An approach to infer the gene regulatory network of a stable cell type

Wing Hung Wong

Cellular state

Let x_i = expression level of gene *i*,
then the cellular state is the vector

$$X = (x_1, x_2, x_3, \ldots, x_n)$$

How are the levels of expression maintained ?
What are the gene regulatory mechanisms?

Our task is to formulate these questions mathematically and find a way to solve them

Dynamical system

Assume X varies in time according to

$$\frac{dX(t)}{dt} = A(X(t))$$

The vector field $A: R^n \rightarrow R^n$ contains detailed information on regulatory information, e.g.

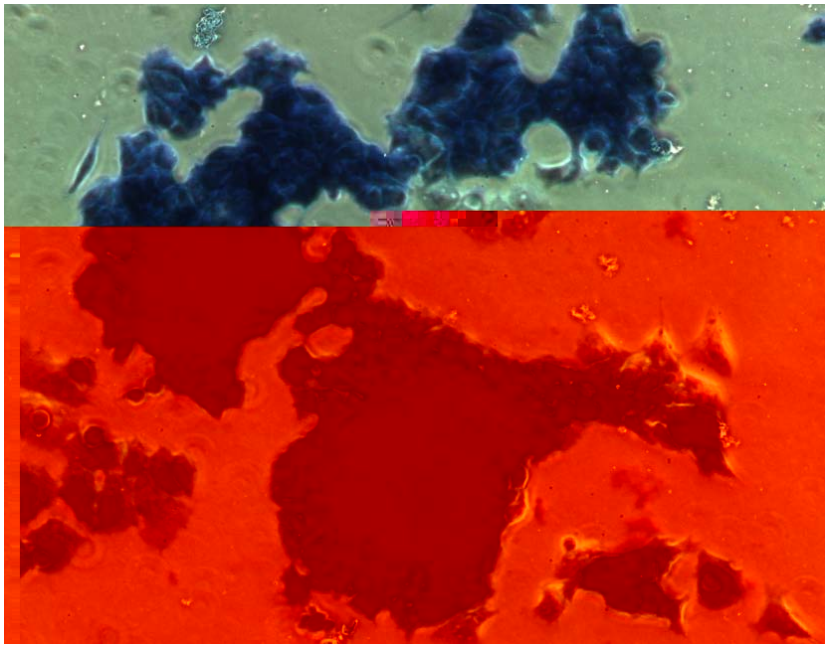
X_j *positively regulates* X_i $\leftrightarrow$ $A_i(x_1 \ldots x_n) \uparrow$ in x_j

However, A is too complex to reconstruct from experiments based on current technology.

Stable cell type

A stable cell type can maintain a characteristic pattern of gene expression through a gene regulatory network.

Example:
Mouse embryonic stem cells (on 0.1% gelatin, with LIF)



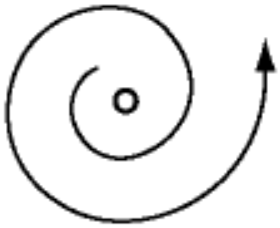
Equilibrium state

A state $\boldsymbol{\mu}$ is an equilibrium state if $\boldsymbol{A(\mu)=0}$

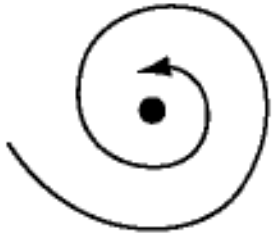
The equilibrium is stable if, once the system comes close to μ, it will stay close to μ from then on.

We identify stable cell types with stable equilibrium states of the dynamical system

unstable



stable



Regulatory network

Suppose $X(0)= \mu+\delta$, then for small t

$X(t)-X(0) \approx t\, A(\mu+\delta) \approx t\, [A(\mu) + T\, \delta\,] = t\, T\, \delta$

where T is the Jacobian matrix: $T_{ij}=(\partial A_i/x_j)(\mu)$

We propose to regard T as the regulatory network that maintains the equilibrium μ

Stability imposes a global constraint on the network: T must be negative definite to ensure stability

An approach to network reconstruction

- Use RNA-interference to knockdown each regulator in the stable cell type
- Measure gene expression after the perturbation
- Infer network based on a regression model
- Incorporate sparsity & stability into the regression
- Incorporate regulator binding data when available

