COMPUTATIONAL ANALYSIS OF CELL-TO-CELL HETEROGENEITY IN SINGLE-CELL RNA-SEQUENCING DATA REVEALS HIDDEN SUBPOPULATIONS OF CELLS

Buettner et al., (2015) Nature Biotechnology, 1–32. doi:10.1038/nbt.3102

Saket Choudhary [1]

[1]University of Southern California

OUTLINE

- Motivation
- Methods
- Results
- Conclusion

MOTIVATION

SINGLE CELL SEQUENCING: CHALLENGES

- Limited amount of sample: Prone to noise and contamination
- Profiling hunderds of cells: How to identify subpopulations
- Identify regulatory landscape of each cell population
- Account for hidden confounding factors that might explain cell heterogeneity

EXPERIMENT

WHAT IS IT ALL ABOUT?

- Single cell transcriptomics heterogeneity: many single cells at the same time
- Accounting for technical noise was a solved problem; how do you account for *other* sources of variability: cell cycle, differentiation state etc.
- Given expression levels, how do you infer the effect of *latent* variables

Key focus: How does cell cycle affect expression levels? Given the apriori nature of genes(association with cell-cycle), is it possible to remove the effect of cell cycles?

GOALS

Separate out different sources of variation : *technical*, *biological*, *cell cycle*, *other hidden factors* with the idea of studying the underlying *interesting* biology.

Variations such as cell cycle can mask out more physiologically important differences.

Decomposition of this variance by splitting it for different confounding factors can really useful. If the original interest is in studying effect of differentiation, it makes sense to factor out the effect of cell cycle.

Discovering previously unidentified cell types?!

METHODS

METHOD IN A GIST

- Two step approach: Reconstruct state of unobserved factors; Use this information to perform correction over gene expression values
- Learn from a pool of well annotated gene sets, those related to cell cycle
- Uses bayesian technique to infer effect of latent variables
- Fit a *low-rank* covariance matrix to a set of **predefined** marker genes
- Estimate the state of hidden factors using maximum likelihood approach
- Decompose the variability of expression levels *across* single cells: expression = mean effect + random effect
- Regress out effects of hidden factors
- Corrected gene expression level

Details later...

RESULTS

CELL CYCLE AFFECTS GLOBAL GENE EXPRESSION

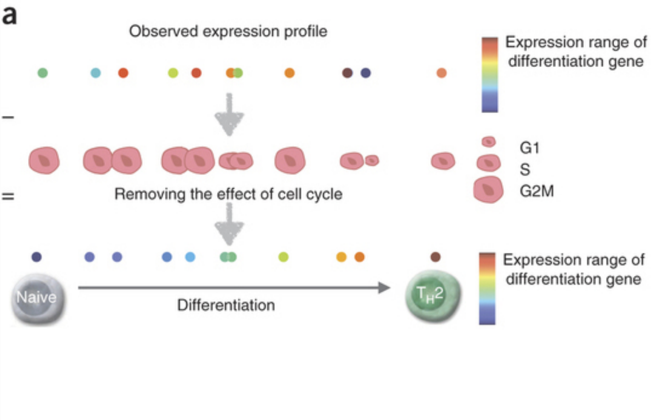
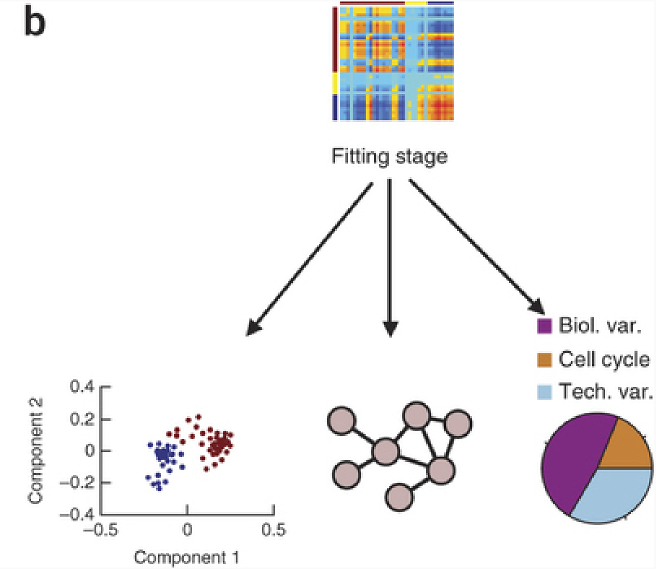

a
Observed expression profile
Expression range of differentiation gene
–
G1
S
G2M
=
Removing the effect of cell cycle
Expression range of differentiation gene
Naive
Differentiation
T_H2

