The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells

Trapnell, Cole, et al (Nature biotechnology,2014)

Saket Choudhary

November 9, 2015

INTRODUCTION

RNA-SEQ

- rna-seq involved *direct* sequencing of transcripts

Pachter, Lior. "Models for transcript quantification from RNA-Seq." arXiv preprint arXiv:1104.3889 (2011).

RNA-SEQ

- rna-seq involved *direct* sequencing of transcripts
- Resolution at the level of individual isoform of genes

Pachter, Lior. "Models for transcript quantification from RNA-Seq." arXiv preprint arXiv:1104.3889 (2011).

RNA-SEQ

- rna-seq involved *direct* sequencing of transcripts
- Resolution at the level of individual isoform of genes
- Area with ample scope of insights: biological and computational(RNA-Seq is *challenging!*)

Pachter, Lior. "Models for transcript quantification from RNA-Seq." arXiv preprint arXiv:1104.3889 (2011).

RNA-SEQ

- rna-seq involved *direct* sequencing of transcripts
- Resolution at the level of individual isoform of genes
- Area with ample scope of insights: biological and computational(RNA-Seq is *challenging!*)
- NOTE(A common misbelief): RNA-Seq doesn't measure what is *technically* gene expression: Measures *relative transcript abundances*

Pachter, Lior. "Models for transcript quantification from RNA-Seq." arXiv preprint arXiv:1104.3889 (2011).

