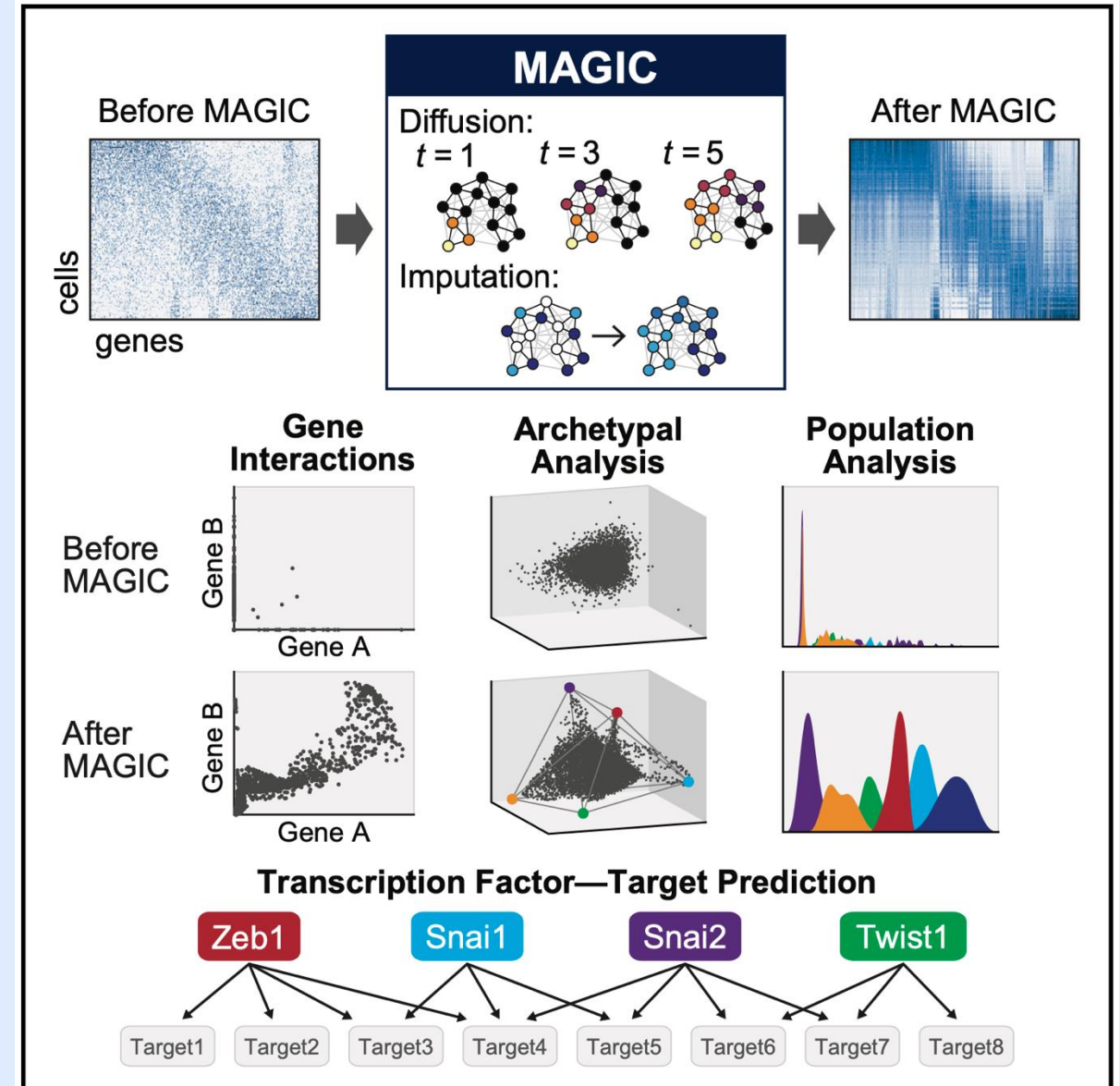


# Recovering Gene Interactions from Single-Cell Data Using Data Diffusion

David van Dijk, Roshan Sharma, Juozas Nainys, ..., Guy Wolf, Smita Krishnaswamy, Dana Pe'er (2018)

## ABSTRACT

- Algorithm to overcome the **Sparsity** of the single cell data
- This is one of the **Pioneering** papers that applied the concept of **Diffusion** to single-cell data.
- Based on this algorithm, various downstream applications and methods have since been developed.
- Published at Cell (2018), cited 1,720 times



## INDEX

1. Target Problem
2. Algorithm
3. Applications

# Target Problem

# TARGET PROBLEM

Early scRNA-seq data (~10 years ago) suffered from **low read depth**,

→ resulted in **severe sparsity** in gene expression matrices

This made even basic tasks like **DEG detection** and **cell type classification** difficult

➤ **Sparsity-aware algorithms** became essential



# TARGET PROBLEM

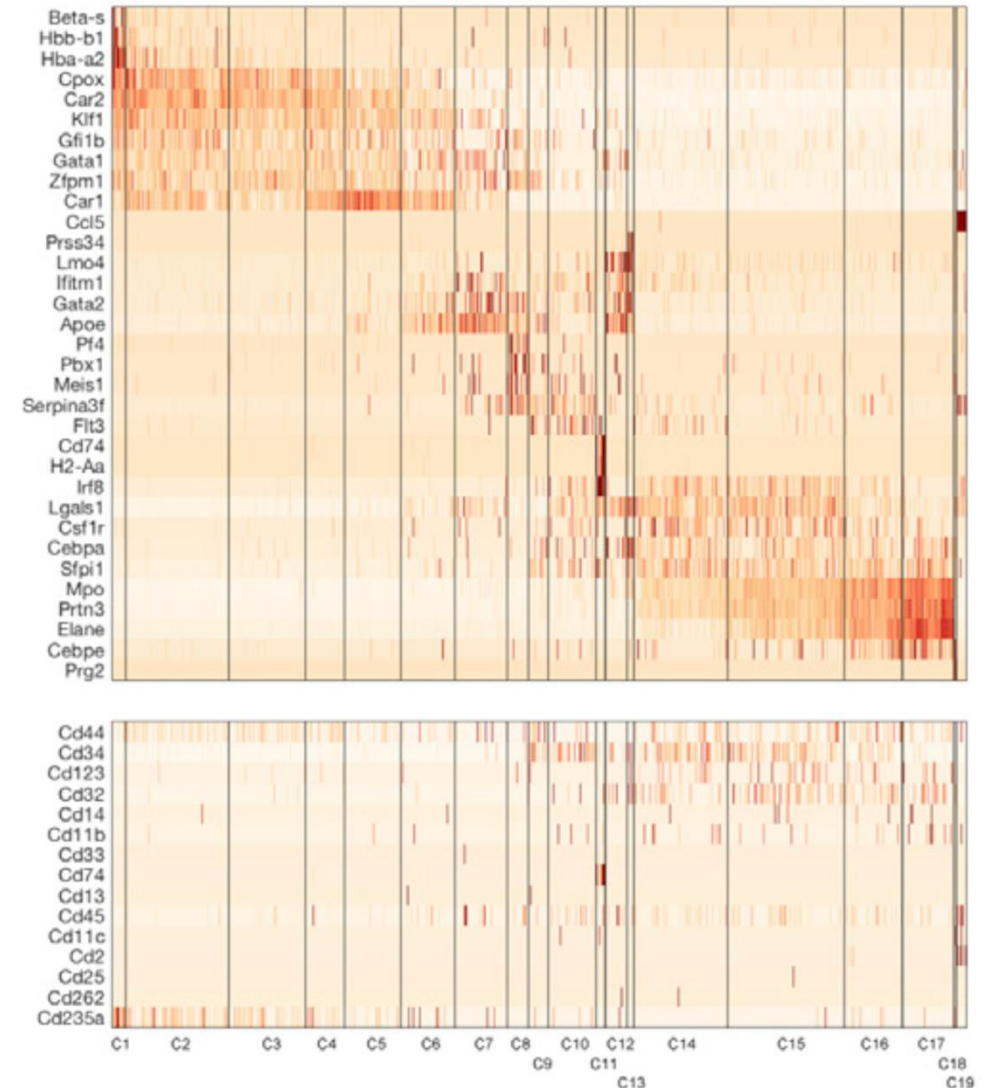
Early scRNA-seq data (~10 years ago) suffered from **low read depth**,

→ resulted in **severe sparsity** in gene expression matrices

This made even basic tasks like **DEG detection** and **cell type classification** difficult

➤ **Sparsity-aware algorithms** became essential

Gene Expression Heatmap of mouse bone marrow dataset (Paul et al., 2015) (2730 cells)



# TARGET PROBLEM

Early scRNA-seq data (~10 years ago) suffered from **low read depth**,

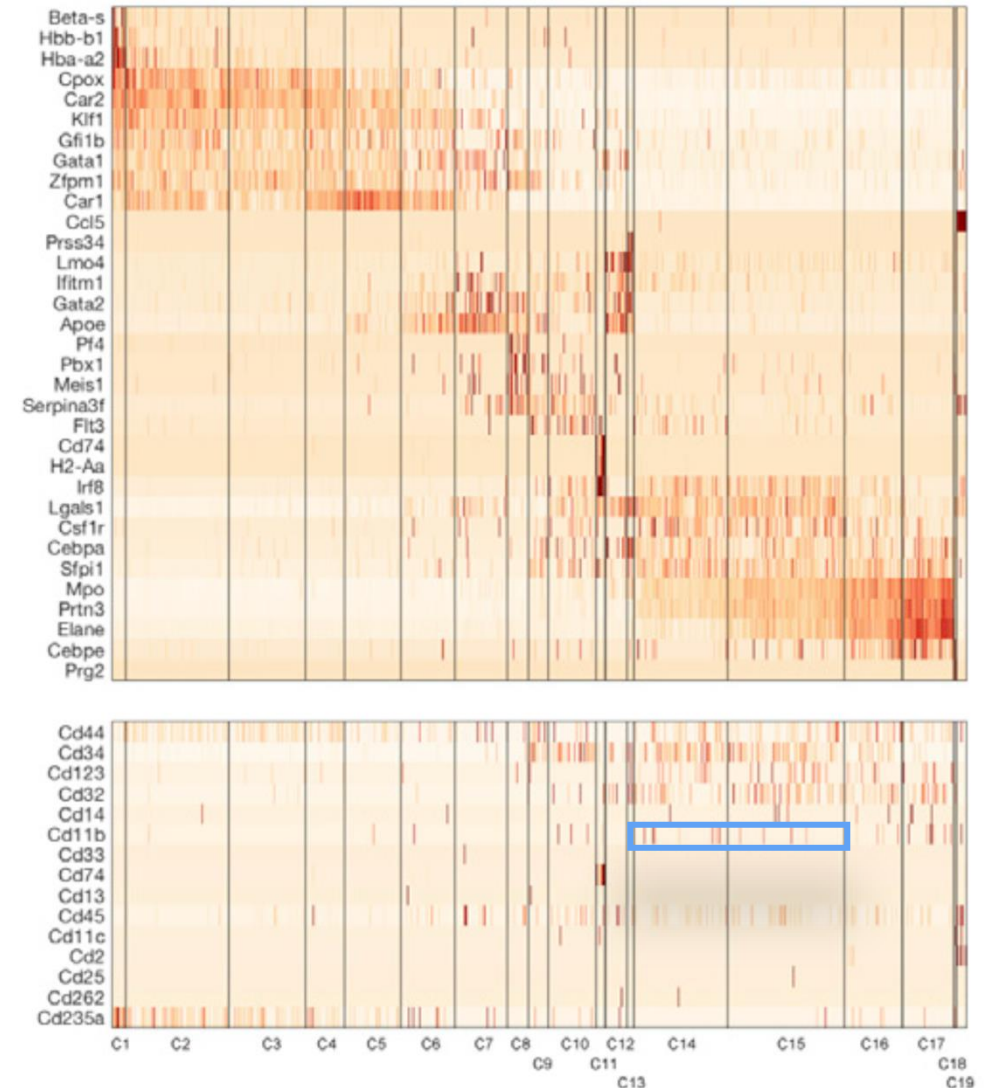
→ resulted in **severe sparsity** in gene expression matrices

This made even basic tasks like **DEG detection** and **cell type classification** difficult

➤ **Sparsity-aware algorithms** became essential

Ex) C14,C15 cluster is monocyte but only 1.6% of cells expressed CD14 and 5.8% cells expressed CD11b which are marker genes of Monocyte

Gene Expression Heatmap of mouse bone marrow dataset (Paul et al., 2015) (2730 cells)



# Why It Still Matters?

**Modern scRNA-seq** can achieve higher read depth  
→ clustering & DEG analysis now feasible without imputation

**BUT:**

**Sparsity remains** a core challenge in many domains

This algorithm is **still broadly applicable** to other **sparse matrix problems**

It serves as a foundational algorithm for many genomic diffusion methods developed later

(Personally, this was a field I was very curious about)

# Algorithm

MAGIC (Markov Affinity-based Graph  
Imputation of Cells)

# THE CONCEPTUAL FOUNDATION OF THE ALGORITHM

1. If two cells exhibit similar global gene expression patterns, then it is likely that they also share similar expression values at individual genes.

|        | Gene A | Gene B | Gene C | Gene D | Gene F | Gene G |
|--------|--------|--------|--------|--------|--------|--------|
| Cell A | 6      | 2      | 4      | 0      | 3      | 7      |
| Cell B | 5      | 1      | 3      | 0      | 1      | ???    |
| Cell C | 1      | 6      | 1      | 7      | 2      | 0      |
| Cell D | 5      | 1      | 0      | 3      | 1      | 4      |

Gene Expression Matrix

# THE CONCEPTUAL FOUNDATION OF THE ALGORITHM

1. If two cells exhibit similar global gene expression patterns, then it is likely that they also share similar expression values at individual genes.

|        | Gene A | Gene B | Gene C | Gene D | Gene F | Gene G |
|--------|--------|--------|--------|--------|--------|--------|
| Cell A | 6      | 2      | 4      | 0      | 3      | 7      |
| Cell B | 5      | 1      | 3      | 0      | 1      | ???    |
| Cell C | 1      | 6      | 1      | 7      | 2      | 0      |
| Cell D | 5      | 1      | 0      | 3      | 1      | 4      |

Gene Expression Matrix

|       | cellA | cellB | cellC | cellD |
|-------|-------|-------|-------|-------|
| cellA | 0     | 2.6   | 10    | 5.6   |
| cellB | 2.6   | 0     | 9.7   | 4.2   |
| cellC | 10    | 9.7   | 0     | 7.7   |
| cellD | 5.7   | 4.2   | 7.7   | 0     |

Distance Matrix

Ex) Euclidean Distance :  $((6-5)^2 + (2-1)^2 + (4-3)^2 + (0-0)^2 + (3-1)^2)/5 = 2.6$

# THE CONCEPTUAL FOUNDATION OF THE ALGORITHM

1. If two cells exhibit similar global gene expression patterns, then it is likely that they also share similar expression values at individual genes.
2. In an expression matrix, a zero value can either represent a true biological zero or result from insufficient sequencing depth or technical limitation.

(True Zero vs Technical Dropout)

|        | Gene A | Gene B | Gene C | Gene D | Gene F | Gene G |
|--------|--------|--------|--------|--------|--------|--------|
| Cell A | 6      | 2      | 4      | 0      | 3      | 7      |
| Cell B | 5      | 1      | 3      | 0      | 1      | 0/?    |
| Cell C | 1      | 6      | 1      | 7      | 2      | 0      |
| Cell D | 5      | 1      | 0      | 3      | 1      | 4      |

Gene Expression Matrix

# THE CONCEPTUAL FOUNDATION OF THE ALGORITHM

1. If two cells exhibit similar global gene expression patterns, then it is likely that they also share similar expression values at individual genes.
2. In an expression matrix, a zero value can either represent a true biological zero or result from insufficient sequencing depth or technical limitation.  
(True Zero vs Technical Dropout)

But,  
They cannot be discriminated theoretically.

# THE CONCEPTUAL FOUNDATION OF THE ALGORITHM

1. If two cells exhibit similar global gene expression patterns, then it is likely that they also share similar expression values at individual genes.
2. In an expression matrix, a zero value can either represent a true biological zero or result from insufficient sequencing depth or technical limitation.  
(True Zero vs Technical Dropout)

But,  
This cannot be discriminated theoretically.

So,  
The MAGIC allows analysts to control the magnitude of diffusion.

# THE CONCEPTUAL FOUNDATION OF THE ALGORITHM

If two cells exhibit similar global gene expression patterns, then it is likely that they also share similar expression values at individual genes.

2. In an expression matrix, a zero value can either represent a true biological zero or result from insufficient sequencing depth or technical limitation.

