Scran Method

scran Method: Steps

1. **Constructing Pools of Cells**
   - Group cells into pools to mitigate zero counts.
   - Pooling similar cells averages out dropouts.
2. **Summing Counts Across Pools**
   - Sum gene counts across cells in each pool.
   - Reduces impact of zero counts and technical noise.
3. **Calculating Pool-Based Size Factors**
   - Compute size factors for each pool.
   - Normalize summed counts for library size differences.

scran Method: Steps

4. **Deconvolution for Cell-Specific Factors**
   - Deconvolute pool-based factors to infer cell-specific factors.
   - Solve a linear system relating pooled and individual factors.
5. **Normalization of Individual Cell Counts**
   - Normalize original counts using cell-specific size factors.
   - Adjusts for library size differences at the cell level.

Scran Normalization: A Mathematical Overview

Y_{ij}: Count of gene i in cell j.
Expected Count: $E(Y_{ij}) = \theta_j \lambda_{i0}$

- θ_j: Cell-specific bias.
- λ_{i0}: Expected transcript count for gene i.

Adjusting the Count (Z_{ij}): $Z_{ij} = \frac{Y_{ij}}{t_j}$

- t_j: Adjustment factor for cell j.

Expected Adjusted Count ($E(Z_{ij})$): $E(Z_{ij}) = \frac{\theta_j \lambda_{i0}}{t_j}$

Pooling Cells

Consider a pool k

$$E(V_{ik}) = \lambda_{i0} \sum_{j \in S_k} \frac{\theta_j}{t_j} \tag{1}$$

where:

- $E(V_{ik})$: Expected summed of Z_{ij} expression value for gene i in pool k.
- S_k: Set of cells in pool k.

Reference Pseudo-Cell in Scran Normalization

- Averaged Reference Pseudo-Cell U_i, define U_i as the mean of Z_{ij} across all N cells in the entire dataset, with S_0 referring to the set of all cells in the data set.

$$E(U_i) = \lambda_{i0} N^{-1} \sum_{j \in S_0} \frac{\theta_j}{t_j} \tag{2}$$

Normalization Against Reference Pseudo-Cell

Normalization Process

- Each cell pool k is normalized against the reference pseudo-cell. For a non-DE gene i, define R_{ik} as the ratio of V_{ik} to U_i.

$$R_{ik} = \frac{V_{ik}}{U_i} \tag{3}$$

Expectation of R_{ik}

- The expectation of R_{ik} represents the true size factor for the pooled cells in S_k.

Calculation of Size Factor

$$E(R_{ik}) \approx \frac{E(V_{ik})}{E(U_i)} = \frac{\sum_{j \in S_k} \frac{\theta_j}{t_j}}{N^{-1} \sum_{j \in S_0} \frac{\theta_j}{t_j}} \tag{4}$$

Simplifying the expectation, we get:

$$E(R_{ik}) = \frac{\sum_{j \in S_k} \frac{\theta_j}{t_j}}{C} \tag{5}$$

- The approximation assumes that the variance of U_i is small, which is valid for datasets with hundreds of cells.
- C is a constant that does not depend on the gene, cell, or S_k.

Estimation of Pool-Based Size Factor

- Denote the realizations of Y_{ij}, V_{ik}, U_i, and R_{ik} as y_{ij}, v_{ik}, u_i, and r_{ik}, respectively.
- The pool-based size factor $E(R_{ik})$ is estimated by taking a robust average (e.g., the median) of r_{ik} across all genes.

Calculating r_{ik}

- **Observed Values:**
  - y_{ij}: Observed count of gene i in cell j.
  - v_{ik}: Observed sum of adjusted expression values for gene i across all cells in pool k.
  - u_i: Observed mean of adjusted expression values for gene i across all cells in the dataset.
- **Calculating r_{ik}:**
  - Calculated as $r_{ik} = \frac{v_{ik}}{u_i}$.

Using Estimates of $E(R_{ik})$ and Setting Up Linear Equations

- Estimates of $E(R_{ik})$ are derived from various cell pools.
- These estimates are used to estimate θ_j for each cell.

Linear Equation Formation

- For each cell pool k, linear equations are formed using the estimates of $E(R_{ik})$.

$$E(R_{ik}) = \frac{\sum_{j \in S_k} \frac{\theta_j}{t_j}}{C} \tag{6}$$

- The process is repeated with different cell pools.
- This leads to a system of linear equations.

Solving the System and Final Estimation

Least-Squares Method

- The system is solved using least-squares methods.
- This provides estimates of $\frac{\theta_j}{t_j}$ for all cells.

Deconvolution and Estimating θ_j

- The process represents deconvolution of cell pool factors to individual cell factors.
- By multiplying the estimated $\frac{\theta_j}{t_j}$ by t_j, an estimate of θ_j is obtained for each cell.

Constructing the Linear System by Selecting Cell Pools

- **Grouping Cells by Library Size:**
  - Cells are ordered by total counts and partitioned into odd and even groups.
- **Arranging Cells in a Ring:**
  - Cells are arranged in a ring, with odd cells on one side and even cells on the other.
  - Starts with largest libraries at 12 o'clock, moving clockwise to smallest at 6 o'clock, then through odd cells.
- **Using a Sliding Window:**
  - A sliding window moves across the ring, each window containing the same number of cells.
  - Each window defines a single instance of S_k.
- **Defining Separate Equations:**
  - Each window of cells defines a separate equation in the linear system.
- **Advantages of the Ring Structure:**
  - Ensures uniform selection of cell pools.