Varying-Censoring Aware Matrix Factorization (VAMF) for scRNA-seq

yhou@Single.Cell.Batches

08/03/2017

Motivation and basic idea

- Issue in “Smushing”:
  - high proportion of non-biological zeroes -> Clustering based on detection ability
- Previous approach: Consider existence of zero as the result of **censoring**

censoring is a condition in which the value of a measurement or observation is only partially known.

  - Model unobserved expression levels with one factor across all the cells -- Zero Inflated Factor Analysis (ZIFA. Paper & SourceCode)
  - But, are all cells equal?

Detection rate varies across cells

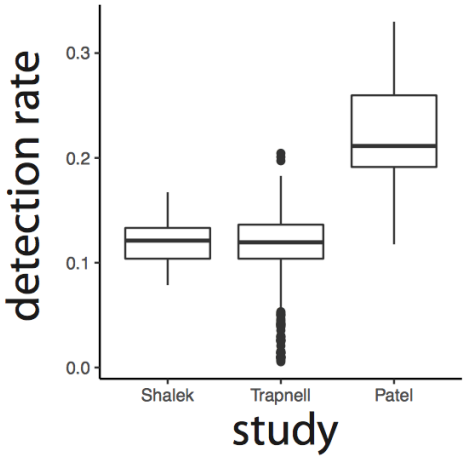
- Binary indicator (Z_{ng}): Whether a gene is detected or not:

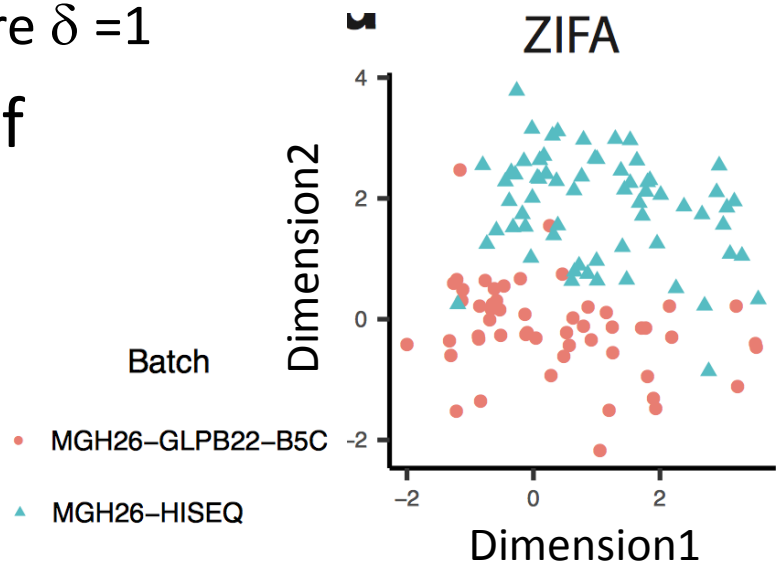
$$Z_{ng} = 1_{Y_{ng} > \delta}$$

Y_{ng} is the normalized scRNA-seq data, where n = 1, …, N are cell indexes, g =1, …, G are gene indexes. δ is the threshold. Here δ =1

- Detection rate (P_n): Ratio of genes detected

$$P_n \equiv \frac{1}{G} \sum_g Z_{ng}$$





VAMF: Cell-specific censoring

- Censoring mechanism:
  - $f_n(\eta_{ng}) \equiv \Pr(Z_{ng} = 1 \mid \eta_{ng})$
  - $E[\log(M_g)] = E[\eta_{ng}]$
  - $f_n(\eta_{ng}) \approx \Pr(Z_{ng} \mid M_g)$
- Unobserved log-transformed expression level (η_{ng}):
  - $\eta_{ng} = y_0 + wg + u'_n v_g$

η_{ng}: unobserved log-transformed expression level
M_g: normalized bulk RNA-seq data

y_0: global intercept
w_g: gene/feature-specific effect
$u_n \sim$ principal components
$v_g \sim$ loadings (correlation coefficients)

