# 生体ビッグデータ・AI（とくにDeep　Learning）を用いた創薬・Drug Repositioningの実際

東京医科歯科大学　データ科学推進室
東北大学　東北メディカル・メガバンク機構
田中　博

# 本日の話題

- 第 I 部： 医療ビッグデータ時代の到来
- 第 II 部：ビッグデータ創薬/DR
  - 生命情報ビッグデータを用いた創薬・DR
  - 「**生体分子プロファイル型創薬/DR**」とは
- 第 III 部：AI創薬/DR
  - 人工知能とくにDeep Learningを用いた
    創薬・DRと我々の研究室での成果
  - おわりに
    - ゲノム・オミックス医療の次世代の展開

# 第 Ⅰ 部

# 医療ビッグデータ時代の到来

TMDU

# 医療ビッグデータ時代

(1) 次世代シーケンサ（Clinical Sequencing)による
「ゲノム/オミックス医療」における網羅的分子情報収集/蓄積
(2) **Biobank**/ゲノムコホート普及による分子・環境情報の蓄積
(3) モバイルヘルス(mHealth）によるWearable センサの連続計測による生理データの蓄積（unobstructed monitoring)

急激な大量データの出現
コストレス化かつ高精度化

!!

ゲノム：13年→1日(1/5000) 3500億→10万円(1/350万)

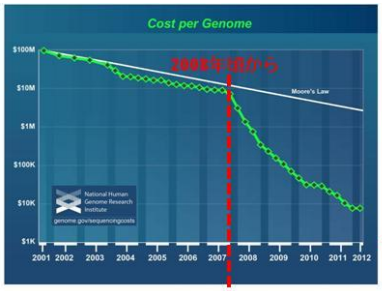
個別化医療・医療の国民レベルの向上
医療/ヘルスケアの適確性の飛躍的な増大

# ゲノム医療の２つの流れ

- 米国の流れ
  - 次世代シーケンサの急激な発展による「シーケンス革命」からの怒涛の展開（2010から）
  - 「治療医学」レベル質的向上のためにゲノム情報を取り入れた臨床実装の推進
    - 稀少疾患の原因遺伝子変異の同定
    - がんのドライバー遺伝子変異の同定と分子標的薬の選択
    - 薬剤代謝酵素の多型性の同定と個別化投与
- 欧州の流れ
  - 社会福祉国家の理念より国民医療（医療の国民レベル）の向上
  - 「予防医学」レベル質的向上のためにゲノム情報を取り入れたバイオバンク推進
  - 大規模前向きpopulation型バイオバンク/ゲノム・コホートの確立
    - 遺伝的素因だけでなく環境要因（生活習慣）との相互作用を解明し、「ありふれた疾患」発症を予測し、これに基づいて個別化予防する。
    - 疾患を発症前に対応して発症を防ぐ「先制医療(preemptive medicine)」や「予測医療（predictive medicine）の実現を目的

TMDU

# 米国ゲノム医療の流れ




千ドル
ゲノム

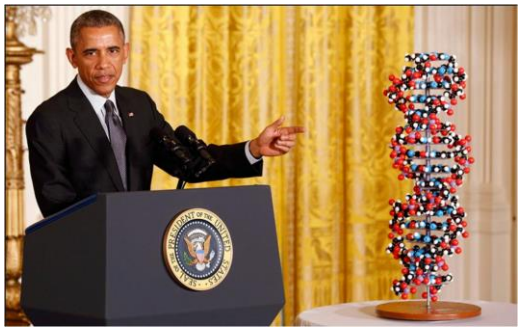
DNA Sequencing Cost: the National Human Genome Research Institute

シーケンス革命　2007/8

2005～ NGS 454 (LS,Roche)
2007/8～454, Solexa (Illumina), SOLiD (LT,TF)
シーケンス革命

| | HiSeq2500 | | Ion Proton |
|---|---|---|---|
| 本体価格 | 約1億円 | | 約3500万円 |
| モード / チップ | ハイアウトプット | ラピッドラン | Ion Proton I |
| 解析時間 | 11日 | 27時間 | 2時間 |
| リード長（bp） | 2 x 100 | 2 x 150 | 200 |
| データ産出量(Gb) | 約600 | 約120 | 10 |
| 試薬コスト（ヒト1人全ゲノム） | 数十万円 | | 不可 エクソームのみ |

急速な高速化と廉価化
ヒトゲノム解読計画13年,3500億円
⇒1日,10万円



先陣争いの時代

国家政策の時代

精密医療普及期

第一期

ゲノム多型性の認識
Hapmap 計画(2002)
GWAS研究など

薬剤代謝酵素の多型性の判別・電子カルテで警告・Preemptive PGx
Vanderbilt大病院

2005～ NGS 登場
(454,Solexa,SOLID)
2007/8～
シーケンス革命

Undiagnosed Genetic Diseaseの原因遺伝子POC同定
MCW小児病院

国際がんコンソーシアム開始
ICCG (2008年)
2011頃からがん変異成果報告

Cancer Driver Geneの同定と抗がん剤治験
Dana Faber CC

第二期

ゲノム・オミックス医療の臨床実装の普及
ゲノム・オミックス情報のビッグデータの出現

ゲノム医療の国家的取組み
NIH “BD2K “計画 ・各種ゲノムコンソーシアム開始

オバマ大統領　年頭教書
Precision Medicine initiative 政策の発表

第三期

100万人コホート:バンダービルト大学設開始(2016-2020)
NCI“National Cancer MoonShot” 10年計画開始
各州でのプレシジョン医療計画開始(カリフォルニア、ペンシルバニア)

2007年
2009年
2010年
2011年
2012年
2013年
2014年
2015年
2016年
2017年

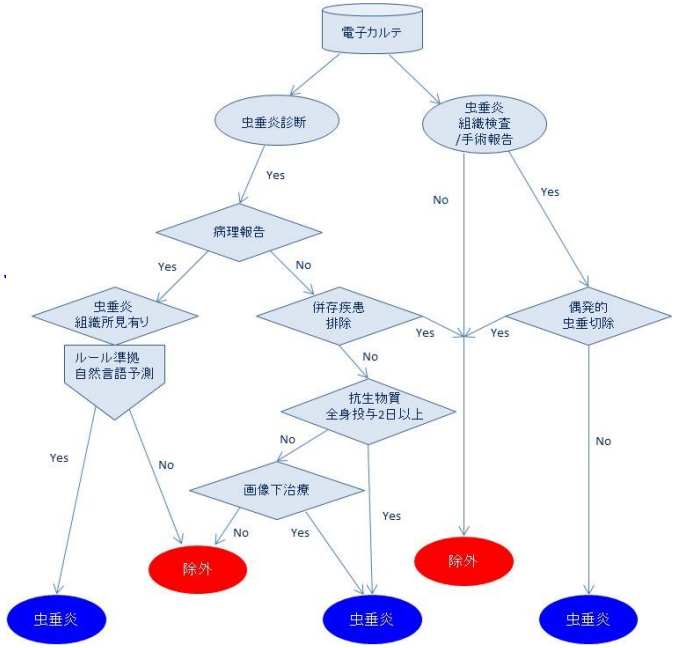
オバマ大統領　Precision Medicine Initiativeを開始
2015年1月　大統領一般年頭教書演説

TMDU

# 臨床表現型との統合 *eMERGE* 計画

electronic MEdical Records + GEnomics (NHRI-funded)

- phase I (2007-2011) EMR-bsedゲノム研究の探求
  - EMR(臨床phenotyping)とbiorepositoryに基づく
  - GWAS等（EMR-based GWAS）が可能か。
  - 開始時はGWAS全盛時代。ゲノム医療の臨床実装は始まら
  - 電子カルテより臨床表現型情報 phenotypingルール
  - 計画開始時参加施設：Mayo, Vanderbilt Univ., Marshfield, Univ. Washington, Northwestern Univ.など５施設,
- phase II (2011-2015) 臨床実装へ舵を切る
  - MCWの臨床実装のインパクト、Vanderbiltの先制PG x
  - 電子カルテと遺伝情報の統合
    - 電子カルテへのゲノム情報の統合
    - **PheKB** (Phenotype Knowledge Base)
    - ゲノム医療の実装、PGxの臨床応用
    - 結果回付 **Return of Result,** ELSI等
  - 4つのサイトが新しく加わる
    - 小児病院グループとMount Sinai, Geisinger
- phase III : 2015より始まる
- NHGRI予算化のコンソーシアムと連携
  - **CSER consortium等**と連携
    - “Clinical Sequencing Exploratory Research”




PheKB: phenotyping ルール

# 個別化医療からPrecision Medicine

個人の遺伝素因・環境要因に合わせた (tailored) 医療
One size fits for all のPopulation 医療とは異なる

趣旨：基本は、個別化医療 Personalized Medicine の概念と変わらないが、目的は診断/ 治療の個人化ではなく層別化を明確化

概念の拡張：Personalized Medicineが標榜された時から10数年経っている

## 医療ビッグデータ時代の到来による個別化医療の拡張

**（１）遺伝素因 X 環境(生活習慣)要因のスキーマ重視**
遺伝要因(SNPや変異：Genome) だけでなく環境・生活習慣要因(Exposome) の重視
疾患発症は2つの要因の相互作用と明快に強調。電子カルテの臨床表現型（Clinical Phenome)情報の重要性認識。

**（２）ゲノムコホート・Biobankの重視**
Precision Medicineを実現する「情報基盤」として、ゲノムコホート/Biobankが必要であることを認識。Real world dataの重視

**（３）日常生理モニタリング情報の包摂**
モバイルヘルス(mHealth)・ wearableセンサーによる大量継続情報収集の重視

# 医療ビッグデータ時代の到来（米国）

ゲノム医療の実践

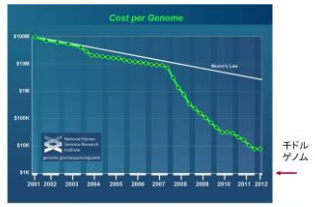
**第１段階 ゲノム医療の発展**

次世代シーケンシングの臨床普及(2010～)

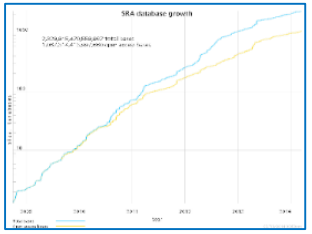
全ゲノム（x30：100Gb)・エキソーム解析（x100：6Gb）

米国では数十の著名病院で実施

ゲノム・オミックス情報の蓄積




2000兆塩基 (2 Pb)
が登録(NCBI:SRA)




医療ビッグデータ

**第２段階 医療ビッグデータ時代**

医療情報との統合

電子カルテからの
臨床フェノタイプ

医療ビッグデータ

学習アルゴリズム

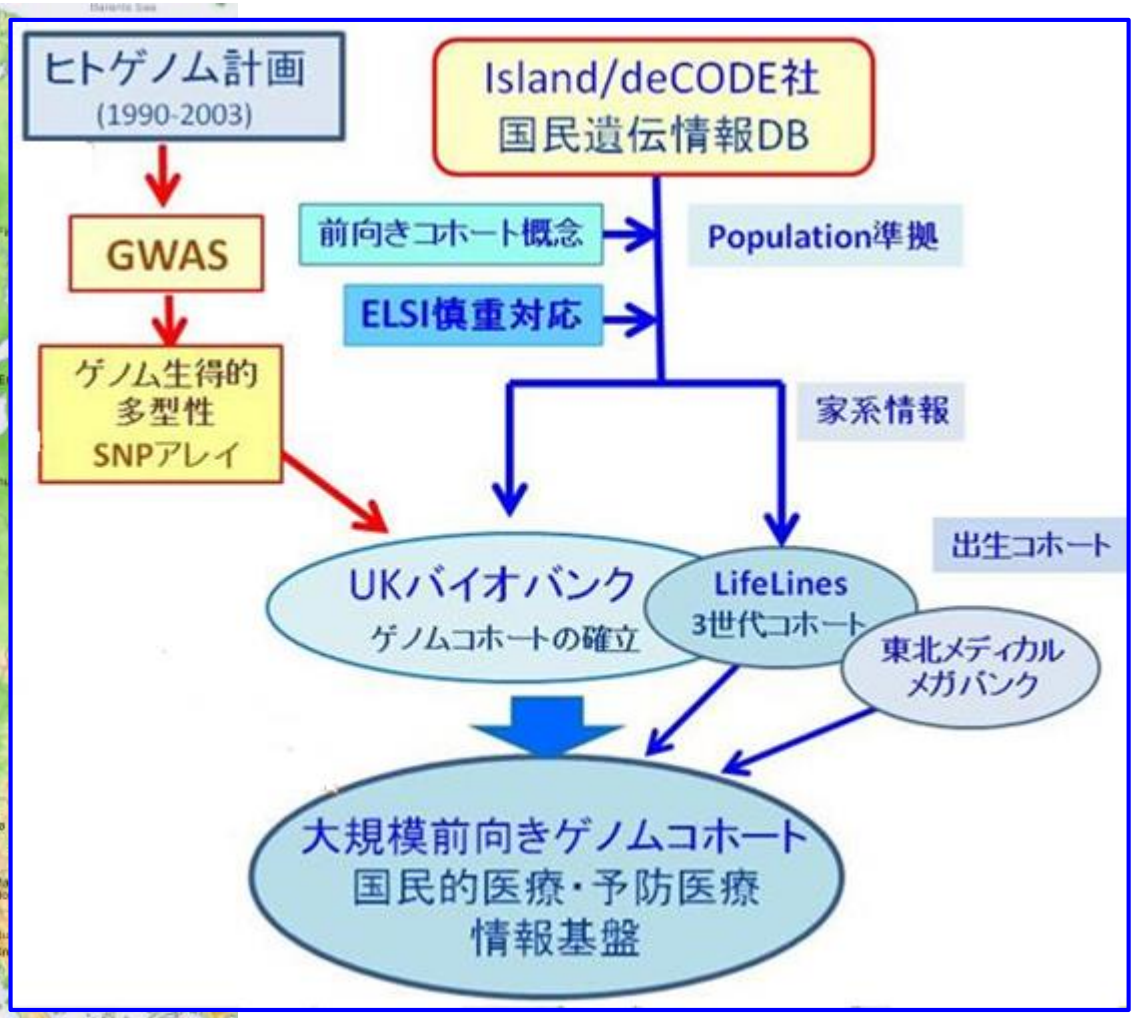
ゲノム医療知識

人工知能AI

MayoClinicでは
10万人患者WGS

TMDU

# 欧州のバイオバンクの流れ

# ビッグデータ医学/医療の第２の流れ
# Biobankとゲノムコホートの世界的興隆

## バイオバンクの目的･機能の変化

- 従来は**稀少疾患組織標本や臨床研究の**資料保存、近年は**ゲノム医療の基盤**としての役割が認識され、**世界的に普及**
- **ゲノム/オミックス個別化医療、創薬**の情報基盤：**疾患**BioBank（**疾患ゲノムコホート**）
  - **疾患罹患患者の網羅的分子情報**（ゲノムなど）とそれに対応する**臨床表現型情報**（臨床検査、医用画像、処方歴、手術歴、病態経過、転帰など）の収集。
  - **疾病の分子機序や治療戦略、予後予測、創薬科学への貢献**
- **予防**の情報基盤：Population**型**BioBank（一般住民ゲノムコホート）
  - 「健常者」前向きコホート。調査開始時の網羅的分子情報（ゲノム）と臨床・環境情報（exposome）を集めて、長期間（生涯）を追跡するゲノム・コホート
  - 主に遺伝子素因情報も含めた「ありふれた病気」の**疾患の発症リスク予測、重症化予測**

## 欧米のBiobank

- **英国　UK biobank**
  - 50万人の健常者。40～69歳（2006-2010, 62Mポンド), 追加調査（2011-16, 25Mポンド）
  - 健診データ（血液・尿・唾液サンプル、生活情報）とゲノム情報（SNPアレイを集め、健康医療状況を追跡する。その淵源は、アイスランド、deCODE社の「国民遺伝子情報データベース」プロジェクト
- **英国　Genomics England**,
  - 2013開始、2017年までに 10万人のゲノム 配列収集。全ゲノム次世代シーケンス
  - 最初の対象は稀少疾患（患者・家族）、がん患者、最初はEnglandのみ。企業とのコンソーシアム
- **欧州　BBMRI**（Biobanking and Biomole Research. Infrastructure.)
  - 250以上の欧州各国のBioBankを統合
- **オランダ　Lifeline**
  - 165000人北部オランダ　2006年開始　30年間の追跡、3世代コホート（世界初）
- **Precision Medicine Initiative, Genome Cohort：“All of Us”コホート**
  - これまでのBiobank（例えばBioVUなど）を集めて100 万人のゲノムを集める

TMDU

# ビッグデータ医学/医療の第３の流れ
# 大規模な生命情報DB/KBの出現と利用

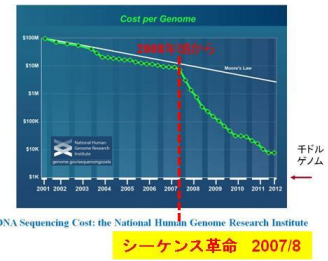
- ヒトゲノム解読計画以降急速に進展
  - Hapmapプロジェクト,1000 genome,がんICGC,TCGA等
  - ゲノム変異・多様体
    - dbSNP, HGMD, **Clinvar**, **Clingen**, OMIM,GWAS catalog
    - 表現型との対応: dbGaP, EGA
  - 遺伝子発現プロファイル
    - 疾患特異的transcriptome: **GEO, ArrayExpress**,
    - 薬剤特異的trancriptome: **c-Map, LINCS**
  - タンパク質
    - ３次元構造：PDB, Swiss-Prot,
    - タンパク質間相互作用：**HPRD, STRING,** BIND
  - 分子ネットワーク、パスウェイ
    - KEGG, TRANSFAC, BioCyc, Reactome
- 各種バイオバンク症例ベース（制限アクセス）
  - UK biobank, BMBRI, 東北メディカル・メガバンク
- これらの大規模DB/KBを組合せてゲノム医療・創薬を推進

TMDU

# 医学/医療へのビッグデータの衝撃


次世代シーケンサの登場

**シーケンス革命（2007)**




コストレスで高精度な網羅的分子情報の出現

1. ゲノム・オミックス医学/医療の進展
   - Clinical Sequencingによるゲノム・オミックス医療の臨床実装の急速な進展
2. Biobank/ゲノムコホートの世界的普及
   - 個別化医療/予防の情報基盤として普及
3. 大規模な生命情報DB/KBの出現
   - ゲノム・オミックスによるDB/KBの膨大化

# 医療ビッグデータ

- 臨床ゲノム・オミックス医療の進展
  - Clinical Sequenceのインパクト
  - 網羅的分子情報、臨床表現型情報の統合
  - 個別化医療、Precision Medicine
- Biobank, 疾患レジストリの拡充
  - 疾患型：個別化医療の情報基盤
  - 住民型：慢性疾患発症予測・個別化予防
  - レジストリ準拠ランダム臨床治験
- 網羅的分子情報DBの大規模化と利用
  - 1000genome, GWAS, ICGC, Clinvar, Clingen, dbGaP, Cmap, LINCS, HPRD,STRINGなど

TMDU

# 医療の「ビッグデータ」革命は どんな既存のパラダイムに挑戦しているか

- Population medicineのパラダイム転換
  - <One size fits for all>のPopulation医療はもはや成り立たない
  - 個別化医療 “Personalized (Precision) medicine”
    - 個別化医療実現のために＜個別化・層別化パターン＞がどれだけ有るか網羅的に調べる：どこまでの粒度で個別化・層別化すればよいか
- Clinical research（臨床研究）のパラダイム転換
  - 臨床研究を科学にする従来の範型RCTは,個別化概念を取扱えない
  - <statistical evidence based>呪縛からの解放
  - 「標本」統計・「推測」統計学に制約されない臨床研究
  - Real World Data・ビッグデータからの知識生成（BD2K)
- 創薬の戦略パラダイムの転換
  - ＜ビッグデータ創薬＞へ:分子特性準拠からビッグデータ準拠へ
  - 網羅的分子データ/ネットワーク情報からの計算論的創薬・DR
  - Disease, Drug, Targetの各ネットワーク間の相互写像関係

TMDU

# 医療の「ビッグデータ革命」
# ～何が新しいのか～

**1）臨床診療情報**
- 従来型の医療情報 電子化による蓄積
  - 臨床検査、医用画像、処方、レセプトなど

**2）社会医学情報**
- 従来型の社会医学情報 電子化による蓄積
  - 疫学情報・集団単位での疾患罹患情報

旧来のタイプの
医療データの
大容量化

**3）新しい種類の医療ビッグデータ**
- **網羅的分子情報・個別化医療**
  - **ゲノム・オミックス医療の網羅的分子情報**
  - **Biobank,ゲノムコホートによる分子/環境情報**
- **生涯型モバイル健康管理**（mHealth）
  - **ウェアラブル・生体センシング**

新しいタイプの
医療ビッグデータ

TMDU

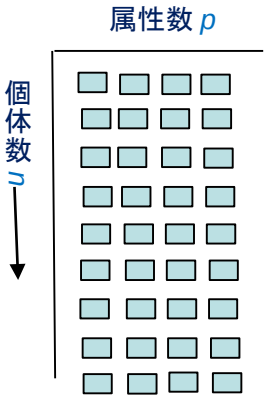
# 医療の「新しいビッグデータの革命性」

## ～ゲノム・オミックスデータの基軸的な特徴～

＜目的もデータ特性も従来型と違う＞

### 従来の医療情報の「ビッグデータ」

**Big “Small Data”　(n≫p)**



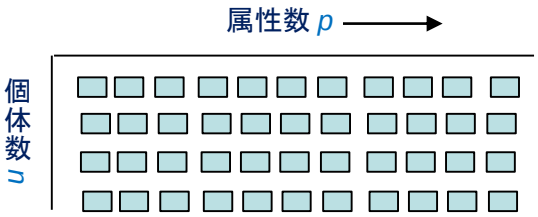

医療情報・疫学調査では 属性数：数十項目程度

– 目的：Population MedicineのBig Data

⇒個別を集めて「集合的法則」を見る

### 網羅的分子情報などのビッグデータ

**Small “Big Data”(p≫n)**




1個体に関するデータ属性種類数が膨大

属性(*p*)に比べて個体数(*n*)少数:従来の統計学が無効

「新 ***np*** 問題」: GWASは単変量解析の羅列

– 目的：例えば医療の場合 個別化医療 Personalized Medicine

⇒大量データを集めて「個別化パターン」の多様性を摘出

## 新しいデータ科学の必要性

TMDU

# ゲノム医療時代の
# ビッグデータ解析・人工知能

- ゲノム医療の2つの流れ
  - どちらにおいても超多次元相関ネットワークから「革新的知(innovative insight)」発見の必要性
- 治療医学：米国型
  - 〈網羅的分子情報と臨床表現型情報〉の相関ネットワークより革新的知の発見
  - 分子画像やオミックス情報により複雑化
- 予防医学：欧州型
  - 〈遺伝的素因と環境/生活様式要因〉の相互作用と発症の相関ネットワークより革新的知の発見
- 医療ビッグデータ
  - 超多次元ネットワークから
    如何に「innovative knowledge」を獲得するか

TMDU

# わが国での現状「ゲノム医療元年 (2016)」

■「ゲノム医学実現推進協議会」(中間報告）2015.7

研究費を用いた試行的ゲノム医療であるが、いくつかの医療施設で
ゲノム・オミックス医療が試行されている

- 例：がんの網羅的分子診断と個別化治療
  - 国立がん研究センター　（Top-gear,SCRUM-Japan)
  - ドライバー遺伝子の診断。分子標的薬の治験グループに割当て
  - 静岡県立がんセンター 上記と同様の内容のプロジェクト
  - 京大腫瘍内科（OncoPrime),岡大,北大,千葉大 診療施設併設型BB

■AMED（日本医療研究開発法人）がゲノム医療を推進

- IRUD（Initiative on Rare and Undiagnosed Disease)
  未診断疾患の原因遺伝子をIRUD拠点病院が審査して解析センターがシーケンシング。その後、DB化する。
- ゲノム医療実現推進プラットフォーム事業
- 臨床ゲノム情報統合DB事業

ゲノム医療では、米国と水を空けられている。しかし、Biobank Genomic Cohortでは我が国の状況はそれほど遅れていはいない。
Biobank準拠のゲノム医療/創薬推進を行うべきである。
また、第２世代のゲノム医療：多因子疾患に着手すべきである

# 第II部

# ビッグデータ創薬/DR

# ビッグデータと医薬品開発

**ビッグ網羅的分子知識・DBを創薬・DRへ利用**

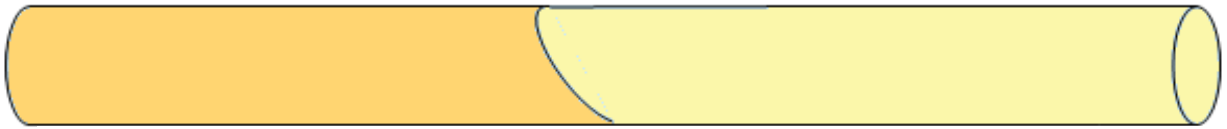
1．オミックス創薬/DR（遺伝子発現プロファイル）
2．疾患ネットワーク創薬/DR
3．階層的ネットワーク創薬/DR

**ビッグデータ・疾患レジストリの治験への利用**

1．疾患レジストリ準拠臨床無作為化治験
(RRCT: registry-based clinical randomized trial)

**ビッグデータ・BioBankの利用**

大規模知識・DBの利用　疾患BD/Registryの利用



**創薬・DR（研究開発）**　**治験・市販後調査（臨床試験）**

# 生体分子プロファイル型創薬/DRの基本概念

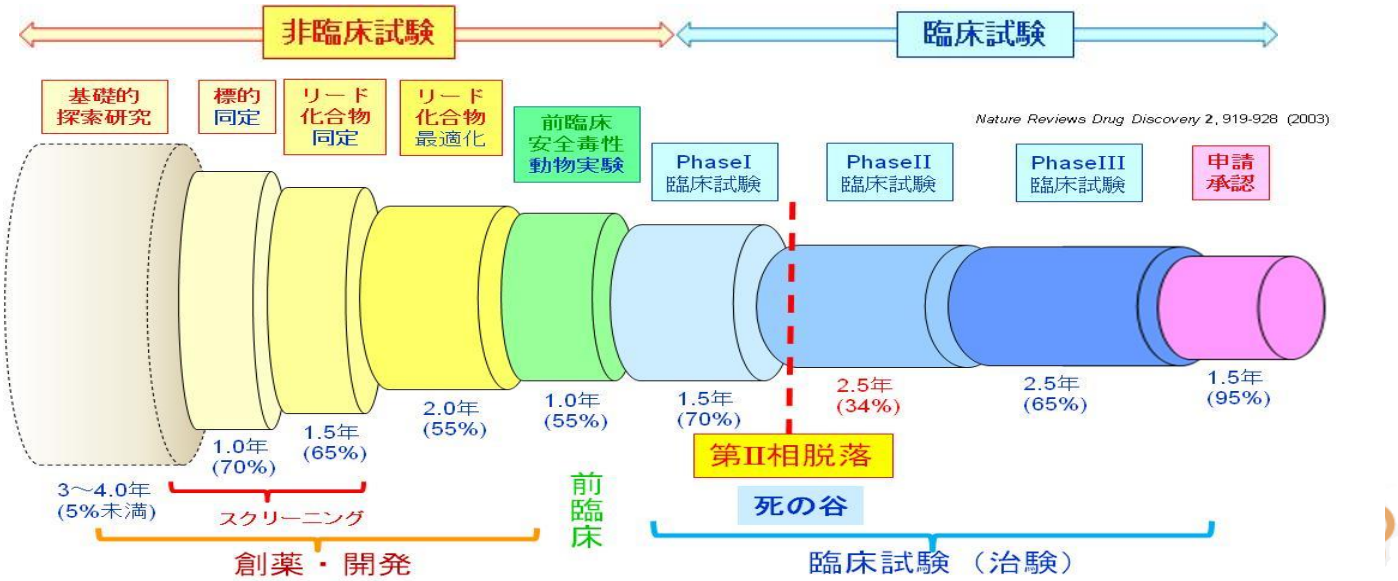
TMDU

# 創薬をめぐる状況と解決の方向

- 医薬品の開発費の増大
  - １医薬品を上市するのに約1000億円以上
- 開発成功率の減少
  - 2万~3万分の１の成功率
  - とくに**非臨床試験**から**臨床試験**への間隙
  - **phase II attrition (第２相脱落)**
- 臨床的予測性
  - 医薬品開発過程の**できるだけ早い段階**での**有効性・毒性の予測**
- **臨床予測性の早期での実施**
  - **罹患者のiPS細胞を使う**


*Nature Reviews Drug Discovery* (2012)

ヒトの＜薬剤－疾患－生体系＞のビッグデータを早期R&D段階で使う



# ドラッグ・リポジショニング

## 薬剤適応拡大

**ヒトでの安全性と体内動態が十分に分かっている**
**既承認薬**の標的分子や作用パスウェイなどを、体系的・論理的・網羅的に解析することにより**新しい薬理効果**を発見し、その薬を**別の疾患治療薬として開発する**創薬戦略

## 利　点

（1）既承認薬なので、ヒトでの安全性や体内動態などが既知で臨床試験で予想外の副作用や体内動態の問題により開発が失敗するリスクが少なく**開発の成功確率が高い**

（2）既にあるデータや技術（動物での安全性データや製剤のGMP製造技術など）を再利用することで、**開発にかかる時間とコストを大幅に削減できる**

（3）**DR候補探索**に**疾患生命情報ビッグデータ知識DB**を使用できる。

Tanaka K…,**Tanaka H**, ..,Mizushima T, *Nature Comm*, 4:2686, 2013

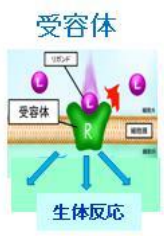
# 疾患・薬剤・標的の関係

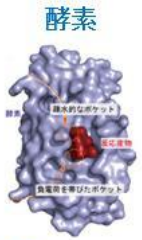
病気の主要な要因

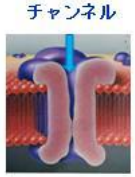
疾患関連タンパク質（複数）

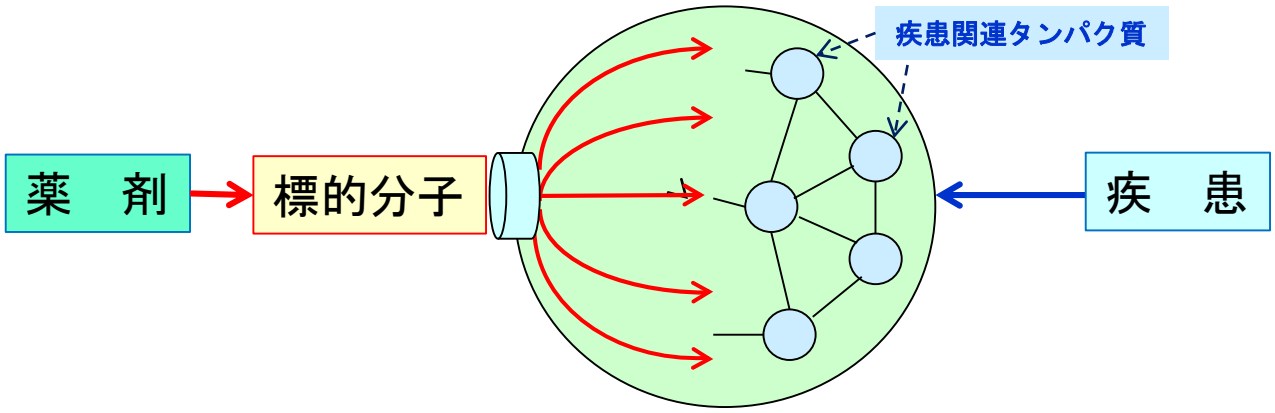
薬：疾患関連タンパク質に影響を示す
標的タンパク質に作用し阻害する

薬剤の標的分子
受容体・酵素・チャンネルなど



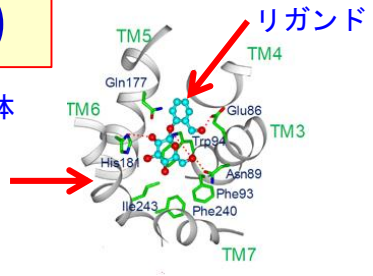



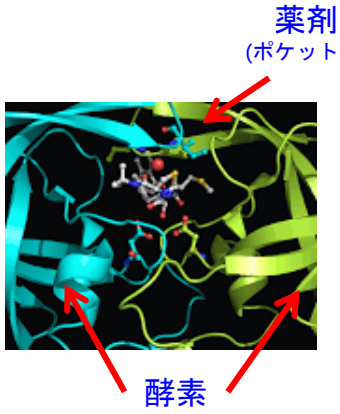

生体システム/ネットワーク

# ビッグデータ計算創薬１

## 計算創薬(computational drug discovery) の新しい方向

### これまでの計算創薬 (*in silico* 創薬)

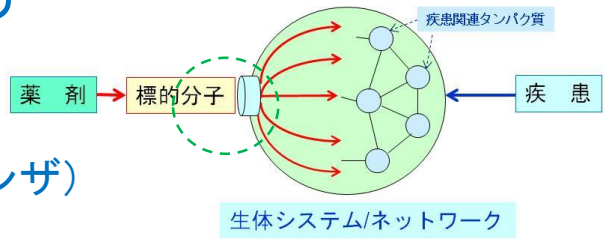



**分子(結合構造)中心**

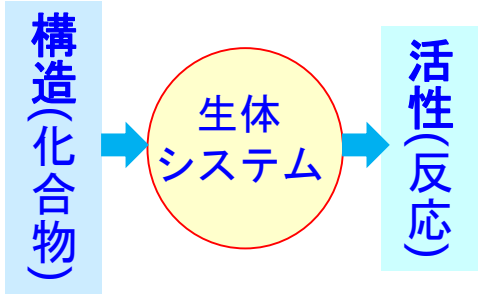
- **分子構造解析・分子設計**
- **Structure-based rational drug design**
- **標的分子**(受容体・酵素)**と薬剤**(リガンド)との**結合構造**（ポケット）の分子構造を根拠に
- **リガンドの分子設計(量子化学等を使う)**
  - 成功例：インフルエンザ薬　ザナミビル（リレンザ）
- 標的に結合する**リード化合物・構造最適化**
- **結合後の生体システム**の反応・振舞い

