Network inference from Single-cell data

Jean-Philippe Vert
Google Brain / MINES ParisTech

Plan

Network inference from bulk data

Network inference from single-cell data

Challenges and opportunities

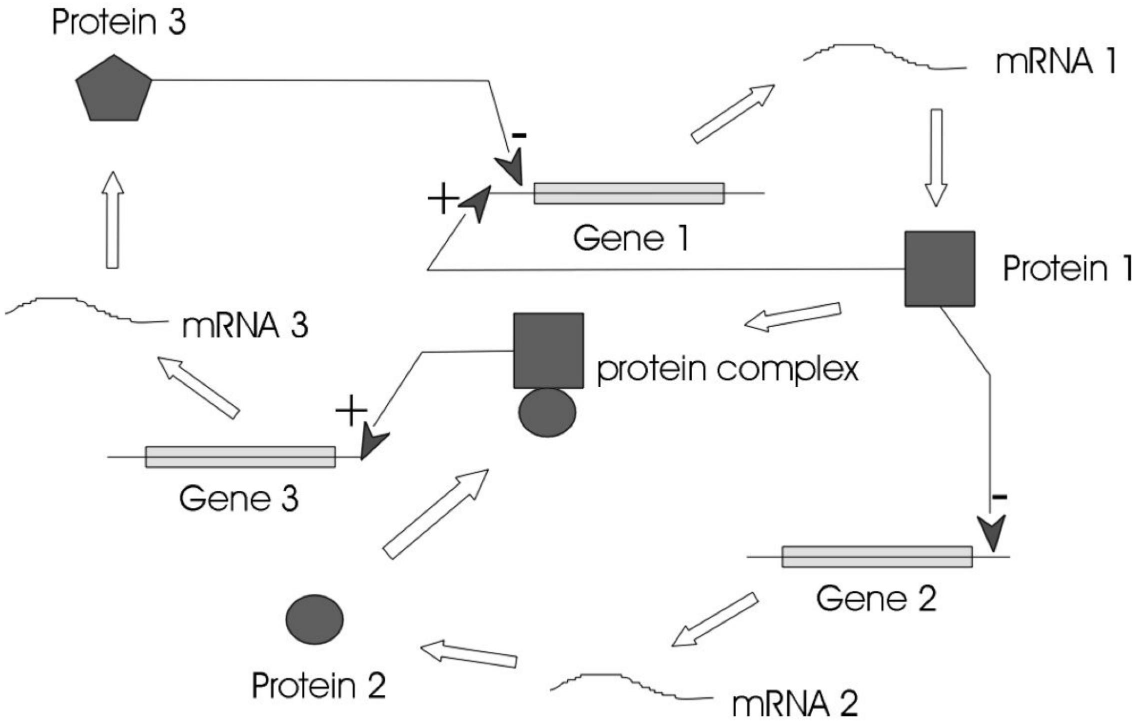
Plan

Network inference from bulk data

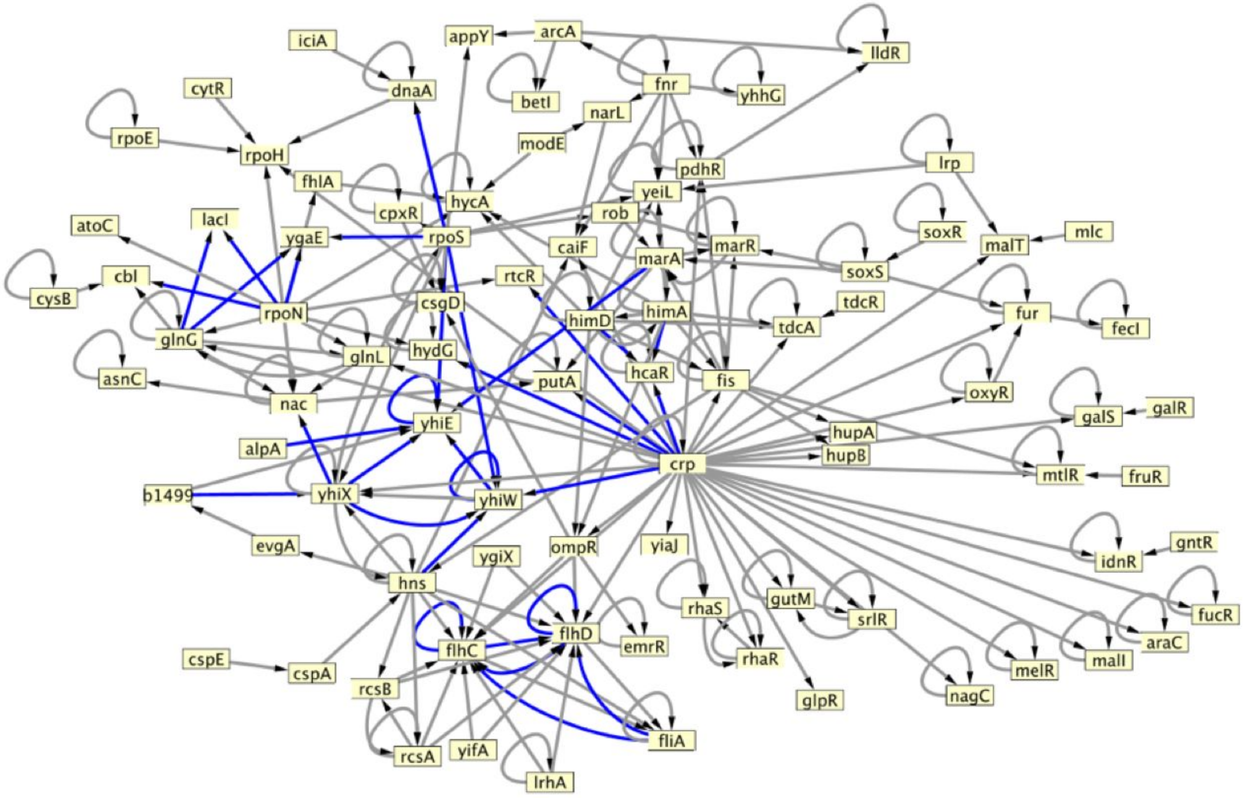
Network inference from single-cell data

Challenges and opportunities

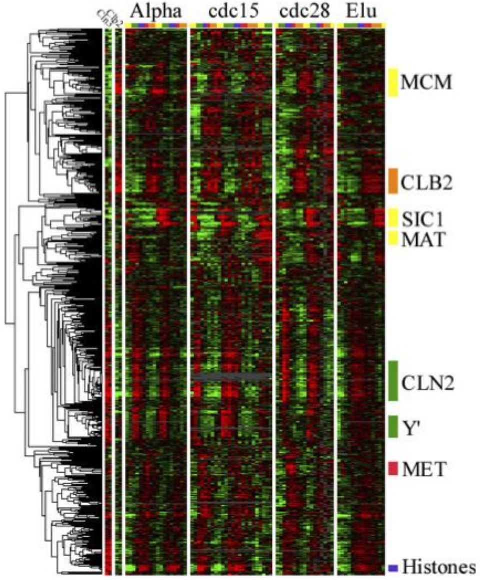
Gene regulatory network (GRN)

