CS-E5875 High-Throughput Bioinformatics
Single-cell sequencing

Harri Lähdesmäki

Department of Computer Science
Aalto University

November 24, 2020

Contents

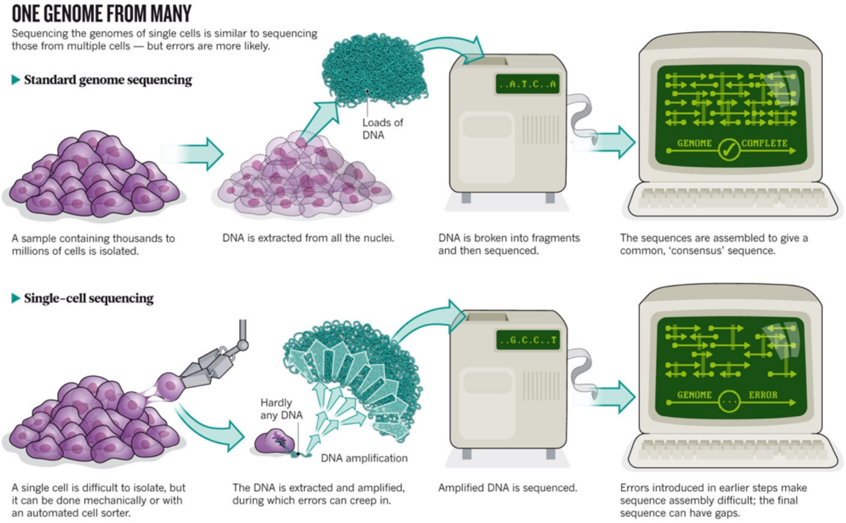
- Background & Motivation
- Single cell sequencing technologies
- Single cell sequencing data analysis: overview
- Single cell sequencing data analysis: emerging methods

Background & Motivation

- Most genomic profiling methods analyze cell populations
- We know that even cells of the same cell type can be different
  - Genome: somatic mutations
  - Transcriptome
  - Epigenome
  - ...
- Recent technology development has made it possible to characterize individual cells at the level of
  - Transcriptome/RNA
  - DNA
  - Proteome
  - DNA methylation
  - Histone modifications
  - Chromatin accessibility

Background & Motivation

- Bulk sequencing vs. single-cell sequencing



ONE GENOME FROM MANY

Sequencing the genomes of single cells is similar to sequencing those from multiple cells — but errors are more likely.

Standard genome sequencing

A sample containing thousands to millions of cells is isolated.

DNA is extracted from all the nuclei.

DNA is broken into fragments and then sequenced.

The sequences are assembled to give a common, 'consensus' sequence.

Single-cell sequencing

A single cell is difficult to isolate, but it can be done mechanically or with an automated cell sorter.

The DNA is extracted and amplified, during which errors can creep in.

Amplified DNA is sequenced.

Errors introduced in earlier steps make sequence assembly difficult; the final sequence can have gaps.

Figure from https://scitechdaily.com/images/one-genome-from-many.jpg

Background & Motivation

- Single cell analysis has many important applications in molecular biology, biomedicine, etc.
- For example, blood is a complex organ
  - Molecular level profiling of whole blood sample provides average measurements across about 20 cell types present in blood

  → Single-cell technologies can extract information separately for different blood cell types

Background & Motivation

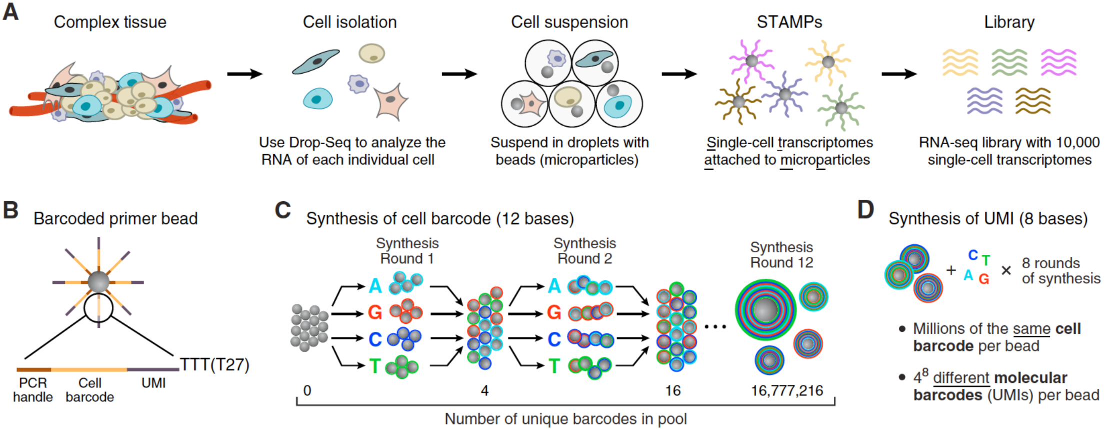
- Single cell analysis has many important applications in molecular biology, biomedicine, etc.
- For example, blood is a complex organ
  - Molecular level profiling of whole blood sample provides average measurements across about 20 cell types present in blood

  → Single-cell technologies can extract information separately for different blood cell types
- Cancer research can greatly benefit from single cell technologies because
  - Cancer can originate from a single cell
  - Cancer progression can involve rare cell types that are difficult to quantify otherwise
  - Tumour biopsies are heterogeneous, contain infiltrating cell types, etc.
- . . .

