Learning from single-cell genomic data

Jean-Philippe Vert
“Bulk” genomics is great!

- Mutations
  - WGS (whole genome)
  - WES (whole exome)
- Gene expression (RNA-seq)
- DNA accessibility
- DNA methylation
- Histone modification
- ....

https://www.slideshare.net/nurialopezbigas/identification-of-cancer-drivers-across-tumor-types
But sometimes, not enough
From “bulk” to “single-cell” genomics

Inspired from slides of A. Regev
Eg: single-cell genomics to study intra-tumor heterogeneity
Single cell datasets are getting large enough for machine learning
1. Extracting signal from raw data
2. Gene regulatory network inference
3. Integration of multi-omics data
1. Extracting signal from raw data

2. Gene regulatory network inference

3. Integration of multi-omics data
Some challenges

<table>
<thead>
<tr>
<th></th>
<th>SRR1275356</th>
<th>SRR1274090</th>
<th>SRR1275251</th>
<th>SRR1275287</th>
<th>SRR1275364</th>
<th>SRR1275269</th>
<th>SRR1275263</th>
<th>SRR1275242</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1BG</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A1BG-AS1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A1CF</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A2M</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>31</td>
<td>0</td>
<td>0</td>
<td>46</td>
<td>0</td>
</tr>
<tr>
<td>A2M-AS1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A2ML1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A2MP1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>A3GALT2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A4GALT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A4GNT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AA06</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>35</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AAAS</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>312</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>AACS</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AACSP1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AADAC</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Some challenges
A general and flexible method for signal extraction from single-cell RNA-seq data
Some benefits

Better lineage reconstruction

Robustness to drop-out

Robustness to batch effects
Hot topic!

Deep generative modeling for single-cell transcriptomics

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan and Nir Yosef

scVAE: Variational auto-encoders for single-cell gene expression data

Christopher H Grønbech, Maximillian F Vording, Pascal N Timshel, Capser K Sønderby, Tune H Pers and Ole Winther

Single-cell RNA-seq denoising using a deep count autoencoder

Gökcen Eraslan, Lukas M. Simon, Maria Mircea, Nikola S. Mueller and Fabian J. Theis

A general and flexible method for signal extraction from single-cell RNA-seq data

Davide Risso, Fanny Perradeau, Svetlana Gribkova, Sandrine Dudoit and Jean-Philippe Vert

AutoImpute: Autoencoder based imputation of single-cell RNA-seq data

Divyangshu Talwar, Aanchal Mongia, Debarka Sengupta and Angshul Majumdar

Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics

Qiwen Hu and Casey S. Greene
1. Extracting signal from raw data

2. Gene regulatory network inference

3. Integration of multi-omics data
GRN inference from bulk expression data

- Connect “similar” genes (co-expression, mutual information…)
- Causal inference (Bayesian network, causal networks...)
- Sparse regression (Random forests, lasso..)
Steady-state hypothesis for regression methods (Genie3, TIGRESS…)

- The dynamic equation of the mRNA concentration of a gene is of the form:

\[
\frac{dX}{dt} = f(X, R)
\]

where \( R \) represent the set of concentrations of transcription factors that regulate \( X \).

- At steady state, \( \frac{dX}{dt} = 0 = f(X, R) \)

- If we linearize \( f(X, R) = 0 \) we get linear relation of the form

\[
X = \sum_{i \in R} \beta_i X_i
\]

- This suggests to look for transcription factors whose expression is sufficient to explain the expression of \( X \) across different experiments.
Steady-state hypothesis for single-cell data?

From p. 17 of T. Lönnberg et al. Single-cell RNA-seq and computational analysis using temporal mixture modelling resolves Th1/Tfh fate bifurcation in malaria, Sci Immunol. 2(9), March 24, 2017
Pseudo-time

Trapnell (2015)
From steady-state to dynamical model

\[
\frac{dx}{dt} = Ax
\]

- Given cells \((X_i, t_i)\) for \(i=1,\ldots,N\)
  - \(X_i\) vector of expression
  - \(t_i\) inferred pseudo-time

- How to infer a **sparse model** \(A\)?
SCODE (Matsumoto et al 2017)

\[
\min_{A \in \mathcal{M}_n(\mathbb{R})} \sum_i \| X_{t_i} - \exp(t_i A)X_0 \|^2_2
\]

- Hard to solve (nonconvex…)
- Sensitive to noise for large pseudo-time
GRISLI (Aubin and V., 2018)

- Solve instead

$$\min_{A \in \mathcal{M}_n(\mathbb{R})} \sum_i \|X'_{t_i} - AX_{t_i}\|_2^2$$

- Pro:
  - easy to solve (convex, sparse regression)
  - Not sensitive to outliers for large t

