ゲノムワイド遺伝子ネットワーク解析

東京大学医科学研究所 ヒトゲノム解析センター
宮野　悟

2013年1月16日　15:05ー15:25
第5回HPCI戦略プログラム合同研究交流会
理化学研究所　計算科学研究機構　6階講堂

「世界に一つだけの花」
私たちは一人一人異なる
ゲノムを持っている

「がんは日本の国民病」日本人
の半分が罹り、3分の1が亡く
なっている

四千万歩の男･伊能忠敬

- 伊能忠敬は，江戸幕府の事業として，1800年から1816年にかけて全国を歩いて測量をし，1821年に「大日本沿海輿地全図」が幕府に納められたといいます．伊能忠敬はその完成を見ずに1818年に死去しましたが，その後，仕上げの編纂作業が行われ，全部で21年の歳月をかけてこの地図は完成しています．

- それから200年後の現在，Google™マップに象徴されるように，だれもが世界中をナビゲーションして，目的の場所に行くことができます．

2001年のNHKの正月時代劇「四千万歩の男･伊能忠敬」（原作：井上ひさし「四千万歩の男」（講談社））として放映

地図を作った人びと

ー古代から観測衛星最前線にいたる地図製作の歴史ー

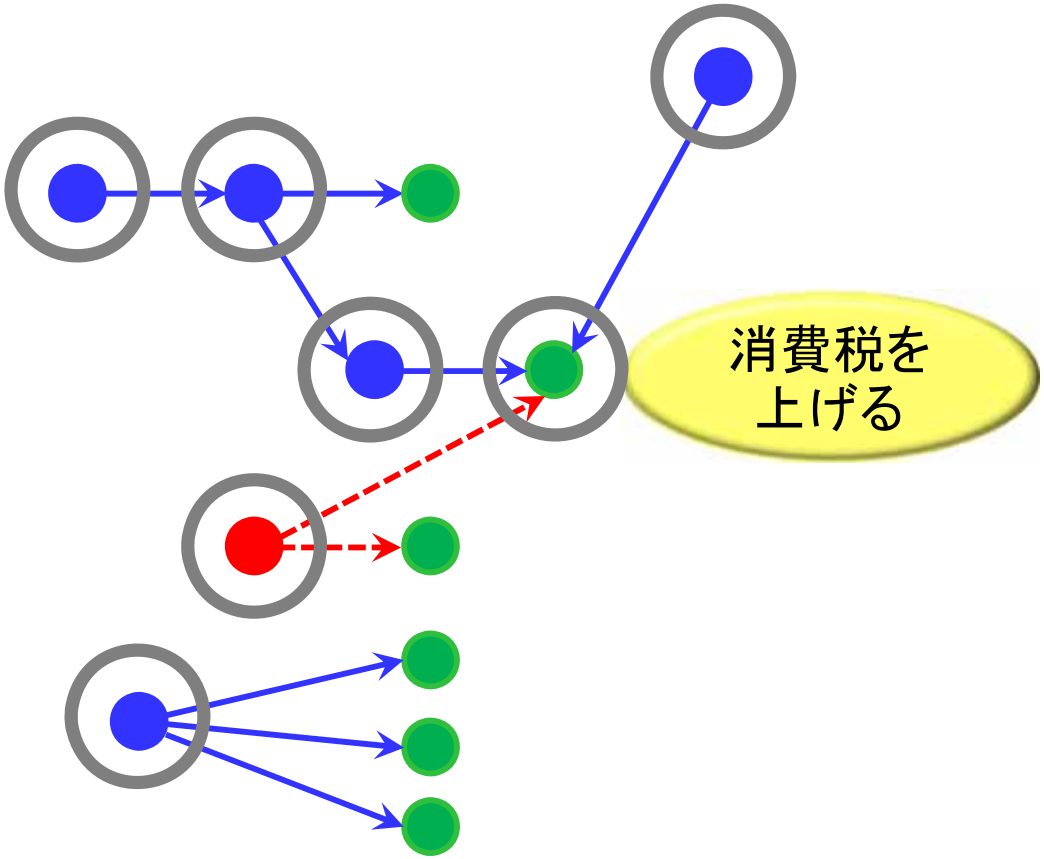
- John Noble Wilford著
- ピューリッツァー賞受賞作

遺伝子ネットワーク？ /!

遺伝子ネットワーク

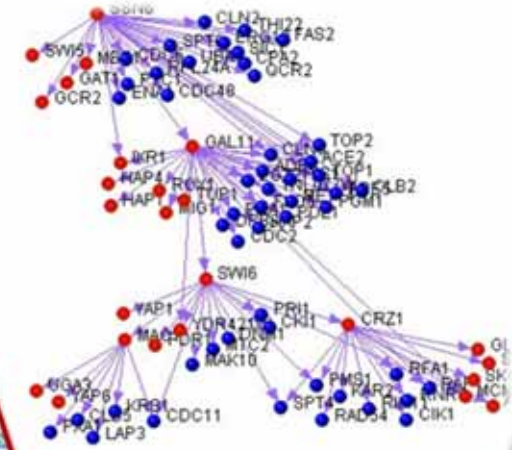
～「職場・社会における人間関係」のようなもの～

- 「働け」と命令する人
- 仕事のじゃまをする人

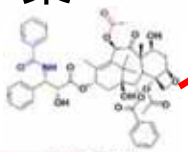


がんの悪性度や治療応答性、副作用の出やすさなどを規定している。

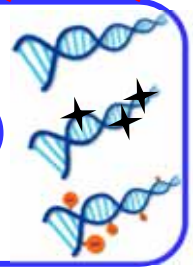
遺伝子の転写を
コントロールするのが
遺伝子ネットワーク



薬



- 個人個人で異なっている遺伝的要因（DNA）
- 腫瘍細胞に蓄積した遺伝子変異（がんゲノム）
- 環境要因によるDNAの修飾（エピゲノム）



遺伝子ネットワーク推定

- Gene Expression Profile Data (mRNA and microRNA)（DNAマイクロアレイ、RNA-Seq）
- それぞれの人の働きぶりを数値化したようなもの。
- 遺伝子ネットワークを推定するアルゴリズムは、職場・社会における人の働きぶりデータから、人間の制御構造とその影響力を推定。

遺伝子発現プロファイルデータからの遺伝子ネットワークの抽出法

◇◆地図の作り方・・・レシピー◆◇

- システムを捉えるための観測データ（地図作りの材料）
  - 刺激後の時系列データ
  - 遺伝子KD後のデータ、など

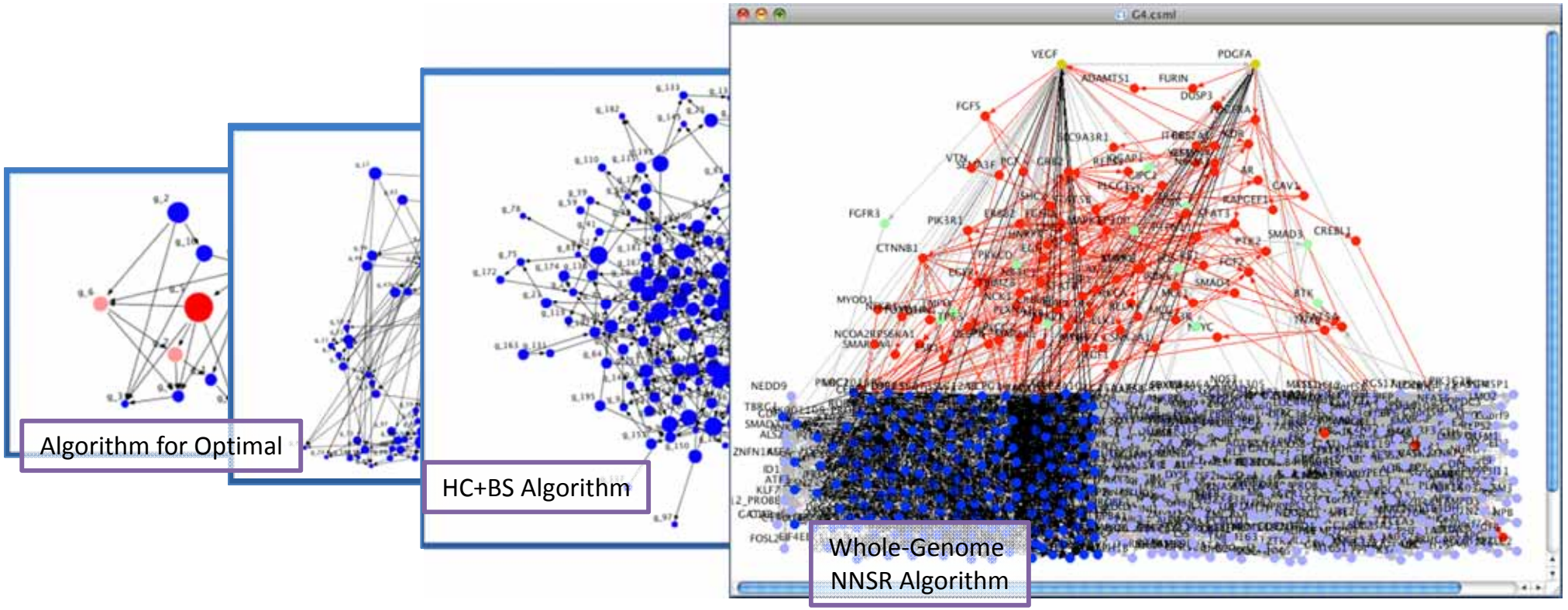
- 因果・制御関係の数理モデル
  1. BN: Bayesianネットワーク＋非線形回帰
     ⇒因果のネットワーク
  2. SSM: 状態空間モデル＋次元圧縮
     ⇒ シミュレーション
  3. L1正則化法に基づくネットワーク推定法
     ⇒ NetComparator: 複数時系列データに基づくネットワーク比較法
     ⇒ NetworkProfiler:

◇◆地図を作る道具◆◇

A Series of Programs on Supercomputer for mining gene networks of size from 30 to 20,000 (genome-wide)

Optimal to Locally Optimal

WR: Optimal Bayesian Networks of 32 Nodes



Q

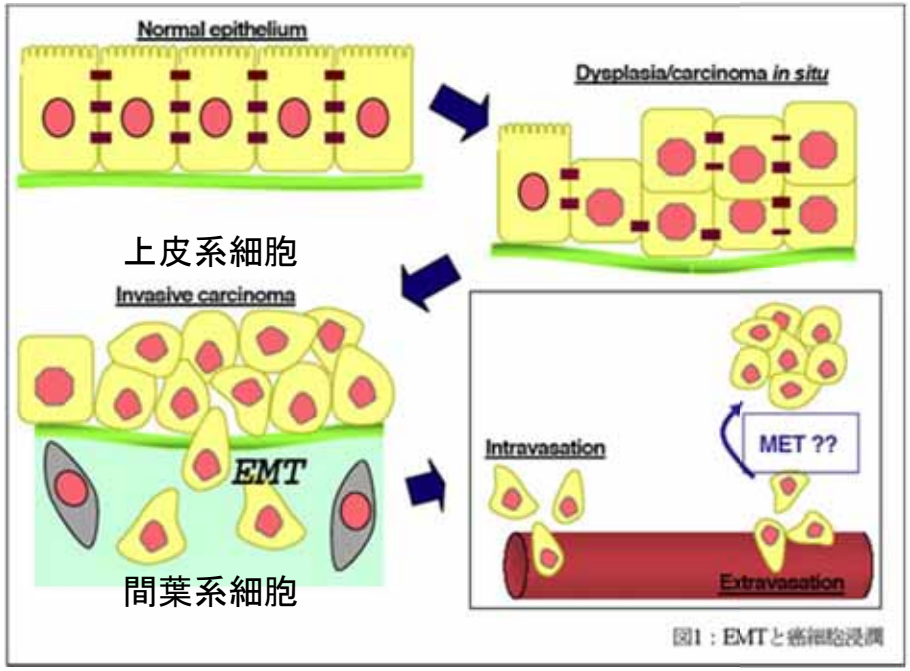
１セットの遺伝子発現プロファイルデータ（複数サンプルの遺伝子発現データ）から、スパコンで大規模遺伝子ネットワークを推定するといったい何が見えてくるのか？

- 数百のがんサンプルの遺伝子発現プロファイルデータから個々のがんの個性・多様性を分子ネットワーク（推定）として抽出する

スパコンで捉えたがんの黒幕たちのネットワーク
“上皮間葉転換を引き起こす遺伝子たち”

上皮間葉転換(Epithelial-Mesenchymal Transition **EMT**)

- 上皮系細胞が間葉系細胞に形態変化する現象
- がんの悪性度、がんの浸潤性、線維症に深く関連
- ゲフィチニブなどの抗がん剤を効かなくする原因



http://gantoku3.umin.jp/topics/saito.html

？

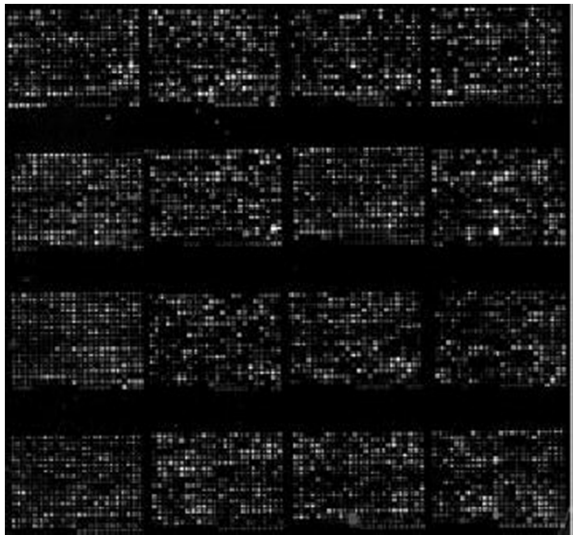
EMTを引き起こすいくつかの遺伝子は見つかっているが、そのメカニズムはよくわかっていない