Contents

- Background & Motivation
- Single cell sequencing technologies
- Single cell sequencing data analysis: overview
- Single cell sequencing data analysis: emerging methods

DROP-seq



A
Complex tissue
Cell isolation
Cell suspension
STAMPs
Library
Use Drop-Seq to analyze the RNA of each individual cell
Suspend in droplets with beads (microparticles)
Single-cell transcriptomes attached to microparticles
RNA-seq library with 10,000 single-cell transcriptomes
B
Barcoded primer bead
TTT(T27)
PCR handle
Cell barcode
UMI
C
Synthesis of cell barcode (12 bases)
Synthesis Round 1
Synthesis Round 2
Synthesis Round 12
A
G
C
T
0
4
16
16,777,216
Number of unique barcodes in pool
D
Synthesis of UMI (8 bases)
C T A G
+
×
8 rounds of synthesis

- Millions of the same **cell barcode** per bead
- 4^8 different **molecular barcodes** (UMIs) per bead

Figure from (Macosko et al, 2015)

DROP-seq

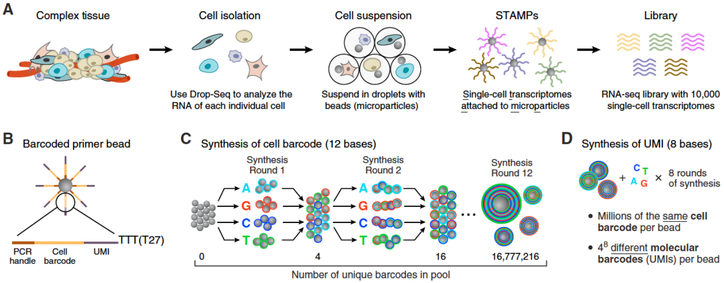

A
Complex tissue
Cell isolation
Use Drop-Seq to analyze the RNA of each individual cell
Cell suspension
Suspend in droplets with beads (microparticles)
STAMPs
Single-cell transcriptomes attached to microparticles
Library
RNA-seq library with 10,000 single-cell transcriptomes
B
Barcoded primer bead
TTT(T27)
PCR handle
Cell barcode
UMI
C
Synthesis of cell barcode (12 bases)
Synthesis Round 1
Synthesis Round 2
Synthesis Round 12
A
G
C
T
0
4
16
16,777,216
Number of unique barcodes in pool
D
Synthesis of UMI (8 bases)
+ C T A G × 8 rounds of synthesis
• Millions of the same cell barcode per bead
• 4^8 different molecular barcodes (UMIs) per bead

Figure 1. Molecular Barcoding of Cellular Transcriptomes in Droplets

(A) Drop-Seq barcoding schematic. A complex tissue is dissociated into individual cells, which are then encapsulated in droplets together with microparticles (gray circles) that deliver barcoded primers. Each cell is lysed within a droplet; its mRNAs bind to the primers on its companion microparticle. The mRNAs are reverse-transcribed into cDNAs, generating a set of beads called "single-cell transcriptomes attached to microparticles" (STAMPs). The barcoded STAMPs can then be amplified in pools for high-throughput mRNA-seq to analyze any desired number of individual cells.

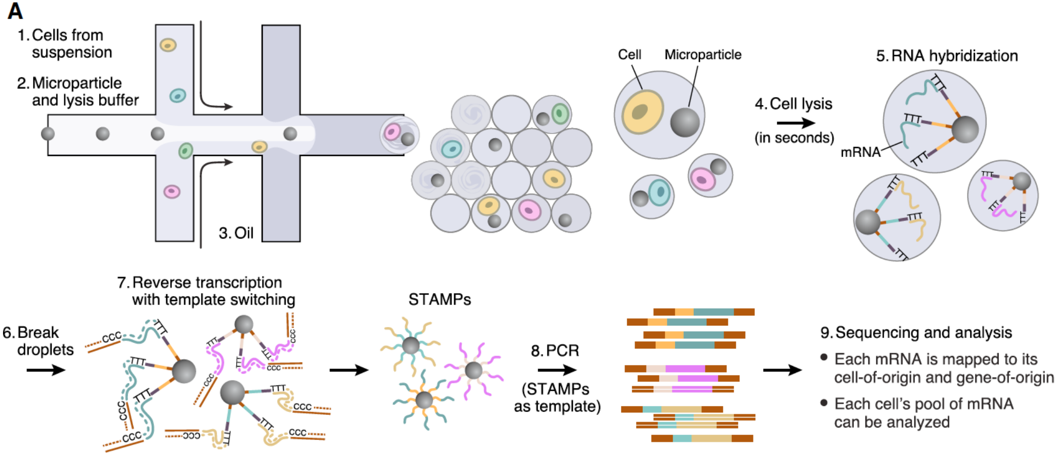
(B) Sequence of primers on the microparticle. The primers on all beads contain a common sequence ("PCR handle") to enable PCR amplification after STAMP formation. Each microparticle contains more than 10^8 individual primers that share the same "cell barcode" (C) but have different unique molecular identifiers (UMIs), enabling mRNA transcripts to be digitally counted (D). A 30-bp oligo dT sequence is present at the end of all primer sequences for capture of mRNAs.

(C) Split-and-pool synthesis of the cell barcode. To generate the cell barcode, the pool of microparticles is repeatedly split into four equally sized oligonucleotide synthesis reactions, to which one of the four DNA bases is added, and then pooled together after each cycle, in a total of 12 split-pool cycles. The barcode synthesized on any individual bead reflects that bead's unique path through the series of synthesis reactions. The result is a pool of microparticles, each possessing one of 4^{12} (16,777,216) possible sequences on its entire complement of primers (see also Figure S1).

(D) Synthesis of a unique molecular identifier (UMI). Following the completion of the "split-and-pool" synthesis cycles, all microparticles are together subjected to eight rounds of degenerate synthesis with all four DNA bases available during each cycle, such that each individual primer receives one of 4^8 (65,536) possible sequences (UMIs).

