RUV-III-NB: A robust scRNA-seq normalization methods

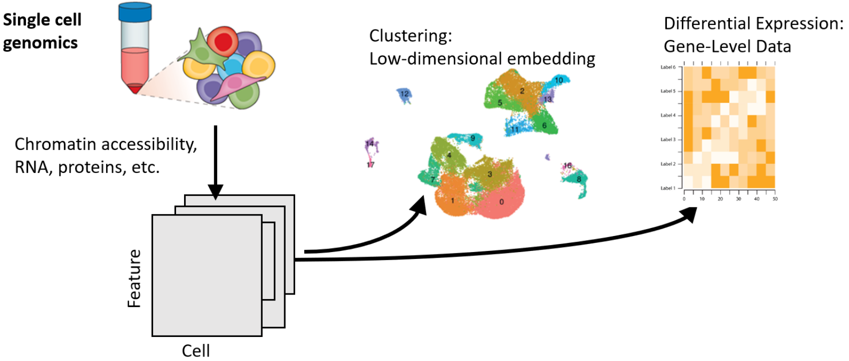
Agus Salim

email: salim.a@unimelb.edu.au
twitter:@asalim_hint

MSPGH and School of Mathematics and Statistics
The University of Melbourne

BIRS Workshop, 4 July 2023

Single-Cell Sequencing



Single cell genomics

Chromatin accessibility, RNA, proteins, etc.

Clustering:
Low-dimensional embedding

Differential Expression:
Gene-Level Data

Feature

Cell

Figure adapted from Longo, Guo, Ji and Khavari (2021, *Nat. Rev. Genetics*)

Clustering is used to identify cell states; DE is used to identify marker genes that differentiate states

Motivations

Definition

Normalization = removal of all kinds of unwanted variation, not limited only to library size

Motivations

Definition

Normalization = removal of all kinds of unwanted variation, not limited only to library size

- Motivation I: Current normalization methods remove biology when unwanted variation (UV) are associated with biology.

Motivations

Definition

Normalization = removal of all kinds of unwanted variation, not limited only to library size

- Motivation I: Current normalization methods remove biology when unwanted variation (UV) are associated with biology.
- Motivation II: Most methods only return dimensional reduction (*cell embedding*) unsuitable for downstream analyses.

Motivations

Definition

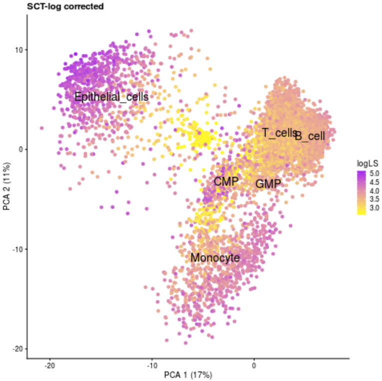
Normalization = removal of all kinds of unwanted variation, not limited only to library size

- Motivation I: Current normalization methods remove biology when unwanted variation (UV) are associated with biology.
- Motivation II: Most methods only return dimensional reduction (*cell embedding*) unsuitable for downstream analyses.
- RUV-III-NB takes into account biology × UV association and return adjusted data for all genes.

NSCLC Study

Non-small cell lung carcinoma ($\sim$ 6,000 cells) study using 10x platform (from one batch)

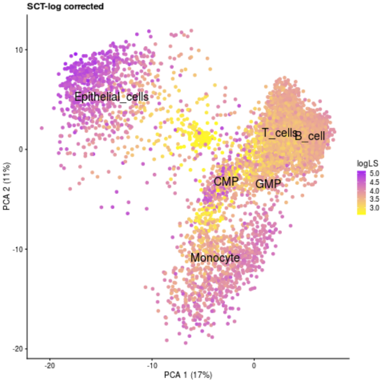
- sctransform seems to separate all major cell-types adequately.



NSCLC Study

Non-small cell lung carcinoma ($\sim$ 6,000 cells) study using 10x platform (from one batch)

- sctransform seems to separate all major cell-types adequately.