１サンプルのトランスクリプトームデータから因果・制御のネットワークを抽出するのは無理

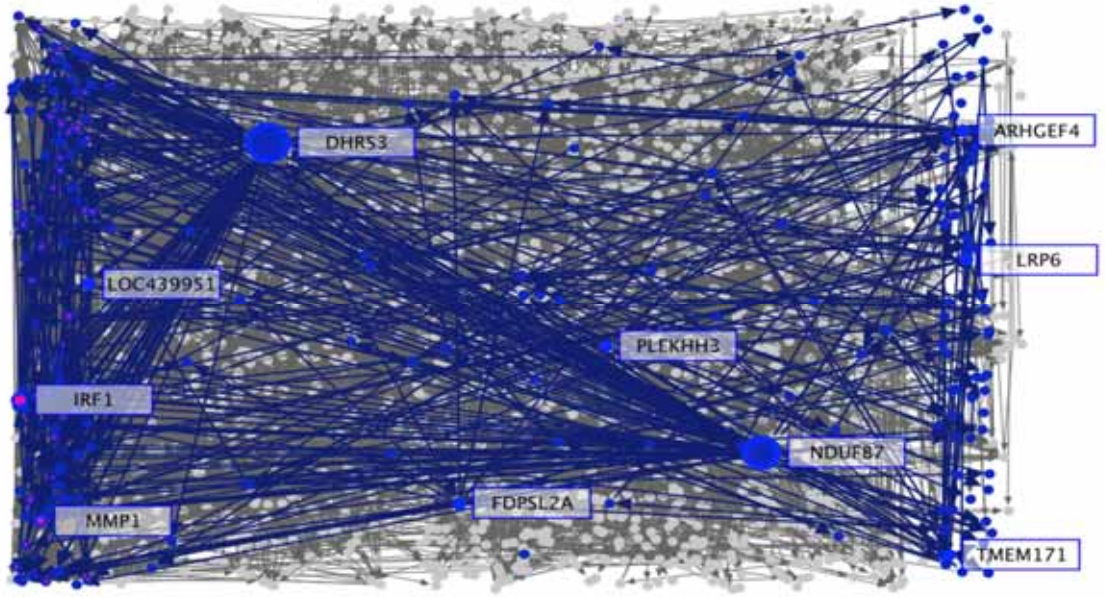
クラスタリング

２サンプルだとFold-Change Analysis

１サンプル

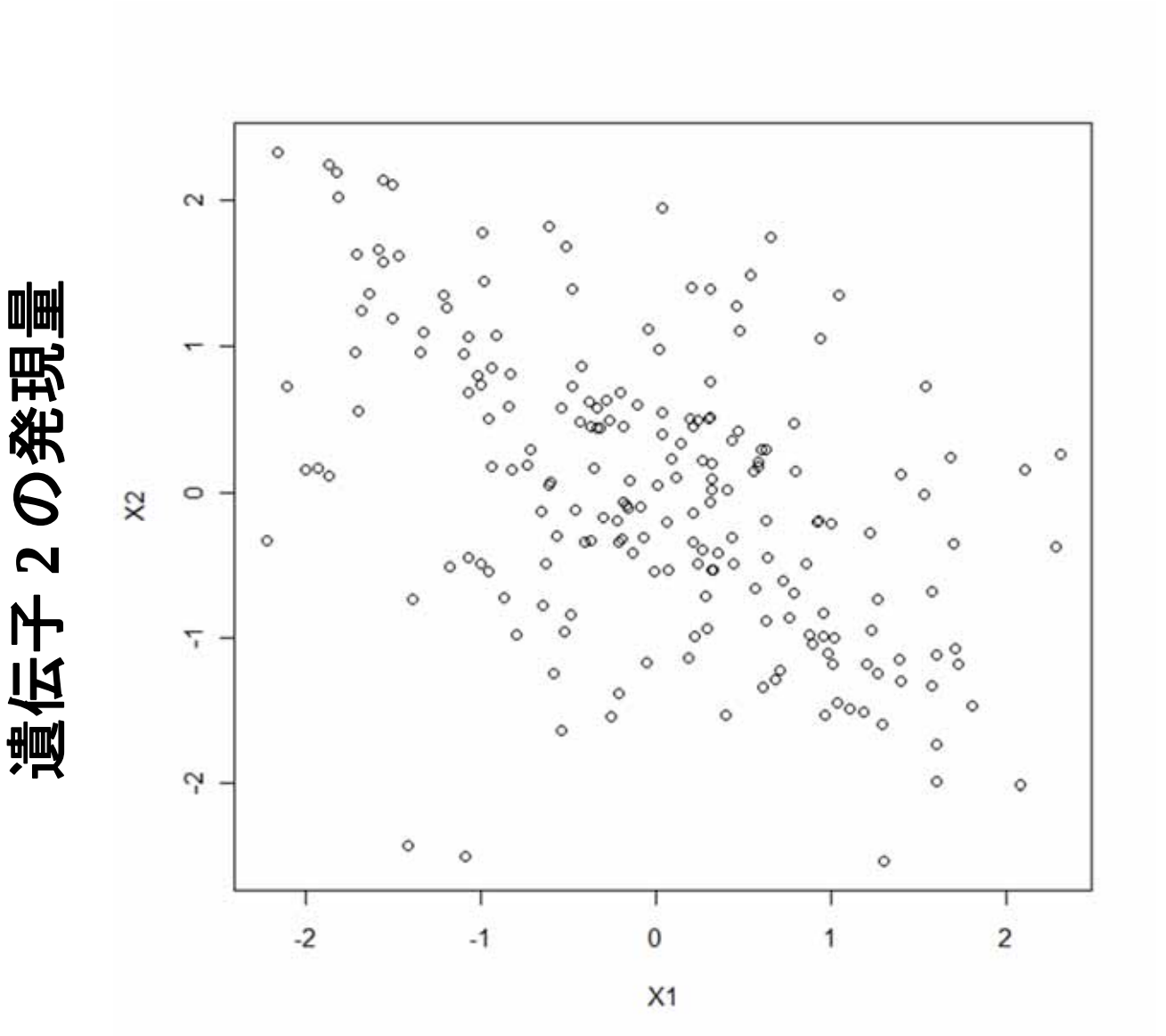


因果・制御のネットワーク



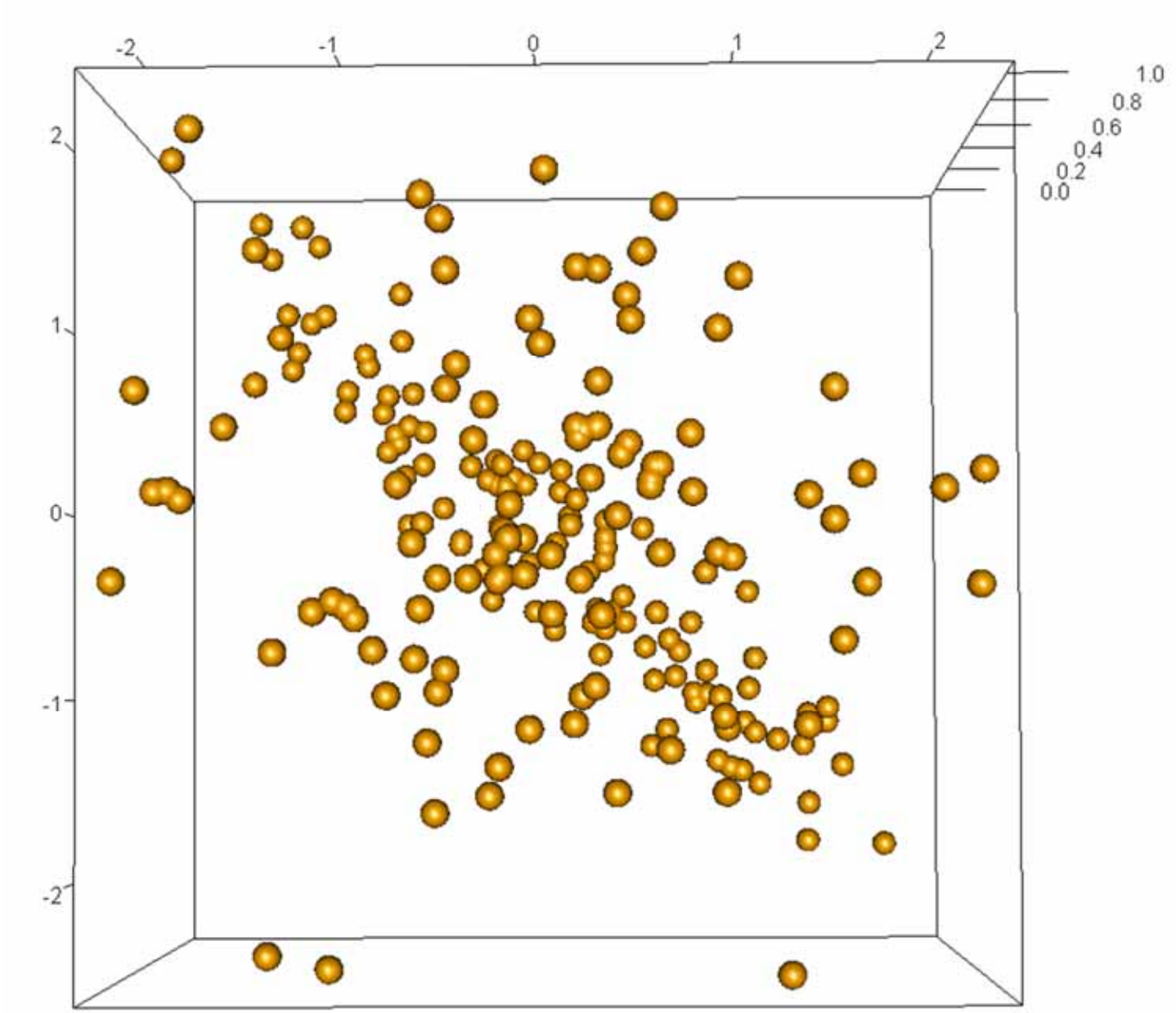
しかし、たくさんのサンプルがあると・・・

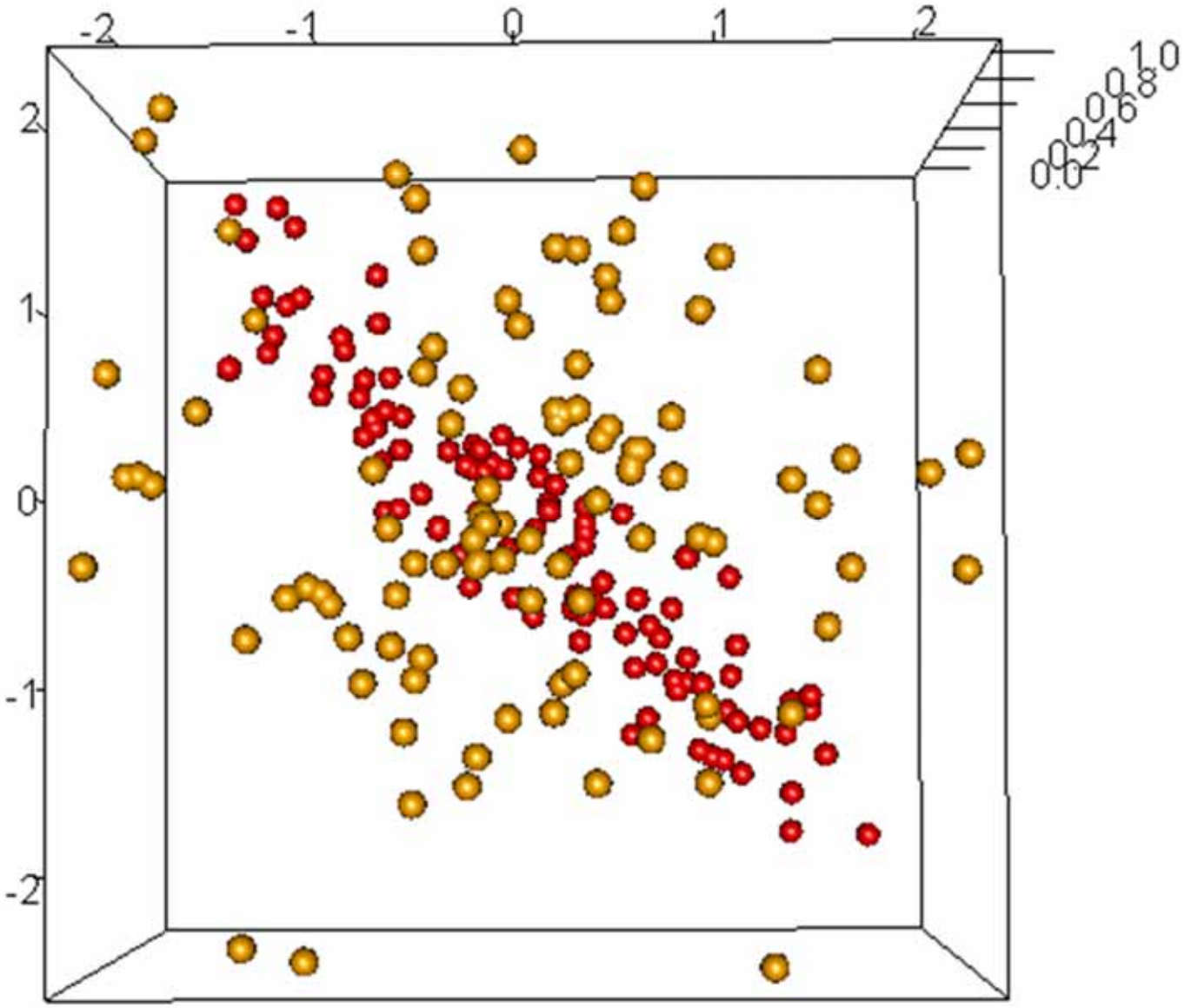
そのままデータを見ると相関はなさそう



がんの多様性に関係する因子

- 原発部位の違い（肺癌、胃癌、乳癌、大腸癌、…）
- 進行度の違い（Stage I, Stage II, …）
- 薬剤の効きやすさの違い（耐性と感受性）
- 性質の違い（運動性、接着性、浸潤性）
- 形状の違い（上皮細胞と間葉細胞）
- 生存リスクの違い
- ドライバー変異遺伝子の違い

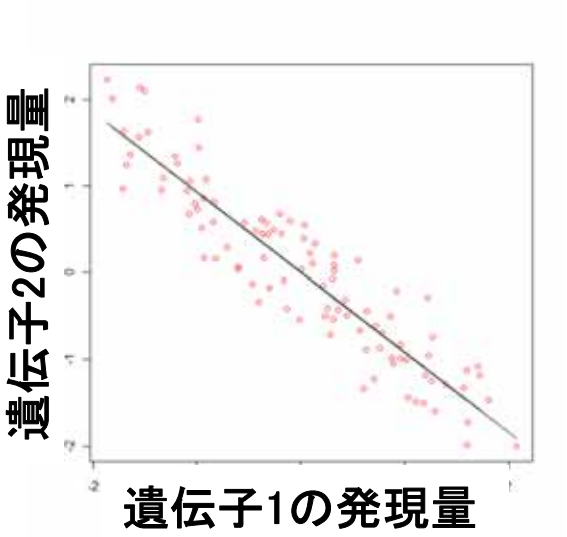

-2
-1
0
1
2
2
1
0
-1
-2
1.0
0.8
0.6
0.4
0.2
0.0

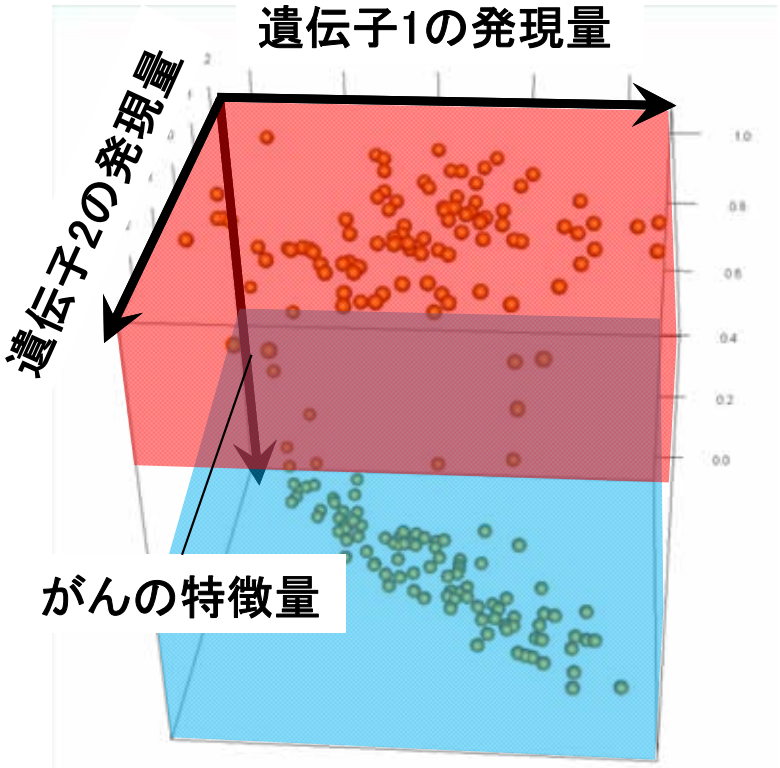

-2
-1
0
1
2
2
1
0
-1
-2
0.0
0.2
0.4
0.6
0.8
1.0

データの層別化による相関構造の違い

$$X_2 = -0.92 \times X_1 + \epsilon_2$$

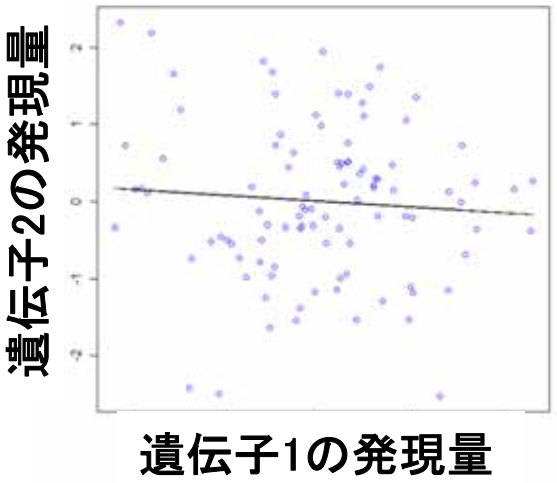
特徴量が低いとき





$$X_2 = -0.07 \times X_1 + \epsilon_2$$

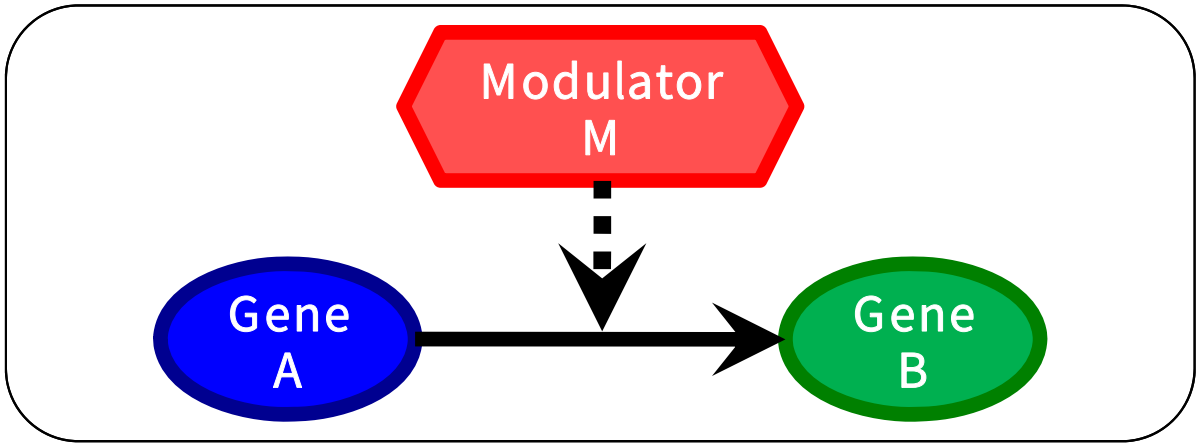
特徴量が高いとき



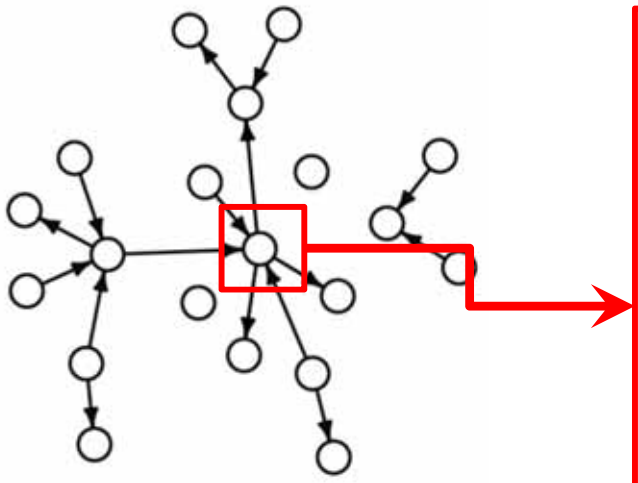
データの層別化

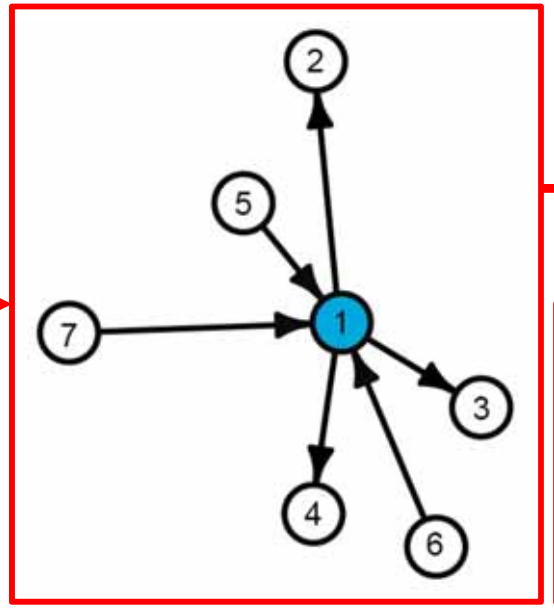
がんの「特徴量」を導入すれば、関係がでてくる
⇒サンプルそれぞれのネットワークを推定できる。