Figure from (Macosko et al, 2015)

DROP-seq



9. Sequencing and analysis

- Each mRNA is mapped to its cell-of-origin and gene-of-origin
- Each cell's pool of mRNA can be analyzed

Figure from (Macosko et al, 2015)

DROP-seq

- Paired-end sequencing reads the barcodes and the actual RNA fragment/gene

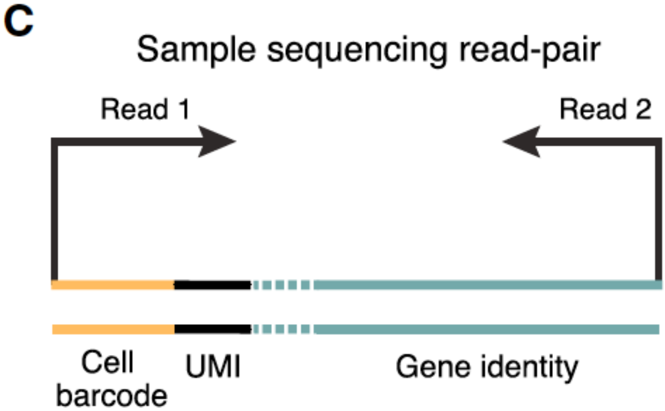
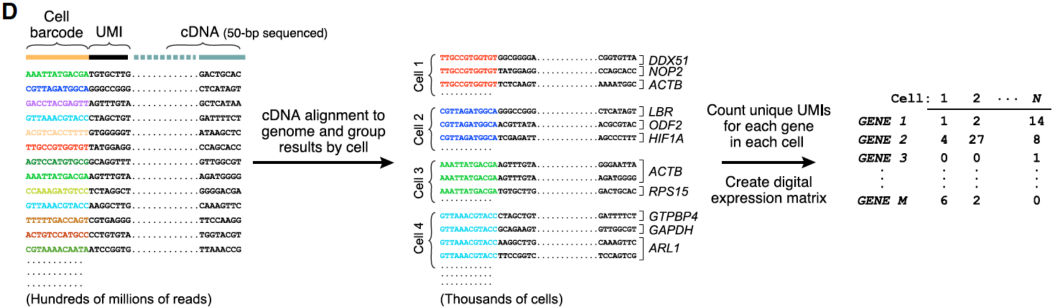


Figure from (Macosko et al, 2015)

DROP-seq

- Analysis of the paired-end sequencing reads from DROP-seq distinguishes cells and UMIs
- Assign each read to the "closest" cell based on the cell barcode


D
Cell barcode
UMI
cDNA (50-bp sequenced)
AAATTATGACGATGTGCTTG.............GACTGCAC
CGTTAGATGGCAGGGCCGGG.............CTCATAGT
GACCTACGAGTTAGTTTGTA.............GCTCATAA
GTTAAACGTACCCTAGCTGT.............GATTTTCT
ACGTCACCTTTTGTGGGGGT.............ATAAGCTC
TTGCCGTGGTGTTATGGAGG.............CCAGCACC
AGTCCATGTGCGGCAGGTTT.............GTTGGCGT
AAATTATGACGAAGTTTGTA.............AGATGGGG
CCAAAGATGTCCTCTAGGCT.............GGGGACGA
GTTAAACGTACCAAGGCTTG.............CAAAGTTC
TTTTTGACCAGTCGTGAGGG.............TTCCAAGG
ACTGTCCATGCCCCTGTGTA.............TGGTACGT
CGTAAAACAATAATCCGGTG.............TTAAACCG
(Hundreds of millions of reads)
cDNA alignment to genome and group results by cell
Cell 1
TTGCCGTGGTGTGGCGGGGA.............CGGTGTTA DDX51
TTGCCGTGGTGTTATGGAGG.............CCAGCACC NOP2
TTGCCGTGGTGTTCTCAAGT.............AAAATGGC ACTB
Cell 2
CGTTAGATGGCAGGGCCGGG.............CTCATAGT LBR
CGTTAGATGGCAACGTTATA.............ACGCGTAC ODF2
CGTTAGATGGCATCGAGATT.............AGCCCTTT HIF1A
Cell 3
AAATTATGACGAAGTTTGTA.............GGGAATTA
AAATTATGACGAAGTTTGTA.............AGATGGGG ACTB
AAATTATGACGATGTGCTTG.............GACTGCAC RPS15
Cell 4
GTTAAACGTACCCTAGCTGT.............GATTTTCT GTPBP4
GTTAAACGTACCGCAGAAGT.............GTTGGCGT GAPDH
GTTAAACGTACCAAGGCTTG.............CAAAGTTC
GTTAAACGTACCTTCCGGTC.............TCCAGTCG ARL1
(Thousands of cells)
Count unique UMIs for each gene in each cell
Create digital expression matrix

Cell:	1	2	...	N
GENE 1	1	2		14
GENE 2	4	27		8
GENE 3	0	0		1
⋮	⋮	⋮		⋮
GENE M	6	2		0

Figure from (Macosko et al, 2015)

DROP-seq

- (Figure caption from (Macosko et al, 2015))

