人工知能（AI）とビッグデータの医療と創薬への応用

東京医科歯科大学　名誉教授
医学部　臨床腫瘍学分野　生命情報学
東北大学　東北メディカル・メガバンク機構
田中　博

医療ビッグデータ時代の到来

医療ビッグデータ時代の到来

(1) 次世代シーケンサ（Clinical Sequencing)による
「ゲノム/オミックス医療」の網羅的分子情報蓄積
(2)モバイルヘルス(mHealth）によるWearable センサ情報の継続的蓄積（unobstructed monitoring)
(3) GWAS, Biobankによるゲノム・コホート情報の蓄積

大量データの急激な
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・予測医療
健康・医療の適確性の飛躍的な増大

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
  - 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
  - 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**
  - **Biobank,GWASによるゲノム情報**
- **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

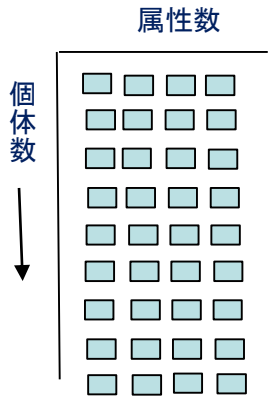
医療の「ビッグデータ革命」

～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)



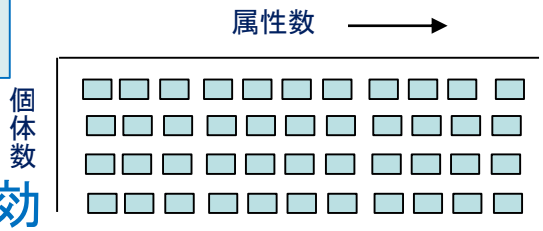
医療情報・疫学調査では 属性数：10項目程度

– 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

属性に比べて個体数 少数:従来の統計学が無効

「新NP問題」：多変量解析:GWASで単変量解析の羅列

– 目的：例えば医療の場合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

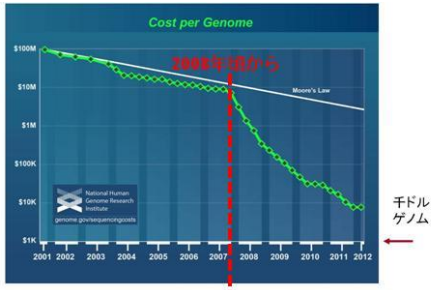
TMDU

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”
    - 個別化医療を実現するために＜個別化・層別化パターン＞を網羅的に調べる：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,個別化概念に破綻した
  - <statistical evidence based>呪縛からの解放
  - 「標本」統計・「推測」統計学に限定されない臨床研究
  - Real World Data:ビッグデータ知識生成（BD2K)
- 創薬の戦略パラダイムの転換
  - ＜ビッグデータ創薬＞の可能性
  - 網羅的分子データからの計算機創薬・システム創薬
  - Disease network, Drug networkの双対network による創薬

TMDU

ビッグデータとゲノム医療の流れ



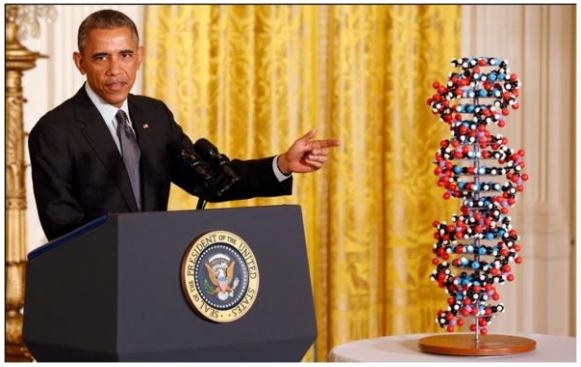
DNA Sequencing Cost: the National Human Genome Research Institute

