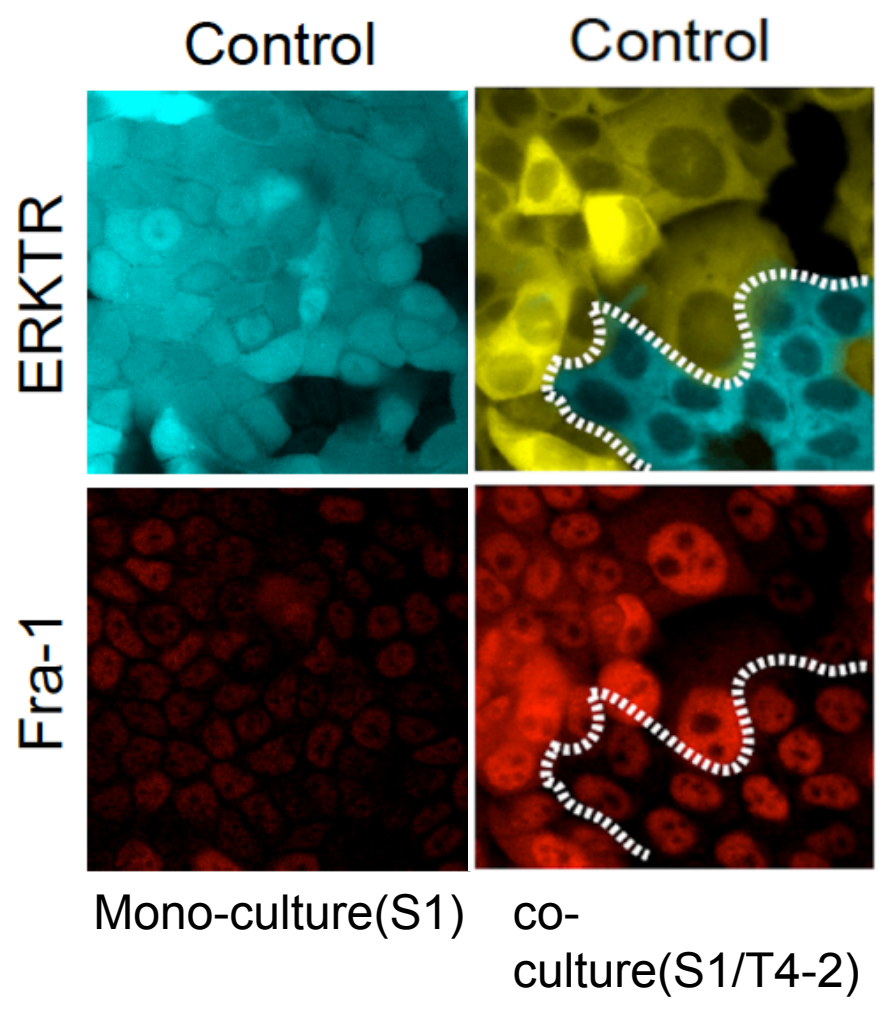
Deep nested generative model to identify the cell-cell interactions using single-cell RNA sequencing

Heqiao Ruan [1,2], Alexander Davis[3], John Albeck[3], Gerald Quon [1,3]
[1]Genome Center, University of California, Davis. [2]Department of Statistics, University of California, Davis.
[3]Department of Molecular and Cellular Biology, University of California, Davis.
hruan@ucdavis.edu, gquon@ucdavis.edu