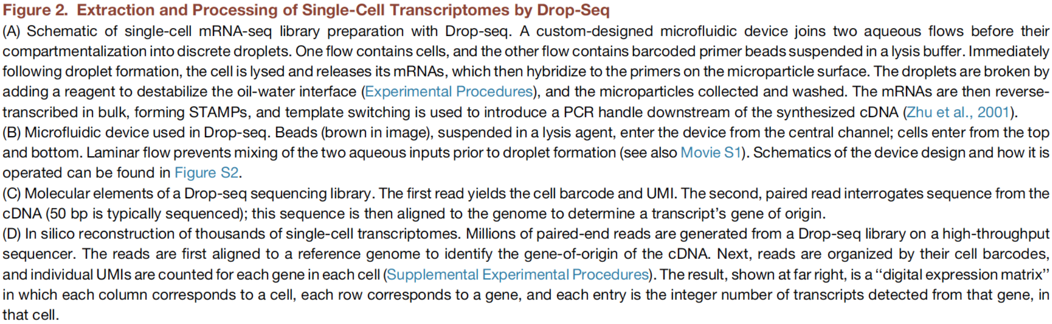


Figure 2. Extraction and Processing of Single-Cell Transcriptomes by Drop-Seq

(A) Schematic of single-cell mRNA-seq library preparation with Drop-seq. A custom-designed microfluidic device joins two aqueous flows before their compartmentalization into discrete droplets. One flow contains cells, and the other flow contains barcoded primer beads suspended in a lysis buffer. Immediately following droplet formation, the cell is lysed and releases its mRNAs, which then hybridize to the primers on the microparticle surface. The droplets are broken by adding a reagent to destabilize the oil-water interface (Experimental Procedures), and the microparticles collected and washed. The mRNAs are then reverse-transcribed in bulk, forming STAMPs, and template switching is used to introduce a PCR handle downstream of the synthesized cDNA (Zhu et al., 2001).

(B) Microfluidic device used in Drop-seq. Beads (brown in image), suspended in a lysis agent, enter the device from the central channel; cells enter from the top and bottom. Laminar flow prevents mixing of the two aqueous inputs prior to droplet formation (see also Movie S1). Schematics of the device design and how it is operated can be found in Figure S2.

(C) Molecular elements of a Drop-seq sequencing library. The first read yields the cell barcode and UMI. The second, paired read interrogates sequence from the cDNA (50 bp is typically sequenced); this sequence is then aligned to the genome to determine a transcript's gene of origin.

(D) In silico reconstruction of thousands of single-cell transcriptomes. Millions of paired-end reads are generated from a Drop-seq library on a high-throughput sequencer. The reads are first aligned to a reference genome to identify the gene-of-origin of the cDNA. Next, reads are organized by their cell barcodes, and individual UMIs are counted for each gene in each cell (Supplemental Experimental Procedures). The result, shown at far right, is a "digital expression matrix" in which each column corresponds to a cell, each row corresponds to a gene, and each entry is the integer number of transcripts detected from that gene, in that cell.

Figure from (Macosko et al, 2015)

Contents

- Background & Motivation
- Single cell sequencing technologies
- Single cell sequencing data analysis: overview
- Single cell sequencing data analysis: emerging methods

scRNA-seq data analysis

- While single-cell RNA sequencing (scRNA-seq) is structurally similar with data from bulk RNA-seq, scRNA-seq has distinct characters:
  - Abundance of zeros (both biological and technical): only ∼20% of gene expression counts are non-zero
  - Increased variability
  - Complex expression distributions

→ scRNA-seq requires specific analysis methods

Unique molecular identifiers (UMI)

- Due to a very small amount of starting material, RNA library needs to be amplified with PCR
- Many of the sequenced reads are multiple PCR-copies of the original transcripts
- The experimental protocol incorporates so-called unique molecular identifiers (UMI) for each RNA fragment, which can be used to recover the counts of unique RNA molecules
  - The DROP-seq protocol described above has $4^8 = 65536$ different UMIs

→ Align the sequencing read corresponding to the RNA fragment (not the UMI) and then count the unique UMIs for aligned sequencing reads

Unique molecular identifiers (UMI)

- Align the sequencing read corresponding to the RNA fragment (not the UMI) and then count the unique UMIs for aligned sequencing reads
- Because there are “only” $4^8 = 65536$ different UMIs, some truly different RNA fragments can have the same UMI by chance and one of them would be removed if UMI control was applied before alignment

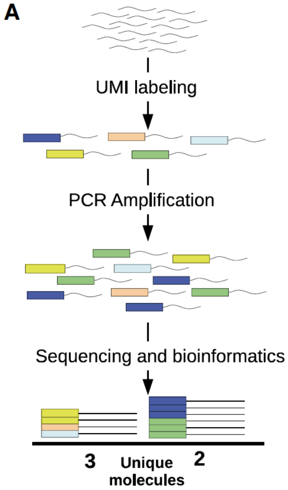


Figure from (Smith et al, 2017)

scRNA-seq analysis to identify cell types

- Single-cell sequencing protocols and analysis methods are under active research and development
- The standard practices and methods have not yet been established
- Lets illustrate how scRNA-seq data can be analyzed using Seurat tool, following a guided tutorial from http://satijalab.org/seurat/, to identify cell types from whole blood sample
- Data is from peripheral blood mononuclear cells (PBMC)
  - → Lots of different cell types
- scRNA-seq from 2700 single PBMC cells
- One of the goals is to identify cells types from the PBMC scRNA-seq data

scRNA-seq analysis: cell and UMI identification

- Sequencing read data is grouped by cells using the cell barcode
- Transcript part of each read is aligned to the genome and unique UMIs are counted for each gene in each cell
- Distributions of cell-specific count data: the number of genes and UMIs

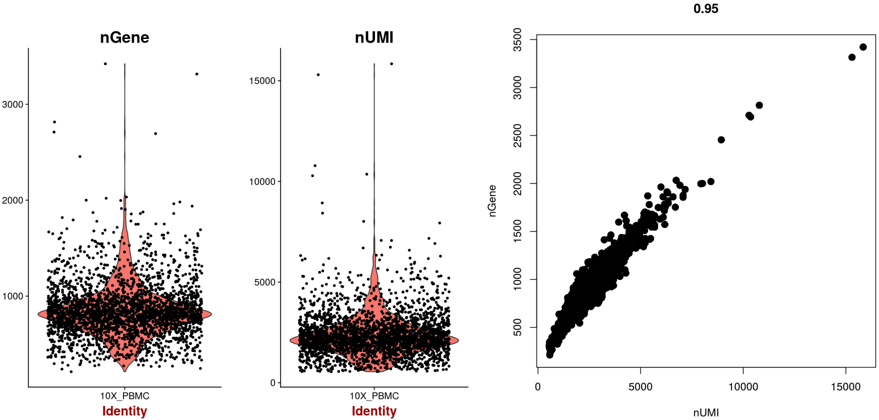


Figure from http://satijalab.org/seurat/

scRNA-seq analysis: normalization

- Recall normalization methods for bulk RNA-seq (e.g. RPKM)
- scRNA-seq data has more complex characteristics

