Imputation Methods For Single Cell Data

Jee Dong Jun

19.9.2020

Single Cell RNA sequencing

- Recent development allows isolation of single cells.

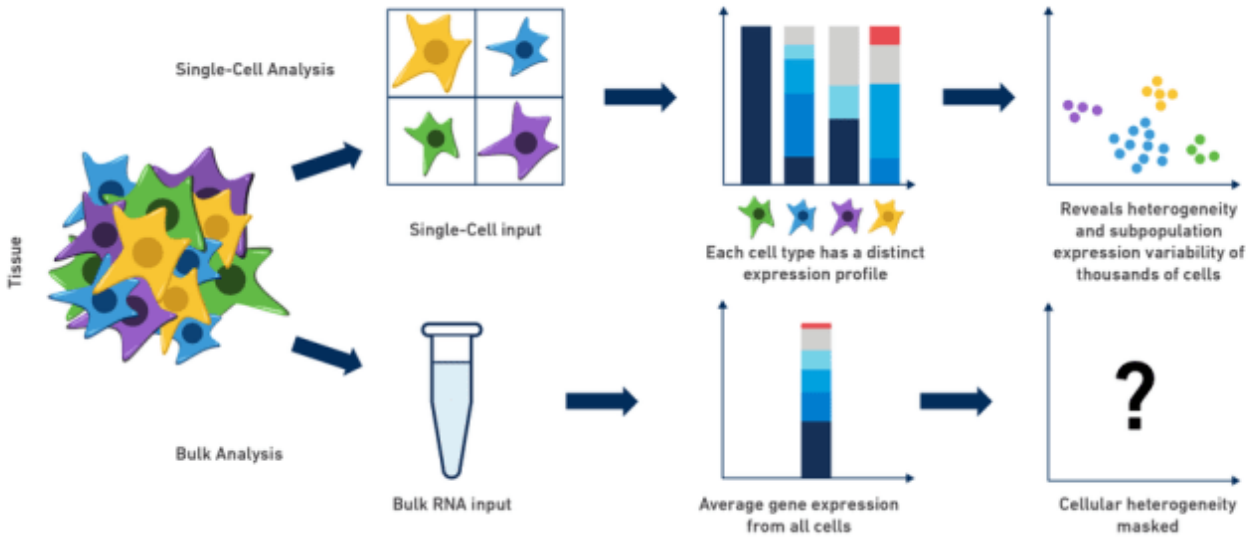


Figure 1: single cell vs bulk cell

Flow of scRNA sequencing

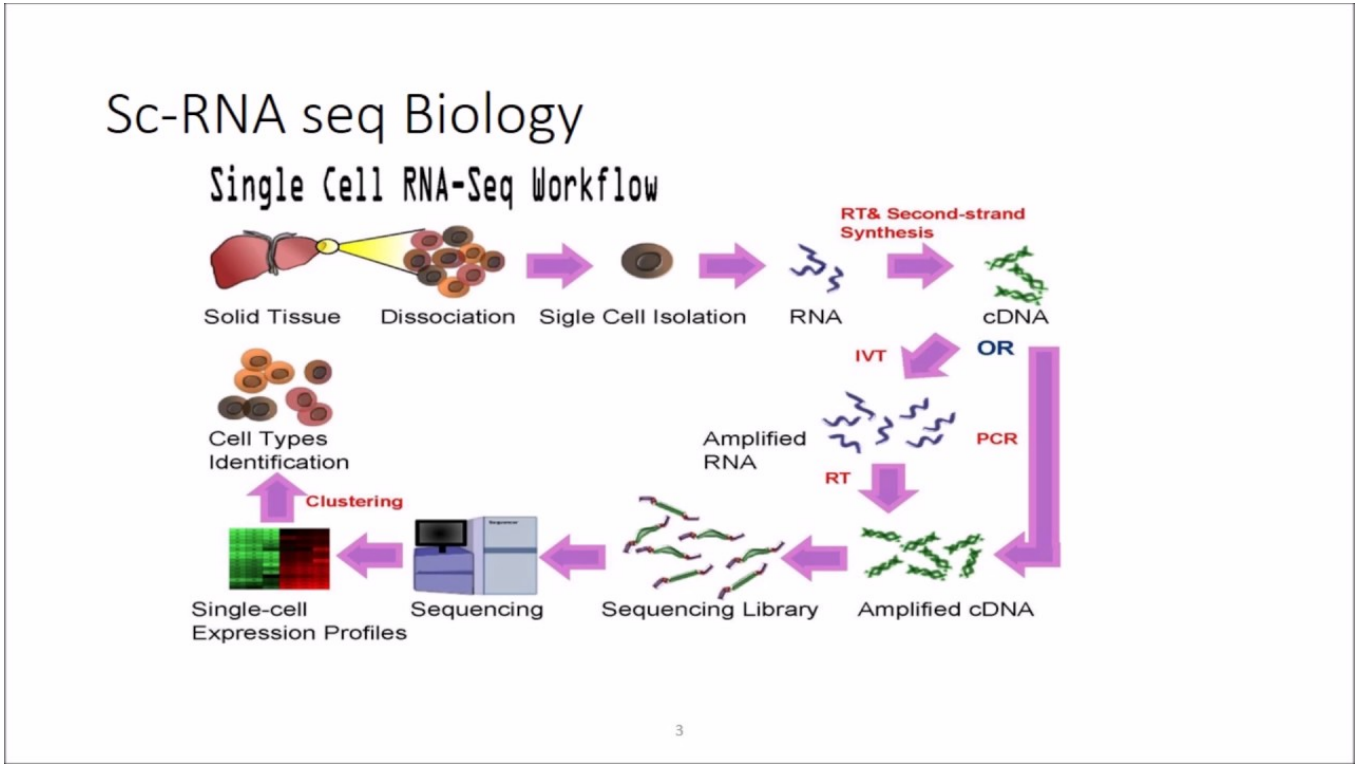


Figure 2: Method of Single Cell

Problem of Sequencing

- Amplification Bias
- Low RNA capture rate
- Dropout events happen because of low RNA capture rate .
- To increase capture rate, we could increase sequencing depth but this requires cost.

Challenge in Imputation

- Not all 0 are missing values.
- Some are 0 because gene is not expressed and some are 0 because RNA was not captured.
- These are difficult to differentiate so traditional method of imputation may not be applicable.
- There are no true values to check whether imputation is done well.

Single-cell analysis via expression recovery SAVER

- Method designed for UMI counts
- It uses gene to gene relationship to recover the true expression level.
- Model UMI count as Poisson-gamma mixture, negative binomial.

Model of SAVER

- $Y_{gc} \sim Poisson(s_c \lambda_{gc})$, where Y is observed UMI count, is true expression and s is size factor.
- $\lambda_{gc} \sim gamma(\alpha, \beta)$ where α, β are reparametrization of mean and variance.
- Data without UMI counts are subject to more amplification bias and would violate poisson distribution.
- Goal is to find posterior distribution $\lambda_{gc} | Y_{gc}$

