Imputation of scRNA-seq data

David van Dijk

Internal Medicine & Computer Science @ Yale

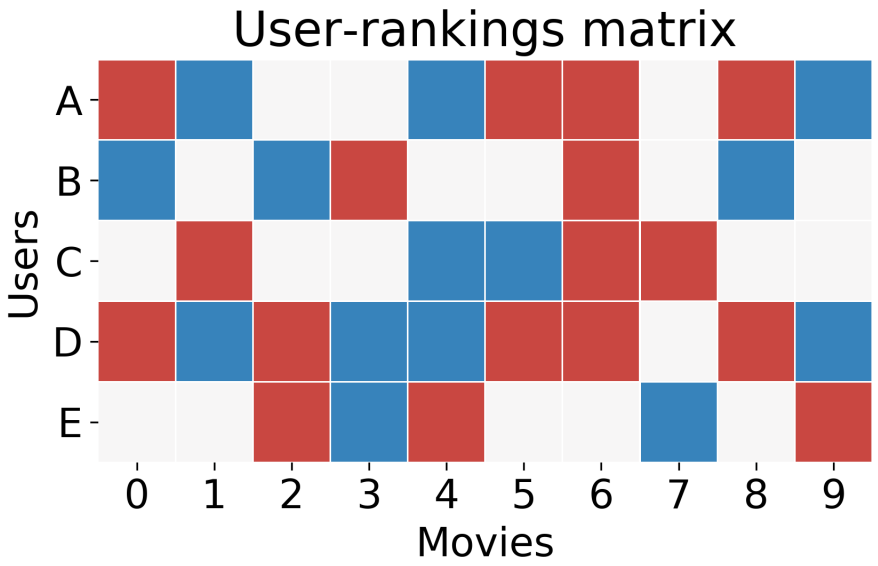
vandijklab.org

What do we do with missing data?

The Netflix Problem

- Rows are users
- Columns are “+1” for like and “-1” for dislike
- We know which values are missing (unwatched)

- Goal: predict missing values to suggest movies to each user

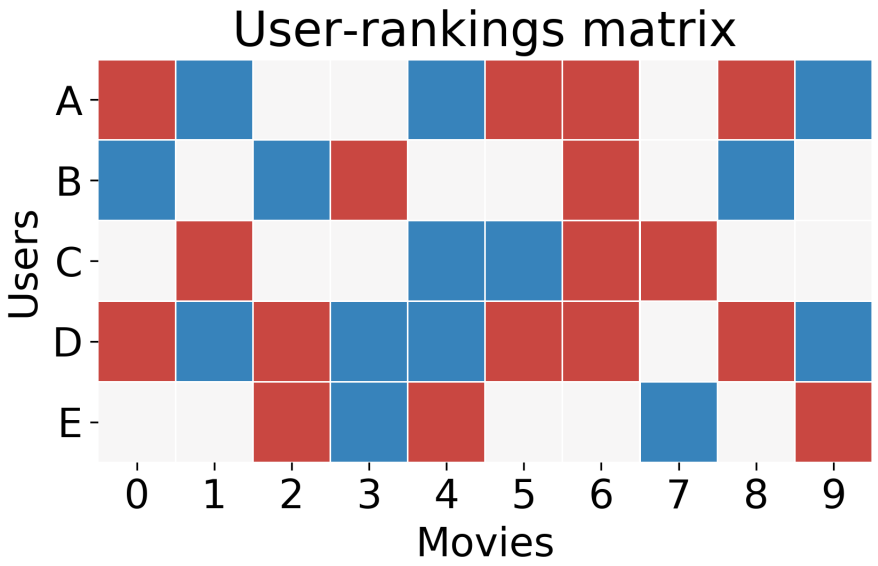


What do we do with missing data?

The Netflix Problem

- Rows are users
- Columns are “+1” for like and “-1” for dislike
- We know which values are missing (unwatched)

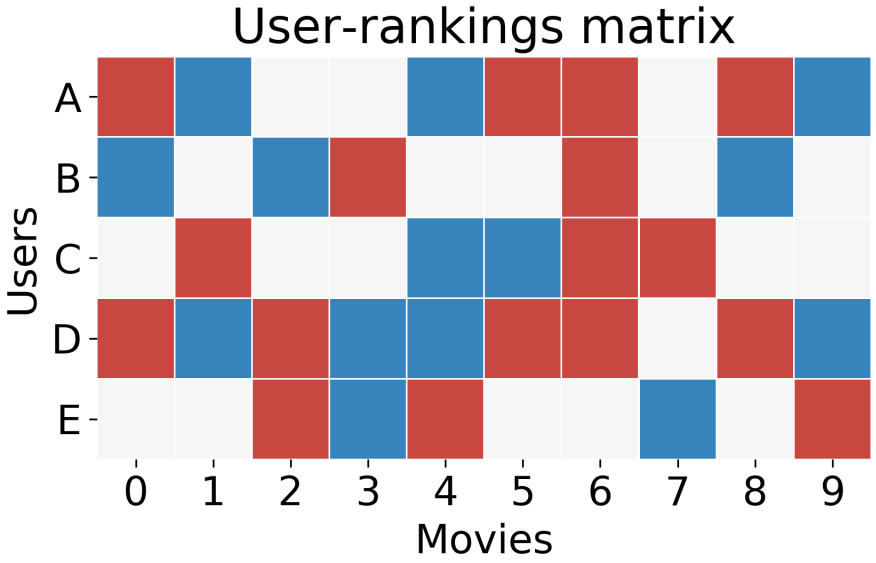
- Goal: predict missing values to suggest movies to each user



How can we guess missing values?

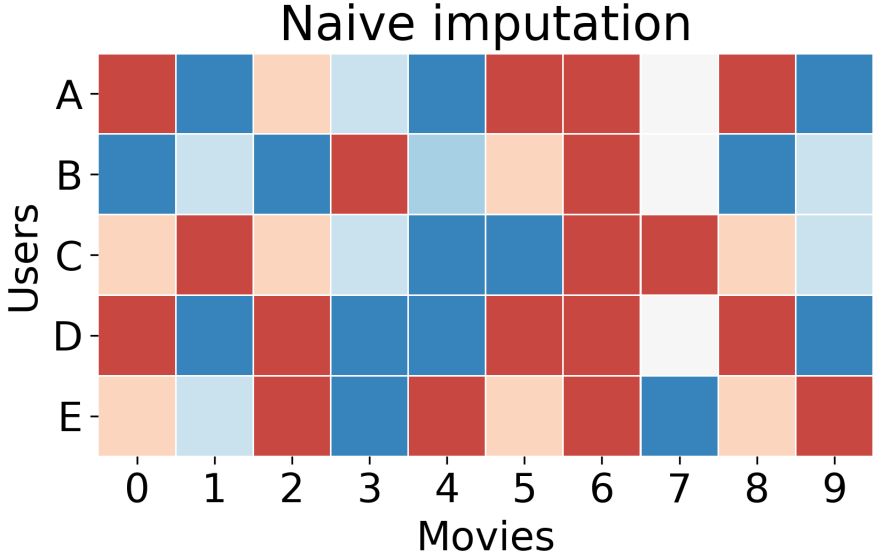
What do we do with missing data?

Raw data



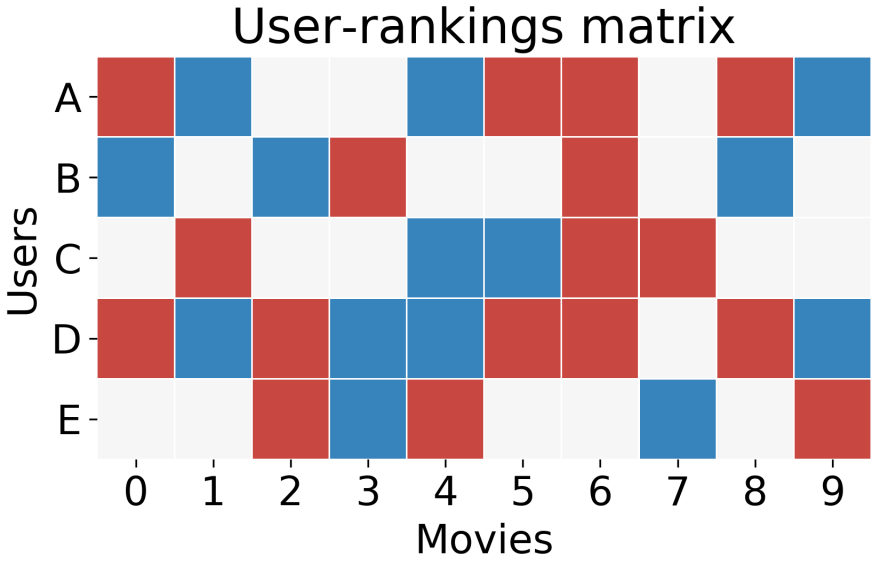
Naïve solution

Use the average ranking



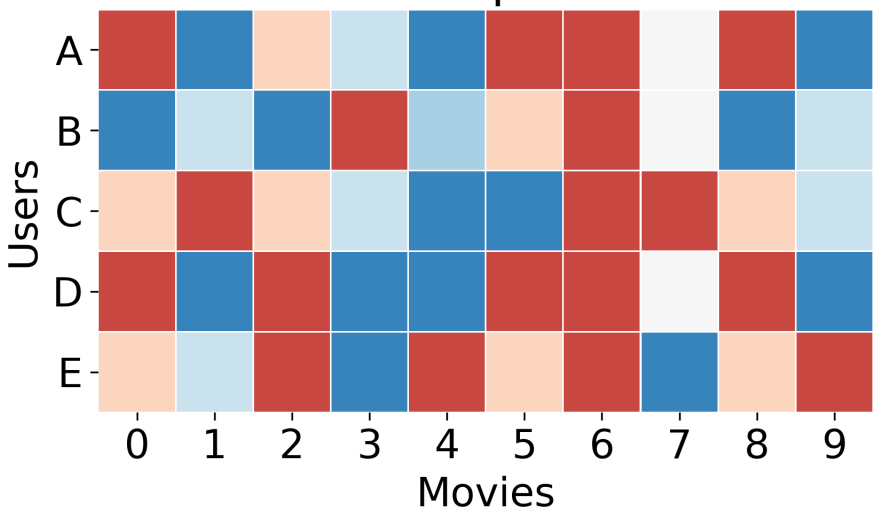
What do we do with missing data?

Raw data



Naïve solution

Use the average ranking

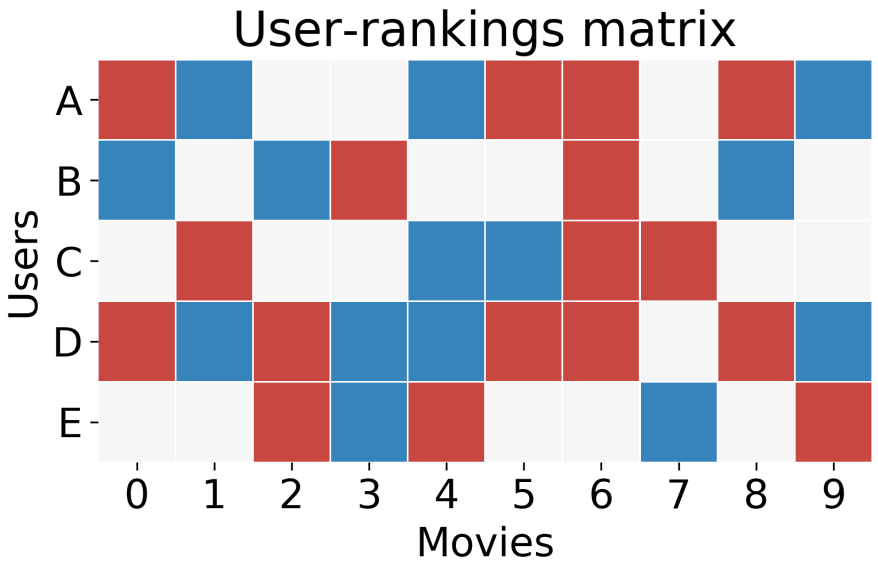


How can we improve on this?

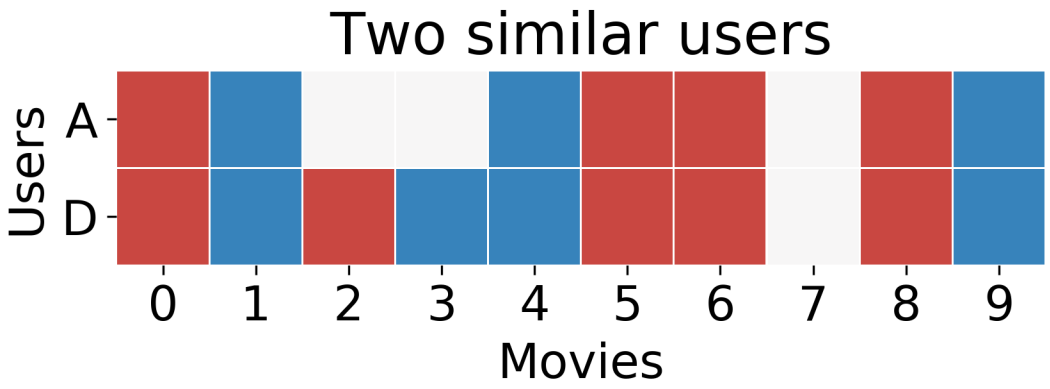
Information redundancy in the user-rankings matrix

Two sources of redundancy

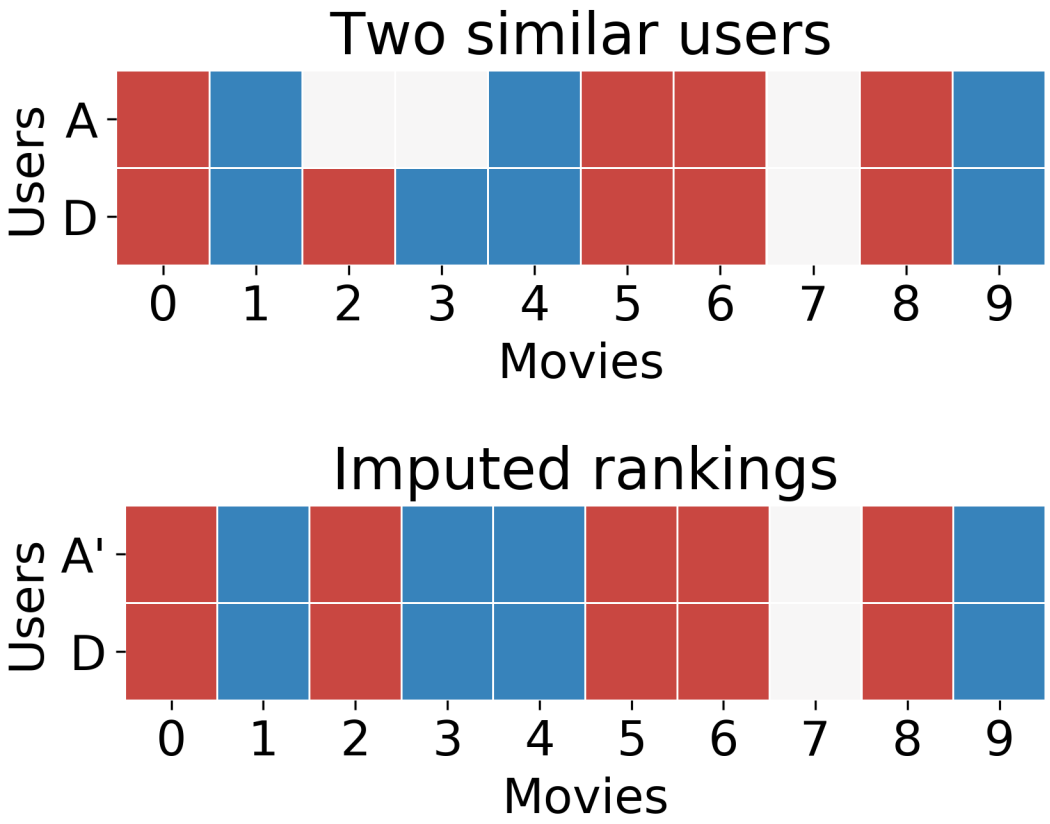
- Similarity between movies (e.g. rom-coms, horror movies)
- Similarity between users (e.g. “Likers of rom-coms”)



