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

cell innovator

1. マネーボール：統計学の応用

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。

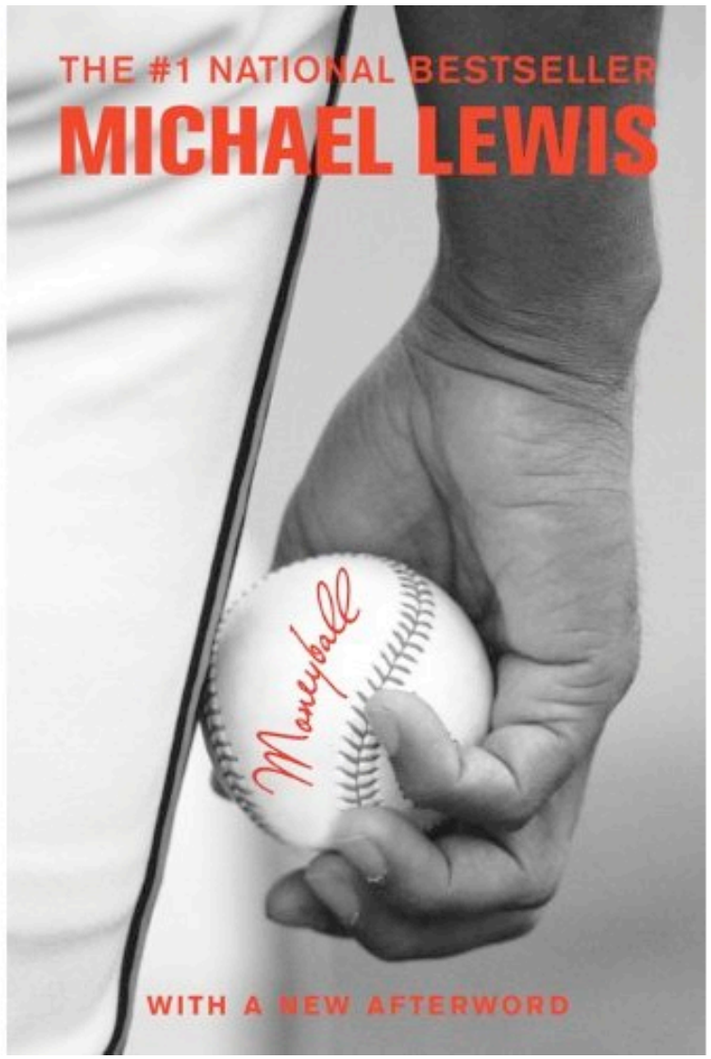
- データアナリスト、データサイエンティストが25万人不足。 http://www.nikkei.com/article/DGXNZO57421630X10C13A7EA1000/

「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

cell innovator

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。
- 送りバントされたら、2塁に投げろ。

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、、、

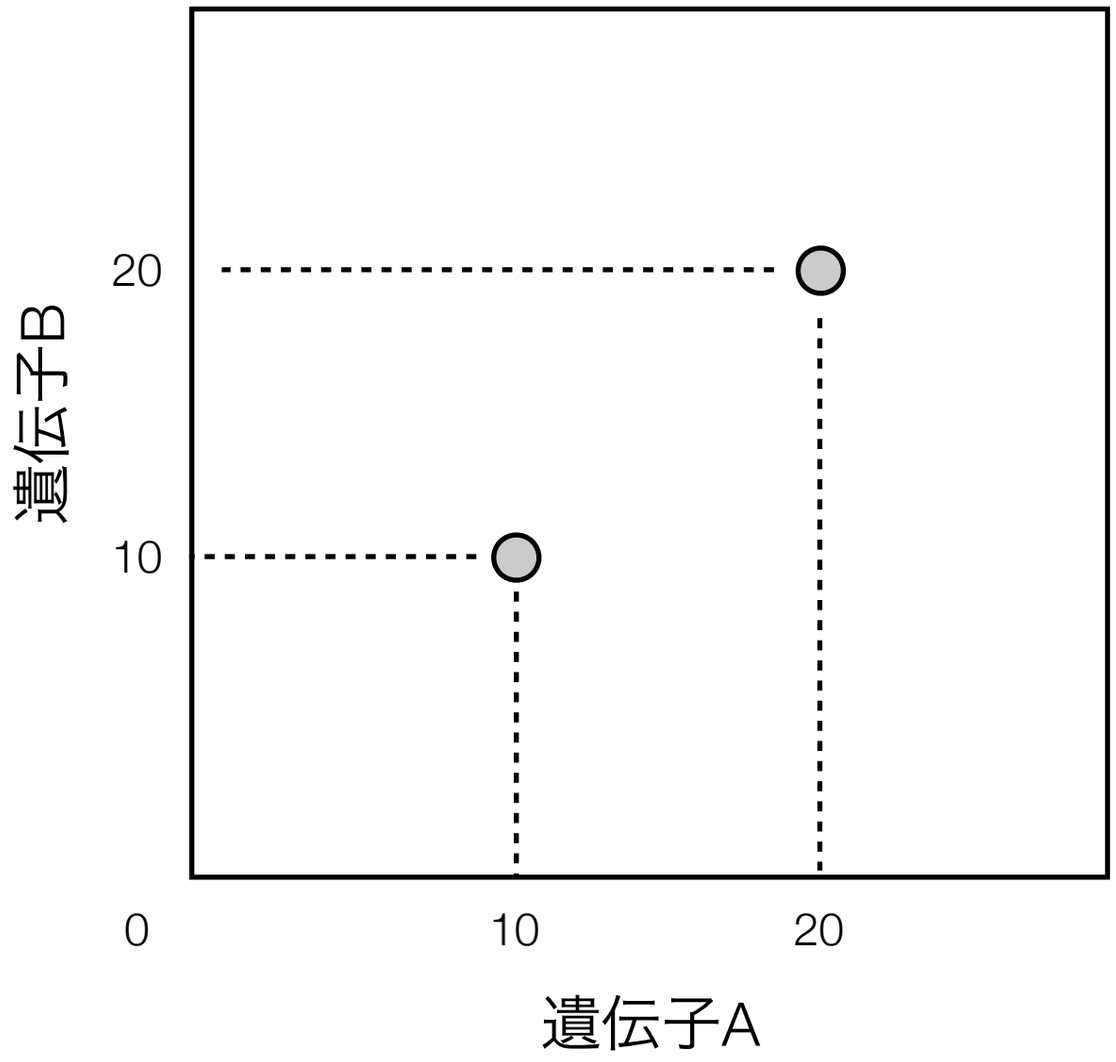
- 近年、統計学的なアプローチが、よく用いられるようになった。

- 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

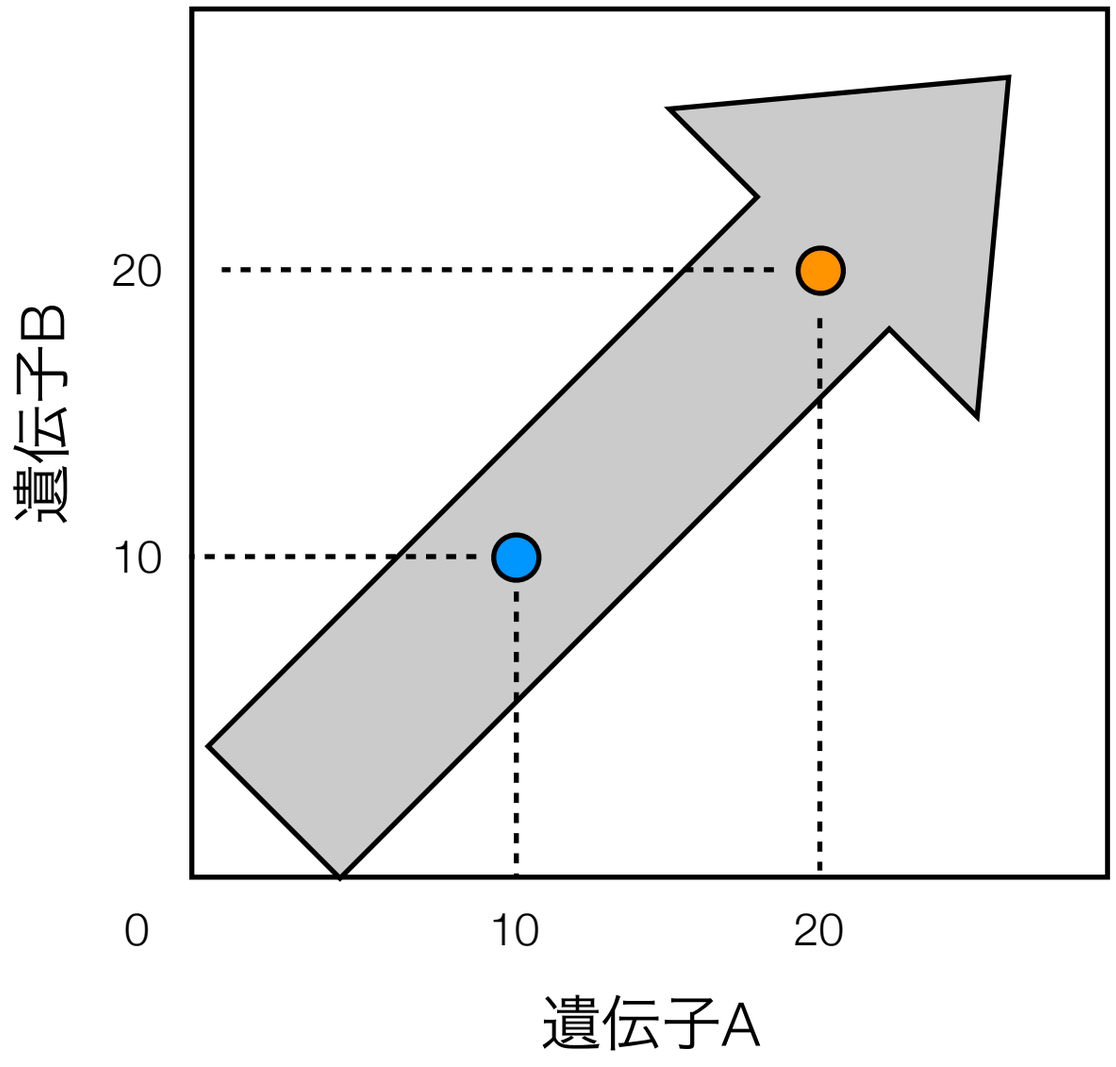
遺伝子発現と散布図



- 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)

