Single-cell sequencing

Background

- Most of the biological experiments are performed on "bulk" samples, which contains a large number of cells (millions).
- The high-throughput data we introduced so far are all "bulk" data, which measures the average (gene expression, TF binding, methylation, etc.) of many cells.
- The bulk measurement ignores the inter-cellular heterogeneities:
  - Different cell types.
  - Biological variation among the same type of cell.

Single-cell biology

- The study of individual cells.
- The cells are isolated from multi-cellular organism.
- Experiment is performed for each cell individually.
- Provides more detailed, higher resolution information.
- High-throughput experiments on single cell is possible.

Single cell sequencing

- Perform different types of sequencing at the single-cell level:
  - DNA-seq
  - ATAC-seq
  - BS-seq
  - RNA-seq
- Very active research field in the past few years.
- Major challenges:
  - Cell isolation.
  - Amplification of genomic material.
  - Data analysis.

Basic experimental procedure

- Isolation of single cell. Techniques include
  - Laser-capture microdissection (LCM)
  - Fluorescence-activated cell sorting (FACS)
  - Microfluidics
- Open the cell and obtain DNA/mRNA/etc.
- PCR amplification to get enough materials.
- Perform sequencing.
- Note that single cell sequencing usually has higher error rates than bulk data.

Single cell DNA-seq (scDNA-seq)

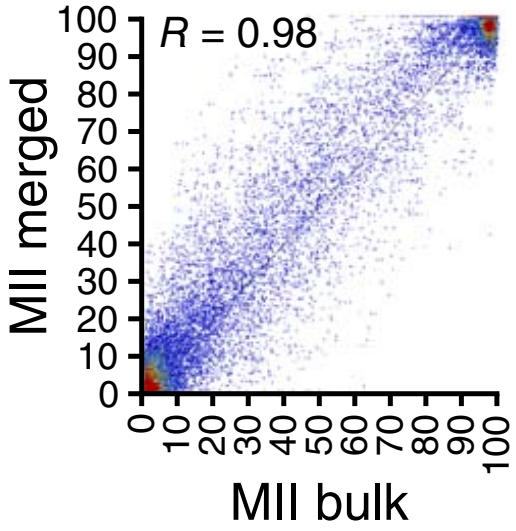
- For a comprehensive review, read Gawad *et al.* (2016) NRG.
- Examples of biological applications:
  - Identify and assemble the genome of unculturable microorganisms.
  - Determine the contribution of intra-tumor genetic heterogeneity in cancer development of treatment response.

scDNA-seq data analysis

- Single cell variant calling:
  - Bulk data can be used as reference to reduce false positives.
  - Combine data from several cells.
  - Software: Monovar (Zafar *et al.* 2016 Nat. Method.)
- Determining genetic relationship among single cells:
  - This is a clustering problem. Cells can be put into groups or a phylogenetic tree based on similarity of variants.
  - Methods are mostly ad hoc.

Single cell BS-seq (scBS-seq)

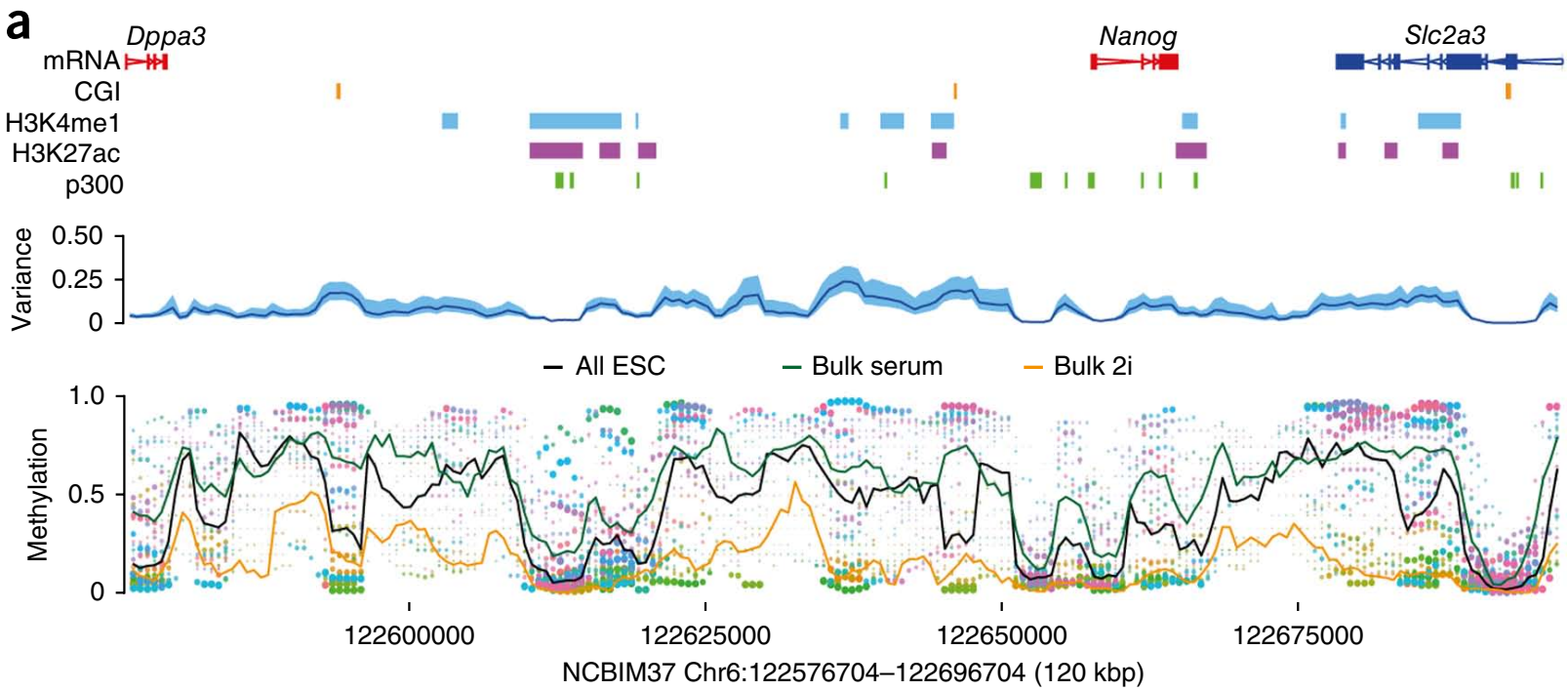
- Similar to scDNA-seq, but with bisulfite treatment before sequencing.
- There's scWGBS and scRRBS.
- The methylation levels from scBS-seq should be 0/1, with some exceptions caused by technical artifacts.
- Merged single cell and bulk data have good correlation.