Regression model

- Response: gene expression changes on I genes

$$Y = \{Y_1, Y_2, ..., Y_I\}$$

where $Y_i = X_i(t) - X_i(0)$

- Predictor: perturbation on J regulators

$$Z = \{Z_1, Z_2, ..., Z_J\}$$

e.g. $Z = ((0.5)\mu_1, 0, 0, ...0)'$

- Model: $E(Y_i) = \sum_{j=1}^{J} T_{ij} Z_j$, for $i = 1, ...I$

- Goal: identify non-zero elements in T_{ij}

Sparsity

- The true network is likely to be sparse
- Lasso-type regularization with L_1 penalty
- Penalized loss function

$$L(T,\lambda_1)=\sum_{per}\sum_{i=1}^{I}\left\|Y_i-\sum_{j=1}^{J}T_{ij}Z_j\right\|_2^2+\lambda_1\|T\|_1$$

here the outer sum is over all perturbation experiments

Stability

- Stability imposes useful constraints on *T*
- Lyapunov stability

$$\left\| X(t) - \mu \right\|^2 = \left\| (I + T)(X(0) - \mu) \right\|^2 \le \left\| X(0) - \mu \right\|^2$$

choose $X(0)-\mu$ to get an necessary condition

$$T_{jj} \le 0; \qquad \left\| {}^{(-j)}T_j \right\|_2^2 \le 1, \text{for } j = 1, \ldots, J$$

- This leads to the optimization of

$$L(T, \lambda_1, \lambda_2) = \sum_{per} \sum_{i=1}^{I} \left\| Y_i - \sum_{j=1}^{J} T_{ij} Z_j \right\|_2^2 + \lambda_1 \sum_{j=1}^{J} \left\| T_j \right\|_1 + \lambda_2 \sum_{j=1}^{J} \left\| {}^{(-j)}T_j \right\|_2^2$$

- Alternative formulations are possible

Incorporate TF binding location data

- TF association strength (TFAS) integrates the ChIP-seq peak intensities of TF *j* in the vicinity of gene *i*

$$a_{ij} = \sum_k g_k e^{-d_k / d_0}$$

- Define the TFAS weighting factor

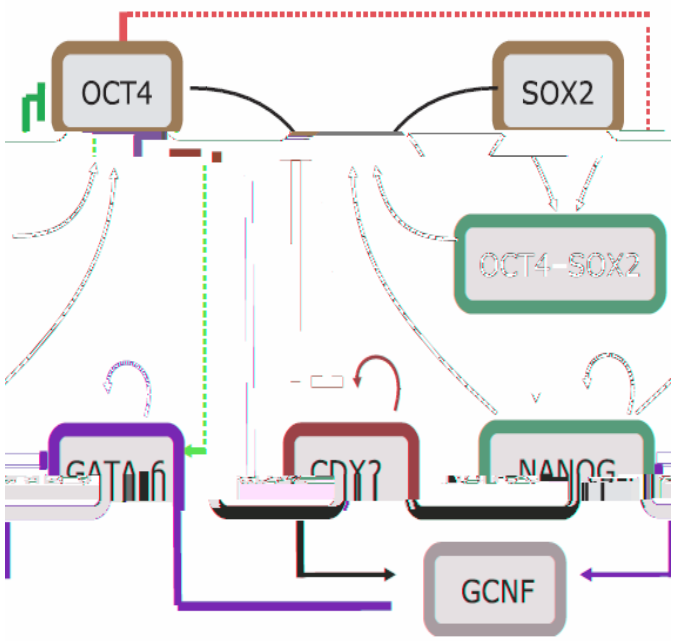
$$c_{ij} = 1 / a_{ij}$$

- Penalized loss function

$$L(T, \lambda_1, \lambda_2, c) = \sum_{per} \sum_{i=1}^{I} \left\| Y_i - \sum_{j=1}^{J} T_{ij} Z_j \right\|_2^2 + \lambda_1 \sum_{j=1}^{J} \left\| c_j \cdot T_j \right\|_1 + \lambda_2 \sum_{j=1}^{J} \left\| c_j \cdot^{(-j)} T_j \right\|_2^2$$

Simulated data:

Manually constructed by Chickarmane et al., (2008) PloS One.



$$\frac{d[O]}{dt} = \frac{a_0 + a_1[A] + a_2[O][S] + a_3[O][S][N]}{1 + b_0[A] + b_1[O] + b_2[O][S] + b_3[O][S][N] + b_4[C][O] + b_5[GC]} - \gamma_1[O] \qquad (2)$$

$$\frac{d[S]}{dt} = \frac{e_0 + e_1[O] + e_2[O][S] + e_3[O][S][N]}{1 + f_0[O] + f_1[O][S] + f_2[O][S][N] + f_3[O][G]} - \gamma_3[N]$$

