計算生命科学の基礎：
1.3 遺伝子ネットワーク解析

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

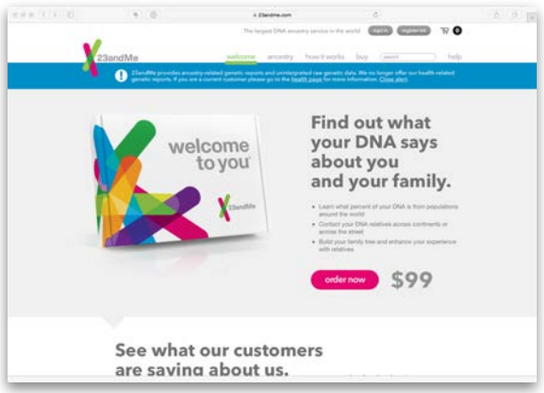
福岡市東区箱崎6-10-1
九州大学 産学連携棟I アントレプレナーシップ・センター 2階
http://www.cell-innovator.com

cell innovator

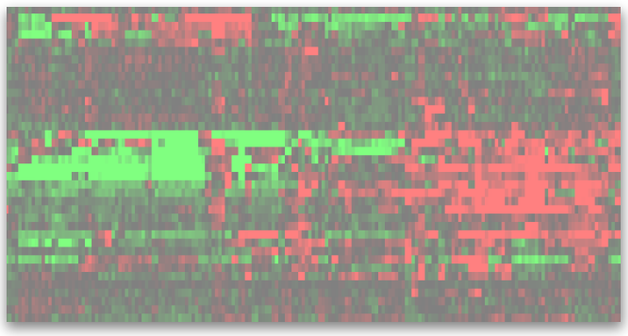
- 前半：遺伝子発現情報の可視化（パスウェイ図、ネットワーク図）
- 後半：（ベイジアンネットワークによる）遺伝子ネットワーク解析

遺伝子発現情報とは？

- 23andMe, GeneQuest, MYCODE 何かと話題の遺伝子診断ビジネス。
  これらが対象としているのは、DNA。



- これに対し、「**遺伝子発現情報**」とは、RNA （またはタンパク）の情報を指す。



DNAとRNA 情報の性質の違い

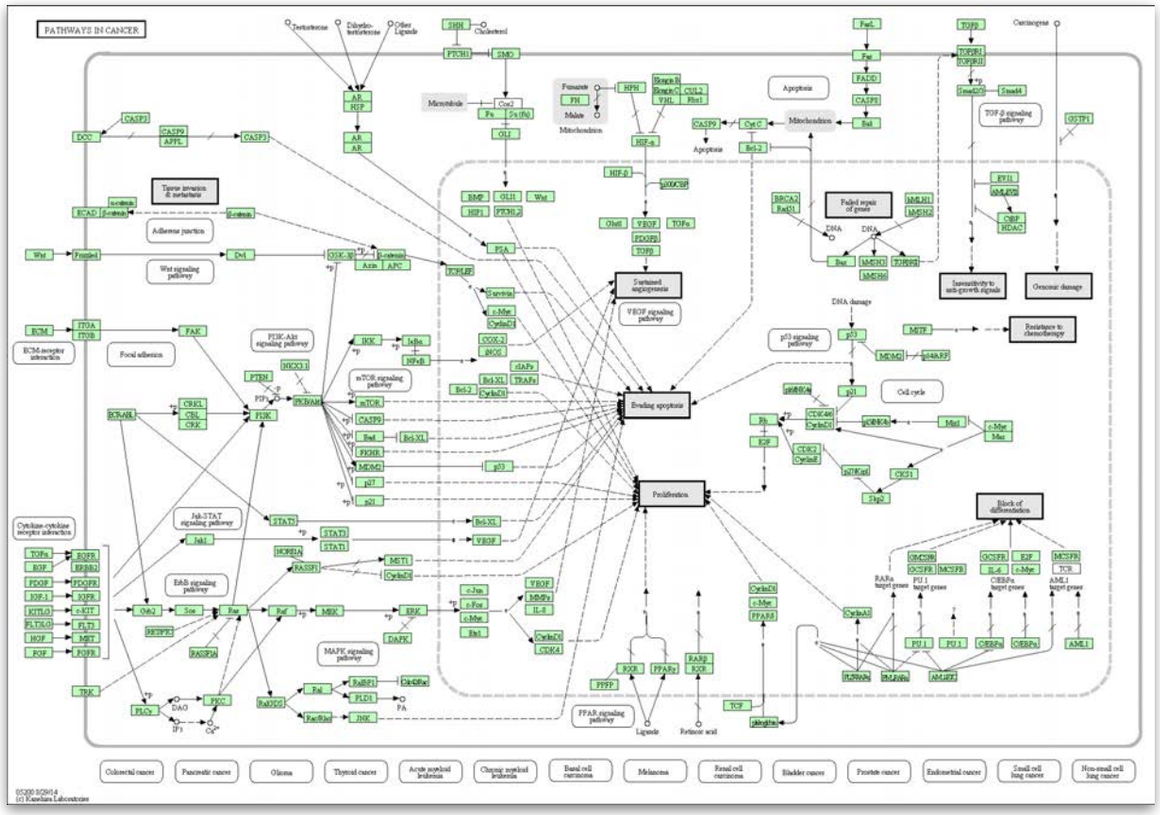
- DNA ＝塩基配列。A, T, G, C からなる文字列。基本的に、特定の遺伝子のアリ、ナシ。または、文字列の異常（変異）のアリ、ナシ。

- RNA or タンパク = 遺伝子の働いている（発現している）度合いの情報。高い、低い、といった連続的な数値情報。どれと一緒に働くのかということも重要。

今日はこちらの話

遺伝子と遺伝子の関係

- 遺伝子は、単独で機能しているわけではなく、その他の遺伝子と相互作用している。
- 伝統的に、関係を矢印で表現。



たくさんの関係の集まり＝パスウェイの図

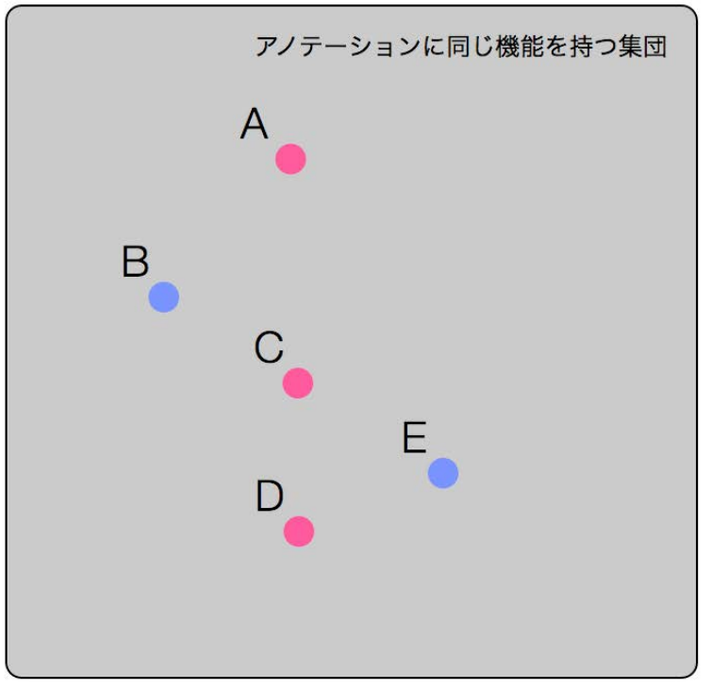
代表的なパスウェイデータベース

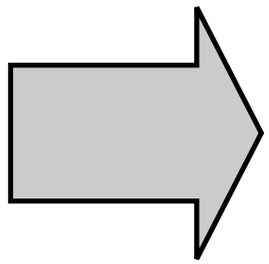
- 無償のデータベース
    - BioCyc (EcoCyc)
    - KEGG
    - Reactome
    - WikiPathways

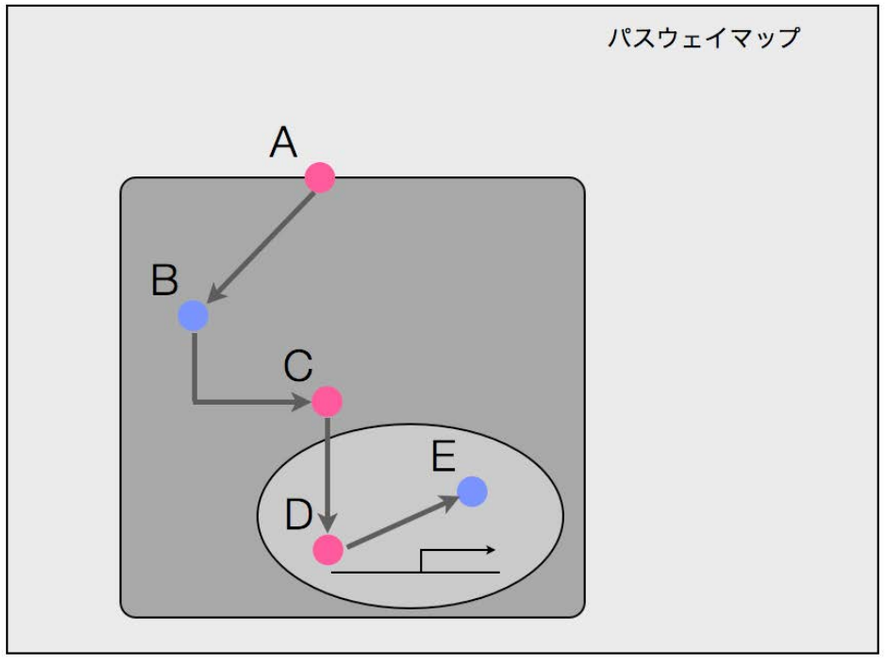
- 有償のデータベース
    - TRANSPATH (BIOBASE)
    - Ingenuity Pathway Analysis (IPA)　　-> キアゲンに買収。
    - NextBio　　-> イルミナに買収。
    - PathwayStudio　　-> Elsevier に買収。

なぜ？可視化（グラフ表現）するのか？

- 文字では分かりづらい。
- 数式では分かりづらい。
- 酵素反応は数式化できるが、まだ、数式化できない知識も多い。
- 全体のイメージを把握したい。







cell innovator

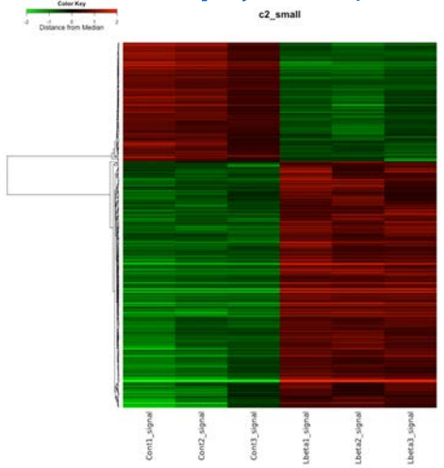
cell innovator

実際の使われ方（パスウェイ解析）

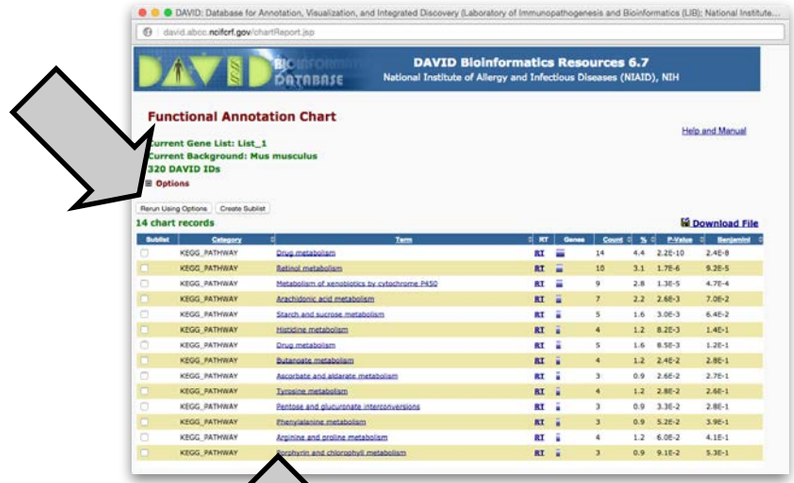
- マイクロアレイ解析などで、遺伝子発現が増減している遺伝子（変動遺伝子）を見つける。
- データベースに問い合わせ。
- 「変動遺伝子が多く見つかるパスウェイ」が分かる。
- どうやら、そのパスウェイに影響があるらしい。。。

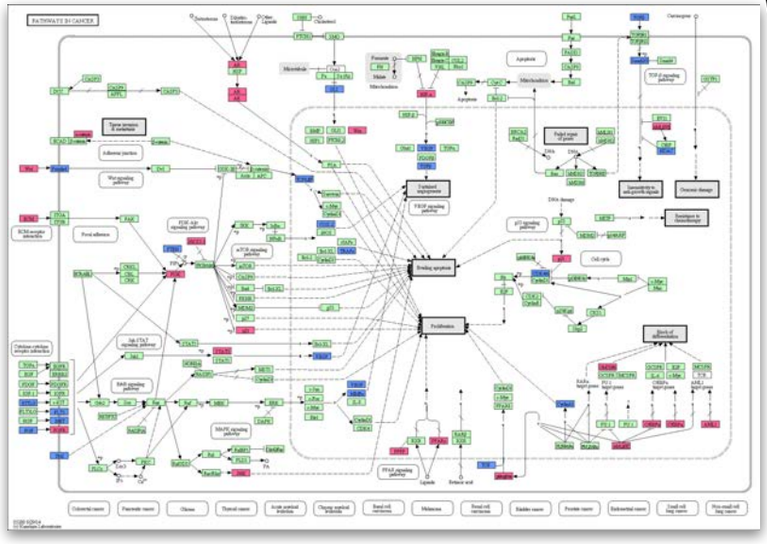
  (Pathways in Cancer, Drug metabolism, ECM receptor interaction …)