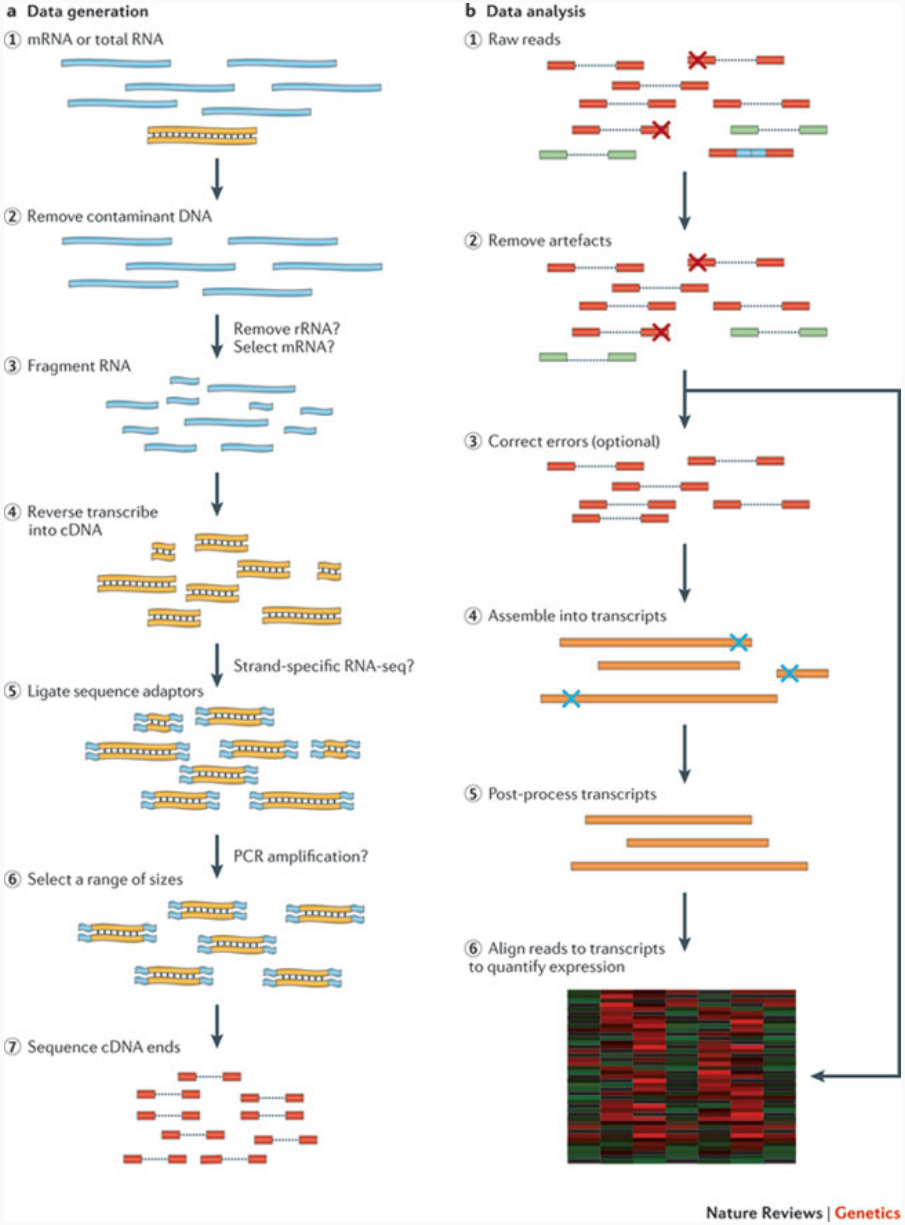


Figure 1: A typical RNA-Seq experiment

[1] Next-generation transcriptome assembly, *Martin et al. Nature Reviews Genetics(2011)*

SINGLE CELL RNA-SEQ MOTIVATION

’ Context: Studying differentiation

Transcriptional dynamics of a temporal process like cell differentiation is challenging

- Time-series analysis of bulk cell data : hard to distinguish early and late phases of transcriptional cascade

SINGLE CELL RNA-SEQ MOTIVATION

’ Context: Studying differentiation

Transcriptional dynamics of a temporal process like cell differentiation is challenging

- Time-series analysis of bulk cell data : hard to distinguish early and late phases of transcriptional cascade
- Difficult to capture cell-to-cell variability

SINGLE CELL RNA-SEQ MOTIVATION

’ Context: Studying differentiation

Transcriptional dynamics of a temporal process like cell differentiation is challenging

- Time-series analysis of bulk cell data : hard to distinguish early and late phases of transcriptional cascade
- Difficult to capture cell-to-cell variability

SINGLE CELL RNA-SEQ MOTIVATION

’ Context: Studying differentiation

Transcriptional dynamics of a temporal process like cell differentiation is challenging

- Time-series analysis of bulk cell data : hard to distinguish early and late phases of transcriptional cascade
- Difficult to capture cell-to-cell variability

High variability arises due to high cell-to-cell variability: Simple averages don’t work!

Which is a better treatement?

Treatments for kidney stones

	Treatment A	Treatment B
Small Stones	Group1: **93%(81/87)**	Group2: 87% (234/270)
Large Stones	Group3: **73%(192/263)**	Group 4 69% (55/80)

DIGRESSION

Which is a better treatement?

Treatments for kidney stones

	Treatment A	Treatment B
Small Stones	Group1: **93%(81/87)**	Group2: 87% (234/270)
Large Stones	Group3: **73%(192/263)**	Group 4 69% (55/80)

Hint: Is your sample size *enough* to draw causality relations?

Which is a better treatement?

Treatments for kidney stones

	Treatment A	Treatment B
Small Stones	Group1: **93%(81/87)**	Group2: 87% (234/270)
Large Stones	Group3: **73%(192/263)**	Group 4 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

Reversal of inequality! : Simpson's Paradox. Sizes of the groups being combined are not same. Large stones patients were offered *better* treatment A, small stones patients were offered treatment B.
$\implies$ Stone size is a 'confounding variable'.

METHOD

HYPOTHESIS

- RNA-Seq experiment constitutes a time-series: each cell is a discrete time point (during its differentiation)

HYPOTHESIS

- RNA-Seq experiment constitutes a time-series: each cell is a discrete time point (during its differentiation)
- Using an unsupervised algorithm 'Monocle', we want to study the temporal development of single cell

HYPOTHESIS

- RNA-Seq experiment constitutes a time-series: each cell is a discrete time point (during its differentiation)
- Using an unsupervised algorithm 'Monocle', we want to study the temporal development of single cell
- Cell type: Skeletal myoblasts $\Longrightarrow$ known to undergo well-characterised sequence of transcriptional changes

HYPOTHESIS

- RNA-Seq experiment constitutes a time-series: each cell is a discrete time point (during its differentiation)
- Using an unsupervised algorithm 'Monocle', we want to study the temporal development of single cell
- Cell type: Skeletal myoblasts $\Longrightarrow$ known to undergo well-characterised sequence of transcriptional changes
- Cultured in high serum medium, then shifted to low serum $\Longrightarrow$ induces differentiation