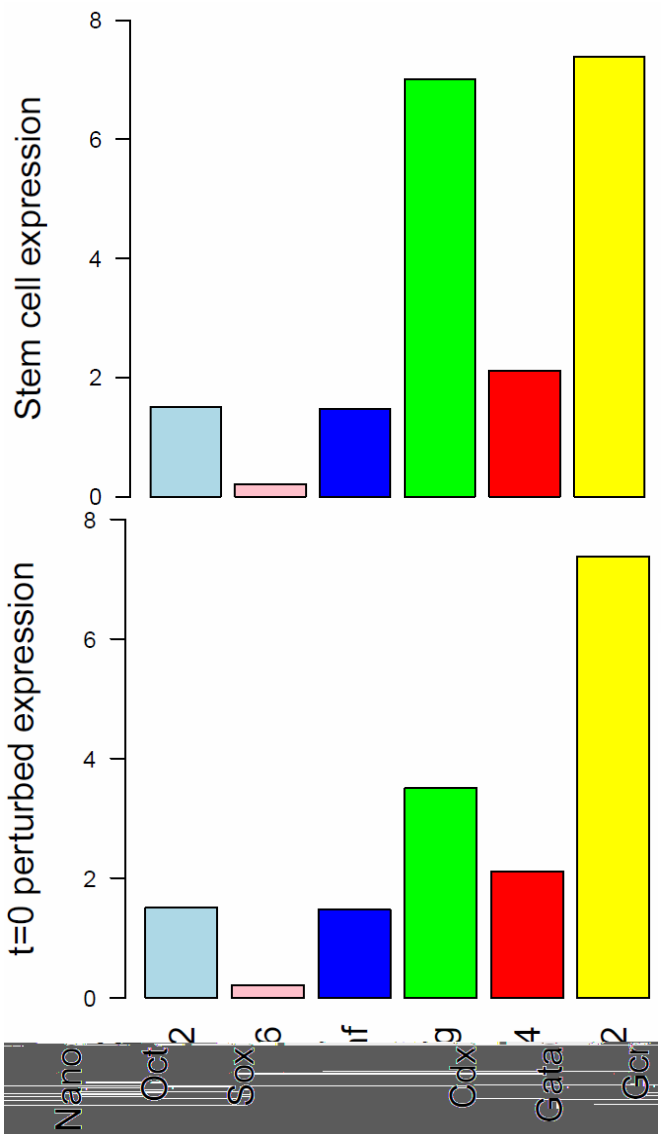
$$\frac{d[GC]}{dt} = \frac{i_0 + i_1[C] + i_2[G]}{1 + j_0[C] + j_1[G]} - \gamma_5[GC]$$

$$\frac{d[G]}{dt} = \frac{p_0 + p_1[O] + p_2[G]}{1 + q_0[O] + q_1[G] + q_2[N]} - \gamma$$

