Normalization and **variance stabilization** of single-cell RNA-seq data using **regularized** negative binomial regression

Christoph Hafemeister et al. (2020/04)

1-introduction

Characteristics of an effective normalization

- In general, the normalized expression level of a gene should not be correlated with the total sequencing depth of a cell.
- The variance of a normalized gene (across cells) should primarily reflect biological heterogeneity, independent of gene abundance or sequencing depth.

SC RNA-seq Counts are confounded by seq depth

- 10X 22k PBMC is used
- In B, divide the genes into 6 groups by gene mean cross cell.
- C, D show the trends of gene count vs. cell count.
- E divides cells into 5 groups based seq-depth.
- E calculate the contribution of each cell group to the variations within each gene group.

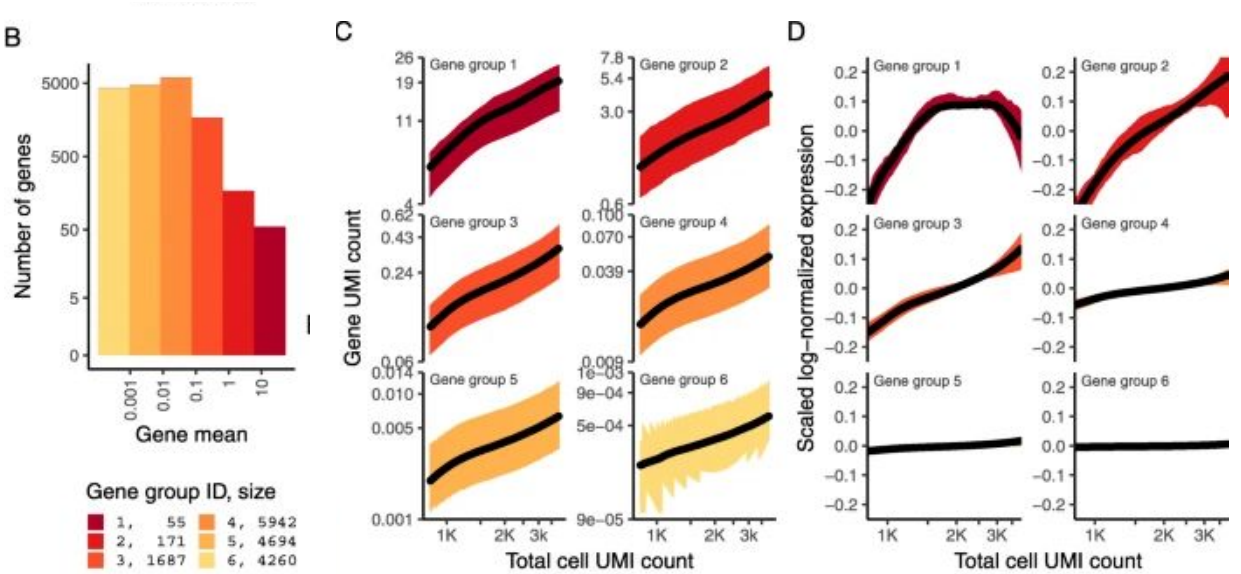
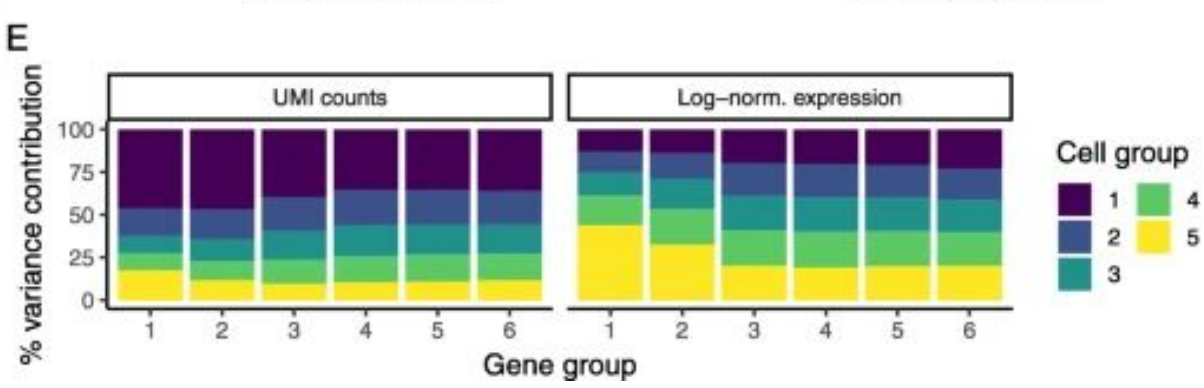