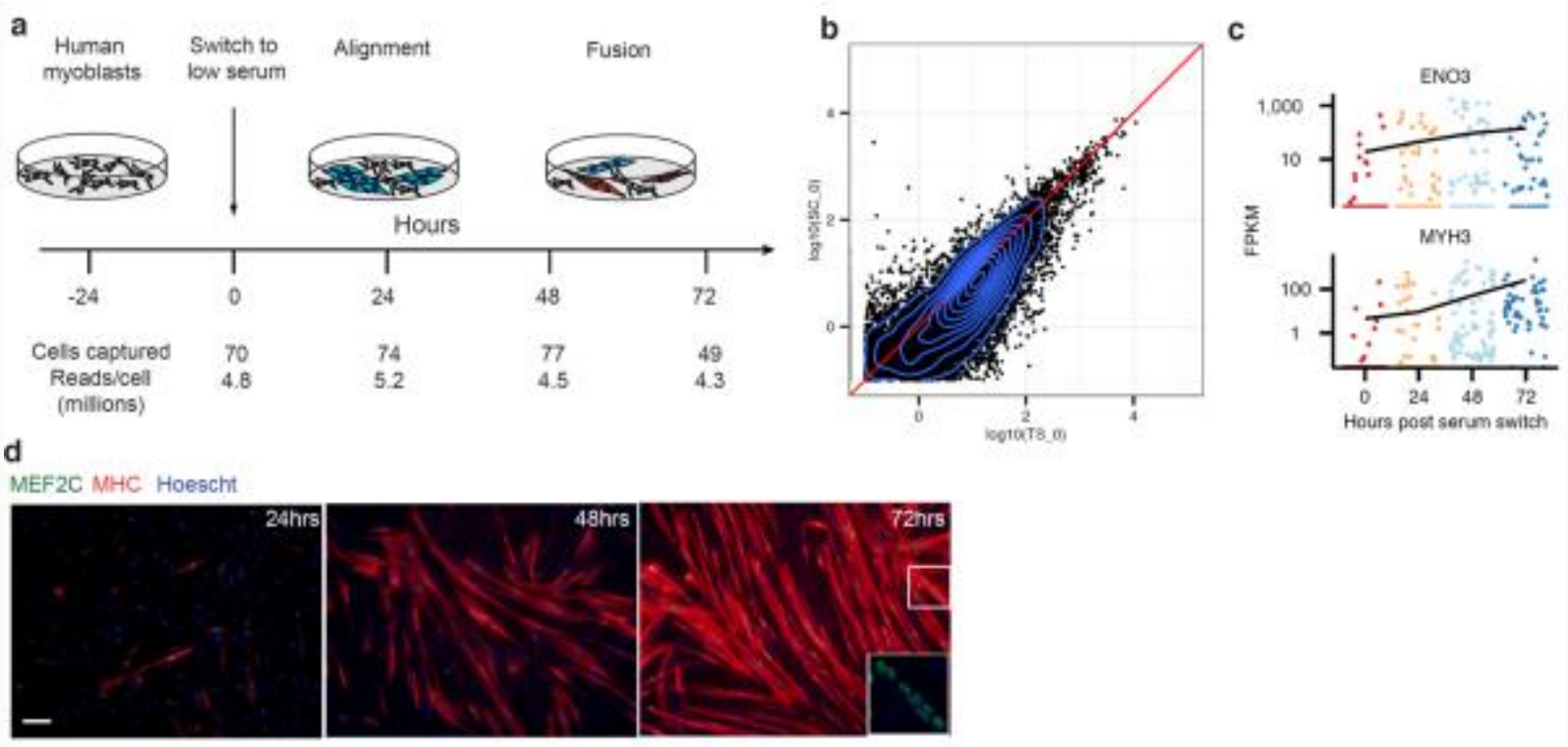


Figure 2: a. Cell Culturing b. Sanity check with bulk RNA-seq(time=0) c.Late stage differentiation markers d.