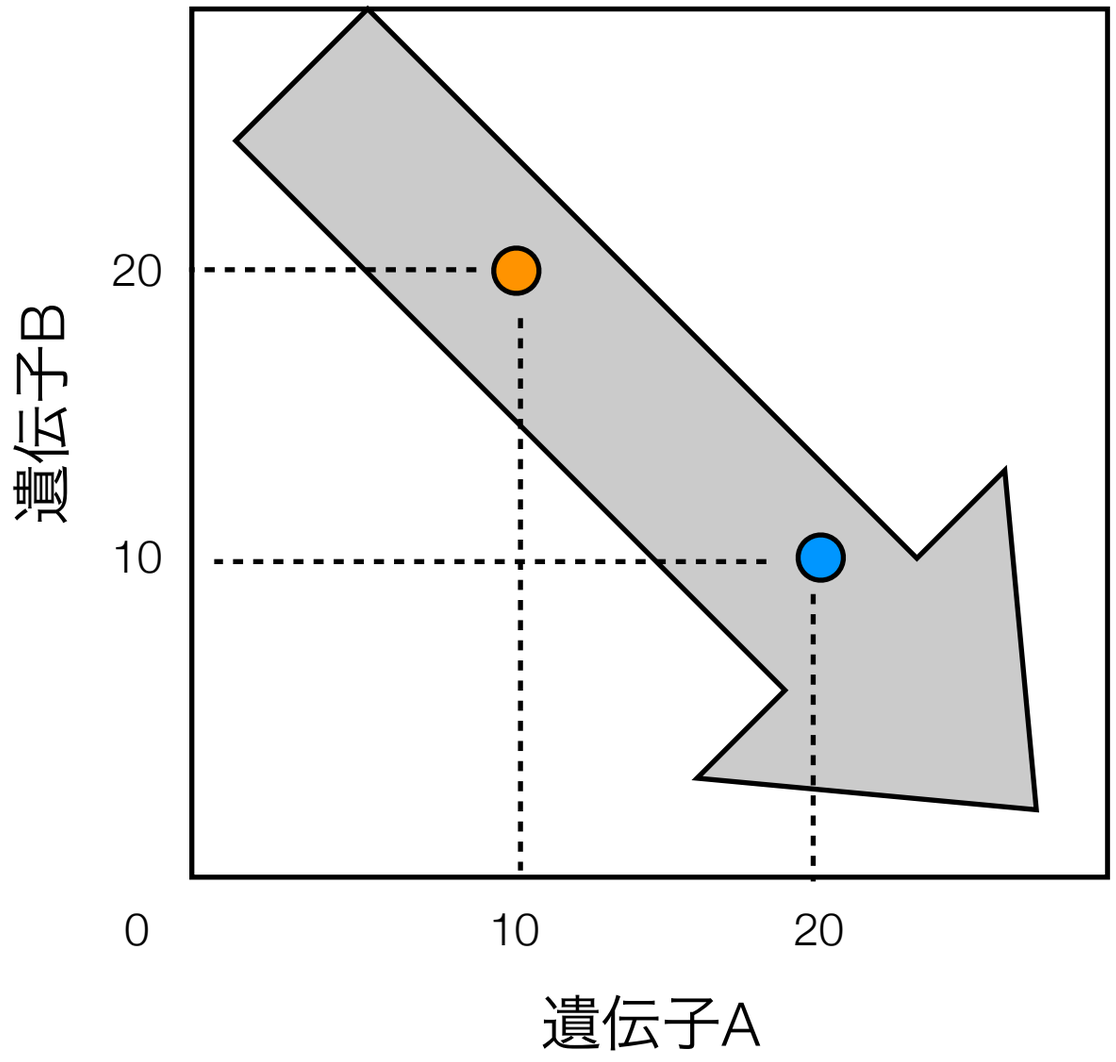


- つまり、遺伝子Aの発現量が低いとき、遺伝子Bの発現量も低い。
- また、遺伝子Aの発現量が高いとき、遺伝子Bの発現量も高い。

- 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)

