

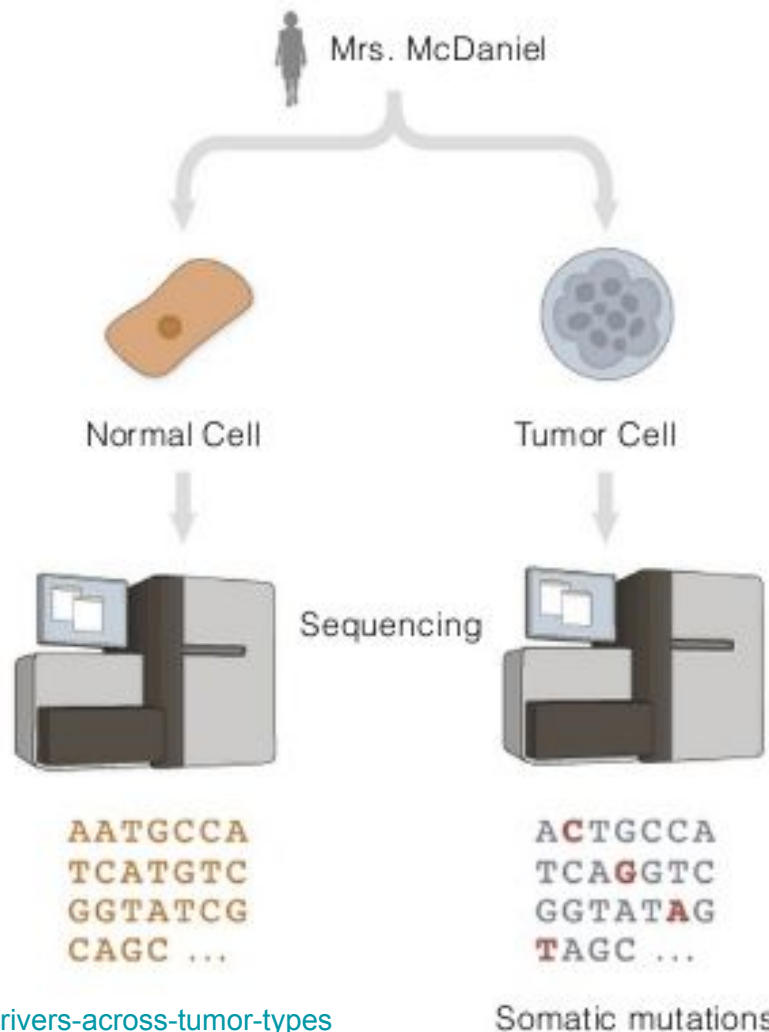
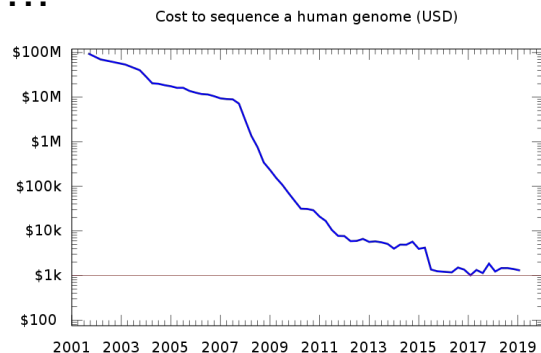
Learning from single-cell genomic data

Jean-Philippe Vert

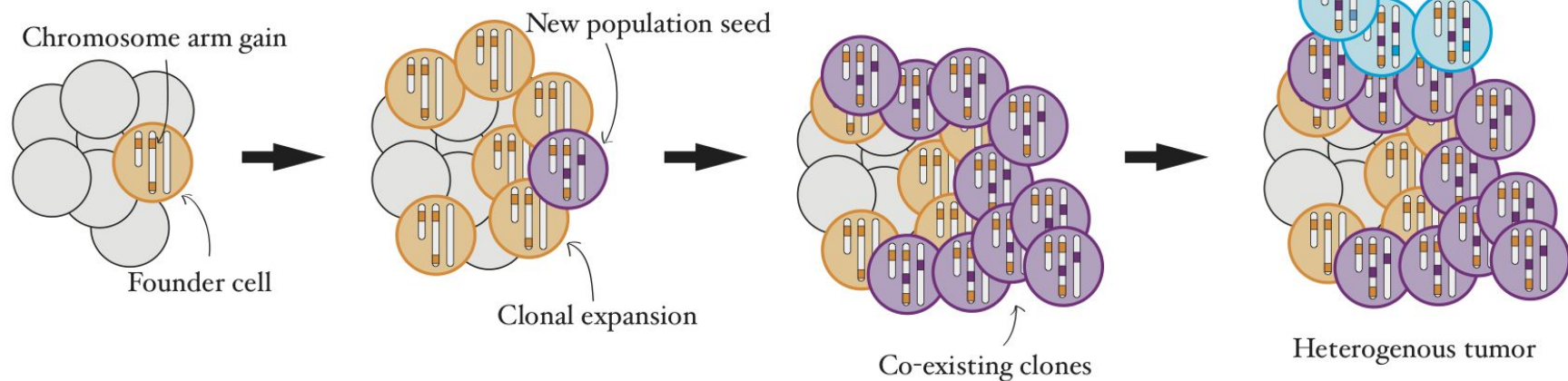
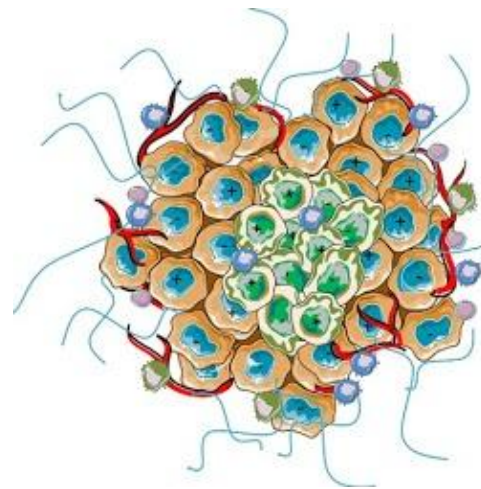


“Bulk” genomics is great!

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



But sometimes, not enough



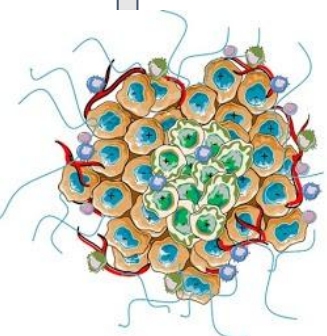
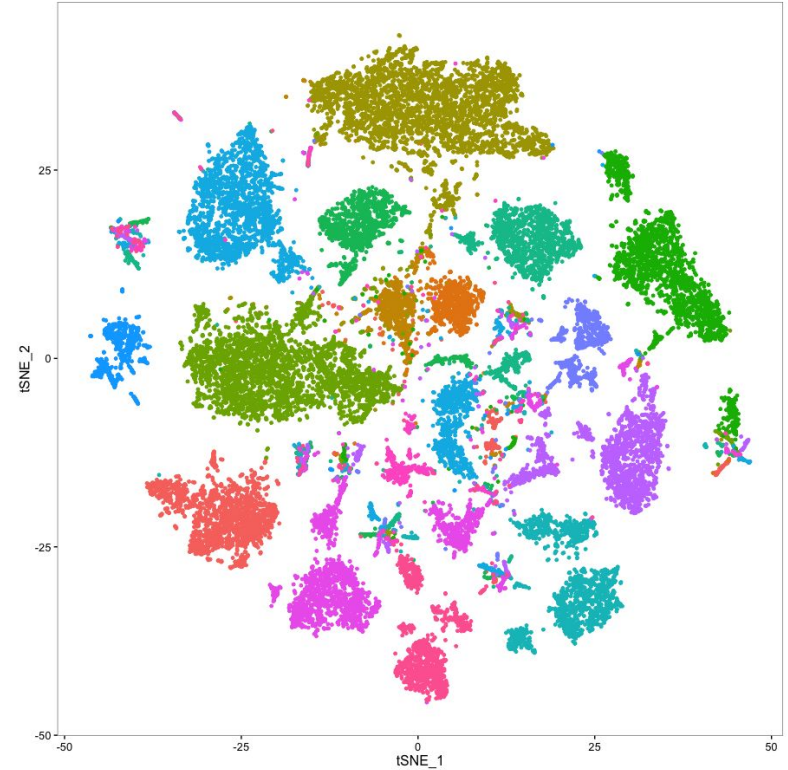
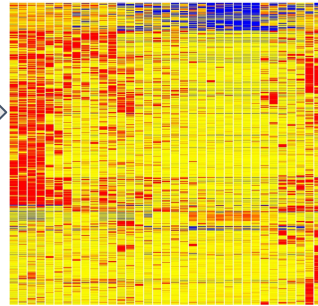
From “bulk” to “single-cell” genomics



