Analytic Dropout Recovery via Pairwise Log-Ratio Bridges for RNA-Seq

Daniel A. Rosenthal
Department of Cell Biology
SUNY Downstate Health Sciences University
Faculty Mentor: Tade Souaiaia, PhD

Abstract

Robert F. Furchgott's Nobel Prize–winning work at SUNY Downstate showed that scrutiny of unexpected results can transform biological understanding. In that spirit, this project targets a common but overlooked laboratory artifact that distorts RNA sequencing experiments. Modern RNA-seq methods such as Patch-seq and laser capture microdissection analyze extremely small samples, where limited RNA causes many genes to go undetected. PCR amplifies detected genes but cannot recover missing ones, leaving zeros unchanged. When standard normalization rescales these samples, detected genes are forced to represent the transcriptome, producing false differences driven by technical imbalance rather than biology. To address this problem, I developed a measurement-level solution that operates on the observed gene counts themselves, rather than on normalized summaries. It compares shared expression patterns across samples and treats zeros as missing technical measurements rather than true absence, preventing detected genes from artificially representing the entire transcriptome. Mathematically, the method estimates sample scaling from pairwise log-ratios of shared detected genes under a multiplicative model, avoiding global normalization. By examining overlooked technical assumptions, this study reinforces that biological progress depends as much on questioning measurements as on generating data.

Analytic Dropout Recovery via Pairwise Log-Ratio Bridges

Model. Let genes $g \in \{1, \dots, G\}$ and samples $c \in \{1, \dots, C\}$. Observed counts follow

$$x_{cg} \;=\; a_c\, m_{cg}\, t_g\, \varepsilon_{cg},$$

with per-sample scale $a_c > 0$, dropout mask $m_{cg} \in \{0, 1\}$, shared profile $t_g > 0$, and multiplicative noise $\varepsilon_{cg} > 0$ with $\mathbb{E}[\log \varepsilon_{cg}] = 0$ and $\mathrm{Var}(\log \varepsilon_{cg}) = \sigma_g^2$. Define $y_{cg} = \log x_{cg} = \alpha_c + \beta_g + e_{cg}$ on observed entries, with $\alpha_c = \log a_c$, $\beta_g = \log t_g$, $e_{cg} = \log \varepsilon_{cg}$.

Pairwise offsets. For any two samples c, d and any shared observed gene h, $y_{ch} - y_{dh} = (\alpha_c - \alpha_d) + (e_{ch} - e_{dh})$. Let H_{cd} be the set of shared genes. The inverse-variance weighted mean of these log-ratios estimates $\Delta_{cd} \equiv \alpha_c - \alpha_d$:

$$\widehat{\Delta}_{cd} \;=\; \frac{\sum_{h \in H_{cd}} w_h\, (y_{ch} - y_{dh})}{\sum_{h \in H_{cd}} w_h}, \qquad w_h = \frac{1}{\mathrm{Var}(e_{ch} - e_{dh})} = \frac{1}{2\sigma_h^2}.$$

In practice $w_h = 1$ suffices; robust Huber/trimmed means are straightforward.

Theorem 1 (Exactness in the zero-noise limit)**.** *If $e_{cg} \equiv 0$ for all observed entries, then for any missing (c, g) such that there exists a neighbor d with $x_{dg} > 0$ and a shared gene h with $x_{ch}, x_{dh} > 0$,*

$$\widehat{x}_{cg} \;=\; x_{dg}\, \frac{x_{ch}}{x_{dh}}$$

recovers the true value $a_c t_g$ exactly. Moreover, using $\widehat{\Delta}_{cd} = \log(x_{ch}) - \log(x_{dh})$ yields $\log \widehat{x}_{cg} = \log x_{dg} + \widehat{\Delta}_{cd}$, identical to the 2×2 ratio.

Noisy case (MLE/BLUE). With $e_{cg} \sim \mathcal{N}(0, \sigma_g^2)$ and weights $w_g = 1/\sigma_g^2$, the generalized least-squares estimator of $\{\alpha_c, \beta_g\}$ solves

$$\min_{\{\alpha, \beta\}} \sum_{(c,g)\ \mathrm{obs}} w_g \left(y_{cg} - \alpha_c - \beta_g \right)^2.$$

Its normal equations are the alternating weighted means:

$$\alpha_c \leftarrow \frac{\sum_{g \in G_c} w_g (y_{cg} - \beta_g)}{\sum_{g \in G_c} w_g}, \quad \beta_g \leftarrow \frac{\sum_{c \in C_g} w_g (y_{cg} - \alpha_c)}{\sum_{c \in C_g} w_g},$$

with an identifiability constraint (e.g. $\sum_c \alpha_c = 0$). For any missing (c, g), $\widehat{x}_{cg} = \exp(\widehat{\alpha}_c + \widehat{\beta}_g)$.

Lemma 1 (Graph identifiability)**.** *Form a bipartite graph with sample nodes $\{c\}$ and gene nodes $\{g\}$, edges for observed pairs (c, g). If this graph is connected, then the GLS solution is unique up to a global additive constant ($\alpha \to \alpha + \delta$, $\beta \to \beta - \delta$), which does not affect $\widehat{x}_{cg}$ nor column-normalized data.*

Fallback for ultra-sparse regimes. If for a target (c, g) there is *no* neighbor d with both $x_{dg} > 0$ and at least one shared gene, we proceed as follows: (i) solve the GLS by alternating ridge least squares on $y_{cg} = \alpha_c + \beta_g$ over observed entries (small ℓ_2 penalties on α, β resolve degenerate components), then (ii) impute $\widehat{x}_{cg} = \exp(\widehat{\alpha}_c + \widehat{\beta}_g)$. If the bipartite graph has multiple connected components, the ridge yields component-wise solutions; any global scaling between components cancels after per-column normalization.

Normalization. After imputation, inferred totals $\hat{T}_c = \sum_g \widehat{x}_{cg}$ estimate a_c. The normalized matrix is $\widetilde{x}_{cg} = \widehat{x}_{cg}/\hat{T}_c$, which equals $t_g / \sum_{g'} t_{g'}$ under the model (no additional transform is required for correctness).