『戦略分野1「予測する生命科学：医療および創薬基盤」』

2014-03-14

SiGN講習会資料：ベイジアンネットワークを用いた遺伝子ネットワークの推定と解析

土井 淳
atsushi_doi@cell-innovator.com

株式会社セルイノベーター
研究開発部

福岡市東区箱崎6-10-1
九州大学 産学連携棟I アントレプレナーシップ・センター 2階
http://www.cell-innovator.com

cell innovator

1. マネーボール：統計学の応用

2. 遺伝子発現とベイジアンネットワーク

3. 遺伝子ネットワーク

1. マネーボール：統計学の応用

cell innovator

近年の統計学にまつわるトピック

- マネーボール理論：経営論の参考にも。日経BP -- http://special.nikkeibp.co.jp/ts/article/aaaa/114314/

- ビッグデータ：Google、Facebook、Amazon などの企業によるイメージ。

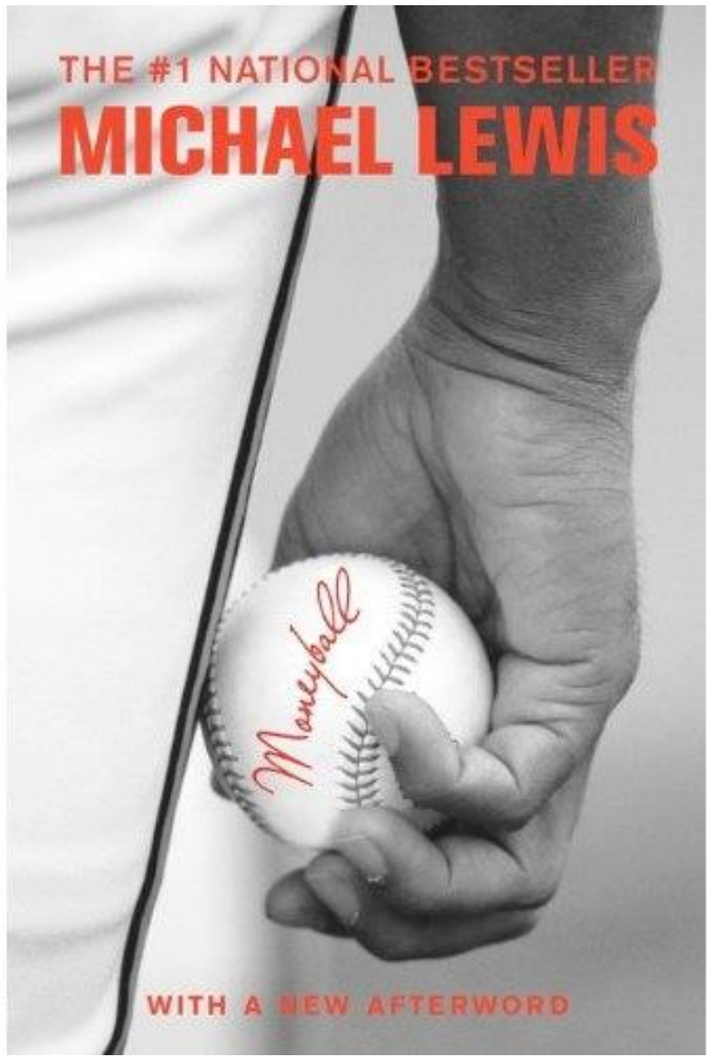
- データアナリスト、データサイエンティストが25万人不足。 http://www.nikkei.com/article/DGXNZO57421630X10C13A7EA1000/

「大量のデータを統計学を使って、なんとかしよう」というのがトレンド

マネーボール理論とは？

- 野球をアウトを取られないようにするゲームと定義。過去のデータをもとに導きだされた理論。

- バントをするな。
- フォアボールでいい。
- 初球に手を出すな。
- 盗塁もダメ。
- バントされても、2塁に投げるな。



安い選手で効率よく勝つための理論

cell innovator

ここまでの話で、、、

- 近年、統計学的なアプローチが、よく用いられるようになった。

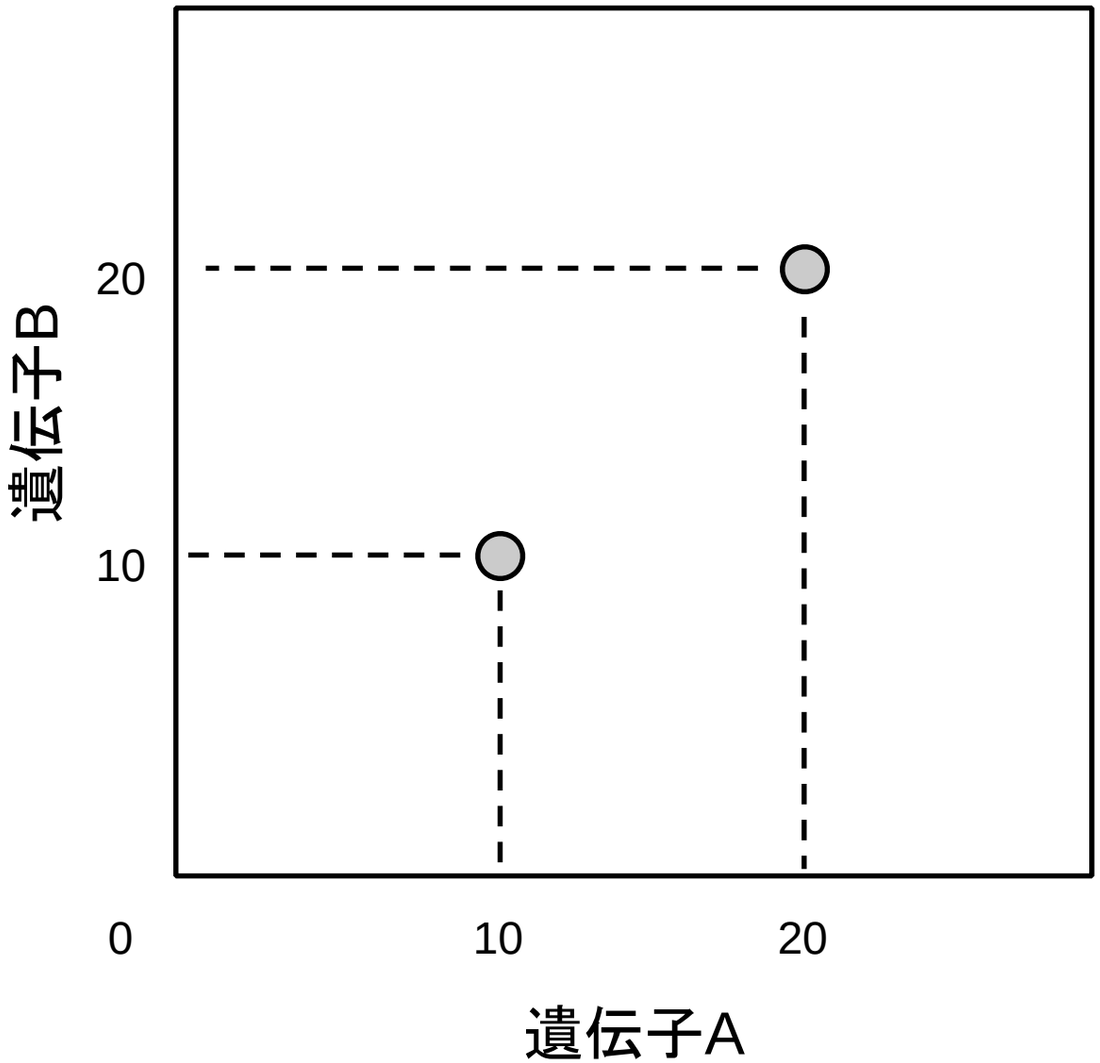
- 統計学的なアプローチから得られたものが、必ずしも人間の直感に合わない。（裏、裏、裏と来たら、次は表と思いたいのが心情。）

- 直感に合わなくても、役に立つかもしれない。（マネーボール理論のアスレチックスは、シーズン中に20連勝。レッドソックスは、ワールドシリーズ優勝。）

2. 遺伝子発現とベイジアンネットワーク

cell innovator

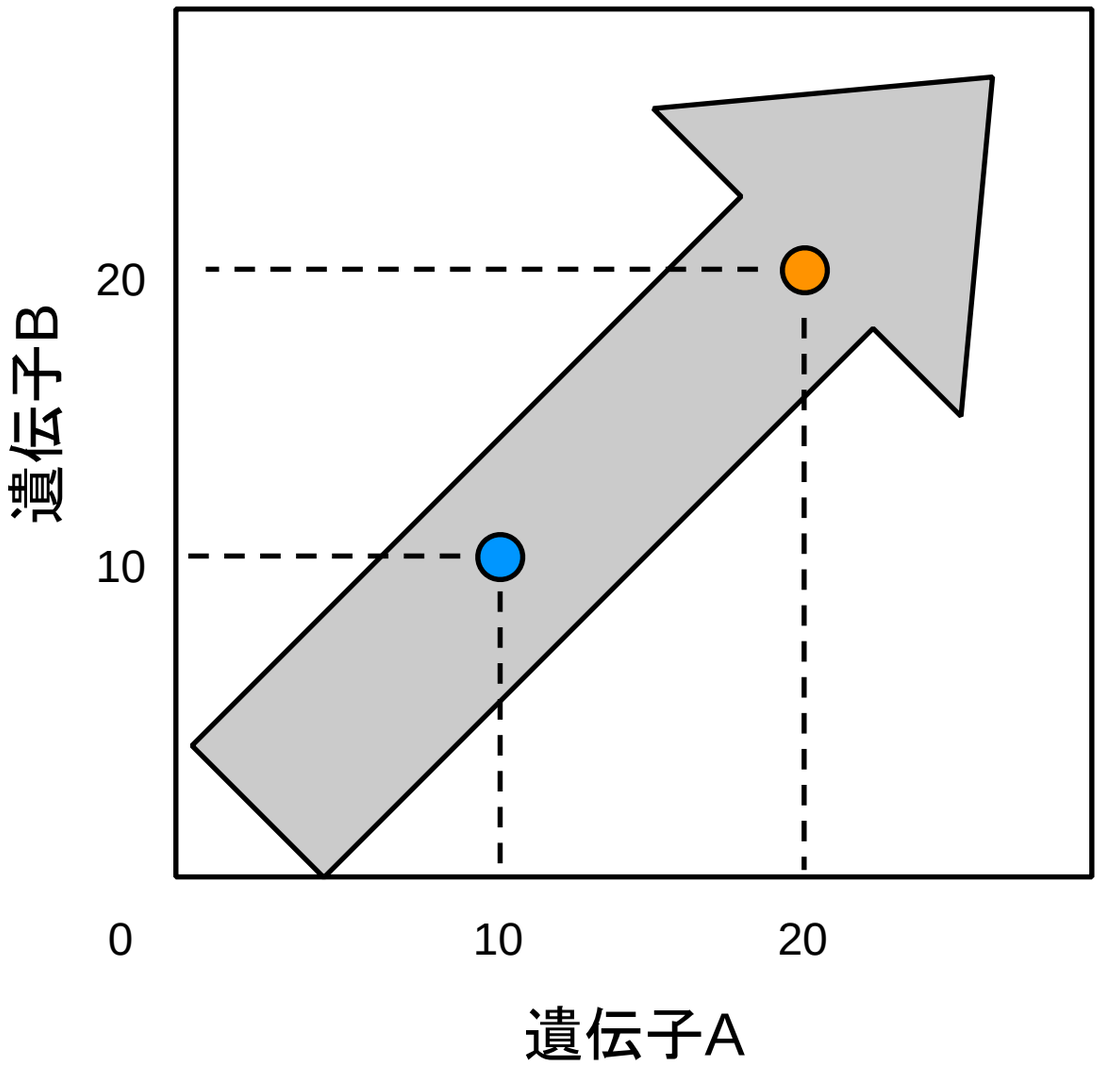
遺伝子発現と散布図



- 遺伝子Aの発現量が、10のとき、
- 遺伝子Bの発現量が、10なら、
- 散布図に表すと、(x, y) = (10, 10)