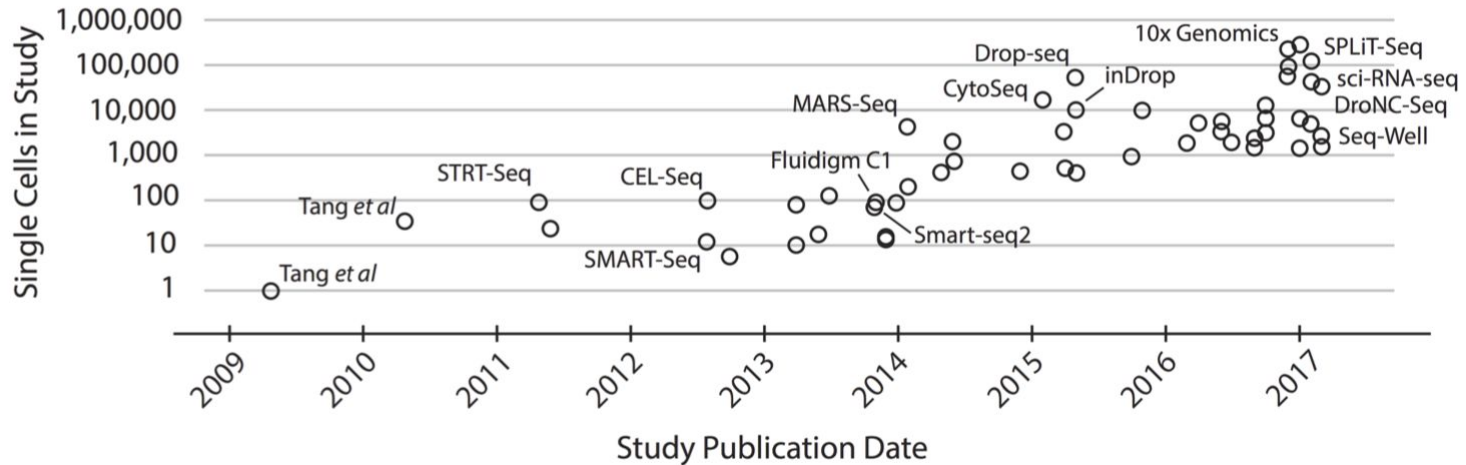
Inspired from slides of A. Regev



Eg: single-cell genomics to study intra-tumor heterogeneity



Single cell datasets are getting large enough for machine learning

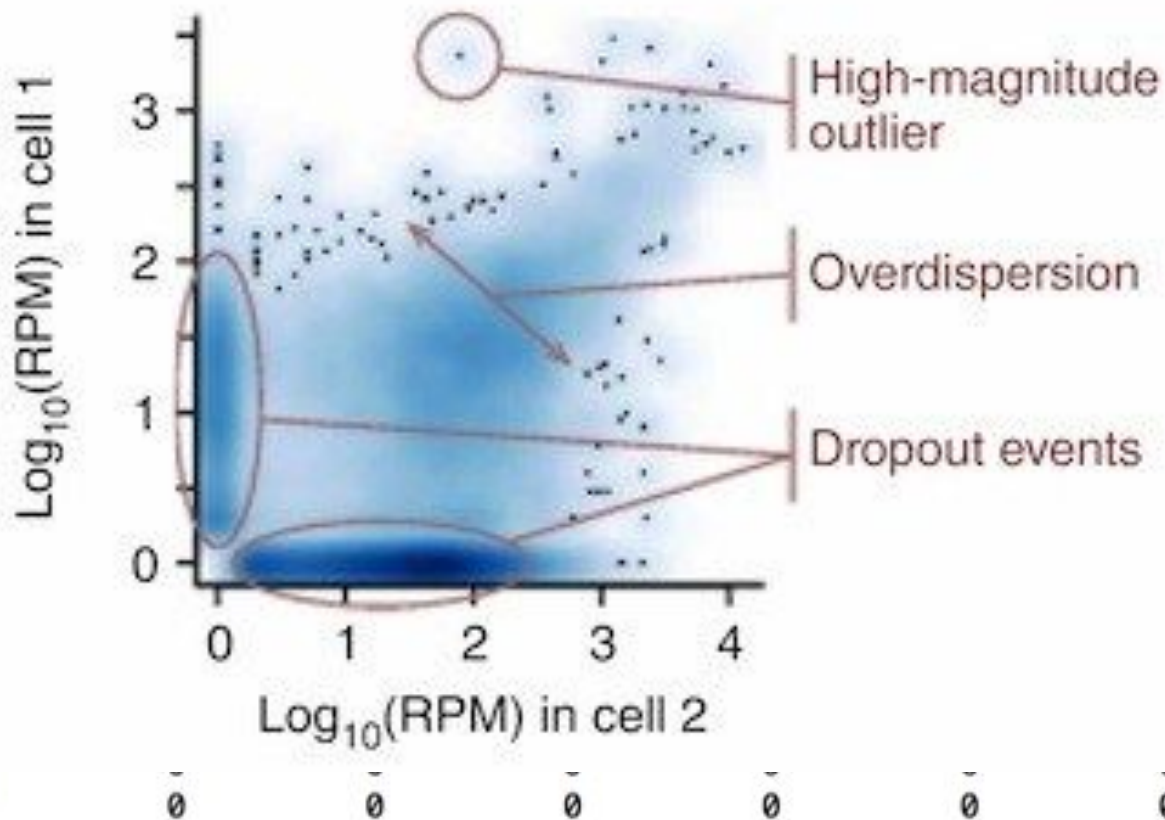


1. Extracting signal from raw data
2. Gene regulatory network inference
3. Integration of multi-omics data

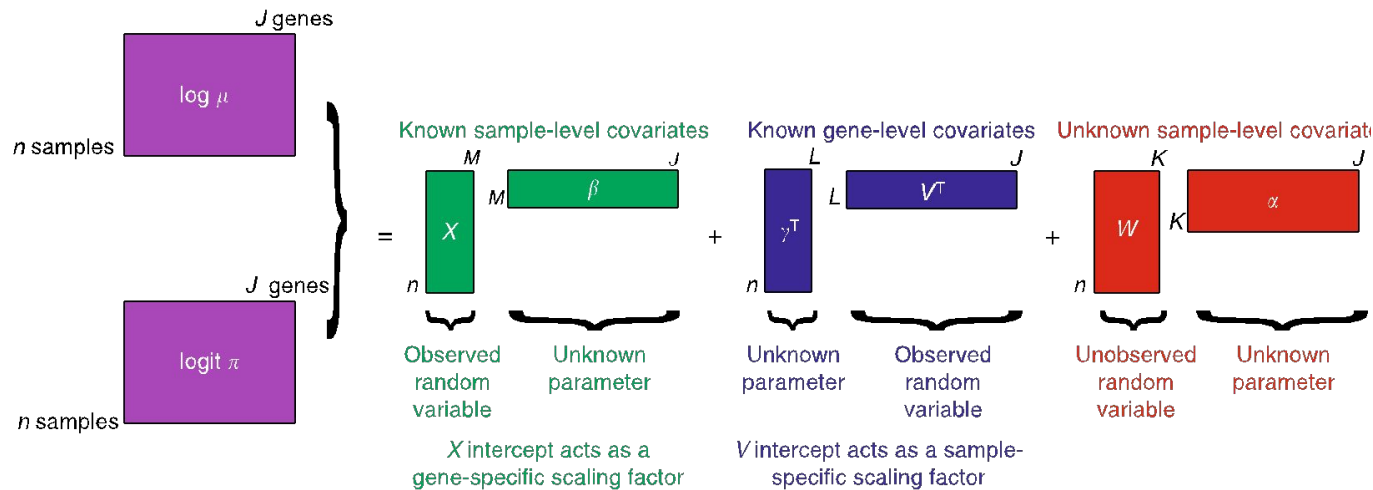
1. Extracting signal from raw data
2. Gene regulatory network inference
3. Integration of multi-omics data

Some challenges

	SRR1275356	SRR1274090
A1BG	0	0
A1BG-AS1	0	0
A1CF	0	0
A2M	0	0
A2M-AS1	0	0
A2ML1	0	0
A2MP1	0	0
A3GALT2	0	0
A4GALT	0	0
A4GNT	0	0
AA06	0	0
AAAS	0	0
AACS	1	0
AACSP1	0	0
AADAC	0	0

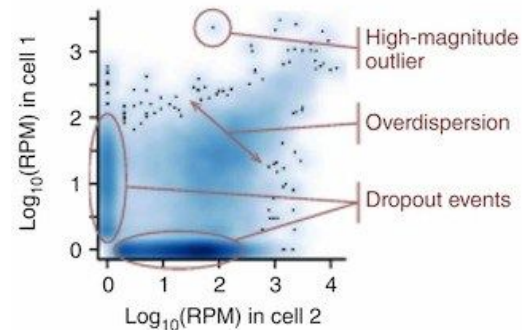


ZINB-WaVe



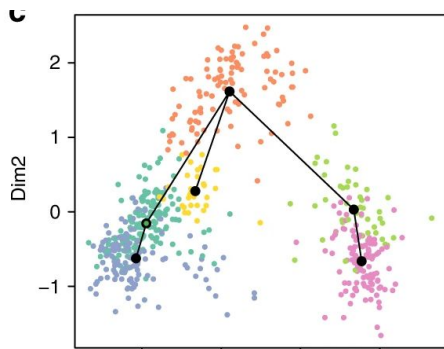
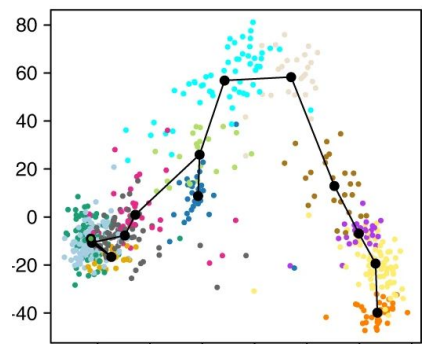
$$f_{ZINB}(y; \mu, \theta, \pi) = \pi \delta_0(y) + (1 - \pi) f_{NB}(y; \mu, \theta), \quad \forall y \in \mathbb{N},$$