マイクロアレイデータ



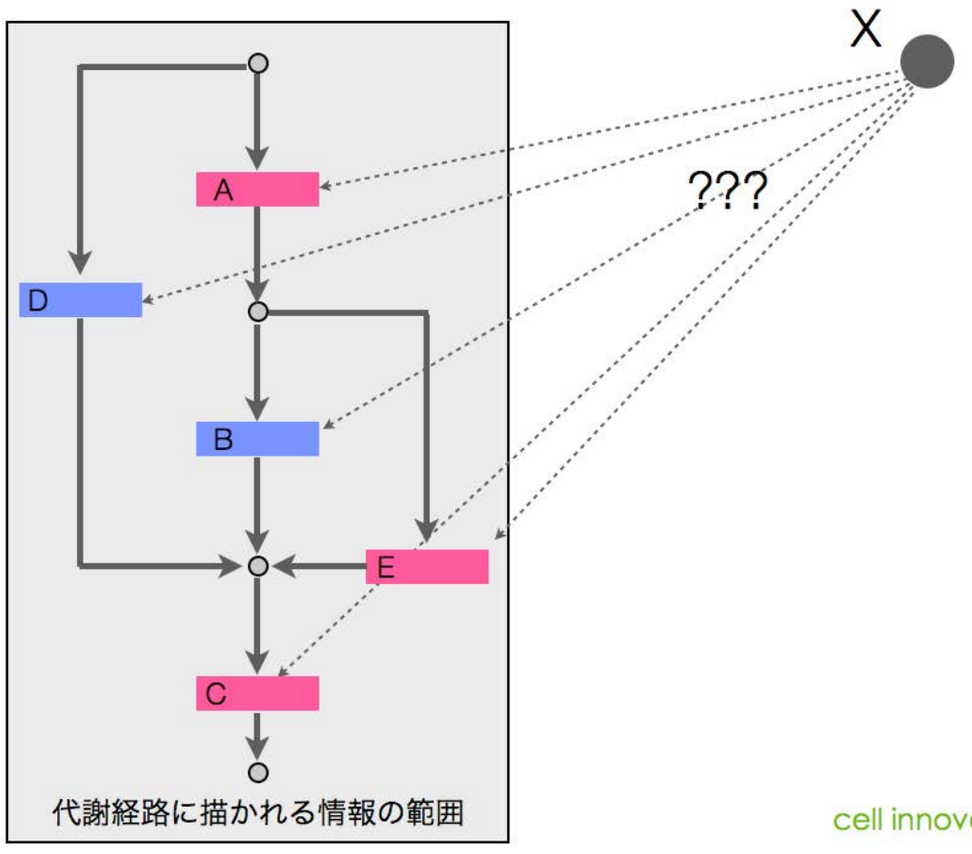
データベース（DAVIDなど）





変動遺伝子の多いパスウェイ

パスウェイやネットワーク情報の制約



- 当然ながら、図上に載ってないものはわからない。
- 新規発見のためには、他の情報を考慮して拡張することになる。

cell innovator

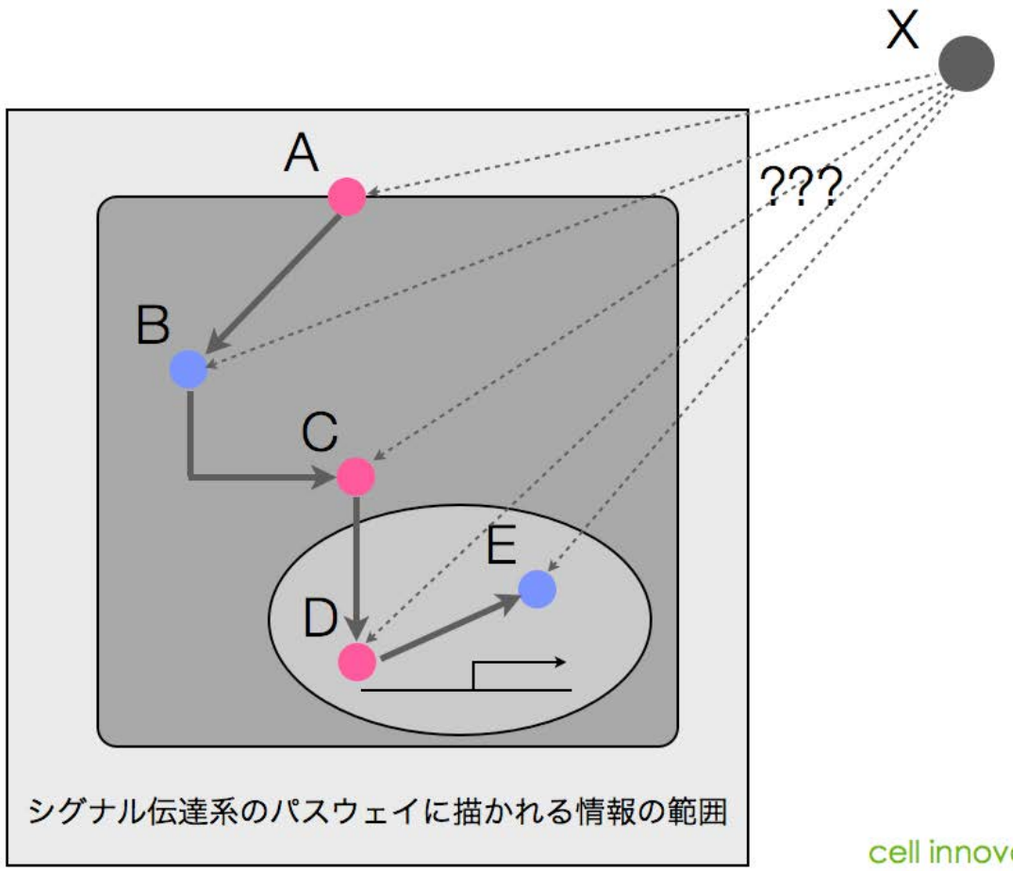
パスウェイ情報の偏り

- 無償のデータベース
  - BioCyc (EcoCyc)：代謝経路
  - KEGG：代謝経路、シグナル伝達系
  - Reactome：シグナル伝達系
  - WikiPathways：シグナル伝達系
- 有償のデータベース
  - TRANSPATH (BIOBASE)：転写制御、シグナル伝達系
  - Ingenuity Pathway Analysis (IPA)：代謝経路、シグナル伝達系、転写制御
  - PathwayStudio：シグナル伝達系、転写制御

拡張できる情報（利用できる情報）

- **タンパク間相互作用 (PPI: protein-protein interaction)**：あるタンパクとあるタンパクが結合するかどうかの情報。EBI や BioGRID データベースから入手可能。比較的、入手が容易。エッジに方向はない。
- **遺伝子発現制御**：転写因子と、その結合部位の情報。IPAやBIOBASE, JASPAR, MSigDB から入手可能。有償のデータベースが使えれば入手は容易。エッジに方向がある。
- **共発現**：ある遺伝子とある遺伝子が、ある状況で共に発現しているかどうかの情報。GeneMANIA や COXPRESdb から入手。エッジに方向はない。
- **文献情報**：論文から、キュレーターまたは自然言語処理によって抽出された情報。KEGG, BIOBASE, IPA, PathwayStudio, GeneSpring から取得可能。エッジに方向がある。

新規性のある発見のためには？



シグナル伝達系のパスウェイに描かれる情報の範囲

- 新たな情報源からデータを取り入れたとしても、ほかのデータベースに登録されているという意味では、既知情報では？
- 新規性のある発見、見つかっていない関係を探すには？

cell innovator

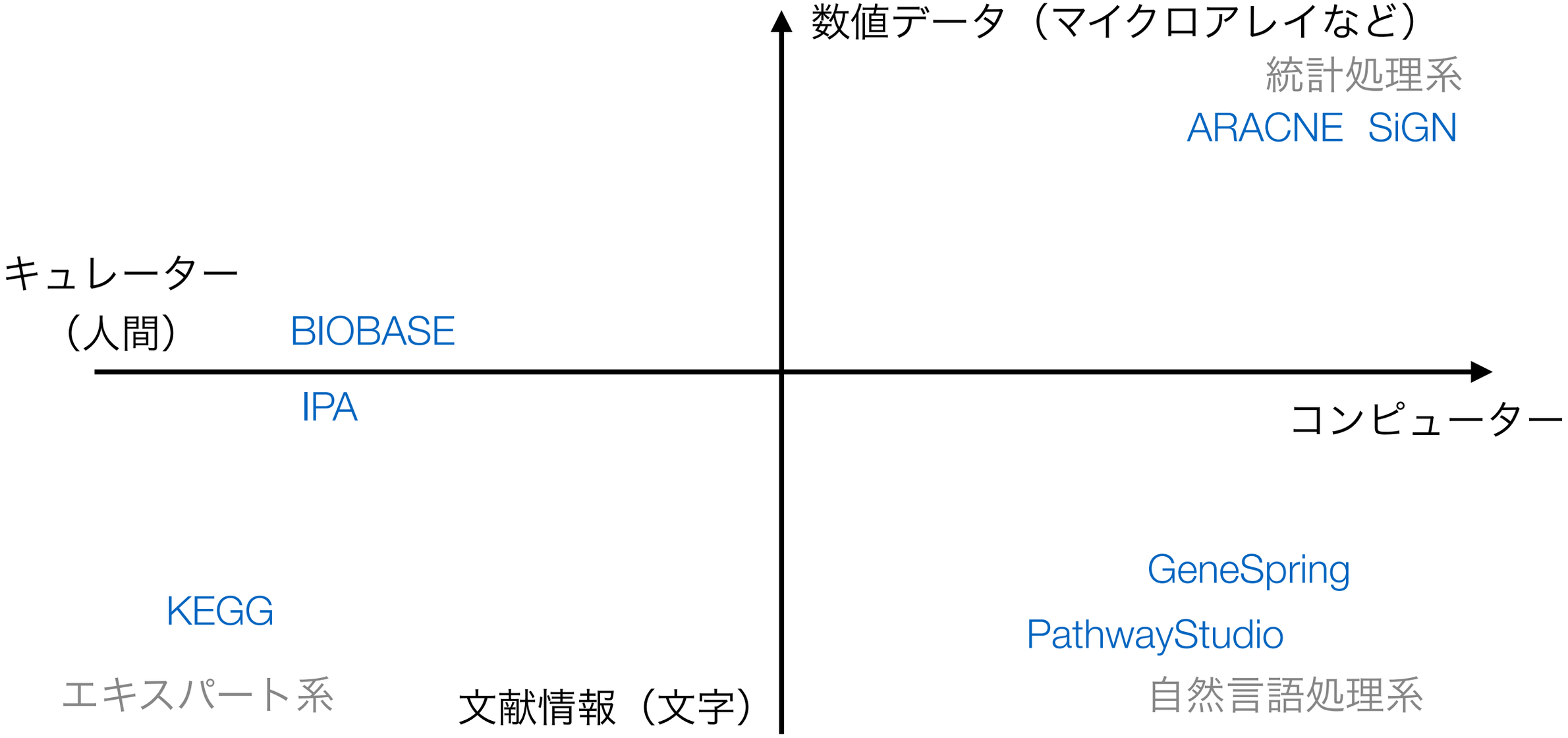
データドリブン、つまり、データのみから関係を推定。

cell innovator

ネットワーク解析（データドリブン）

- 情報源は、おもに、遺伝子発現データ（マイクロアレイ）。
- 情報をもとに、機械的にネットワークを生成する。
- ネットワークを生成するために、さまざまなアルゴリズムが開発されている。
  - ARACNE: Mutual information
  - BANJO: Bayesian inference
  - CLR: Mutual information
  - MIKANA: Dynamic systems
  - SiGN-BN: Bayesian inference
  - その他、相関係数をつなぐもの。

パスウェイ or ネットワーク



- 情報源と作成方法により性質が異なる。

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（ベイジアンネットワークによる）

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/

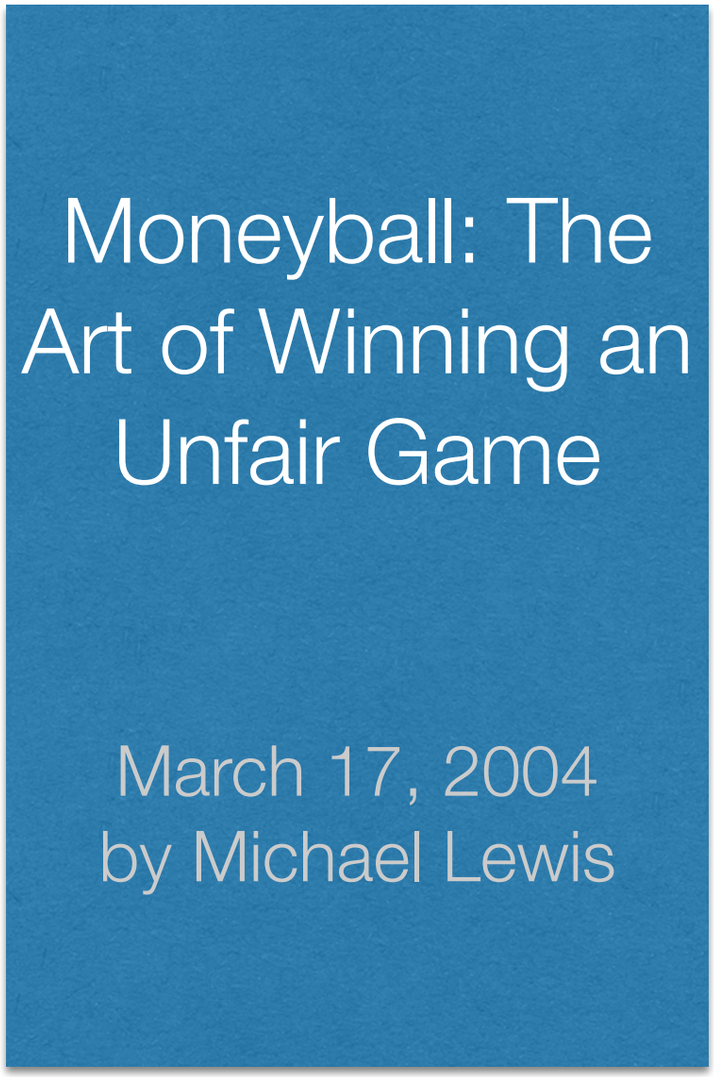
- ビッグデータ：Google、Facebook、Amazon などの企業によるイメージ。