Model of SAVER

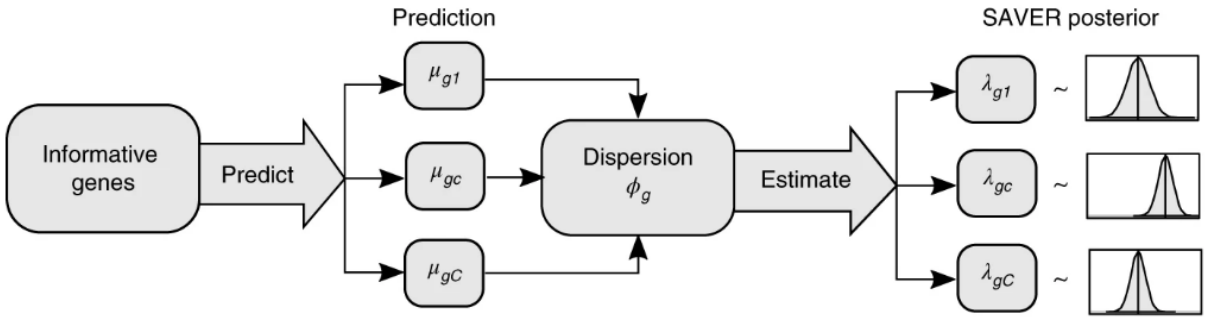


Figure 3: from the original paper of saver

Estimate Prior Mean

- We estimate prior mean with expression of other genes in the same cell.
- Use GLM with Y_{gc}/s_c as response and $Y_{g'c}$ as predictor.
- We are using Poisson regression model with a log link function.

$$\log E(Y_{gc}/s_c) = \log \mu_{gc} = \gamma_{g0} + \sum \gamma_{gg'} \log(\frac{Y_{g'c} + 1}{s_c})$$

- A penalized poisson LASSO regression is used.
- Only few genes affect.

Estimate Prior Variance

- Prior Variance is estimated by assuming constant noise model across cells ϕ_g
- Three possible models, where we assume constant variance v_g, or constant α_g or constant β_g
- We can find MLE of all three models and choose the one with highest MLE.

Posterior distribution

- Once we have both estimated value of μ and v, reparametrize them to α and β
- The posterior is then $\lambda_{gc}|Y_{gc} \sim Gamma(Y_{gc} + \alpha_{gc}, s_c + \beta_{gc})$
- The recovered expression is posterior mean.
- $\lambda_{gc} = \frac{s_c}{s_c+\beta_{gc}} \frac{Y_{gc}}{s_c} + \frac{\beta_{gc}}{s_c+\beta_{gc}} \mu_{gc}$

How to test result

- From a melanoma cell Drop-seq was used to sequence.
- 26 drug-resistance markers and housekeeping genes RNA FISH measurement were obtained from same cell line.
- After filtering only 15 genes are common
- Gini coefficient is a measure of gene expression variability