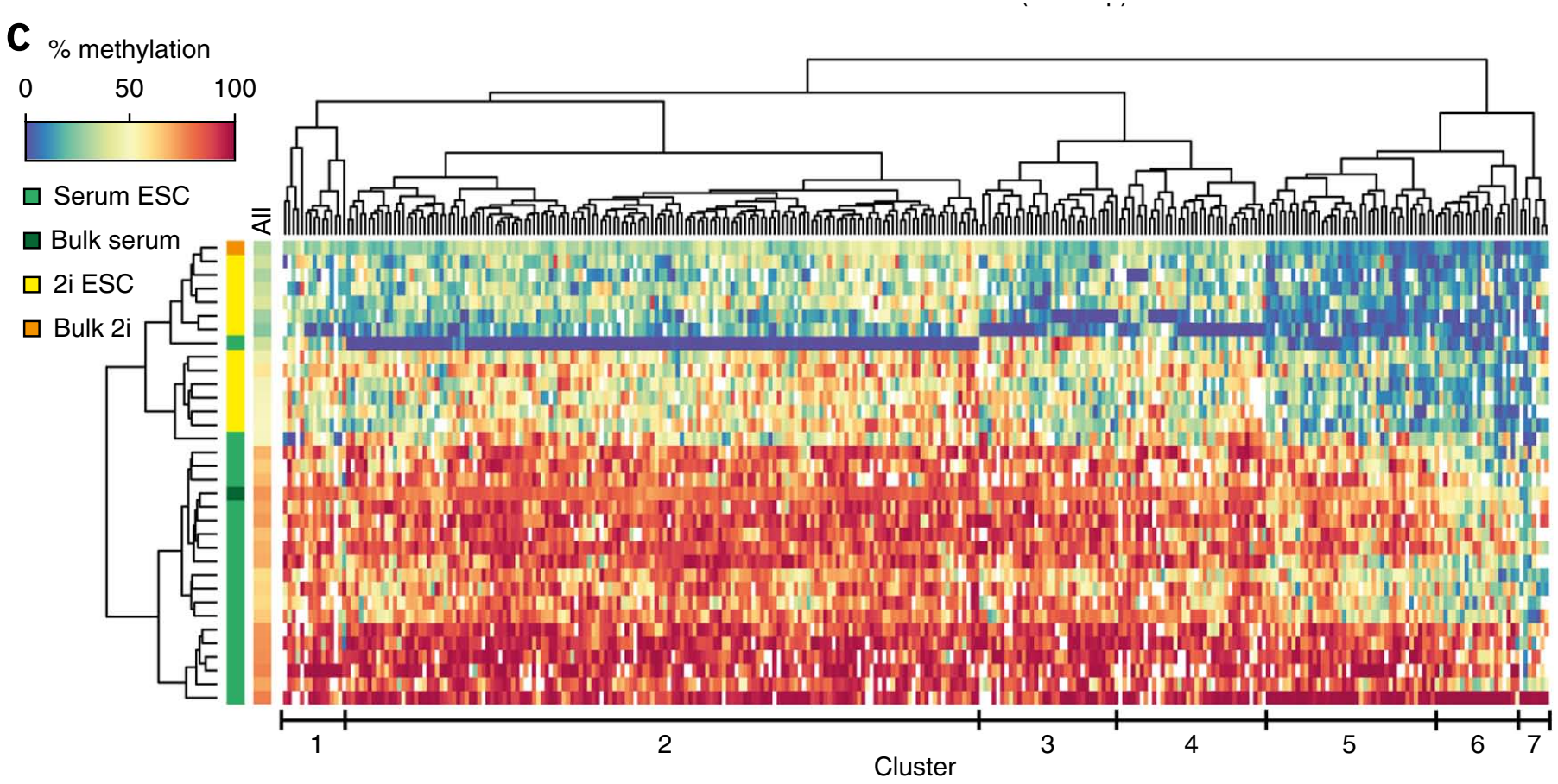
Smallwood et al. 2014, NM

scBS-seq data analysis

- So far the data analysis are mostly descriptive:
  - compute variations among cells
  - Cell clustering
- Lots of rooms for method development



Smallwood et al. 2014, NM



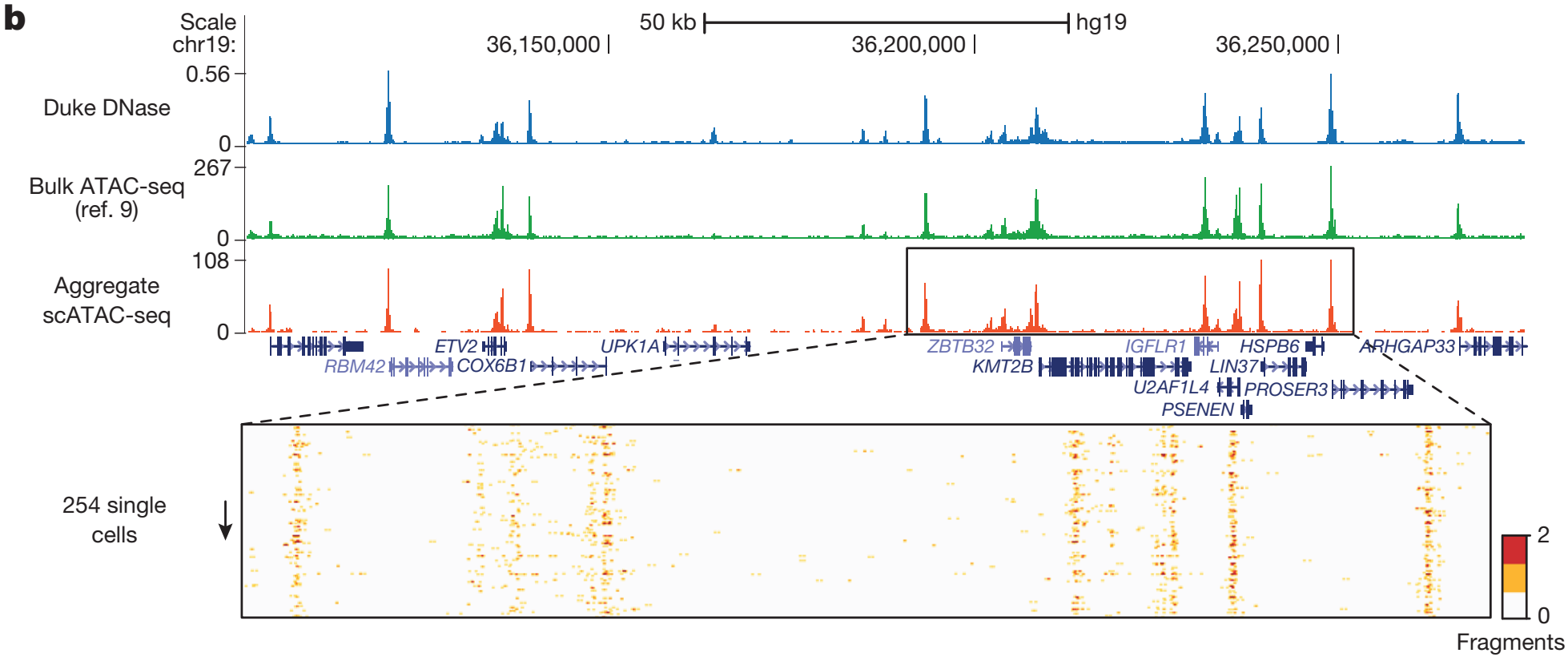
Smallwood et al. 2014, NM

Single cell ChIP/ATAC-seq

- ATAC-seq: similar to DNase-seq, profile the active genomic regions. Data look like ChIP-seq.
- A few papers:
  - Rotem et al. (2015) NBT: scChIP-seq
  - Buenrostro et al. (2015) Nature: scATAC-seq

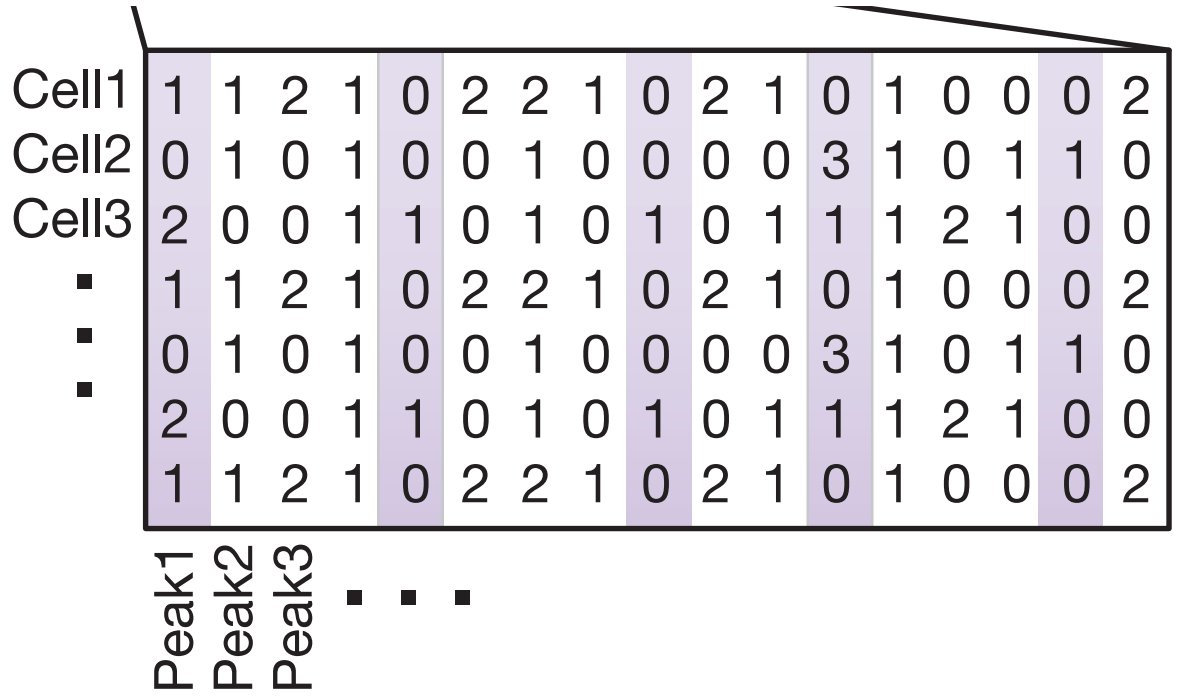
scChIP/scATAC-seq data

- Aggregated sc data has good agreement with bulk.



Buenrostro et al. 2015 Nature

- Very sparse: one a few reads at peak regions.
  - Extremely low signal to noise ratio.
  - Peak calling have to be based on combined data, or rely on other prior information

