Short Course: Statistical Methods for Single-cell RNA-seq Analysis
Joint Statistical Meetings, Vancouver, 2018

1:00-1:20: Logistics and Outline (CKendziorski)

1:20-1:50: Overview of scRNA-seq technologies and QC (RBacher)

1:50-2:20 Single-cell expression distributions (NZhang)

2:20-2:50: Normalization methods including SCnorm (RBacher)

2:50-3:20: Analysis of allele-specific gene expression (MLi)

3:20-3:50: Single-cell expression denoising and imputation (NZhang)

3:50-4:20: Bulk tissue cell type deconvolution with scRNA-seq gene expression reference (MLi)

4:20-4:50: Identifying differential distributions and pseudotime reordering (CKendziorski)

4:50: Closing comments/ questions/ discusssion

https://github.com/rhondabacher/JSM2018_ShortCourseMaterials

Single-cell RNA-seq technologies and quality control

Rhonda Bacher
Department of Biostatistics
University of Florida

JSM 2018
Short Course

UF UNIVERSITY of FLORIDA

Single-cell RNA-seq platforms

Fluidigm C1

Drop-seq
inDrop (similar)

10X Chromium

Bio-Rad ddSEQ

2014 2015 2017

…

What are people using?

- Google scholar since Jan. 1, 2018:
  - 50 papers for **10X Chromium.**
  - 137 papers for **Fluidigm C1.**
  - 162 papers for **Drop-seq.**

- bioRxiv since Jan. 1, 2018:
  - 71 results for **10X Chromium.**
  - 34 results for **Drop-seq.**
  - 23 results for **Fluidigm C1.**

(As of June 22, 2018)

Platforms and Protocols

- Single-cell *platforms/instruments* isolate cells, extract mRNA, and efficiently prepare mRNA for sequencing.
- Single-cell *protocols* relate to specific chemistries used to prepare the mRNA for sequencing.
- Platforms may be protocol specific (proprietary) or allow multiple protocols (open-source).

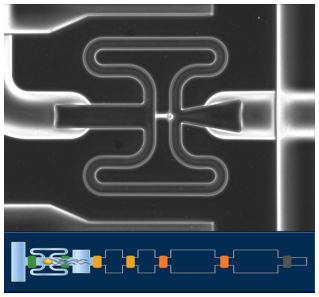
Drop-seq
inDrop
Fluidigm C1

vs.

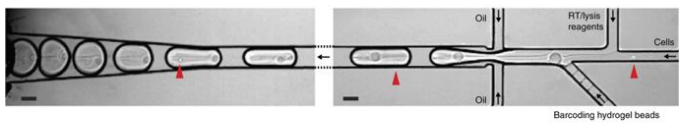
10X
Bio-rad

Differences in cell capture

- **Microwells**: high precision; low throughput



- **Droplets:** low precision, high throughput



Key platform differences

- Throughput: Number of cells prepared in a fixed amount of time.
  - C1 Fluidigm: 10^2 - 10^3
  - Droplet methods: 10^3-10^5
- Cost: Cost per cell (including all reagents and sequencing).

Cost

Key platform differences

- Initial cell numbers required:
  - C1 Fluidigm: 200 – 1000
  - Droplet methods: 500 – 20,000
- Cell capture efficiency: Percent of cells successfully isolated.
  - C1 Fluidigm ~ 90% (size-specific selection)
  - Drop-seq ~ 10%
  - 10X ~ 50%

Key protocol differences

- Molecular identifiers (UMI): Combination of cell- and mRNA- specific barcodes allow for a unique transcript count.
- Transcript data: Full length or 3’ end only.
  - Use of UMI’s currently prevents full length sequencing of the transcript.
- Droplet platforms tend to use UMIs.

Additional differences

- Different biological applications lend themselves to different protocols.
  - Isoform analysis and allele specific expression require full length transcripts.
  - Droplet based methods are not ideal for precious samples with a small number of cells.
- All methods detect highly and even moderately expressed genes well.
- Experimental designs, pre-processing, and quality control are technology specific.

Reviews/Comparisons of technologies

- Baran-Gale, J., Chandra, T., & Kirschner, K. (2017). Experimental design for single-cell RNA sequencing. *Briefings in functional genomics*.
- Svensson, V., Vento-Tormo, R., & Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nature protocols*, *13*(4), 599.
- Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., ... & Enard, W. (2017). Comparative analysis of single-cell RNA sequencing methods. *Molecular cell*, *65*(4), 631-643.

Quality Control

Bacher, Rhonda, and Christina Kendziorski. "Design and computational analysis of single-cell RNA-sequencing experiments." *Genome biology* 17.1 (2016): 63.

Quality Control

QC on reads:

- Total reads per sample
- Base quality scores
- GC content
- Adapter sequences
- Over represented sequences

QC on alignment:

- Total transcripts
- Uniquely mapping reads
- Reads mapping to mitochondria
- Reads mapping to spike-ins
- Coverage bias

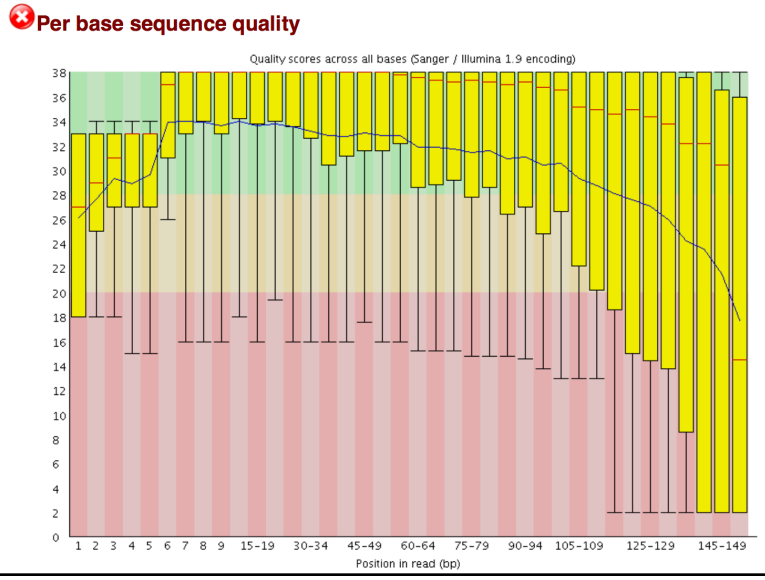
QC across cells:

- Batch effects
- Detection rate

QC on reads

- Quality control on reads for non-UMI single-cell RNA-seq can use similar tools as those for bulk, including:
  - FASTQC
  - Kraken
  - RNA-SeQC
- Low-quality scores in the initial positions of many reads may indicate a problem with the sequencing run.
- A decrease in quality in the last positions often indicates a general degradation. Reads may be trimmed before aligning.

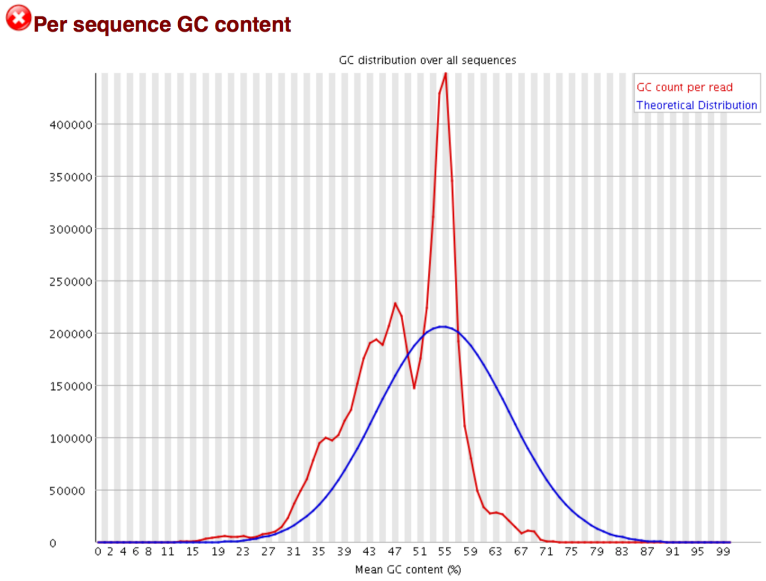
QC on reads—example



QC on reads

- High quality libraries should be free of biases in:
  - Nucleotide composition: proportion of nucleotides across each position should be evenly distributed.
  - Overall GC content: should be normally distributed.
  - Overrepresented sequences: large duplicated rates usually indicates bias in PCR or a library with low complexity.

QC on reads—example



QC on reads

- UMI data requires additional processing and is often platform specific. Differences across platforms include:
  - UMI length.
  - Order of read structure.
  - Barcode errors.

- UMI-specific software tools for Drop-seq and inDrop:
  - UMI-tools
  - zUMIs
  - PoissonUMIs
  - DropRNA

QC on aligned reads

- Ensure each cell index represents mRNA from a **single high-quality** cell. Identify and remove:
  - Empty cells
  - Doublet cells
  - Degraded/Broken cells
  - Leaky cells
- Visual inspection of capture sites in Fluidigm C1 is possible, but may not always be feasible especially for large scale experiments.

QC on aligned reads

- Metrics that correlate with 'low-quality' cells common that are common to all protocols:
  - Lower % reads aligned.
  - Higher % reads mapping to spike-ins.
  - Higher % reads mapping to mitochondria.
  - Lower % features detected.
- QC software tools:
  - cellity
  - scater

QC on aligned reads

- High quality cells should have reads distributed across many genes.

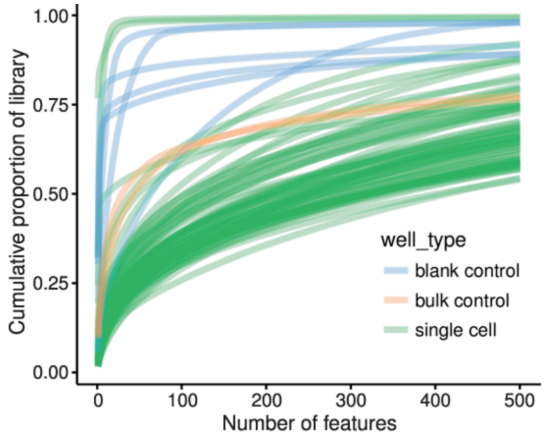
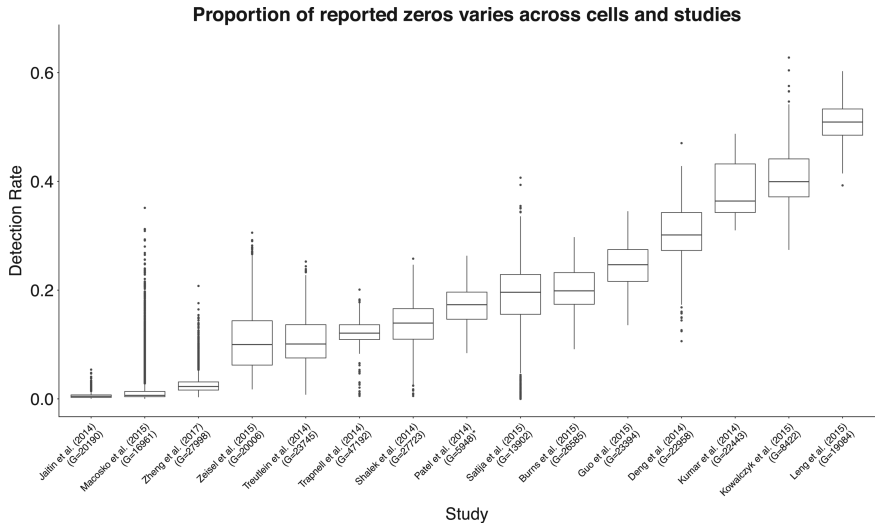


Figure 2a of McCarthy, D. J., Campbell, K. R., Lun, A. T., & Wills, Q. F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, *33*(8), 1179-1186.

QC across cells—detection rate



From Hicks, Stephanie C., et al. "Missing data and technical variability in single-cell RNA-sequencing experiments." *Biostatistics* (2017).

QC across cells—batch effects

- The process of generating single-cells inevitably leads to processing in batches.
- For different biological conditions, cells must be captured separately in order to preserve their identity.
- Careful experimental design is required to avoid confounding of biological signals with batch effects.
  - Replication of biological conditions.
  - Multiplexing of cells during capture.

QC across cells—batch effects

- Two methods for cell multiplexing across conditions to mitigate batch effects:
  - demuxlet: Multiplex genetically diverse cells across individuals then demultiplex using genotypes.
  - Add cellular bar codes to different samples, demultiplex after sequencing (https://www.biorxiv.org/content/early/2017/12/21/237693).

QC across cells—batch effects

- Methods for batch effect correction or data integration (across protocols):
  - MNN Batch Correction: https://www.nature.com/articles/nbt.4091
  - seurat: https://www.biorxiv.org/content/early/2017/07/18/164889

The Gene Expression Distribution

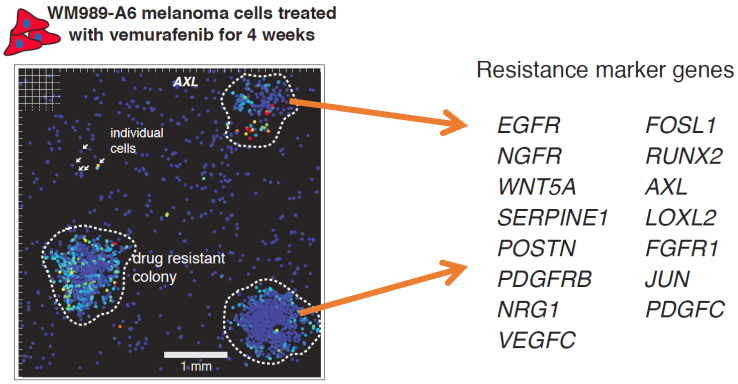
Short Course: Statistical methods for single-cell RNA-seq analysis
Joint Statistical Meetings, Vancouver, 2018

Nancy R. Zhang
Dept. of Statistics
University of Pennsylvania

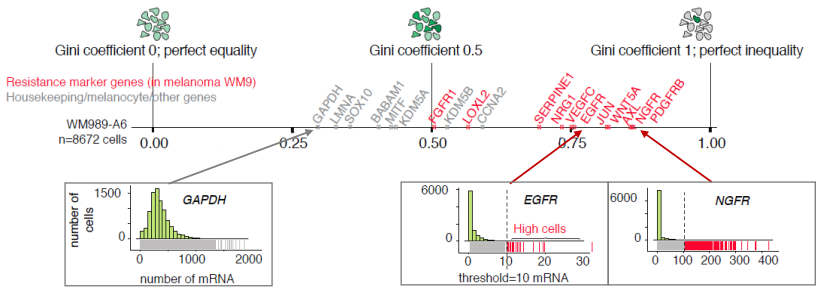
Motivating Example

Drug Resistance in Melanoma

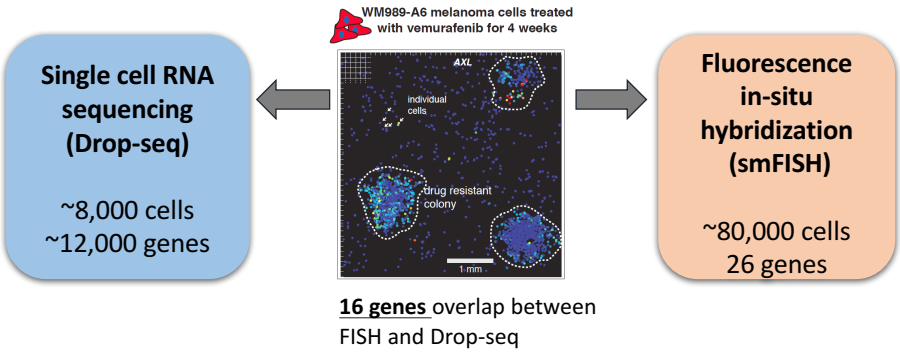
RNA-level heterogeneity → drug-resistance in melanoma (smFISH study by Shaffer et al., 2017)



RNA-level heterogeneity → drug-resistance
(smFISH study by Shaffer et al., 2017)



DROP-Seq data from same melanoma cell line

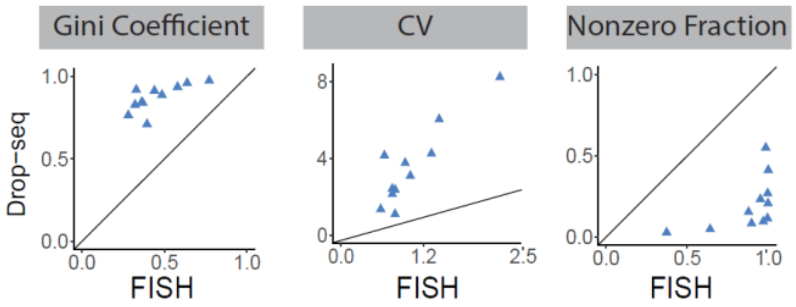


16 genes overlap between FISH and Drop-seq

FISH image from Shaffer et al. (2017)

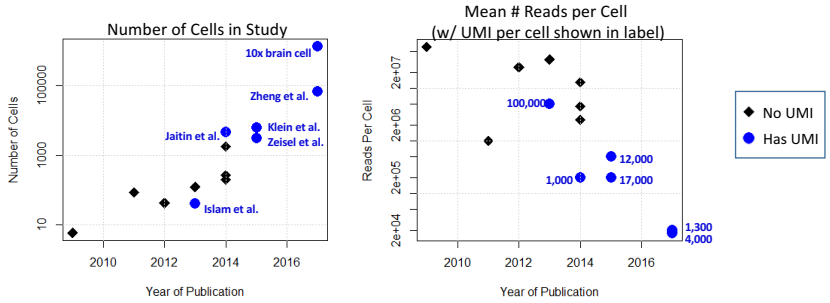
Drop-seq data from Torre & Dueck(2018)

Comparing FISH and scRNA-seq data:



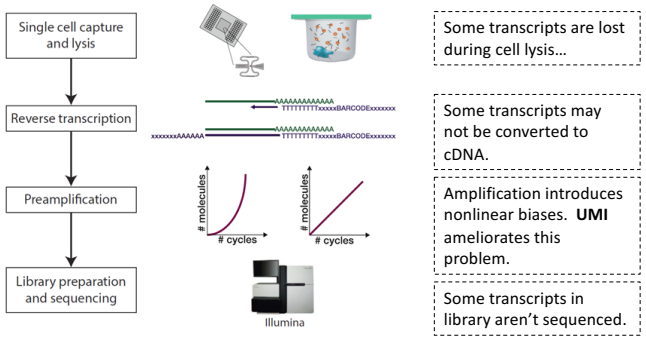
Distribution Recovery in scRNA-seq

Massively-parallel single cell sequencing



~**200,000** RNA molecules in a typical mammalian cell (*Shapiro, Biezuner and Linnarsson, 2013*)

Experimental Procedure



Some transcripts are lost during cell lysis…

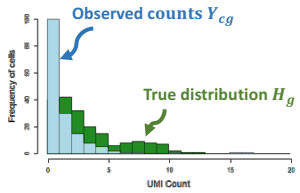
Some transcripts may not be converted to cDNA.