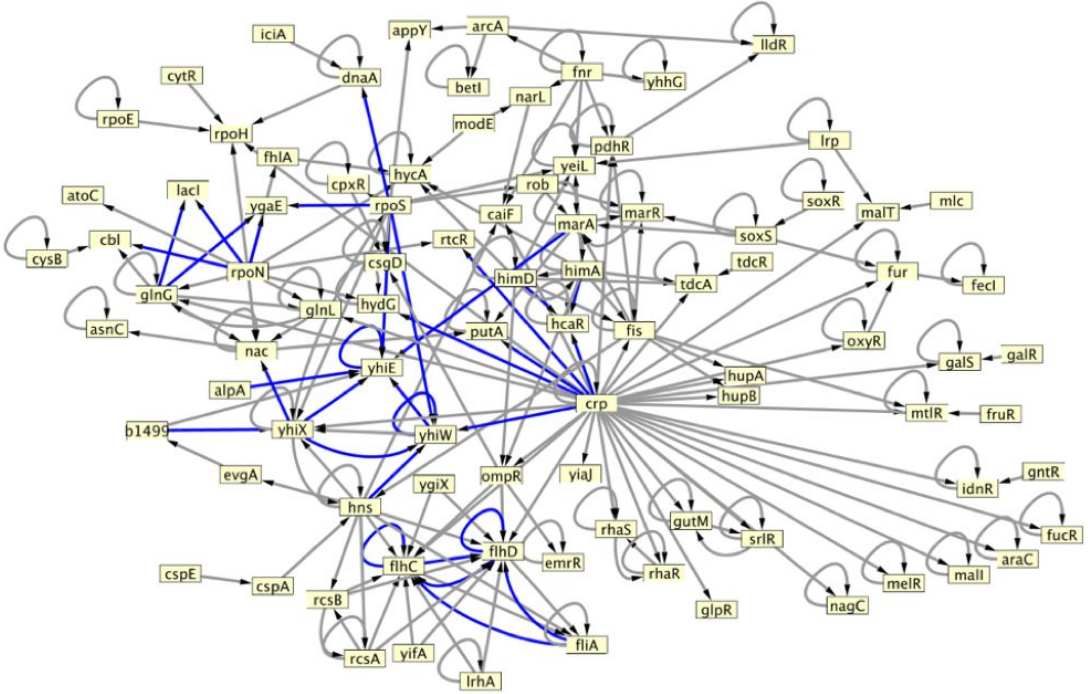


GRN of E. coli


iciA
appY
arcA
lldR
dnaA
betI
fnr
yhhG
cytR
narL
rpoE
modE
rpoH
pdhR
lrp
fhlA
yeiL
hycA
cpxR
rob
atoC
lacI
ygaE
rpoS
caiF
marR
soxR
mlc
marA
malT
soxS
cbl
rtcR
cysB
csgD
tdcR
himA
himD
fur
rpoN
tdcA
fecI
glnG
hydG
glnL
hcaR
asnC
putA
fis
oxyR
nac
galS
galR
yhiE
hupA
alpA
hupB
crp
mtlR
fruR
b1499
yhiX
yhiW
evgA
ygiX
ompR
yiaJ
idnR
gntR
hns
rhaS
gutM
srlR
fucR
flhD
flhC
emrR
araC
cspE
rhaR
malI
cspA
melR
rcsB
glpR
nagC
fliA
rcsA
yifA
lrhA