  ⟹ 明確な取扱いがない




**定量的構造活性相関**(QSAR)

- 化合物の分子**構造**と生体**活性**の**関係**
- **しかし両者の間**には**生体システム**がある

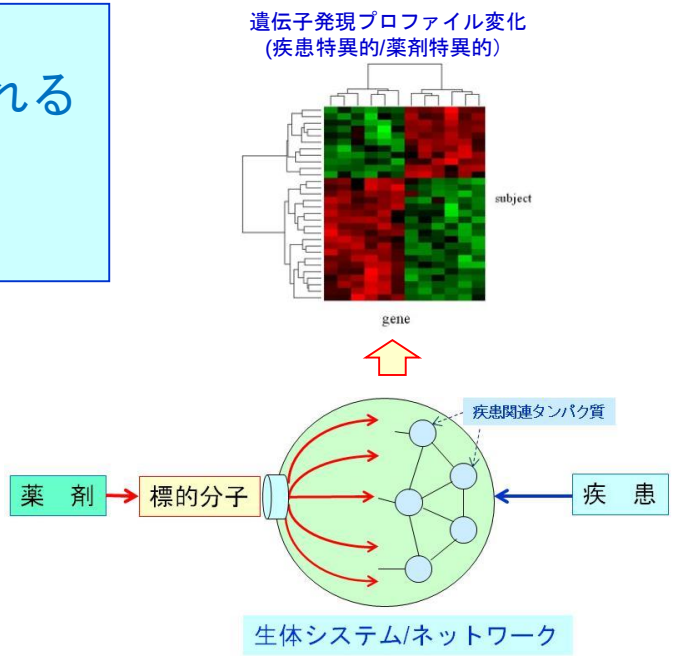


# ビッグデータ計算創薬２

**新しい計算論的創薬**のアプローチ（生体分子プロファイル型創薬）

**疾患罹患状態における**
疾患関連遺伝子（タンパク質）に起因し決定される
疾患時の**生体のゲノムワイドな特異状態**
疾患特異的な網羅的分子プロファイル変化

**薬剤投与による**
**標的分子と薬剤**分子の結合に起因し起こる
**投与時の生体のゲノムワイドな反応/振舞い**
薬剤特異的な網羅的分子プロファイル変化




**網羅的分子プロファイル⇒分子ネットワーク全体変化**

**＜疾患状態の生体＞に＜薬剤ー標的分子の結合＞から**引き起される作用によって
**ゲノムワイドな 生体分子環境がどう変化するか「生命システム観点からの理解」**

化合物, 標的分子, 疾患間の関係の「ビッグデータ」DBを利用

TMDU

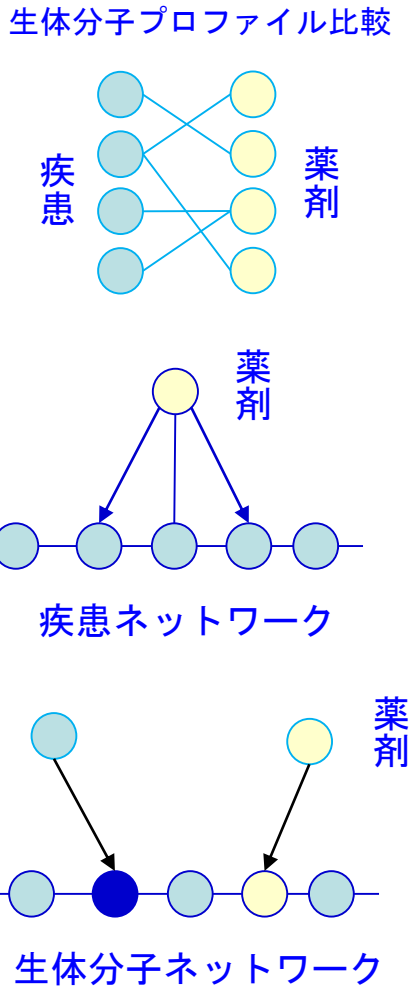
# 生体分子プロファイル型創薬/DR 方法論の深化

深化

**第１段階：疾患・薬剤プロファイル直接比較**
- 疾患罹患時と薬剤投与時の生体反応の遺伝子発現プロファイルを比較。
- パターン正負相関性に基づく有効性毒性予測

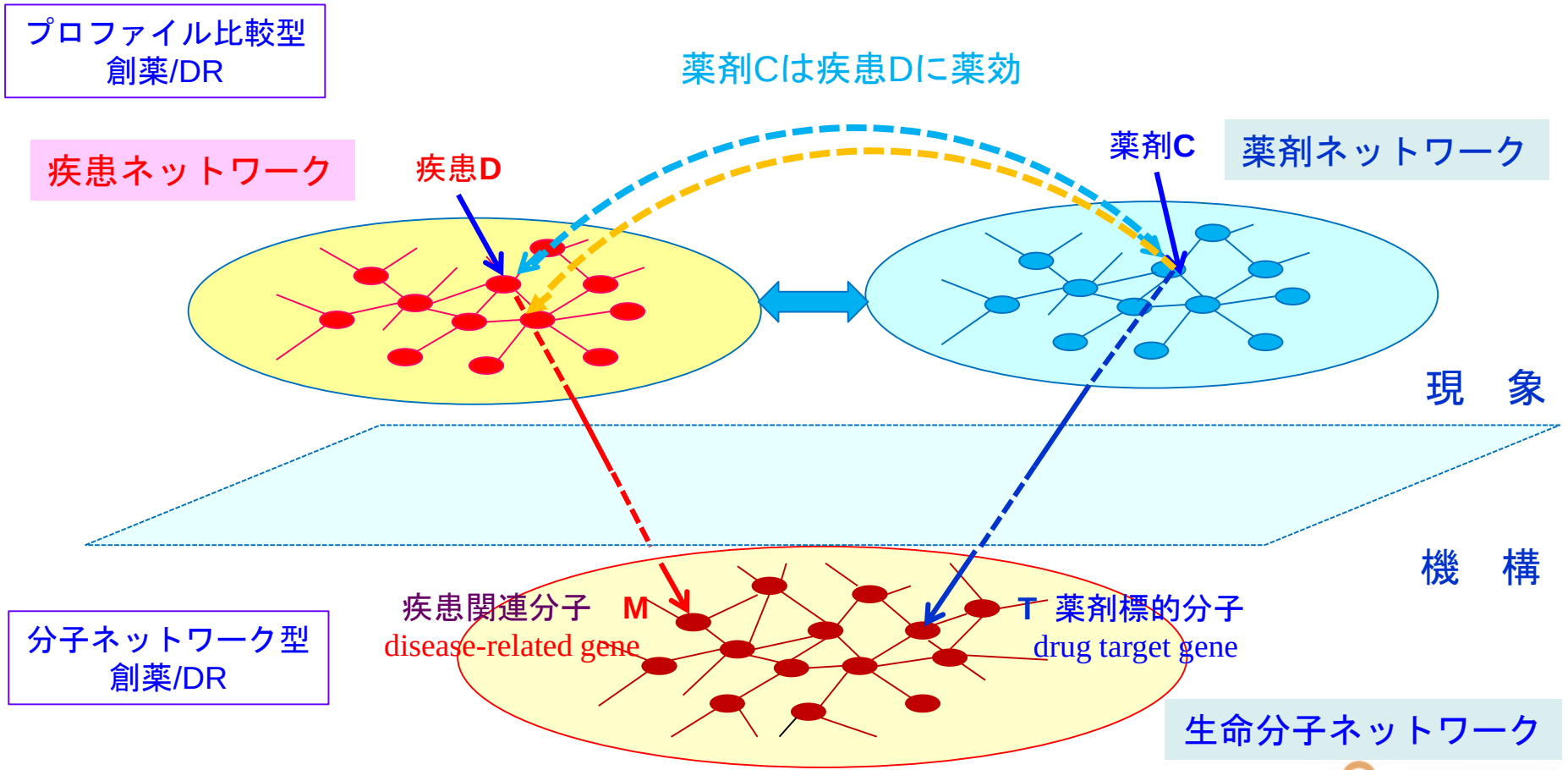
**第２段階：疾患・薬剤ネットワーク近接解析**
- 疾患あるいは薬剤の集合をネットワーク表現
- ネットワーク近接性に基き有効性・毒性予測

**第３段階：生体ネットワーク媒介型比較**
- 生体分子ネットワークを＜場＞として、疾患・薬剤の作用の足場分子を同定
- 足場分子間の相互作用（総合的距離）の評価に基づき有効性・毒性予測



# 生体分子プロファイル型計算創薬/DRの基本的枠組み

## 3層の生体・薬剤のネットワーク間の関係図式



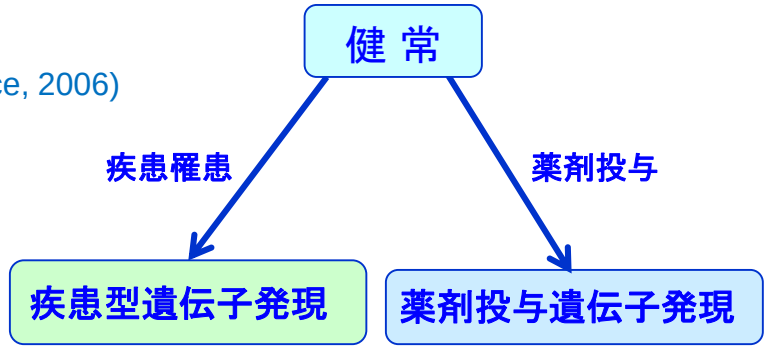
# ビッグデータ創薬/DRの方法

１．遺伝子発現プロファイル 創薬/DR
２．疾患ネットワーク創薬/DR
３．階層的ネットワーク創薬/DR
(Interactome 創薬/DR)

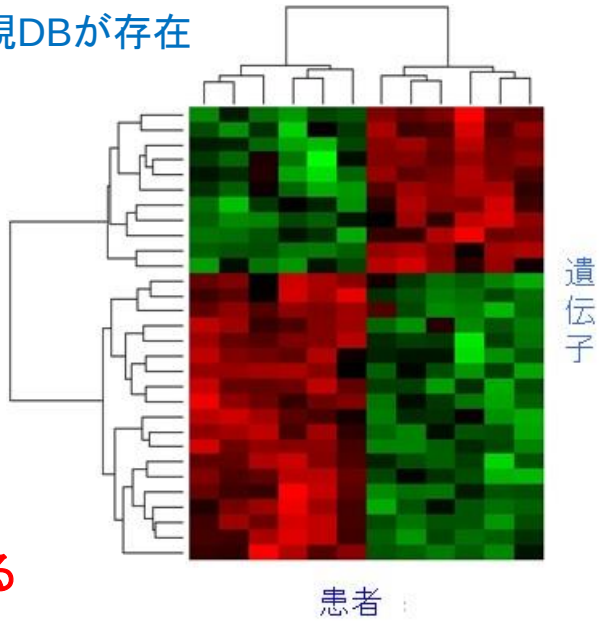
# １．遺伝子発現プロファイル比較型創薬・DR

# ビッグデータ計算創薬
# 発現プロファイル比較型創薬・DR

- **薬剤特異的遺伝子発現**
  - CMAP(Connectivity Map) (Lamb J, science, 2006)
    - 薬剤投与による遺伝子発現プロファイル変化
    - 米国 ブロード研究所,**1309化合物,**
      **5 種類のがんの培養細胞**
      約7000 遺伝子発現プロファイル
    - シグネチュア (“刻印”) 差異的発現遺伝子代表群
    - DB利用：シグネチャアを「問合せ」として投入
      類似性の高い順に化合物を提示
    - 最近はLINCSデータベース：100万種の薬剤特異的発現DBが存在



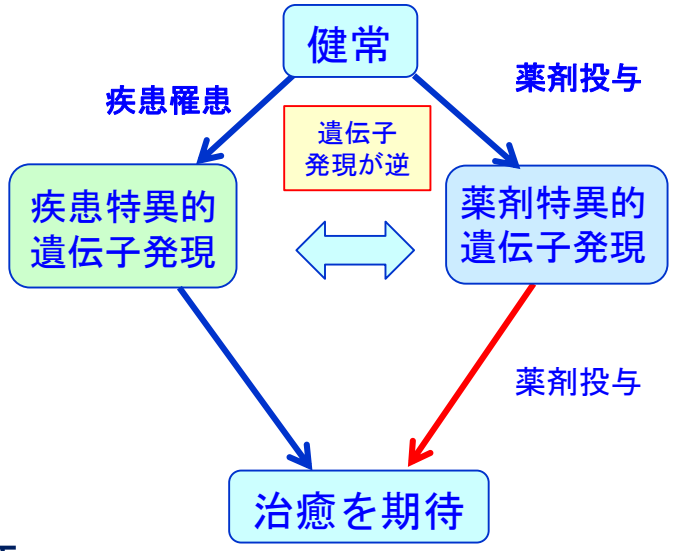

- **疾病特異的遺伝子発現**
  - GEO（Gene Expression Omnibus) (Barrett T, 2007)
    - 疾病罹患時の遺伝子発現プロファイルの変化
    - 米国NCBI作成・運用　2万5千実験,
      70万プロファイル（欧州　ArrayExpress）
    - EBIが作成、サンプル数同程度



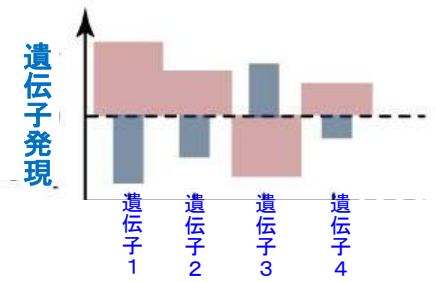

**基礎には分子ネットワークの疾病/薬剤特異的変化**
**遺伝子発現プロファイル変化**
**≈　分子ネットワーク活動構造変化を反映する**

# 遺伝子発現プロファイルによる有効性予測

- 遺伝子発現シグネチャア逆位法
  - Signature revision法（Irio, 2012）
  - 疾患によって健常状態から変異
    「**疾患特異的遺伝子発現プロファイル**」
  - これに薬剤投与の変化を起こす
    「**薬剤特異的遺伝子発現プロファイル**」
  - 両者のパターンが負に相関する
  - ノンパラメトリックな相関尺度で評価

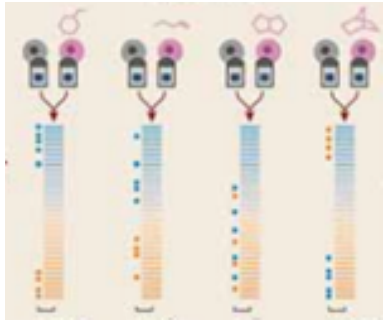


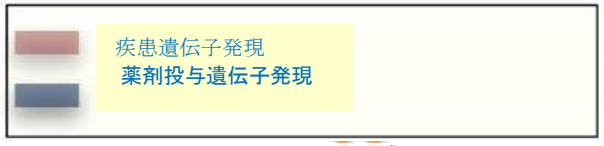

- 効果がお互いに打ち消すなら**有効性**が期待される
  - 例：炎症性腸疾患に 抗痙攣剤(topiramate),
  - 骨格筋委縮にウルソール酸




リストは疾患遺伝子発現の差異的発現順
横の点は薬剤遺伝子発現のシグネチャア

青は発現が**上昇**した遺伝子
赤は発現が**下降**した遺伝子

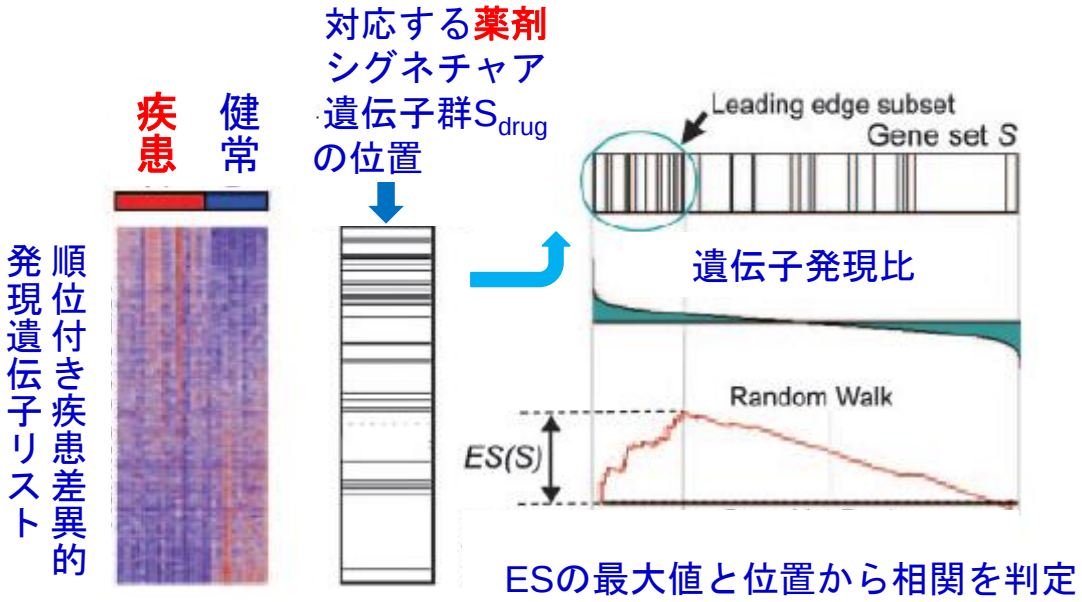


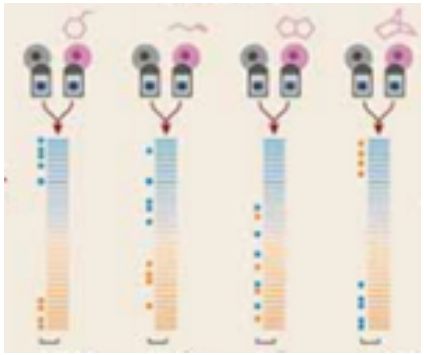



# Non-parametric な相関尺度で評価

**Gene Set Enrichment Analysis（GSEA)**

– 対照と比較して順位づけられた遺伝子リストの上位に密集している度合いの尺度

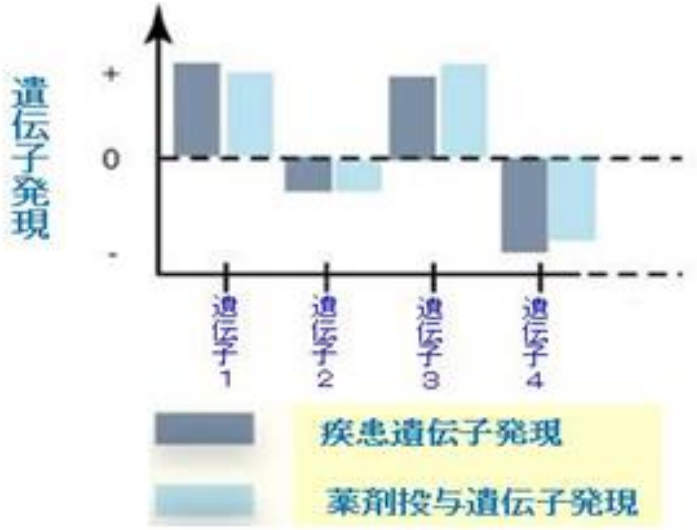


発現比ランクの高い順から遺伝子を調べ、薬剤特異的発現遺伝子シグネチャアリスト$S_{drug}$中の**該当する遺伝子のところで遺伝子発現比を加算**、無ければ減算して、最大累積値を**ES (Enrich Score)**とする

# 遺伝子発現プロファイルによる毒性予測

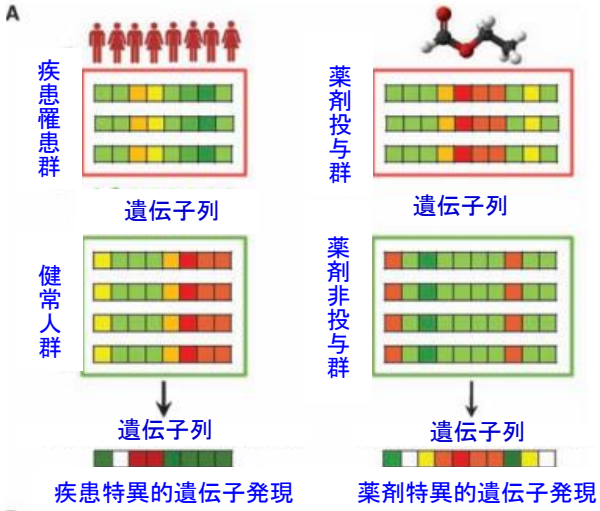
- 連座法 guilt-by-association ：
- **薬剤ー疾患間　副作用予測**
  - 薬剤特異的シグネチュアと
  - 疾患特異的シグネチャアが
  - ノンパラメトリック相関　正
  - 毒性・副作用の予測
- 薬剤ー薬剤間
  - 薬剤ネットワークからのDR
  - Connectivity map から薬剤特異的遺伝子発現の薬剤間の親近性をノンパラメトリック親近性尺度（GSEA)で評価
  - この類似性のもとに薬剤ネットワーク構築
    近隣解析によりDR
  - 例：抗マラリア剤をクローン病に適応



# 遺伝子発現プロファイルの正・負のパターン相関性に基づく計算DR

(Sirota, Butte 2011)

- NCBI・**GEO**から**100疾患**のシグネチャアを取得
- **c-Map**より得た**164の薬剤・化合物**の
  薬剤特異的遺伝子発現プロファイル
  疾患-薬剤間で類似性スコアを計算
- 約16000組の疾患-薬剤間の2664組が
  有意、半数以上が治療的関連(負)あり
- 100疾患内, 53疾患が有意に164薬剤と関連



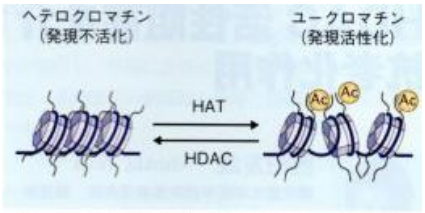

適応の多い薬剤と疾患

| Drugs with most indications | | Diseases with most indications | |
|---|---|---|---|
| Vorinostat **HDAC阻害剤** | 21 | Transitional cell carcinoma | 95 |
| Gefitinib | 18 | Melanoma | 79 |
| HC toxin | 18 | Cardiomyopathy | 73 |
| Colforsin | 17 | Adenocarcinoma of lung | 72 |
| 17-Dimethylamino-geldanamycin | 16 | Multiple benign melanocytic nevi | 68 |
| Trichostatin A | 16 | Squamous cell carcinoma of lung | 67 |
| 3-Hydroxy-DL-kynurenine | 15 | Malignant neoplasm of stomach | 66 |
| 5114445 | 15 | Dermatomyositis | 63 |
| Dexverapamil | 15 | Malignant mesothelioma of pleura | 53 |
| Prochlorperazine | 15 | Primary cardiomyopathy | 48 |

(Sirota, Butte 2011)

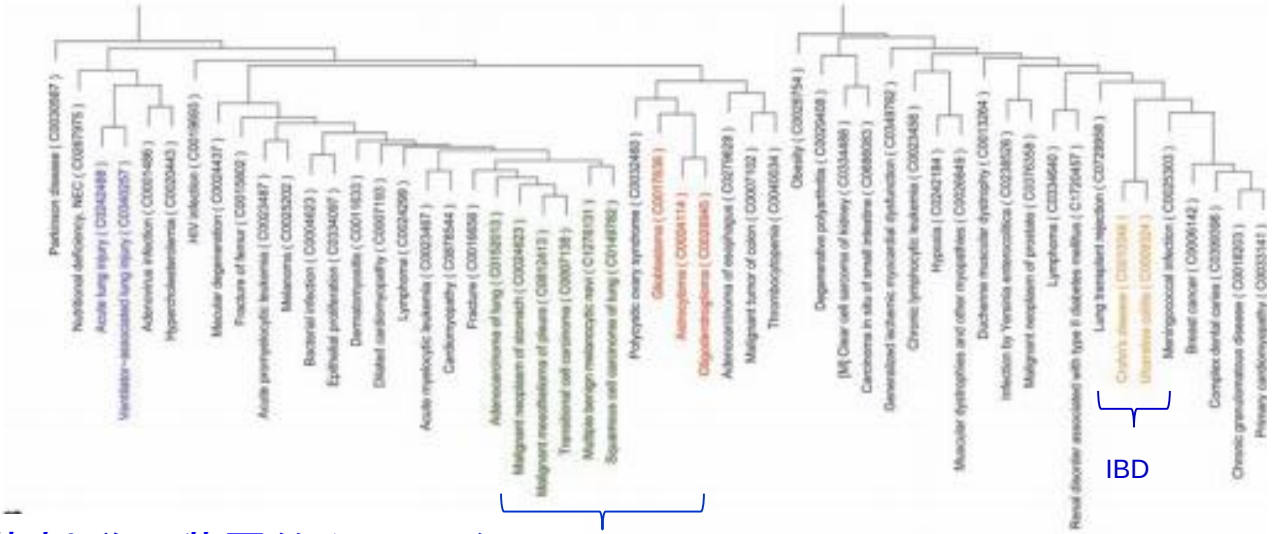
| Drug group | Drugs |
|---|---|
| PI3K inhibitors | LY-294002 and wortmannin |
| HSP90 inhibitors | Geldanamycin, raloxifene, monorden, and sodium phenylbutyrate |
| HDAC inhibitors | Vorinostat, HC toxin, and trichostatin A |
| Salicylate anti-inflammatory agents | Sulfasalazine, mesalazine, and acetylsalicylic acid |

| Canonical | Noncanonical |
|---|---|
| Cancers | Crohn's disease and lung transplant |
| Ulcerative colitis and Crohn's disease | Polycystic ovary and glioblastoma |
| | Cardiomyopathy and cancer |



# 遺伝子発現プロファイル比較による
# 疾患－薬剤関係に基づく計算/DR

## 疾患群の階層的クラスター

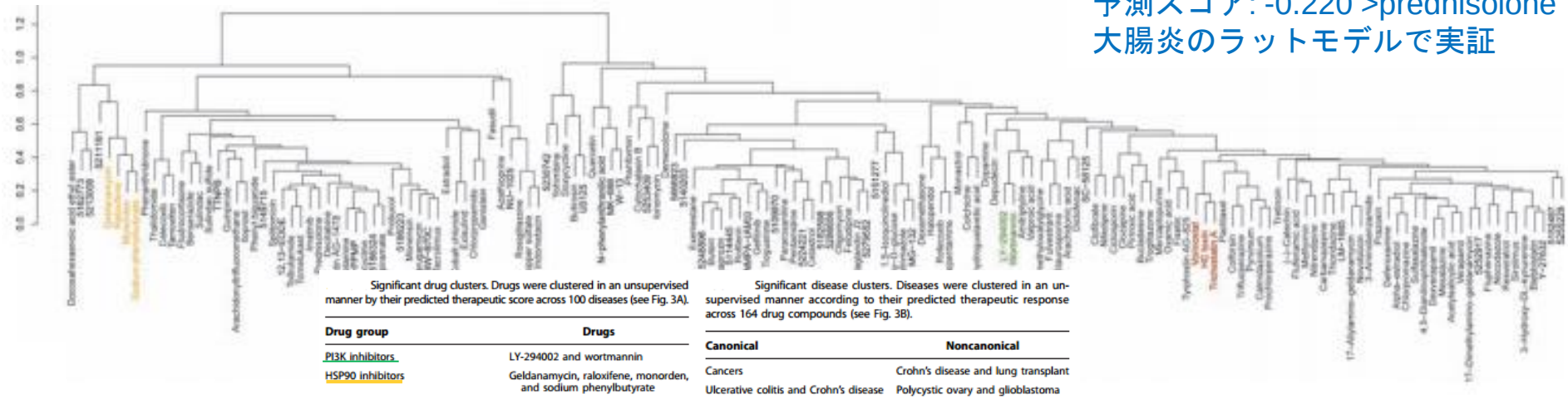



**疾患のクラスター解析**
- がんの大クラスター
- IBD: 潰瘍性大腸炎、クローン病のクラスター

**薬剤のクラスター解析**
- HDAC阻害剤
  HC toxin, trichostatin
- PI3K, 抗炎症剤クラスター
- HSP90関連薬剤

**DR候補**：
Topiramate(抗痙攣剤) IBDに有効
予測スコア: -0.220 >prednisolone
大腸炎のラットモデルで実証

## 薬剤群の階層的クラスター



Significant drug clusters. Drugs were clustered in an unsupervised manner by their predicted therapeutic score across 100 diseases (see Fig. 3A).

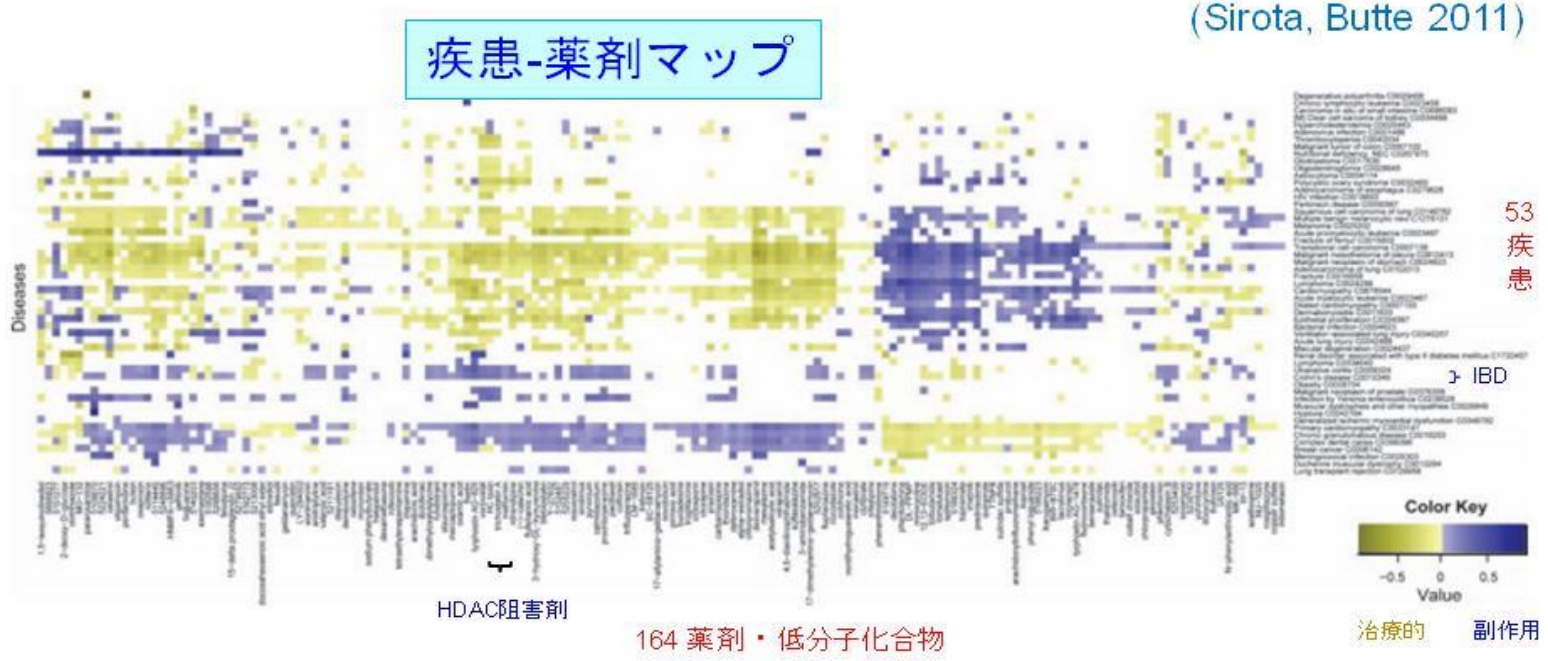
| Drug group | Drugs |
|---|---|
| PI3K inhibitors | LY-294002 and wortmannin |
| HSP90 inhibitors | Geldanamycin, raloxifene, monorden, and sodium phenylbutyrate |
| HDAC inhibitors | Vorinostat, HC toxin, and trichostatin A |
| Salicylate anti-inflammatory agents | Sulfasalazine, mesalazine, and acetylsalicylic acid |

Significant disease clusters. Diseases were clustered in an unsupervised manner according to their predicted therapeutic response across 164 drug compounds (see Fig. 3B).

| Canonical | Noncanonical |
|---|---|
| Cancers | Crohn's disease and lung transplant |
| Ulcerative colitis and Crohn's disease | Polycystic ovary and glioblastoma |
| | Cardiomyopathy and cancer |

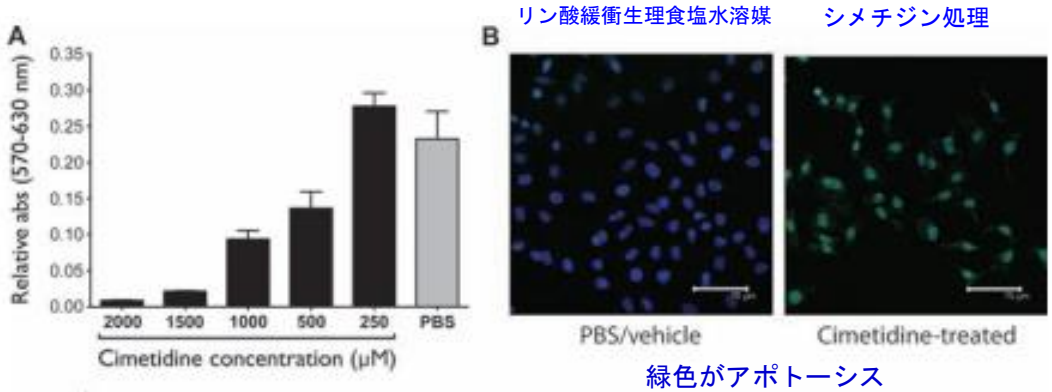
# 遺伝子発現プロファイルによる薬剤の有効性・毒性の俯瞰

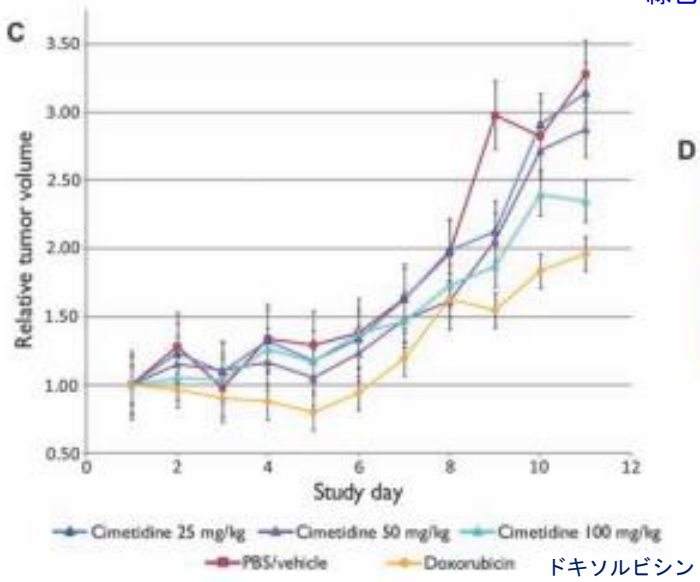
- 疾患罹患時と薬剤投与時の遺伝子発現プロファイルの
  正・負のパターン相関性を基準
- ＜疾患－薬剤＞の**「有効性・毒性」予測俯瞰図**



# 動物実験での実証

シメチジン(cimetidine:ヒスタミンH2受容体拮抗薬）→肺腺癌(LA)に有効か
予測スコア -0.088 であったが gefinitib (イレッサ）の-0.075より高い





左：シメチジン処理　右：対照（マウス）

# 遺伝子発現プロファイル による疾患ー薬剤ネットワーク

遺伝子発現プロファイルの類似性を相関係数、ESによってリンク （Hu, Agarwal, 2009）

疾患ー疾患、薬剤ー薬剤、疾患ー薬剤の
ネットワークを発現プロファイルより構成

**疾患**（disease-disease）645 組
**疾患- 薬**（disease-drug) 5008 組
**薬 - 薬**（drug-drug) 164,374 組

**結果**

①**疾患-疾患NWの60%はMeSH** (既知体系)
その他は**分子レベル疾患分類学**
遺伝子発現の類似性による疾患体系

②**主な発見**

**＜疾患 - 疾患＞**

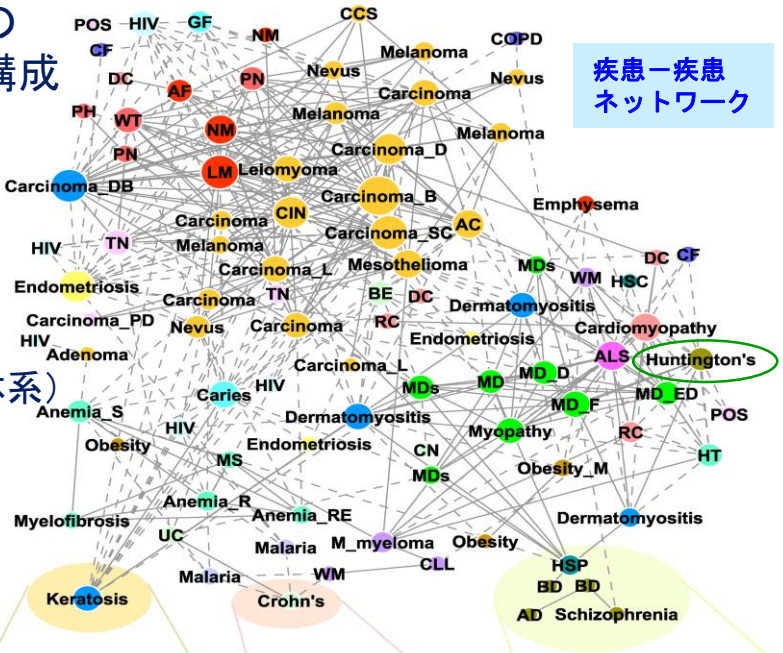
HSP（Hereditary Spastic Paraplegia
（遺伝性痙攣性対麻痺）
⇒bipolar 双極性障害

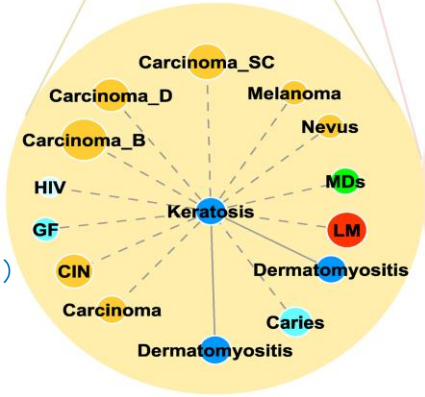
Solar keratosis 日光性角化症
⇒ cancer(squamous)

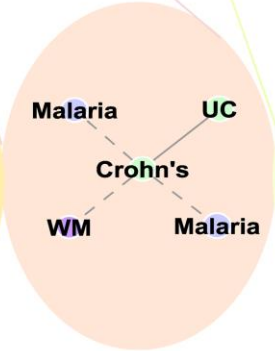
**＜疾患 - 薬＞**

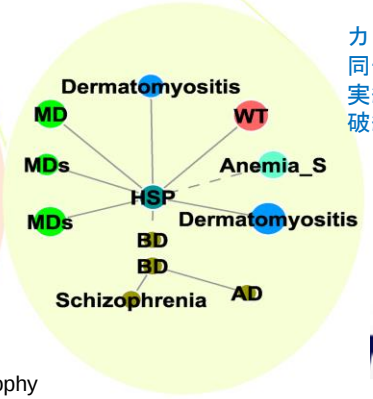
有効性：マラリア治療薬
⇒ Crohn's disease

（ベトナム経験：クローン病罹患保護効果）
ハンチントン病に種々の薬剤



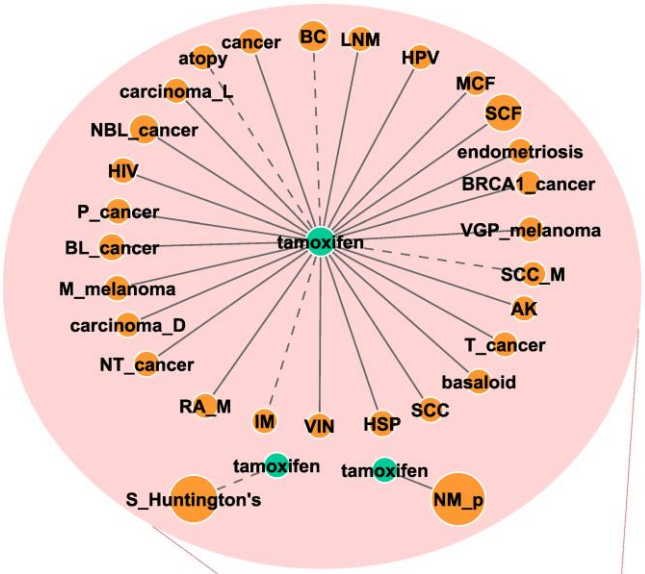

カラーはMeSH
同一カテゴリー
実線はMeSH内
破線はMeSH外

MD: muscular dystrophy

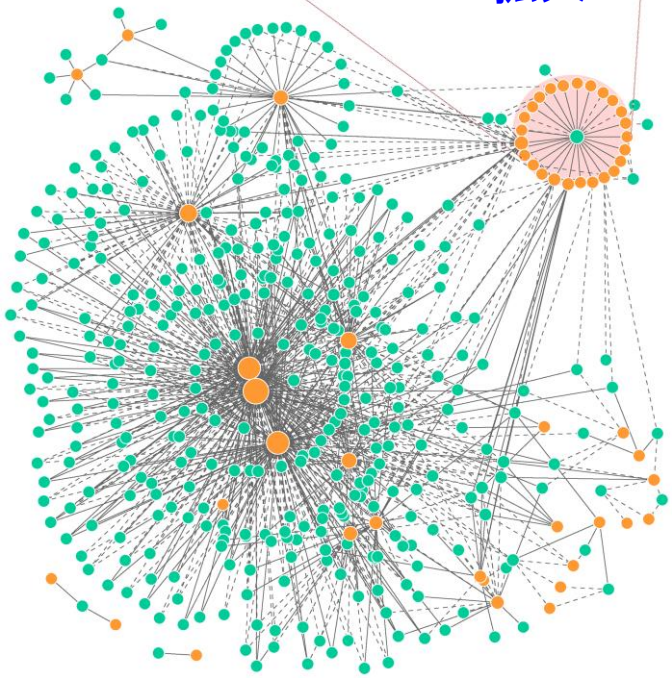
# 遺伝子発現プロファイルによる疾患-薬剤ネットワーク

疾患－薬剤ネットワーク
（Disease-drug network:右図）
橙色　49 疾患
緑色　216 薬剤
(全体で906対　疾患－薬剤結合）

Tamoxifen (乳がんのホルモン療法薬)
有効性（破線：負の値をもっている）
⇒アトピー,
マスト細胞分泌抑制、
アレルギー抑制
⇒ Hunting病に多数のDR薬
副作用（実線：正の値をもっている）
副作用の予測⇒多くのがん
⇒ 発癌性