Amplification introduces nonlinear biases. **UMI** ameliorates this problem.

Some transcripts in library aren't sequenced.

~ 60 to > 90% zeros, some are real zeros, some are "drop outs"

True versus Observed Expression Distributions



Observed UMI counts:

$$Y_{cg} \sim F_{cg}(\lambda_{cg}), \qquad \lambda_{cg} \sim H_g$$

True count

Noise distribution:

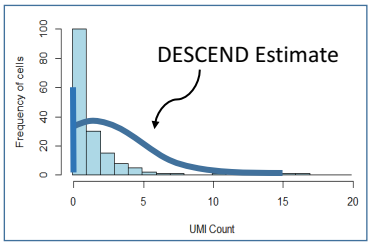
$$F_{cg}(\lambda_{cg}) = Poisson(\alpha_c \lambda_{cg})$$

True gene expression distribution:

Exponential family with spline basis

$$H_g = \begin{cases} \tilde{H}_g, & p_g \\ 0, & 1 - p_g \end{cases}$$

True versus Observed Expression Distributions



Nonzero fraction: p_g

Nonzero mean: $E[\lambda_{cg} | \lambda_{cg} > 0]$

Observed UMI counts:

$$Y_{cg} \sim F_{cg}(\lambda_{cg}), \qquad \lambda_{cg} \sim H_g$$

True count

Noise distribution:

$$F_{cg}(\lambda_{cg}) = Poisson(\alpha_c \lambda_{cg})$$

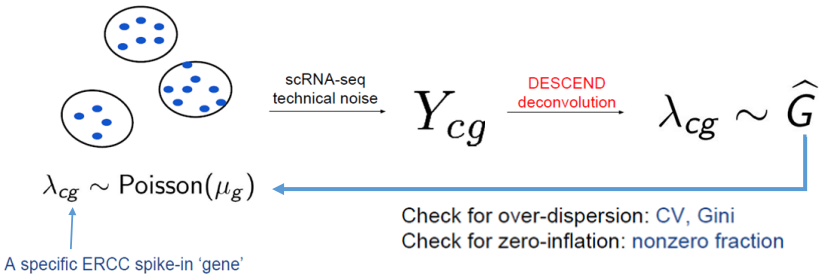
True gene expression distribution:

Exponential family with spline basis

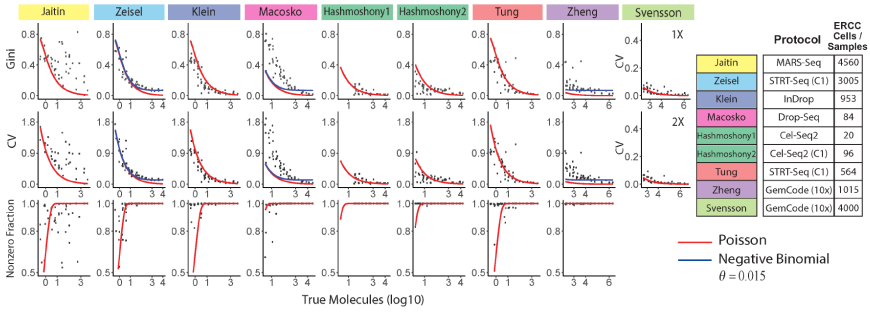
$$H_g = \begin{cases} \tilde{H}_g, & p_g \\ 0, & 1 - p_g \end{cases}$$

Checking the technical noise model

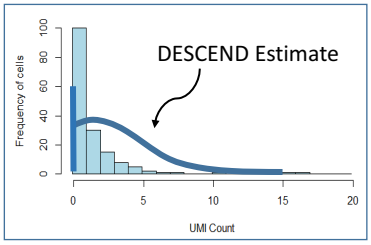
The assumption that $Y_{cg} \sim Poisson(\alpha_c \lambda_{cg})$ is critical.
How do we check it?



Is technical noise Poisson (after controlling for efficiency) ?



Cell-level covariates: U_c (e.g. cell size, cell type)



Nonzero fraction: p_g

Nonzero mean: $E[\lambda_{cg}|\lambda_{cg} > 0]$

Observed UMI counts:

$$Y_{cg} \sim F_{cg}(\lambda_{cg}), \qquad \lambda_{cg} \sim H_{gc}$$

True count

True gene expression distribution:

$$\lambda_{gc} = \begin{cases} \tilde{\lambda}_{gc}, & p_{gc} \\ 0, & 1 - p_{gc} \end{cases}$$

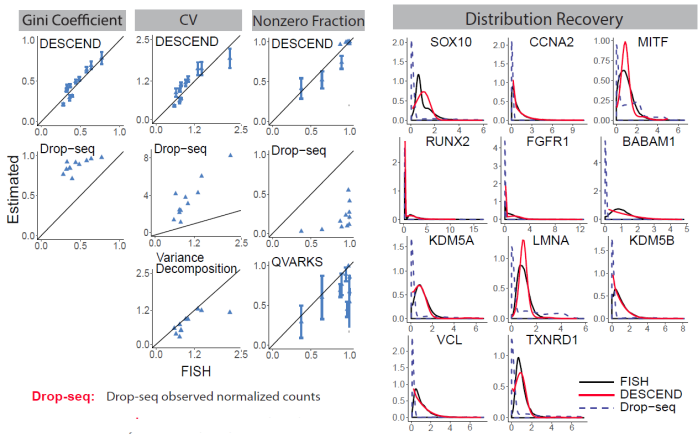
$$logit(p_{gc}) = \beta_{0g} + \beta_g U_c$$

$$\log(\tilde{\lambda}_{gc}) = \alpha_g U_c + \epsilon_{gc}$$

Spline exponential family

Benchmarking and Validation

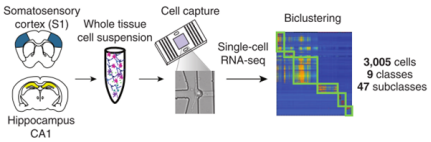
Comparisons to RNA-FISH on Melanoma Cell Line

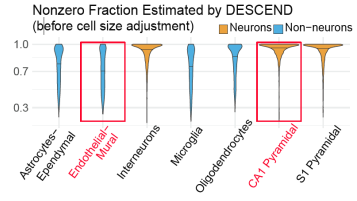


Case Studies

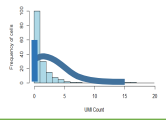
Differential Bursting Analysis

Figure 1A from Zeisel et al. (2015)



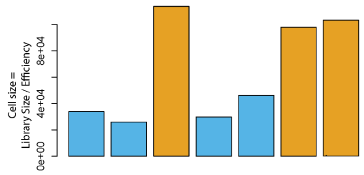


DESCEND deconvolution for each gene:

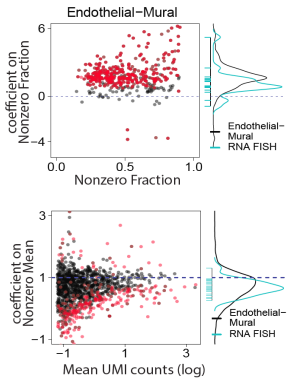


Total RNA abundance for each cell:

$$\frac{Total\ UMI\ barcode\ count}{Efficiency}$$



Cell-level covariates: U_c (e.g. cell size, cell type)



Observed UMI counts:

$$Y_{cg} \sim F_{cg}(\lambda_{cg}), \qquad \lambda_{cg} \sim H_{gc}$$

True count

True gene expression distribution:

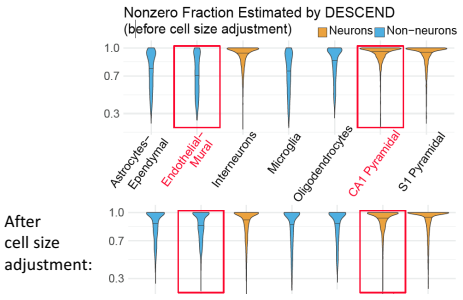
$$\lambda_{gc} = \begin{cases} \tilde{\lambda}_{gc}, & p_{gc} \\ 0, & 1 - p_{gc} \end{cases}$$

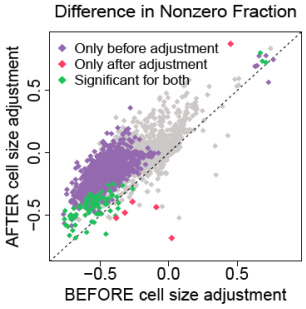
$$logit(p_{gc}) = \beta_{0g} + \beta_g U_c$$

$$\log(\tilde{\lambda}_{gc}) = \alpha_g U_c + \epsilon_{gc}$$

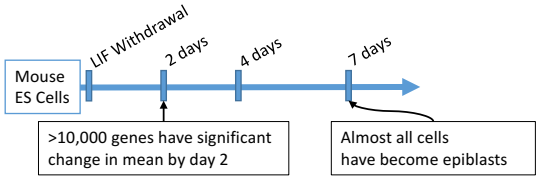
Spline exponential family

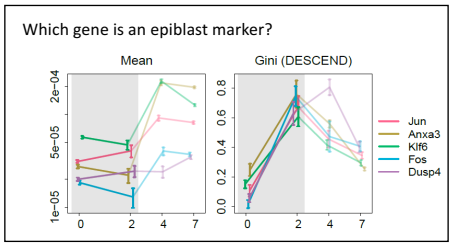
After adjusting for cell size

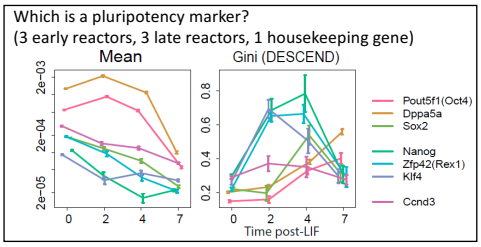




Marker Gene Identification (Klein et al. 2015)







Summary

- Noise for UMI data can be modeled by $Poisson(\alpha_c \lambda_{cg})$
- Lots of zeros seen in single cell data are due to technical loss.
- Adjusting for cell size differences is important during differential expression analyses
- Gene-level dispersion estimates, such as Gini index, are useful for marker identification.

DESCEND: Wang et al. (2018) Gene Expression Distribution Deconvolution in Single Cell RNA Sequencing, *to appear in* PNAS.

https://www.biorxiv.org/content/early/2017/12/01/227033

https://github.com/jingshuw/descend

Acknowledgements

Mo Huang, Statistics, University of Pennsylvania

Jingshu Wang, Statistics, University of Pennsylvania

Mingyao Li, Biostatistics, University of Pennsylvania

Arjun Raj and John Murray Labs, University of Pennsylvania

Normalization methods for single-cell RNA-seq data

Rhonda Bacher
Department of Biostatistics
University of Florida

JSM 2018
Short Course

UF UNIVERSITY of FLORIDA

Why do we need normalization ?

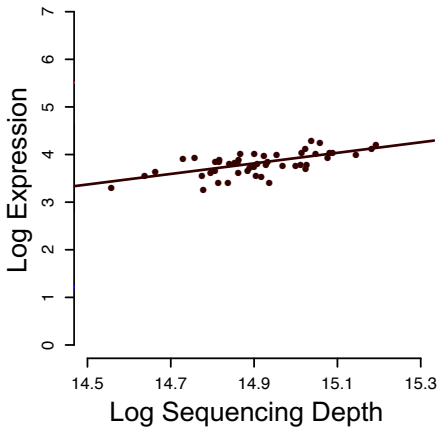
- Given two cells of scRNA-seq data we can not directly compare the expression of a gene across the cells.
- Some cells are sequenced more than other cells.
- A cell sequenced twice as much will on average have twice as high expression for every gene.

Data

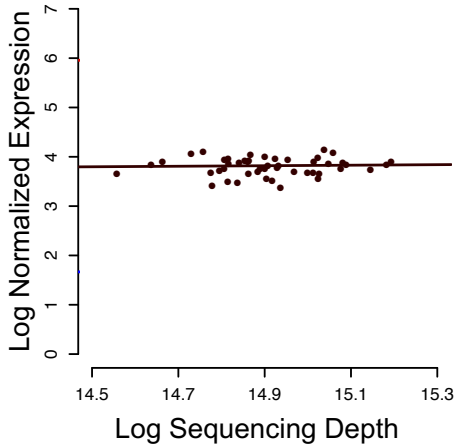
- Let Y be a matrix of expression estimates, for m genes and n cells (g = 1,…,m and j = 1,…,n).

	Cell 1	Cell 2	…	Cell n
Gene 1	62	124	…	42
Gene 2	10	20	…	10
Gene 3	316	632	…	322
…	…	…	$Y_{g,j}$	…
Gene m	85	170	…	73
Sequencing Depth	$\sum_{g=1}^{m} Y_{g,1}$	$\sum_{g=1}^{m} Y_{g,2}$	…	$\sum_{g=1}^{m} Y_{g,n}$

Count-depth relationship in bulk RNA-seq data



Count-depth relationship in bulk RNA-seq data—after normalization



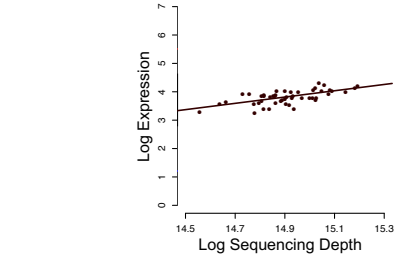
Normalization methods

- Normalization typically adjusts for differences in sequencing depth by calculating global scale factors (one per cell).