Identifying similar users to impute missing values



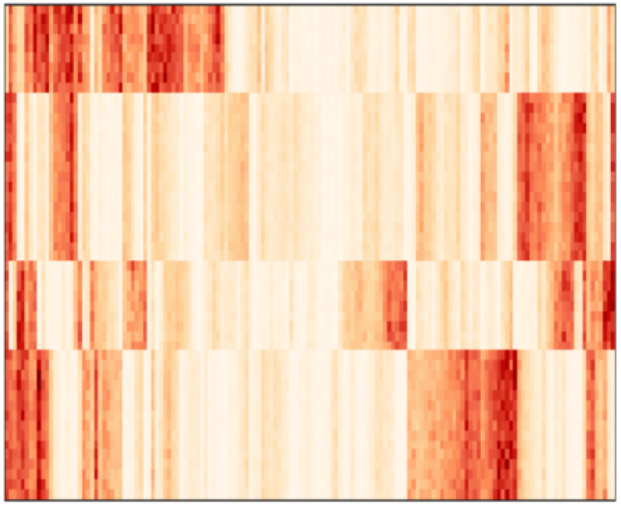
Identifying similar users to impute missing values



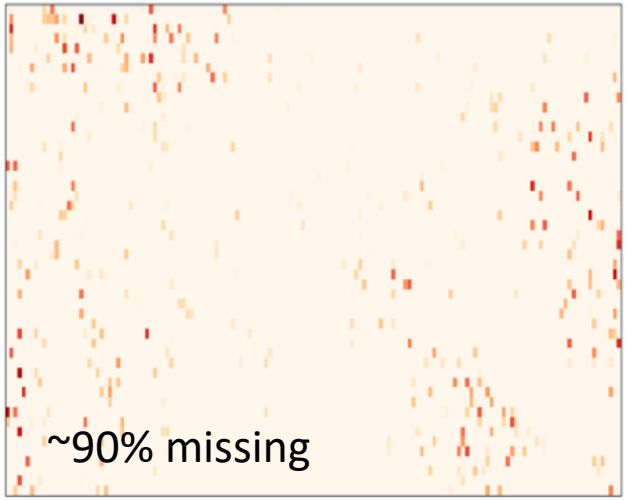
Problem: scRNA-seq data is sparse and noisy

How do we we know which values to impute?

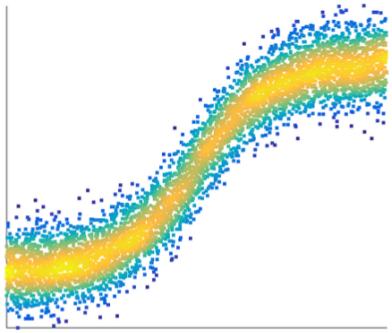
True structure



Noisy measurement



We only capture ~5-15% of the transcripts per cell due to inefficient RT in small volumes

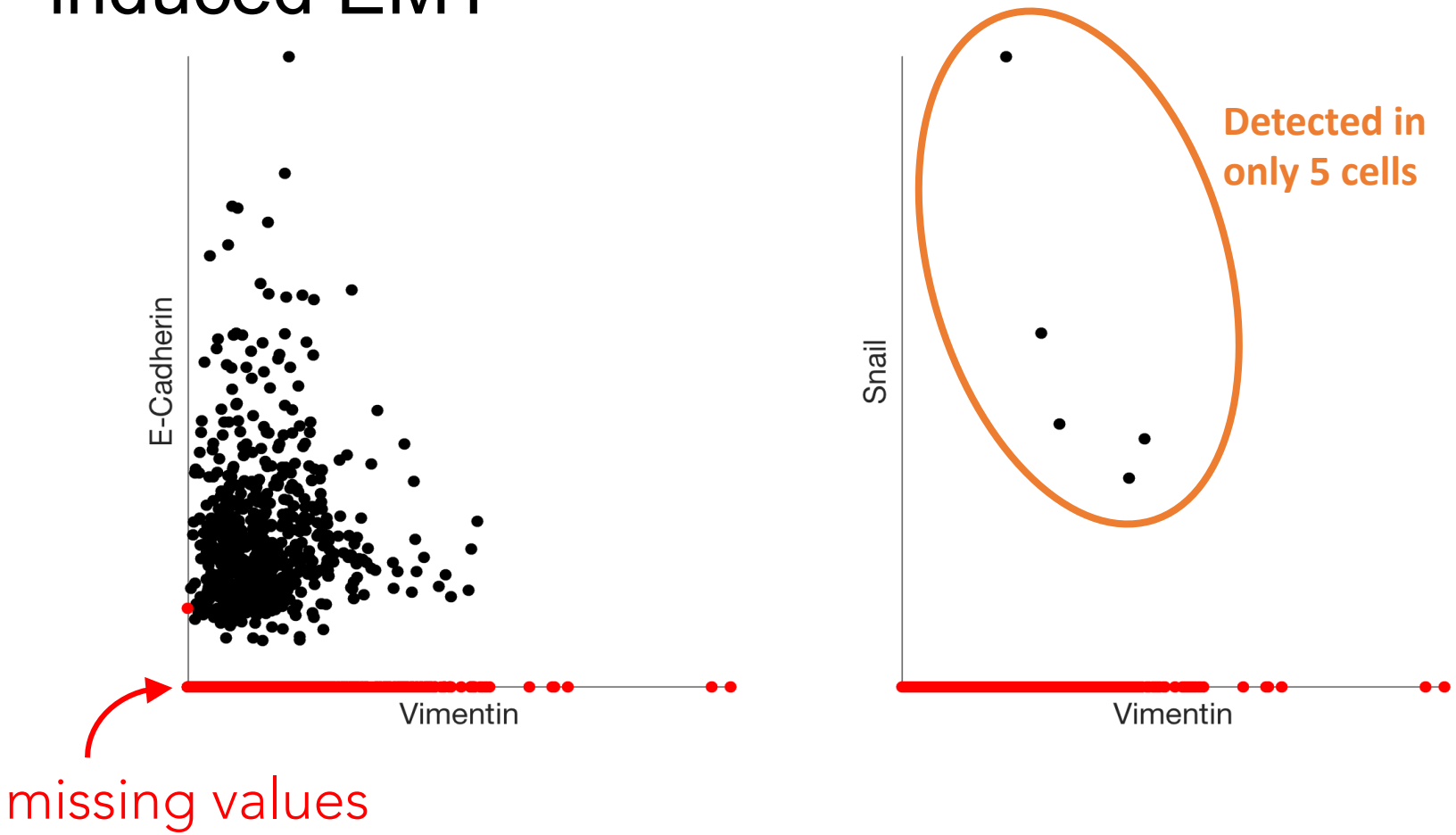


MAGIC

Markov Affinity-based Graph Imputation of Cells
(van Dijk et al. *Cell*, 2018)

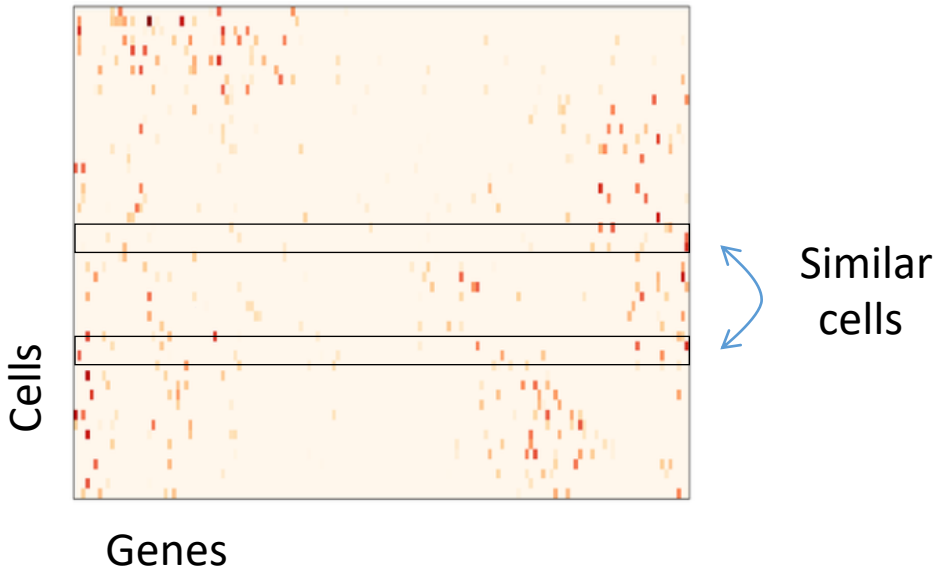
Main idea: cells learn values from their neighbors, gaining information from similar cells while retaining their individual identity.

HMLE breast cancer cell line, TGF-β induced EMT



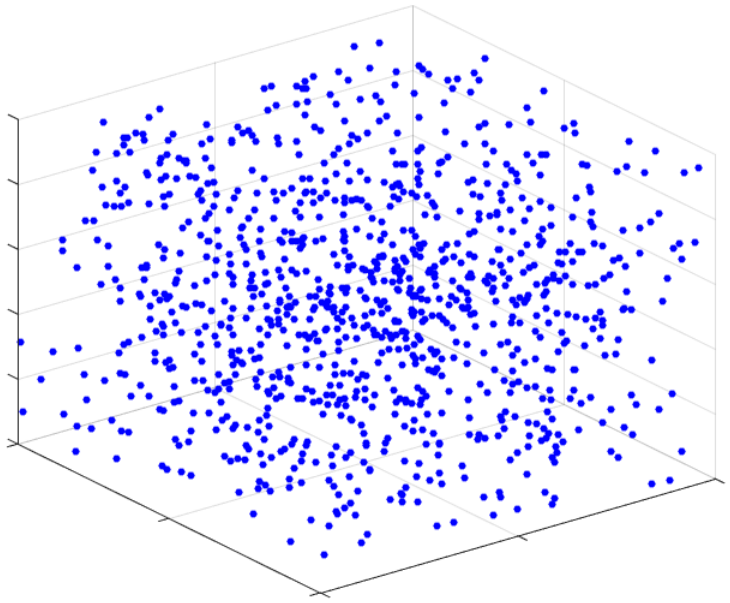
dropout obscures gene-gene relationships

Solution: Recover data by learning from similar cells



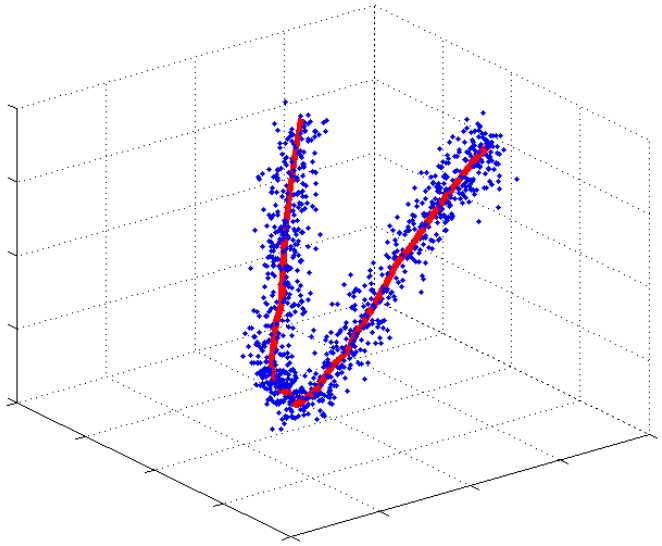
- Biologically similar cells will have similar expression
- Cells “exchange” information

Low dimensional structure



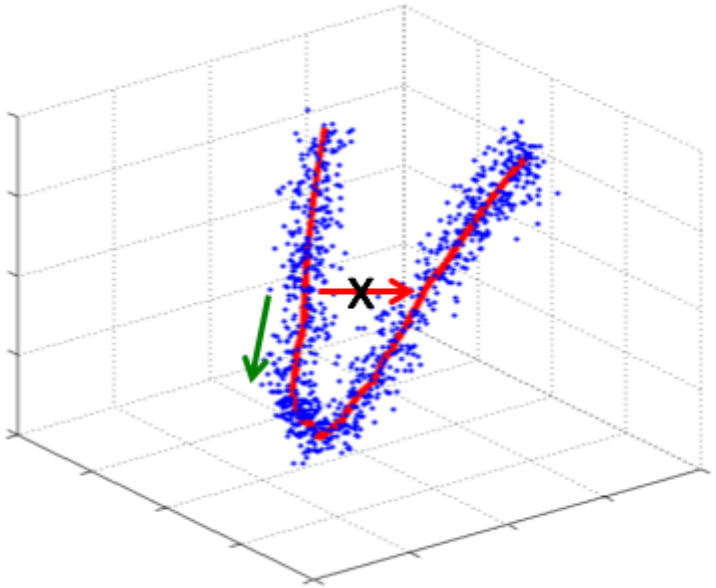
- Many features (e.g. genes) but high degree of redundancy
- Features are correlated
- States (e.g. cell types) are constrained

Manifold

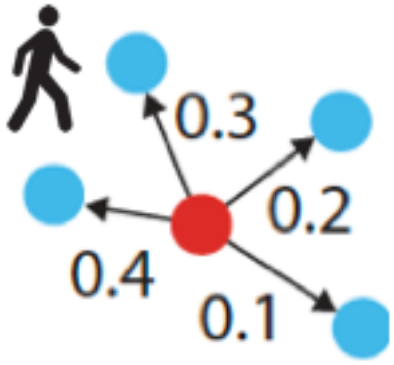


- Locally Euclidean
- Smooth
- Low dimensional
- Data has a “shape”