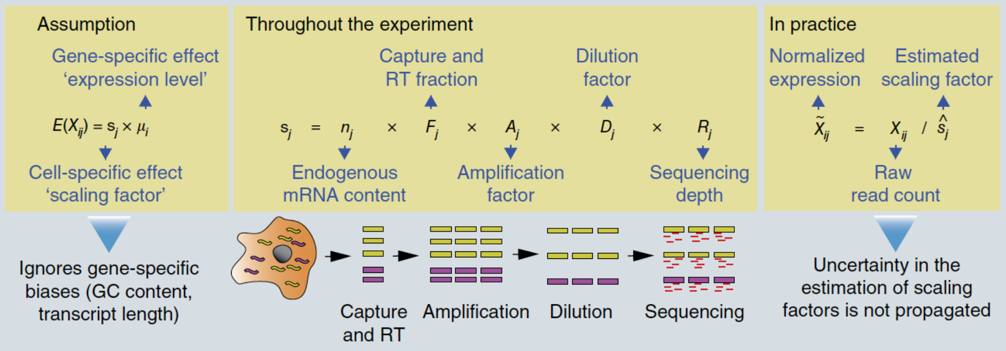
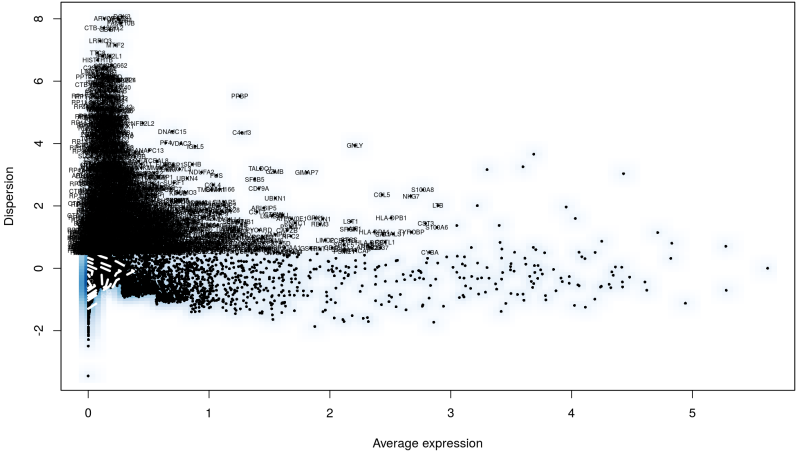

Assumption
Gene-specific effect 'expression level'
$E(X_{ij}) = s_j \times \mu_i$
Cell-specific effect 'scaling factor'
Ignores gene-specific biases (GC content, transcript length)
Throughout the experiment
Capture and RT fraction
Dilution factor
$s_j = n_j \times F_j \times A_j \times D_j \times R_j$
Endogenous mRNA content
Amplification factor
Sequencing depth
Capture and RT
Amplification
Dilution
Sequencing
In practice
Normalized expression
Estimated scaling factor
$\tilde{X}_{ij} = X_{ij} / \hat{s}_j$
Raw read count
Uncertainty in the estimation of scaling factors is not propagated

Figure from (Vallejos et al, 2017)

- Seurat implements a standard normalization: scale each cell by the total read count, multiply by 10000, and take logarithm

scRNA-seq analysis: highly variable genes

- Focus analysis on highly variable genes (across cells)
  - Compute empirical means and dispersions/variances
  - Focus e.g. on ∼2000 genes
  - This is kind of ad-hoc (more principled statistical methods exist)



scRNA-seq analysis: remove unwanted variation

- Remove (or account for) unwanted variation, if needed, from measured read count y_{cg} of gene g and cell c using linear regression and use the regression residuals e_{cg} for downstream analysis
- Possible sources of unwanted variation for cell c
  - Batch effects: $x_{c,\text{batch}}$
  - Biological sources of variation (e.g. cell cycle stage): $x_{c,\text{cycle}}$
  - Sequencing read alignment rate per cell: $x_{c,\text{rate}}$
  - The number of detected molecules $x_{c,\text{UMI}}$ and mitochondrial gene expression $x_{c,\text{mito}}$ per cell c
- For example

$$y_{cg} = a_0 + a_1 x_{c,\text{batch}} + a_2 x_{c,\text{cycle}} + a_3 x_{c,\text{rate}} + a_4 x_{c,\text{UMI}} + a_5 x_{c,\text{mito}} + e_{cg},$$

- Denote the expression residuals for cell c and d genes as $\mathbf{x}_c = [e_{c1}, \ldots, e_{cd}]^T \in \mathbb{R}^d$

scRNA-seq analysis: dimensionality reduction

- Reduce dimensionality further by using principle component analysis (PCA)
- Intuition: find a new basis vector representation and represent the data points in that new basis
- Find the basis vectors so that they are oriented along the largest variation in the data

