Introduction to the statistical analysis of single cell transcritomic data

Franck Picard

Laboratoire Biométrie et Biologie Evolutive, CNRS Univ. Lyon

Netbio workshop - December 2018

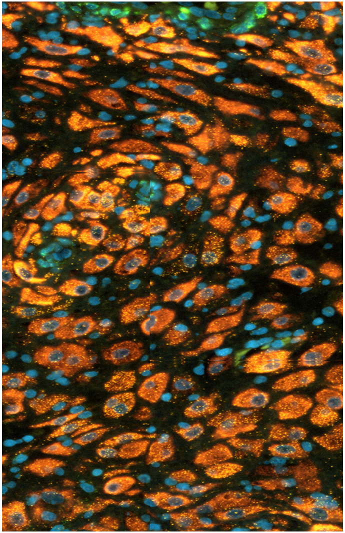
franck.picard@univ-lyon1.fr

Outline

1. Brief Presentation of single cell sequencing
2. Differential Expression Analysis for sequencing data
3. Differential Expression Analysis for single cell data
4. Linear Dimension reduction and data visualization
5. Alternatives to PCA, non linear embedding methods
6. Conclusions

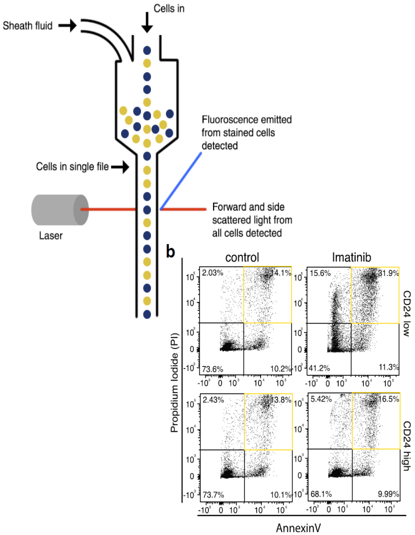
Cell biology revolution

- The cell has been discovered in the 17th century
- Cells are the basic unit of structure and function in living organisms
- Physiology emerges as the meta-cellular science (interaction between cells)

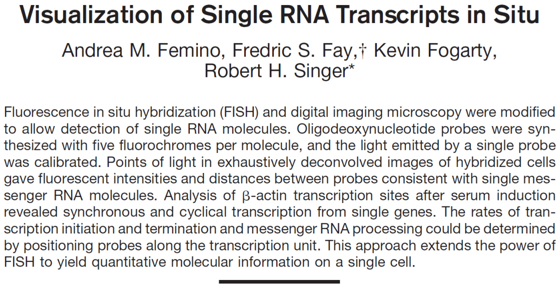


Cell sorting and the investigation of between-cell variations

- Development of monoclonal antibodies ($\sim$ 70s)
- Cell sorting by fluoresence (FACS)
- Cell specific proteins, DNA content
- Limited to a few markers



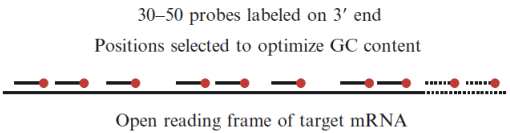
Cell biology is going molecular



Visualization of Single RNA Transcripts in Situ

Andrea M. Femino, Fredric S. Fay,† Kevin Fogarty,
Robert H. Singer*

Fluorescence in situ hybridization (FISH) and digital imaging microscopy were modified to allow detection of single RNA molecules. Oligodeoxynucleotide probes were synthesized with five fluorochromes per molecule, and the light emitted by a single probe was calibrated. Points of light in exhaustively deconvolved images of hybridized cells gave fluorescent intensities and distances between probes consistent with single messenger RNA molecules. Analysis of β-actin transcription sites after serum induction revealed synchronous and cyclical transcription from single genes. The rates of transcription initiation and termination and messenger RNA processing could be determined by positioning probes along the transcription unit. This approach extends the power of FISH to yield quantitative molecular information on a single cell.



30–50 probes labeled on 3′ end

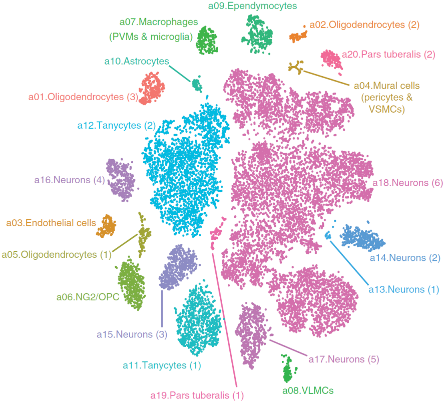
Positions selected to optimize GC content

Open reading frame of target mRNA

SCIENCE • VOL. 280 • 24 APRIL 1998 • www.sciencemag.org

Cell biology goes genome-wide

- Classify cells into distinct cell types
- Shape, location, interactions, function
- Recent technological breakthroughs allow the molecular characterization of cells

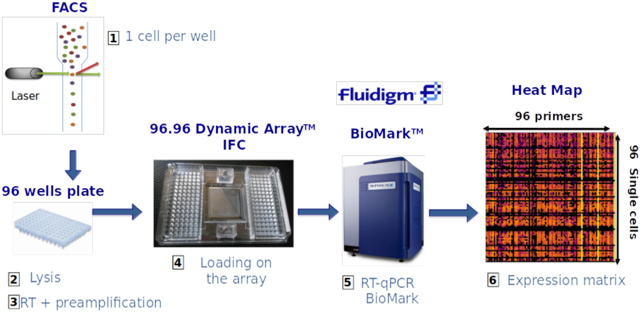


[2]

The single-cell rule

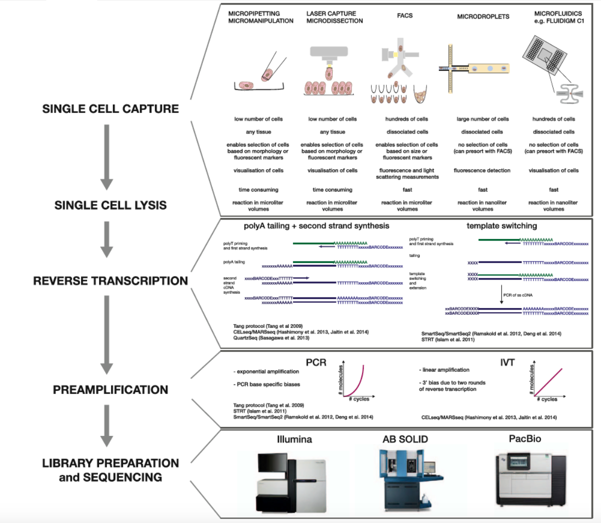
if IT exists, there is a single-cell version of IT (sooner or later)