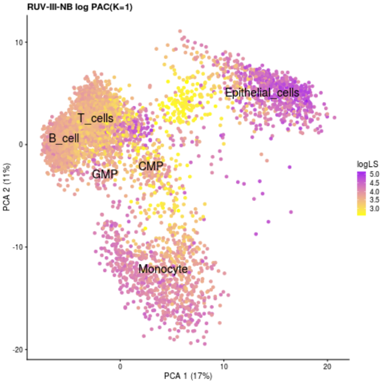


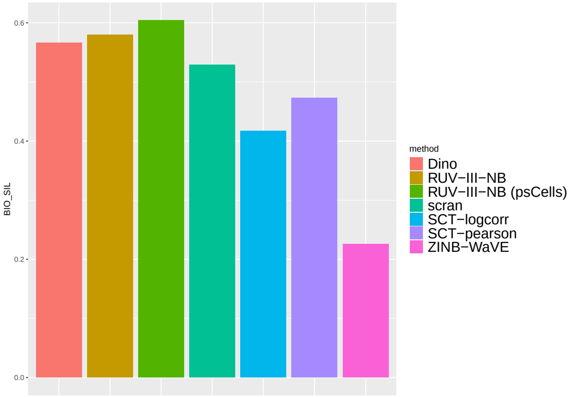
- **Biology (cell-type) is associated with library size (UV)**, with the larger Epithelial cells and Monocytes have higher LS.

NSCLC Study

Non-small cell lung carcinoma ($\sim$ 6,000 cells) study using 10x platform (from one batch)

- RUV-III-NB separates Monocytes better and makes Epithelial cells cluster tighter.





- Only RUV-III-NB and Dino improve biological silhouette score relative to scran.

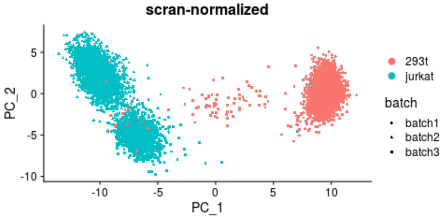
Cell line Study

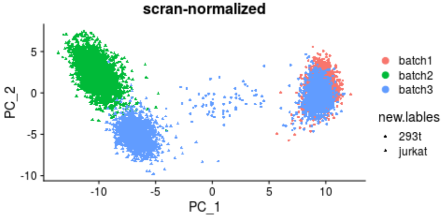
Jurkat and 293t cells ($\sim$ 9,000 cells) from 3 batches (10x protocol), but only one batch contains both cell types.

Cell line Study

Jurkat and 293t cells ($\sim$ 9,000 cells) from 3 batches (10x protocol), but only one batch contains both cell types.

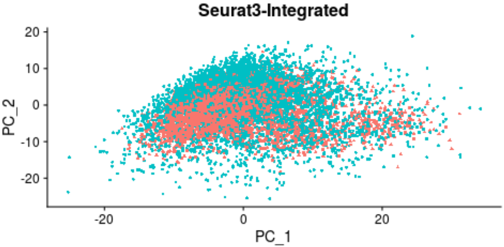
- There's a strong batch effects for Jurkat cells and biology (cell-type) is associated with batch (UV).