Fig1

SC RNA-seq variance are confounded by seq depth

- E divides cells into 5 groups based seq-depth.
- E calculate the contribution of each cell group to the variations within each gene group.



D and E shows that being normalized by single size factor will be confounded by sequence depth.

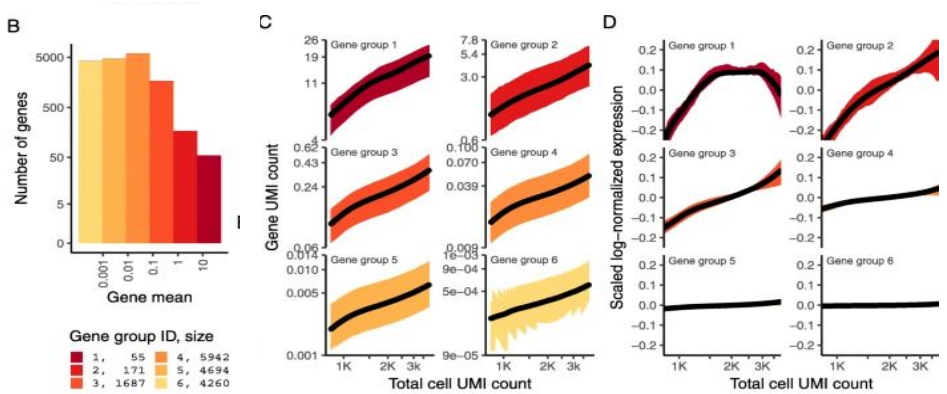


Fig1

Main idea of SCTransform

- That different groups of genes cannot be normalized by the same constant factor.
- A generalized linear model (GLM) for each gene with UMI counts as the response and sequencing depth as the explanatory variable.
- We can regularize parameter by pooling information across genes with similar abundances.

2-method

Review General Linear Model

Linear Model

$$Y|X \sim \mathcal{N}(\mu(X), \sigma^2 I)$$

$$\mathbb{E}(Y|X) = \mu(X) = X^\top \beta$$

General Linear Model

1. Random component:

$$Y \sim \text{some exponential family distribution}$$

2. Link: between the random and covariates:

$$g(\mu(X)) = X^\top \beta$$

where g called link function and $\mu = \mathbb{E}(Y|X)$.

Review General Linear Model

Linear Model

$$Y|X \sim \mathcal{N}(\mu(X), \sigma^2 I)$$

$$\mathbb{E}(Y|X) = \mu(X) = X^\top \beta$$

MASS: glm.nb()

General Linear Model

1. Random component:

$$Y \sim \text{Negative Binomial}$$

2. Link: between the random and covariates:

$$\log(\mathbb{E}(x_i)) = \beta_0 + \beta_1 \log_{10} m$$

Variance: $\mu + \frac{\mu^2}{\theta}$

Regularize the GLM

Fig A is the dot plot of model parameters vs. gene mean.

Interestingly, they use bootstrap to calculate the uncertainty of each dot and color the dots by degree of uncertainty.

Fig B is the distribution of parameters' standard deviation during bootstrap.

So the estimated parameters were not reproducible across bootstraps.