シーケンス革命　2007/8

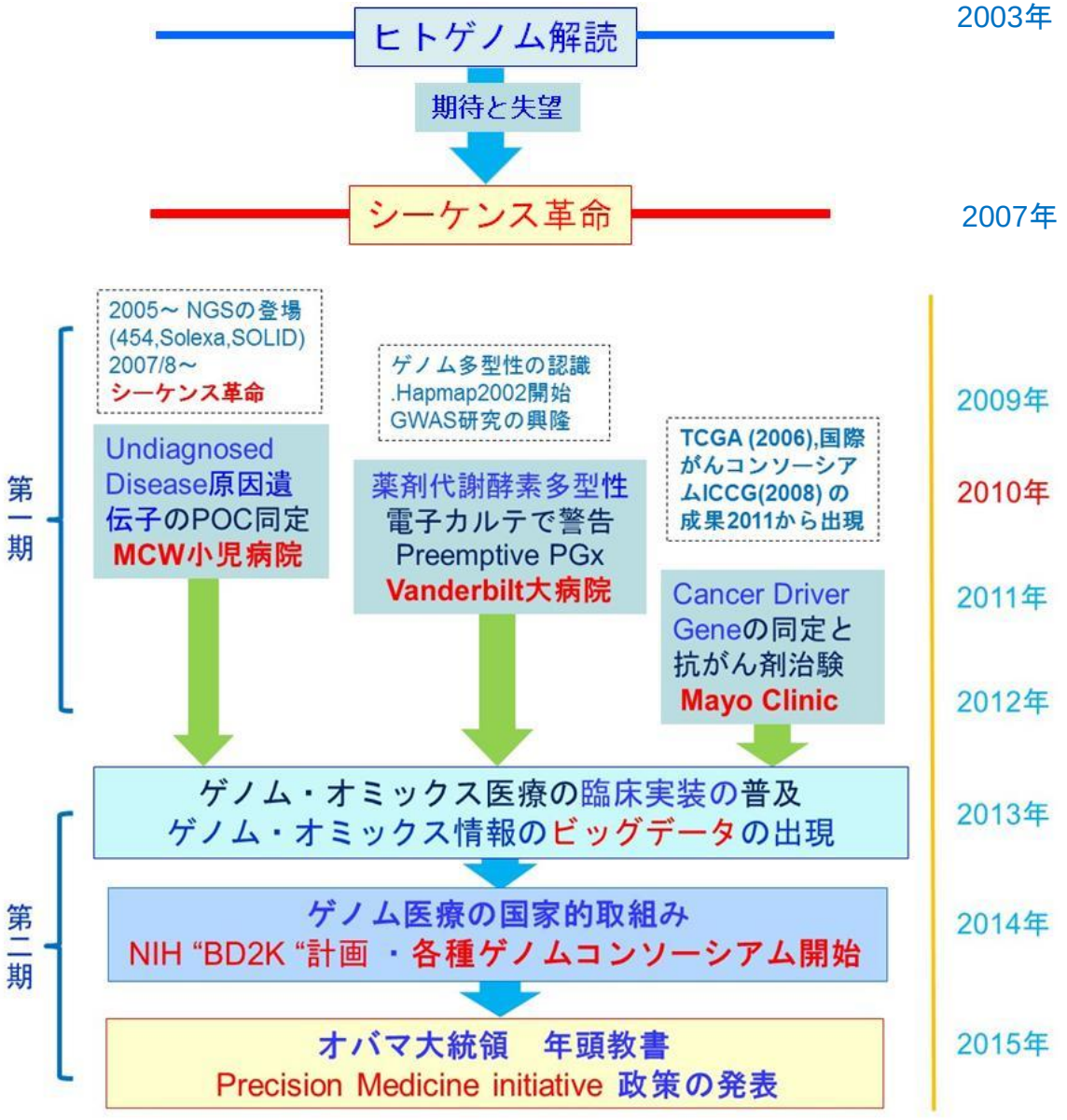
2005～ NGS 454 (LS,Roche)
2007/8～454, Solexa (Ilumina),
SOLiD (LT,TF)
シーケンス革命

	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



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



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

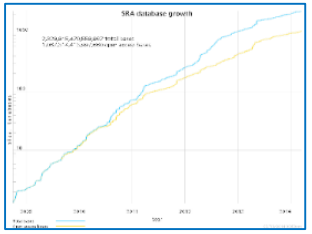
次世代シーケンシングの臨床普及(2010～)