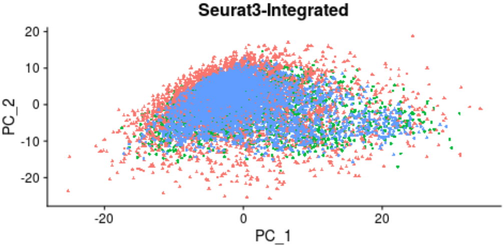

scran-normalized


scran-normalized

Cell line Study

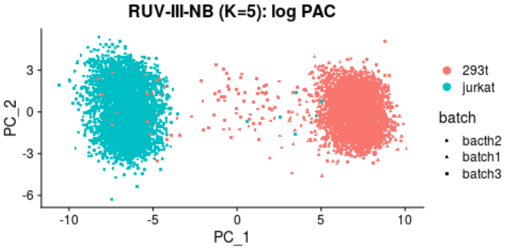
Seurat completely removes biology

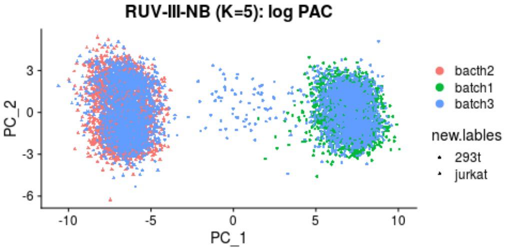




Cell line Study

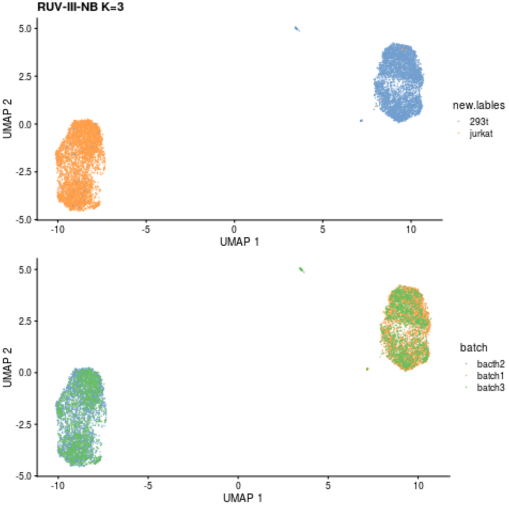
RUV-III-NB removes batch effects without removing biology





Cell line Study

RUV-III-NB removes batch effects without removing biology



RUV-III-NB: Model

- Let $\mu_\mathbf{g} = (\mu_\mathbf{g1}, \mu_\mathbf{g2}, \ldots, \mu_\mathbf{gN})^\mathbf{T}$ be the vector of NB mean parameter for gene g across N cells, we assume $\mathbf{y_g} \sim \mathbf{NB}(\mu_\mathbf{g}, \psi_\mathbf{g})$, with

RUV-III-NB: Model

- Let $\mu_{\mathbf{g}} = (\mu_{\mathbf{g1}}, \mu_{\mathbf{g2}}, \ldots, \mu_{\mathbf{gN}})^{\mathbf{T}}$ be the vector of NB mean parameter for gene g across N cells, we assume $\mathbf{y_g} \sim \mathbf{NB}(\mu_{\mathbf{g}}, \psi_{\mathbf{g}})$, with
- We assume that we have cell state information for $n_a \leq 3,000$ cells. This cell state information can come from:
    - For cell types: highly-confident annotation after initial LS normalization
    - For other factors, e.g. treatment, we have this information from experimental design.

RUV-III-NB: Model

- For cells with annotation,

$$\log \boldsymbol{\mu}_g^a = \boldsymbol{\zeta}_g + \mathbf{M}\beta_g + \boldsymbol{W}_a\boldsymbol{\alpha}_g,$$

$\mathbf{M}(\mathbf{n_a} \times \mathbf{m})$ matrix that contains dummy variables for cell states, $\mathbf{W_a}(\mathbf{n_a} \times \mathbf{K})$ is rows subset of a K-dimensional *unknown* unwanted factors $\boldsymbol{W}$ associated with annotated cells, $\beta_{\mathbf{g}} \sim \mathbf{N}(\mathbf{0}, \lambda_\beta^{-\mathbf{1}}\mathbf{I_m}), \boldsymbol{\alpha}_{\mathbf{g}} \sim \mathbf{N}(\mathbf{0}, \lambda_\alpha^{-\mathbf{1}}\mathbf{I_k})$

RUV-III-NB: Model

- For cells with annotation,

$$\log \boldsymbol{\mu}_g^a = \boldsymbol{\zeta}_g + \mathbf{M}\beta_g + \boldsymbol{W}_a \boldsymbol{\alpha}_g,$$

  $\mathbf{M}(\mathbf{n_a} \times \mathbf{m})$ matrix that contains dummy variables for cell states, $\mathbf{W_a}(\mathbf{n_a} \times \mathbf{K})$ is rows subset of a K-dimensional *unknown* unwanted factors $\boldsymbol{W}$ associated with annotated cells, $\beta_\mathbf{g} \sim \mathbf{N}(\mathbf{0}, \lambda_\beta^{-\mathbf{1}}\mathbf{I_m}), \boldsymbol{\alpha}_\mathbf{g} \sim \mathbf{N}(\mathbf{0}, \lambda_\alpha^{-\mathbf{1}}\mathbf{I_k})$

- For cells without annotation,

$$\log \boldsymbol{\mu}_g^u = \boldsymbol{\zeta}_g + \beta_{gc} + \boldsymbol{W}_u \boldsymbol{\alpha}_g,$$

  $\mathbf{W_u}$ is rows subset of $\boldsymbol{W}$ associated with the un-annotated cells and $\beta_{gc} \sim N(0, \lambda_\beta^{-1})$

RUV-III-NB: Model

- We also assume there is a negative control gene set (C) so that for any genes in this set,

  $$\log \boldsymbol{\mu}_g = \zeta_g + \boldsymbol{W}\boldsymbol{\alpha}_g,$$

  W(**N** × **k**) is a K-dimensional *unknown* unwanted factors for all cells

Adjusted data: log percentile-invariant adjusted count (PAC)