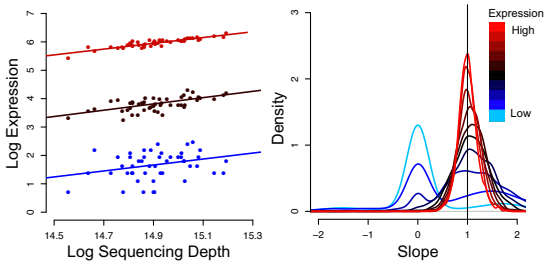
- Two ways:
  - Counts Per Million (CPM): $SF_j = \frac{10^6}{\sum_{g=1}^{m} Y_{g,j}}$
  - Median-Ratio (MR): $SF_j = \underset{g}{\text{median}} \frac{Y_{g,j}}{(\prod_{j=1}^{n} Y_{g,j})^{1/n}}$
- Normalized expression is given as:

$$Y'_{g,j} = Y_{g,j}/SF_j$$

Count-depth relationship in bulk RNA-seq data



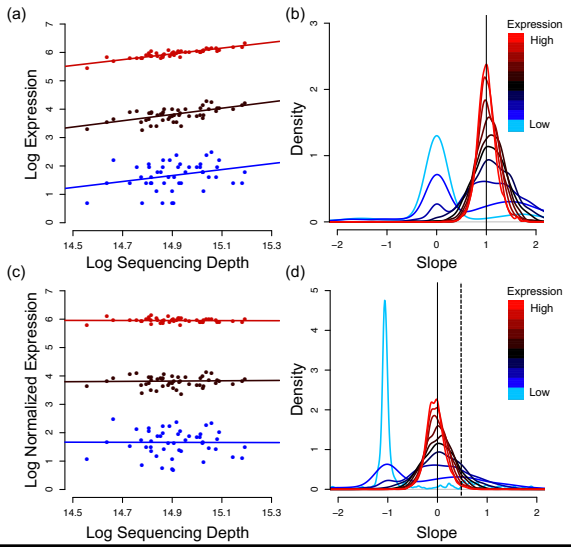
Unnormalized:



Count-depth relationship in bulk RNA-seq data—after normalization

Unnormalized:

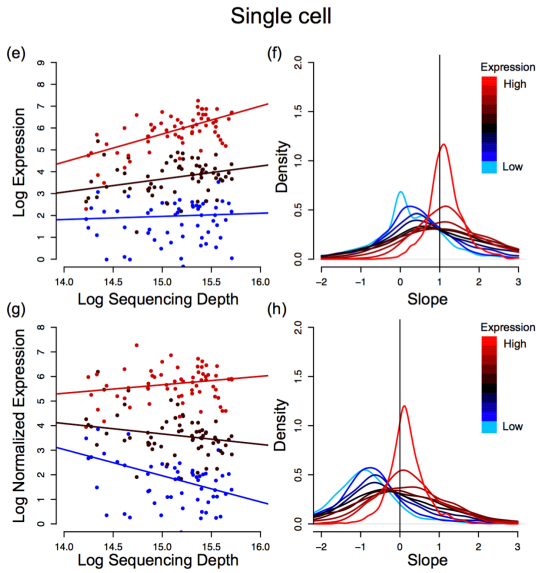
Global scale factor:



Count-depth relationship varies in scRNA-seq data

Unnormalized

Global scale factor:



Methods developed for single-cell normalization

- BASiCS: uses control transcripts known as spike-ins to estimate scale factors per cell.
- scran: pools cells and sums counts within each pool to obtain stable scale factor estimates.
- SCDE: internal normalization of sequencing depth while accounting for technical variability.

All are still estimating *global* scale factors!

SCnorm: robust normalization of single-cell RNA-seq data

Bacher R, Chu LF, Leng N, Gasch AP, Thomson JA, Stewart RM, Newton M, Kendziorski C. "SCnorm: robust normalization of single-cell RNA-seq data." *Nature methods*. 2017 Jun;14(6):584.

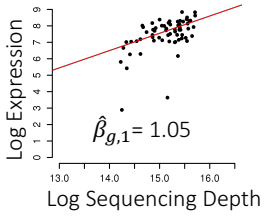
SCnorm: An Overview

- Step 1: Quantify each gene's relationship with sequencing depth (count-depth relationship) and cluster genes into k groups.
- Step 2: Estimate within group scaling factors and normalize each group separately.
- Step 3 : Evaluate the sufficiency of *k* groups.
  - If the evaluation suggests more groups are needed, step 2 is repeated using *k* + 1 groups until convergence.

SCnorm: Step 1

- $Y_{g,j}$ = log expression for gene g in cell j (*g = 1…m* and *j = 1…n*).
- X_j = log sequencing depth for cell *j* ($X_j = \log \sum_{g=1}^{m} e^{Y_{g,j}}$).
- Identify gene groups based on their count-depth relationship, estimated as $\hat{\beta}_{g,1}$ using median quantile regression:

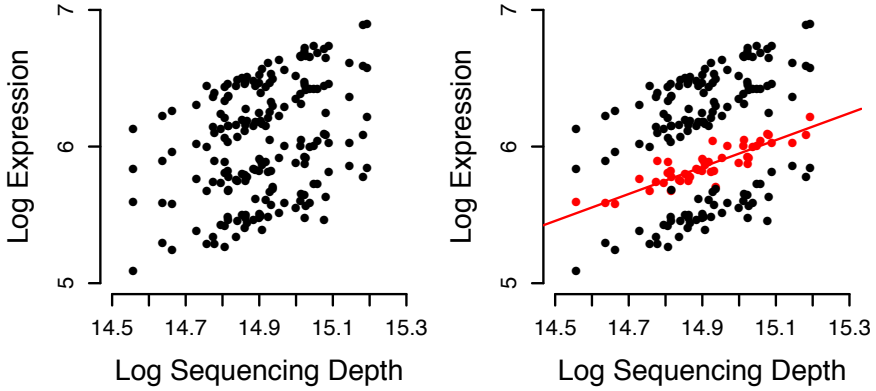
$$Q^{.5}(Y_{g,j}|X_j) = \beta_{g,0} + \beta_{g,1}X_j$$



- Genes are grouped using the *K*-medoids algorithm.

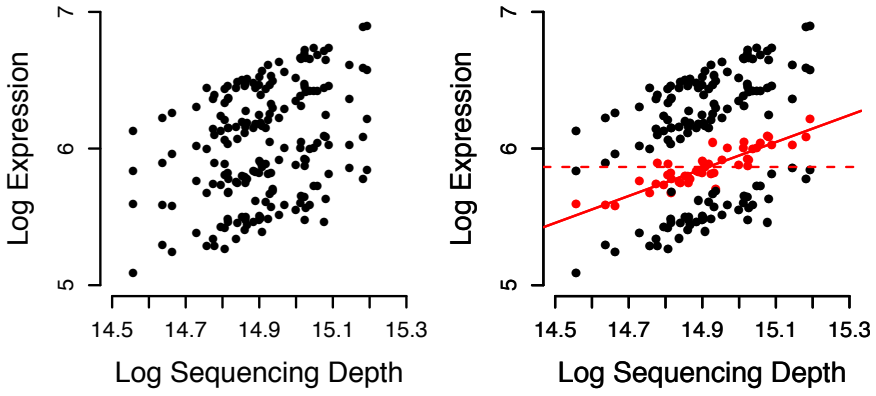
SCnorm: Step 2

- Within each group (ideally):



SCnorm: Step 2

- Within each group (ideally):



SCnorm: Step 2

- Within each group :
  - A τ^{th} quantile polynomial regression of degree *d* is fit over all genes in the group:

$$Q^{\tau_k,d_k}\left(Y_{g_k,j}\middle|X_j\right) = \beta_0^{\tau_k} + \beta_1^{\tau_k}X_j + \cdots + \beta_d^{\tau_k}X_j^{d_k} \qquad (1)$$

  - The predicted values from this regression, $\hat{Y}_j^{\tau_k,d_k}$, can be viewed as values from a stable gene from which we estimate scale factors.

SCnorm: Step 2 (estimation)

- However, values of *d* and τ are unknown. Equation (1) is fit on a grid of possible values of *d* and τ.
- For each fit, the predicted values are regressed again sequencing depth by median quantile regression:

$$Q^{.5}\left(\hat{Y}_j^{\tau_k,d_k}\middle|X_j\right) = \eta_0^{\tau_k,d_k} + \eta_1^{\tau_k,d_k}X_j$$

- The best model is chosen the the one best representing the overall group's count-depth relationship and minimizes:

$$F(\tau_k,d_k) = |\hat{\eta}_1^{\tau_k,d_k} - \underset{g\epsilon K}{\text{mode}}\,\hat{\beta}_{g,1}|$$

SCnorm: Step 2 (estimation)

- Scale factors for each cell may be be estimated as:

$$SF_j = e^{\hat{Y}_j^{\tau_k^*,d_k^*}} / e^{Y^{\tau_k^*}}$$

where $Y^{\tau_k^*}$ is the τ^{th} quantile of expression counts in the k^{th} group.

- Normalized counts are given as: $Y'_{g,j} = \frac{e^{Y_{g,j}}}{SF_j}$

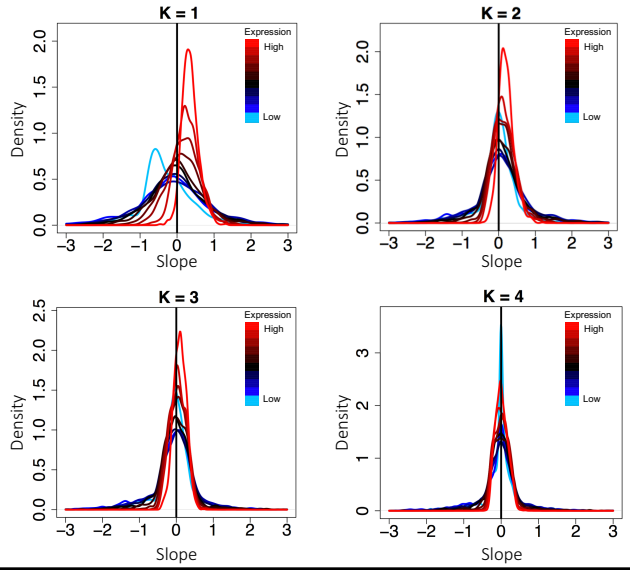
SCnorm: Step 3

- *K* is also unknown, SCnorm starts at *K = 1.*
- Genes in normalized data are divided into 10 equally sized groups based on their non-zero median un-normalized expression.
- For each gene, the normalized count-depth relationship $\beta'_{g,1}$ is estimated from:

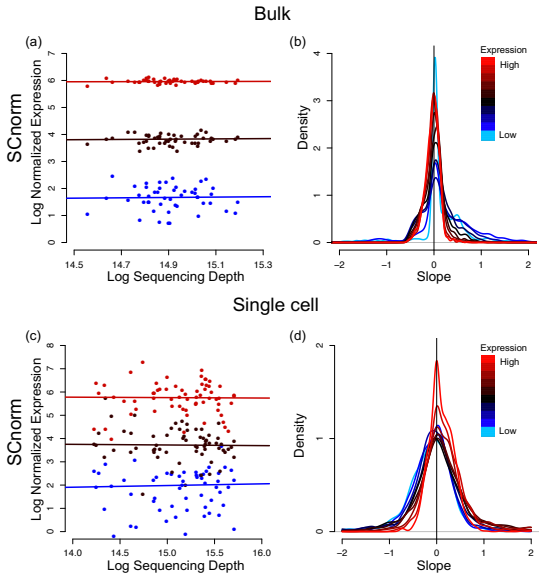
$$Q^{.5}\left(\log Y'_{g,j} \middle| X_j\right) = \beta'_{g,0} + \beta'_{g,0} X_j$$

- If the absolute value of the slope mode of all 10 groups is less than a threshold, then the value of *K* is sufficient.

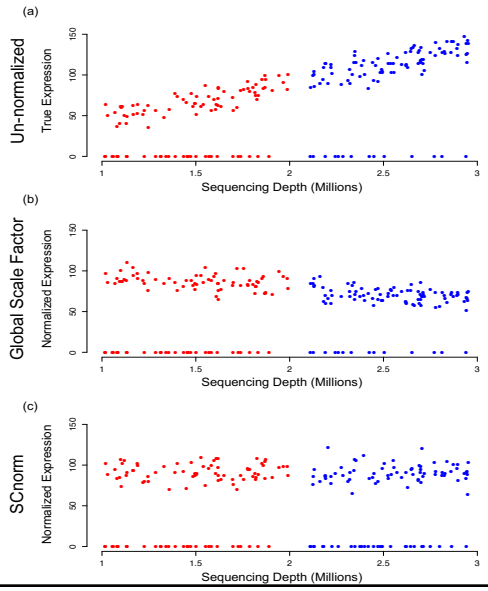
SCnorm: Step 3 (in pictures)



SCnorm on bulk and single-cell RNA-seq data



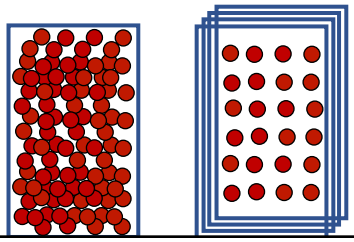
Implications in DE analysis



Case study: H1 data

- Experiment on H1 embryonic stem cells was performed by collaborators at the Thomson Lab (Morgridge Institute for Research).
- Prior to sequencing, the fragmented and indexed cDNA for each cell was split into two groups.
  - Group H1-1M: 96 cells per lane (sequencing depth ~ 1M)
  - Group H1-4M: 24 cells per lane (sequencing depth ~ 4M).

Reads per lanes is fixed ~ 100 million

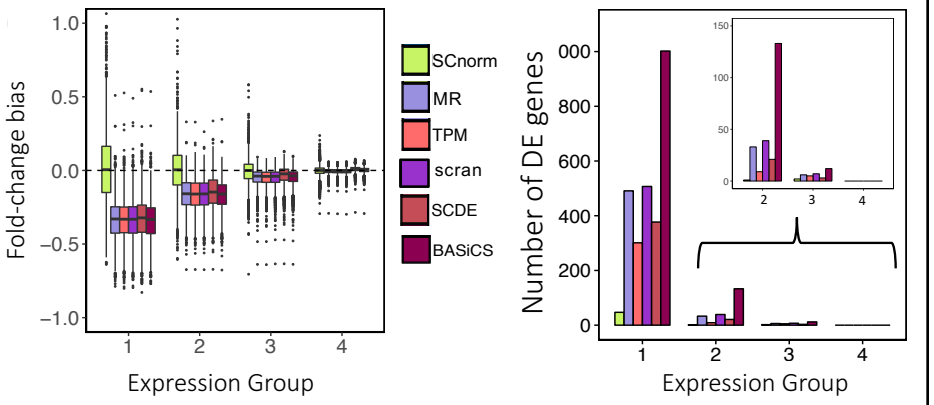


Case study: H1 data

- Prior to normalization, gene expression in H1-4M will appear on average four times higher than expression in H1-1M.
- Effective normalization should remove results in all genes appearing equivalently expressed (EE) since the cells are identical.
- Evaluate performance of SCnorm and other normalization methods based on the fold-change bias and number of DE genes.
- Fold-change bias = $\frac{\text{mean(H1−4M)}}{\text{mean(H1−1M)}} - 1$

Case study: H1 data results

Differential expression analysis using MAST



Using SCnorm:

R package is available on Bioconductor:
https://bioconductor.org/packages/release/bioc/html/SCnorm.html

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Acknowledgements

Kendziorski Lab
Christina Kendziorski, PhD
Ning Leng, PhD
Ziyue Wang

Michael Newton, PhD
Audrey Gasch, PhD

Thomson Lab
Li-Fang Chu, PhD
Ron Stewart, PhD
James Thomson, PhD

Analysis of Allele-Specific Gene Expression in Single-Cell RNA Sequencing

Mingyao Li, PhD
Professor of Biostatistics
Department of Biostatistics, Epidemiology & Informatics
University of Pennsylvania Perelman school of medicine

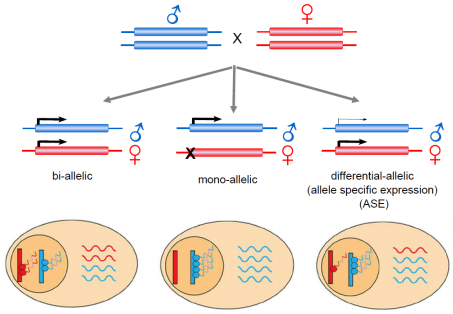
DBEI Department of Biostatistics Epidemiology & Informatics

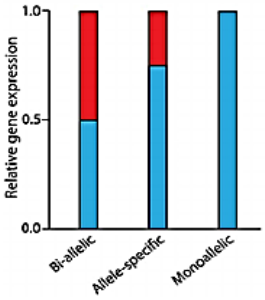
July 30, 2018 JSM Short Course

Perelman School of Medicine University of Pennsylvania

Allele-Specific Expression (ASE)

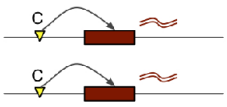
In diploid organisms, two copies of each autosomal gene are available for gene transcription.