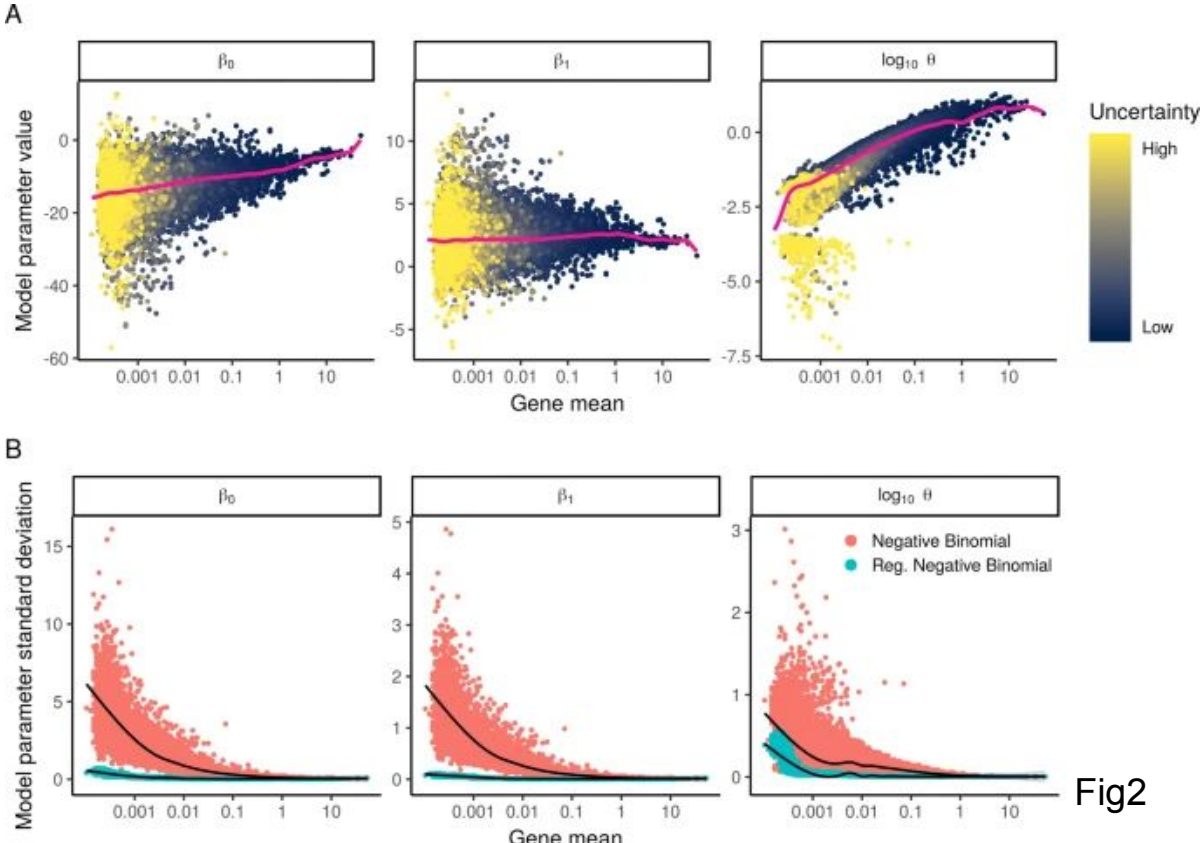


Fig2

Regularize the GLM

So the estimated parameters were not reproducible across bootstraps.

kernel regression is used to smoothing the distribution by considering the effect of neighbors within the kernel window.

It can be seen as a regularization to avoid overfitting to the technical noise.

The blue dots in fig B show that the parameters become stable during bootstrap after regularization.