1. Calculate percentile under full fitted model: $r_{gc} = \frac{a_{cg}+b_{cg}}{2}$, where

$$\begin{aligned} a_{gc} &= F_{NB}(y_{gc}; \mu_{gc} = e^{\hat{\zeta}_g + \hat{\beta}_{gc} + \hat{\mathbf{w}}_{\mathbf{c}}^T \hat{\boldsymbol{\alpha}}_g}, \hat{\psi}_g) \\ b_{gc} &= F_{NB}(y_{gc} + 1; \mu_{gc} = e^{\hat{\zeta}_g + \hat{\beta}_{gc} + \hat{\mathbf{w}}_{\mathbf{c}}^T \hat{\boldsymbol{\alpha}}_g}, \hat{\psi}_g) \end{aligned}$$

and ˆw_c the cth row of the matrix $\hat{W}$

Adjusted data: log percentile-invariant adjusted count (PAC)

1. Calculate percentile under full fitted model: $r_{gc} = \frac{a_{cg}+b_{cg}}{2}$, where

$$\begin{aligned} a_{gc} &= F_{NB}(y_{gc}; \mu_{gc} = e^{\hat{\zeta}_g + \hat{\beta}_{gc} + \hat{\mathbf{w}}_{\mathbf{c}}{}^T \hat{\boldsymbol{\alpha}}_g}, \hat{\psi}_g) \\ b_{gc} &= F_{NB}(y_{gc} + 1; \mu_{gc} = e^{\hat{\zeta}_g + \hat{\beta}_{gc} + \hat{\mathbf{w}}_{\mathbf{c}}{}^T \hat{\boldsymbol{\alpha}}_g}, \hat{\psi}_g) \end{aligned}$$

   and ˆw_c the cth row of the matrix $\hat{W}$

$$\mathrm{PAC}_{gc} = F_{NB}^{-1}(r_{gc}; \mu_{gc} = \exp(\hat{\zeta}_g + \hat{\beta}_{gc} + \bar{\mathbf{w}}^T \hat{\boldsymbol{\alpha}}_g), \hat{\psi}_g)$$

2. Invert the percentile under NB distribution where the mean is shifted to have average unwanted variations, where ¯w is vector of entries equal to the average level $N^{-1} \sum_{c=1}^{N}$ˆw_c of unwanted variation.

Adjusted data: log percentile-invariant adjusted count (PAC)

1. Calculate percentile under full fitted model: $r_{gc} = \frac{a_{cg}+b_{cg}}{2}$, where

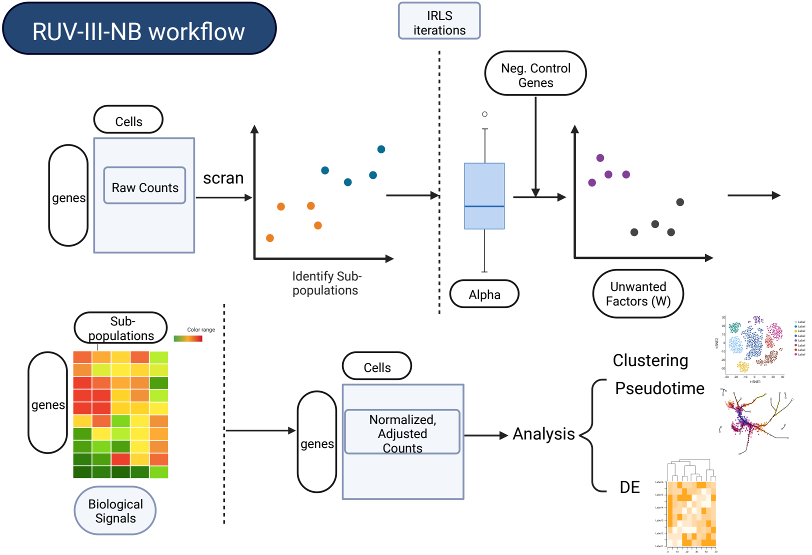
$$\begin{aligned} a_{gc} &= F_{NB}(y_{gc}; \mu_{gc} = e^{\hat{\zeta}_g + \hat{\beta}_{gc} + \hat{\mathbf{w}}_{\mathbf{c}}{}^T \hat{\boldsymbol{\alpha}}_g}, \hat{\psi}_g) \\ b_{gc} &= F_{NB}(y_{gc}+1; \mu_{gc} = e^{\hat{\zeta}_g + \hat{\beta}_{gc} + \hat{\mathbf{w}}_{\mathbf{c}}{}^T \hat{\boldsymbol{\alpha}}_g}, \hat{\psi}_g) \end{aligned}$$