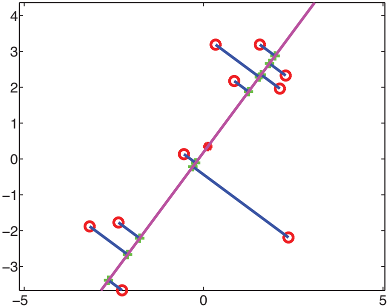


Figure from (Murphy, 2012)

scRNA-seq analysis: dimensionality reduction

- Reduce dimensionality further by using principle component analysis (PCA)
- Normalized expression vectors for C cells $\mathbf{x}_1, \ldots, \mathbf{x}_C$, $\mathbf{x}_i \in \mathbb{R}^d$ (d genes)
- Estimate the covariance matrix

  $$S = \frac{1}{C-1} \sum_{i=1}^{C} (\mathbf{x}_i - \mu_{\mathbf{x}})(\mathbf{x}_i - \mu_{\mathbf{x}})^T,$$

  where $\mu_{\mathbf{x}} = \frac{1}{C} \sum_{i=1}^{C} \mathbf{x}_i$
- The real-valued symmetric covariance matrix S can be written in a diagonalized form $S = V \Lambda V^T$, where $V = [\mathbf{v}_1, \ldots, \mathbf{v}_d]$ contains the orthogonal eigenvectors $\mathbf{v}_i$ as columns and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ is the diagonal matrix with eigenvalues on its diagonal
  - Columns of V and Λ are typically ordered in decreasing order of eigenvalues $\lambda_i \geq \lambda_{i+1}$

scRNA-seq analysis: dimensionality reduction

- Take the $k \leq d$ largest eigenvalues and use the corresponding eigenvectors to form a $d \times k$ matrix $W_k = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$
  - Typically $k \ll d$
- The PCA transformed data are $\mathbf{y}_i = W_k^T \mathbf{x}_i \in \mathbb{R}^k$
- Orthogonal transformation, each component chosen to have the largest variance
- 9 PCA components in this example
  - That is, each cell is represented by a 9-dimensional expression vector

scRNA-seq analysis: visualization

- Visualization of the two most important PCA components

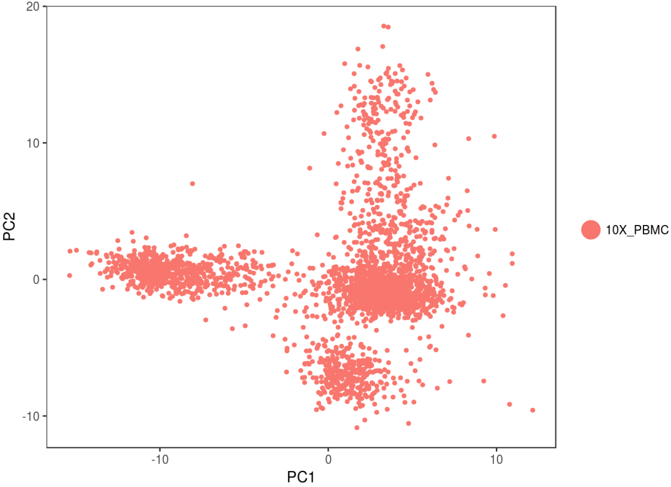


Figure from http://satijalab.org/seurat/

scRNA-seq analysis: clustering

- The final clustering for the 9-dimensional representation of cells using a graph-based clustering method
  - The Euclidean distance between two cells in the PCA space
  - K-nearest neighbor (KNN) graph: edges drawn between cells with similar gene expression profiles
  - The edge weights between any two cells is based on the shared overlap in their local neighborhoods (Jaccard distance)
  - Optimize modularity in the network
- Visualize the clustering result and the data in 2-D using the t-distributed stochastic neighbor embedding (tSNE)

scRNA-seq analysis: visualization

- Visualize the clustering result and the data in 2-D using the t-distributed stochastic neighbor (tSNE) embedding (tSNE)
- tSNE is a nonlinear dimensionality reduction technique
- Input: data in the k-dimensional PCA space $\mathbf{y}_1, \ldots, \mathbf{y}_C$
- Probability distribution centered on $\mathbf{y}_i$: probability of sampling data item $\mathbf{y}_j$

$$p_{j|i} = \frac{\exp(-||\mathbf{y}_j - \mathbf{y}_i||^2/2\sigma_i^2)}{\sum_{k \neq j} \exp(-||\mathbf{y}_k - \mathbf{y}_i||^2/2\sigma_i^2)},$$

  where σ_i^2 is a parameter
- A probability distribution over data item pairs: $\mathbf{y}_i$ and $\mathbf{y}_j$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2C}$$

scRNA-seq analysis: visualization

- tSNE then tries to learn a new map in a lower dimensional space (typically 2-D) such that the similarities between the cells are preserved
  - Similar cells are modeled by nearby points and dissimilar cells are modeled by distant points
  - Distances between cells cannot be maintained exactly in a lower dimensional space, so we need to accept some errors between maps
  - Model such errors robustly using a heavy-tailed distribution
- Motivated by heavy-tailed t-distribution, similarities q_{ij} are defined as

$$q_{ij} = \frac{(1 + ||\mathbf{z}_j - \mathbf{z}_i||^2)^{-1}}{\sum_{k \neq j}(1 + ||\mathbf{z}_k - \mathbf{z}_i||^2)^{-1}}$$

- The locations of the cells $\mathbf{z}_i$ are optimized using e.g. gradient descent such that the (non-symmetric) Kullback-Leibler divergence of the distribution Q from the distribution P is minimized

$$KL(P||Q) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

scRNA-seq analysis: clustering

- Visualization of the clustering results in 2-D using tSNE

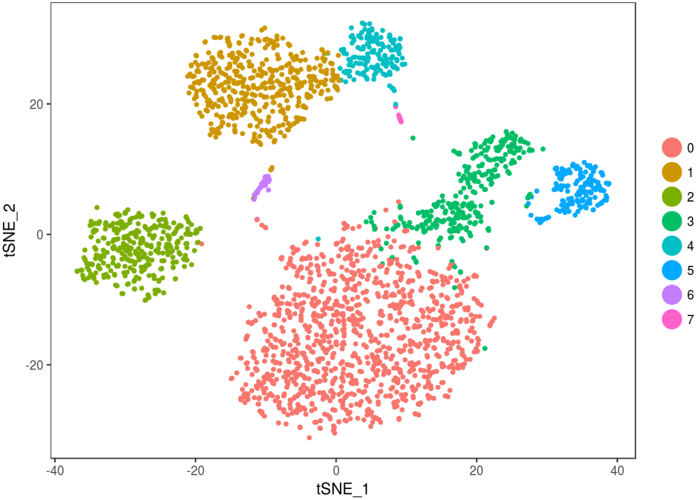


Figure from http://satijalab.org/seurat/

scRNA-seq analysis: cluster biomarkers

- Identify genes differentially expressed between clusters (=cell types)
- Differential expression in one cell type relative to all other cell types