Result

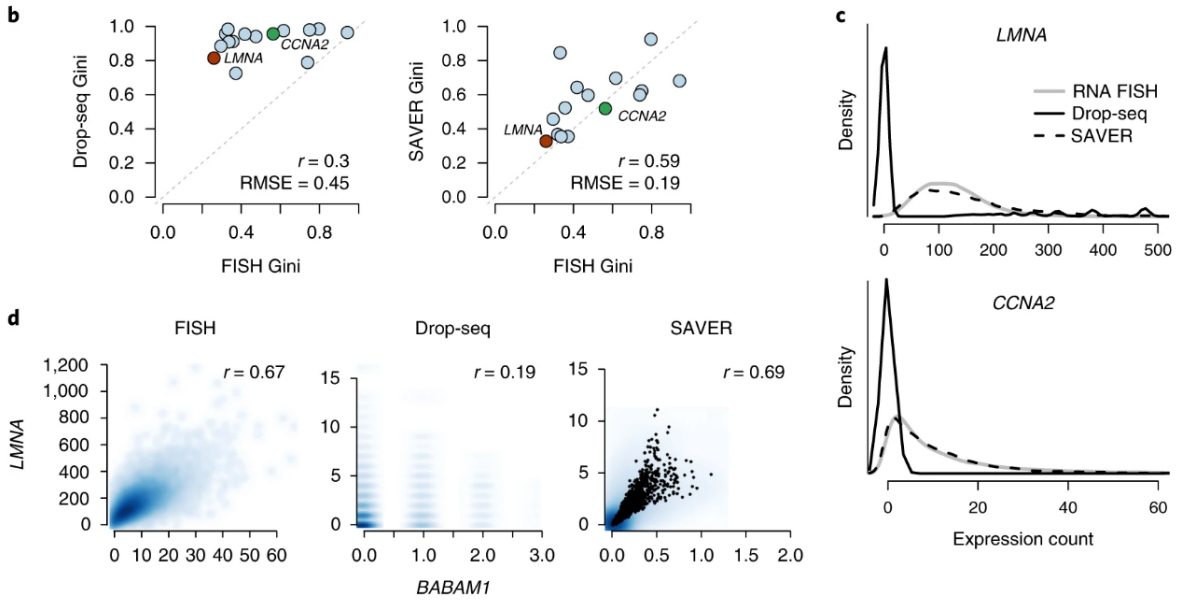


Figure 4: Result from original SAVER paper

Downsampling

- Generate a reference dataset from real data.
- Select high-quality cells and genes with high expression from the orignal dataset to treat as the true expression.
- Downsampled observed dataset by drawing form a Poisson distribution with mean parameter $\tau_c \lambda_{gc}$ where τ_c is the cell specific efficiency loss.
- Calculate gene to gene correlation and cell to cell correlation and compare with reference dataset.

tsne result

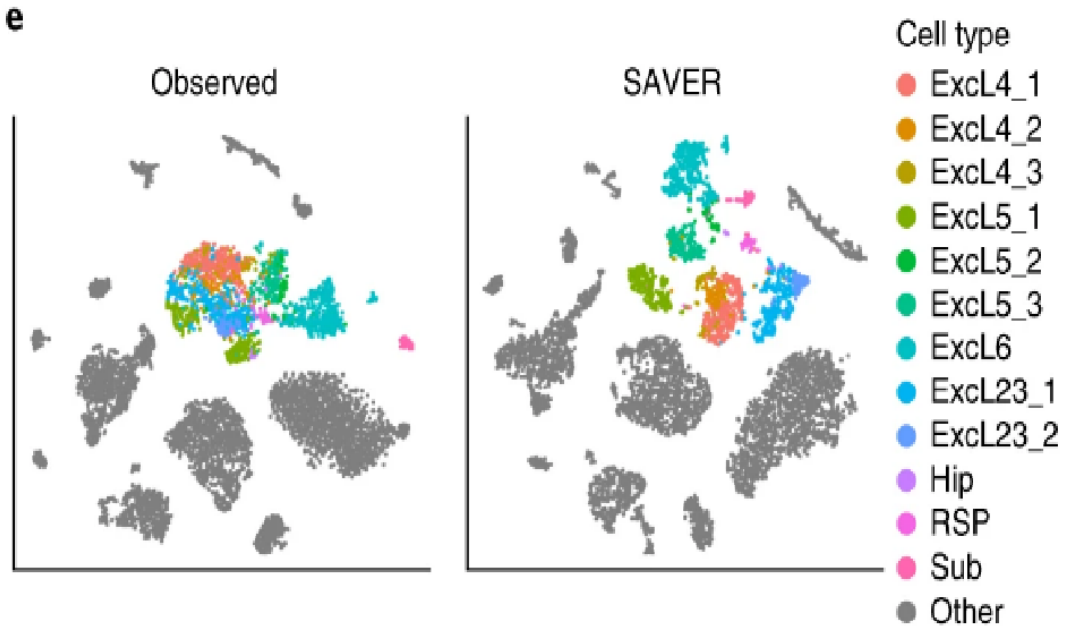


Figure 5: Result from original SAVER paper

Summary

- Captures gene to gene correlation and cell to cell correlation well.
- Only applicable to UMI model
- Every cells are imputed.
- All zero counts are considered as missing values (?)
- Scalability

scImpute

- Determine which values of zero counts are really missing values.
- Based on mixture model, learn each gene's dropout probability in each cell.
- Imputes the dropout values in a cell by borrowing information of similar cells.

scImpute

- Determine which values of zero counts are really missing values.
- Based on mixture model, learn each gene's dropout probability in each cell.
- Imputes the dropout values in a cell by borrowing information of similar cells.