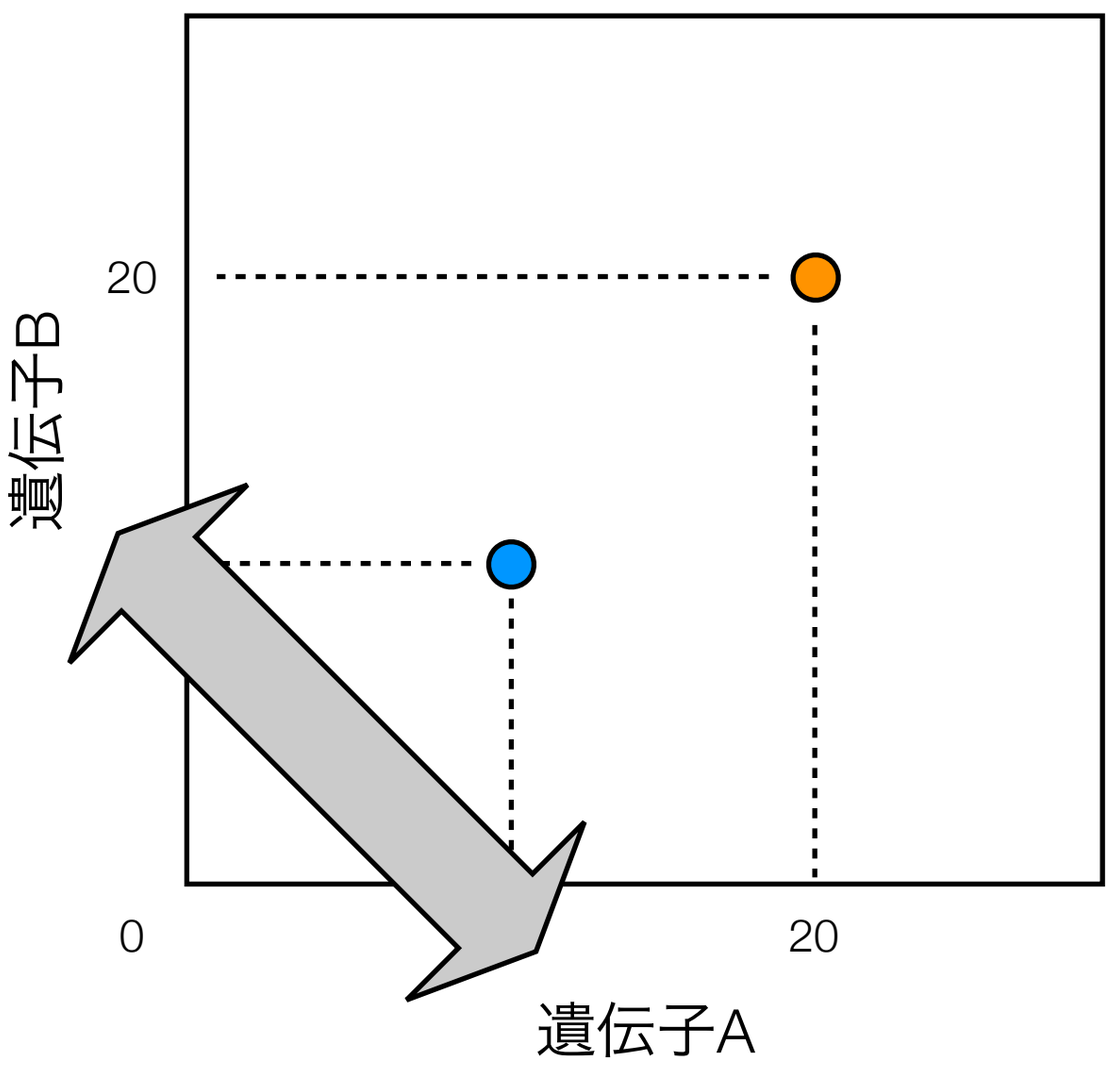


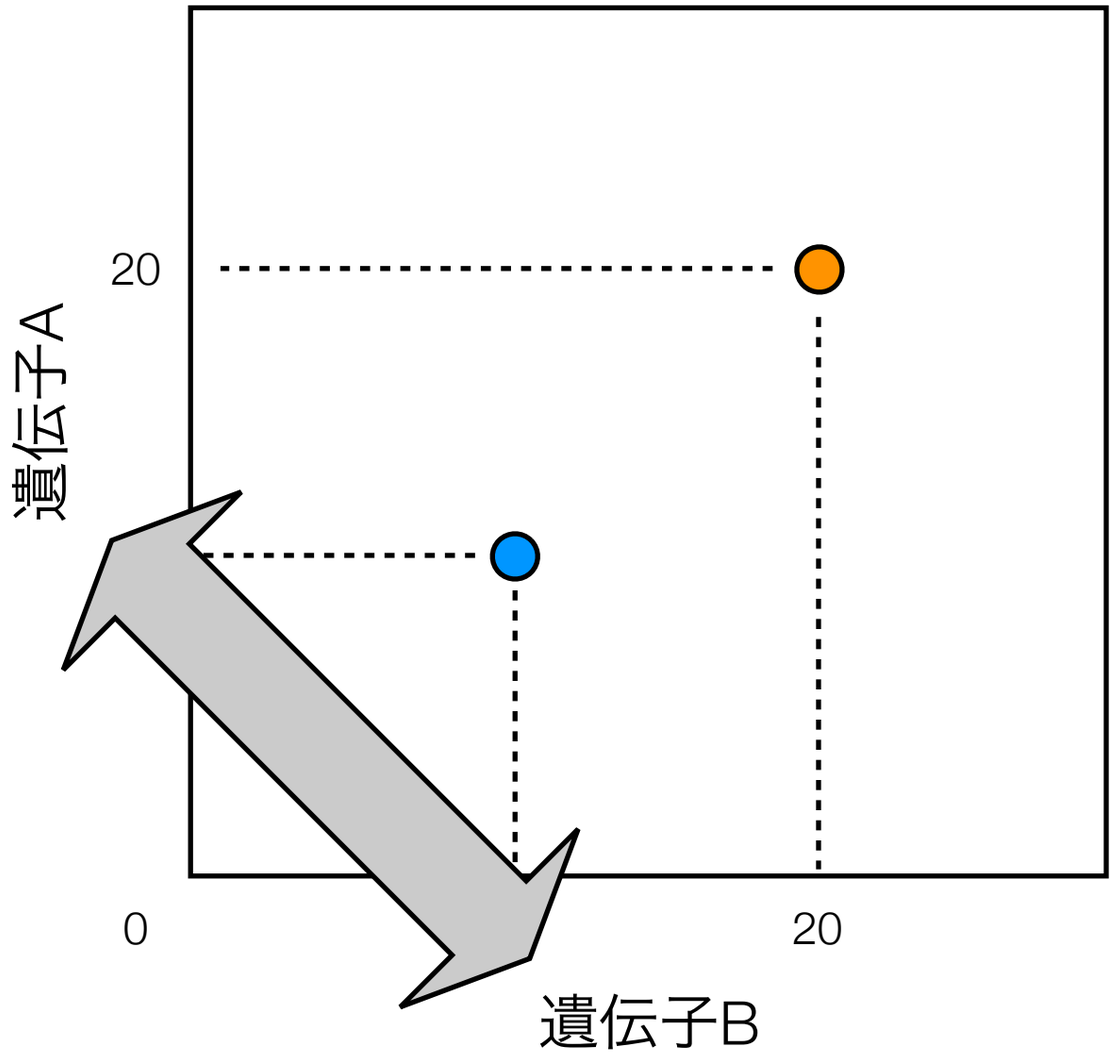
- その逆なら、遺伝子Aの発現量が低いとき、遺伝子Bの発現量は高い。
- また、遺伝子Aの発現量が高いとき、遺伝子Bの発現量は低い。