Motivation

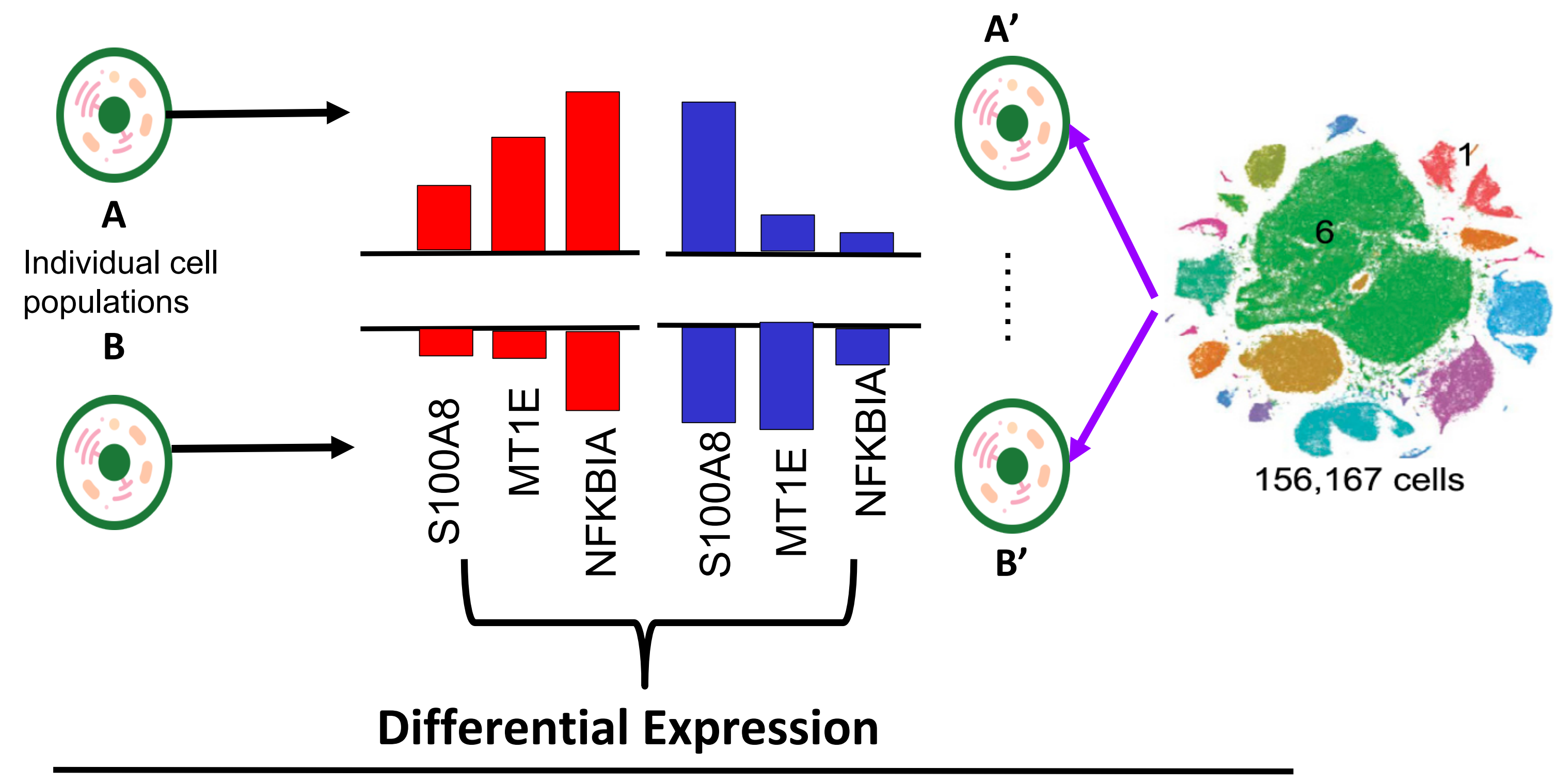

Control
Control
ERKTR
Fra-1
Mono-culture(S1)
co-culture(S1/T4-2)

The plot shows the differential behavior for gene expression of mono-culture tumor cells (S1) and the tumor cells near the malignant tumor cells(S1/T4-2).

The recent technological development enable us to characterize the whole transcriptional profile of whole biological samples and measure precise transcriptional changes during previously difficult-to-observe processes.

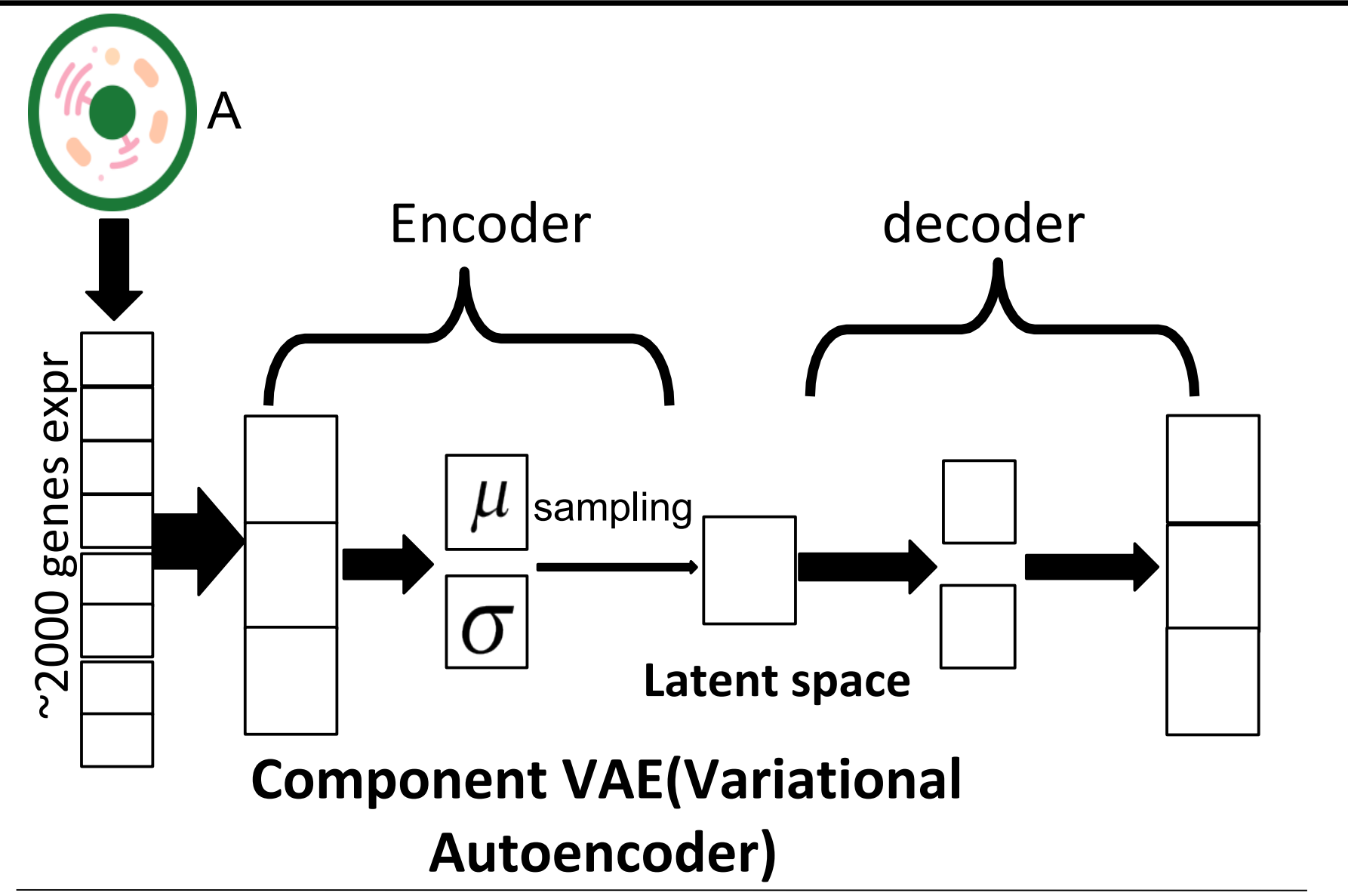
We want study the co-culture behavior of gene expressions using single cell RNA-seq(sc-RNA seq) data

Research Aims


A
Individual cell populations
B
S100A8
MT1E
NFKBIA
S100A8
MT1E
NFKBIA
A'
B'
6
1
156,167 cells
Differential Expression