Detecting Neighbourhood

- Normalize count matrix and take log transformation.
- If label is known, we can utilize them.
- If unknown carry out PCA and calculate the distance matrix D
- Based on D, determine which cells are outliers.
- After removing outliers, cells are clustered into K groups.

Identification of Dropout values

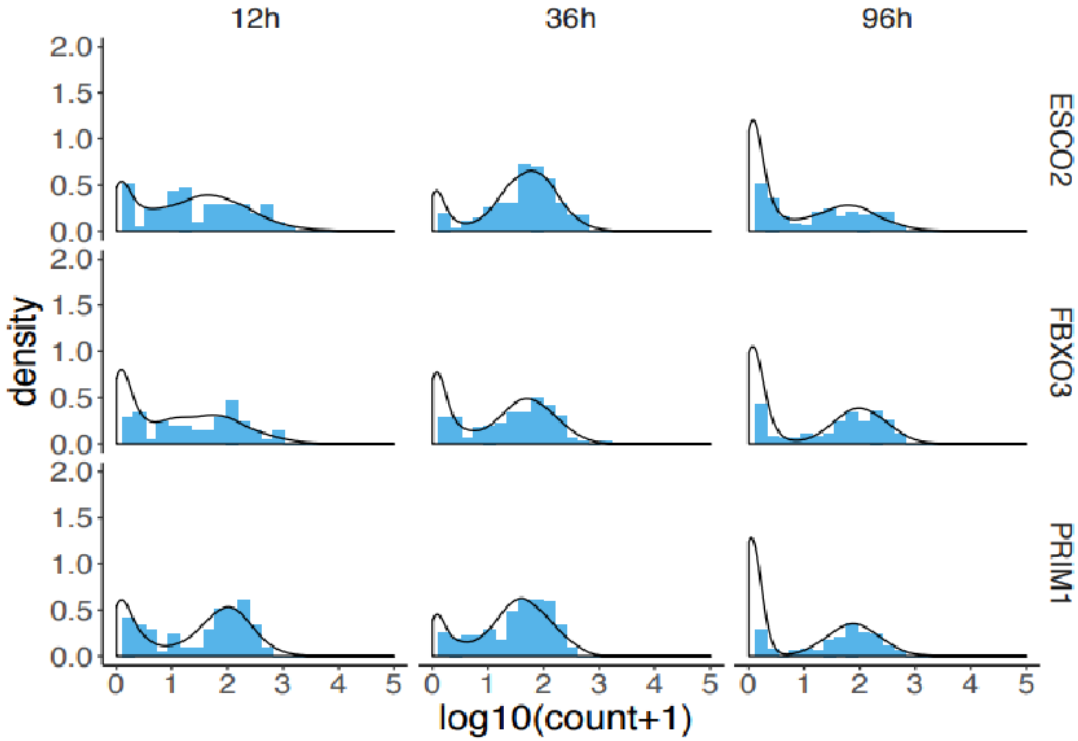


Figure 6: Example gene distribution in scImpute paper

- Model dropout events with gamma distribution and expression as normal distribution.

Dropout Model

- For each gene i, in cell subpopulation k,

$$\lambda_i^{(k)} Gamma(x; \alpha_i^{(k)}, \beta_i^{(k)}) + (1 - \lambda_i^{(k)}) Normal(x; \mu, \sigma)$$

- Dropout probability is $\lambda_i^{(k)}$
- All parameters are estimated by EM algorithm.
- Dropout probability of gene i in cell j which belongs to subpopulation k is

$$d_{ij} = \frac{\lambda_i^{(k)} Gamma(X_{ij}; \alpha_i^{(k)}, \beta_i^{(k)})}{\lambda_i^{(k)} Gamma(X_{ij}; \alpha_i^{(k)}, \beta_i^{(k)}) + (1 - \lambda_i^{(k)}) Normal(X_{ij}; \mu, \sigma)}$$

Imputation of likely dropout values

- Impute cell by cell
- For each cell, using d_{ij} determine gene set A_j that need imputation and B_j that does not need imputation.
- We use B_j to determine similarity of cells.
- Carry out non-negative least square

$$mininimze||X_{B_j,j} - X_{B_j,N_j}\beta^{(j)}||\ \beta \geq 0$$

- Note, N_j are candidate neighbours of cell j.
- Do not impute B_j cells and impute $X_{i,N_j}\beta$ for genes in A_j

ERCC spike in

- ERCC spike-ins are synthetic RNA molecules with known concentrations
- Therefore, we can use it to compare read count with true concentrations.