- 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？

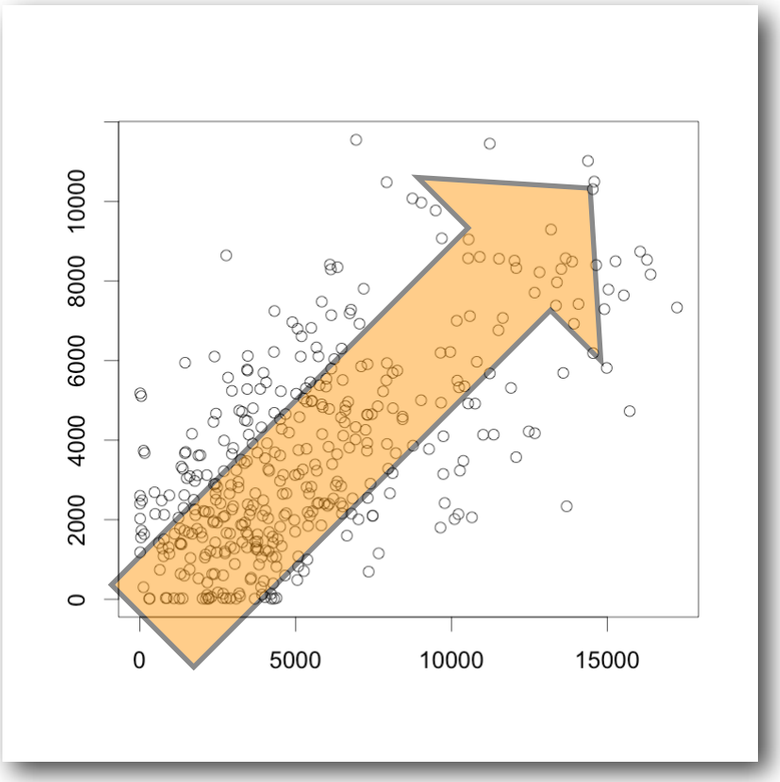


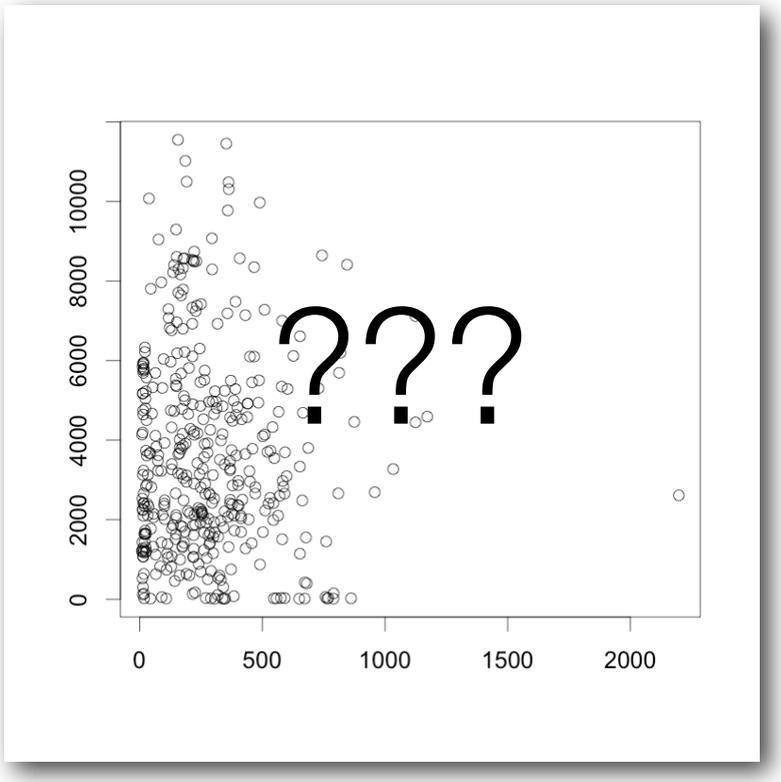


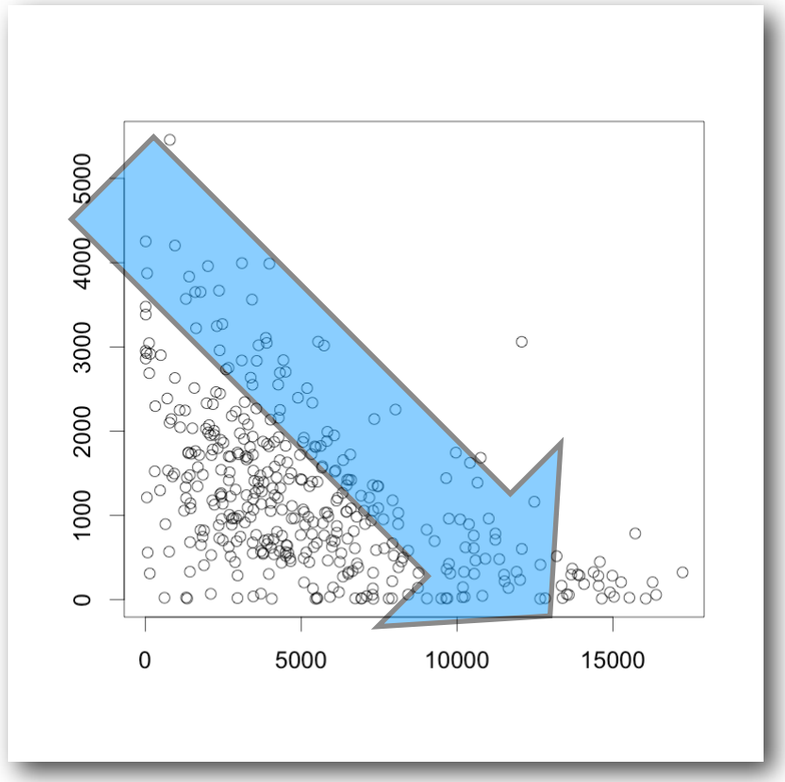
- X軸とY軸を入れ替えても同じなので、どちらが上流か分からない？？

cell innovator

データを増やしていくと見えてくるもの







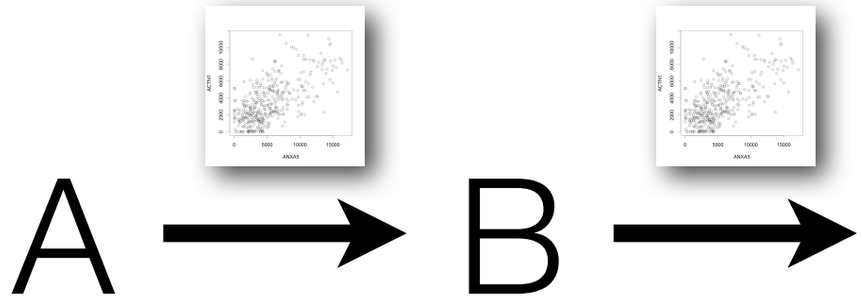
- 上記は、400サンプル＝400個の点における関係を見たもの。
- サンプル数を増やしていくと、「関係の度合い」（＝確率）も見えそう。

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）

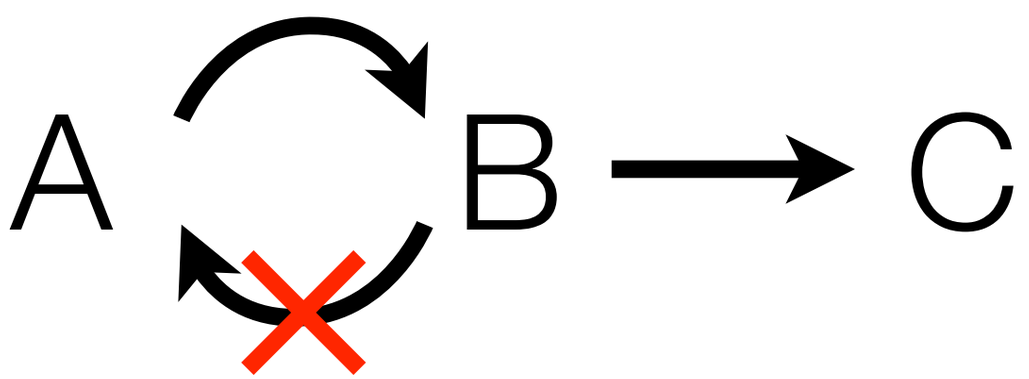
- 遺伝子Aが、ある確率で、遺伝子Bを制御していて、
- 遺伝子Bが、ある確率で、遺伝子Cを制御している。



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？

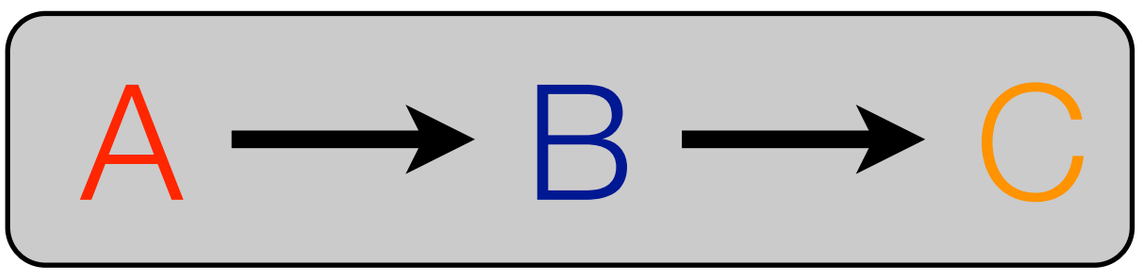
A ← B → C

cell innovator

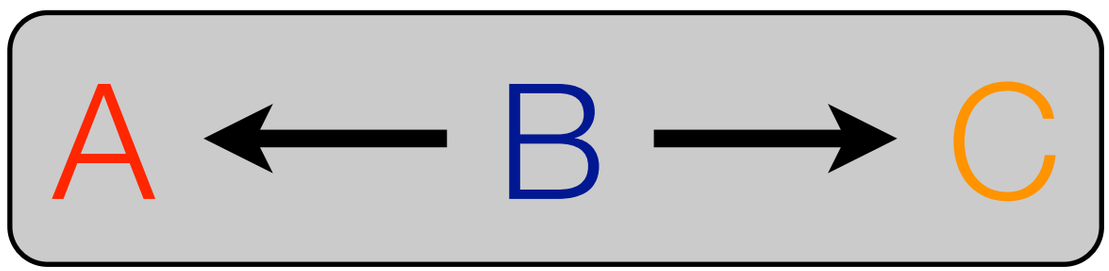
ベイジアンネットワーク（モデル）

- 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）



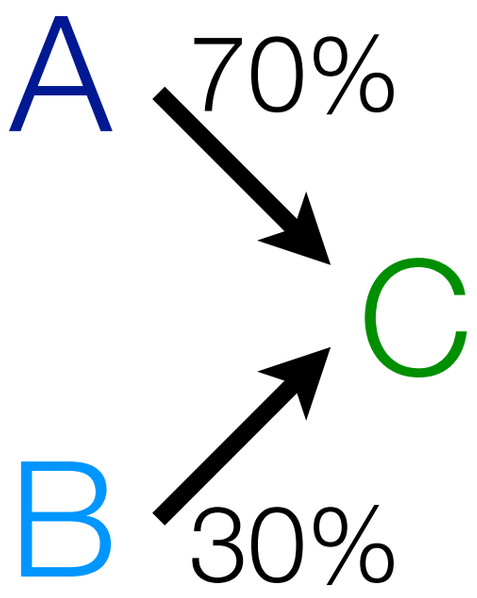
原因はどちら？



cell innovator

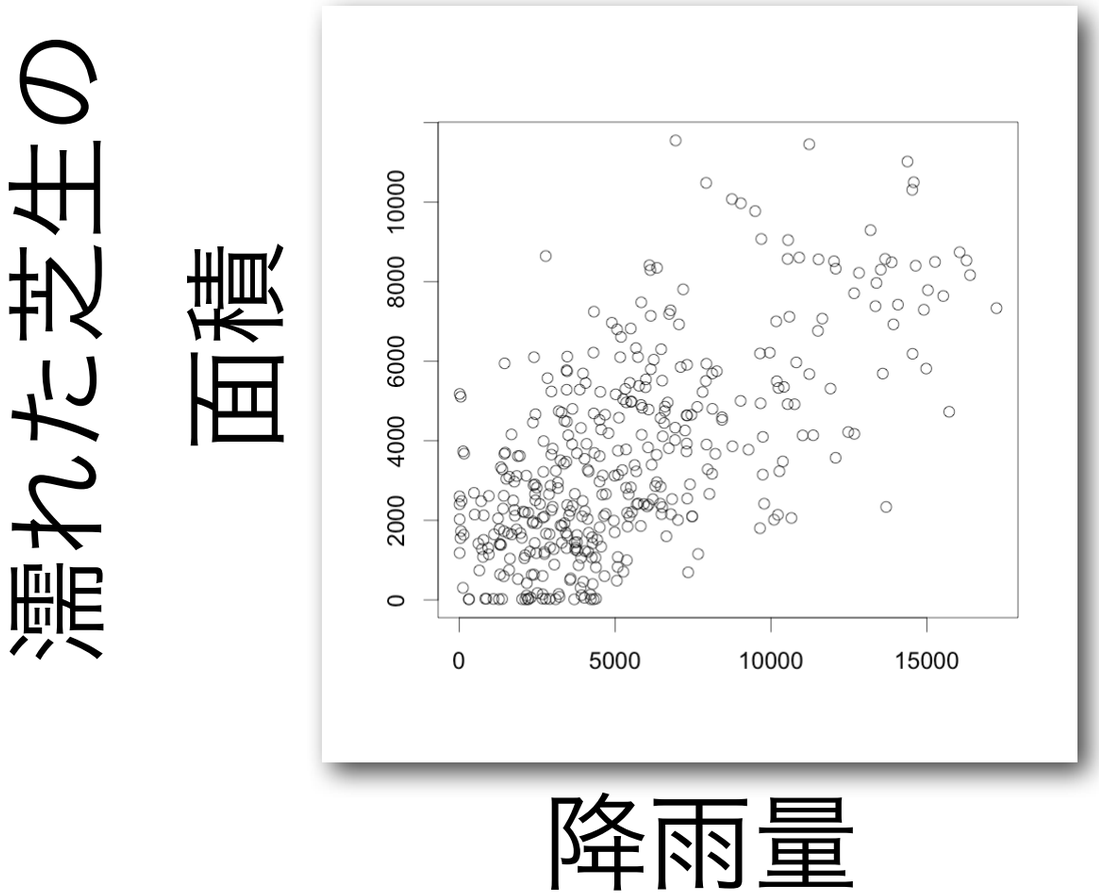
例えば、雨とスプリンクラーと芝生の関係は？

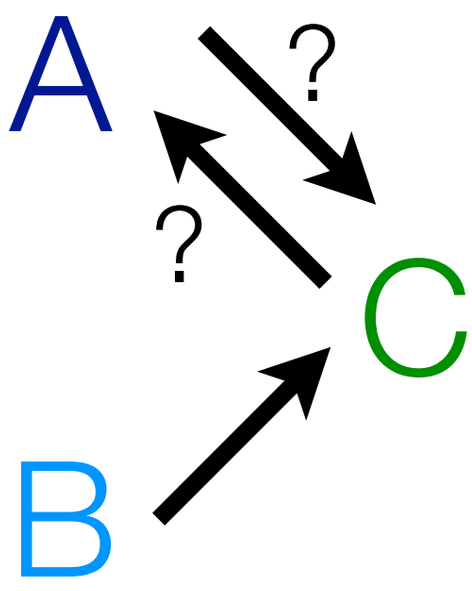
- A: 雨が降る（降雨量）。
- B: スプリンクラーが作動する。
- C: 芝生が濡れる。