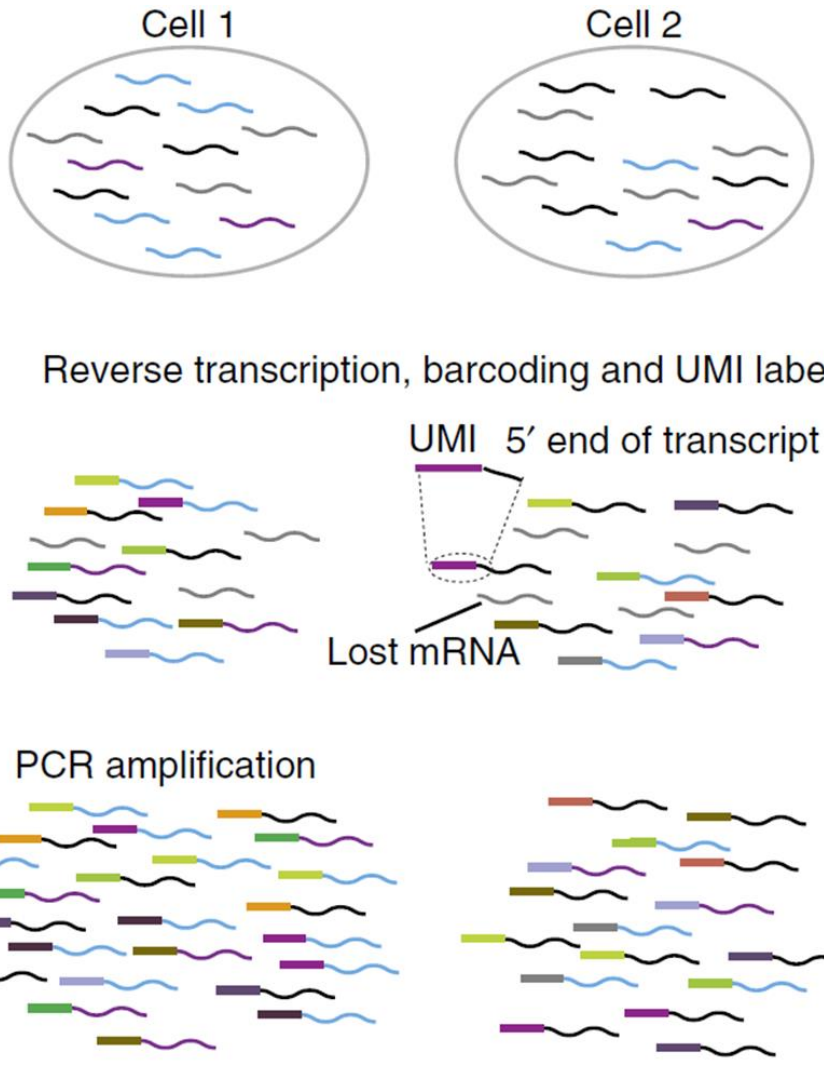


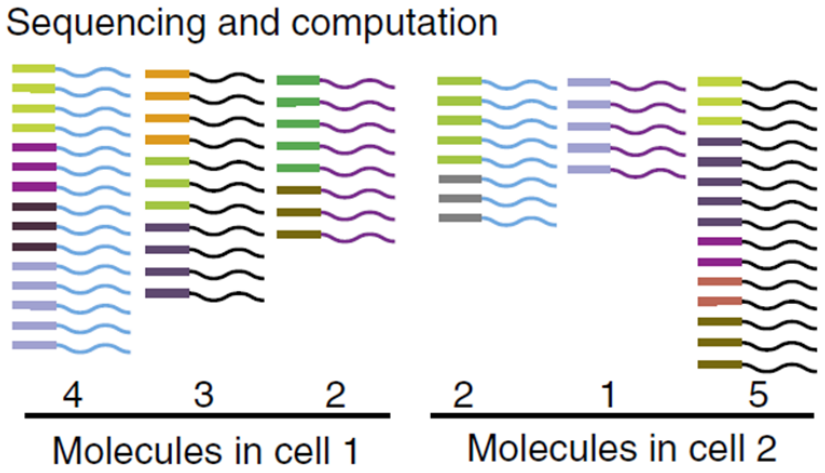
Buenrostro et al. 2015 Nature

Single cell RNA-seq (scRNA-seq)

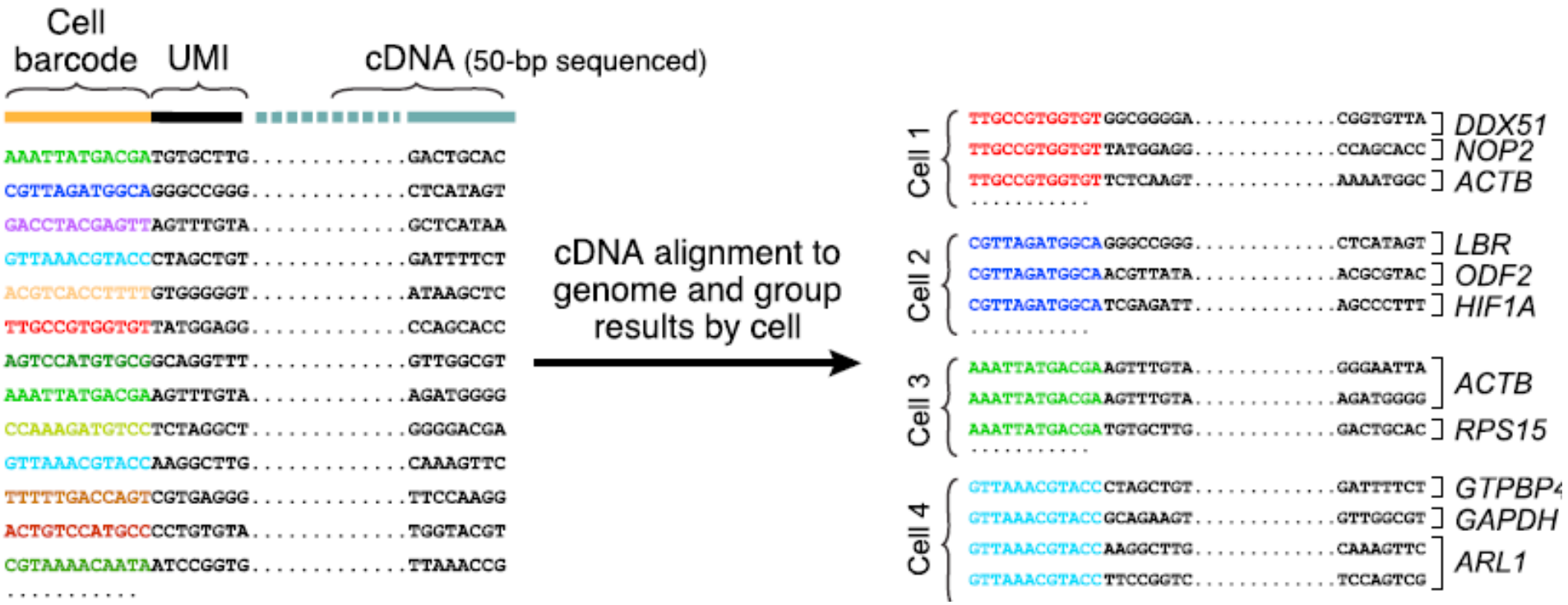
- The most active in the sc field.
- Scientific goals:
  - Understand the gene expression heterogeneity within the same sample.
  - Composition of different types of cell in complex tissues, such as brain, cancer, etc.
  - Above can be explored spatially, temporally, or under different biological condition.
- Raw data are the same as bulk RNA-seq, can be aligned using the same software.

Experimental procedure





Saiful Islam ··· Sten Linnarsson



Macosko et al. 2015 Cell

Some data characteristics

- Number of transcripts detected is much lower compared to bulk RNA-seq, perhaps due to low capture and reverse transcription efficiencies.

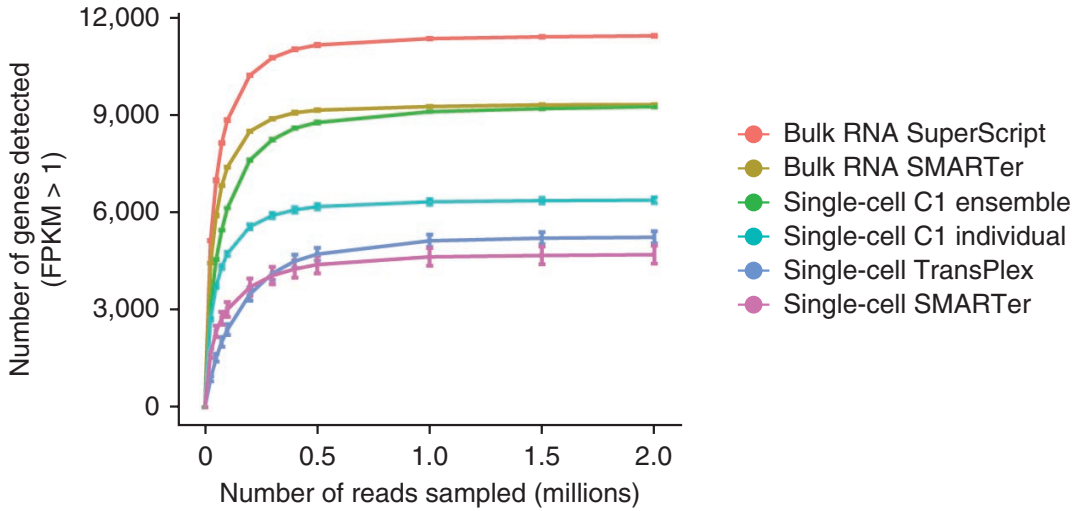
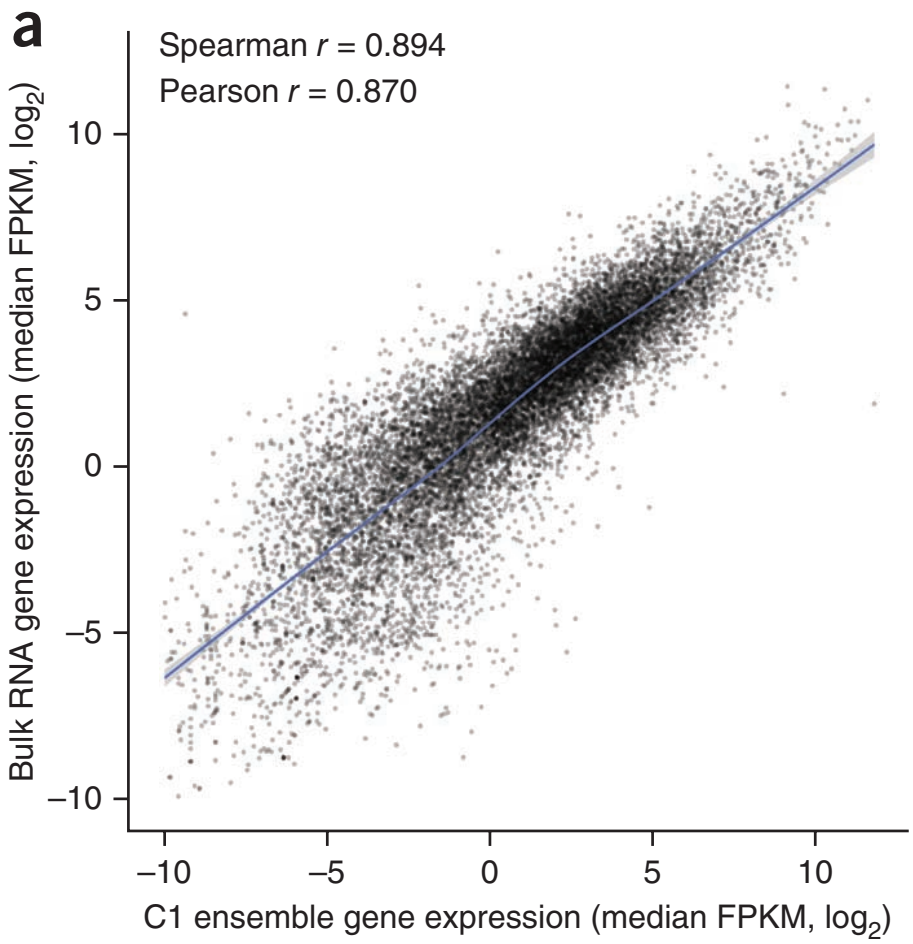