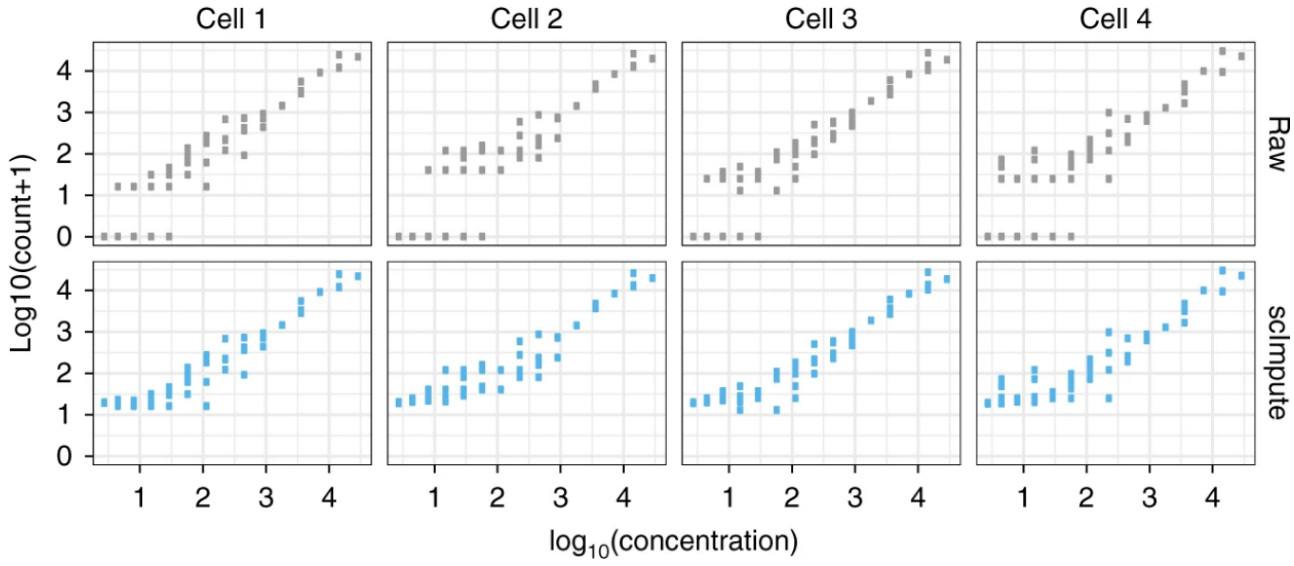


Figure 7: count vs concentration from scImpute paper

Cell Cycle Gene

- Sequence embryonic stem cells that had been staged for cell cycle phases (G1, G2M, and S)
- Cell cycle genes are known to modulate the cell cycle and are expected to have non-zero expression in different stages of cell cycle.
- But before imputation 25 % of them are 0.

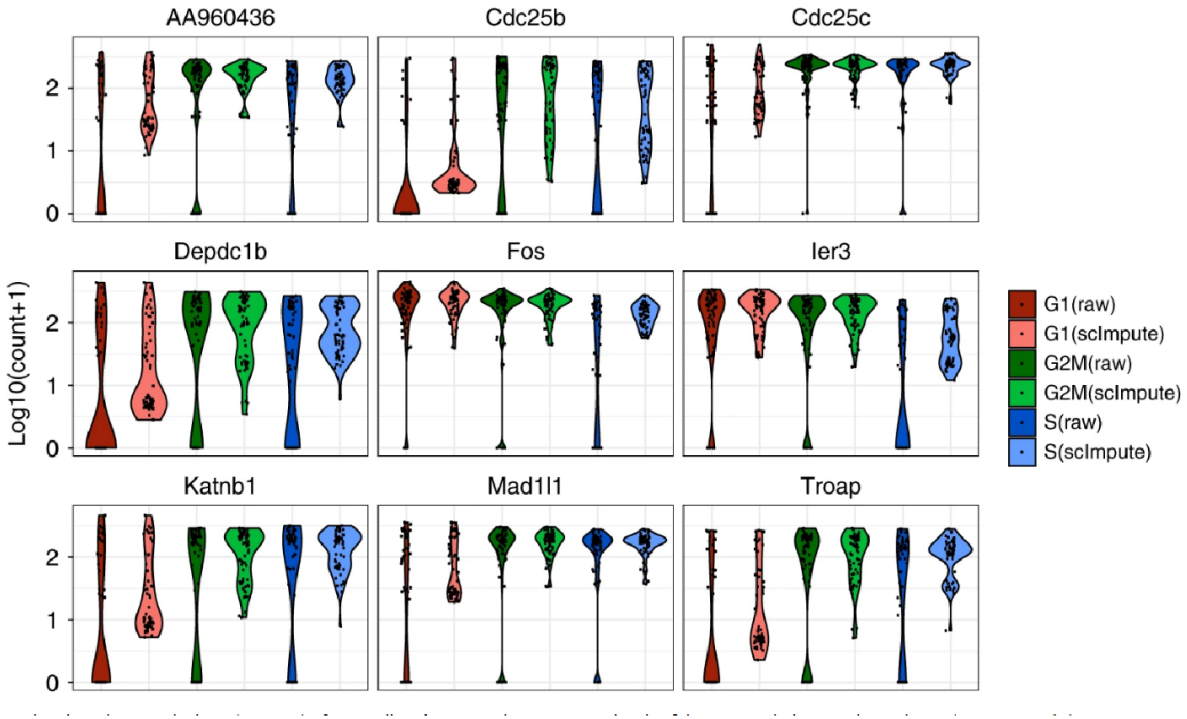


Figure 8: count of cell cycle gene before and after imputation from scImpute paper

Other ways to show their results

- Simulate the gene expression data from the scratch.
- Suppose there are 3 cell types, and only 810 genes are truly differentially expressed.
- Dropout for each gene follows a double exponential function.
- Similar to SAVER, carry out t-sne and first 2 pcs.