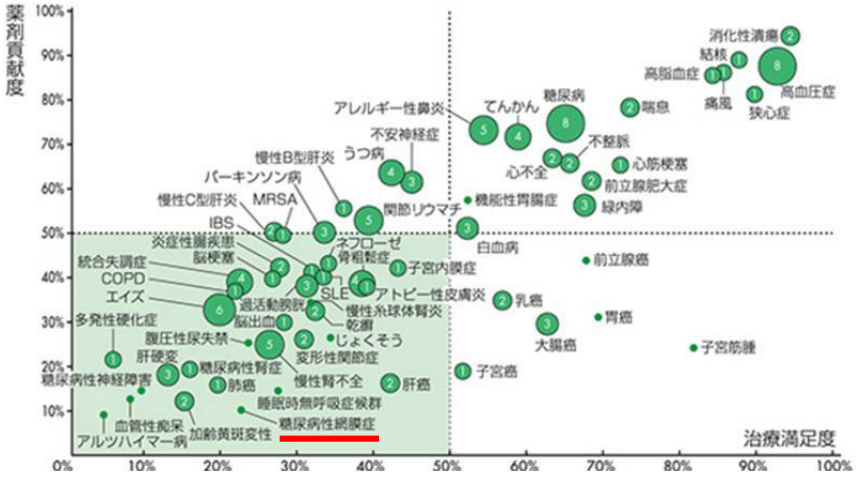

拡大



疾患－薬剤ネットワーク

# 我々の研究室での成果

- 対象疾患　(Shibata et al. 2016)
  - 薬剤貢献度と治療満足度がともに低い糖尿病性網膜症 (diabetic retinopathy) の薬剤探索
- 方法
  - Signature revision法を適用
  - **疾患特異的遺伝子発現**
    - GEOから糖尿病性網膜症の遺伝子発現プロファイルを収集 (GSE53257)
    - 対照：16サンプルの健常例
    - 206遺伝子 疾患signature を確定
      - 130 up-regulated
      - 76 down-regulated genes
  - cMAPより疾患と負値ESの薬剤特異的発現を提示する有意な薬剤を探索
- 結果
  - 1600組のなかで37組の＜疾患 - 薬剤＞が有意、その中でも**11剤が負値のES**
  - FDR（*q*値）＜0.005
  - thapsigargin (score -0.983, p-value 0.00002),
    **タプシガルギン（小胞体ストレス誘導）**
  - alprenolol (score -0.892, p-value 0.00026),
    ionomycin (score -0.896, p-value 0.00208),
    phenylpropanolamine (score -0.814, p-value 0.00219)
    など
- 考察
  - thapsigargin ：endoplasmic reticulum (ER) ストレスに関与。ER stress は NF-kB を活性化
  - 糖尿病性網膜症は本質的には炎症反応
  - NF-kB は the unfolded protein response (UPR) で制御されている .
  - ER stress がこの炎症の制御に役立つ可能性がある




ライフサイエンス振興財団

**糖尿病性網膜症のDR候補化合物**

| | 薬剤 | SCORE | p 値 |
|---|---|---|---|
| 1 | thapsigargin | -0.983 | 0.00002 |
| 2 | alprenolol | -0.892 | 0.00026 |
| 3 | ionomycin | -0.896 | 0.00208 |
| 4 | phenylpropanolamine | -0.814 | 0.00219 |
| 5 | etiocholanolone | -0.621 | 0.00961 |
| 6 | kinetin | -0.72 | 0.01249 |
| 7 | triflupromazine | -0.706 | 0.0155 |
| 8 | vanoxerine | -0.681 | 0.02274 |
| 9 | cicloheximide | -0.657 | 0.03185 |
| 10 | khellin | -0.579 | 0.03975 |
| 11 | rotenone | -0.625 | 0.04852 |

TMDU

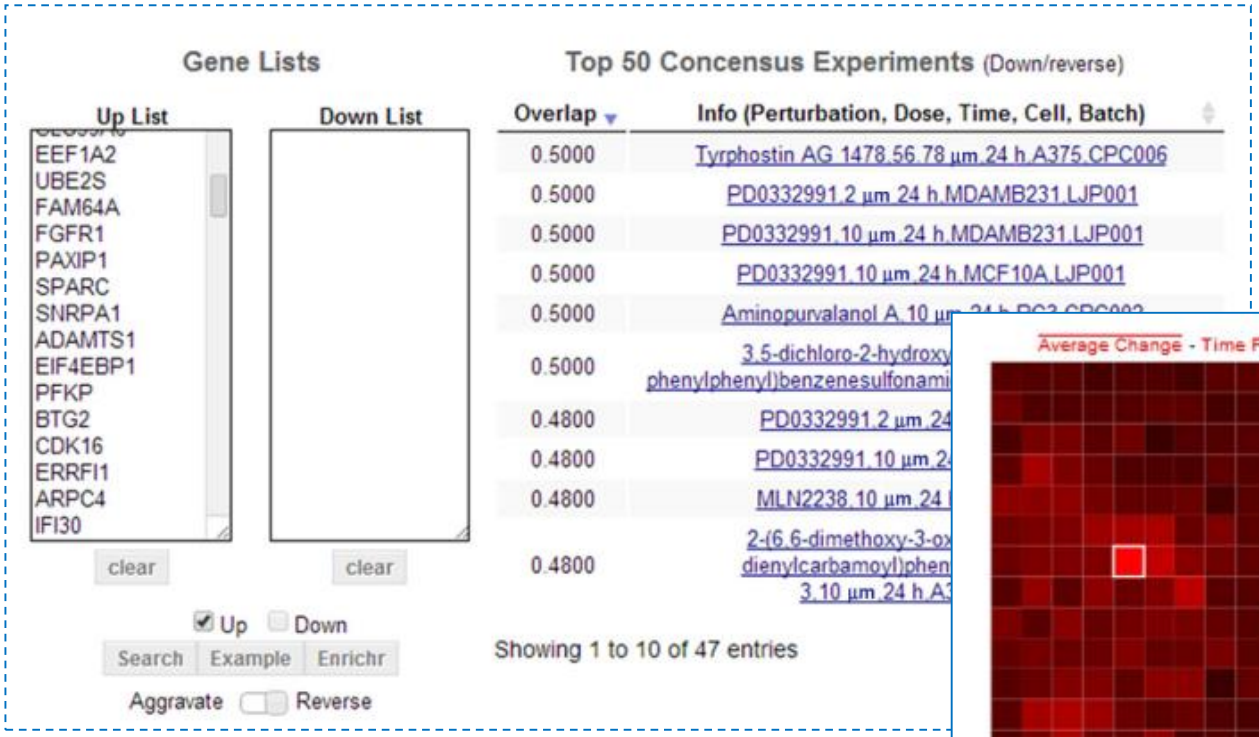
# 近年のビッグデータ化
# LINCS

- **LINCS** (library of Integrated network-based cellular signatures)
  - GE-HTS(gene expression high throughput screening)の１つ
  - 摂動（化合物添加）を与え調節系を介して、細胞表現型を観察する
  - 遺伝子発現変化⇒差異的発現　**signature**
  - cMAP (2006, Lamb)に比べてスケール拡大 (Duan, 2014)
    - cMAPは、４つの細胞系列～1300化合物　FDA認可薬剤
      Micro array (mRNA)  Affymetrix U 113で遺伝子発現測定
- NIHから助成, **百万の遺伝子発現プロファイル**を **L1000 技術**で測る
  - Broad Institute cMAPと同じメンバーが考案
  - 1000遺伝子の発現しか測定しない　ゲノムワイドな遺伝子発現プロファイル（～全遺伝子22000 genesの発現）をGEOから作ったモデルで推定する
  - 相互依存性高い⇒1000遺伝子にすべて情報が含まれている
- **L1000技術**
  - 細胞溶液からリガンド媒介増幅によってmRNA増幅
  - 遺伝子特異的なProbeはcDNA（mRNA) にtaqリガーゼでアニールする
  - ProbeはPCRで増幅され、ルミネックスビーズと遺伝子特異的部分で対形成する
  - 対形成した差異染色ビーズはレーザを用いて検出され定量化される
  - ビーズの上の対形成したprobeの密度を測る　80の恒常的発現校正遺伝子
- **22412  摂動遺伝子発現**
  - 56　細胞コンテキスト（ヒト初代培養細胞、がん培養細胞）について
  - 16425　化合物、薬剤
  - 5806　遺伝子ノックアウト(RNAi, miRNA)、過剰発現
  - 総計で100万ぐらい遺伝子発現プロファイルがある
- **Genometry がL1000™ Expression Profiling技術でヤンセンと契約**
  - 25万種類の化合物

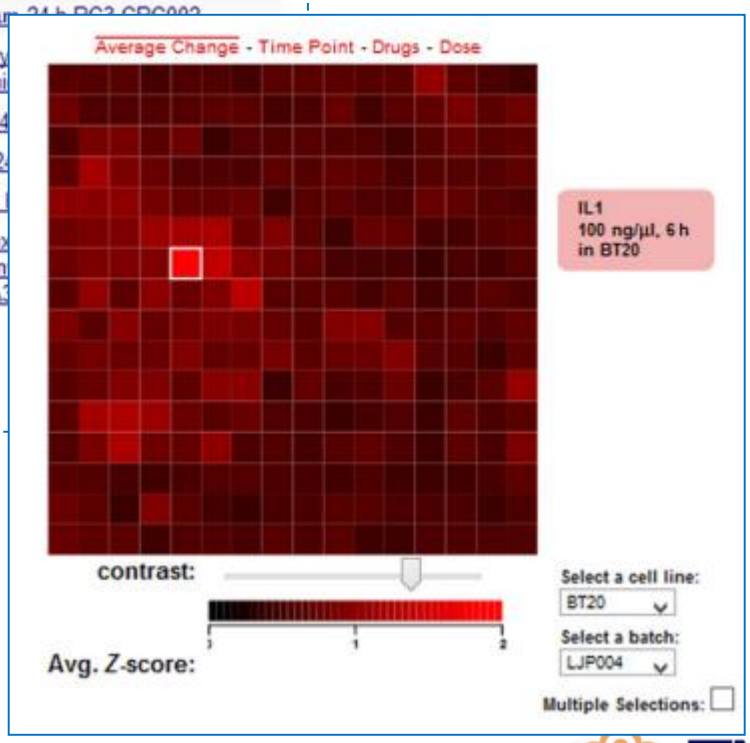
TMDU

# LINCSの問合せ画面

## --- LINCS Canvas Browser ---




実験キャンバス表示



# 2. 疾患ネットワーク創薬/DR

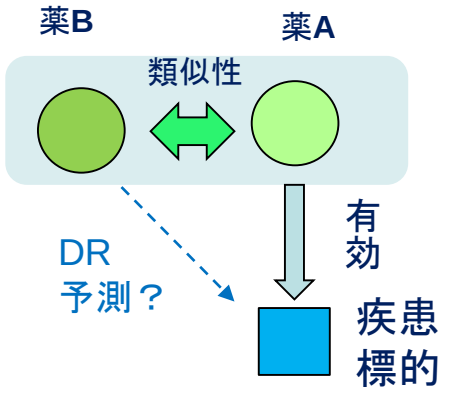
疾患ネットワーク空間を基礎にした
ビッグデータ創薬/DR
＜疾患ネットワークでの近接性＞

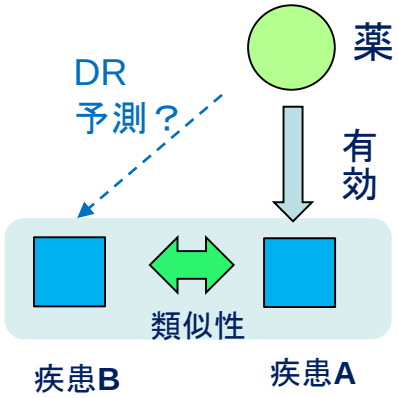
# 遺伝子発現プロファイルから疾患・薬剤ネットワークへ

- 遺伝子発現プロファイルを基礎にした「シグネチャア逆位法」
  - 広くDR候補を**枚挙する**方法としては有効
  - もう少し**対象を絞り込む**方法はないか
  - 疾患の**内因的機序**(ゲノム・オミックス機序)によるDRは可能か
- 疾患のゲノム・オミックス機序による**疾患間のネットワーク化**
  - **同様に薬剤間のネットワーク化も可能**

# 疾患・薬剤ネットワークへのアプローチ

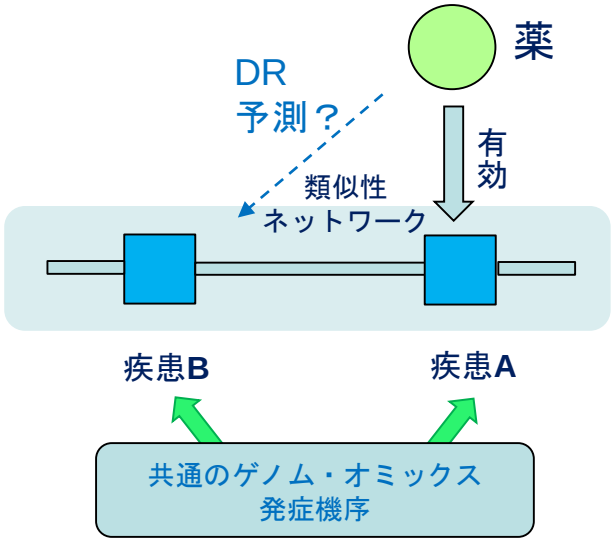
- 医薬品中心 Drug-based (drug-centric)
  - 医薬品の構造・特徴の類似性に基づいて別の医薬品の適応を予測
    ① 化合物の化学的構造・特徴の類似性
    ② 薬物投与時の遺伝子発現プロファイル
- 疾患中心 Disease-based(disease-centric)
  - 疾患の発症機序の類似性に同一の医薬品が別の疾患の適応を予測
    ① 疾患原因/感受性遺伝子の共有
    ② 疾病遺伝子発現プロファイル
    ③ 疾患を起こす分子ネットワークの類似性
- 両者の融合的アプローチ



# ビッグデータ創薬/DRの基本原理２
# 疾患ネットワーク準拠創薬/DR

- 従来の疾患体系 nosology
  - Linne以降300年に亙って表現型による疾病分類
  - 臓器別・病理形態学別の疾患分類学
- **ゲノム・オミックスレベルでの発症機構での疾患分類**
  - 発症の**内在的 (intrinsic)機構の類似性**を**基準に疾患ネットワーク**（疾患マップ）をつくる
  - ゲノム・オミックスによる内在的疾病機序の概念が基礎



# 疾患形成のゲノム・オミックス機序

- 疾患関連遺伝子型（第一世代型）
  - 原因遺伝子、疾患感受性遺伝子の変異・多型が主要発症機序
- 疾患オミックス型（第２世代型）
  - 疾患オミックスプロファイルの変容が主要発症機序、「分子表現型 molecular phenome」
  - Trans-disease omics
- 疾患分子ネットワーク型（第３世代型）
  - 「分子ネットワークの歪み」が主要発症機序
  - 大半の疾患（先天的稀少疾患を除く）, ”common disease”はネットワークの歪み

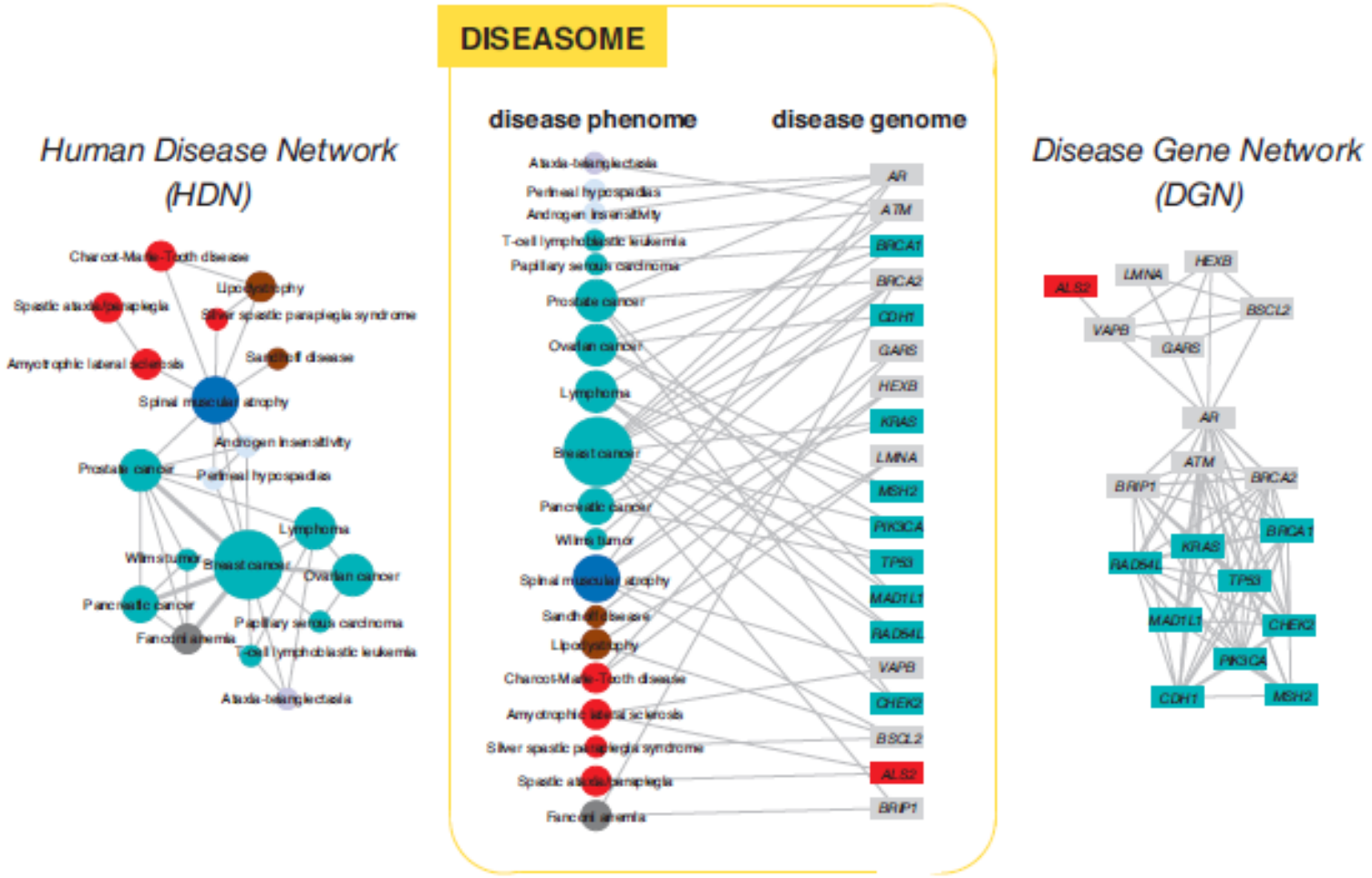
TMDU

# 第１世代型
# Diseasomeと疾患遺伝子

- **OMIM**から 1,284 疾患と 1,777 疾患遺伝子を抽出
- **ヒト疾患ネットワーク（HDN)**
  - 867疾患は他疾患へリンクを持つ　細胞型や器官に非依存
  - 516疾患が巨大クラスターを形成
    - 大腸がん、乳がんがハブ形成
    - がんはP53 やPTENなどにより最結合疾患　がんなどは後天的変異
  - 疾患を網羅的に見る見方：臓器や病理形態学に非依存
  - リンネ（12疾患群分類）以来300年続いた分類学を越える
- **疾患遺伝子ネットワーク（DGN)**
  - 1377遺伝子は他の遺伝子へ結合
  - 903遺伝子が巨大クラスター
    - P53がハブ
- ランダム化した疾患/遺伝子ネットワークに比べ
  - 巨大クラスターのサイズが**有意に小さい**
- 疾患遺伝子は機能的なモジュール構造
  - 同じモジュールに属する遺伝子は相互作用し
  - 同一の組織で共発現し、同じ**GO**（遺伝子オントロジー）を持つ

TMDU

# 疾患ネットワーク Diseasome




１つ以上の疾患関連遺伝子を共有する疾患

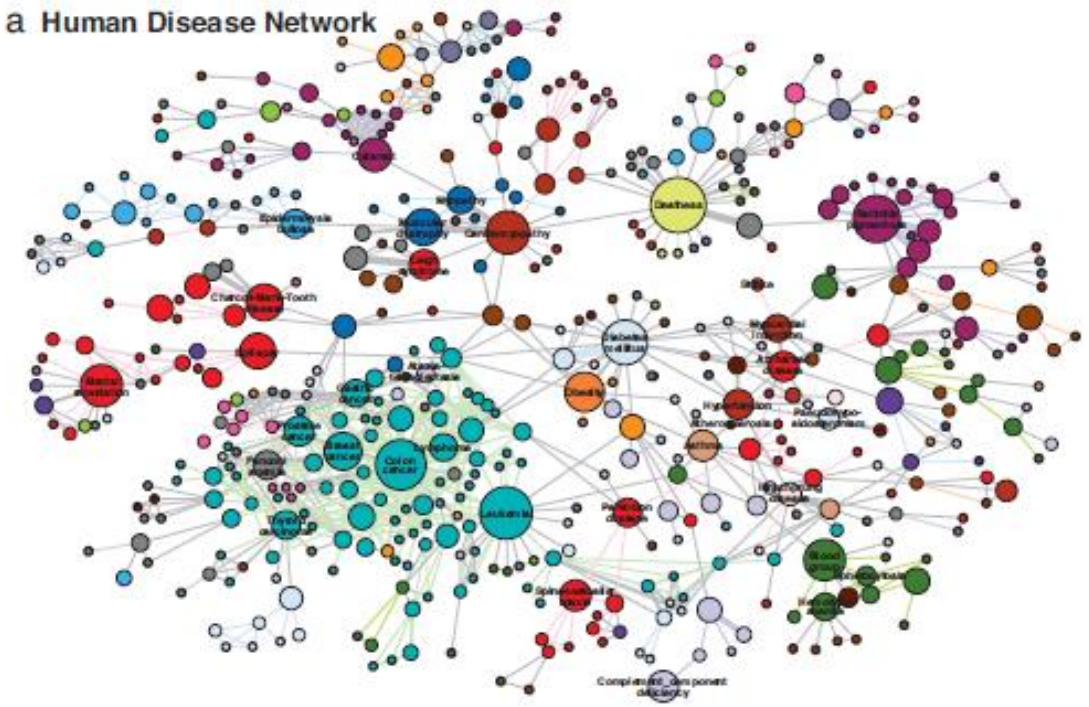
１つ以上の疾患を共有する疾患関連遺伝子

Kwang-Il Goh*, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-Laszlo Barabasi The human disease network　PNAS, 2007

TMDU

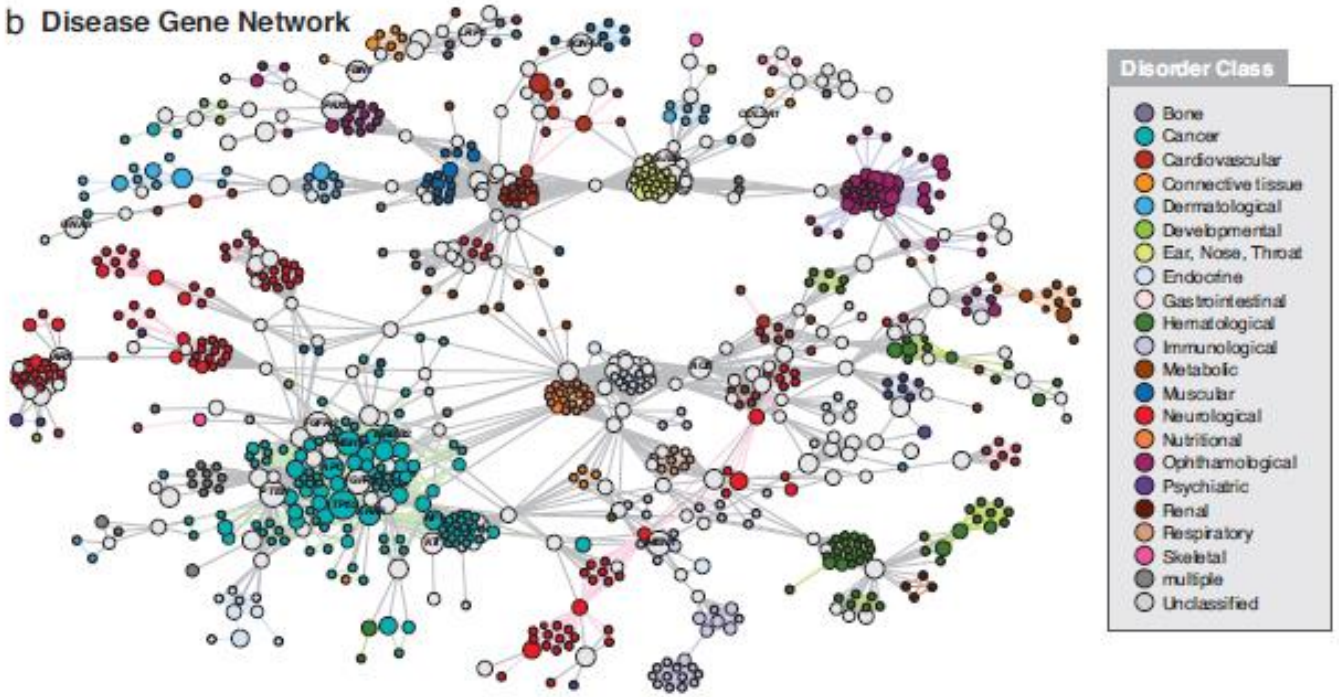
疾患
ネットワーク
(HDN)

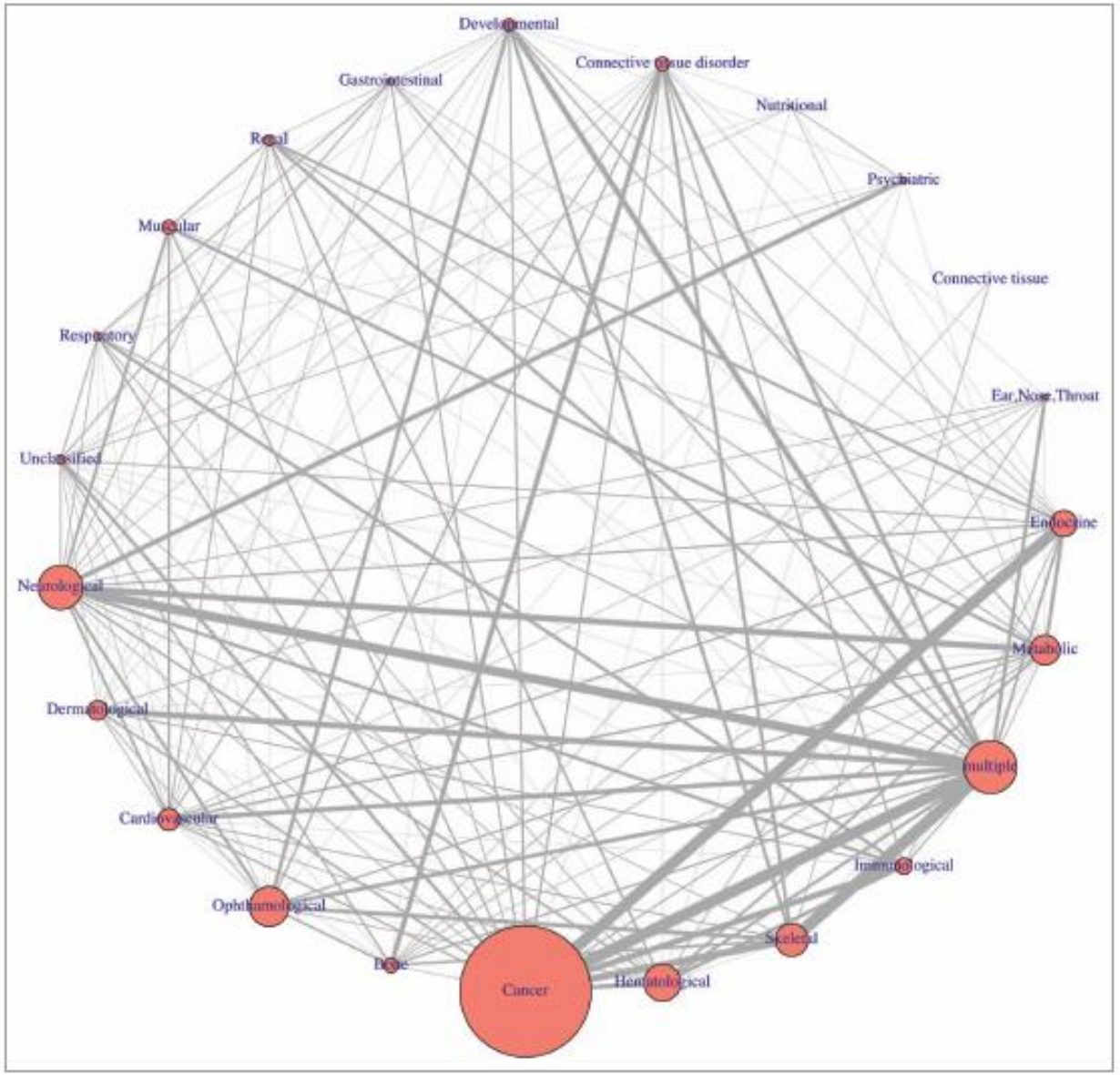
Nodeの直径
疾患に関与している原因遺伝子の数に比例
リンクの太さ
疾患間で共有している原因遺伝子の数




疾患遺伝子
ネットワーク
(DGN)

Nodeの直径
その遺伝子を原因にしている疾患の数に比例
2つ以上の疾患に関与していると明灰色の遺伝子ノード






23グループのOMIM疾患のネットワーク

ノードの径：共通リンクの合計、リンクの太さ：共通遺伝子数

TMDU

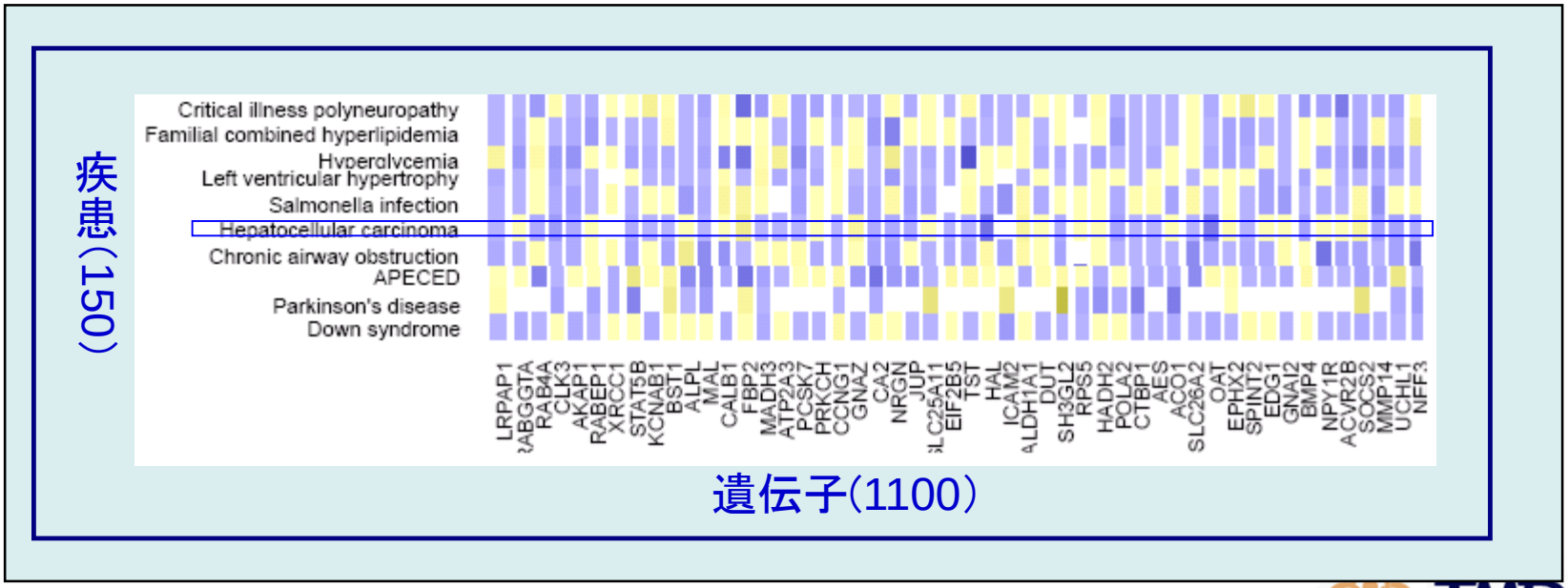
# Diseasomeを巡る状況

- Mendel疾患、複雑疾患、環境疾患へと発展
- **他のネットワークと融合**
  - タンパク質相互ネットワーク、代謝ネットワーク
    - PPIの近傍(Vanunu)、代謝網での酵素の基質の共有
  - GWAS（WTCCC, NIH-GAD)のSNPの共有
    - すべてがつながり偽陽性のネットワークで有効性低い
  - miRNA, 環境因子（annotation MEDLINE）
    - 電子カルテから時系列病歴収集
    - 進化的直系的表現型性（他の動物も利用）
  - パスウェイ準拠型の疾患ネットワークも
- **表現型疾患ネットワークも存在する**
    - Phenotype：MeSHの頻度をベクトルとする(van Driel)
- Diseasomeは、臨床表現型NWと分子NWを繋げる機構
    - 遺伝子を通して疾患間を移動できる
    - Systems pathobiology, nosology, personalized medicine

# 第２世代型

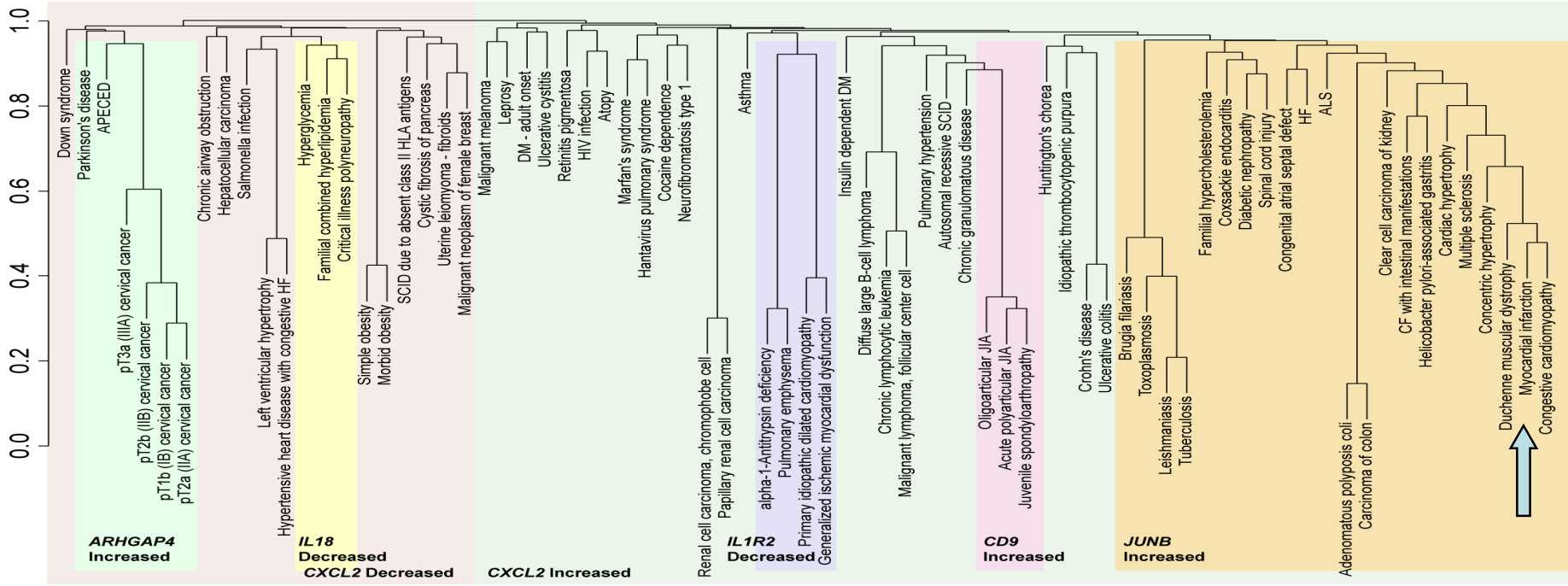
# **GENOMED** (A.Butte et al)

