

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

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

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 $A: \mathbb{R}^n \rightarrow \mathbb{R}^n$ contains detailed information on regulatory information, e.g.

X_j positively regulates X_i $\frac{\partial A_i(x_1 \dots x_n)}{\partial x_j} > 0$

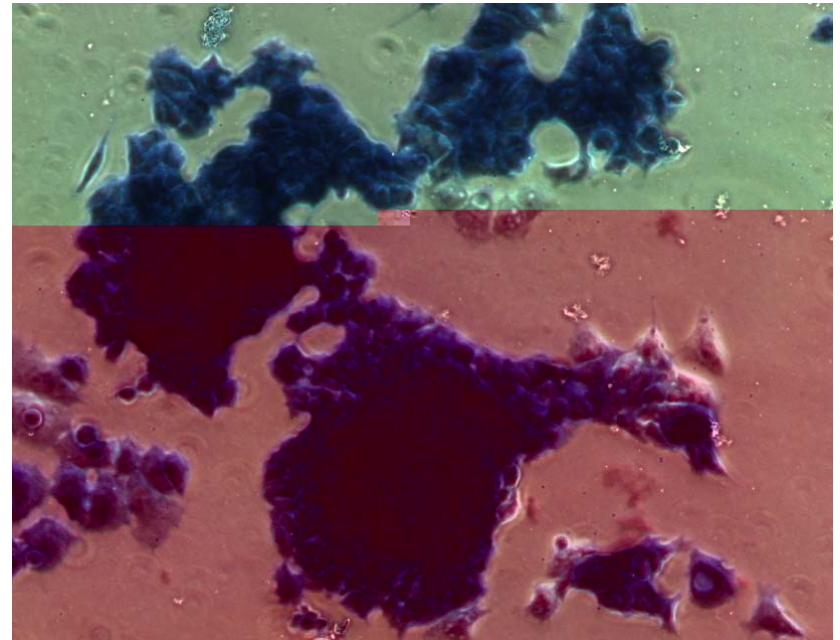
However, A is too complex to reconstruct from experiments based on current technology.

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)



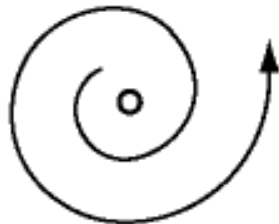
Equilibrium state

A state μ is an equilibrium state if μ

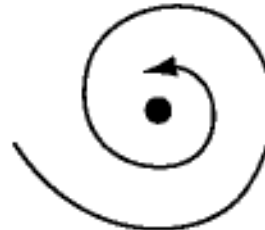
The equilibrium is stable if, once the system comes close to μ , it will stay close to μ from then on.

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

unstable



stable



Regulatory network

Suppose $X(0) = X^*$, then for small t

$$X(t) - X(0) \approx J(t - 0) \quad J = t$$

where J is the Jacobian matrix: $J_{ij} = \partial f_i / \partial x_j$

We propose to regard J as the regulatory network that maintains the equilibrium

Stability imposes a global constraint on the network: J 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

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

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

- Predictor: perturbation on J regulators

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

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

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

- 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
- Lyapunov stability

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

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

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

- This leads to the optimization of

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

- 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

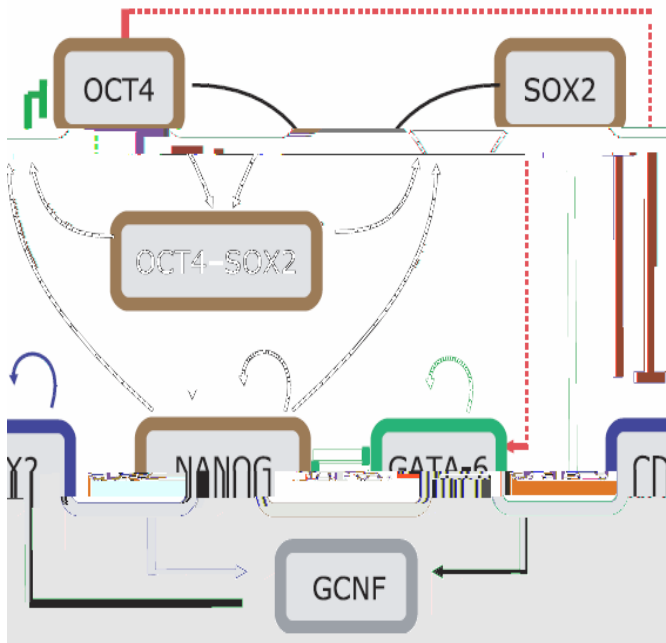
$$c_{ij} = 1 / a_{ij}$$

- Penalized loss function

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

Simulated data:

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



2 stable equilibrium states:
stem cell & endoderm

Use symbolic solver to get the
two networks

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

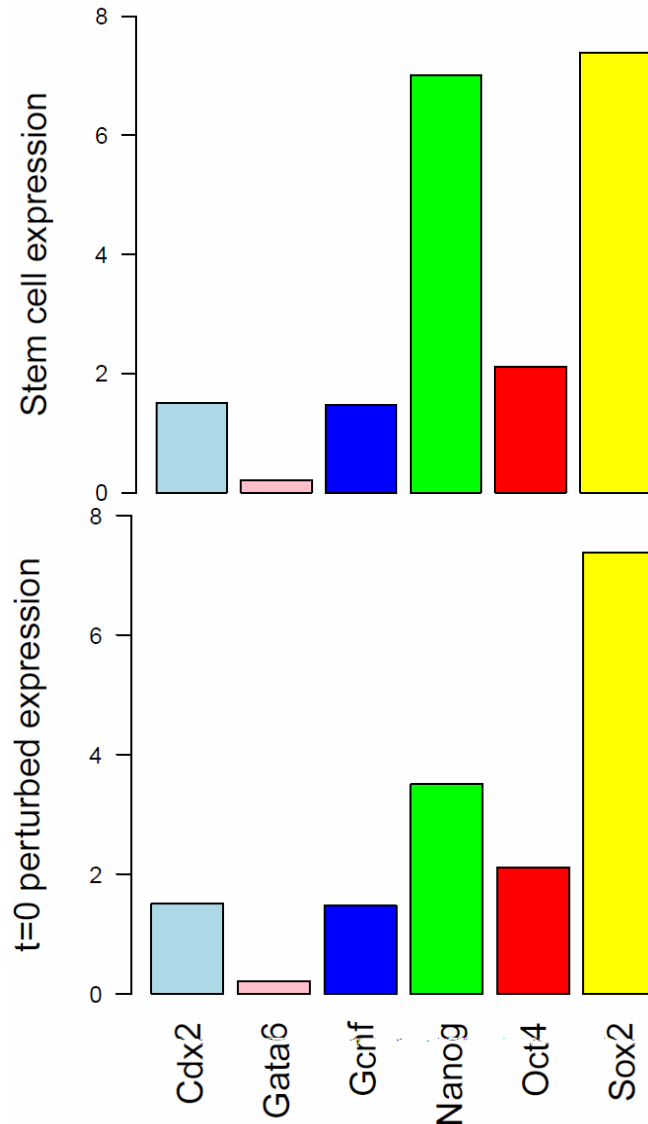
$$\frac{d[S]}{dt} = \frac{e_0 + e_1[O][S] + e_2[O][S][N]}{1 + d_0[O] + d_1[O][S] + d_2[O][S][N]} - \gamma_2[S]$$

$$\frac{d[N]}{dt} = \frac{e_3 + e_4[O][S][N]}{1 + f_0[O] + f_1[O][S] + f_2[O][S][N] + f_3[O][G]} - \gamma_3[N]$$