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

ゲノムに対応する臨床表現型 *eMERGE*

electronic MEdical Record + GEnome (NIH grand)

- phase I (2007-2011) 臨床表現型情報のタイピング
  - 電子カルテを通して臨床phenotypingの形式
  - EMR：臨床phenotypingとbiorepositoryに基づくGWASが可能か（EMR-based GWAS). ELSI側面も検討
  - eMERGE-I: Mayo Clinic, Vanderbilt大学, Northwestern大学など５施設, **PheKBを構築**
- phase II (2011-2015) 臨床実装
  - 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
    - PGxの臨床応用に関する試行プロジェクト
    - 結果回付 Return of Result (RoR)
  - 4施設がeMERGE-IIより加わる
    - いくつかの小児病院とMount Sinai/Gesinger
- phase III : 2015より始まる
- **CSER consortium**と連携
  - “Clinical Sequencing Exploratory Research” コンソーシアム NHGRIにより予算化

個別化医療からPrecision Medicine

個人の遺伝素因・環境素因に合わせた (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（臨床バイオマーカ探索）

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
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を明快に強調。電子カルテの臨床表現型（Clinical Phenome)抽出も疾患発症後には不可欠。３つの成因の重視
（２）日常生理モニタリング情報の包摂
モバイルヘルス(mHealth)・ wearable sensorによる大量継続情報収集の重視
（３）ゲノムコホート・Biobankの重視
Precision Medicineを実現する基礎として、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

もう一つのゲノム医学の流れ
Biobankとゲノムコホート

- **バイオバンクの目的･機能の変化**
  - 従来は再生医療ための生体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
  - **ゲノム/オミックス個別化医療、創薬**の情報基盤
    - **疾患型BioBank**：全国的・全世界規模で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（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。**疾患ゲノムコホート**
  - **個別化予防**の情報基盤
    - **Population型BioBank**：「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、生涯を追跡するゲノム・コホート
- **欧米のBiobank**
  - **英国　UK biobank**
    - 50万人の健常者。40～69歳（2006-2010, 62Mポンド), 2011-16, 25Mポンド
    - 健診データ（血液・尿・唾液サンプル、生活情報）を集め、健康医療状況を追跡する。
  - **英国　Genomics England**,
    - 2013開始、2017年までに 10万人のゲノム 配列収集。
    - 最初の対象は稀少疾患（患者・家族）、がん患者、最初はEnglandのみ
  - **欧州　BBMRI**（Biobank/Biomole. Res. Infra.)
    - 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**
    - 100 万人のゲノムを集める

TMDU

大型プロジェクトによる知識データベースの大規模化

- ヒトゲノム解読計画以降急速に進展
  - Hapmapプロジェクト,1000genome,GWAS等
  - ゲノム変異
    - dbSNP, **Clinvar**, HGMD, GWAS catalog, **Matchmaker Exchange**
  - 遺伝子発現プロファイル
    - GEO, ArrayExpress, **c-Map, LINCS**
  - タンパク質
    - PDB, Swiss-Prot, HPRD, 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

TMDU

医療ビッグデータ

- 臨床ゲノム・オミックス情報
  - Clinical Sequenceのインパクト
  - 網羅的分子情報も含めた臨床症例データ
  - 個別化医療、Precision Medicine
- Biobank, GWAS,疾患レジストリ情報
  - ゲノム患者対照分析、ゲノムコホート
  - Population型は個別化予防
- 網羅的分子情報DBの大規模化
  - Clinvar,LINCS, HPRD