- 遺伝子発現DBのGEO（Gene Expression Ominibus)利用
  - 約20 万のサンプル
- 疾患名は注釈文より用語集UMLSを用いて抽出
- 疾患ごとに多数の遺伝子発現パターンを平均化



# Gene-Expression Nosology of Medicine

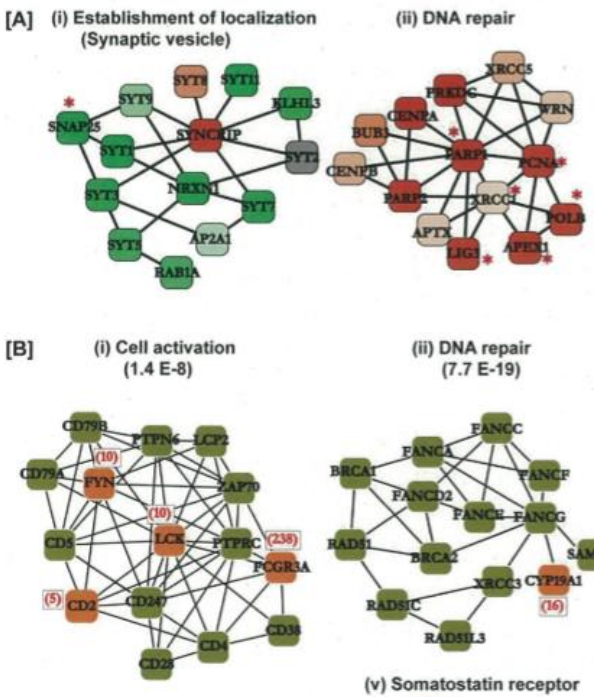
- 疾患を平均遺伝子発現パターンよりクラスター分類
  - 臓器別疾患分類では予想できない疾患間の親近性
  - 分類項目はサイトカインの遺伝子発現と相関
  - 疾患の再体系化に基づいた医薬の repositioning
- さらに656種類の臨床検査を結合した分析
- 心筋梗塞・デュシャンヌ型筋ジストロフィーが近い



# 遺伝子発現プロファイル変化をPPIに投影した疾患ネットワーク

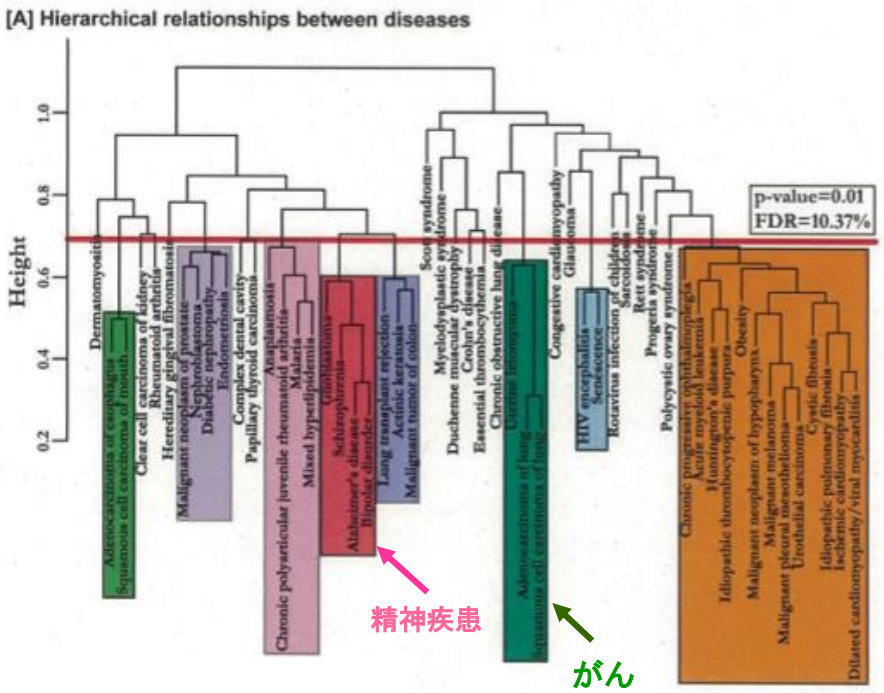
(Suthram,Butte, 2010)

- **ネットワークモジュール**
  遺伝子発現プロファイルではなく4620 に分解したPPIネットワークの機能moduleでの疾病での平均発現変化をもとに疾患ネットワーク構築
- **基本方法**
  - GEOから信頼性などより54の疾患を選択
  - 各疾患について各moduleに含まれる遺伝子群の
    **疾患時と健常時の発現差の*t* 統計量の平均**
  - MRS: Molecular Response Score 各疾患に各モジュールで定義（ベクトル量）
  - 疾患間の相関は、両疾患の健常時発現を制約とした
    MRSの偏相関係数
- **疾患ネットワークの性質**
  - 138の有意な相関：ランダム化ネットに対し有意
  - p<0.01,FDR=0.1
  - 疾患類似性: 肺がん群：修復パスウェイ,精神疾患:SNP-25
- 138の疾患相関のうち
  - 17は少なくとも1つの共通薬（14は共通の薬剤に有意）
  - Flourarcil（日光性角化）⇒大腸がん、ほかDoxorubicin
- 疾患の大半を占める59モジュール
  - **「共通疾患シグネチャア」**

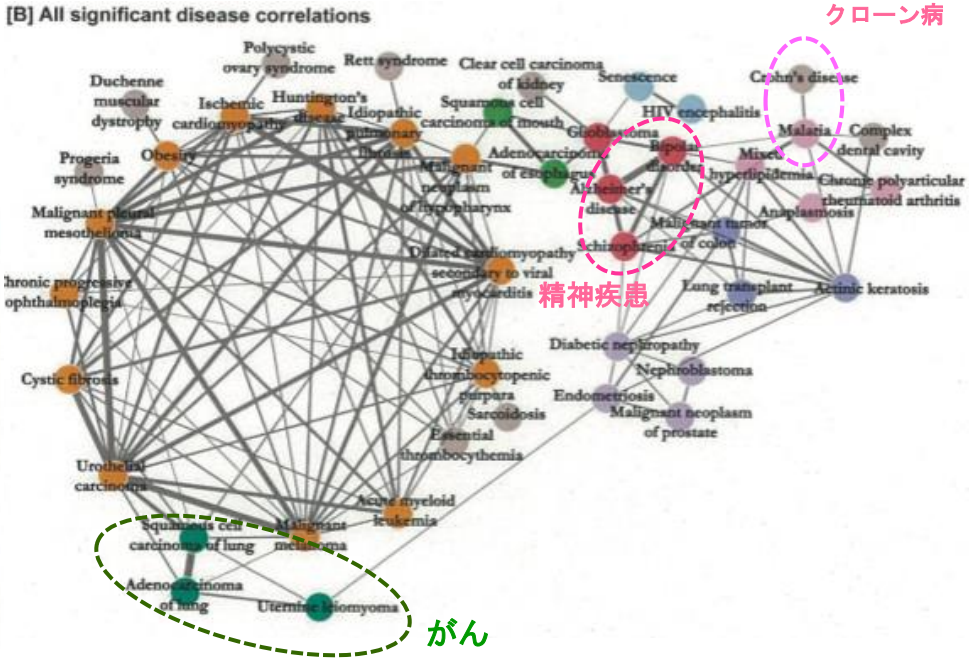


# 遺伝子発現の変化をPPIに投影した
# 疾患ネットワーク

- アルツハイマー症、統合失調症、双極性障害がグループ化
- 子宮筋腫と肺がん、マラリアとクローン病
- 17のがんが１つの群ではない。がんの異質性
- 疾患ネットワーク間の遺伝子共有は高くない（遺伝子外効果）




階層的クラスタリング

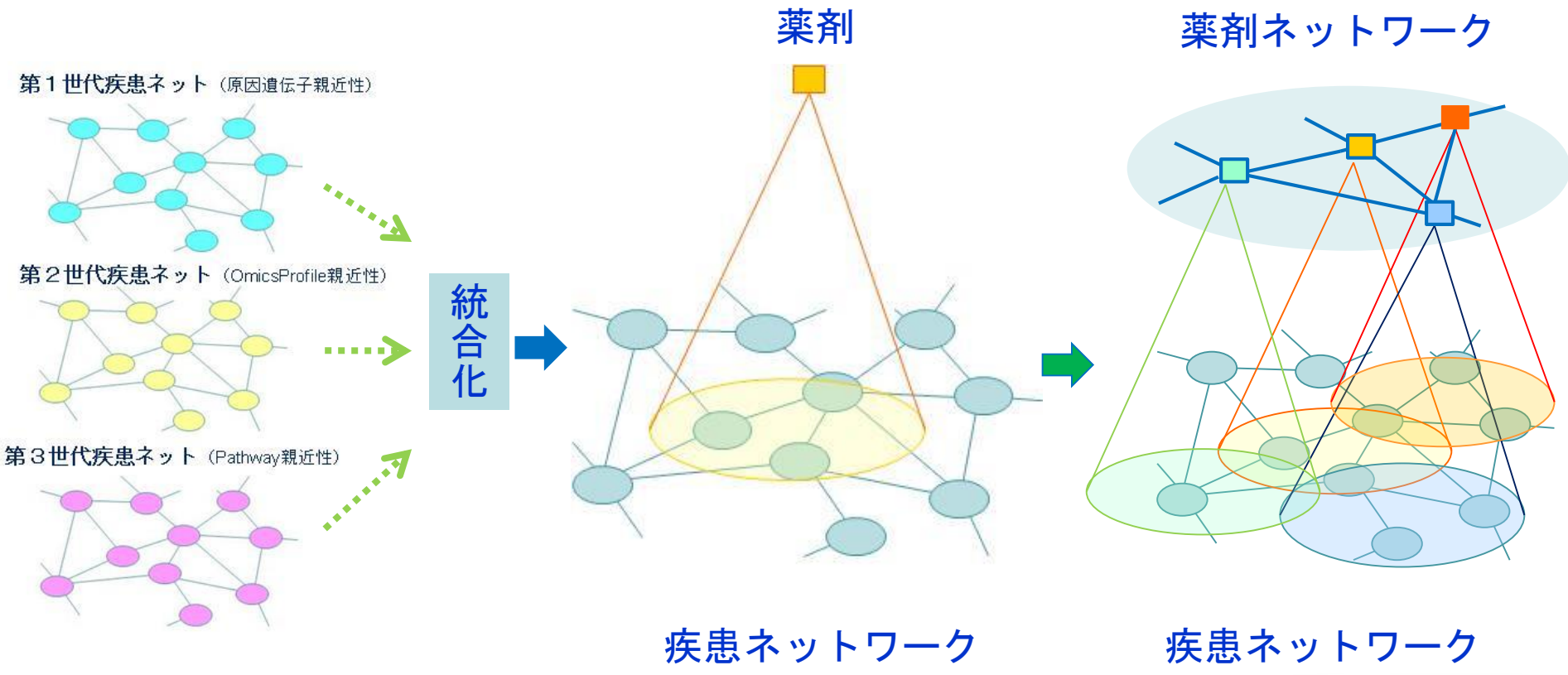



疾患ネットワーク

TMDU

# 疾患ネットワークと薬剤(化合物）投影マップ

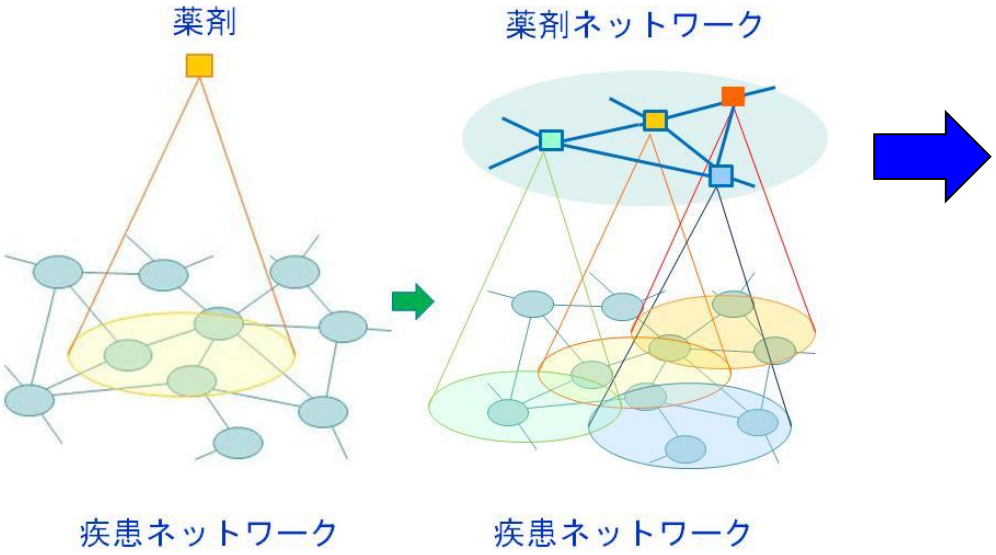
DR informaticsの構築

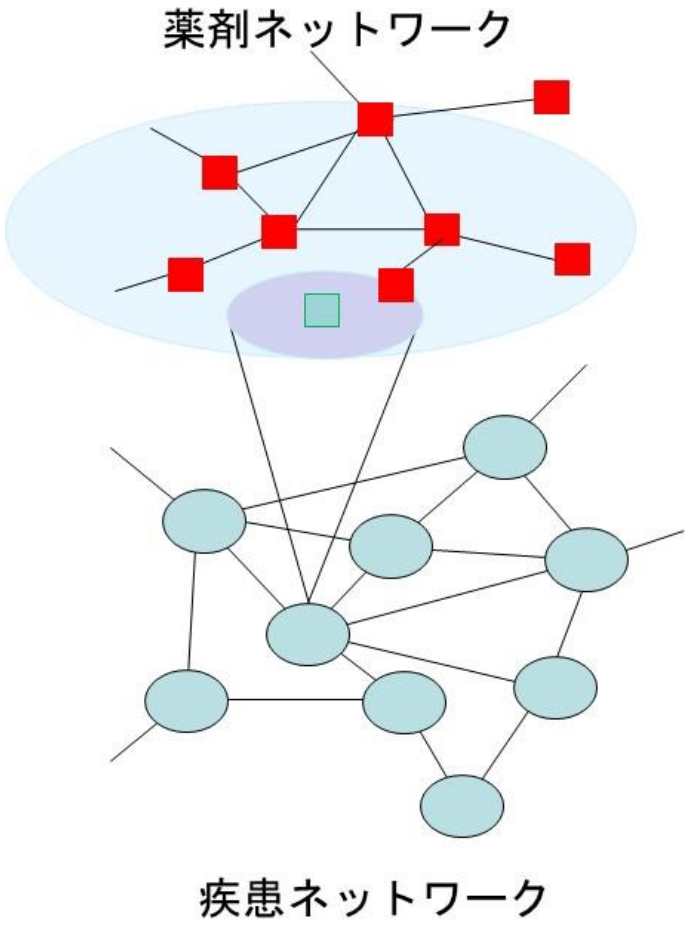


# DRの方法論から創薬方法論へ

- 疾患ネットワークの十全な形成
  - 多層的な疾患ネットワークとその統合
  - 医薬品の有効性・毒性の近傍Projection
    ⇒ DRにおけるfeasibilityは証明
- 創薬への発展
  - 薬剤・化合物空間のネットワークは既に確立
  - cMapでは不十分・LINCS(2014)
  - 疾患ネットワークの確立が重要
  - 疾患から逆投影。創薬の可能性探索
- 疾患トポロジーと薬剤トポロジーの双対性
  - 双対トポロジー準拠計算創薬

疾病から薬剤ネットワークへの逆投影
Dual Topology 双対写像　創薬方法論



# 3. 階層的ネットワークによる創薬/DR

＜疾患-薬剤-標的＞の多層ネットワーク
生体分子ネットワークを基盤とする創薬・DR
ビッグデータ創薬/DRの基本的枠組み

# 疾患ネットワークから
# 生体分子ネットワークを基盤とする創薬/DRへ

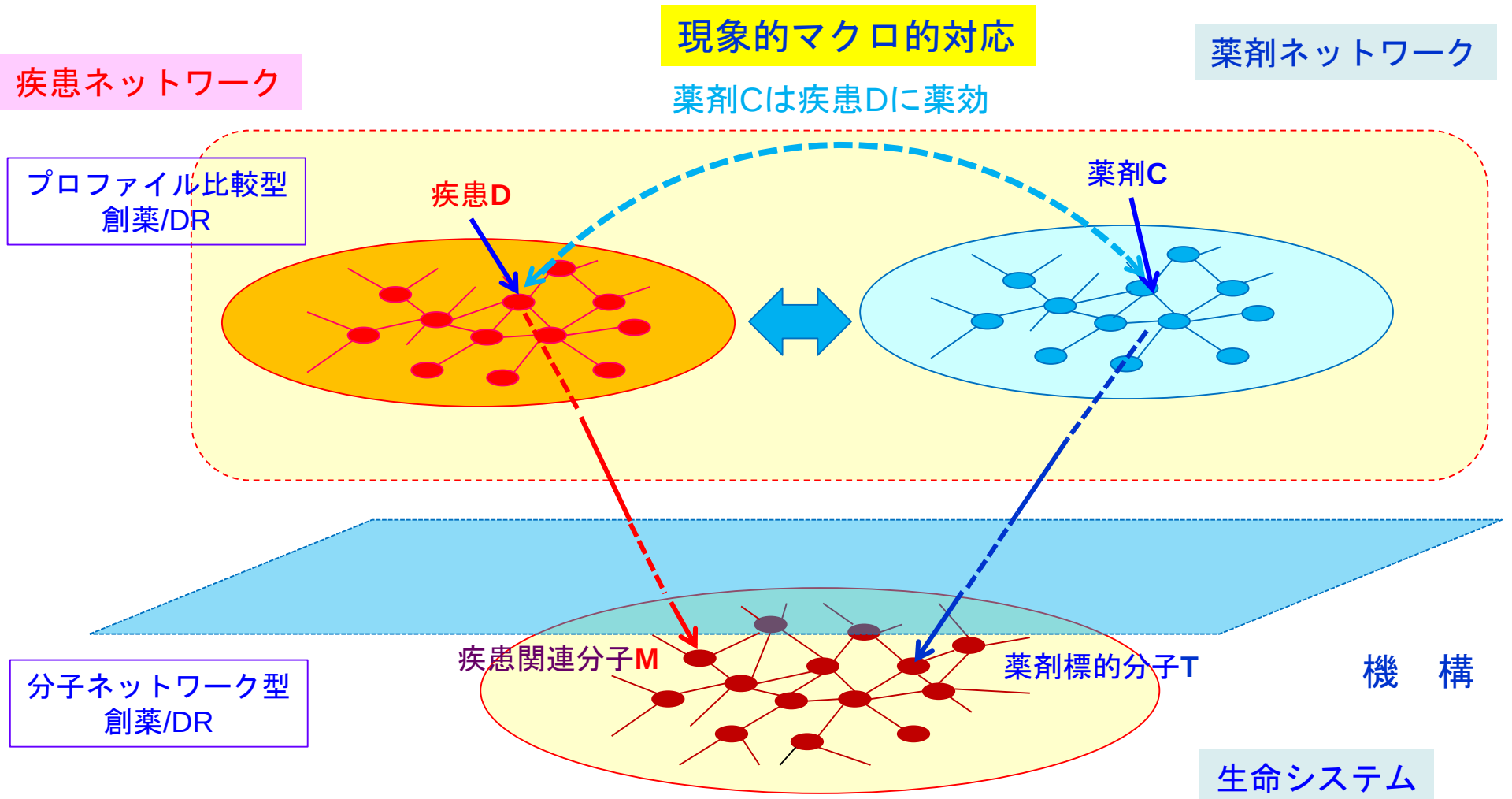
## 疾患ネットワークに準拠した創薬/DR

- 疾患のゲノム・オミックス機序に基づいている→内因的機序を考慮した点で評価
- しかし「疾患関連遺伝子」と「薬剤」の相互作用の関係が明示的ではない

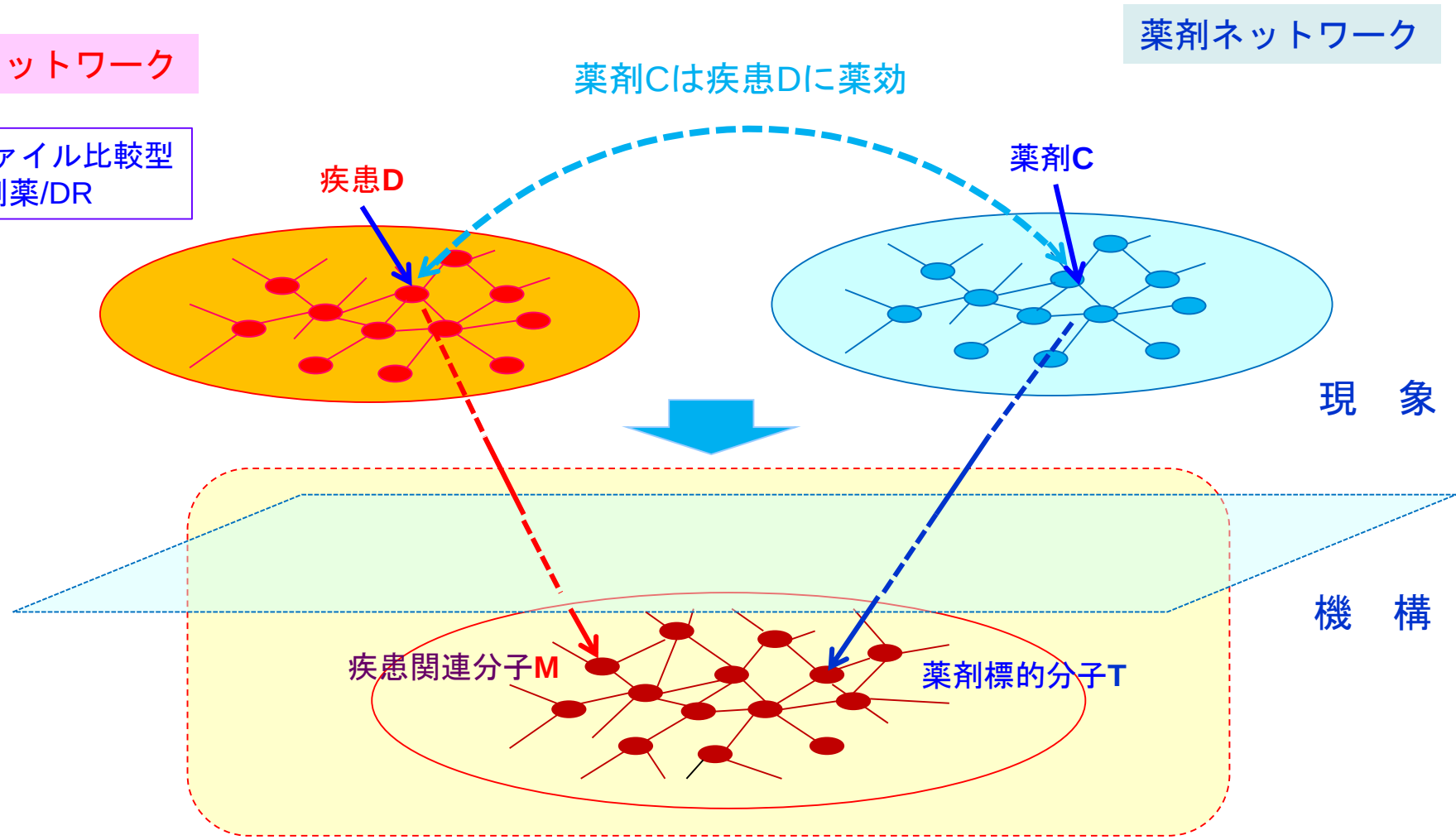
## 集成的な創薬/DRのフレームワーク

- 生命分子ネットワークを作用の＜場＞とする
- 薬剤の足場である＜標的分子＞と疾患の足場である＜疾患関連分子＞との＜相互作用＞を基礎とする枠組み

# 3層の生体・薬剤のネットワーク間の関係図式



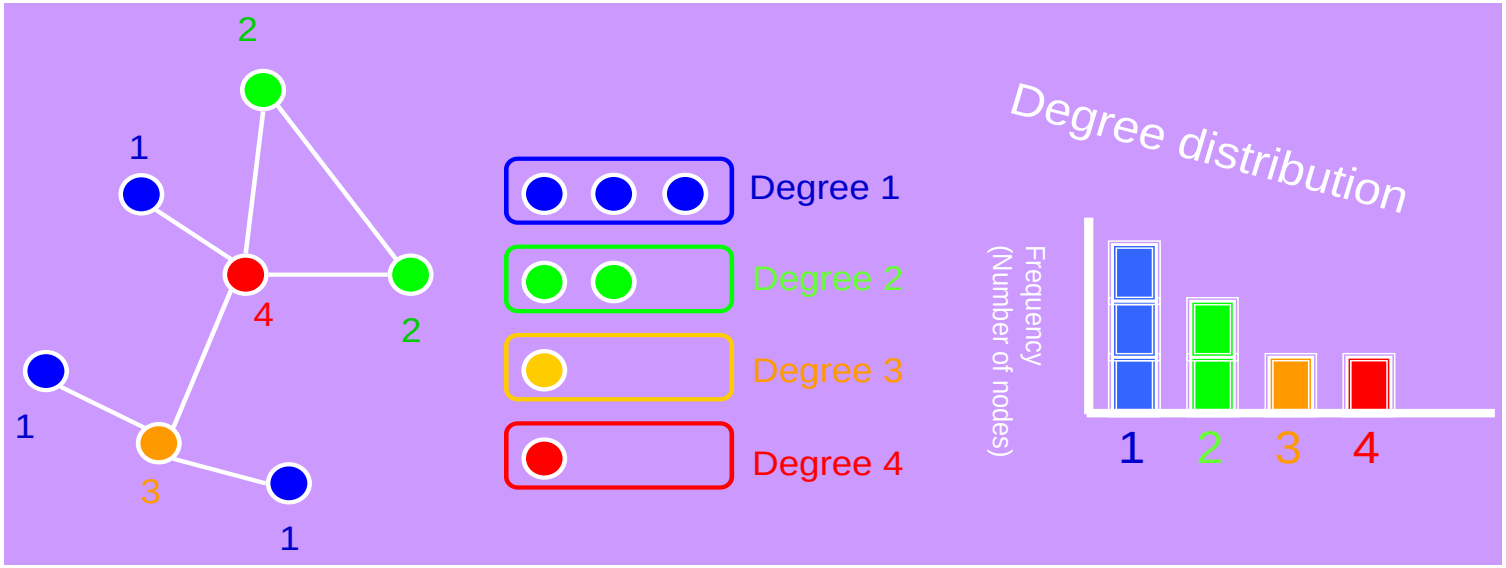
# 3層の生体・薬剤のネットワーク間の関係図式

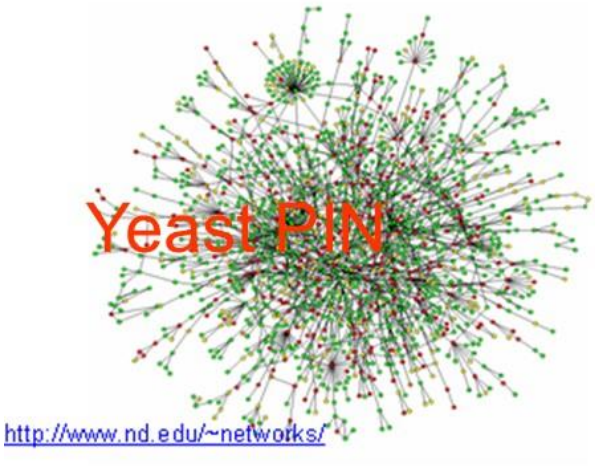


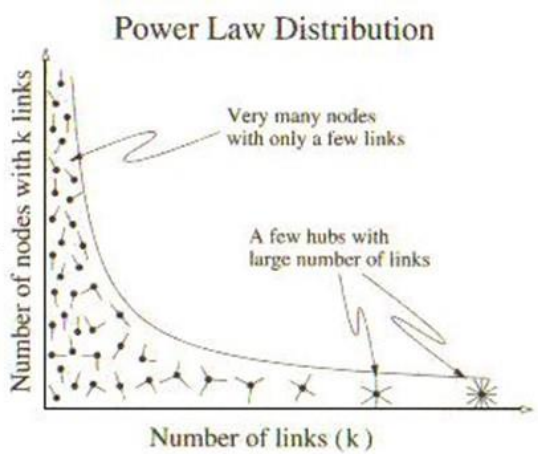
# タンパク質相互作用
# ネットワークの構造と薬剤標的分子

TMDU

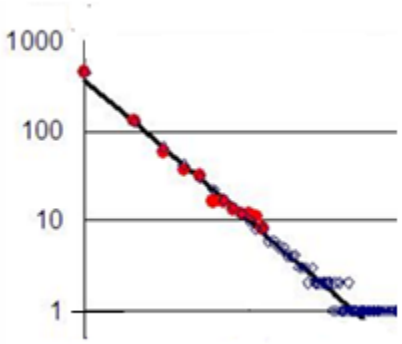
# タンパク質相互作用ネットワーク(PIN)では数少ない相互作用が集中したタンパク質(hub)と相互作用が1や2の多数の末端タンパク質(branch)が存在する



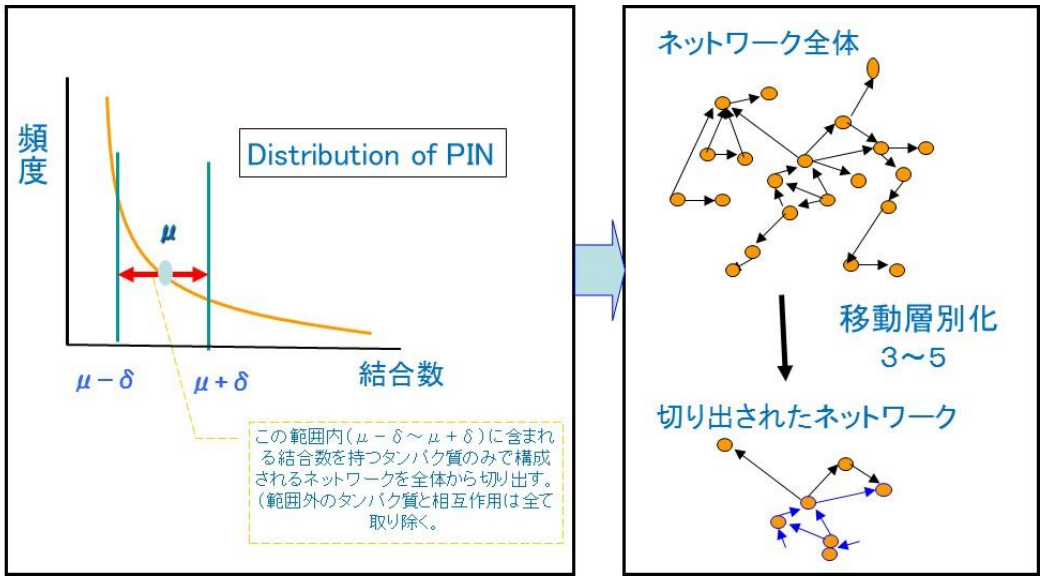




http://www.macs.hw.ac.uk/~pdw/topology/ScaleFree.html

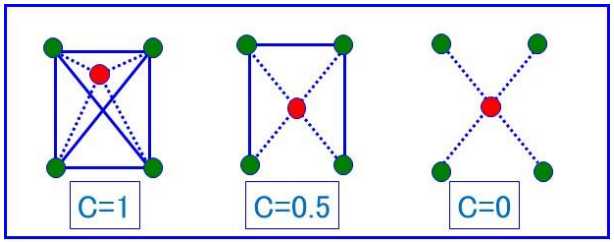



Log-log変換で直線

# 結合次数ごとの部分ネットワーク構造の結合密度の解析

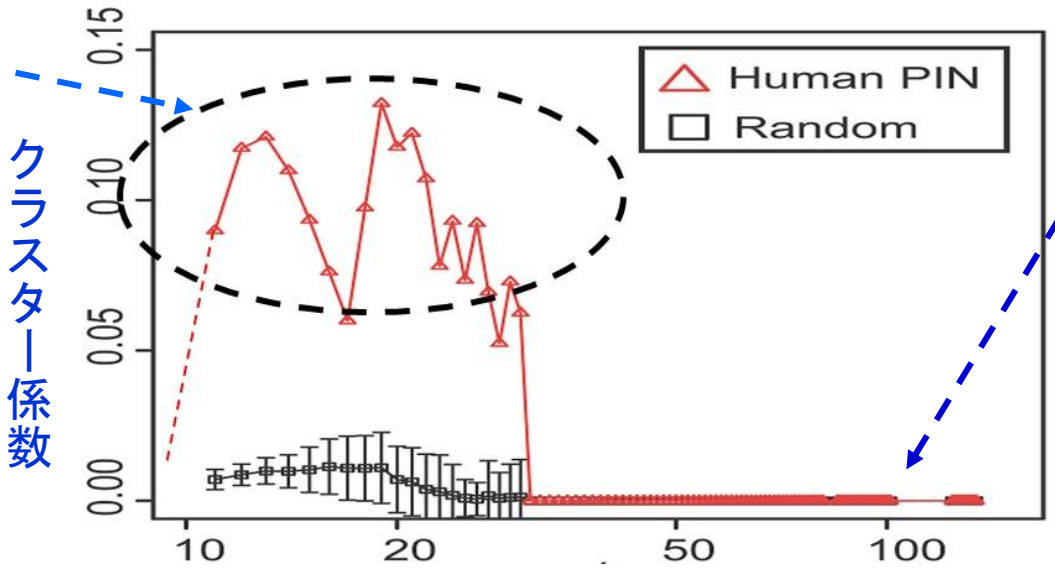



**クラスター係数**




**Hase, T., Tanaka, H et.al (2009)**
Structures of protein protein interaction network and their implications on drug design. ***PLoS Compt Biol***. 5(10):

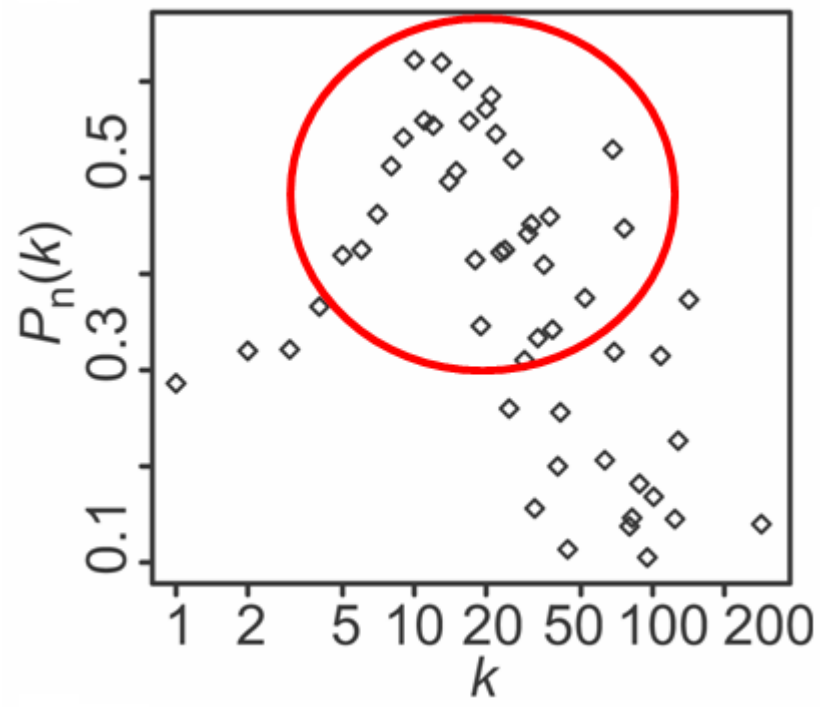
中程度の次数（7～42）を持つタンパク質は多数の密なモジュールを構成

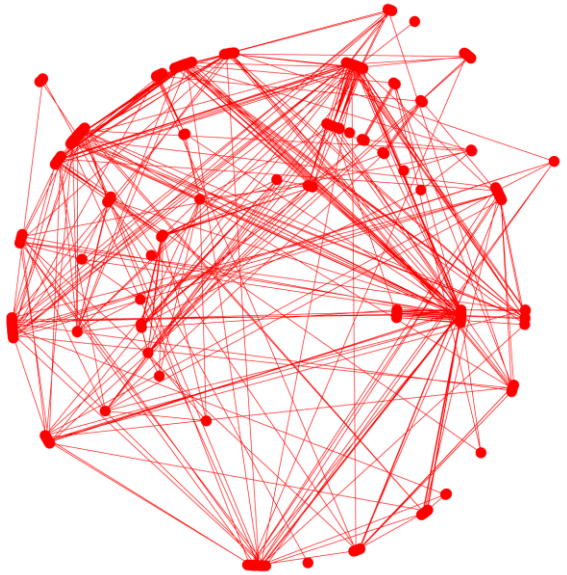



高い次数を持つノード（スーパーハブ）はお互いに密に結合しない

TMDU

# 中層は互いに相互作用する

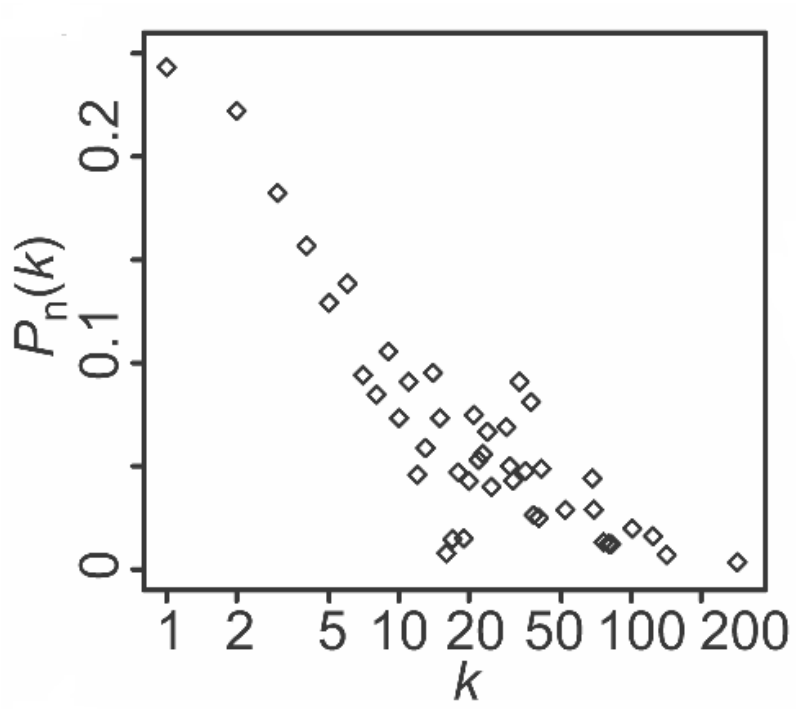


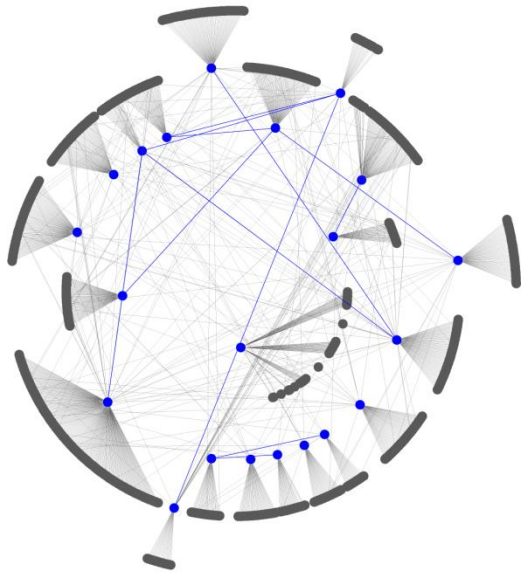



$P_n(k)$　$k$-次数 のノードのリンクが、中間層の次数のノードへ結合している確率

中程度次数のノード間ではPINのバックボーンを構成している.

