Resistant Fit Regression Normalization for Single-cell RNA Sequencing Data

Da Kuang, Junhyong Kim

Penn Engineering

1 Introduction

- Motivation
- Current Methods
- Potential Issues

Penn Engineering

1-1 Motivation

- scRNA-seq is used to estimates the molecular counts of all the mRNA types in a cell.
- Many factors can affect the measurement.
  - Technical variations: noise, reduced sensitivities, etc.
  - Biological variations: different cell types, life cycles, etc.
- **Normalization** corrects variations and makes measurements of two or more samples comparable.

1-2 Current Methods

Three types of normalization methods

- Estimate scaling factors
  - Log-scale norm, SCRAN
- Approximate probabilistic model
  - scVI, DCA
- Normalize to matched distribution
  - SCNorm, SCTransform

1-3 Potential Issues

All three classes of methods are focused on technical variations but that is not enough for transcriptome analysis.

1-3 Potential Issues: Example

Example

Suppose we have 101 genes in two cells, g_1 ~ g_{101}.

- In $Cell_1$, all gene counts are 100 except g_{101}= 0.
- In $Cell_2$, all gene counts are 80 except g_{101} = 2000.

Two cell have the same sum of counts.

Discussion

The measured molecular numbers of g_1 to g_{100} are different in two cells.

- But it is not necessary to mean different expression levels.
- The variations are caused by the highly differential expression of g_{101}.

Correct the first normalized cell by a 0.8 size factor, making genes g_1 to g_{101} equivalent in both datasets.

	Measured	Normalized	Better Normalized
$Cell_1$	(100,100, …,100, 0)	(0.01, …, 0.01, 0)	(0.008, …, 0.008, 0)
$Cell_2$	(80, 80, …, 80, 2000)	(0.008, …, 0.008, 0.2)	(0.008, …, 0.008, 0.2)

2 Normalization Framework

- Overview
- Two Transcriptomes
- Multi-transcriptomes

Penn Engineering

2-1 Overview

Here we propose

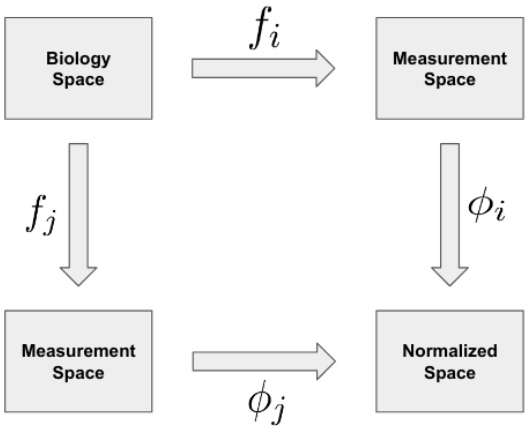
- For a given cell, $\boldsymbol{\phi}_i$ is the same among all the genes.
- For different cells, $\boldsymbol{\phi}_i$ are only differed by a scalar size factor.

If we estimate the size factor based on all the genes, then it makes no difference with count per million (CPM).

Biological Feature Set (S)

Inspired by example, to estimate size factor, we will only use a subset of genes that did not experience localized large deviation.

Schematic of Framework



Normalizing Implicit Rule (NI)

If $x_1 = x_2$ in biology space,
then $\boldsymbol{\phi}_1(f_1(x_1)) = \boldsymbol{\phi}_2(f_2(x_2))$.

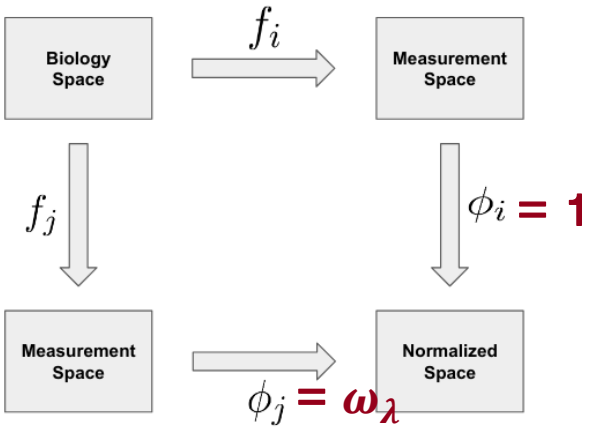
2-2 Two Transcriptomes

Object

To find the normalizing function, we set $\boldsymbol{\phi}_i$ to an identical function and $\boldsymbol{\phi}_j$ to some function ω with parameter λ.