GRN inference from bulk expression data



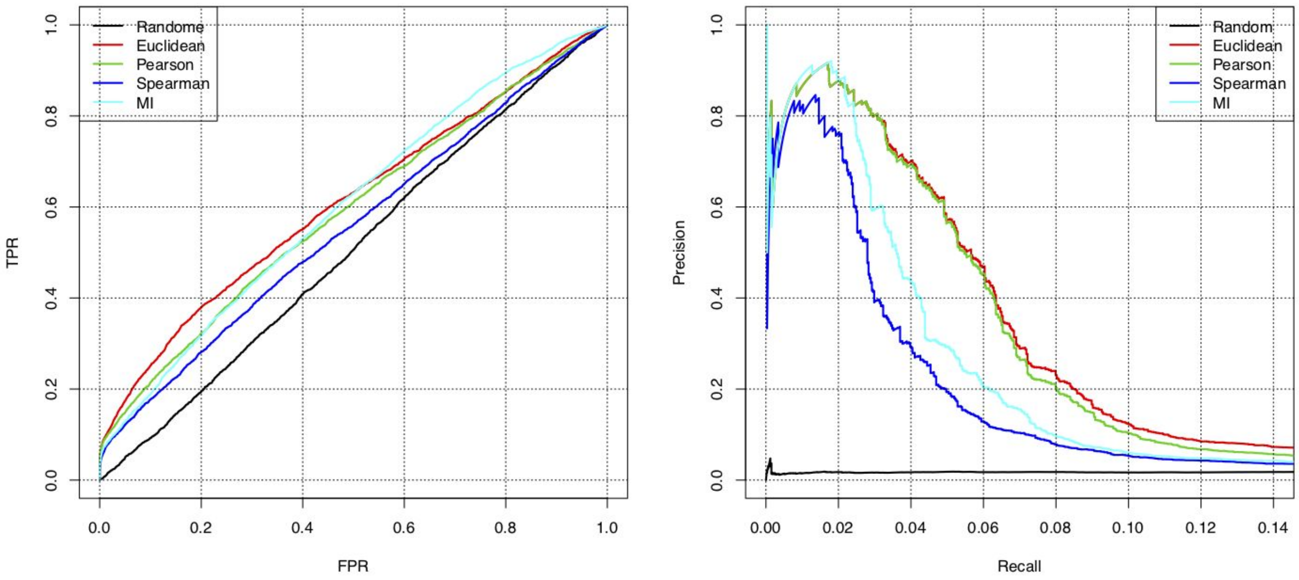


Inference principles

- Connect “similar” genes
  - Co-expression, correlation, mutual information...
- Causal inference
  - Bayesian network, causal networks...
- Sparse regression
  - Random forests, lasso..

Example: co-expression inference

Application: E coli regulatory network : 154 TF targeting 1164 genes through 3293 regulations



Steady-state hypothesis

- 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.

GRN inference by sparse regression

- Treat each target in turn
- Let Y the expression of a target, and $X_1, \ldots, X_p$ the expression of all TFs. We look for a model

$$Y = \sum_{i=1}^{p} \beta_i X_i + \text{noise}$$

  where β is sparse, i.e., only a few β_i are non-zero
- Examples:
  - GENIE: feature selection by random forest (Huynh-Thu et al., 2010)
  - Feature selection by Lasso + stability selection (Haury et al., 2011)
- Both methods were ranked 1st and 2nd (out of 28) at the DREAM5 in silico network inference challenge