2 stable equilibrium states:
stem cell & endoderm

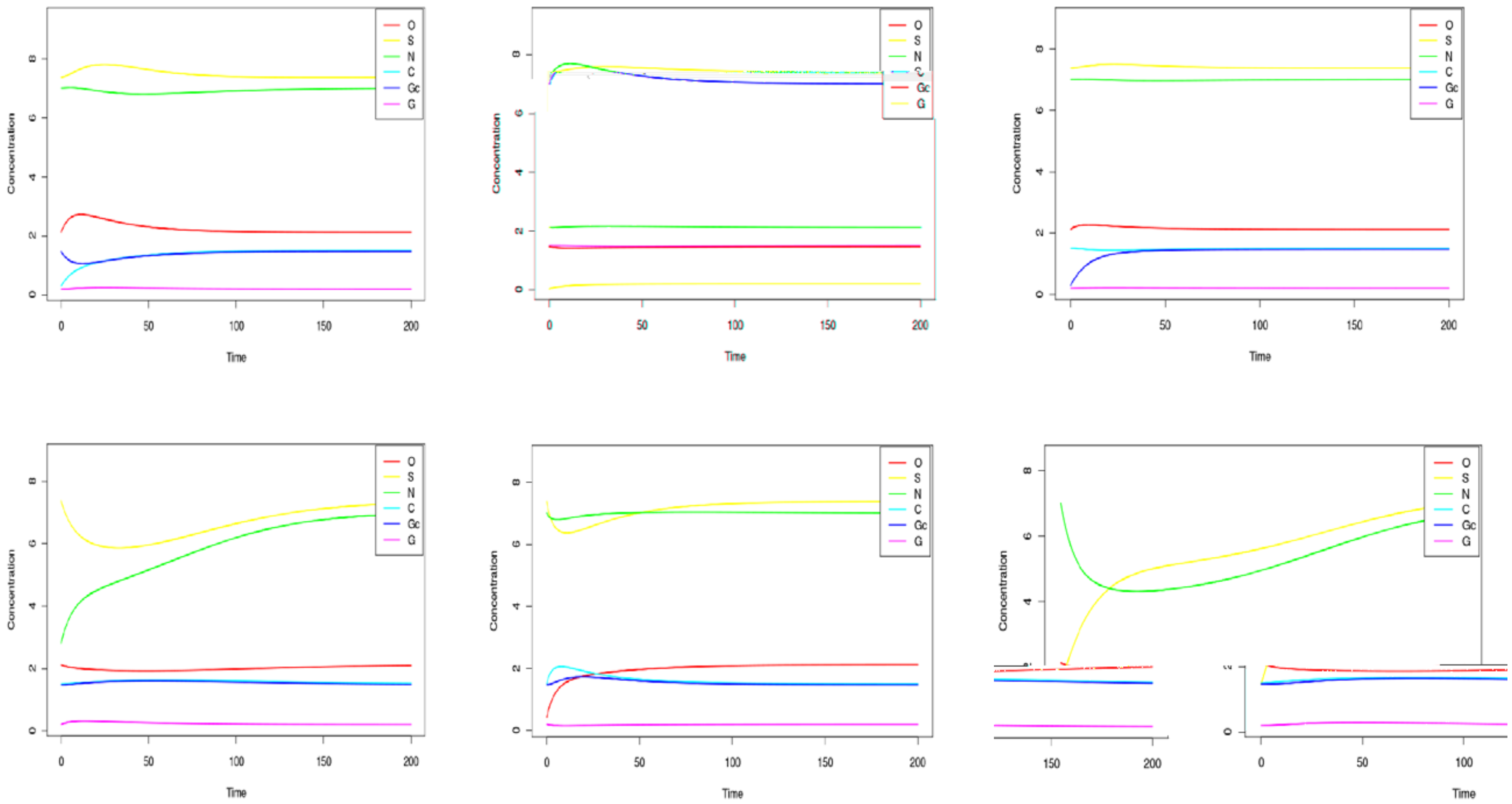
Use symbolic solver to get the two networks

Perturbation of stem cell state



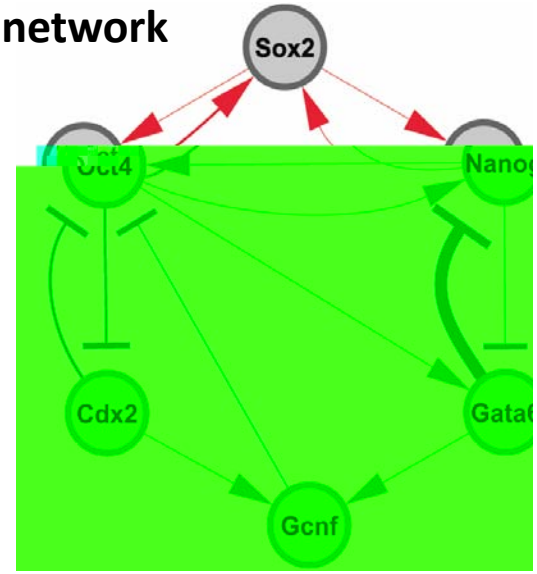
- Knockdown one of the six TFs in each experiment
- The TF expression is reduced by 50% (Nanog) at time *t*=0
- Simulate evolution of expression after perturbation