- Suppose we have two observations of two cells,
  $$x = \{x_1, \dots, x_g\} \text{ and } y = \{y_1, \dots, y_g\}.$$
- We find a normalizing function by minimizing
  $$O(\lambda) = \|x - \omega_\lambda(y)\|^2$$
- For simplicity, we only consider ω_λ as a two-parameter linear function.

Schematic of Framework



Normalizing Implicit Rule (NI)

If $x_1 = x_2$ in biology space,
then $\boldsymbol{\phi}_1(f_1(x_1)) = \boldsymbol{\phi}_2(f_2(x_2))$.

2-3 Two Transcriptomes

Optimization

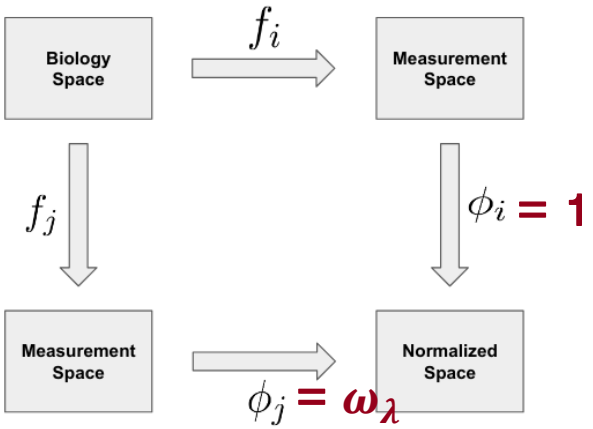
We use robust regression to fit the normalizing function ω_λ while excluding extremely distorted data.

- Regress on the biological feature set S with p genes.
- By default $p = 0.7 \times n$, where n is the number of genes .
- The types of genes in S can be different from cell to cell.

Expectation-maximization (EM) algorithm for optimization

- [Init E step] Linear fit on all genes.
- [M step] Rank genes by squared residual. Put top p genes in S.
- [E step] Liner regression on S. Repeat M, E steps until converge.

Schematic of Framework



Normalizing Implicit Rule (NI)

If $x_1 = x_2$ in biology space,
then $\boldsymbol{\phi}_1(f_1(x_1)) = \boldsymbol{\phi}_2(f_2(x_2))$.

2-4 Multi-transcriptomes

For more than two transcriptomes, we suggest an iterated normalization to match all the measurements to a centroid transcriptome.

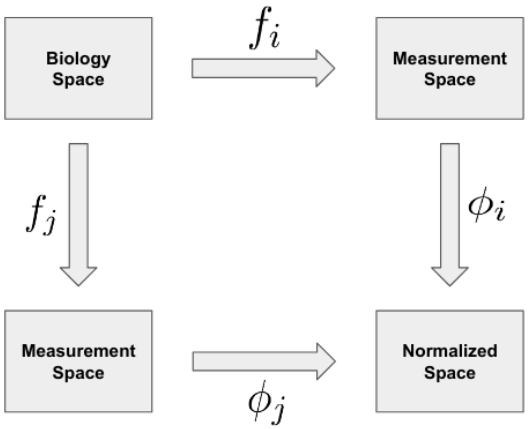
Let $T_1, T_2, \ldots, T_n$ be n different transcriptomes for normalization and T_c is defined as the centroid transcriptome.

Algorithm 1 Normalizing Multiple Transcriptomes

```
for j in epoch do
    T_c = arithmetic average (T_1, T_2, ··· , T_n)
    for i in 1:n do
        Normalize T_i to T_c
    end for
end for
```

In our experiment, usually 1 to 3 epoch would be sufficient to normalize a dataset.

Schematic of Framework



Normalizing Implicit Rule (NI)

If $x_1 = x_2$ in biology space,
then $\boldsymbol{\phi}_1(f_1(x_1)) = \boldsymbol{\phi}_2(f_2(x_2))$.