Plan

Network inference from bulk data

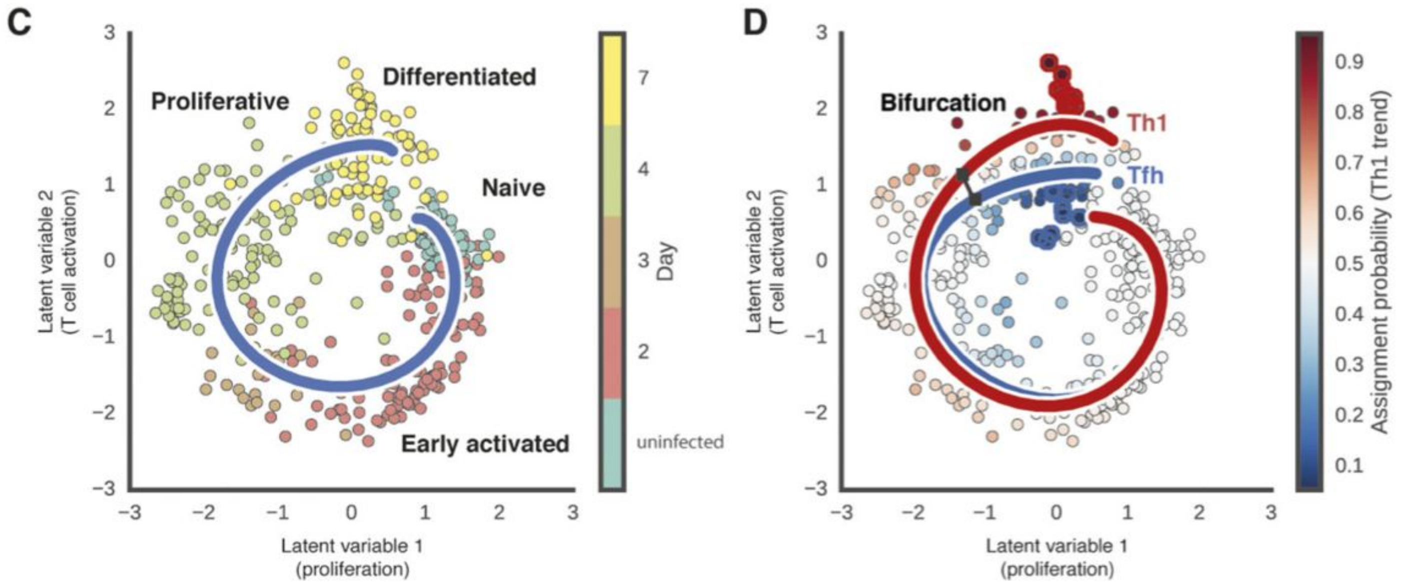
Network inference from single-cell data

Challenges and opportunities

From bulk to single-cell

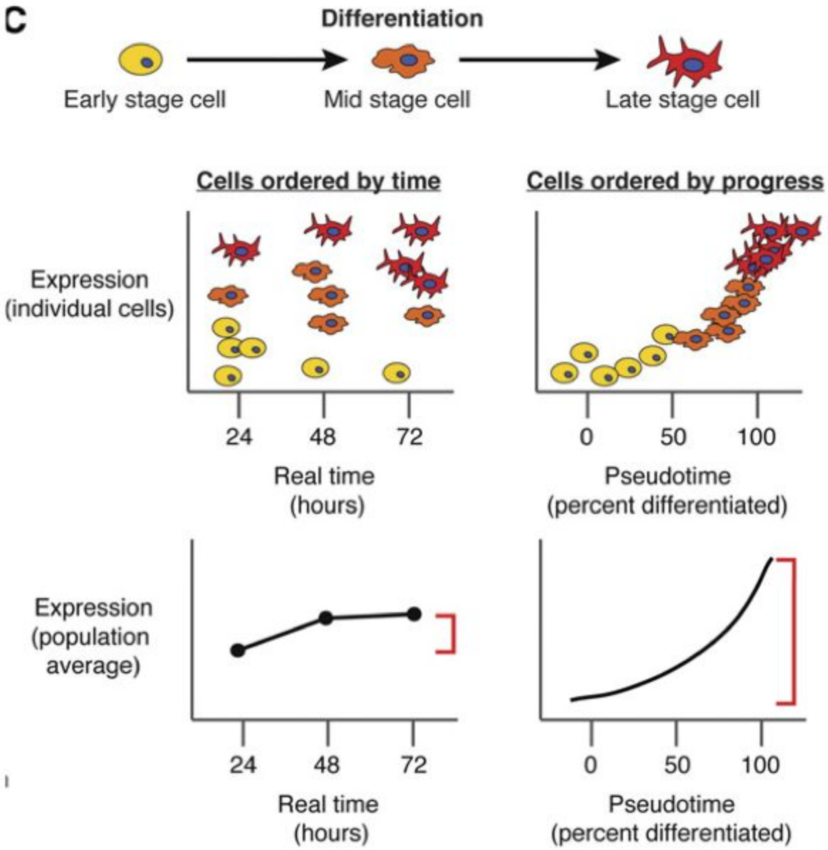
Inspired from slides of A. Regev

Steady-state?



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

$$dX/dt = A*X$$

- Given cells (X_i, t_i) for i=1,...,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

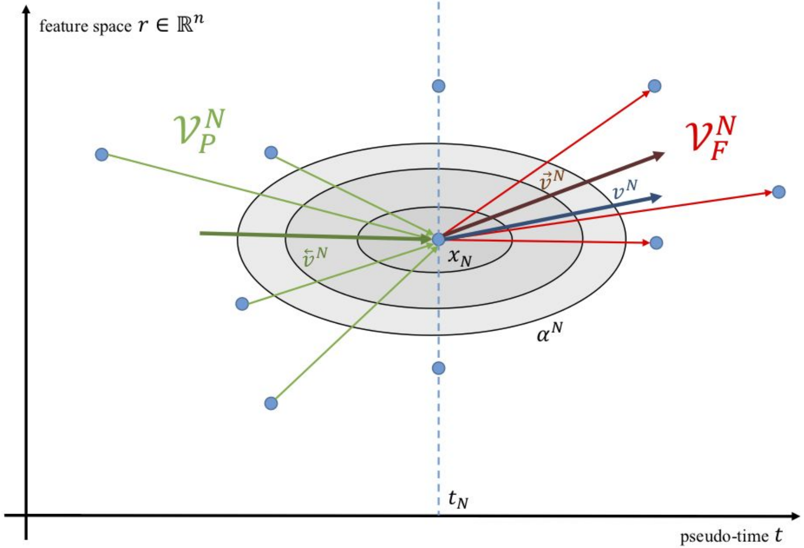
GRISLI (Aubin and V., 2018)