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

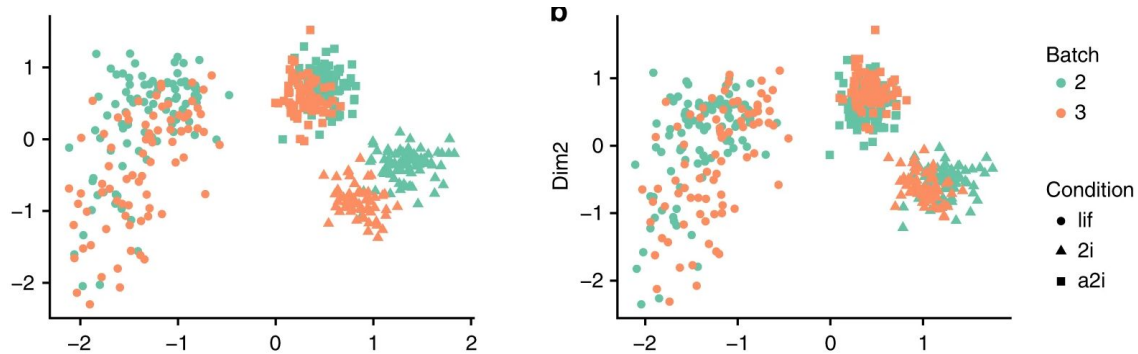


Some benefits

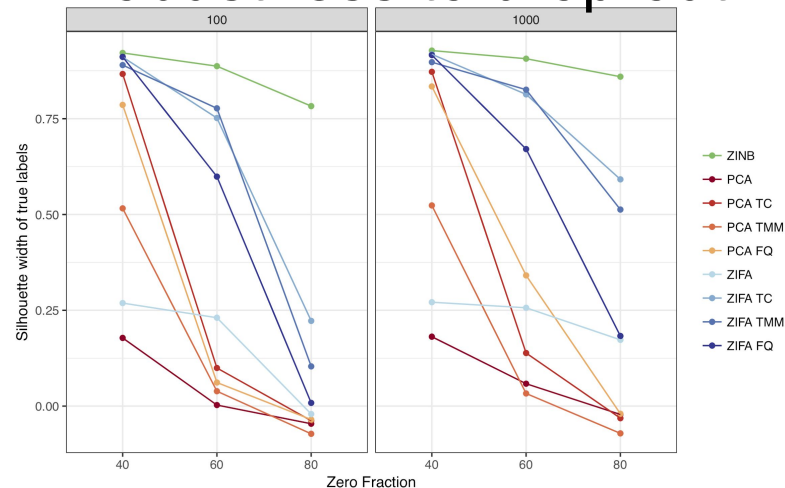
Better lineage reconstruction



Robustness to batch effects



Robustness to drop-out



Hot topic!

Deep generative modeling for single-cell transcriptomics

Romain Lopez¹, Jeffrey Regier¹, Michael B. Cole², Michael I. Jordan^{1,3} and Nir Yosef^{1,4,5*}

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

Christopher H Grønbech¹, Maximilian F Vording¹, Pascal N Timshel^{2,3}, Capser K Sønderby⁴, Tune H Pers^{2,3}, and Ole Winther^{1,4}

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

Gökçen Eraslan^{1,2}, Lukas M. Simon¹, Maria Mircea¹, Nikola S. Mueller¹ & Fabian J. Theis^{1,2,3}

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

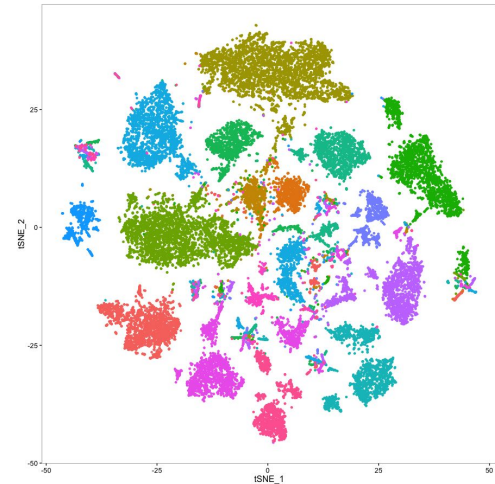
Davide Risso¹, Fanny Perraudeau², Svetlana Gribkova³, Sandrine Dudoit^{2,4} & Jean-Philippe Vert^{1,5,6,7,8}

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

Divyanshu Talwar¹, Aanchal Mongia¹, Debarka Sengupta^{1,3} & Angshul Majumdar²

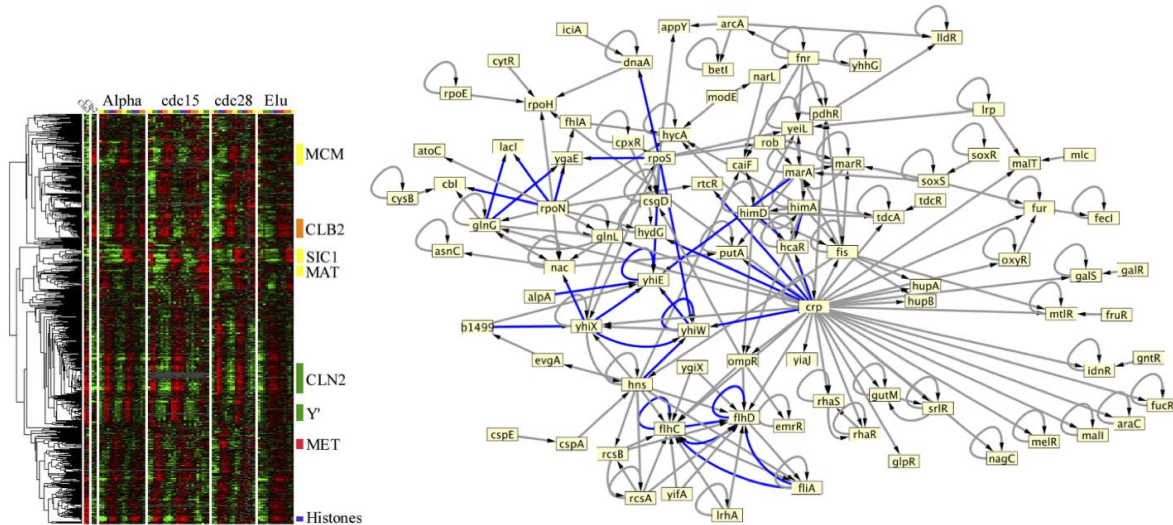
Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics

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:

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

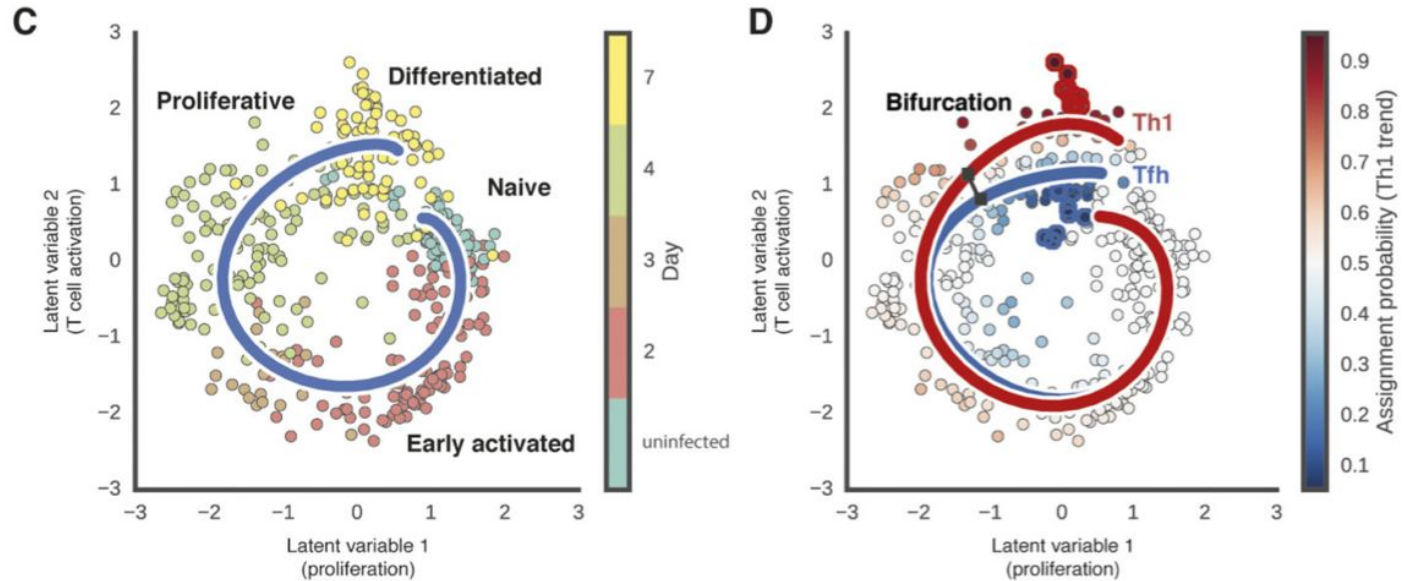
where R represent the set of concentrations of transcription factors that regulate X .