Time evolution after perturbation

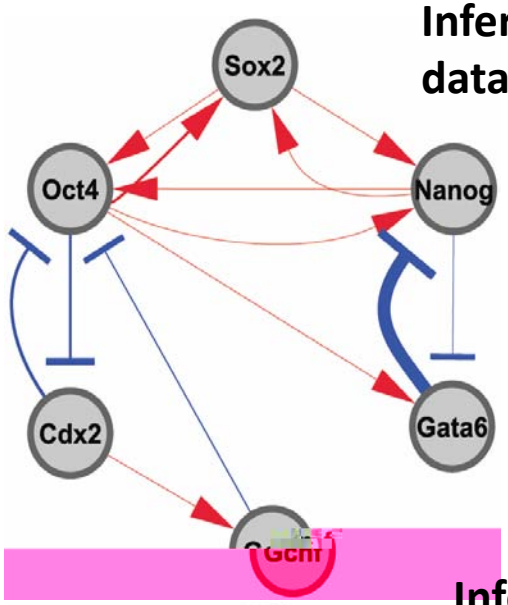


Network reconstruction results

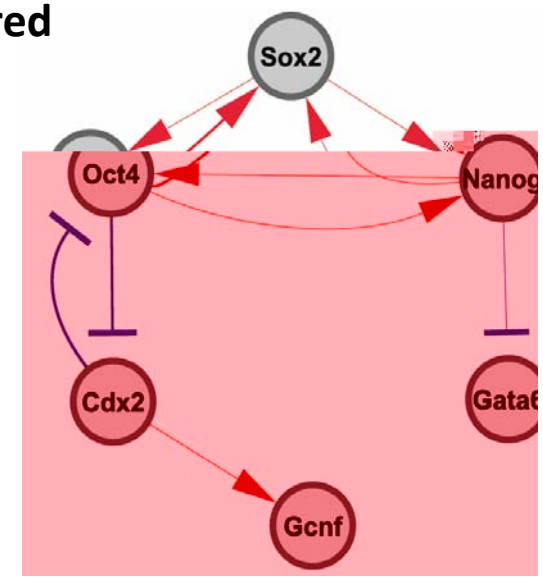
True network



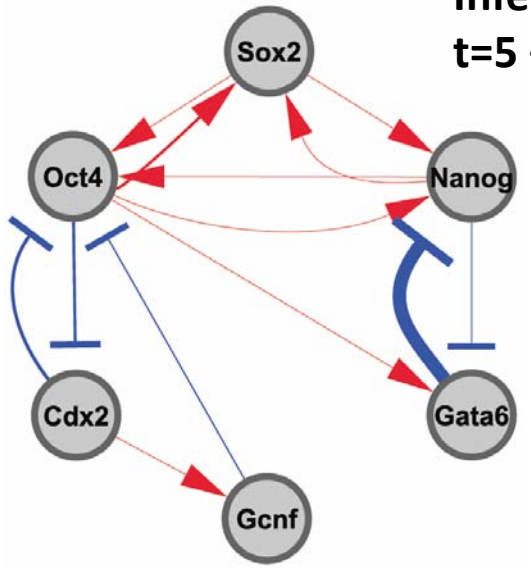