- 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’_ti of each cell

Velocity inference



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

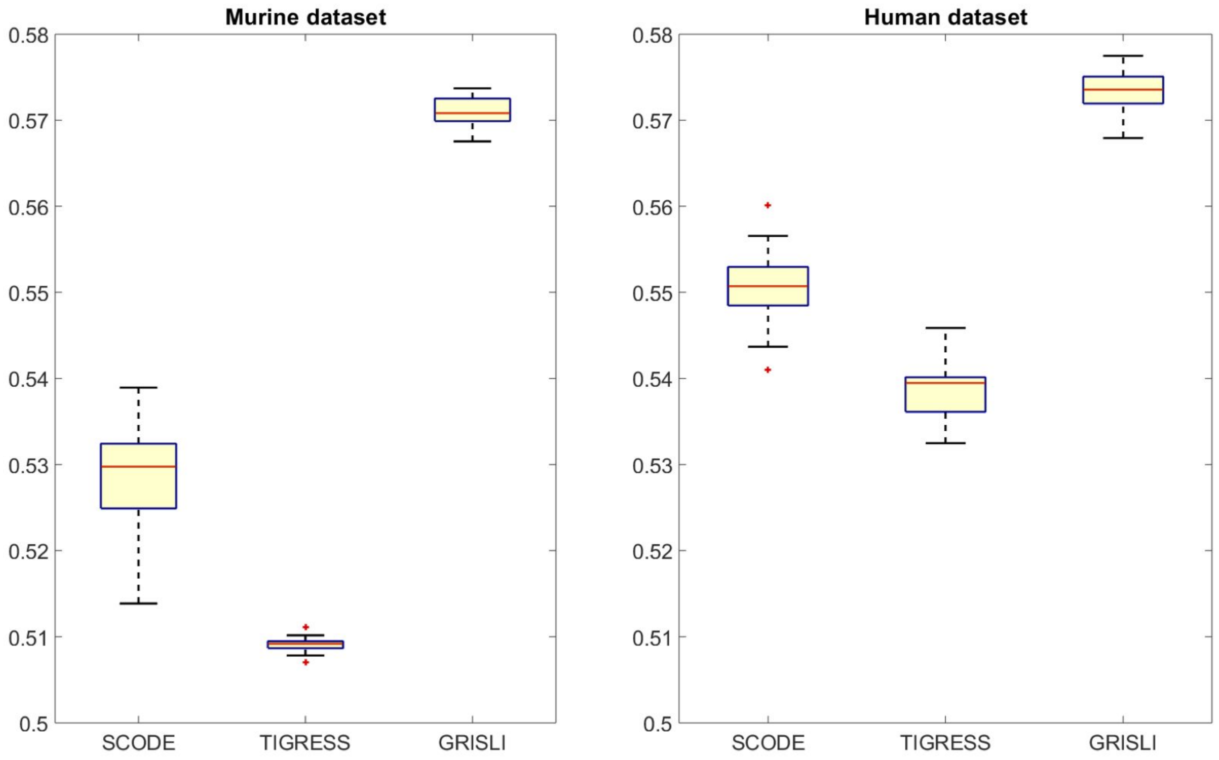
$$K(x,t,x',t') = (t-t')^2 \exp\left(-\frac{(t-t')^2}{2\sigma_t^2}\right) \times \exp\left(-\frac{\|x-x'\|^2_{\mathbb{R}^G}}{2\sigma_x^2}\right)$$

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

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)