- Cons
  - Need to infer velocity $v_i = X'_t - ti$ of each cell
Velocity inference

\[ \hat{v}_{i,j} = \frac{x_j - x_i}{t_j - t_i}. \]

\[ K(x, t, x', t') = (t - t')^2 \exp \left( -\frac{(t - t')^2}{2\sigma_t^2} \right) \times \exp \left( -\frac{||x - x'||^2_{\text{GR}}}{2\sigma_x^2} \right) \]

\[ \hat{v}_i = \frac{1}{2} \sum_{j \mid t_j > t_i} K(x_i, t_i, x_j, t_j) \hat{v}_{i,j} \]

\[ + \frac{1}{2} \sum_{j \mid t_j < t_i} K(x_i, t_i, x_j, t_j) \hat{v}_{i,j} \]
**Validation (AUC)**

**Murine**: 373 cells, direct reprogramming of murine embryonic fibroblasts to myocytes at days 0, 2, 5, 22 (Treutlein et al. 2016)

**Human**: 758 cells, differentiation of human ES cells to definitive endoderm cells at 0, 12, 24, 36, 72, 96h (Chu et al. 2016)
New velocity inference...

Resource

Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming

Geoffrey Schiebing er1,11,16, Jian Shu1,2,16, X.2,16, Marcin Tabaka1,16, Brian Cleary1,3,16, Vidya Subramanian1, Ayeh Solomon1,17, Joshua Gould1, Siyan Liu1,13, Stacie Lin1,6, Peter Berube1, Lia Lee1, Jenny Chen1,4, Justin Brumbaugh5,7,8,9,10, Philippe Rigollet11,12, Konrad Hochdorferg7,8,9,13, Rudolf Jaenisch2,3, Aviv Regev1,6,13

Eric S. Lander1,6,14,18,28

RNA velocity of single cells

Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochger, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson & Peter V. Kharchenko

Nature 560, 494–498 (2018) | Download Citation ↓
1. Extracting signal from raw data

2. Gene regulatory network inference

3. Integration of multi-omics data
Flood of single-cell data
Data integration is important

- Cells
- Measurements:
  - Gene Expression
  - Open regions
  - 3D Structure
  - Microscopy Images

Data integration
Integration of single data is challenging

**sci-RNA-seq**
- Gene Expression
- No 1-1 correspondence between cells

**sci-ATAC-seq**
- Open regions
- No 1-1 correspondence between features
Integrate single-cell data by projecting to a shared manifold.
Related work

- Joint Laplacian Manifold Alignment (JLMA; Wang 2011)
  - Construct a joint Laplacian across multiple domains and perform eigenvalue decomposition.
  - Relies on k-nearest neighbor graph to characterize local geometry.
- Generalized unsupervised manifold alignment (GUMA; Cui NIPS 2014)
  - Optimize a function with three terms: geometry matching term across domains, feature matching, and geometry preserving term within domains.
  - Assumes that instances in the two domains can be matched one-to-one.
- Manifold Alignment Generalized Adversarial Network (MAGAN; Amodio ICML 2018)
  - Two generative adversarial networks that learn reciprocal mappings between two domains
  - In practice, requires prior information about correspondence between features.
An approach: MMD-based algorithm to align single-cell data
(Lui, Huang, Ritambhara, V. and Noble, WABI 2019)

Assumption: Data shares a projection to a common manifold structure
Maximum mean discrepancy (MMD) measures the distance between two distributions

\[
\text{MMD}^2_u[\mathcal{F}, X, Y] = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(y_i, y_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j)
\]

MMD manifold alignment (MMD-MA) minimizes the distance between two or more distributions

$$\min_{\alpha_1, \alpha_2} MMD(K_1 \alpha_1, K_2 \alpha_2) + \lambda_1 (pen(\alpha_1) + pen(\alpha_2)) + \lambda_2 (distortion(\alpha_1) + distortion(\alpha_2))$$

- **Parameters learned during training**
- **Penalty term to avoid trivial solution**
- **Distortion term to preserve structure**

- sci-RNA-seq
- sci-ATAC-seq
MMD-MA works well for simulated data

Original manifold
(n=2000)

Sampled points
(dimension = 2)

Random projection +
Gaussian noise

Learned manifold

MMD-MA
(dimension = 5)

1000-dimension
ional space

2000-dimension
ional space

Set 1
(n=300)

Set 2
(n=300)

View 1
(n=300)

View 2
(n=300)
Comparing to the baseline (JLMA)

Branching structure

Branching structure + Swiss roll

Circular frustrum in 3D

Fraction of samples closer than true match

1 2 3

JLMA (k=5) JLMA (k=6) MMD (k=5)
Aligning single-cell RNA-seq and DNA methylation data

MMD-MA aligns single-cell RNAseq and DNA methylation data
MMD-MA correctly matches cells

- For >50% of the cells, the nearest neighbor is the correct match.
- On average, only 2.4% of the cells are closer than the true match.
Summary

- MMD-MA is an unsupervised algorithm
- Uses MMD measure to match two distributions
- Does not require sample or feature correspondence
- Performs well for both simulated and biological data
Conclusion

- Single cell genomics moving the field to “big data”
- Many exciting perspectives
- Many challenges as well
  - Data with largely unknown structure, trade-off quality/quantity
  - Cells communicate