# ハブ(高層）は低層と相互作用する



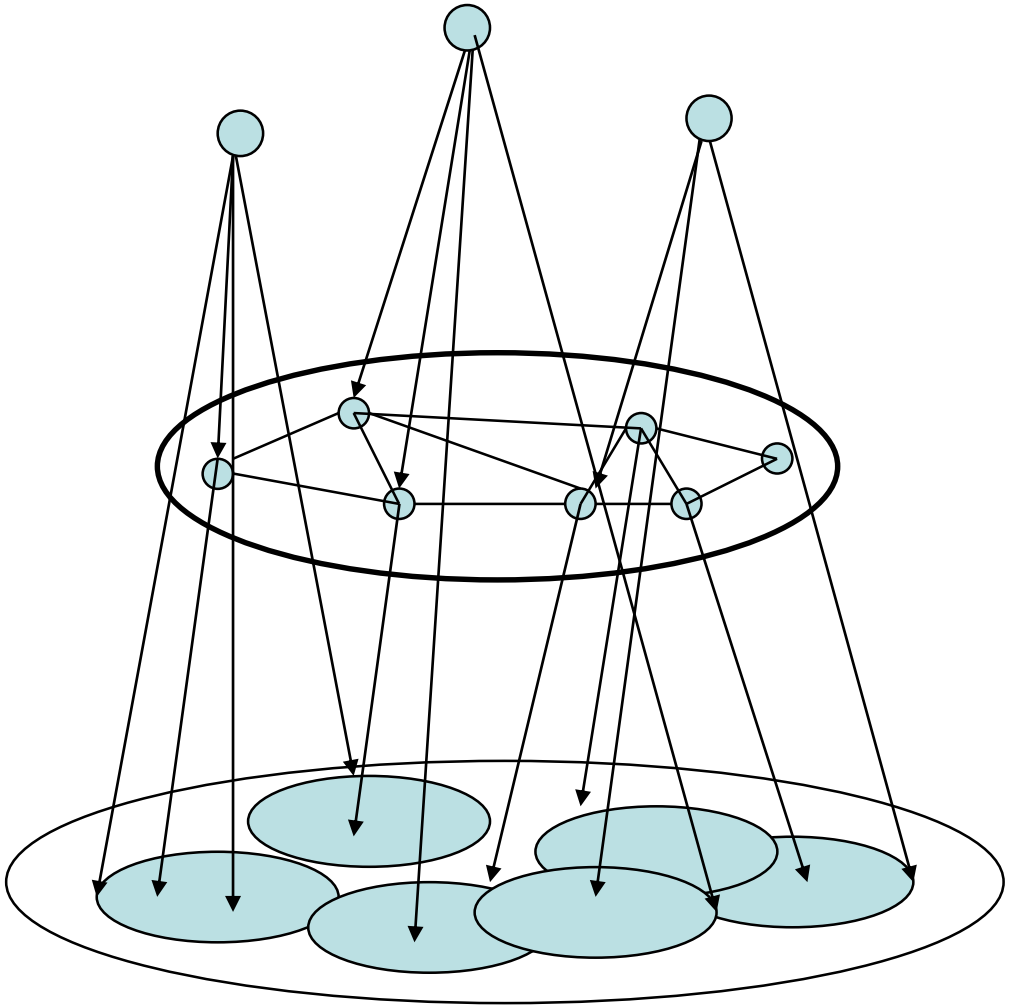
$P_n(k)$　$k$-次数 のノードのリンクが、高層の次数のノードへ結合している確率

高層の次数のハブが、低層の次数のノードに結合し、ハブ間の結合は抑えられている.

TMDU

タンパク質相互作用から見られる

# 生命情報ネットワークの構造



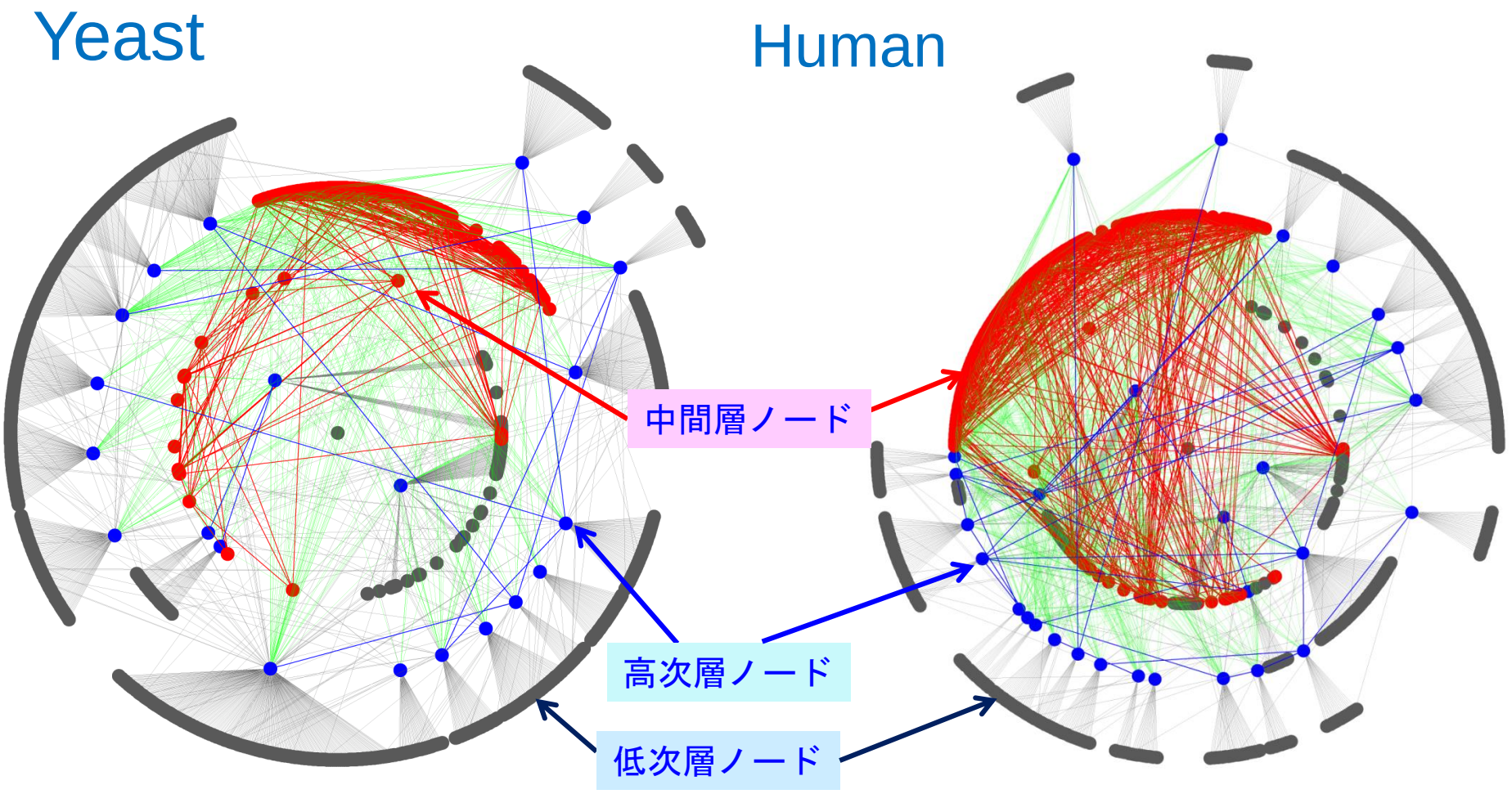
**高層**
**高次数　ハブ**
次数
＞31 ヒト
＞39 酵母

**中間層**
**中程度次数**
次数
6 ～30　ヒト
6 ～38　酵母

**低層**
**低次数　ブランチ**
次数＜6

TMDU

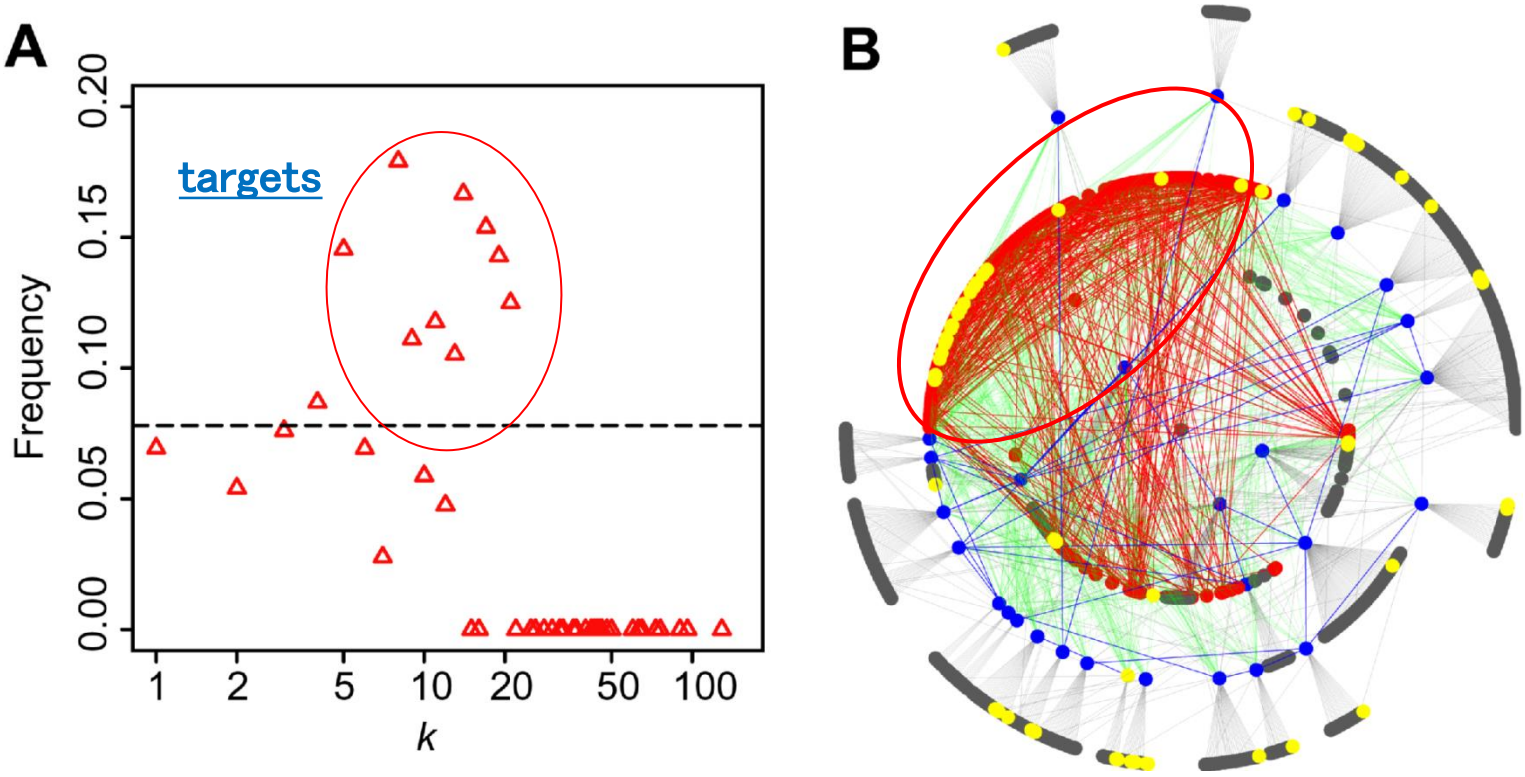
# タンパク質相互作用ネットワークの Cloud Topology（3環トポロジー）




**中間層の次数 ノードは PPI バックボーン を形成する**

灰色, 赤, 青は、それぞれ低層、中層、高層の次数のノードをそれぞれ表す。

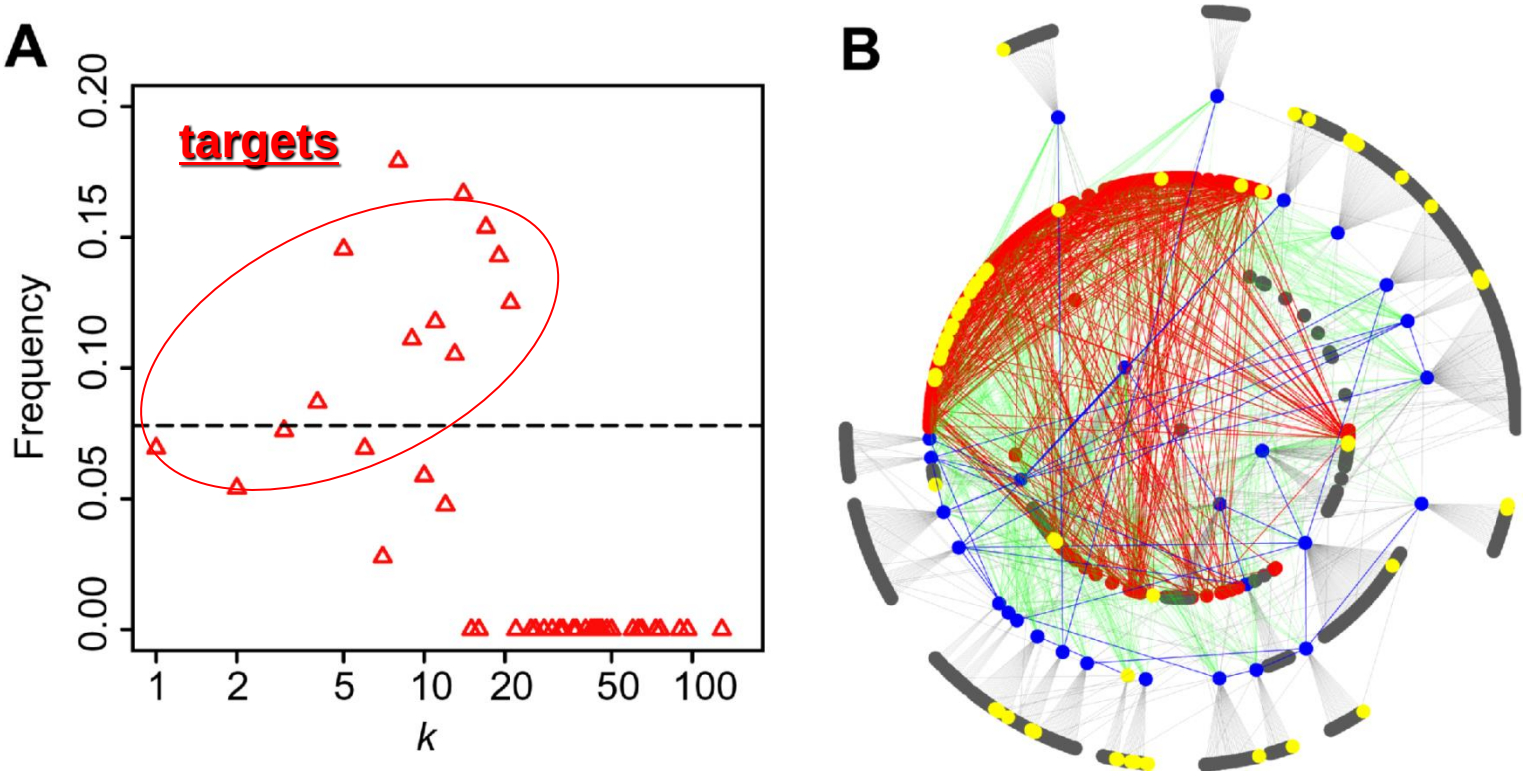
# 薬剤標的分子と結合度数




中層レベルのノードは治療薬として最適な標的である。それゆえ、多くの市場にある薬剤標的は、ヒトのバックボーンタンパク質に集中している

TMDU

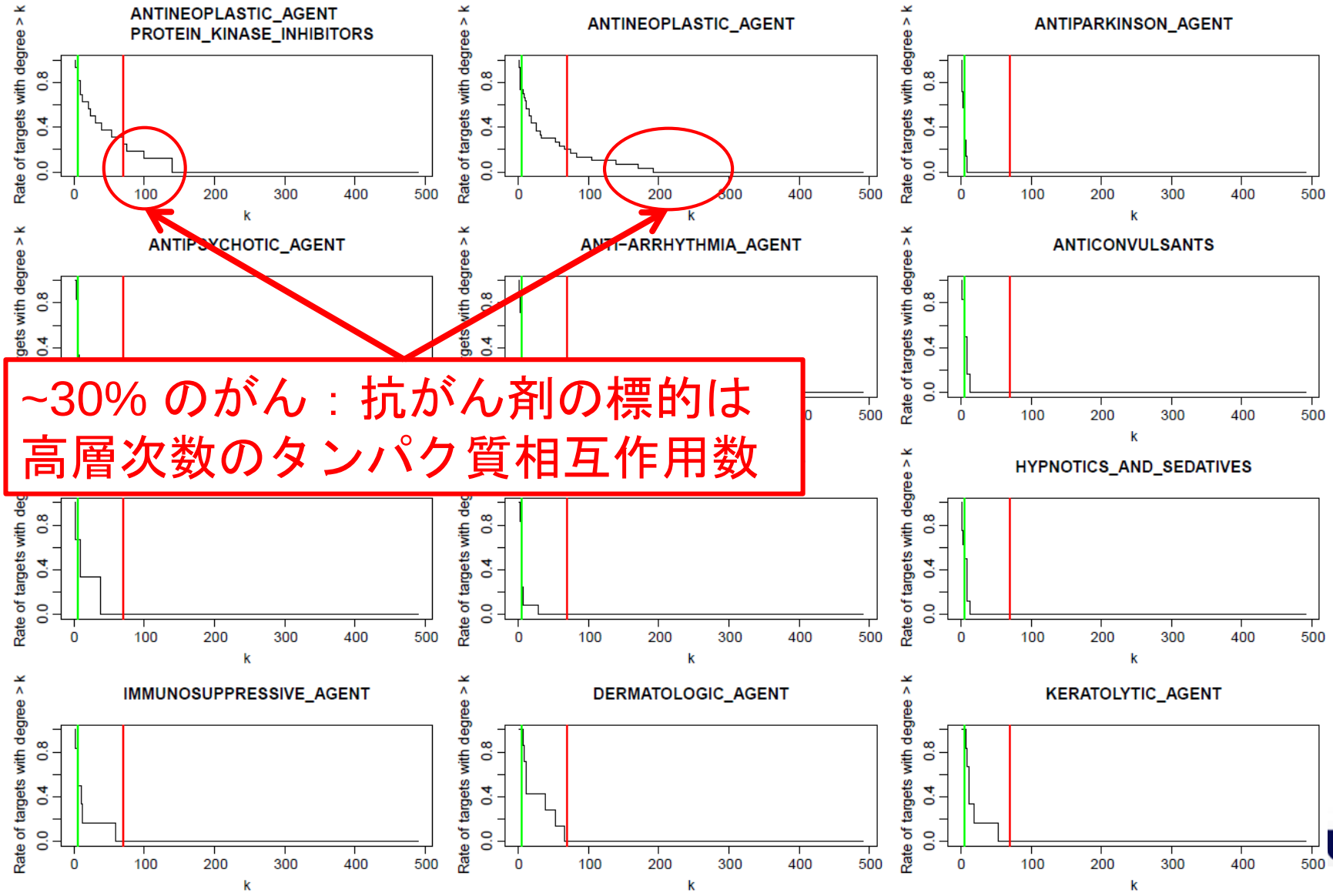
# 薬剤標的分子と結合度数




中間層の次数のノードは治療薬の標的として最適であるが
低層の次数のノードにも薬剤標的分子が存在する

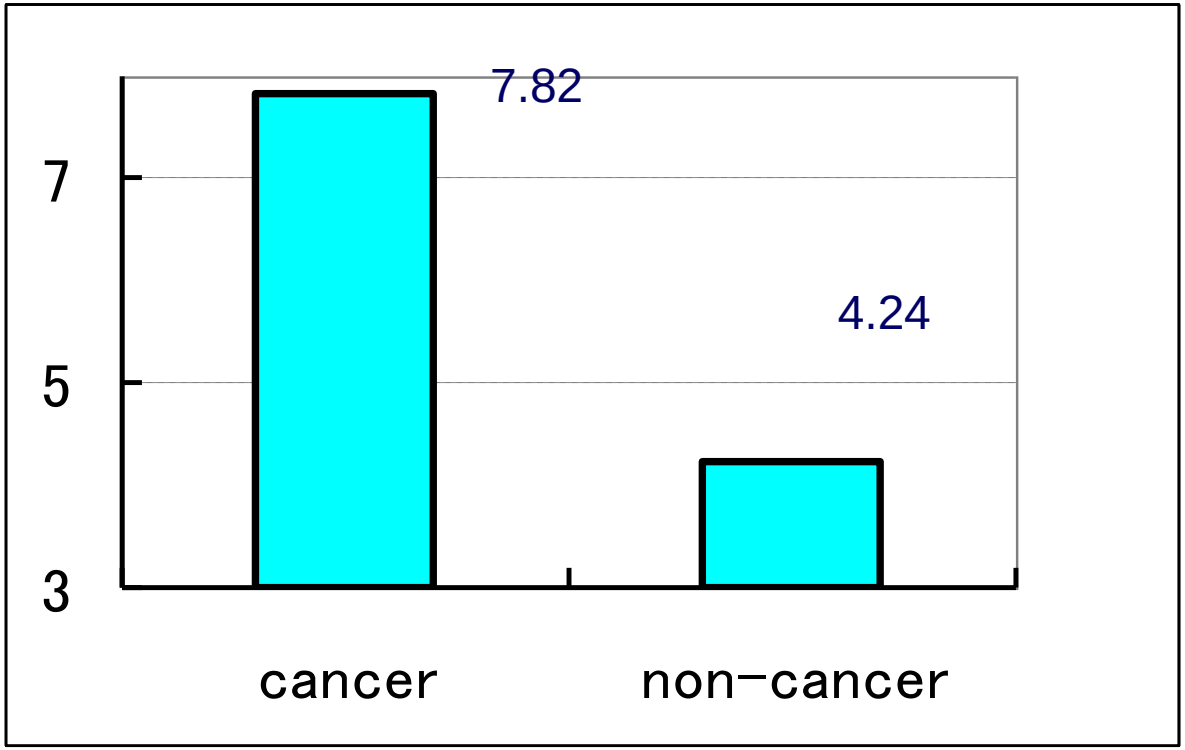
TMDU

# がん疾患遺伝子は高層次数ハブのタンパク質が多い




~30% のがん：抗がん剤の標的は
高層次数のタンパク質相互作用数

# 薬剤標的分子と結合次数

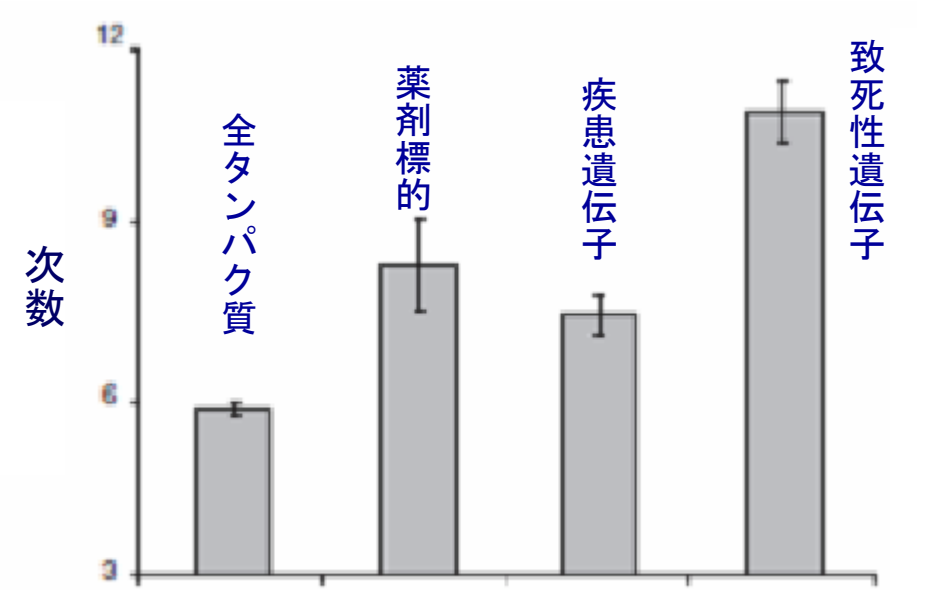



薬剤標的
全平均
4.74

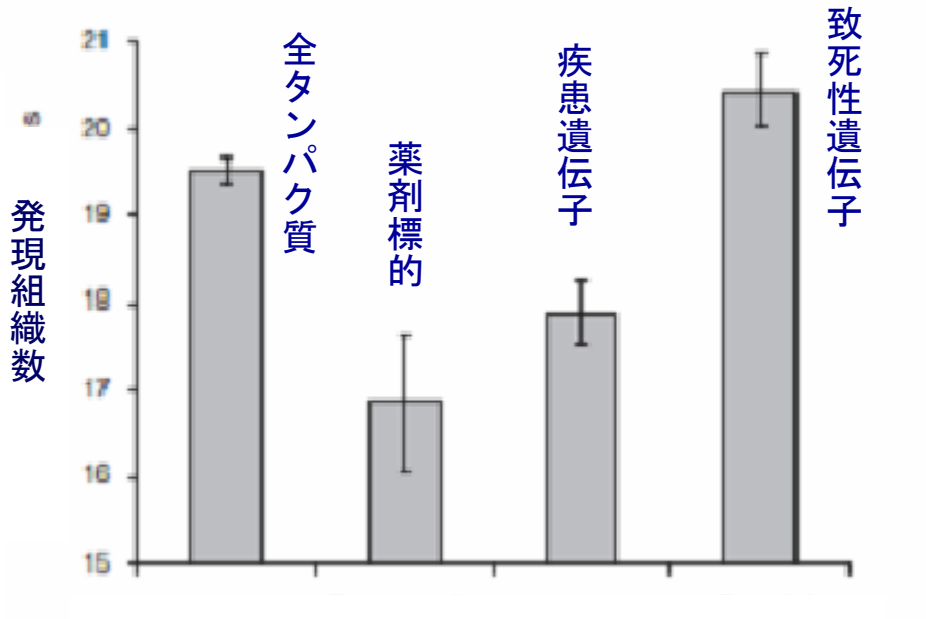
タンパク質
全平均
4.06

抗がん剤 (P=0.01)の標的分子は平均的に次数が高い。
抗がん剤がより厳しい副作用を起こす理由である.

# 薬剤標的分子と結合次数




**結合次数**
薬剤標的タンパク質
致死的タンパク質
疾患関連遺伝子タンパク質



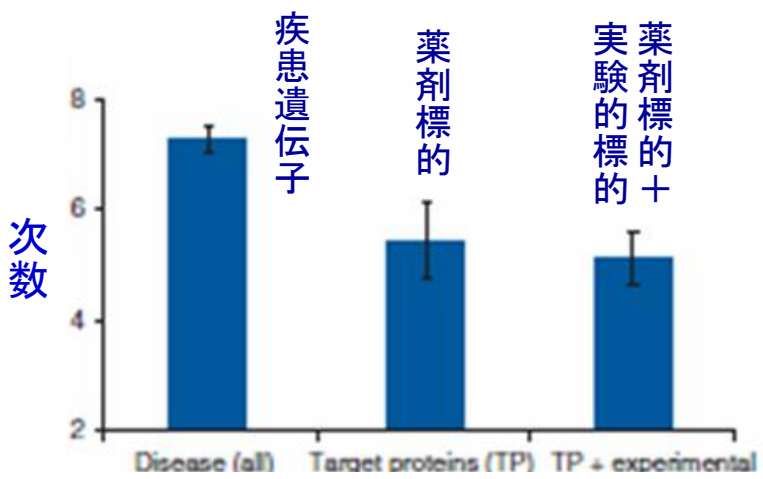

**発現組織数**
薬剤標的タンパク質
致死的タンパク質
疾患関連遺伝子タンパク質

( YıldırımM A, et al, NATURE Biotechnology 2007)

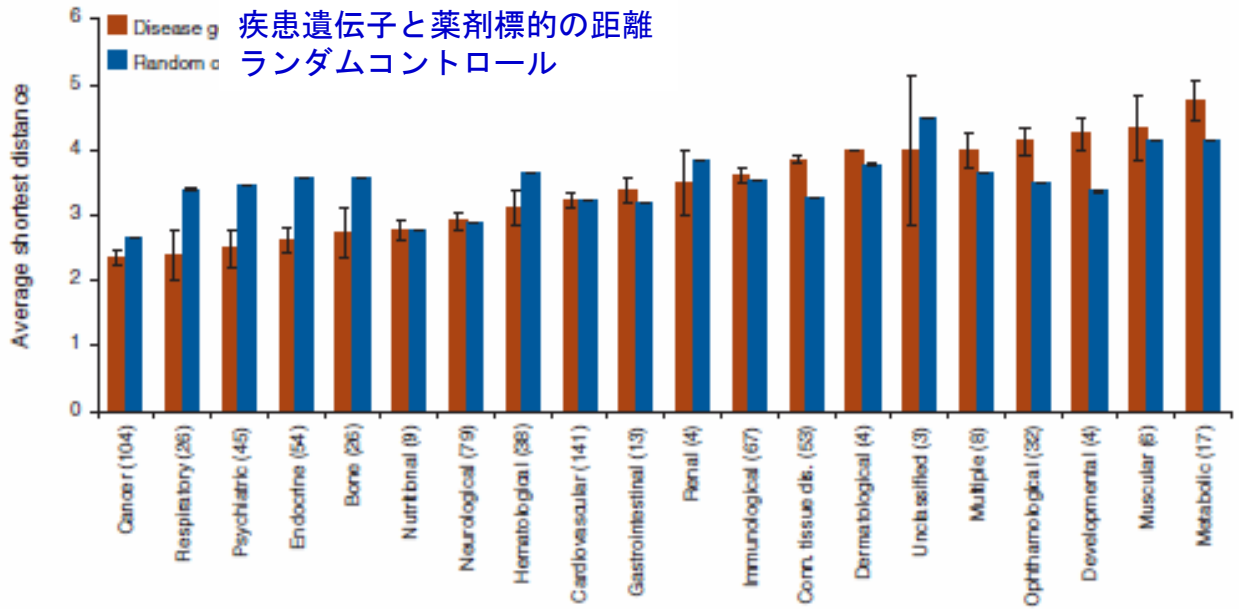
TMDU

# 標的タンパク質と疾患遺伝子の距離

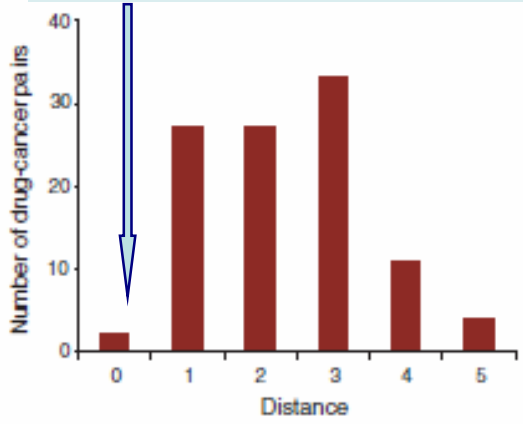
薬剤標的タンパク質と
疾患関連タンパク質の
間の距離：２～４リンク




Yıldırım　M A, et al,
NATURE Biotechnology
2009




**抗がん剤の場合**
**疾患遺伝子と距離０の標的**

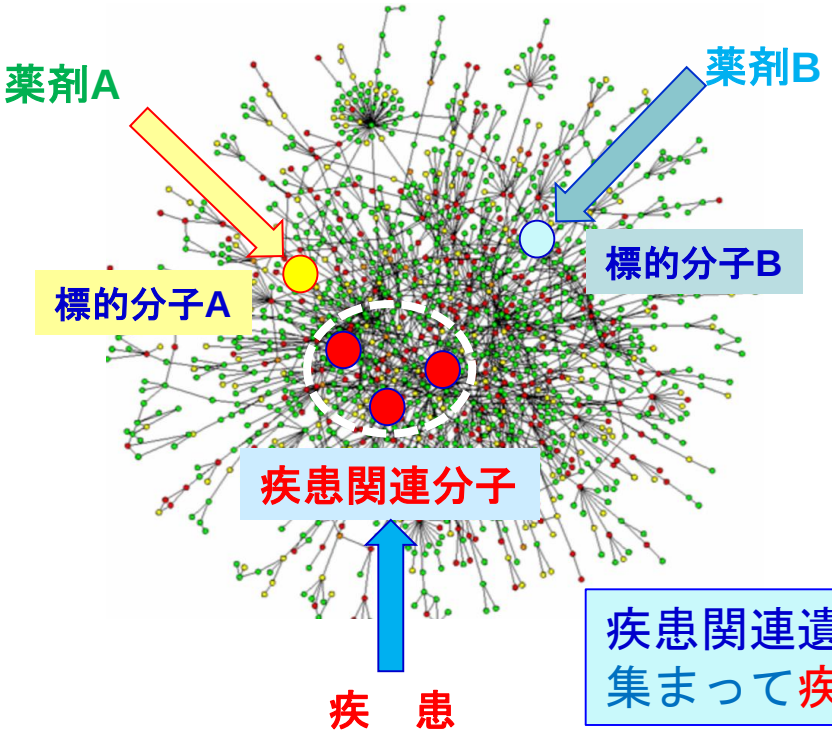



抗がん剤の標的分子と
疾患遺伝子の間に距離

# タンパク質相互作用ネットワークを基盤にした計算創薬/DR

# 標的分子や疾患関連分子の
# タンパク質相互作用ネットワーク（PPIN）

- 薬剤ネットワークと疾患ネットワークの基盤：生体分子ネットワーク
- タンパク質相互作用ネットワーク（PPIN）での創薬/DR戦略
- PPIネットワーク場を基礎にして距離（近接性）を検討
- **薬　剤**：**薬剤の標的分子**（タンパク質）によって　PPI場と繋がる
- **疾　患**：**疾患特異的発現遺伝子**を疾患関連分子（**タンパク質**）へ翻訳、
- PPIN場内での**薬剤（標的分子）と疾患（疾患関連遺伝子）の「代理人」**の**距離・近接性**を基準に、薬理作用のインパクト力を評価