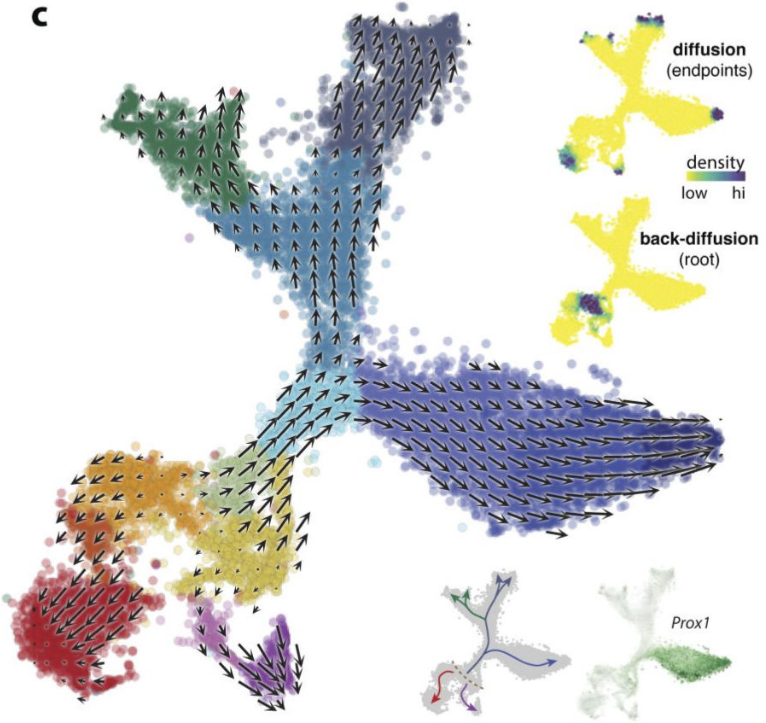
Plan

Network inference from bulk data

Network inference from single-cell data

Challenges and opportunities

Velocity inference



Cell

CellPress

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

Resource

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

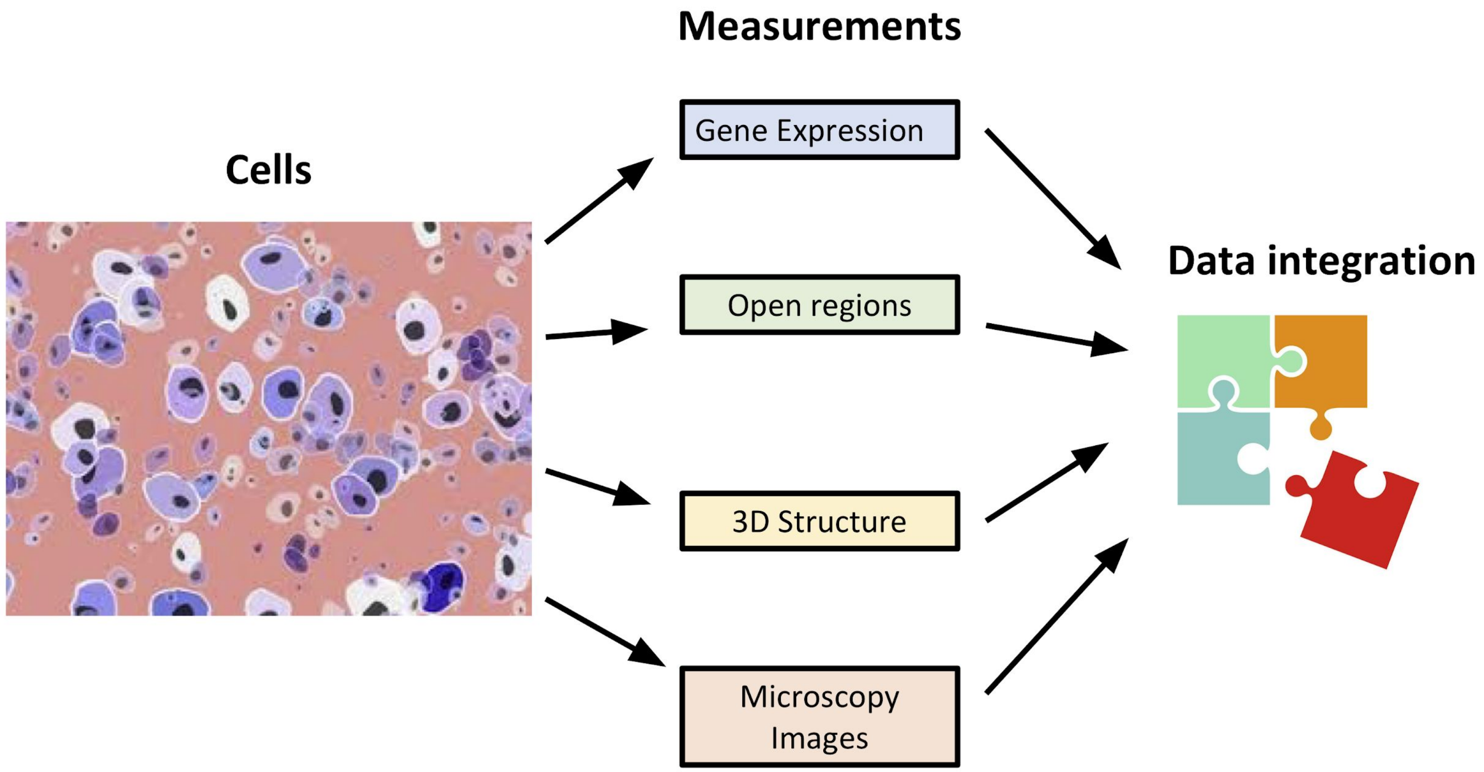
Geoffrey Schiebinger [1, 11, 16], Jian Shu [1, 2, 16], Marcin Tabaka [1, 16], Brian Cleary [1, 3, 16], Vidya Subramanian [1], Aryeh Solomon [1, 17], Joshua Gould [1], Siyan Liu [1, 15], Stacie Lin [1, 6], Peter Berube [1], Lia Lee [1], Jenny Chen [1, 4], Justin Brumbaugh [5, 7, 8, 9, 10], Philippe Rigollet [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