- 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

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

ここまでの話で、、、

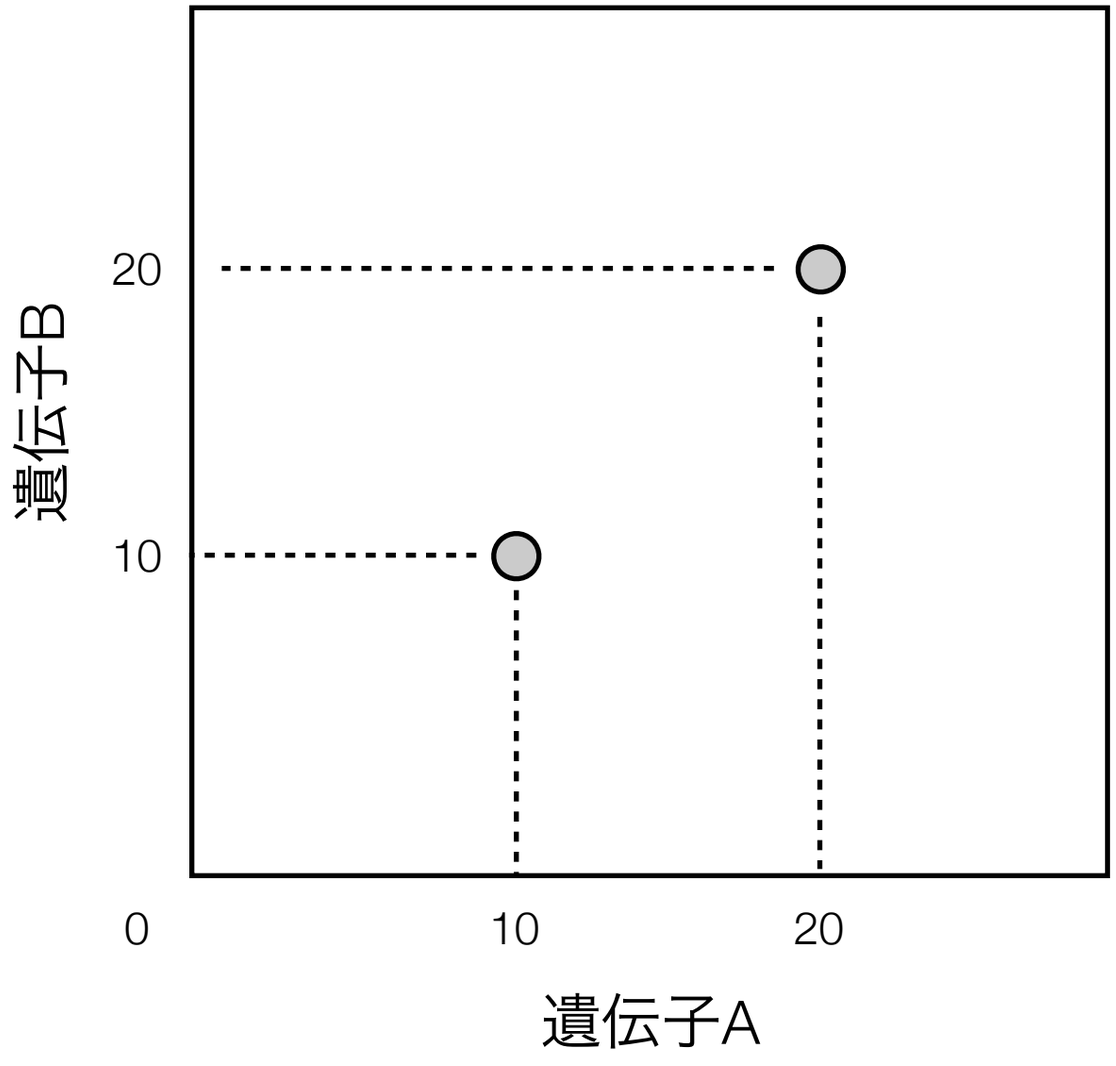
- 近年、統計学的なアプローチが、よく用いられるようになった。

- 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

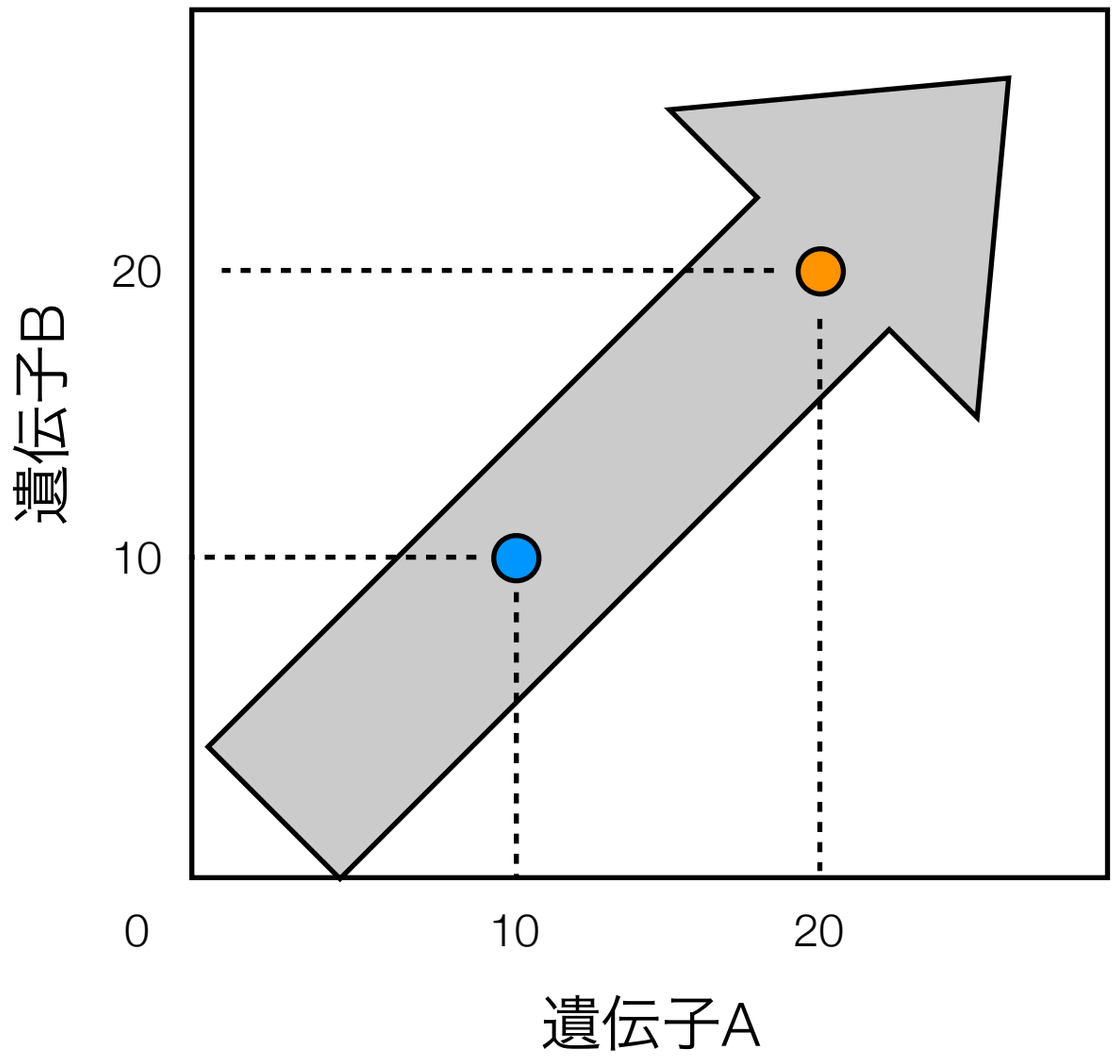
遺伝子発現と散布図



- 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)

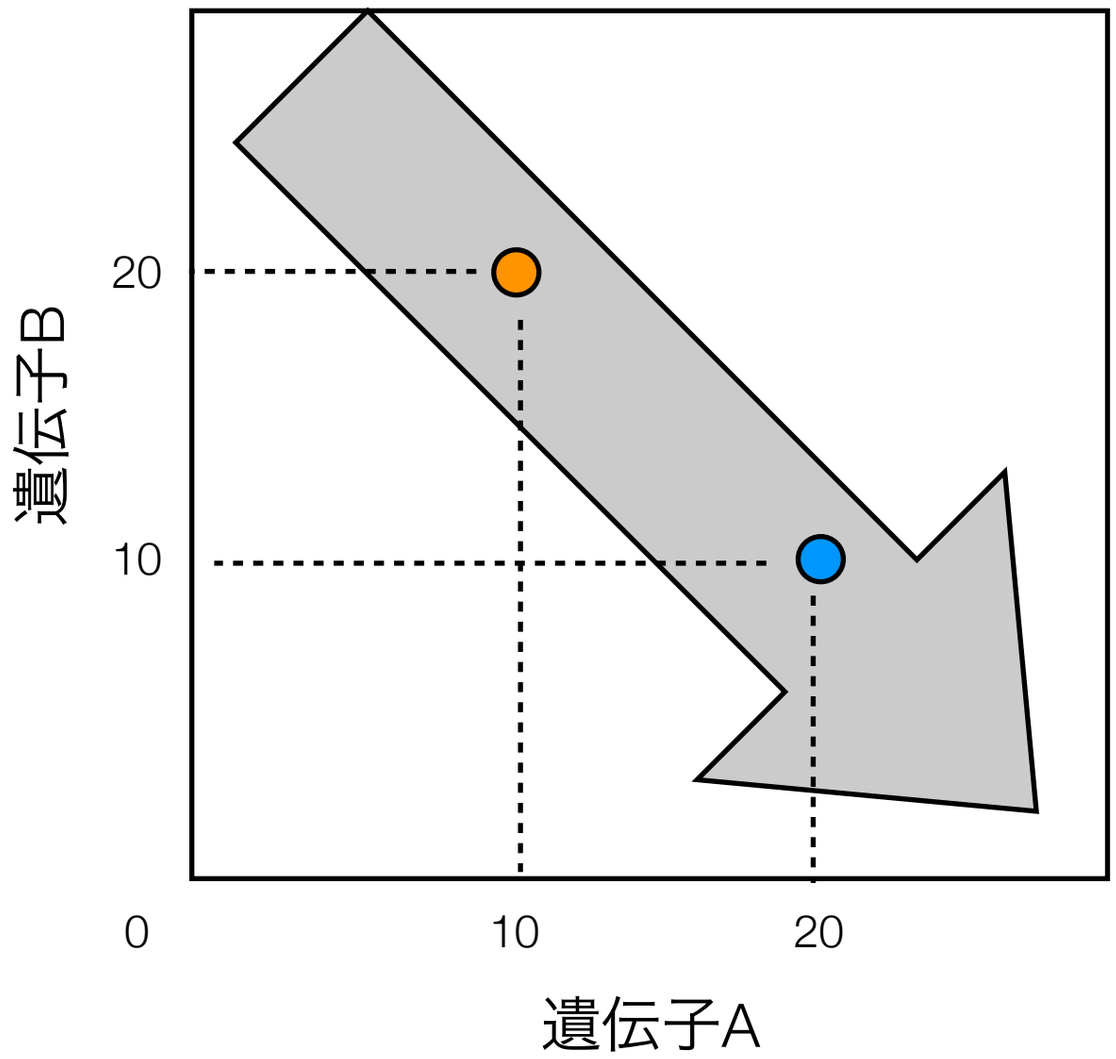


- つまり、遺伝子Aの発現量が低いとき、遺伝子Bの発現量も低い。
- また、遺伝子Aの発現量が高いとき、遺伝子Bの発現量も高い。

- 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)