Figure: Observed Expression = Effect of differentiation + Effect of state of cell($G1 > S > G2$)

VARIANCE MATRIX



RESULTS IN A GIST

- Before applying *scLVM* the cells looked like a variable population
- scLVM corrected expression data showed there existed two sub-populations
- These two sub-populations were infact found to be associated with T-cell differentiation stages

The method is not about single cell transcriptomics. It is a general approach to isolate, model and understand sources of variability

RESULTS IN A GIST

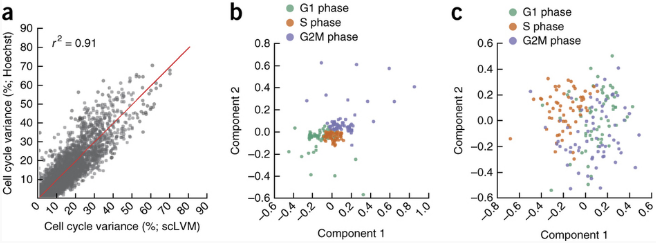


Figure:

RESULTS IN A GIST

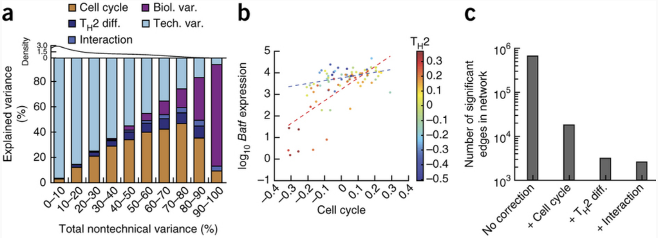


Figure:

VALIDATION

- Generate single-cel RNA-seq data from mouse embryonic stem cells(mESCs)
- The status of cell-cycle of each such cell is known apriori
- Perform scLVM fitting of the final expression values using an annotated gene set (from GO/Cellbase) known to be associated with cell cycle
- The final results are independent of the source of annotation(either GO or CellBase)
- The results were consistent even if the size of annotated genes was reduced to 10.

EFFICACY: HOW ABOUT JUST THROWING OUT THOSE KNOWN GENES?

If the genes are known to be associated with cell cycle, why not simple throw them out to rule out the effect of cell cycle?

A non-linear PCA of datasets with cell-cylce associated genes thrown out gave a clear separation, and this separation was later validated to be two different cell cycles $=>$ *scLVM* accounts for the latent factors which cannot be simply accounted by throwing away those *informative* genes.

TIATING t_h2 CELLS

- scLVM corrected gene expression levels have significantly low between cell correlations: Majority of the varianceis attributable to cell cycle
- Significantly corrleated genes post csLVM application were found to be involved in glycolysis and cellular response to IL-4 stimulus(triggers differentiation)
- To further validate: non linear PCA with and without scLVM correction. Without correction: no obvious subgroups
- One of the two sub-populations post scLVM correction were found to have genes that marked completion of differentiation

POSSIBILITES AND LIMITATIONS

- It is possible to account for more than one factor, as long as the corresponding annotated gene sets are available
- When multiple confounding factors are considered, in order to ensure robust analysis it is important to ensure the statistical significance if the factors are weak or nonindependent
- No formal tests exist for testing the presence of a particular factor

THE UNDERLYING MODEL

Let N = Number of cells
G = Number of variable genes(determined using a T-test on pre-processed count data) G_h = Set of marker genes(cell-cycle related) Y_h = $[y_1, y_2, ...y_h]$ = Vector of gene expressions where y_g represents gene g's expression acorss cells

$$Y_h = \mu + CU + XW + \psi \tag{1}$$

U, W = Weight of linear covariance model where C models Q known covariates and W models unknown covariates.

ψ models the rest of noise

C, X are determined using a bayesian approach assuming both U,W as gaussians prior.

$$P(Y_h|\sigma_u^2, \nu^2, X, C) = \pi_{g=1}^{G} N(y_g|\mu_g, \sigma_u^2 CC^T + XX^T + \nu^2) \tag{2}$$