Figure 5 | Saturation curves for the different sample preparation methods. Each point on the curve was generated by randomly selecting a number of raw reads from each sample library and then using the same alignment pipeline to call genes with mean FPKM >1. Each point represents four replicate subsamplings. Error bars, standard error.

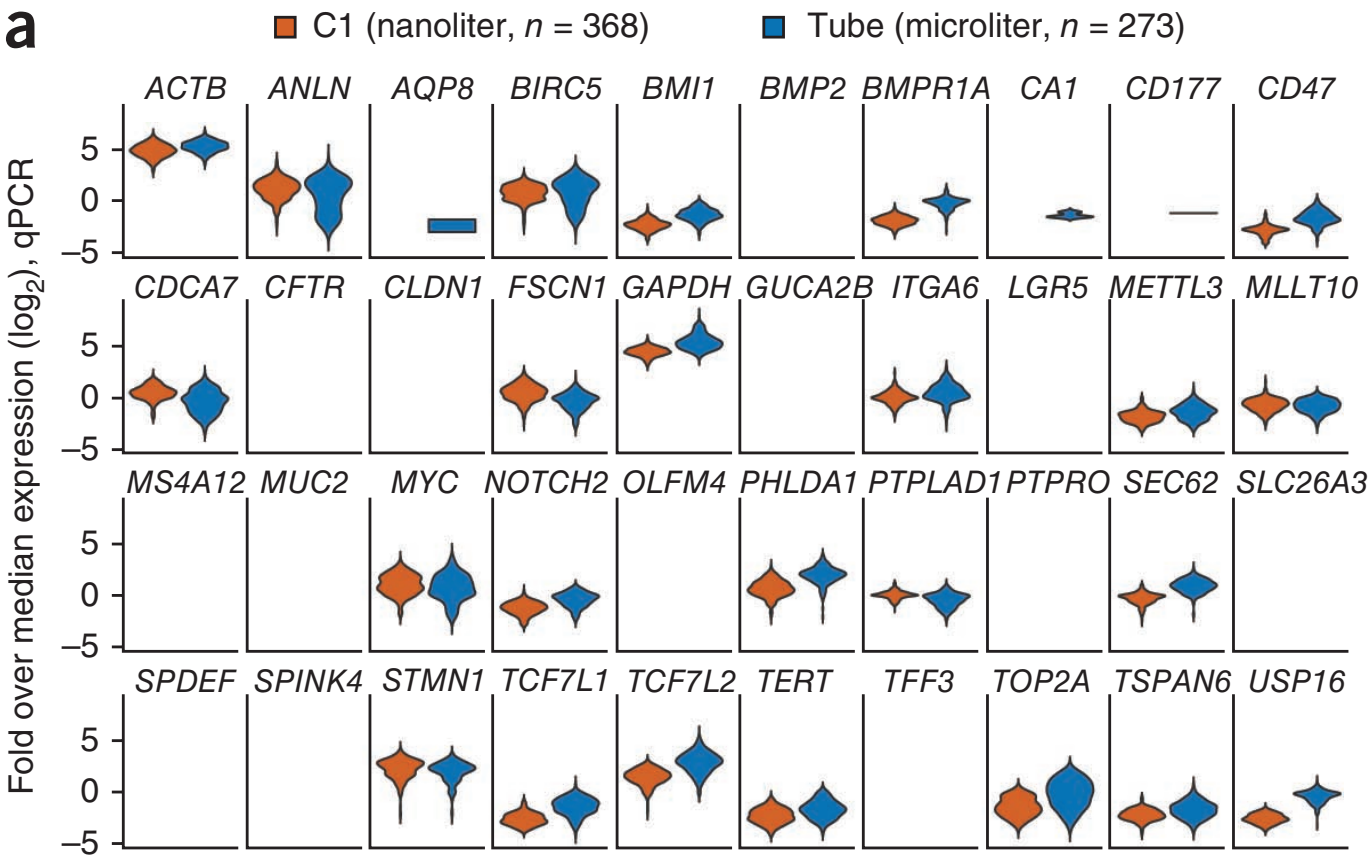
Wu et al. 2013 Nature Method

- Bulk and aggregated single cell expressions have good correlation.



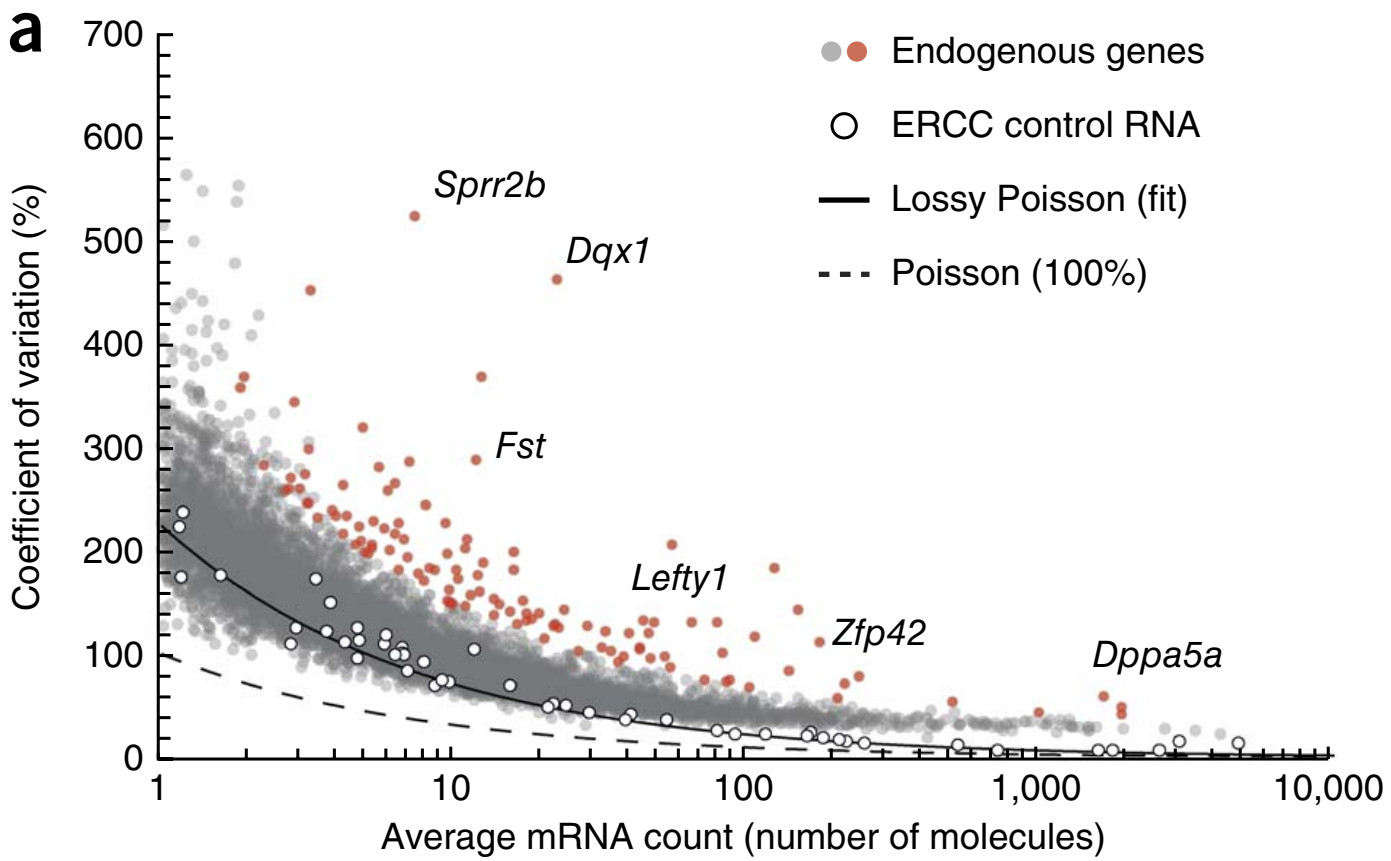
Wu et al. 2013 Nature Method

- Expression levels for a gene in different cells sometimes show bimodal distribution.