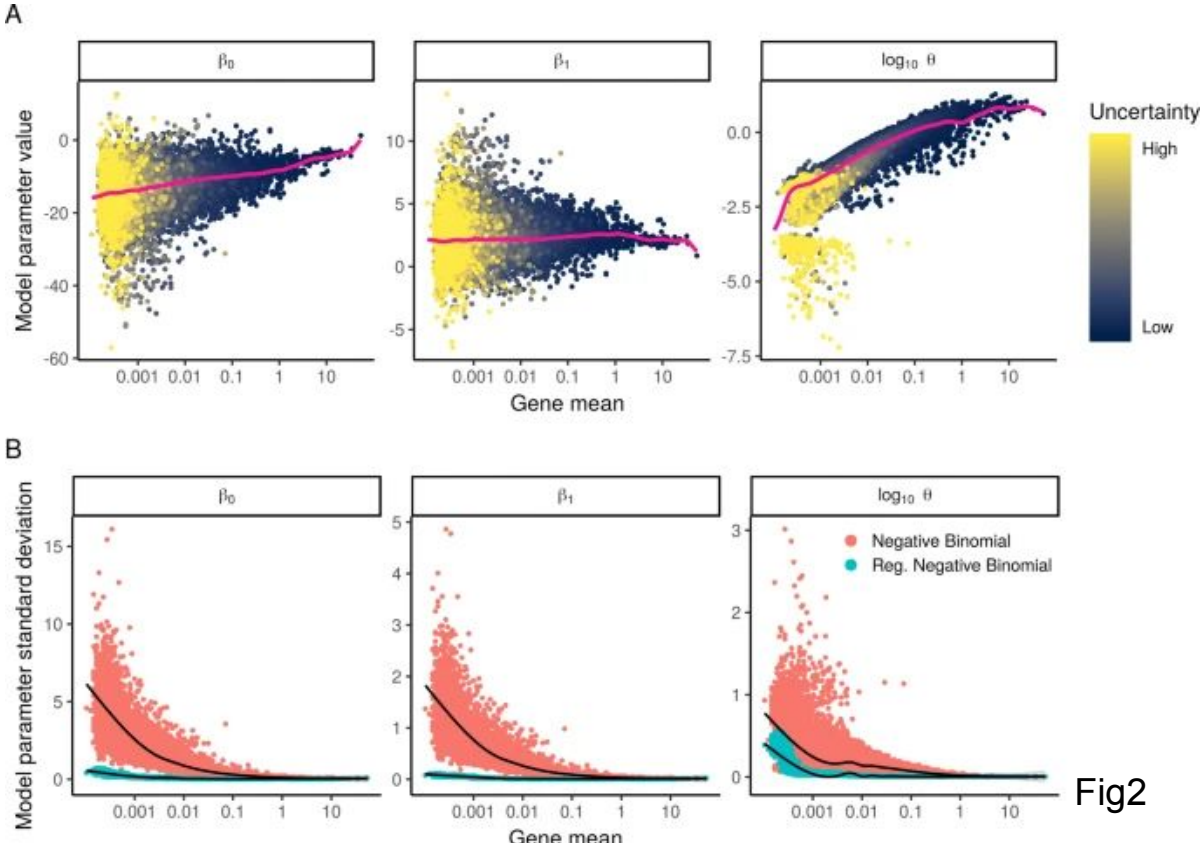


Fig2

Pearson residual effectively normalized scRNA-seq

Pearson residuals: response residuals divided by the expected standard deviation.

In fig A, it shows that Pearson residuals are independent with the total UMI counts in most of the cells.

In figure B, it shows that the contribution of variance become even if Pearson residuals are used as the corrected expression level.