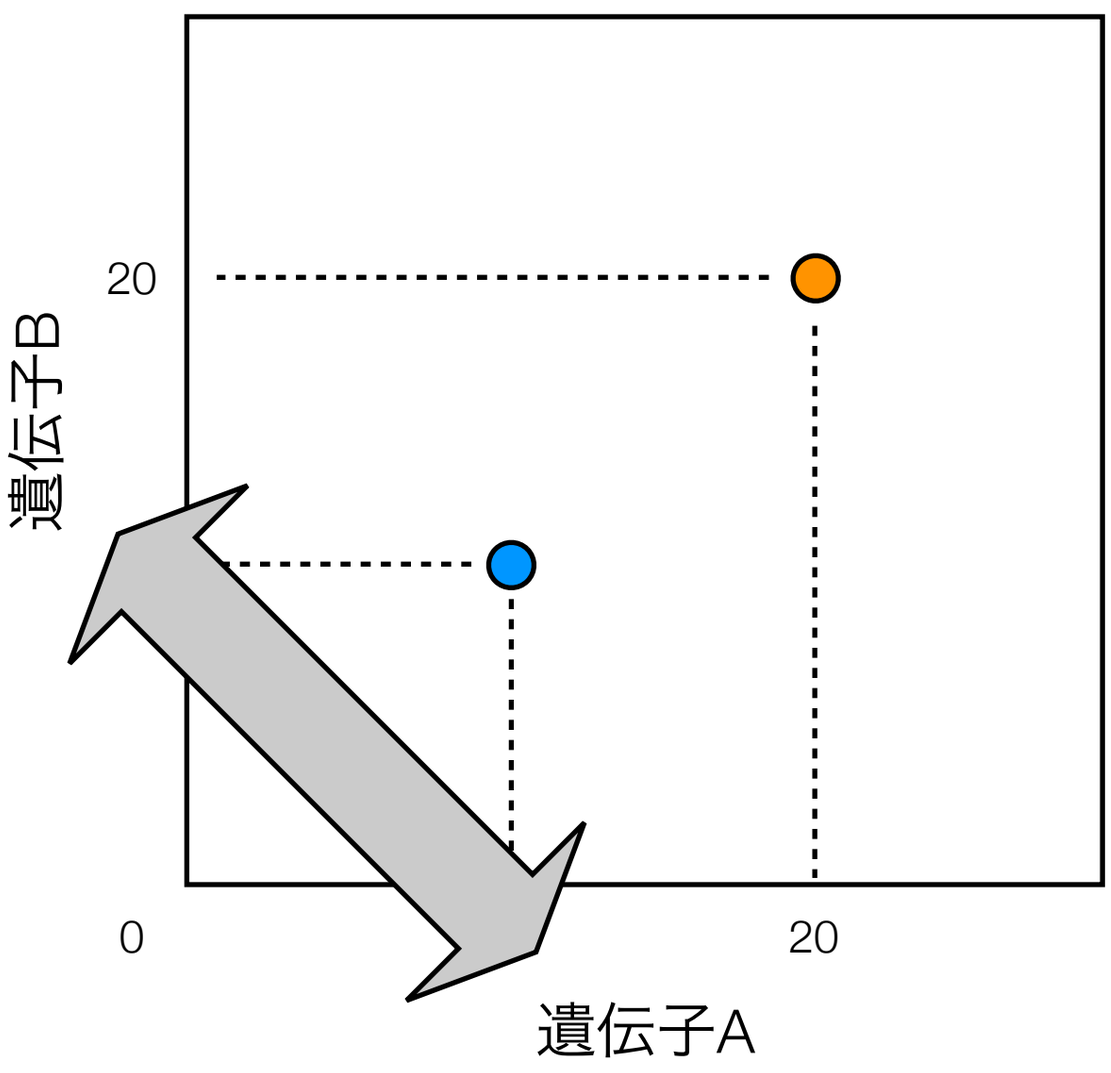


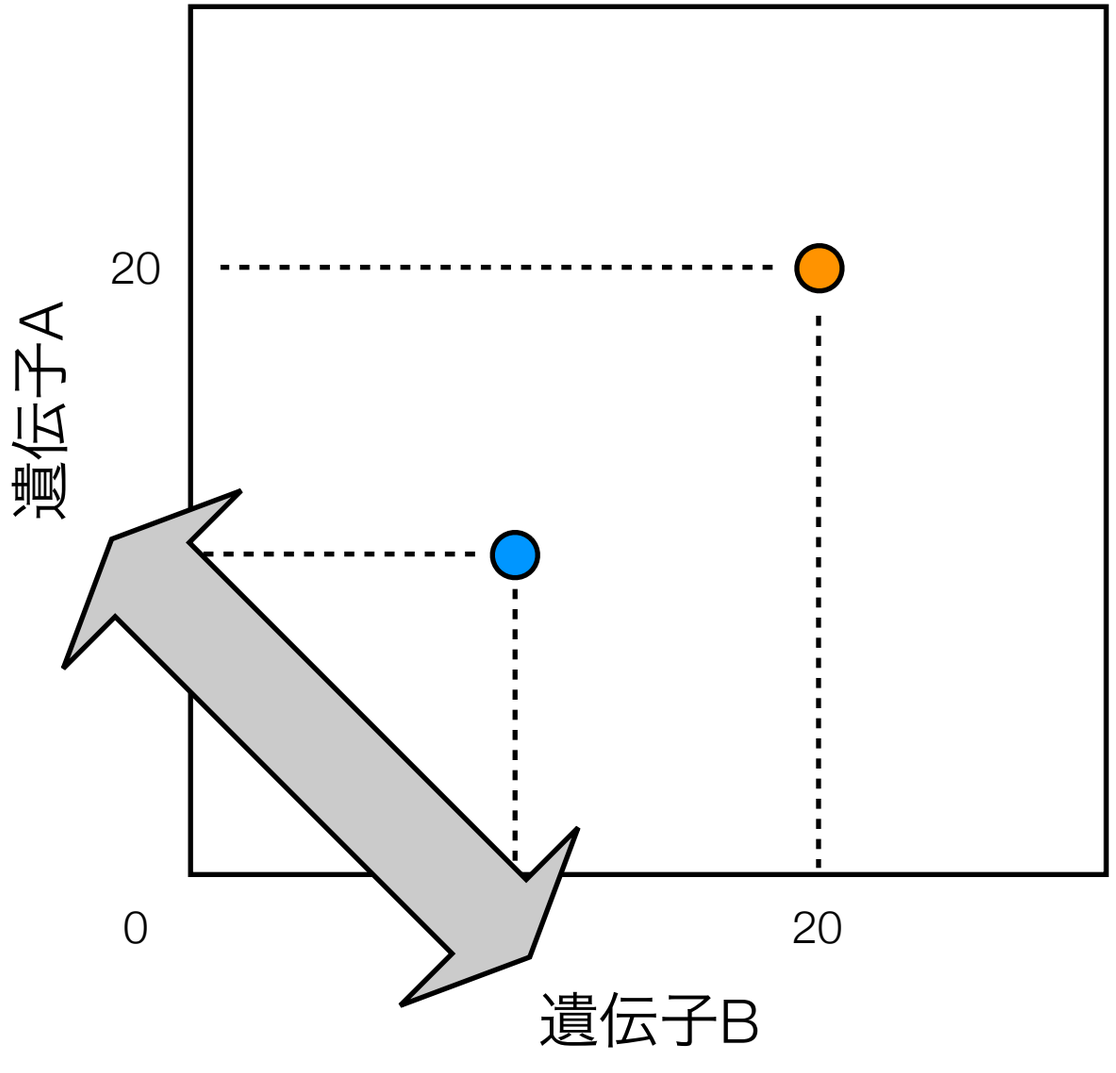
- その逆なら、遺伝子Aの発現量が低いとき、遺伝子Bの発現量は高い。
- また、遺伝子Aの発現量が高いとき、遺伝子Bの発現量は低い。

- 遺伝子AとBの発現量には、負の相関が見られる。

A ⟞ B

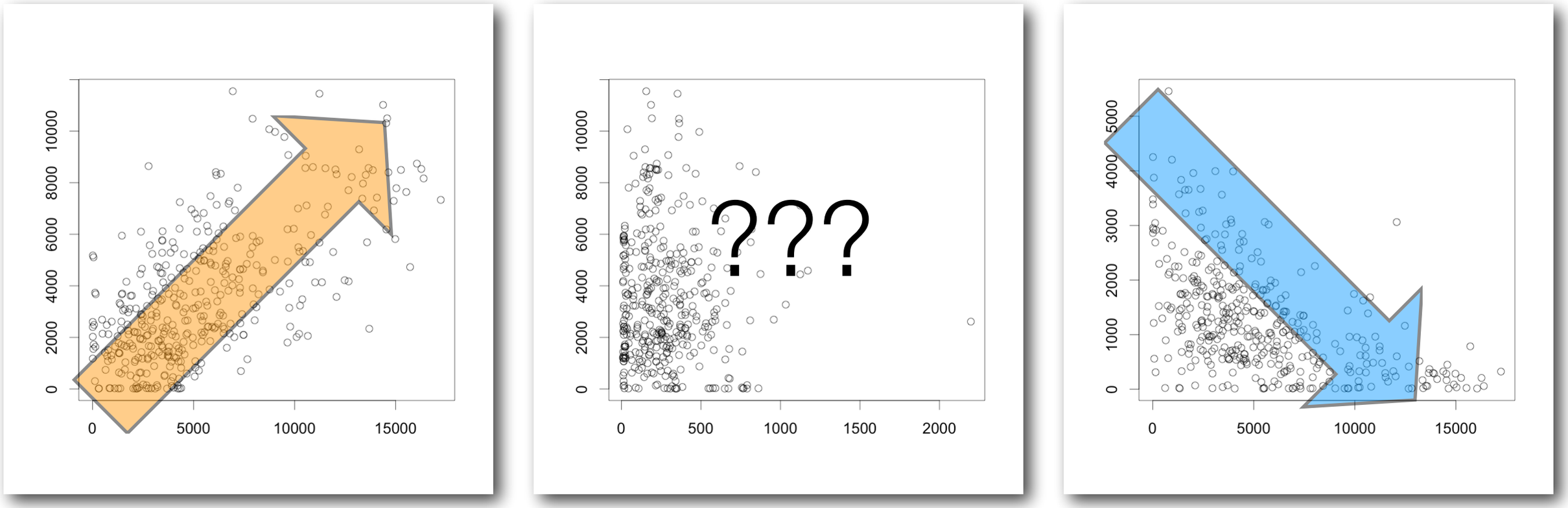
どちらが上流？





- 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？？

データを増やしていくと見えてくるもの

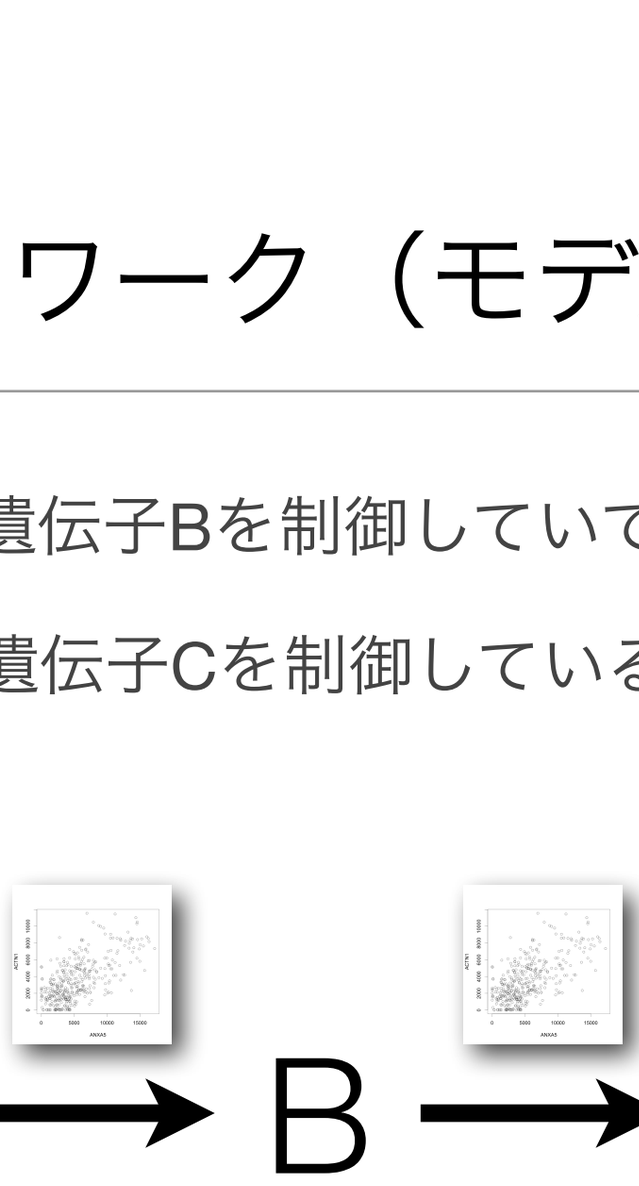


- 上記は、400サンプル＝400個の点における関係を見たもの。
- サンプル数を増やしていくと、「関係の度合い」（＝確率）も見えそう。

遺伝子発現にも統計学的なアプローチを。

ベイジアンネットワーク（モデル）

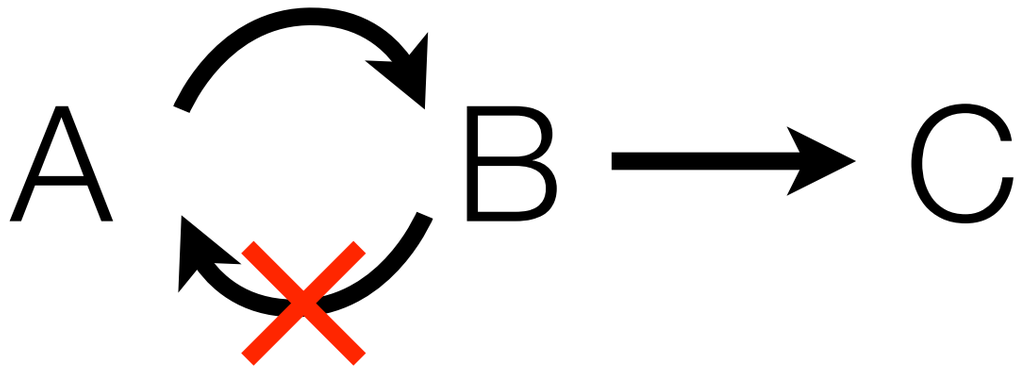
- 遺伝子Aが、ある確率で、遺伝子Bを制御していて、
- 遺伝子Bが、ある確率で、遺伝子Cを制御している。



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

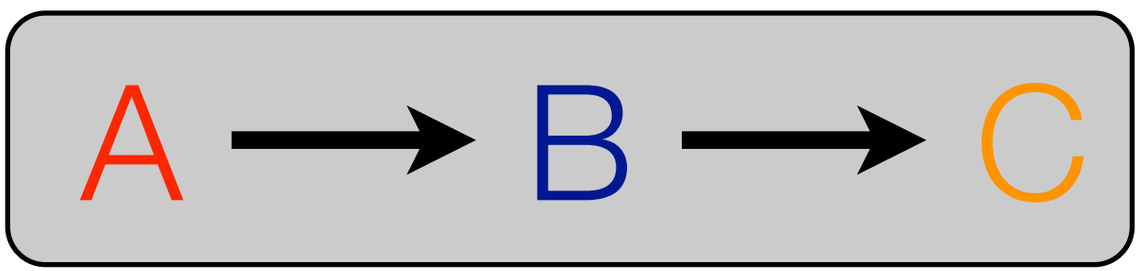
どちらか？

A ← B → C

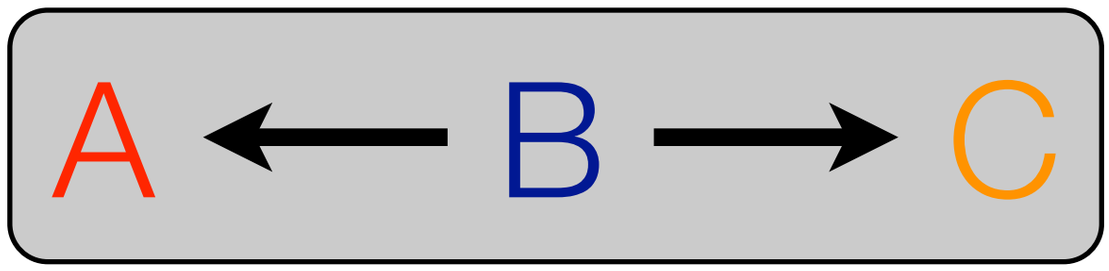
ベイジアンネットワーク（モデル）

- 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）

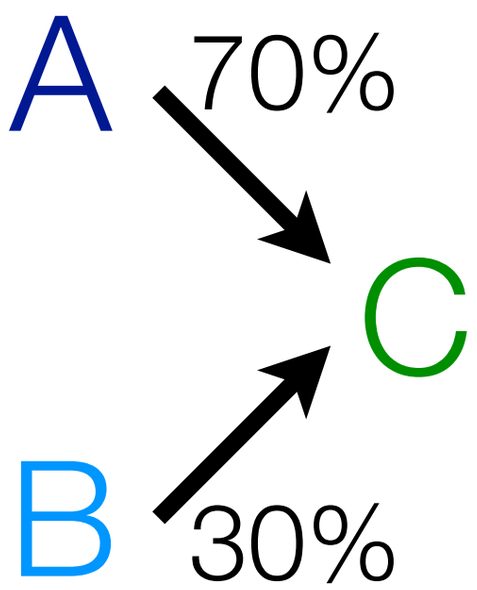


原因はどちら？



例えば、雨とスプリンクラーと芝生の関係は？

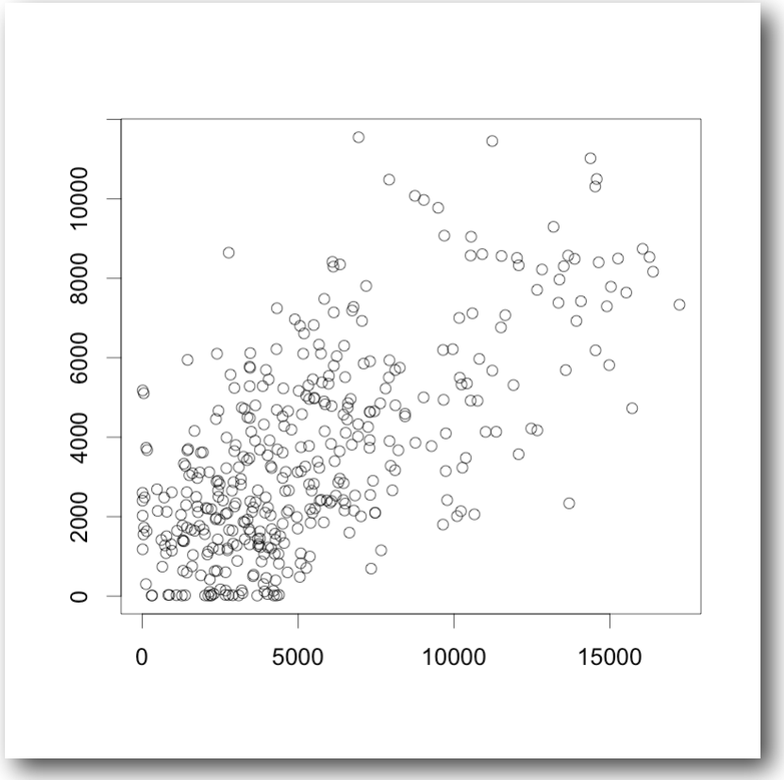
- A: 雨が降る（降雨量）。
- B: スプリンクラーが作動する。
- C: 芝生が濡れる。