Wu et al. 2013 Nature Method

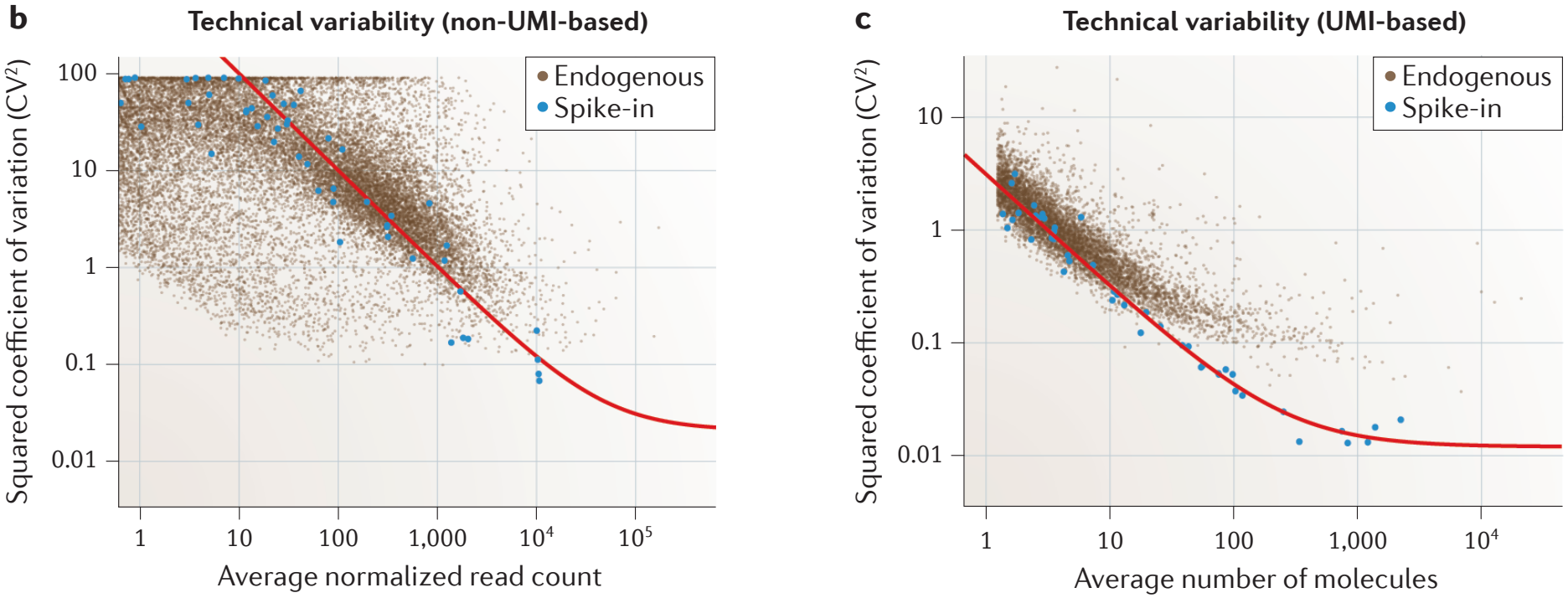
- Negative correlation between mean expression and biological variation (same as in bulk).



Islam et al. 2013 Nature Method

Normalization issues

- scRNA-seq is very noisy.
- Spike-in data is usually available.
  - Spike-ins from the external RNA Control Consortium (ERCC) panel, which contains 92 synthetic spikes based on bacterial genome.
- UMI (unique molecule identifier) is sometimes used to barcode the molecules for estimating amplification noise.
- A combination of spike-in and UMI can potentially be used for data normalization.



Stegle et al. 2015 NRG

Existing work for scRNA-seq normalization

Application Note

Normalization and noise reduction for single cell RNA-seq experiments

Bo Ding[1,#], Lina Zheng[1,#], Yun Zhu[1], Nan Li[1], Haiyang Jia[1,2], Rizi Ai[1], Andre Wildberg[1] and Wei Wang[1,3*]

[1]Department of Chemistry and Biochemistry, University of California, La Jolla, CA 92093, USA,

[2] College of Computer Science and Technology, Jilin University, Changchun 130012, China.

[3]Department of Cellular and Molecular Medicine, University of California, La Jolla, CA 92093, USA,

[#]Equal contribution

Associate Editor: Dr. Ziv Bar-Joseph

ABSTRACT

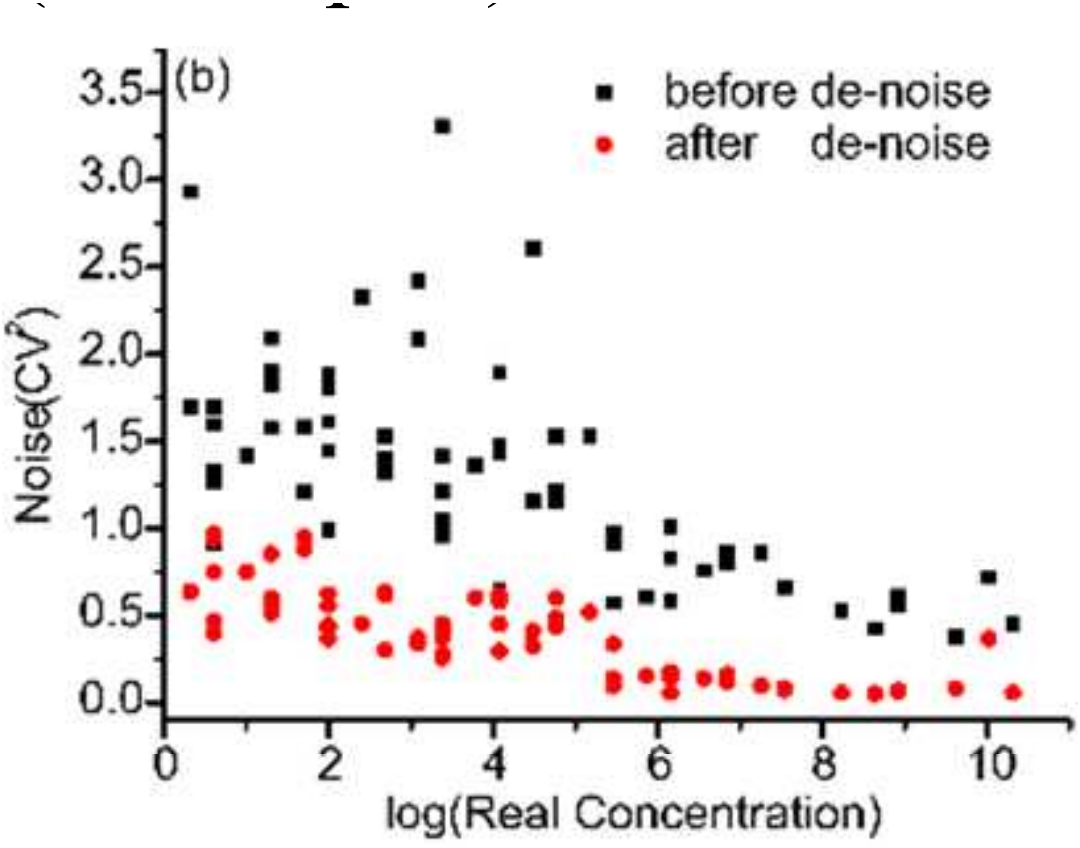
- Log-transform FPKM values, denoted by x.
- Assume the expression value, y, follow Gamma distribution. The mean of Gamma is a polynomial function of x: $y = \mu(x)$.

$\mu(x) = \sum_{i=0}^{n} \beta_i x^i$. The model is the following:

$$y \sim Gamma(y; \mu(x), \varphi)$$

- Use MLE to estimate parameters based on ERCC data. Then the fitted model is applied to all genes to estimate concentration.

- Results: reduce CV cross cells.



Lun *et al. Genome Biology* (2016) 17:75
DOI 10.1186/s13059-016-0947-7

Genome Biology

METHOD Open Access

Pooling across cells to normalize single-cell RNA sequencing data with many zero counts

Aaron T. L. Lun[1*], Karsten Bach[2] and John C. Marioni[1,2,3*]

- Works for data without spike-in.
- The goal is to estimate a size factor for each cell.
- The idea is to normalize on summed expression values from pools of cells – it's more stable than using individual cell.