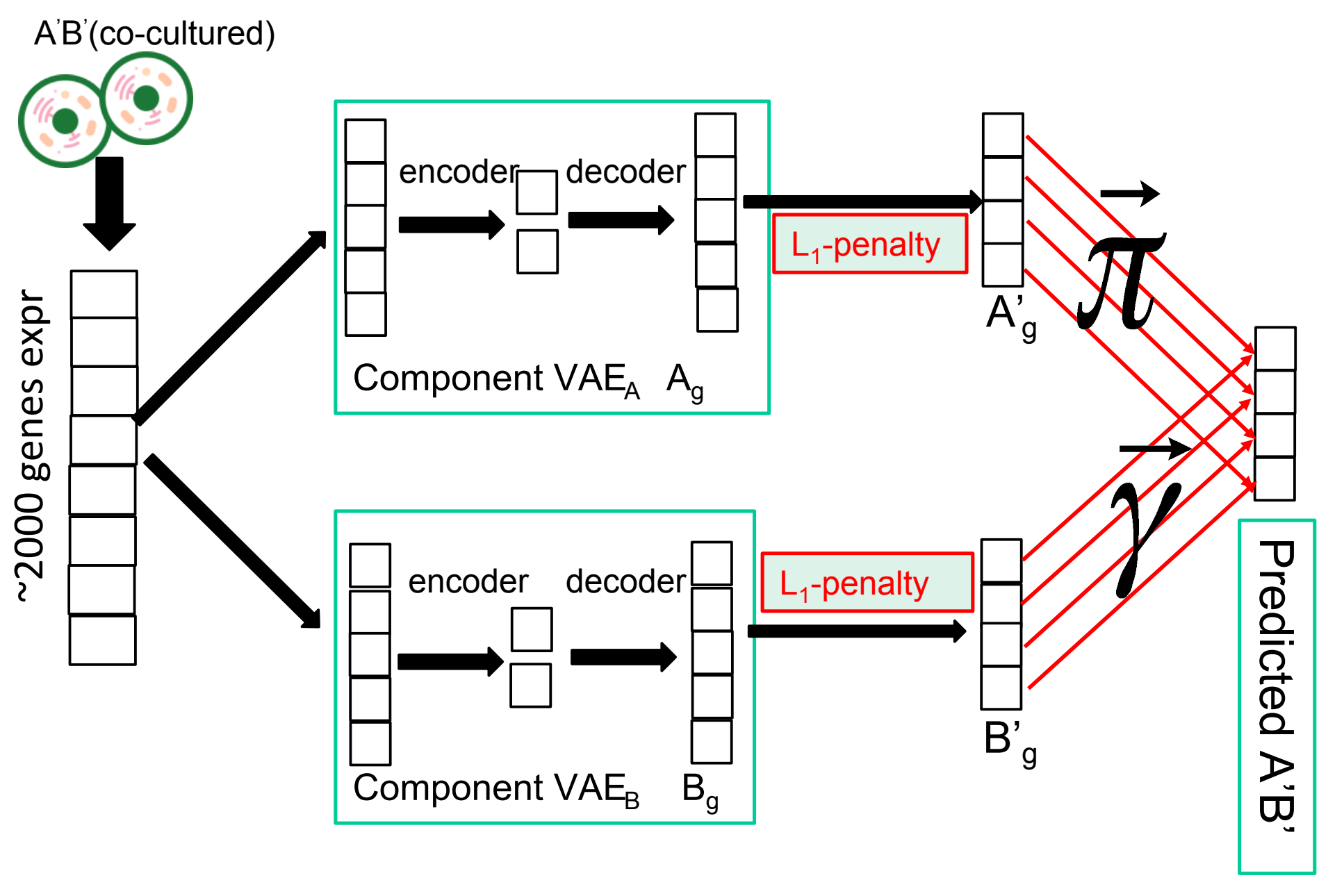
Our research aim to **(1)** Extract the gene expression profile A' and B' representing population A and B respectively from the co-cultured cell populations

Methods


A
Encoder
decoder
~2000 genes expr
μ
sampling
σ
Latent space
Component VAE(Variational Autoencoder)

We propose a deep nested generative model to study the single cell RNA-seq data

Step 1: Train the two component VAE on cell population A and B respectively


A'B'(co-cultured)
~2000 genes expr
encoder
decoder
Component VAE_A
A_g
L_1-penalty
A'_g
encoder
decoder
Component VAE_B
B_g
L_1-penalty
B'_g
$\vec{\pi}$
$\vec{\gamma}$
Predicted A'B'

- Step 2: Train the nested generative model with two component VAE fixed.
- Propose an intermediate layer with L_1 penalty to characterize the differential behavior of individual cell population and co-cultured population.

Step 3: Differentially Expressed Genes(DEG) Analysis
The set differentially expressed genes represent the differential behavior of gene expression levels across populations

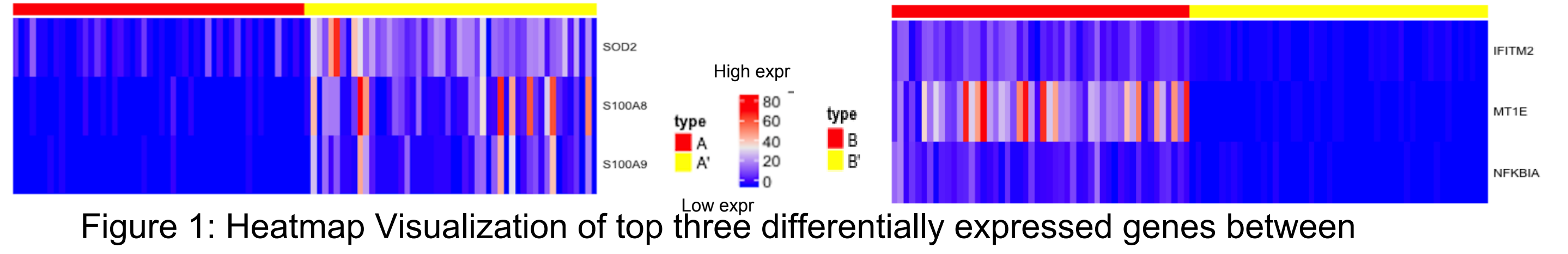


Figure 1: Heatmap Visualization of top three differentially expressed genes between different cell populations