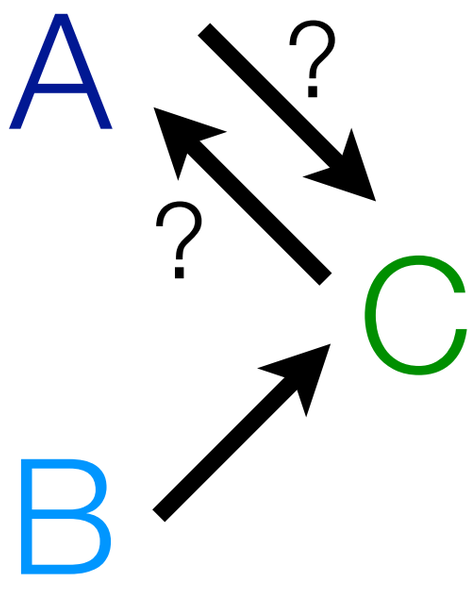
- 芝生が濡れるのは、雨が降ったか、**また**は、スプリンクラーが作動したから。

芝生が濡れたら、雨が降る？

濡れた芝生の面積



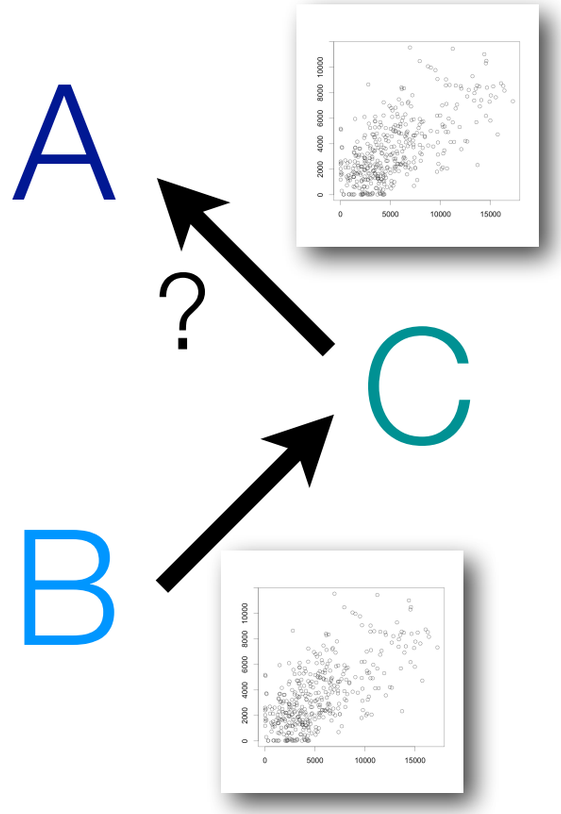
降雨量



- 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。
  - http://www.scls.riken.jp/scruise/software/sign-bn.html

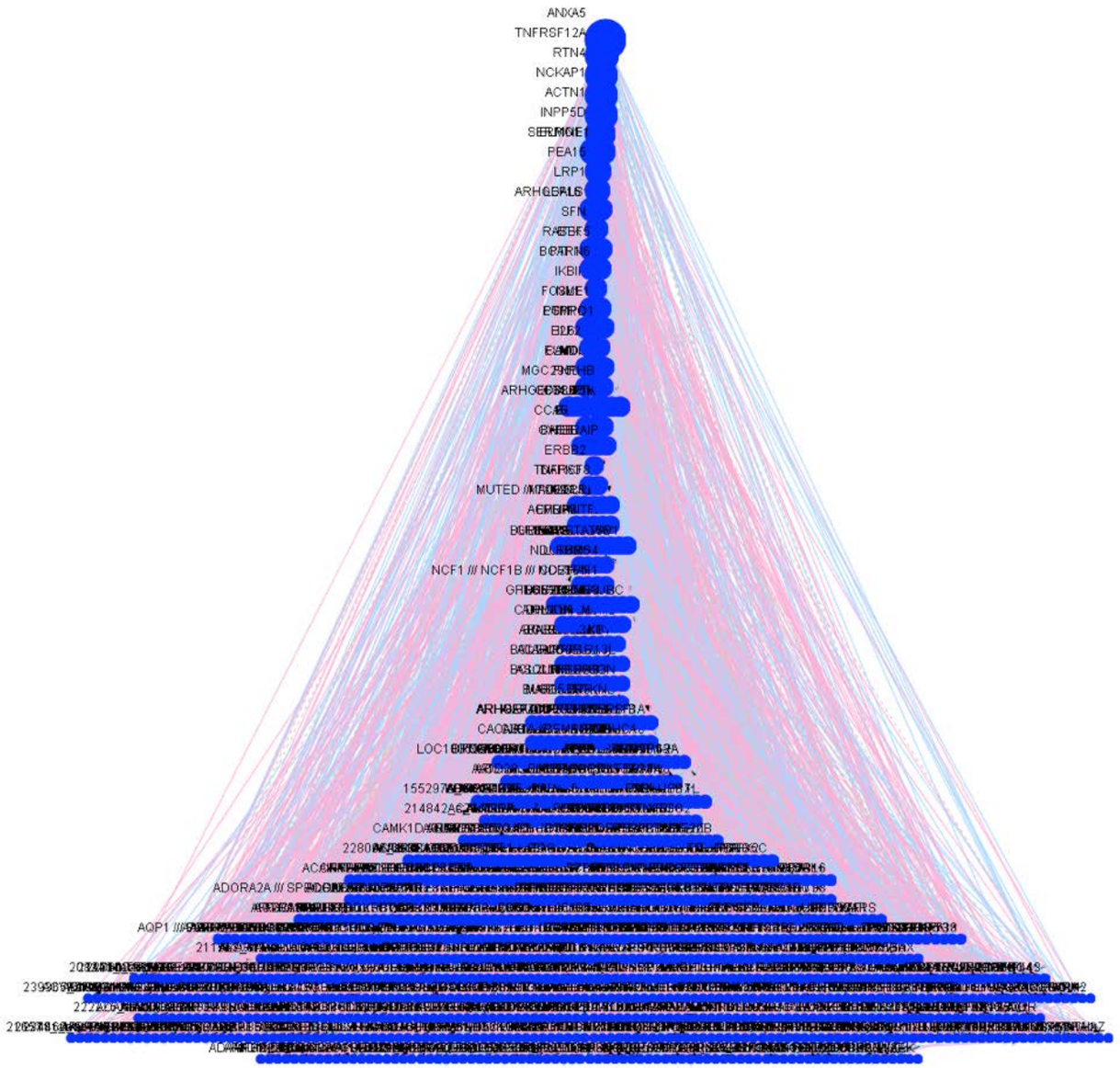
イメージ的には、とにかく総当たりで、
すべてのネットワークのパターンをチェックして、
もっともらしいネットワークの状態を推定

3. 遺伝子ネットワーク（ベイジアンネットワークによる）

遺伝子ネットワーク（ベイジアンネットワークによる）

- 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。

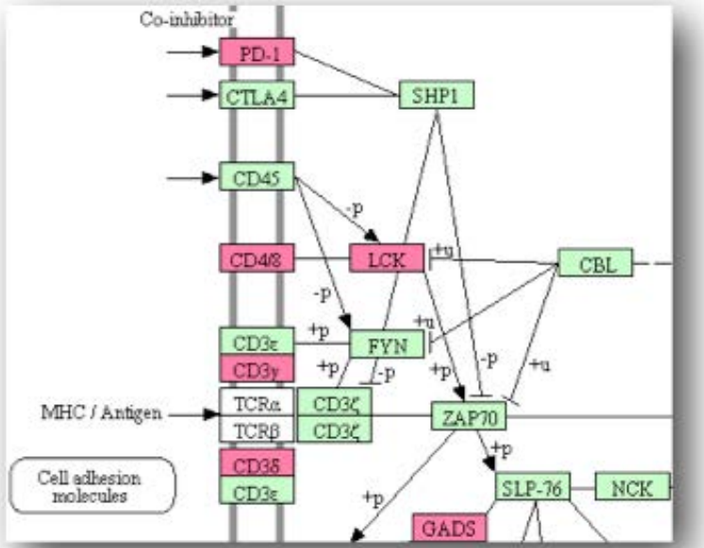


遺伝子ネットワークの意味するもの

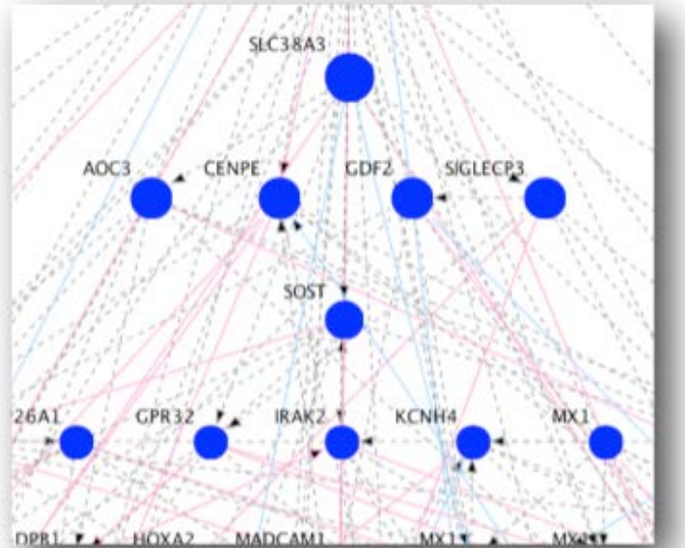
- 遺伝子ネットワークは、いわゆる「パスウェイ」**ではない。**
- いわゆる「パスウェイ」は、下記の情報のいずれか。
  - タンパク間相互作用 = Protein-Protein Interaction (PPI) network。
  - 遺伝子発現制御 = 転写因子と、その転写制御領域を持つ遺伝子の関係。
  - 共発現 = ともに発現している遺伝子の関係。
  - 文献情報 = 文献に、「制御関係あり」と報告された関係。

- 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。

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）

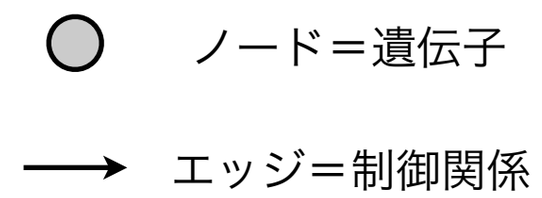
現在では、推定時の問題を回避可能

- 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]
  - The Cancer Genome Atlas (TCGA) のデータも利用可能。

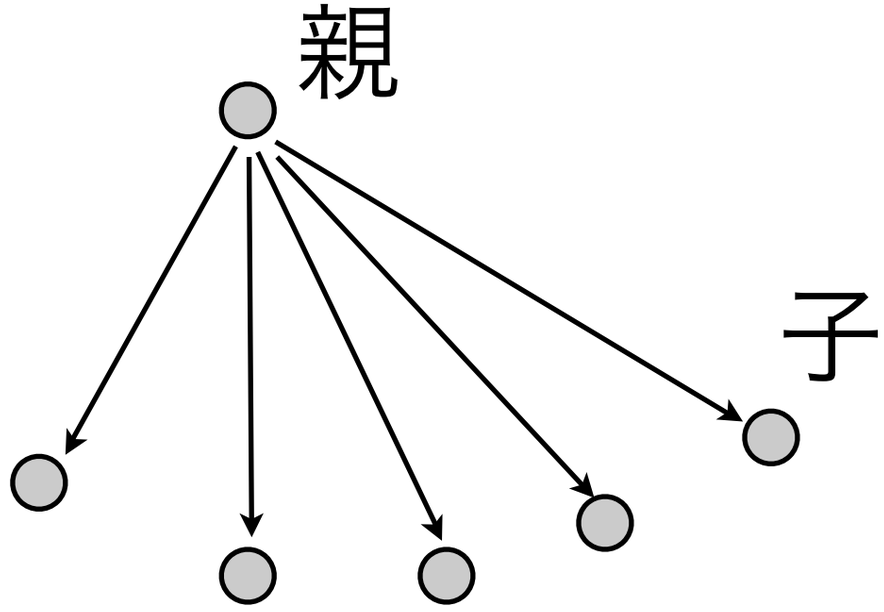
- 計算には、「東大医科研ヒトゲノム解析センター」、「京（SCLS）」などのスーパーコンピューターを利用。 **--> 計算機環境の問題をクリア。**