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

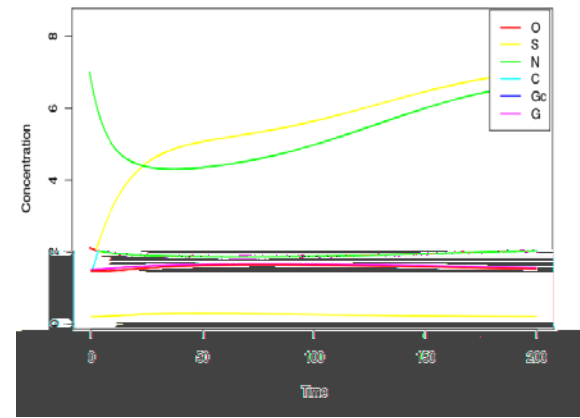
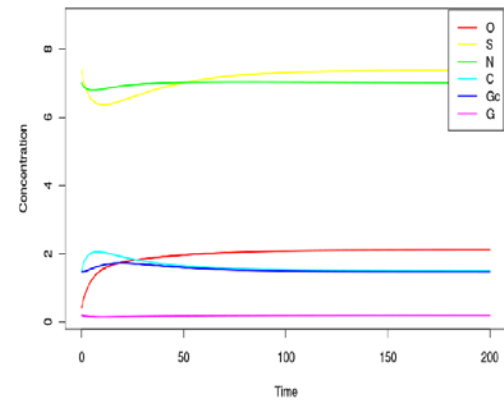
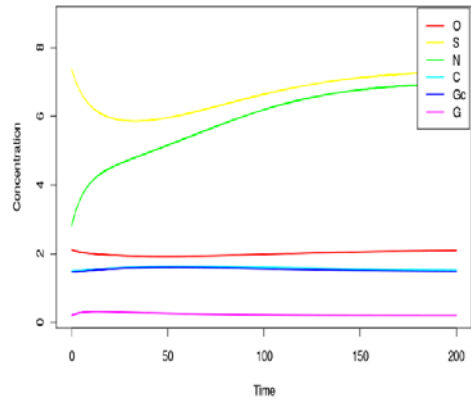
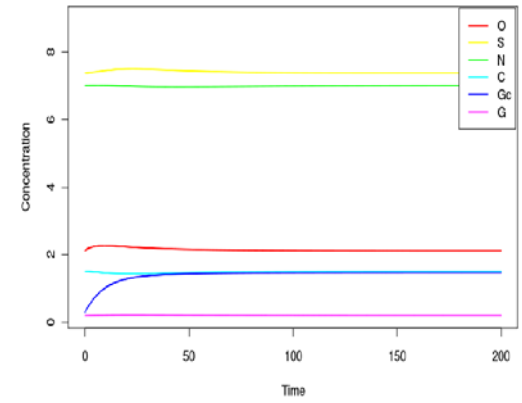
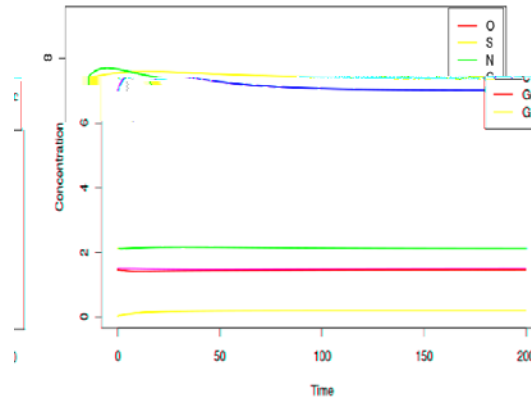
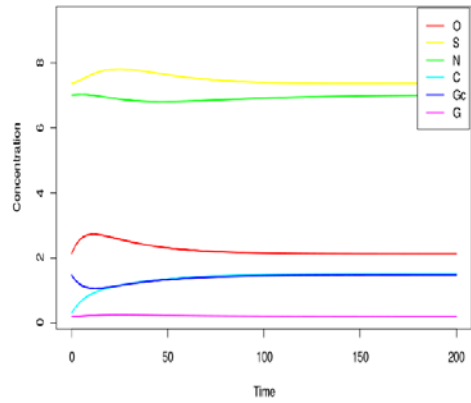
$$\frac{d[G]}{dt} = \frac{p_0 + p_1[Q] + p_2[G]}{1 + q_0[Q] + q_1[G] + q_2[N]} - \gamma_g[G]$$