The RT-qPCR version

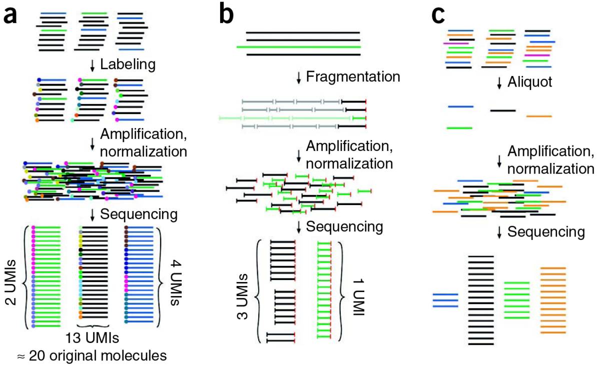


- Precise for a small number of genes
- not too expensive (100 cells)
- pros and cons of RT-qPCR are well known

The Sequencing version [6]

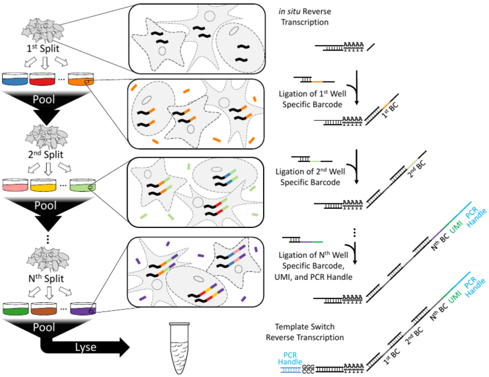

SINGLE CELL CAPTURE
SINGLE CELL LYSIS
REVERSE TRANSCRIPTION
PREAMPLIFICATION
LIBRARY PREPARATION and SEQUENCING
MICROPIPETTING MICROMANIPULATION
LASER CAPTURE MICRODISSECTION
FACS
MICRODROPLETS
MICROFLUIDICS e.g. FLUIDIGM C1
low number of cells
any tissue
enables selection of cells based on morphology or fluorescent markers
visualisation of cells
time consuming
reaction in microliter volumes
low number of cells
any tissue
enables selection of cells based on morphology or fluorescent markers
visualisation of cells
time consuming
reaction in microliter volumes
hundreds of cells
dissociated cells
enables selection of cells based on size or fluorescent markers
fluorescence and light scattering measurements
fast
reaction in microliter volumes
large number of cells
dissociated cells
no selection of cells (can presort with FACS)
fluorescence detection
fast
reaction in nanoliter volumes
hundreds of cells
dissociated cells
no selection of cells (can presort with FACS)
visualisation of cells
fast
reaction in nanoliter volumes
polyA tailing + second strand synthesis
polyT priming and first strand synthesis
AAAAAAAAAAAA
TTTTTTTTTxxxxxBARCODExxxxxxx
polyA tailing
xxxxxxxAAAAAA
second strand cDNA synthesis
xxxxBARCODExxxTTTTTT
AAAAAAAxxxxxBARCODExxxxxxx
Tang protocol (Tang et al 2009)
CELseq/MARSseq (Hashimony et al. 2013, Jaitin et al. 2014)
QuartzSeq (Sasagawa et al. 2013)
template switching
polyT priming and first strand synthesis
tailing
XXXX
template switching and extension
PCR of ss cDNA
xxBARCODEXXXX
SmartSeq/SmartSeq2 (Ramskold et al. 2012, Deng et al. 2014)
STRT (Islam et al. 2011)
PCR
- exponential amplification
- PCR base specific biases
# molecules
# cycles
Tang protocol (Tang et al. 2009)
STRT (Islam et al. 2011)
SmartSeq/SmartSeq2 (Ramskold et al. 2012, Deng et al. 2014)
IVT
- linear amplification
- 3' bias due to two rounds of reverse transcription
# molecules
# cycles
CELseq/MARSseq (Hashimony et al. 2013, Jaitin et al. 2014)
Illumina
AB SOLID
PacBio

The UMI version [6]



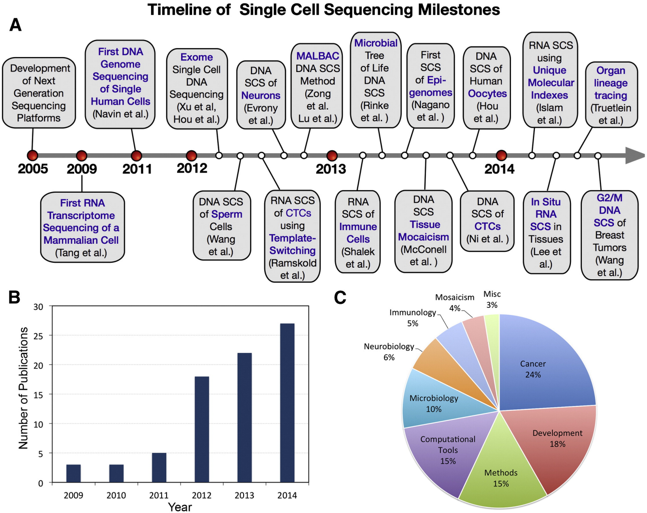
- up to 20,000 genes analyzed
- static snapshot
- expensive (for hundreds of cells)

The split version [12]


1st Split
Pool
2nd Split
Pool
Nth Split
Pool
Lyse
in situ Reverse Transcription
AAAAA
TTTTT
Ligation of 1st Well Specific Barcode
1st BC
Ligation of 2nd Well Specific Barcode
2nd BC
Ligation of Nth Well Specific Barcode, UMI, and PCR Handle
Nth BC
UMI
PCR Handle
Template Switch Reverse Transcription
PCR Handle
GGG
CCC

- up to 20,000 genes analyzed
- for millions of cells
- cheap

A timeline: technologies [15]

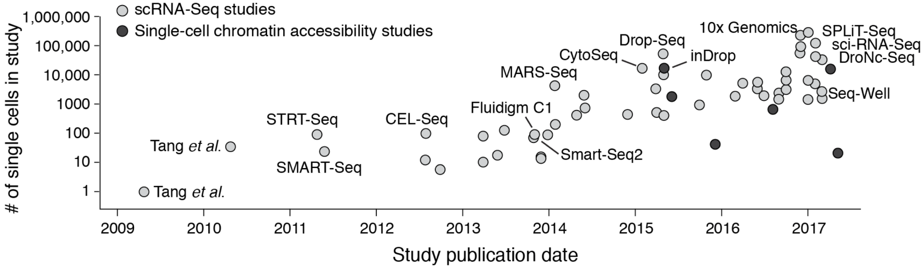


Timeline of Single Cell Sequencing Milestones

A

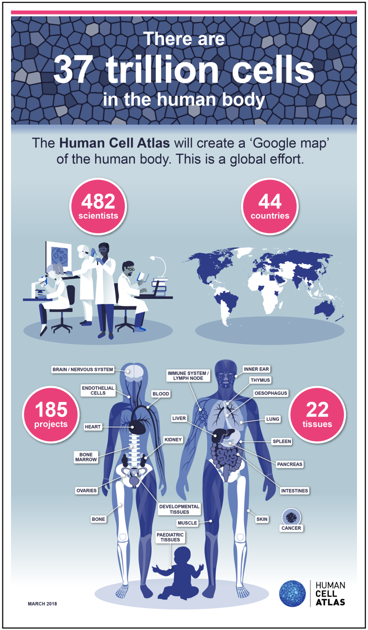
- Development of Next Generation Sequencing Platforms
- First DNA Genome Sequencing of Single Human Cells (Navin et al.)
- Exome Single Cell DNA Sequencing (Xu et al, Hou et al.)
- DNA SCS of Neurons (Evrony et al.)
- MALBAC DNA SCS Method (Zong et al. Lu et al.)
- Microbial Tree of Life DNA SCS (Rinke et al.)
- First SCS of Epi-genomes (Nagano et al.)
- DNA SCS of Human Oocytes (Hou et al.)
- RNA SCS using Unique Molecular Indexes (Islam et al.)
- Organ lineage tracing (Truetlein et al.)
- First RNA Transcriptome Sequencing of a Mammalian Cell (Tang et al.)
- DNA SCS of Sperm Cells (Wang et al.)
- RNA SCS of CTCs using Template-Switching (Ramskold et al.)
- RNA SCS of Immune Cells (Shalek et al.)
- DNA SCS Tissue Mocaicism (McConell et al.)
- DNA SCS of CTCs (Ni et al.)
- In Situ RNA SCS in Tissues (Lee et al.)
- G2/M DNA SCS of Breast Tumors (Wang et al.)