わが国でのゲノム医療の臨床実装

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

「ゲノム医学実現推進協議会」(中間報告）2015.7

AMED：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

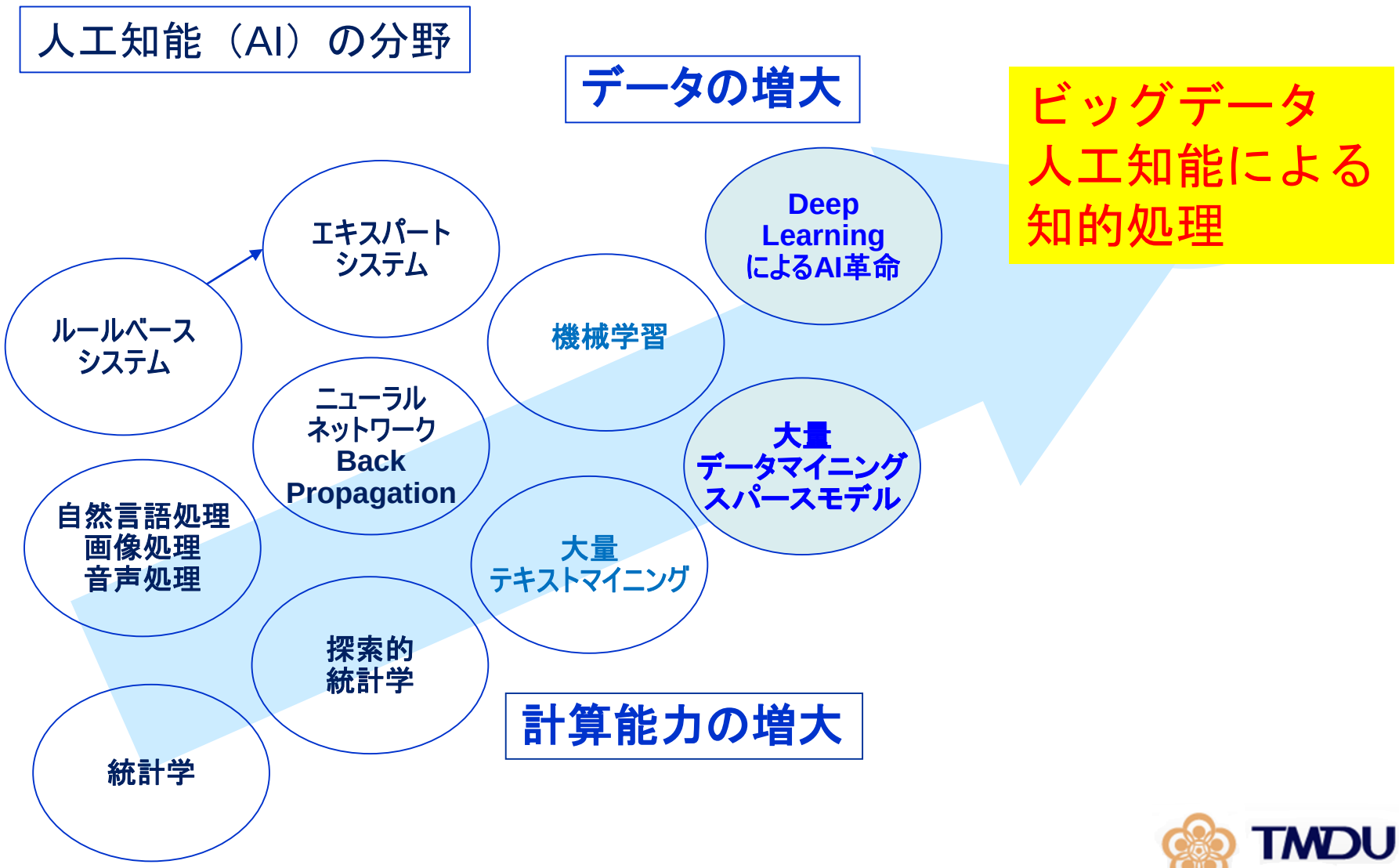
がんの網羅的分子診断と個別化治療

- 国立がん研究センター東病院
  - 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

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。
最後に述べる多因子疾患のゲノム医療に着手すべきである