Methods

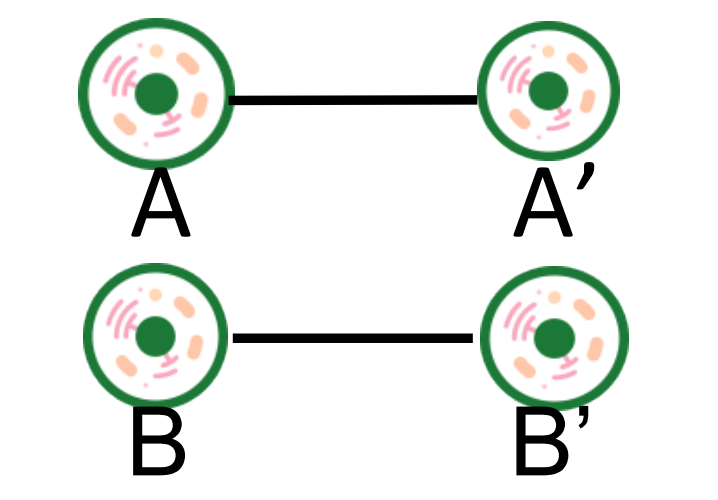


Figure 2:Comparing with ground truth

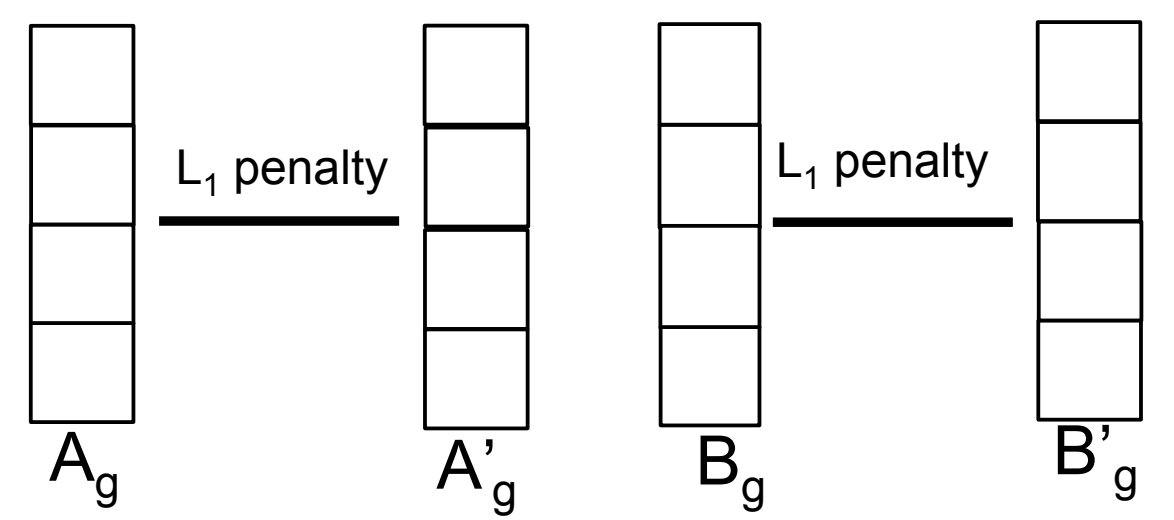

L_1 penalty
L_1 penalty
A_g
A'_g
B_g
B'_g

Figure 3:Intermediate layer of nested generative model

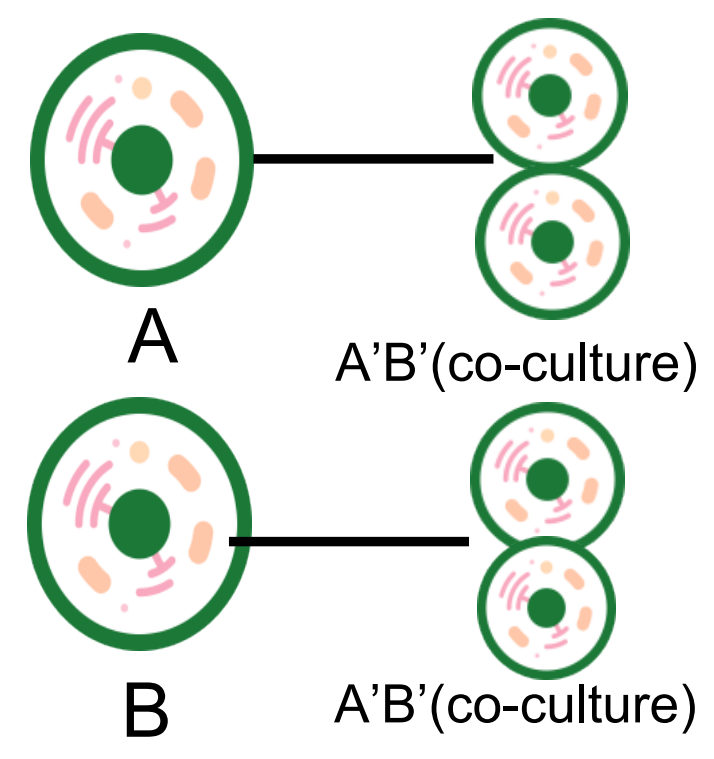

A
A'B'(co-culture)
B
A'B'(co-culture)