The Moore's law of single cell [10]

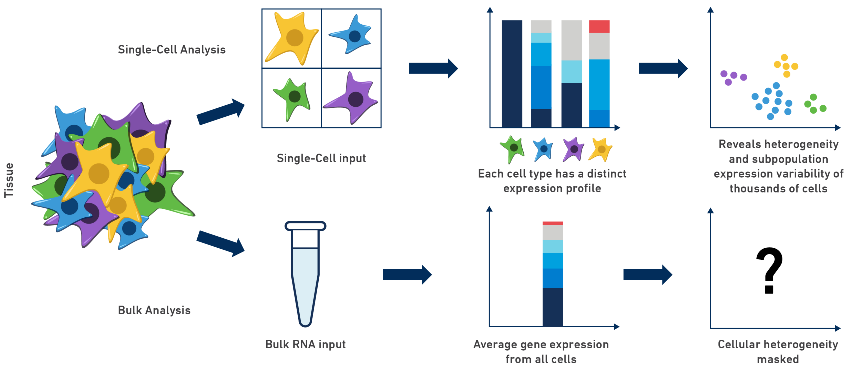


The human cell Atlas project

- comprehensive reference catalog of all human cells
- use stable properties, transient features, locations and abundances.
- describe each human cell by a defined set of molecular markers
- based on DNA variations, RNA, Epigenome at the single-cell resolution



Single-Cell from a statistician's perspective


Single-Cell Analysis
Tissue
Single-Cell input
Each cell type has a distinct expression profile
Reveals heterogeneity and subpopulation expression variability of thousands of cells
Bulk Analysis
Bulk RNA input
Average gene expression from all cells
?
Cellular heterogeneity masked

From 10X Genomics

Outline

1. Brief Presentation of single cell sequencing
2. Differential Expression Analysis for sequencing data
3. Differential Expression Analysis for single cell data
4. Linear Dimension reduction and data visualization
5. Alternatives to PCA, non linear embedding methods
6. Conclusions

Let's adopt the ANOVA framework

- Y_{ijr} : expression (continuous) for gene i in condition j at replicate r
- Perform DE between conditions using model

$$Y_{ijr} \sim \mathcal{N}\left(\mathbb{E}(Y_{ijr}), \sigma^2\right)$$

$$\mathbb{E}(Y_{ikr}) = \mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

- The parameters of the model are interpreted as :
  - α_i : mean expression of gene i (across conditions),
  - β_j : mean expression in condition j (across genes),
  - $(\alpha\beta)_{ij}$: interaction effect gene x condition
- Allows to integrate normalization while testing

Testing framework

- **Hypothesis** : no expression difference between conditions

$$\mathcal{H}_0^i : \{(\alpha\beta)_{i1} = (\alpha\beta)_{i2}\}$$

- The classical statistic for gene i is the Student statistic

$$T_i = \frac{|\widehat{\alpha\beta}_{i1} - \widehat{\alpha\beta}_{i2}|}{\widehat{\sigma}} \times \sqrt{2R-2} \underset{\mathcal{H}_0}{\sim} \mathcal{T}(2R-2)$$

- Estimation of mean fixed effects is done by **Maximum Likelihood**
- Multiple testing issues are assessed using the FDR control

What about the estimation of the dispersion parameter ?

- Refinements / difficulties concern the **estimation of** σ, the dispersion parameter
- A **common variance** to all genes σ^2 : robust but lacks of power
- A **specific variance** to every gene σ_i^2 : powerful but sensitive to outliers,
  - Large sampling variance
  - To be stabilized empirically
- **Groups of variances** (combination of both)

The Generalized Linear Model framework

- Y_{ijr} : the **read count** (positive integer), for gene i in condition j
- Define the **Generalized Linear Model** (GLM) by setting

$$\begin{aligned} Y_{ijr} &\sim \mathcal{P}(\mu_{ij}) \\ \log \mathbb{E}(Y_{ijr}) = \log(\mu_{ij}) &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \end{aligned}$$

- Parameters have the same interpretation
- Testing hypotheses are similar : $\mathcal{H}_0^i : \{(\alpha\beta)_{i1} = (\alpha\beta)_{i2}\}$
- Dispersion parameter ? Test statistics ?

The Exponential family of distributions

- Family of distributions that share common mathematical properties
- The **Exponential Family** is one of the most widely used
- Consider two types of parameters :
    - θ the **natural parameter**, related to the location parameter
    - ϕ the **dispersion parameter**
- If Y belongs to the exponential family, its density is of the form

$$p(y;\theta,\phi) \propto \exp\left(\frac{y\theta - b(\theta)}{a(\phi)}\right)$$

	θ	ϕ
Gaussian	μ	σ^2
Poisson	$\log(\mu)$	1
Binomial	$\text{logit}(\mu)$	1

The Generalized Linear Model (GLM)

- Suppose that $\mathbb{E}(Y) = \mu$ linearly depends on some covariates X

$$g(\mathbb{E}(Y)) = g(\mu) = X\beta$$

- $\eta = g(\mu)$ is often called the **linear predictor**
- Most of the time the **canonical link** is used $g(\mu) = g(b'(\theta)) = \theta$

	$g(\mu)$	$V(\mu)$
Gaussian	μ	1
Poisson	$\log(\mu)$	μ
Binomial	$\text{logit}(\mu)$	$\mu(1-\mu)$

- In GLMs, overdispersion ϕ is not used (exponential dispersion family)

Location / Dispersion relations

- The first two moments of the exponential family are :

$$\begin{aligned} \mathbb{E}[Y] &= b'(\theta) = \mu \\ \mathbb{V}[Y] &= b''(\theta) \times a(\phi) \end{aligned}$$

- The **expectation μ is a function of θ only** (location)
- The variance is a function of **both** (θ, ϕ) (location and dispersion)
- $b''(\bullet)$ is called the **variance function** (also denoted by $V(\mu)$), and describes how the variance relates to the mean
- Gaussian : (exception!) $a(\phi) = \sigma^2$ and $b''(\theta) = 1$ (cst curvature)
- Poisson, Binomial : $a(\phi) = 1$ (no freedom)