- At steady state, $dX/dt = 0 = f(X, R)$
- If we linearize $f(X, R) = 0$ we get linear relation of the form

$$X = \sum_{i \in R} \beta_i X_i$$

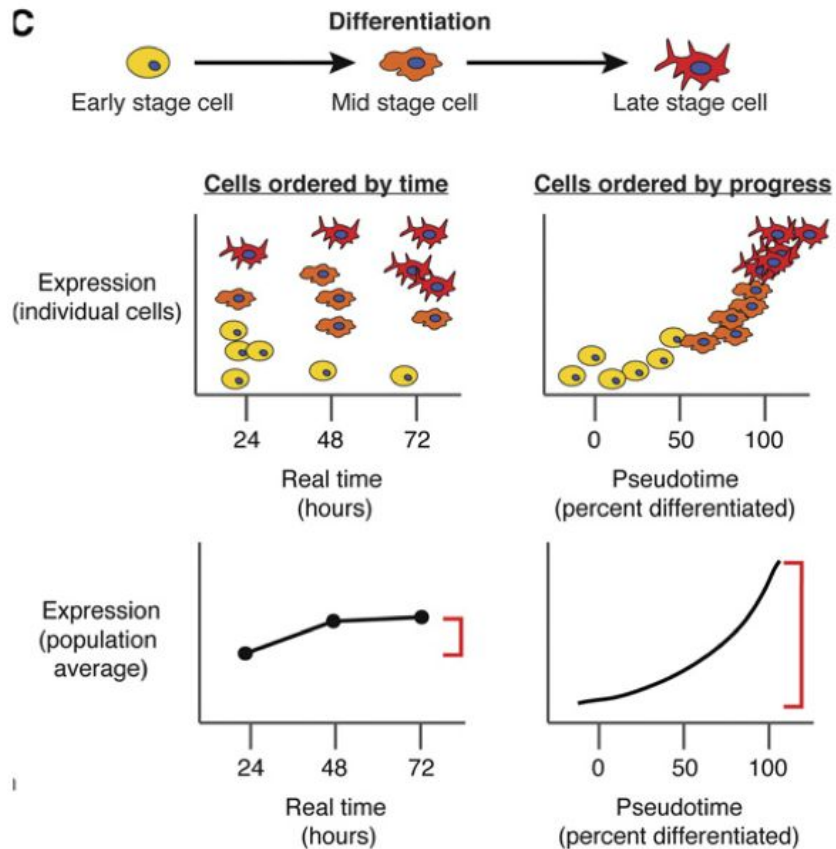
- This suggests to look for **transcription factors whose expression is sufficient to explain the expression of X across different experiments.**

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 $i=1, \dots, N$
 - 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

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

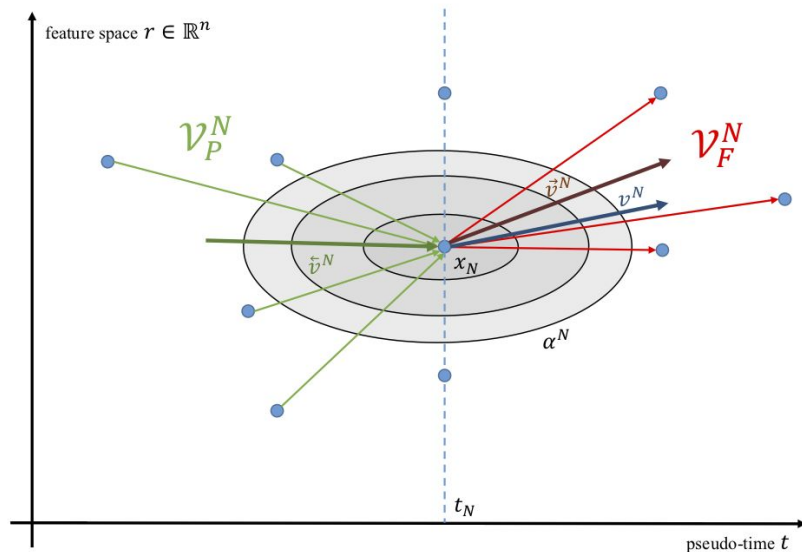
- Pro:
 - easy to solve (convex, sparse regression)
 - Not sensitive to outliers for large t
- Cons
 - Need to infer velocity $v_i = X'_{t_i}$ of each cell

Velocity inference

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

$$K(x, t, x', t') = (t - t')^2 \exp\left(-\frac{(t - t')^2}{2\sigma_t^2}\right) \times \exp\left(-\frac{\|x - x'\|_{\mathbb{R}^G}^2}{2\sigma_x^2}\right)$$

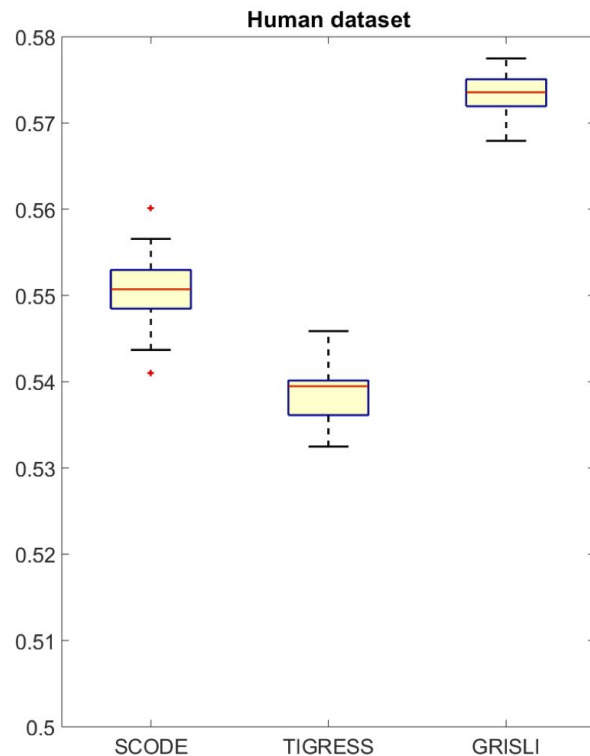
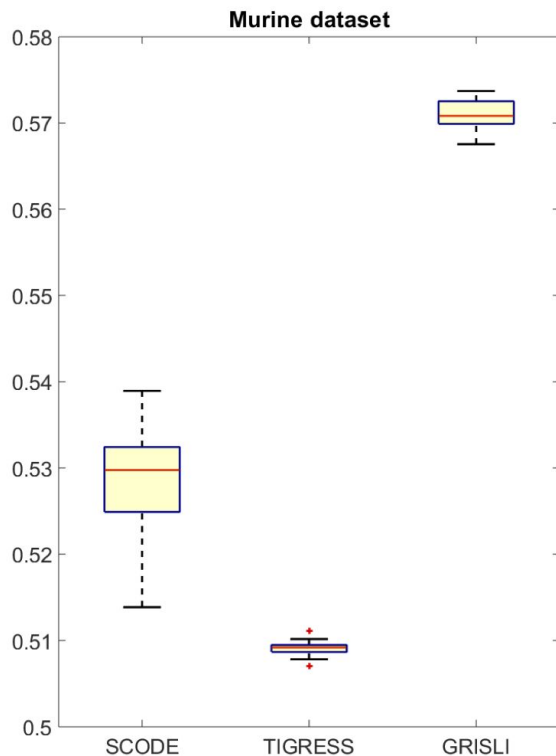
$$\hat{v}_i = \frac{1}{2} \frac{\sum_{j | t_j > t_i} K(x_i, t_i, x_j, t_j) \hat{v}_{i,j}}{\sum_{j | t_j > t_i} K(x_i, t_i, x_j, t_j)} + \frac{1}{2} \frac{\sum_{j | t_j < t_i} K(x_i, t_i, x_j, t_j) \hat{v}_{i,j}}{\sum_{j | t_j < t_i} K(x_i, t_i, x_j, t_j)}.$$



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)

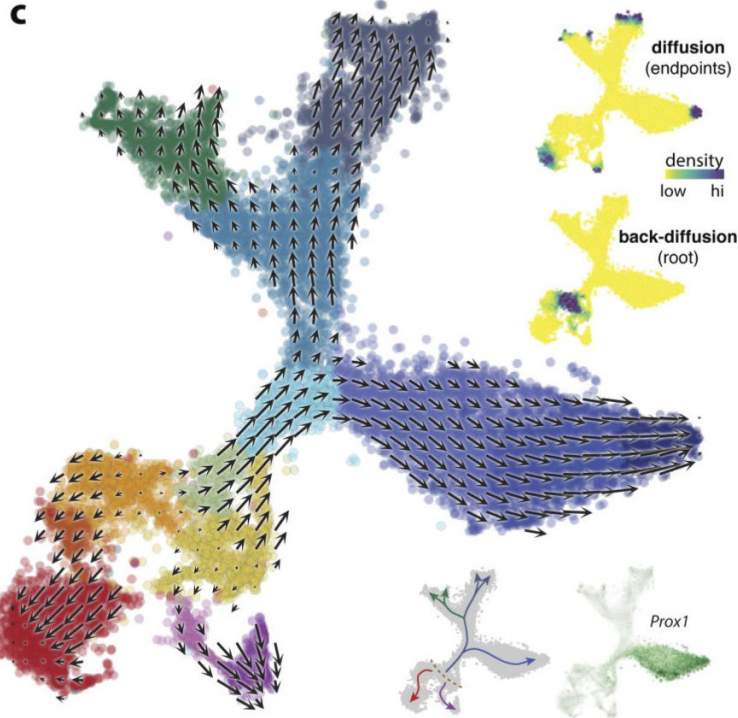


New velocity inference...

Cell

Volume 176, Issue 4, 7 February 2019, Pages 928-943.e22

CellPress






Resource

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

Geoffrey Schiebinger^{1, 11, 16}, Jian Shu^{1, 2, 16}  , Marcin Tabaka^{1, 16}, Brian Cleary^{1, 3, 16}, Vidya Subramanian¹, Aryeh Solomon^{1, 17}, Joshua Gould¹, Siyan Liu^{1, 15}, Stacie Lin^{1, 6}, Peter Berube¹, Lia Lee¹, Jenny Chen^{1, 4}, Justin Brumbaugh^{5, 7, 8, 9, 10}, Philippe Rigollet^{11, 12}, Konrad Hochedlinger^{7, 8, 9, 13}, Rudolf Jaenisch^{2, 3}, Aviv Regev^{1, 6, 13}  , Eric S. Lander^{1, 6, 14, 18}  

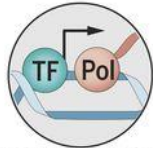
RNA velocity of single cells

Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastrioti, 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 

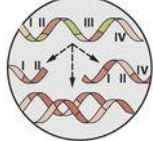
Nature **560**, 494–498 (2018) | [Download Citation](#) 

1. Extracting signal from raw data
2. Gene regulatory network inference
3. Integration of multi-omics data

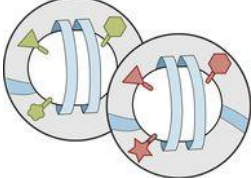
Flood of single-cell data



Transcription factor binding
 TF binding interacts with DNA methylation and chromatin accessibility