  → Biomarkers for cell types
- Several possible methods
  - t-test
  - Likelihood-ratio test based on zero-inflated models
  - Receiver operating characteristics (ROC) analysis measures classification power for individual genes (used in an example below)

scRNA-seq analysis: cluster biomarkers

- Visualization of cell type specific biomarkers

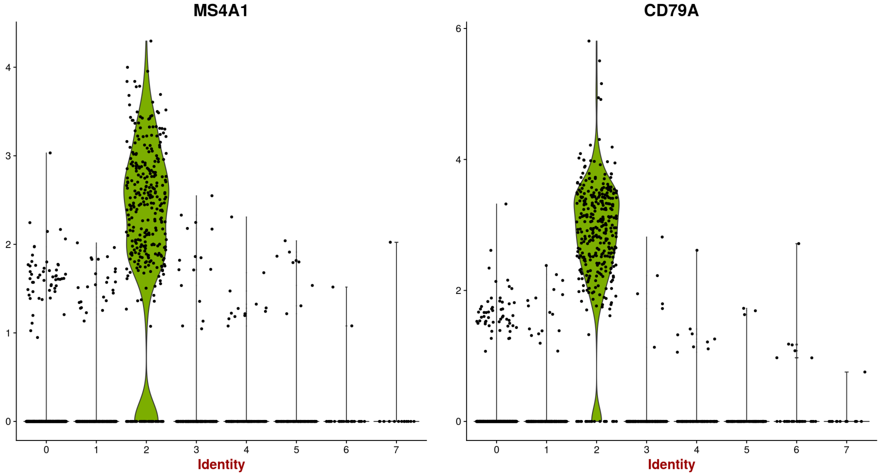


Figure from http://satijalab.org/seurat/

scRNA-seq analysis: cluster biomarkers

- Visualization of cell type specific cluster biomarkers

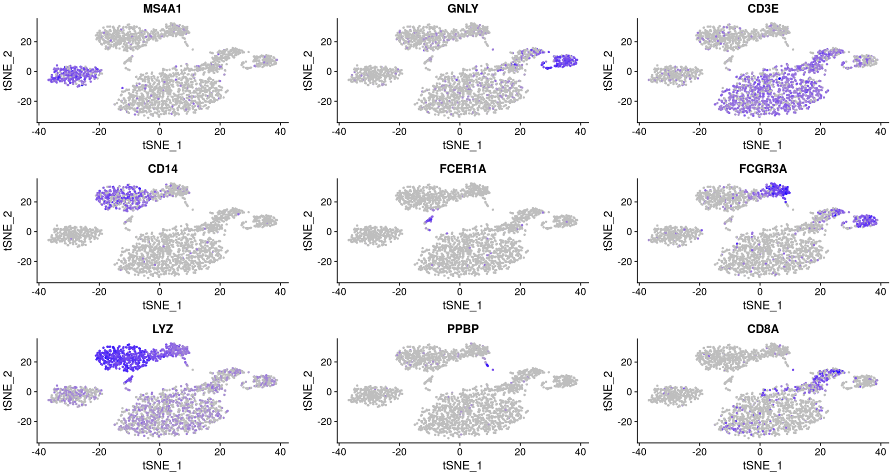


Figure from http://satijalab.org/seurat/

scRNA-seq analysis: clustering

- Assign cell types based on the biomarkers

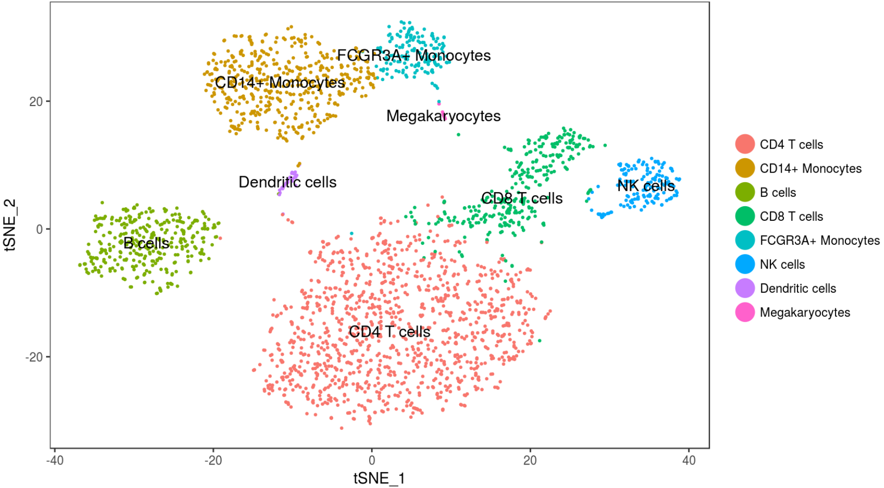


Figure from http://satijalab.org/seurat/

Contents

- Background & Motivation
- Single cell sequencing technologies
- Single cell sequencing data analysis: overview
- Single cell sequencing data analysis: emerging methods