A matter of vocabulary

- Recall that $\mathbb{V}[Y] = a(\phi) \times V(\mu)$ where ϕ is the dispersion parameter
- The **Poisson distribution has no dispersion parameter**
- The only possible Discrete Exponential Dispersion model with a disperson parameter are additive models such as **Negative Binomial** or **Poisson-Tweedie**
- Parameter α may be called dispersion parameter

	$a(\phi)$	$V(\mu)$
Poisson	1	μ
QuasiPoisson	ϕ	μ
Negative Binomial	1	$\mu + \kappa\mu^2$
Tweedie Poisson	1	μ^p

Dealing with dispersion estimates

- If we choose the NB model, $V(\mu) = \mu + \kappa\mu^2$
- Then we can follow the same procedure compared with the Gaussian case:
- Use genes as replicates to uncover the mean/variance relationship (one gene=one point)
- perform a regression $V(\mu) = \mu + \kappa_{Global}\mu^2$ to estimate κ_{Global} that would be **common to every gene**
- In Anders et al., the final "dispersion" estimate is for each gene : $\max(\widehat{\kappa}_{Global}, \widehat{\kappa}_i)$ (little loss in power)

Testing Strategies based on LRT

- Compare different models, for instance

$$\begin{aligned}\log(\mu_{ij}) &= \mu + \alpha_i + \beta_j \\ \log(\mu_{ij}) &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}\end{aligned}$$

- Use the Ratio of log likelihoods as a Statistics, which incorporates all infos:

$$LRT = -2\log\left(\frac{\mathcal{L}(\widehat{\mu}, \widehat{\alpha}, \widehat{\beta}, \widehat{\alpha\beta})}{\mathcal{L}(\widehat{\mu}, \widehat{\alpha}, \widehat{\beta})}\right) \underset{\mathcal{H}_0}{\sim} \chi^2(\Delta df)$$

- This has been shown to be the best strategy on Sequencing data

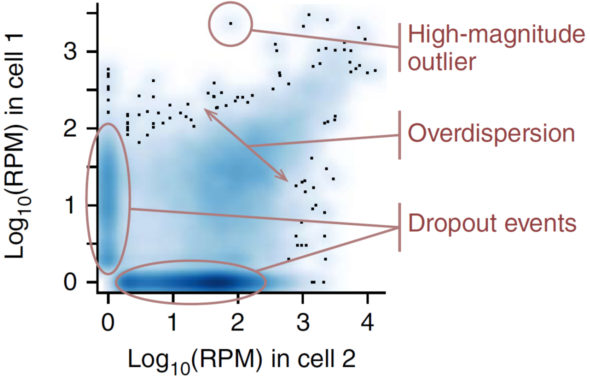
Conclusion: don't think Normal !

- Use **Generalized Linear Models** to perform Count regression, and not Gaussian regression on the log-counts
- Incorporate effects in the model to perform a global analysis that **accounts for distributional characteristics**
- Do not perform tests that imply Poisson distribution when data are over-dispersed
- Use **Likelihood Ratio Tests** to compare models
- Overdispersion leads to estimation issues due to **numerical problems**

Outline

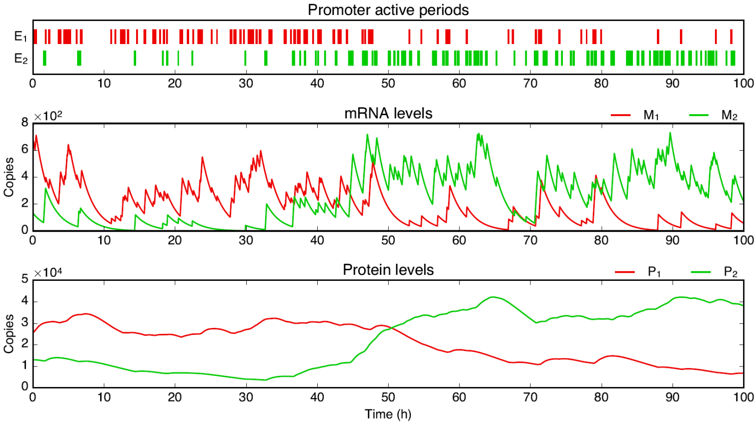
1. Brief Presentation of single cell sequencing
2. Differential Expression Analysis for sequencing data
3. Differential Expression Analysis for single cell data
4. Linear Dimension reduction and data visualization
5. Alternatives to PCA, non linear embedding methods
6. Conclusions

How bad is the situation in single cell data ?



Overdispersion is mainly biological because diversity is high between cells [5]

Expression is a stochastic bursty process: biological zeros



The curse of Dropouts

- Low starting amount of RNAs: transcripts will be missed during RT
- Amplification is needed ($\times 10^6$), which creates distortions
- Stochasticity of gene expression (bursty process) sparsity of the data, high proportion of zeros
- Dropout depends on cells (different in different wells),
- Lowly expressed genes : sampling / amplification issues
- Highly expressed genes: is more likely to indicate a burst

Consequences for the analysis of multiple single cells

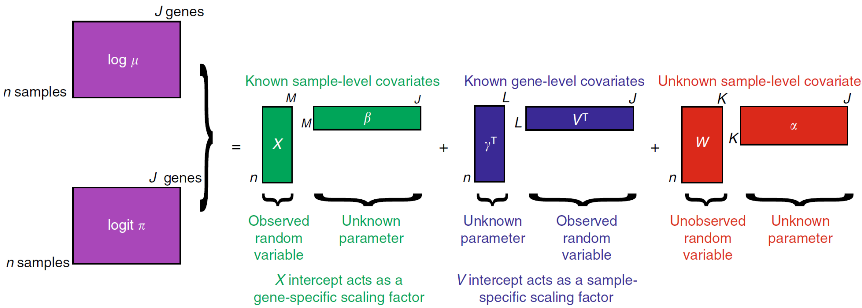
- The data consist of a **snapshot** : all cells are not synchronized
- No technical replicate per cell (invasive experiment)
- A lot of zeros in the data : zero inflated count distributions
- For cell r, gene i, condition j, the expression value is modelled by

$$Y_{ijr} \sim \pi_i \delta_0 + (1 - \pi_i)\mathrm{NB}(\mu_{ijr})$$

- Difficulty to discriminate between low expression / no expression

$$\pi_i = f(\mathbb{E}(Y_{ijr}))$$

Two-component Generalized Linear Model [11]


J genes
log μ
n samples
J genes
logit π
n samples
=
Known sample-level covariates
M
J
M
X
β
n
Observed random variable
Unknown parameter
X intercept acts as a gene-specific scaling factor
+
Known gene-level covariates
L
J
L
γ^T
V^T
n
Unknown parameter
Observed random variable
V intercept acts as a sample-specific scaling factor
+
Unknown sample-level covariates
K
J
K
W
α
n
Unobserved random variable
Unknown parameter

Inference for ZINB models

- Optimization can be based on IRWLS (iterative) combined with the EM algorithm
- EM is used to retrieve the ZI compartment
- Then IRWLS is used to estimate the parameters in the NB compartement
- Quite challenging from the numerical point of view
- Use Bayesian strategies thanks to the Poisson-Gamma representation of NB distributions

$$\mu \sim \Gamma(a, b), \ \ Y|\mu \sim \mathcal{P}(\mu), \ \ Y \sim NB(a, \frac{b}{a+1})$$