Transcription and RNA maturation



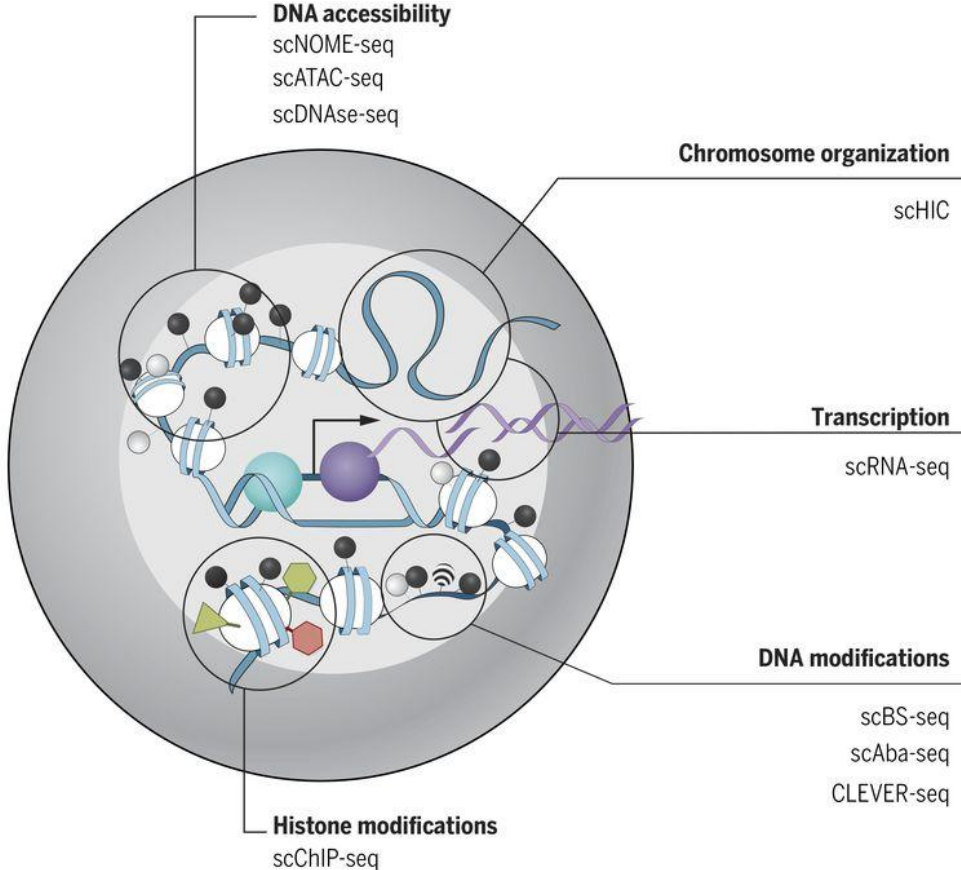
Histone modifications
 Modifications can be active marks (e.g., H3K4me3 in green) or repressive marks (e.g., H2K27m3 in red)



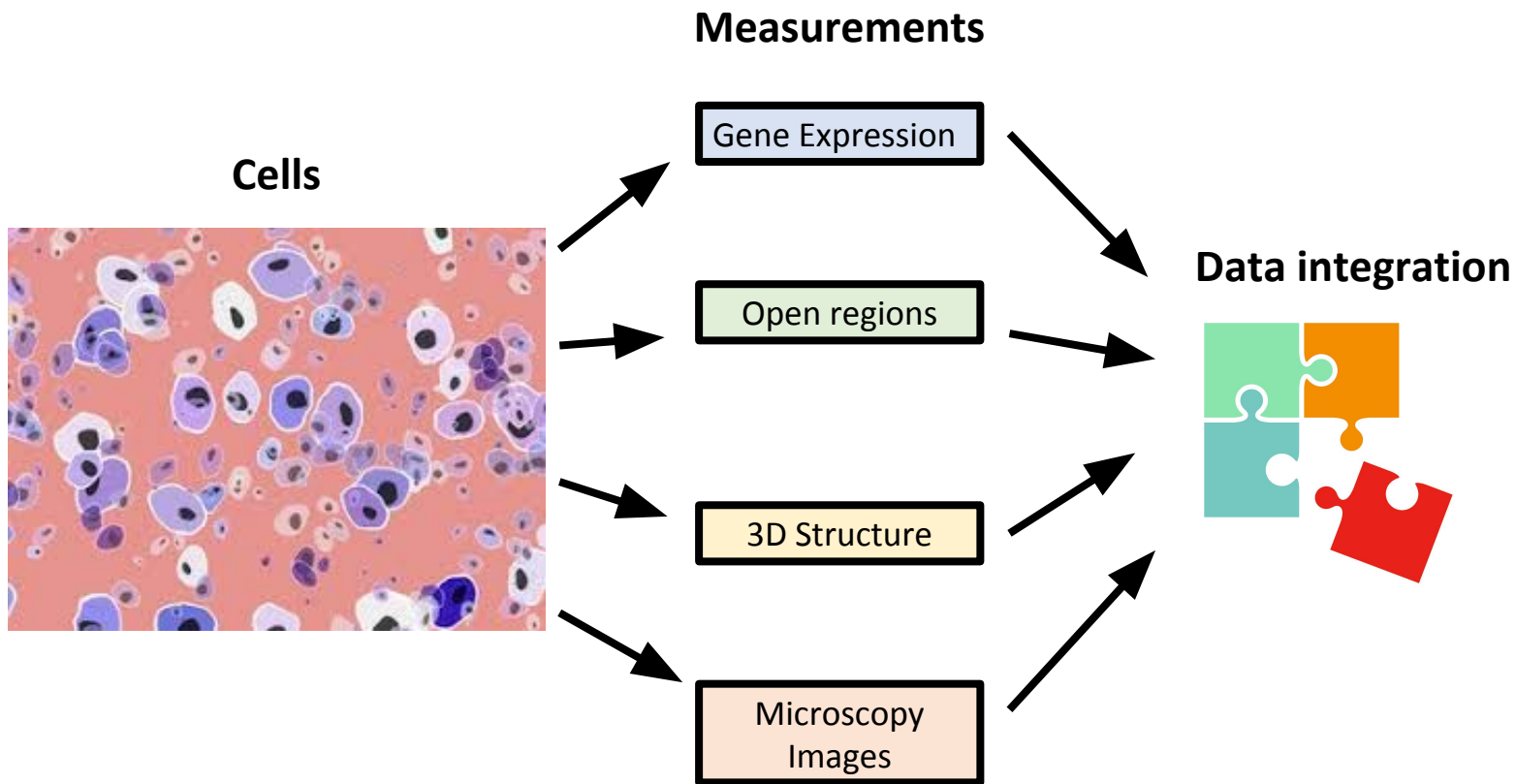
DNA modifications
 ● C ● 5mC
 ☹ 5hmC / 5fC / 5caC



Chromosome organization
 Higher-order chromatin organization into LADs and TADs



Data integration is important



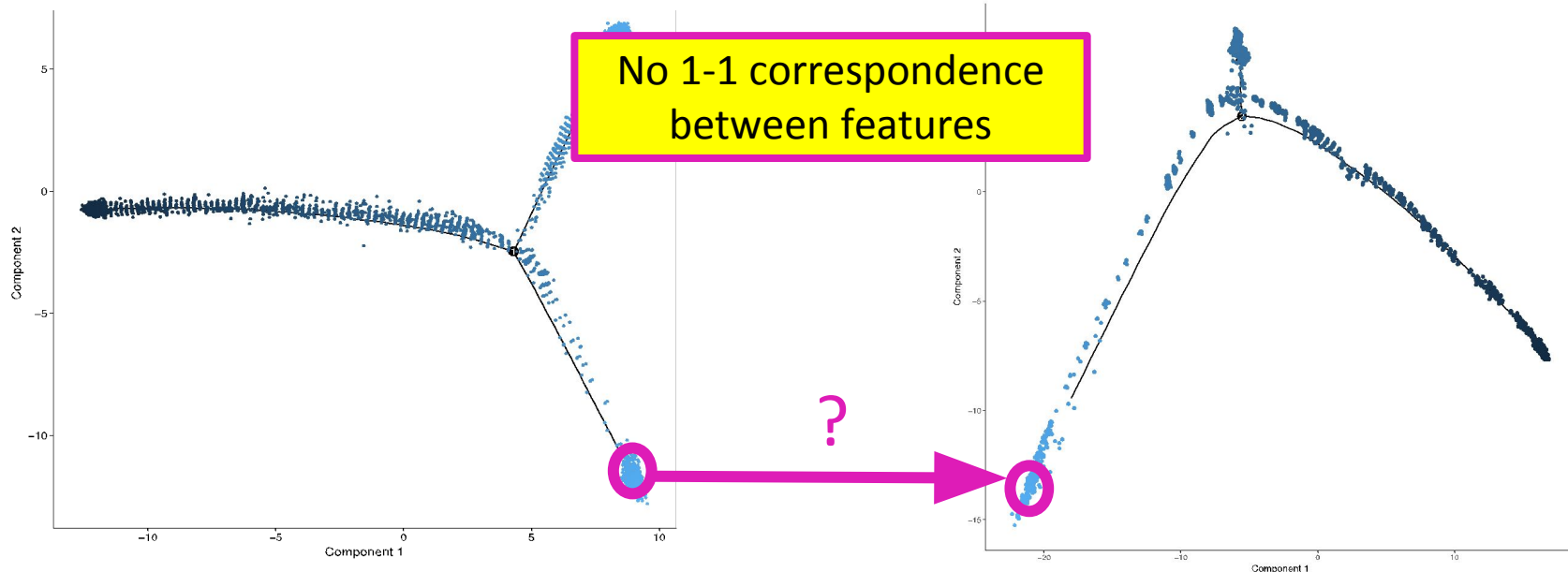
Integration of single data is challenging