   and $\hat{}w_c$ the cth row of the matrix $\hat{W}$

$$\text{PAC}_{gc} = F_{NB}^{-1}(r_{gc}; \mu_{gc} = \exp(\hat{\zeta}_g + \hat{\beta}_{gc} + \bar{\mathbf{w}}^T \hat{\boldsymbol{\alpha}}_g), \hat{\psi}_g)$$

2. Invert the percentile under NB distribution where the mean is shifted to have average unwanted variations, where $\bar{}w$ is vector of entries equal to the average level $N^{-1}\sum_{c=1}^{N} \hat{}w_c$ of unwanted variation.
3. Add 1 and take log → $\log(\text{PAC}_{gc}+1)$

RUV-III-NB workflow



Cells

genes

Raw Counts

scran

Identify Sub-populations

IRLS iterations

Neg. Control Genes

Alpha

Unwanted Factors (W)

Sub-populations

Color range

genes

Biological Signals

Cells

genes

Normalized, Adjusted Counts

Analysis

Clustering

Pseudotime

DE

Parameter Estimation

- Iterative reweighted least squares (IRLS)-based
- Parameters ζ_g,ψ_g,$\boldsymbol{W}_a$ and $\boldsymbol{\alpha}_g$ are estimated using annotated cells
- Parameters β_{gc} and $\boldsymbol{W}_u$ are estimated using un-annotated cells.

To run RUV-III-NB, we need:

- Cell states information (**M** matrix): some cells need to have known cell states.

To run RUV-III-NB, we need:

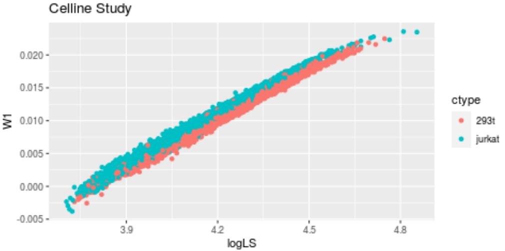
- Cell states information (**M** matrix): some cells need to have known cell states.
- Negative control gene sets: RUV-III-NB is a robust against a degree of miss-specification

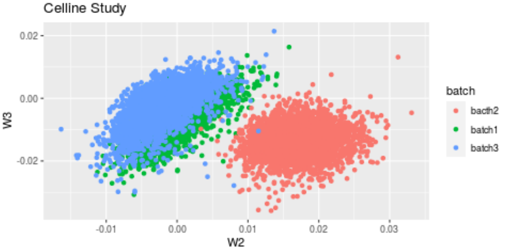
To run RUV-III-NB, we need:

- Cell states information (**M** matrix): some cells need to have known cell states.
- Negative control gene sets: RUV-III-NB is a robust against a degree of miss-specification
- The number of unwanted factors (**K**): slight overestimation does not remove biological signals.

Cell line Study: **W** estimates

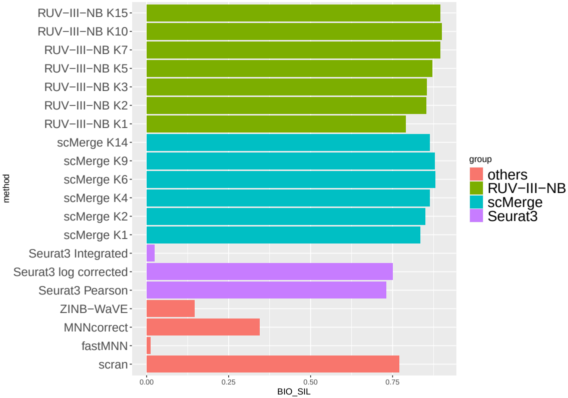
RUV-III-NB correctly identifies logLS and batch as the unwanted factors.