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**」

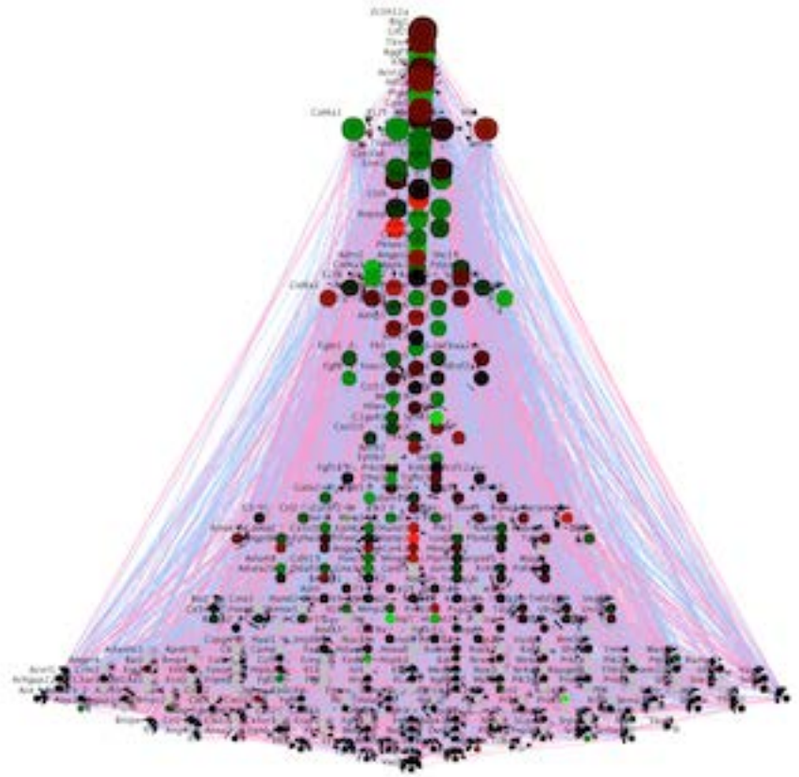


- ネットワークの構造としては、一部の親に多数の子が集中するという構造になることが多い。（スケールフリー）
- 特に「子が多いノード」は、「ハブ」と呼ばれる。

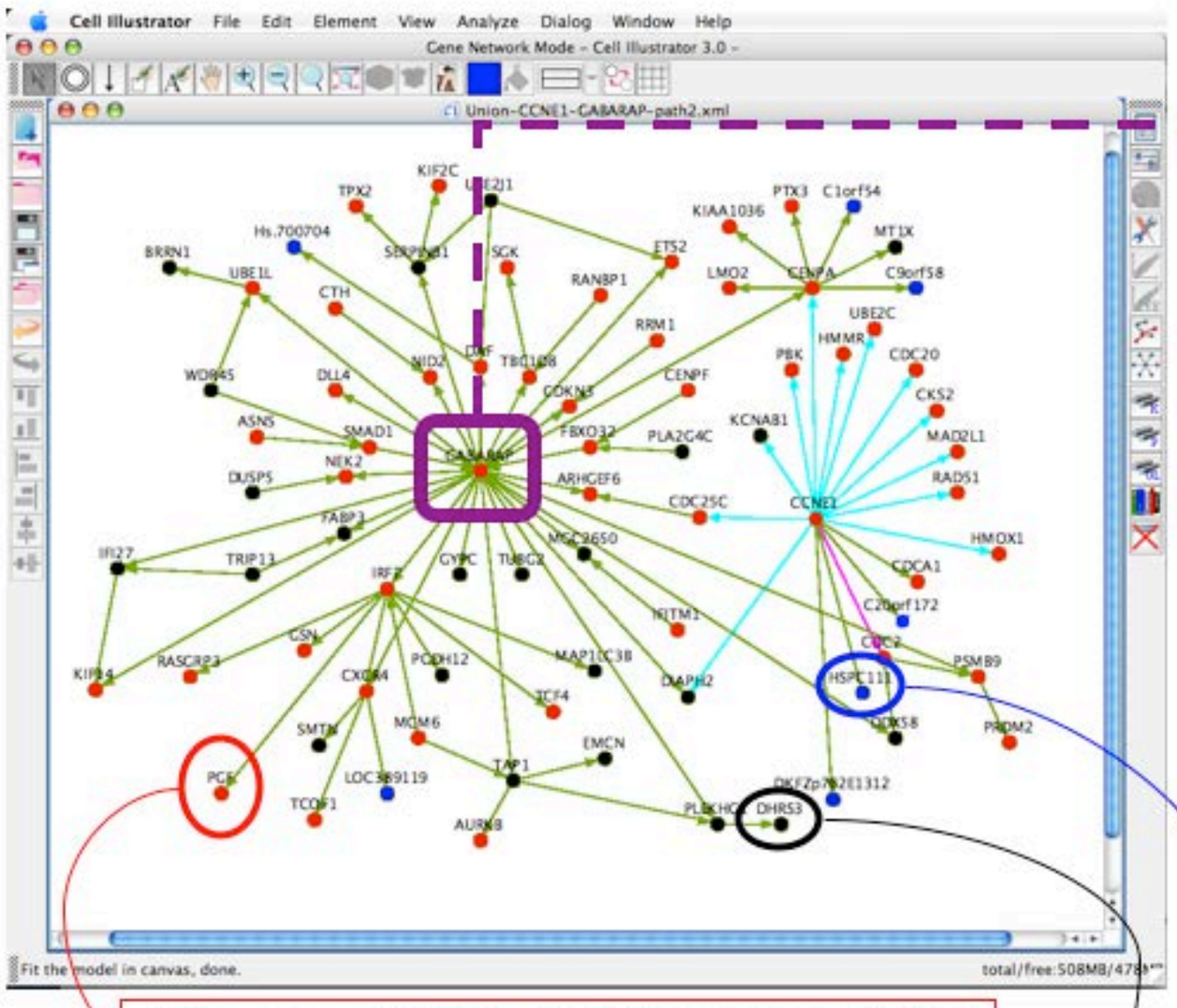


遺伝子ネットワークの利用方法

- 「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）



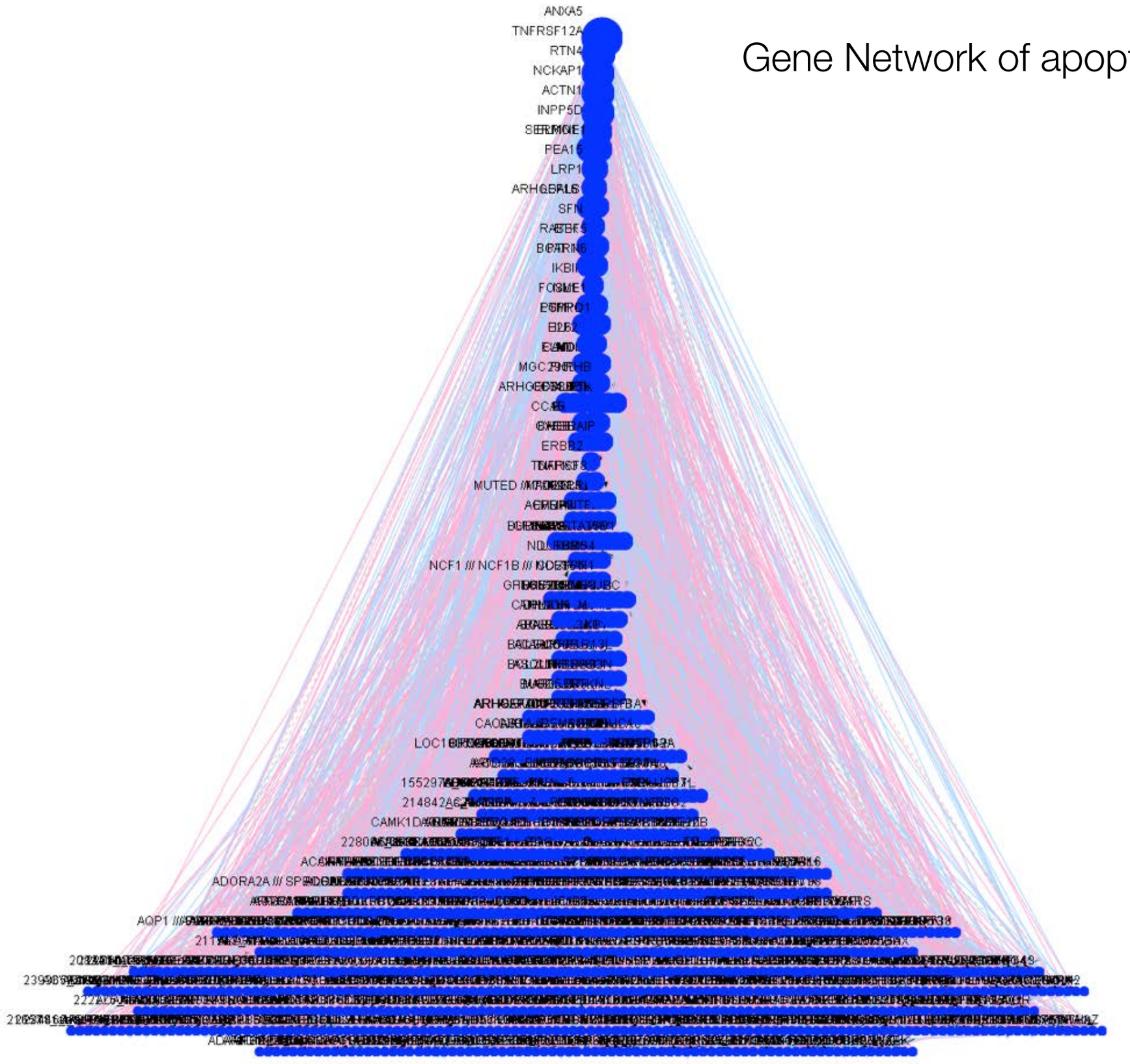
Affara, M., Dunmore, B., Savoie, C., Imoto, S., Tamada, Y., Araki, H., Charnock-Jones, D. S., et al. (2007). Understanding endothelial cell apoptosis: what can the transcriptome, glycome and proteome reveal? Philosophical transactions of the Royal Society of London Series B, Biological sciences, 362(1484),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

よくある質問、疑問

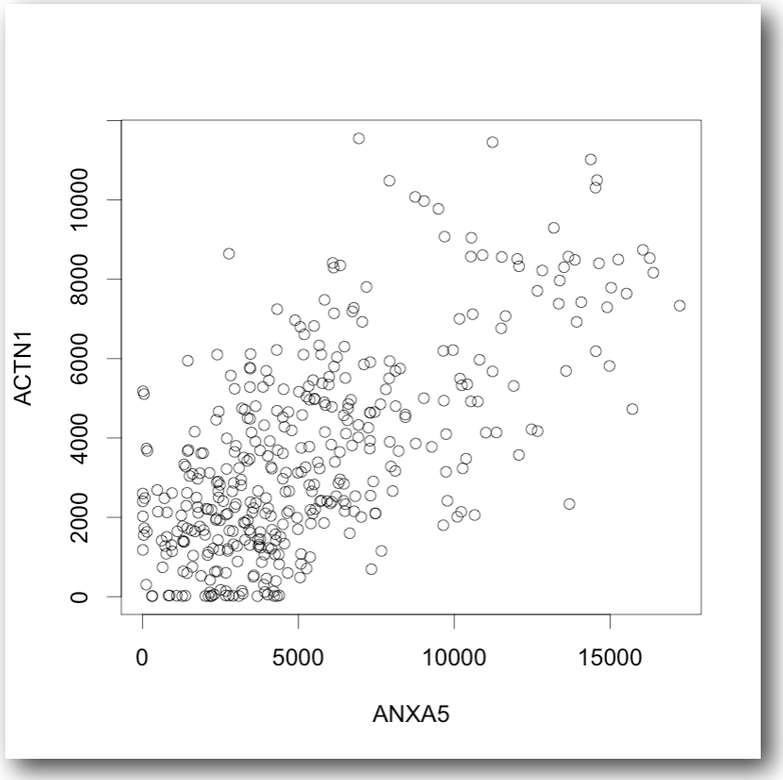
- エッジの何パーセントが当たっているのか？エッジの何パーセントが既知で、何パーセントが未知の情報か？
- シグナル伝達系の活性化される順序は、分からないのか？
- レセプターが、リガンドを活性化しているように見えるが？
- 「ハブ」といっても、ただのキナーゼでは？転写因子でないから、転写は制御できないハズ。