- Model the sampling process of genes

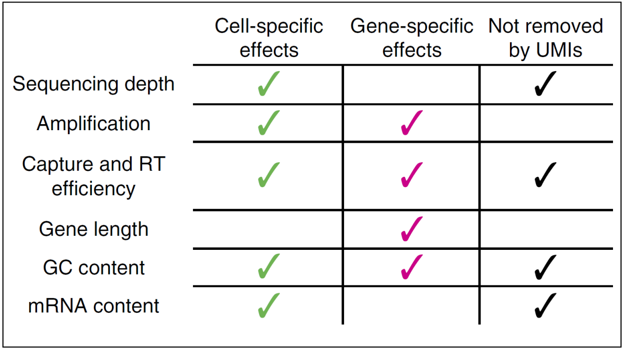
Normalizing single cell expression data

- Adjusts for effects related to distributional differences in read counts between cells (sequencing depth)
- Scaling factor for cell r to rescale all cell specific measures on a common scale

$$\mathbb{E}(Y_{ijr}) = s_r \times \mu_{ij}$$

- How to estimate the scaling factor ? RPKM ?
- Library size normalization can be dominated by a handful of highly expressed genes, which can bias downstream analysis .
- Quantile matching ? but difficult to apply with zero Inflation
- Litterature suggests to use the trimmed means proposed by DESeq [13]

Normalizing single cell expression data [13]



	Cell-specific effects	Gene-specific effects	Not removed by UMIs
Sequencing depth	✓		✓
Amplification	✓	✓	
Capture and RT efficiency	✓	✓	✓
Gene length		✓	
GC content	✓	✓	✓
mRNA content	✓		✓

Hypothesis testing becomes more intricate

- A Compartment model defined by (π, μ)
- Testing hypotheses are similar : $\mathcal{H}_0^i : \{(\pi_j, \mu_j) = \big(\pi_{j'}, \mu_{j'}\big)\}$
- What is differential expression ? Differential dropout ?
- Difficulty to define $\mathcal{H}_1$

Outline

1. Brief Presentation of single cell sequencing
2. Differential Expression Analysis for sequencing data
3. Differential Expression Analysis for single cell data
4. Linear Dimension reduction and data visualization
5. Alternatives to PCA, non linear embedding methods
6. Conclusions

An unprecedented challenge

- Genomics was precursor for data representation and visualization

Publication	cells	tissue	Seq. protocol	clusters
Cadwell et al. (2016)	46	visual cortex	Smart-seq2	2
Tasic et al. (2016)	1,679	visual cortex	SMARTer	49
Macosko et al. (2015)	44,808	retina	Drop-seq	39
10x Genomics	1,306,127	brain cells	10x Gen.Chrom.	39

- We are far beyond the few clusters / some points
- Dimension reduction is mandatory for any analysis (clustering, visualization, GRN inference, etc)

High-dimensional count data

$$x_{ij} = \text{expression of gene } j \text{ in cell } i$$

$$\mathbf{X}_{n\times p} = \begin{array}{c} \left[\begin{array}{c|c|c|c|c|c|c} \quad & \quad & \quad & \quad & \quad & \quad & \quad \\ \hline & & & x_{ij} & & & \\ \hline & & & & & & \end{array}\right] \\ \underbrace{1 \quad \dots \quad \dots \quad \dots \quad \dots \quad p}_{\text{genes}} \end{array} \left.\begin{array}{c} 1 \\ \vdots \\ n \end{array}\right\} \text{cells}$$

- **High dimension:** n grows but $\ll p$
- **Count data** with dropouts

A quickie on PCA 1

- PCA represents multivariate data $\mathbf{X}_{n\times p}$ using projection in a lower dimensional space of dimension $K < p$.
- PCA is non supervised in the sense that the projection is done without guidance
- The strategy is global : there may exist correlations between variables that could be used to summarized the data, and to visualize $\mathbf{X}$ despite its dimensionality.
- PCA provides orthogonal principal components that best explain the variability of the data globally.
- The key ingredient of PCA is then the empirical covariance matrix

$$\mathbf{S}_{p\times p} = \frac{1}{n-1}\mathbf{X}_c^T\mathbf{X}_c$$

A quickie on PCA 2

- PCs $\mathbf{t}_1, \dots, \mathbf{t}_K$, are defined s.t. $t_{ik} = \sum_{j=1}^{p} w_{jk} X_{ij}$
- w_{jk} quantifies the weight of variable $X_{\cdot,j}$ in the constitution of PC k.
- $\mathbf{w}_1, \dots, \mathbf{w}_K$ are determined iteratively by finding the PCs that carry most inertia

$$\mathbb{V}(\mathbf{t}_k) = \mathbf{w}_k^T \mathbf{X}^T \mathbf{X} \mathbf{w}_k.$$

- Solve iteratively the following optimization problem:

$$\mathbf{w}_k = \arg \max_{\mathbf{w} \in \mathbb{R}^p, \|\mathbf{w}\|=1} \left\{ \mathbf{w}^T \mathbf{X}_c^T \mathbf{X}_c \mathbf{w} \right\}, \text{ with } \mathbf{t}_k = \mathbf{X} \mathbf{w}_k \perp \mathbf{t}_1, \dots \mathbf{t}_{k-1}.$$

- $\mathbf{w}_1, \dots \mathbf{w}_r$ are the associated eigen vectors of $\mathbf{w}^T \mathbf{X}_c^T \mathbf{X}_c \mathbf{w}$ associated with eigen values $\lambda_1, \dots \lambda_r$.

A quickie on PCA 3

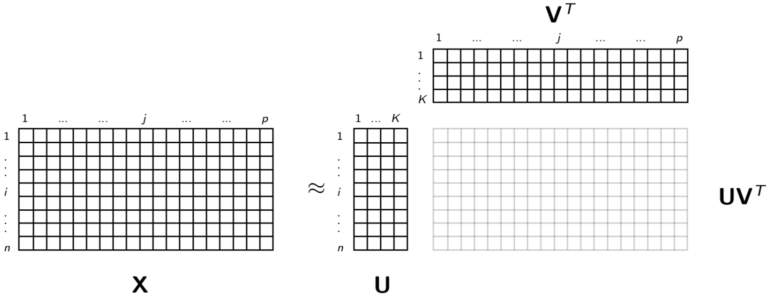
- A singular value decomposition is the decomposition of a matrix $\mathbf{X}_c$ such that $\mathbf{X}_c = \mathbf{UDV}^T$,
- $\mathbf{D}_{r \times r} = \text{diag}(\delta_1, ...\delta_r)$ is the diagonal matrix of singular values of $\mathbf{X}_c$.
- $\mathbf{U}$ is orthonormal, whose columns are eigen vectors of $(\mathbf{X}_c \mathbf{X}_c^T)$
- $\mathbf{V}$ is orthonormal whose columns are eigen vectors of $(\mathbf{X}_c^T \mathbf{X}_c)$
- PCA can be rephrased as a minimization problem:

$$\min_{\mathbf{u} \in \mathbb{R}^n, \mathbf{v} \in \mathbb{R}^p, \|\mathbf{u}\| = \|\mathbf{v}\| = 1} \|\mathbf{X}_c - \mathbf{UV}^T\|_F^2$$