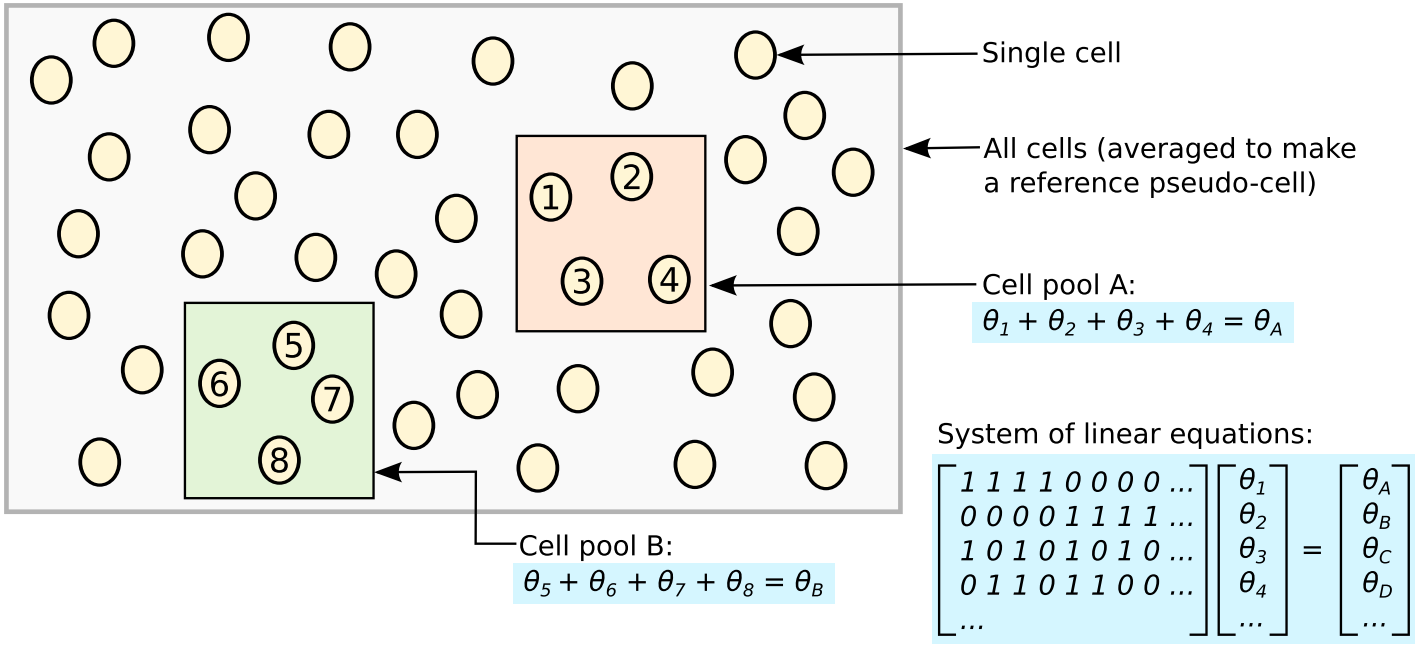


Fig. 3 Schematic of the deconvolution method. All cells in the data set are averaged to make a reference pseudo-cell. Expression values for cells in pool A are summed together and normalized against the reference to yield a pool-based size factor θ_A. This is equal to the sum of the cell-based factors θ_j for cells $j = 1$–4 and can be used to formulate a linear equation. (For simplicity, the t_j term is assumed to be unity here.) Repeating this for multiple pools (e.g., pool B) leads to the construction of a linear system that can be solved to estimate θ_j for each cell j

Differential expression

- Traditional methods test mean changes.
- Due to the biomodal distribution of the GE in scRNA-seq, the consideration and modeling of "drop-out" event (non-expressed) is very important.
- A few existing work, but lots of room for method development.

Bayesian approach to single-cell differential expression analysis

740 | VOL.11 NO.7 | JULY 2014 | NATURE METHODS

Peter V Kharchenko[1–3], Lev Silberstein[3–5] & David T Scadden[3–5]

- SCDE (single-cell differential expression).
- Use a mixture of a Poisson with small rate (dropout) and negative binomial (expressed) to model the expression: $p(x \mid r_c, \Omega_c) = p_d(x) p_{Poisson}(x) + (1 - p_d(x)) p_{NB}(x \mid r_c)$
- The DE is based on Bayesian inference. But the derivation in this paper is messy.

Finak *et al. Genome Biology* (2015) 16:278
DOI 10.1186/s13059-015-0844-5

Genome Biology

METHOD Open Access

MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data

Greg Finak[1†], Andrew McDavid[1†], Masanao Yajima[1†], Jingyuan Deng[1], Vivian Gersuk[2], Alex K. Shalek[3,4,5,6], Chloe K. Slichter[1], Hannah W. Miller[1], M. Juliana McElrath[1], Martin Prlic[1], Peter S. Linsley[2] and Raphael Gottardo[1,7*]

- MAST: “Model-based Analysis of Single- cell Transcriptomics.”

MAST for DE

- Main ideas:
  - Use log2(TPM+1) as input data
  - Both dropout probability and expression level depends on experimental conditions.

$$logit\left(Pr(Z_{ig}=1)\right) = \mathrm{X_i}\,\beta_g^{\mathrm{D}}$$

$$\Pr\left(Y_{ig}=y|Z_{ig}=\ 1\right) =\ \mathrm{N}\left(\mathrm{X_i}\beta_g^{\mathrm{C}},\ \sigma_g^2\right)$$

  - Model fitting with some regularization.
  - DE is based on chi-square or Wald test.

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

Cole Trapnell[1,2,6], Davide Cacchiarelli[1–3,6], Jonna Grimsby[2], Prapti Pokharel[2], Shuqiang Li[4], Michael Morse[1,2], Niall J Lennon[2], Kenneth J Livak[4], Tarjei S Mikkelsen[1–3] & John L Rinn[1,2,5]

- Monocle: part of “tuxedo suite” for scRNA-seq analysis.
- Works for DE and clustering.
- Main idea for DE:
  - Model data with observed and dropout: $Y = \begin{cases} Y^* & \text{if } Y^* > \lambda \\ \lambda & \text{if } Y^* \leq \lambda \end{cases}$
  - Use a generalized additive model (GAM) for design:

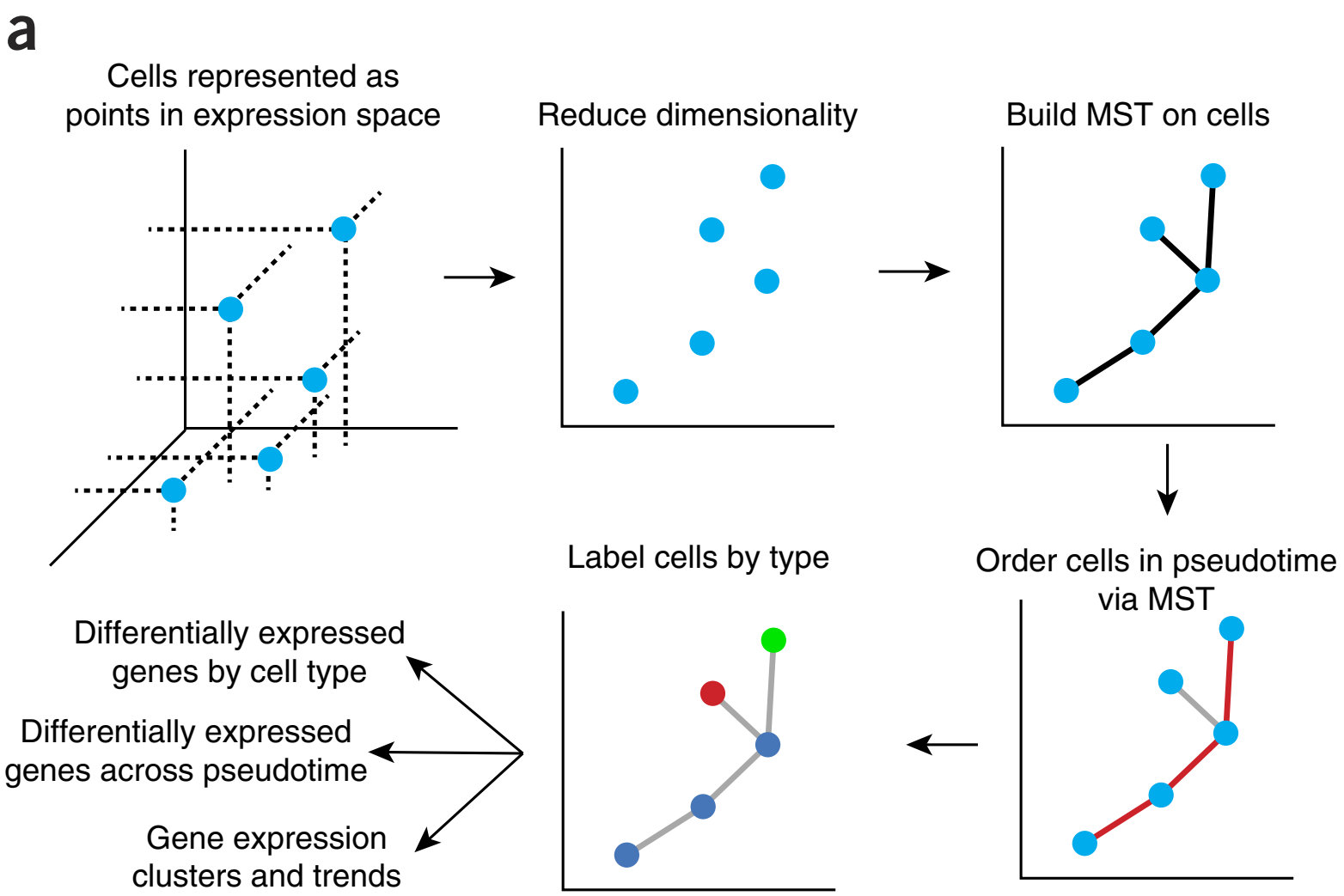
    $$g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)$$

  - DE is tested from the GAM.