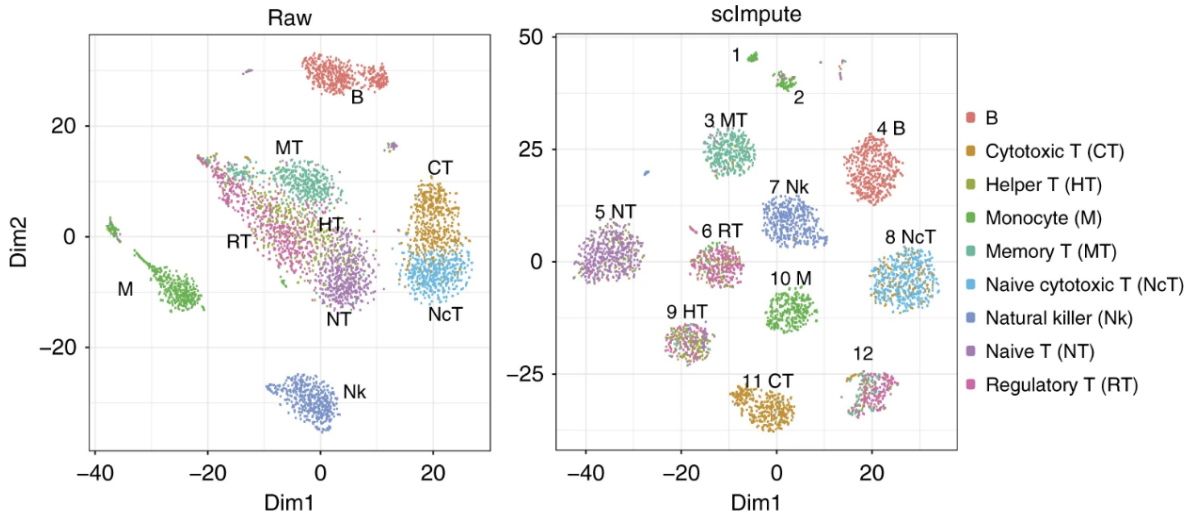


Figure 9: tsne from scImpute paper

Summary

- Impute only highly likely missing values
- Impute based on similar cells.
- Can use prior knowledge.
- May smooth cell stochasticity.
- Does not impute outliers.
- Scalability.

Deep Count Autoencoder

- Use Autoencoder to denoise model.
- One only need to choose appropriate noise model for count data.
- Either ZINB or NB

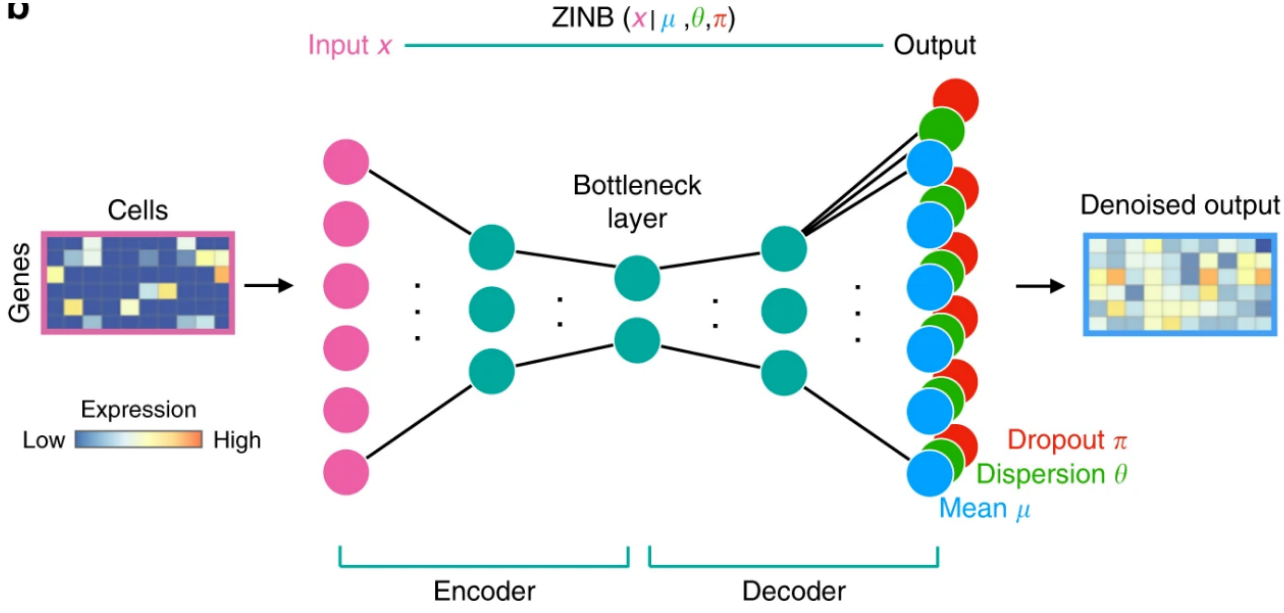


Figure 10: autoencoder structure from DCA paper

Simulation Result

- In DCA, they used package called 'Splatter' to simulate which provide both with and without dropout data.
- By computing likelihood ratio test of NB and ZINB fits the user can determine whether zero-inflation is present or not.
- MSE with normalized data does not work well.