- 同様に遺伝子Aの発現量が、20のとき、遺伝子Bの発現量が、20なら、(x, y) = (20, 20)

遺伝子の相関関係 (1)

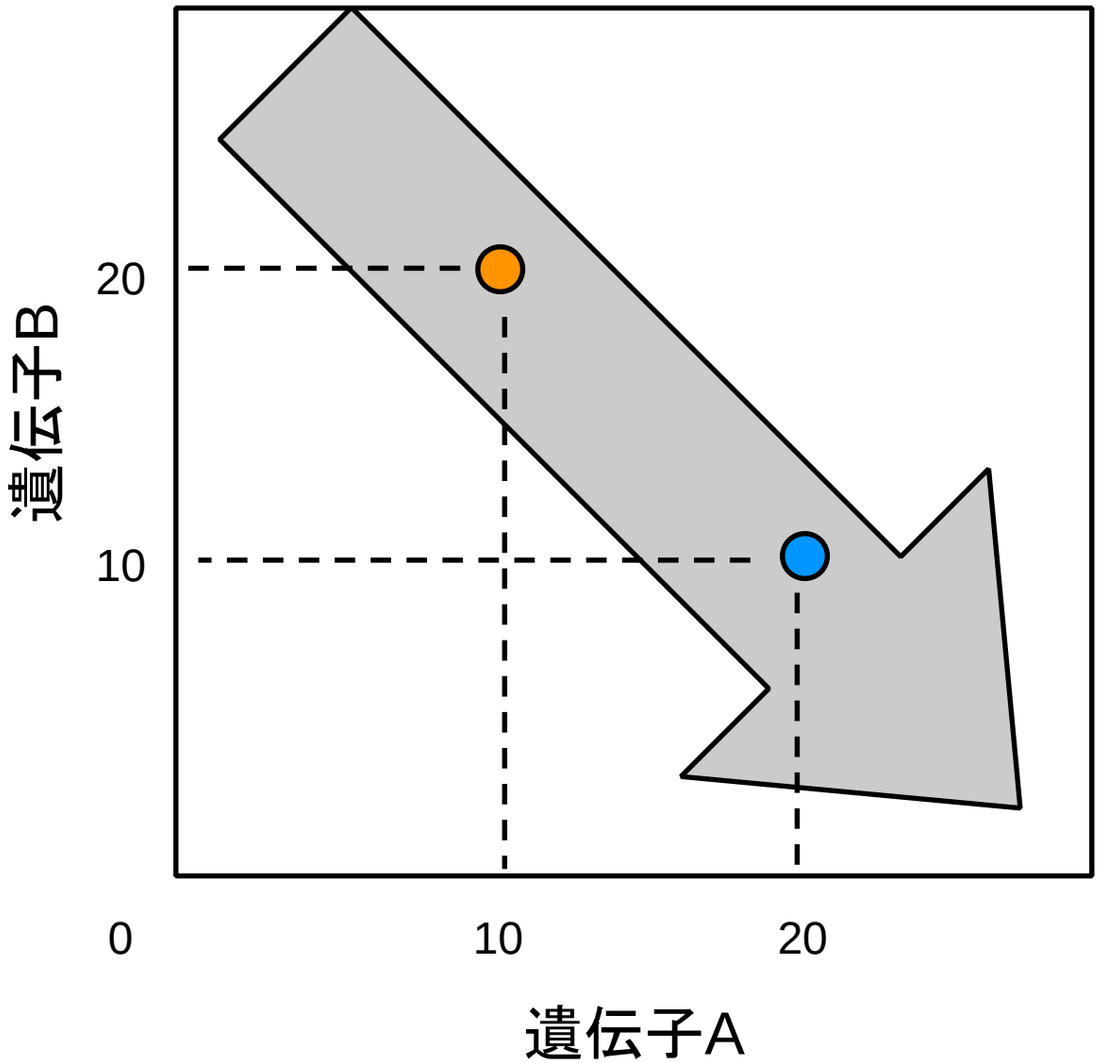


- つまり、遺伝子Aの発現量が低いとき、遺伝子Bの発現量も低い。
- また、遺伝子Aの発現量が高いとき、遺伝子Bの発現量も高い。

- 遺伝子AとBの発現量には、正の相関が見られる。

A ⟶ B

遺伝子の相関関係 (2)