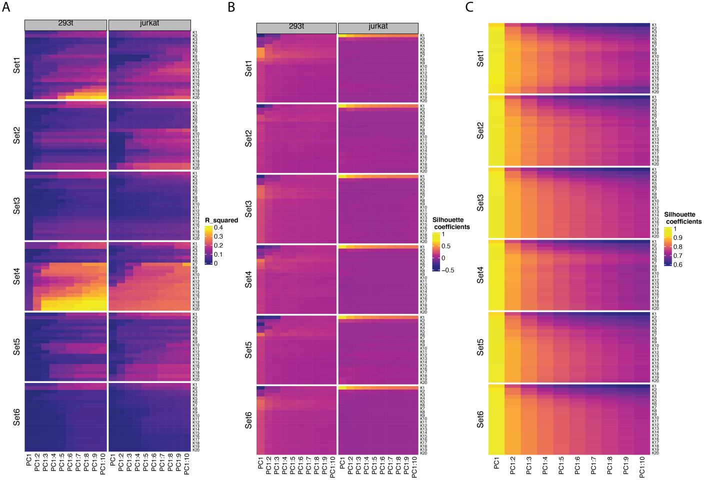
Cell line Study

RUV-III-NB's performance is quite robust for a range of assumed unwanted factors (K)



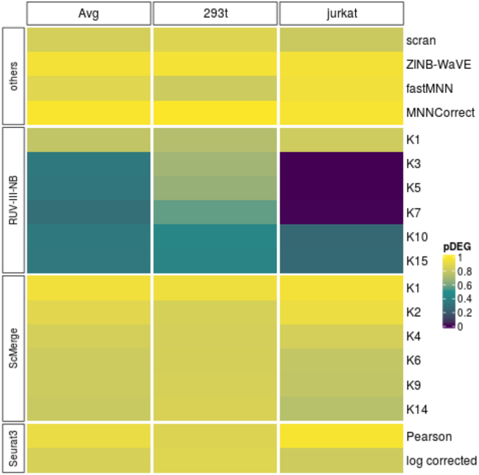
Cell line study

Robust performance with different sets of negative control genes



Cell line Study: DEG

DEG of the same cell types located in different batches. RUV-III-NB adjusted data has the smallest amount of batch effects



ZINB extension

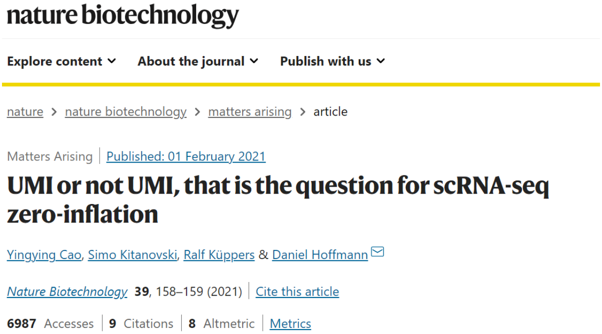
- UMI data dominates in scRNA-seq world but there are still platforms without UMI

ZINB extension

- UMI data dominates in scRNA-seq world but there are still platforms without UMI
- Non-UMI data are known to often exhibit zero inflation

ZINB extension

- UMI data dominates in scRNA-seq world but there are still platforms without UMI
- Non-UMI data are known to often exhibit zero inflation



nature biotechnology

Matters Arising | Published: 01 February 2021

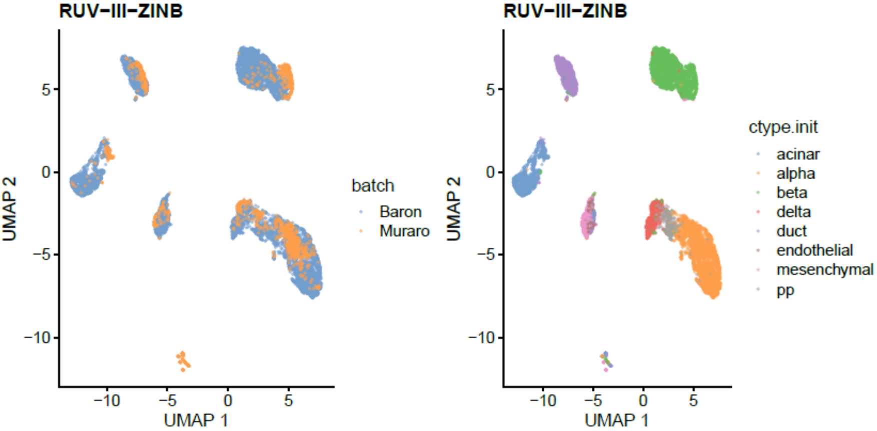
UMI or not UMI, that is the question for scRNA-seq zero-inflation