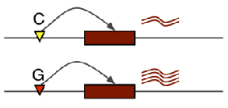


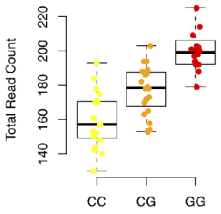
ASE, eQTL and Disease

Sample 1: genotype CC

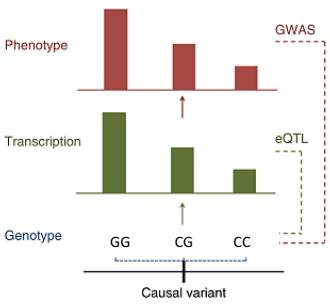


Sample 2: genotype CG



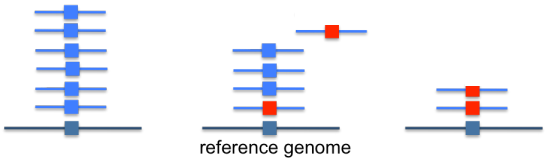


Sun and Hu (2013) Statistics in Biosciences



Zhu et al. (2016) Nature Genetics

ASE Detection in Bulk RNA-Seq



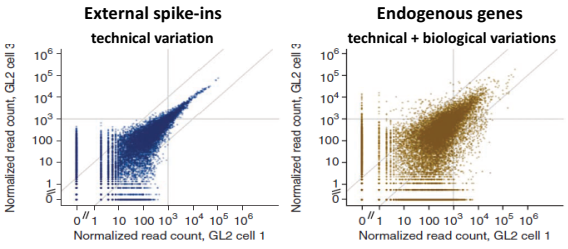
Reference allele

Alternative allele

Chi-square goodness-of-fit test	
Observed allele ratio	$X_{\text{ref,RNA}} / (X_{\text{ref,RNA}} + X_{\text{alt,RNA}})$
Expectation	0.5
H_0	True allele ratio is 0.5
H_1	True allele ratio is not 0.5
Test statistics	$\chi^2(1) = \frac{(X_{\text{ref,RNA}} - 0.5)^2}{0.5}$

ASE Analysis is Challenging in Single-Cell RNA-Seq

Technical noise, in particular, gene dropout may lead to inflated estimates of monoallelic expression

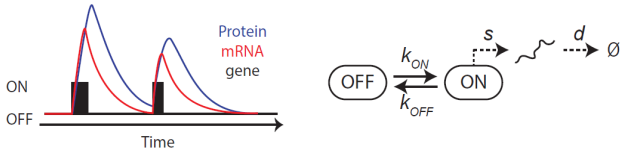


Brennecke et al. Nature Methods 10, 1093-95 (2013)

ASE Analysis is Challenging in Single-Cell RNA-Seq

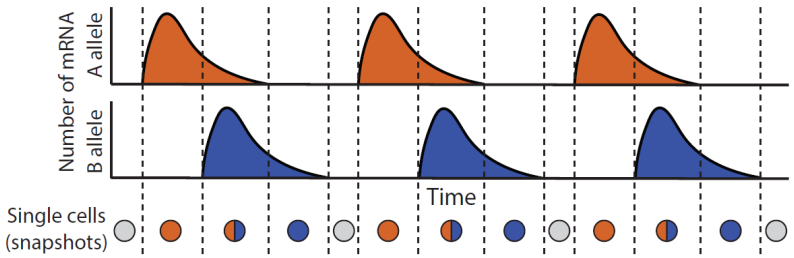
Stochasticity of gene expression induced by **transcriptional bursting** can also lead to inflated monoallelic expression.

Transcriptional bursting, is a fundamental property of genes in which transcription from DNA to RNA can occur in "bursts".



Allele-Specific Transcriptional Bursting

Allele-specific transcriptional bursting further complicates the analysis

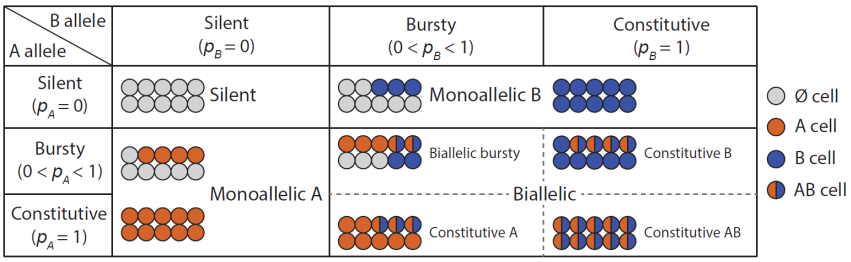


ASE Analysis in Single-Cell RNA-Seq

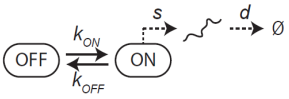
- Characterize ASE patterns across cells;
- Estimate allele-specific bursting kinetic parameters;
- Identify genes with differential bursting btw paternal and maternal alleles;
- Do the two alleles burst independently?

Gene Classification based on ASE Patterns

For each gene, developed an Empirical Bayes procedure to classify each cell into one of the four categories {silent, expression of A only, expression of B only, expression of both A and B}



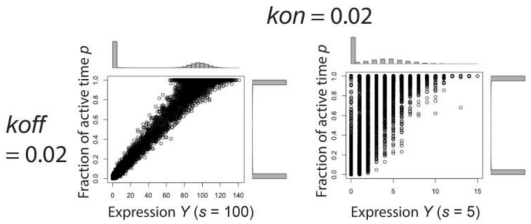
Statistical Model for Transcriptional Bursting



$$Y \sim \text{Poisson}(sp)$$
$$p \sim \text{Beta}(k_{on}, k_{off})$$

Burst frequency: k_{on}

Burst size: s/k_{off}



(Kepler & Elston 2001, Biophysical Journal; Raj et al. 2006, PLoS Biology)

Modeling Allele-Specific Transcriptional Bursting

True expression (unobserved):

cell-size factor

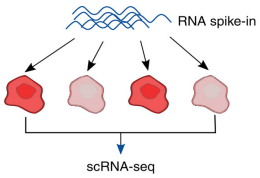
$$Y_{cg}^A \sim \text{Poisson}\left(\boldsymbol{\phi_c} s_g^A p_{cg}^A\right) \qquad Y_{cg}^B \sim \text{Poisson}\left(\boldsymbol{\phi_c} s_g^B p_{cg}^B\right)$$
$$p_{cg}^A \sim \text{Beta}\left(k_{on,g}^A, k_{off,g}^A\right) \qquad p_{cg}^B \sim \text{Beta}\left(k_{on,g}^B, k_{off,g}^B\right).$$

Observed expression (noisy):

$$Q_{cg}^A \sim Z_{cg}^A \text{Poisson}\left(\alpha_c \left(Y_{cg}^A\right)^{\beta_c}\right) \qquad Q_{cg}^B \sim Z_{cg}^B \text{Poisson}\left(\alpha_c \left(Y_{cg}^B\right)^{\beta_c}\right)$$
$$Z_{cg}^A \sim \text{Bernoulli}\left(\pi_{cg}^A\right) \qquad Z_{cg}^B \sim \text{Bernoulli}\left(\pi_{cg}^B\right)$$
$$\pi_{cg}^A = \text{expit}\left(\kappa_c + \tau_c \log\left(Y_{cg}^A\right)\right) \qquad \pi_{cg}^B = \text{expit}\left(\kappa_c + \tau_c \log\left(Y_{cg}^B\right)\right).$$

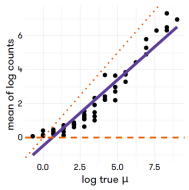
$\{\alpha_c, \beta_c, \kappa_c, \tau_c\}$ are cell-specific parameters that characterize technical noise.

Use Spike-Ins to Measure Technical Noise

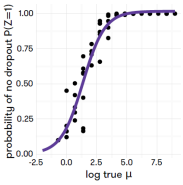


- A set of external RNA transcripts with known concentrations ;
- Added as internal controls.

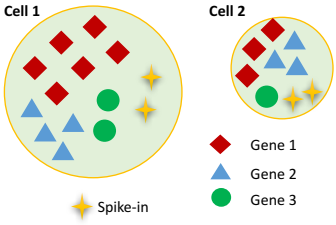
Capture, amplification, sequencing efficiency



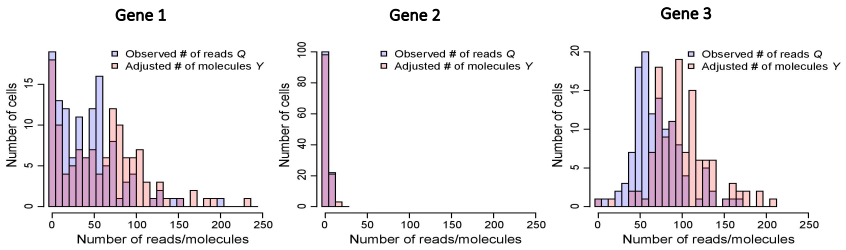
Dropout probability



Estimate cell size

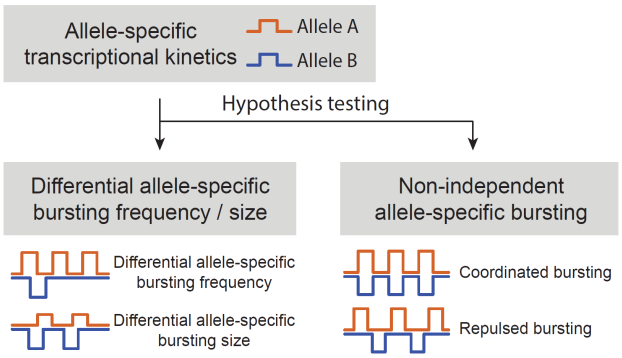


Estimating Bursting Kinetic Parameters



- Adjust for dropout, amplification, and sequencing bias by a **deconvolution algorithm**;
- Get the bursting kinetic parameters by **moment estimators**;
- Compute confidence intervals by **nonparametric Bootstrap**.

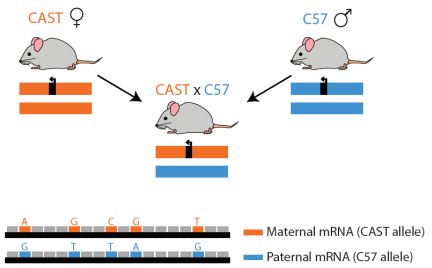
Detecting Allele-Specific Transcriptional Bursting

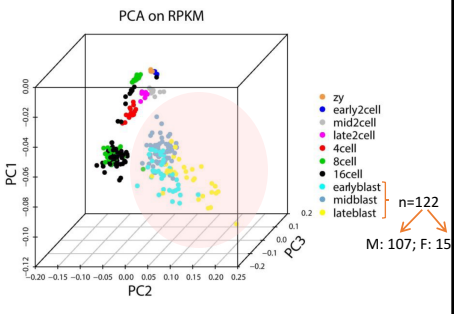


Single-Cell RNA-Seq of Mouse Blastocysts

317 single cells dissociated from F1 mouse embryos (Deng *et al.* 2014, Science)

Obtained allele-specific read counts at heterozygous loci.

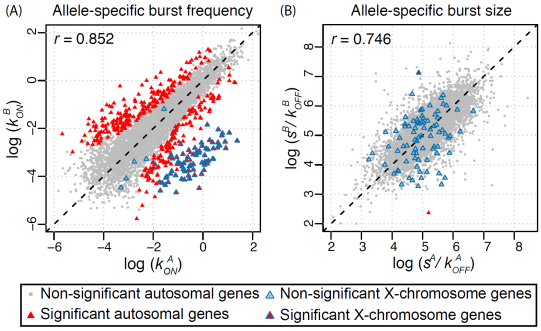




Allele-Specific Transcriptional Bursting Kinetics

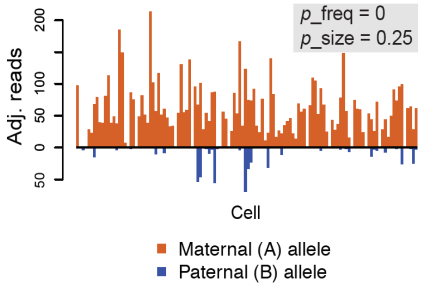
Do the two alleles share the same bursting kinetics?

Differential allele-specific bursting frequency / size

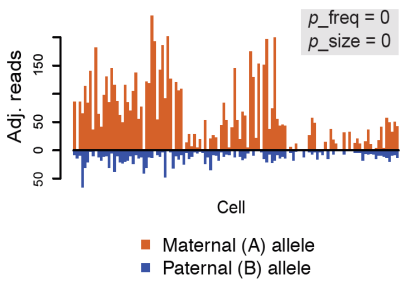


Examples of Significant Genes

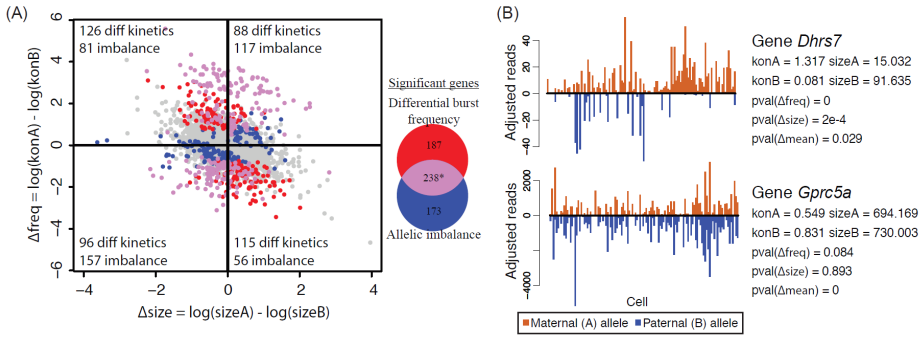
Differential burst frequency
gene *Btf314*



Differential burst size
gene *Fdps*

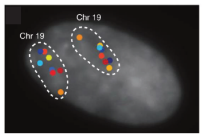


Comparison with Allelic Imbalance Test



Do the Two Alleles Burst Independently?

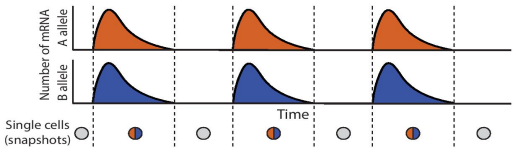
RNA-FISH



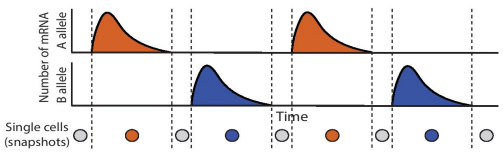
(Levesque *et al.* 2013, Nature Methods)

Independent bursting based on 20 genes from chr19

Are the two alleles *coordinated* in bursting?

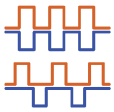


Are the two alleles *repulsed* in bursting?



Do the Two Alleles Burst Independently?

Non-independent allele-specific bursting



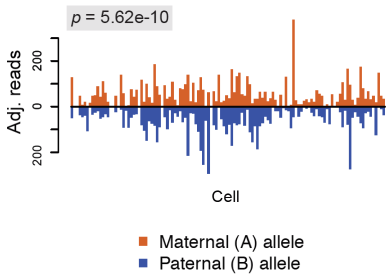
Observed

	A off	A on
B off	8	5
B on	5	101

Expected

	A off	A on
B off	1.4	11.6
B on	11.6	94.4

- Detected **424** genes with **dependent bursting**, most show *coordinated* bursting
- Example: gene *Nfatc2ip*



Summary

- Single-cell RNA sequencing allows systematic characterization of transcriptional bursting, in an allele-specific manner.
- Important to account for technical noise in analysis.
- A mouse embryonic development study:
  - On the genome-wide scale, *cis* control in gene expression acts through modulation of burst frequency, not burst size;
  - A significant number of genes exhibit coordinated bursting between alleles.

Acknowledgements

Joint work with
- Nancy Zhang
- Yuchao Jiang

Software: https://github.com/yuchaojiang/SCALE

Jiang Y, Zhang NR*, Li M* (2017) Genome Biology, 18:74

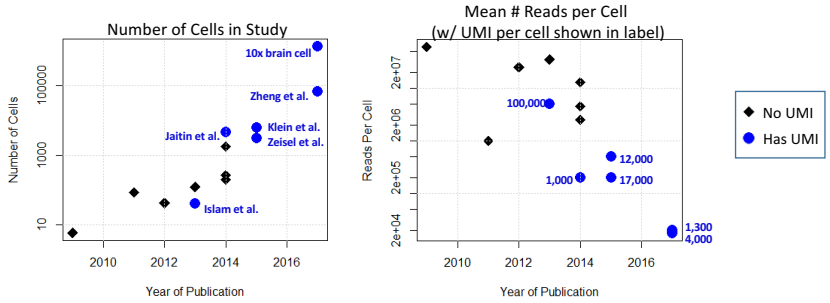
Gene Expression Recovery

(Denoising and Imputation)

Short Course: Statistical methods for single-cell RNA-seq analysis
Joint Statistical Meetings, Vancouver, 2018

Nancy R. Zhang
Dept. of Statistics
University of Pennsylvania

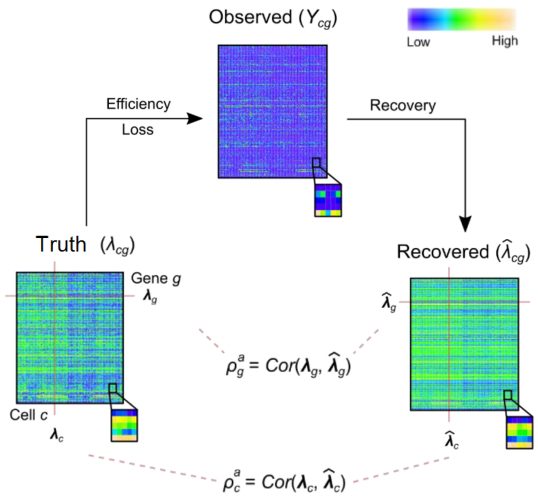
Massively-parallel single cell sequencing



~**<u>200,000</u>** RNA molecules in a typical mammalian cell (*Shapiro, Biezuner and Linnarsson, 2013*)

Goal:

Recover the ***true expression count for each gene in each cell***.