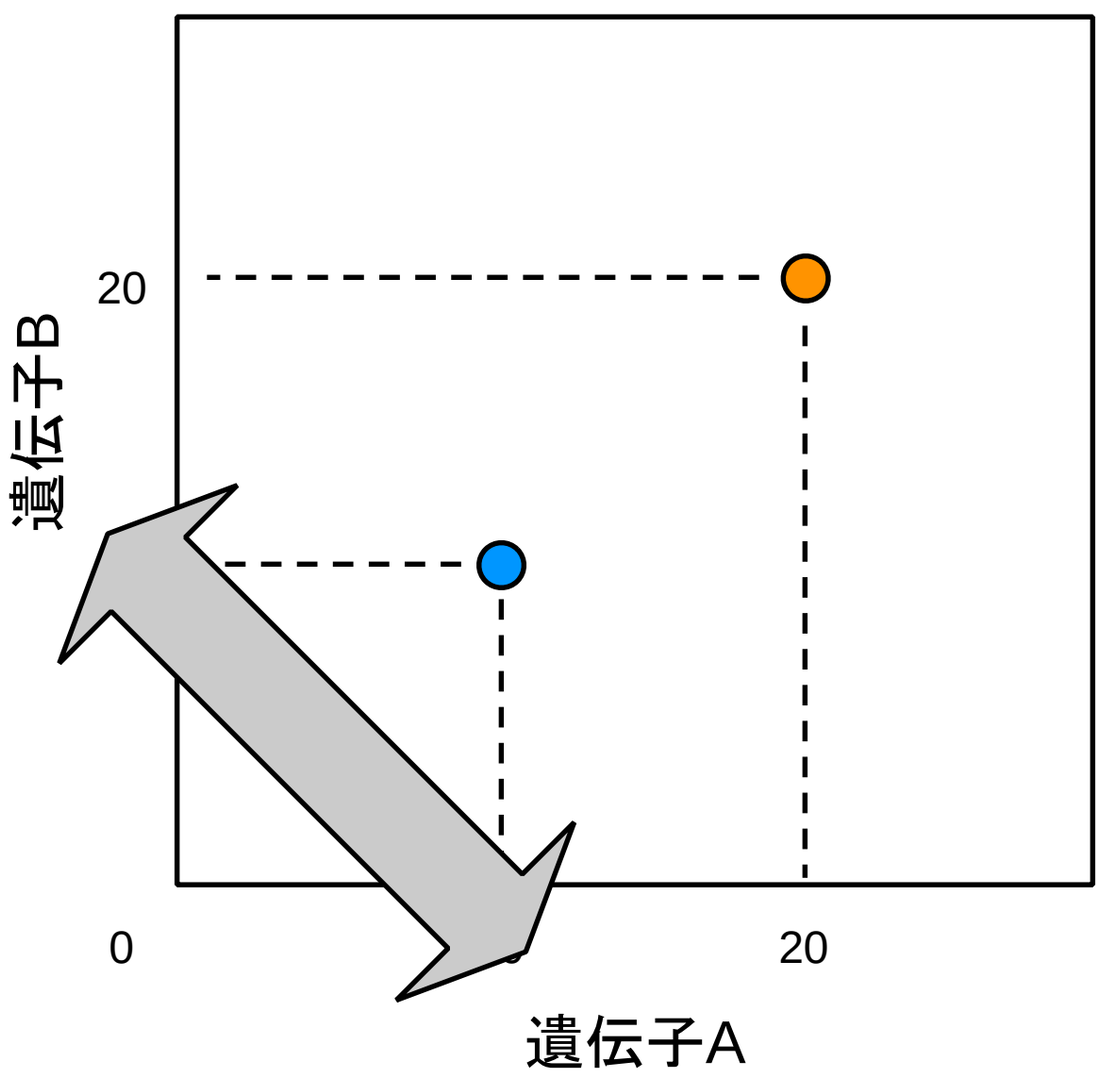


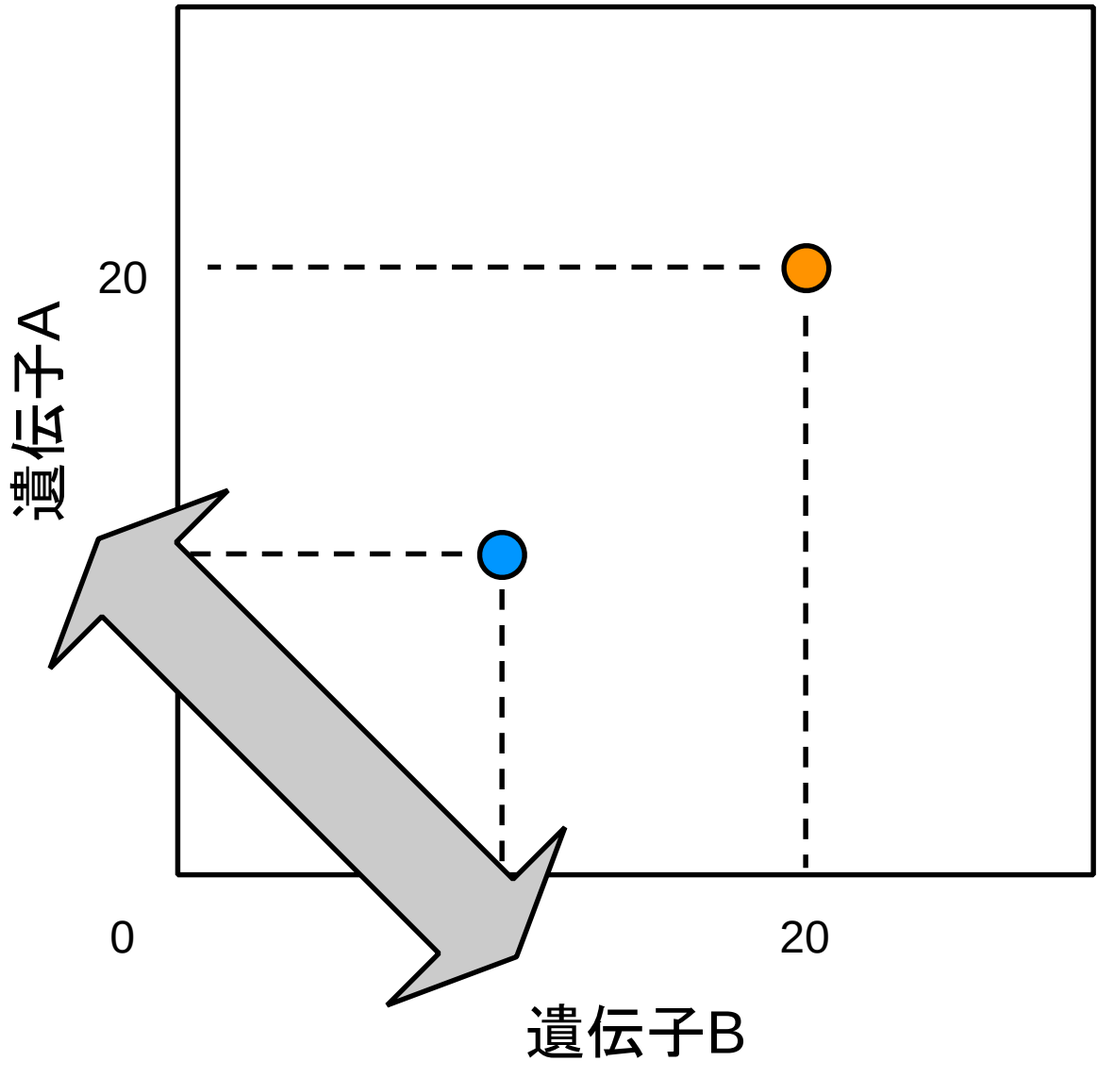
- その逆なら、遺伝子Aの発現量が低いとき、遺伝子Bの発現量は高い。
- また、遺伝子Aの発現量が高いとき、遺伝子Bの発現量は低い。