モジュレータ: 遺伝子AとBの条件付き従属性に影響を与える因子を導入



構造方程式モデル





構造方程式

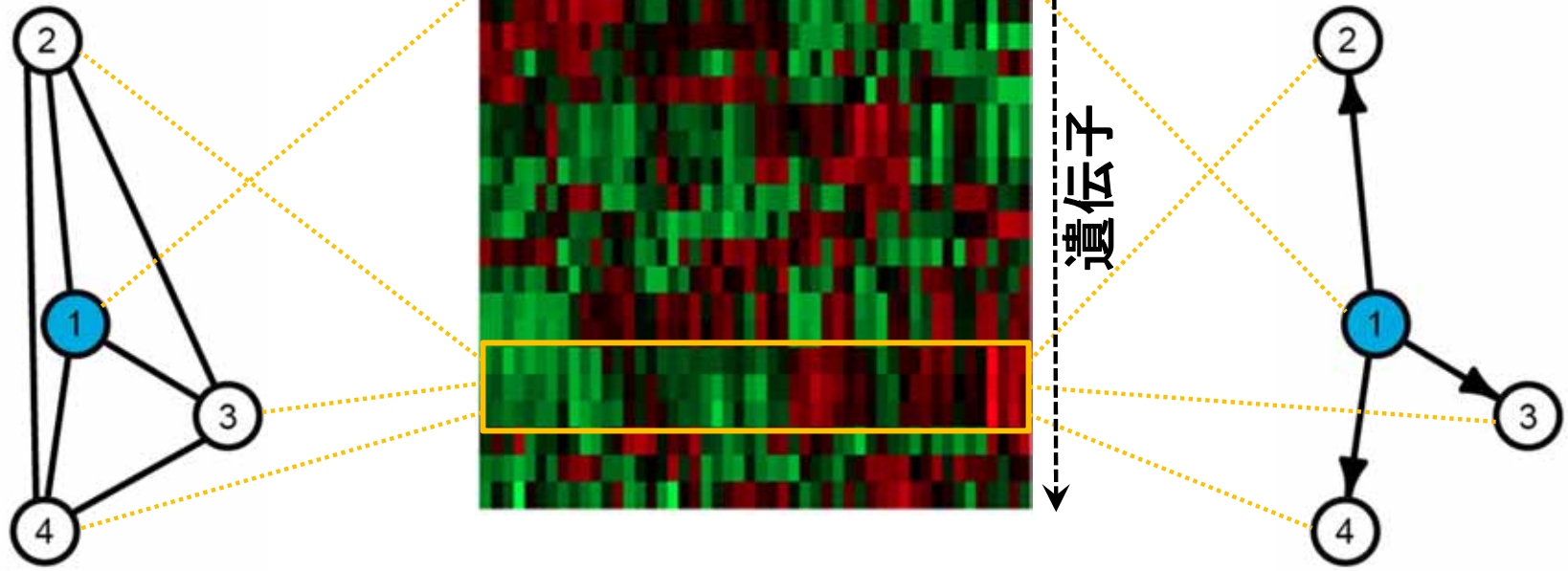
$$
\begin{aligned}
x_1 &= 4.6 \times x_5 + 4.2 \times x_6 + 0.7 \times x_7 + \epsilon_1 \\
x_2 &= -2.7 \times x_1 + 0.95 \times x_8 + \epsilon_2 \\
x_3 &= -1.4 \times x_1 + \epsilon_3 \\
x_4 &= -0.5 \times x_1 + \epsilon_4
\end{aligned}
$$

○：遺伝子（トランスクリプト）
→：条件付き依存関係
(相関ではない)

サンプル

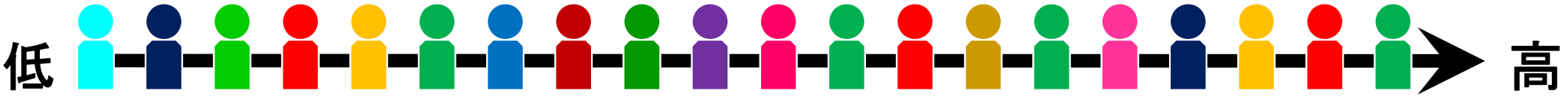
共発現ネットワーク

遺伝子ネットワーク



がんの個性を捉えるためのネットワーク推定法

モジュレータによる特徴空間にサンプルを射影し、
サンプルそれぞれの遺伝子ネットワークを推定する



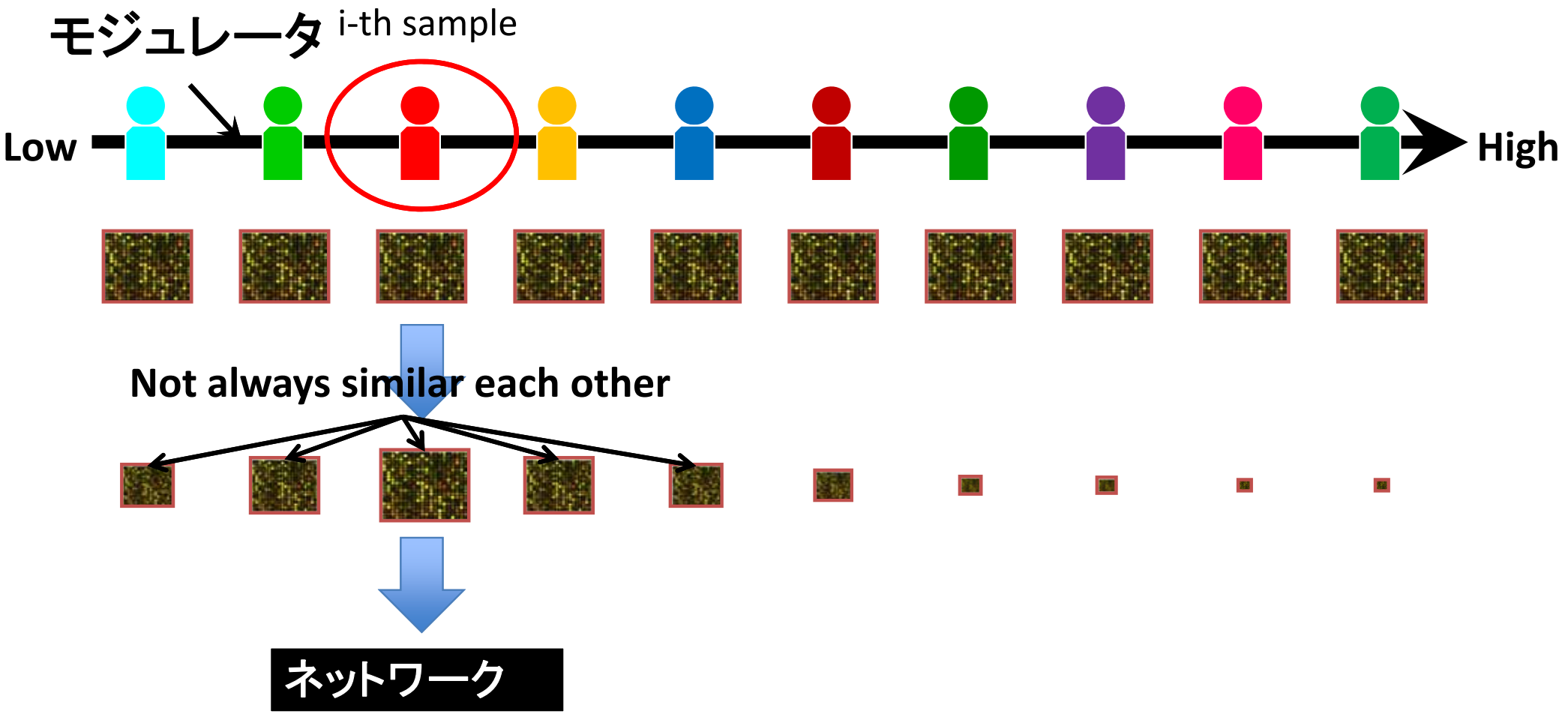
がんの特徴ベクトル(モジュレータ)

NetworkProfiler

L1正則化に基づくネットワーク推定法

i番目のサンプルの構造方程式モデルの係数を推定

Sample Weightingを使う

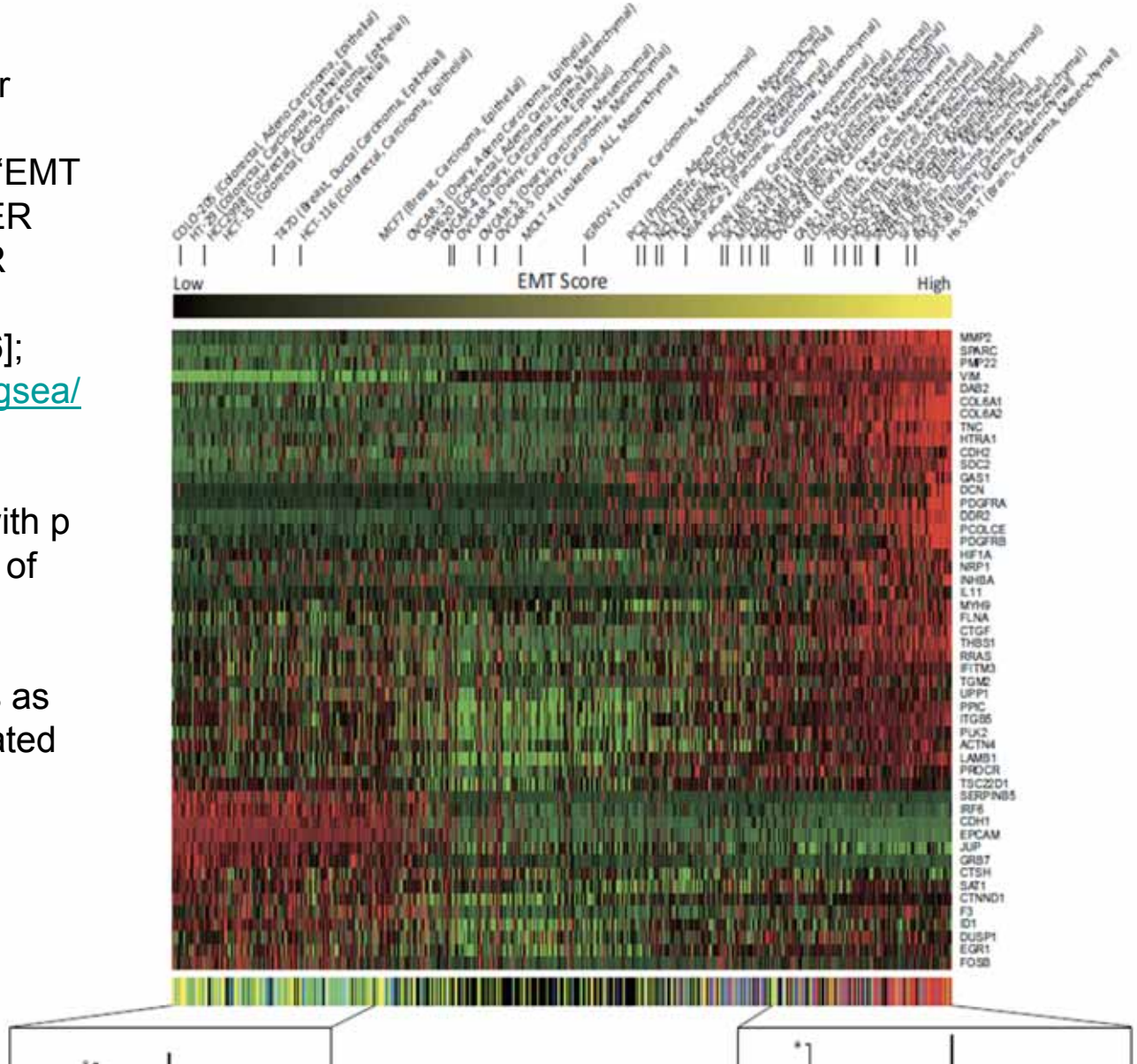


Modulator for EMT

We selected coherent 50 genes from 122 EMT signature genes to define the modulator for EMT (EEM, Niida et al., Bioinformatics, 2009)

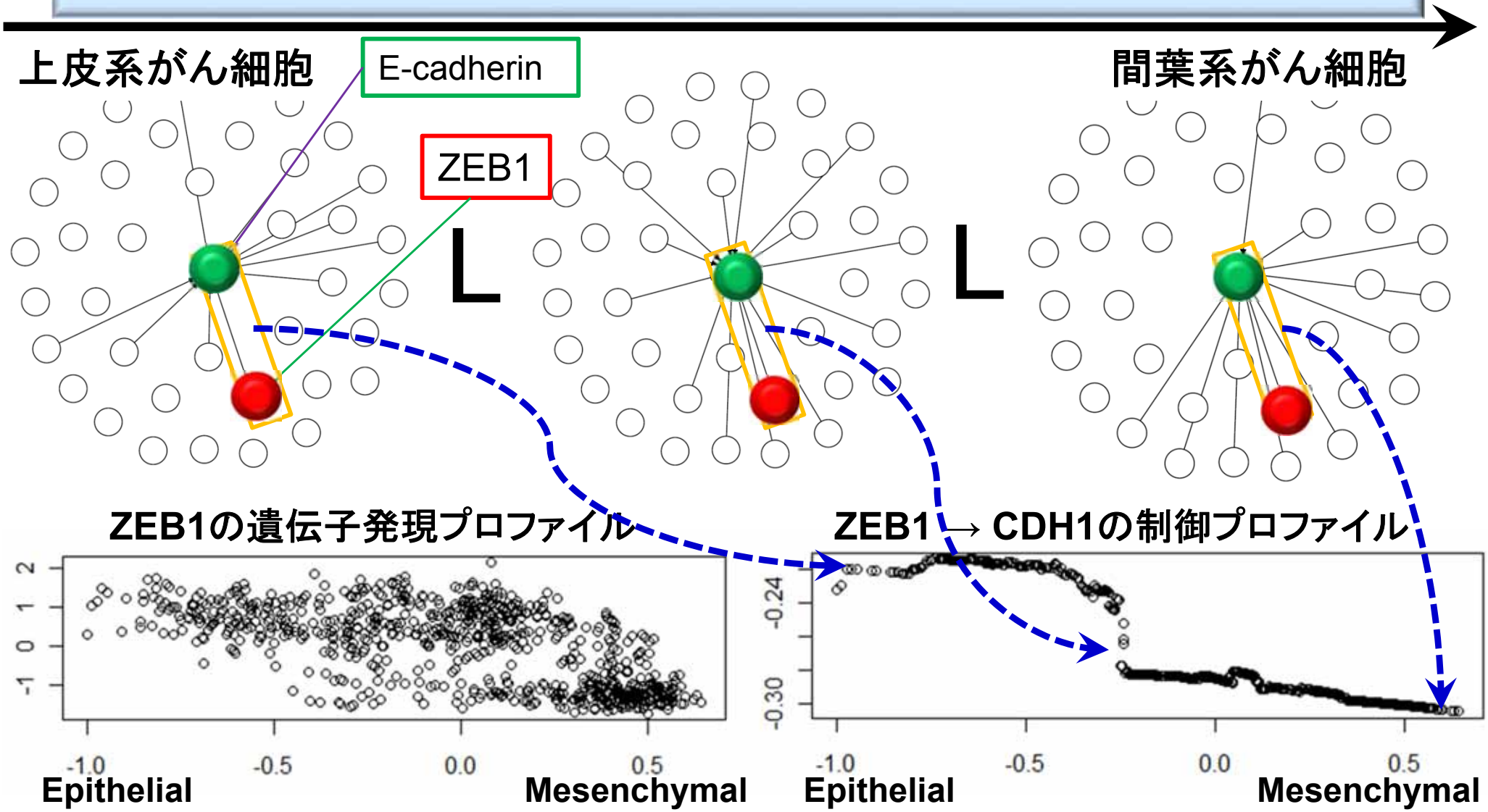
Signature-based hidden modulator extraction algorithm