Figure 4:Baseline Measurement

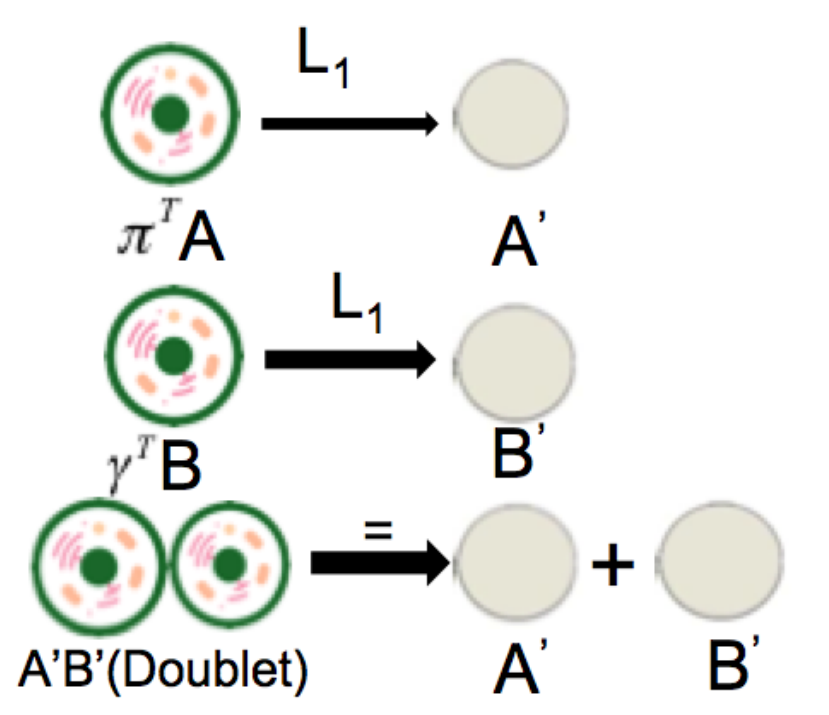

L_1
$\pi^T A$
A'
L_1
$\gamma^T B$
B'
=
+
A'B'(Doublet)
A'
B'

Figure 5:Linear baseline model

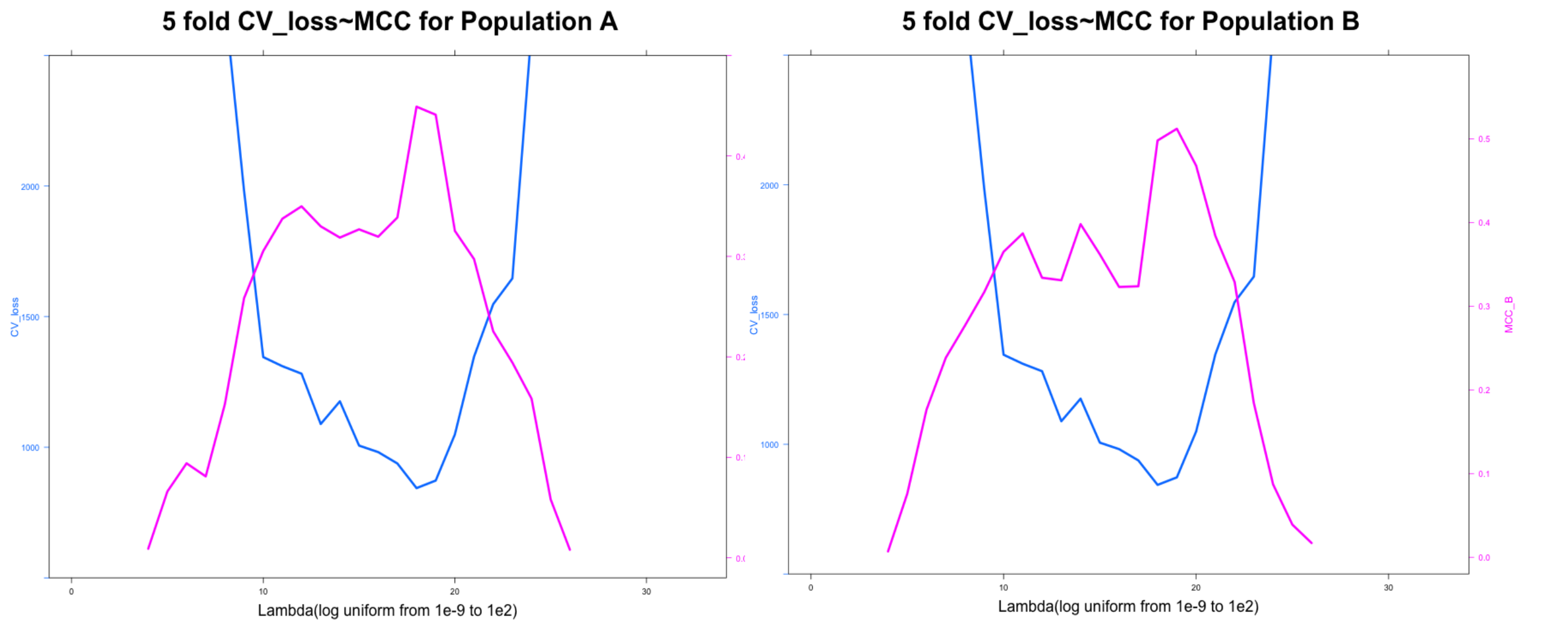
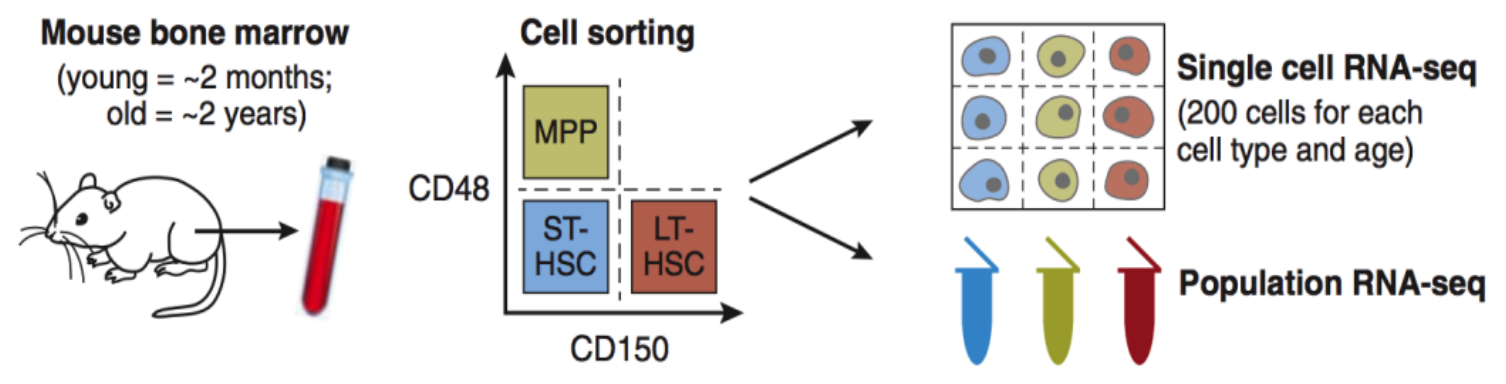