(True Zero vs Technical Dropout)

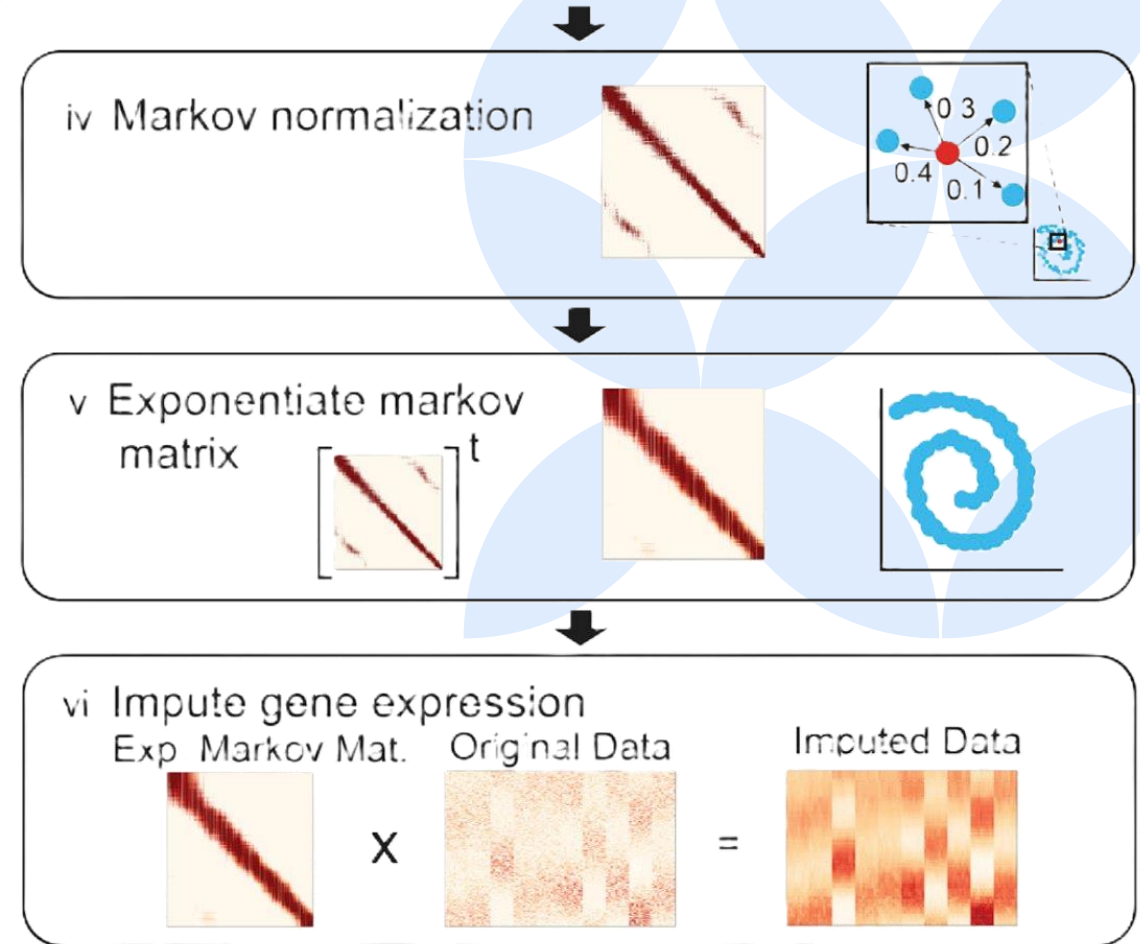
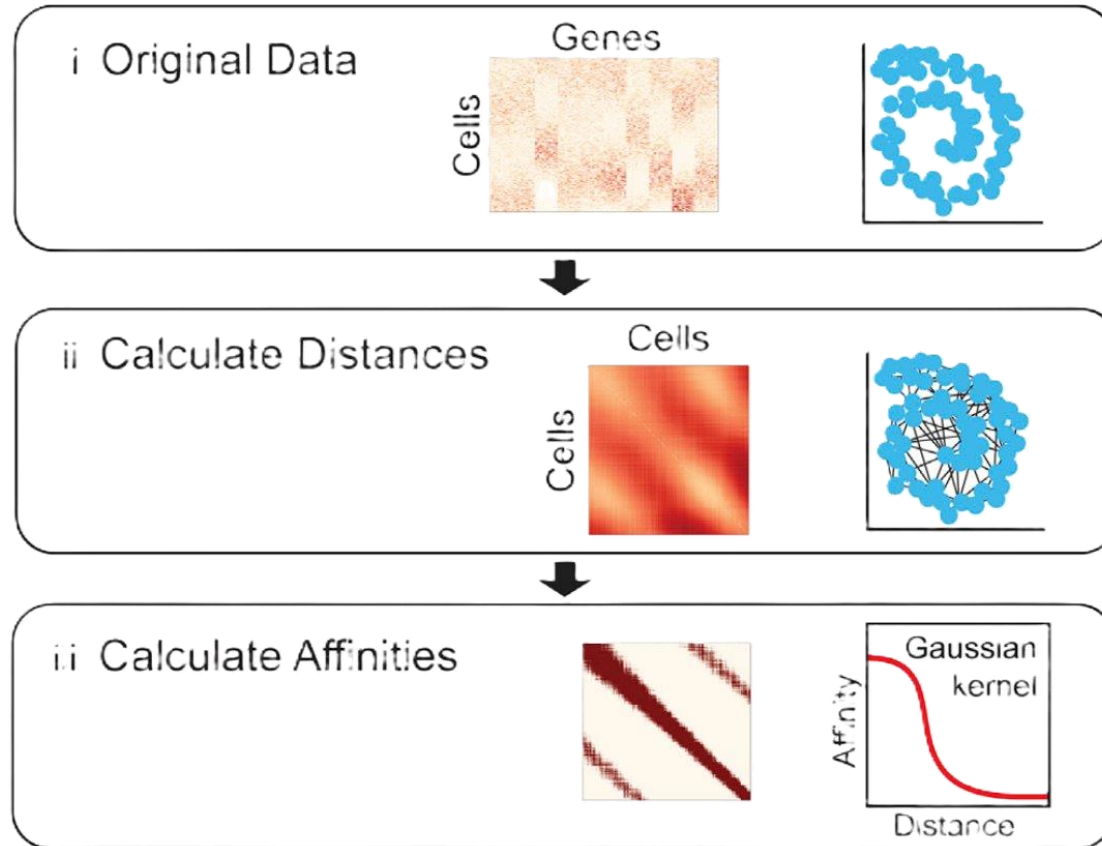
And actually this problem is much more general concept because every number in a Gene expression matrix is Count and they are sum of the measured zero/one.

$$\text{Ex) } 7 = 1 + \cancel{1} + 1 + 0 + \cancel{1} + 0 + 0 + 1 + 0 + 1 + \cancel{1} + 1 + 1 + 1 + 0$$

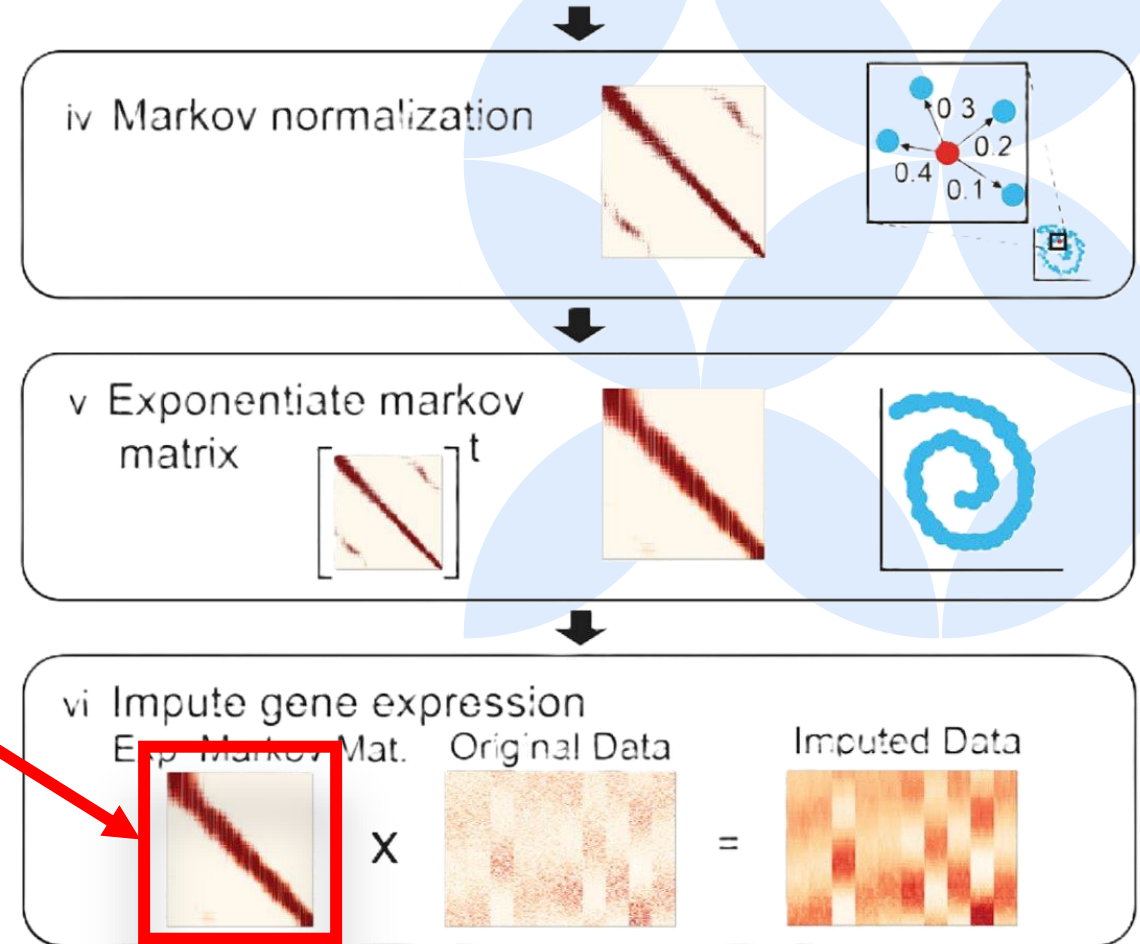
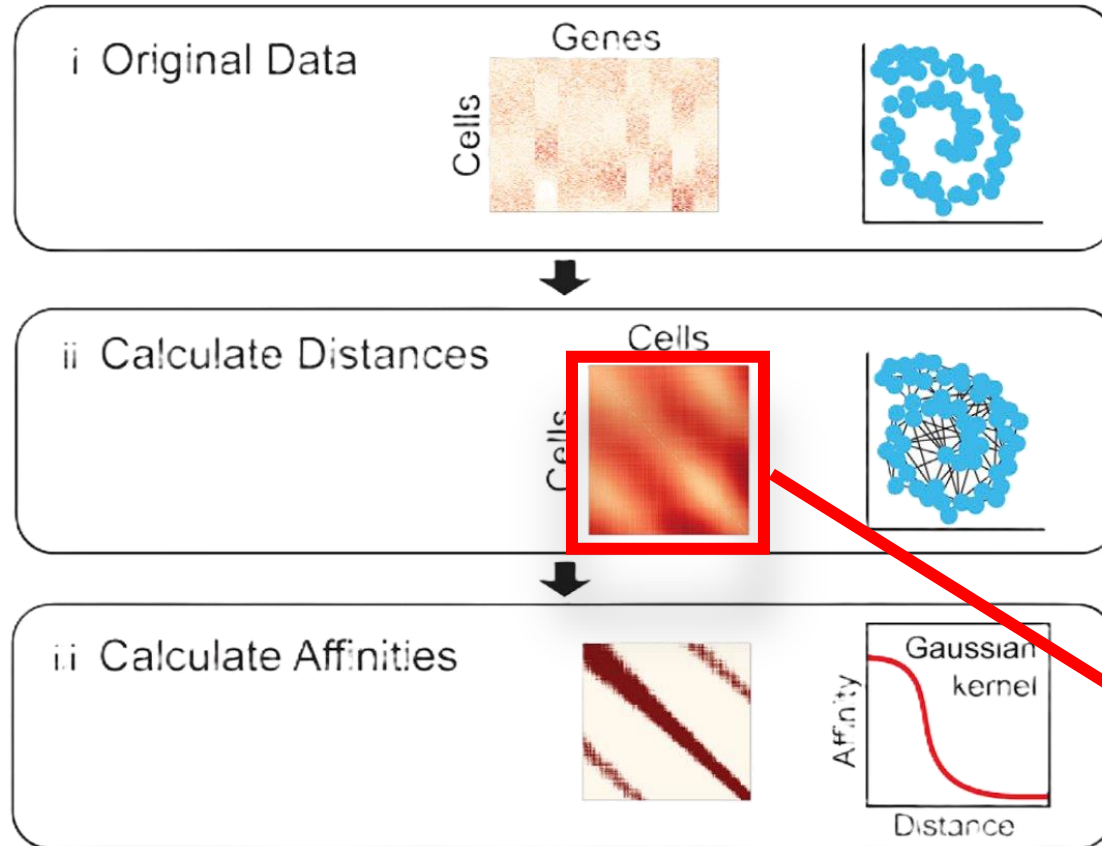
# THE CONCEPTUAL FOUNDATION OF THE ALGORITHM

1. If two cells exhibit similar global gene expression patterns, then it is likely that they also share similar expression values at individual genes.
2. In an expression matrix, a zero value can either represent a true biological zero or result from insufficient sequencing depth or technical limitation.  
(True Zero vs Technical Dropout)
3. Background Statistical theories : Diffusion Probability & Markov Chain Process  
(To construct the algorithm, not to explain the real phenomena)

# ALGORITHM ABSTRACT



# ALGORITHM ABSTRACT



## GETTING TRANSITION MATRIX FROM DISTANCE MATRIX

|        | Gene A | Gene B | Gene C | Gene D | Gene F | Gene G |
|--------|--------|--------|--------|--------|--------|--------|
| Cell A | 6      | 2      | 4      | 0      | 3      | 7      |
| Cell B | 5      | 1      | 3      | 0      | 1      | 3      |
| Cell C | 1      | 6      | 1      | 7      | 2      | 0      |
| Cell D | 5      | 1      | 0      | 3      | 1      | 4      |

Gene Expression Matrix

|       | cellA | cellB | cellC | cellD |
|-------|-------|-------|-------|-------|
| cellA | 0     | 4.79  | 12.2  | 6.3   |
| cellB | 4.79  | 0     | 10.2  | 4.4   |
| cellC | 12.2  | 10.2  | 0     | 8.7   |
| cellD | 6.3   | 4.4   | 8.7   | 0     |

Distance Matrix

$$? = \sum \text{Similarity} \times \text{Gene G} = 4.79 \times 7 + 0 \times 3 + 10.2 \times 0 + 4.4 \times 4$$

### 1. Calculating Distance Matrix among cells

## GETTING TRANSITION MATRIX FROM DISTANCE MATRIX

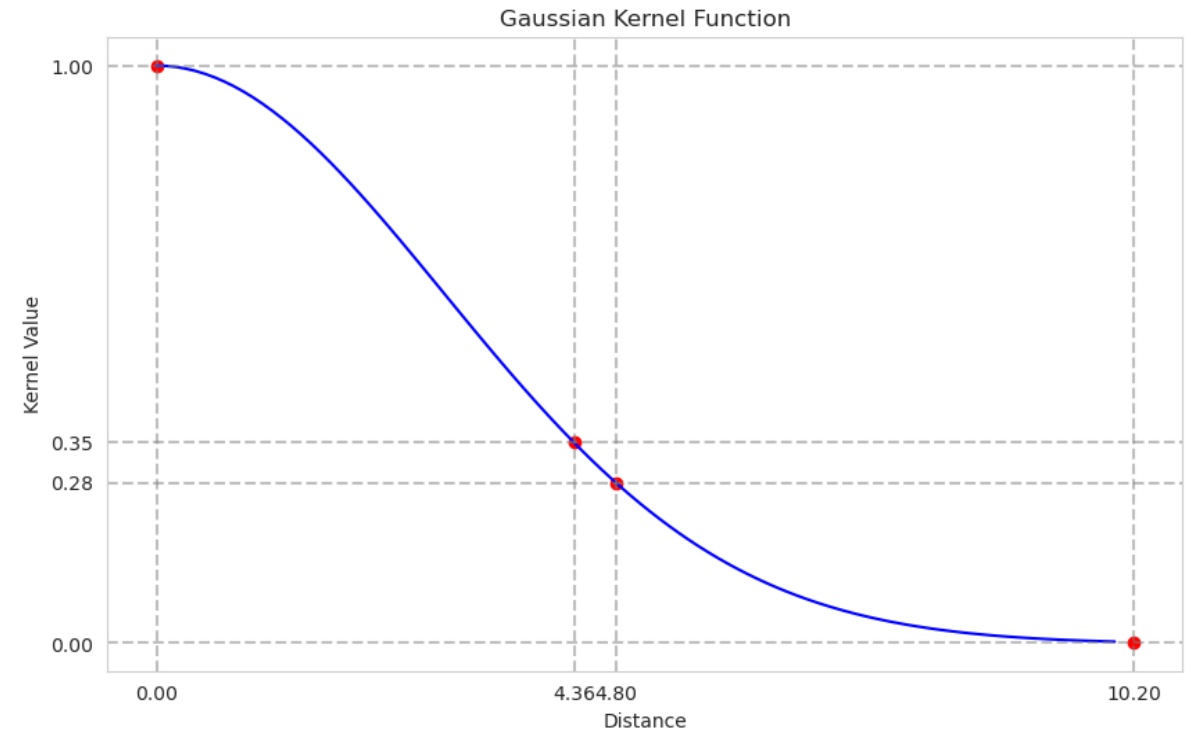
$$? = \sum \text{Similarity} \times \text{Gene G} = 4.79 \times 7 + 0 \times 3 + 10.2 \times 0 + 4.4 \times 4$$

$$? = \sum \text{Similarity} \times \text{Gene G} = 0.35 \times 7 + 1 \times 3 + 0.00 \times 0 + 0.28 \times 4$$

## 2. Applying Gaussian Kernel, and transforming the distance into affinity

It increases the weight for nearby points and decreases it for distant ones.

MAGIC used adapted Gaussian Kernel, which assign different sigma per values.