Yingying Cao, Simo Kitanovski, Ralf Küppers & Daniel Hoffmann

Nature Biotechnology **39**, 158–159 (2021) | Cite this article

6987 Accesses | **9** Citations | **8** Altmetric | Metrics

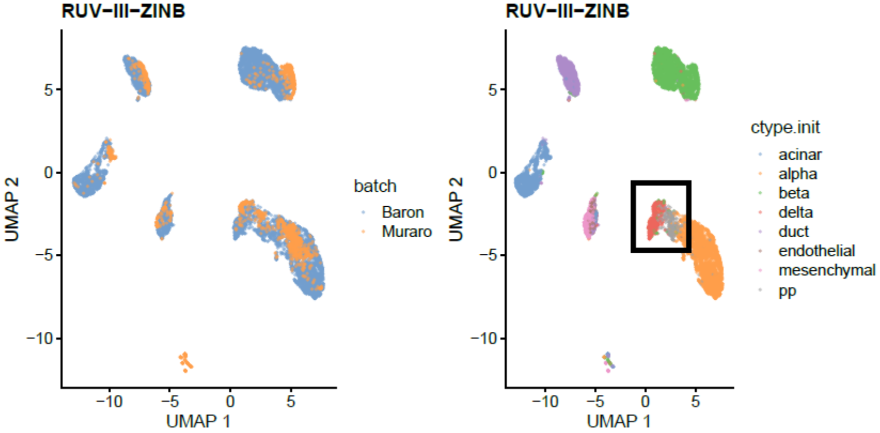
RUV-III-ZINB



This is achieved with only 5% of the cells having known annotations.

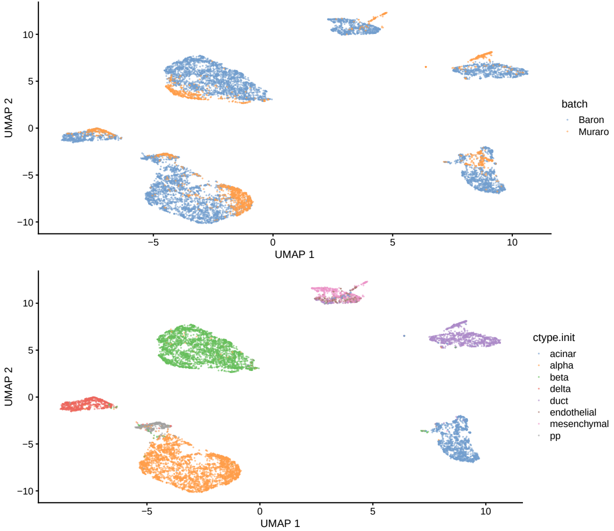
Robustness against incorrect annotation?

We rerun RUV-III-ZINB assuming that the delta and PP cells are of the same cell-type



Robustness against incorrect annotation?

RUV-III-ZINB can still separate the two cell-types



Conclusion

- RUV-III-NB removes UV and preserves biology when biology and UV are associated.

Conclusion

- RUV-III-NB removes UV and preserves biology when biology and UV are associated.
- RUV-III-NB returns percentile adjusted count (PAC) that can be readily used for downstream analyses without further normalisation.

Conclusion

- RUV-III-NB removes UV and preserves biology when biology and UV are associated.
- RUV-III-NB returns percentile adjusted count (PAC) that can be readily used for downstream analyses without further normalisation.
- The method has a degree of robustness against overestimation of unwanted factors, negative control gene sets and miss-specified initial annotation.
- R package is available from `https://github.com/limfuxing/ruvIIInb/`.

Conclusion

- RUV-III-NB removes UV and preserves biology when biology and UV are associated.
- RUV-III-NB returns percentile adjusted count (PAC) that can be readily used for downstream analyses without further normalisation.
- The method has a degree of robustness against overestimation of unwanted factors, negative control gene sets and miss-specified initial annotation.
- R package is available from `https://github.com/limfuxing/ruvIIInb/`.
- Future works: extensions to scMultiOmics and spatial transcriptomics.

Acknowledgments

- Terry Speed, Ramyar Molania, Jianan Wang (WEHI)
- Alysha de Livera (La Trobe)
- Hsiao-chi Liao, Muhammad Fachrul (UoM)
- Jean Yang, Yingxin Lin (USyd)