Table of Contents

Introduction

This gitbook is used to introduce the detailed algorithm of some important bioinformatics paper, which only include a very brief introduction of the algorithm in the methods.

Normalization

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$$

Each row represents a gene, each column represents a single cell. X is the raw read count matrix.

To normalize X, we mean get the count per ten thousands read, and the +1, and log transform it. This is to adjust the influence of sequence depth.

$$x_{ij} = log(\frac{x_{ij}}{\sum_{k=1}^{n} x_{kj}} \cdot 10000 + 1)$$

In the equation above, x_ij on RHS are raw read count.

Standardization

To make each row has mean 0, and standard deviation 1.

$$x_{ij} = \frac{x_{ij} - \bar{x_{i\cdot}}}{\sigma_{i\cdot}}$$

$$\bar{x_{i\cdot}} = \frac{\sum_{k=1}^{m} x_{ik}}{m}$$

$$\sigma_{i\cdot} = \frac{\sum_{k=1}^{m} (x_{ik} - \bar{x_{i\cdot}})^2}{m-1}$$

In the equation above, x_{ij} on RHS are normalized according to step 1.

Feature Selection

Individual Dataset

Based on unnormalized raw read counts.

For each gene i, we can calculate its mean $\bar{x_{i\cdot}}$ and standard deviation $\sigma_{i\cdot}$ for raw read counts, log transform them. We then fit a curve for these two variables across all cells, by calculating a local fitting of polynomials of degree 2 (R function loess, span = 0.3). For a given gene i, we can know from the fitted curve about its expected standard deviation: σ_i

Denote: $z_{ij} = \frac{x_{ij} - \bar{x_{i\cdot}}}{\sigma_i}$.

We can calculate $\sum_{k=1}^{m} max(z_{ik}^2, \sqrt{m}) \cdot 1_{xik \neq 0} + 1_{xik=0} z_{ik}^2)$, which is nearly the ratio of real standard deviation and the expected standard deviation, namely, $\frac{\sigma_{i\cdot}}{\sigma_i}$. By choosing genes with the highest ratio, we can eliminate the influence of mean on the variability of the gene.

Multiple Datasets

1. Features must exist in every dataset (not necessarily HVG) that are going to be integrated.
2. Features must be highly variable in at least one individual dataset.
3. Select features that most frequently exist in HVG of each individual dataset (Let’s assume there are 50 datasets, a gene is an HVG for 48 of them, then its frequency is 48.).
4. For features with the same frequency in all datasets, calculate the median of their rank in the HVG list (if a gene does not have a rank for a specific list, then just ignore this list and calculate the median rank in all other list that include this gene), select the lower ones. Identification of anchor correspondence bewteen two datasets

Identification of anchor correspondence bewteen two datasets