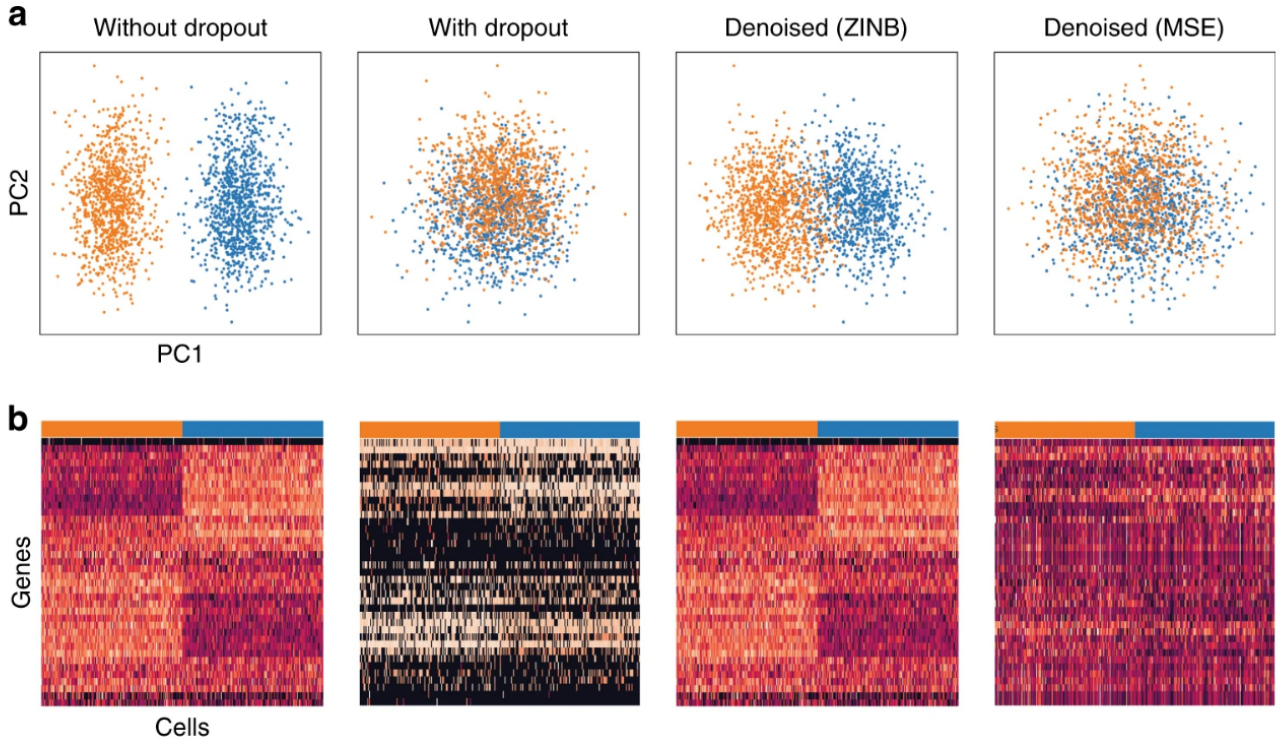
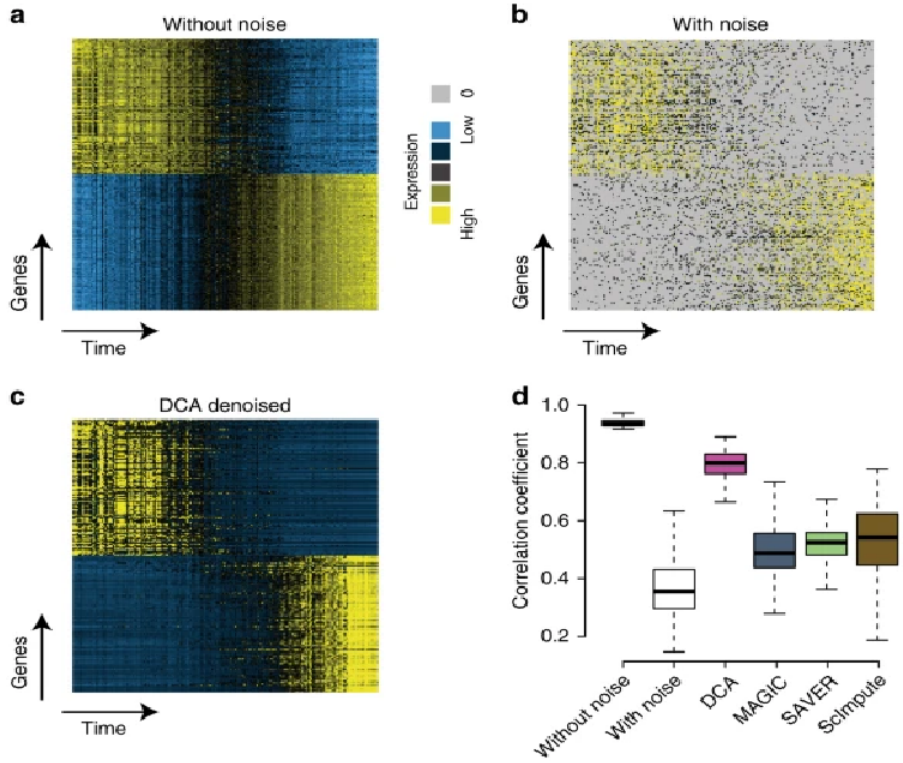


Figure 11: simresult from DCA paper

Using Bulk Cell

- Bulk cell data contains less noise than single cell and do not suffer much from dropout.
- Therefore, bulk cell data can provide good ground truth.
- Single cell specific noise was added by gene-wise subtracting values drawn form the exponential distribution such that 80% of values are 0.