sci-RNA-seq
Gene Expression

No 1-1 correspondence
between cells

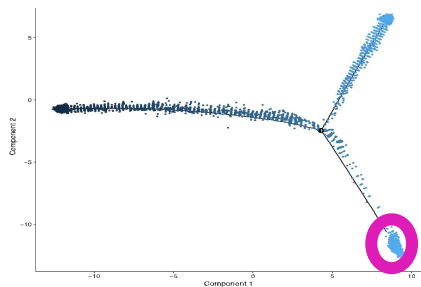
sci-ATAC-seq
Open regions

No 1-1 correspondence
between features

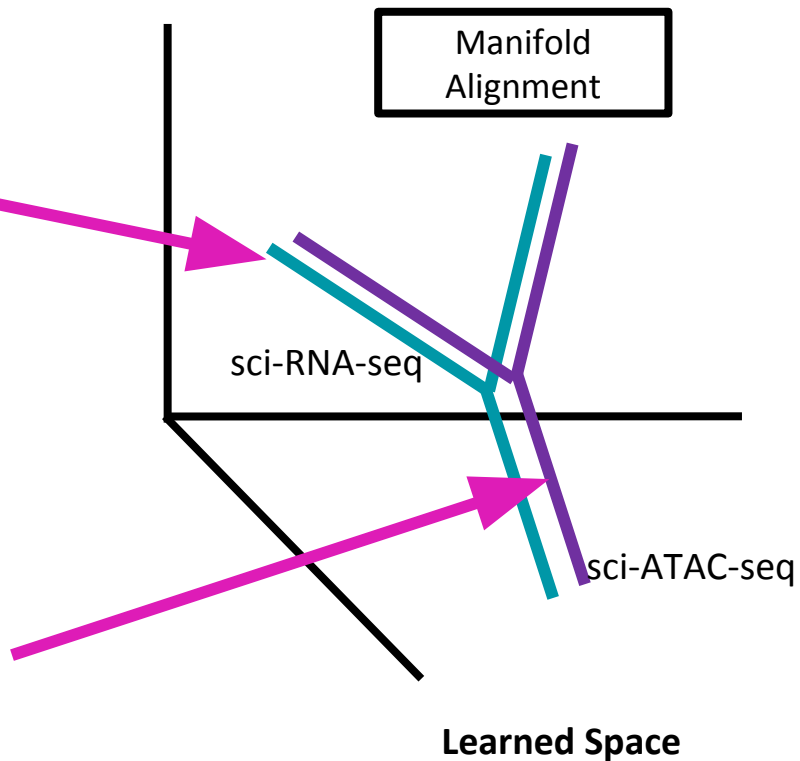
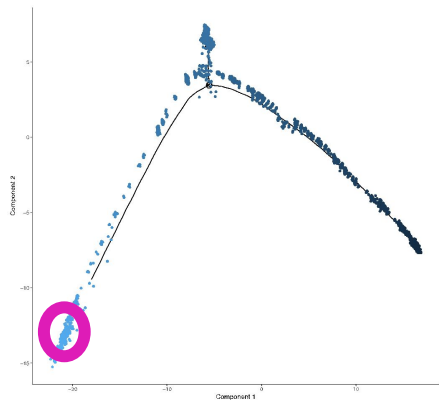


Integrate single-cell data by projecting to a shared manifold

sci-RNA-seq



sci-ATAC-seq

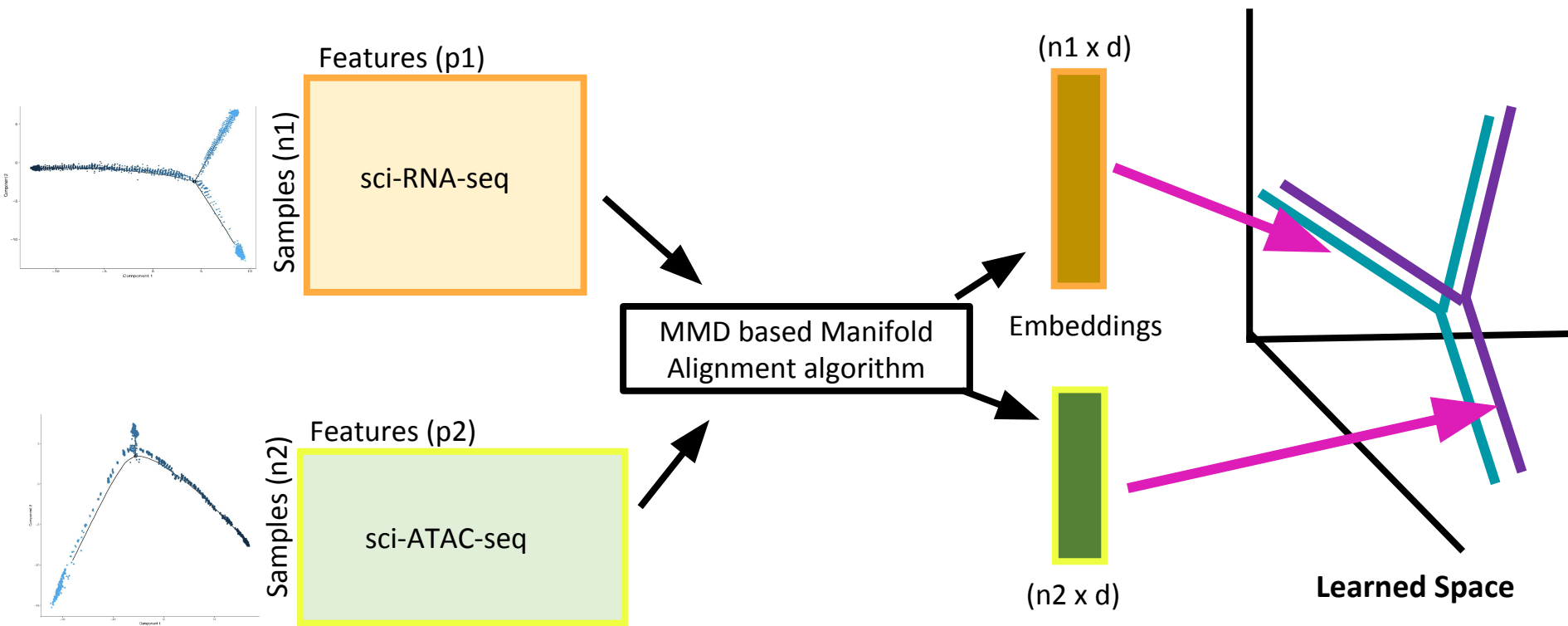


Related work

- 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

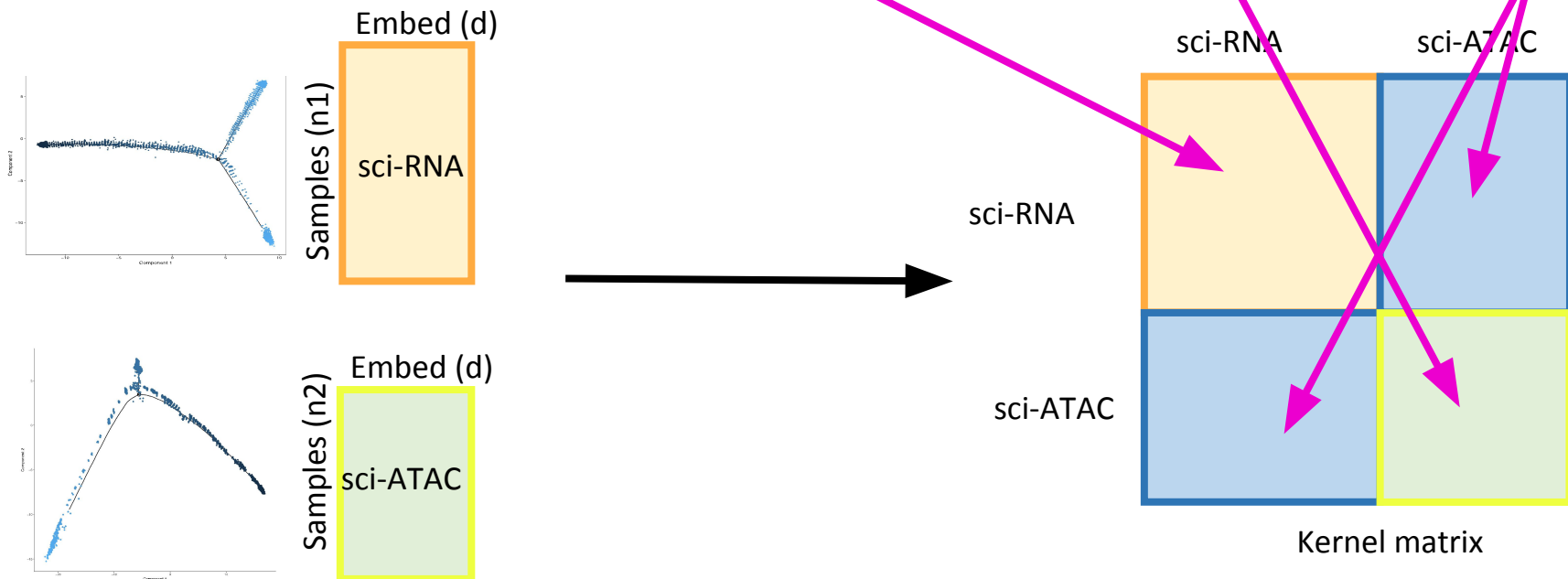
(Lui, Huang, Ritambhara, V. and Noble, WABI 2019)