$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

## GETTING TRANSITION MATRIX FROM DISTANCE MATRIX

$$? = \sum \text{Similarity} \times \text{Gene G} = 0.35 \times 7 + 1 \times 3 + 0.00 \times 0 + 0.28 \times 4$$

$$? = \sum \text{Similarity} \times \text{Gene G} = 0.17 \times 7 + 0.61 \times 3 + 0.00 \times 0 + 0.21 \times 4$$

### 3. Transforming the affinity matrix into Markov transition matrix

a.k.a. Row wise normalization

$$[0.35, 1, 0, 0.28] / (0.35+1+0+0.28) = [0.17, 0.61, 0.00, 0.21]$$

Cf). By normalizing the affinity matrix into a Markov transition matrix, the process is modeled as a Markov chain, allowing the application of theoretical tools such as stationary distributions, spectral analysis (via eigenvalues), and convergence behavior.

## GETTING TRANSITION MATRIX FROM DISTANCE MATRIX

$$? = \sum \text{Similarity} \times \text{Gene G} = 0.17 \times 7 + 0.61 \times 3 + 0.00 \times 0 + 0.21 \times 4 = 3.86$$

### 4. Applying Transition Matrix at Original Matrix

|       | cellA | cellB | cellC | cellD |
|-------|-------|-------|-------|-------|
| cellA | 0.72  | 0.20  | 0     | 0.08  |
| cellB | 0.17  | 0.61  | 0     | 0.21  |
| cellC | 0.00  | 0.00  | 0.98  | 0.02  |
| cellD | 0.07  | 0.24  | 0.01  | 0.68  |

Transition Matrix



|        | Gene A | Gene B | Gene C | Gene D | Gene F | Gene G |
|--------|--------|--------|--------|--------|--------|--------|
| Cell A | 6      | 2      | 4      | 0      | 3      | 7      |
| Cell B | 5      | 1      | 3      | 0      | 1      | 3      |
| Cell C | 1      | 6      | 1      | 7      | 2      | 0      |
| Cell D | 5      | 1      | 0      | 3      | 1      | 4      |

Original Matrix

$$? = \sum \text{Similarity} \times \text{Gene G} = 0.17 \times 7 + 0.61 \times 3 + 0.00 \times 0 + 0.21 \times 4 = 3.86$$

|       | cellA | cellB | cellC | cellD |
|-------|-------|-------|-------|-------|
| cellA | 0.72  | 0.20  | 0     | 0.08  |
| cellB | 0.17  | 0.61  | 0     | 0.21  |
| cellC | 0.00  | 0.00  | 0.98  | 0.02  |
| cellD | 0.07  | 0.24  | 0.01  | 0.68  |

Transition Matrix



|        | Gene A | Gene B | Gene C | Gene D | Gene F | Gene G |
|--------|--------|--------|--------|--------|--------|--------|
| Cell A | 6      | 2      | 4      | 0      | 3      | 7      |
| Cell B | 5      | 1      | 3      | 0      | 1      | 3      |
| Cell C | 1      | 6      | 1      | 7      | 2      | 0      |
| Cell D | 5      | 1      | 0      | 3      | 1      | 4      |

Original Matrix



|       | Gene A | Gene B | Gene C | Gene D | Gene F | Gene G |
|-------|--------|--------|--------|--------|--------|--------|
| cellA | 5.72   | 1.72   | 3.48   | 0.24   | 2.44   | 5.96   |
| cellB | 5.12   | 1.16   | 2.51   | 0.63   | 1.33   | 3.86   |
| cellC | 1.08   | 5.90   | 0.98   | 6.92   | 1.98   | 0.08   |
| cellD | 5.03   | 1.12   | 1.01   | 2.11   | 1.15   | 3.93   |

Diffused/Imputed Matrix

## APPLYING TRANSITION MATRIX T-TIMES

$t$

|       | cellA | cellB | cellC | cellD |
|-------|-------|-------|-------|-------|
| cellA | 0.72  | 0.20  | 0     | 0.08  |
| cellB | 0.17  | 0.61  | 0     | 0.21  |
| cellC | 0.00  | 0.00  | 0.98  | 0.02  |
| cellD | 0.07  | 0.24  | 0.01  | 0.68  |

Transition Matrix



|        | Gene A | Gene B | Gene C | Gene D | Gene F | Gene G |
|--------|--------|--------|--------|--------|--------|--------|
| Cell A | 6      | 2      | 4      | 0      | 3      | 7      |
| Cell B | 5      | 1      | 3      | 0      | 1      | 3      |
| Cell C | 1      | 6      | 1      | 7      | 2      | 0      |
| Cell D | 5      | 1      | 0      | 3      | 1      | 4      |

Original Matrix

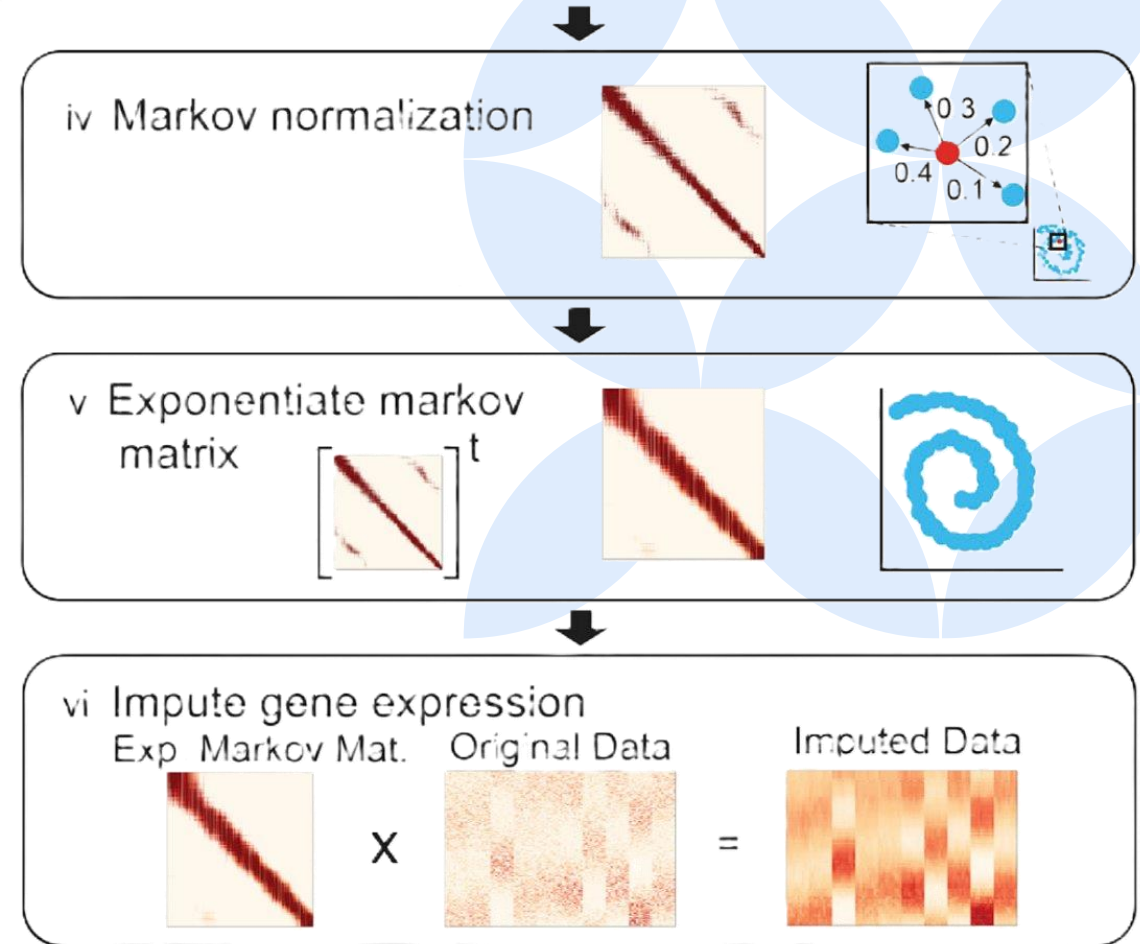
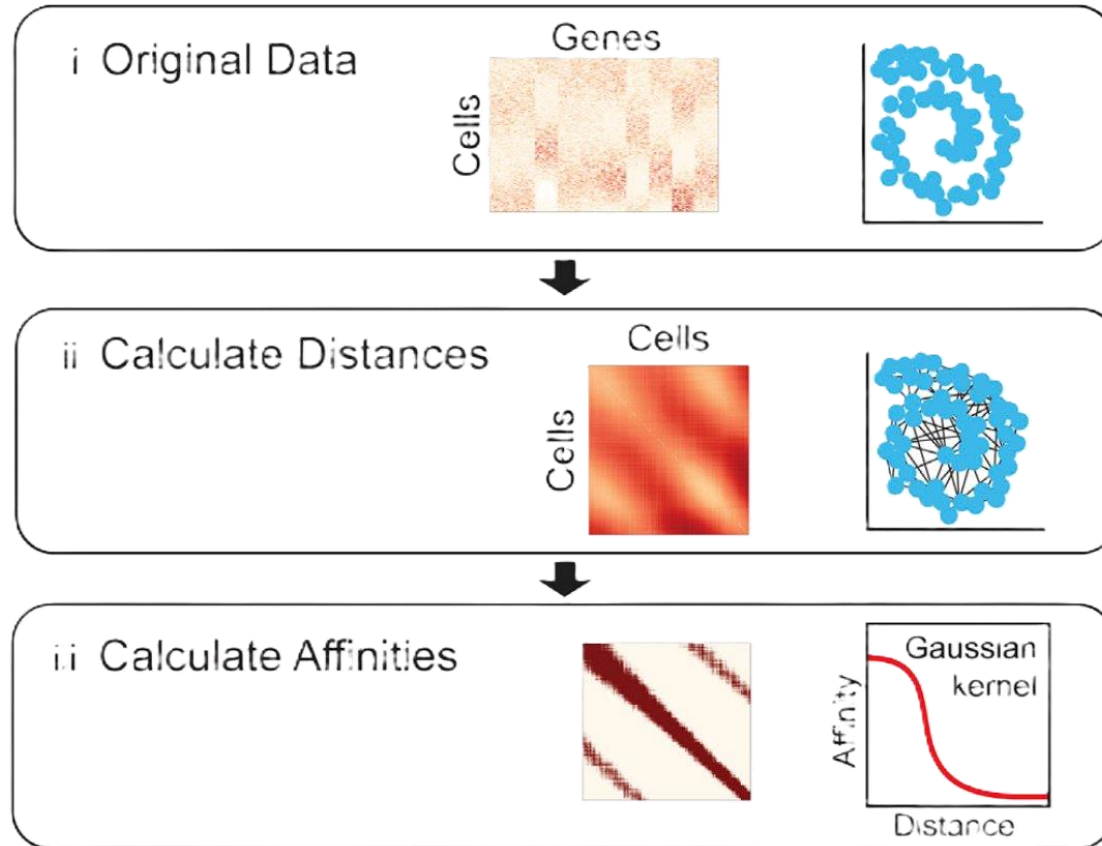
$$D = M \times O$$

$$D_2 = M \times D = M \times M \times O = M^2 \times O$$

...

$$D_t = M^t \times O$$

# ALGORITHM ABSTRACT



# Applications

Evaluating MAGIC's ability

## Before looking at the example..

The purpose of this algorithm is **not to perfectly reconstruct the data** and draw absolutely reliable conclusions.

When the data are **sparse**, many analytical tools fail to even capture the inherent potential of the data itself.

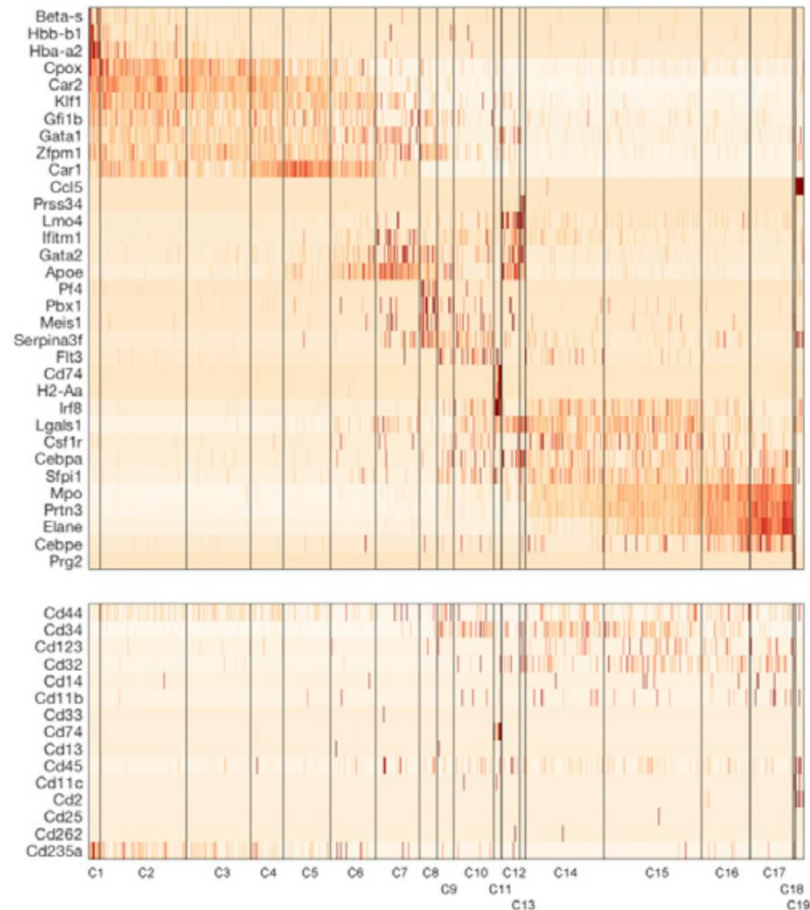
This algorithm is intended to prevent such situations.

In other words, it serves as a kind of **smoothing tool** designed to observe **general trends**, rather than focusing on every small detail.

|        | Gene A | Gene B | Gene C | Gene D | Gene F | Gene G |
|--------|--------|--------|--------|--------|--------|--------|
| Cell A | 6      | 2      | 4      | 0      | 3      | 7      |
| Cell B | 5      | 1      | 3      | 0      | 1      | 3      |
| Cell C | 1      | 6      | 1      | 7      | 2      | 0      |
| Cell D | 5      | 1      | 0      | 3      | 1      | 4      |

|        | Gene A | Gene B | Gene C | Gene D | Gene F | Gene G |
|--------|--------|--------|--------|--------|--------|--------|
| Cell A | 5.72   | 1.72   | 3.48   | 0.24   | 2.44   | 5.96   |
| Cell B | 5.12   | 1.16   | 2.51   | 0.63   | 1.33   | 3.86   |
| Cell C | 1.08   | 5.90   | 0.98   | 6.92   | 1.98   | 0.08   |
| Cell D | 5.03   | 1.12   | 1.01   | 2.11   | 1.15   | 3.93   |

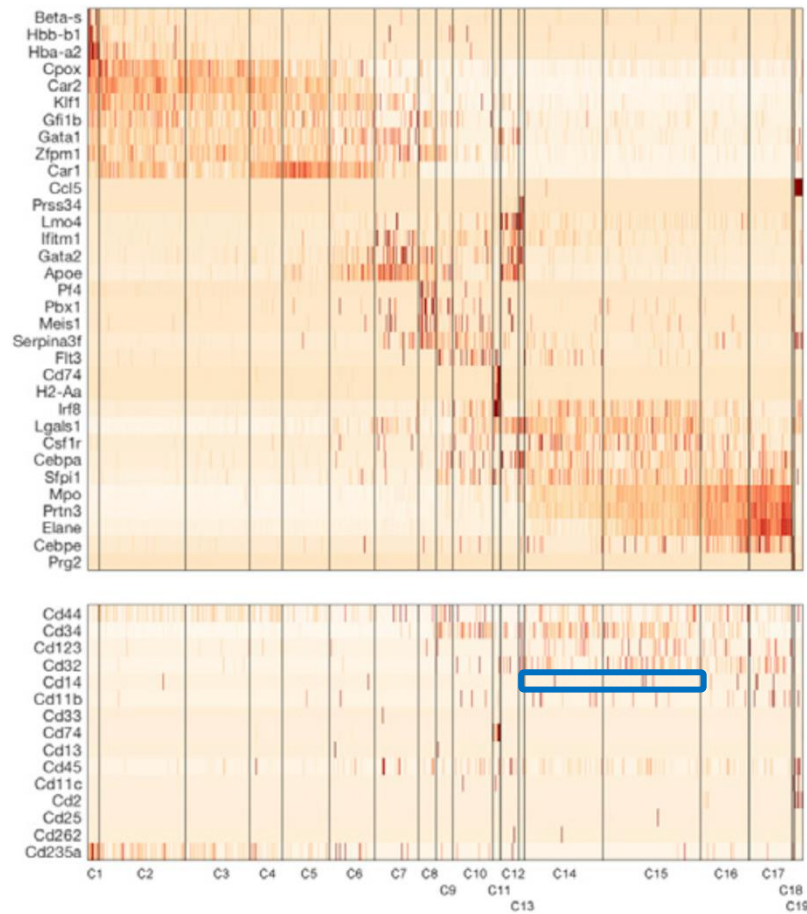
Figure 2. Application on scRNA seq of mouse bone marrow (Paul et al., 2015)



CD14 (Monocyte Marker Genes) appears only  
1.6% cells at Monocyte Cluster (C14,C15)