- 遺伝子AとBの発現量には、負の相関が見られる。

A ⟶ B

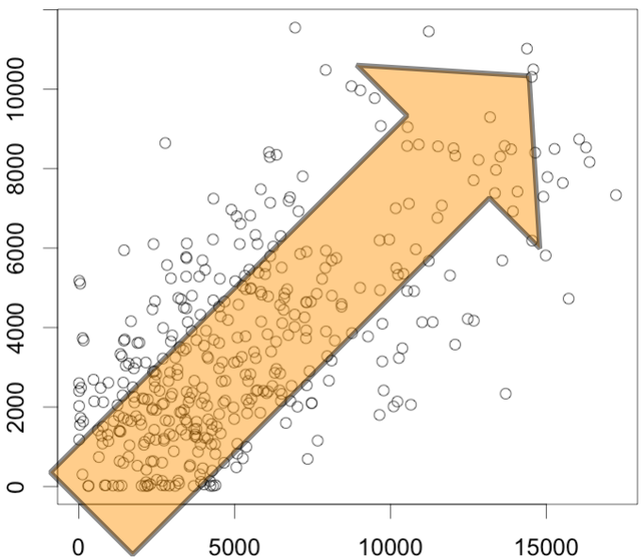
どちらが上流？

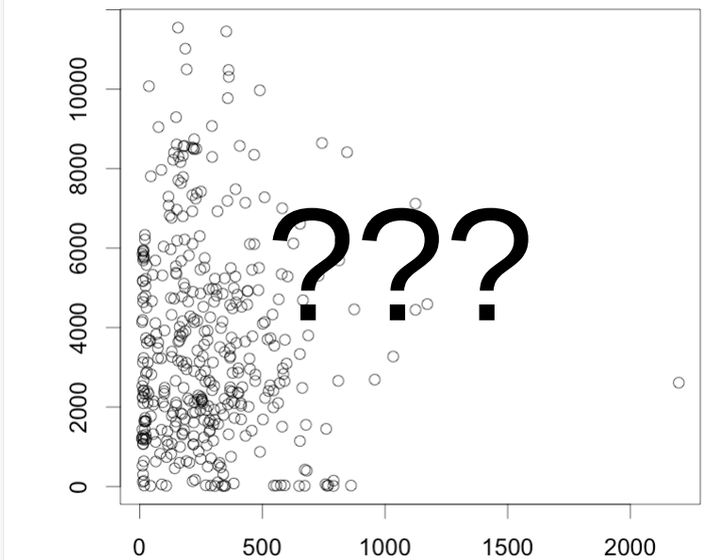


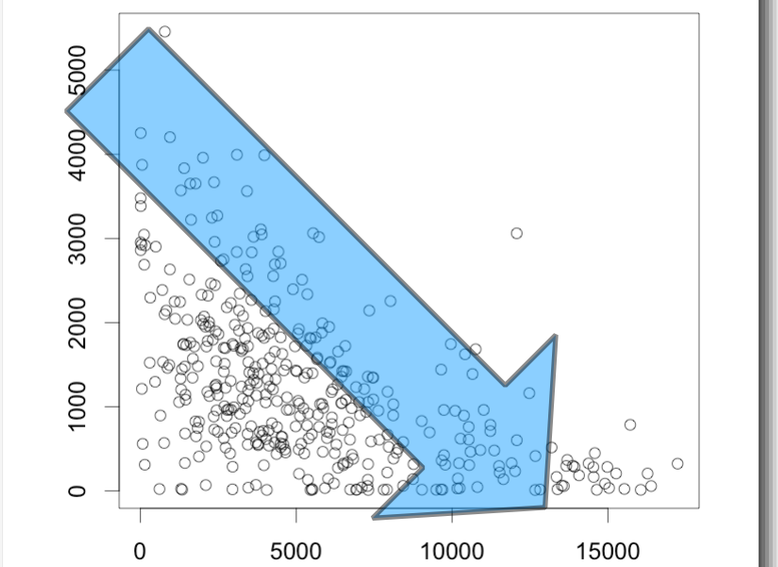


- X軸とY軸を入れ替えても同じなので、どちらが上流か分からない？？

データを増やしていくと見えてくるもの







- 上記は、400サンプル＝400個の点における関係を見たもの。
- サンプル数を増やしていくと、「関係の度合い」（＝確率）も見えそう。

遺伝子発現にも統計学的なアプローチを。

cell innovator

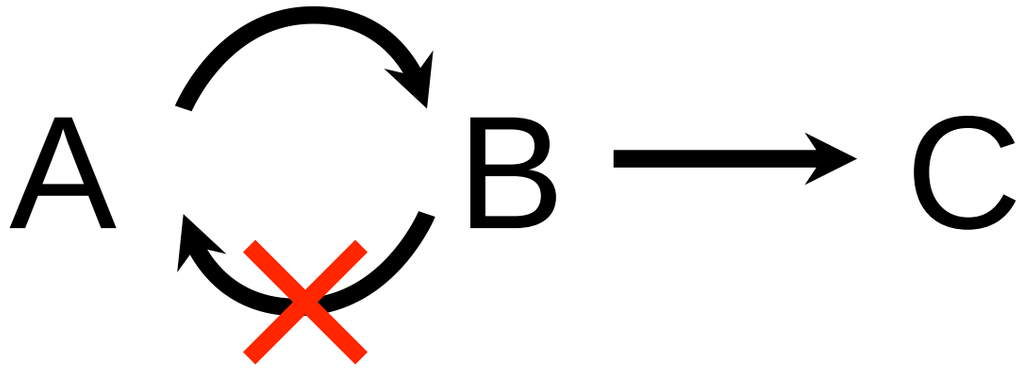
ベイジアンネットワーク（モデル）

- 遺伝子Aが、ある確率で、遺伝子Bを制御していて、
- 遺伝子Bが、ある確率で、遺伝子Cを制御している。

条件付き確率で表されたネットワークが書ける。

ベイジアンネットワーク（モデル）

- ベイジアンネットワーク＝条件付き確率で表されたネットワークのうち、ループ構造がないもの。



ループは、なし

A ⟶ B ⟶ C

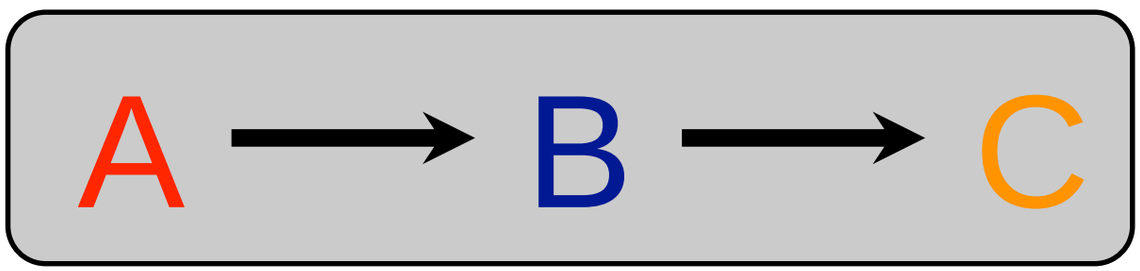
どちらか？

A ⟵ B ⟶ C

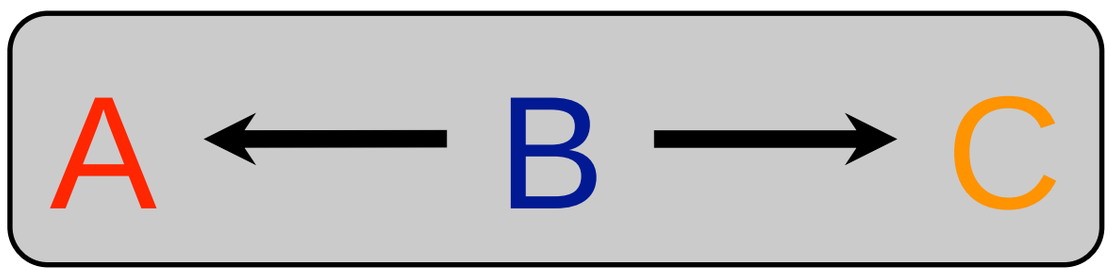
ベイジアンネットワーク（モデル）

- Aが起こってから、Bが起こり、Cになるのか？
- Bが起こってから、AとCが起こるのか？

- 言い換えると、Aが原因なのか、Bが原因なのか？
- どちらのモデルか分かれば、どちらが原因か分かる。（因果推定）

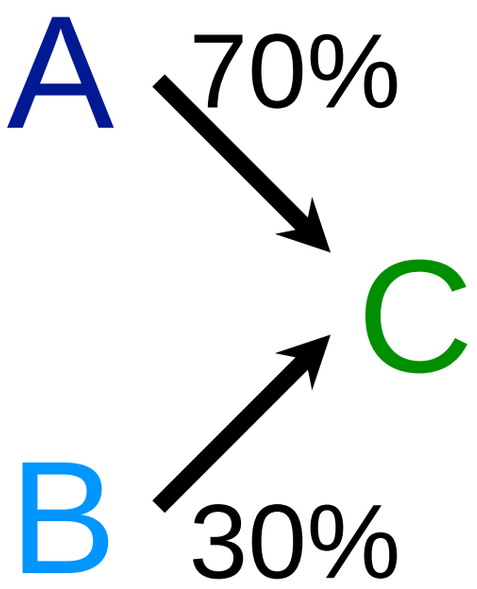


原因はどちら？



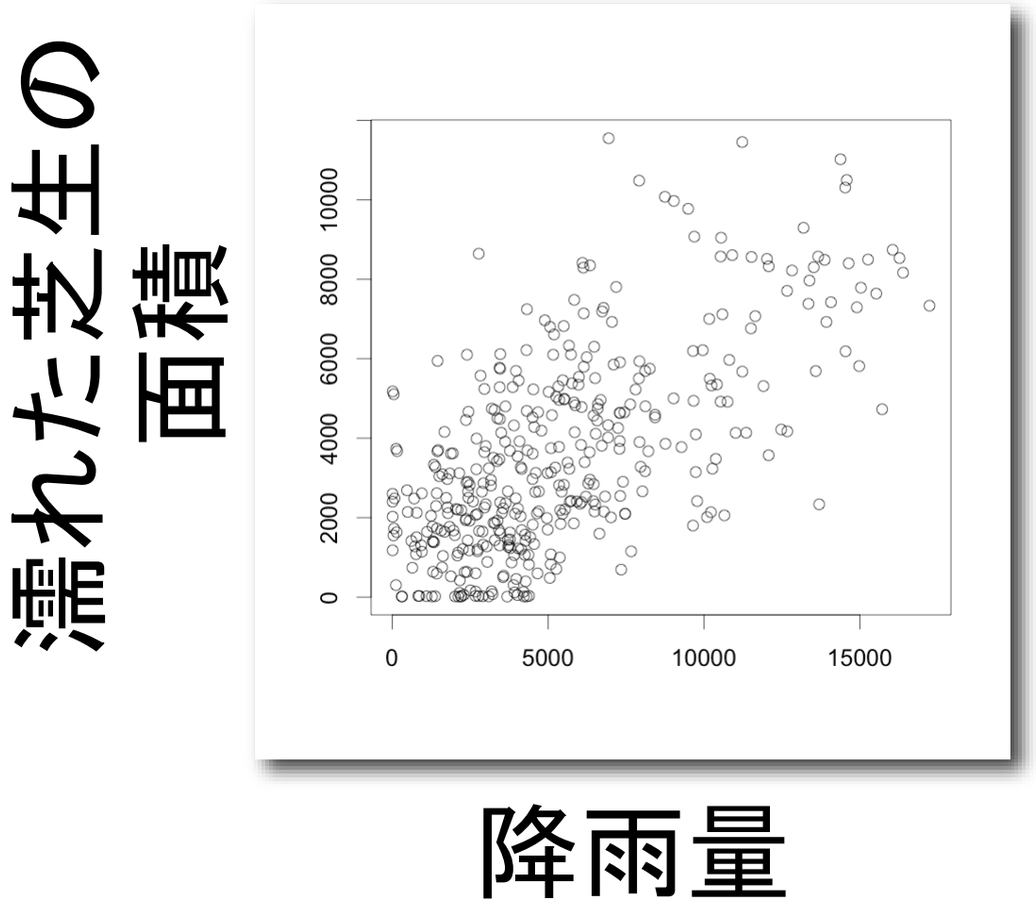
例えば、雨とスプリンクラーと芝生の関係は？

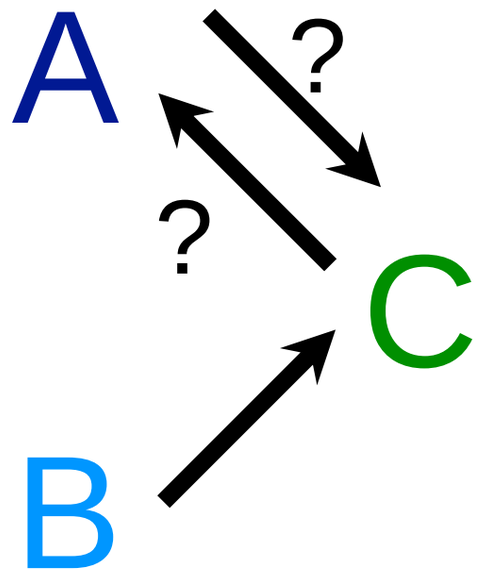
- A: 雨が降る（降雨量）。
- B: スプリンクラーが作動する。
- C: 芝生が濡れる。



- 芝生が濡れるのは、雨が降ったか、**または、**スプリンクラーが作動したから。

芝生が濡れたら、雨が降る？

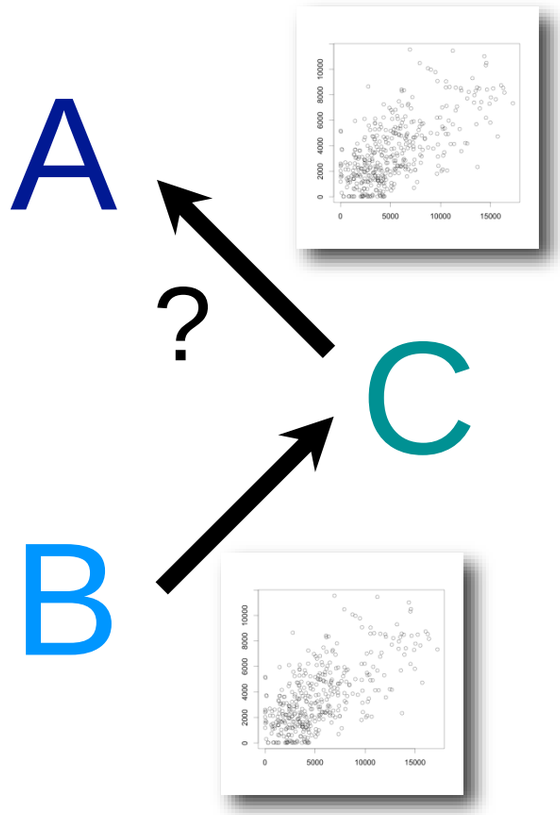




- 雨が降ったから、芝生が濡れたのか？ A --> C
- 芝生が濡れたから、雨が降ったのか？ C --> A

スプリンクラーの影響を考慮

- もし、芝生が濡れたから、雨が降ったのなら、B --> C --> A
- つまり、スプリンクラーが作動すると、雨に何らかの影響があることになる。



- これは調べれば分かる。スプリンクラーが作動しても、天気に影響はない。

すべてのパターンを調べれば、どちらのモデルが適切か分かる！

実際は、、、

- Bスプラインによるノンパラ回帰
- DAG 探索問題
- Greedy Hill Climbing アルゴリズム
- BNRC スコア、オーバーフィッティング
- 、、、、、（詳細は玉田さんの資料をご覧ください）

イメージ的には、とにかく総当たりで、
すべてのネットワークのパターンをチェックして、
もっともらしいネットワークの状態を推定

補論
簡単な数式で紹介する
ベイジアンネットワーク

この補論で取り扱うベイジアンネットワーク

ベイジアンネットワークとは

1. 大量の遺伝子発現解析データから求められる
   同時確率分布 $P(G_1,G_2,\ldots,G_N)$ を用いて、
2. 遺伝子 $G_1, G_2, \ldots, G_N$ 間の統計的依存関係を見出し,
3. 非循環型の有向グラフでネットワーク構造を表現する方法

注1） ここでは、個々の実験で得られるN個の遺伝子の発現量 $G_1, G_2, \ldots, G_N$ を確率変数とみなし,
細胞の状態をそれら確率変数の同時確率分布 $P(G_1, G_2, \ldots, G_N)$ で表現している。
注2） 非循環型とは、「フィードバックループなどの循環構造がネットワークの中には存在しない」という制限がこの方法にはあらかじめ課せられていることを示している。
ただし、この制限はダイナミック・ベイジアンネットワークでは取り除かれる。

条件付き確率と有向グラフ

ベイジアンネットワークでは

1. 遺伝子間（例：G_1 と G_2 ）の条件付確率 $P(G_1|G_2)$ がグラフに付与され，
2. 遺伝子 G_2 から遺伝子 G_1 への統計的依存関係を $G_2 \rightarrow G_1$ と方向性を持って，

すなわち因果関係として表現する。

$$P(G_1|G_2) \quad \Leftrightarrow \quad G_2 \rightarrow G_1$$

条件付き確率での子遺伝子 G と親遺伝子 $\pi(G)$

$$P(G_1|G_2) \quad \Leftrightarrow \quad G_2 \rightarrow G_1$$

1. ここで G_1 は子と呼ばれ，G_2 は G_1 の親 $\pi(G_1)$と呼ばれ、
   $\pi(G_1)=(G_2)$ とリスト表記される。
2. 遺伝子 G_1 の親遺伝子が複数，例えば G_2 と G_3 である場合、
   $\pi(G_1)$ は $\pi(G_1) = (G_2, G_3)$ とリスト表記される。
3. なお，遺伝子 G_1 に親がいない場合には便宜的に，
   $\pi(G_1) = (\emptyset)$ と空集合 ø でリスト表記される.