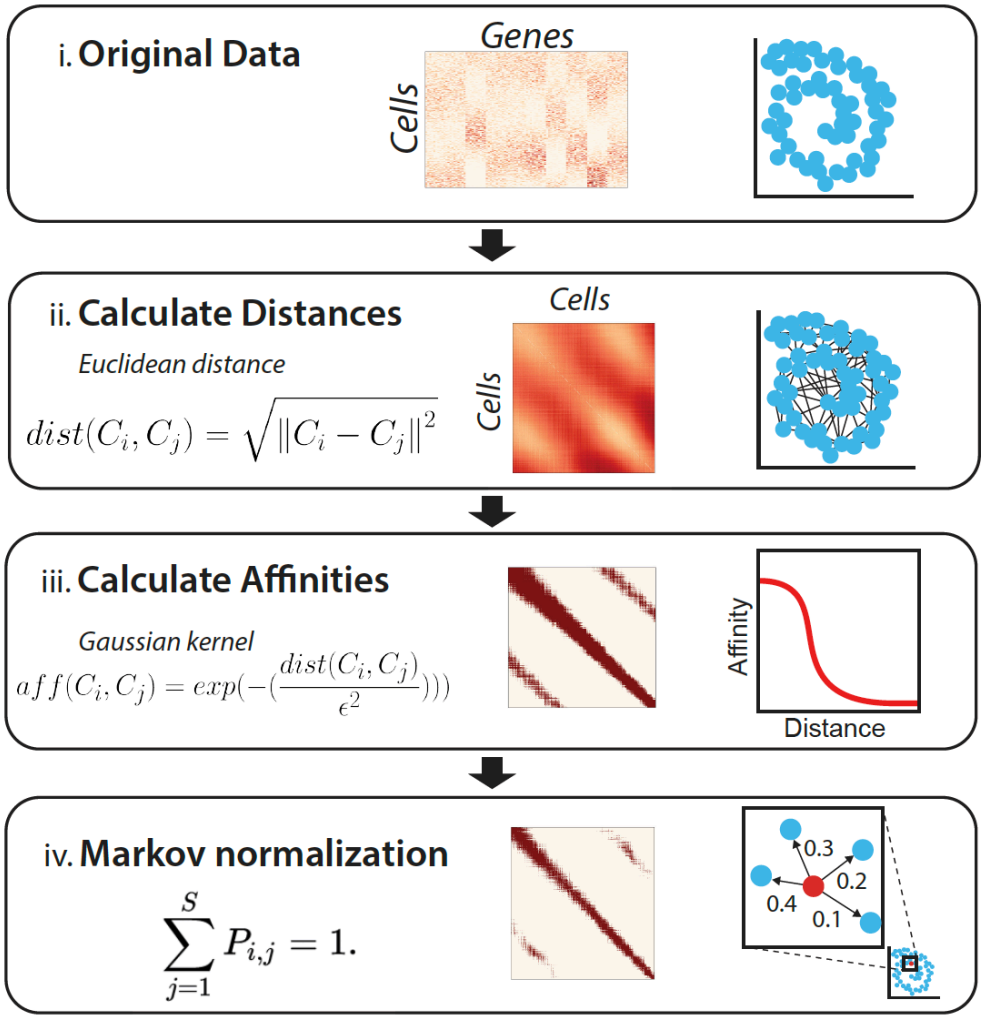
Manifold learning via data diffusion



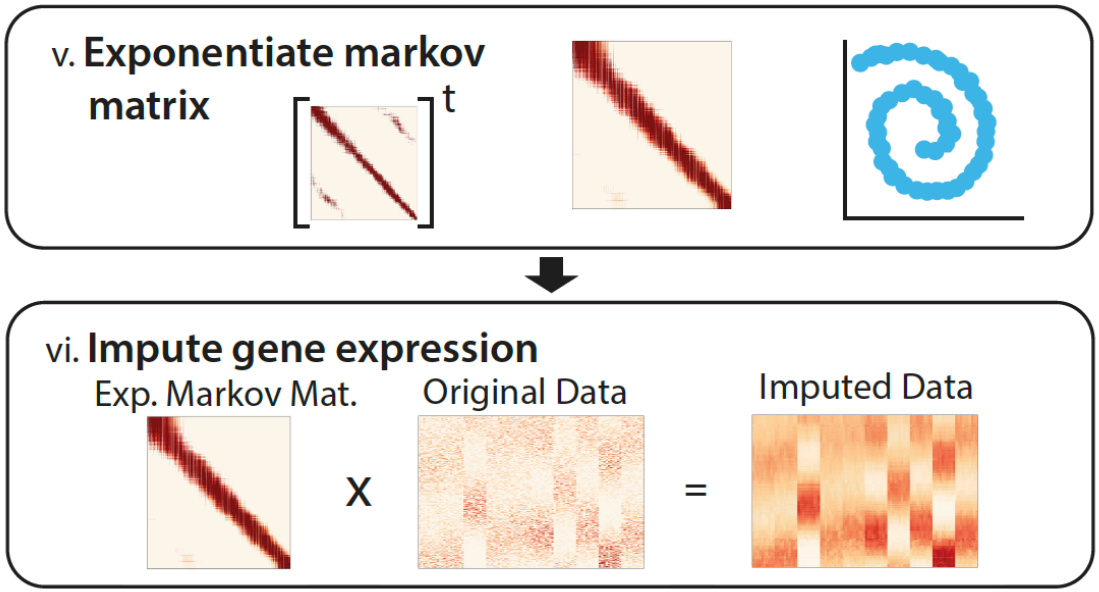
- Biological neighborhood = manifold
- Walk via small local steps
- Random walk = diffusion