医療ビッグデータを
人工知能を用いて知識抽出する

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

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

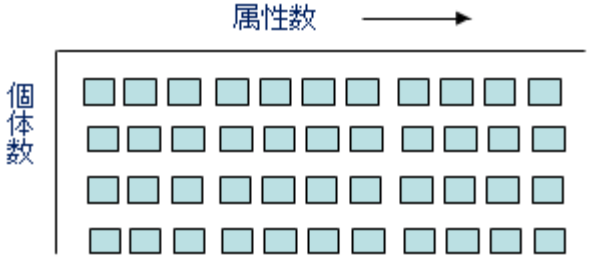
2000

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

「ビッグデータ」のData 原理

問題点　　属性値数(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**)

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

- 数理的知識処理：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AutoEncode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

数理的知識処理
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 $\{(y_i, \mathbf{x}_i); i = 1, \ldots, n\}$

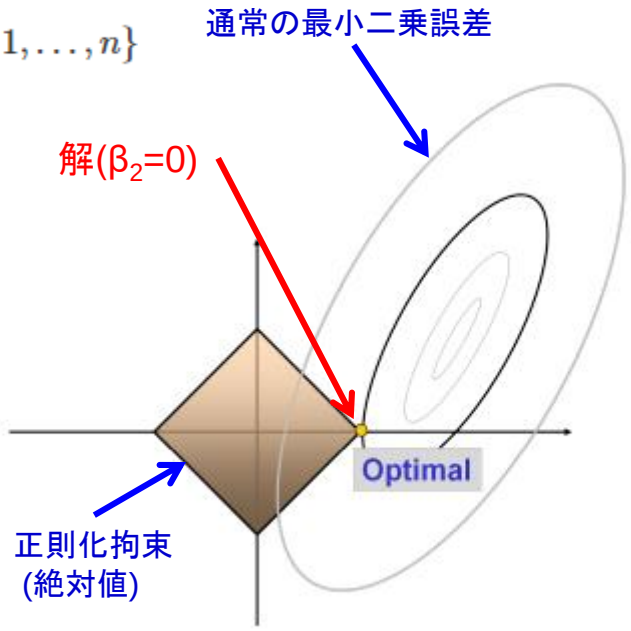
$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

Lasso(L_1型正則化重回帰分析)

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

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

通常の最小二乗　　正則化項 (絶対値)



寄与の低いβj 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

TMDU

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(主効果と相互作用）
  - Correlated SNPs (Ayers and Cordell, 2010)
  - More power while having a lower false-discovery rate (FDR) （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

TMDU

ビッグデータと機械学習

- **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
  - 新しい臨床治験へのガイドライン作成
  - 17万人のがん症例データベースを構築。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
- **IBM Watsonのがんセンターへの普及**
  - 質問・応答（QA)システム、知識探索、ライト・オントロジー
  - Memorial Sloan-Kettering Cancer Center（MSKCC）などと共同
  - Watsonを母体に**The Oncology Expert Adviser software** (OEA)開発
  - 他にNew York Genome Centerとglioblastoma (グリア芽細胞腫）知識生成
- **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
- Google X project, “Human Longevity Inc.”