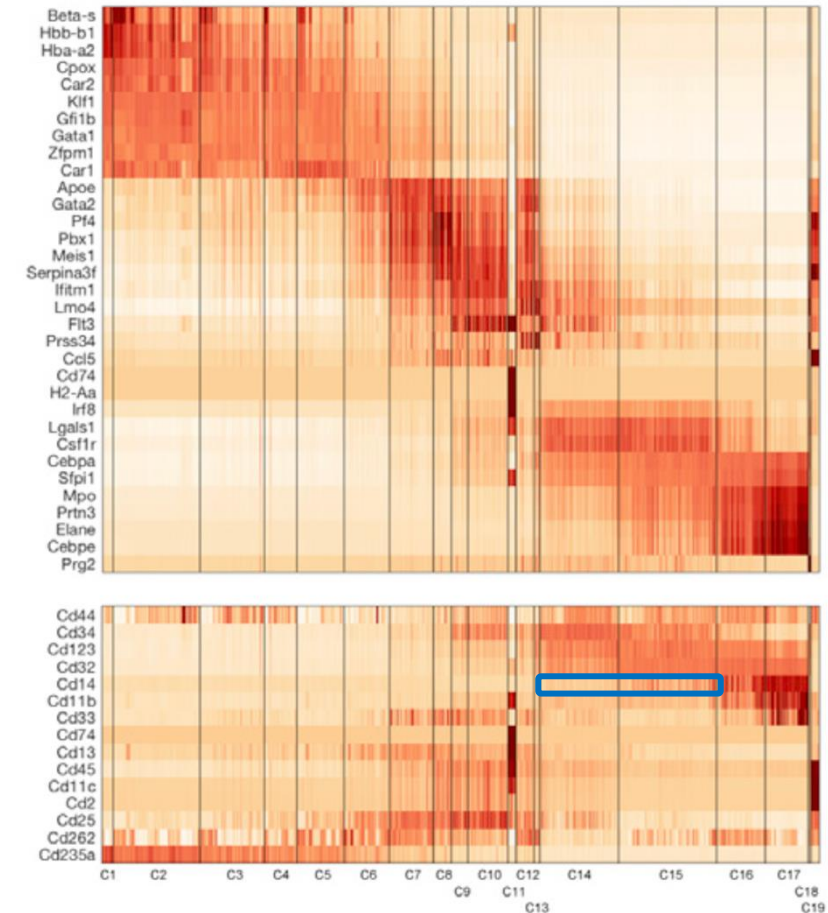
CD32 (Dendritic Marker Genes) appears only  
10% cells at Dendritic Cluster (C11)

Figure 2. Application on scRNA seq of mouse bone marrow (Paul et al., 2015)



CD14 (Monocyte Marker Genes) appears only 1.6% cells at Monocyte Cluster (C14,C15)

CD32 (Dendritic Marker Genes) appears only 10% cells at Dendritic Cluster (C11)



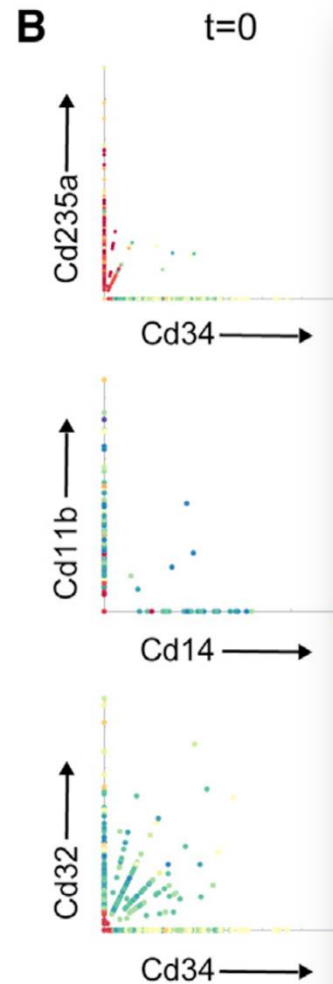
CD14 (Monocyte Marker Genes) appears only 94% cells at Monocyte Cluster (C14,C15)

CD32 (Dendritic Marker Genes) appears only 97% cells at Dendritic Cluster (C11)

Figure 2. Application on scRNA seq of mouse bone marrow (Paul et al., 2015)

**B. Scatter Plot of the known marker genes.**

Magic restores missing values and relationships, recreating the biaxial plots typically seen in flow cytometry.



- erythrocyte C1
- erythrocyte C2
- erythrocyte C3
- erythrocyte C4
- erythrocyte C5
- erythrocyte C6
- early erythrocyte C7
- megakaryocyte C8
- early neutrophil C9
- early monocyte C10
- dendritic cells C11
- early basophil C12
- basophil C13
- monocyte C14
- monocyte C15
- neutrophil C16
- neutrophil C17
- eosinophil C18
- lymphoid progenitors (NK) C19

Figure 2. Application on scRNA seq of mouse bone marrow (Paul et al., 2015)

**B. Scatter Plot of the known marker genes.**

Magic restores missing values and relationships, recreating the biaxial plots typically seen in flow cytometry.

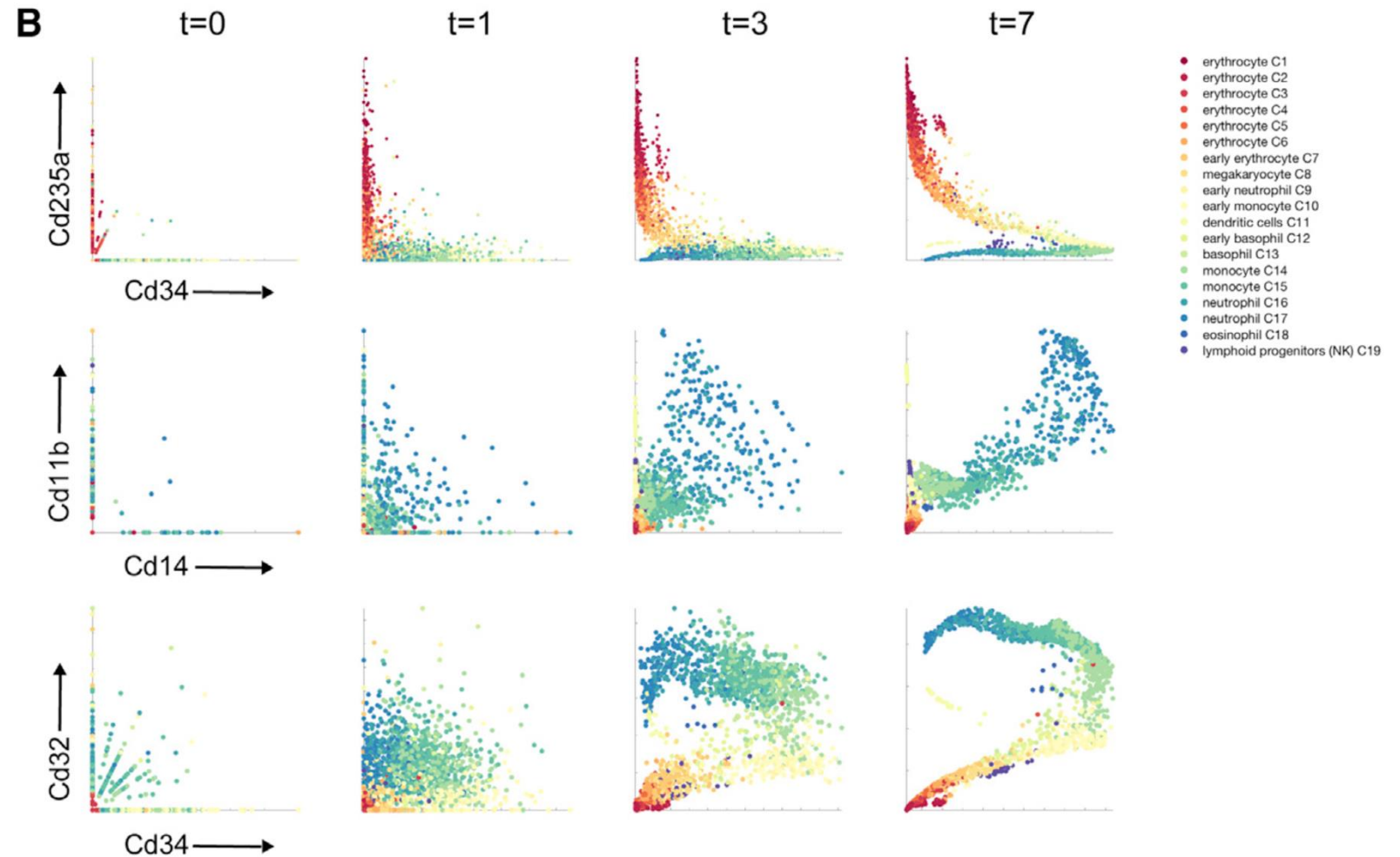


Figure 2. Application on scRNA seq of mouse bone marrow (Paul et al., 2015)

**C. 3-dimensional scatter plot of key genes.** Trajectory of three genes become more clear with MAGIC

**D. The scatter plot of mRNA and FACS of marker genes.** MAGIC successfully restores this relationship by denoising the gene expression data.

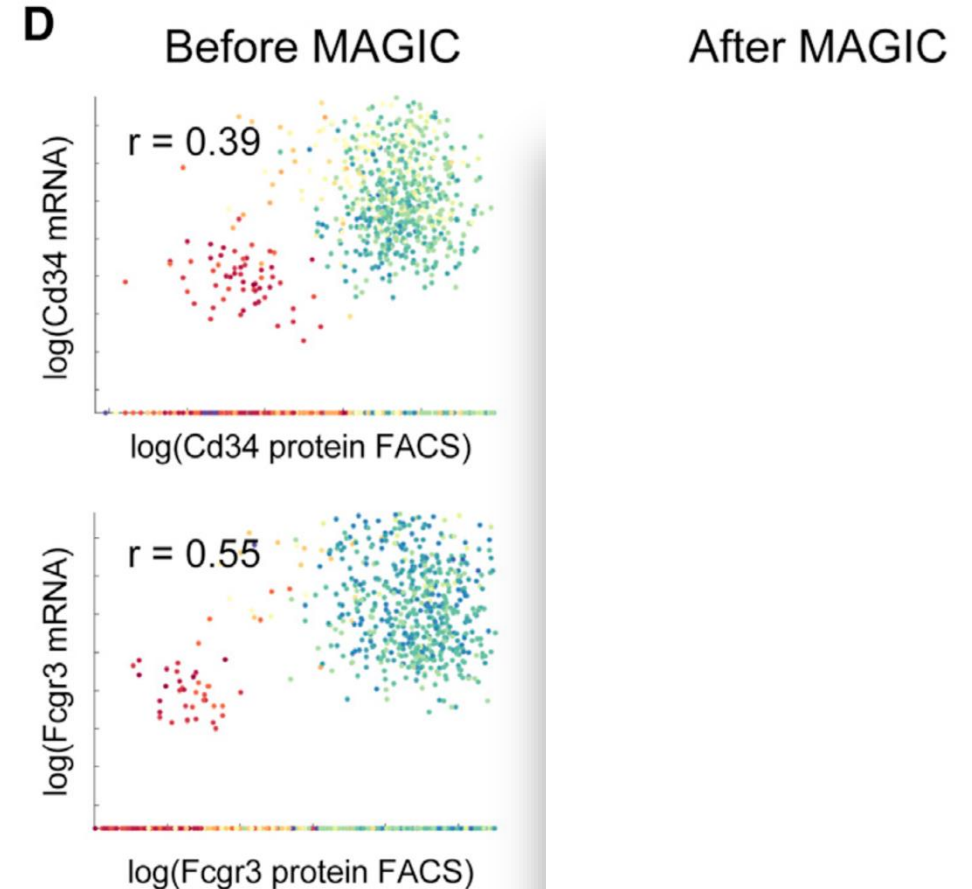
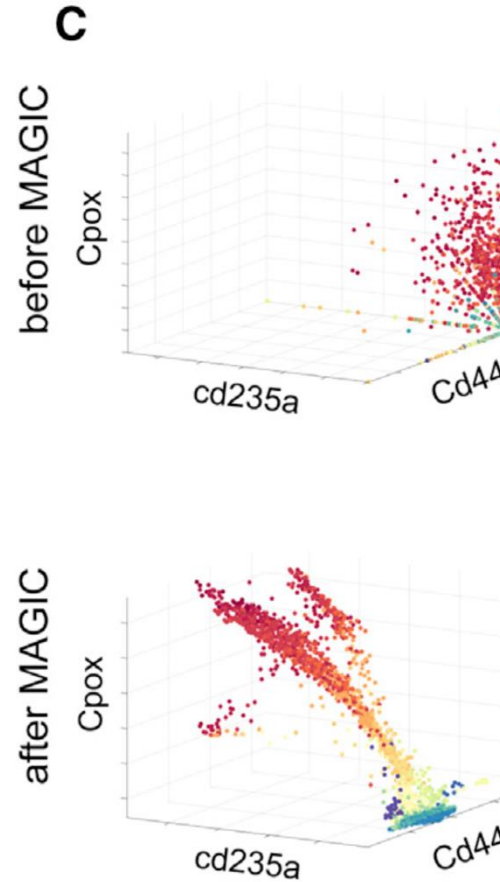
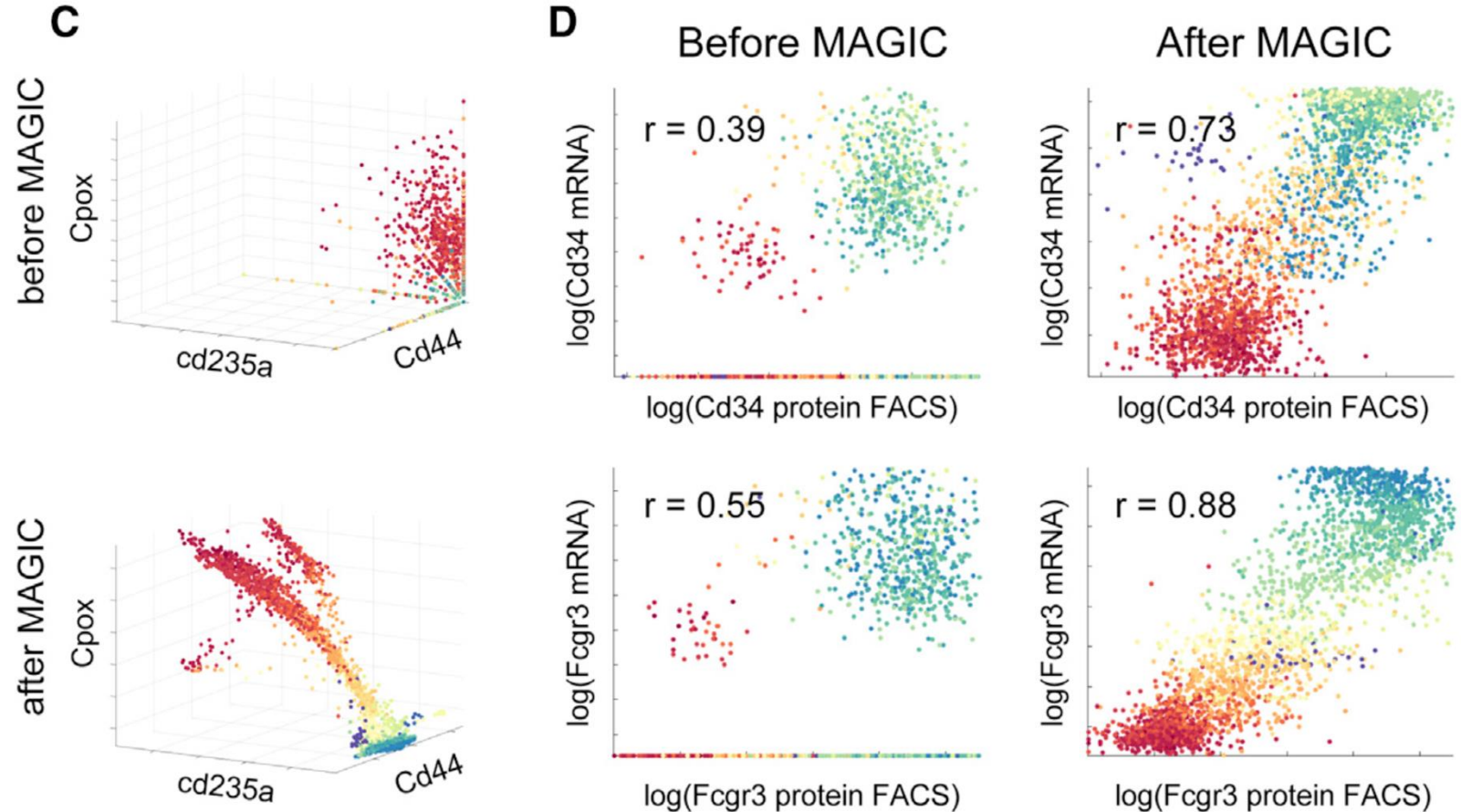


Figure 2. Application on scRNA seq of mouse bone marrow (Paul et al., 2015)

**C. 3-dimensional scatter plot of key genes.** Trajectory of three genes become more clear with MAGIC

**D. The scatter plot of mRNA and FACS of marker genes.** MAGIC successfully restores this relationship by denoising the gene expression data.