- with $\|\mathbf{A}\|_F^2 = \sum_{ij} a_{ij}^2$,

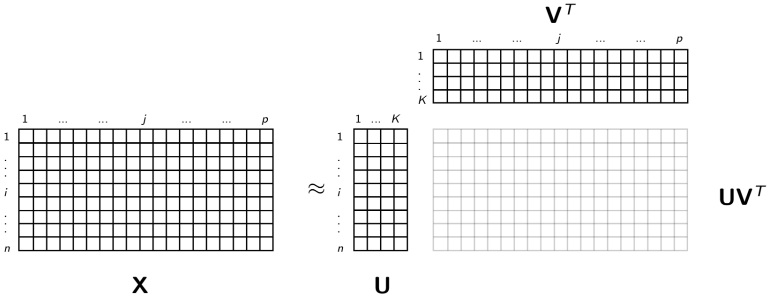
Matrix factorization: $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$

$$\left.\begin{array}{ll} \text{Cells:} & \mathbf{U} \in \mathbb{R}^{n \times {\color{red}K}} \\ \text{Genes:} & \mathbf{V} \in \mathbb{R}^{p \times {\color{red}K}} \end{array}\right\} \text{Low dimensional representation}$$

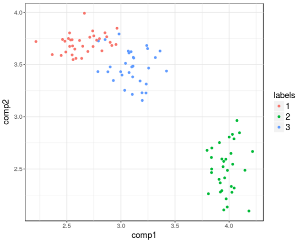


→ Low-rank representation of $\mathbf{X}$

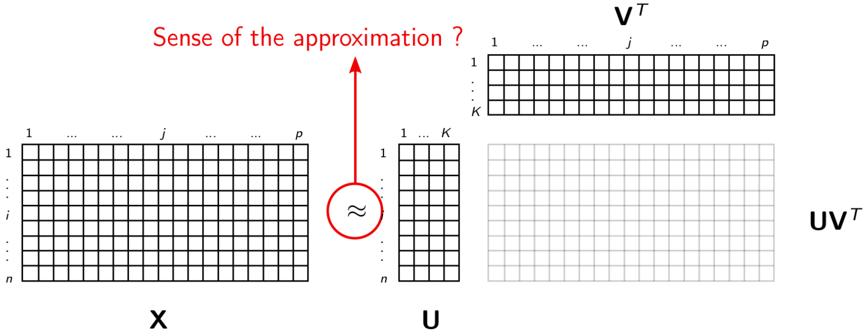
Matrix factorization: $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$



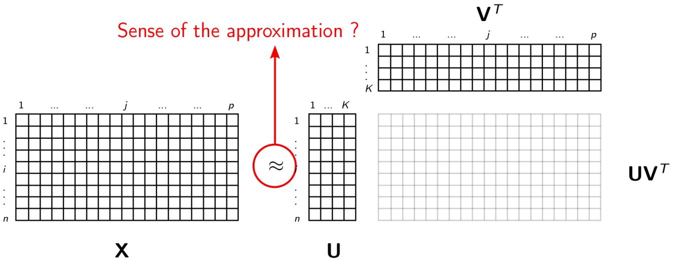
Data visualization:
scatter plot $(u_{i1}, u_{i2})_{i=1:n}$



Approximation $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$?



Approximation $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$?



Principal Component Analysis:

- Find a linear projection of $\mathbf{X}$ with maximum variance
- SVD algorithm: $\underset{\mathbf{U}\in\mathbb{R}^{n\times K},\mathbf{V}\in\mathbb{R}^{p\times K}}{\operatorname{argmin}} \left\|\mathbf{X}-\mathbf{U}\mathbf{V}^T\right\|_F^2$
- **Least squares approximation**

RNA-seq data = Counts

Relation between geometry and underlying model

$\|\cdot\|_2 \leftrightarrow$ Gaussian distribution

- First idea: $X_{ij} \sim \mathcal{P}(\lambda)$
- Highly expressed genes

  $\hookrightarrow$ large λ

  $\hookrightarrow$ Gaussian approximation

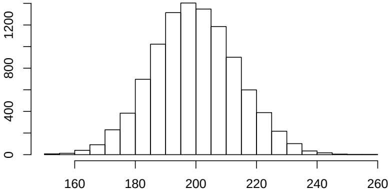


Figure: $\mathcal{P}(200)$ empirical distribution

RNA-seq data = Counts

Relation between geometry and underlying model

$\|\cdot\|_2 \leftrightarrow$ Gaussian distribution

- First idea: $X_{ij} \sim \mathcal{P}(\lambda)$
- Highly expressed genes

  $\hookrightarrow$ large λ

  $\hookrightarrow$ Gaussian approximation

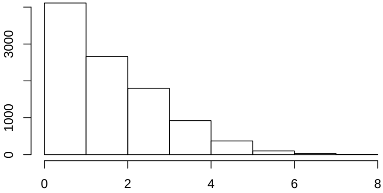


Figure: $\mathcal{P}(2)$ empirical distribution

Need for a probabilistic PCA

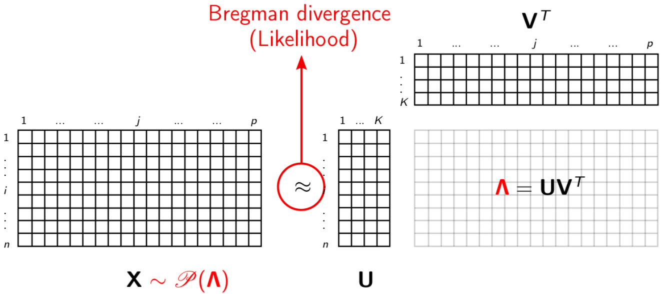
- **Over-dispersion** in RNA-seq data → $\text{Var}(X_{ij}) > \mathbb{E}[X_{ij}]$
- Single-cell data: **zero-inflation** → $\mathbb{P}(X_{ij} = 0) > e^{-\lambda}$

Embed PCA with a **probabilistic model**

- $X_{ij} \sim$ probability distribution in the exponential family
- Factorization of $\mathbb{E}[\mathbf{X}]$ rather than $\mathbf{X}$
- Replace $\| \cdot \|_2$ approximation by likelihood-based approaches

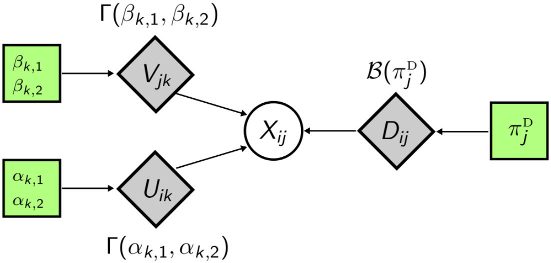
Generalized PCA[3] and Poisson NMF [8]

- $X_{ij} \sim \mathcal{P}(\lambda_{ij})$ with the Poisson rate matrix $\mathbf{\Lambda} = [\lambda_{ij}]_{n \times p}$
- Decompose $\mathbb{E}[\mathbf{X}] = \mathbf{\Lambda}$ such that $\lambda_{ij} = \sum_k U_{ik} V_{kj}$