Perturbation of stem cell state



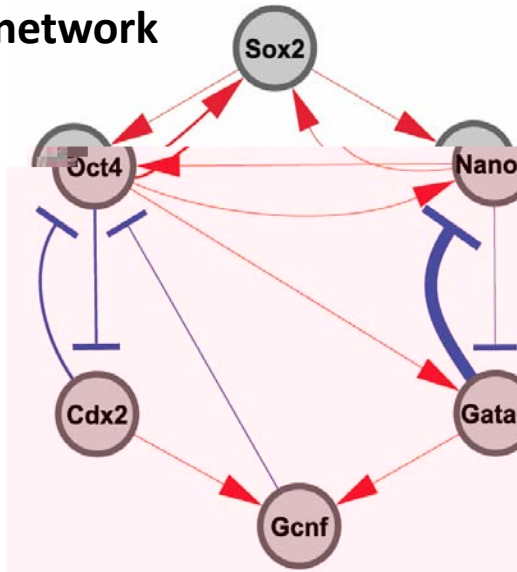
- Knockdown one of the six TFs in each experiment
- The TF expression is reduced by 50% (Nanog) at time $t=0$
- Simulate evolution of expression after perturbation

Time evolution after perturbation

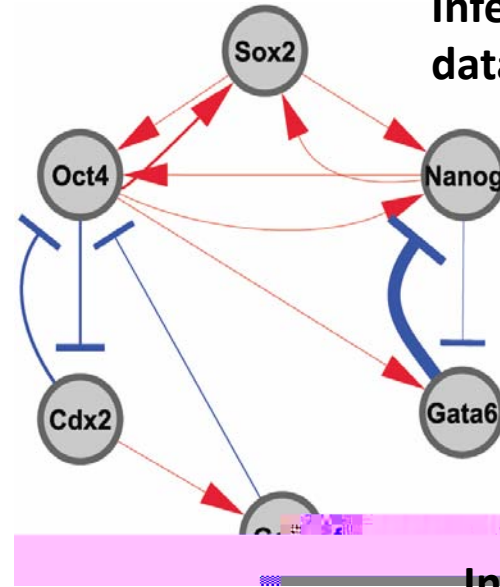