1. Selected 122 genes labeled “EMT UP”, “EMT DN”, “JECHLINGER EMT UP”, and “JECHLINGER EMT DN”from Molecular Signatures Database v2.5 ([6]; http://www.broadinstitute.org/gsea/msigdb/index.jsp).
2. Then, narrowed the set to 50 functionally coherent genes with $p < 10^{-5}$ by using the extraction of expression module (EEM) .
3. Computed the first principal component of these 50 genes as hidden values of the EMT-related modulator

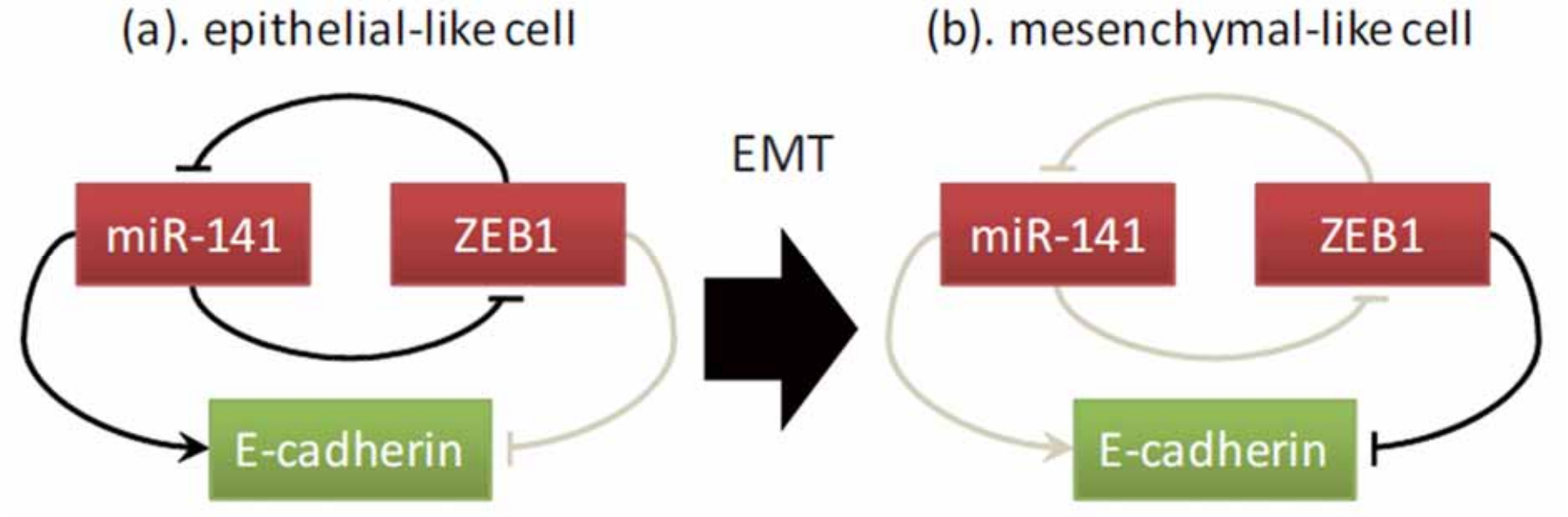


EMTで変化する大規模遺伝子ネットワークを 1024コアで論理的には3週間の計算・実際は3ヶ月かかった

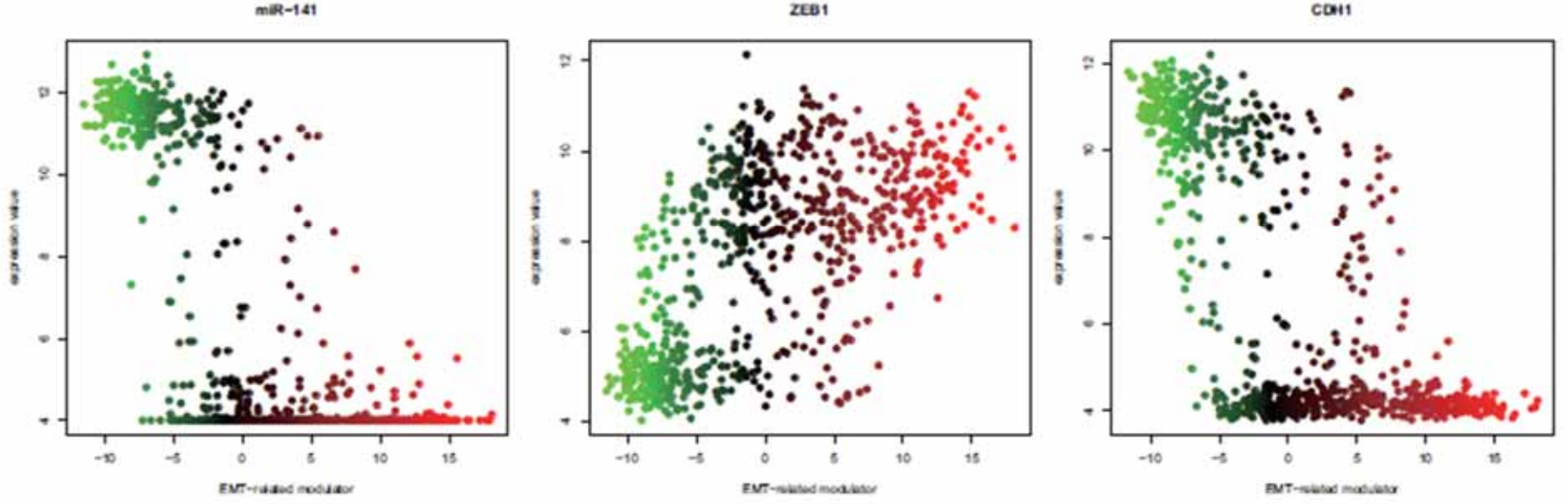
- **入力**:762のがん細胞の遺伝子発現プロファイルデータ。各がん細胞に対するEMTモジュレータ値（EMTの度合を定義したもの）。
- **出力**:13,508個の遺伝子から構成される遺伝子ネットワークを762個出力。EMTの度合で遺伝子ネットワークの構造がどのように変化するかが暴き出された。

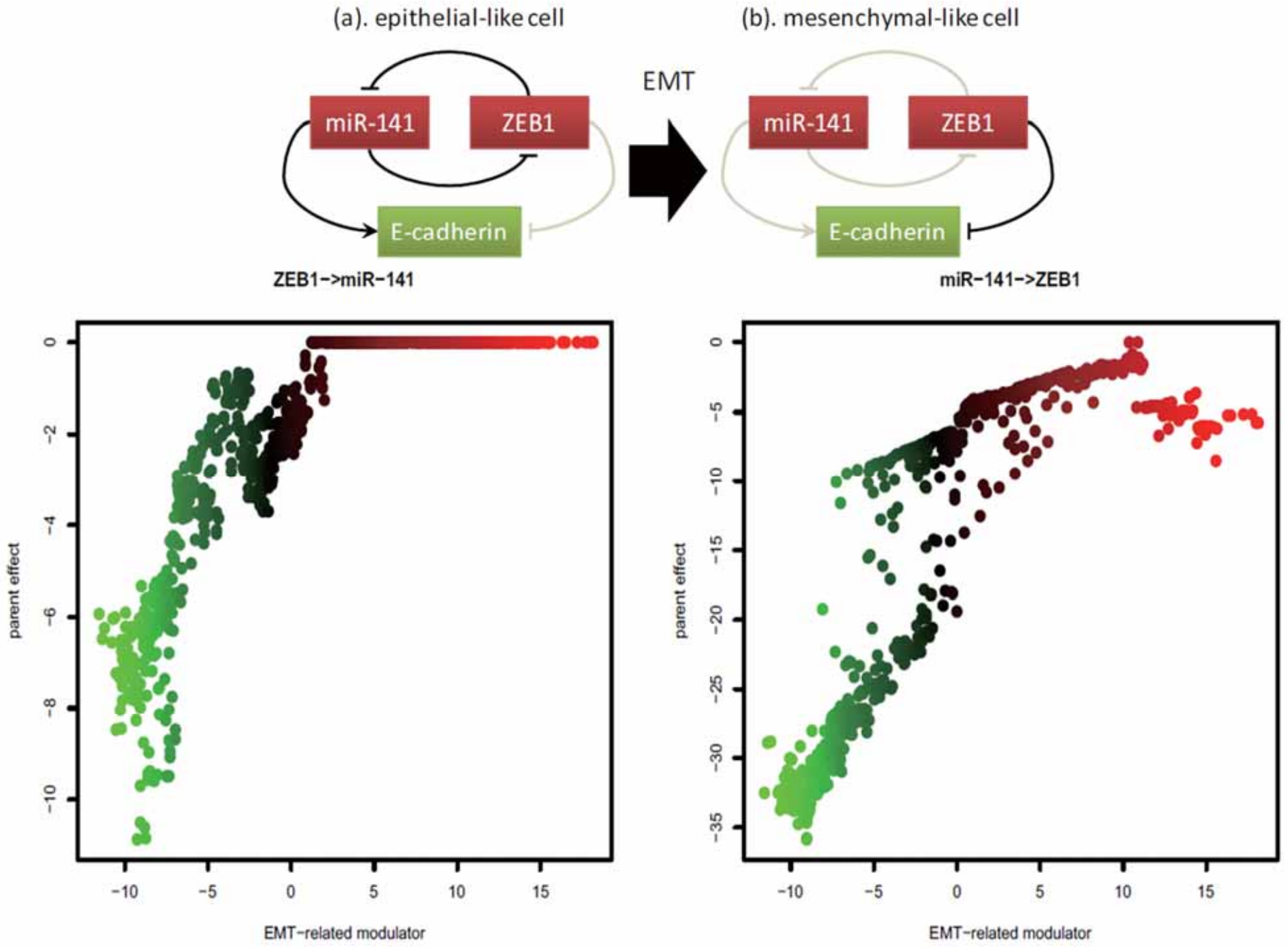


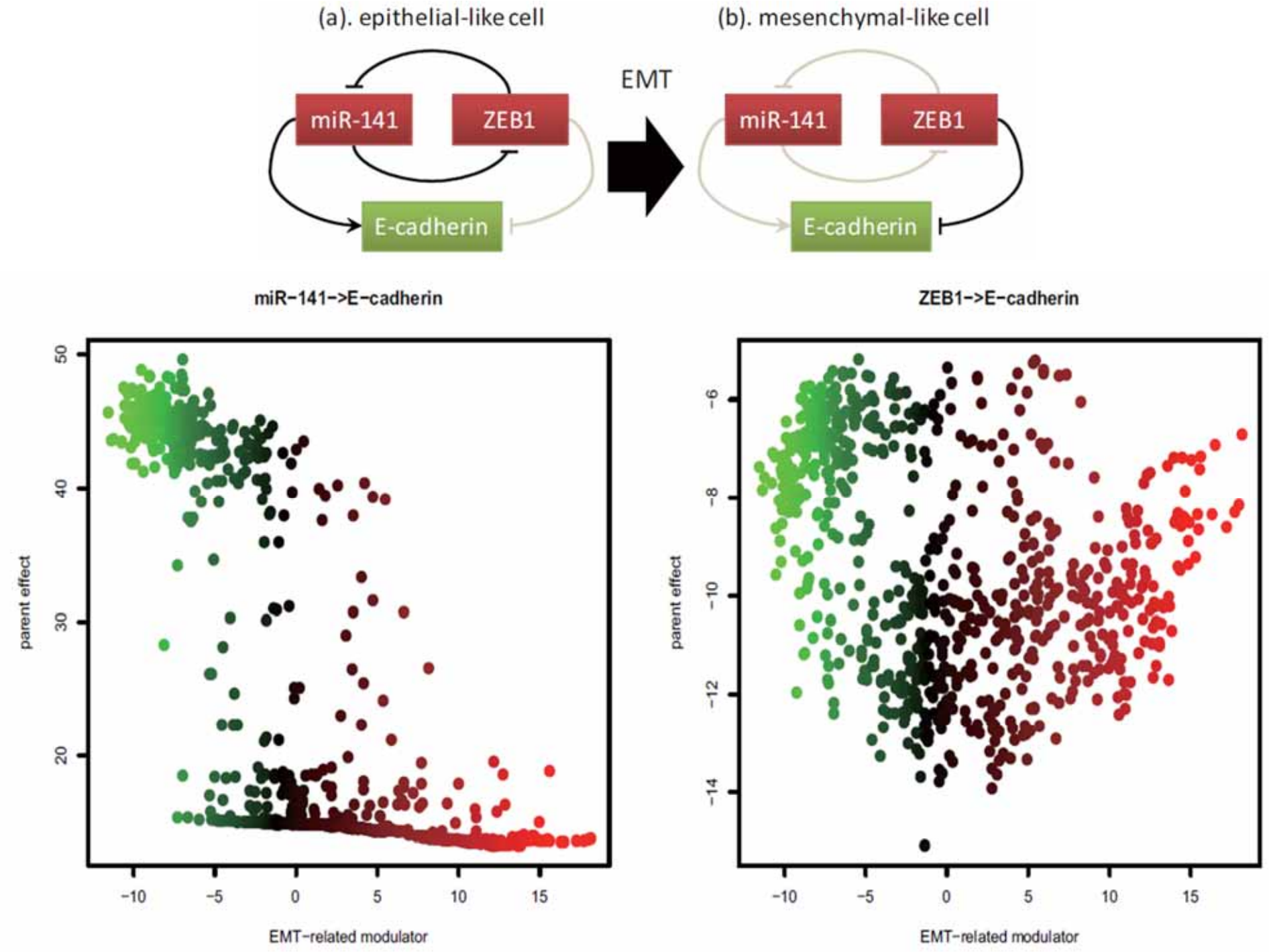
Upstream Regulatory Changes of E-cadherin



(c). The green and red colors indicate epithelial- and mesenchymal-like cells, respectively.




(a). epithelial-like cell
(b). mesenchymal-like cell
EMT
miR-141
ZEB1
E-cadherin
miR-141
ZEB1
E-cadherin
ZEB1->miR-141
miR-141->ZEB1
parent effect
EMT-related modulator
parent effect
EMT-related modulator


(a). epithelial-like cell
(b). mesenchymal-like cell
EMT
miR-141
ZEB1
E-cadherin
miR-141
ZEB1
E-cadherin
miR−141−>E−cadherin
ZEB1−>E−cadherin
parent effect
EMT−related modulator
parent effect
EMT−related modulator

スパコンが予測したEMT制御因子と新規遺伝子KLF5

スパコンによる大規模データ解析の有効性

Shimamura et al. PLoS ONE. 2011.

regulator	type	regulatory effect change	Evidence
IRF6	A	101.04	
miR-141	A	87.58	Nat Cell Biol 10(5): 593-601, 2008
GRHL2	A	64.13	Cancer Res 72(9):2440-53, 2012
ZEB1 (SIP1)	I	50.72	Mol Cell 7(6): 1267-78, 2001
LSR	I	46.89	
miR-200b	A	31.55	Nat Cell Biol 10(5): 593-601, 2008
KLF4	A	26.28	J Biol Chem 285(22):16854-63, 2010
OVOL2	A	22.08	
miR-200a	A	17.70	Nat Cell Biol 10(5): 593-601, 2008
FOXA2	A	17.26	Cancer Res 70(5):2115-25, 2010
TCF4 (E2-2)	I	14.15	J Cell Sci 122(Pt 7): 1014-24, 2009
ELF3	A	13.58	
ZNF217	A	13.53	
MYB	A	12.50	
KLF5	A	12.42	PLoS ONE 6(6): e20804, 2011
miR-192	A	12.30	PNAS 104(9): 3432-7, 2007
FOXA1	A	11.69	Cancer Res 70(5):2115-25, 2010
ZNF165	A	11.39	
NKX2-1	A	11.21	Cancer Res 69(7):2783-91, 2009
HNF1B	A	11.08	
TFE3	A	11.01	
ZEB2 (δEF)	I	10.66	Oncogene 24(14):2375-85, 2005
TRIM29	I	9.87	
SNAI2	I	9.74	Cancer Res 62(6): 1613-8, 2002

- スパコンがEMT制御因子として予測したトップ24遺伝子のうち12遺伝子が「当たっていた」。残り12遺伝子についてはEMTに関する直接的な結果は無し。
- システムがん・高橋隆がKLF5のノックダウンでEMTが起こることを実験で証明。
- その後、米国でGRHL2がEMT誘導することが発表される。