Model for Expression Recovery

Model (for UMIs):

g: gene, c: cell

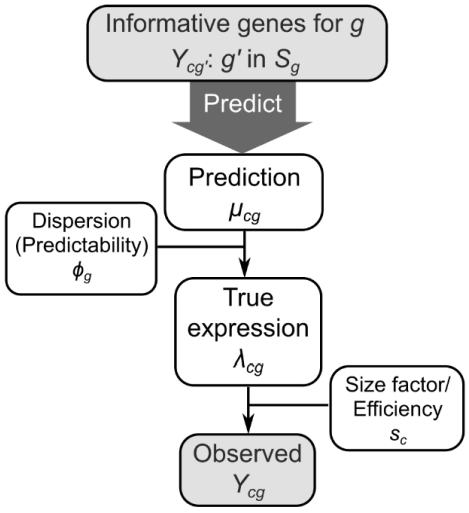
$$\mu_{cg} = \beta_{g0} + \sum_{g' \in S_g} \beta_{gg'} \log \frac{Y_{cg'}+1}{s_c}$$

$\lambda_{cg} \sim \Gamma(\mu_{cg}, \phi_g)$

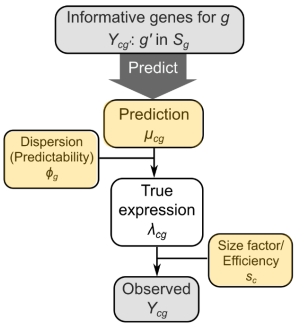
$Y_{cg} \sim Poisson(s_c \lambda_{cg})$

s_c library size
Recover relative expression

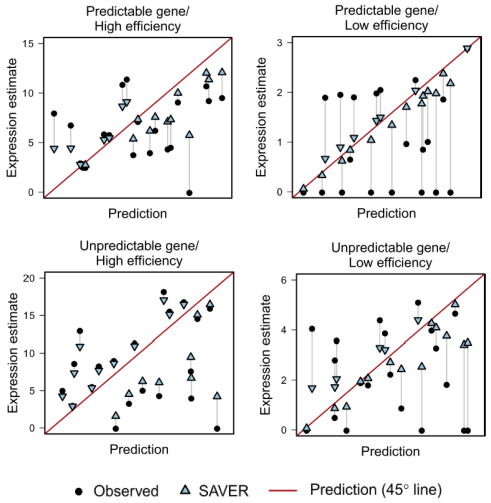
This noise model was discussed in in slides on single cell gene expression distributions.



Balancing s_c and ϕ_g



$$\hat{\lambda}_{cg} = w_{cg}\frac{Y_{cg}}{s_c} + (1 - w_{cg})\mu_{cg}$$



Validation and Benchmarking

FISH Experiment Set-up

Single cell RNA sequencing (Drop-seq)

~8,000 cells
~12,000 genes

Melanoma Cell Line

Fluorescence in-situ hybridization (smFISH)

~80,000 cells
26 genes

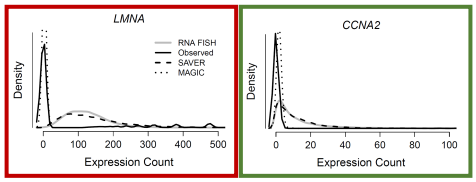
16 genes overlap between FISH and Drop-seq

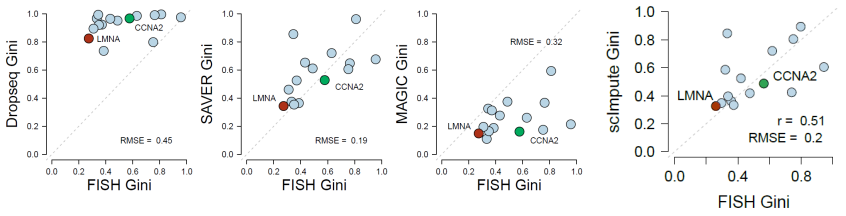
Compare:

- Each gene's expression distribution across cells
- Gene-gene correlations

Torre & Dueck (2017) bioRxiv

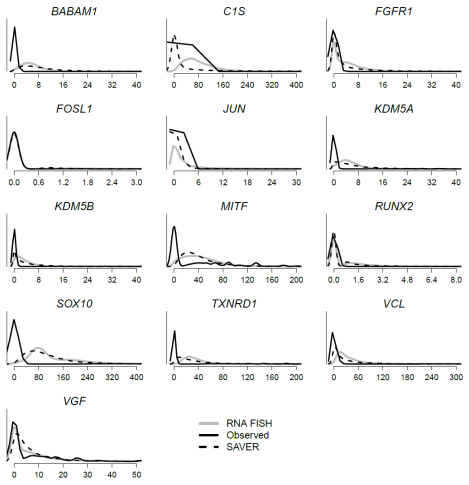
Distributions match with FISH





MAGIC: Van Dijk et al. (2017, Biorxiv)
scImpute: Li and Li (2017, Biorxiv)

And the rest of the genes

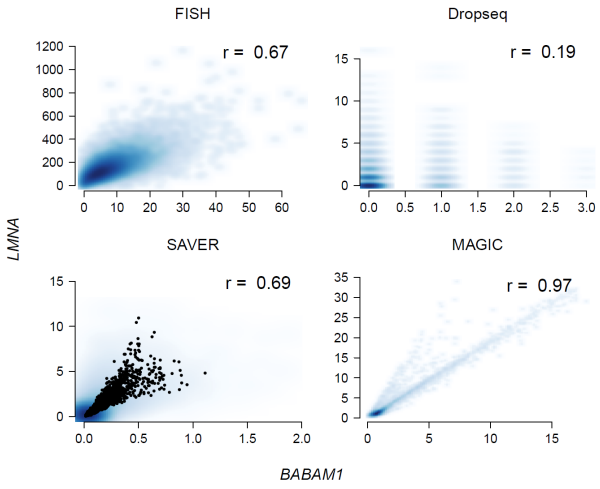


Can we recover gene-gene relationships?

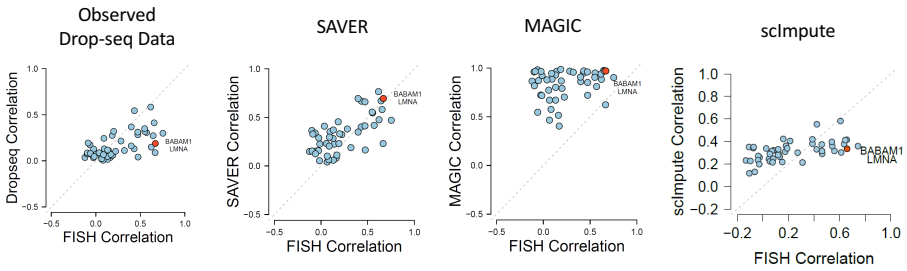
$$\lambda_{cg} | \, Y_{cg}, \mu_{cg}, \hat{\sigma}_{cg} \sim \Gamma(\hat{\lambda}_{cg}, v_{cg})$$

$$Cor(\lambda_{cg}, \lambda_{cg'}) = Cor(\hat{\lambda}_{cg}, \hat{\lambda}_{cg'}) \times f_g \times f_{g'}$$

f_g has simple analytical formula.



All pairwise gene-gene correlations



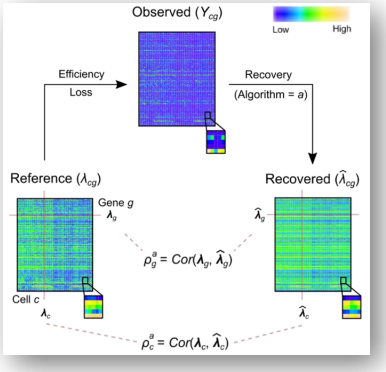
Data down-sampling experiment

Zeisel et al. (2015):
- 3005 cells from mouse brain
- Used Unique Molecular Identifiers

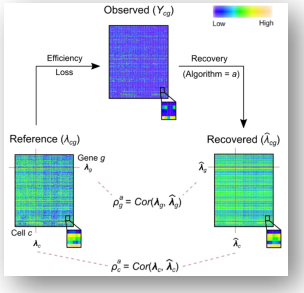
- Choose high coverage cells and genes,
  - This results in 1799 cells and 3529 gene
- Down-sample:

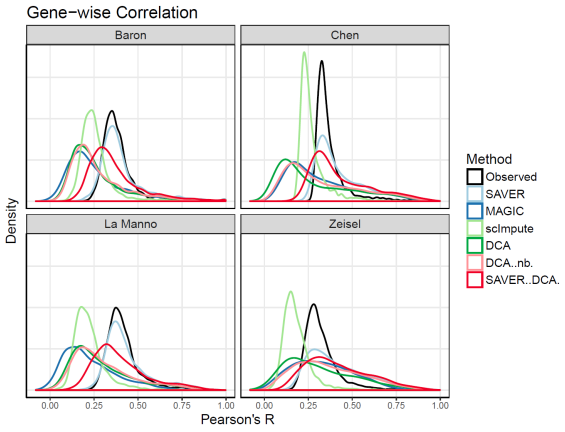
$$Y_{cg} \sim Poisson(e_c X_{cg}),$$
e_{cg} centered around $\{0.25, 0.1, 0.05\}$

- How well can we recover X from Y ?
- Extensive benchmarking on **4 data sets** in paper

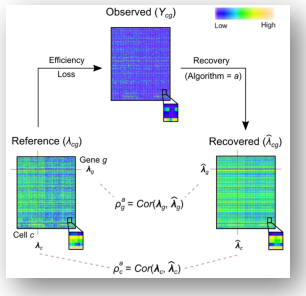


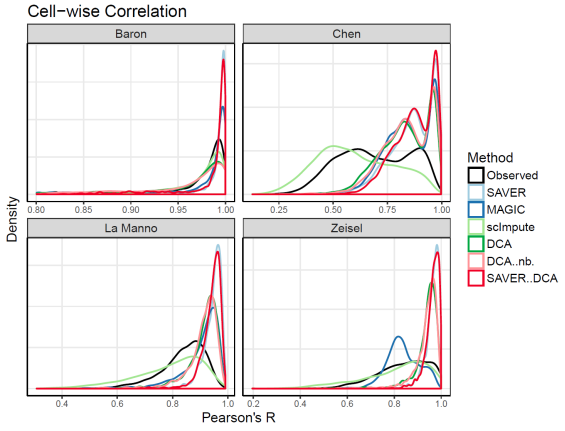
Recovery of cell-specific gene expression





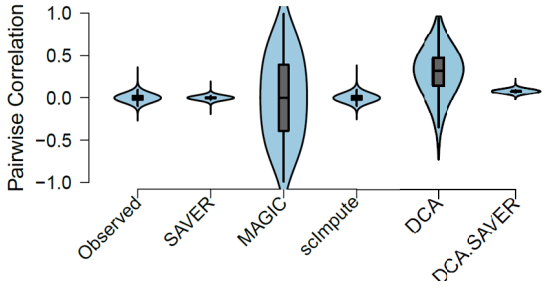
Comparisons with single-cell imputation methods





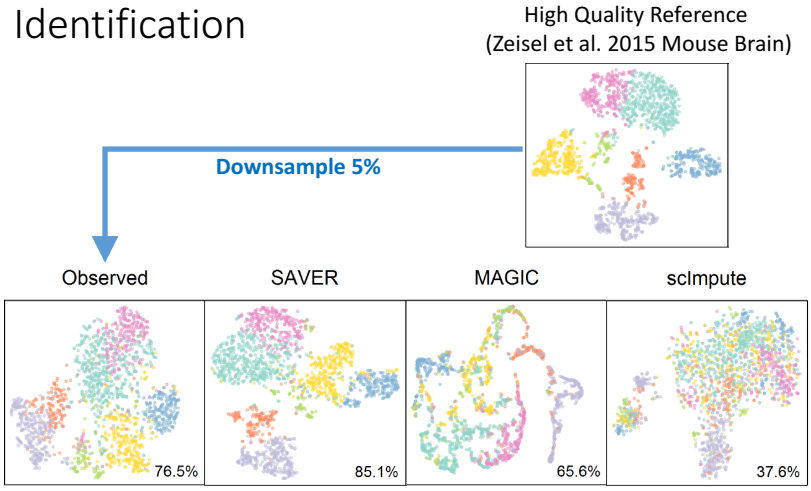
Do we introduce spurious structure?

- Start with real scRNA-seq matrix, permute cell labels for each gene
- Computed gene-gene pairwise correlations after applying algorithm

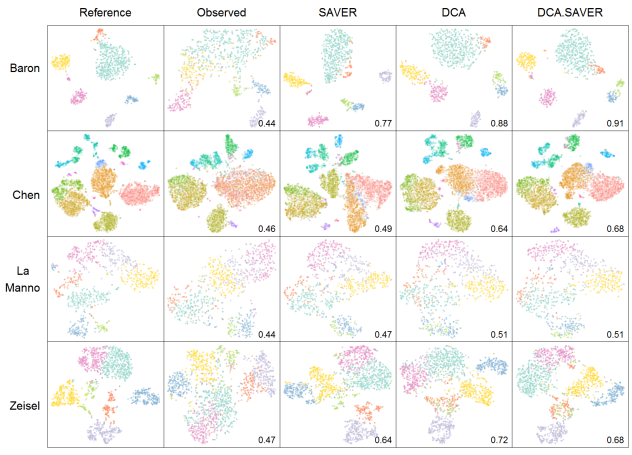


Impact on Downstream Analysis

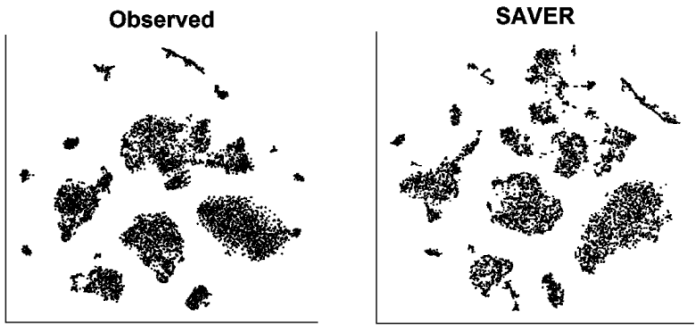
Impact on Cell Type Identification



Comparison with DCA

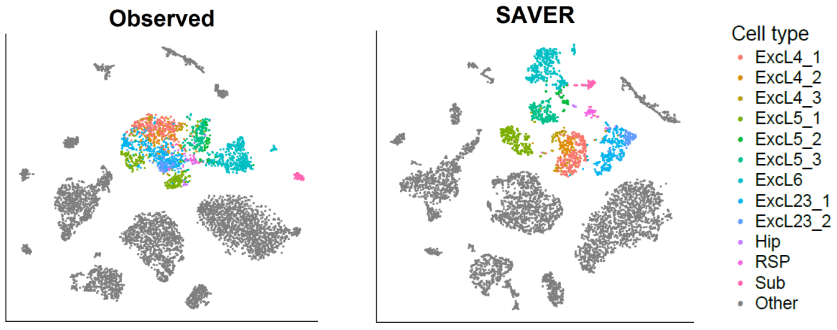


Cell types in mouse visual cortex



Hrvatin, S. *et al.* Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat. Neurosci.* **21,** (2017).

Cell types in mouse visual cortex

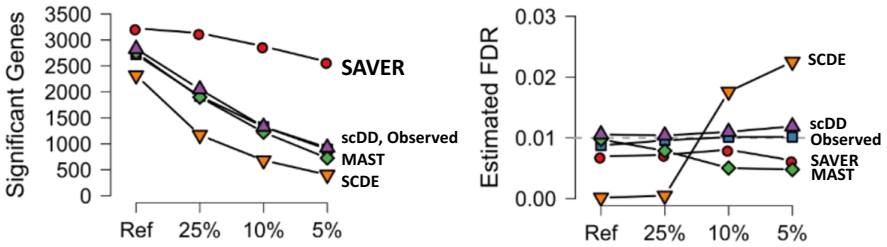


Hrvatin, S. *et al.* Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat. Neurosci.* **21,** (2017).