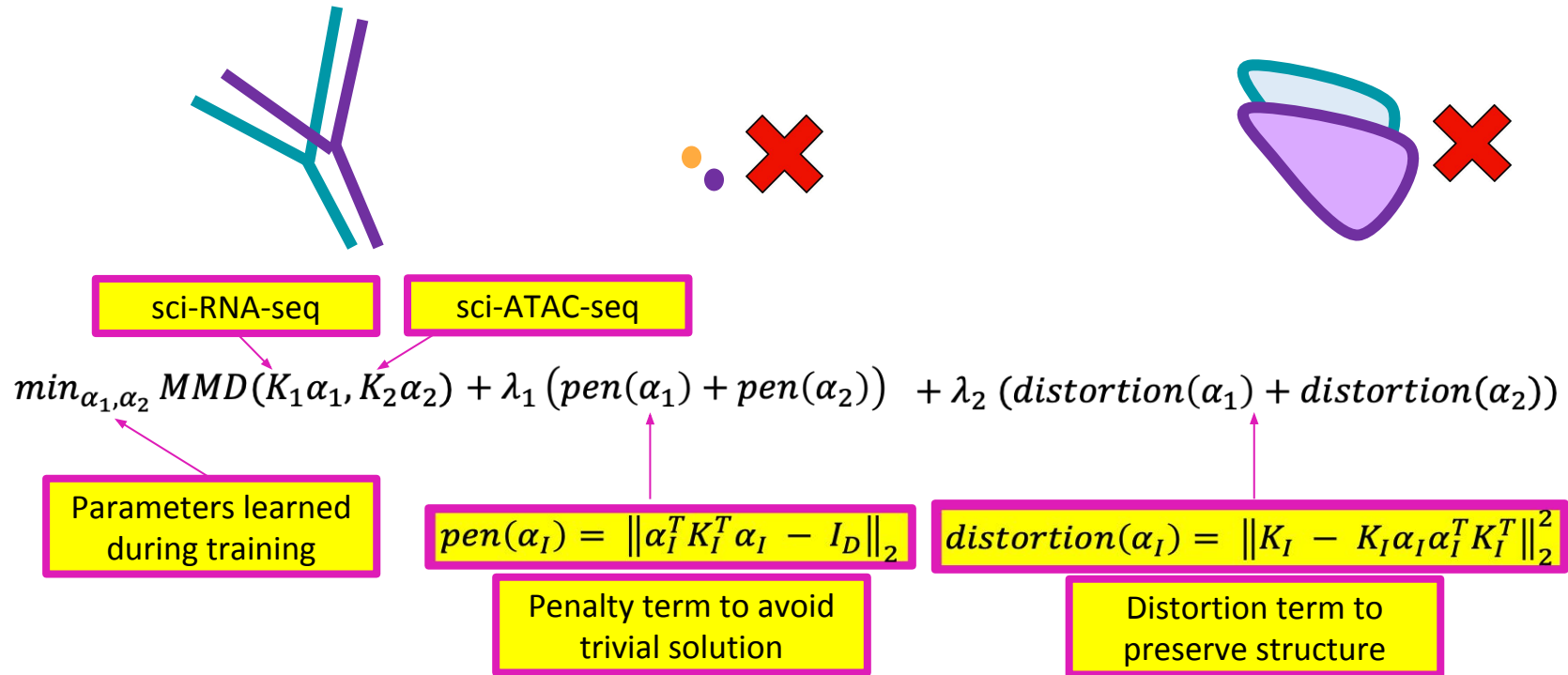
Assumption: Data shares a projection to a common manifold structure

Maximum mean discrepancy (MMD) measures the distance between two distributions

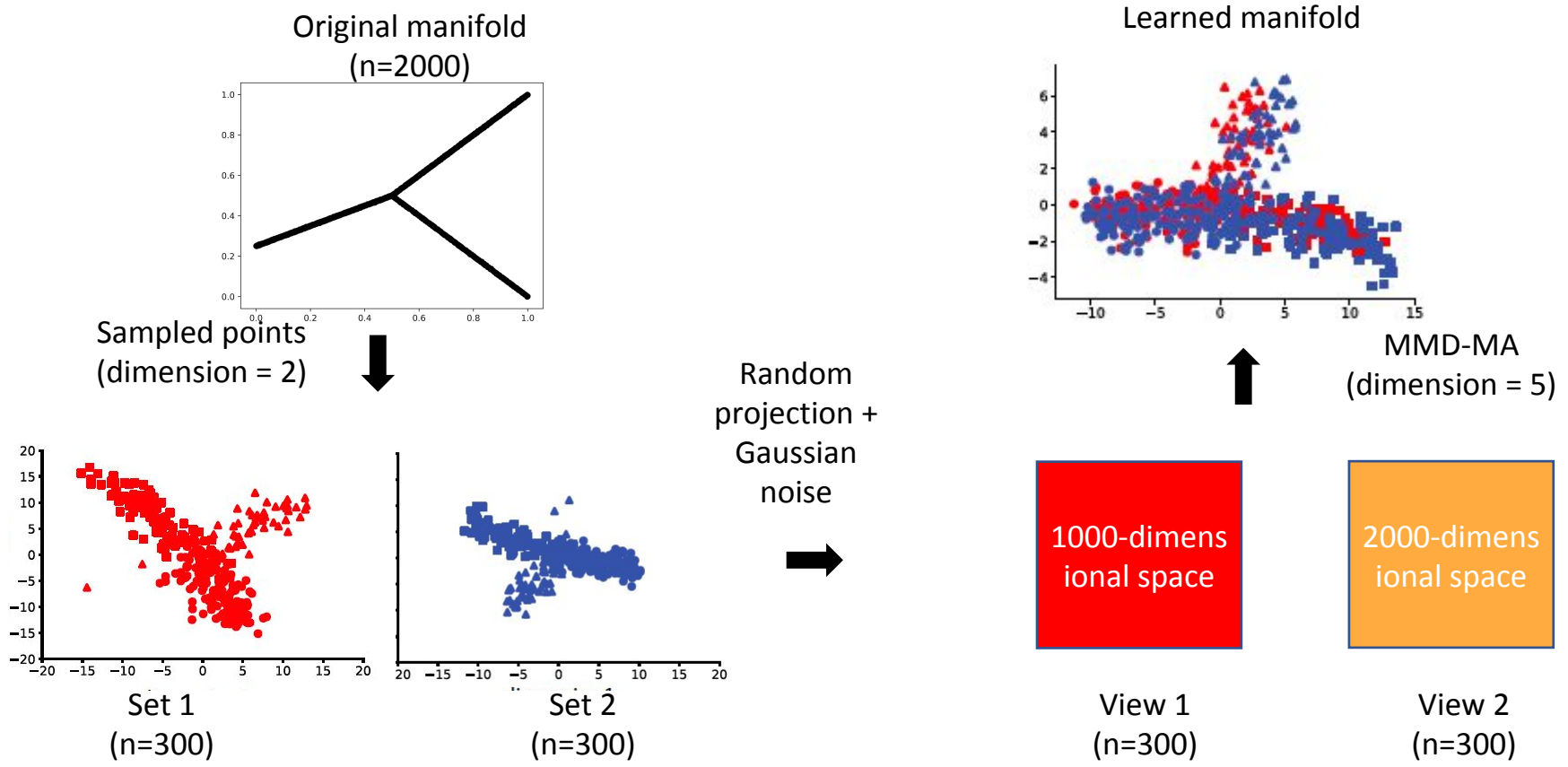
$$\text{MMD}_u^2[\mathcal{F}, X, Y] = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n k(y_i, y_j) - \frac{2}{mn} \sum_{i=1}^m \sum_{j=1}^n k(x_i, y_j)$$



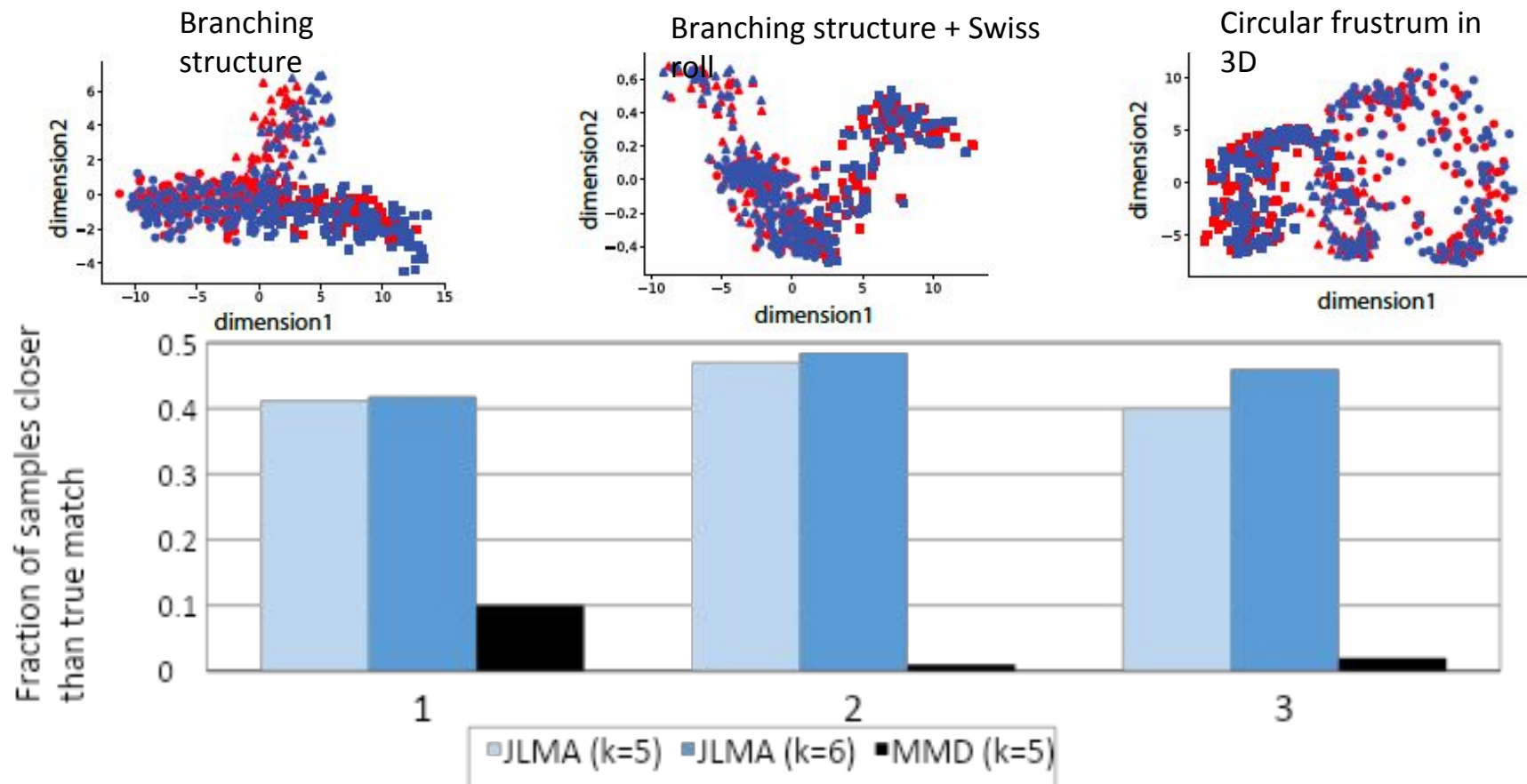
MMD manifold alignment (MMD-MA) minimizes the distance between two or more distributions



