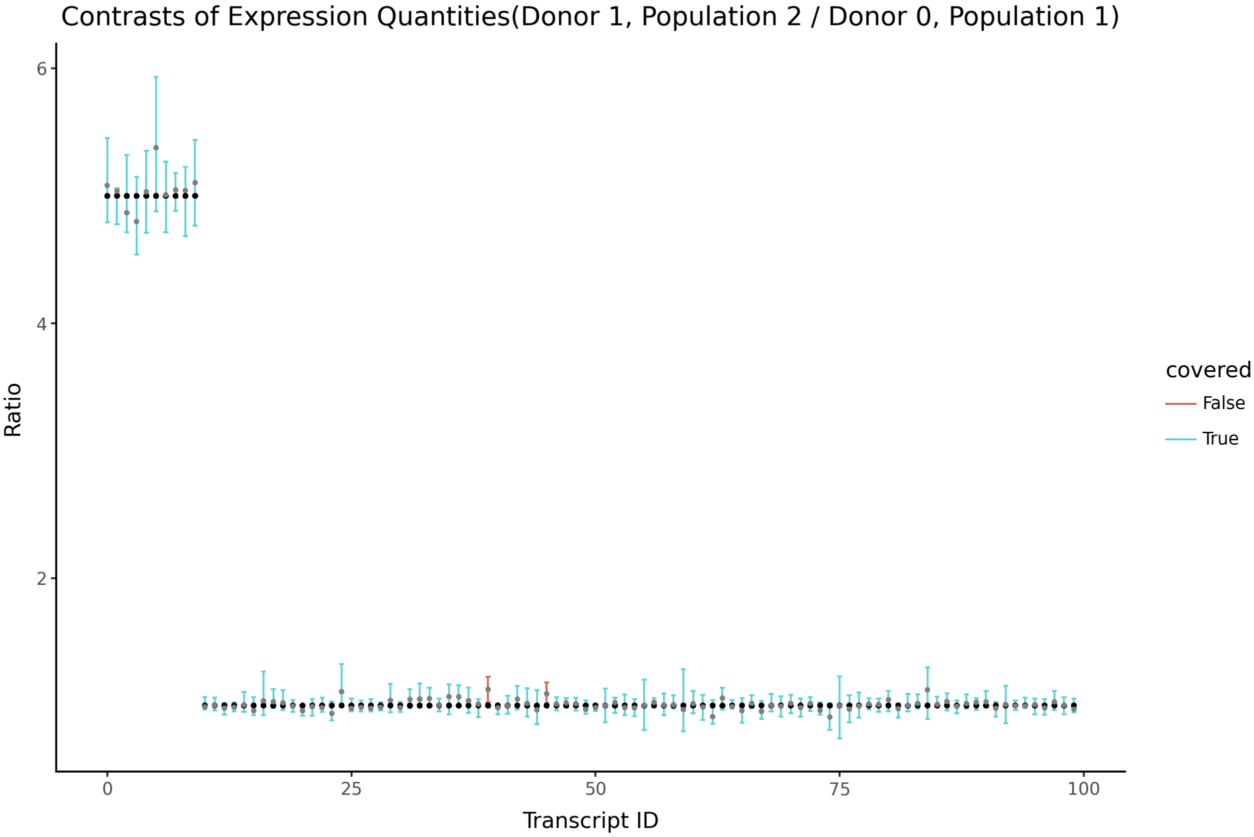

Figure 2: 95% credibility intervals for the ratio of relative expression quantities between populations 2 and 1. Coloured dots represent the ground truth ratios(black) and size factor-adjusted sample mean ratios(grey). Only the first 100 transcripts are shown here.

An Outlook on Previous Methods

There is no common method for the joint modelling of scRNA-seq data for differential expression analysis as I've described above. The closest that I know of is ZINBMM [2], which is essentially a finite mixture model with a zero-inflated negative binomial distributional assumption. Disregarding the zero-inflation assumption(see [3] for some commentary if you're interested), I do like the idea of fitting a finite mixture model on scRNA-seq data. Unfortunately, 1) there are often continuous trajectories that can mess up the identification of discrete cell populations, and 2) most people are already used to clustering being separate from differential expression analysis.

After clustering, the most common procedure for differential expression analysis is to pseudobulk(sum counts per-gene, per-batch, per-condition, etc) cell types of interest and compute however many pairwise comparisons with existing bulk-RNAseq methods(edgeR, DESeq2, etc). These methods use heuristics to calculate size factors, so I would be interested in seeing how well they match up to my proposed method.

I also realized a few things while writing this, it would be relatively easy to extend the above model and handle arbitrary covariates. The only reason donor(d) and population(p) are the main focus here is because the original goal was to infer cell type-specific expression quantities without molecular spikes. I think it would be interesting to try and write a general framework for jointly modelling RNAseq data, but the major appeal I see is that I'm not sure if pseudo-bulking is the right choice particularly for scRNA-seq(it's better to just treat them as repeated measurements/i.i.d observations rather than summing because I think the number of cells collected will bias you in some way, last I checked people are still summing) so that might be a knowledge gap to build a tool for or at least correct people.

On a final note, I find the idea of placing a prior on a subset of genes that are more likely to be conserved interesting. For example, if we know *a priori* that a certain set of genes are known to not differ between two or more groups(whether the groups be donors, populations, conditions, or otherwise), then this is useful information for estimating relative size factors.

Bibliography

[1] C. M. Carvalho, N. G. Polson, and J. G. Scott, "Handling Sparsity via the Horseshoe," in *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, PMLR, Apr. 2009, pp. 73–80. Accessed: Aug. 16, 2024. [Online]. Available: https://proceedings.mlr.press/v5/carvalho09a.html

[2] Y. Li, M. Wu, S. Ma, and M. Wu, "ZINBMM: a general mixture model for simultaneous clustering and gene selection using single-cell transcriptomic data," *Genome Biology*, vol. 24, no. 1, p. 208, Sep. 2023, doi: 10.1186/s13059-023-03046-0.

[3] V. Svensson, "Droplet scRNA-seq is not zero-inflated," *Nature Biotechnology*, vol. 38, no. 2, pp. 147–150, Feb. 2020, doi: 10.1038/s41587-019-0379-5.