Modeling the censoring mechanism

- From: $f_n(\eta_{ng}) \approx \Pr(Z_{ng} \mid M_g)$
  - Use scRNA-seq/bulk RNA-seq pair to compare Z_{ng} vs. $\log_2(M_g)$

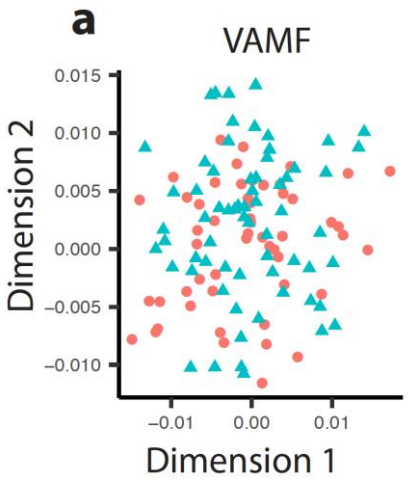
$$\Pr(Z_{ng} = 1 \mid \eta_{ng}) = \frac{1}{1 + \exp\{-(\beta_{0n} + \beta_{1n}\eta_{ng})\}}$$

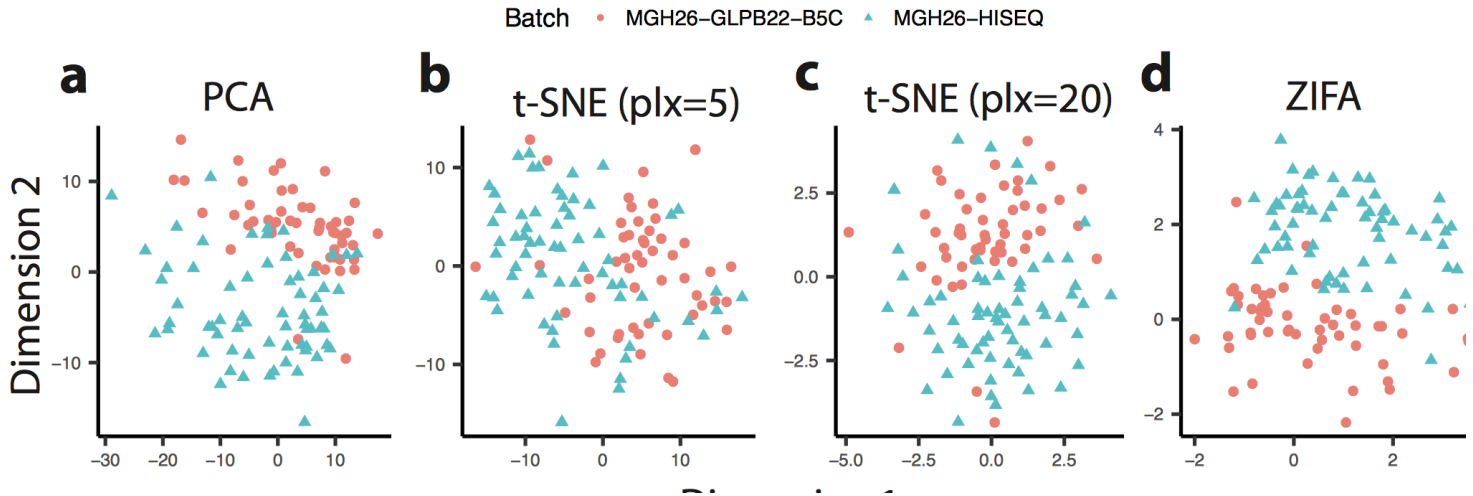
β_{0n} and β_{1n} accounts for the cell-specific censoring

- Model fitting:
  - Borrow info across cells through empirical Bayesian hierarchical model
  - Learn correct latent dimensionality through Automatic Relavance Determination (ABD)
  - (Parameter est. relies on paired bulk RNA-seq data)