Inferred from data at t=1



Inferred t=5

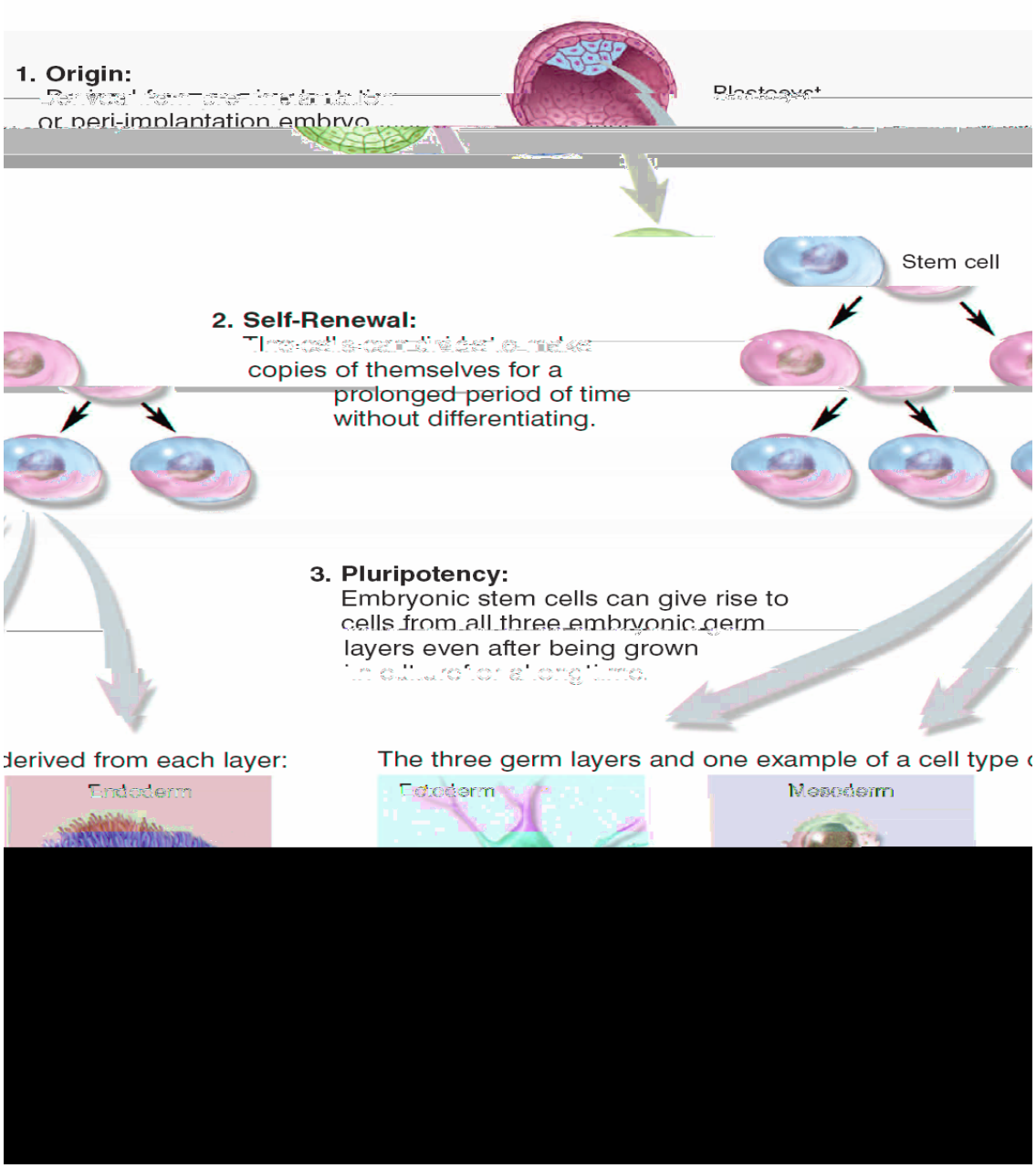


Inferred t=5 + prior



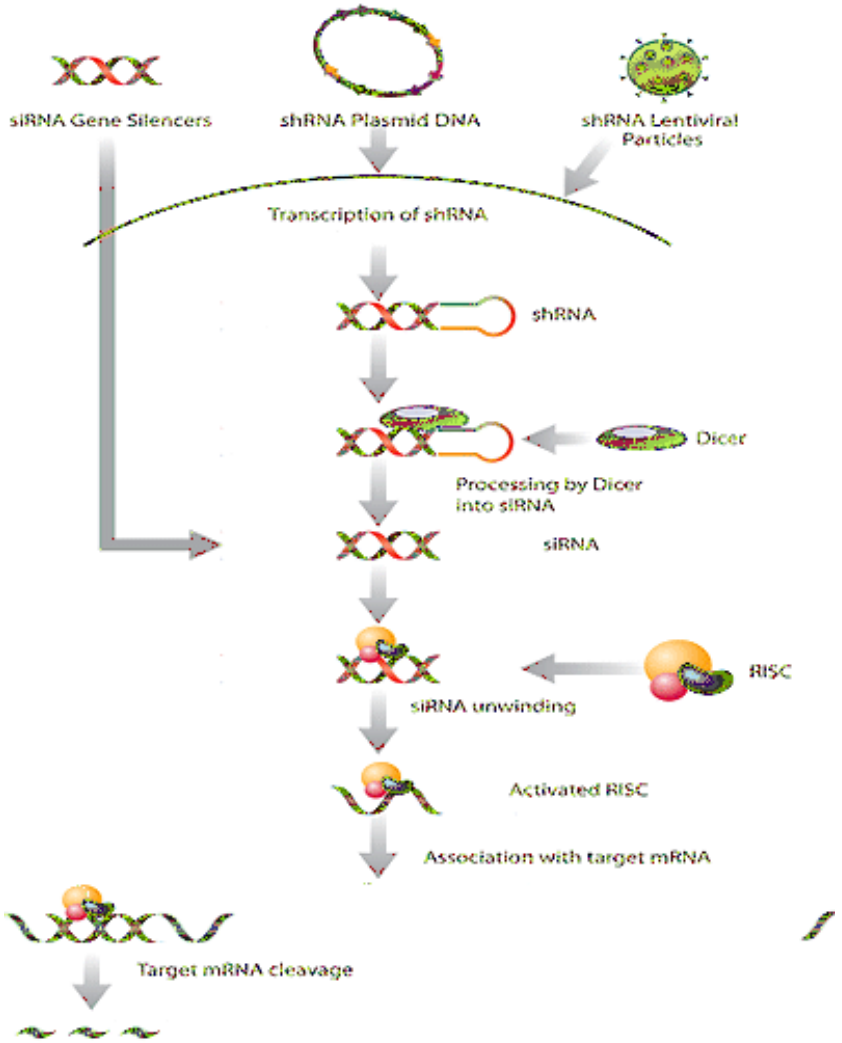
Real data

Embryonic stem cell



Regenerative Medicine. August 2006.