私の肺がんのシステムはどうなっているのか？

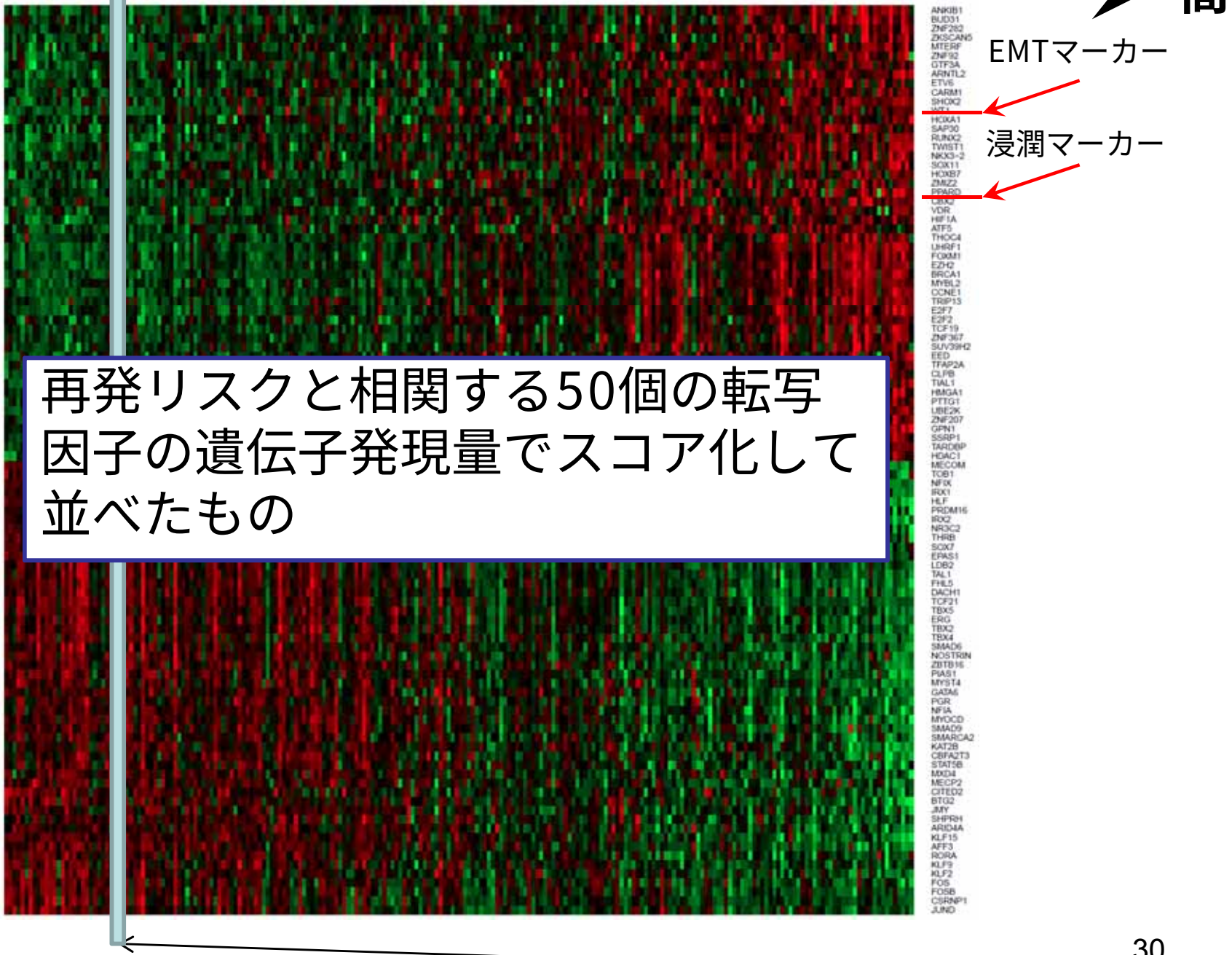
再発リスクスコアをモジュレータにすると

低

再発リスク

高

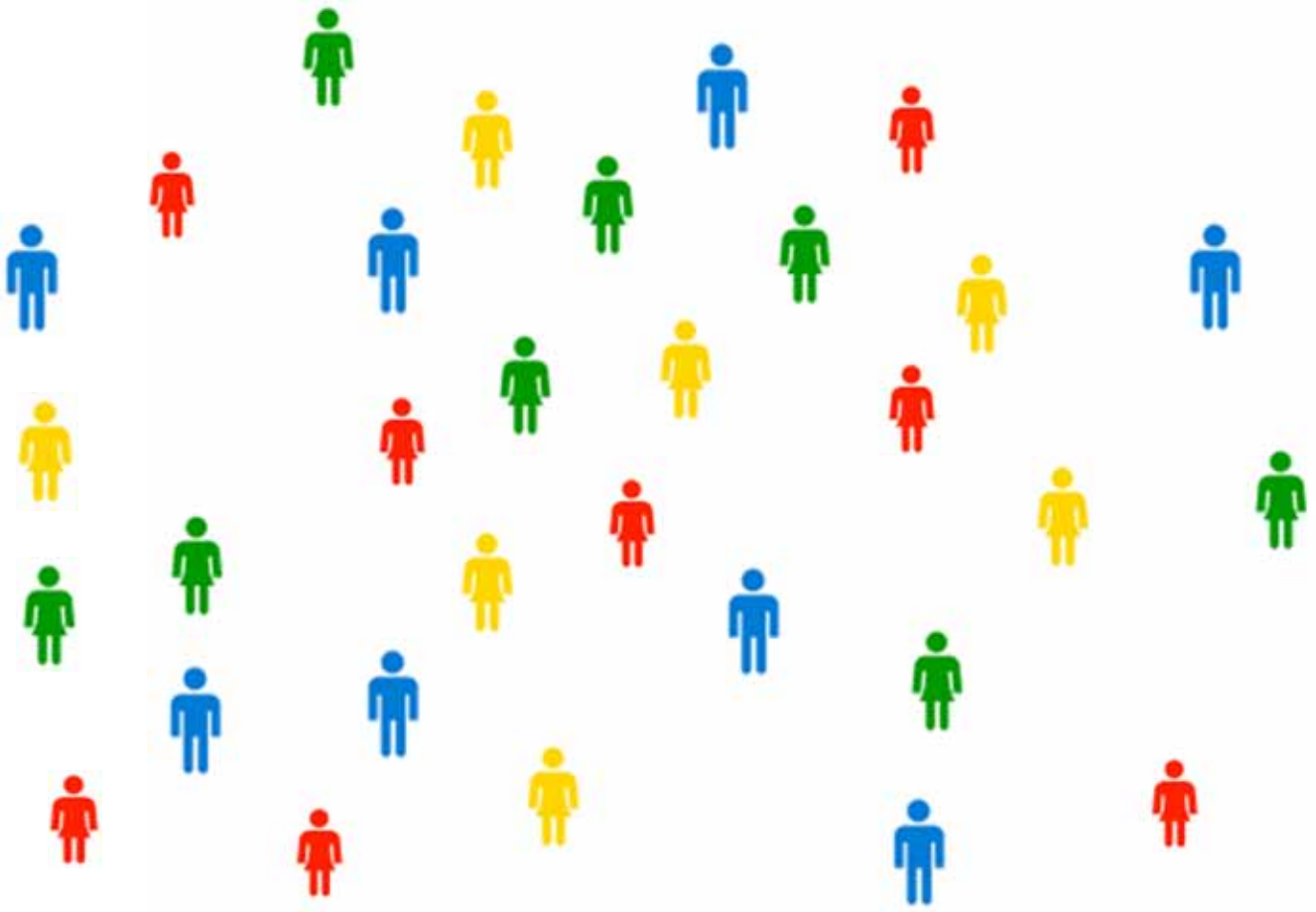
マイクロアレイで遺伝子発現を計測



再発リスクと相関する50個の転写因子の遺伝子発現量でスコア化して並べたもの

これが一人のひとのすべての遺伝子発現データになります。

数百人のがん検体



低 ⟷ 高

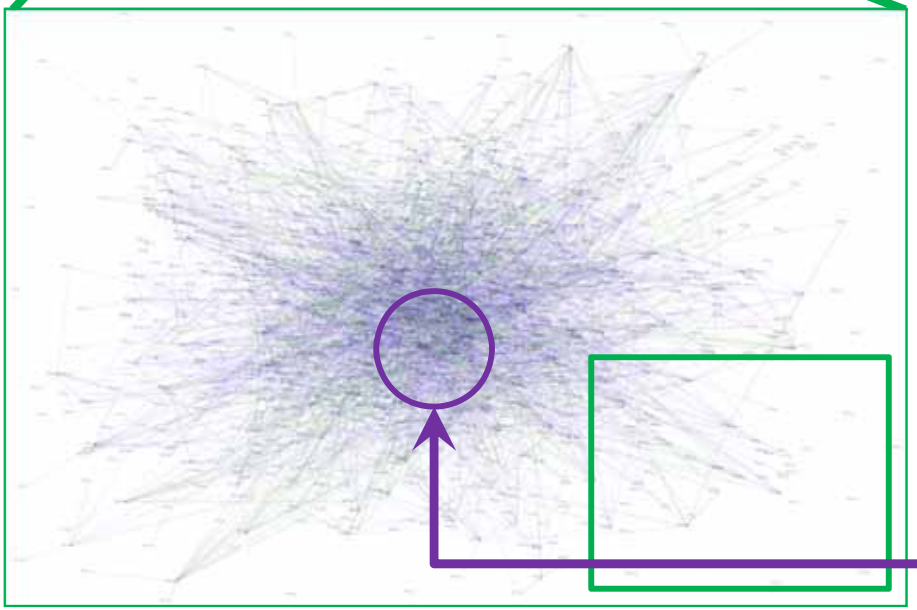
再発リスク

スパコンがあぶり出した再発リスクに関わる肺がん患者のパーソナル遺伝子ネットワーク

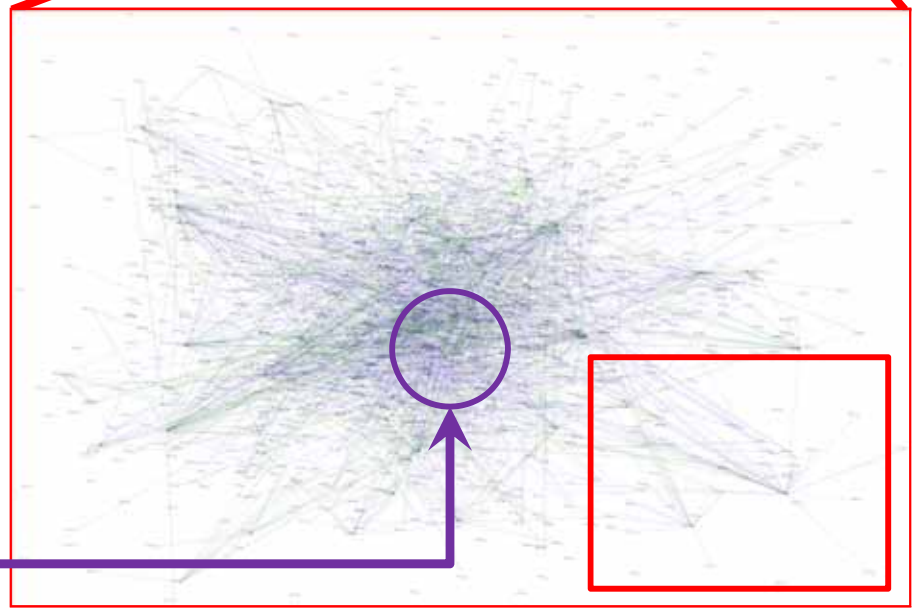
低 高

再発リスクをモジュレーターとしたときの 226 症例各々の遺伝子ネットワークを推定し、再発リスクが高い患者と低い患者のシステムの違いを比較

再発リスク



データ:
226症例の肺線がん患者の遺伝子発現データ（国立がん研究センター横田淳先生、河野隆志先生）



再発リスクが最も低い患者のシステム

再発リスクが最も高い患者のシステム

CTGF (Connective Tissue Growth Factor)
増殖･分化を制御するTGF-βの下流因子で、TGF-βの間質繊維化促進作用を仲介
増殖促進、遊走、細胞外基質算出、血管新生作用を呈する

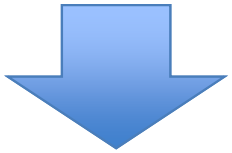
薬剤に対して強くしている遺伝子群をネットワークで探す

エルロチニブの薬剤耐性に関わる遺伝子ネットワークの推定

160 個の NSCLC 細胞株 に対する
22277 プローブのマイクロアレイデータ

モジュレータ: 各細胞株のエルロチニブに対するGI50

- 薬剤未処理群に比して細胞増殖を 50 %抑制する濃度
- GI50が小⇒感受性、GI50が大⇒耐性



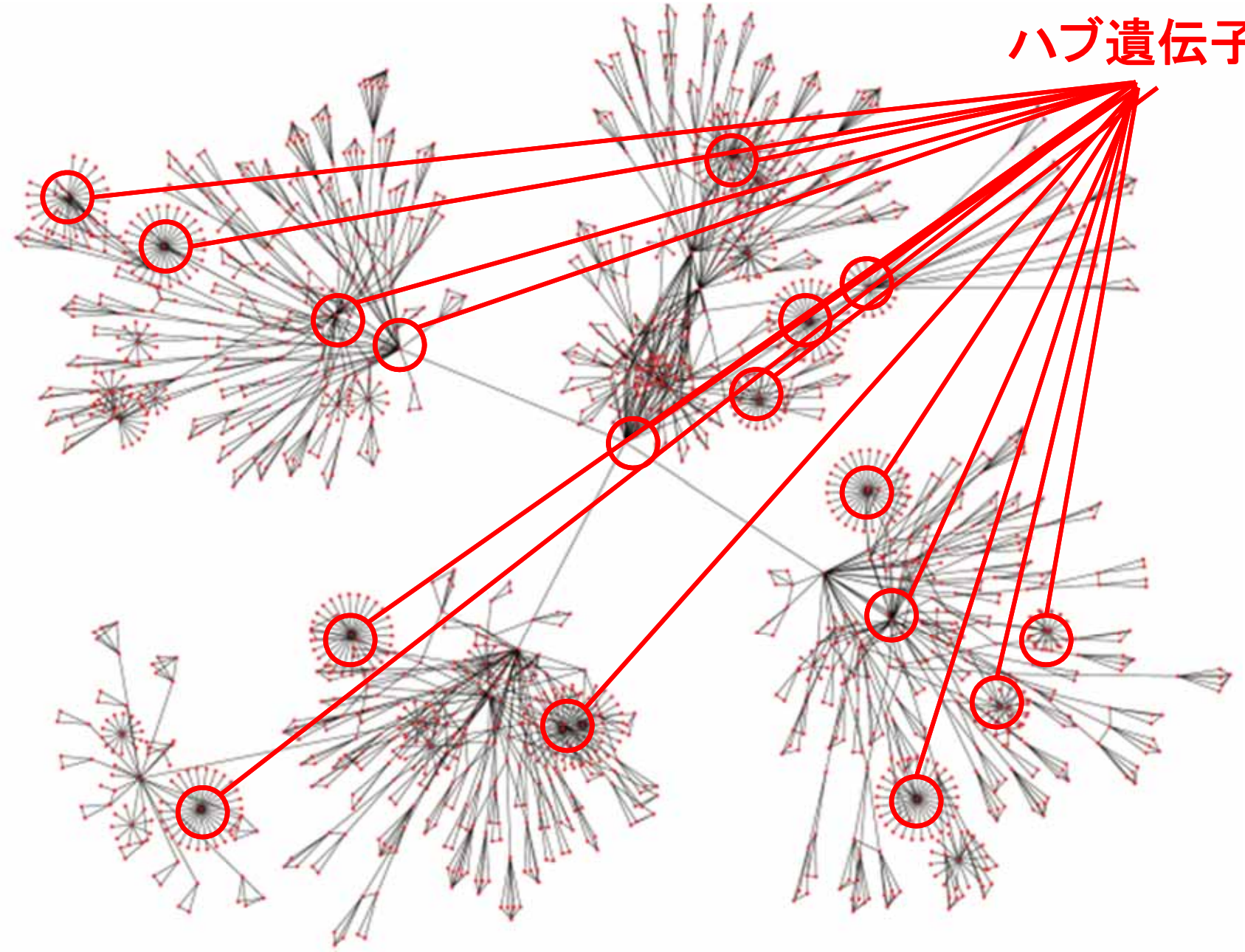
EMT に関わる 160 NSCLC 細胞株の遺伝子ネットワーク

目的

エルロチニブ薬剤耐性・感受性の違いに関わる鍵分子を見つける

Hubness の差異による鍵分子の探索

多くの遺伝子を制御する/に制御される遺伝子



Hubness の差異による鍵分子の探索

遺伝子 *A* がどれくらいハブになっているか

$$\mathrm{OP}_{A\alpha} = \sum_{k \in P_{A\alpha}} \hat{\beta}_{Ak\alpha}, \quad P_{A\alpha} = \left\{ j; \hat{\beta}_{jA\alpha} > 0 \right\}$$

α番目のサンプルにおいて
遺伝子Aが正に制御する遺伝子の個数

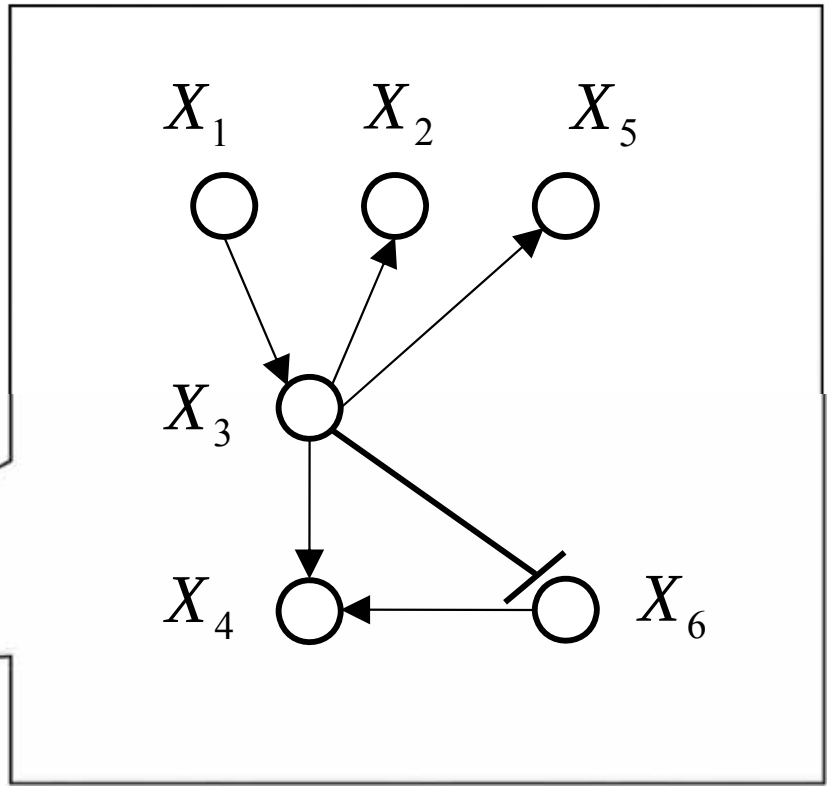
$$\mathrm{OI}_{A\alpha} = \sum_{k \in I_{A\alpha}} \hat{\beta}_{Ak\alpha}, \quad I_{A\alpha} = \left\{ j; \hat{\beta}_{jA\alpha} < 0 \right\}$$