- 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？



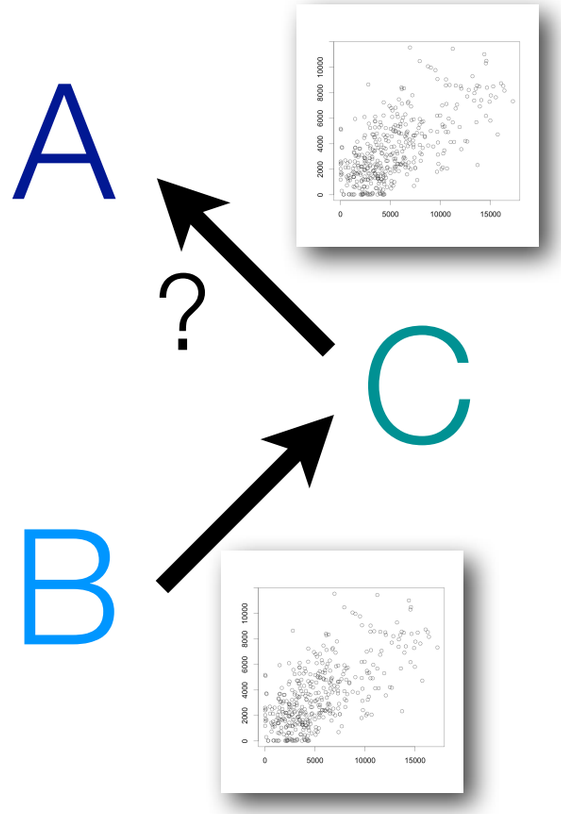


- 雨が降ったから、芝生が濡れたのか？ A --> C
- 芝生が濡れたから、雨が降ったのか？ C --> A

cell innovator

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！

cell innovator

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）

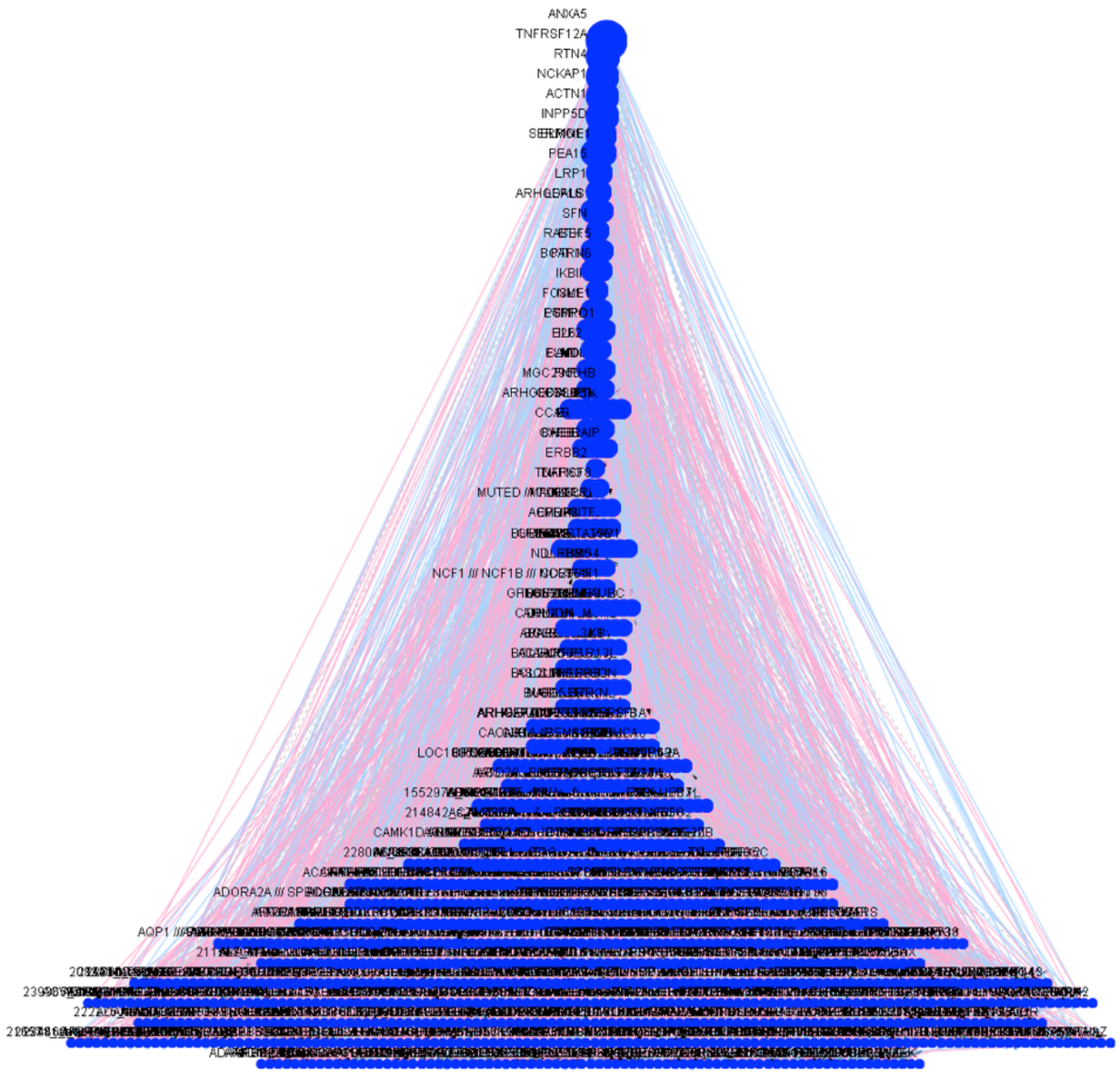
イメージ的には、とにかく総当たりで、
すべてのネットワークのパターンをチェックして、
もっともらしいネットワークの状態を推定

cell innovator

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。



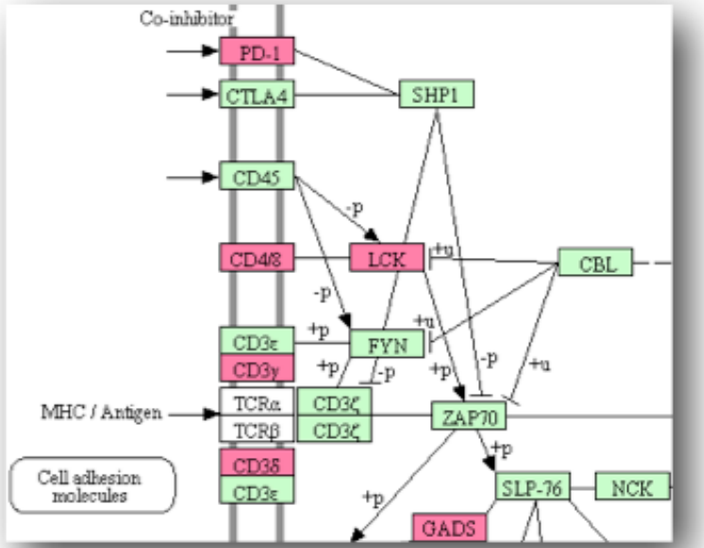
cell innovator

遺伝子ネットワークの意味するもの

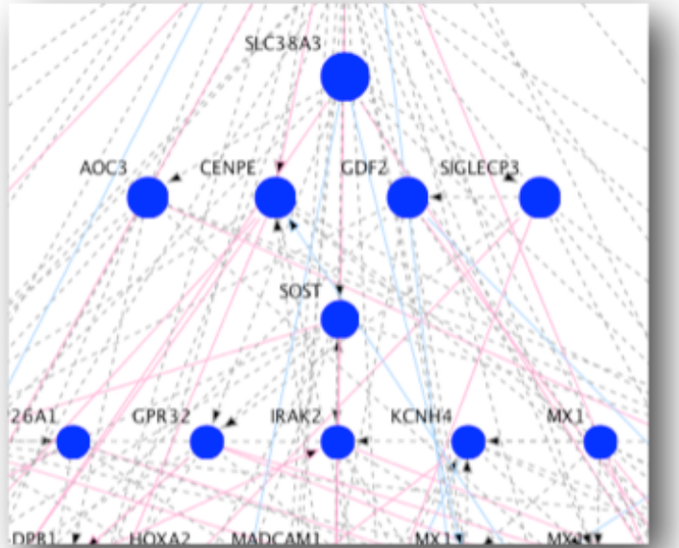
- 遺伝子ネットワークは、いわゆる「パスウェイ」**ではない**。
- いわゆる「パスウェイ」は、下記の情報のいずれか。
  - タンパク間相互作用 = Protein-Protein Interaction (PPI) network。
  - 遺伝子発現制御 = 転写因子と、その転写制御領域を持つ遺伝子の関係。
  - 共発現 = ともに発現している遺伝子の関係。
  - 文献情報 = 文献に、「制御関係あり」と報告された関係。