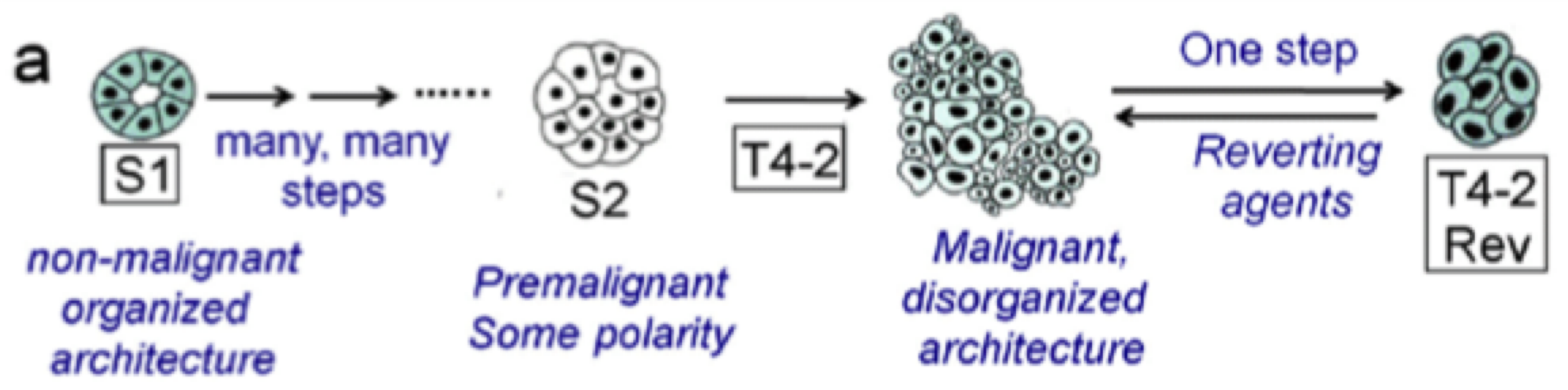
Fig 6A,B:5-fold cross validation loss and MCC for population A and B with respect to log-uniform design of tuning parameter Lambda

We detect the genes involved in interactions from the L_1-penalty layer and compare with the ground truth comparison.
We use the MCC(Matthew's coefficient of correlation) between two sets of DE genes to quantify the efficiency of identification
Two baseline measurements for MCC, first is prior-knowledge based(Fig 4), second is a simple linear based model(Fig 5).
Tuning parameter λ controls the amount of L_1 penalty

Datasets


Mouse bone marrow
(young = ~2 months; old = ~2 years)
Cell sorting
CD48
MPP
ST-HSC
LT-HSC
CD150
Single cell RNA-seq
(200 cells for each cell type and age)
Population RNA-seq

Shalek et.al(2014) studied the dynamic paracrine control of cellular variation in mouse bone marrow cells


a
S1
many, many steps
non-malignant organized architecture
S2
Premalignant Some polarity
T4-2
Malignant, disorganized architecture
One step
Reverting agents
T4-2 Rev

The research study the tumor cells for predicting reversion protocols to explain changes in gene expression in malignant tumor cells.