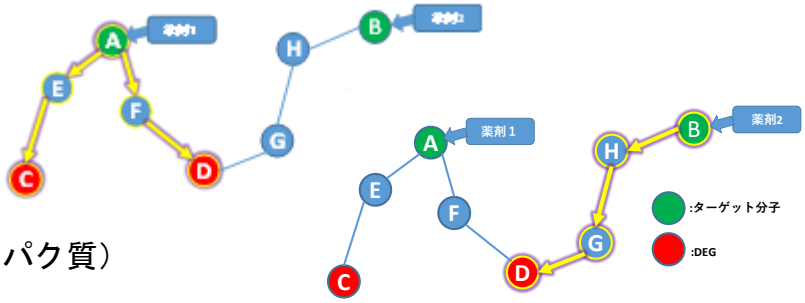



タンパク質相互作用
ネットワーク（PPIN）

疾患関連遺伝子はネットワーク上の近傍に集まって疾患モジュールを形成する

# PPIの基づくDR（肺腺癌の例）

- **Interactome**(タンパク質相互作用)ネットワーク（Sun, 2016）
  - HPRD
    - 37,070 PPI, 9465 タンパク質
  - STRINGS
- **薬剤⇒標的分子**：DrugBank
  - 7,759 薬剤、4300タンパク質
  - 12,604 薬剤ー標的分子 （4,452薬剤, 1,617タンパク質）
- **疾患遺伝子（肺腺癌）**
  - TCGA (The Cancer Genome Atlas)より差異的発現遺伝子を同定
    - 445 肺腺癌例, 19 正常例, 疾患遺伝子 FC >2.0 or <0.5, 927 差異的発現遺伝子
- **薬剤の疾患遺伝子への影響力 評価IPS** (Impact power score)
  - **薬剤の標的分子**と**疾患遺伝子**の間のネットワーク距離の総合評価
    **「再出発ありランダム歩行」**RWRでネットワーク距離を評価
  - **標的分子からランダム歩行を繰り返す**（出発点から再出発あり）
    **s時点後, 疾患遺伝子のノードにどれだけの確率で滞在**しているかをIPSとする
  - 一定の時間が過ぎると、**定常状態**になり、歩行で滞在確率分布は変化しない。
    定常状態での**疾患遺伝子ノードに滞在している確率**の総和が薬剤の評価になる




$$\mathbf{P}^{s+1} = (1-\gamma)\mathbf{M}\mathbf{P}^{s} + \gamma\mathbf{P}^{0}$$

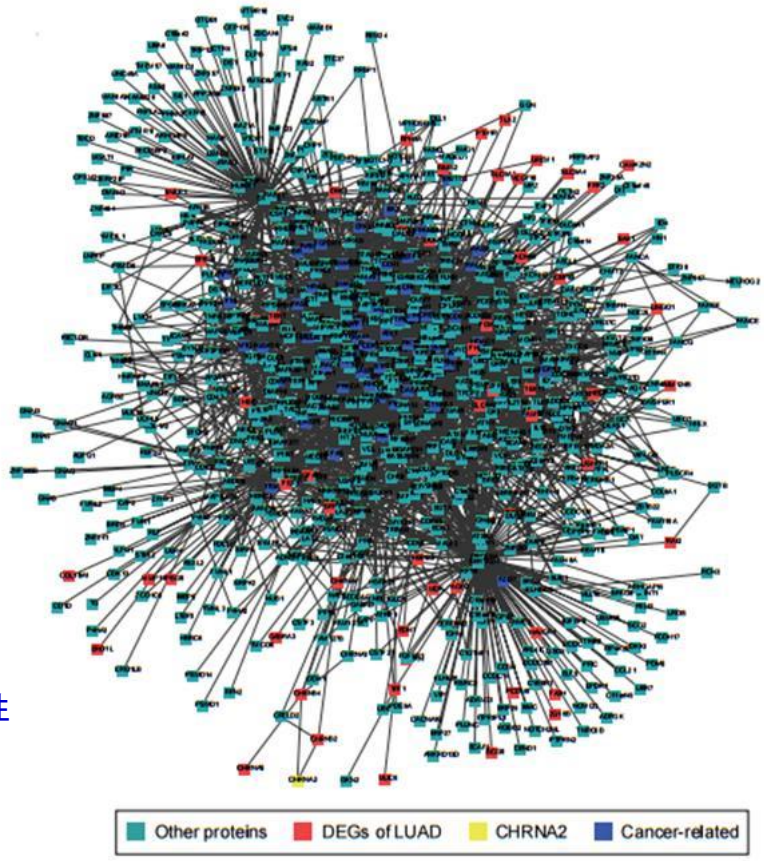
$\mathbf{P}^{s}$：時点sでの各ノードでの滞在確率　$\mathbf{M}$：各ノードへの遷移確率　$\gamma$：再出発確率

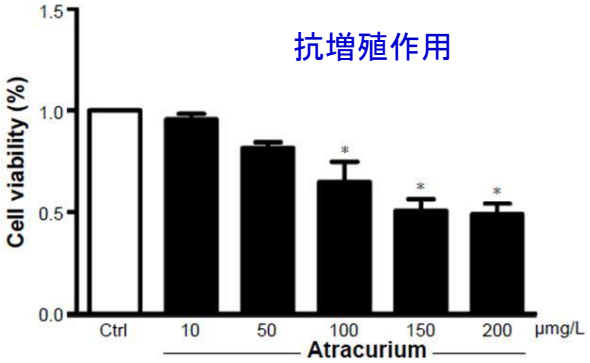
TMDU

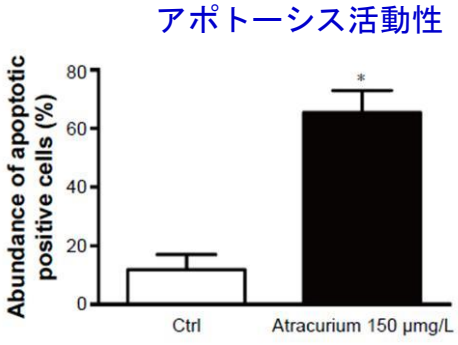
# タンパク質相互作用ネットワーク DR 結果の検証

| Drug ID | Drug name | Target | Score | Rank |
|---|---|---|---|---|
| DB00416 | Metocurine Iodide | CHRNA2 | 0.966581 | 1 |
| DB00565 | Cisatracurium besylate | CHRNA2 | 0.966581 | 1 |
| DB00732 | Atracurium | CHRNA2 | 0.966581 | 1 |
| DB00657 | Mecamylamine | CHRNA2 | 0.966581 | 1 |
| DB02457 | Undecyl-phosphinic acid butyl ester | LIPF | 0.953846 | 5 |

- **HPRDとSTRINGS**の両方のランダム歩行で**145薬剤・化合物**が共通
- 最高スコアを挙げた**Atractium**を選択
- 標的は**CHRNA2**(Cholinergic Receptor Nicotinic Alpha 2）**でアポトーシス経路である**
- **培養細胞A549**（ヒト肺胞基底上皮腺癌細胞）の抗増殖作用を確認



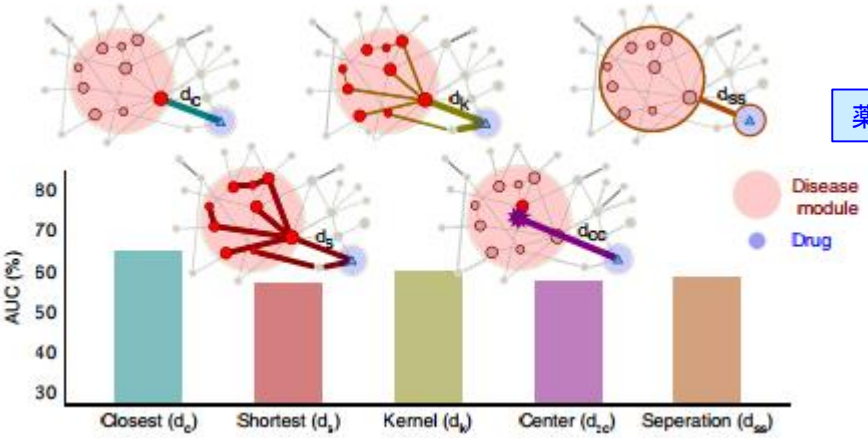




TMDU

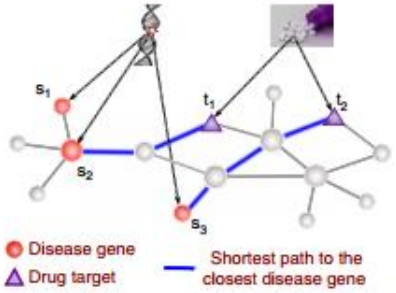
# タンパク質相互作用ネットワークでの近接性によるDR

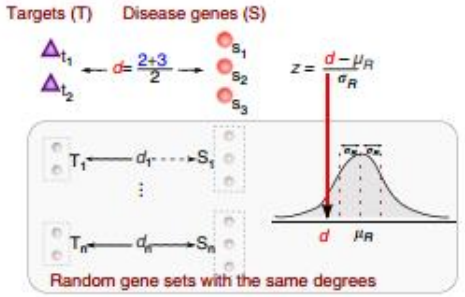
**相対近接指標$d_c$**：
①最近接の疾病関連分子との最短経路長の平均
②同じサイズで度数の分布より近接指標を計算して規格化⇒zスコア
（z < - 0.15⇒近接）
②様々な近接指標の中ではclosest measure $d_c$ が一番薬効を予測する

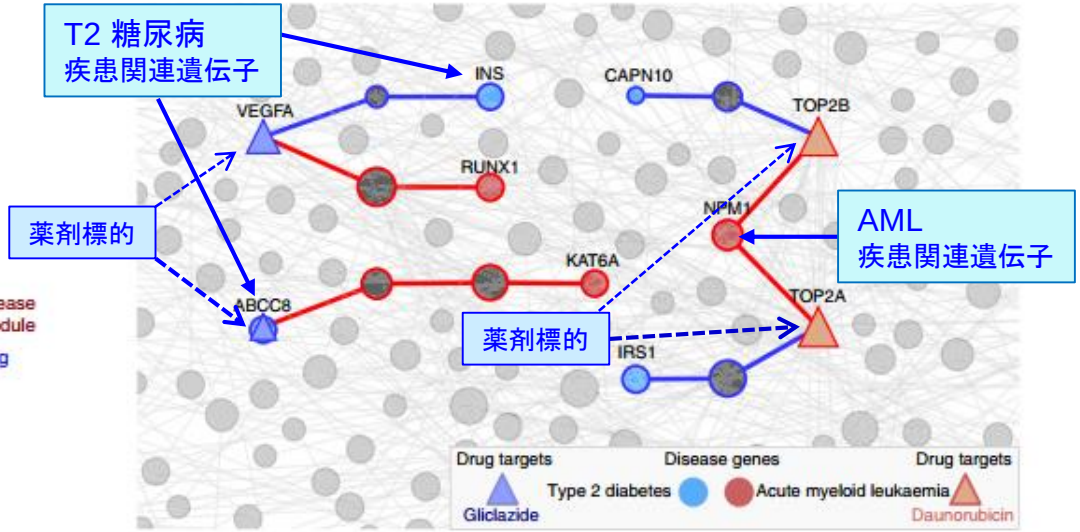


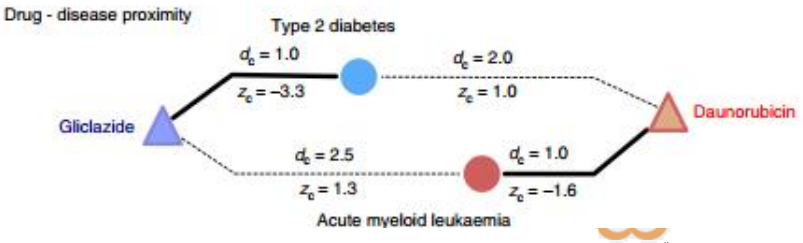

大半の薬剤は標的と疾患関連分子
２リンク離れている

(Guney, Barabasi,2016, Nat. Com)



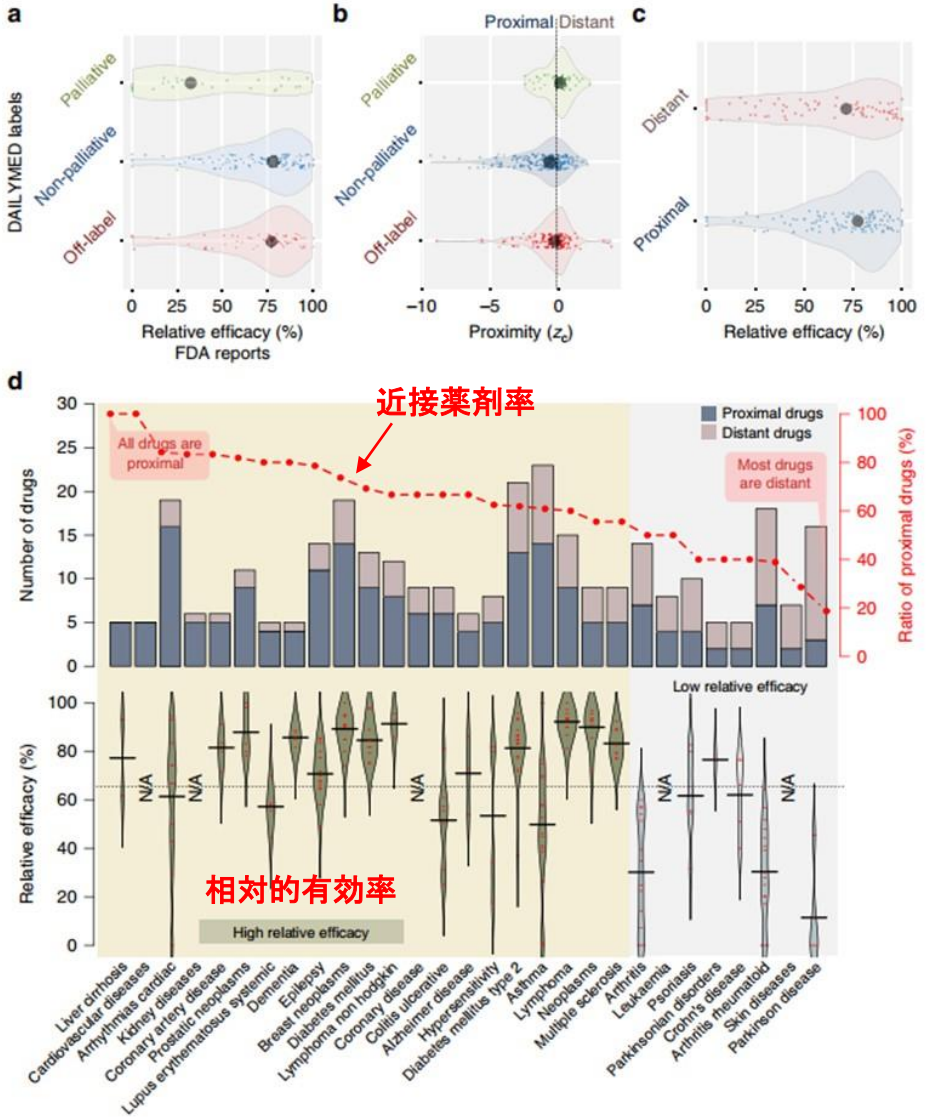



# 相対近接性による薬効予測

- 疾患モジュールの内部/近接に標的分子を持つ必要がある
- これまでの研究では疾患関連分子と標的分子の距離が大
  - 対症療法・緩和療法：疾患原因ではなく症状を標的としている
  - 標的分子が疾患関連分子の数は少ない（402対のうち62）
- 既成の薬は疾患と近接的である
- 緩和療法は遠隔的である
- Off-labelは緩和より近接的である
- 近接薬剤の治験の頻度は高い
- 薬剤は選択的であるが排他的ではない
- 相対的有効性と近接指標は相関する
- 平均の標的分子の数は3.5個である



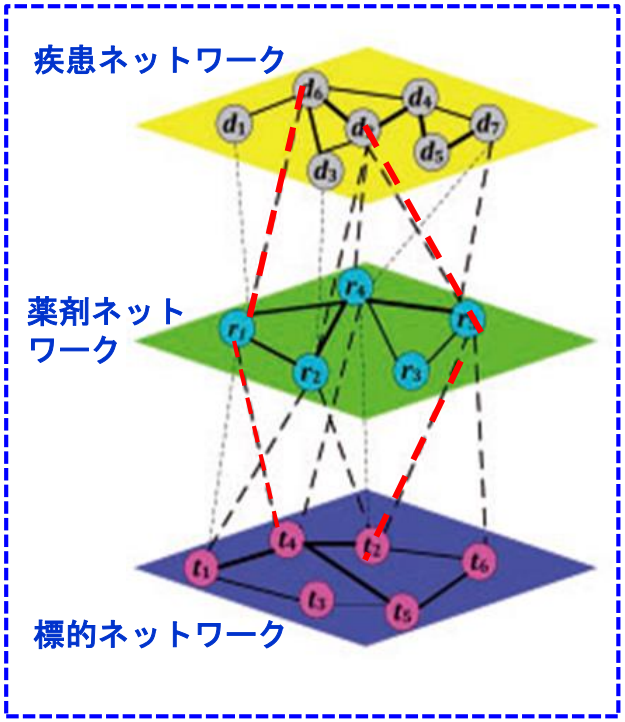
# 3階層生命ネットワークでの創薬/DR

- 3階層の生体ネットワーク
  - 疾患ネットワーク：網羅的分子による内在的機序
  - 薬剤ネットワーク：化学構造によってネットワーク
  - 標的ネットワーク：薬剤と標的（DrugBank参照）
- 各層のネットワーク内結合
  - 稠密に自己完結的に構築可能
- 各層ネットワーク間のリンク
  - 成功した＜疾患－薬剤＞の事実の根拠のみ
  - 階層間はスパースな結合である

**創薬/DRとは**
未発見の階層間リンクを
既存の階層間リンクの事実と
各層のネットワークから推測

Wang et al. 2014は
- 階層間リンク（事実）と各階層内のリンクより階層間のリンクの強さを計算する方法を提案している

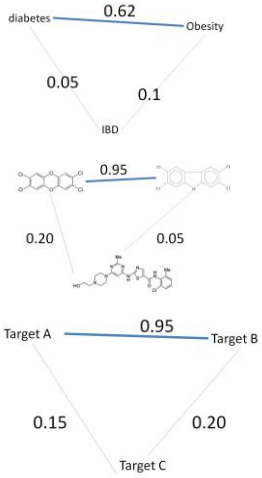
(Wang et al. 2014)

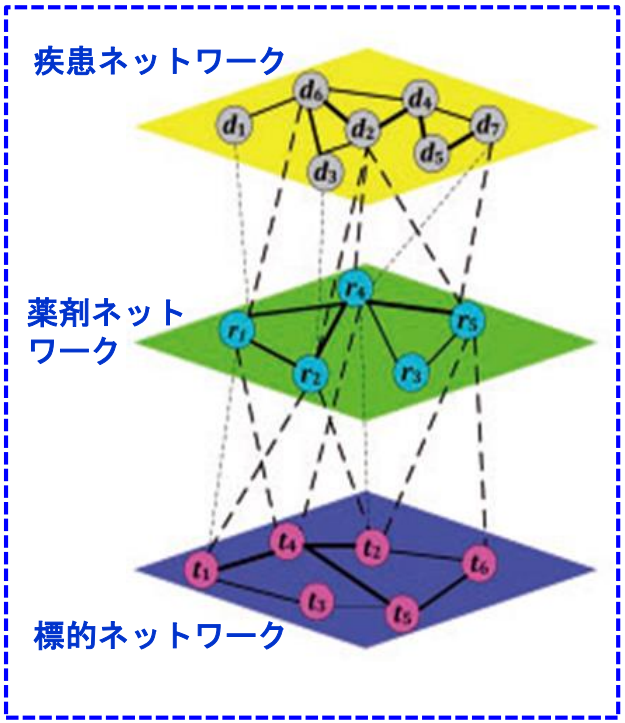


# 異質 ネットワーク 創薬/DR

(Wang et al. 2014)

- **３層ネットワーク構成**
  - 疾患ネットワーク($d_i$)
  - 薬剤ネットワーク($r_i$)
  - 標的ネットワーク($t_i$)
- **各ネットワークで距離定義**
  - 疾患ネット：MeSHの共通項数
  - 薬剤ネット：Tanimotoスコア
  - 標的ネット：Smith-Waterman法
- **結合係数w(*i*,*j*)更新法**




$$w(d,r)=\sum_{d_i\in D}\sum_{r_j\in R} w(d,d_i)\times w(d_i,r_j)\times w(r,r_j)$$

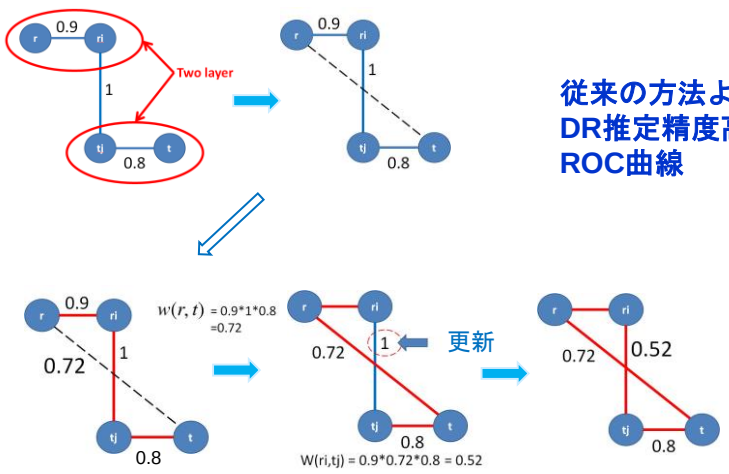
$$w(d,t)=\sum_{r_i\in R}\sum_{r_j\in R} w(d,r_i)\times w(r_i,r_j)\times w(r_j,t)$$

$$w(d,r)=\sum_{t_i\in T}\sum_{t_j\in T} w(d,t_i)\times w(t_i,t_j)\times w(t_j,r)$$

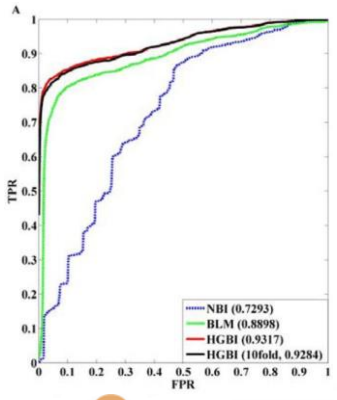
結合係数更新のマトリックス表示

$$W_{dr}^{k+1}=\alpha W_{dr}^{k}\times(W_{rr}\times W_{rt}^{k}\times W_{tt}\times W_{rt}^{kT})+(1-\alpha)W_{dr}^{0}$$

$$W_{rt}^{k+1}=\alpha(W_{dr}^{kT}\times W_{dd}\times W_{dr}^{k}\times W_{rr})\times W_{rt}^{k}+(1-\alpha)W_{rt}^{0}$$



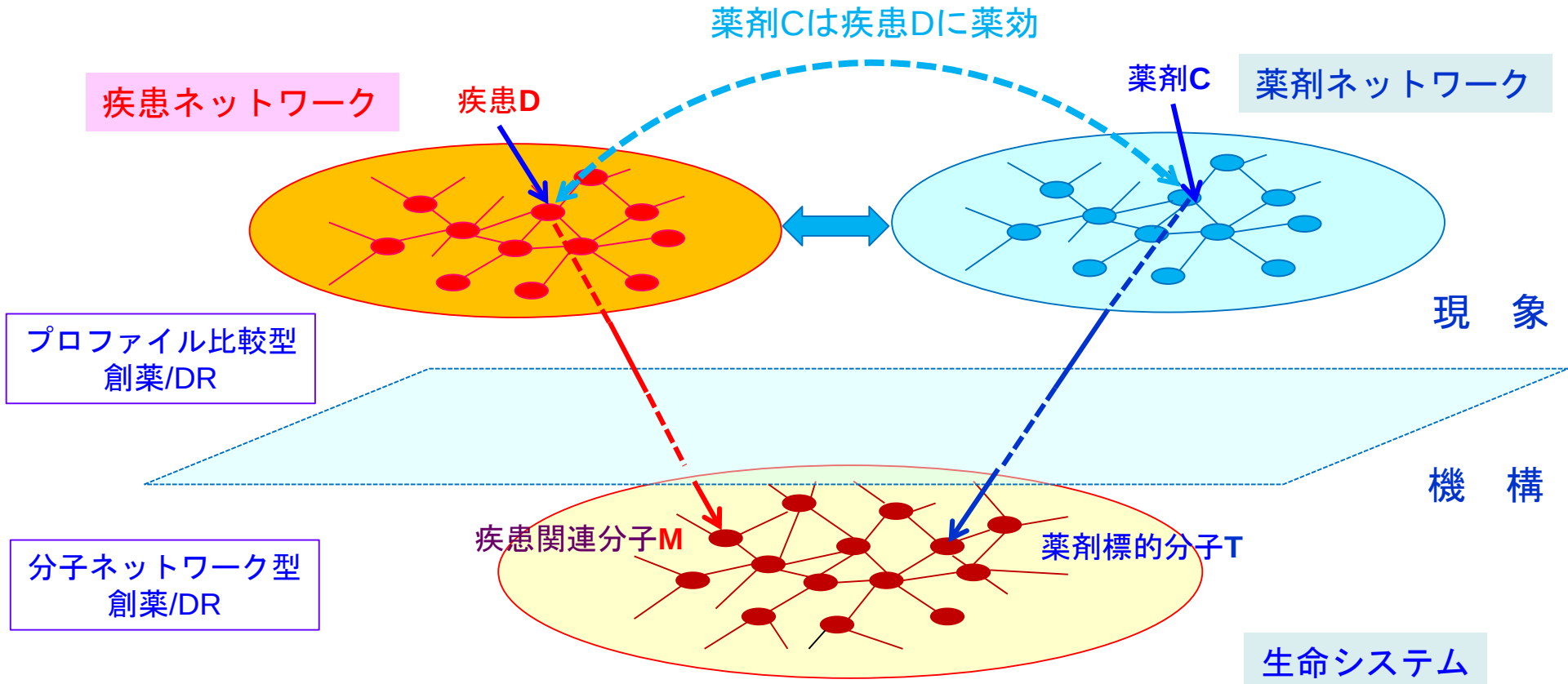

従来の方法より
DR推定精度高い
ROC曲線




TMDU

# プロファイル型計算創薬の原理

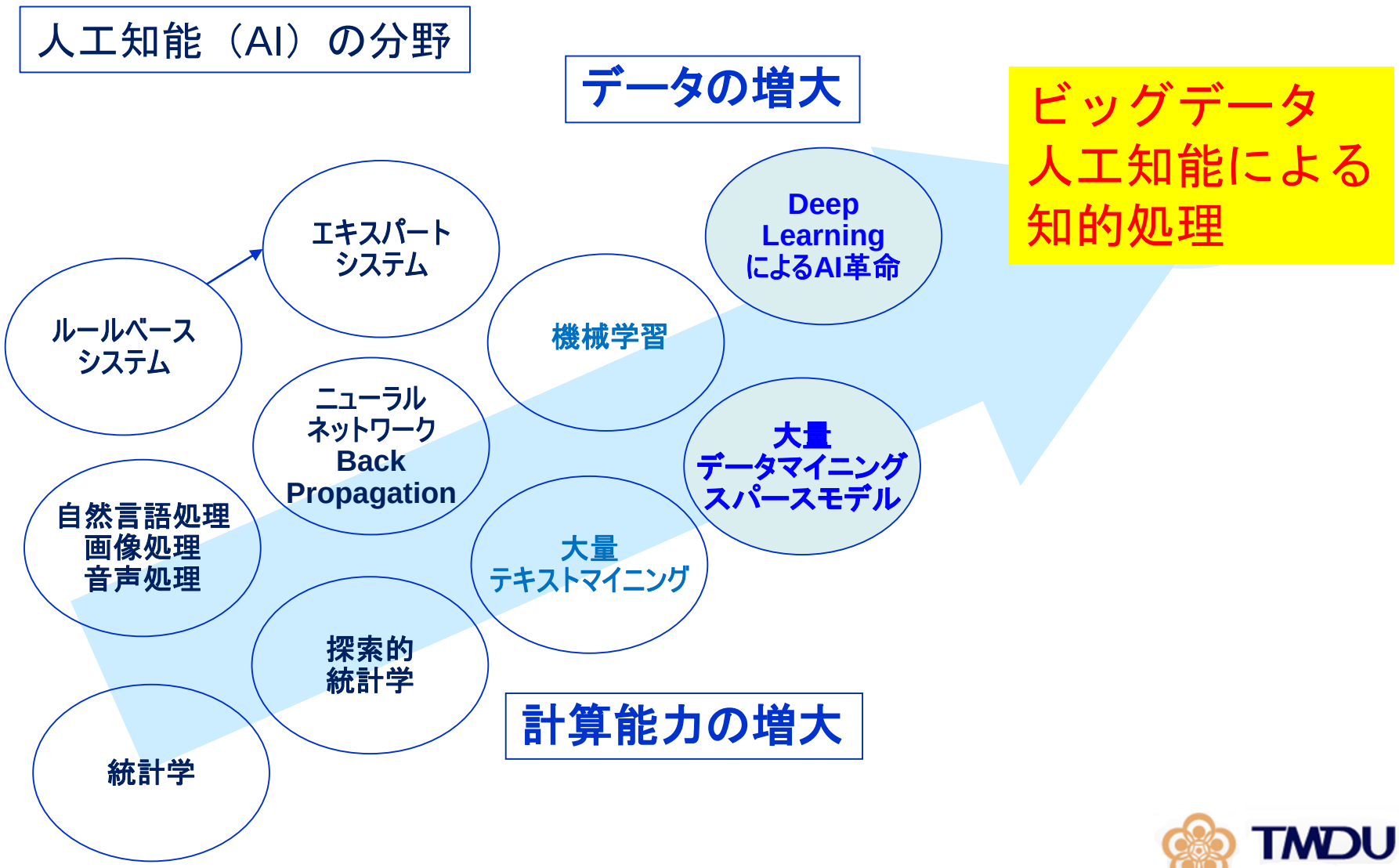
## 3層生体・薬剤ネットワークのFramework



# 第 III 部

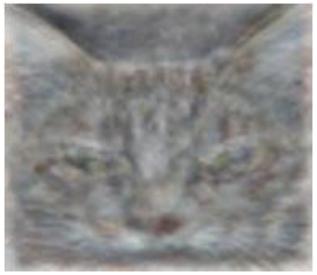
# AI 創薬

# 人工知能への期待


人工知能（AI）の分野
データの増大
ビッグデータ
人工知能による
知的処理
ルールベース
システム
エキスパート
システム
ニューラル
ネットワーク
Back
Propagation
自然言語処理
画像処理
音声処理
探索的
統計学
統計学
機械学習
大量
テキストマイニング
Deep
Learning
によるAI革命
大量
データマイニング
スパースモデル
計算能力の増大

# 人工知能の最近の話題

- **「アルファ碁」**（Google DeepMindによるコンピュータ囲碁プログラム）が2016年3月に数多くの世界戦優勝経験のあるプロ棋士李世乭（Lee Sedol：九段）に挑戦し、**4勝1敗と勝ち越した**
  - チェス：IBM「Deep Blue」が1997年に当時の世界champion, カスパロフ氏（ロシア）に勝利
  - 将棋：ボンクラーズ, 2012年米長永世棋聖に勝利
- 人工知能が1000万枚の画像を与えて**「猫」を認識するニューロン**をできたと2012年に発表



# アルファ碁

- 「アルファ碁」にはニューラルネットワーク（Deep Learning)が使われた。評価経験則が人間によってコードされていない
- 最初、棋譜に記録された熟練した棋士の手と合致する手をさすように訓練され、次に、ある程度の能力に達すると、強化学習を用いて自分自身と多数の対戦（3000万回）を行うことで上達した。
- 2017年初頭、その改良版が日中韓のトップ棋士を相手に60戦無敗というな驚異的戦績を挙げた。

# 人工知能とは

定義：人工的にコンピュータ上など知的な振舞いを行うシステムを実現すること、あるいはそのための技術

## ２つのアプローチ

記号的AI：人間の知識を明示的に表現し、それを用いた推論などの知的課題を解決する。「人間を真似る」
「古いAI」。知識準拠型（エキスパート）システムなど

非記号的AI：人間の用いる思考・知識内容とは関係なく最良の計算方式により知的課題を解決する。「人間を超える？」
「計算AI」ニューロネットワークなど

TMDU

# 人工知能研究の推移


（出典）総務省「ICTの進化が雇用と働き方に及ぼす影響に関する調査研究」（平成28年）

# 医療分野の人工知能の歴史

記号（シンボル）的知識処理

ニューロネットワーク処理

1970

問題解決の一般探索手法 GPS
解決木の高速探索（ゲーム）

ニューロネットワーク
3層の学習機械 Perceptron
入力層、隠れ層、出力層

1980

推論システム（if-thenルールシステム）
知識の表現と利用（専門家システム）
医療診断システム（Mycin, Internist-I)
大ブーム　医療から産業応用の期待波及

多層型ニューロネット
後方伝播 Back Propagation
結合係数修正アルゴリズム

期待消滅！

1990

しばらく停滞！

知識発見 機械学習
Machine Learning, KDD
診断知識のDBからの学習

ニューロネットワーク型
多層型ニューロネット
**深層学習 Deep Learning**
結合係数修正アルゴリズム
画像処理から創薬まで

2000

知識準拠診療支援（DSS)
医療ターミノロジー
医療オントロジー

# 自己紹介と医療人工知能の歴史

人工知能(AI)を医療・創薬へ応用

田中　博
東京医科歯科大学
生命医療情報学
東北大学
東北メディカル・
メガバンク機構

1980から1995
第１期の
AIブームの時
医療AI研究に従事

私の経歴

東大医学部

医療AI
研究

東京医科歯科大

ゲノム
医療

東北大学

BigData医療
AI創薬

記号知識処理

問題解決の
探索法（GPS）

医学「知識」を
計算機に格納
医療診断システム（MYCIN）
**知識工学：大ブーム**
政府：第５世代コンピュータ
**知識の移植問題**

ブーム消滅！

医療機械学習
診断知識のDBからの学習

診療支援
医学の用語や
概念体系の基礎理論

ニューロネット(NN)

単純NN
パーセプトロン
**判別能力の限界**

1970
以前

1980

多層NN
バックプロ
パゲーション
**重み修正の限界**

1990

ブーム消滅！

2000

Deep Learning
多層NN
「教師なし」特徴学習

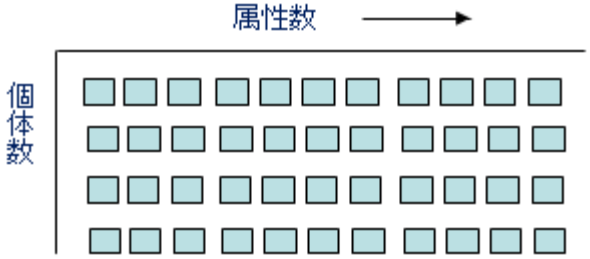


当時の講演者執筆の医療AI成書

# 「ビッグデータ」のData縮約 原理

問題点　　属性項目数(p)≫サンプル数(n)

p: 数億になる場合あり　　n: 多くても数万、通常数千




これら膨大な属性変数がすべて独立ならばビッグデータの構造解析は不可能。単変量解析の羅列（GWASのManhattan Plot)しか可能でない

ビッグデータ・スパース仮説

ビッグデータは、多数であるが属性項目数より少ない独立成分が**基底**となって、相互にModificationして構成されている。

データ次元縮約の原理（**principle of compositionality**)

# ビッグデータ解析に向けた2つの人工知能（AI）方法の適用

- 統計的学習：データマイニング、探索統計学の数理的枠内で次元縮約

  ⇒　スパース推定による従来手法の次元落ちの正則化

- ニューロネットワーク：Deep Learningによる特徴量抽出を用いた次元縮約

  ⇒Deep LearningのAutoEncoder機能を用いた実質的な独立次元抽出に基いた解析・予測

TMDU

# 統計的学習
# スパース推定による次元落ちの正則化

## 従来の重回帰分析

$\mathbf{x} = (x_1, \ldots, x_p)$ と目的変数 $y$ に関して $n$ 組のデータ $\{(y_i, \mathbf{x}_i); i = 1, \ldots, n\}$

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

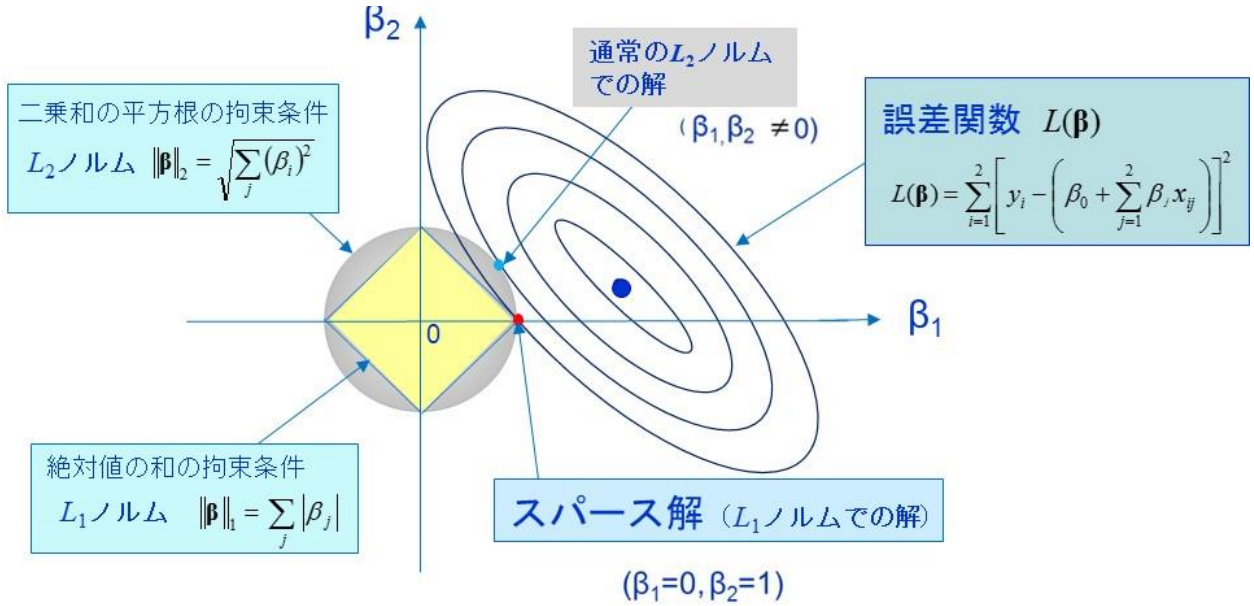
$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n}\|X\beta - Y\|_2^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|.$$

通常の最小二乗　正則化項 (絶対値)

## Lasso($L_1$型正則化重回帰分析)

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T\beta)^2, \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \le t.$$

寄与の低い$\beta j$ は0になる
⇒変数選択と次元落ち
正則化が同時に
達成できる




様々な変法：Larsアルゴリズム
(λを∞から減少), elastic net, adaptive lasso, grouped lasso

# 様々なスパース正則化の利用

- GWASへの応用
  GWASにおけるgene-gene interactionの取り込み
  主単独効果と相互作用
  - Correlated SNPs (Ayers and Cordell, 2010)
  - 検出力増大、FDR(false-discovery rate) 減少
    （He and Lin, 2011）
  - Pathwayに含まれているSNP間だけ相互作用を認める
    （Lu, Latourelle、2013)
- 遺伝子発現プロファイルへの応用
  - Biomarker（差異的発現遺伝子）が明確化
- 主成分分析にスパース正則化
  - 主成分分析にLasso回帰
  - 主成分の解釈が容易になる
  - 次を最小化

$$Q_\lambda(v_1, X) = \frac{1}{2}\mathrm{trace}[(X - z_1 v_1^T)^T (X - z_1 v_1^T)] + \sum_{j=1}^{p} p_\lambda(|v_{1j}|),$$