α番目のサンプルにおいて
遺伝子Aが負に制御する遺伝子の個数

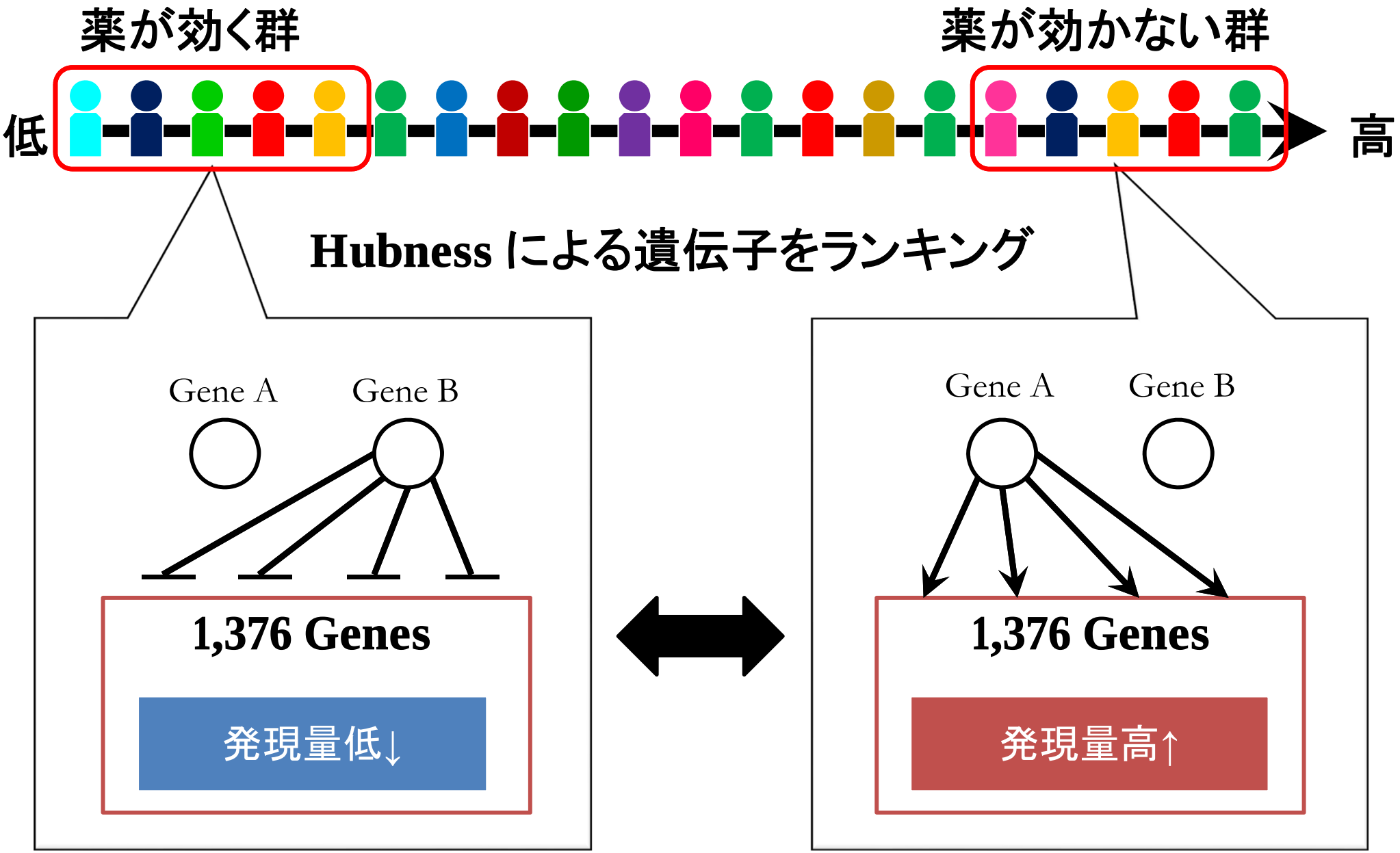
α番目のサンプル

$\mathrm{OP}_{3\alpha}=3$

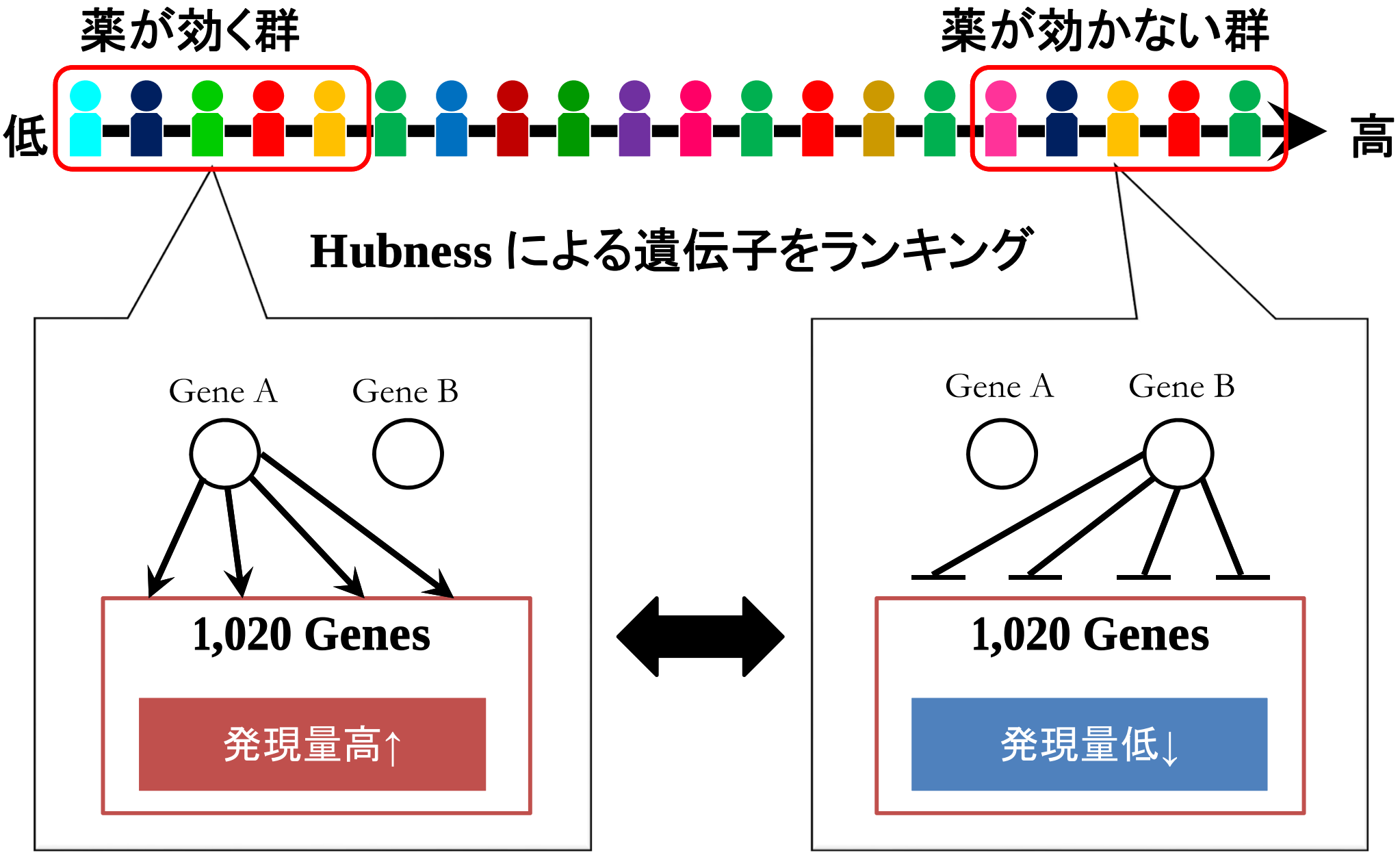
$\mathrm{OI}_{3\alpha}=1$



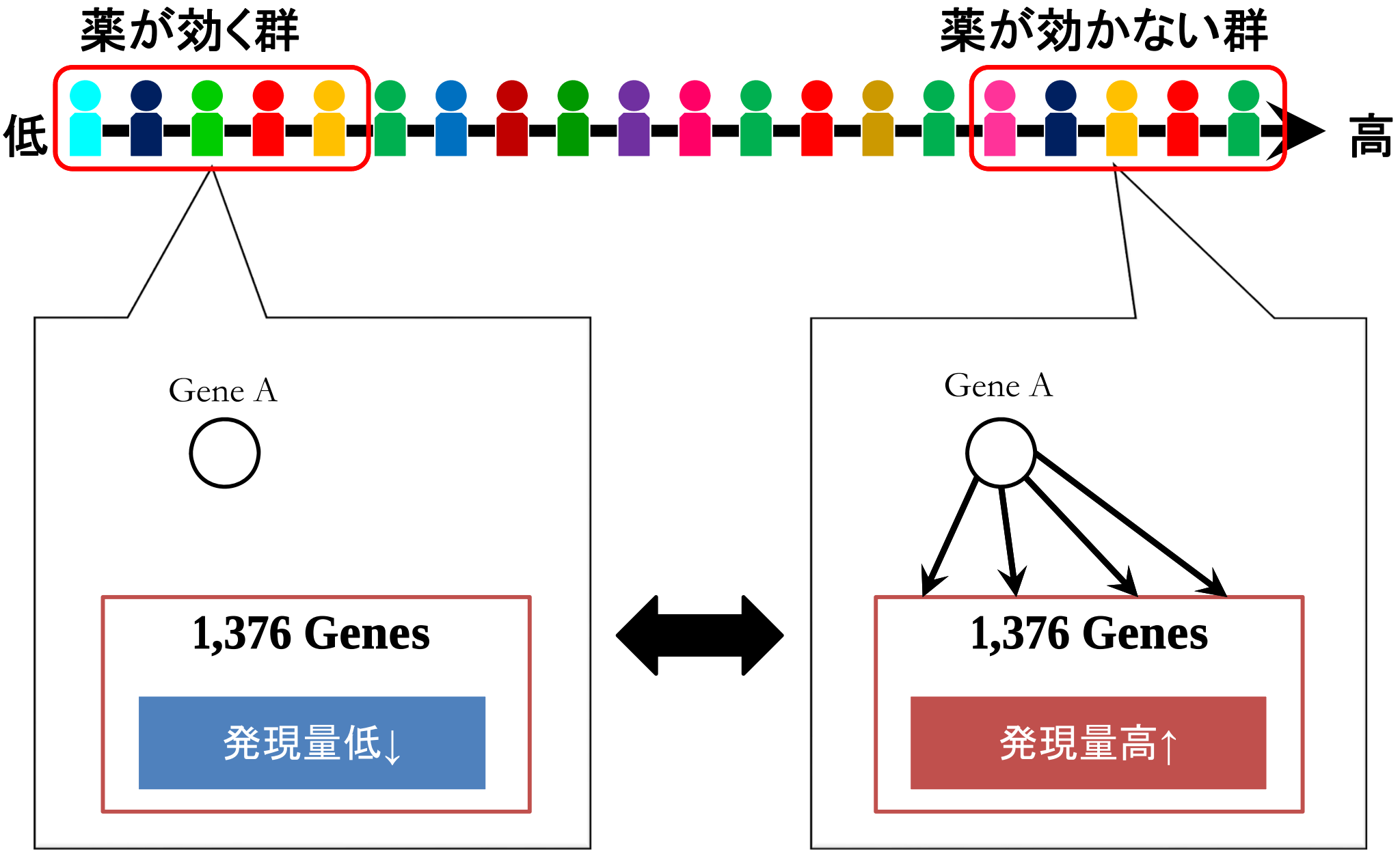
Hubness の差異による鍵分子の探索



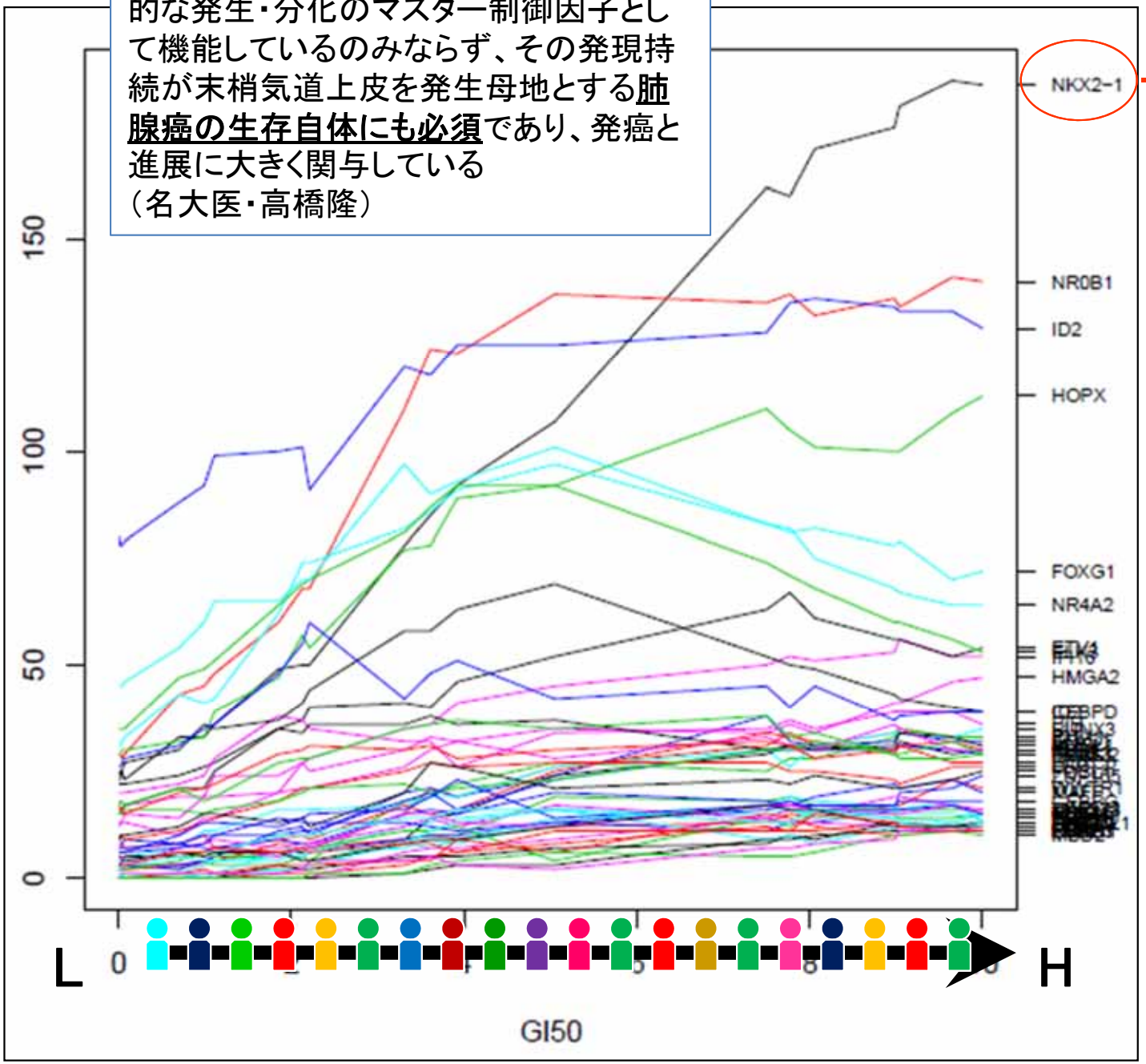
Hubness の差異による鍵分子の探索



結果1

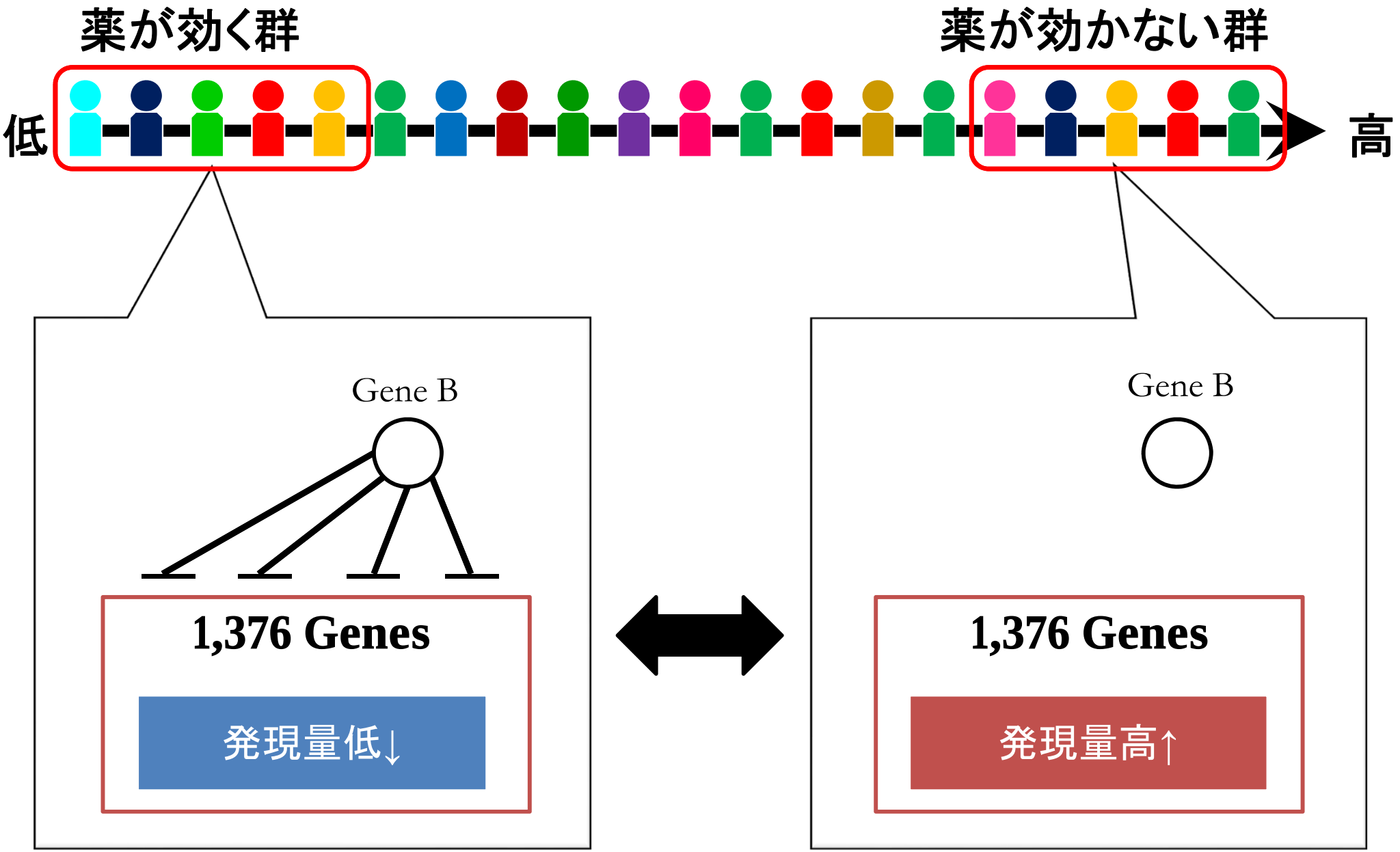

薬が効く群
薬が効かない群
低
高
Gene A
1,376 Genes
発現量低↓
Gene A
1,376 Genes
発現量高↑

TTF-1遺伝子は、末梢肺のリネッジ特異的な発生・分化のマスター制御因子として機能しているのみならず、その発現持続が末梢気道上皮を発生母地とする**肺腺癌の生存自体にも必須**であり、発癌と進展に大きく関与している
（名大医・高橋隆）

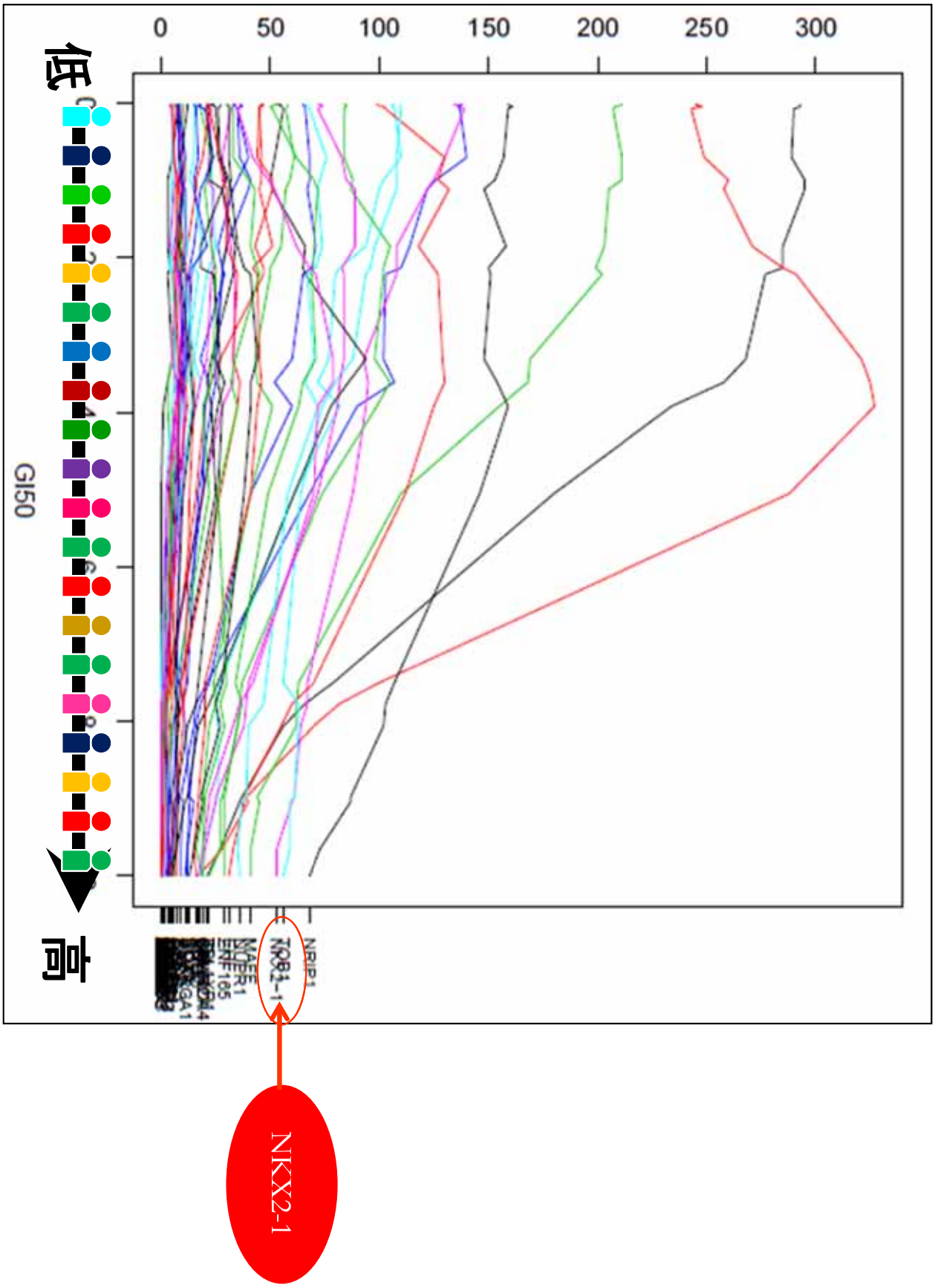


NKX2-1

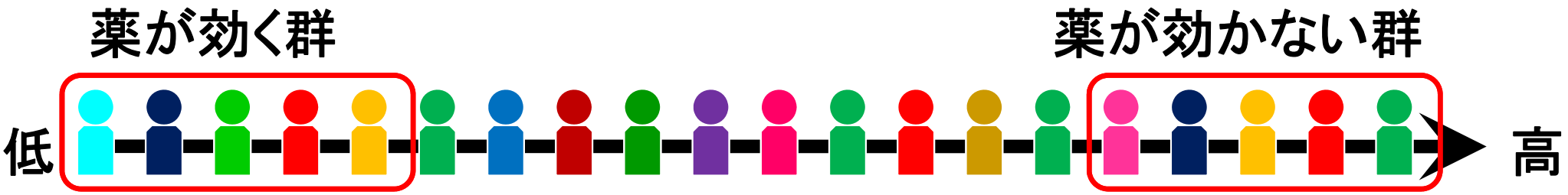
結果2


薬が効く群
薬が効かない群
低
高
Gene B
1,376 Genes
発現量低↓
Gene B
1,376 Genes
発現量高↑

負に制御される遺伝子数



機能比較



上流制御因子がもつ機能

Cell Cycle
DNA Replication, Recombination, and Repair
Connective Tissue Development and Function

上流制御因子がもつ機能

Cellular Movement
Tissue Development
Embryonic Development

“因果応報”

Bayesian Network+予後解析

- たくさんのサンプル（たくさんの人の活動スナップショット）から1つのネットワークを作る

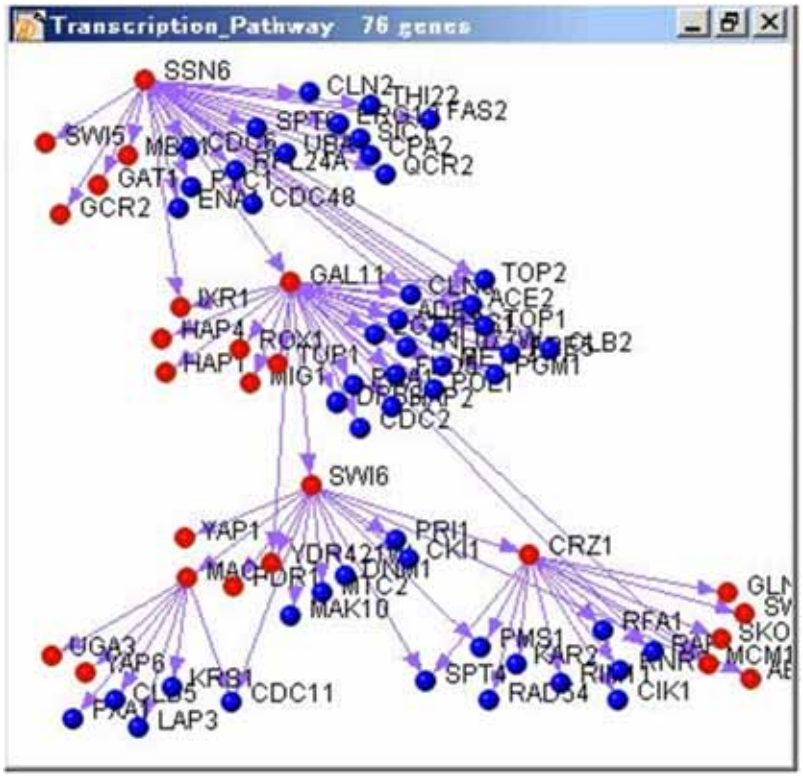
  「遺伝子間の因果の関係図」

- 予後データ（それぞれの人生の結果）から因果の図を解析すると

  「応報」の原因がみえてくる。

Bayesian Network + Nonparametric Regression