遺伝子ネットワークの表現

ベイジアンネットワークによる遺伝子ネットワークの表現は、同時確率分布 $P(G_1, G_2, \ldots, G_N)$ を条件付確率 $P(G_i|\pi(G_i))$ の積として展開することと同等となる。

$$P(G_1, G_2, \cdots, G_N) = \prod_i P(G_i \mid \pi(G_i))$$

仮想的な遺伝子発現の実験結果 *D*

ケース（実験）	G_1	G_2	G_3
1	2	1	1
2	2	2	2
3	1	1	2
4	2	2	2
5	1	1	1
6	1	2	2
7	2	2	2
8	1	1	1
9	2	2	2
10	1	1	1

注）1: 発現していない　2: 発現している

仮想的な遺伝子発現の実験結果 D

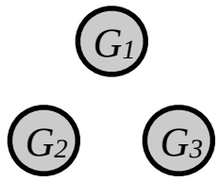
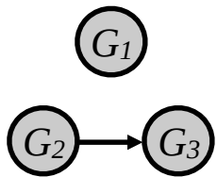
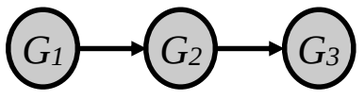
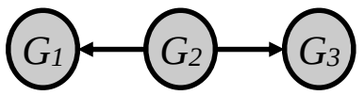
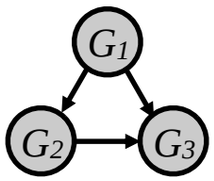
マイクロアレイのデータで考えると, 下の表のようなイメージ.
（10サンプルぶんの実験結果のうち, 3つの遺伝子だけを見ている.）

	exp.1	exp.2	exp.3	exp.4	exp.5	exp.6	exp.7	exp.8	exp.9	exp.10
G_1	2	2	1	2	1	1	2	1	2	1
G_2	1	2	1	2	1	2	2	1	2	1
G_3	1	2	2	2	1	2	2	1	2	1

注）1: 発現していない　2: 発現している

遺伝子が3つの場合のグラフ表現

遺伝子が3つの場合には、全ての有向グラフ、つまり考えられる全ての統計的因果関係を数えきれる。

グラフ番号	構造Ni	同時確率分布P(Ni)=P(G1,G2,G3)の表現
1		$P(G_1,G_2,G_3) = P(G_1)P(G_2)P(G_3)$
2-7		$P(G_1,G_2,G_3) = P(G_3\|G_2)P(G_2)P(G_1)$
8-13		$P(G_1,G_2,G_3) = P(G_3\|G_2)P(G_2\|G_1)P(G_1)$
14-16		$P(G_1,G_2,G_3) = P(G_3\|G_1)P(G_2\|G_1)P(G_1)$
17-19		$P(G_1,G_2,G_3) = P(G_2\|G_1,G_3)P(G_3)P(G_1)$
20-25		$P(G_1,G_2,G_3) = P(G_3\|G_1,G_2)P(G_2\|G_1)P(G_1)$

モデル選択の評価関数

「データD が与えられたとき、遺伝子が3つの場合には25個のうちどのネットワーク構造が最も確からしい構造として推定されるか」
＝
「データD からのモデル選択の問題」
＝
「一般的にはある評価関数を設定し、その値が最大値（あるいは最小値）をとるネットワーク構造 N_i が選択される。」
⇒
例えば、
「データDが与えられたときネットワーク構造 N_i が実現する事後確率，すなわち条件付同時確率 $P(N_i|D)$ を評価関数に用いる。」

Bayes の定理から事後確率 $P(N_i|D)$ を求める

$$P(N_i \mid D) = \frac{P(D \mid N_i)P(N_i)}{\sum_i P(D \mid N_i)P(N_i)} \equiv \frac{P(D \mid N_i)P(N_i)}{P(D)}$$

注） ネットワーク構造 N_i の出現確率 $P(N_i)$ は、データ D 以外の情報から先験的に推測される事前確率で，もしなんらの付加情報がない場合には $P(N_i) = P(N_j)$ となる。

実験データ D から $P(D|N_i)$ を求める

Cooperらは以下の4つの条件
条件１：遺伝子の発現量は離散値（例：1=発現していない、2=発現している）で表現される。
条件２：ベイジアンネットワークのモデルが与えられると，実験データDの各行は相互に
　　独立して現れる。
条件３：実験データDに、欠損値データはない。
条件４：実験データを得る前は、ネットワーク構造 N_i にあらかじめ付与する情報について
　　完全に無知である。
が仮定できる場合には，条件付同時確率 $P(D|N_i)$ が次式で与えられることを示した。

$$P(D \mid N_i) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} (N_{ijk})!$$

ただし、 $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$

Cooper,G.F. et al.(1992) A Bayesian method for the induction of probabilistic networks from data, Mach.Learn., 9, 309

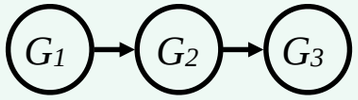
Cooperらの式の説明

$$P(D \mid N_i) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} (N_{ijk})!$$

1） n は遺伝子の数で、今回の場合、$n = 3$。
2） q_i は遺伝子 G_i の親 $\pi(G_i)$ が取りうる状態の数。
今回の場合、親が1つの遺伝子 G_m の場合，すなわち$\pi(G_i) = (G_m)$ の場合には $q_i = 2$。
親が2つ遺伝子 G_m と G_n の場合，すなわち $\pi(G_i) = (G_m, G_n)$ の場合には $q_i = 4$。
3） r_i は遺伝子 G_i が取りうる状態の数で，今回の場合、$r_i = 2$.
4） N_{ijk} は，遺伝子 G_i が「k =1:発現していない」あるいは「k=2:発現している」のどちらか
の値をとり，G_iの親 $\pi(G_i)$ が1 ≦ j ≦ q_i の j 番目の状態を取っている数（ケースの数）。

実験データDから事後確率 $P(N_i|D)$ を計算する

$P(N_8)=P(G_3|G_2)P(G_2|G_1)P(G_1)$



ネットワーク構造N_8:

遺伝子G_1の状態

$\pi(G_1)$	1:非発現	2:発現	計
(なし)	N_{11}=5	N_{12}=5	N_1=10

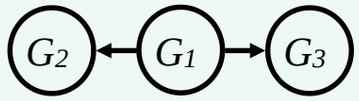
遺伝子G_2の状態

$\pi(G_2)$	1:非発現	2:発現	計
(G_1=1)	N_{211}=4	N_{212}=1	N_{21}=5
(G_1=2)	N_{221}=1	N_{222}=4	N_{22}=5

遺伝子G_3の状態

$\pi(G_3)$	1:非発現	2:発現	計
(G_2=1)	N_{311}=4	N_{312}=1	N_{31}=5
(G_2=2)	N_{321}=0	N_{322}=5	N_{32}=5

$P(N_8)=P(G_3|G_1)P(G_2|G_1)P(G_1)$



ネットワーク構造N_{14}:

遺伝子G_1の状態

$\pi(G_1)$	1:非発現	2:発現	計
(なし)	N_{11}=5	N_{12}=5	N_1=10

遺伝子G_2の状態

$\pi(G_2)$	1:非発現	2:発現	計
(G_1=1)	N_{211}=4	N_{212}=1	N_{21}=5
(G_1=2)	N_{221}=1	N_{222}=4	N_{22}=5

遺伝子G_3の状態

$\pi(G_3)$	1:非発現	2:発現	計
(G_1=1)	N_{311}=3	N_{312}=2	N_{31}=5
(G_1=2)	N_{321}=1	N_{322}=4	N_{32}=5

Cooperらの式から

$P(D|N_8) = 2.23 \times 10^{-9}$　　$P(D|N_{14}) = 2.23 \times 10^{-10}$

もし $P(N_i) = P(N_j)$ とすると、**Bayesの定理から**

$P(N_8|D) = 0.109$　　$P(N_{14}|D) = 0.011$

ただし、最大の事後確率を与えるネットワーク構造は

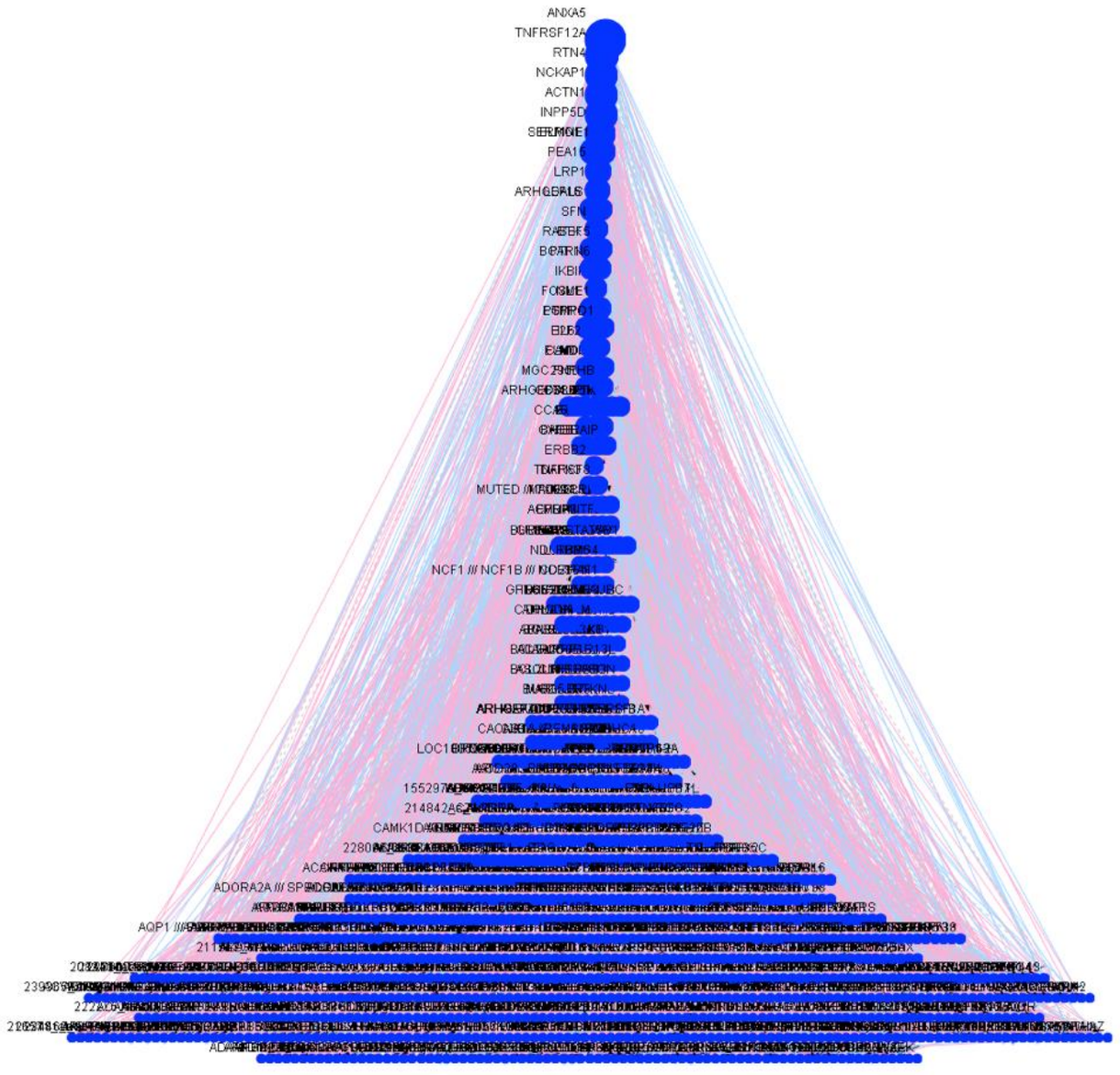
$N_{13}: G_3 \rightarrow G_2 \rightarrow G_1$ で、$P(N_{13}|D) = 0.112$