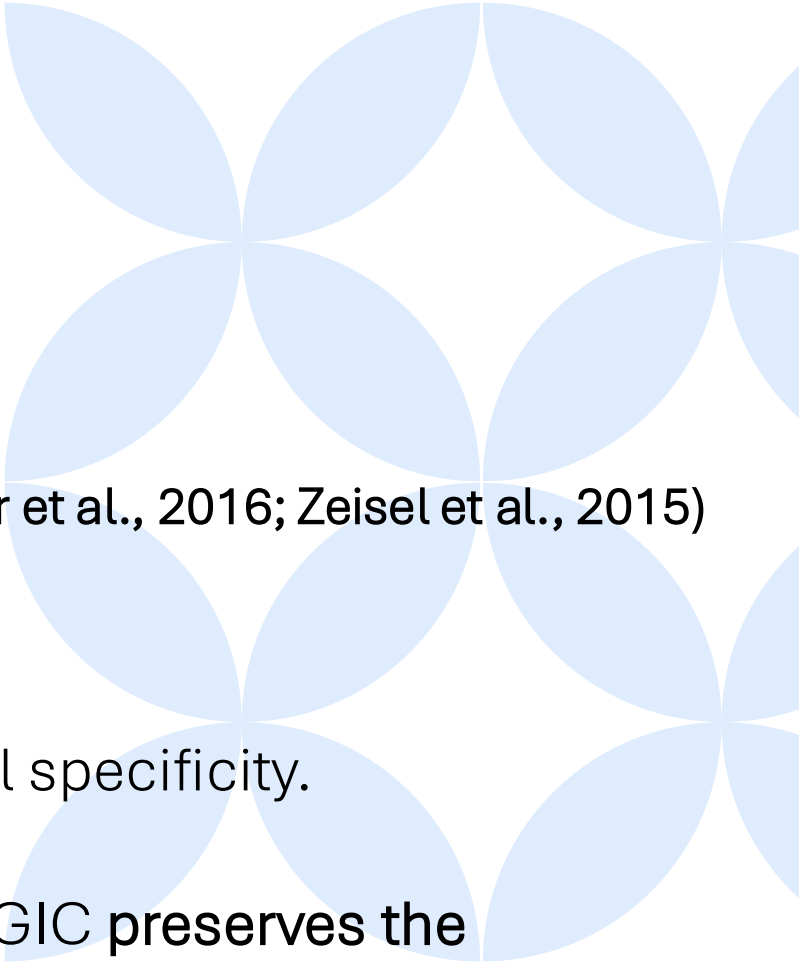
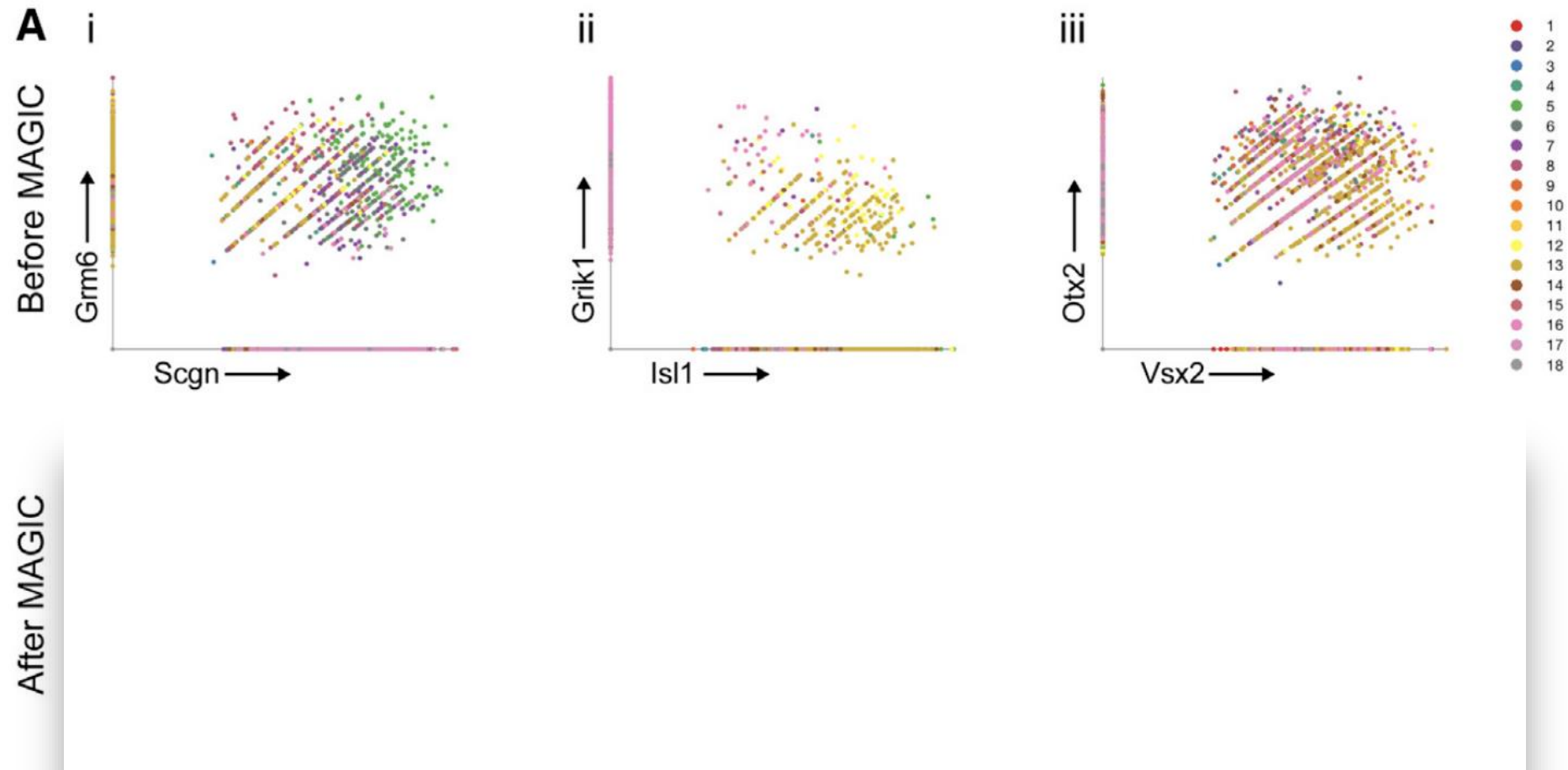


Figure 3. Application on scRNA seq of mouse retina neuronal cells (Shekhar et al., 2016; Zeisel et al., 2015)

The datasets are known to exhibit a high degree of functional specificity.

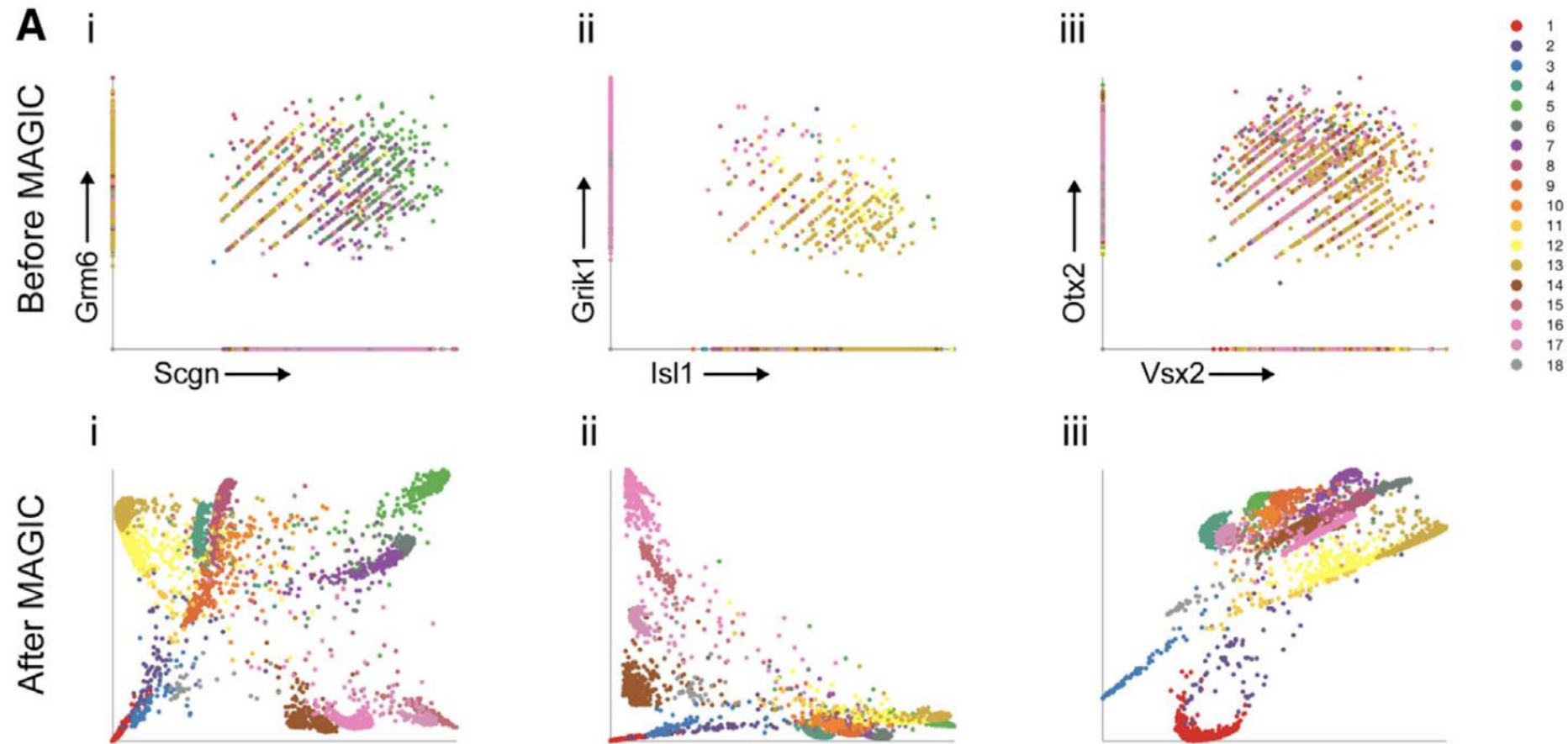
Therefore, the goal of this section is to evaluate whether MAGIC **preserves the underlying cluster structure.**

Figure 3. Application on scRNA seq of mouse retina neuronal cells (Shekhar et al., 2016; Zeisel et al., 2015)



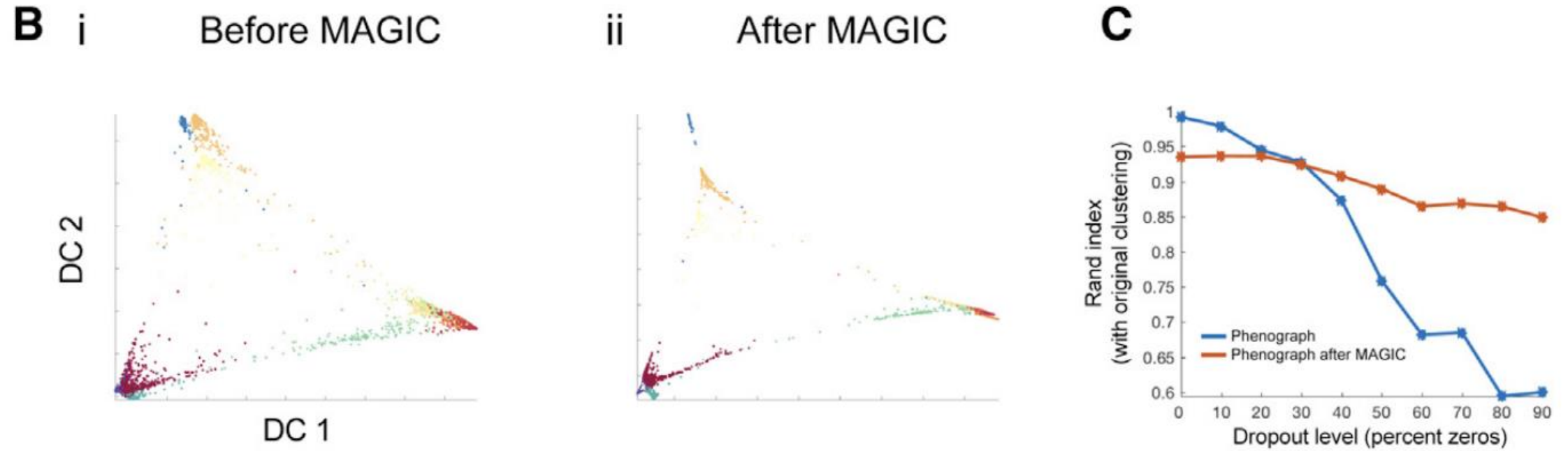
A. The scatter plot of two-marker genes before and after MAGIC

Figure 3. Application on scRNA seq of mouse retina neuronal cells (Shekhar et al., 2016; Zeisel et al., 2015)



A. The scatter plot of two-marker genes before and after MAGIC

Figure 3. Application on scRNA seq of mouse retina neuronal cells (Shekhar et al., 2016; Zeisel et al., 2015)



**B. Diffusion Component before and after magic.**

Diffusion Components are the Eigenvector of Diffusion Map  
Visualization of Cell x Cell Distance

**C. Rand Index as increasing dropout level**

Rand Index is the ratio of agreement of cluster relation.

Figure 3.D Artificial Data to evaluate MAGIC's correcting power.

| Cluster | Gene A | Gene B | Gene C |
|---------|--------|--------|--------|
| A       | 1      | 2      | 3      |
| A       | 1      | 3      | 2      |
| A       | 3      | 2      | 1      |
| ...     |        |        |        |
| B       | 7      | 8      | 9      |
| B       | 9      | 8      | 7      |
| B       | 7      | 9      | 8      |

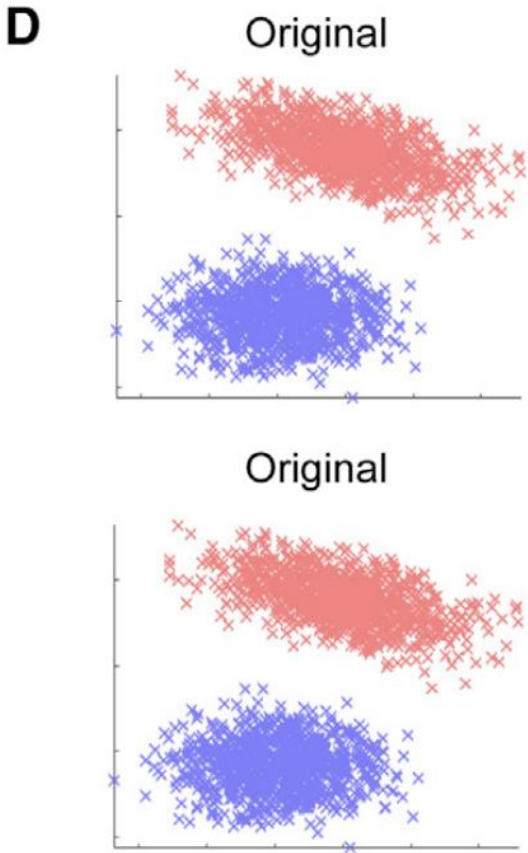


Figure 3. Application on scRNA seq of mouse retina neuronal cells (Shekhar et al., 2016; Zeisel et al., 2015)

| Cluster | Gene A | Gene B | Gene C |
|---------|--------|--------|--------|
| A       | 1      | 2      | 3      |
| A       | 7      | 3      | 8      |
| A       | 3      | 8      | 1      |
| ...     |        |        |        |
| B       | 7      | 2      | 9      |
| B       | 9      | 8      | 7      |
| B       | 1      | 9      | 2      |

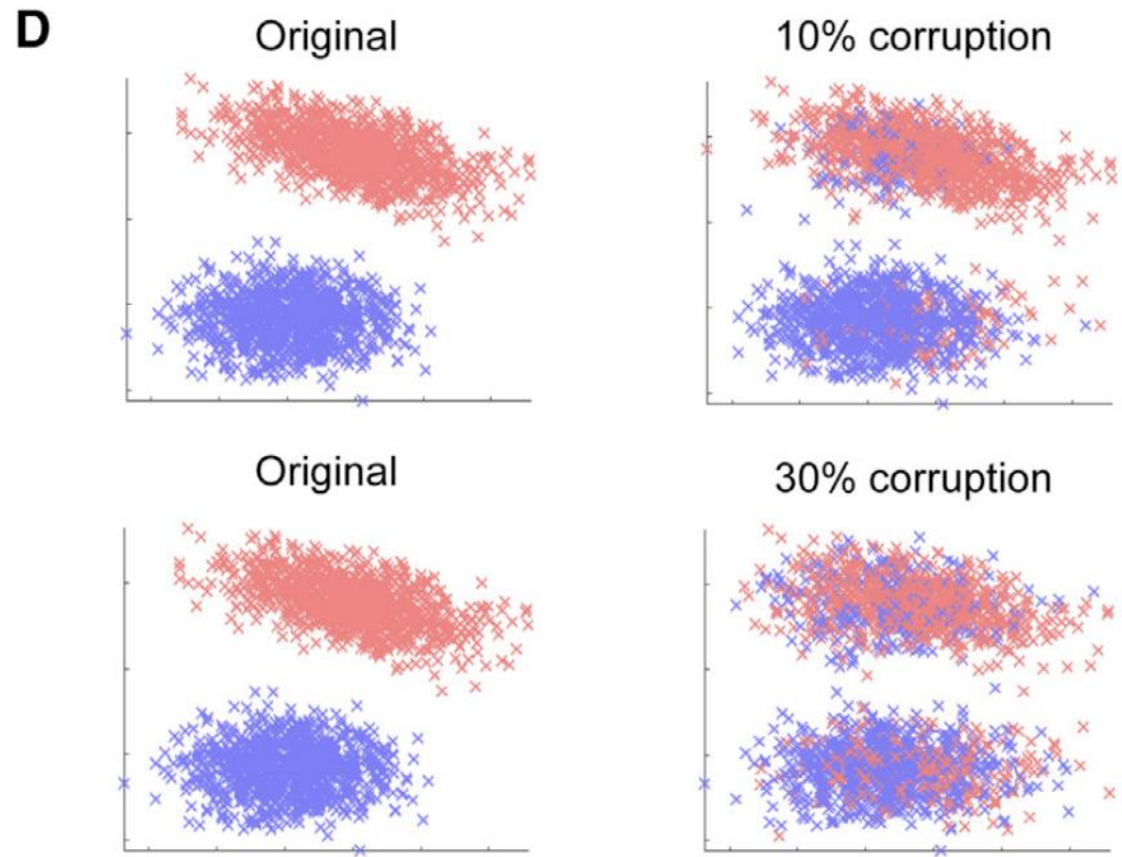


Figure 3. Application on scRNA seq of mouse retina neuronal cells (Shekhar et al., 2016; Zeisel et al., 2015)

| Cluster | Gene A | Gene B | Gene C |
|---------|--------|--------|--------|
| A       | 1      | 2      | 3      |
| A       | 7      | 3      | 8      |
| A       | 3      | 8      | 1      |
| ...     |        |        |        |
| B       | 7      | 2      | 9      |
| B       | 9      | 8      | 7      |
| B       | 1      | 9      | 2      |