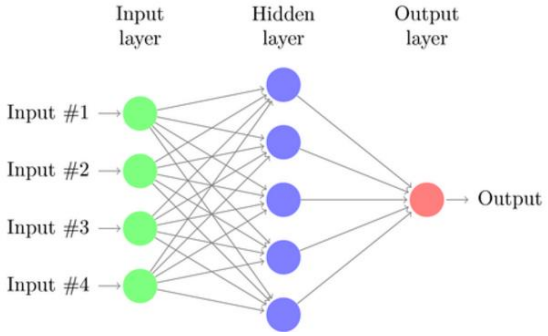
IBM Watson Learning Big Data

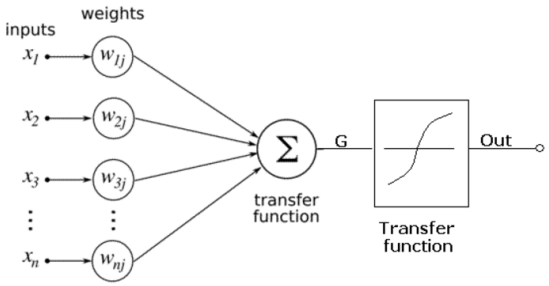
TMDU

Deep Learning 型人工知能の革命性

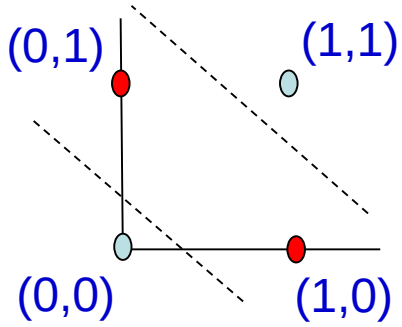
従来のニューロネットワーク

古典的Neural Network・パーセプトロン(1970年代）



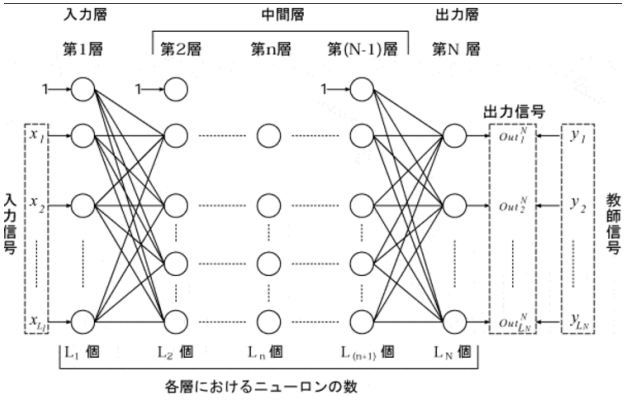


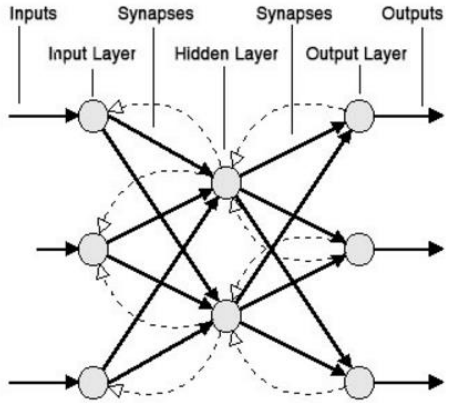
XOR



線形分離できない

多層Neural NetworkとBack projection (1980年代）





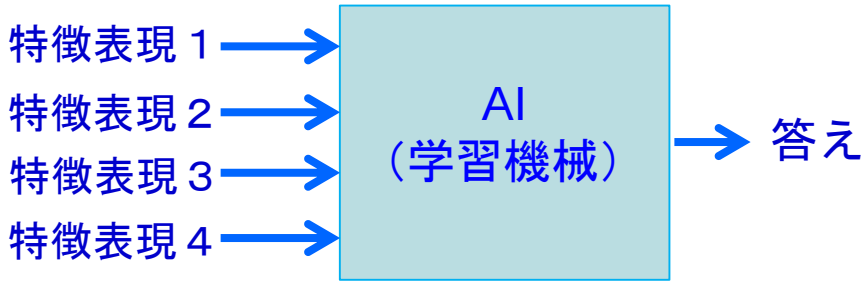
Back Propagation (1986　Rumelhart)
望ましい出力との誤差を教師信号として与える事により、次第に結合係数を変化させ、最終的に正しい出力が得られるようにする。
結合係数を変える事を学習と呼ぶ。
この学習方法には、最急降下法（勾配法）が使われる。出力層へ寄与の高いノードの重みが変更。

多層にわたる逆伝搬で修正感度減衰

TMDU

Deep Learning による人工知能革命

- 機械学習のこれまでの限界
  - 分類・判別する学習機械（システム）
  - 対象の特徴表現ベクトルを与えて分類
  - 与え方に関して細かな技法にとらわれる
- 「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
  - 対象の表現(画像等)と概念を結合できない