3. 遺伝子ネットワーク

遺伝子ネットワーク

- 遺伝子発現レベルのデータから推定されたベイジアンネットワークが、遺伝子ネットワーク。

- ただ、相関係数を調べて、線で結んだわけではない。
- 矢印（エッジ）には方向がある。



cell innovator

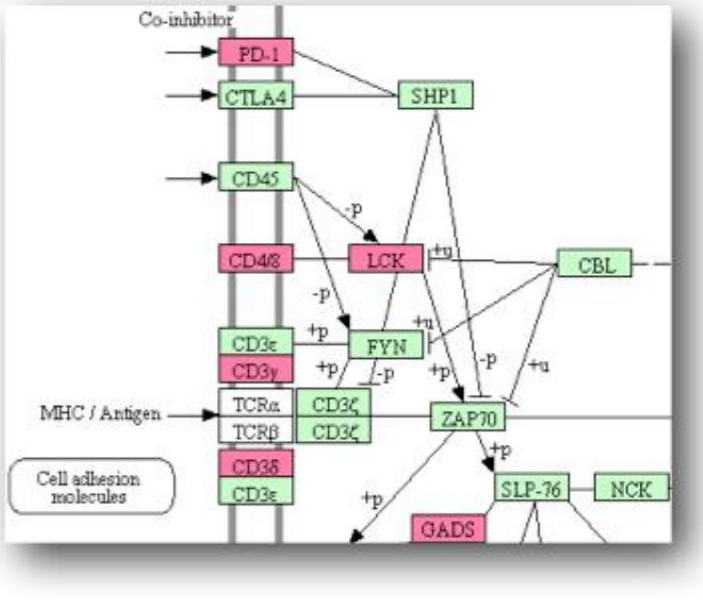
遺伝子ネットワークの意味するもの

- 遺伝子ネットワークは、いわゆる「パスウェイ」**ではない**。
- いわゆる「パスウェイ」は、下記の情報のいずれか。
  - タンパク間相互作用 = Protein-Protein Interaction (PPI) network。
  - 遺伝子発現制御 = 転写因子と、その転写制御領域を持つ遺伝子の関係。
  - 共発現 = ともに発現している遺伝子の関係。
  - 文献情報 = 文献に、「制御関係あり」と報告された関係。

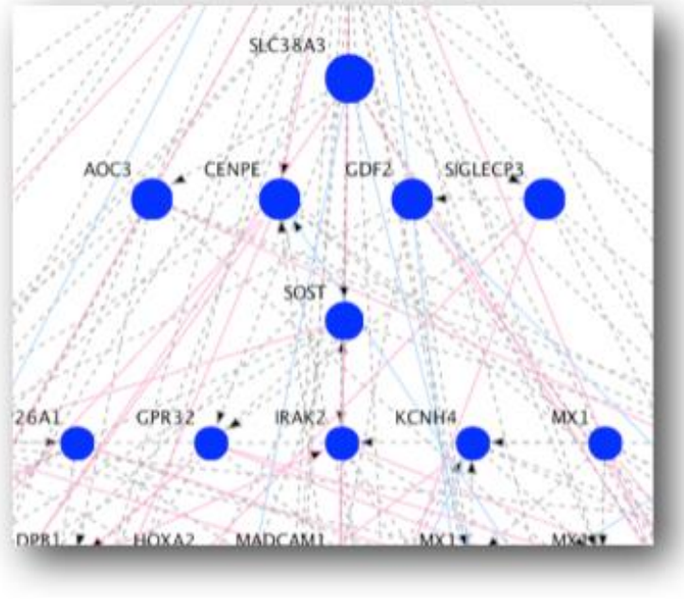
- 遺伝子ネットワークは、パスウェイとは異なる、新たな相互作用の情報。

cell innovator

パスウェイ解析と遺伝子ネットワーク解析の違い



?



- パスウェイ解析は、「どの遺伝子が増加、減少した遺伝子した」のか、既知の情報をもとに**結果**を表示するもの。
- 遺伝子ネットワーク解析は、「どの遺伝子の影響が強い」のか、**原因**を予想するもの。また、未知の情報を含む。

遺伝子ネットワークの利点と欠点

- **利点**
  - 純粋にマイクロアレイデータのみから推定できるため、文献情報や、配列情報などのアノテーション情報を必要としない。（データドリブン）
  - lincRNAなど、機能が不明な遺伝子であっても、制御関係を推定できる。
  - これまでに未知の制御関係を発見できる可能性がある。

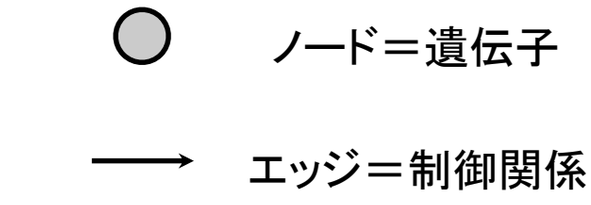
- **欠点**
  - 数十から数百個のマイクロアレイデータが必要。＝高いコスト
  - 高レベルの計算機環境が必要。（スーパーコンピューターなど）

現在では、推定時の問題を回避可能

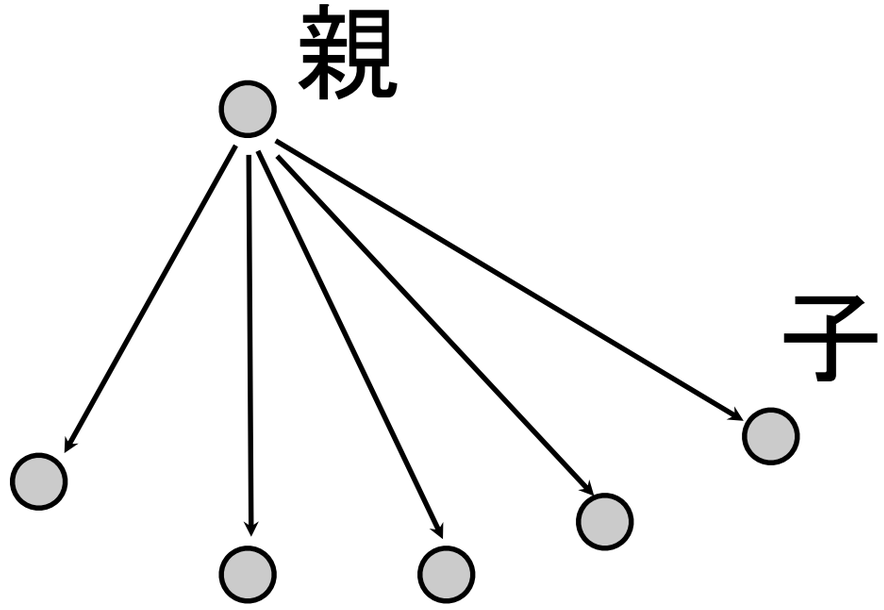
- NCBI の Gene Expression Omnibus (GEO) に公開されているマイクロアレイデータを用いて推定を行う。**--> 高コストの問題を回避。**
  - 例えば、Cancer Cell Line Encyclopedia (CCLE) には、およそ 1000 サンプル分のマイクロアレイデータが公開されている。[GSE36133]
- 計算には、「京（SCLS）」などのスーパーコンピューターを利用。**--> 計算機環境の問題をクリア。**

遺伝子ネットワークのグラフ論的な解釈

- 数学的には、丸を「**ノード**」、矢印を「**エッジ**」と呼ぶ。
- エッジの始点になるノードが「**親**」
- エッジの終点になるノードが「**子**」