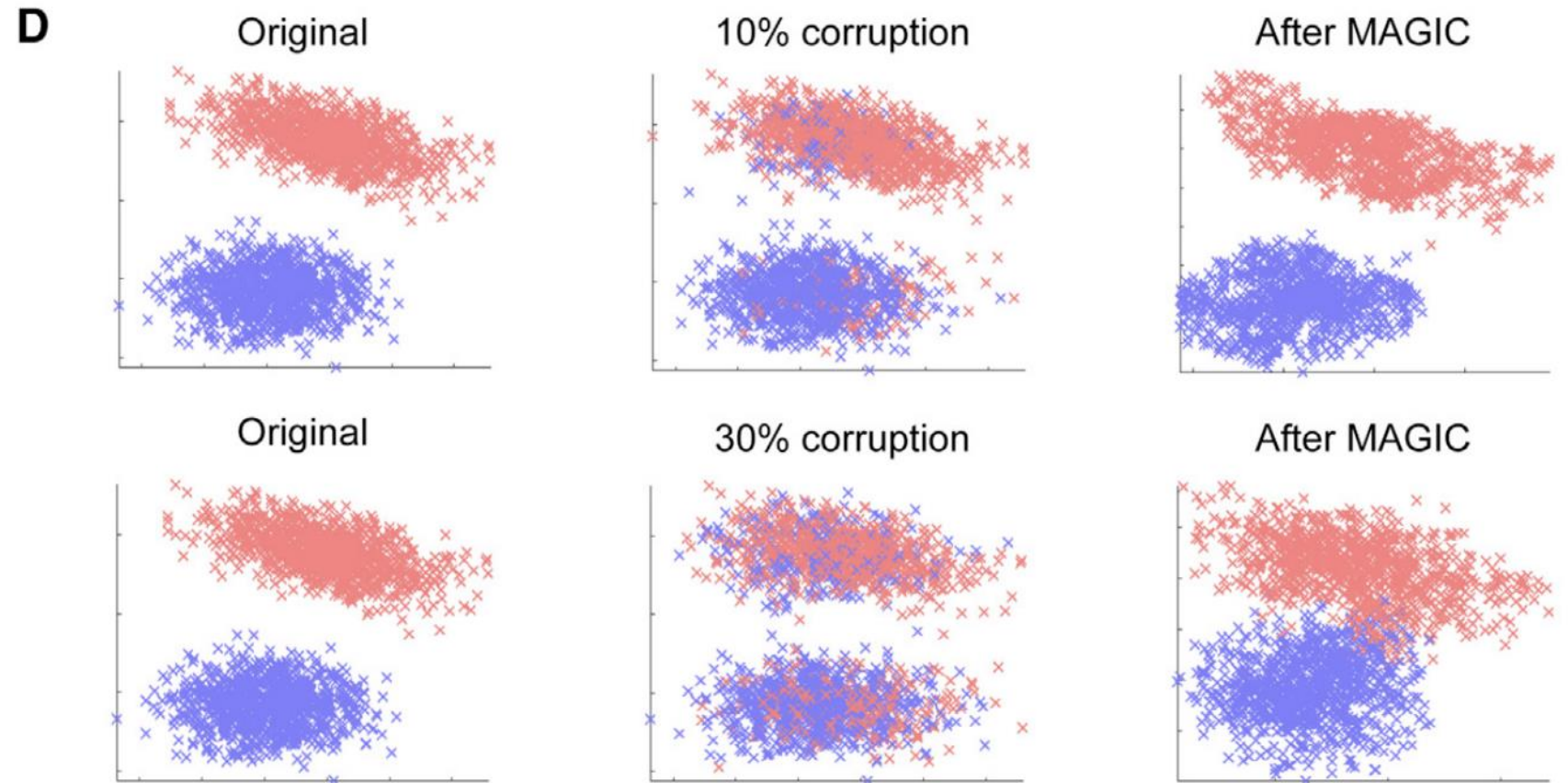
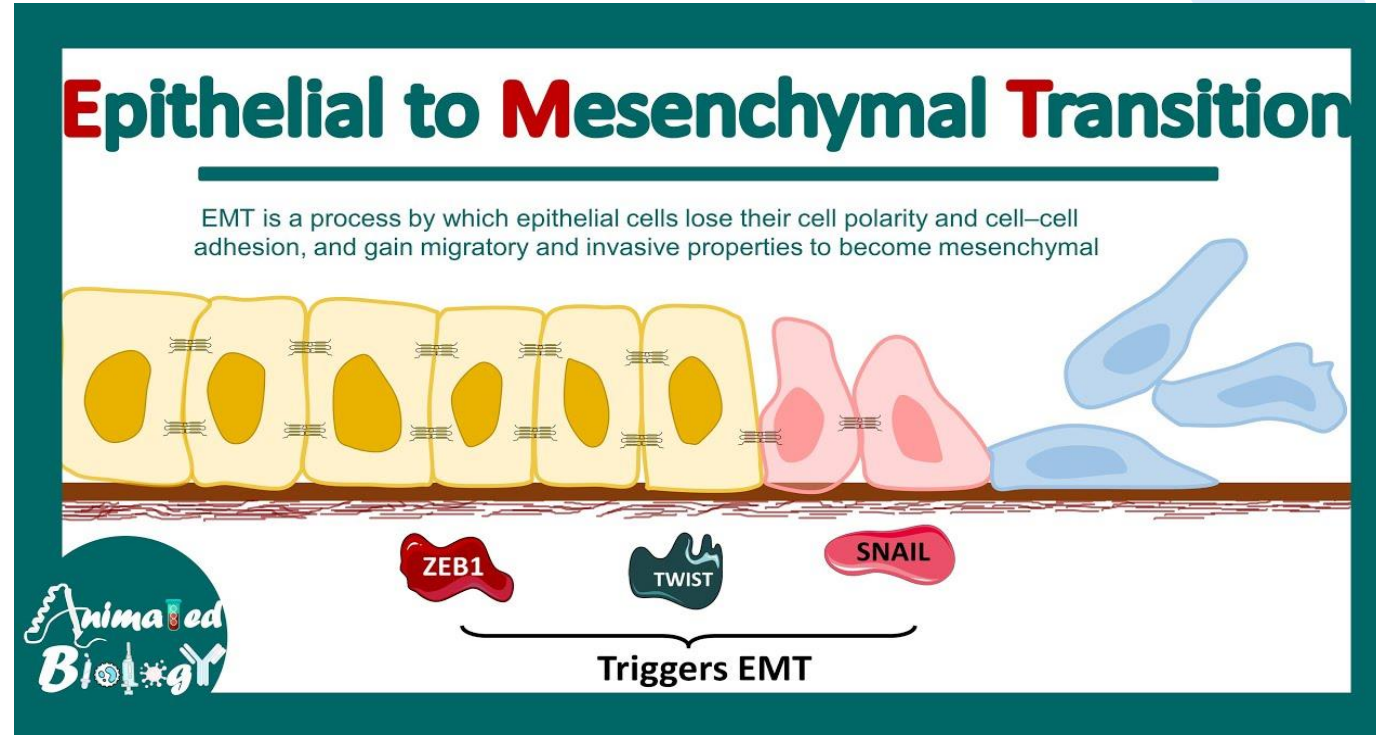


Figure 4. MAGIC revovers a state space in EMT data



### Background

- Previous studies compared epithelial vs. mesenchymal endpoints, overlooking intermediate EMT states.
- Bulk-level analyses limited resolution of EMT progression at the single-cell level.

### Focus

- Analyze gene expression changes across EMT states.

### Data

- HMLE (Human Mammary Luminal Epithelial) cell lines
- Cross-sectional snapshots of cells at different EMT stages (not time-series).