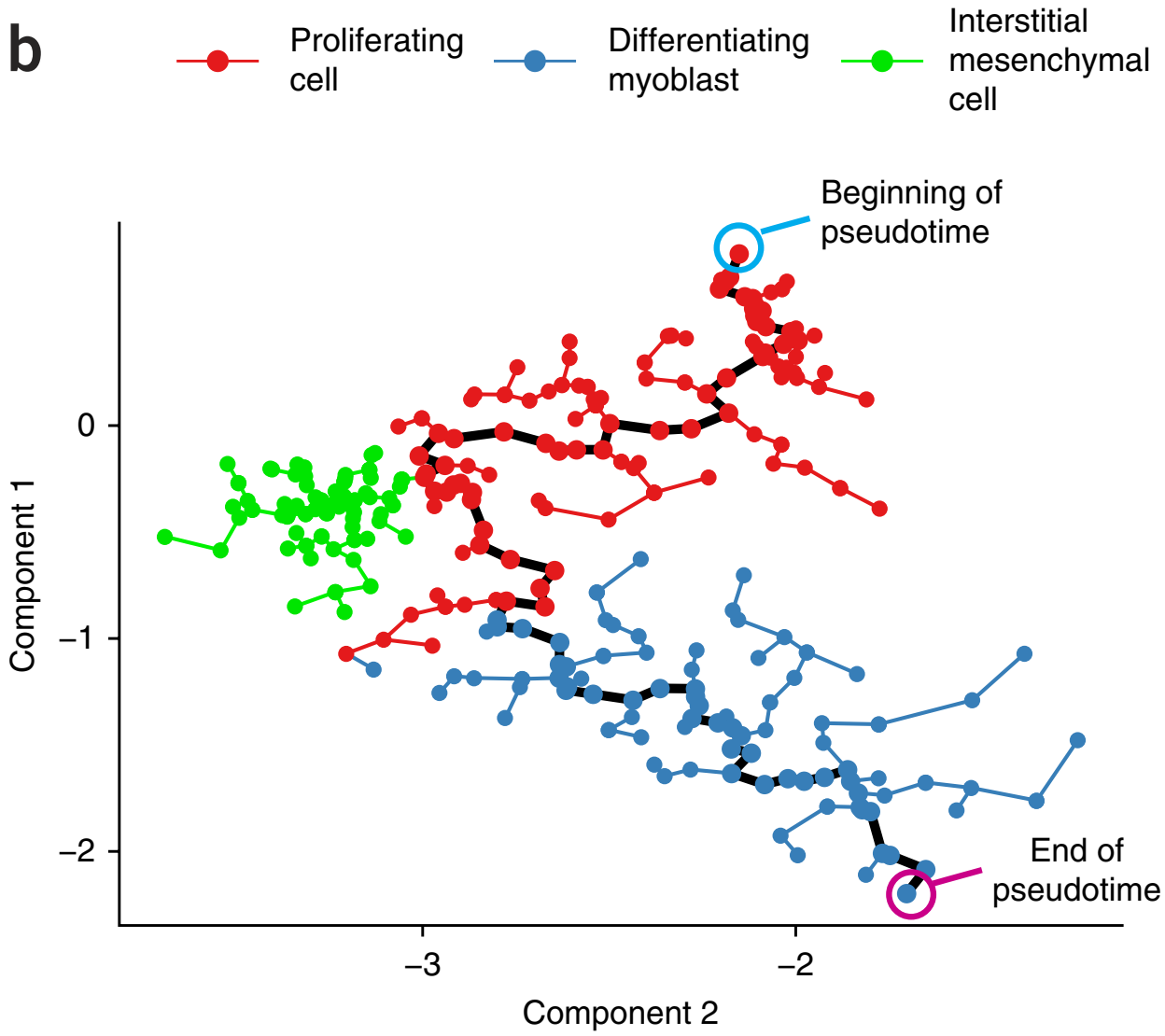
Cell clustering

- Perhaps the most active topic in scRNA-seq.
- The goals include:
  - Cluster cells into subgroups.
  - Model temporal transcriptomic dynamics: reconstruct "pseudo-time" for cells. This is useful for understanding development or disease progression.
- Traditional method like k-means or hierarchical clustering need to be used with caution due to dropout events.

Monocle


a
Cells represented as
points in expression space
Reduce dimensionality
Build MST on cells
Order cells in pseudotime
via MST
Label cells by type
Differentially expressed
genes by cell type
Differentially expressed
genes across pseudotime
Gene expression
clusters and trends

Monocle result



Use Monocle Bioconductor package

First create a `CellDataSet` object using `newCellDataSet` function, then:

- Differential expression using `differentialGeneTest`.
- Cell ordering (pseudo-time estimation). This contains three steps:
  - Select a list of genes (often the DE genes) used for cell ordering. Use `setOrderingFilter` function to set that.
  - Dimension reduction using `reduceDimension` function.
  - Cell ordering using `orderCells` function.

```
### Create data object
pd <- new("AnnotatedDataFrame", data = sample_sheet)
fd <- new("AnnotatedDataFrame", data = gene_annotation)
dataobj <- newCellDataSet(as.matrix(expr_matrix),
                          phenoData = pd, featureData = fd)
### DE test
diff_test_res <- differentialGeneTest(dataobj,
    fullModelFormulaStr=GE~cond”,
    reducedModelFormulaStr=“GE~1”)

### cell ordering
ordering_genes <- row.names(subset(diff_test_res, qval < 0.1))
dataobj <- setOrderingFilter(dataobj, ordering_genes)
dataobj <- reduceDimension(dataobj, use_irlba=FALSE)
dataobj <- orderCells(dataobj, num_paths=2, reverse=TRUE)
plot_spanning_tree(dataobj)
```

Other similar software

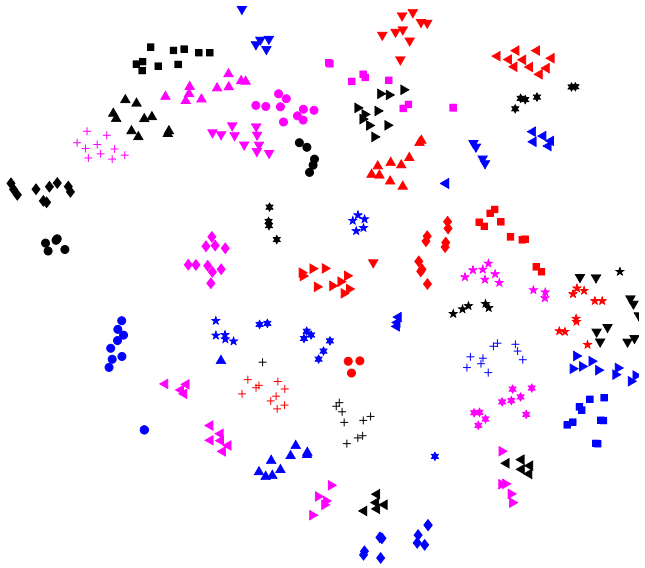
- Waterfall: Shin et al. (2015) Cell Stem Cell
- Wanderlust: Bendall et al. (2014) Cell
- TSCAN: Ji et al. (2016) NAR
- Ideas are similar:
  - Select informative genes.
  - Dimension reduction of GE.
  - Cluster the cells based on reduced data. Often want to over-cluster them to have many groups.
  - Construct a MST (miminum spanning tree) from the clustering results.
  - Map cells to the MST.