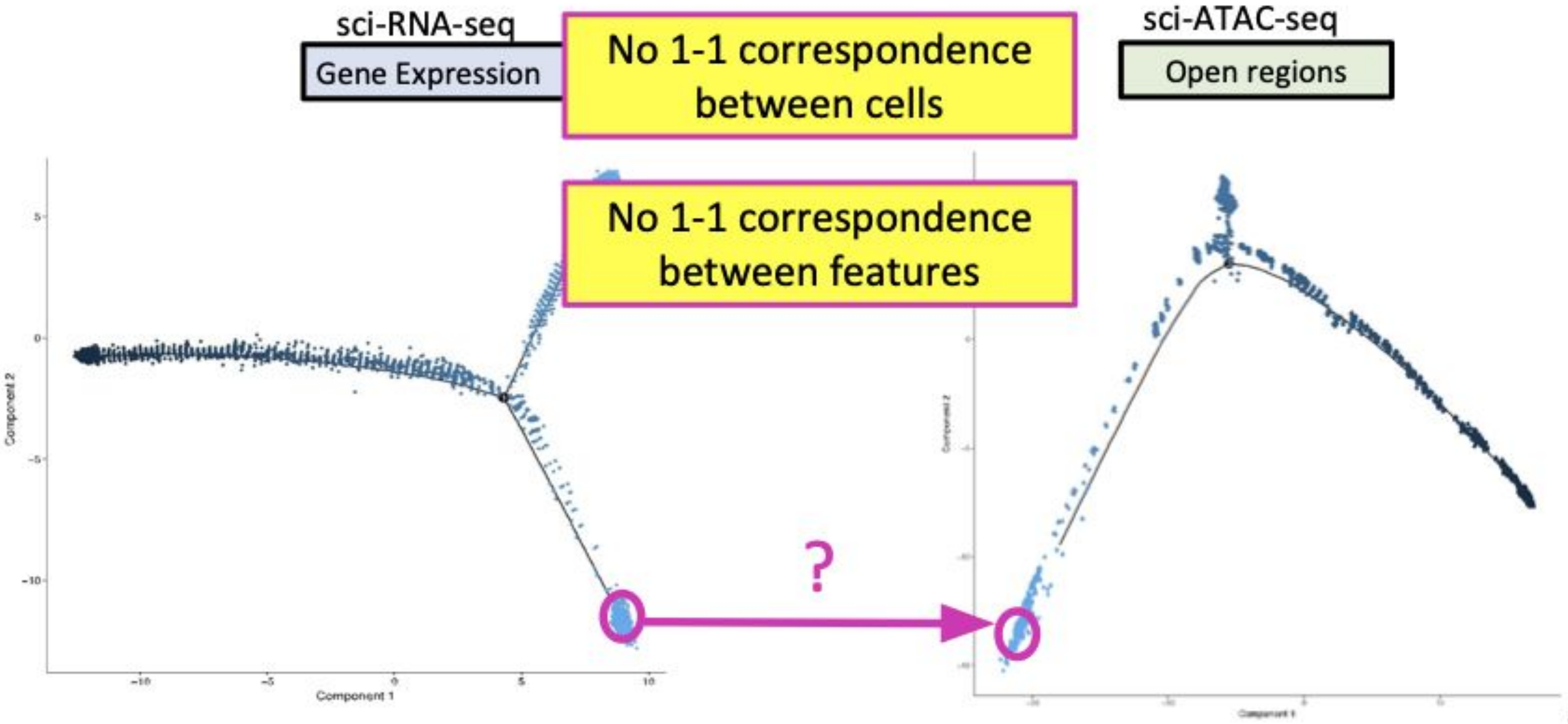
Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco,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