Figure 4. MAGIC revovers a state space in EMT data

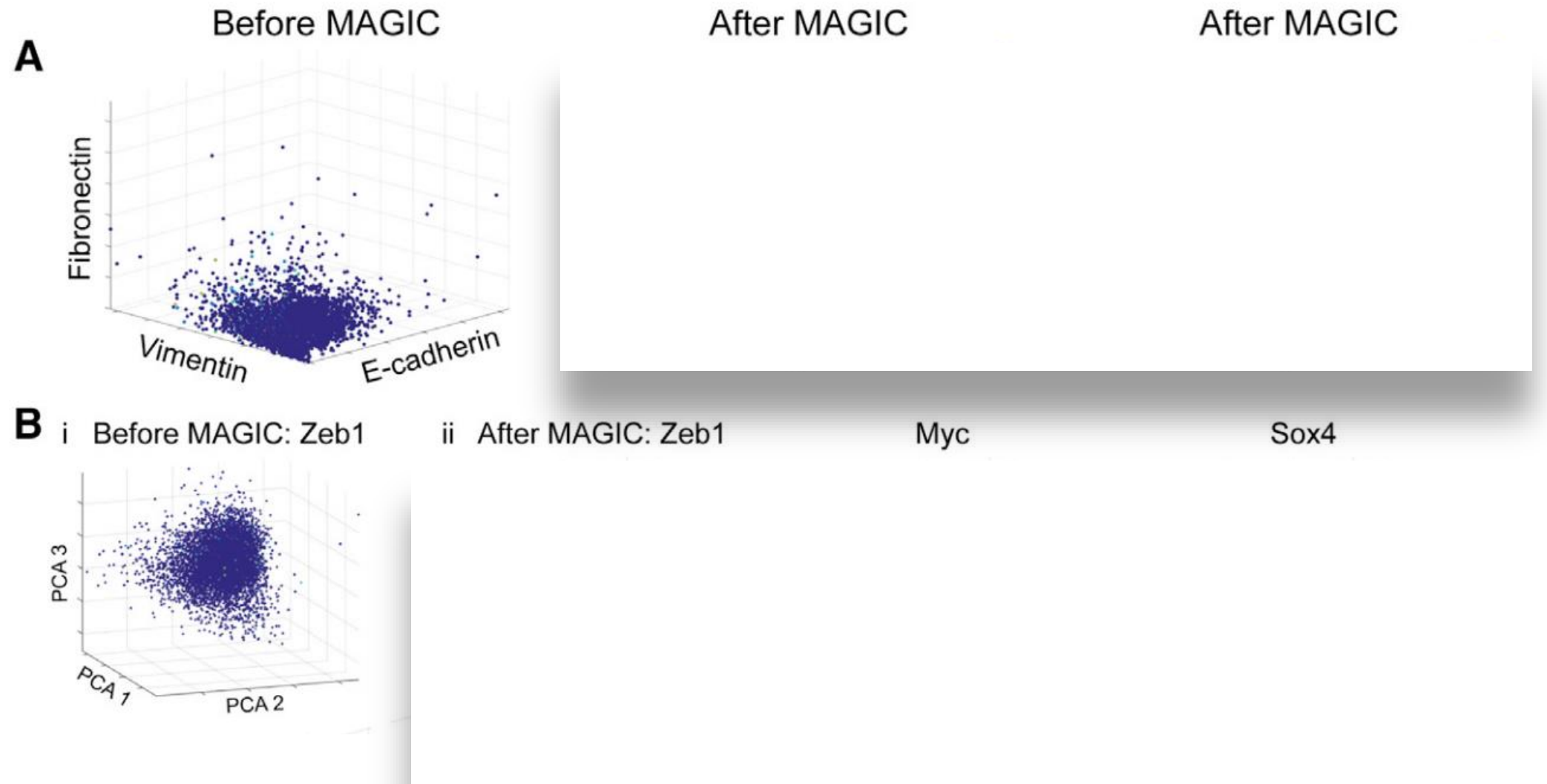


Figure 4. MAGIC revovers a state space in EMT data

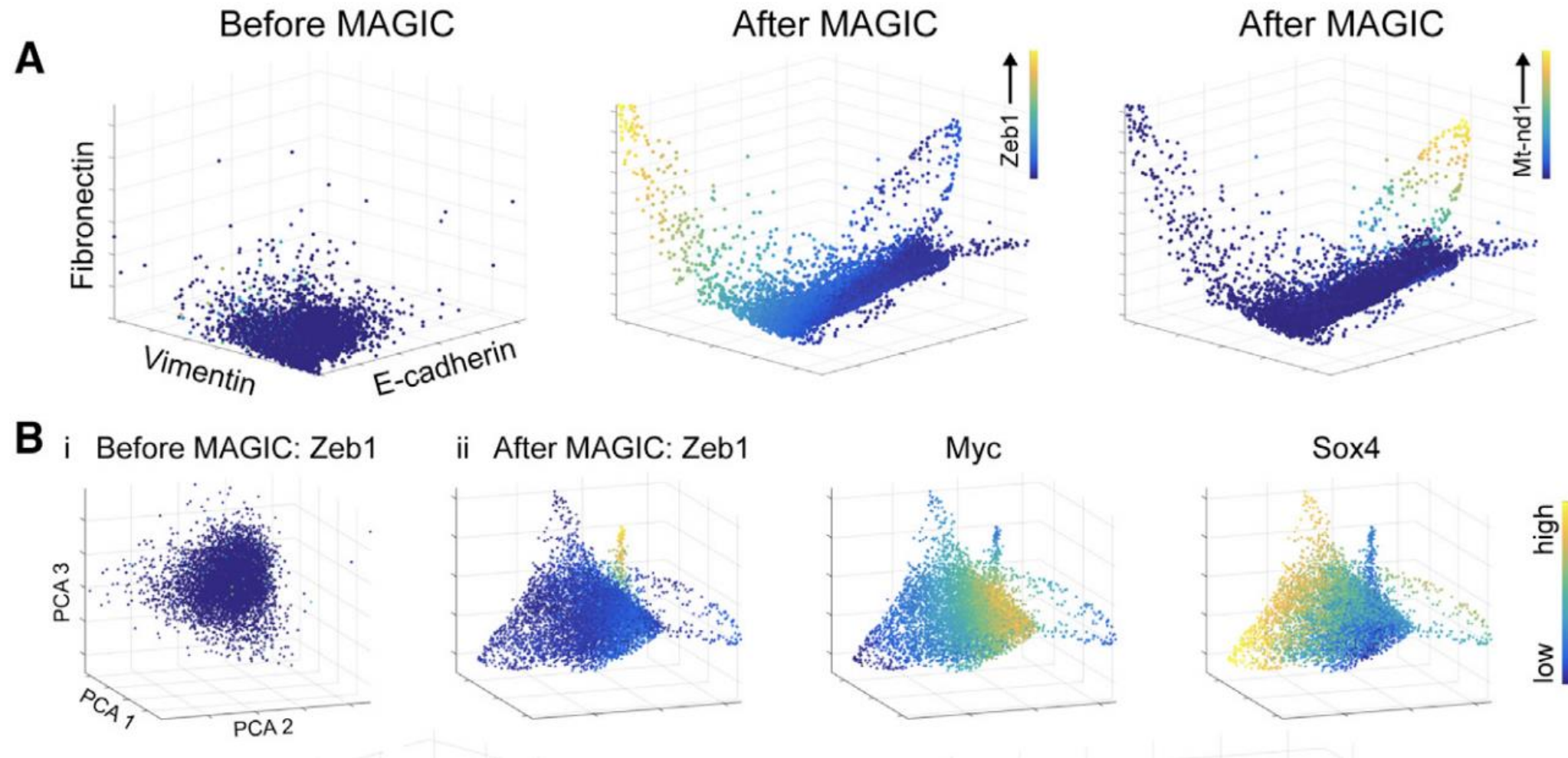


Figure 4. MAGIC revoers a state space in EMT data

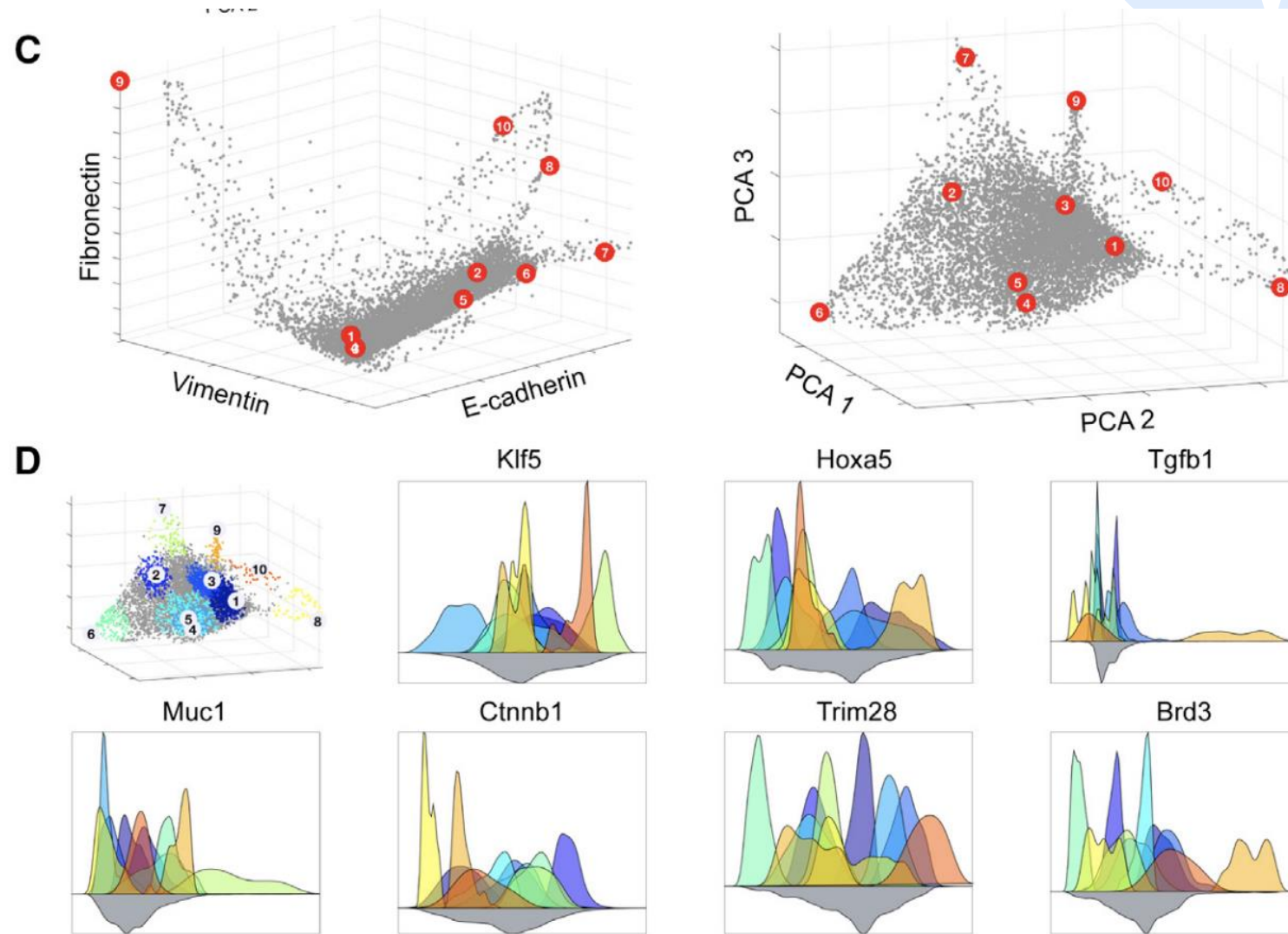


Figure 4. MAGIC revoers a state space in EMT data

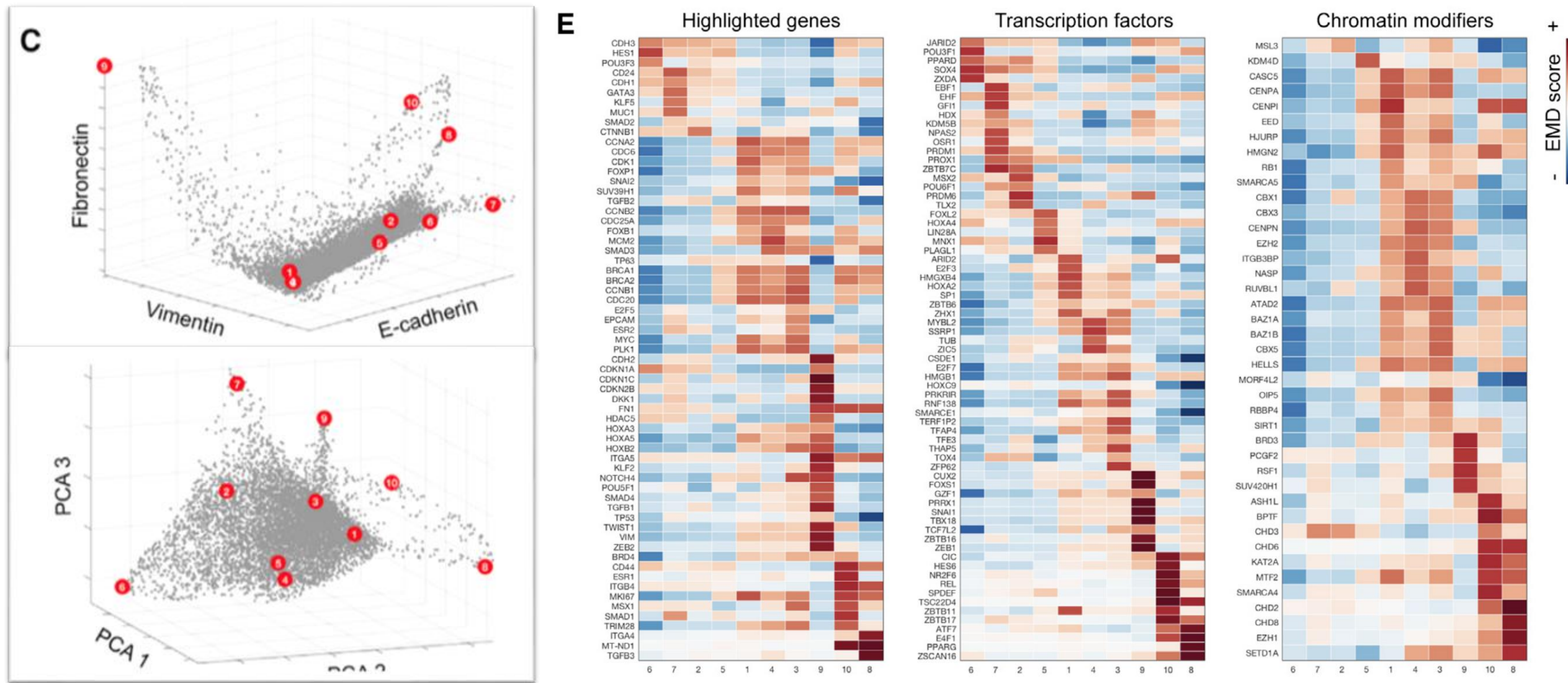
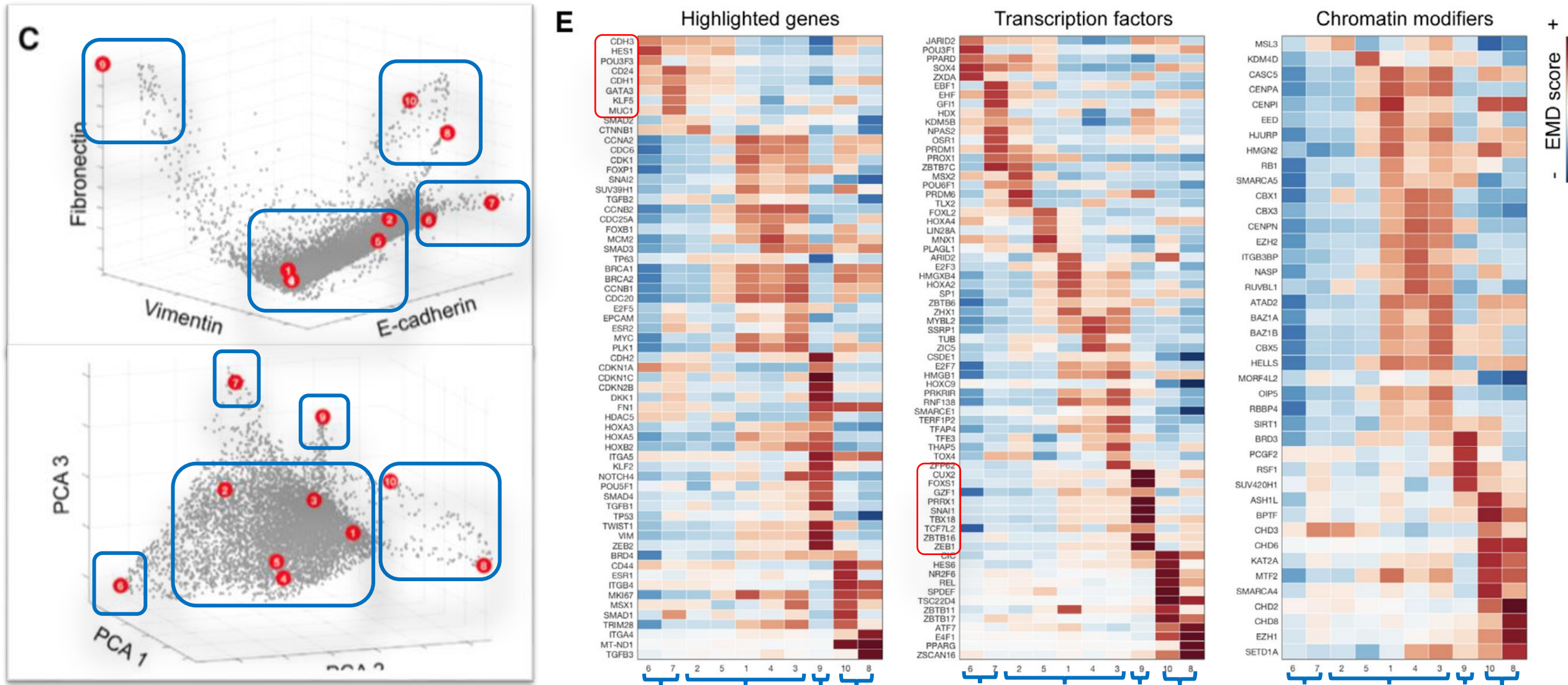


Figure 4. MAGIC revoers a state space in EMT data



AT6,7 : Epithelial, AT1-5 : Intermediate, AT9 : Mesenchymal, AT8,10 : Apoptotic

Figure 5. Intermediate Archetype of EMT data

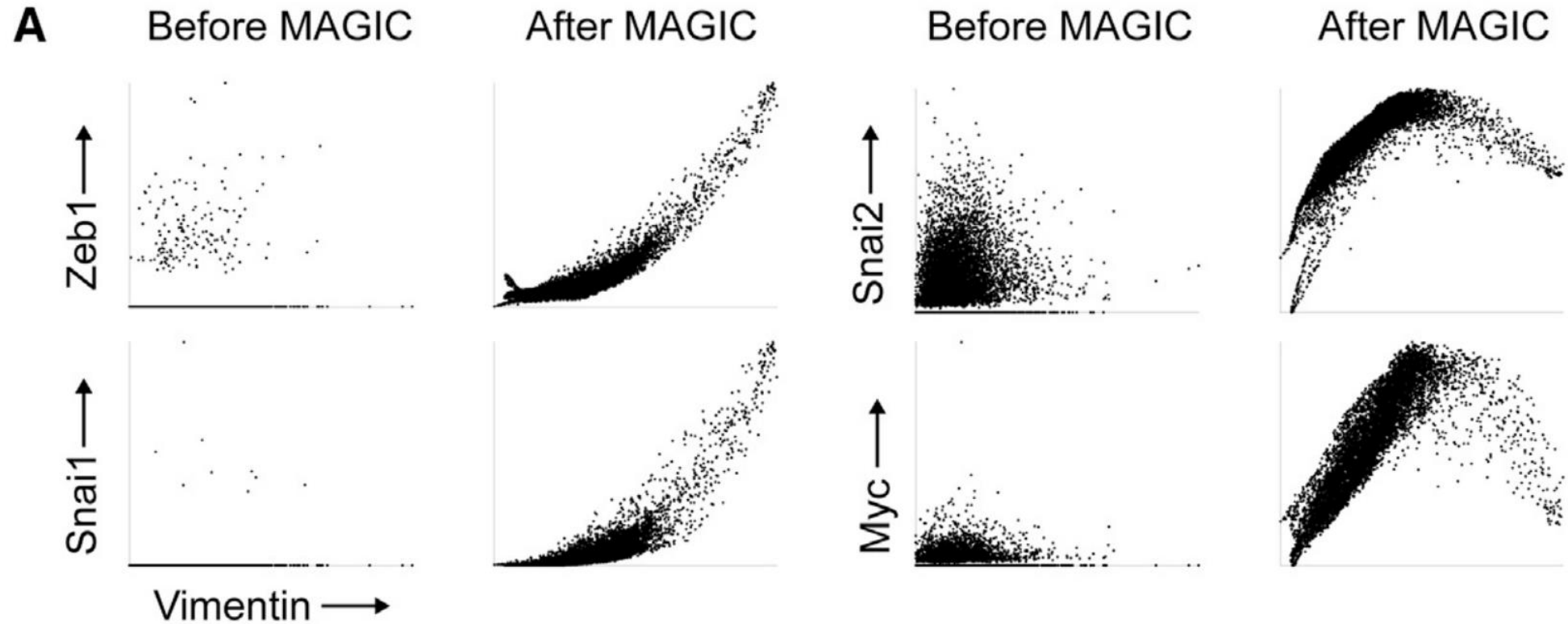
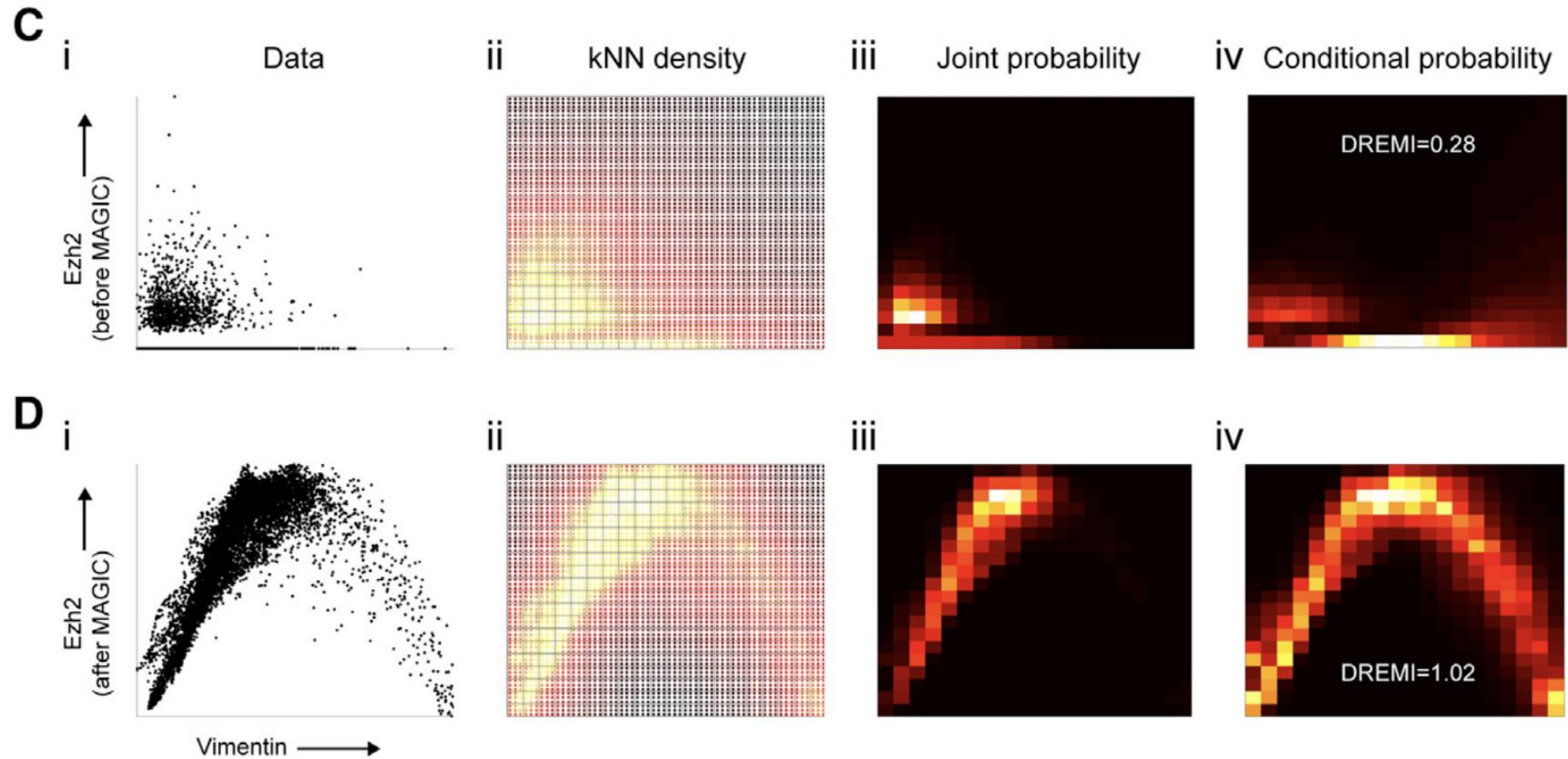


Fig5.A. Gene expression along with VIM expression

Vimentin expression could be work as a proxy (Pseudo-time) for the EMT state

Figure 5. Intermediate Archetype of EMT data



DREVI(Density Resampled Visualization) : Conditional Probability Plot

DREMI(Density Resampled Mutual Information) : Score for dependency of two variables