Visualization

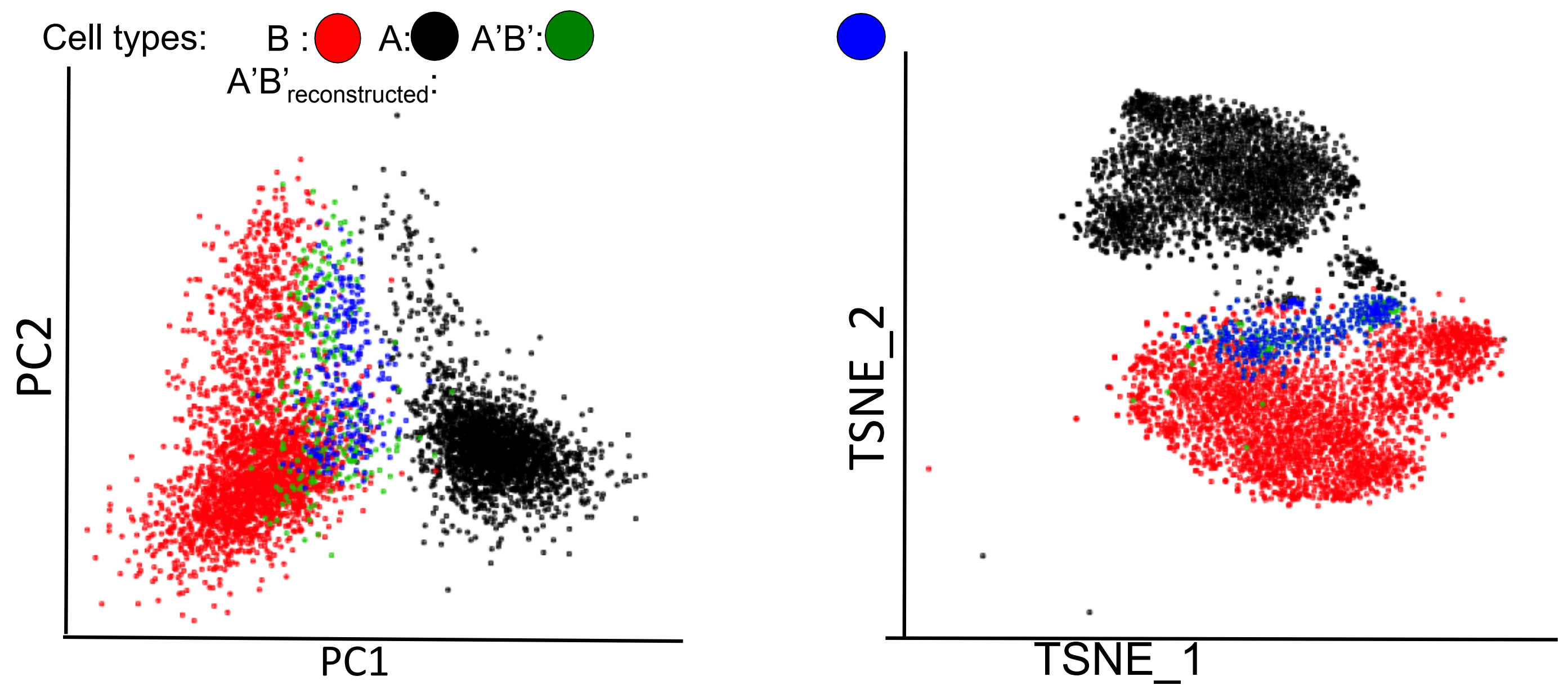


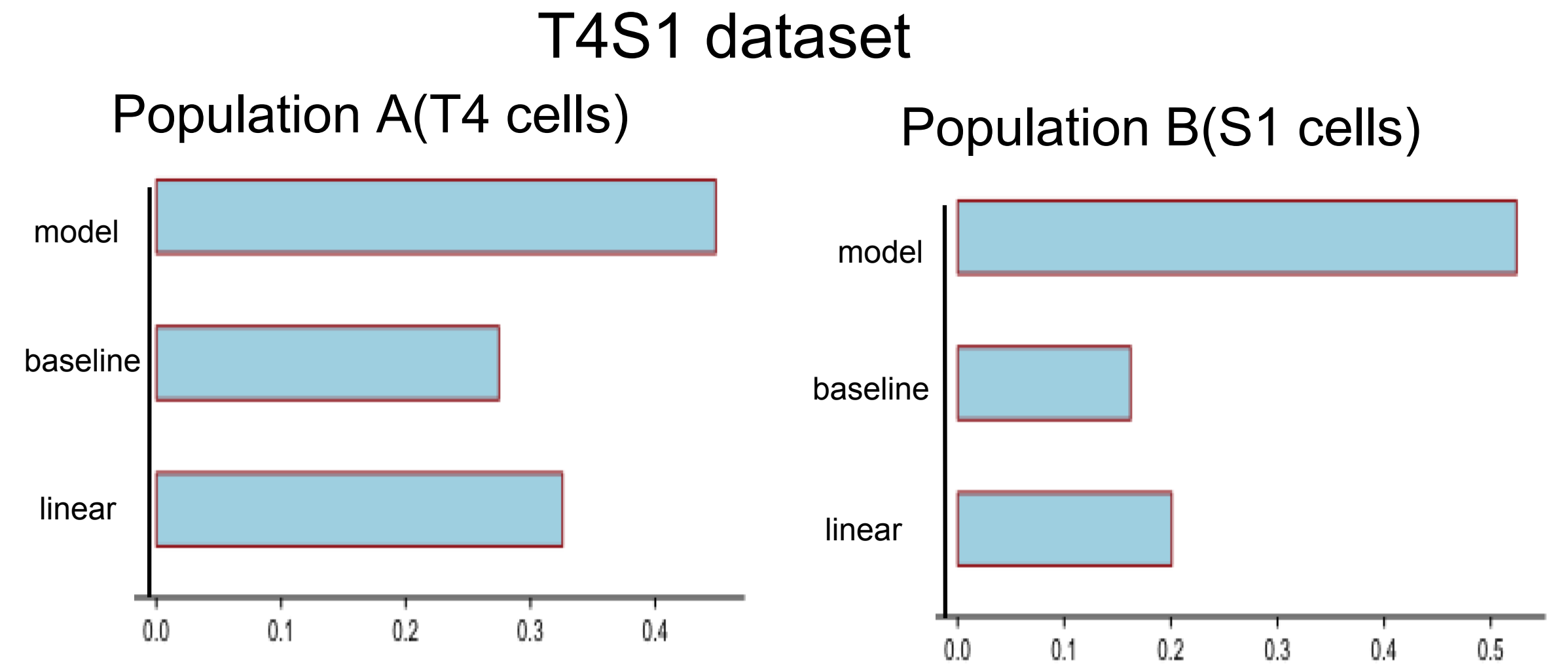
Figure 7,8: Use PCA and TSNE to visualize the low dimensional embedding
Given the model work, the reconstructed A'B' should be closer to the original co-cultured cell population A'B' than the individual cell population A and B in the low dimensional embedding.