EXPERIMENT

- Sanity Check: Single Cell rna-seq should correlate with bulk

EXPERIMENT

- Sanity Check: Single Cell rna-seq should correlate with bulk
- *Known* markers such as ENO3, MYH3 are known to show increased expression with time

- Expression profile $\Longrightarrow$ Represented as points in Euclidean space $\mathbb{R}^d$ where d is the number of genes

- Expression profile $\Longrightarrow$ Represented as points in Euclidean space $\mathbb{R}^d$ where d is the number of genes
- Reduce dimension: Independent Component Analysis: Like PCA, but rather than maximising the variance, the projection ensures that the resulting data is one of the independent components of the data(Orthogonality still holds)

METHODS

- Expression profile $\Longrightarrow$ Represented as points in Euclidean space $\mathbb{R}^d$ where d is the number of genes
- Reduce dimension: Independent Component Analysis: Like PCA, but rather than maximising the variance, the projection ensures that the resulting data is one of the independent components of the data(Orthogonality still holds)
- Construct a Minimum Spanning Tree using these points in 2D.

METHODS

- Expression profile $\implies$ Represented as points in Euclidean space $\mathbb{R}^d$ where d is the number of genes
- Reduce dimension: Independent Component Analysis: Like PCA, but rather than maximising the variance, the projection ensures that the resulting data is one of the independent components of the data(Orthogonality still holds)
- Construct a Minimum Spanning Tree using these points in 2D.
- Find the longest path through the MST which corresponds to the long-sequence of transcriptionally similar cells(Essentially with non significant differential expression)

METHODS

- Expression profile $\implies$ Represented as points in Euclidean space $\mathbb{R}^d$ where d is the number of genes
- Reduce dimension: Independent Component Analysis: Like PCA, but rather than maximising the variance, the projection ensures that the resulting data is one of the independent components of the data(Orthogonality still holds)
- Construct a Minimum Spanning Tree using these points in 2D.
- Find the longest path through the MST which corresponds to the long-sequence of transcriptionally similar cells(Essentially with non significant differential expression)
- It is possible to create a trajectory using *pseudotime* values: So there are now branches + a main trajectory(ref figure)