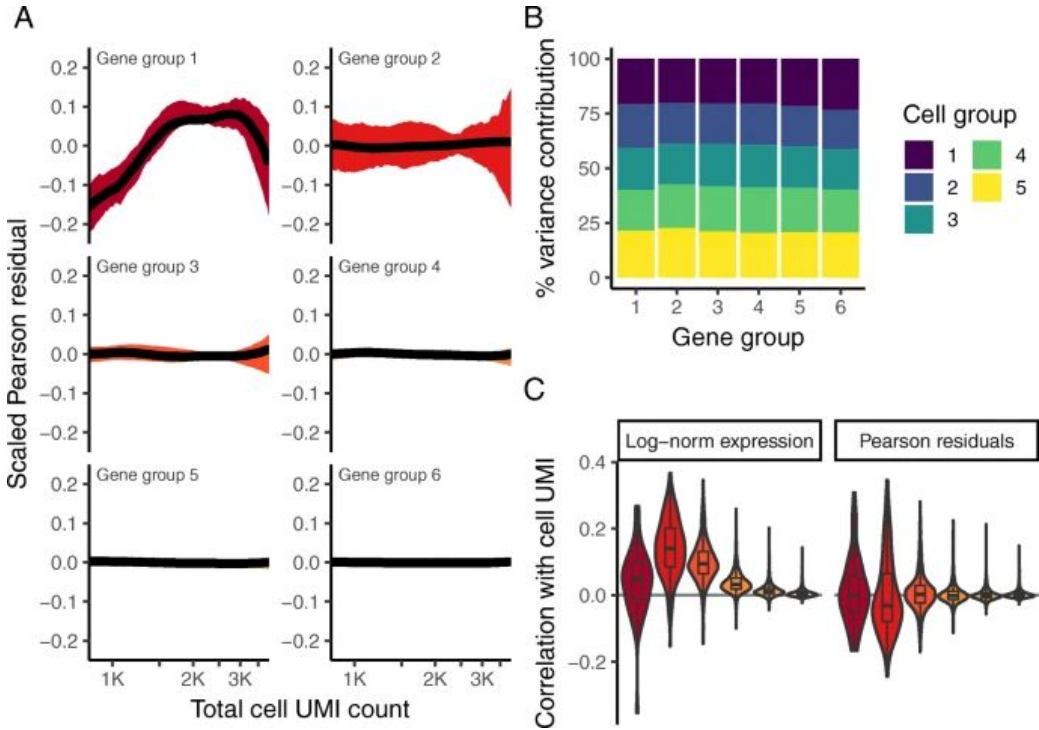


Fig3

3-result

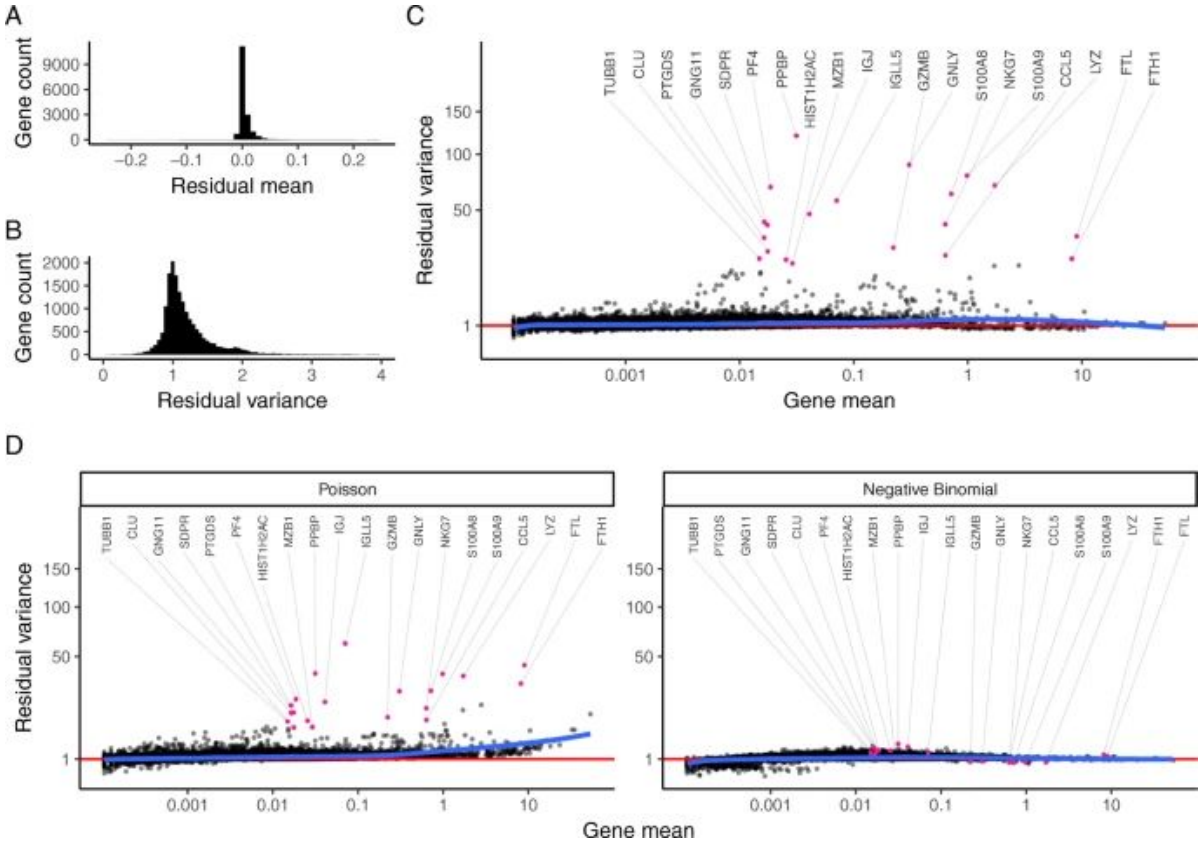

A
Gene count
9000
6000
3000
0
−0.2
−0.1
0.0
0.1
0.2
Residual mean
B
Gene count
2000
1500
1000
500
0
0
1
2
3
4
Residual variance
C
Residual variance
150
100
50
1
TUBB1
CLU
PTGDS
GNG11
SDPR
PF4
PPBP
HIST1H2AC
MZB1
IGJ
IGLL5
GZMB
GNLY
S100A8
NKG7
S100A9
CCL5
LYZ
FTL
FTH1
0.001
0.01
0.1
1
10
Gene mean
D
Poisson
Negative Binomial
Residual variance
150
100
50
1
TUBB1
CLU
GNG11
SDPR
PTGDS
PF4
HIST1H2AC
MZB1
PPBP
IGJ
IGLL5
GZMB
GNLY
NKG7
S100A8
S100A9
CCL5
LYZ
FTL
FTH1
0.001
0.01
0.1
1
10
Gene mean