Gene Knockdown/Knockout
Time-Course Measurement

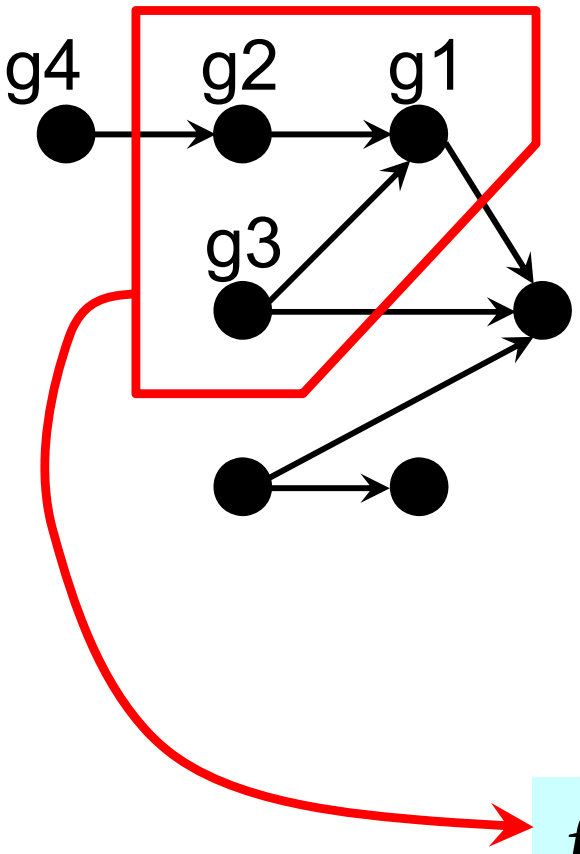


遺伝子ネットワーク

1. Imoto, S., Goto, T., Miyano, S. Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression. Pacific Symposium on Biocomputing. 7:175-186, 2002.
2. Imoto, Kim, Goto, Aburatani, Tashiro, Kuhara, Miyano (2003). Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic networkJ. *Bioinformatics and Comp. Biol.*, 1(2), 231-252

“仕事場における人間関係のようなもの”

Bayesian Network



- Markov assumption

  「直上の言うことしか耳を傾けない」

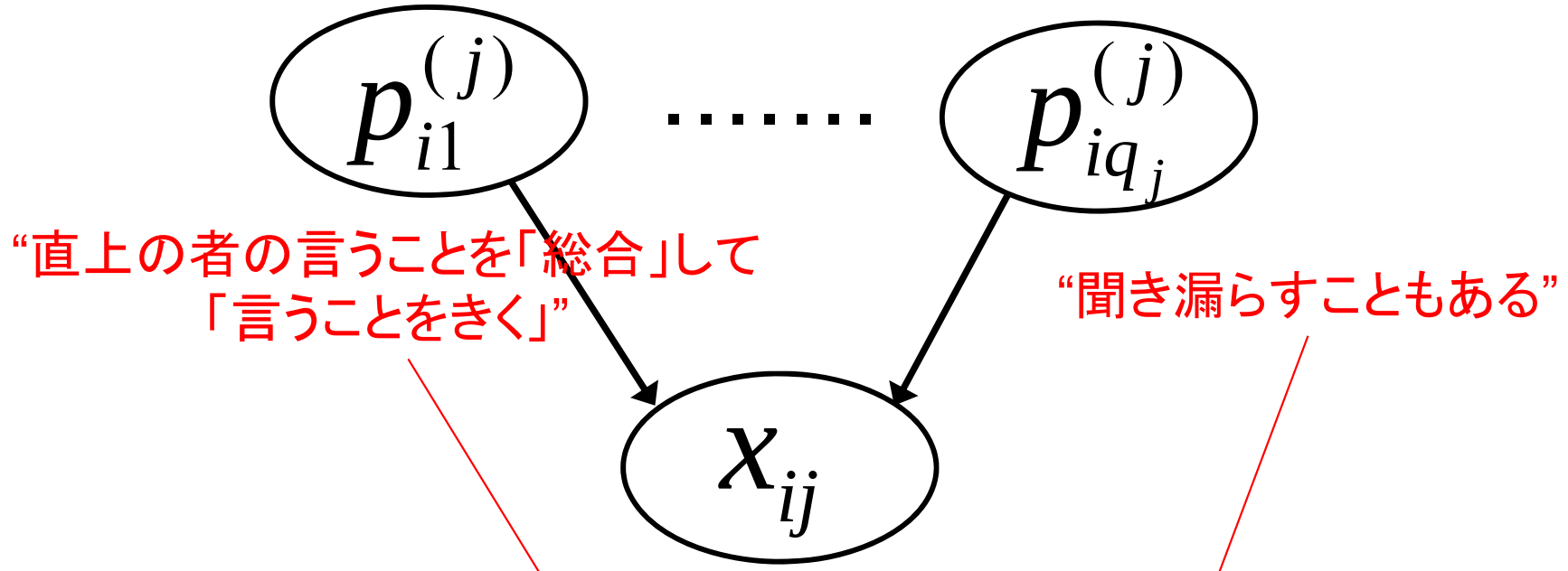
- ネットワーク構造がわかっていれば、
  データがたくさん観測されると、

  「言うことをきく」確率がわかってくる。

$$x_{i1} \quad \Leftarrow \quad \mathbf{p}_{i1} = (x_{i2}, x_{i3})^T$$

$$f(x_{i1},\ldots,x_{ip} \mid \boldsymbol{\theta}_G) = \prod_{j=1}^{p} f_j(x_{ij} \mid \mathbf{p}_{ij}, \boldsymbol{\theta}_j)$$

Nonparametric Regression



We consider the additive regression model:

$$x_{ij} = m_1(p_{i1}^{(j)}) + \cdots + m_{q_j}(p_{iq_j}^{(j)}) + \varepsilon_j ,$$

$$where \quad \varepsilon_j \sim N(0, \sigma_j^2) \; and \; \mathbf{p}_{ij} = (p_{i1}^{(j)}, \ldots, p_{iq_j}^{(j)}).$$

Here $m_k(\cdot)$ is a smooth function from R to R.