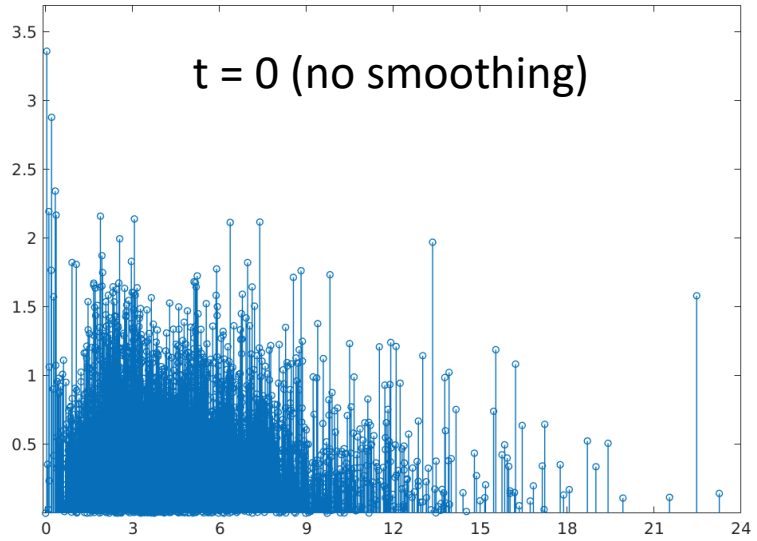
The Algorithm

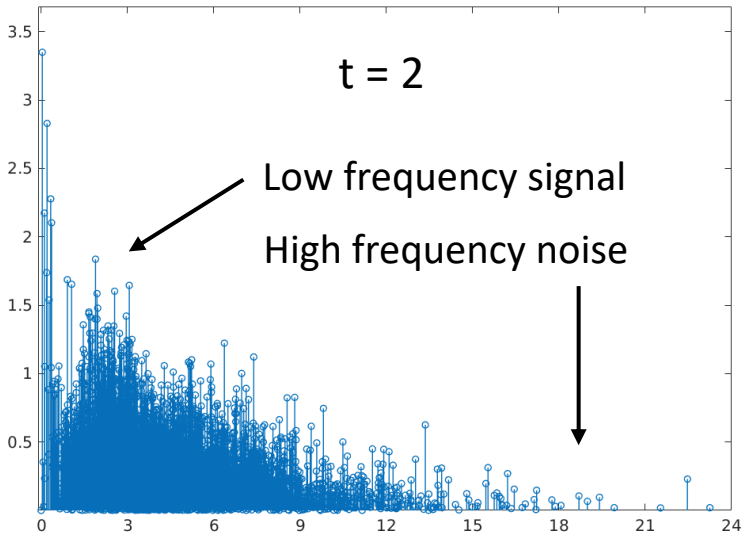


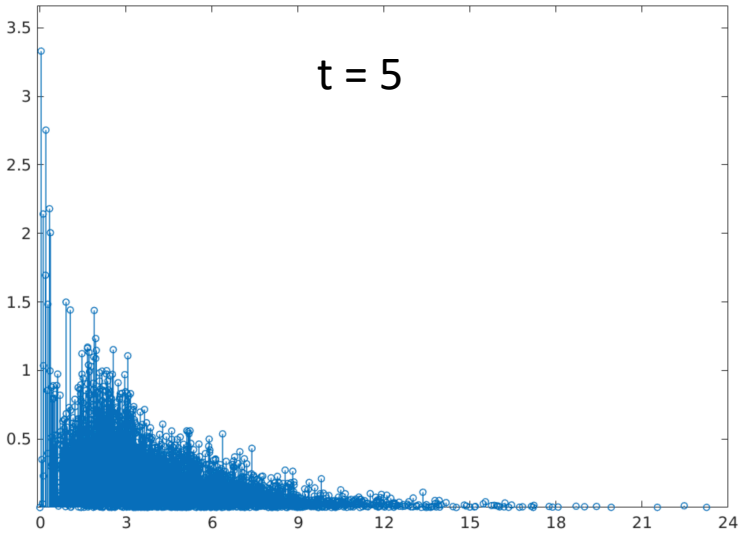
Diffusion and imputation step

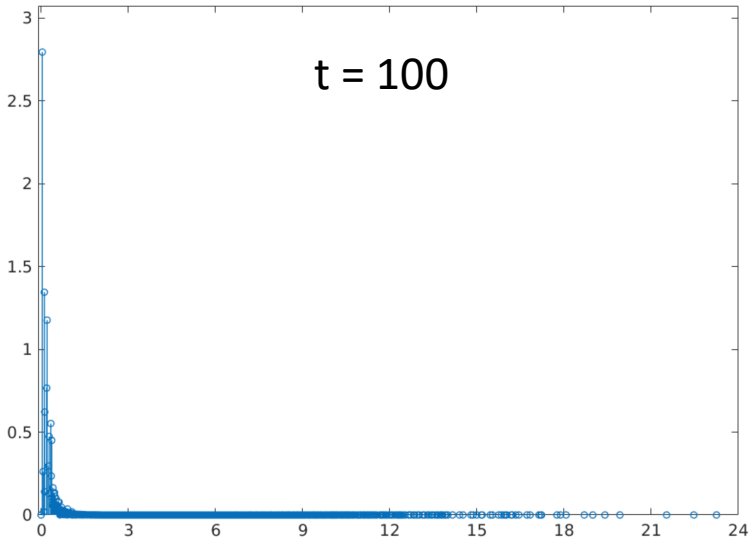


Graph diffusion = low pass filter on graph spectrum

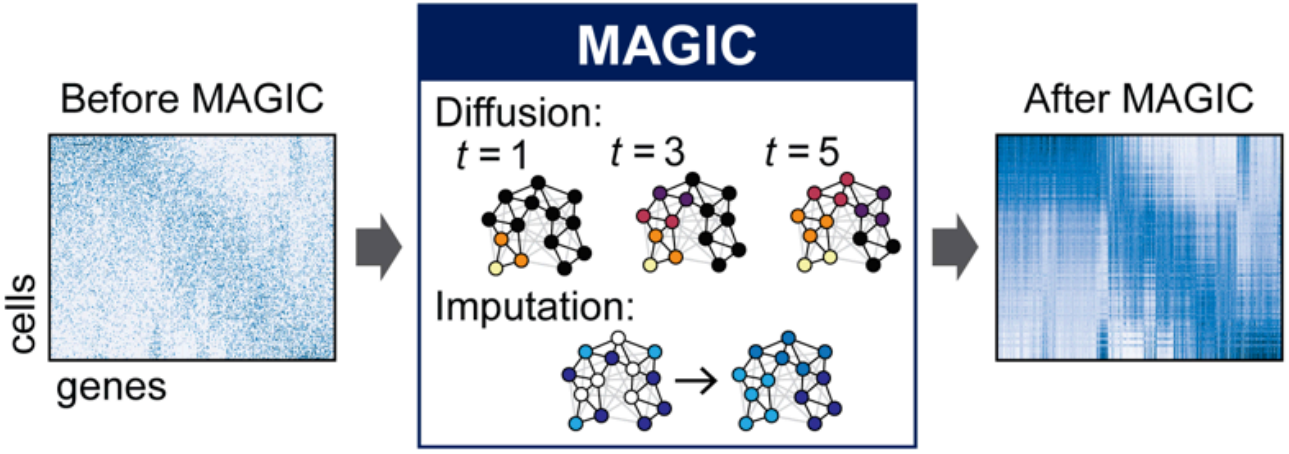


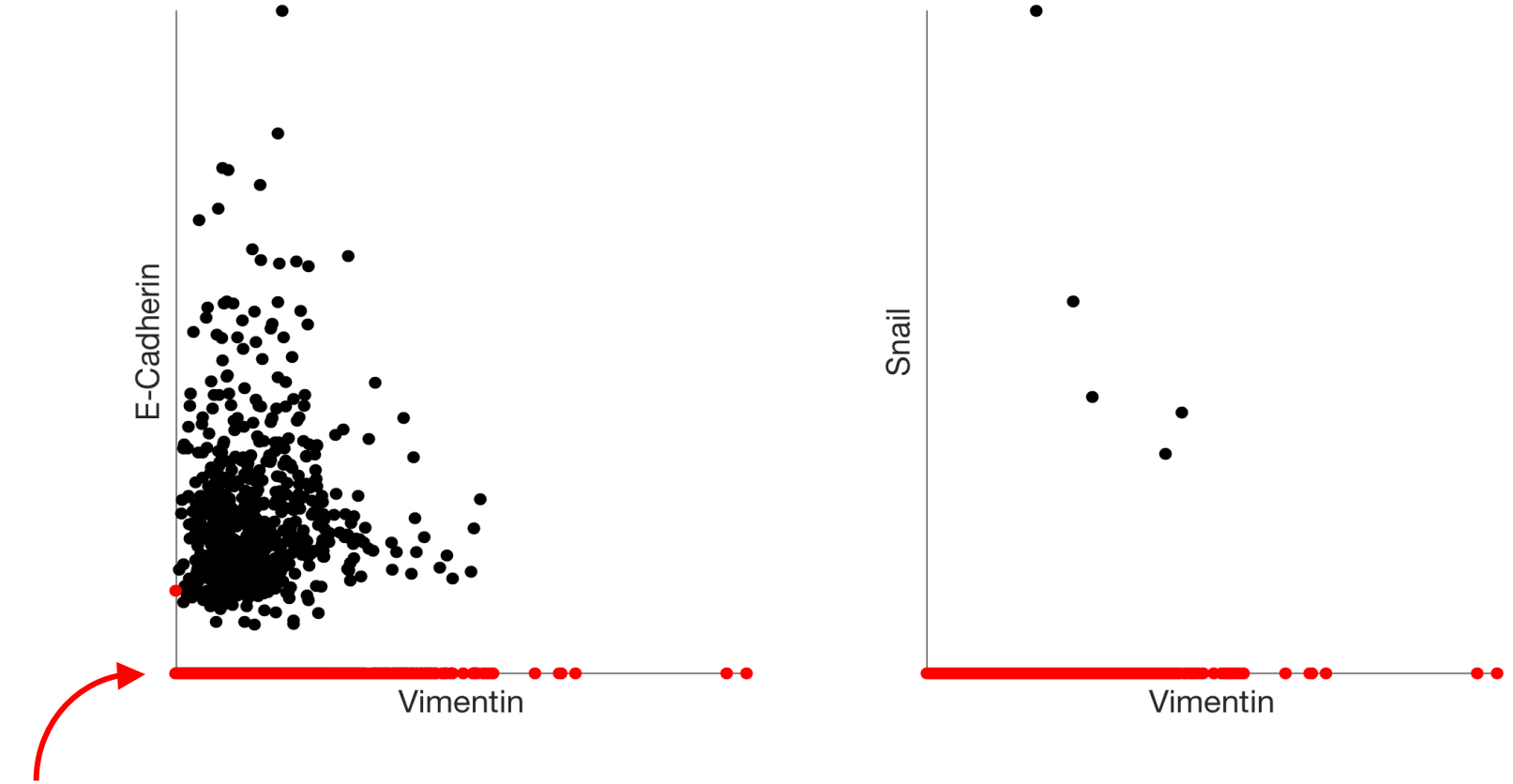




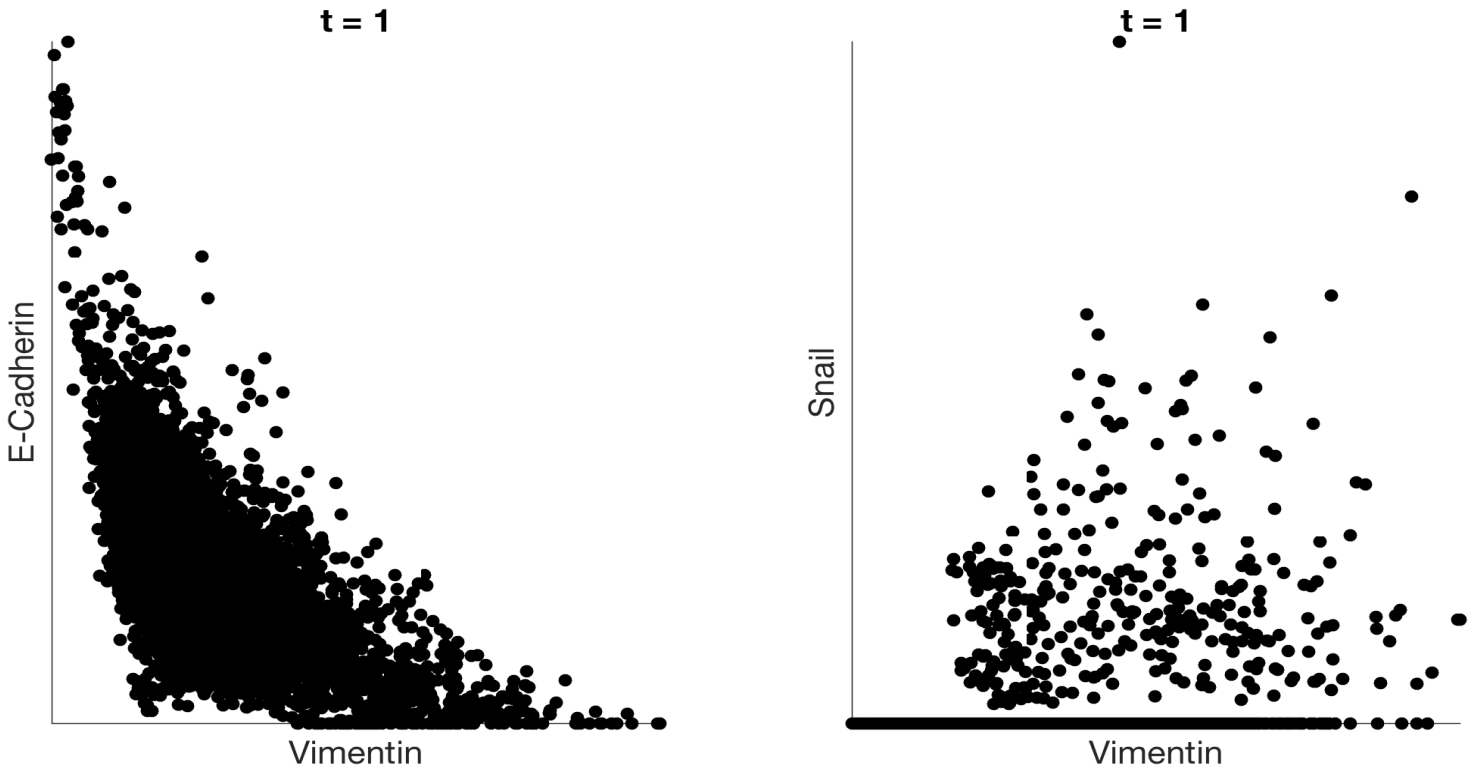


Markov Affinity-based Graph Imputation of Cells (MAGIC)