Fig4

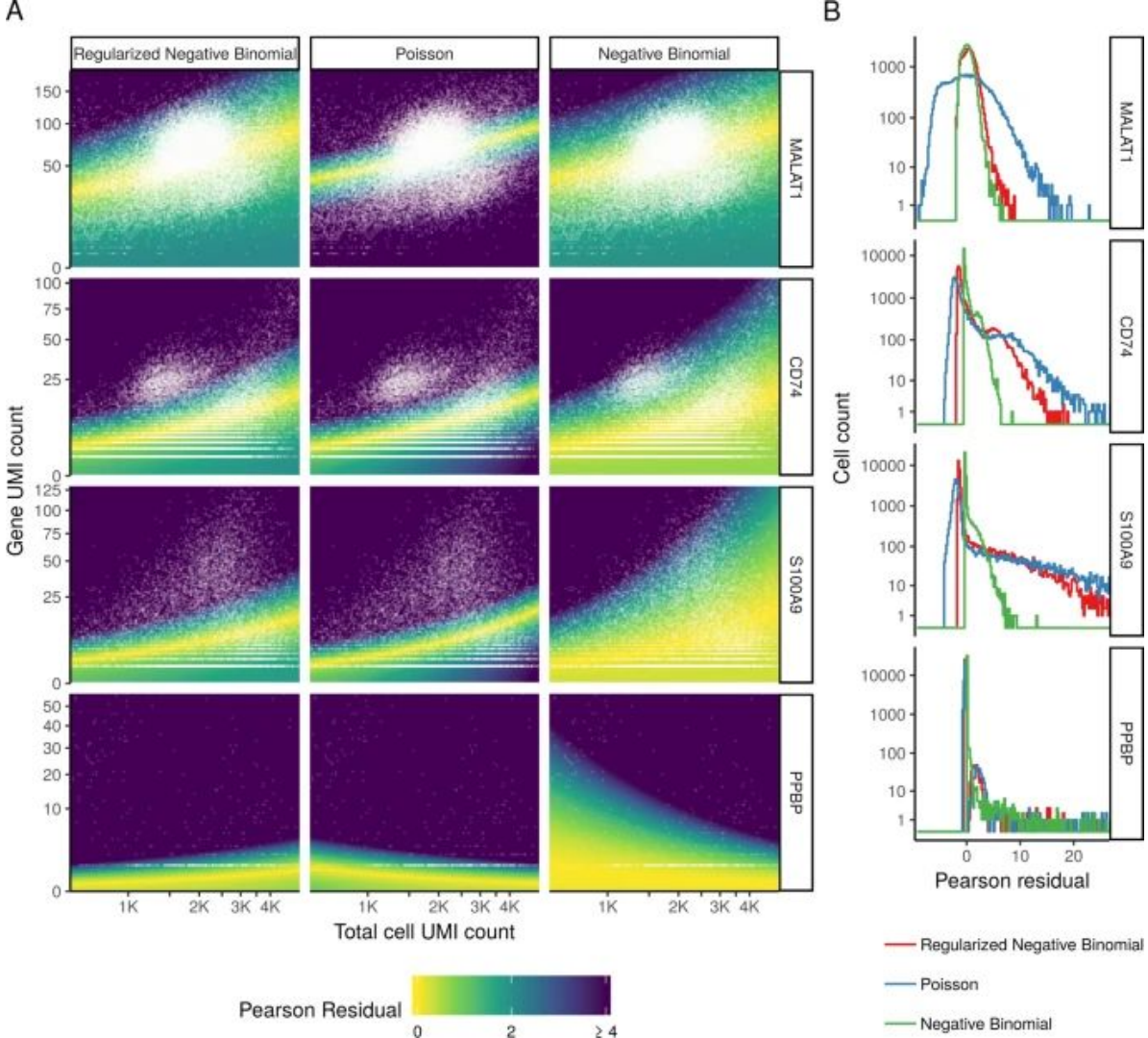