- 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。

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。

cell innovator

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）

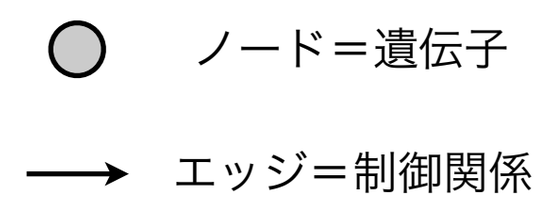
現在では、推定時の問題を回避可能

- 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]

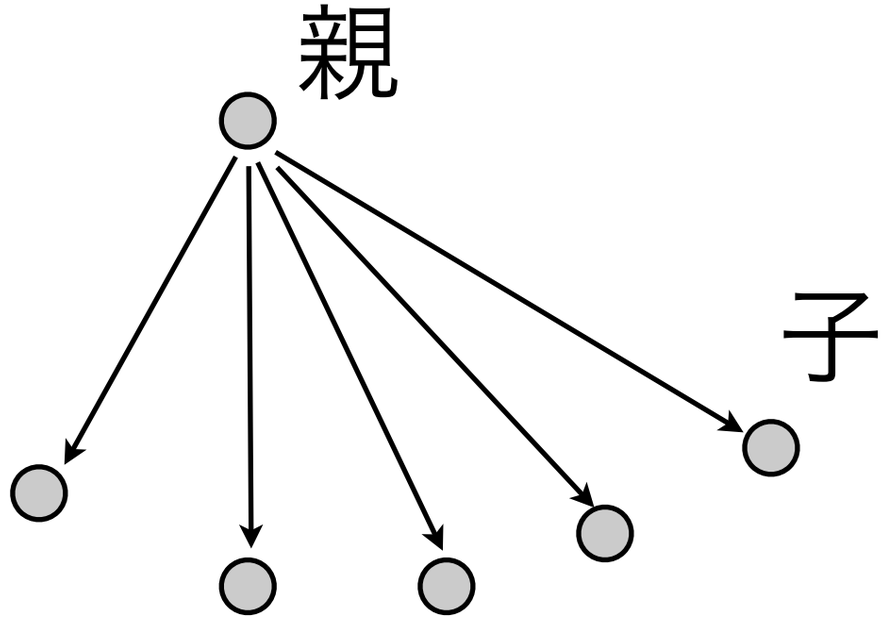
- 計算には、「京（SCLS）」などのスーパーコンピューターを利用。 **--> 計算機環境の問題をクリア。**

