Sctransform Method

Definitions

- **UMI Counts**
  - Measure the number of unique original molecules in a sample.
- **Library Size**
  - Total number of sequencing reads obtained from a sample.
- **Sequencing Depth**
  - Number of times a particular region of the genome is sequenced.

Sctransform Approach

- Introduction of a new statistical approach for modeling, normalization, and variance stabilization of UMI count data.
- Proposes a generalized linear model (GLM) for each gene, with UMI counts as response and sequencing depth as explanatory variable.

Regularized Negative Binomial Regression

- Unconstrained NB models tend to overfit scRNA-seq data.
- Solution: Pool information to regularize parameters across genes with similar expression levels.

Model Equation

- Generalized Linear Model (GLM) with a log link function:

$$\log(E(x_i)) = \beta_0 + \beta_1 \log_{10}(m) \tag{1}$$

  - x_i: UMI (Unique Molecular Identifier) counts for gene i in a single cell.
  - m: Total molecule count or sequencing depth for each cell.
  - β_0: Intercept of the regression model, representing the baseline expression level.
  - β_1: Slope. .

Negative Binomial Distribution

- The UMI counts x_i are assumed to follow a Negative Binomial distribution.
- The mean (μ) and variance of the NB distribution are given by:

$$\mu = E(x_i) \tag{2}$$

$$\text{Variance} = \mu + \frac{\mu^2}{\theta} \tag{3}$$

- Here, θ is the dispersion parameter of the NB distribution. It captures the degree of overdispersion in the count data (variance greater than the mean).

Understanding the Statistical Approach

- **Regression Model for UMI Counts:** Utilizes a regression model to analyze UMI counts, correcting for sequencing depth differences and standardizing data.
- **Issue with Modeling Each Gene Separately:** Separate modeling for each gene can lead to overfitting, especially for low-abundance genes, resulting in high variance.
- **Overestimation of True Variance:** The high variance for low-abundance genes is likely overestimated, influenced more by cell-type heterogeneity than by variability in sequencing depth.
- **Regularization of Model Parameters:** To prevent overfitting and variance overestimation, regularization is applied to model parameters, including the dispersion parameter of the Negative Binomial distribution, by sharing information across genes.

Procedure Overview

Step 1: Fit independent regression models per gene.
Step 2: Each model parameter is regularized based on the relationship between parameter values and gene mean (Use kernel regression).
Step 3: Use regularized regression parameters to transform UMI counts into Pearson residuals.

$$z_{ij} = \frac{x_{ij} - \mu_{ij}}{\sigma_{ij}},$$
$$\mu_{ij} = \exp(\beta_{0i} + \beta_{1i} \log_{10} m_j),$$
$$\sigma_{ij} = \sqrt{\mu_{ij} + \frac{\mu_{ij}^2}{\theta_i}},$$

where z_{ij} is the Pearson residual of gene i in cell j, x_{ij} is the observed UMI count, μ_{ij} is the expected UMI count, and σ_{ij} is the expected standard deviation in the regularized NB regression model. Parameters β_{0i}, β_{1i}, and θ_i are linear model parameters after regularization.

Average Expression Calculation

Geometric Mean for Average Expression:

- To avoid the influence of outlier cells and respect the exponential nature of count distributions, the geometric mean is used.
- The average abundance or gene mean is defined as:

$$\text{Mean} = \exp\left(\text{amean}\left(\log(x + \delta)\right)\right) - \delta,$$

where:

- x is the vector of UMI counts of the gene.
- amean is the arithmetic mean.
- δ is a small fixed value to avoid log(0), set to 1 in this study.