Figure 6. Gene Expression Dynamics

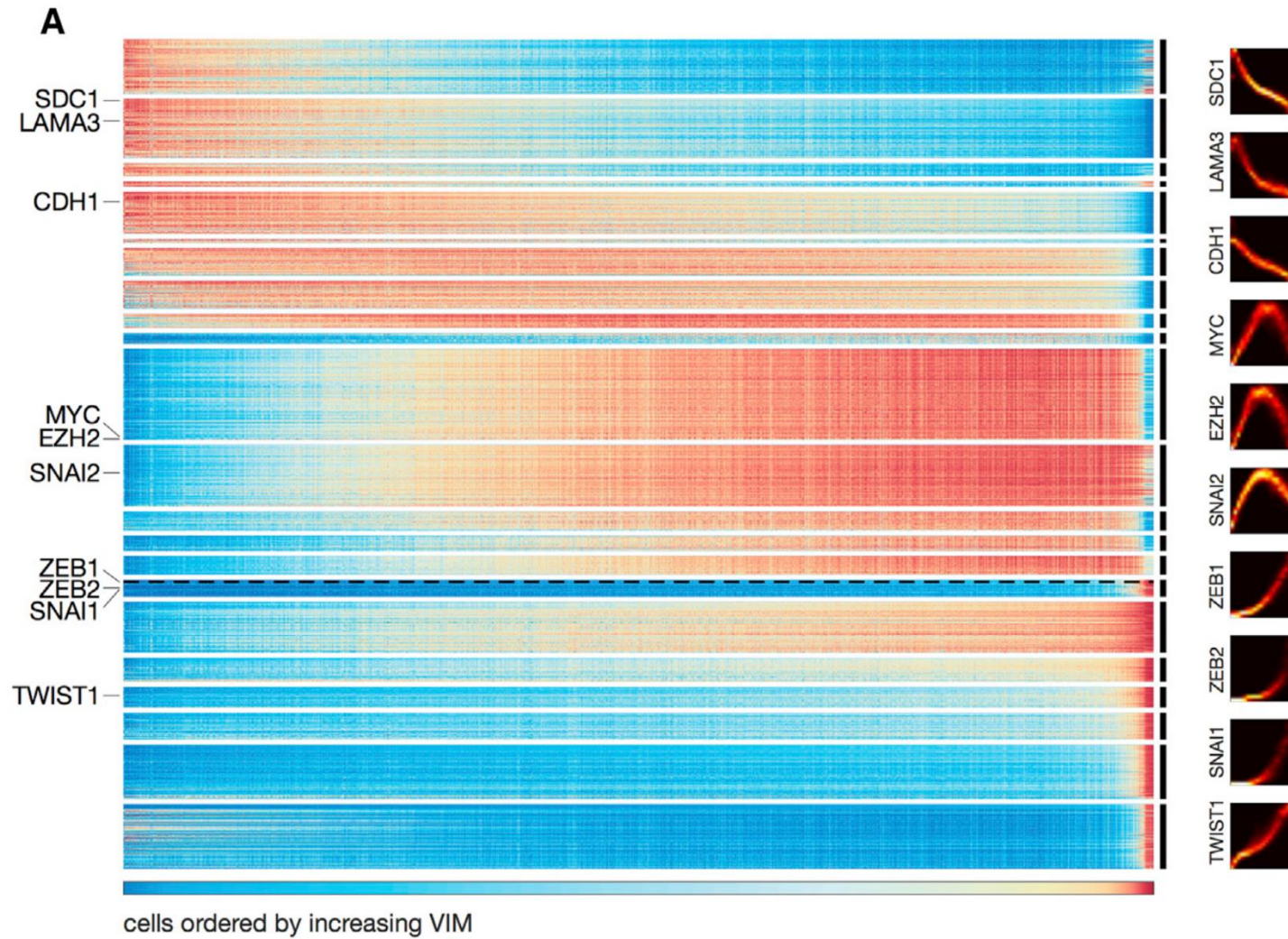
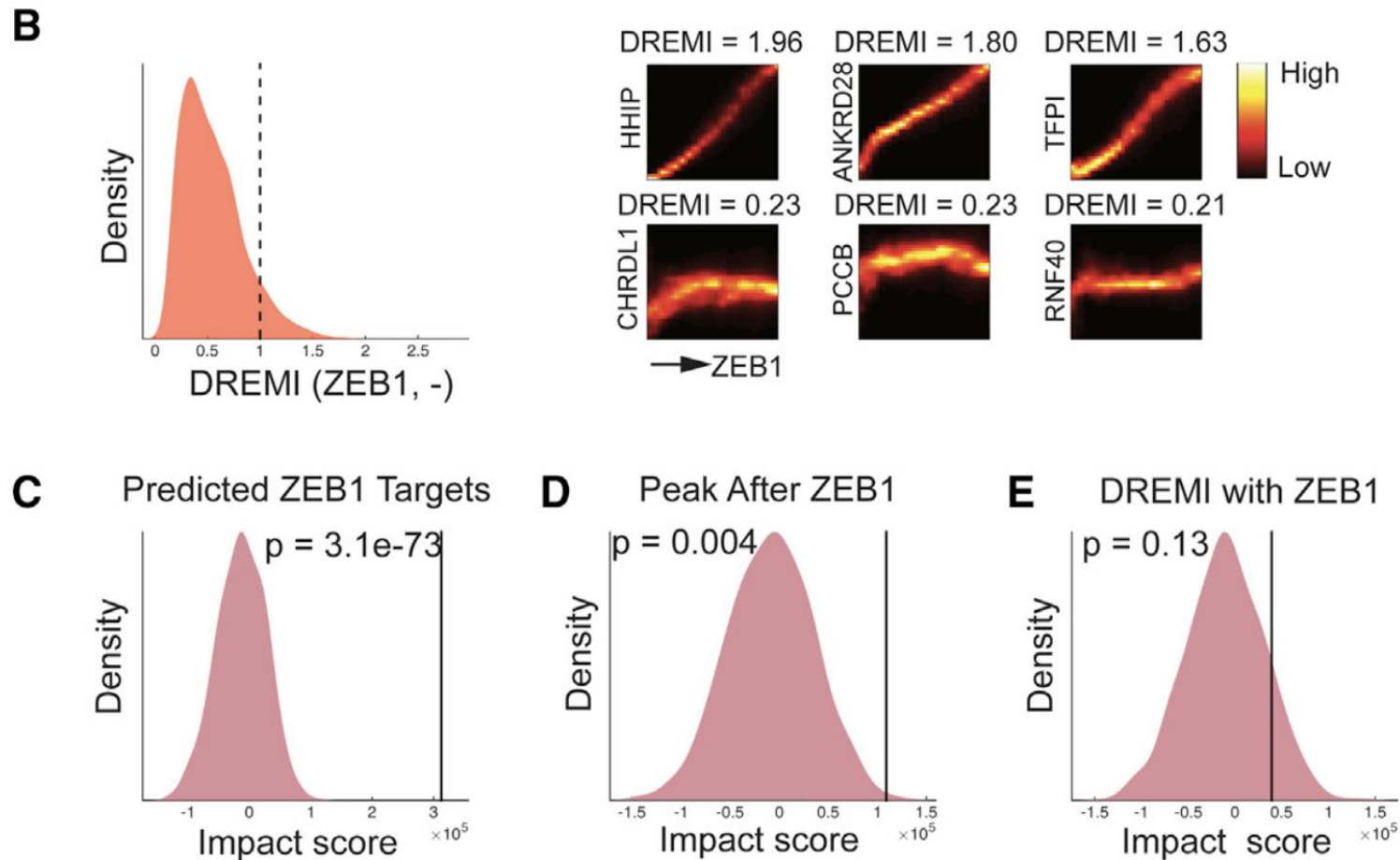


Fig6.A. Gene Expression Dynamic along with VIM expression (Pseudo-time)

Figure 6. Gene Expression Dynamics



B. DREMI score Description

C. Permutation Test comparing the ZEB1 and TGF-beta inductions

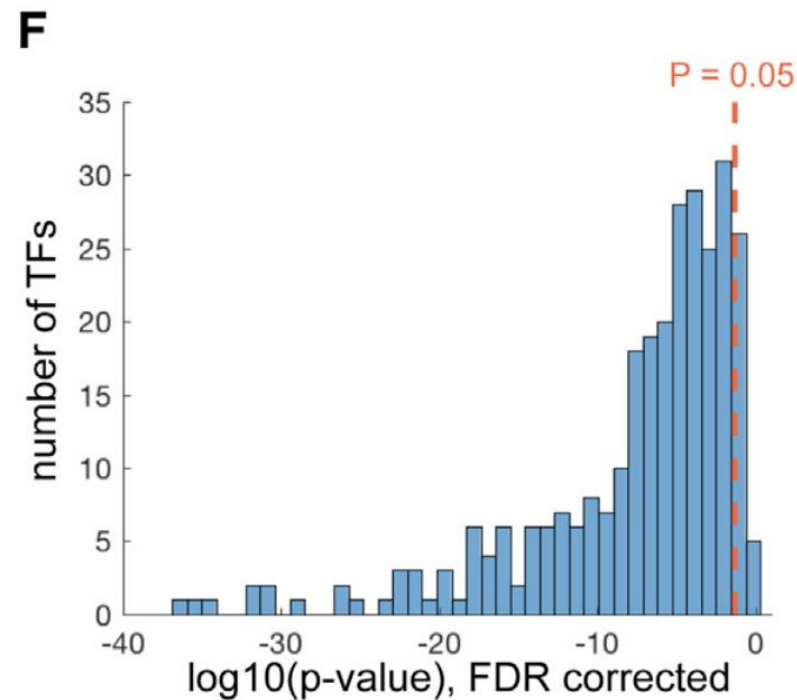
- Impact Score compares the relative ranking of gene expression between two scRNA set

## Figure 6. Gene Expression Dynamics

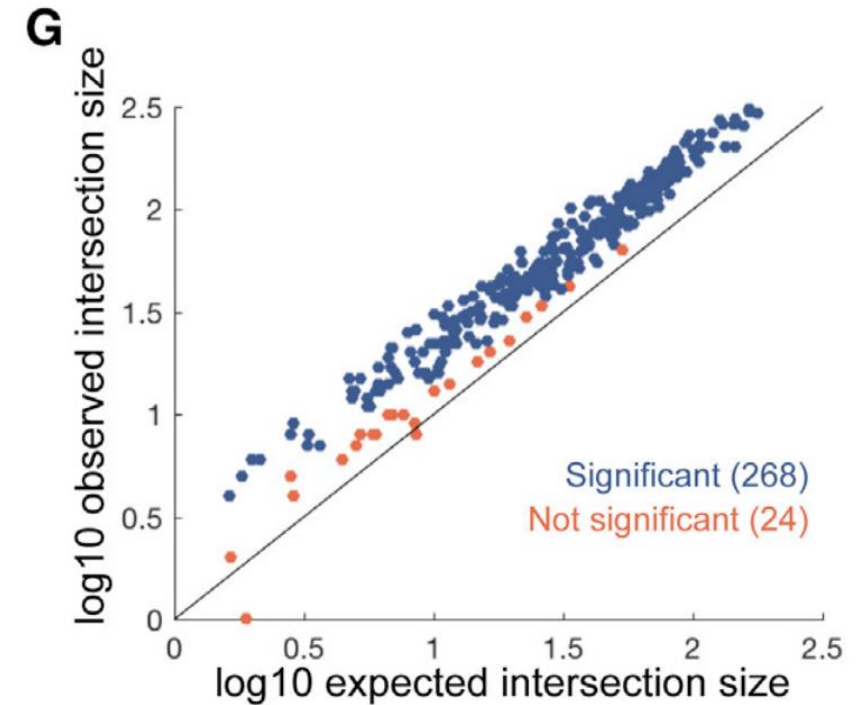
Choose 292 TFs and find the target genes using the same approach with ZEB1

Find target genes of each TF using ATAC seq and Motif Analysis.

Compare two sets using Fisher Test

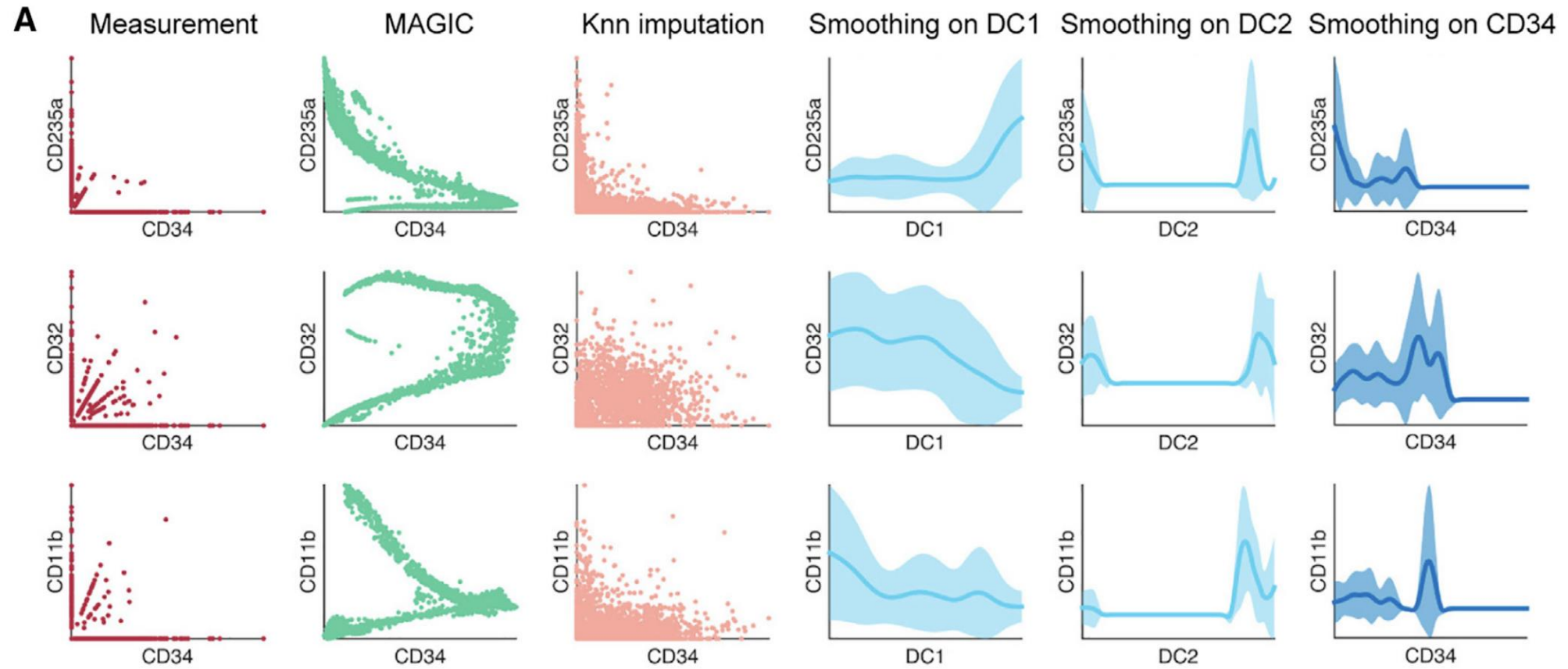


F. Distribution of FDR  
268 TF passed the FDR test.



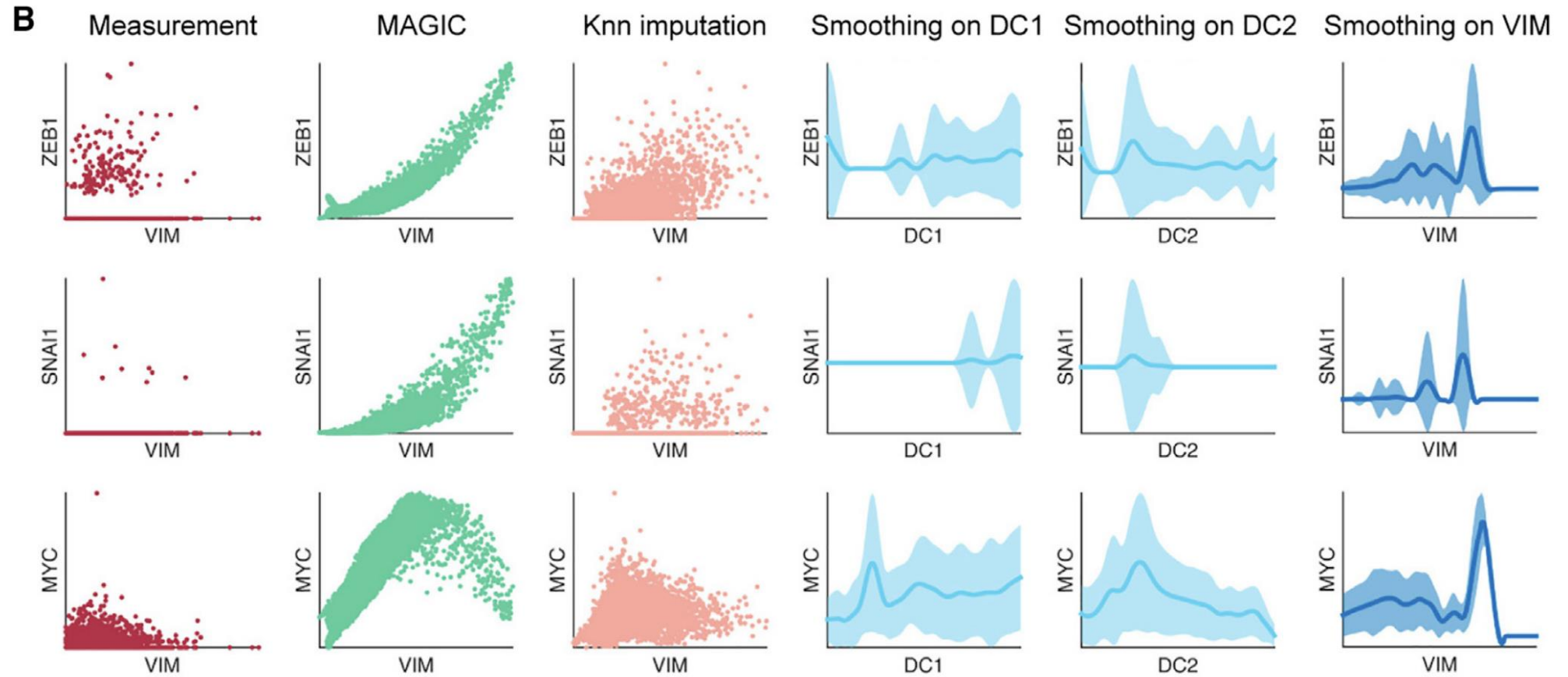
F. Expected vs Observed Intersection Size  
291/292 over the x=y line  
(without FDR correction)

Figure 7. Compare with other methods



A. MAGIC, KNN-imputation and DC-analysis with Mouse Bone Marrow dataset

Figure 7. Compare with other methods



B. MAGIC, KNN-imputation and DC-analysis with EMT dataset

Thank  
you