VAMF can remove the batch effect

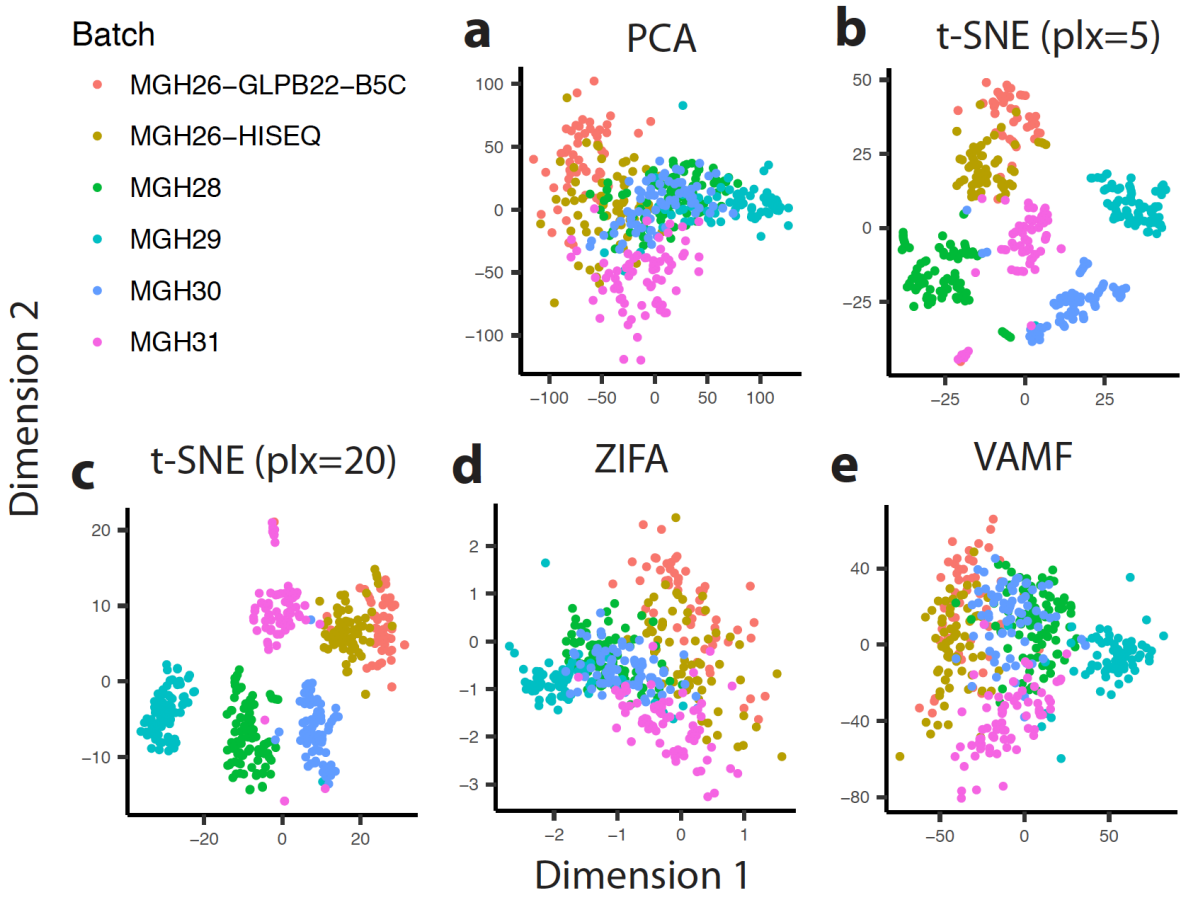
- Two-batches dataset:





VAMF can remove the batch effect

- Five-tumor dataset



Removed technological differences, while maintained biological effect (Though not as separable as other methods).

VAMF outperforms other approaches

- Simulation:
  - 2 generated datasets.
  - Mimicked batch effect with difference in detection rate.