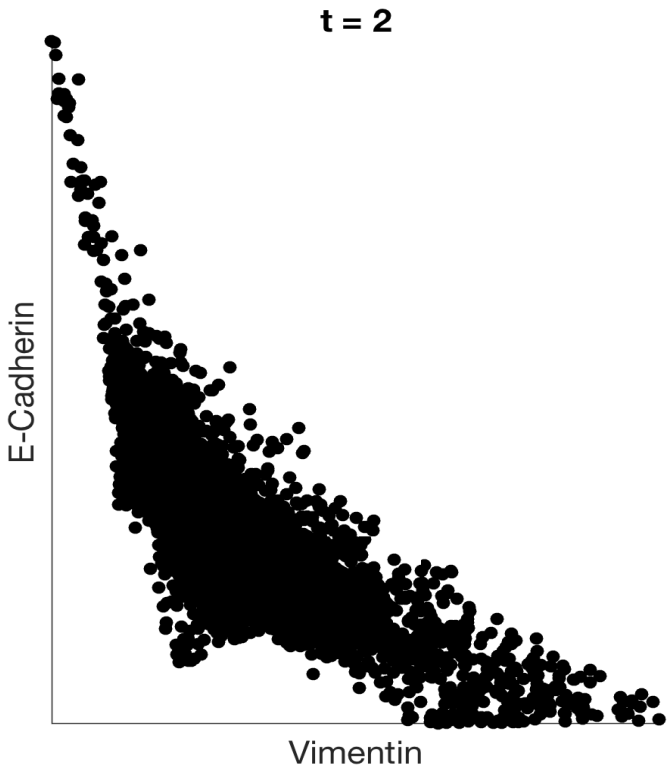


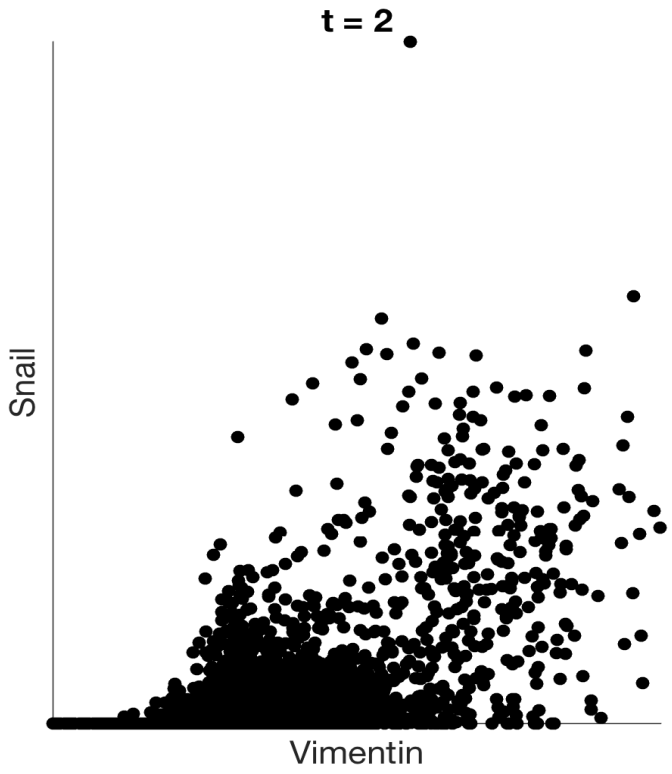


original data with dropout

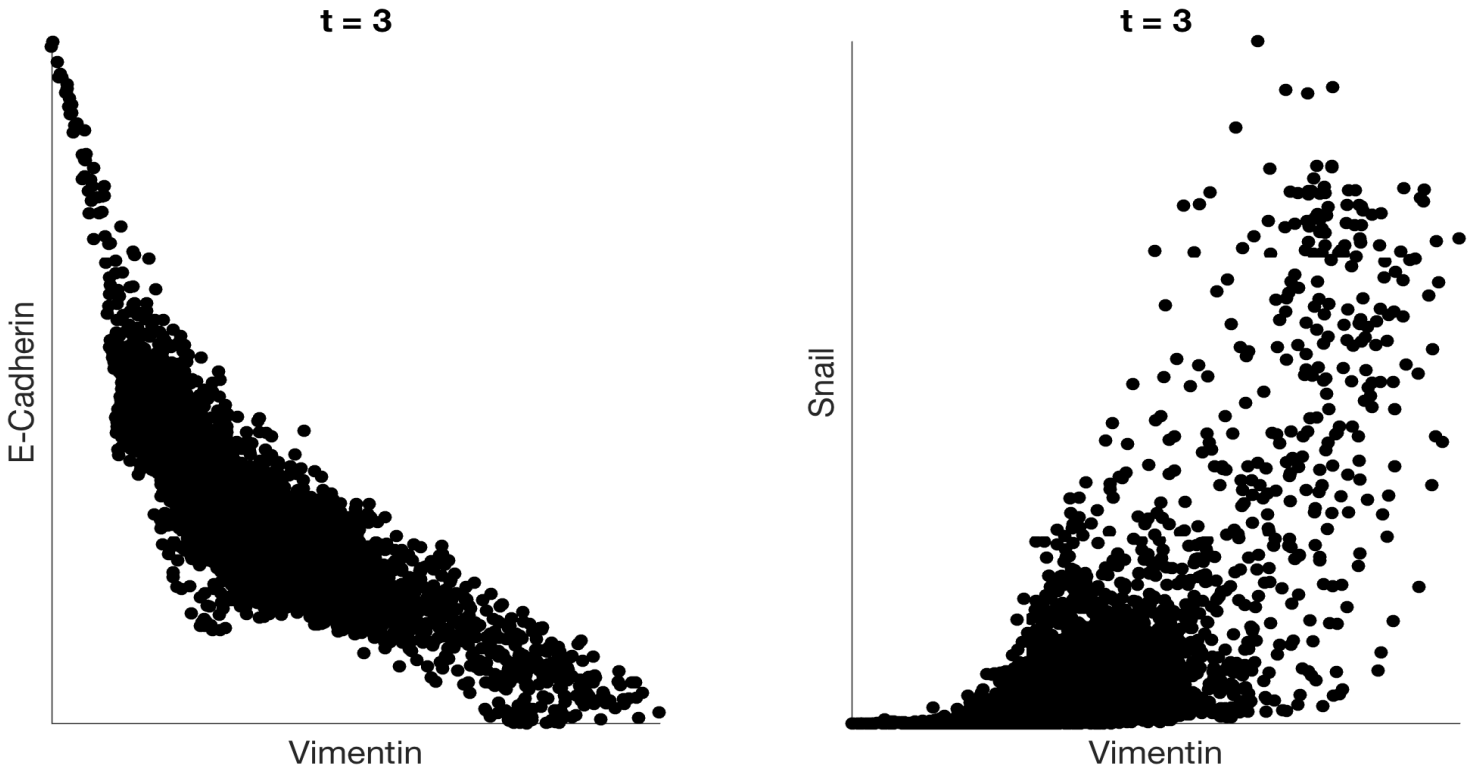


imputation with MAGIC

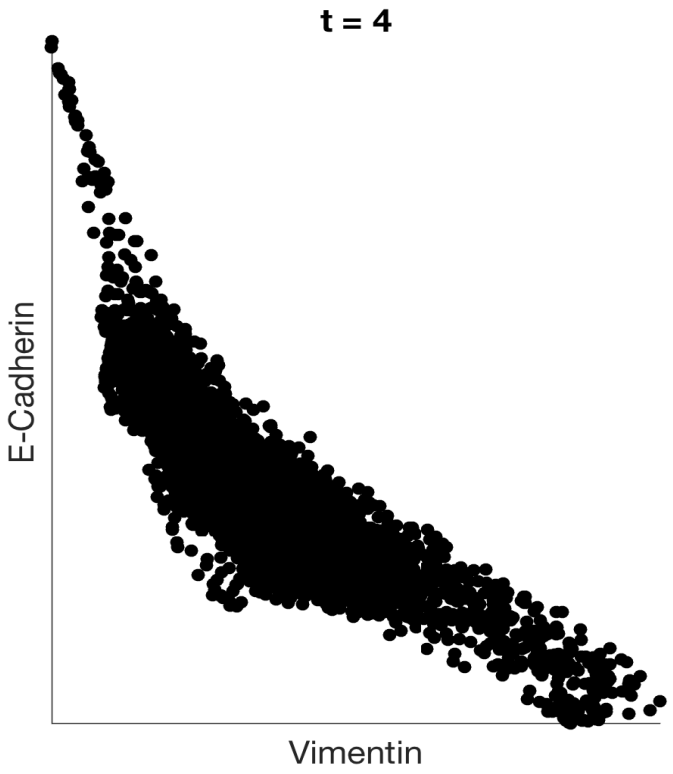


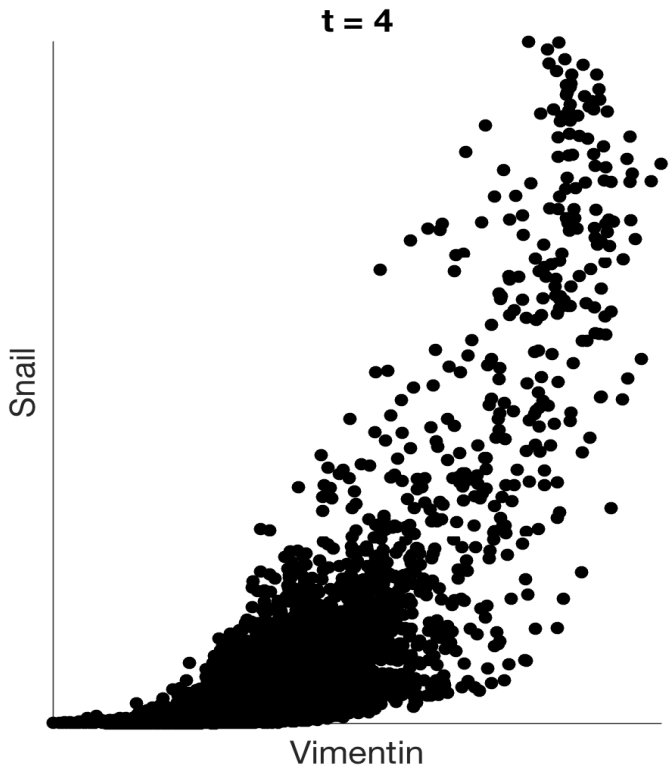


imputation with MAGIC

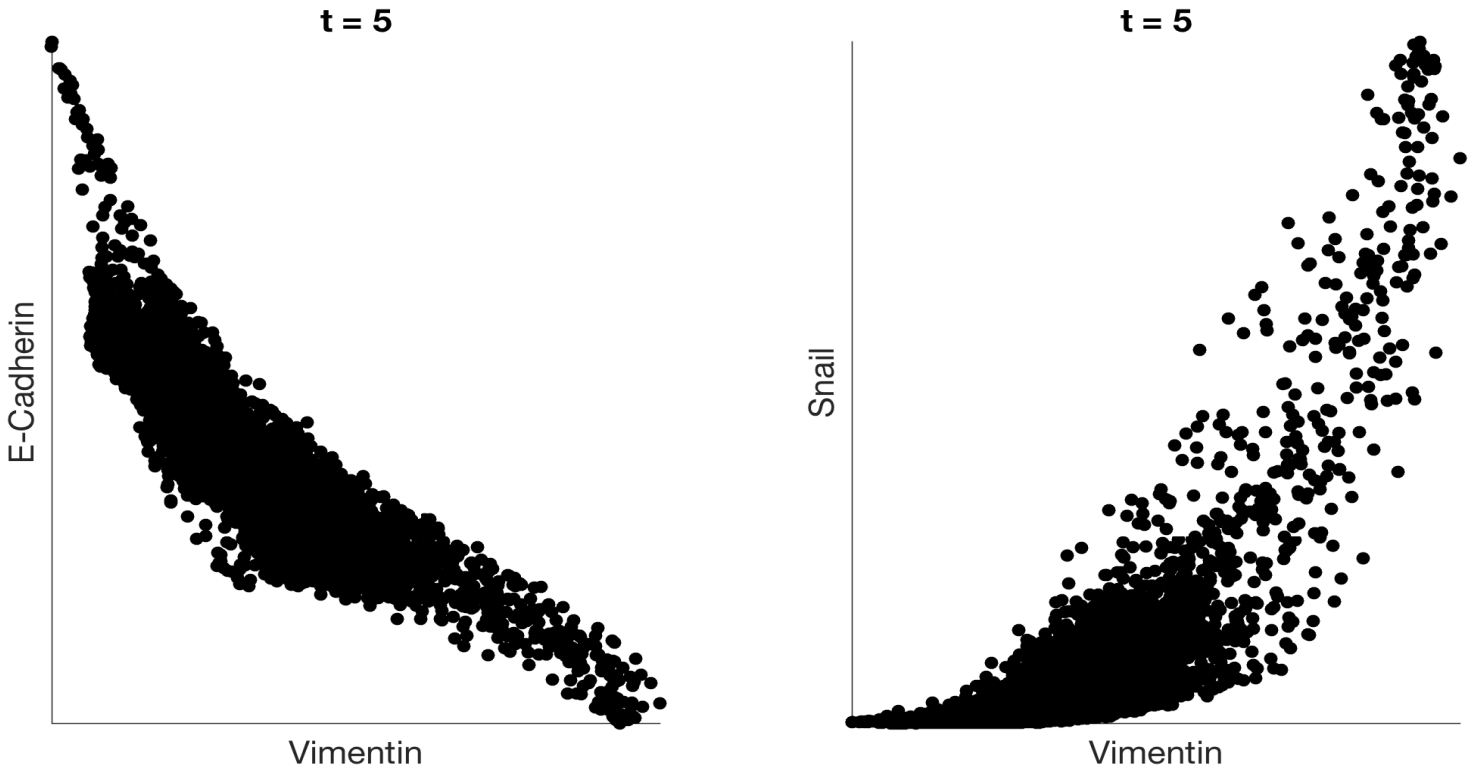


imputation with MAGIC

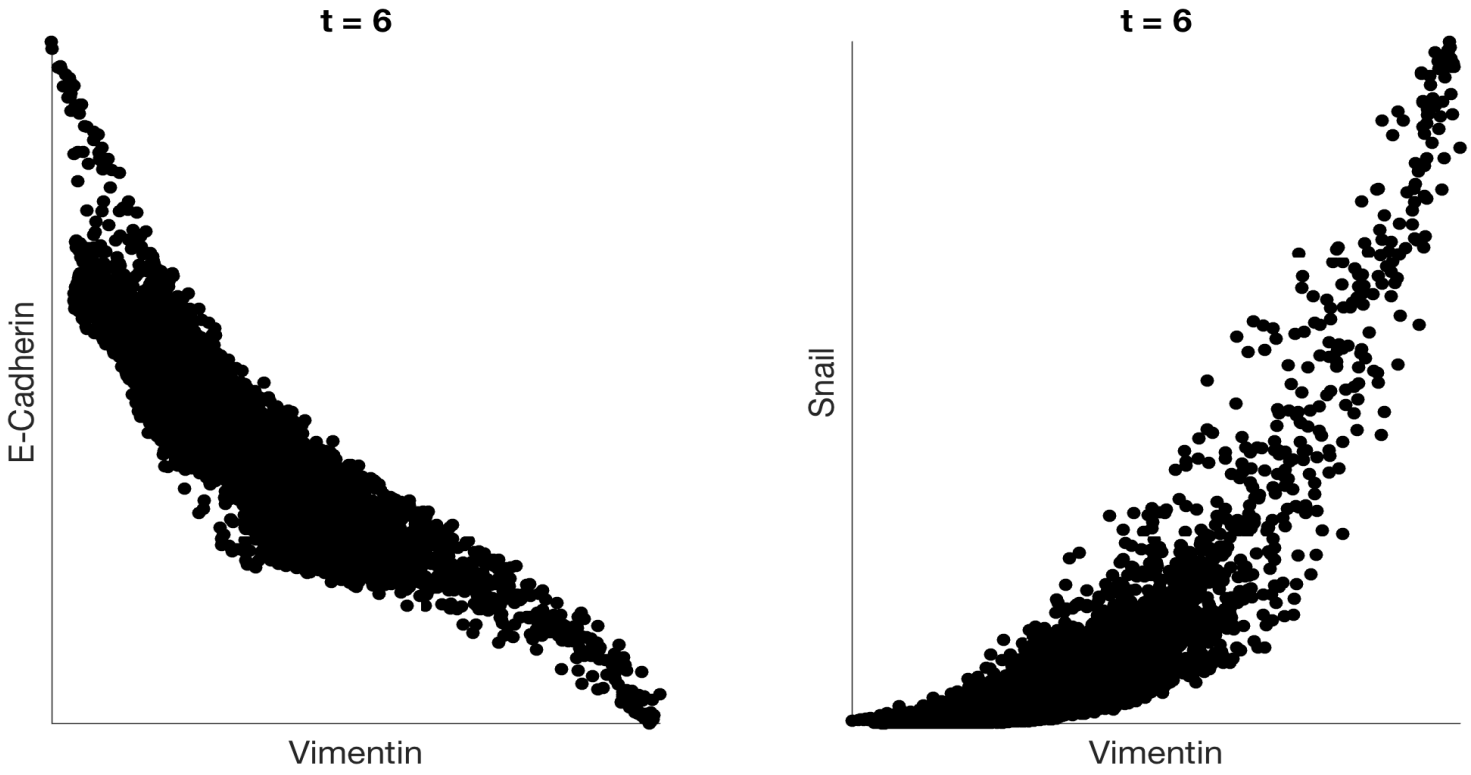




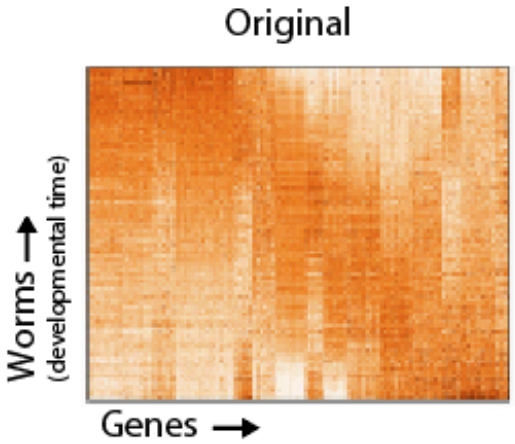
imputation with MAGIC

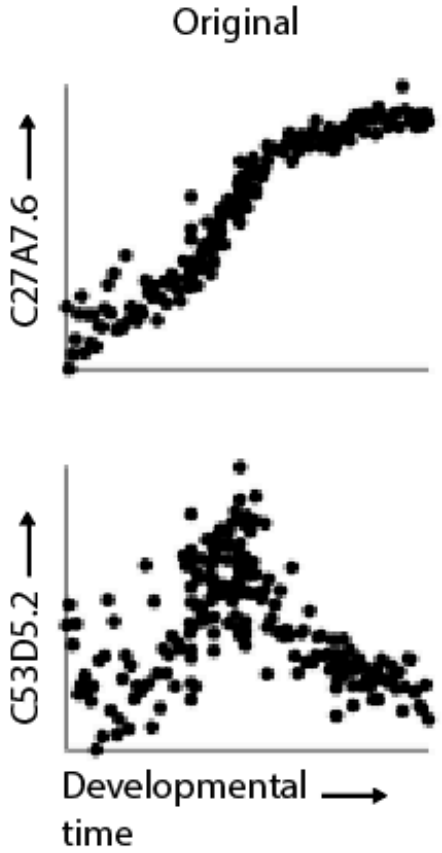


imputation with MAGIC

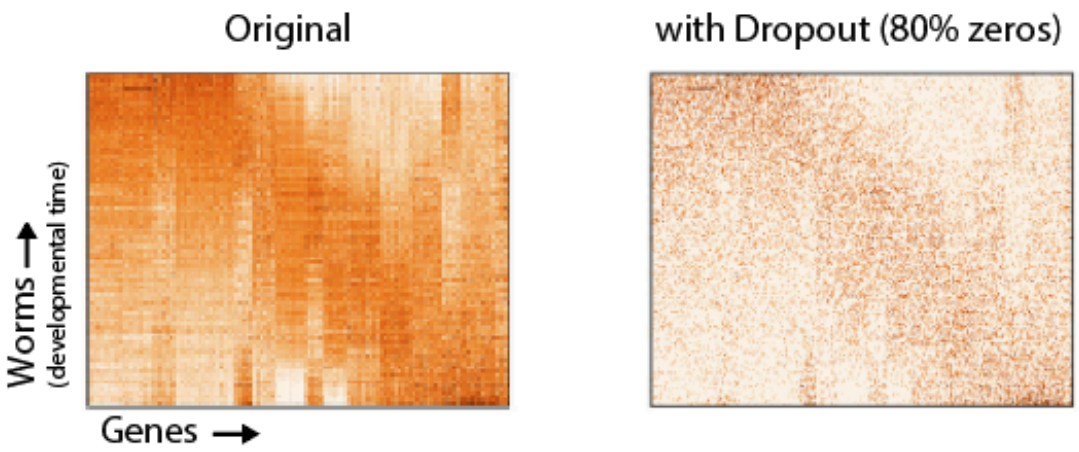


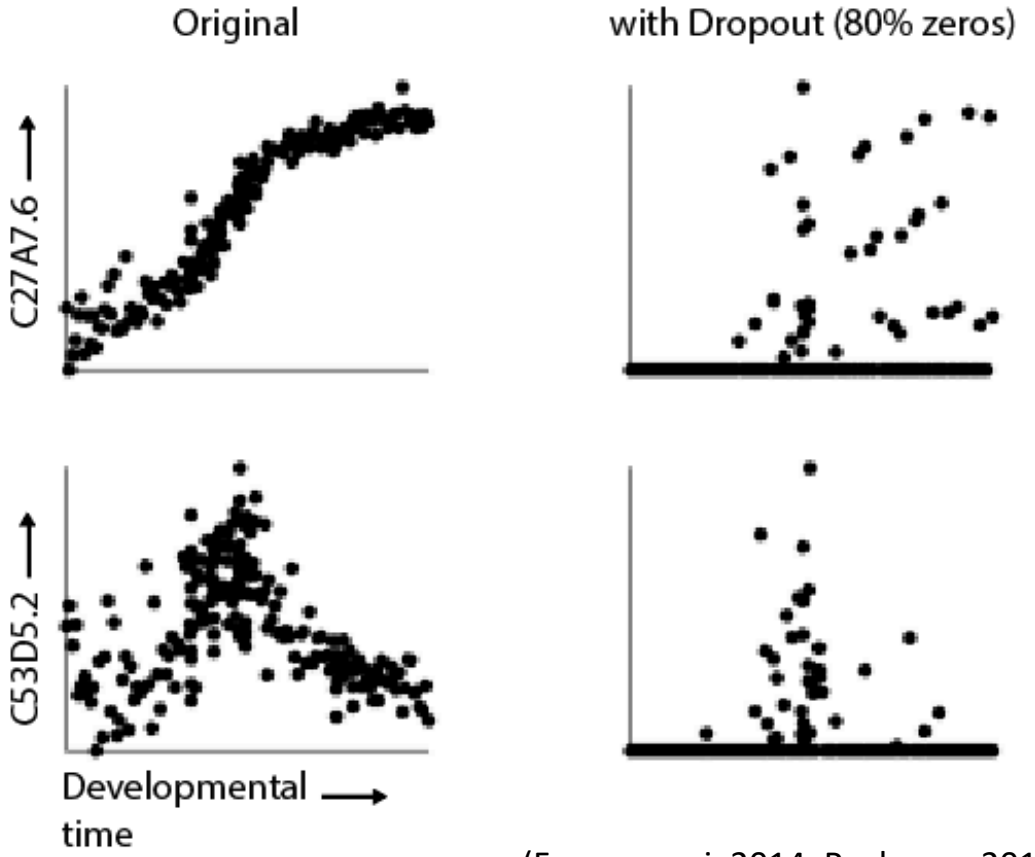
imputation with MAGIC



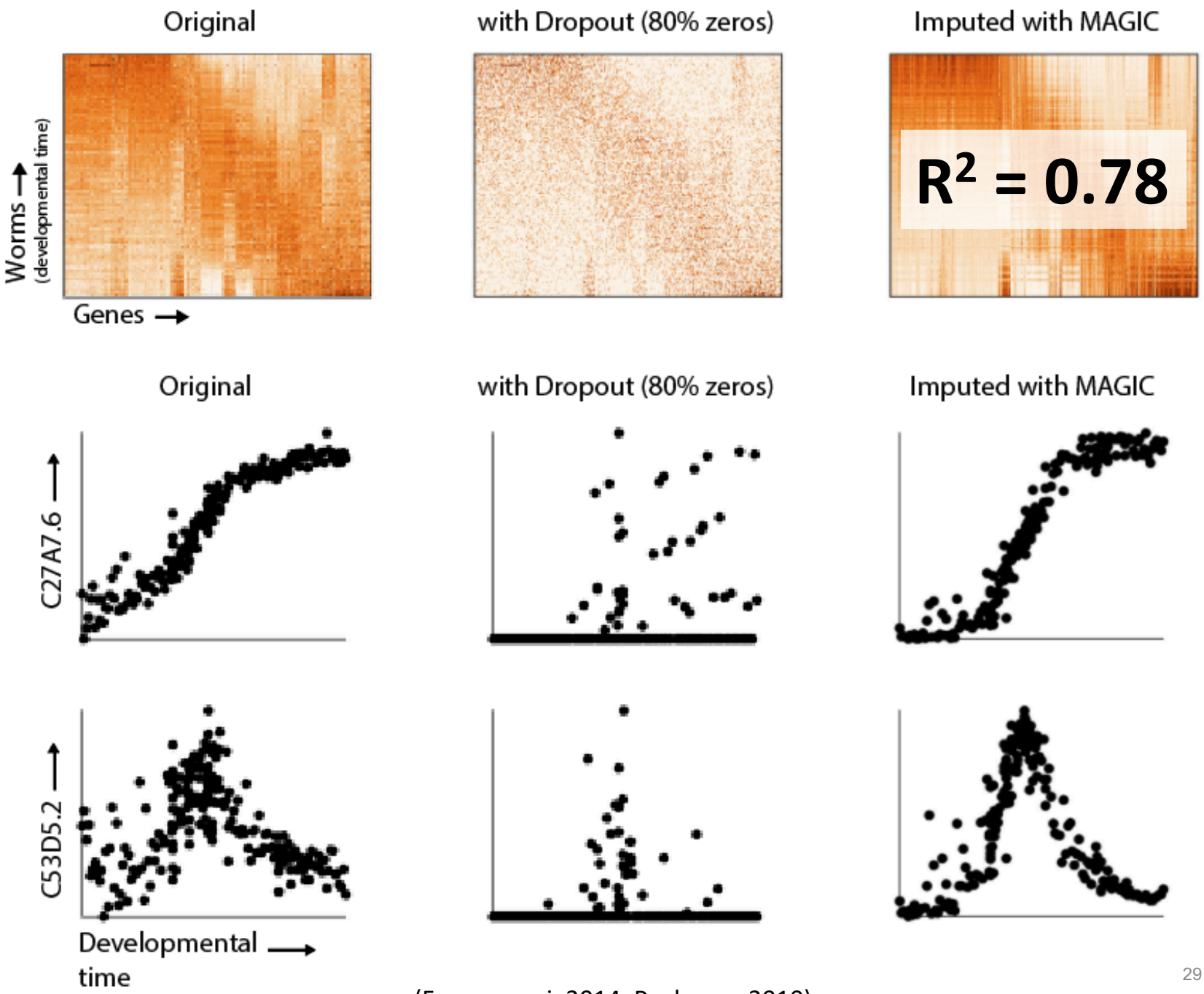


(Francesconi, 2014, Rockman, 2010)