Differential Expression

- Zeisel et al. (2015): 3005 cells from mouse brain
- CA1 Pyramidal neuron cells further classified into two main sub-types: 351 CA1Pyr1, 389 CA1Pyr2
- Same Poisson sub-sampling experiment
- Performed SAVER, then applied Wilcoxon Rank-Sum test to the imputed values to test for differential expression between CA1Pyr1, and CA1Pyr2. FDR control by Benjamini Hochberg.
- Compare with other approaches relying on original counts

Differential Expression Analysis



Summary

Gene-gene relationships can be learned from data and harnessed for denoising.

SAVER is based on shrinking to cell- specific predictions. The shrinkage step is critical to avoid introducing spurious relationships.

SAVER is tested through:

- Down-sampling experiments using 4 high quality scRNA-seq datasets
- Extensive smFISH validation
- Permutation tests

Impact on downstream analyses:

- Cell type identification
- Finding marker genes
- Estimating gene-gene relationships

Huang et al. (2018) Gene expression recovery for single cell RNA sequencing, *to appear in Nature Methods*. (https://www.biorxiv.org/content/early/2017/05/17/138677)

Acknowledgements

Mo Huang, Statistics, University of Pennsylvania

Jingshu Wang, Statistics, University of Pennsylvania

Mingyao Li, Biostatistics, University of Pennsylvania

Arjun Raj and John Murray Labs, University of Pennsylvania

Bulk Tissue Cell Type Deconvolution with Multi-Subject Single-Cell Expression Reference

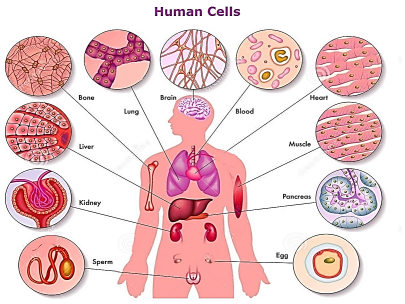
Mingyao Li, PhD
Professor of Biostatistics
Department of Biostatistics, Epidemiology & Informatics
University of Pennsylvania Perelman school of medicine

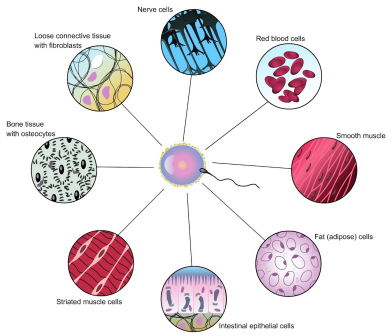
DBEI Department of Biostatistics Epidemiology & Informatics | CCEB

July 30, 2018 JSM Short Course

Perelman School of Medicine University of Pennsylvania

Human and Single Cells

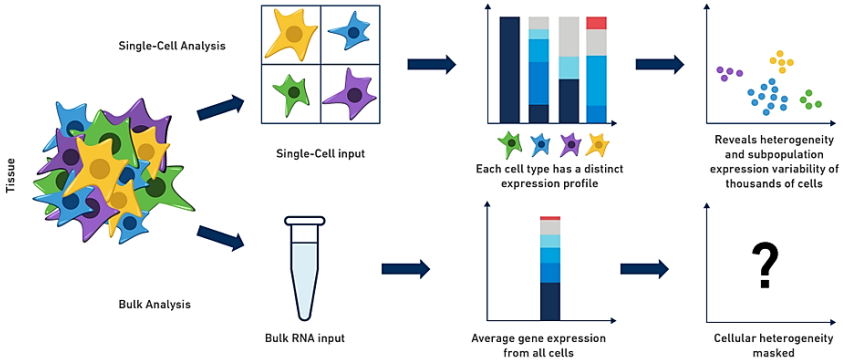




Number of cells in the human body ≈
37,000,000,000,000 = 37 trillion = 3.7×10^{13}

https://humanbodyanatomy.co/various-cells-from-the-human-body

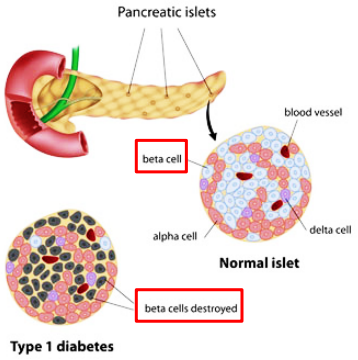
Bulk Tissue v.s. Single-Cell RNA Sequencing



10x Community - 10x Genomics

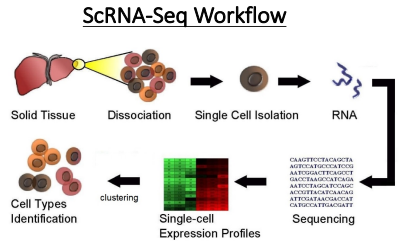
Deconvolution of Cell Type Compositions

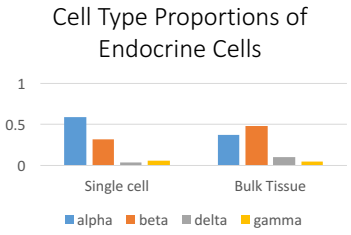
- Cell type composition is a confounding factor in gene expression analysis of bulk tissue.
- Knowledge in cell type composition can eliminate confounding, and also help understand disease progression.
- Ex: T1D and loss of beta cells.



Why Not Just Use Single-Cell Data?

Cell type proportions in single cell data do not reflect the real cell type proportions in intact bulk tissue.

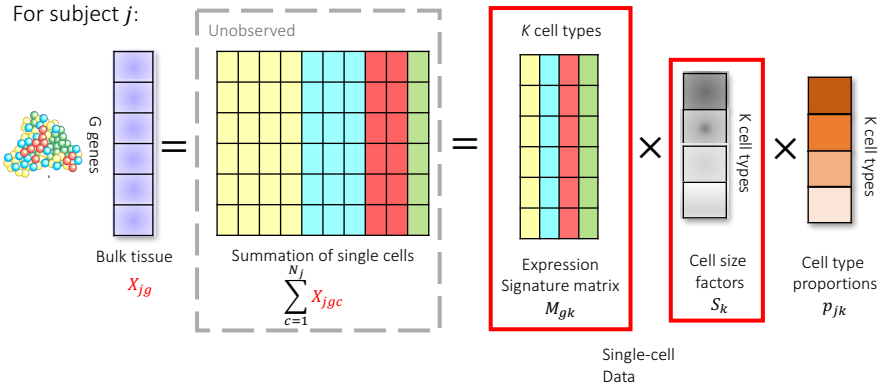




Single cell data from Xin et al.(2016);
Bulk tissue proportions suggested by Steiner et al.(2011)

https://en.m.wikipedia.org/wiki/File:RNA-Seq_workflow-5.pdf

Deconvolution with Single Cells



Existing Deconvolution Methods

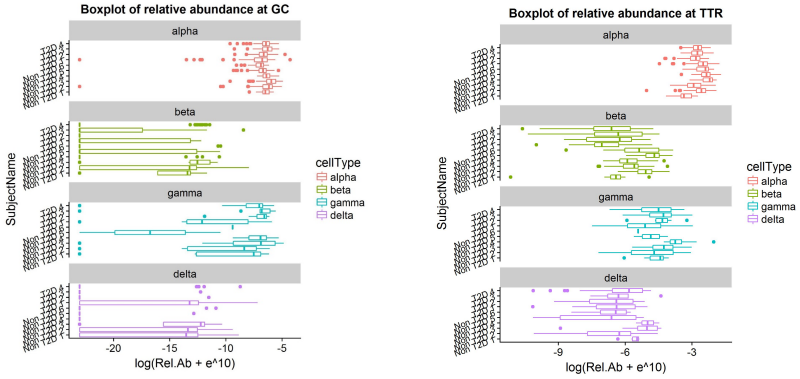
There are many existing deconvolution methods. For example:

- CIBERSORT: Support Vector Regression based estimation for microarray data (Newman et al. 2015 Nature Methods).
- BSEQ-sc: Implementation of CIBERSORT with single cell data as reference (Baron et al. 2016 Cell Systems).

Limitations of existing methods

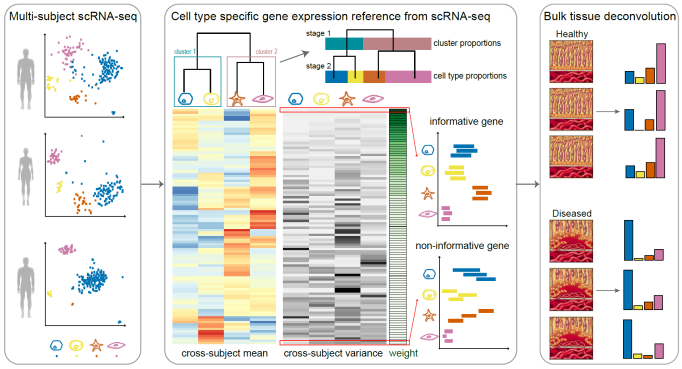
- Pre-select marker genes with hard thresholds when constructing expression signature matrix.
- Ignore subject-to-subject gene expression variation.

Subject-to-Subject Variation is Non-ignorable



Data Source: E-MTAB-5061 from Segerstolpe et al. (2016 Cell Metabolism)

Workflow of MuSiC

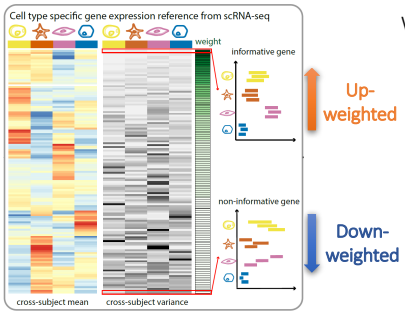


Compare **cell type compositions** between normal and diseased individuals

Cell type-specific **differential expression** analysis

Cell type-specific **eQTL analysis** etc.

Weighted Non-Negative Least Squares (W-NNLS)



Weight is the inverse of

$$Var[X_{jg}|\boldsymbol{p}_j] = E\left[Var[X_{jg}|\boldsymbol{p}_j, \boldsymbol{\theta}_g]\right] + Var\left[E[X_{jg}|\boldsymbol{p}_j, \boldsymbol{\theta}_g]\right]$$

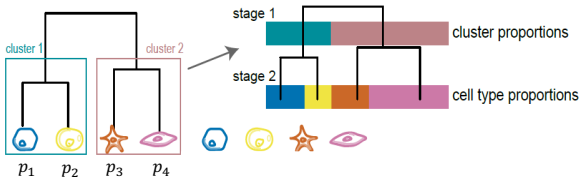
$$= \delta_g^2 + m^2 \sum_{k=1}^{K} p_{jk}^2 \times S_k^2 \times \sigma_{gk}^2$$

Noise in bulk tissue gene expression data

Cross-subject gene expression variance in cell type k

W-NNLS: $X_{jg} = m \sum_{k=1}^{K} p_{jk} \times S_k \times \theta_{gk} + \epsilon_{jg}, \quad \epsilon_{jg} \sim N(0, \delta_g^2)$

Estimation for Closely-Related Cell Types



Stage 1: estimate cluster proportions $\pi_1 = p_1 + p_2$ and $\pi_2 = p_3 + p_4$ using W-NNLS using intra-cluster homogeneous genes.

Stage 2: estimate cell type proportions subject to the constraint that

$\hat{p}_1 + \hat{p}_2 = \hat{\pi}_1$ and $\hat{p}_3 + \hat{p}_4 = \hat{\pi}_2$.

Evaluation using Pancreatic Islet Datasets

Name	Journal	Year	Session #	Tissue Type	Data type	Protocol	# samples	# cells	# genes	# cell types
Segerstolpe et al.	Cell Metabolism	2016	E-MTAB-5061	Pancreatic islet	Single-cell read counts	Smart-seq2	10 (6 H + 4 T2D)	2209	25453	14 + 1 NA
Segerstolpe et al.	Cell Metabolism	2016	E-MTAB-5060	Pancreatic islet	Bulk read counts	Smart-seq2	7 (3H + 4 T2D)	NA	25453	NA
Xin et al.	Cell Metabolism	2016	GSE81608	Pancreatic Islet: endocrine	Single-cell read counts	Illumina HiSeq 2500	18 (12H + 6 T2D)	1492	39849	4
Fadista et al.	PNAS	2014	GSE50244	Pancreatic islet	Bulk read counts	Illumina HiSeq 2000	89	NA	56638	NA

Methods Comparison and Evaluation Metrics

- Weighted-NNLS (MuSiC)
- NNLS (Non-negative least squares)

Input: all genes

- CIBERSORT (Newman et al. 2015 Nature Methods)
- BSEQ-sc (Baron et al. 2016 Cell Systems)

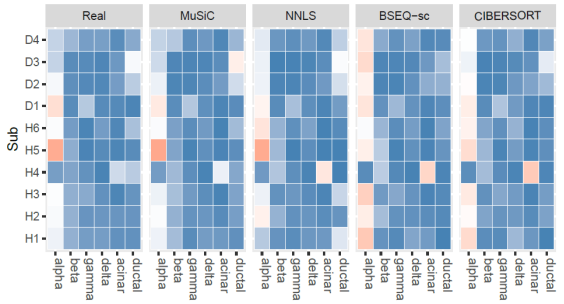
Input: pre-selected marker genes

p: true cell type proportions

$\hat{p}$: estimated cell type proportions

- Pearson correlation: $R = corr(p, \hat{p})$
- Mean Absolute Difference (mAD): $average(|p - \hat{p}|)$
- Root-Mean-Squared Deviation (RMSD): $\sqrt{average((p - \hat{p})^2)}$

Benchmark Evaluation: *in silico* Bulk Data



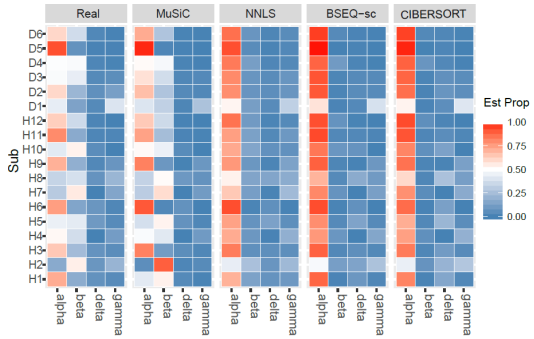
Method	MuSiC	NNLS	BSEQ-sc	CIBERSORT
RMSD	0.040	0.098	0.099	0.085
mAD	0.029	0.064	0.068	0.061
R	0.97	0.85	0.86	0.89

Sergerstolpe et al. data (2016)

For each subject:

- Artificial Bulk data: Summed read counts across cells from the same subject.
- Single-cell reference: 6 healthy subjects. Left out the subject in deconvolution.
- Bulk and single-cell data are from the ***same*** study.

Benchmark Evaluation: *in silico* Bulk Data



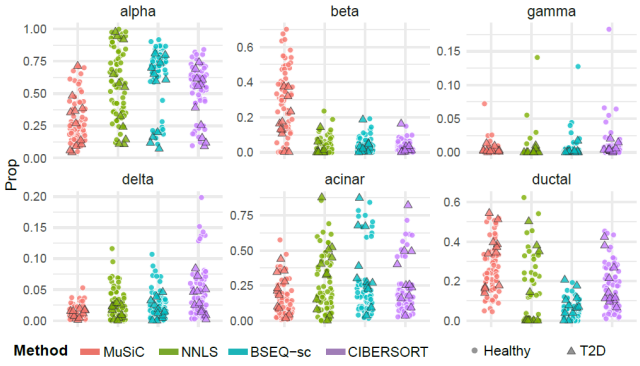
Method	MuSiC	NNLS	BSEQ-sc	CIBERSORT
RMSD	0.10	0.17	0.21	0.21
mAD	0.06	0.12	0.15	0.15
R	0.94	0.82	0.79	0.76

Xin et al. data (2016)

For each subject:

- Artificial Bulk data: Summed read counts across cells from same subject.
- Single-cell reference: 6 healthy subjects from **Sergerstolpe et al. data**.
- Bulk and single-cell data are from ***different*** studyies.

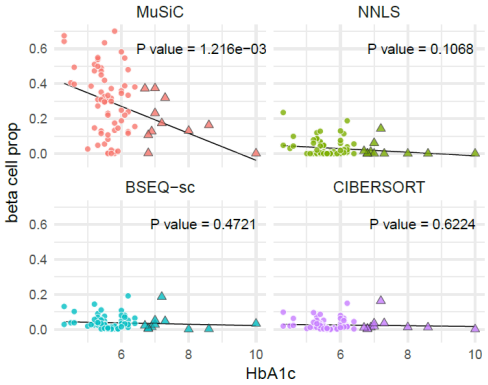
Application to Fadista et al. Bulk RNA-Seq Data



Expected cell-type composition within human islets (Steiner et al. 2011):

- Alpha: 30%-40%
- Beta: 50%-60%
- Less than 10% gamma and delta cells

Application to Fadista et al. Bulk RNA-Seq Data



The HbA1C test is one of the best ways to check diabetes is under control.