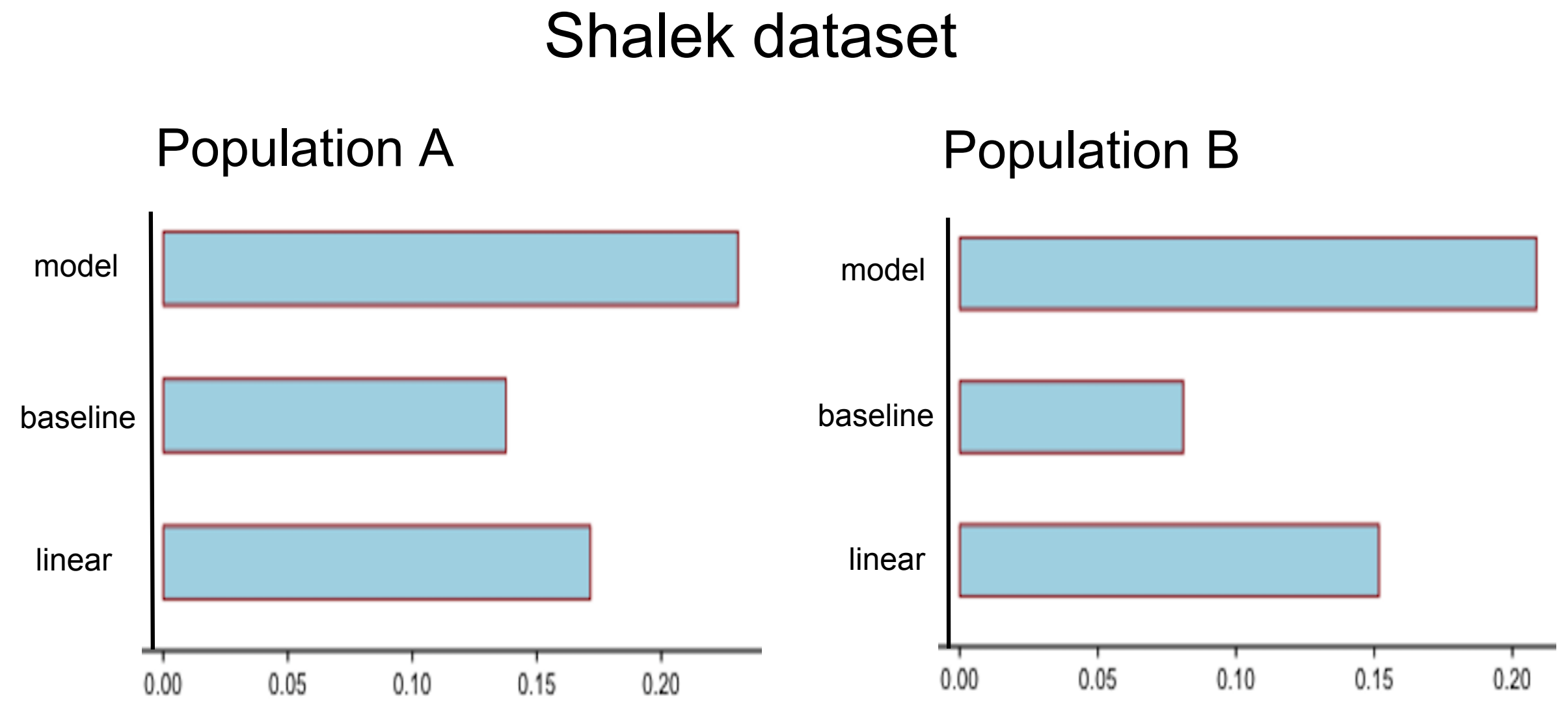
Visualization

- The reconstructed A'B' successfully approximate the original co-cultured cell population A'B'

Quantitative Evaluation

We evaluate our model by calculating the MCC comparing with the baseline measurements

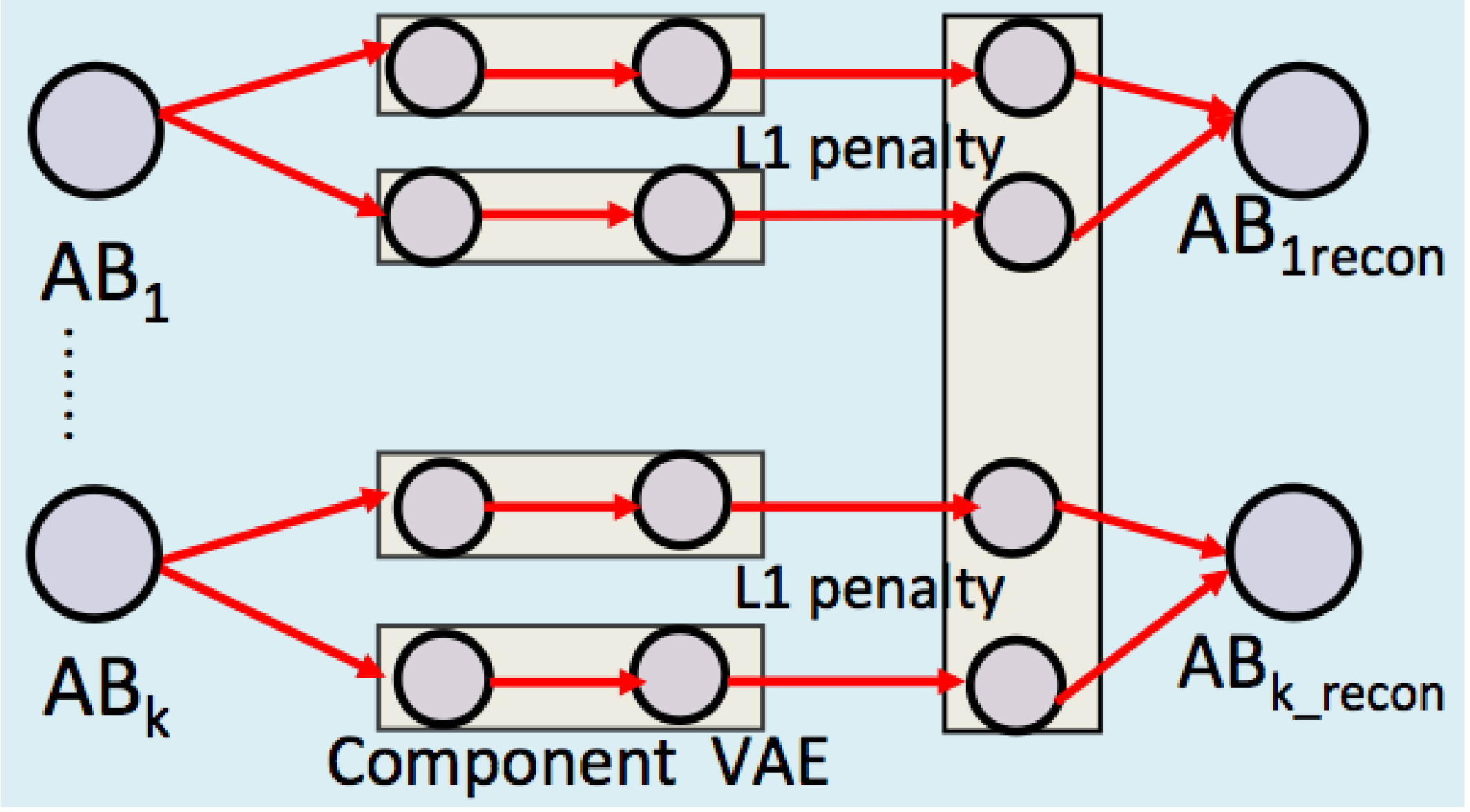




Our proposed model is able to identify the genes involved in interactions among co-cultured cell populations more efficiently than both the prior-based and linear baseline measurements.

Future work

- Extend the model to multiple cell populations:


AB_1
L1 penalty
AB_{1recon}
AB_k
L1 penalty
AB_{k_recon}
Component VAE

- Apply the model to RNA-seq data sequenced from brain tissues to identify genes potentially involved in the schizophrenia.

Reference

[1] Alex K.Shalek et.al. "Single cell RNA Seq reveals dynamic paracrine control of cellular variation" Nature, 2014