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

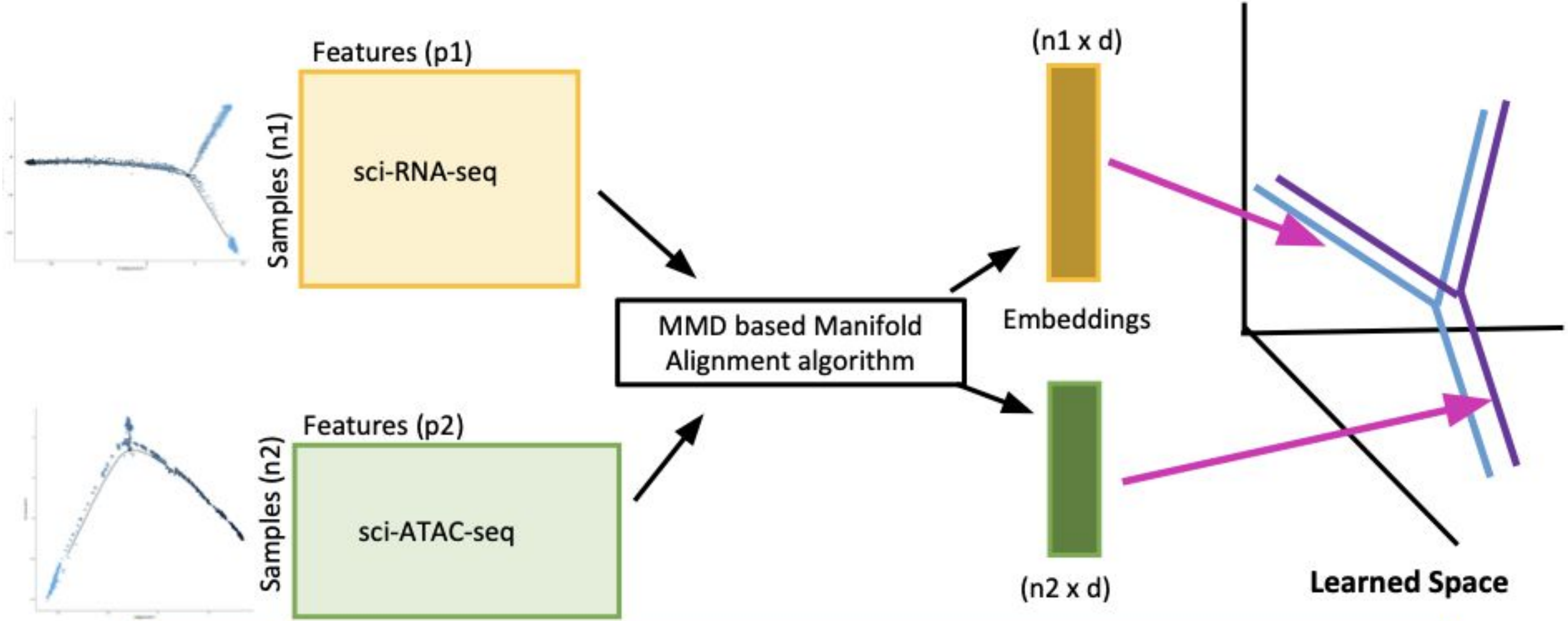
Data integration



Integration of single-cell data is challenging



Learning a shared “representation”



Assumption: Data shares a common manifold structure

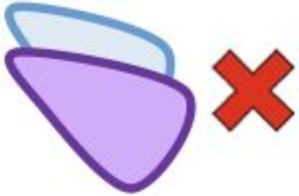
MMD-MA

New Results

Jointly embedding multiple single-cell omics measurements

Jie Liu, Yuanhao Huang, Ritambhara Singh, Jean-Philippe Vert, William Stafford Noble

doi: https://doi.org/10.1101/644310



sci-RNA-seq

sci-ATAC-seq

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

Parameters learned during training

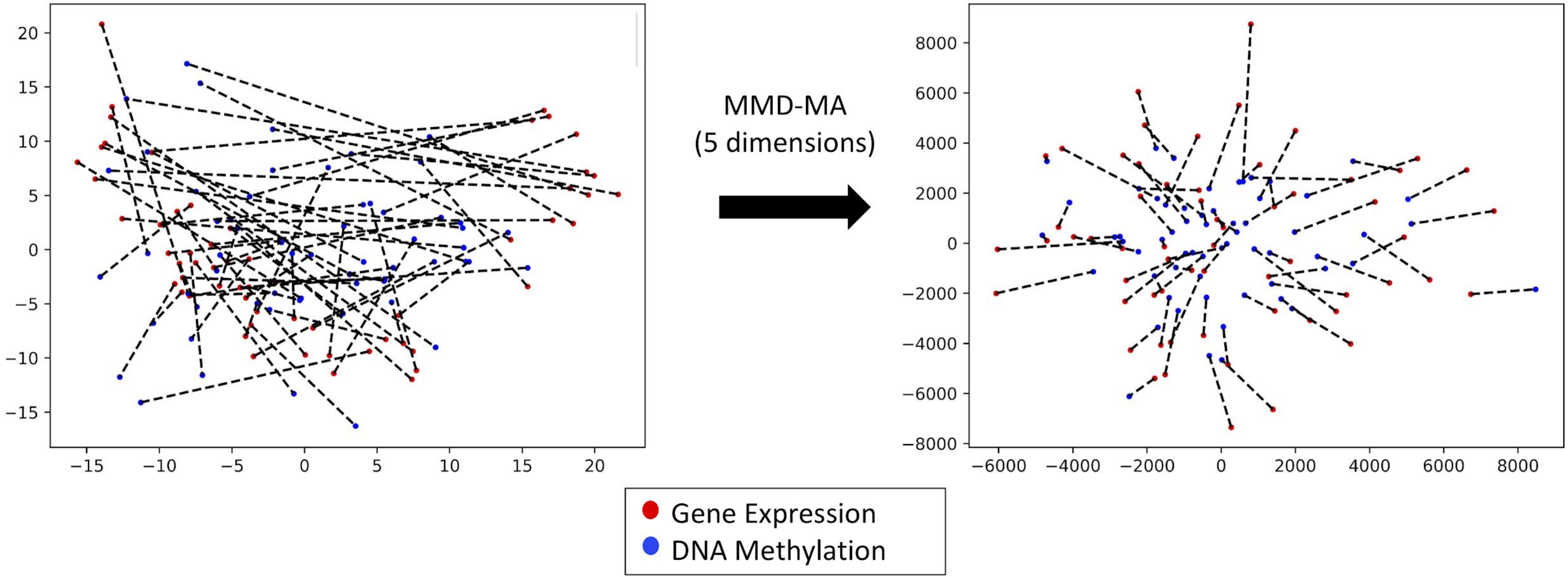
$$\boxed{pen(\alpha_I) = \left\|\alpha_I^T K_I^T \alpha_I - I_D\right\|_2}$$

Penalty term to avoid trivial solution

$$\boxed{distortion(\alpha_I) = \left\|K_I - K_I \alpha_I \alpha_I^T K_I^T\right\|_2^2}$$

Distortion term to preserve structure

Alignment of single cell expression and methylation



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

Conclusion: many opportunities and challenges!

Single-Cell Multiomics:
Multiple Measurements from
Single Cells

Iain C. Macaulay,[1,*] Chris P. Ponting,[2,3,*] and Thierry Voet[2,4,*]