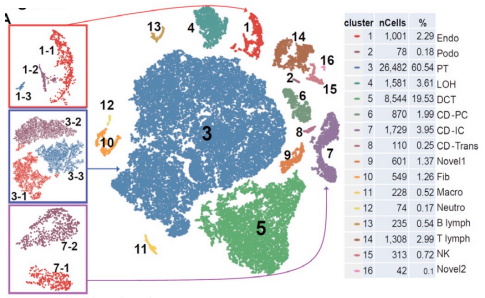
In healthy people: HbA1c level is < 6% of total hemoglobin.

In diabetic people: HbA1c level is > 6.5%.

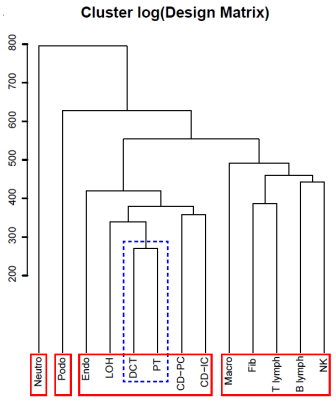
Evaluation using Mouse/Rat Kidney Datasets

Name	Journal	Year	Session #	Tissue Type	Data type	Protocol	# samples	# cells	# genes	# cell types
Park et al.	Science	2018	GSE107585	Kidney	Single-cell c read counts	10x	7 healthy, male	43745	16273	14 + 2 novel
Beckerman et al.	Nature Medicine	2017	GSE81492	Kidney	Bulk read counts	Illumina HiSeq 2500	10 (6 control + 4 APOL1)	NA	19033	NA
Lee et al.	JASN	2015	GSE56743	Kidney tubule	Bulk read counts	Illumina HiSeq 2000	118 replicates (14 segments)	NA	10903	NA
Craciun et al.	JASN	2015	GSE65267	Kidney	Bulk read counts	Illumina HiSeq 2000	18 replicates (6 time points)	NA	25219	NA
Arvaniti et al.	Scientific Reports	2016	GSE79443	Kidney	Bulk read counts	Illumina HiSeq 2000	10 replicates (Sham + 2 time points)	NA	38683	NA

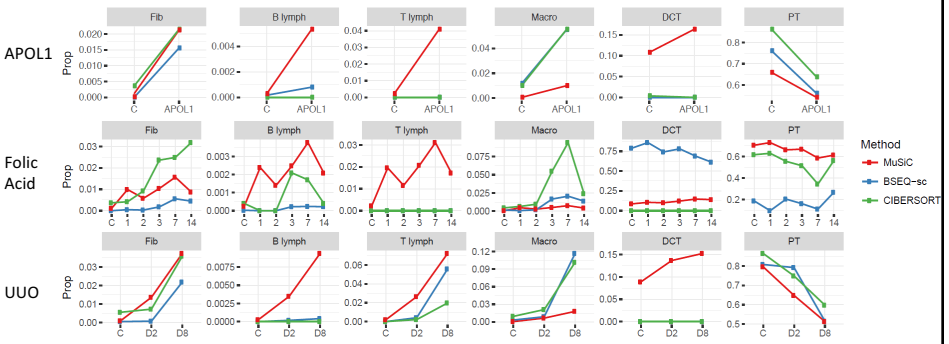
Using Park et al. ScRNA-Seq Data As Reference



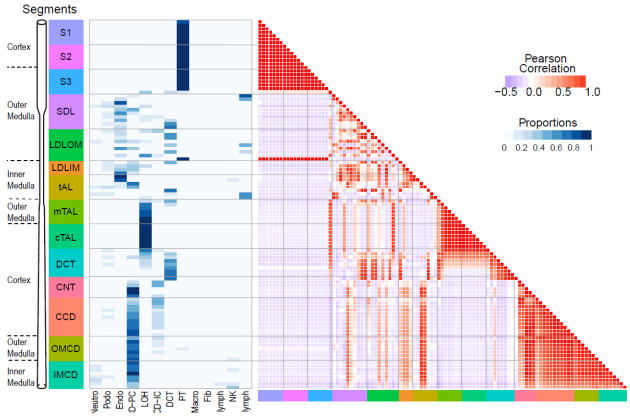
cluster	nCells	%	
1	1,001	2.29	Endo
2	78	0.18	Podo
3	26,482	60.54	PT
4	1,581	3.61	LOH
5	8,544	19.53	DCT
6	870	1.99	CD-PC
7	1,729	3.95	CD-IC
8	110	0.25	CD-Trans
9	601	1.37	Novel1
10	549	1.26	Fib
11	228	0.52	Macro
12	74	0.17	Neutro
13	235	0.54	B lymph
14	1,308	2.99	T lymph
15	313	0.72	NK
16	42	0.1	Novel2



Deconvolution in Bulk Data from 3 Mouse Models



Deconvolution in Bulk Data from Rats



Summary

- Knowledge of cell type composition in disease relevant tissues is an important step towards the identification of cellular targets in disease.
- Although most scRNA-seq data do not reflect true cell type proportions in intact tissues, they do provide valuable information on cell type-specific gene expression.
- Harnessing multi-subject scRNA-seq reference data, MuSiC reliably estimates cell type proportions from bulk RNA-seq.
- As bulk tissue data are more easily accessible than scRNA-seq, MuSiC allows the utilization of the vast amounts of disease relevant bulk tissue RNA-seq data for elucidating cell type contributions in disease.

Acknowledgements

Joint work with
- Nancy Zhang
- Xuran Wang
- Jihwan Park
- Katalin Susztak

Software: https://github.com/xuranw/MuSiC

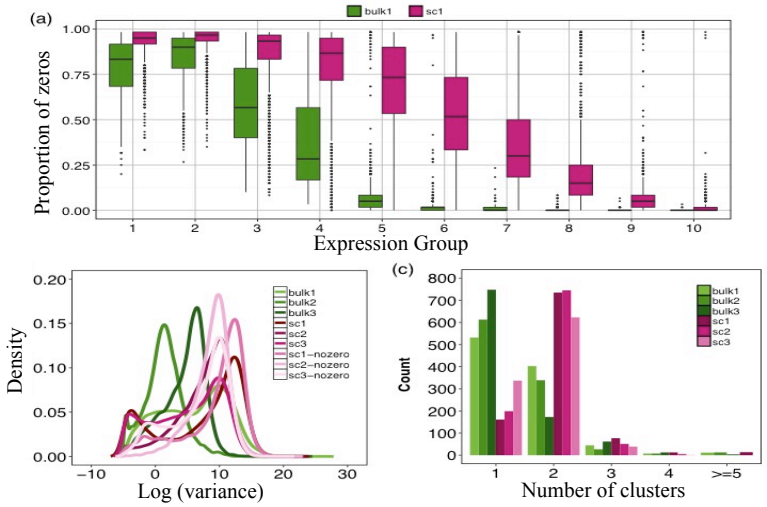
Statistical methods for identifying differentially distributed genes and pseudotime reordering

Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison

http://www.biostat.wisc.edu/~kendzior/

Features of single-cell RNA-seq data

- Abundance of zeros, increased variability, complex distributions

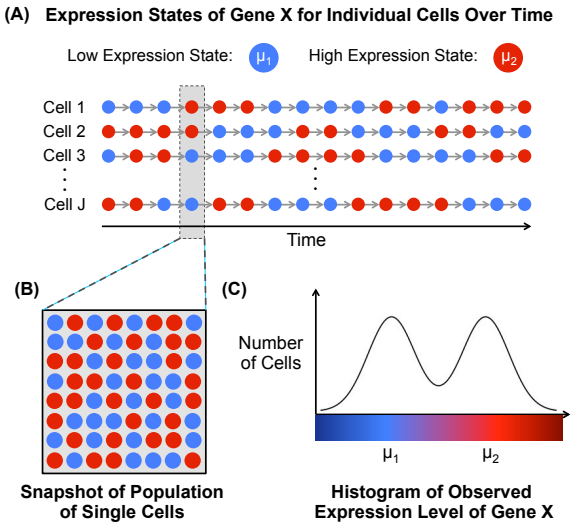


Bacher and Kendziorski, *Genome Biology*, 2016.

scDD: A Dirichlet mixture model based approach for identifying differential distributions in scRNA-seq experiments

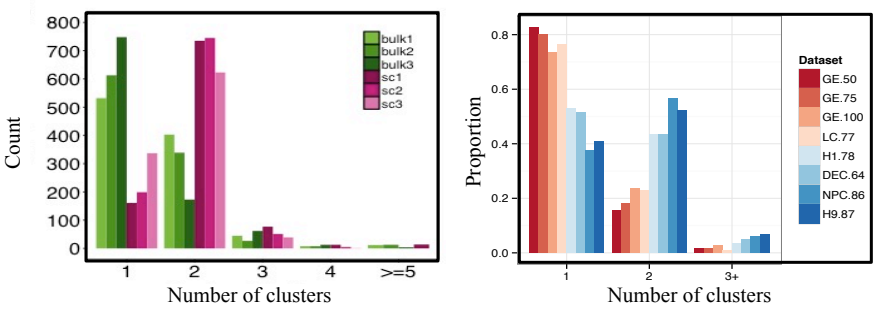
Korthauer *et al*., *Genome Biology*, 2016

Gene-specific multi-modality

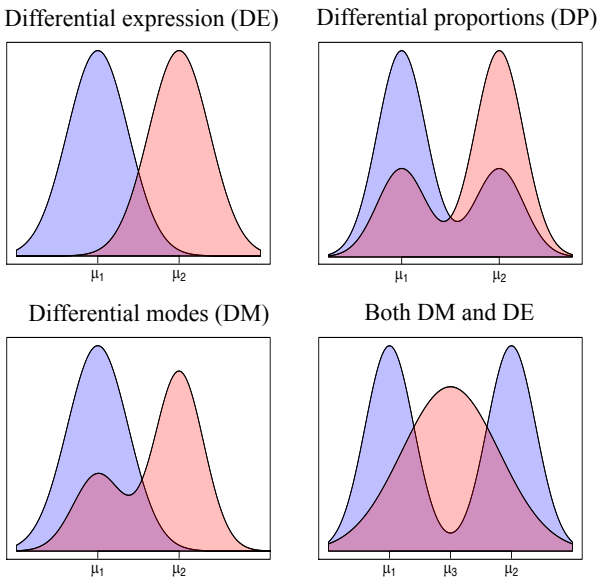


CK JSM 2018

Many genes show multi-modal expression distributions



Opportunity to identify differences beyond traditional DE



scRNA-seq DE Analysis

- Recent methods use mixture modeling to account for 'on' and 'off' components
  - Shalek et al. (2014)
  - SCDE (Kharchenko *et al.*, 2014)
  - MAST (Finak *et al.*, 2015)

- When detected, each gene has a latent level of expression within a biological condition, and measurements fluctuate around that level due to biological and technical sources of variability

scDD: Goal

- Model expression profiles while accommodating the often multimodal distributions in the detected cells
- Find genes with Differential Distributions (DD) of expression across two conditions:
  - differential means
  - differential proportion within modes
  - differential modality (number of modes)
  - combination thereof
  - differential zeroes (detection rate)

scDD: Overview

- Log non-zero normalized, de-noised, expression arises out of a fixed variance Dirichlet Process Mixture of normals model.
- For each gene, obtain maximum a posteriori (MAP) partition of the samples to components using the *modalclust* algorithm (Dahl 2009).
  - fast and deterministic
  - requires point estimate of cluster variance (obtain via *mclust*).
- To evaluate evidence of DD, fit under two different hypotheses:
  - ignoring condition ($\mathcal{M}_{ED}$: equivalent regulation)
  - separately for each condition ($\mathcal{M}_{DD}$: differential regulation)

scDD: Overview (continued)

- Assume that log non-zero normalized, de-noised, expression measurements $Y_g = (y_{g1},...,y_{gJ})$ for gene g in J cells arise from a conjugate Dirichlet Process Mixture (DPM) of normals model:

$$\begin{aligned} y_j &\sim N(\mu_j, \tau_j) \\ \mu_j, \tau_j &\sim G \\ G &\sim DP(\alpha, G_0) \\ G_0 &= NG(m_0, s_0, a_0/2, 2/b_0) \end{aligned}$$

- Let K denote the number of components (unique values in $\{\mu_j, \tau_j\}, j=1,.., J$). Of primary interest is the posterior of (μ,τ), which is intractable for moderate sample sizes.

- Let $Z = (z_1, \ldots, z_J)$ denote component memberships. Then $f(Y \mid Z)$ is a PPM.

$$\begin{aligned} f(Y|Z) &= \prod_{k=1}^{K} f(y^{(k)}) \\ &\propto \prod_{k=1}^{K} \frac{\Gamma(a_k/2)}{(b_k/2)^{a_k/2}} s_k^{-1/2} \end{aligned}$$

scDD: Overview (continued)

- To quantify the evidence of DD for gene g, obtain MAP partition estimate, $\hat{Z}$, and evaluate $f(Y,\hat{Z})$ under competing hypotheses:
  - ignoring condition ($\mathcal{M}_{ED}$: equivalent distributions)
  - separately within condition ($\mathcal{M}_{DD}$: differential distributions)

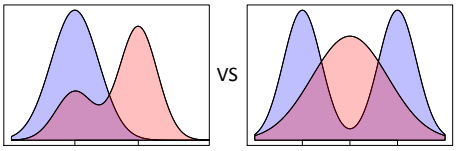
- Evaluate $\mathcal{M}_{DD}$ using a pseudo-Bayes Factor score:

$$Score_g = \log\left(\frac{f(Y_g,\hat{Z}_g|M_{DD})}{f(Y_g,\hat{Z}_g|M_{ED})}\right)$$

- Assess significance via permutation.

scDD: Classification of DD genes

- Classify DD genes into categories based on
  - number of components detected in each condition
  - whether clusters overlap



- Overlap is determined by sampling from the marginal posterior distribution of cluster means

$$\mu_k | Y, Z \sim t_{a_k}\left(m_k, \frac{b_k}{a_k s_k}\right)$$

CK JSM 2018

scDD: Evaluation via simulation studies

- 8000 ED genes:
  - 4000 from single Negative Binomial component
  - 4000 from two component mixture of Negative Binomial
- 2000 DD genes:
  - 500 DE genes
  - 500 DP genes (0.33/0.66 proportion difference)
  - 500 DM genes (0.50 belong to second mode)
  - 500 DB genes (mean in second condition is average of means in the first)
- Sample sizes varied $\in \{50, 75, 100\}$
- Component distances Δ_μ for multimodal conditions varied $\in \{2, 3, 4, 5, 6\}$ SDs
- Means, variances, and detection rates sampled empirically

Evaluate: Power to identify DD genes
Rate at which DD genes are correctly classified
Rate at which correct # components are identified

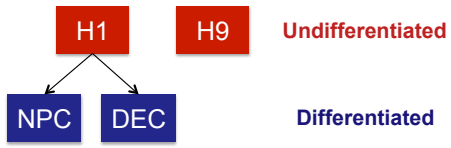
CK JSM 2018

scDD: Power to detect DD genes within each category

Sample Size	Method	True Gene Category DE	DP	DM	DB	Overall (FDR)
50	scDD	0.893	**0.418**	**0.898**	**0.572**	**0.695** (0.030)
	SCDE	0.872	0.026	0.816	0.260	0.494 (0.004)
	MAST	**0.908**	0.400	0.871	0.019	0.550 (0.026)
75	scDD	0.951	0.590	**0.960**	**0.668**	**0.792** (0.031)
	SCDE	0.948	0.070	0.903	0.387	0.577 (0.003)
	MAST	**0.956**	**0.632**	0.942	0.036	0.642 (0.022)
100	scDD	0.972	0.717	**0.982**	**0.727**	**0.850** (0.033)
	SCDE	0.975	0.125	0.946	0.478	0.631 (0.003)
	MAST	**0.977**	**0.752**	0.970	0.045	0.686 (0.022)
500	scDD	**1.000**	0.985	**1.00**	**0.903**	**0.972** (0.034)
	SCDE	**1.000**	0.858	0.998	0.785	0.910 (0.004)
	MAST	**1.000**	**0.992**	**1.00**	0.174	0.792 (0.021)