- 判別分析でも正則化により次元縮約

# Deep　Learning 以外の
# 人工知能の医療・創薬への応用

# IBM Watson

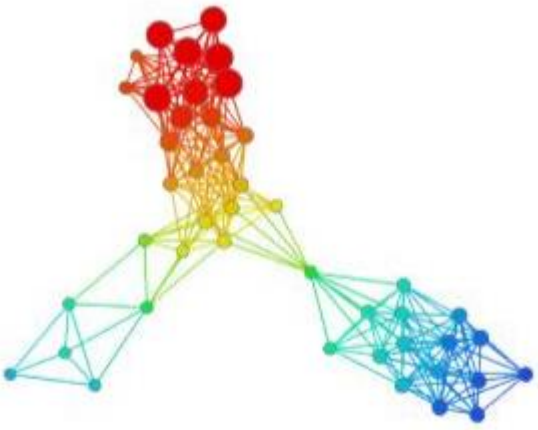
- Learning systemの不可欠性: IBM Watson
  - **自然言語処理、大量データベース探索、確信度付き解答**:　**Deep QAシステム** (jeopardy)
  - MITのSTARTと呼ばれるオンライン自然言語QAシステム:　質問をシンプルな質問に分解
  - CMUのOpen Advancement of Question-Answering Initiative（OAQA）システムが骨格
  - 質問解答に最も適切なテキスト資料を特定する知識源拡張アルゴリズム。テキストから知識を自動的に抽出
  - **大規模情報抽出、構文解析、知識推論により大量の情報資料をシステムの一般知識情報源に変換**
  - 自然言語理解に応用される統計学的学習理論（例えば、カーネル法）が基礎
- **Memorial Sloan-Kettering Cancer Center（MSKCC）**
  - The Oncology Expert Adviser software (OEA)
  - IBMワトソンの計算能力および自然言語処理技術と、MSKCCが持っている**臨床知見**（分子・ゲノムデータ、がん病歴の膨大なリポジトリなど）を組み合わせ、個々の患者にとって最高の治療方針を決定するのに役立つ、最新の研究に基づいた詳細な診断情報や治療の選択肢を見出す
- **New York Genome Center**
  - がん専門医ががん患者に対してより良い**個別ケアを提供できるよう支援するツール**として
    ゲノム研究専用にデザインされたWatsonの試作システム
  - 最初の対象：脳腫瘍のglioblastoma (グリア芽細胞腫）、
    ゲノム配列と医療情報、医学文献から**個別化治療**を提案

# 東京大学医科学研究所の Watson for Genomics

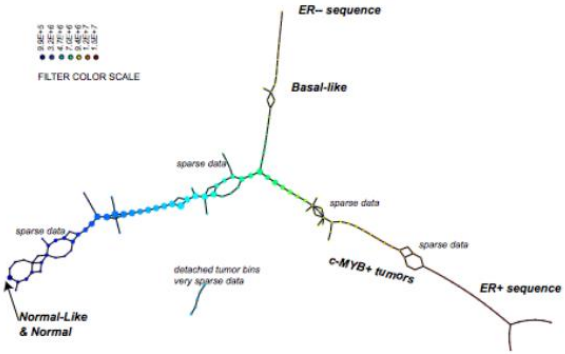
- Watson for Genomics（WfG）と東大医科研と共同研究
  - がん研究に関連する約2000万論文、1500万件以上の薬剤関連情報を学習
- 66歳女性**「急性骨髄性白血病」**と診断されて医科研病院に入院
  - ２種類の抗がん剤治療を半年続けたが回復が遅く敗血症などの危険もあった。
- がんに関係する女性の**1500の遺伝子変異情報**をWatsonに入力。
- STAG2遺伝子（cohesin複合体のサブユニット：染色分体欠損）の変異原因と発見
- WfGは急性骨髄性白血病の**「二次性白血病」**タイプと診断。
  WfGは抗がん剤を別のものに変えるよう提案。
- 女性は**数カ月で回復して退院し、現在は通院治療**を続けていている
- Watsonが治療法を助言した同様な例、医科研で41例

# 疾患のTopological Data Analyis

- 患者の網羅的分子情報や病態ビッグデータ
- Topological Data Analysis（TDA）を用いた機械知能で“病気の形”を描き臨床家に理解を容易にさせ、治療方針を決定させる



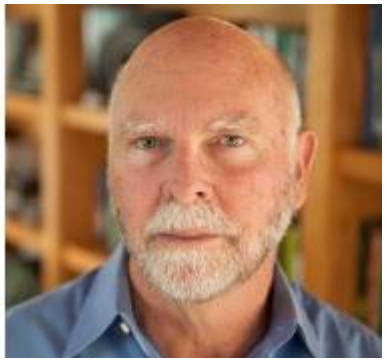
糖尿病のTDA病型




乳がんのTDA病型

# そのほかの機械学習

- **The ASCO (米国臨床癌学) CancerLinQ initiative**
  - 診療の現場(EHR)から大量の診療データを集め分析
  - 17万人のがん症例データベースを構築。新しい臨床治験へのガイドライン作
  - 各がん１～２万人の症例を集める
  - 学習システムを構築し治療知識を統計学習、ニューロネットを駆使して学習。
    BigDataにおけるLearning systemの不可欠性
  - 2013年に、CancerLinQのプロトタイプを完成、10万人以上の乳がんを蓄積、完全規模へ継続構築中
- **Cancer Commons initiative**
  - Rapid learningのインフラ整備
  - 目的：患者の個別症例と最新の知識を更新
  - 個々の患者の”Donate Your Data”(DYD)登録
- **Craig Ventor “Human Longevity Inc.”**
  - 健康寿命伸長のための**ゲノム科学、幹細胞治療**
  - 初期資本7000億円・医療費削減、HiseqX 5sets
  - 一年40000ゲノム（幼児から老人まで、患者・健常者も）収集し
    **最大のゲノムDB**を作る、臨床情報も収集、腸内細菌も含む　一日５人のヒト全ゲノム
  - がん（Mores Cancer Centerと提携）、糖尿病、認知症などの成人疾患に
- **Google X project, “Baseline”**
  - 健康に関する尺度発見、Conrad AのもとにDuke大学やStanford大学が協力
  - 現在175名、先制医療的なバイオマーカ探し、今後拡大



# Artificial Intelligence（AI）と創薬

- 標的分子選択とPOC(Proof of Concept)
  - 適切な分子標的の選択
- Virtual screening と選択
  - 適切な化合物に対するクラス判定
  - 研究例：ChEMBLに対するdeep learning
    - 13 M 化合物特徴量（ECFP12), 1.3M 化合物, 5k 薬剤標的
    - Ligand-based 標的予測,7種の予測法とAUC比較
    - Deep learningは、SVM, k-最近隣法, logistic回帰より優位
    - DLで構造活性相関を学習する
    - 特徴量の抽出、薬理機序への理解
  - リード最適化
- システム薬理学
  - ネットワーク病態学よりの創薬戦略
  - 他のシステムへの影響(毒性, 副作用)

Pharmacophoreの摘出

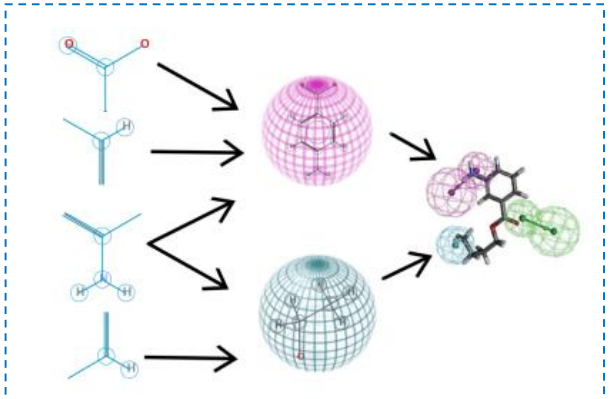


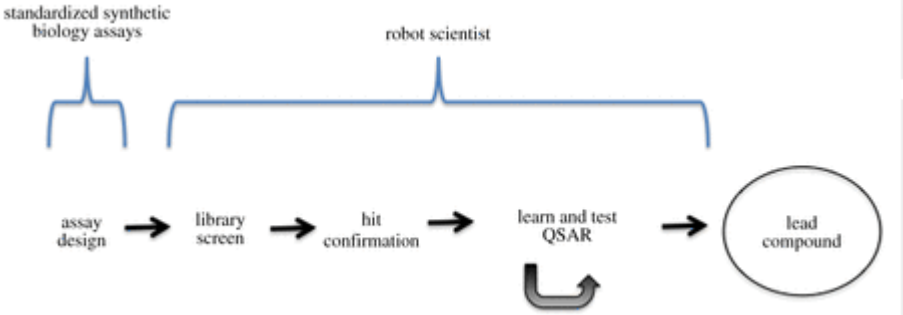
Figure Hierarchical nature of fingerprint features: by combining the ECFP features we can build reactive centers. By pooling specific reactive centers together we obtain a pharmacophore that encodes a specific pharmacological effect.

ECFP(chemical substructure: Enhanced cyan fluorescent)

# AI創薬のDL以外の方法

- Berg社のAI創薬
  - AIを方法として膵臓がんの抗がん剤を開発中
  - 膵臓がんと非患者の14兆のゲノム・オミックス情報を比較。
  - 調節不全パスウェイのシステム推定
  - システム薬理学的AIによる創薬
- マンチェスター大学（ Cambridgeとも共同）
  Artificially-intelligent Robot Scientist for new drugs
  - ライブラリースクリーニング, ヒット化合物の確証, リード化合物などの自動化
  - 構造活性相関（Quantitative Structure Activity Relationship）(QSAR) を反復学習する
  - 熱帯病、寄生体のDHFR（ジヒドロ葉酸還元酵素：薬剤耐性）を標的にして学習、細胞を合成生物学操作
    血管新生阻害因子（抗がん剤）をDR候補を探索
  - 最上位にコンセプト木（“root: assay triple screen”など)

Robot scientist Eve at work



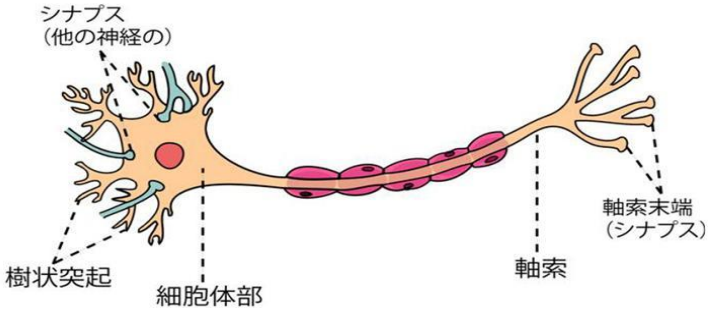

TMDU

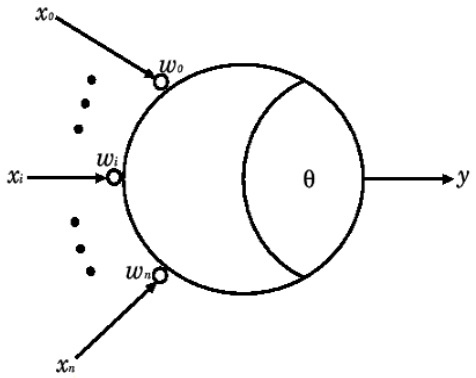
# Deep Learning 型人工知能の革命性

# ニューロ・コンピュータの歴史

- マッカロー･ピッツ（MCP）モデル
- 神経細胞（ニューロン）の機能
- 信号処理素子モデルとして提案
- １９４３年の神経生理学者マッカローと数学者ピッツにより提案される

入力信号　0,1




マッカロー・ピッツのニューロン・モデル

AND回路 θ＝1.5

OR回路　θ＝0.5

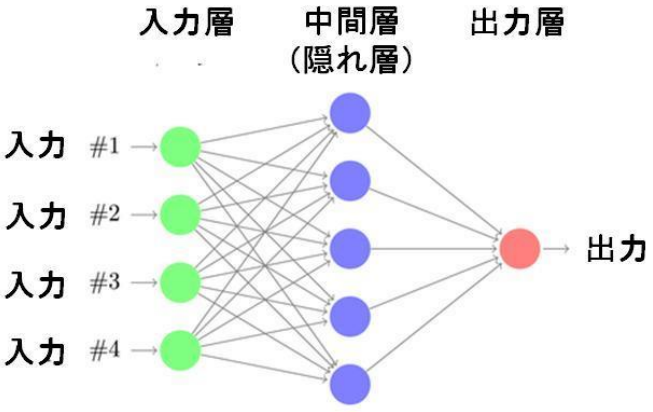
NOT回路

w=　-1.0

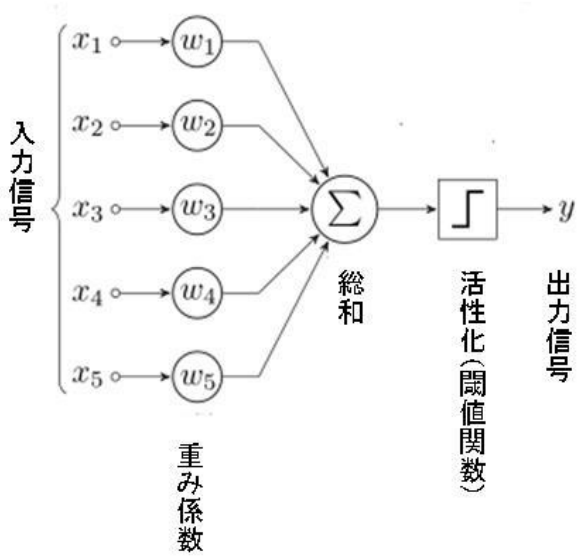
θ =　-0.5

# パーセプトロン・モデル

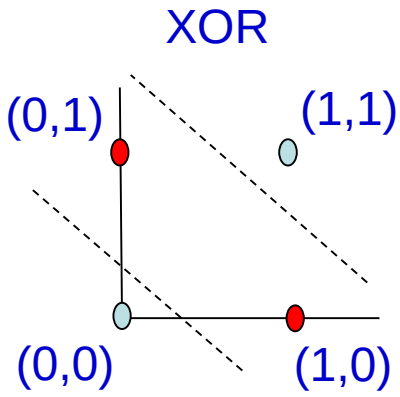
- パーセプトロンモデル
  - ローゼンブラットが1957年に提案 (58年論文)
  - ＭＣＰの神経モデルを用いて脳の**神経網を表すニューラルネットワークモデル**を構築
  - 学習法：出力誤差に合わせて$w_i$を大きさに比例して更新
- MITのMinskyによって線形分離問題しか解けないことを明示される⇒熱意の消失




(a) 3層表現のパーセプトロン



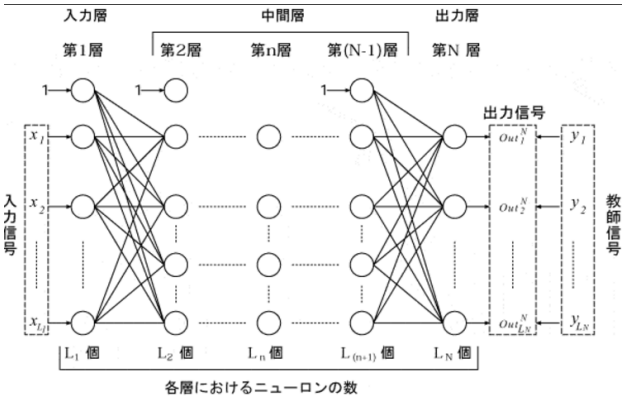

(b)単純パーセプトロン

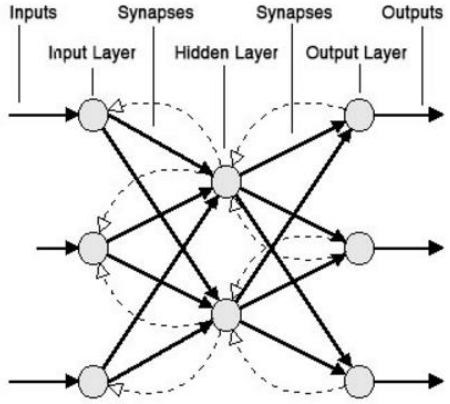



(c) 線形分離できない

# 多層パーセプトロンと逆伝播法

- 多層にすれば線形判別だけでなく非線形判別が可能になり、ニューロネットワークの識別能力も増大
- 多層にしたとき、加重ネットワーク結合の重みの更新法を1986年にルーメルハートが提案
- 出力層から誤差伝播ー逆伝播法（backpropagation法）
- 多層が深くなると入力層まで誤差修正が届かない
  - ブーム鎮静化



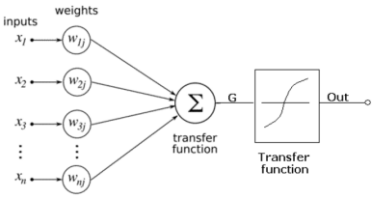




Back Propagation (1986　Rumelhart)
望ましい出力との誤差を教師信号として与える事により、逆方向に結合係数を変化させ、最終的に正しい出力が得られるようにする。
結合係数を変える事を学習と呼ぶ。
この学習方法には、最急降下法（勾配法）が使われる。出力層へ寄与の高いノードの重みが変更。

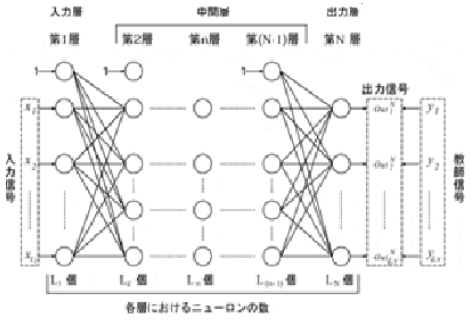
多層にわたる逆伝搬で修正感度減衰

# Deep Learning による人工知能革命

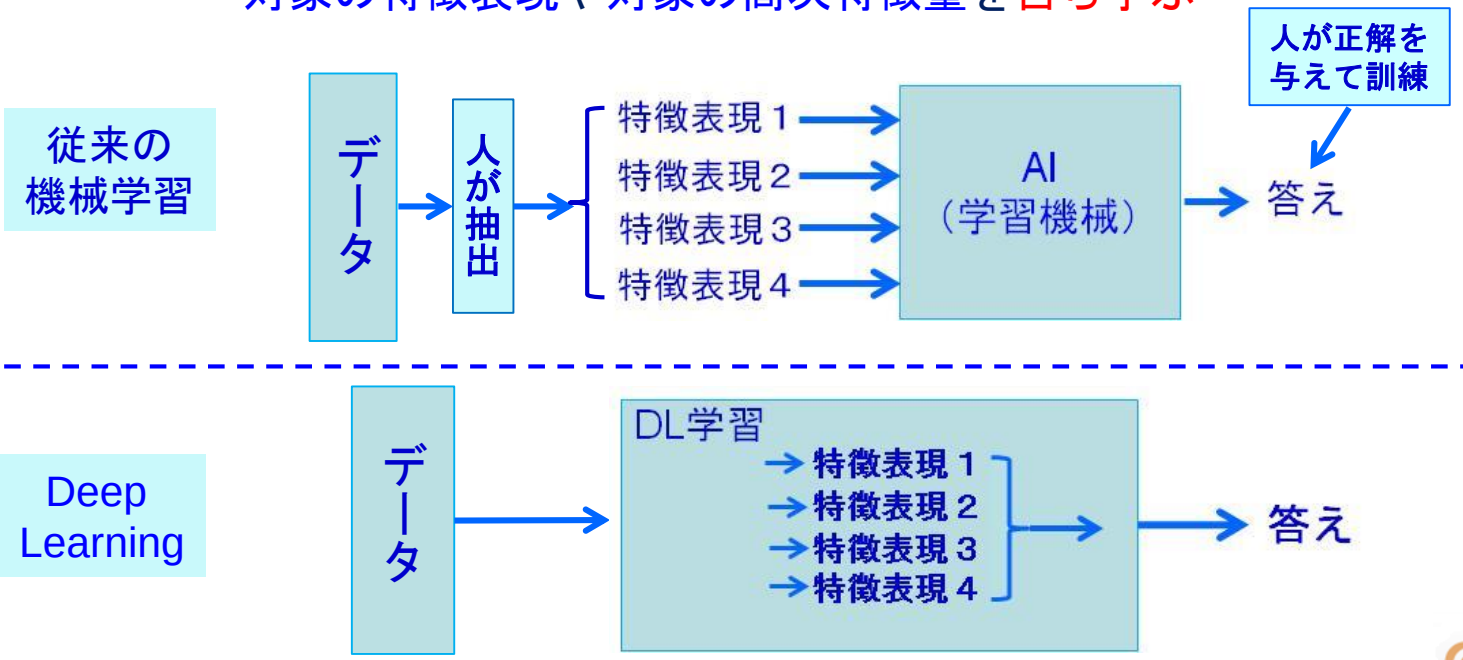
- 機械学習のこれまでの限界
  - 「教師あり学習」
    - 分類対象の特徴と正解を与え学習機械（AI）を構築
- Deep Learningの革命性
  - 「教師なし学習」
    - 対象の特徴表現や対象の高次特徴量を自ら学ぶ




神経情報素子



多層ニューロネットワーク

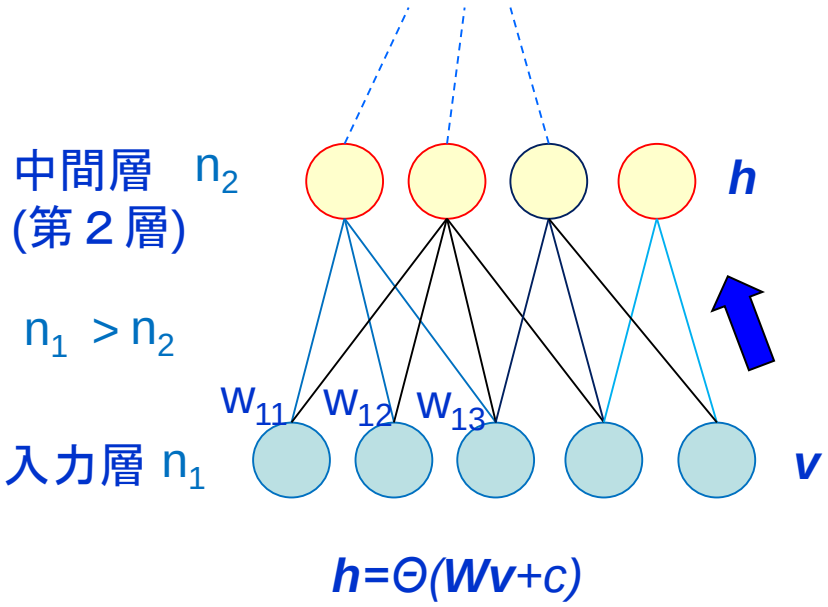


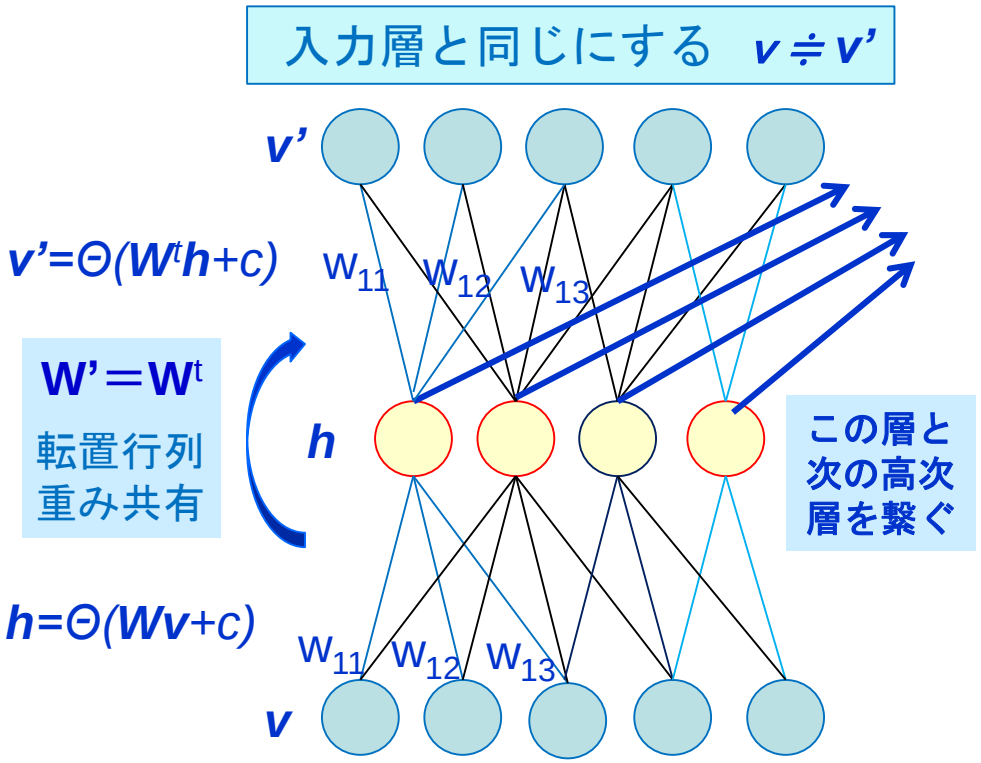
# Deep Learningの革命性

- DLは、まずは対象の固有の構造を記述する特徴表現や対象の高次特徴量を自ら学ぶ「教師なし学習」を行う
- 「内在的な特徴表現の学習」を自動的に行う
  - 自己符号化（Autoencoder）
  - 制限ボルツマンマシン
- 最終層で、人間の概念との相同をとるため「教師あり学習」

# DLの革命点 Autoencoder 1

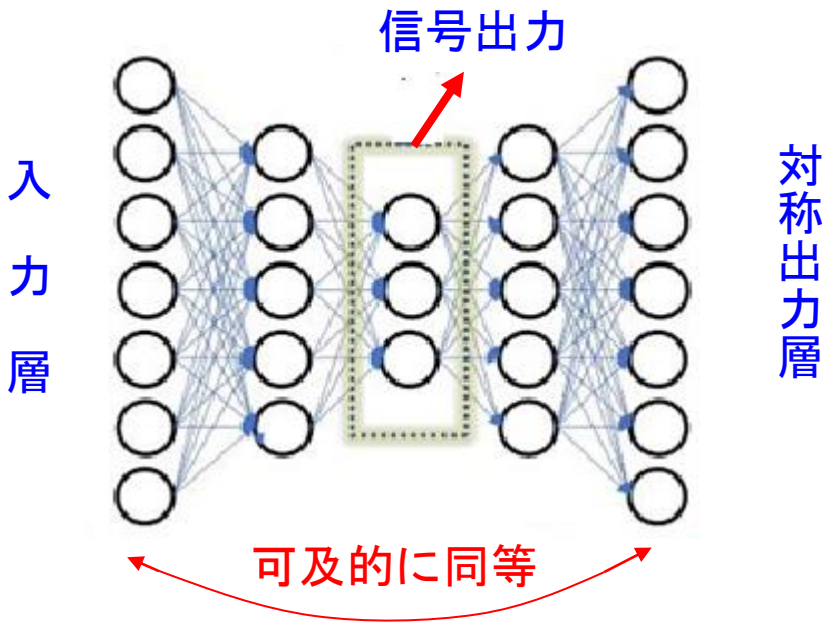
- 対象に固有な内在的特徴を学ぶ自己符号化の原理
- 格段ごとに入力の少ない中間層を入力へ逆投影して復元できるか
- 次元を圧縮され可及的に復元する（1000 nodes ⇒100nodes=?⇒1000nodes）
  →できるだけ復元に効果的な特徴量を探索する
  →内在的な特徴量を見出す



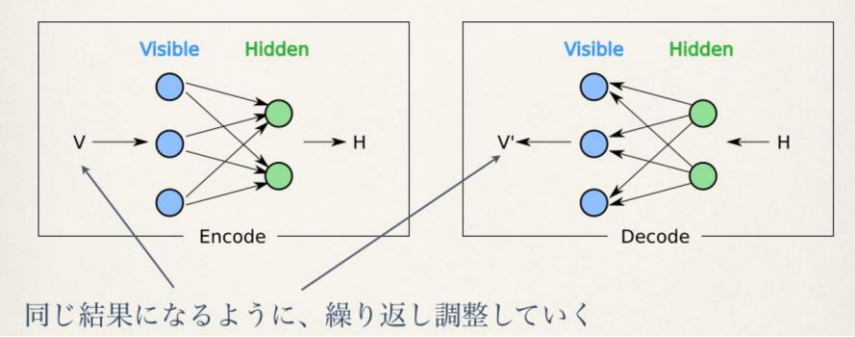
# DLの革命点 Autoencode２

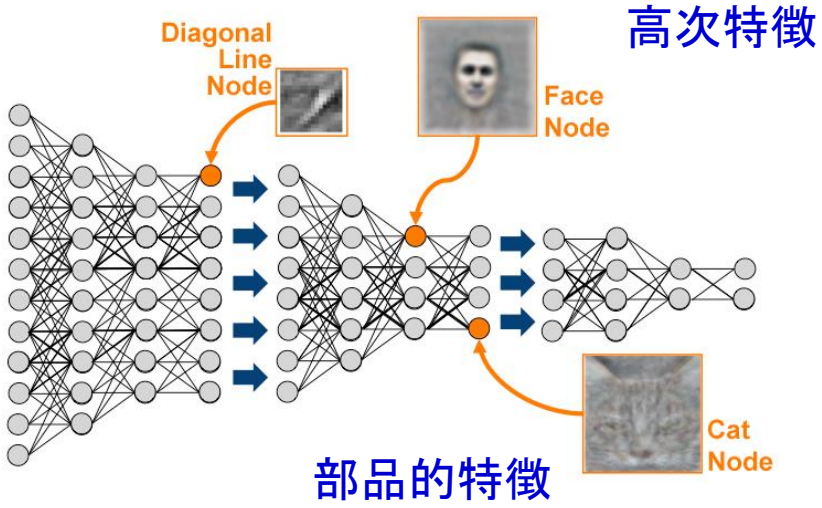
- 自己符号化器を多層に構成する
  - 積層自己符号化器（stacked autoencoder）
- 入力層と出力層を対称に層構成する
  - 深層自己符号化器（deep autoencoder）



# DLの革命点 Autoencoder 3

- 各層ごとに自己符号化を行うので何層でも組める
  → **各層間で「自己符号化」の積上げ**（autoencoder stack)
- 第一層で学習した特徴量を使って次の階層を作るので高次の特徴量が作られる
- 特徴的表現と概念を結びつけるため「教師あり学習」が最後に必要。
- 自動特徴抽出によってこれまでの学習手法の限界を克服した
  → 内在的な特徴量による構造的理解
- 人間の「思考の枠組み」を超えた正解の低次
  – **「アルファGo」が定石にない手で碁の名人に勝つ**

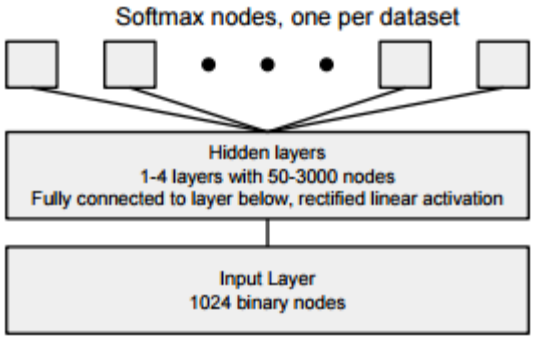


# Deep learning ：創薬からの注目

- 創薬を巡る状況
  - 平均14年、約2000億円（$1.7 B）の費用
  - 市場化された新薬の減少
  - 創薬に費やす期間・コストを低減したい
- Kaggle (データサイエンス競技会)にMerck社が出題 Molecular Activity Challenge (2012).
  - 15データセットから異なった分子の生物学的活動を予測するモデルの開発コンテスト
  - 勝利したモデルは深層学習 deep learning を用いたモデル
- Google in collaboration with Stanford (2015)
  - Stanford 大学の Pande 研究室と共同研究

バーチャルドラッグスクリーニングに対する deep learningによるツール開発
"Massively Multitask Networks for Drug Discovery“　特徴量を多目的に使う




Massively Multitask Networks

# Artificial Intelligenceと創薬

- 標的分子選択とPOC(Proof of Concept) ⇦ ここが重要
  - 適切な分子標的の選択
- Virtual screening と選択
  - 適切な化合物に対するクラス判定
  - 研究例：ChEMBLに対するdeep learning
    - 13 M 化合物特徴量（ECFP12), 1.3M 化合物, 5k 薬剤標的
    - Ligand-based 標的予測,7種の予測法とAUC比較
    - Deep learning: SVM, k-最近隣法, logistic回帰より優位
    - DLで構造活性相関を学習する
    - 特徴量の抽出、薬理機序への理解
  - リード最適化
- システム薬理学
  - ネットワーク病態学よりの創薬戦略
  - 他のシステムへの影響(毒性, 副作用)

Pharmacophoreの摘出

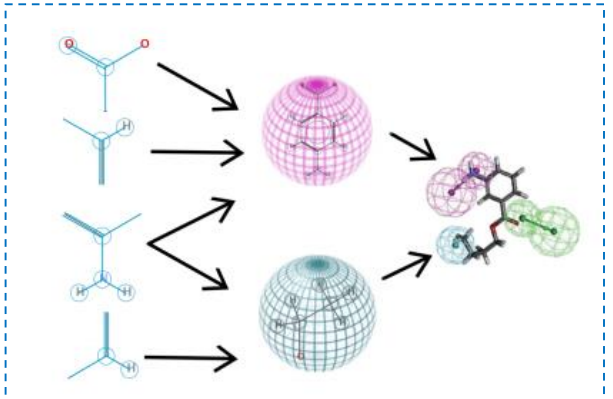


Figure Hierarchical nature of fingerprint features: by combining the ECFP features we can build reactive centers. By pooling specific reactive centers together we obtain a pharmacophore that encodes a specific pharmacological effect.