Network reconstruction results

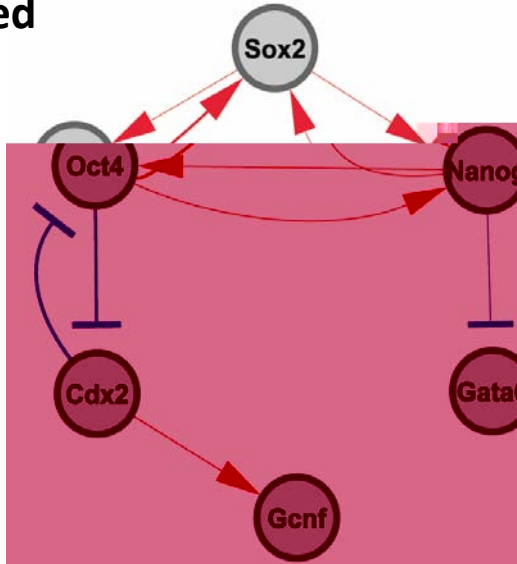
True network



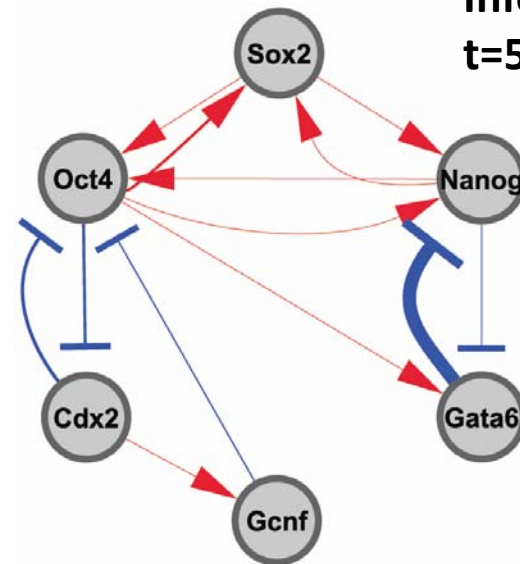
Inferred from data at t=1



Inferred t=5

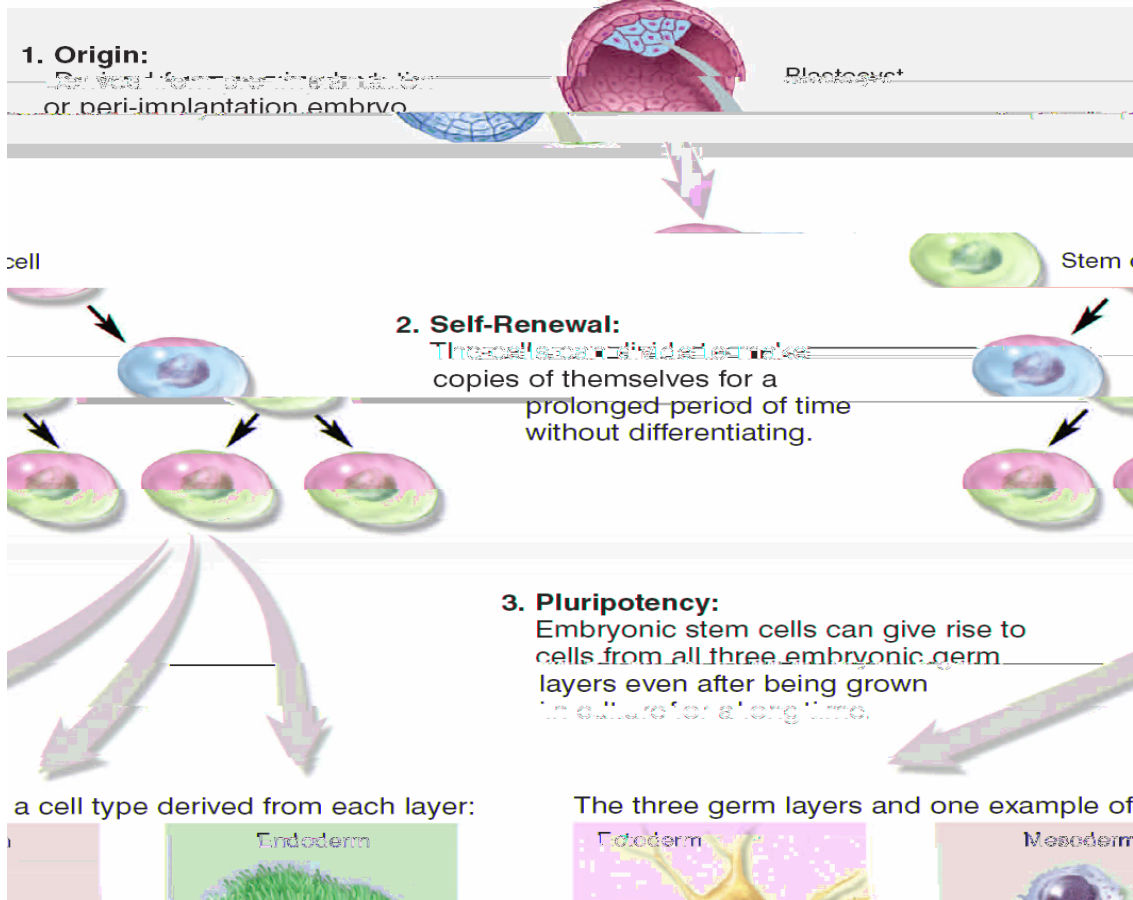


Inferred t=5 + prior

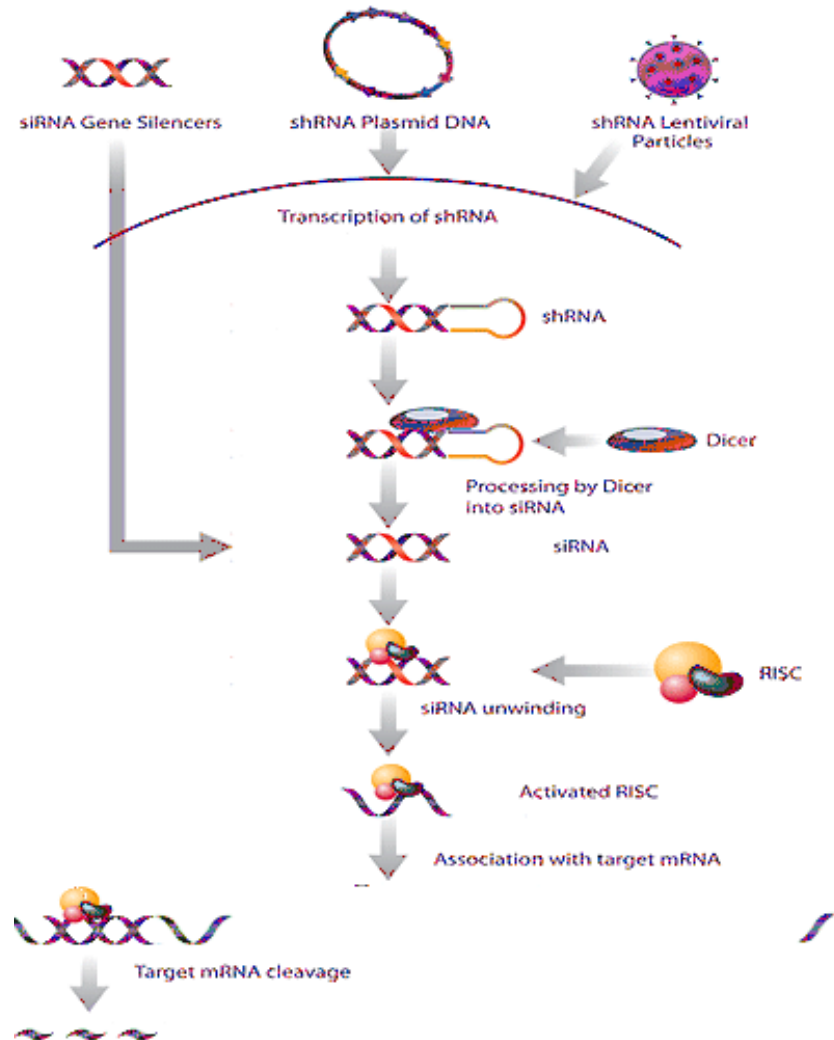


Real data

Embryonic stem cell