Deep generative models for single cell data

- scRNA-seq profiles contain both biological (mostly unknown) and technical (still poorly characterized) uncertainties
  - Challenging to specify a well-motivated probabilistic data generating model
- Recent developments in deep learning field have shown great promise in modeling complex data
- Autoencoder is a (deep) neural network that tries to predict its input x into output r via an internal representation h
- Internal representation typically has a lower dimension and provides a useful characterization for scRNA-seq data

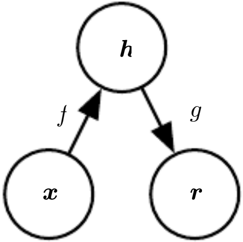


Figure from (Goodfellow et al, 2016)

Deep generative models for single cell data

- A variational autoencoder (VAE) resembles a classical autoencoder, and consists of
  - Encoder
  - Decoder
  - A probabilistic loss function
- Provides a way to design generative (probabilistic) model and inference for complex and large data sets
- Models are end-to-end differentiable: if implemented using a probabilistic programming language, then they can be inferred using automatic differentiation methods (e.g. TensorFlow, Pytorch)

A variational autoencoder for scRNA-seq data

- A probabilistic model for scRNA-seq data

$$
\begin{aligned}
z_n &\sim \mathrm{Normal}(0, I) \\
\ell_n &\sim \mathrm{LogNormal}(\ell_\mu, \ell_\sigma^2) \\
\rho_n &= f_w(z_n, s_n) \\
w_{ng} &\sim \mathrm{Gamma}(\rho_n^g, \theta) \\
y_{ng} &\sim \mathrm{Poisson}(\ell_n w_{ng}) \\
h_{ng} &\sim \mathrm{Bernoulli}(f_h^g(z_n, s_n)) \\
x_{ng} &= \begin{cases} y_{ng} & \text{if } h_{ng} = 0, \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
$$

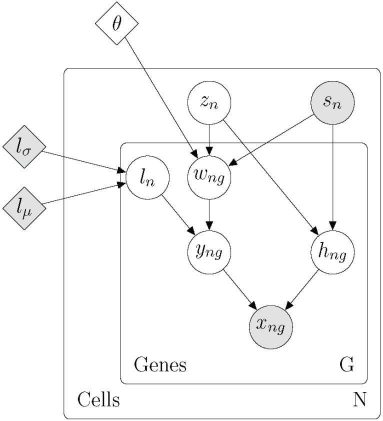


Figure from (Lopez et al, 2019)

A variational autoencoder for scRNA-seq data

- Autoencoder architecture with deep neural networks

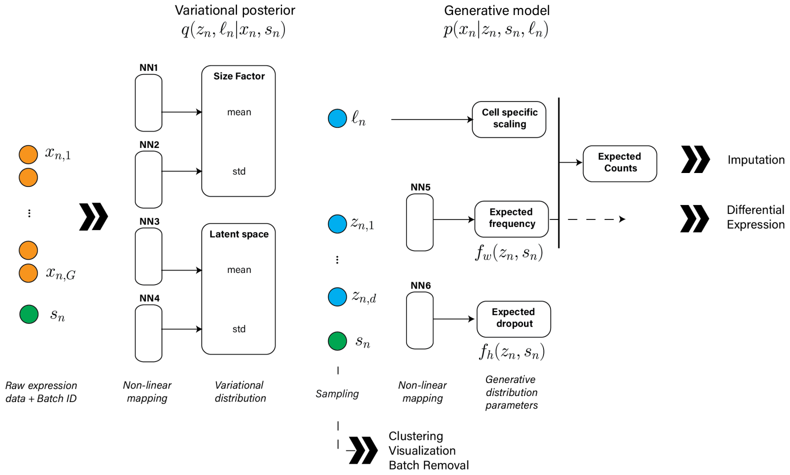


Figure from (Lopez et al, 2019)

A variational autoencoder for scRNA-seq data

- An illustration

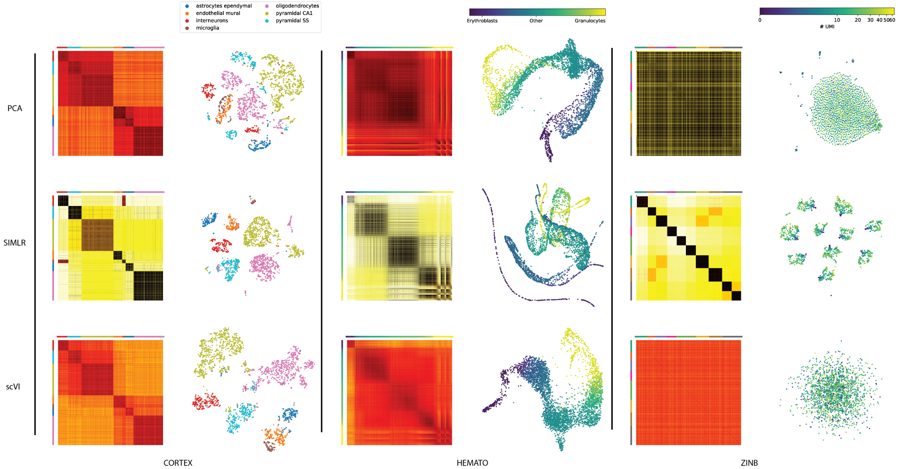


Figure from (Lopez et al, 2019)

References

- Goodfellow I, Bengio Y, Courville A, *Deep Learning*, MIT Press, 2016, `http://www.deeplearningbook.org`
- R. Lopez, J. Regier, MB. Cole, M. Jordan, N. Yosef, Deep Generative Modeling for Single-cell Transcriptomics, *Nature Methods*, 2019
- Macosko EZ et al, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets, Cell 161:1202–1214, 2015
- Smith T, et al., UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy, Genome Research, 27:491–499, 2017.
- Vallejos CA, et al., Normalizing single-cell rna sequencing data: challenges and opportunities, Nature Methods, 14(6):565–571, 2017.
- Seurat tool: http://satijalab.org/seurat/