Probabilistic PCA with a ZI-Gamma-Poisson model [1]

- Factors **U**,**V** become Gamma **latent variables**
- Marginal distribution is over-dispersed: $\mathrm{Var}(X_{ij}) > \mathbb{E}[X_{ij}]$
- D_{ij} = drop-out event indicator

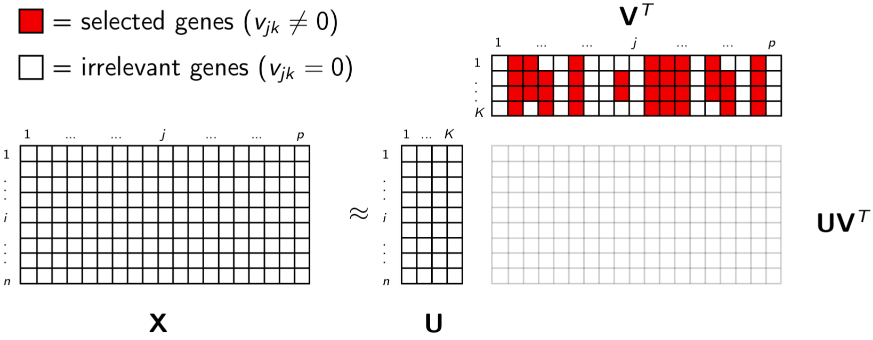


Â

Probabilistic variable selection with a spike and slab model

- Sparsity-inducing priors:

$$V_{jk} \sim \pi_S \delta_0 + (1 - \pi_S)\Gamma(\beta_{k,1}, \beta_{k,2})$$



🟥 = selected genes $(v_{jk} \neq 0)$

□ = irrelevant genes $(v_{jk} = 0)$

Model inference

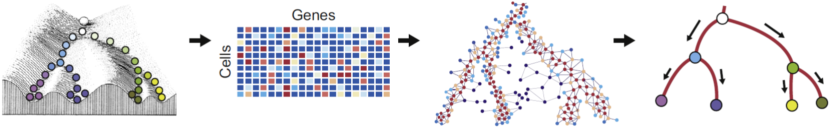
- Recover the posterior distributions $\mathbf{U} \,|\, \mathbf{X}$ and $\mathbf{V} \,|\, \mathbf{X}$
- Estimate the factors as $\widehat{\mathbf{U}} = \mathbb{E}[\mathbf{U} \,|\, \mathbf{X}]$ and $\widehat{\mathbf{V}} = \mathbb{E}[\mathbf{V} \,|\, \mathbf{X}]$
- Posteriors are not explicit
- **Variational inference:** approximation of the posteriors

Outline

1. Brief Presentation of single cell sequencing
2. Differential Expression Analysis for sequencing data
3. Differential Expression Analysis for single cell data
4. Linear Dimension reduction and data visualization
5. Alternatives to PCA, non linear embedding methods
6. Conclusions

Beyond Linear projections

- Linear methods are powerful for planar structures
- High dimensional datasets are characterized by multiscale properties (local / global structures)
- May not be the most powerful for manifolds
- Non Linear projection methods aim at preserving local characteristics of distances



Stochastic Neighbor Embedding 1 [14]

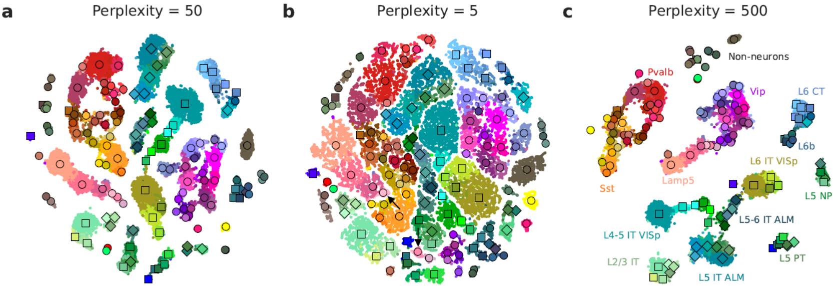
- $(x_1, \dots, x_n)$ are the points in the high dimensional space $\mathbb{R}^p$,
- Consider a similarity between points:

$$p_{i|j} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_k - x_j\|^2/2\sigma_k^2)}, \; p_{ij} = (p_{i|j} + p_{j|i})/2N$$

- σ smooths the data (linked to the regularity of the target manifold)
- σ is chosen such that the entropy of p is fixed to a given value of the so-called perplexity

$$\exp\left(-\sum_{ij} p_{ij} \log(p_{ij})\right)$$

Visual inspection of the influence of σ[7]



Stochastic Neighbor Embedding 2

- Consider $(y_1, \dots, y_n)$ are points in the low dimensional space $\mathbb{R}^2$
- Consider a similarity between points in the new representation:

$$q_{i|j} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_k - y_j\|^2)}$$

- Robustify this kernel by using Student(1) kernels (ie Cauchy)

$$q_{i|j} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i}(1 + \|y_i - y_k\|^2)^{-1}}$$

Stochastic Neighbor Embedding 3

- Minimize the KL between p and q so that the data representation minimizes:

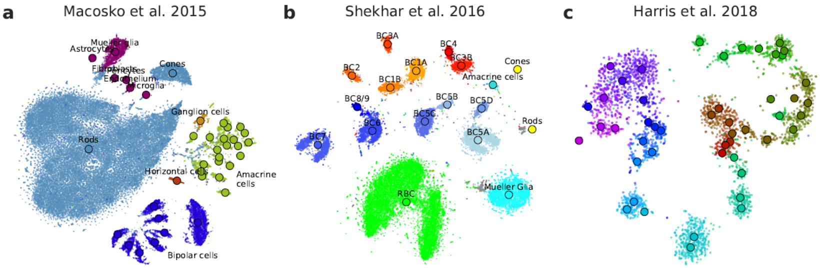
$$C(y) = \sum_{ij} KL(p_{ij}, q_{ij})$$

- The cost function is not convex

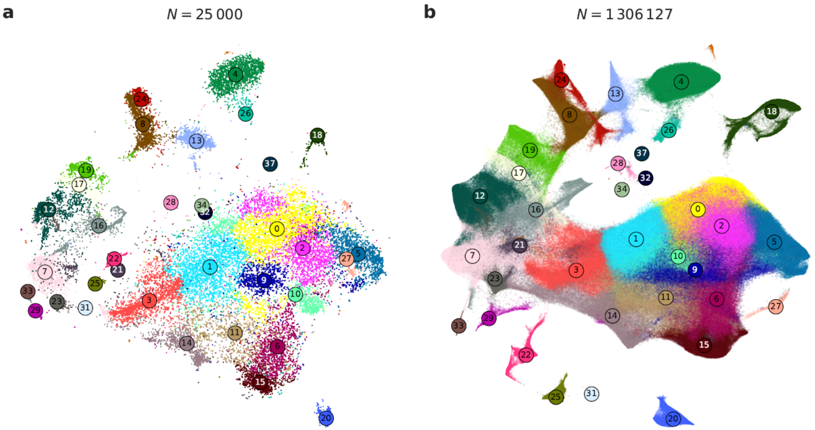
$$\frac{\partial C(y)}{\partial y} = \sum_{ij} (p_{ij} - q_{ij})(y_i - y_j)$$