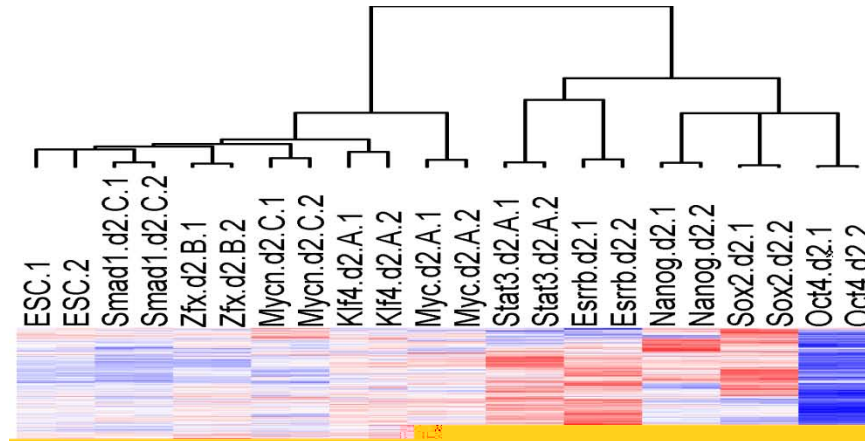
Gene knockdown by RNA interference



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

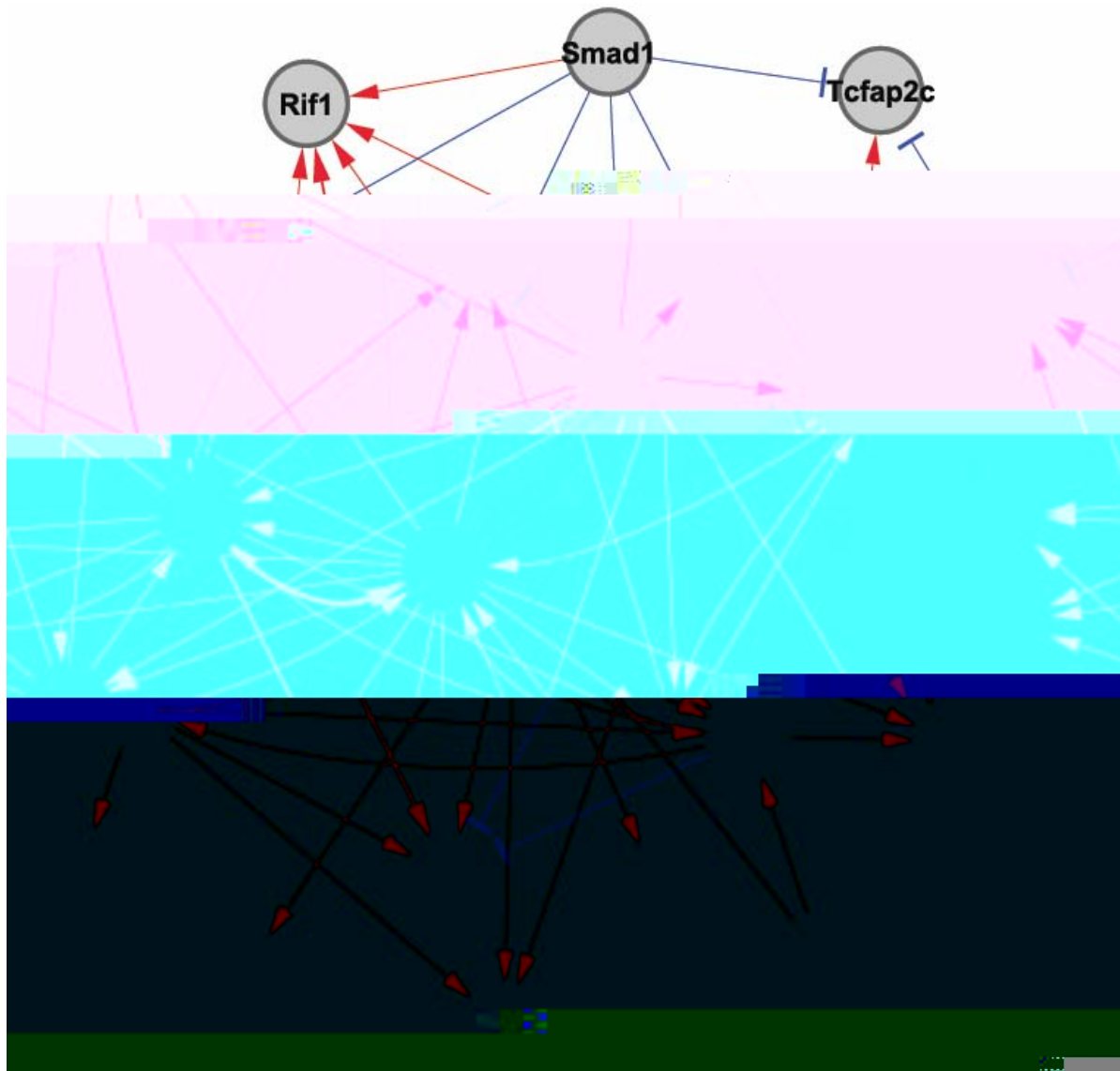
- Network reconstruction with ChIP information (ChIP-seq data from Chen et al 2008)
- Cross-validated for choice of λ

TF targets identified

- 3764 targets regulated by 10 TFs