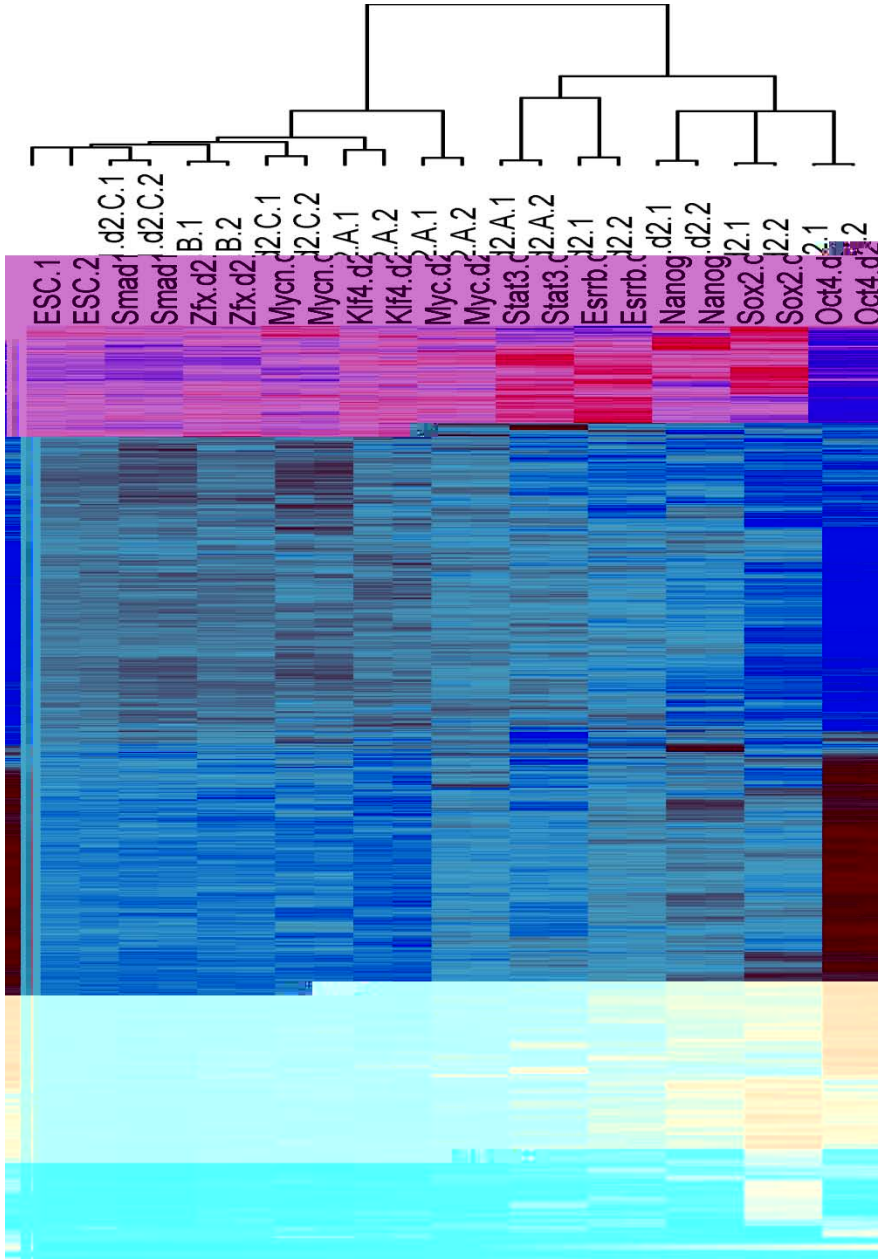
Gene knockdown by RNA interference



Summary of experiments in our lab

-

Sample clustering (after batch effect correction)



some details

- Quantile normalization
- Batch effect modeling
- No gene filtering
  - All 18138 genes entering into the model fitting
  - Perhaps the first attempt on gene regulatory network inference at the whole genome level in a mammalian cell type

- Network reconstruction with ChIP information
  (ChIP-seq data from Chen et al 2008)