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**」

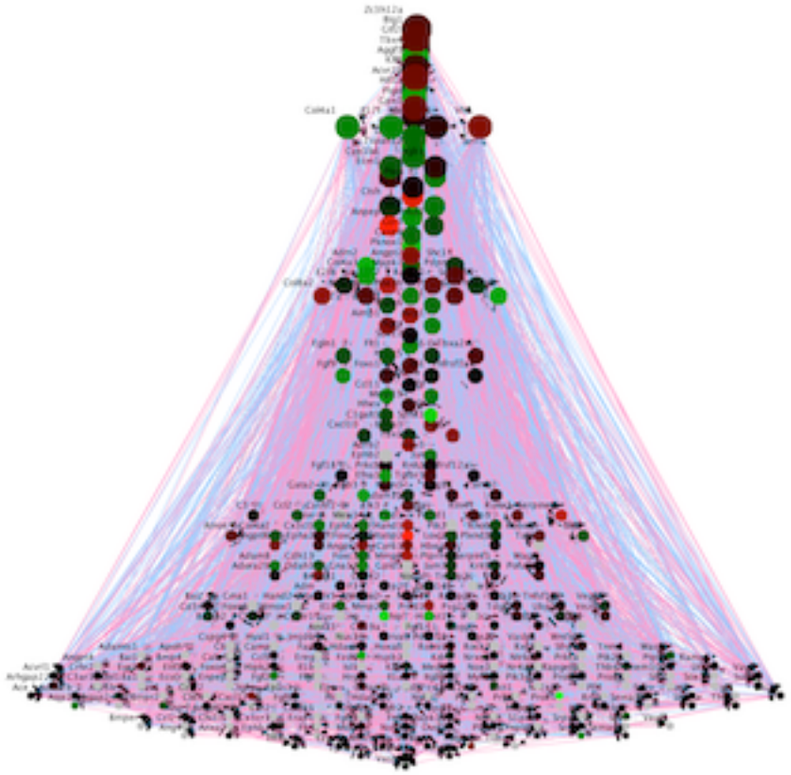


- ネットワークの構造としては、一部の親に多数の子が集中するという構造になることが多い。（スケールフリー）
- 特に「子が多いノード」は、「ハブ」と呼ばれる。



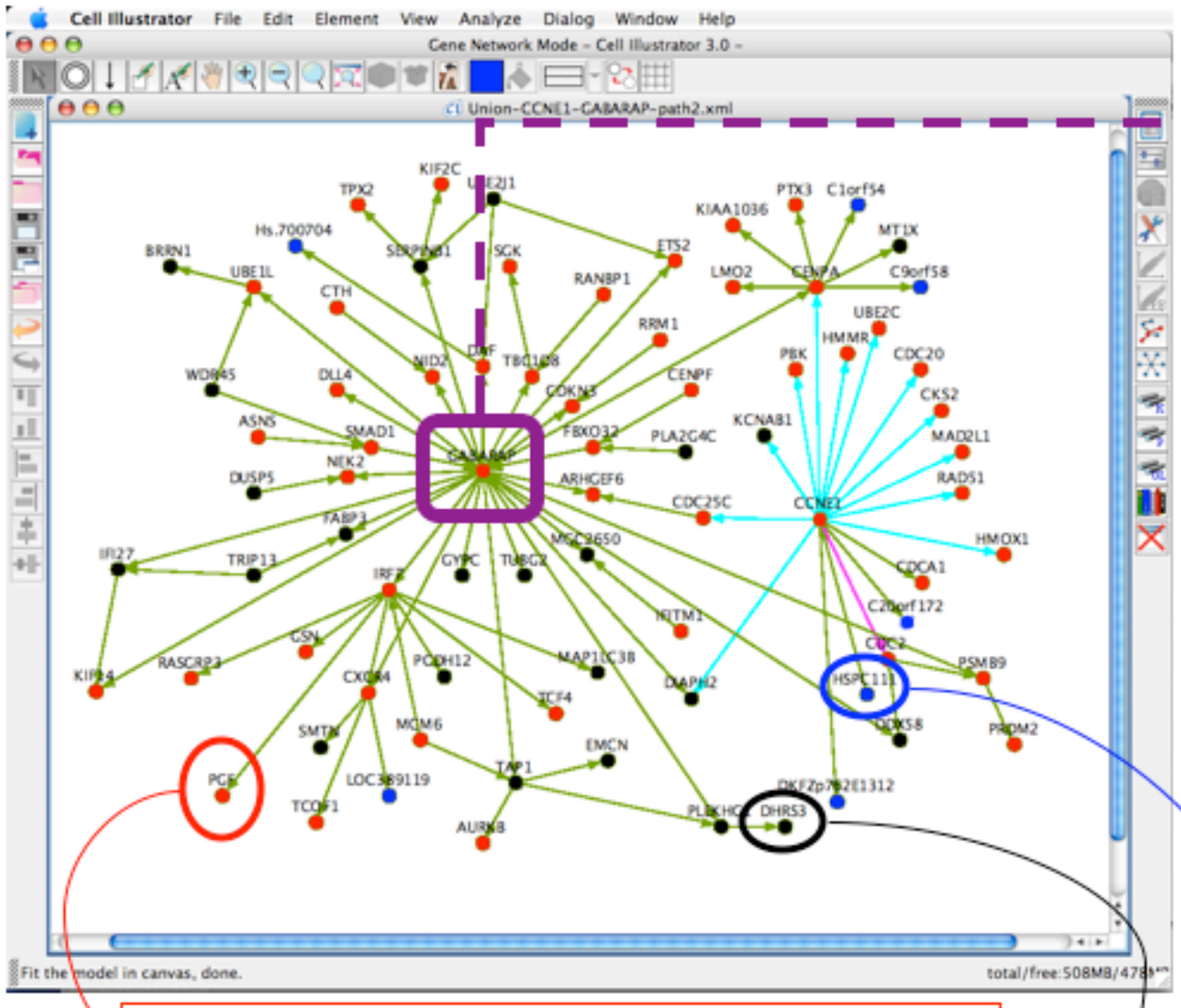
遺伝子ネットワークの利用方法

- 「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？）



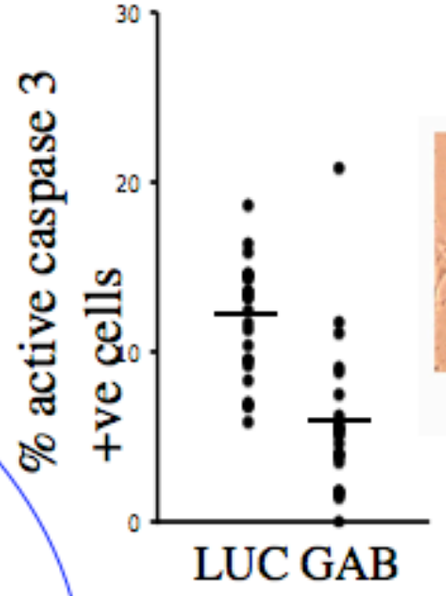
cell innovator

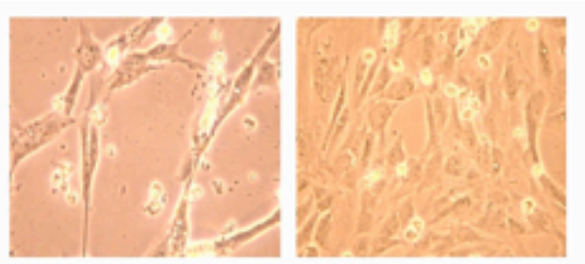
解析事例（ハブをノックダウン）



GABARAPはハブ遺伝子（多数の子供遺伝子をもつ）であり、その下流にはアポトーシス関連遺伝子が多く存在していた。

検証の結果、siRNA GABARAPはアポトーシスを抑制した。





Philos Trans R Soc Lond B Biol Sci (2007)

アポトーシス・細胞増殖との関連が報告されている遺伝子

アポトーシス・細胞増殖との関連が報告されていない遺伝子

機能未知遺伝子

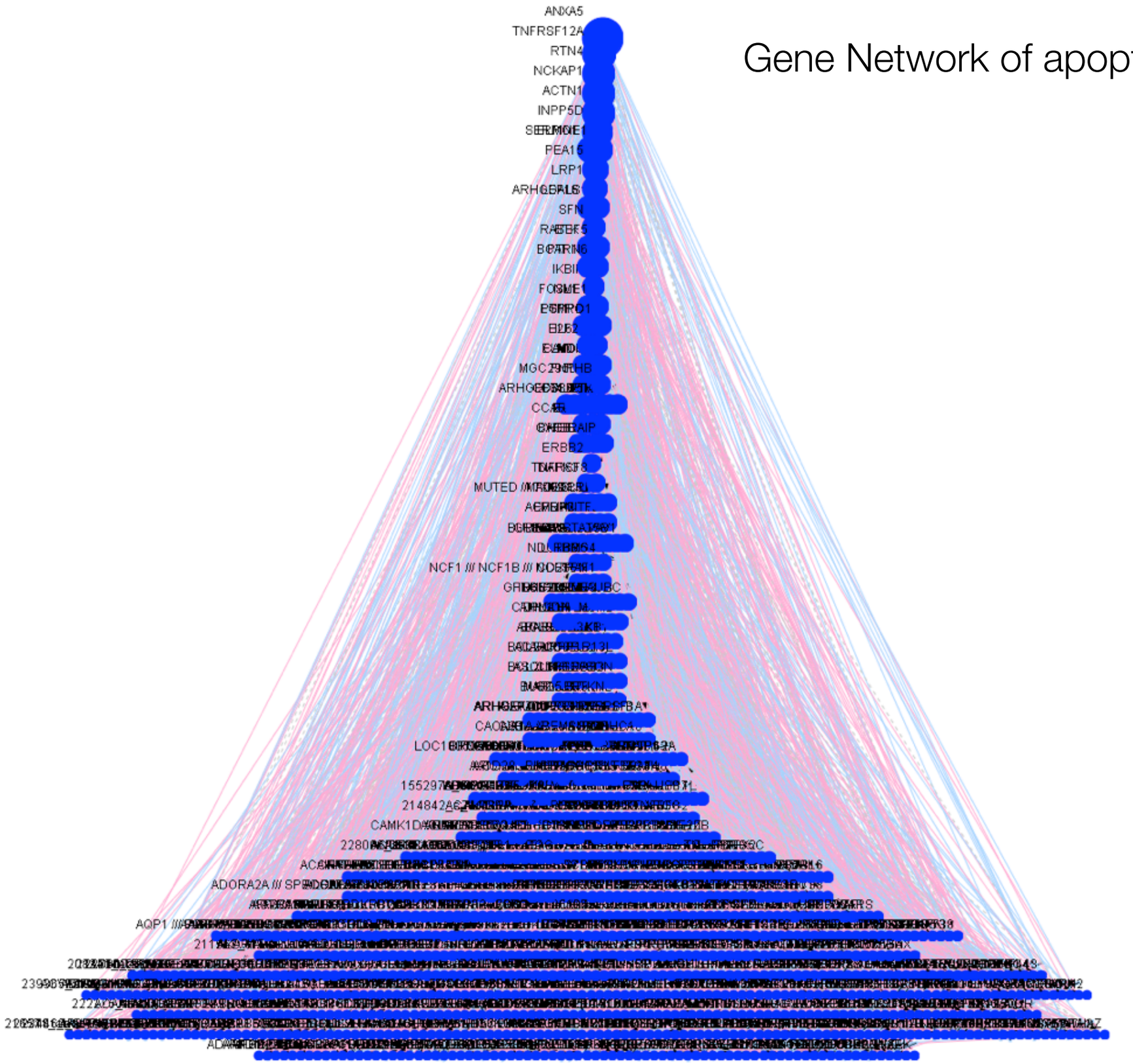
Affara, M., Dunmore, B., Savoie, C., Imoto, S., Tamada, Y., Araki, H., Charnock-Jones, D. S., et al. (2007). Understanding endothelial cell apoptosis: what can the transcriptome, glycome and proteome reveal? Philosophical transactions of the Royal Society of London Series B, Biological sciences, 362(1484),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

よくある質問、疑問

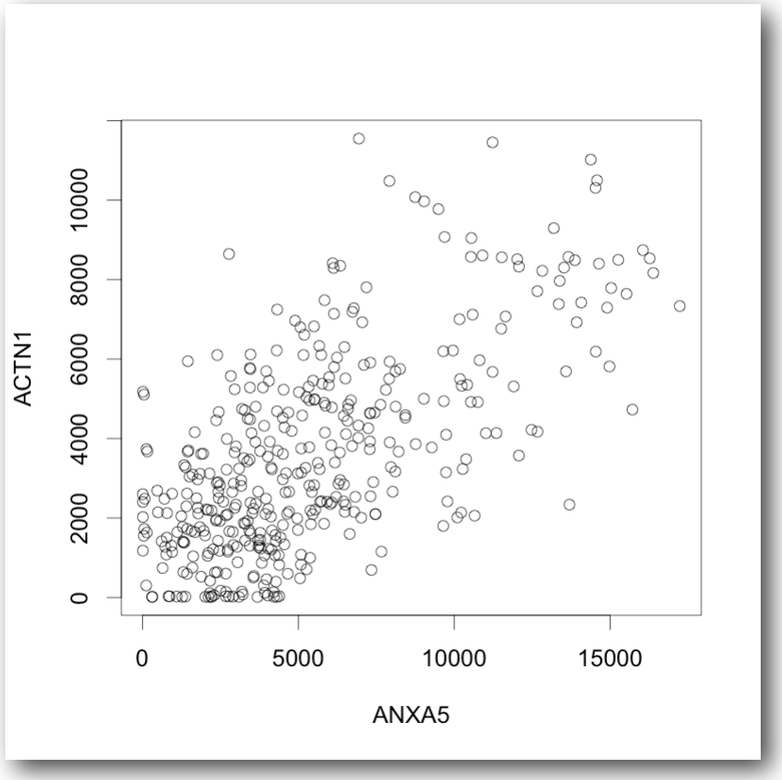
- エッジの何パーセントが当たっているのか？エッジの何パーセントが既知で、何パーセントが未知の情報か？
- シグナル伝達系の活性化される順序は、分からないのか？
- レセプターが、リガンドを活性化しているように見えるが？
- 「ハブ」といっても、ただのキナーゼでは？転写因子でないから、転写は制御できないハズ。