3 Result

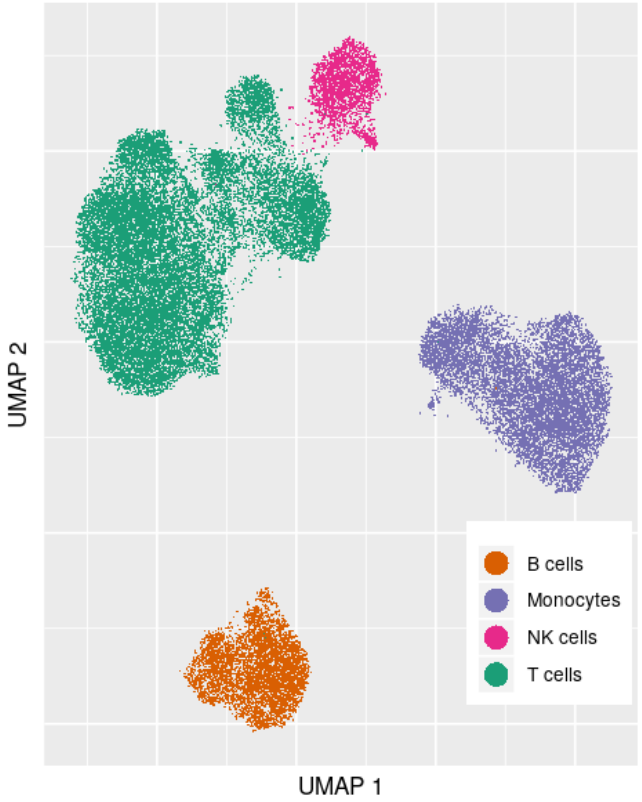
- Biological Application
- Differential Expression
- Biological Sense

Penn Engineering

3-1 Biological Application

Normalize a single-cell RNA sequencing dataset of 33,148 peripheral blood mononuclear cells.

- Four major types
- Fourteen sub-cell types.



3-2 Differential Expression

False Positives in Differential Expression Test

For each sub-cell, cells were evenly separated into two groups.

- In one group, gene counts were randomly down sampled to 50% of the original expression level.
- Therefore, ideally, we expect no DE genes between groups since they are biologically identical.

Resistant Fit Norm has the smallest number of false positives.

Number of False Positives

Cell Types	N Cells	Normalization Methods				
		RF Norm	Log-scale	SCRAN	SCT	scVI
CD4 Memory	5961	0	0	0	0	1
CD4 Naive	4431	0	0	0	0	1
Mono CD14	5551	0	0	0	0	0
B Pre	2886	0	0	0	0	1
CD8 Memory	2067	0	0	0	0	1
CD8 Naive	1917	0	0	0	0	2
NK Dim	1655	3	0	0	0	1
Mono CD16	1475	2	0	1	0	1
CD8 Effector	1453	0	0	0	0	4
B Pro	953	0	1	1	0	0
DC	204	6	57	27	1	1
NK Bright	217	4	59	14	2	0
Mk	140	16	119	74	9	0
pDC	96	18	203	116	30	8

RF Norm = Resistant Fit Normalization
Log-scale = log(1 + CPM)
SCT = SCTransform

3-3 Biological Sense

Biological Meaning Between Cell Types

Normalize the CD4 and CD8 T cells followed by differential expression test.

The number of differentially expressed genes (DEs) are in the table.

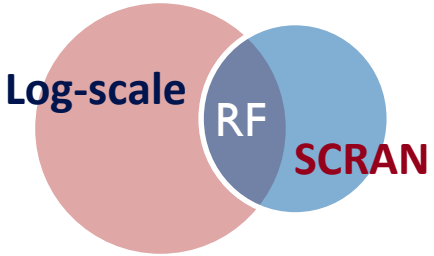
- RF Norm DEs form a subset of Log-scale and SCRAN DEs.
- RF Norm Down DEs form a subset of SCTransform Down DEs .
  - RF Norm Down DEs are enriched in “MHC class I protein binding”.
  - SCTransform Down DEs are enriched in “MHC class II protein binding”.
  - The result of RF Norm is more consistent with the function of CD8 T cells.

CD4 Memory vs. CD8 Effector			
Method	Total	Down	Up
Log-scale	724	161	563
SCRAN	210	70	140
SCTransform	86	45	41
SCVI	87	60	27
Resistant Fit	105	32	73
Shared with Log-scale	104	24	69
Shared with SCRAN	105	32	73
Shared with SCT	61	32	29
Shared with SCVI	29	19	10

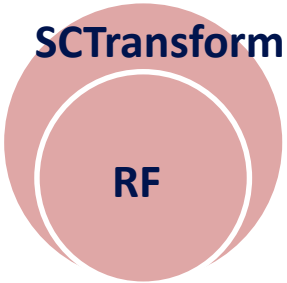
Up refers to increased expression in CD4 cells.

Down refers to decreased expression in CD4 cells.

All genes:



Down genes:



Penn Engineering

4 Conclusion

Penn Engineering

4 Conclusion

We present a new framework for normalization while making the samples biologically comparable.

Our Resistant Fit Normalization shows the less false-positive rate and better biological meaning, demonstrating an effective discovery of a differentially expressed list of genes.

Acknowledgment

This project is funded by the Pennsylvania Commonwealth Health Research Formula Funding.

Thank you!

Penn Engineering