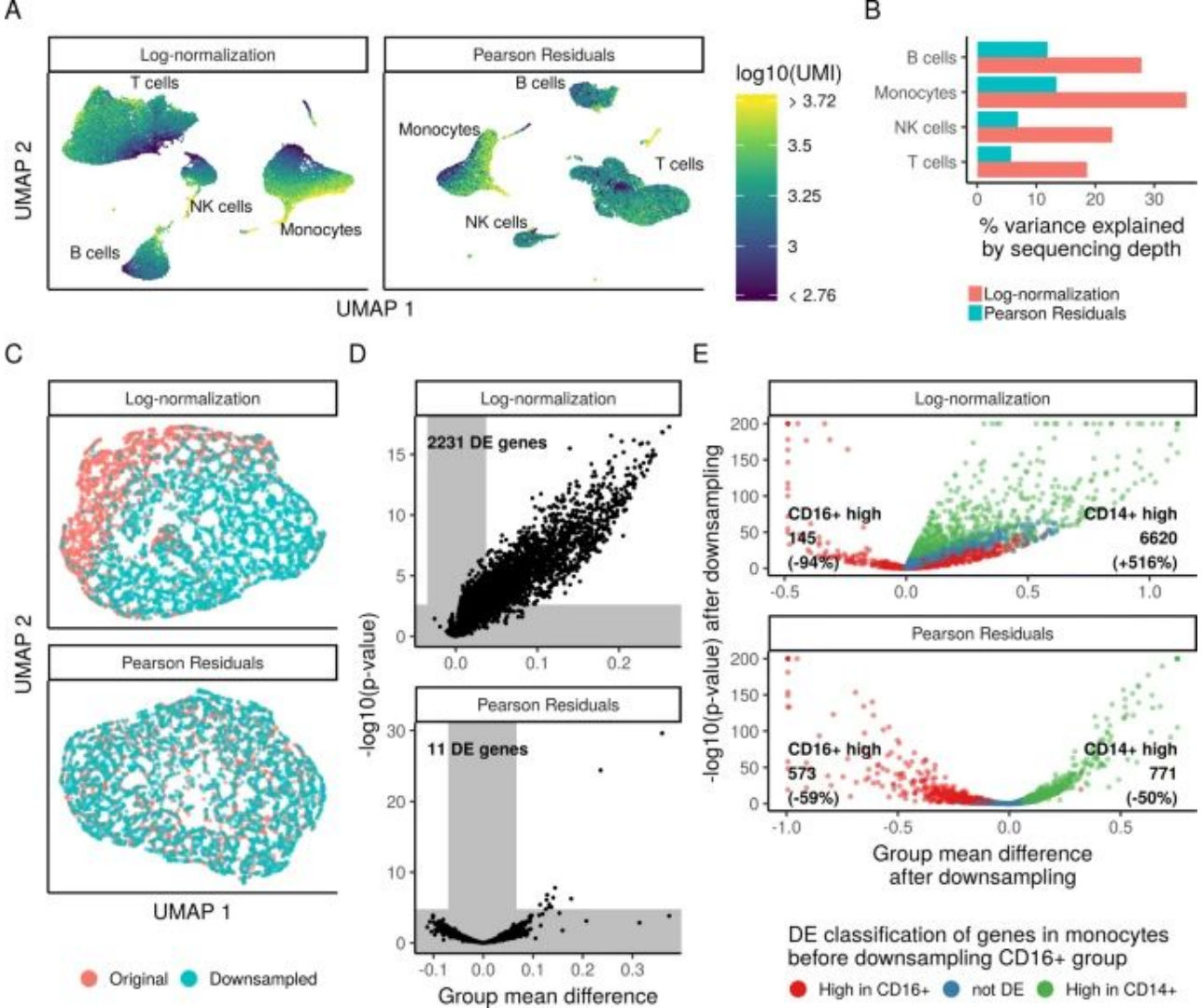