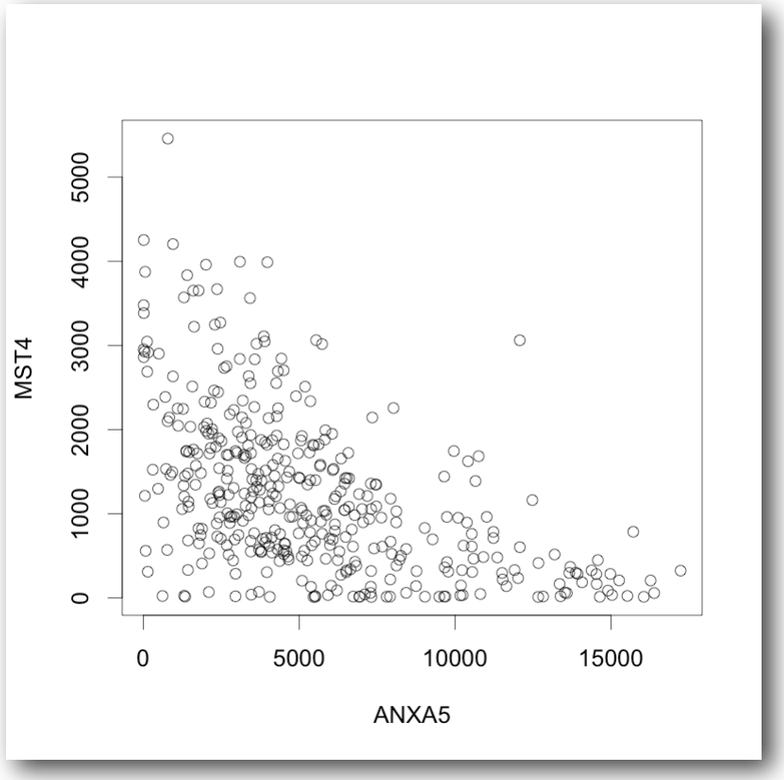
データからはそう見える（バントしないほうがいい）といっているにすぎない。

Gene Network of apoptosis related genes



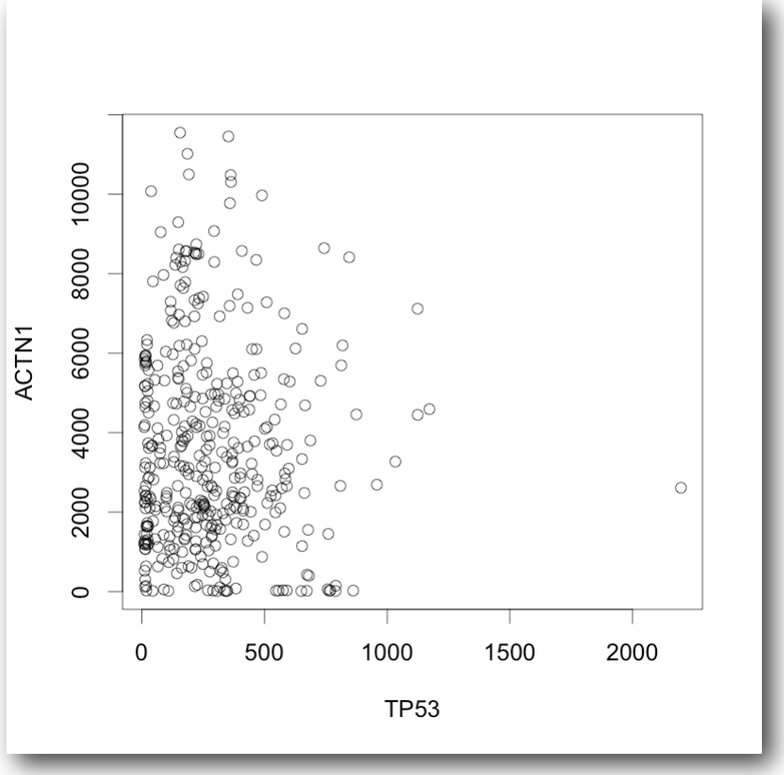
cell innovator

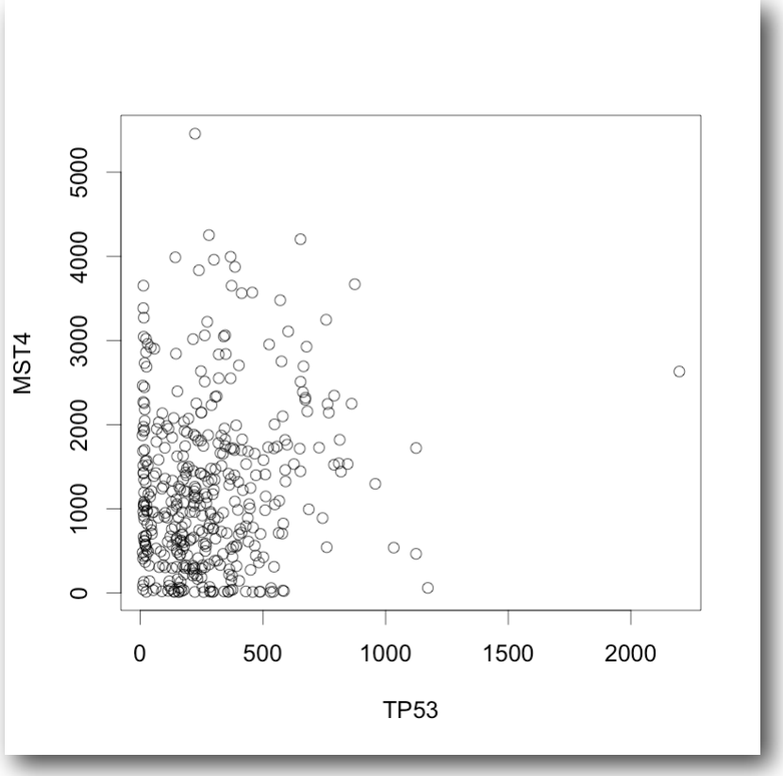




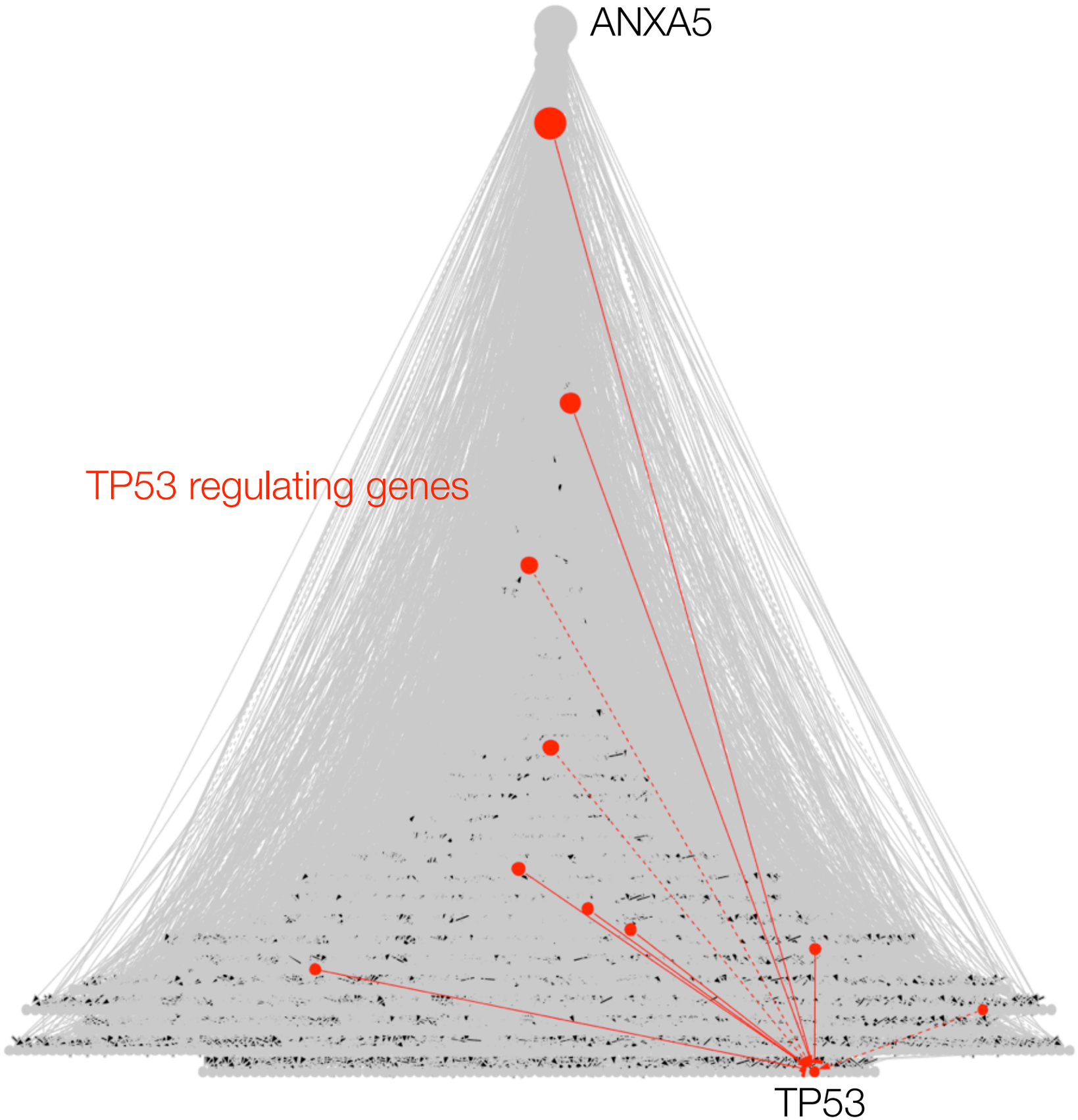
ANXA5 (top gene)

TP53
(bottom gene)





cell innovator



cell innovator

- 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。

- 今年も、レッドソックスが優勝しましたね。。。。