- Very sensitive to starting values

Stochastic Neighbor Embedding 4 [7]



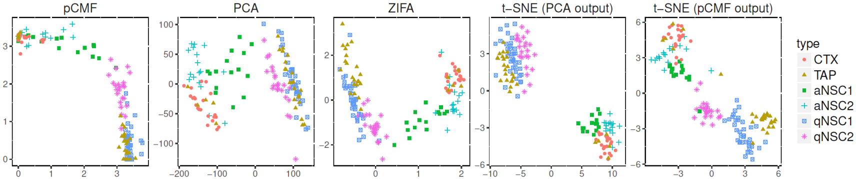
Stochastic Neighbor Embedding 5 [7]



Properties of t-SNE

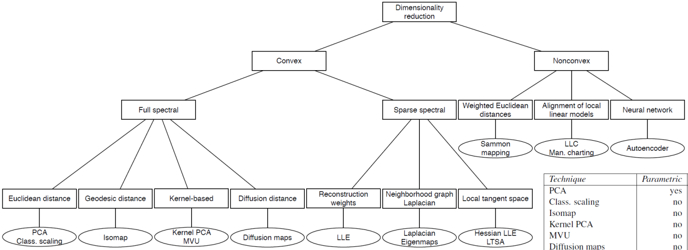
- good at preserving local distances (intra-cluster variance)
- not so good for global representation (inter-cluster variance)
- hence good at creating clusters of points that are close, but bad at positionning clusters wrt each other
- preprocessing very important : initialize with PCA and feature selection plus log transform (non linear transform)
- percent of explained variance ? interpretation of the q distribution ?

Single-cell RNAseq example[9]



Method	**pCMF**	PCA	ZIFA	t-SNE	t-SNE (after pCMF)
Exp. Dev.	**70.3 %**	34.8 %	42.6 %	/	/
Adj. RI	**38.3 %**	24.9 %	38.1 %	37.7 %	**53.6 %**

A taxonomy of Dimension Reduction Methods [4]



Dimensionality reduction

Convex

Nonconvex

Full spectral

Sparse spectral

Weighted Euclidean distances

Alignment of local linear models

Neural network

Sammon mapping

LLC
Man. charting

Autoencoder

Euclidean distance

Geodesic distance

Kernel-based

Diffusion distance

Reconstruction weights

Neighborhood graph Laplacian

Local tangent space

PCA
Class. scaling

Isomap

Kernel PCA
MVU

Diffusion maps

LLE

Laplacian Eigenmaps

Hessian LLE
LTSA

Technique	*Parametric*	*Parameters*	*Computational*	*Memory*
PCA	yes	none	$O(D^3)$	$O(D^2)$
Class. scaling	no	none	$O(n^3)$	$O(n^2)$
Isomap	no	k	$O(n^3)$	$O(n^2)$
Kernel PCA	no	$\kappa(\cdot,\cdot)$	$O(n^3)$	$O(n^2)$
MVU	no	k	$O((nk)^3)$	$O((nk)^3)$
Diffusion maps	no	σ, t	$O(n^3)$	$O(n^2)$
LLE	no	k	$O(pn^2)$	$O(pn^2)$
Laplacian Eigenmaps	no	k, σ	$O(pn^2)$	$O(pn^2)$
Hessian LLE	no	k	$O(pn^2)$	$O(pn^2)$
LTSA	no	k	$O(pn^2)$	$O(pn^2)$
Sammon mapping	no	none	$O(in^2)$	$O(n^2)$
Autoencoders	yes	net size	$O(inw)$	$O(w)$
LLC	yes	m, k	$O(imd^3)$	$O(nmd)$
Manifold charting	yes	m	$O(imd^3)$	$O(nmd)$

Conclusions of a comparative study [4]

- local methods suffer from the choice of the smoothing (neighborhood) parameter
- Kernel PCA suffers from the choice of the Kernel to correctly approximate the manifold.
- Setting the optimization problem is the key (convex or not), trivial solutions, local optima, computationally feasible
- nonlinear techniques for dimensionality reduction are, despite their large variance, often not capable of outperforming traditional linear techniques such as PCA.

Outline

1. Brief Presentation of single cell sequencing
2. Differential Expression Analysis for sequencing data
3. Differential Expression Analysis for single cell data
4. Linear Dimension reduction and data visualization
5. Alternatives to PCA, non linear embedding methods
6. Conclusions

What is missing ?

- Clustering
- Trajectory inference
- Epigenomics
- Networks

[1] *Probabilistic Count Matrix Factorization for Single Cell Expression Data Analysis*, Lecture Notes in Computer Science, Berlin, Germany, 2018. Springer.

[2] J. N. Campbell, E. Z. Macosko, H. Fenselau, T. H. Pers, A. Lyubetskaya, D. Tenen, M. Goldman, A. M. Verstegen, J. M. Resch, S. A. McCarroll, E. D. Rosen, B. B. Lowell, and L. T. Tsai. A molecular census of arcuate hypothalamus and median eminence cell types. *Nat. Neurosci.*, 20(3):484–496, Mar 2017.

[3] Michael Collins, Sanjoy Dasgupta, and Robert E. Schapire. A generalization of principal components analysis to the exponential family. In *Advances in Neural Information Processing Systems*, pages 617–624, 2001.

[4] LJP Van der Maaten, EO Postma, and HJ Van den Herik. Dimensionality reduction: A comparative review. *TiCC*, 2009.

[5] P. V. Kharchenko, L. Silberstein, and D. T. Scadden. Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, 11(7):740–742, 2014.

[6] T. Kivioja, A. Vaharautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, and J. Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, 9(1):72–74, Nov 2011.

[7] Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. *bioRxiv*, 2018.

[8] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999.

[9] Enric Llorens-Bobadilla, Sheng Zhao, Avni Baser, Gonzalo Saiz-Castro, Klara Zwadlo, and Ana Martin-Villalba. Single-Cell Transcriptomics Reveals a Population of Dormant Neural Stem Cells that Become Activated upon Brain Injury. *Cell Stem Cell*, 17(3):329–340, September 2015.

[10] A. Regev, S. A. Teichmann, E. S. Lander, and I. et al. Amit. The Human Cell Atlas. *Elife*, 6, 12 2017.

[11] D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J.-P. Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Comm*, 9(284), 2018.

[12] Alexander B Rosenberg, Charles Roco, Richard A Muscat, Anna Kuchina, Sumit Mukherjee, Wei Chen, David J Peeler, Zizhen Yao, Bosiljka Tasic, Drew L Sellers, Suzie H Pun, and Georg Seelig. Scaling single cell transcriptomics through split pool barcoding. *bioRxiv*, 2017.

[13] C. A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, and J. C. Marioni. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods*, 14(6):565–571, Jun 2017.

[14] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[15] Y. Wang and N. E. Navin. Advances and applications of single-cell sequencing technologies. *Mol. Cell*, 58(4):598–609, May 2015.