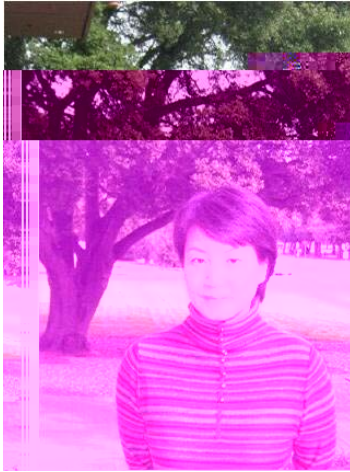
Oct4	Nanog	Sox2	Esrrb	Stat3	Klf4	Myc	Mycn	Zfx	Smad1
2362	588	461	1169	895	277	0	0	72	163

A subnetwork for important TFs

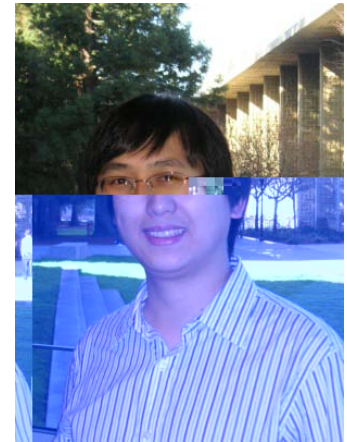


Acknowledgement

Gene perturbation:
Xi Chen



Methods & data analysis:
Zhengqing Ouyang



Methods:
Bokyung Choi