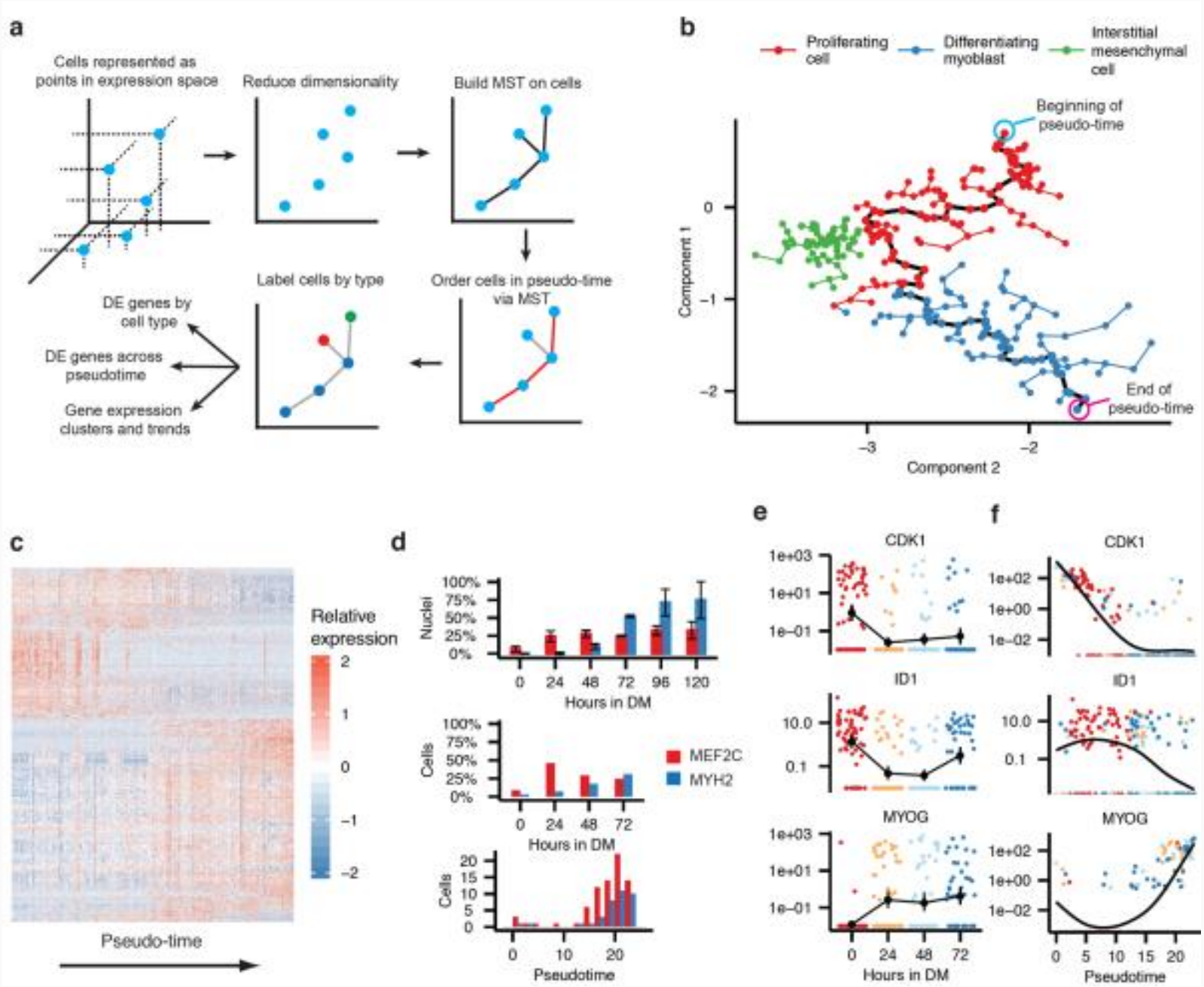


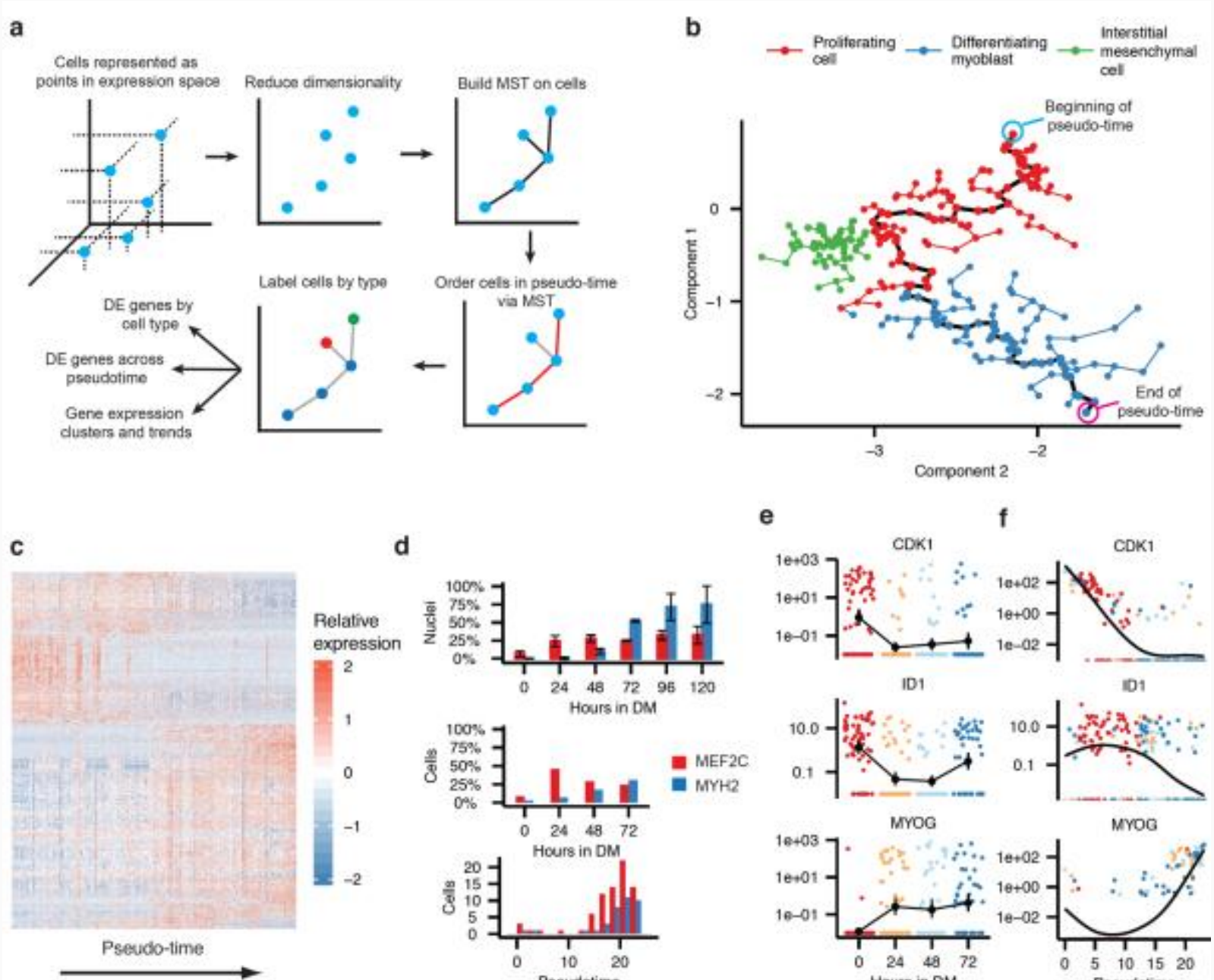
Figure 3: Results

RESULTS

METHOD/RESULTS

- Differentiation = Two phase trajectory(differentiating) + non-differentiating cells
- Trajectory 1: Cells selected under High Mitogen Condition(Mitogen enhances differentiation/division)(time = 0 cells)
- Trajectory 2: 24h, 46h, 72h differentiating cells (MYOG is a known marker for differentiation)
- Trajectory 3: Lacks myogenic markers, possibly did not originate from the myoblasts(known to be stimulants of muscle differentiation)

METHOD/RESULTS



RESULTS

- Sanity Check: Since the collection happened at 24h, 48h. 72h, it is possible to track the expression levels of markers: MEF2C, MYH2 by immunofluorescence and then compare it with pseudotime trajectory (ref Fig (d))

RESULTS

- Sanity Check: Since the collection happened at 24h, 48h. 72h, it is possible to track the expression levels of markers: MEF2C, MYH2 by immunofluorescence and then compare it with pseudotime trajectory (ref Fig (d))
- Outcome: Monocle enables reconstructing the temporal trajectory of differentiation while retaining the in vitro differentiation kinetics

RESULTS

- Sanity Check: Since the collection happened at 24h, 48h. 72h, it is possible to track the expression levels of markers: MEF2C, MYH2 by immunofluorescence and then compare it with pseudotime trajectory (ref Fig (d))
- Outcome: Monocle enables reconstructing the temporal trajectory of differentiation while retaining the in vitro differentiation kinetics
- New Insight: Find differentially expressed genes that would otherwise have been *lost* in bulk RNA-seq:

RESULTS

- Sanity Check: Since the collection happened at 24h, 48h. 72h, it is possible to track the expression levels of markers: MEF2C, MYH2 by immunofluorescence and then compare it with pseudotime trajectory (ref Fig (d))
- Outcome: Monocle enables reconstructing the temporal trajectory of differentiation while retaining the in vitro differentiation kinetics
- New Insight: Find differentially expressed genes that would otherwise have been *lost* in bulk RNA-seq:

RESULTS

- Sanity Check: Since the collection happened at 24h, 48h. 72h, it is possible to track the expression levels of markers: MEF2C, MYH2 by immunofluorescence and then compare it with pseudotime trajectory (ref Fig (d))
- Outcome: Monocle enables reconstructing the temporal trajectory of differentiation while retaining the in vitro differentiation kinetics
- New Insight: Find differentially expressed genes that would otherwise have been *lost* in bulk RNA-seq:

General Additive models:

$$g(E(y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)$$

where x_i are the predictor variables. and g is link function(log/identity) and Y belong to the exponential family, the advantage over GLM being flexibility with nonparametric fits. (ofcourse, less interpretable than GLM)

MORE VALIDATION

- A further validation step involved *clustering* genes with *[similar]* expression levels, with the assumption **similar trends in expression = similar biological function**
- Genes downregulated early or upregulated late were found to be GO enriched in myosis, cell cycle-exit, activation of muscle specific proteins