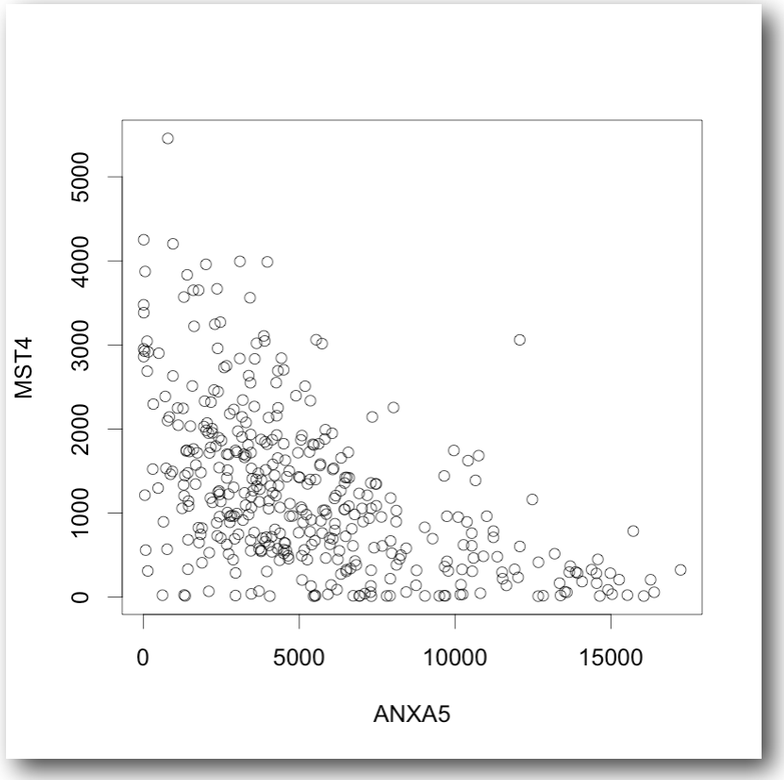
データからはそう見える（バントしないほうがいい）といっているにすぎない。

Gene Network of apoptosis related genes



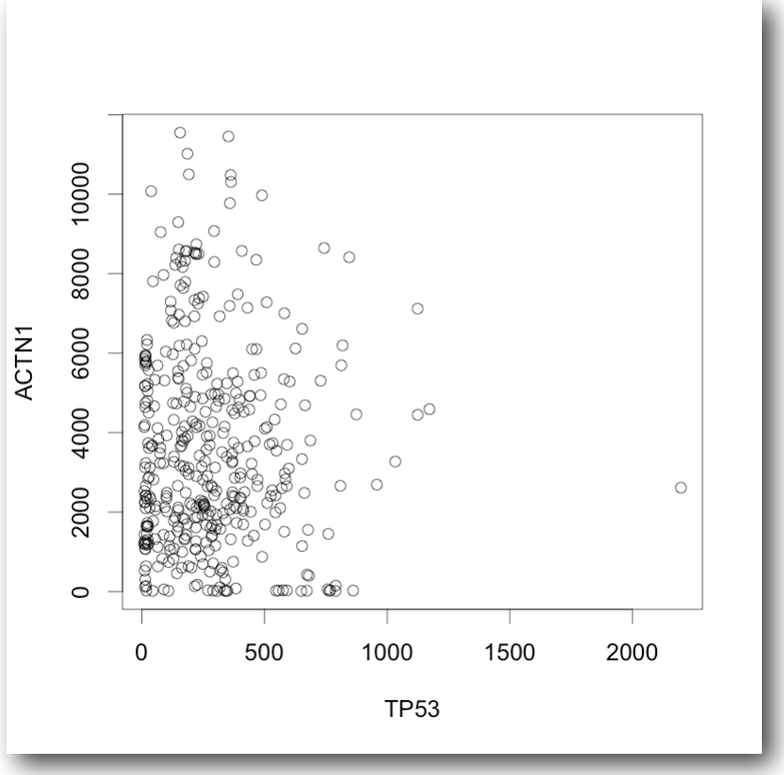
cell innovator

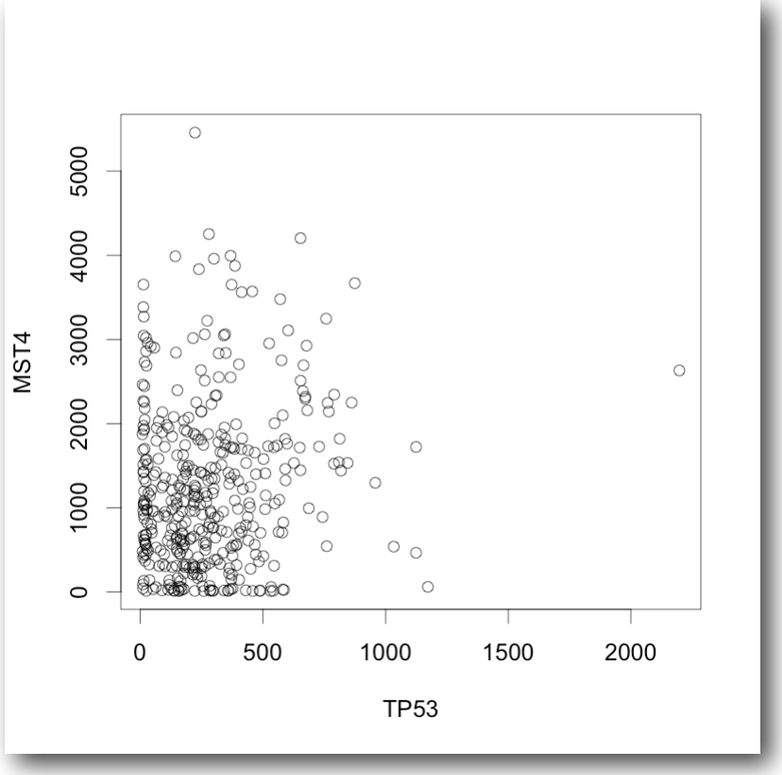


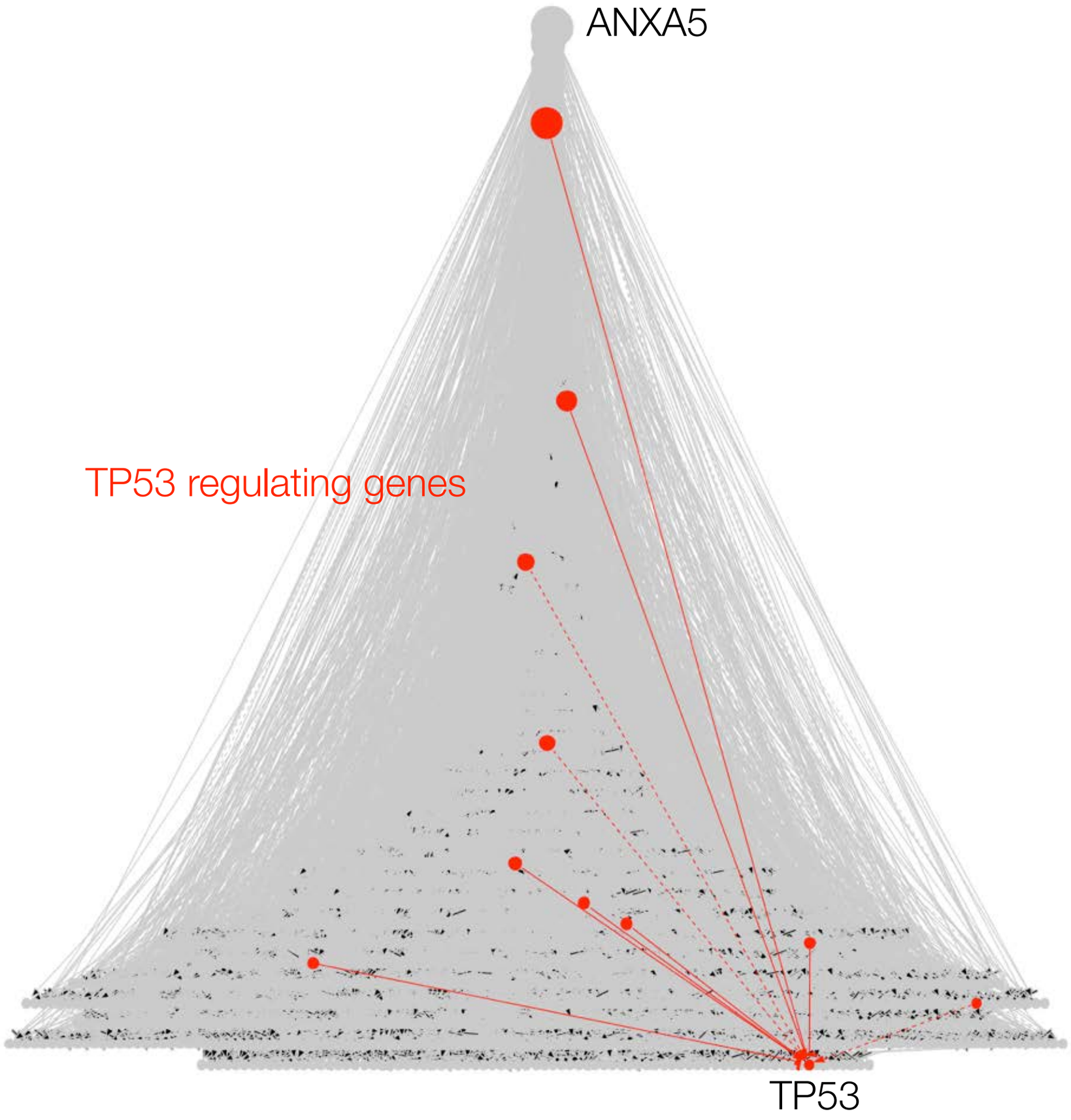


ANXA5 (top gene)

TP53
(bottom gene)






ANXA5
TP53 regulating genes
TP53

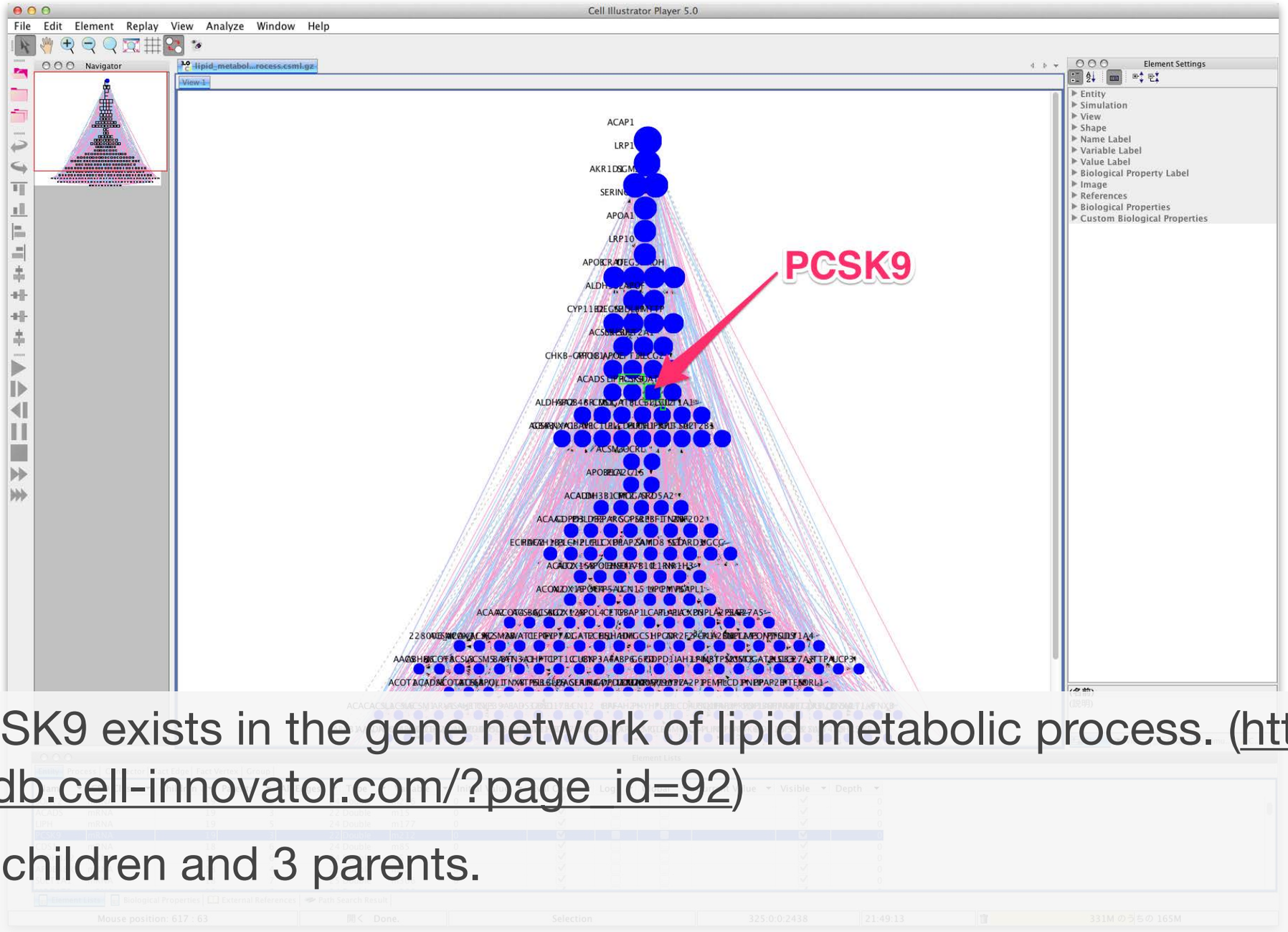
“Big pharma shows signs of renewed interest in RNAi drugs”

— *Nature Medicine* **20**, 109, 2014.

“**PCSK9**” is the target gene of Alnylam RNAi drug.

* *Lancet* **383**, 60-68, 2014.

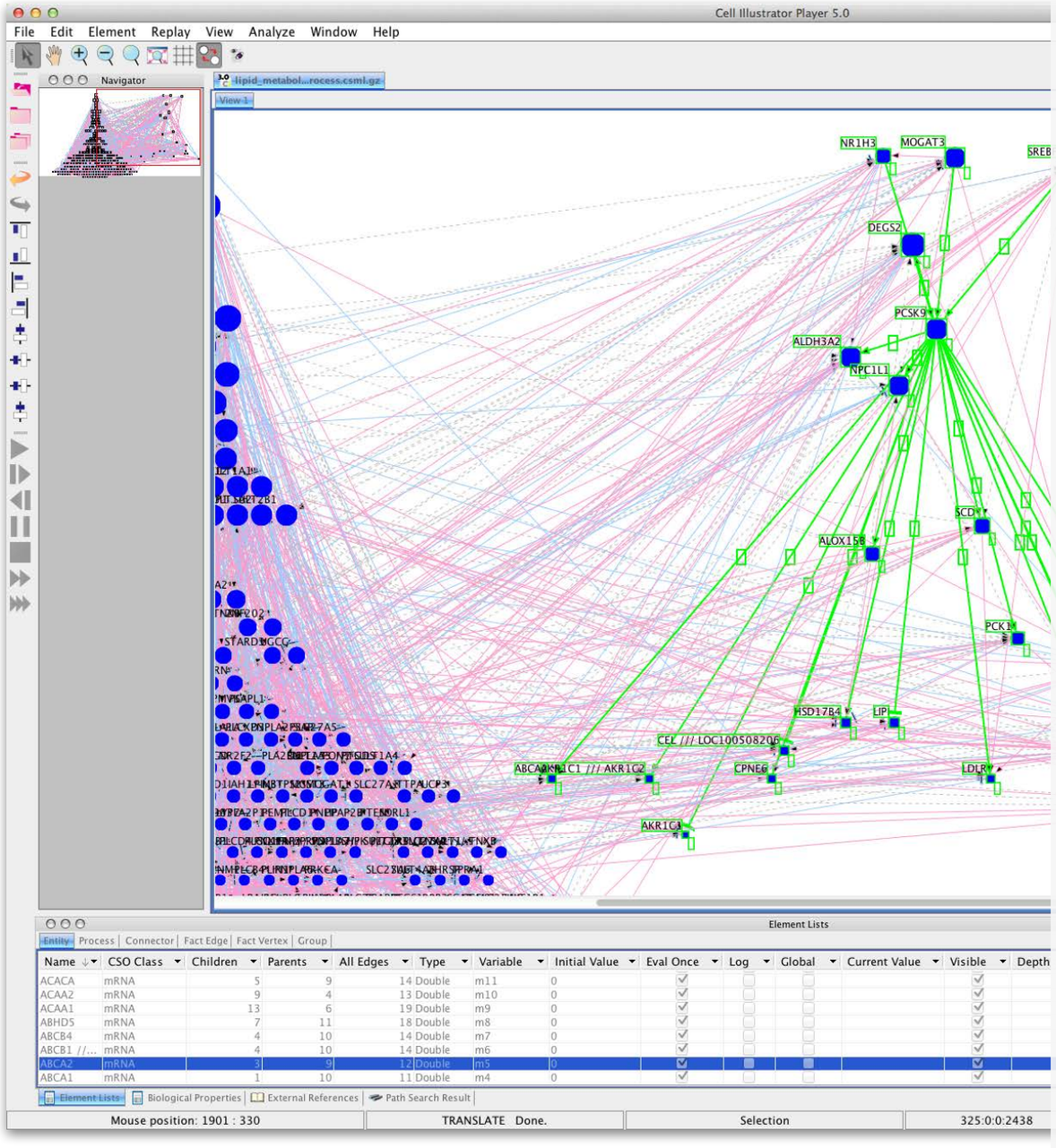
PCSK9 in the gene network



- PCSK9 exists in the gene network of lipid metabolic process. (http://gndb.cell-innovator.com/?page_id=92)
- 19 children and 3 parents.

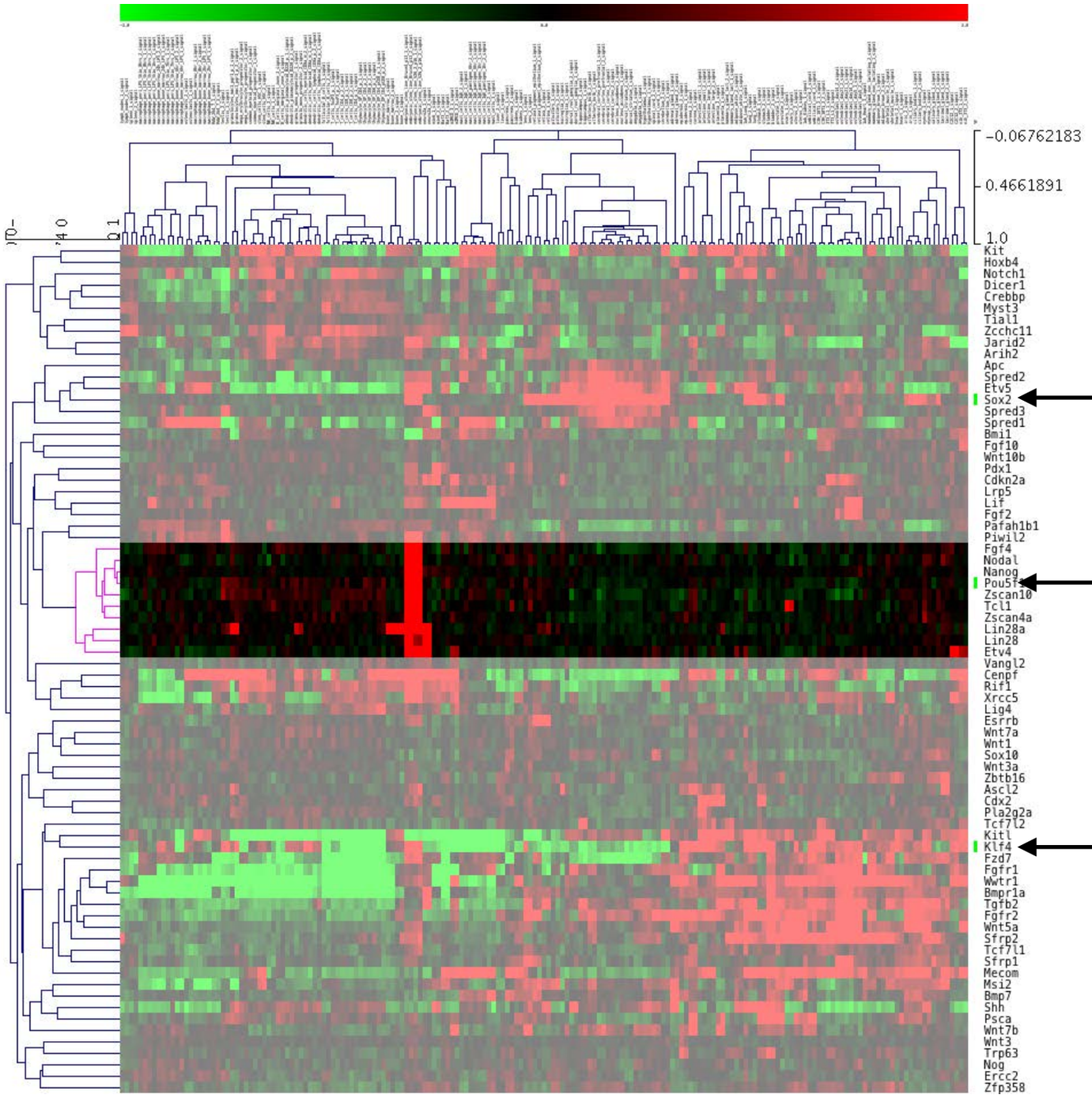
cell innovator

Next genes of PCSK9?