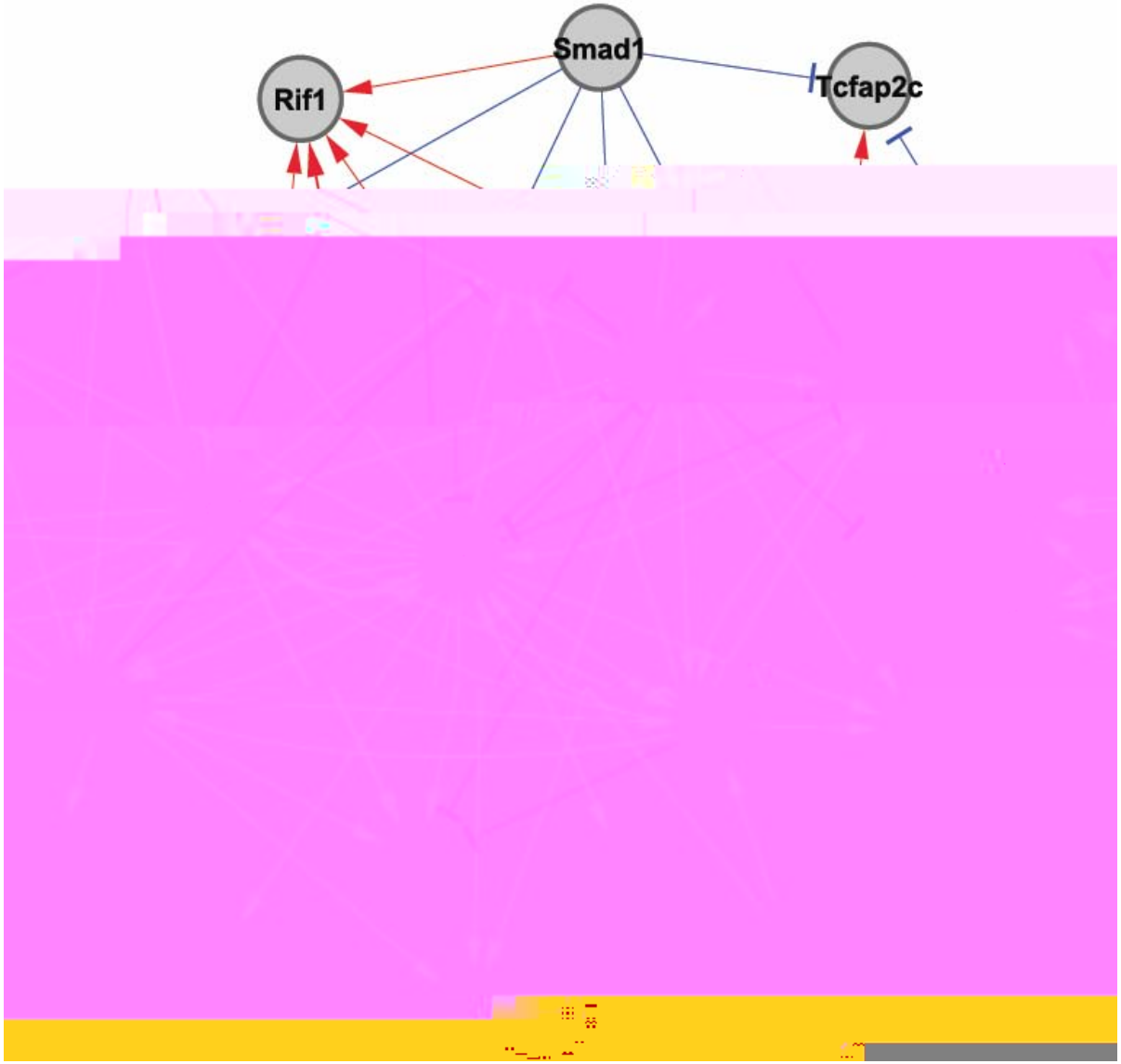
- Cross-validated for choice of λ

TF targets identified

- 3764 targets regulated by 10 TFs

Oct4	Nanog	Sox2	Esrrb	Stat3	Klf4	Myc	Mycn	Zfx	Smad1
2362	588	461	1169	895	277	0	0	72	163

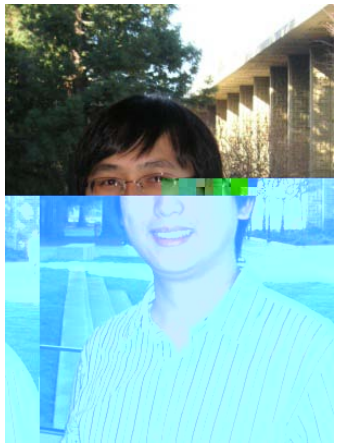
A subnetwork for important TFs



Acknowledgement

Gene perturbation:
Xi Chen

Methods & data analysis:
Zhengqing Ouyang



Methods:
Bokyung Choi