Detect rare cell type

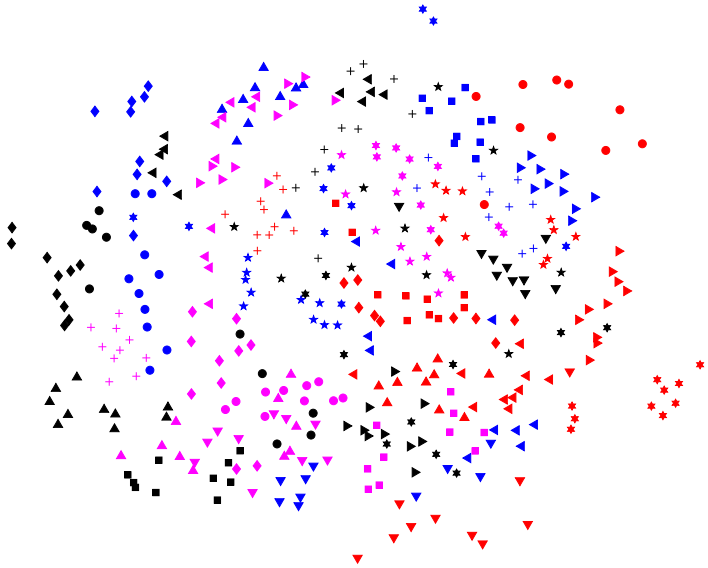
- Rare cell are “outliers” in the data.
- RaceID (Grun et al. 2015 Nature):
  - Normalize and log-transformed data.
  - Filter cells and genes
  - K-means clustering
  - Detect outliers from k-means result.
- GiniClust (Jiang et al. 2016 GB):
  - Difference is the gene filtering. It uses gini-index instead of variance to select genes.

t-SNE: a useful visualization tool

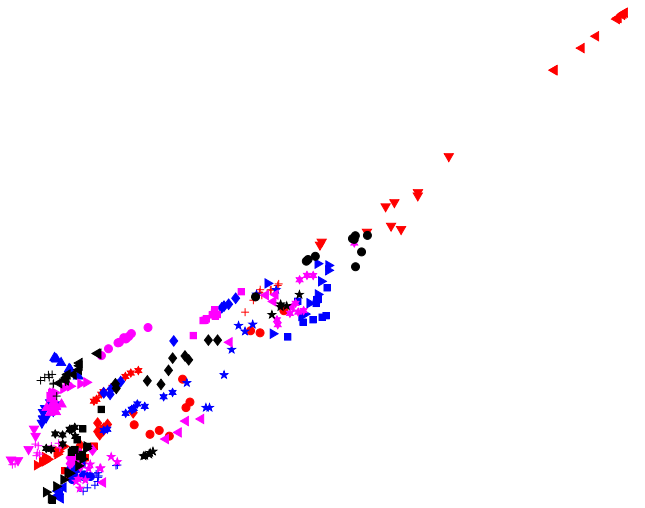
- t-SNE (t-distributed stochastic neighbor embedding): visualize high-dimensional data on 2-/3-D map.
- When project high-dimensional data into lower dimensional space, preserve the distances among data points.
  - This alleviate the problem that many clusters overlap on low dimensional space.
- Try to make the pairwise distances of points similar in high and low dimension.
- This is used in almost all scRNA-seq data visualization.
- Has “tsne” package on CRAN.



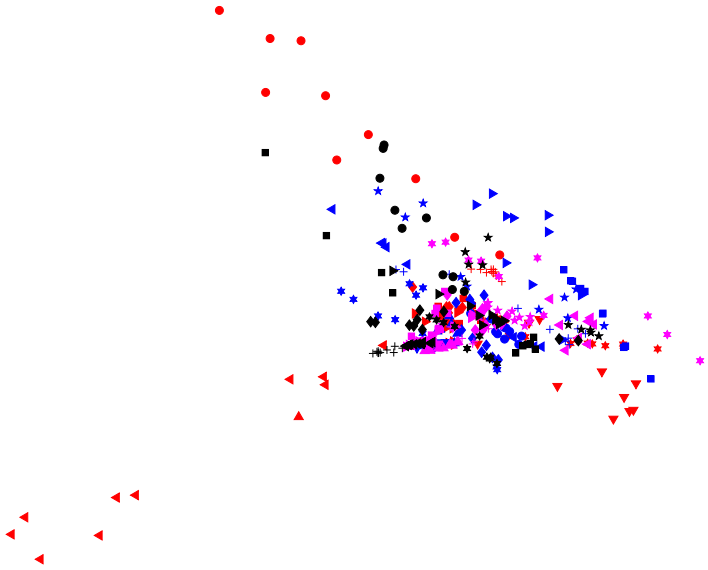
(a) Visualization by t-SNE.



(b) Visualization by Sammon mapping.



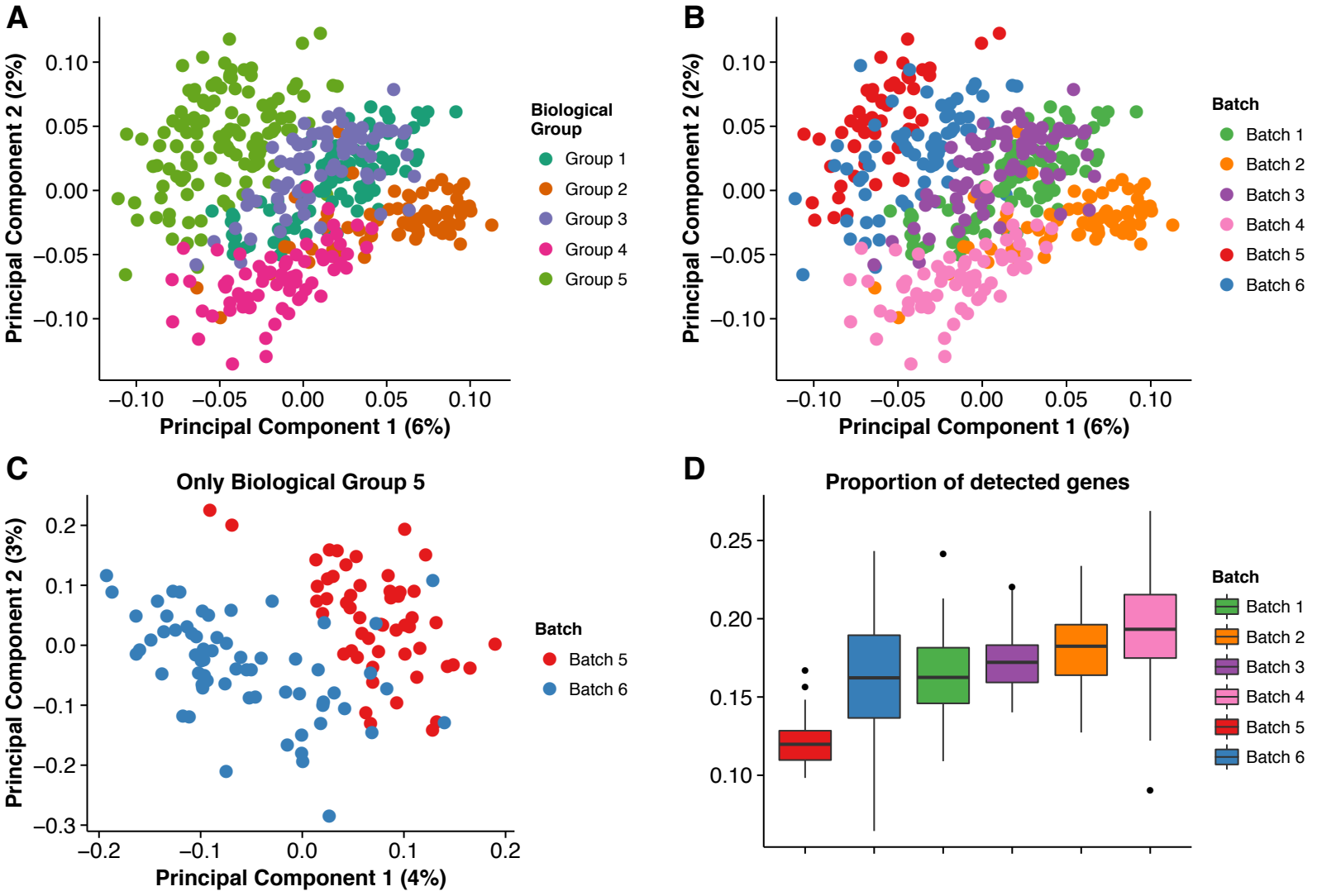
(c) Visualization by Isomap.



(d) Visualization by LLE.

Batch effect in scRNA-seq

(Hicks et al. 2016, bioRxiv)



Summary for scRNA-seq

- The main interests are inter-cellular heterogeneity, expression dynamics, cell type discovery.
- Statistical questions include normalization, differential expression and clustering.
- Batch effect could be a huge problem, and difficult to overcome.
- Rooms for model development.

Single cell GE microarray

PNAS

Single cell-derived clonal analysis of human glioblastoma links functional and genomic heterogeneity

**Mona Meyer[a,1], Jüri Reimand[b,c,1], Xiaoyang Lan[a,b], Renee Head[a], Xueming Zhu[a], Michelle Kushida[a], J
Jessica C. Pressey[e], Anath C. Lionel[b,f], Ian D. Clarke[a,g], Michael Cusimano[h], Jeremy A. Squire[i], Stephe
Mark Bernstein[j], Melanie A. Woodin[e], Gary D. Bader[b,c,2], and Peter B. Dirks[a,b,k,2]**

[a]Division of Neurosurgery, Program in Developmental and Stem Cell Biology, Arthur and Sonia Labatt Brain Tumour Research Cer

Single cell lncRNA

Genome Biology

BioMed Central
The Open Access Publisher

Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution

Genome Biology

doi:10.1186/s13059-015-0586-4

Moran N Cabili (nmcabili@broadinstitute.org)
Margaret C Dunagin (dunagin@seas.upenn.edu)
Patrick D McClanahan (pmccl@seas.upenn.edu)
Andrew Biaesch (biaescha@gmail.com)
Olivia Padovan-Merhar (opadovan@sas.upenn.edu)
Aviv Regev (aregev@broadinstitute.org)
John L Rinn (john_rinn@harvard.edu)
Arjun Raj (arjunraj@seas.upenn.edu)

Published online: 29 January 2015

Grand summary for scSeq

- Single-cell biology reveals a lot of information that can’t be detected from bulk data.
- Data are much noisier, and more difficult to analyze.
- Methods are still under-developed, but quickly catching up.