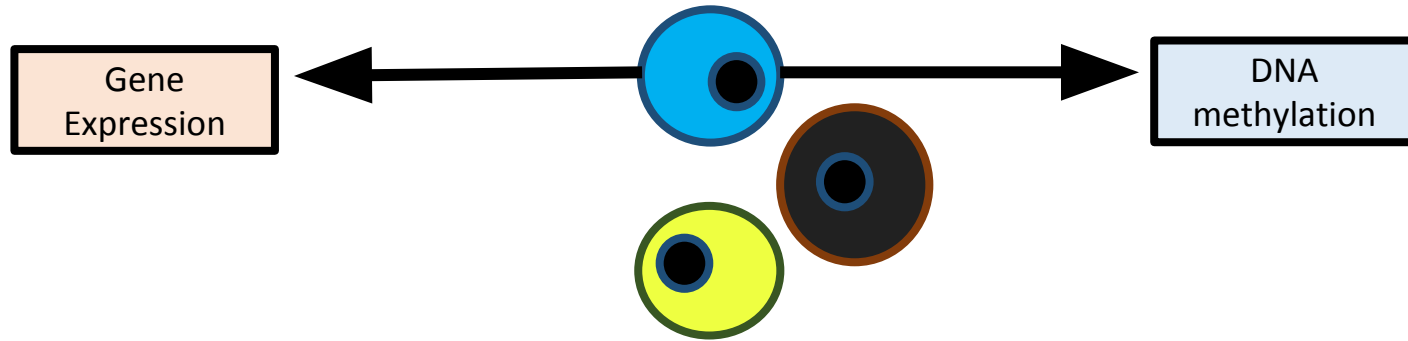
MMD-MA works well for simulated data



Comparing to the baseline (JLMA)

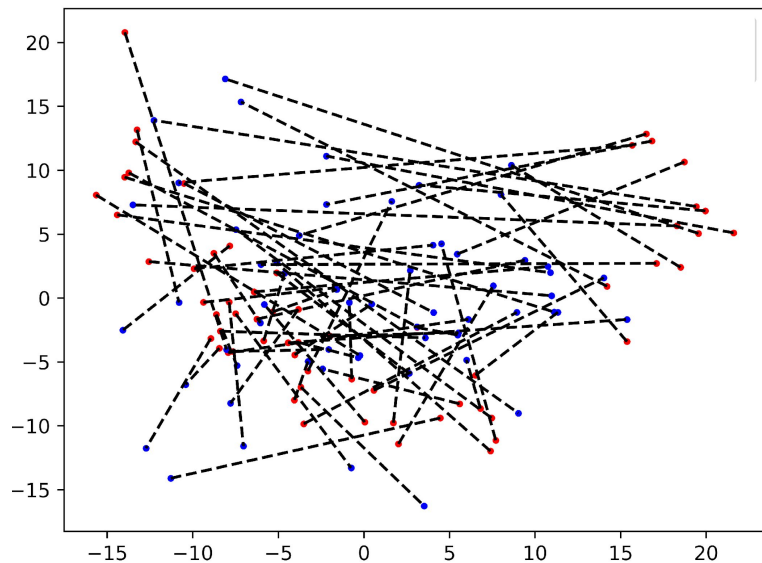


Aligning single-cell RNA-seq and DNA methylation data

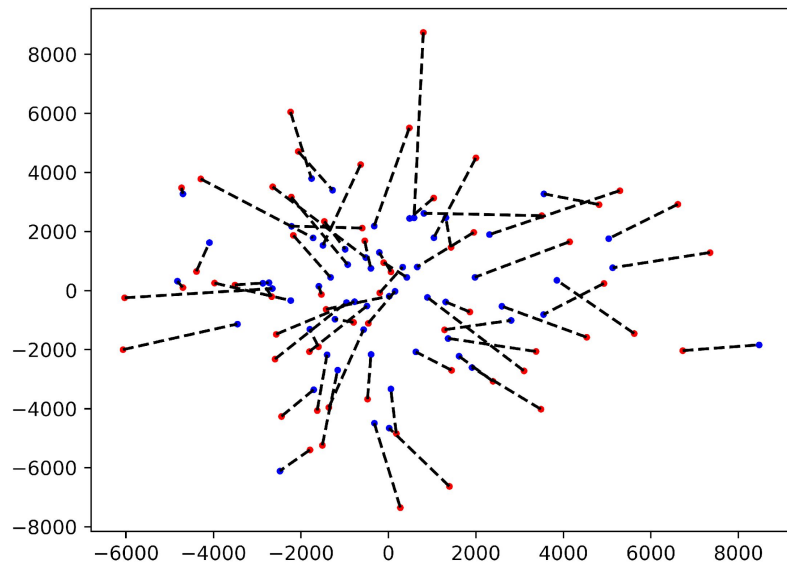


Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Angermueller et al., *Nature Methods* (2016)

MMD-MA aligns single-cell RNAseq and DNA methylation data

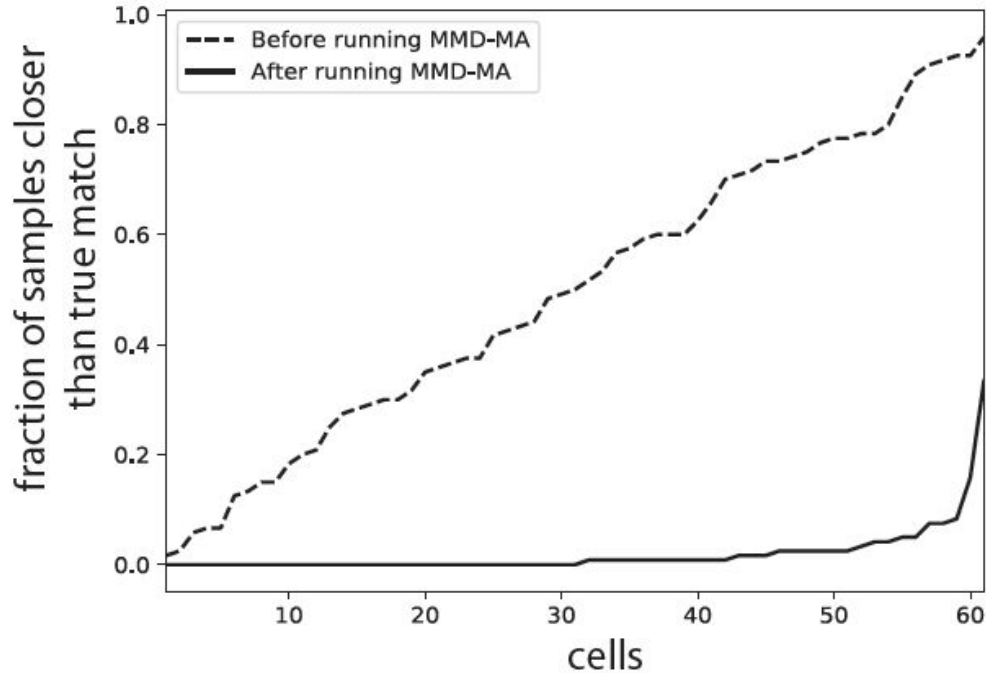


MMD-MA
(5 dimensions)



- Gene Expression
- DNA Methylation

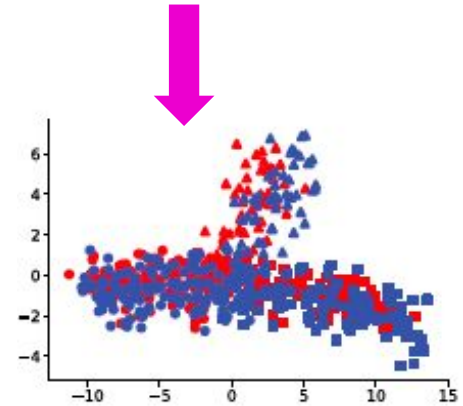
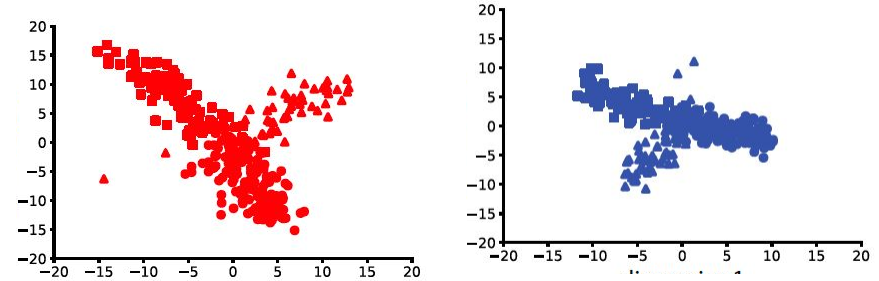
MMD-MA correctly matches cells



- 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