“数学的にはこんな感じになります”

Nonlinear Bayesian network model

$$f(x_{i1},\ldots,x_{ip};\boldsymbol{\theta}_G)=\prod_{j=1}^{p} f_j(x_{ij}\mid \mathbf{p}_{ij};\boldsymbol{\theta}_j),$$

$$f_j(x_{ij}\mid \mathbf{p}_{ij};\boldsymbol{\theta}_j)=\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left\{-\frac{(x_{ij}-\mu_{ij})^2}{2\sigma_j^2}\right\}$$

$$\mu_{ij}=m_1(p_{i1}^{(j)})+\mathrm{L}+m_{q_j}(p_{iq_j}^{(j)})$$

$$=\sum_{k=1}^{q_j}\sum_{m=1}^{M_{jk}}\gamma_{mk}b_{mk}^{(j)}(p_{ik}^{(j)})$$

たくさんのパラメータ

“難しいのはみんなの仕事ぶりデータからネットワーク構造を推定すること”

Criterion for selecting good networks

BNRC Score

Bayesian Network and Nonparametric Regression Criterion

$$\mathrm{BNRC}(G) = -2\log\pi_G \int \prod_{i=1}^{n} f(\mathbf{x}_i;\boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G \mid \boldsymbol{\lambda})d\boldsymbol{\theta}_G$$

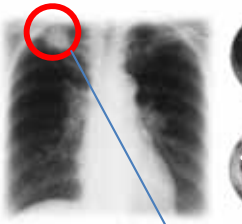
$$= -2\log\pi_G - r\log(2\pi n^{-1})$$

$$+\log\left|J_\lambda(\hat{\boldsymbol{\theta}}_G)\right| - 2nl_\lambda(\hat{\boldsymbol{\theta}}_G \mid \mathbf{X}_n)$$

この”BNRC score”が小さいネットワーク構造とパラメータを探索する。“スパコンがここで必要になる”

肺腺がんの予後の良・不良のスイッチの同定
にシステム的方法論の有効性が示された

マイクロRNAとmRNAのチップ解析データ

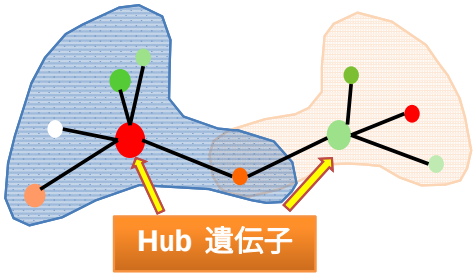


非線形回帰ベイジアンネットワーク

SiGN-BN

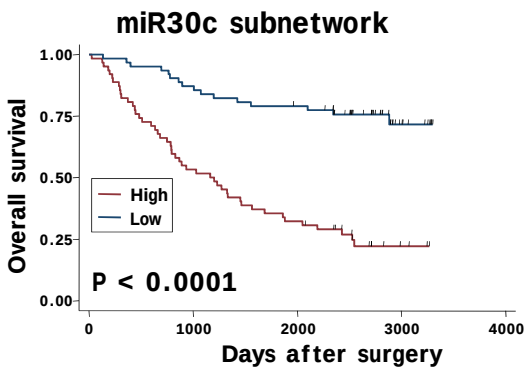
スパコン

肺がんのマイクロRNA遺伝子ネットワーク



予後データ

ハブ機能の実験検証



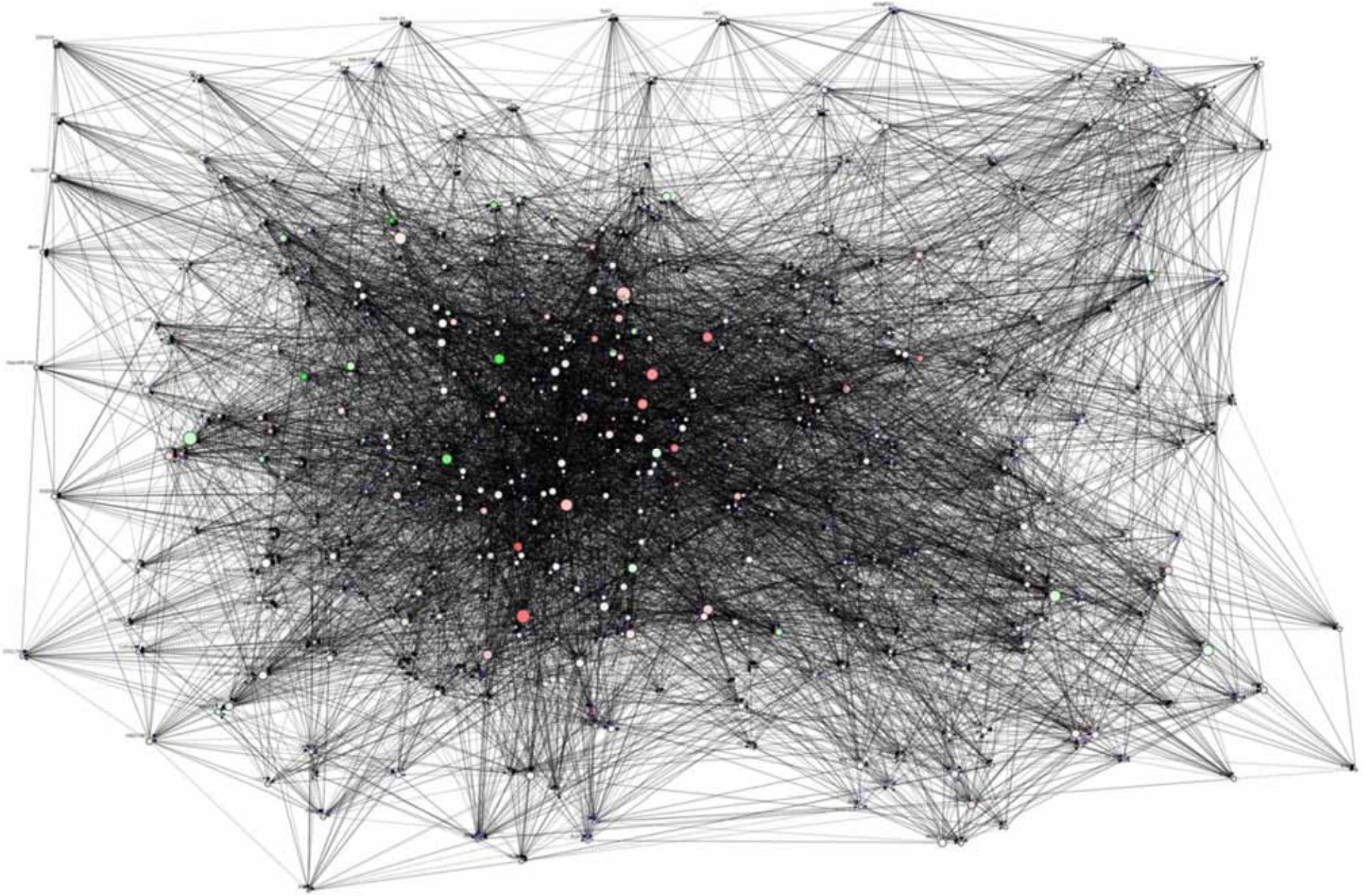
予後予測力プラン・マイヤーダイヤグラムのp値が小さくなるサブネットワークを網羅的に探索

再発・死亡と優位に関連する14個のサブネットワークと、そのハブ遺伝子が見つかる

Cancer Biology

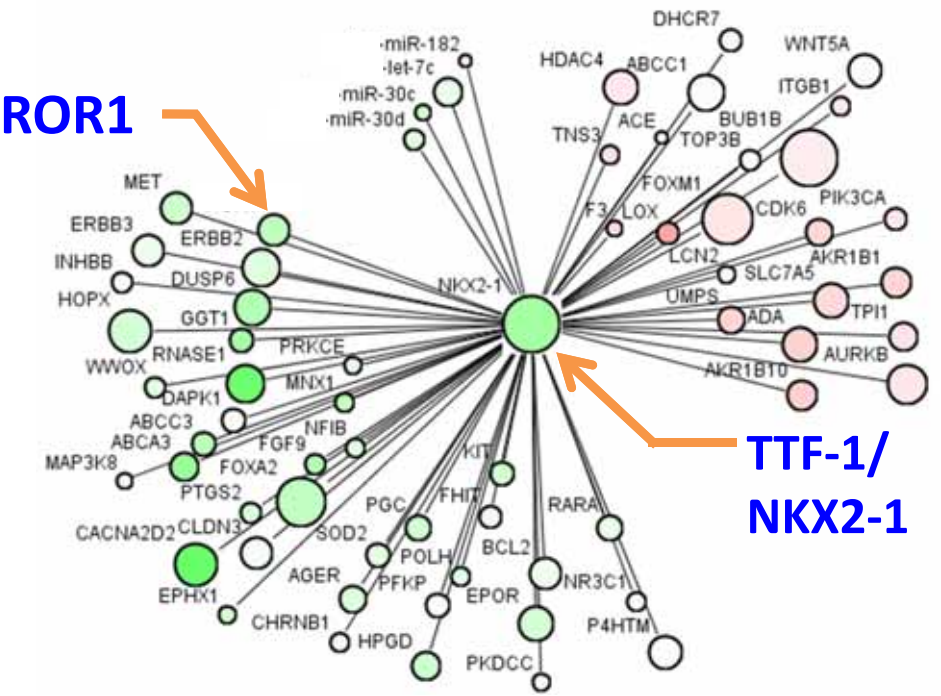
Computatioanl Systems Biology

これが肺腺がんの「遺伝子因果の図」といっても、訳がわからない



実験で確かめてみるとバッチシ！

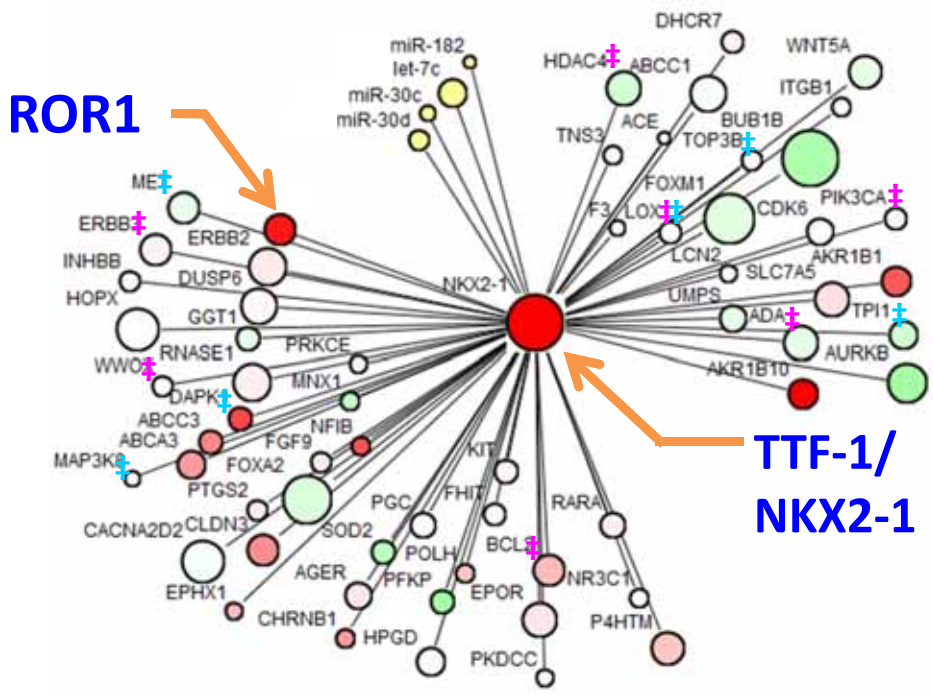
Inferred subnetwork of TTF-1/NKX2-1



Expression ratio

- High in poor prognosis
- Low in poor prognosis

Experimental validation by TTF-1/NKX2-1 introduction

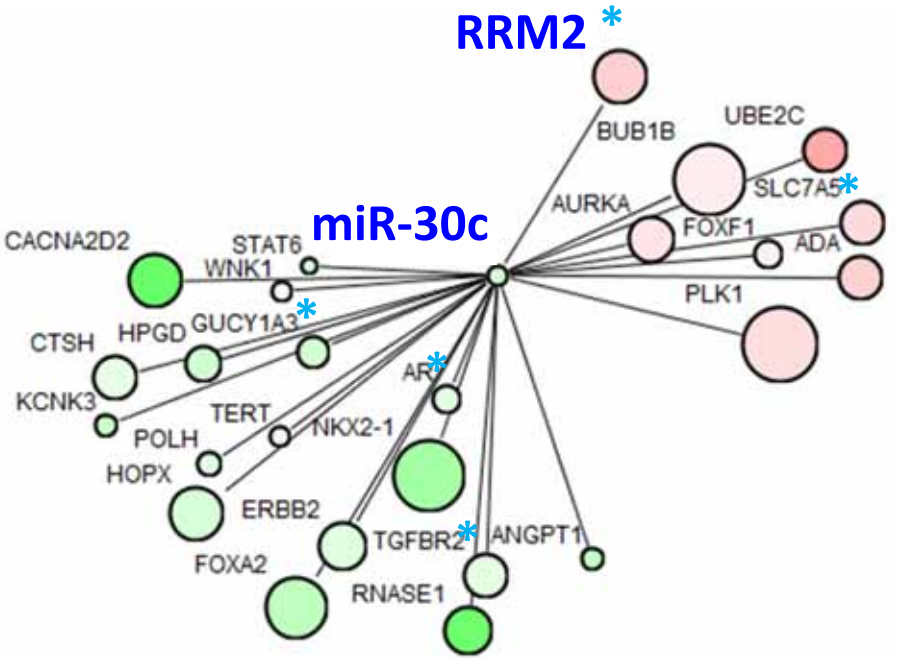


Expression ratio (TTF-1-intorudced / Vector control)

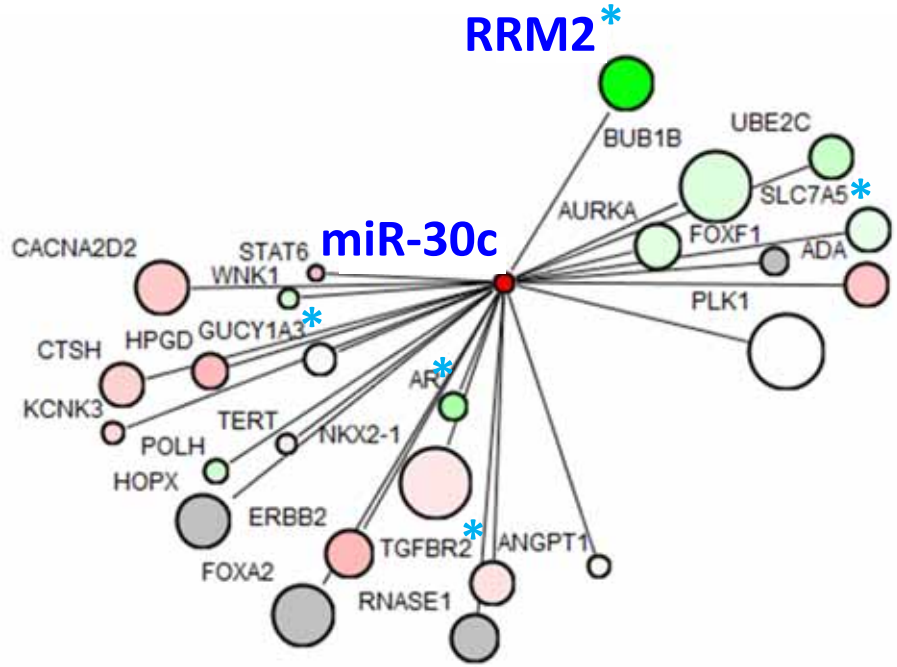
- **High in TTF-1/NKX2-1-introduced**
- **Low in TTF-1/NKX2-1-introduced**
- **: N.A. (miRNA)**
- **: positive binding in E19.5 mouse lung ChIP-chip assay data (Tagne et al., PLoS ONE, 2012)**

新たなマイクロRNAも見つかった miR-30cとその標的遺伝子RRM2

肺癌検体



miR-30cを導入



Ratio(予後不良 / 良好) (log2)

2.4 (予後不良で高発現)

-2.4 (予後不良で低発現)

* : Poorly conserved predicted target (TargetScan)

Ratio(miR-30c導入 / コントロール) (log2)

2.4 (miR-30c導入で高発現)

-2.4 (miR-30c導入で低発現)

○ : not detectable

The Cancer Network Galaxy

がん遺伝子ネットワークデータベース

http://tcng.hgc.jp/ (open Jan 9, 2013)

- 京コンピュータを259CPU年使って、256の様々ながんの遺伝子ネットワークを推定し、データベース化した
- 各遺伝子ネットワークは8000個の遺伝子からなる大規模なもの
- 世界初の大規模がん遺伝子ネットワークデータベース

Where is “My Cancer” in this galaxy?

“The Cancer Gene Network Galaxy” のどこにKLF5が現れるかを探索してみると・・・

- KLF5 activates Perp (TP53 apoptosis effector) in Glioma and Breast Cancer.
- Until now, it is only known that direct targets of KLF5 contribute to the maintenance of embryonic stem cell undifferentiated state. Especially, Perp is one of KLF5 target. (Tommaso Russo’s Group. BMC Biology, 2010)

KLF5 might contribute to the maintenance of cancer stem cell undifferentiated state in Glioma and Breast Cancer

Acknowlegdments

東京大学医科学研究所
ヒトゲノム解析センター
玉田嘉紀（現・情報理工学系研究科）
島村徹平
新井田厚司
山口類
井元清哉

名古屋大学大学院医学系研究科
分子腫瘍学分野　高橋　隆
冨田秀太（現・UCLA)
有馬千夏
竹内俊幸（現・オンコミクス）
島田友香子　　梶野泰祐
山口知也　　細野祥之
鈴木元　　柳澤聖

愛知県がんセンター
研究所・分子腫瘍学部　長田啓隆
病院・遺伝子病理診断部　谷田部恭
同・胸部外科　光冨徹哉（現・近大）

国立がん研究センター研究所
河野隆志　他
横田　淳　他