A
Regularized Negative Binomial
Poisson
Negative Binomial
MALAT1
CD74
S100A9
PPBP
Gene UMI count
Total cell UMI count
1K
2K
3K
4K
Pearson Residual
0
2
≥ 4
B
Cell count
Pearson residual
Regularized Negative Binomial
Poisson
Negative Binomial

Fig5


A
Log-normalization
Pearson Residuals
T cells
B cells
NK cells
Monocytes
UMAP 2
UMAP 1
log10(UMI)
> 3.72
3.5
3.25
3
< 2.76
B
B cells
Monocytes
NK cells
T cells
0
10
20
30
% variance explained by sequencing depth
Log-normalization
Pearson Residuals
C
Log-normalization
Pearson Residuals
UMAP 2
UMAP 1
Original
Downsampled
D
Log-normalization
2231 DE genes
Pearson Residuals
11 DE genes
-log10(p-value)
Group mean difference
E
Log-normalization
CD16+ high
145
(-94%)
CD14+ high
6620
(+516%)
Pearson Residuals
CD16+ high
573
(-59%)
CD14+ high
771
(-50%)
-log10(p-value) after downsampling
Group mean difference after downsampling
DE classification of genes in monocytes before downsampling CD16+ group
High in CD16+
not DE
High in CD14+

Thank you!