Other method for evaluation

- CITE-seq enables simultaneous measurement of protein and RNA levels at cellular resolution.
- Per-cell protein levels are higher than mRNA levels so less prone to dropout events.
- We can use protein levels as ground truth.
- Some correlations between genes are already known. (regulatory correlation.)
- This correlation may not appear in noised data.
- However, after denoising using DCA we can observe these correlation again.

Scalability

- DCA scales linearly with number of cells

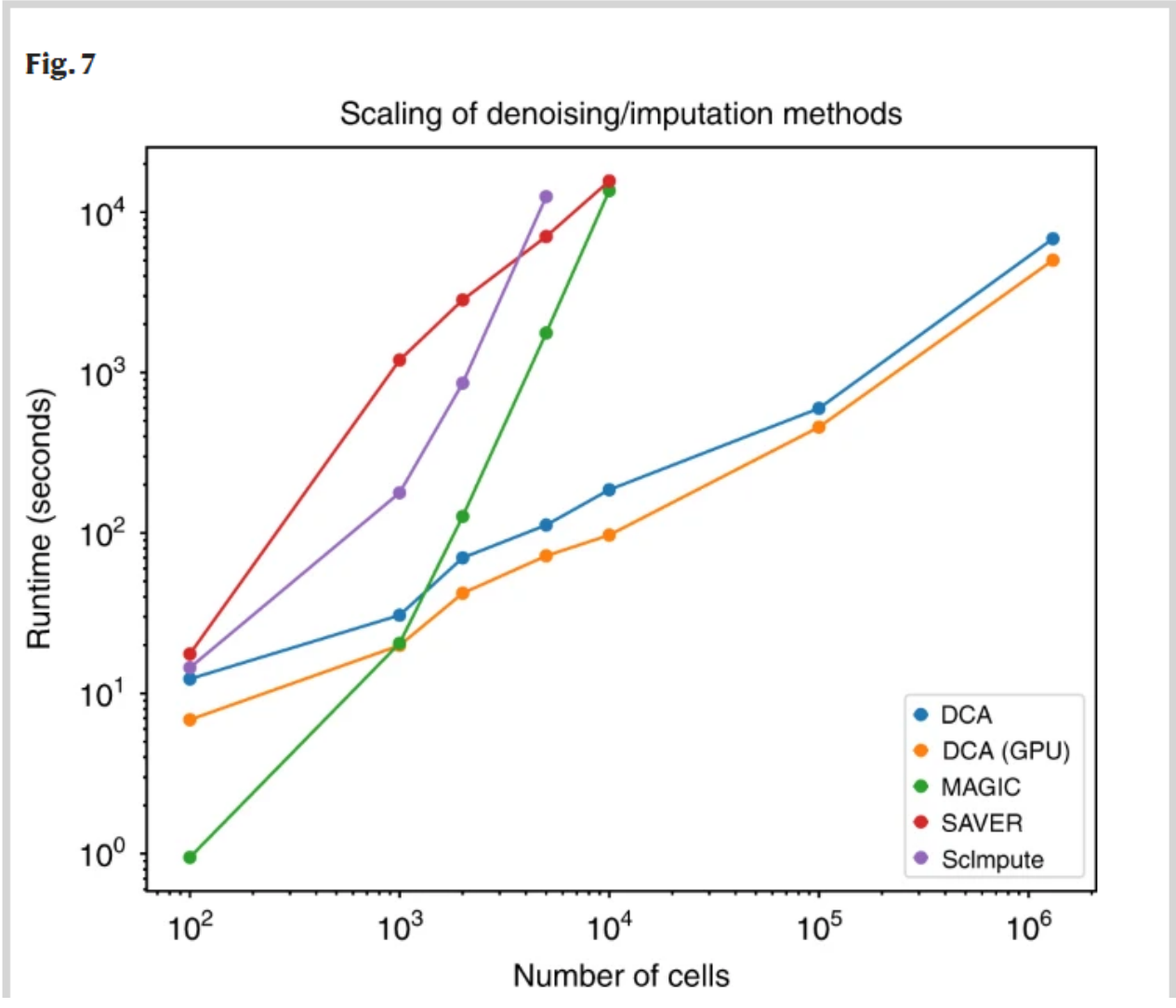


Figure 13: scale result from DCA paper

Reference

- SAVER: gene expression recovery for single-cell RNA sequencing by M Huang 2018
- An accurate and robust imputation method scImpute for single-cell RNA-seq data by WV Li 2018
- Single-cell RNA-seq denoising using a deep count autoencoder by G Eraslan 2019