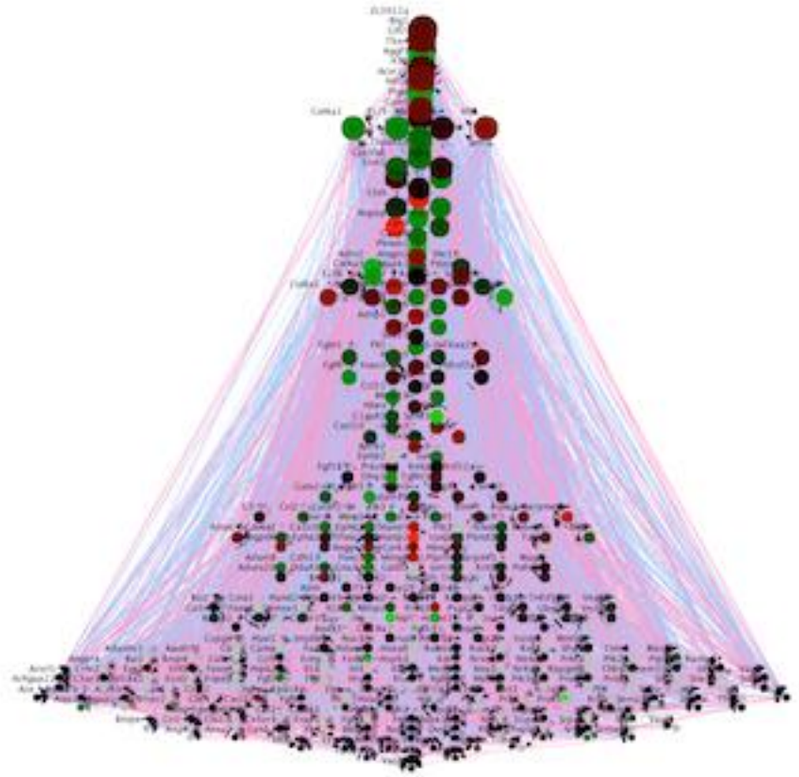


- ネットワークの構造としては、一部の親に多数の子が集中するという構造になることが多い。（スケールフリー）
- 特に「子が多いノード」は、「ハブ」と呼ばれる。

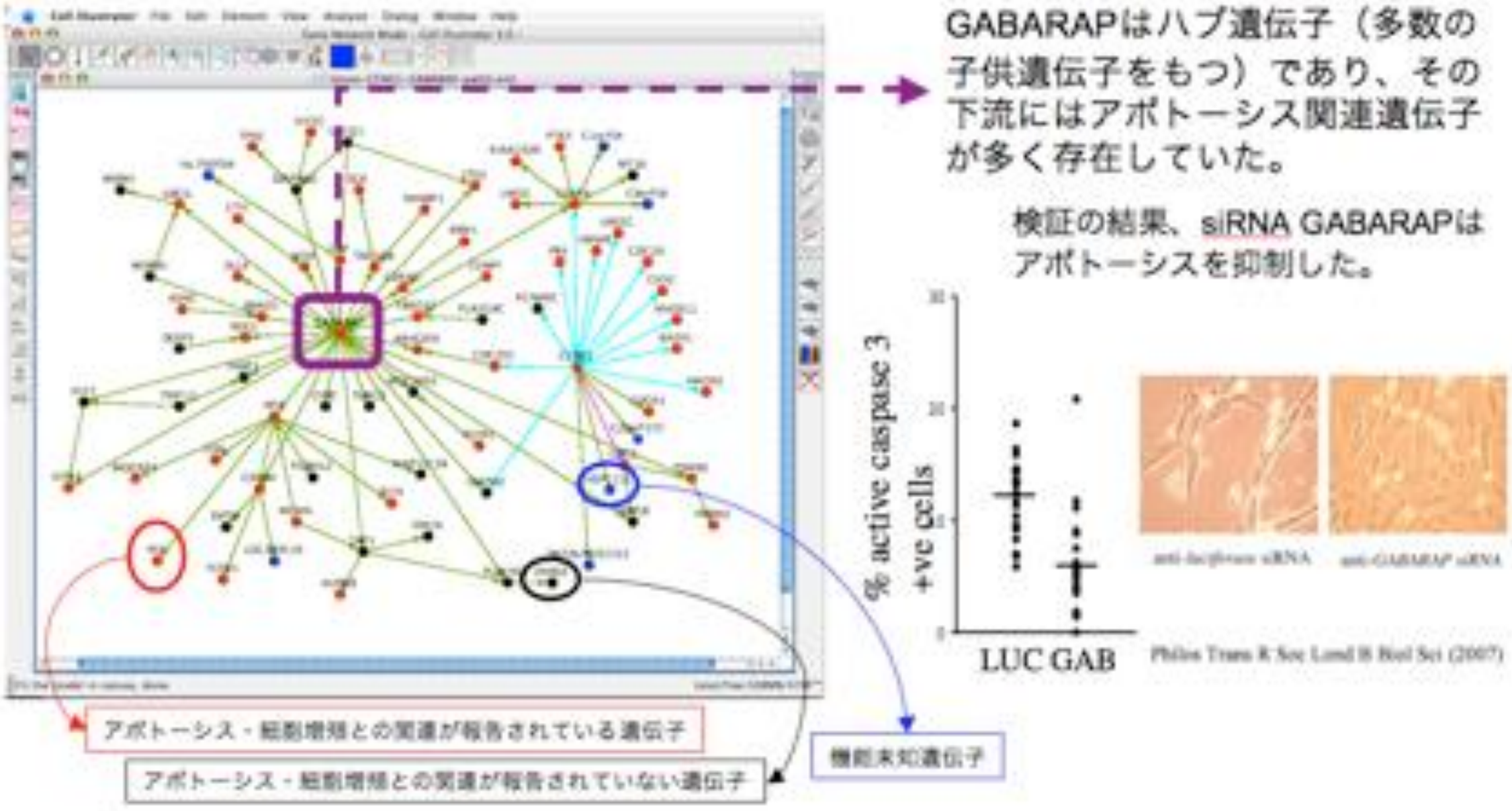


遺伝子ネットワークの利用方法

- 「ハブ」を探す＝ネットワーク中で影響力の強い遺伝子を見つける。（ハブの発現レベルが変化すると、子の発現レベルが変化するはず。）
- 遺伝子ネットワークのノードを、logFCなどで色づけ。（パスウェイと同様、マイクロアレイデータの解析に利用。）
- 上流解析：発現変動遺伝子を制御するのは、どの遺伝子か？（原因はどれか？）



解析事例（ハブをノックダウン）



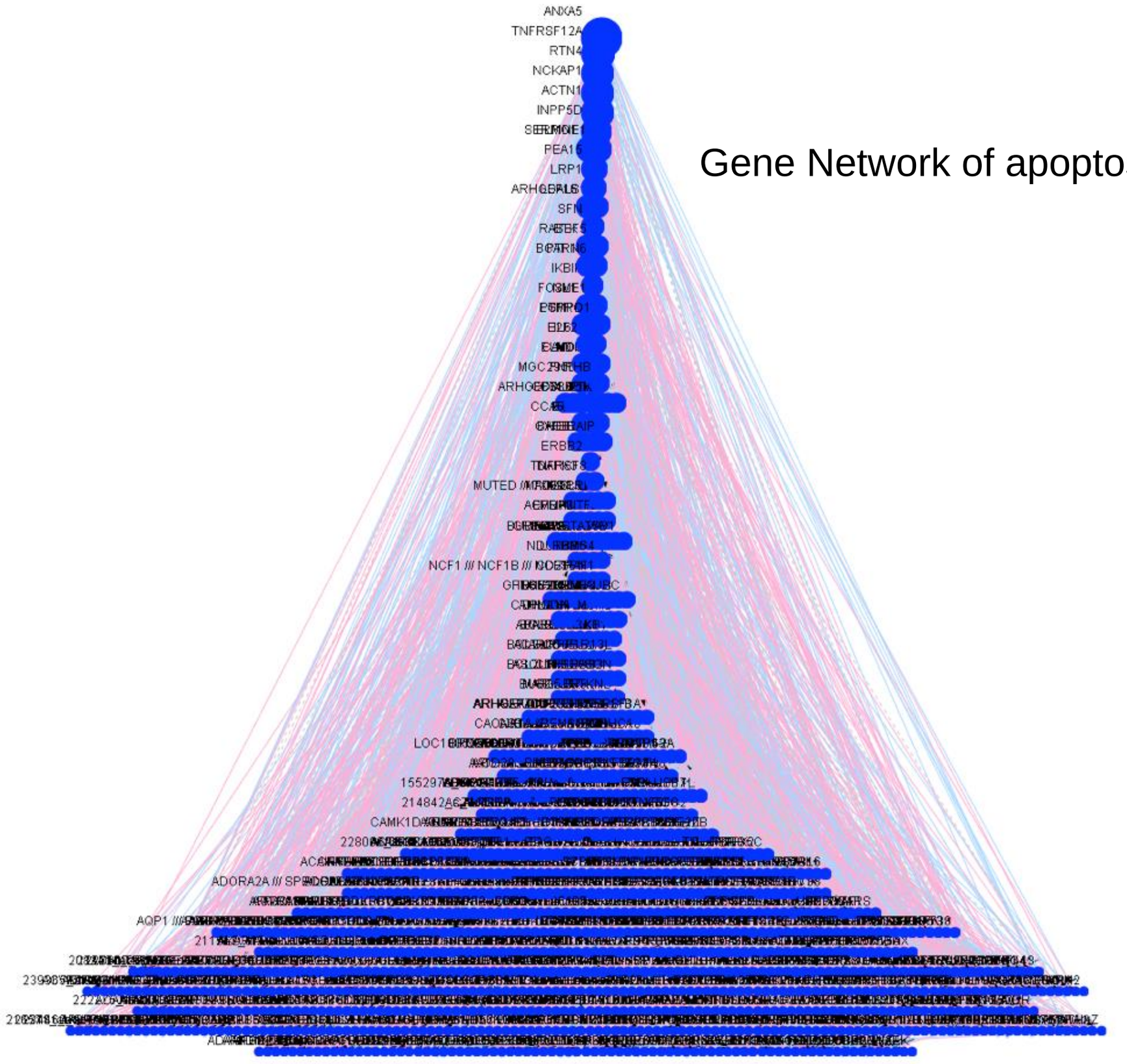
Affara, M., Dunmore, B., Savoie, C., Imoto, S., Tamada, Y., Araki, H., Charnock-Jones, D. S., et al. (2007). Understanding endothelial cell apoptosis: what can the transcriptome, glycome and proteome reveal? Philosophical transactions of the Royal Society of London Series B, Biological sciences, 362(1484), 1469–1487. doi:10.1098/rstb.2007.2129

cell innovator

よくある質問、疑問

- エッジの何パーセントが当たっているのか？エッジの何パーセントが既知で、何パーセントが未知の情報か？
- シグナル伝達系の活性化される順序は、分からないのか？
- レセプターが、リガンドを活性化しているように見えるが？
- 「ハブ」といっても、ただのキナーゼでは？転写因子でないから、転写は制御できないハズ。