(Francesconi, 2014, Rockman, 2010)



(Francesconi, 2014, Rockman, 2010)

TGF-β induced EMT (epithelial-to-mesenchymal transition)

Epithelial markers

Mesenchymal markers

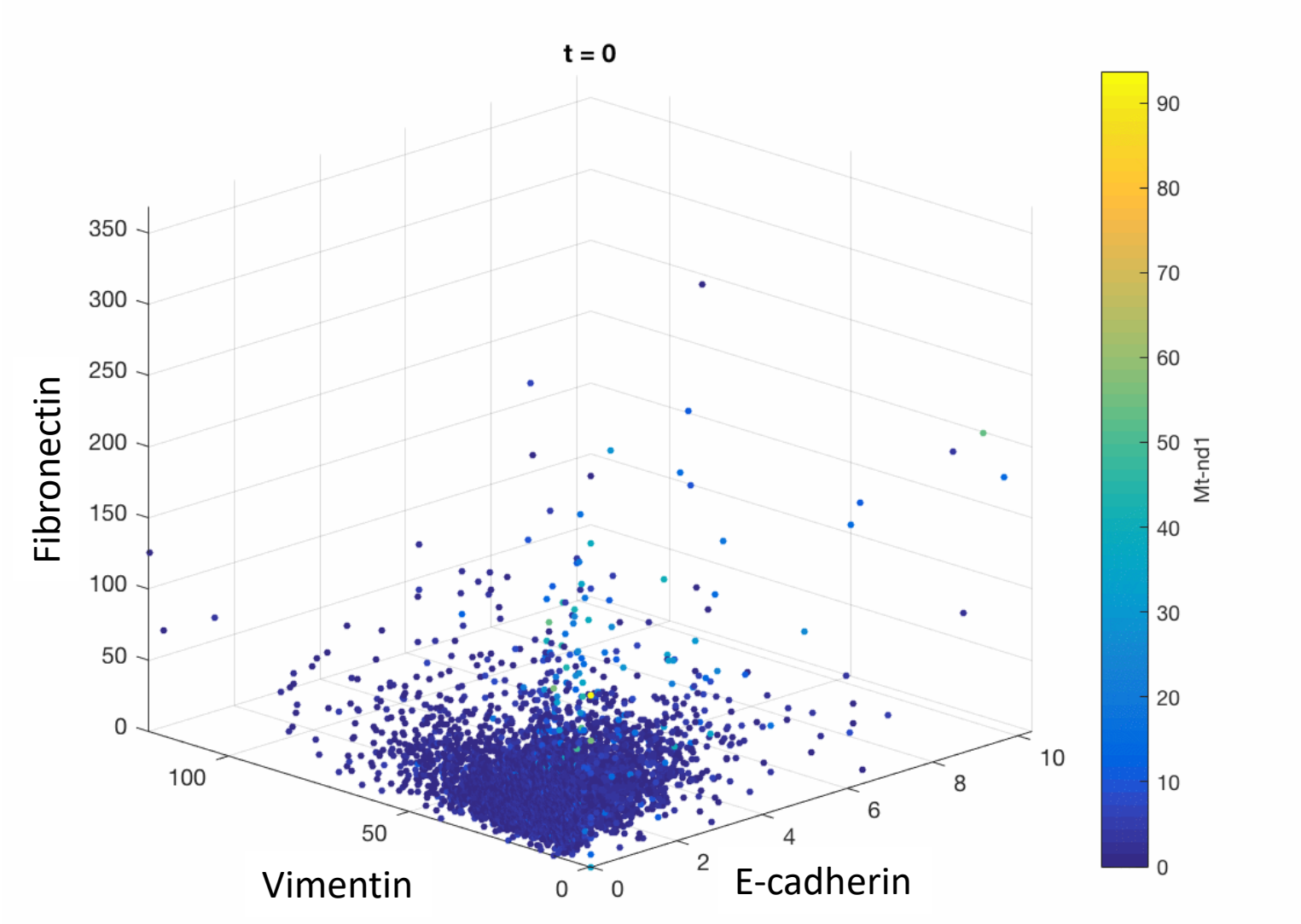
e.g.: E-cadherin, Laminin-1

e.g.: Vimentin, Snail, Zeb

With Weinberg Lab (Whitehead):

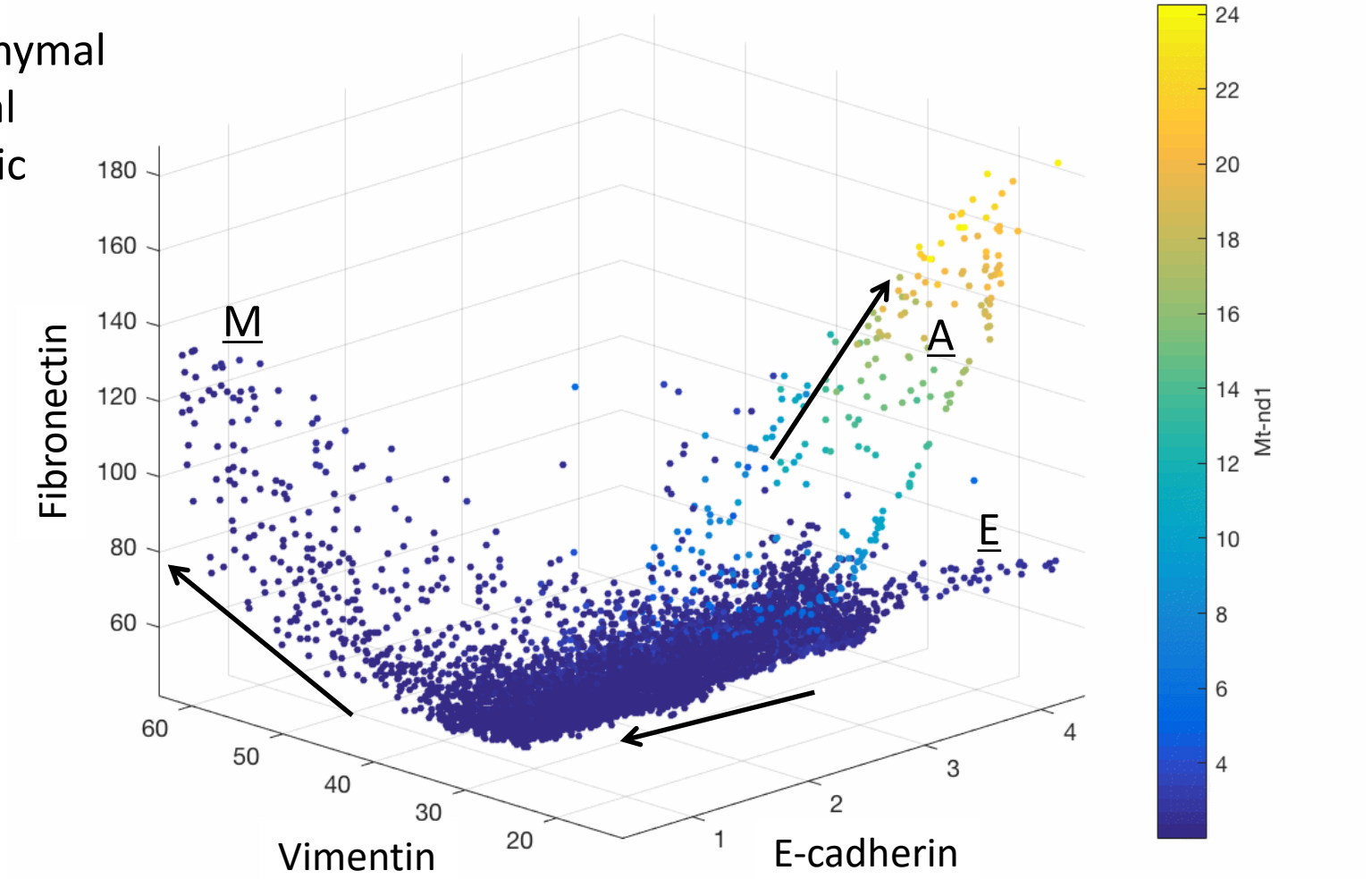
Diwakar Pattabiraman, Brian Bierie, Christine Chaffer