ECFP(chemical substructure: Enhanced cyan fluorescent)

# Deep Learningの医療応用への期待

- DLの医療の応用は開始段階で応用成功例は少ない
- 本質的に「教師なし学習」:人間が思いつかない解を提示
- 画像分類・解釈と文章理解が優れている
  遺伝子発現プロファイル解析や病態推移の理解への応用が期待
- いくつかのDeep Learningを用いた医療応用
  - ヒトmicrobiomeの分類・階層的表現を得た
  - ６つのがんで遺伝子発現をmiRNAとともに分類
  - 異なったMicroarrayを含むがん発現を分類の特徴表現を導き分類
  - Convolution ネットワークを使用して遺伝子発現を画像として分類
  - 遺伝子発現プロファイルの自動アノテーション

TMDU

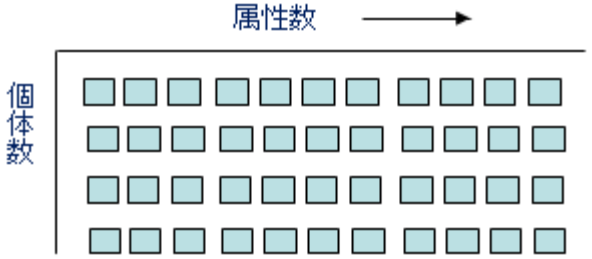
# Deep Learningの創薬へ応用

TMDU

# 「ビッグデータ」のData 縮約原理

問題点　　属性項目数(p)≫サンプル数(n)

p: 数億になる場合あり　　n: 多くても数万、通常数千




これら膨大な属性変数がすべて独立ならばビッグデータの構造解析は不可能。単変量解析の羅列（GWASのManhattan Plot)しか可能でない

ビッグデータ・スパース仮説

ビッグデータは、多数であるが属性値数より少ない独立成分が基底となって、相互にModificationして構成されている。
（独立成分の推定は、サンプル数とともに増加する）

データ次元縮約の原理（**principle of compositionality**)

# Deep Learningによる
# 多次元ネットワーク縮約法
(Hase,Tanaka 2017)

- 医療・創薬ビッグデータへの応用性高い
- 超多次元ネットワーク情報構造の急増
  - ゲノム医療＜網羅的分子情報ー臨床表現型情報＞
  - ゲノムコホートにおける＜遺伝子情報ー環境（生活様式）情報＞
- Deep Learning-based Network Contraction
  「DLネットワーク縮約法」

  超多次元ネットワーク情報構造⇒
  少数の特徴的ネットワーク基底に分解
- 線形分解ではない。非線形分解で基底への射影

TMDU

# タンパク質相互作用ネットワークでの疾患-薬剤-標的分子の学習

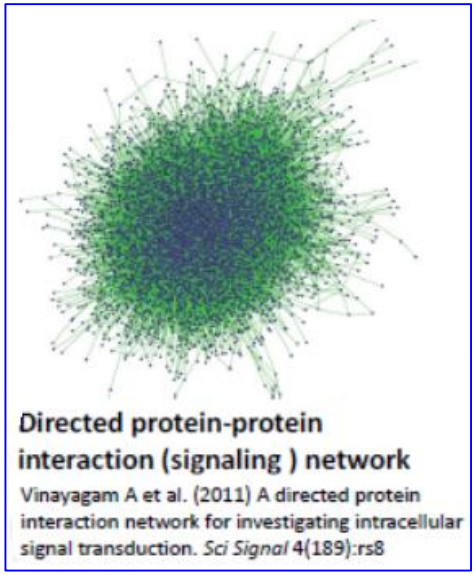
- ビッグデータ創薬/DR
  - タンパク質相互作用ネットワーク上での有効性予測
  - 基準指標：疾患関連分子と薬剤標的分子の距離
  - ネットワーク上のランダム歩行による総合距離（Sun, 2015）
  - 疾患関連遺伝子モジュールと標的分子の標準化近接指標
  - 判定情報量が不足
- AI創薬/DR
  - ビッグデータ創薬/DRの限界（情報の不足）をAI学習で補完
  - 既成の疾患-薬剤-標的分子の正例を学習（DrugBank)
  - 疾患関連分子と標的分子のタンパク質相互作用ネットワークにおけるトポロジカルな関係性を学習
  - 人工知能（AI）によって学習
  - 学習された疾患関連分子と標的分子の関係性のトポロジー特性により各分子の標的分子としての有効性を判定
  - 有力な標的分子を推測

# 特徴的ネットワーク基底への分解

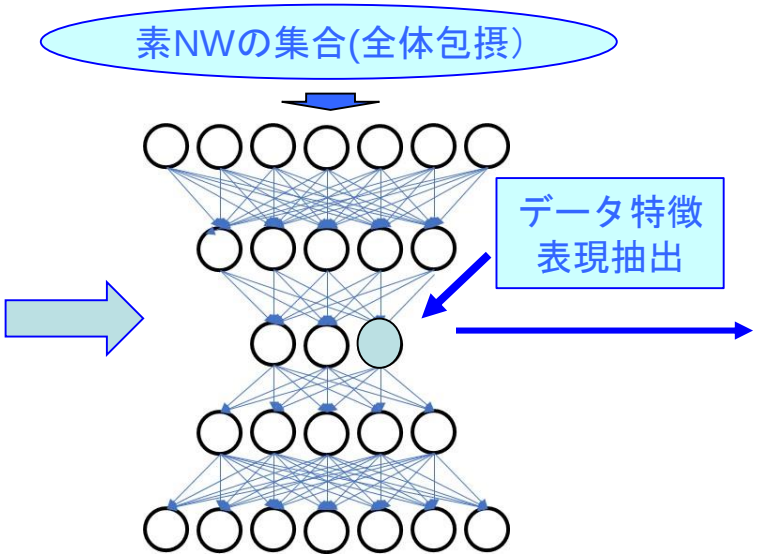
## 特徴的ネットワーク基底の和に縮約

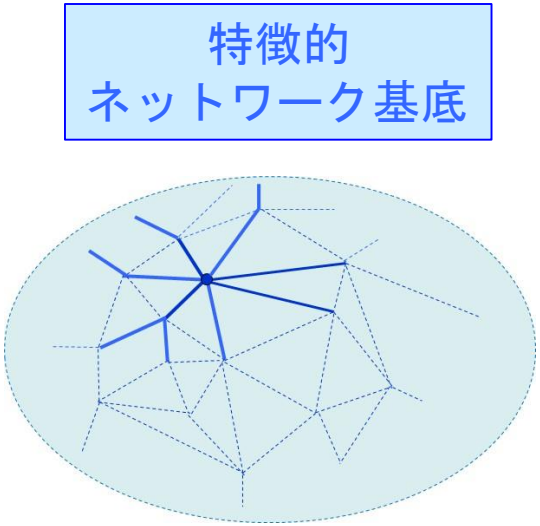
特定のノードを起点とした素NW（部分NW）の集合
全体NWを包摂する集合にDL反復自己学習

特徴的ネットワーク基底：トポロジーのみの構造/頻度構造




PPIネットワーク



# Deep Learningによる創薬・DR

**１）生体ネットワーク（PPIN) 特徴量の抽出**
- **タンパク質相互作用ネットワーク(PPIN)**のNW結合を学習し特徴表現（特徴NW基底）を出力。
- 学習集合を部分ネットワークの集合から決める
- ノードを起点とした素NWでPPIN全体を覆う集合

**２）多層Deep Auto-encoderのDLで学習.**
- 特徴的NW基底の「教師無し」学習
- 次元縮約による特徴的NW基底の抽出

**３）DL特徴NW基底空間における正例補完**
- DrugBankからの正例とその増加（SMOTE法）

**４）DL特徴NW基底量を用いた機械学習分類**
- Xgboot法などを用いたDL特徴量からの判別

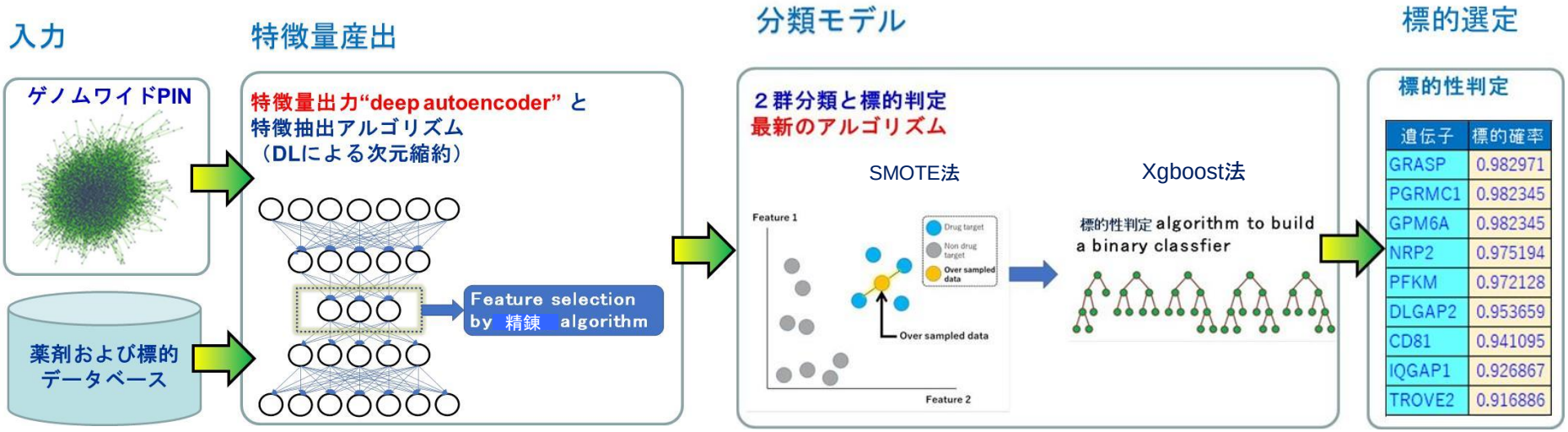
ネットワーク・タンパク質の標的性の判定

TMDU

# Deep Learningによる創薬・DR

分類部　**DrugBankを利用した**
**当該分子を標的とする既製薬剤の探索**

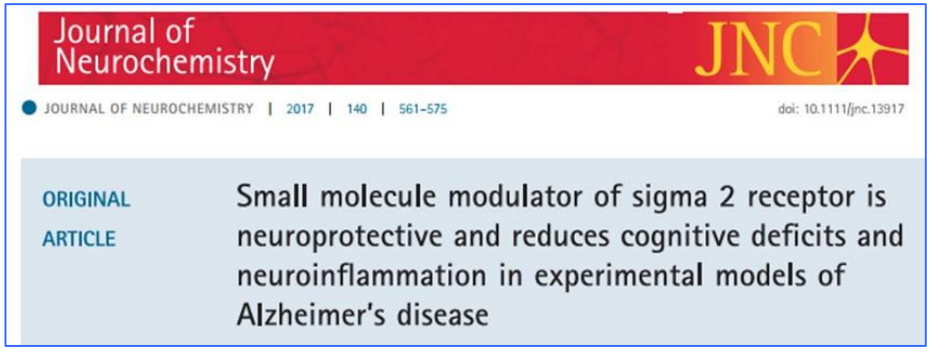
既製薬剤がない→新規薬剤探求（創薬）
既製薬剤がある→DRの検討




従来の機械学習（Random Forrest)と同じ成果は得られている

# 実験的研究との付合 1

## PGCM1：progesterone receptor membrane 1

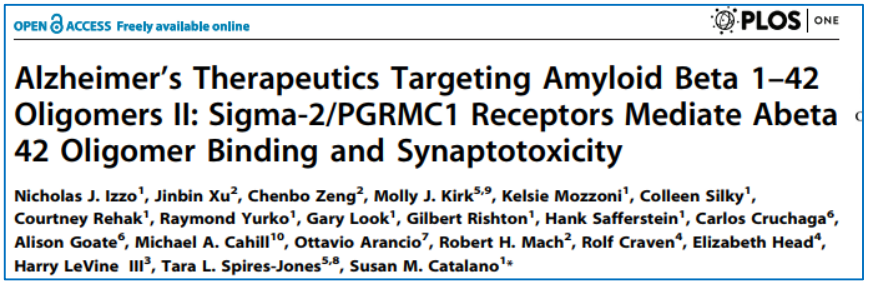

Journal of Neurochemistry

JNC

JOURNAL OF NEUROCHEMISTRY | 2017 | 140 | 561–575 doi: 10.1111/jnc.13917

ORIGINAL ARTICLE

Small molecule modulator of sigma 2 receptor is neuroprotective and reduces cognitive deficits and neuroinflammation in experimental models of Alzheimer's disease

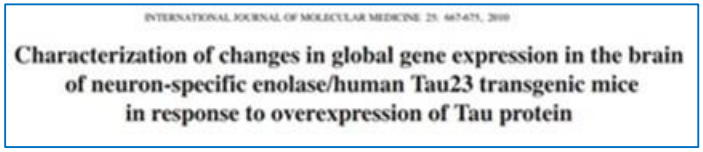

OPEN ACCESS Freely available online

PLOS ONE

Alzheimer's Therapeutics Targeting Amyloid Beta 1–42 Oligomers II: Sigma-2/PGRMC1 Receptors Mediate Abeta 42 Oligomer Binding and Synaptotoxicity

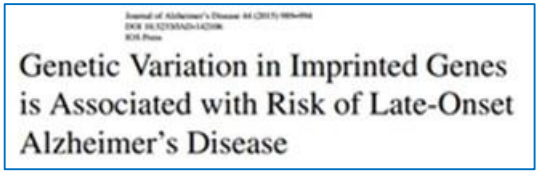
Nicholas J. Izzo[1], Jinbin Xu[2], Chenbo Zeng[2], Molly J. Kirk[5,9], Kelsie Mozzoni[1], Colleen Silky[1], Courtney Rehak[1], Raymond Yurko[1], Gary Look[1], Gilbert Rishton[1], Hank Safferstein[1], Carlos Cruchaga[6], Alison Goate[6], Michael A. Cahill[10], Ottavio Arancio[7], Robert H. Mach[2], Rolf Craven[4], Elizabeth Head[4], Harry LeVine III[3], Tara L. Spires-Jones[5,8], Susan M. Catalano[1]*

## GPM6A：Glycoprotein M6A

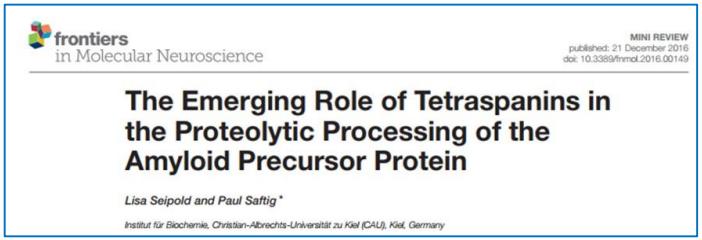

INTERNATIONAL JOURNAL OF MOLECULAR MEDICINE 25: 667-675, 2010

Characterization of changes in global gene expression in the brain of neuron-specific enolase/human Tau23 transgenic mice in response to overexpression of Tau protein

## DLGAP2：DLG-Associated Protein 2


Genetic Variation in Imprinted Genes is Associated with Risk of Late-Onset Alzheimer's Disease

## CD81:Tetraspanins family


frontiers in Molecular Neuroscience

MINI REVIEW

published: 21 December 2016

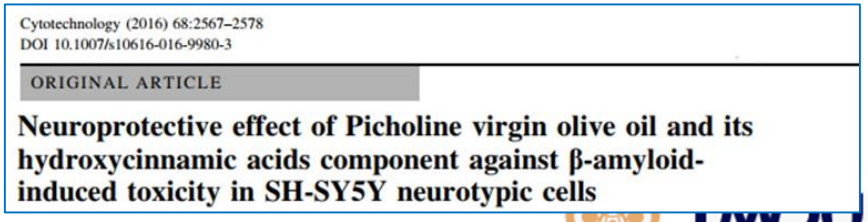
doi: 10.3389/fnmol.2016.00149

The Emerging Role of Tetraspanins in the Proteolytic Processing of the Amyloid Precursor Protein

Lisa Seipold and Paul Saftig*

Institut für Biochemie, Christian-Albrechts-Universität zu Kiel (CAU), Kiel, Germany

## PFKM: Phospofructokinase


Cytotechnology (2016) 68:2567–2578

DOI 10.1007/s10616-016-9980-3

ORIGINAL ARTICLE

Neuroprotective effect of Picholine virgin olive oil and its hydroxycinnamic acids component against β-amyloid-induced toxicity in SH-SY5Y neurotypic cells

# 実験的研究との付合 2

WISP-2/CCN5：WNT1 inducible signaling pathway protein 2

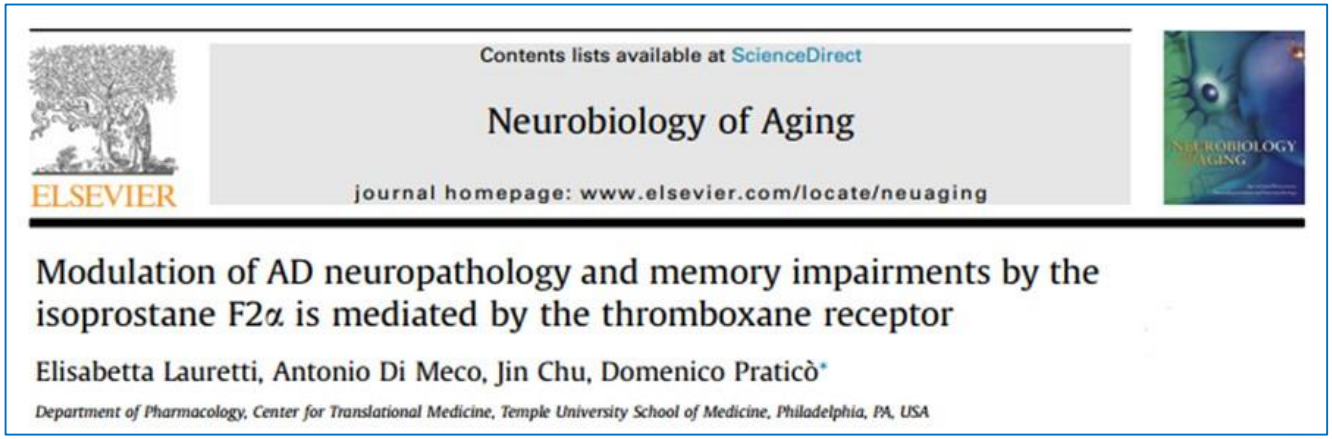
frontiers in Cellular Neuroscience

REVIEW
published: 24 June 2015
doi: 10.3389/fncel.2015.00237

**Matricellular proteins of the Cyr61/CTGF/NOV (CCN) family and the nervous system**

*Anna R. Malik, Ewa Liszewska and Jacek Jaworski**

TBXA2R: thromboxane A2 receptor



Contents lists available at ScienceDirect

Neurobiology of Aging

journal homepage: www.elsevier.com/locate/neuaging

ELSEVIER

**Modulation of AD neuropathology and memory impairments by the isoprostane F2α is mediated by the thromboxane receptor**

Elisabetta Lauretti, Antonio Di Meco, Jin Chu, Domenico Praticò*

*Department of Pharmacology, Center for Translational Medicine, Temple University School of Medicine, Philadelphia, PA, USA*

TMDU

# DL型NNへの期待と困難点

- 医療・創薬の応用は大きく期待される
  - 本質的に「教師なし学習」:人間が思いつかない解を提示
  - **現状では**、画像分類・解釈と文章理解が優れているので、遺伝子発現プロファイル解析や病態推移の理解への応用
    - 例：ヒトmicrobiomeの分類・階層的表現を得た
    - ６つのがんで遺伝子発現をmiRNAとともに分類した。
    - 異なったMicroarrayを含むがん発現を分類の特徴表現を導き分類した。
    - Convolution ネットワークを使用して画像としての遺伝子発現を分類した。
    - 遺伝子発現プロファイルの自動アノテーション
  - **期待される本質的な寄与**
  - **超多次元（生命医学）ネットワークから革新的知の発見**
- DL型ニューラルネットは困難点もある
  - 特徴表現を自己学習するが基本的にはBlack Boxで解析が必要
  - 大量のデータを必要とする
  - DL型NNには、**ハイパーパラメータ**が多種類があり、使用に関して選択問題が残る
  - 計算時間が長くコストが大きい。

TMDU

# Real-World- Dataを用いた創薬/DR戦略

## ―RCT, EBMからの呪縛の解放―

TMDU

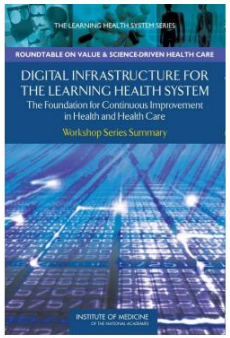
# 「学習する医療システム」
# Learning Health System

新しい生物医学知識が臨床実践に給されるまで17年
臨床データを用いて医療を実施しながら医療を改善

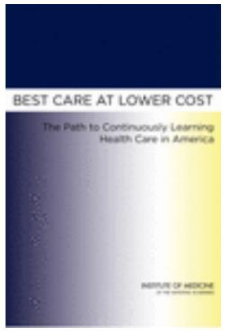
- IOM “Clinical Data as a Basic Staple of Health Learning”
- 医療システムのデジタル化（IT化）は必然の傾向である
- 「ルーチンの医療活動から集められたデータ（形式的臨床研究と違って）がLHSを支える鍵である」
- データを共有することによって学習して医療システムを改善
- RCTは「黄金基準」であるが、通常の医療システムの外で実施されている。医療が実際対象とする患者集団を代表しているのか。
- RCTは時間が掛かり費用もかかる
- 有効な知識の蓄積の速度が加速する

IOM(Institute of Medicine)のレポート
2007年にEBM/RCT（無作為試験）に変わるパラダイムとして提案

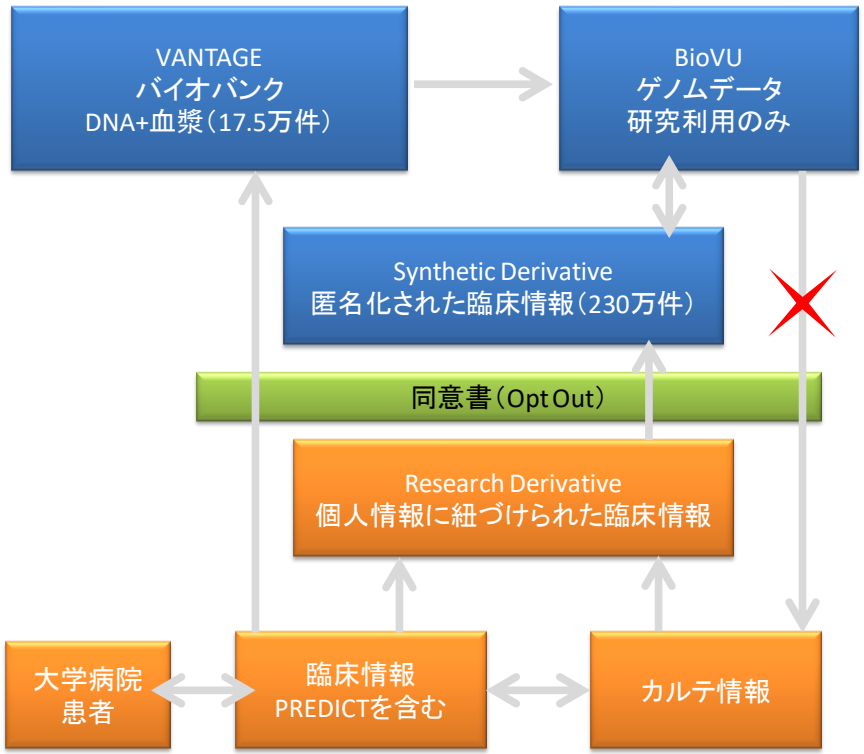
***Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care***




***Best Care at Lower Cost: The Path to Continuously Learning Health Care in America***



# LHSの代表例　BioVU
# ゲノム情報と電子カルテ情報を用いた Vanderbilt大学病院の医療情報システム




## 電子カルテ

Synthetic Derivative：電子カルテから匿名化臨床表現型のデータベース　230万件。Opt out 形式

## バイオバンクと遺伝子解析

BioVU ：Synthetic Derivativeと連結可能なGenome DNA情報

VANTAGE Core ：検体17.5万件、血液検からDNA抽出・ゲノム解析、バイオバンク運営

PREDICT ：臨床レベルの遺伝子解析情報により、薬物副作用防止などを実現するシステムを自らの医療システムにより知識抽出して実現する

クロビドグレル（抗血栓剤）の遺伝子多型関してABCB1, CYP2C19、さらにPON1の多型が知られていたが、ヒトを対象とした臨床実験の報告はなかった。SDから循環器疾患でclopidogrelの投与歴の対象者（ケース群）およびコントロール群を選出。BuiVUから遺伝型を決定する。この条件に合致するケース群は255件。解析の結果、CYP2C19*2とABCB1の関与は有意。PON1は非有意が判明した。

TMDU

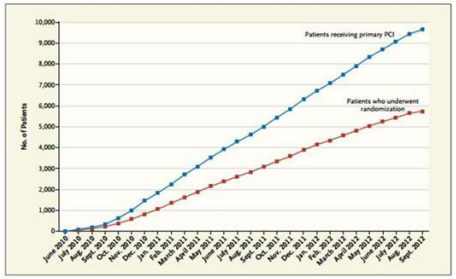
# 個別化（層別化）医療の概念の普及とRCTの限界

- 個別化・層別化の概念の浸透
- RCTの治験集団とReal World Dataの乖離
  - すべての個別化パターンを包摂した治験集団は現実には不可能
  - 現在の治験集団
    - 大半のRCTは医療現実の外の「人工的な環境」
    - 高齢者・妊婦はいない、欧米では黒人とくに青年は含まれない
- 将来へ向けたプラットフォームの確立
  - 母集団に近いReal World 医療データが収集可能
    ⇒データの大規模化の「**n = All**」の実際
  - Real World Data時代の臨床研究のプラットフォームを形成。
    ⇒ RCTとReal World Dataの融合としての
    registry-based clinical randomized trial
  - 我が国の戦略　段階的移行

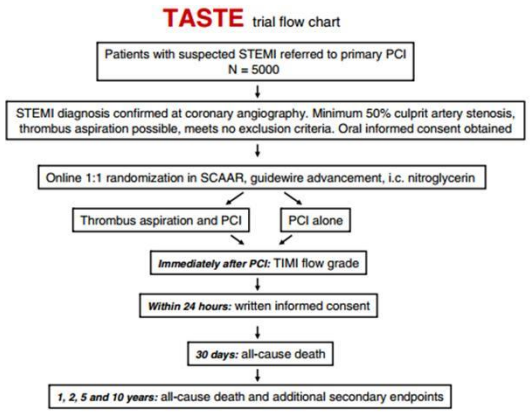
# Biobank (registry) 準拠の創薬過程・治験

- スウェーデンの**TASTE**(STsegment-Elevation MI in Scandinavia)
  **Registry-based randomized clinical trial（RRCT）**
- 国家的網羅的なRegistry登録者から治験対象者を選ぶ
  - **SCAAR**(Swedish Coronary Angiography Registory)
- 選んだ集団で心筋梗塞のPCI*療法で
  - 血栓吸引を行った後、PCIを行う
  - PCIのみを行うの
  - **2群にランダム化割付し**
  - 術後30日での生存をendpointとして治験を行う
- **治験のエンドポイントは疾患レジストリーの追跡**
- これまで小規模の治験では相反する結論
- Observational研究：Population 型コホートでは困難
- 経費は通常なら100万円程度を**50ドル**で済んだ。

*経皮的冠動脈形成術（percutaneous coronary intervention)



Rapid Randomization in the TASTE Trial, with Enrollment of Most Patients Receiving Primary Percutaneous Coronary Intervention (PCI).
Adapted from the Institute of Medicine (www.iom.edu/~/media/Files/Activity%20Files/Quality/VSRT/LST%20Workshop/Presentations/Granger.pdf). The incremental cost of the Thrombus Aspiration in ST-Elevation Myocardial Infarction in Scandinavia (TASTE) trial was $300,000, or $50 for each participant who underwent randomization.



# RRCTの特徴

- 質の高い大規模臨床レジストリーと前向き無作為化治験の長所を結合
  - **被治験者選択が容易**
  - **迅速な登録**
  - **非登録患者の制御**
  - **非常に長期にわたる追跡**が可能
    - アルツハイマー症など
  - **経費が掛からずデザインが単純**
- **疾患レジストリーの方で**
  - 患者鑑別
  - 無作為化
  - ベースライン情報の収集
  - エンドポイントの探索を行ってくれる

RRCTを創設したJames教授と筆者
（ウプサラ大学にて, 2017）

**ビッグデータ研究とRCTの融合: 将来の治験方式**

# ゲノム・オミックス医療の次世代の展開

# ゲノム医療の第2世代

**成功した臨床実装**

1. **希少先天遺伝疾患**の**原因遺伝子**を**病院の現場で**シーケンサにより同定
2. **がんのドライバー遺伝子変異**を同定、適切な**分子標的薬**を処方
3. 患者の**薬剤の代謝酵素の多型性**を**先制的に同定**し、副作用を防ぐ

しかし

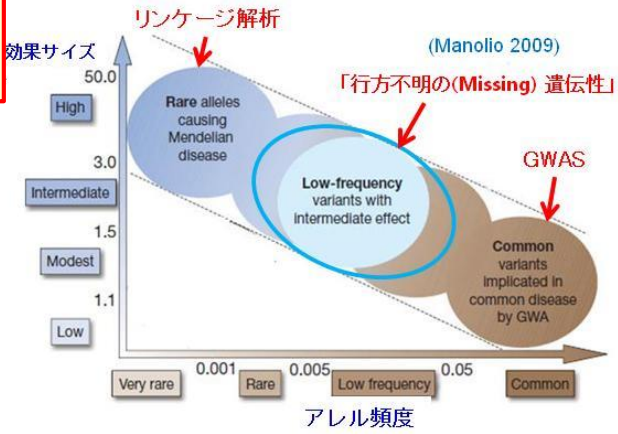
## 多因子疾患の機序/発症予測は無着手である

- 「単一遺伝的原因」帰着アプローチの限界
- 「行方不明の遺伝力」の主要な原因
  - 複数の疾患関連遺伝子間の相互作用: G x G
  - 環境と遺伝子の相互作用が：G x E

SNPの相対リスク
低い(1.1～1.3)理由
GxE組合せ特異的効果を環境要因の平均


多因子疾患は個人の＜遺伝的体質と環境要因＞の＜相互作用の結果。シーケンスだけでは解明不能

疾患発症の遺伝要因と環境要因の相互作用は
加算的 $(G \oplus E)$でもなく乗算的 $(G \otimes E)$ でもない
＜(G,E) 組合せ特異的な効果＞である
例　大腸がんの遺伝要因と環境（生活習慣）要因



# 大半の疾患の基礎としての「遺伝素因X環境要因」の相互作用

一部の単一遺伝病を除き、大半の疾患（Common diseases)の発症は

疾患発症の相対リスク=

遺伝要因(G:genome) X 環境要因(E:exposome)

相互作用は加算的 でもなく乗算的 でもない

＜**(G,E) 組合せ特異的な効果**＞である

GWASでSNPの相対リスクが低い(1.1～1.3)理由： **GxE組合せ特異的効果を環境要因の全てに亘って平均しているからである**

# 発達プログラム説 DOHaD

## Developmental Origin of Health and Disease)

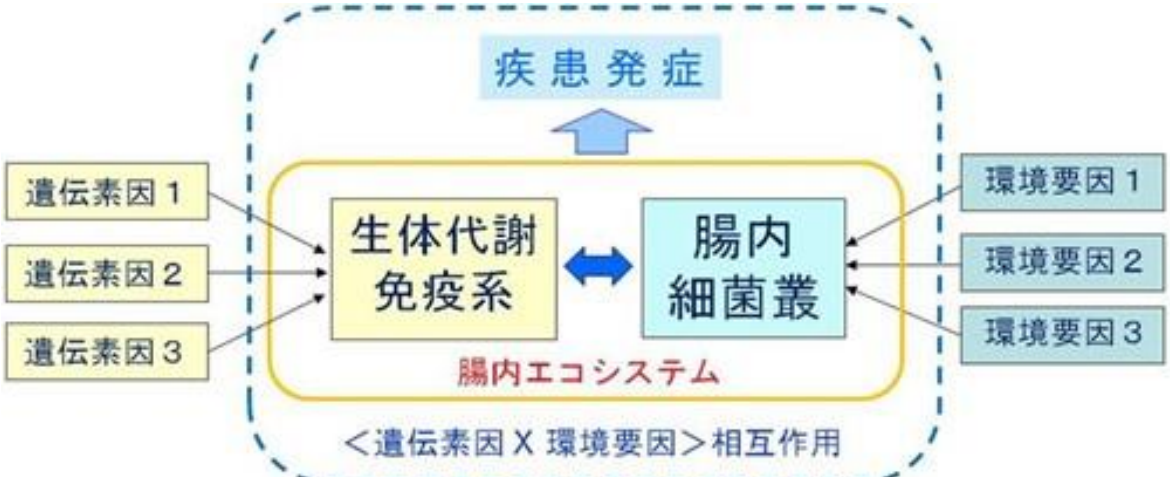
- オランダ飢饉
  - 第2次大戦末期、ナチスの封鎖、約半年間酷い飢饉
  - 飢饉の期間に胎児、戦後30年
  - 成人期:肥満,糖尿病,心筋梗塞,統合失調
- Baker仮説：英国心筋梗塞増加
- エピジェネティク機構
  - 過度な低栄養：肝臓のPPARα/γ（倹約遺伝子）メチル化低下・遺伝子発現がオン
  - エピジェネティック変化は**可変**：短期的変化、**長期的「記憶」**次の世代も

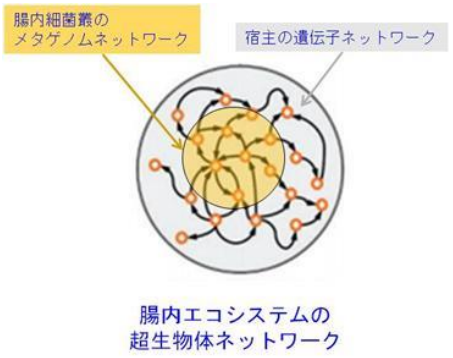
オランダ
飢饉 (1944)

環境因子 → Epigenome変化 → 遺伝子発現調節 → 疾病発症

# 腸内細菌叢microbiome：メタゲノム

- **疾患の環境発症要因**（exposome）
  - 腸内microbiome：環境要因の最大の１つ
- **腸管微生物叢**（gut microbiome）
  - 約1000種類、100兆個、総重量１～1.5kg,「**実質的な臓器**」
  - 遺伝子数個人あたり**約50万遺伝子**、総数：数100万遺伝子
- **免疫系、炎症系、粘膜免疫細胞群との相互作用**
  - **食物の難消化性の食物繊維**：腸内細菌によって嫌気的に代謝、酪酸などの「**短鎖脂肪酸**」がエネルギー源となる
  - 食事・栄養物質による環境要因は、腸内細菌叢の代謝物（短鎖脂肪酸やTMAOなど）から宿主の生体機構に相互作用

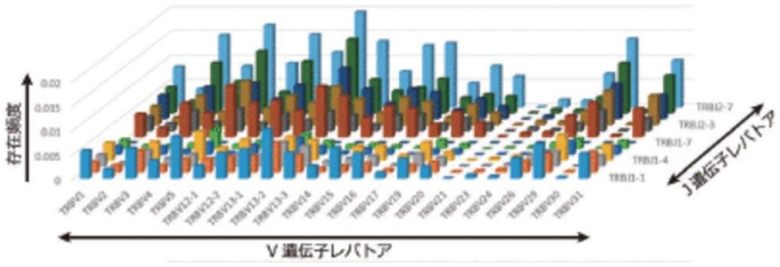



**メタゲノム**
**超生物ネットワーク**

# 免疫ゲノム

- 可変領域や相補性決定領域（特にCDR３）のDNAやRNAを次世代シーケンサ(HTS)で解析
- レパトア解析
  - 抗原受容体全体のプロファイルを俯瞰的に把握できる
  - V(D)Jなどの成分を基軸として3次元表示可能。
  - 疾患罹患ともに瞬時に全体像が変化する。
  - 網羅的病態全体像を提示する
    - VDJの使用頻度
    - 多様性(diversity)の変化
    - 疾病/加齢レパトア分布変化
- 臨床シーケンスに含まれる
- ３次元分布の特徴分析




（レパトア・ジェネシス社）

# 第2世代のゲノム・オミックス医療

- 生涯的全体性においてその個人の疾患可能性の全体性を把握し、個別化予防、個別化治療に取り組む
- ゲノム・オミックス情報と医療・健康
  - **Clinical Sequencingのインパクト**
- 第１世代ゲノム医療
  - ゲノムの変異・多型性の個別性に基づく
- 第２世代のゲノム医療
  - 多因子疾患が対象、環境情報との相互作用
  - エピゲノム、メタゲノム・免疫ゲノムなど

**疾患メタ・オミックス修飾**

TMDU

# 今後の戦略・方向

- 第２世代のゲノム医療・創薬
- Deep Learningによる〈多次元ネットワーク情報構造〉の縮約
  - 創薬だけでなく、ビッグデータ医療への適応可能
  - ゲノム医療の〈網羅的分子情報ー臨床表現型〉の相関ネットワーク構造
  - バイオバンクの〈遺伝素因ー環境要因〉と発症
- AI創薬の「枠組み」実行方向は「見えてきた」
- 本年中に、いよいよAI創薬の実装に着手しなければならない。米国に持って行かれる。
  - 製薬企業、IT企業、医療機関を束ねた集中的プロジェクトを推進するために「ビッグデータ医療・AI創薬コンソーシアム」を設立する

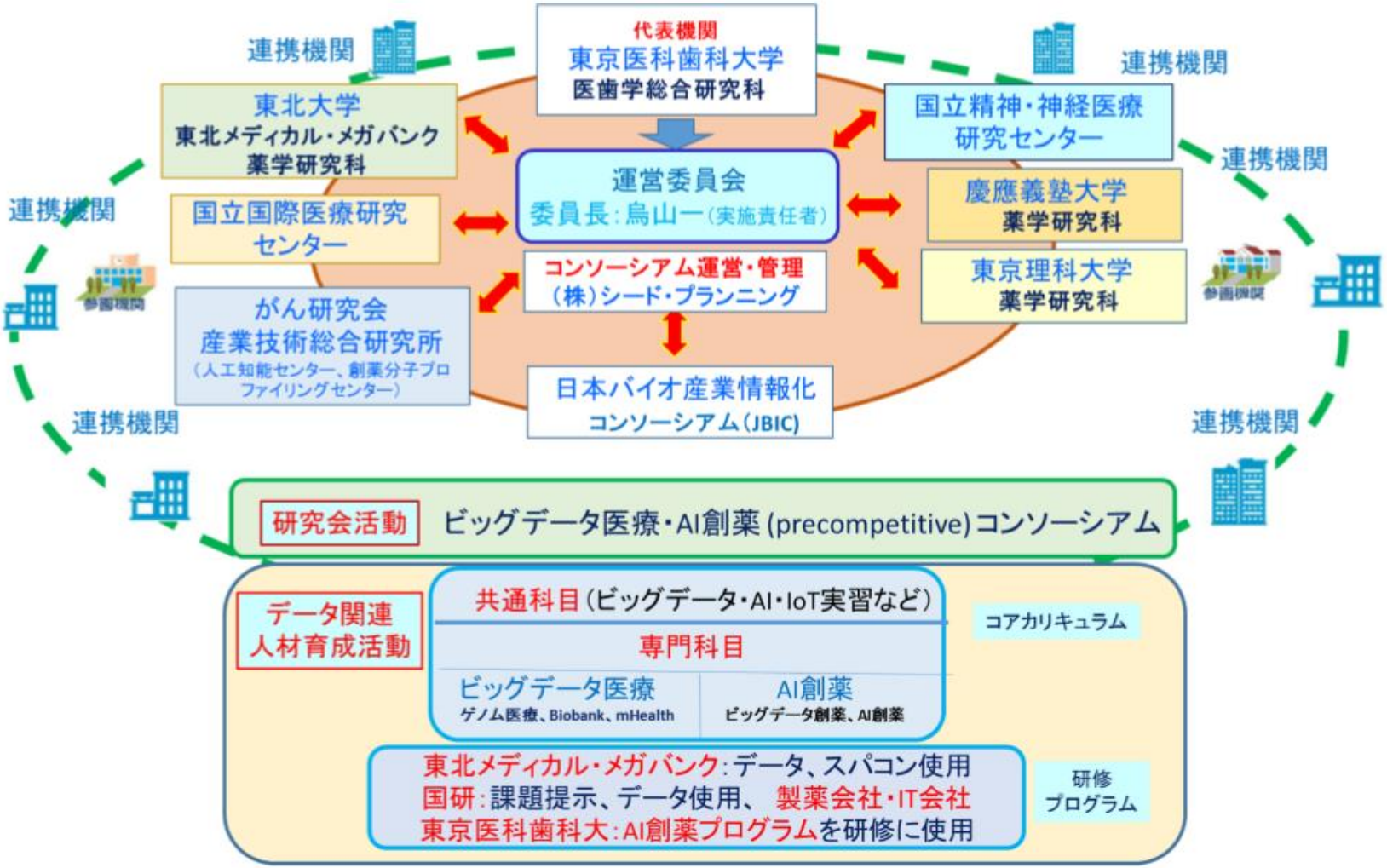
TMDU

田中　博　著

「AI創薬・ビッグデータ創薬」

薬事日報社　6月19日刊行

# ビッグデータ医療・ＡＩ創薬コンソーシアム



# ご清聴ありがとうございました