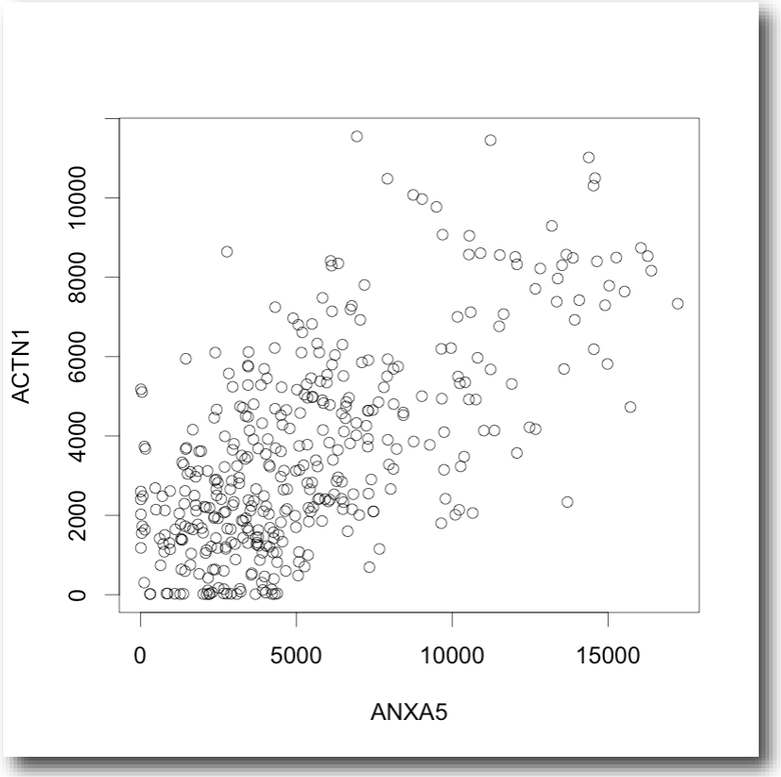
データからはそう見える（バントしないほうがいい）といっているにすぎない。

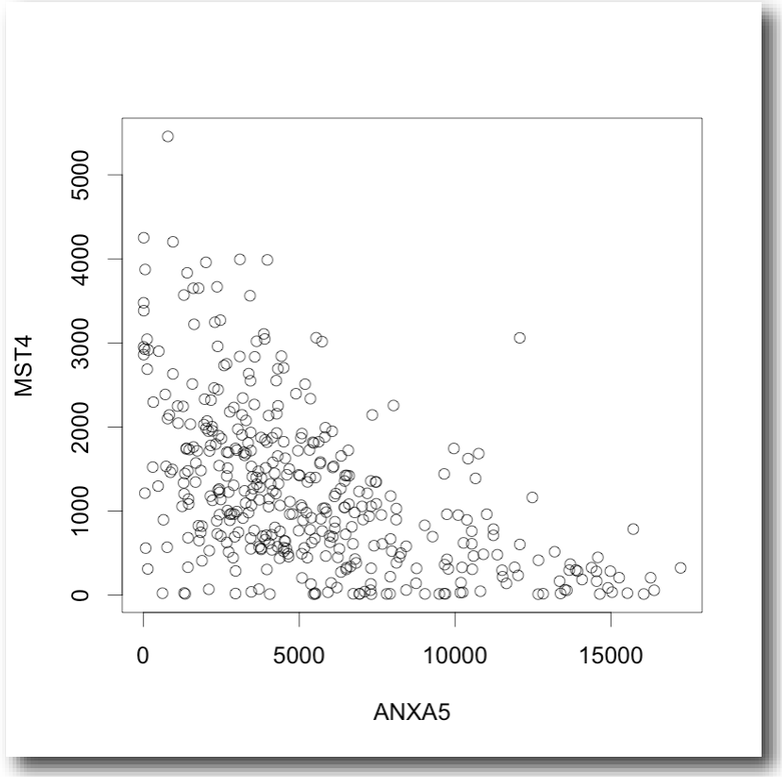


Gene Network of apoptosis related genes

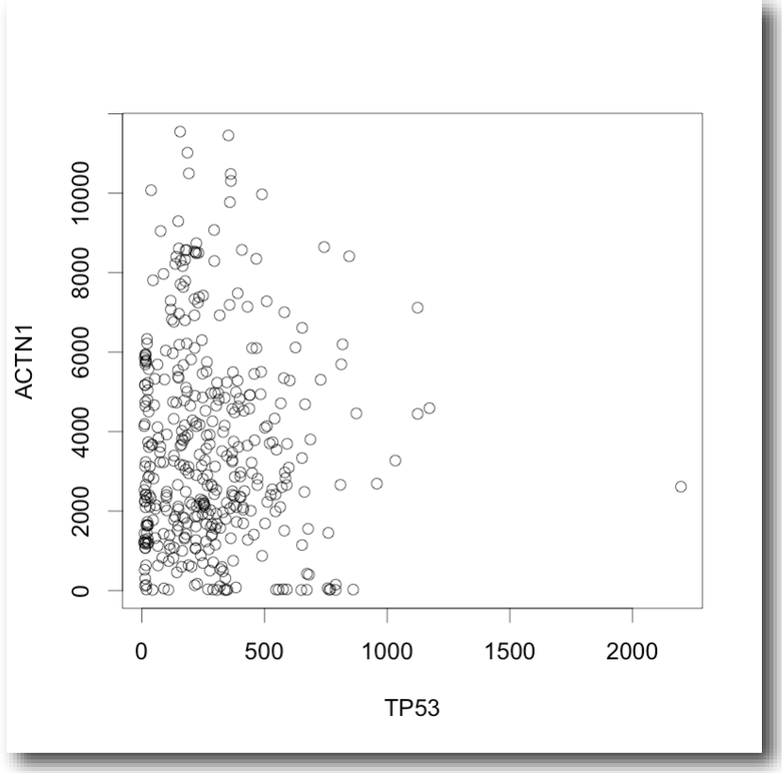
cell innovator

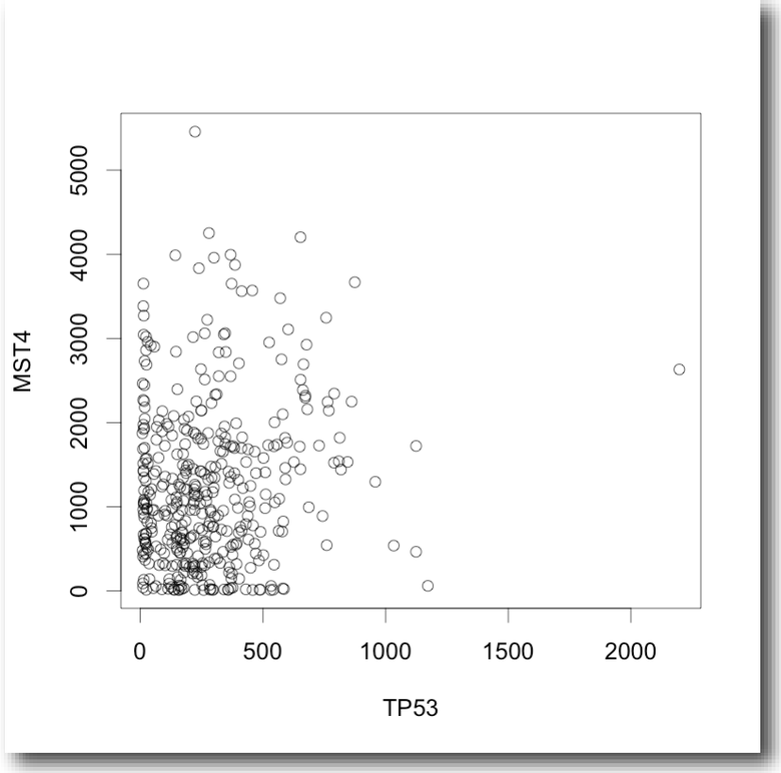
ANXA5 (top gene)



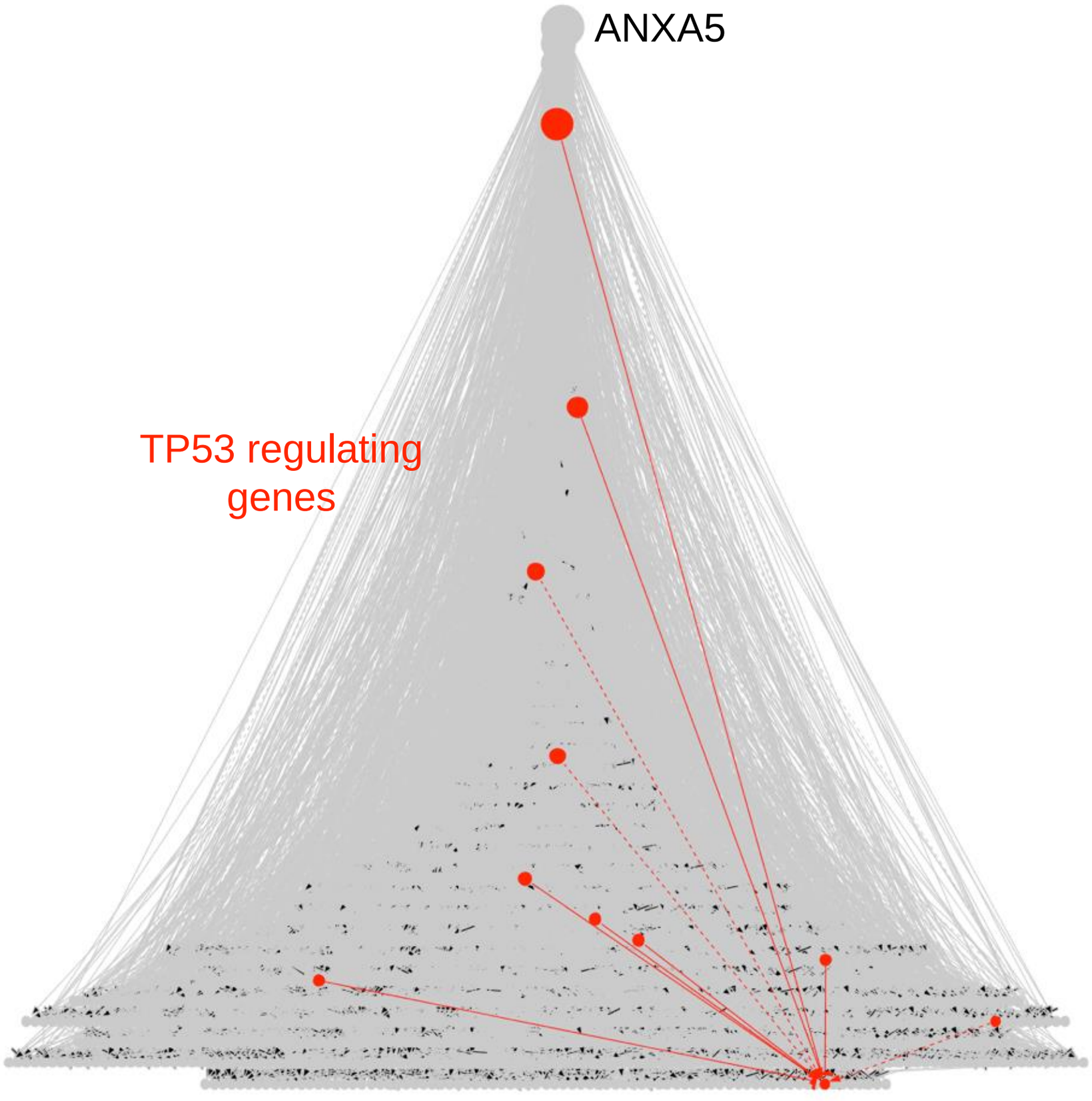


TP53
(bottom gene)





cell innovator


ANXA5
TP53 regulating genes
TP53

- 統計学的に得られた結論は、感覚的には合わないかもしれませんが、参考にするのはどうでしょうか？

- 絶対、バントしてはダメだとか、言うつもりではありません。

- 昨年も、レッドソックスが優勝しましたね。。。。