MAGIC reveals EMT continuum

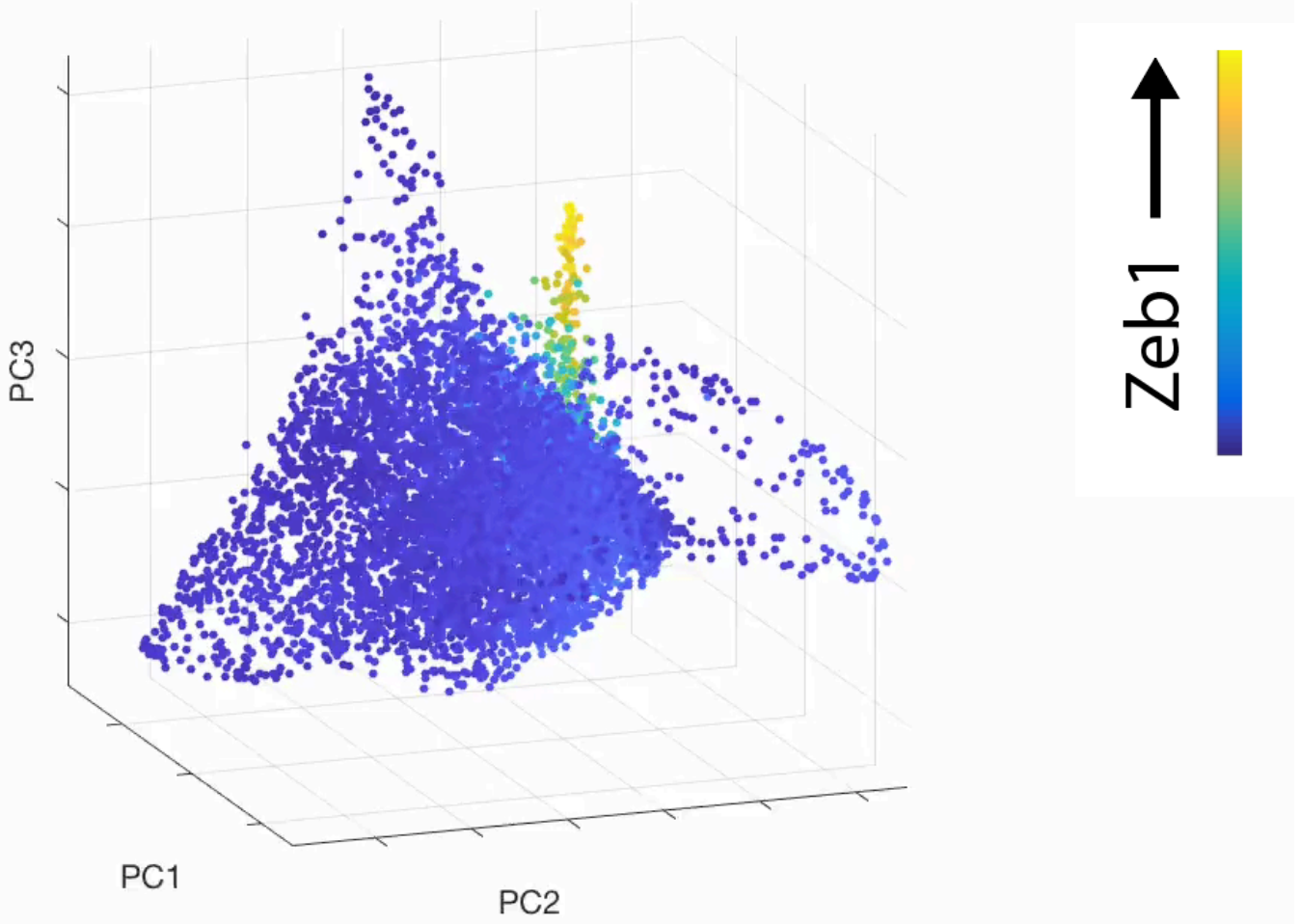


MAGIC reveals EMT continuum

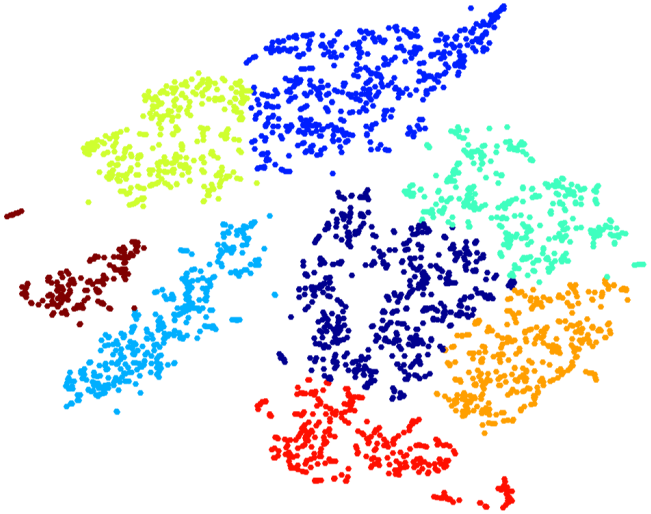
Mesenchymal
Epithelial
Apoptotic

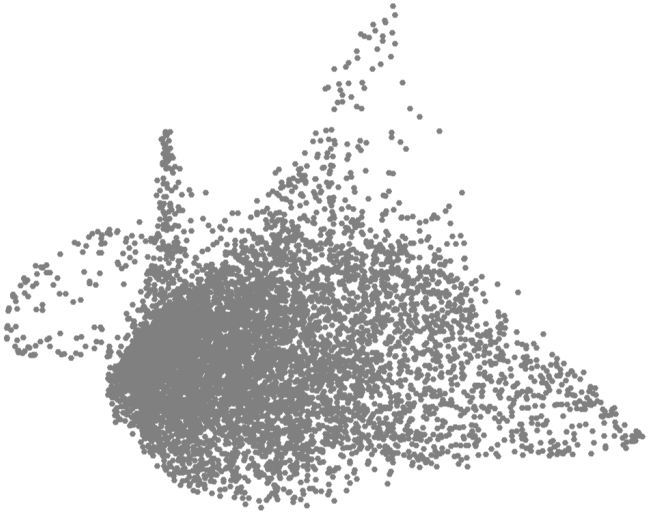


MAGIC reveals complex cellular state-space

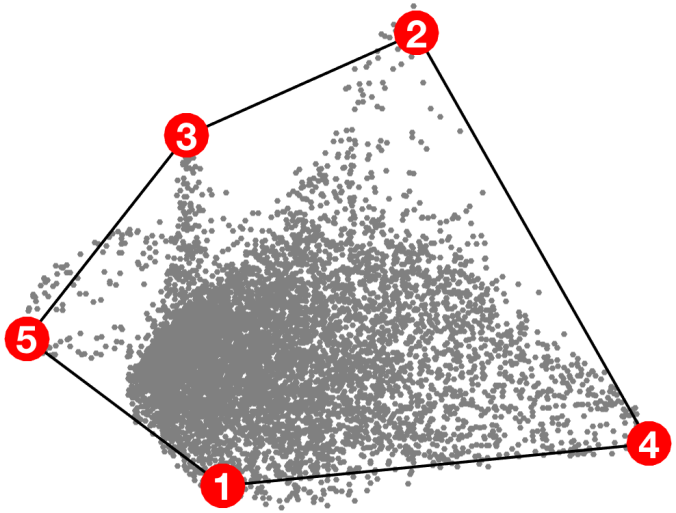


Clustering





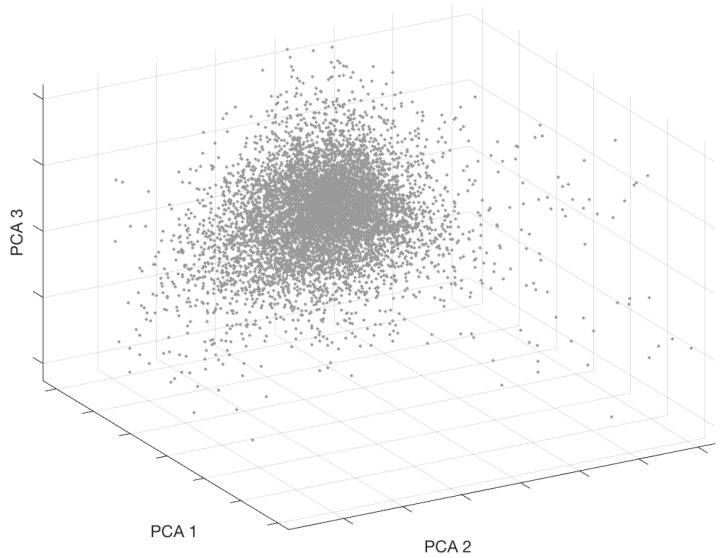
Archetypal analysis



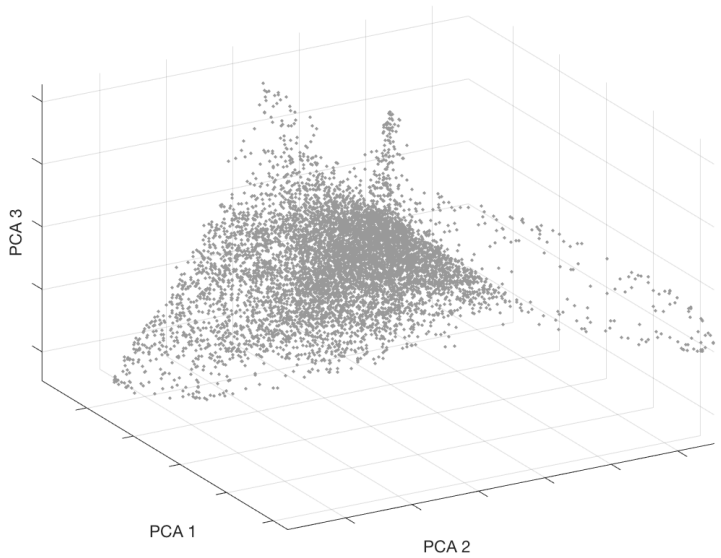
(Uri Alon lab: Shoval et al. 2012, Hart et al. 2015)

Archetypal Analysis

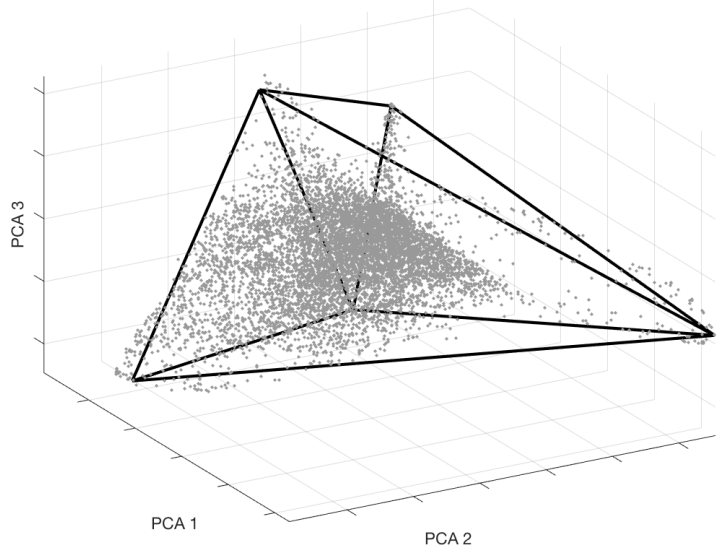
Data before MAGIC



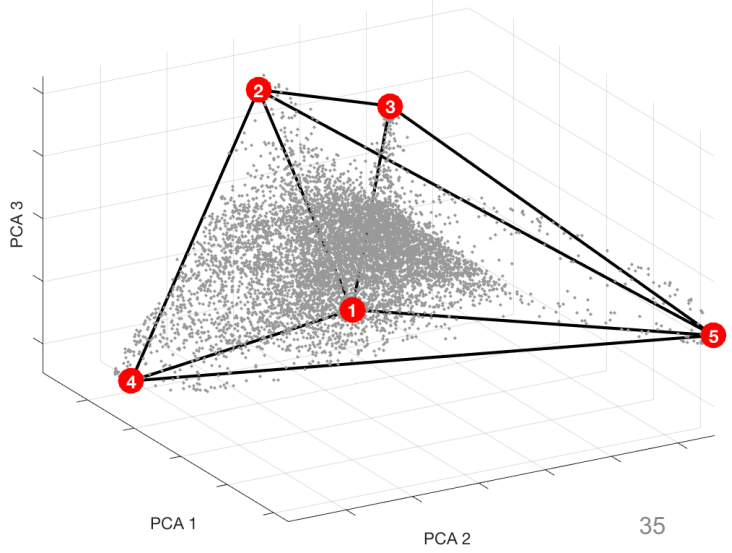
Data after MAGIC

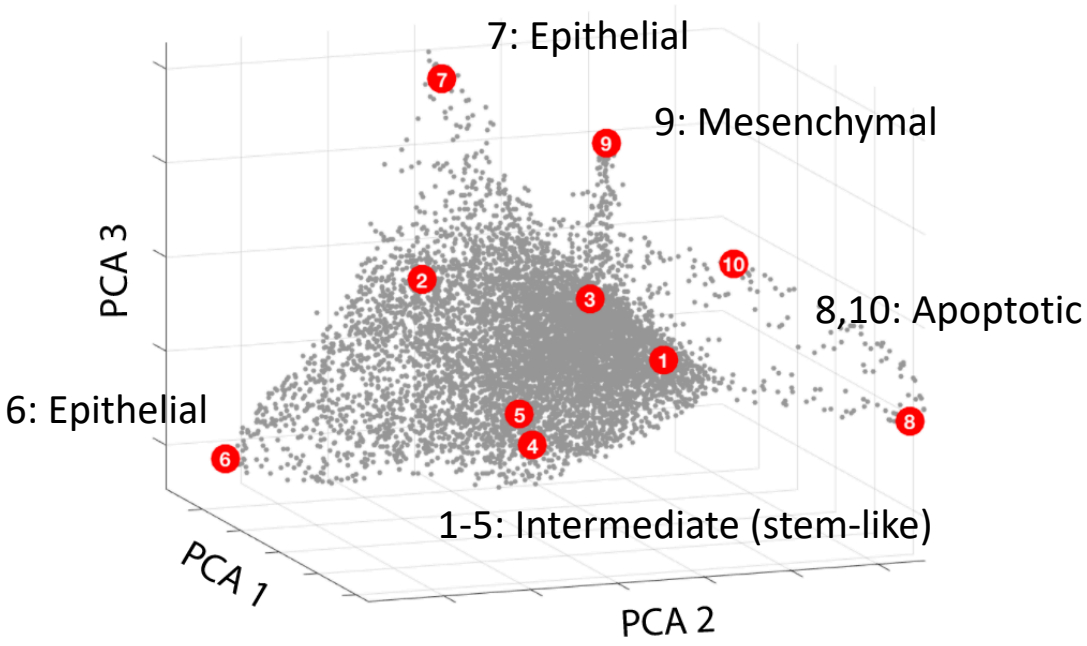


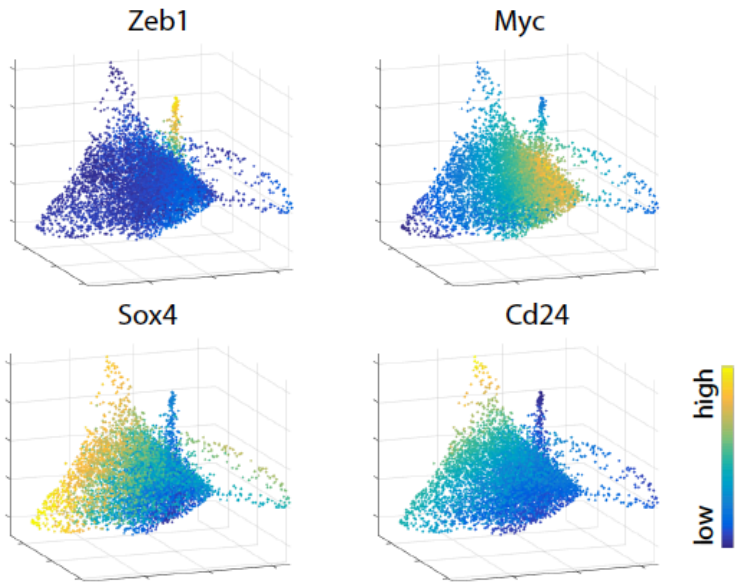
Fit polytope

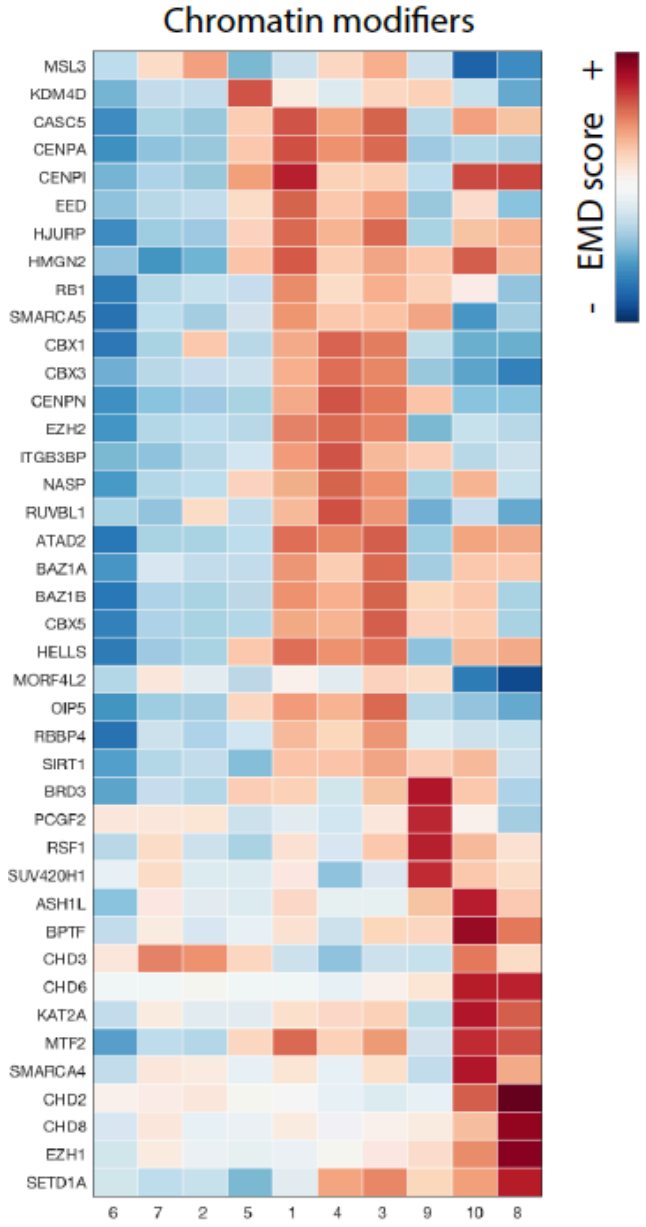


Infer archetypes

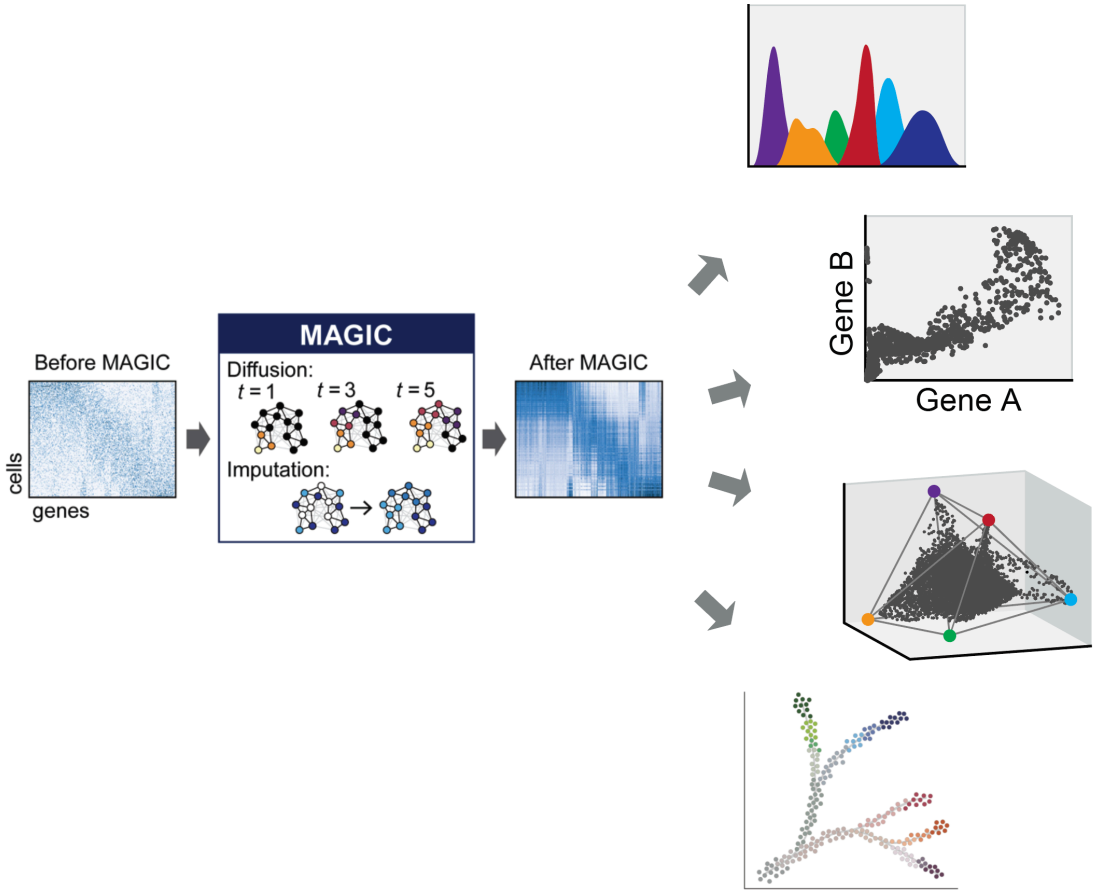


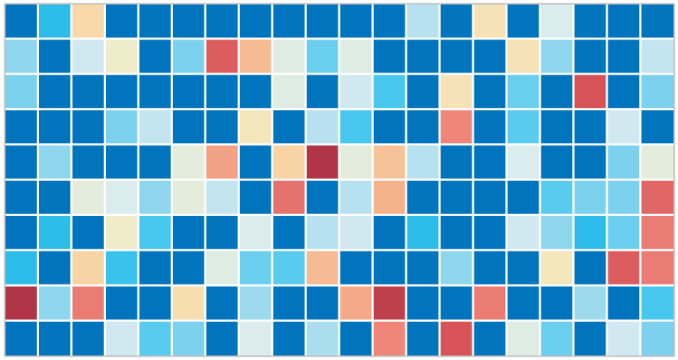

7: Epithelial
9: Mesenchymal
8,10: Apoptotic
6: Epithelial
1-5: Intermediate (stem-like)
PCA 3
PCA 1
PCA 2
7
9
10
2
3
1
5
4
6
8