MORE VALIDATION

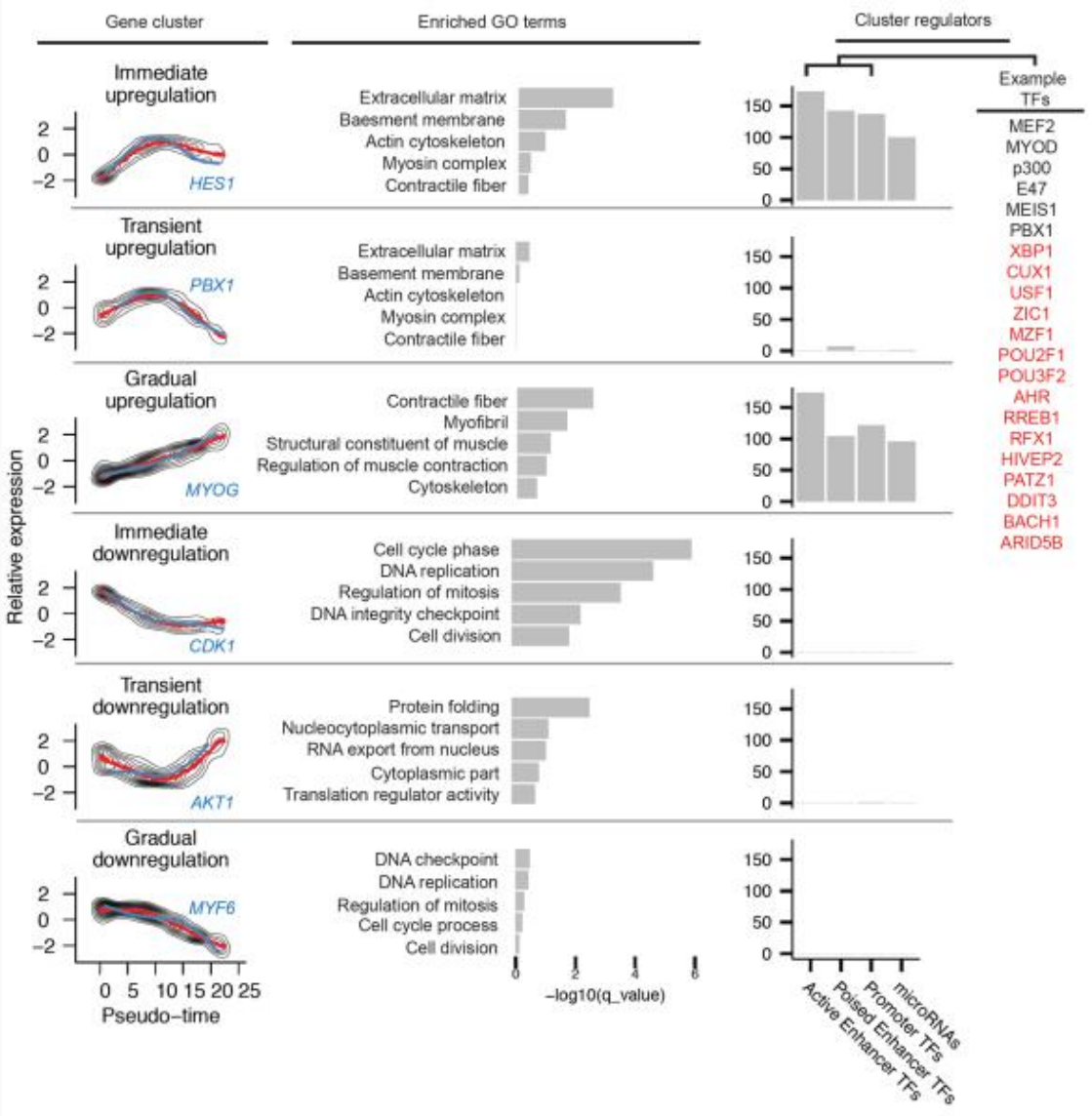


Figure 4: GO analysis of gene clusters

CONCLUSION

SUMMARY

- Monocle takes an unsupervised approach to determine the temporal trajectory for differentiating cells
- Makes it possible to identify the *latent variables* that often get shadwoed with bulk RNA-Seq studies
- The GO validation step is not very convincing, the method otherwise looks solid(there were more experimental validations performed)

Yanai, Itai, et al. "Similar gene expression profiles do not imply similar tissue functions." TRENDS in Genetics 22.3 (2006): 132-138.

Questions?