- PCSK9 の次に狙うのは？
  - PCSK9 の親?
  - PCSK9 の子における、PCSK9以外の共通の親？
- PCSK9 の子の発現が低い人に薬効はあるのか？
- PSCK9 が低下することで、ほかに影響を受けそうな機能は？（副作用）
  PSCK9は、apoptosis の遺伝子ネットワークにも含まれている。

クラスタリング（Stem Cell 関連遺伝子）

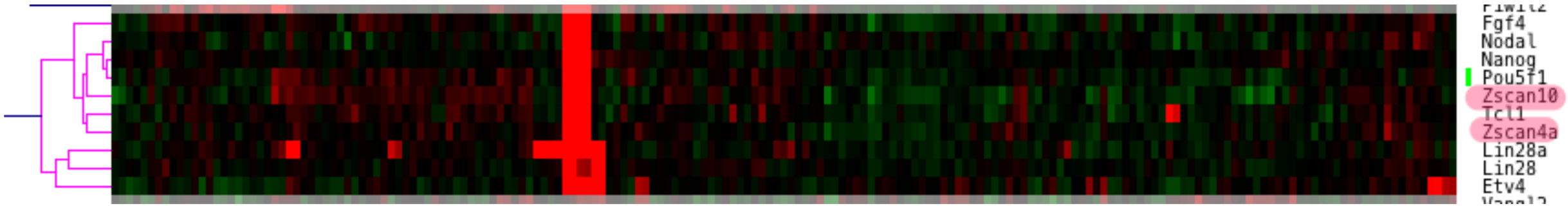


Sox2

Oct3/4

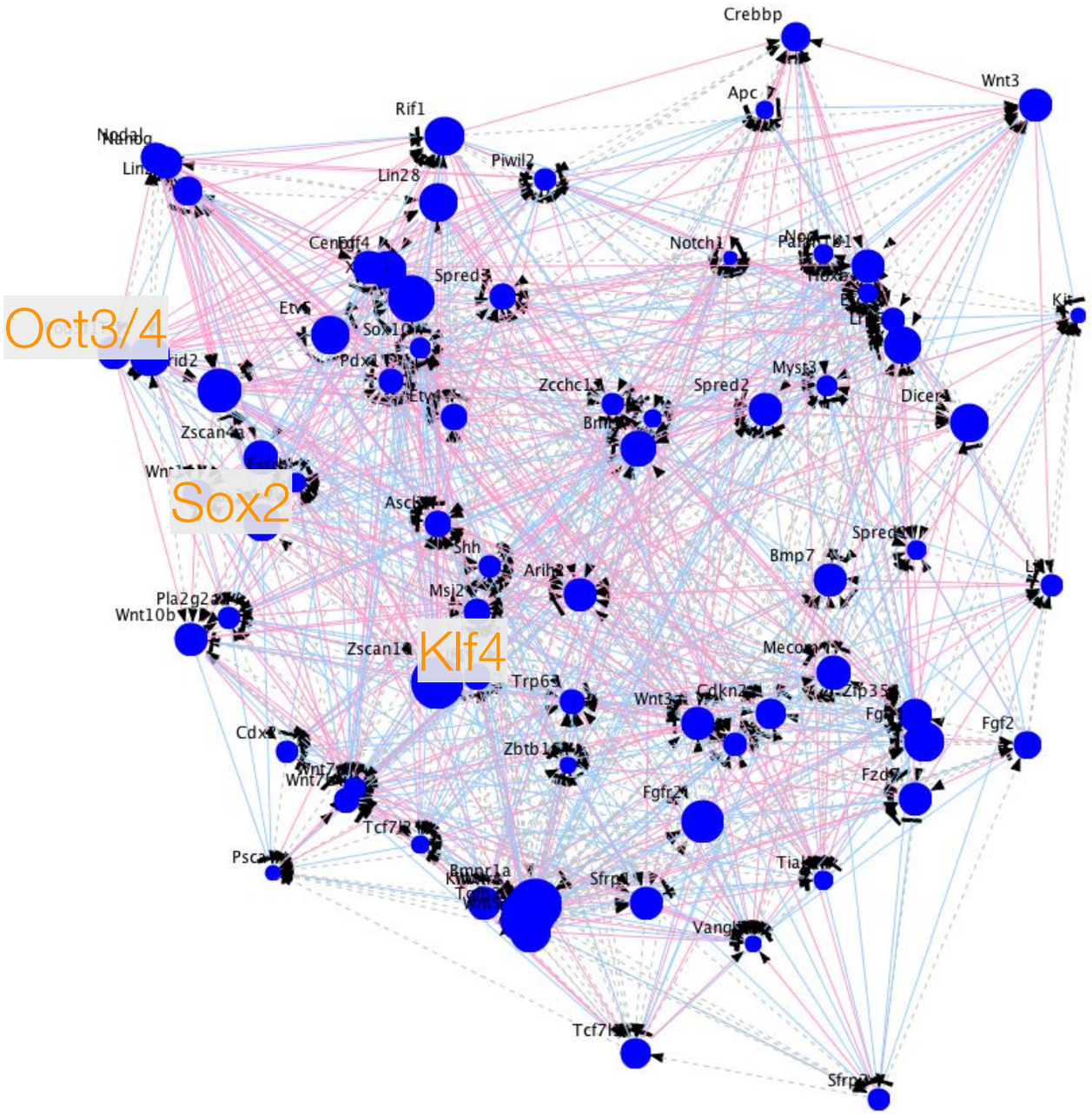
Klf4

次のターゲットは？



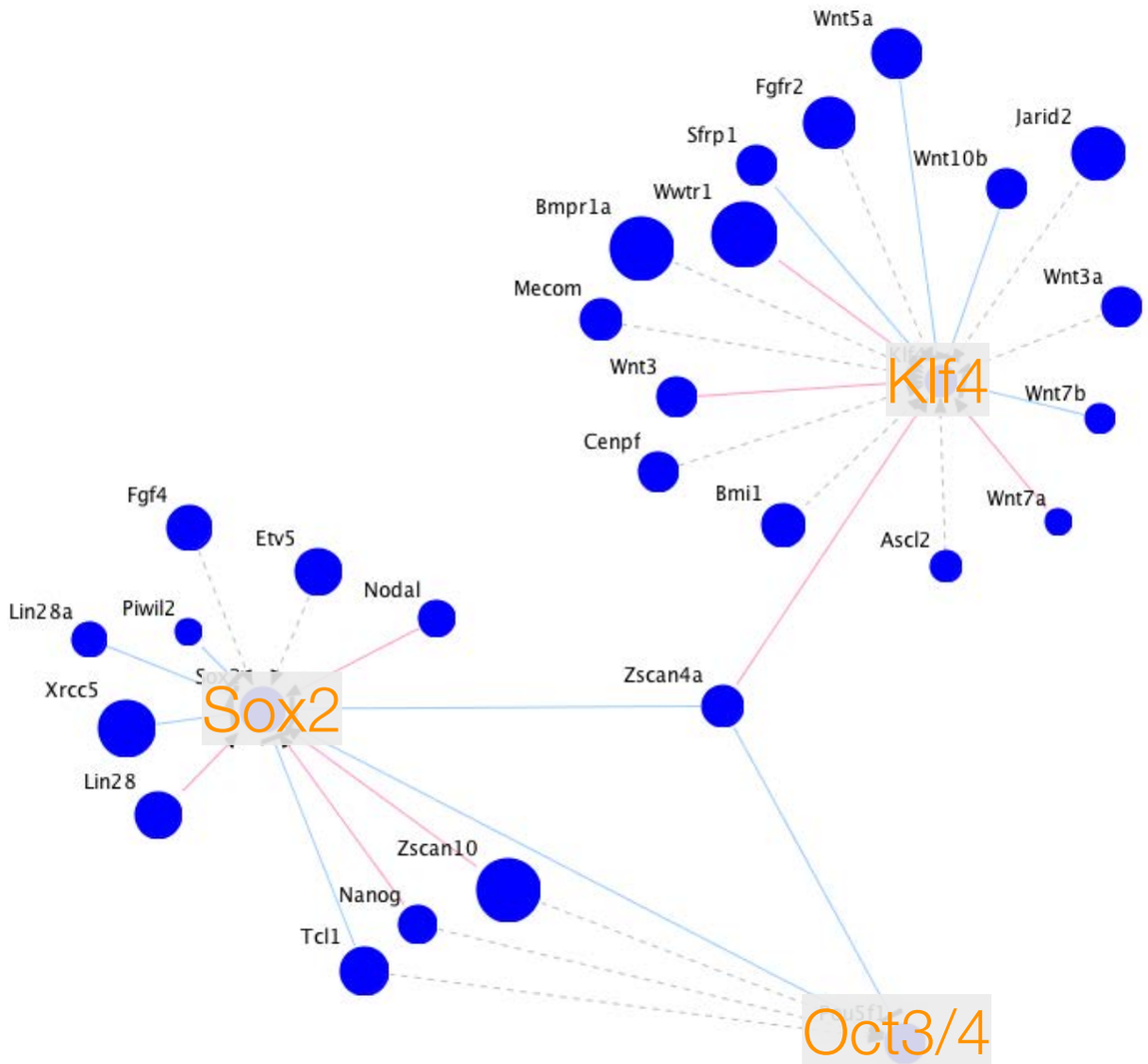
- マイクロアレイによる発現変動パターンをクラスタリングした結果。
- Pou5f1 = Oct3/4 と近いクラスターに分類された遺伝子がある。
- 研究対象として、あなたなら、どちらを選ぶ？？
  - Zscan10
  - Zscan4a
  - それとも Nanog? Tcl1?

Stem Cell 関連遺伝子の遺伝子ネットワーク



- GEO より取得した BioGPS のマイクロアレイデータ（180サンプル＝前述のヒートマップのデータ）を利用。
- アノテーションに Stem Cell を持つ遺伝子について、SiGN-BN により遺伝子ネットワークを推定。

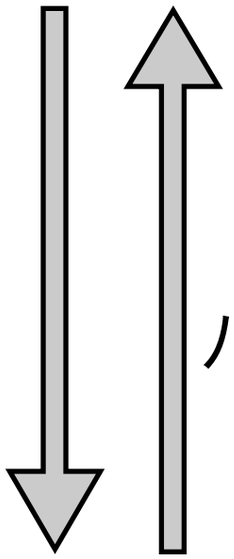
Oct3, Sox2, Klf4 の共通の親は？



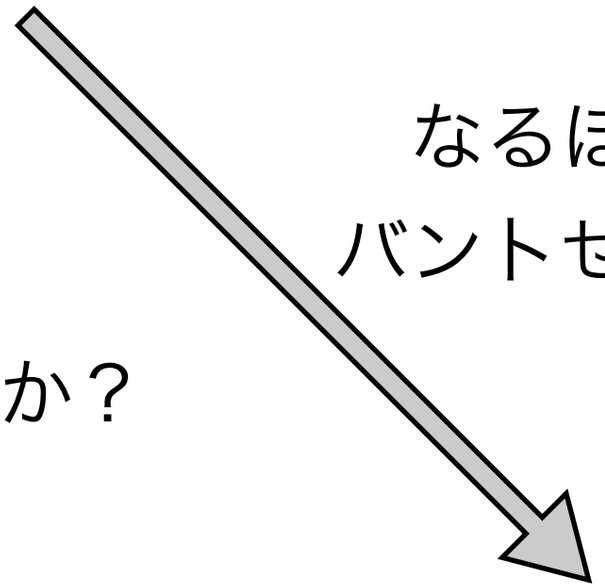
- ネットワークから、Oct3, Sox2, Klf4 の親になっている遺伝子群を抽出。
- Zscan4a のみが、3遺伝子に共通の親。
- ｏ ｏ ｏ
- 既報*でしたが。

* Jing Jiang et. al., Zscan4 promotes genomic stability during reprogramming and dramatically improves the quality of iPS cells as demonstrated by tetraploid complementation. Cell Res. (2012) : 1-15.

バントしないほうがいい



バントなしで勝てるのか？

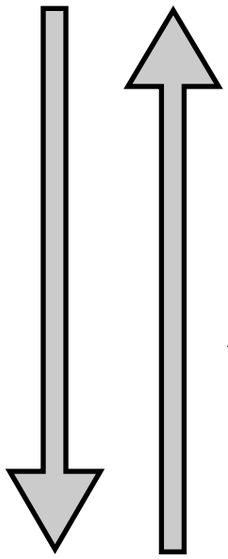


なるほど、それなら、
バントせずに勝つ方法は？

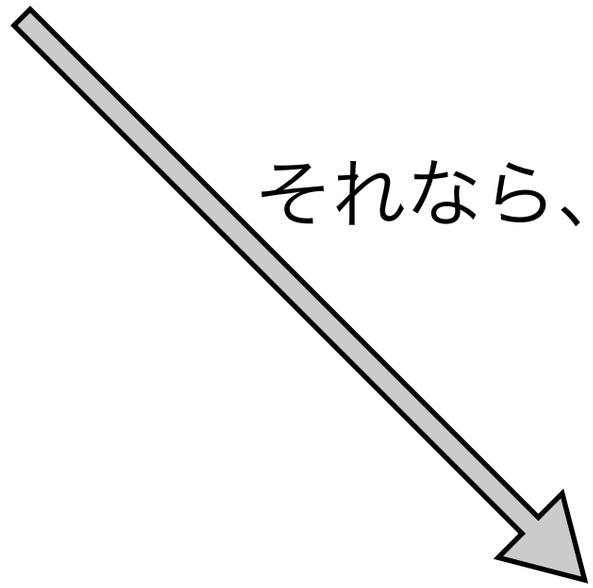
スカウト、監督

マネーボール理論

XXX が影響力ありそう



聞いたことがない。
変異もない。



それなら、何かに使えないか？

エキスパート

？？？