Zeb1
Myc
Sox4
Cd24
high
low

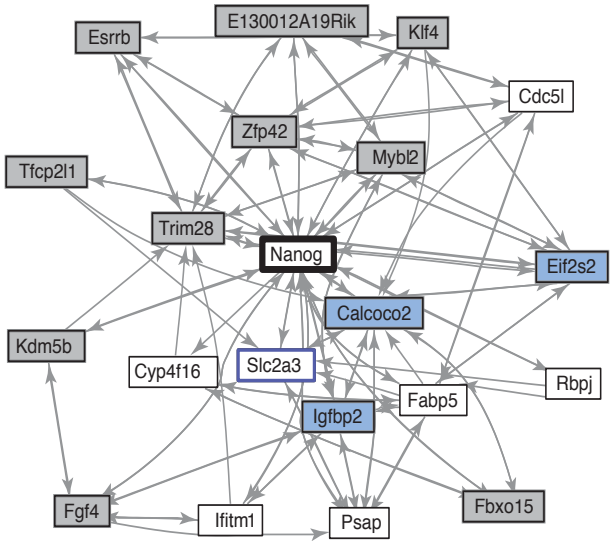

Chromatin modifiers
MSL3
KDM4D
CASC5
CENPA
CENPI
EED
HJURP
HMGN2
RB1
SMARCA5
CBX1
CBX3
CENPN
EZH2
ITGB3BP
NASP
RUVBL1
ATAD2
BAZ1A
BAZ1B
CBX5
HELLS
MORF4L2
OIP5
RBBP4
SIRT1
BRD3
PCGF2
RSF1
SUV420H1
ASH1L
BPTF
CHD3
CHD6
KAT2A
MTF2
SMARCA4
CHD2
CHD8
EZH1
SETD1A
6 7 2 5 1 4 3 9 10 8
+
EMD score
-

MAGIC enables downstream analyses



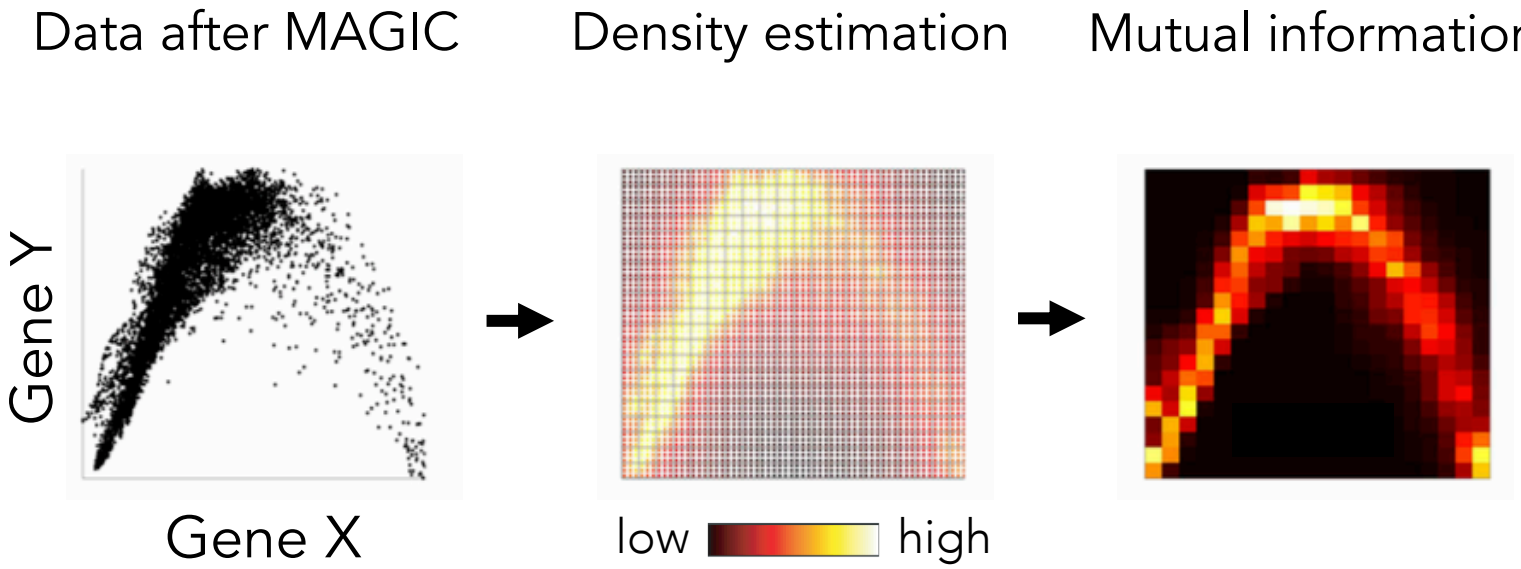


scRNA-seq data

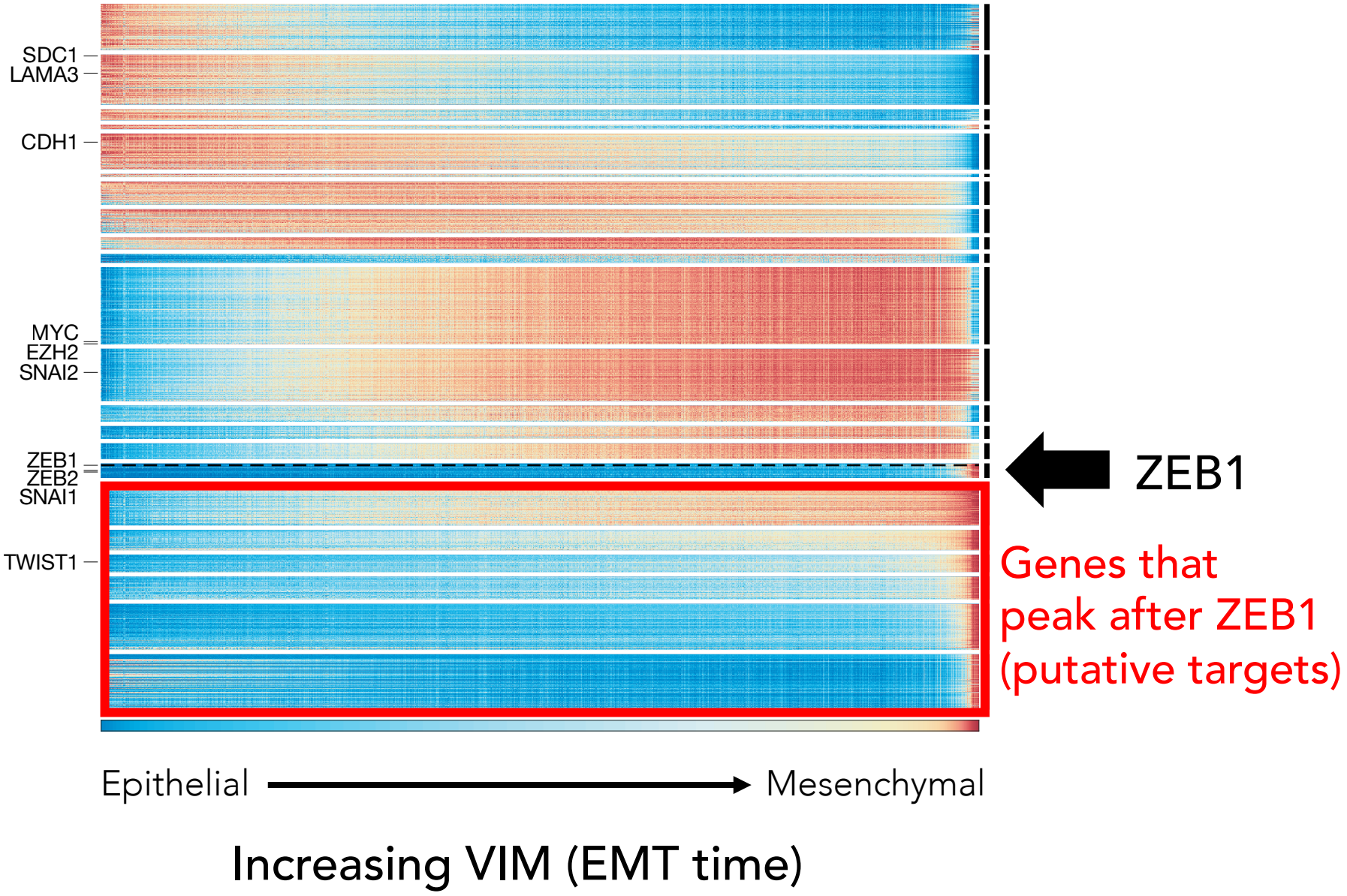


regulatory network

Using “snapshot” scRNA-seq data to predict transcription factor-target interactions involved in the epithelial—mesenchymal transition.

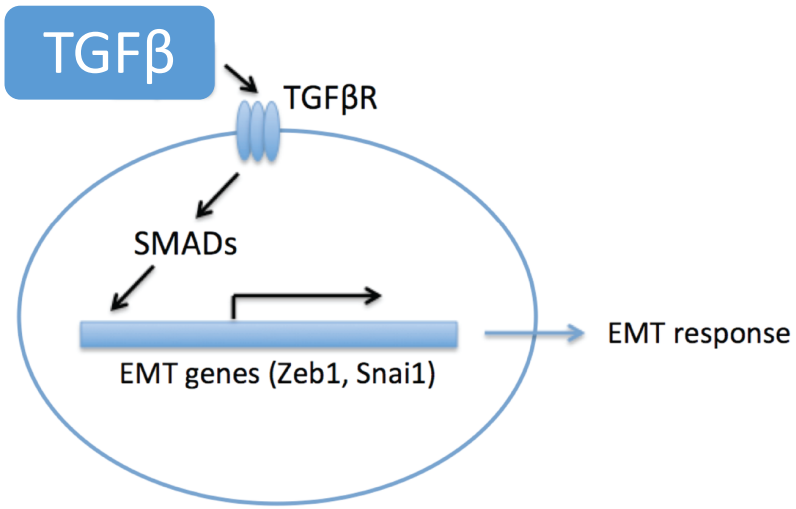


quantifying gene-gene relationships

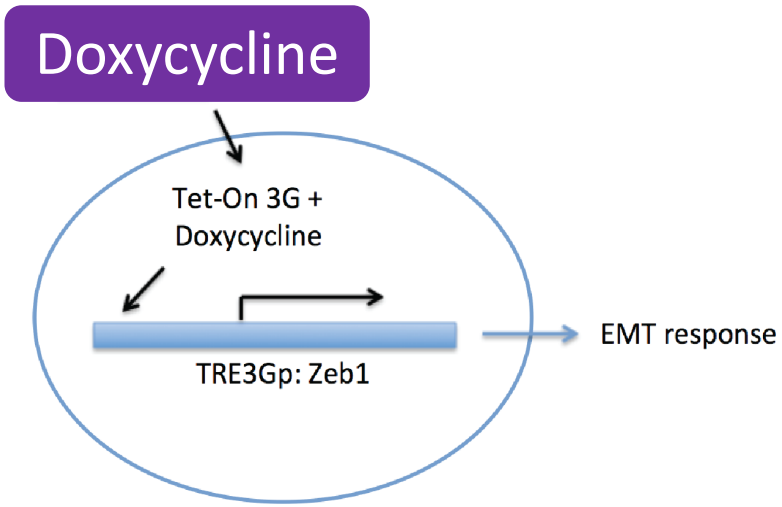

SDC1
LAMA3
CDH1
MYC
EZH2
SNAI2
ZEB1
ZEB2
SNAI1
TWIST1
ZEB1
Genes that peak after ZEB1 (putative targets)
Epithelial
Mesenchymal
Increasing VIM (EMT time)

Validating predictions with ZEB1 induction

TGFβ induction of EMT



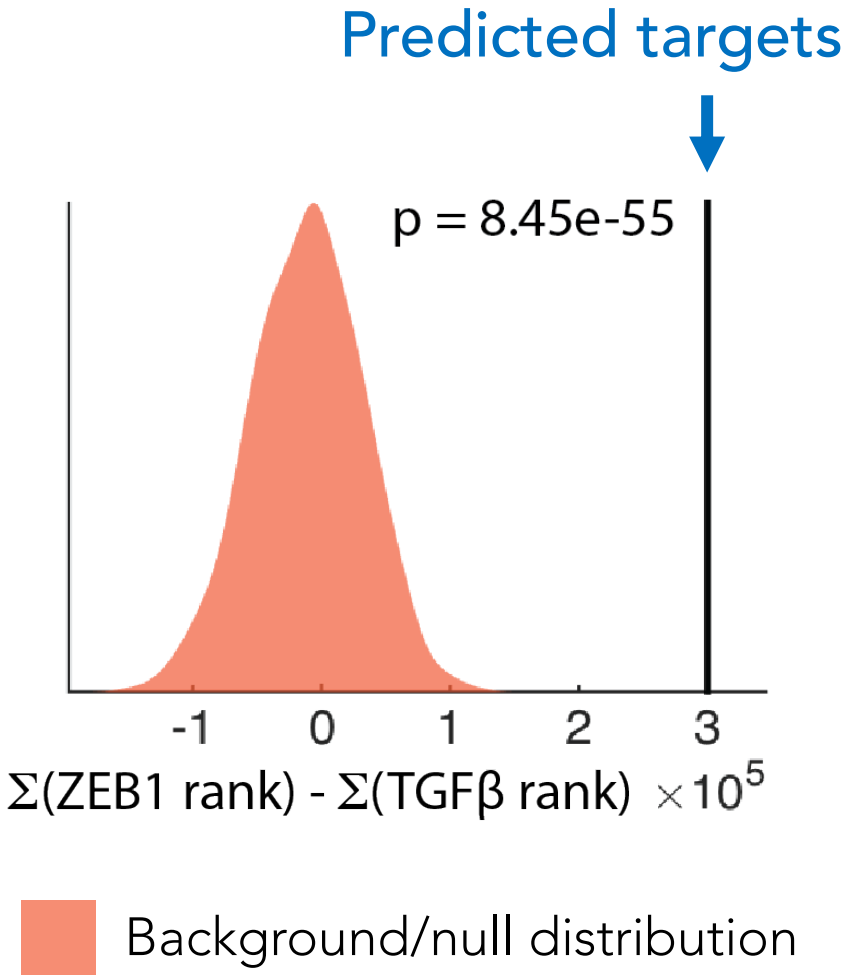
ZEB1 induction of EMT



With Weinberg Lab (Whitehead):

Diwakar Pattabiraman, Brian Bierie, Christine Chaffer

Validating predictions with ZEB1 induction



Summary

- scRNA-seq data is sparse and noisy
- Sparsity obscures biology
- Imputation via manifold learning (MAGIC)
- Imputed data enables down stream analyses