CK JSM 2018

Comparison of hESCs



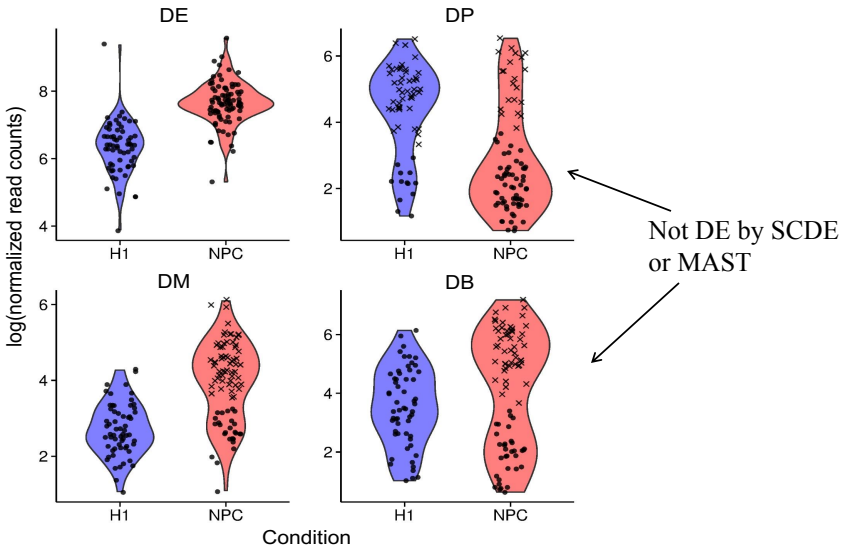
Number of DD genes identified in each cell type comparison

Comparison	scDD DE	DP	DM	DB	DZ	Total	SCDE	MAST
H1 vs NPC	1342	429	739	406	1590	4506	2938	5729
H1 vs DEC	1408	404	939	345	880	3976	1581	3523
NPC vs DEC	1245	449	700	298	2052	4744	1881	5383
H1 vs H9	194	84	55	32	145	510	102	1091

scDD only: 2% 21% 38% 24% 15%

CK JSM 2018

Genes identified in H1 vs. NPC comparison



Variability induced by oscillatory genes is substantial in single-cell RNA-seq and can mask effects of interest

We developed a pseudotime reordering approach called Oscope to identify and characterize oscillations in single-cell RNA-seq experiments

Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments

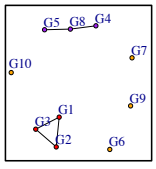
Leng *et al*., *Nature Methods,* 2015

Oscope: Identify and characterize oscillatory genes in an scRNA-seq experiment

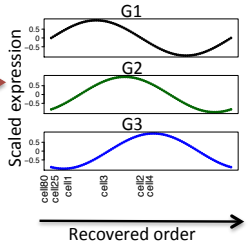
Step 1: Paired-sine model identifies candidate oscillators

Top genes pairs
G1-G2
G3-G2
G1-G3

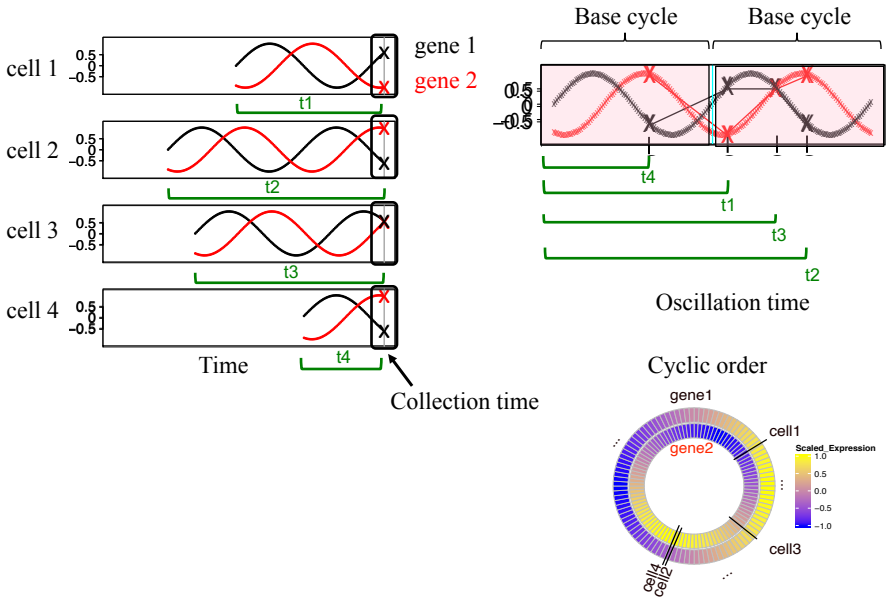
Step 2: Residuals from paired-sine model are used in clustering to group genes with similar frequency, varying phase



Step 3: Extension of the nearest insertion algorithm reconstructs one base cycle for every group



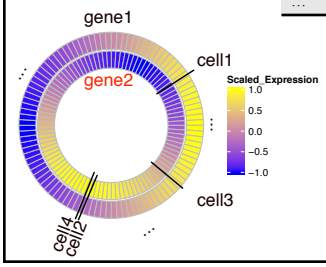
Oscope: Identifying oscillatory genes using scRNA-seq



Oscope: Connection to TSP

- Given a list of cities, only know distances between each pair of cities
- Goal is to find an optimal route that visits each city exactly once and returns to the origin city
- The optimal route will minimize overall distance travelled

	Distance (mile)
Madison-Chicago	148
Madison-Iowa City	175
Madison-Atlanta	847
Madison-Minneapolis	273
Chicago-Iowa City	223
Chicago-Atlanta	716
…	…

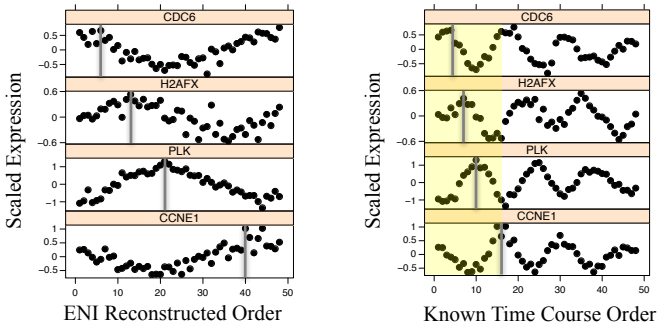


- In our case, for each gene cluster, we want to find an optimal "route" through all cells and return to the first cell
- The optimal route will minimize expression differences between observed and baseline oscillation

CK JSM 2018

Oscope: Results from Whitfield data

- Whitfield data: microarray time course of HeLa cells synchronized for cell cycle. 48 samples; one every hour (~3 cell cycles).
- Applied Oscope on Whitfield data with permuted sample order
  - Top cluster has 69 genes (65 of 69 validated as oscillating in Whitfield).

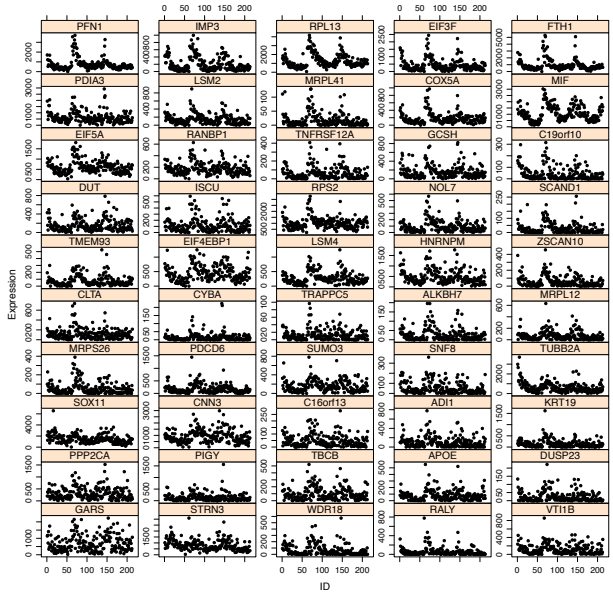


CK JSM 2018

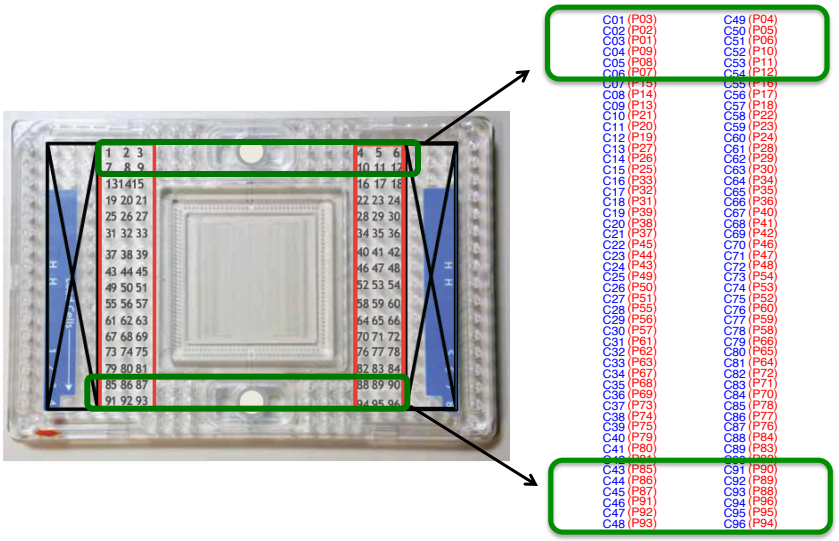
Oscope: Results from H1 hESCs (with Thomson lab)

- Oscope applied to 213 H1 hESCs identified a 29 gene group
  - 21 of 29 genes annotated as cell-cycle by GO.
- To investigate this group, Oscope was reapplied to 460 H1 hESCs
  - 213 unlabeled and 247 FUCCI labeled (cell cycle phase is known).

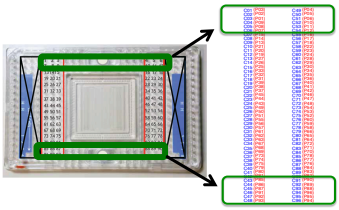
Oscope identifies potential artifact in Fluidigm C1 platform



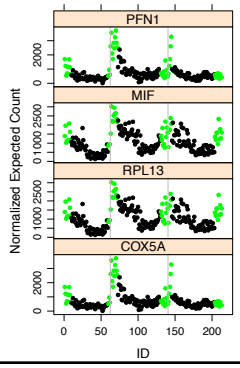
Schematic of Fluidigm's C1 platform

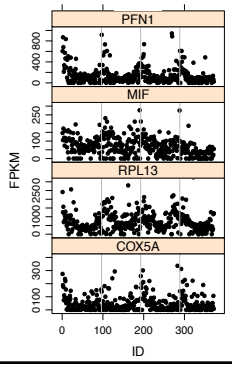


Increased expression related to capture site ID

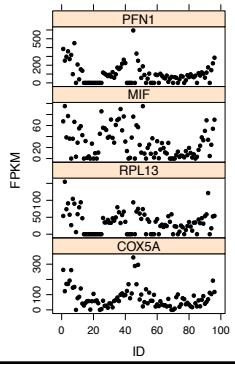


Trapnell *et al*., 2014





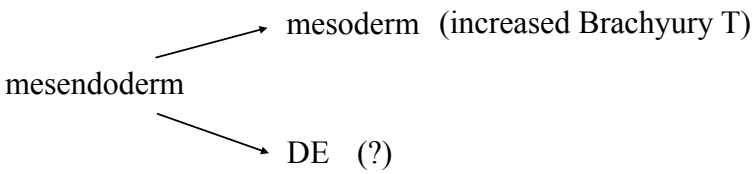
Wu *et al*., 2014



Study of human embryonic stem cell (hESC) differentiation

Definitive endoderm

- DE cells
  - give rise to specialized cell types that line the developing gut tube and contribute to liver, stomach, lungs…
  - are an instrumental resource for regenerative medicine.
  - are not well characterized wrt transition to DE state.



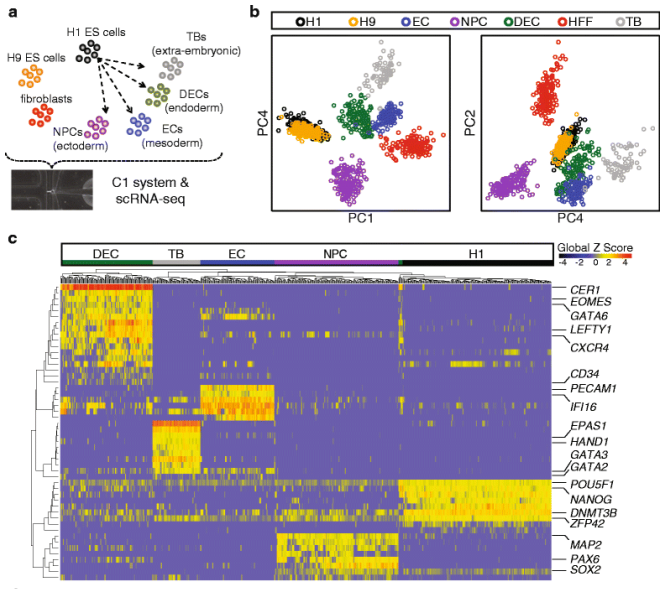
- What signals promote DE differentiation?
- Can we identify the window during which mesendoderm -> DE?

Data

- scRNA-seq snapshot of progenitor cell types
  - 1018 cells – 7 cell types (H1, H9, DEC, EC, HFF, NPC, TB)
- scRNA-seq time course
  - 758 H1 cells – 6 time points (0, 12, 24, 36, 72, 96 hours)
- Bulk RNA-seq snapshot of progenitor cell types
  - 19 samples – 7 cell types (H1, H9, DEC, EC, HFF, NPC, TB)
- Bulk RNA-seq time course
  - Triplicates of H1 cells – 6 time points (0, 12, 24, 36, 72, 96 hours)

CK JSM 2018

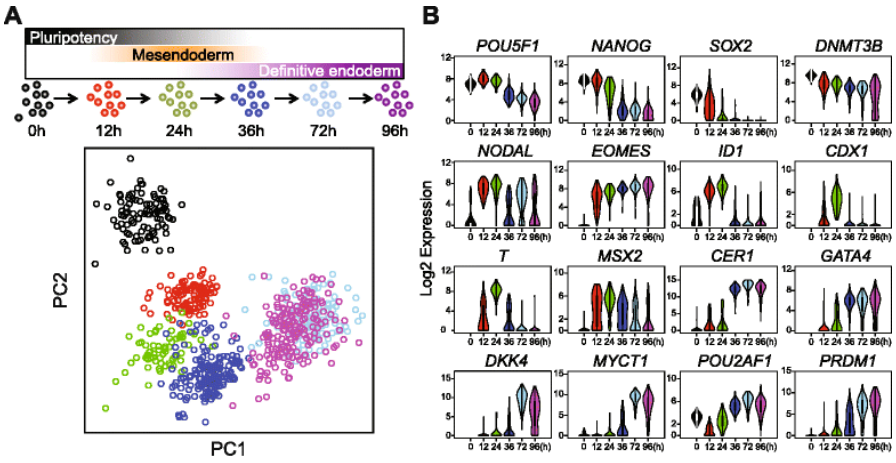
Overview of results



CK JSM 2018

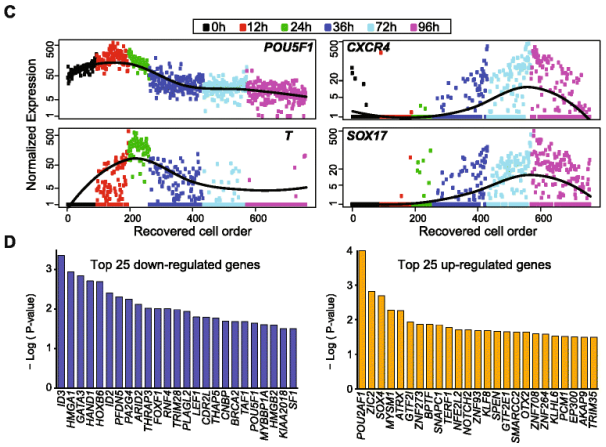
Overview of results

- Early version of Trendy identifies genes that differ among stages



Overview of results

- Wavecrest reconstructs temporal order of differentiation genes
- KLF8 identified/ experimentally validated as novel positive regulator of DE differentiation



CK JSM 2018

Summary

- OEFinder was used to remove genes with ordering effects.
- Trendy and Wavecrest applied to scRNA-seq time course experiments identify KLF8 as a novel regulator of DE (knockdown/overexpression experiments confirm that this is the case).

Chu, Leng *et al.*, *Genome Biology*, 2016

CK JSM 2018

Acknowledgements

Rhonda Bacher, PhD
Ning Leng, PhD
Keegan Korthauer, PhD
Jared Brown
Ziyue Wang
Ying Li
Zijian Ni

Jamie Thomson, VMD, PhD
Ron Stewart, PhD
Li-Fang Chu, PhD

Biomedical Computing Group
NIH R01GM076274
NIH U54 AI117924