TMDU

Deep Learningの革命性

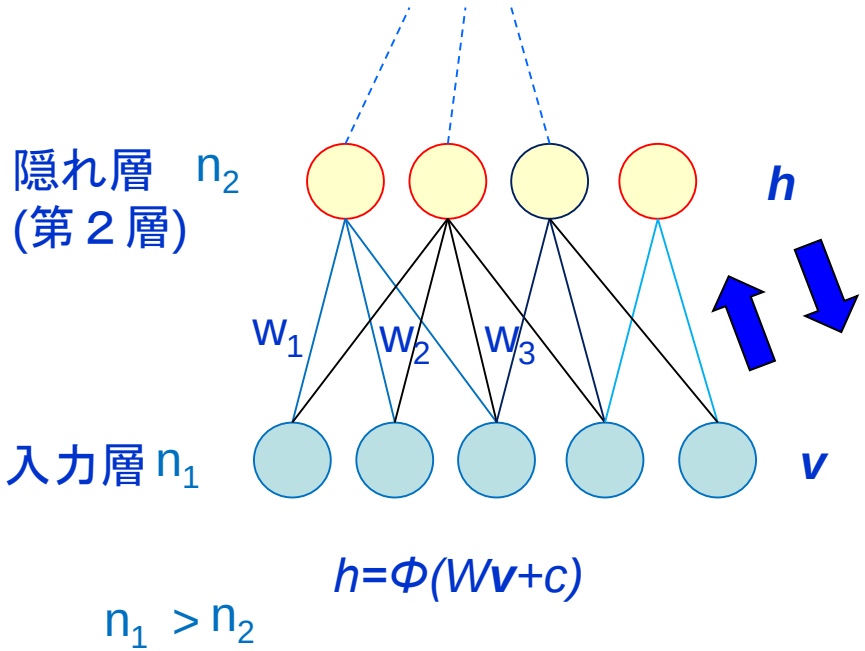
- DLは、まずは対象の固有の構造を記述する特徴表現や対象の高次特徴量を自ら学ぶ「教師なし学習」を行う
- 「内在的な特徴表現の学習」を自動的に行う
  - 自己符号化（Autoencoder）
  - 制限ボルツマンマシン
- 最終層で、人間の概念との相同をとるため「教師あり学習」

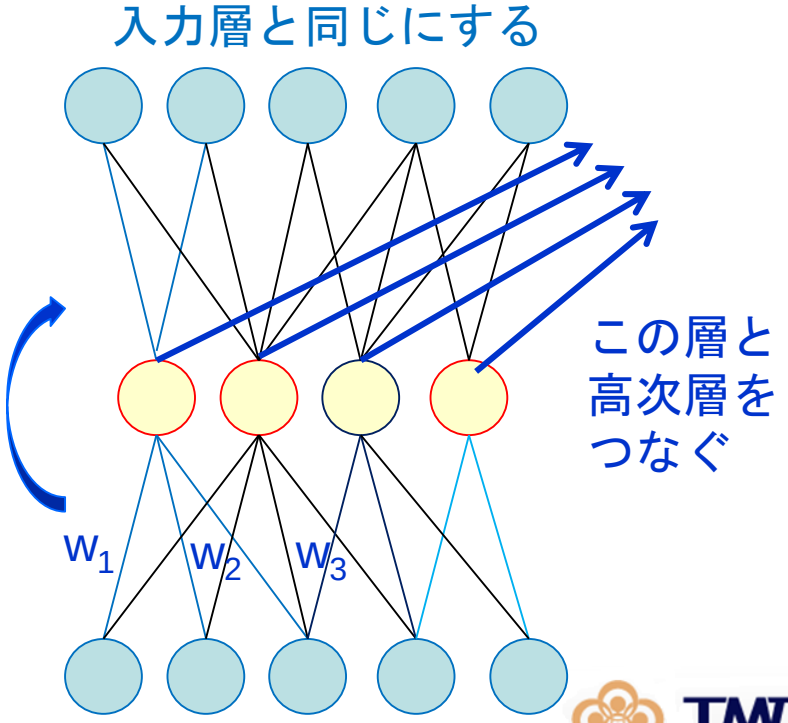
DLの革命点 Autoencode 1

- 対象に固有な内在的特徴を学ぶ自己符号化の原理
- 格段ごとに入力を少ない中間層を介して復元できるかを行なう
- 次元を圧縮されて可及的に復元する

→できるだけ復元に効果的な特徴量を探索する

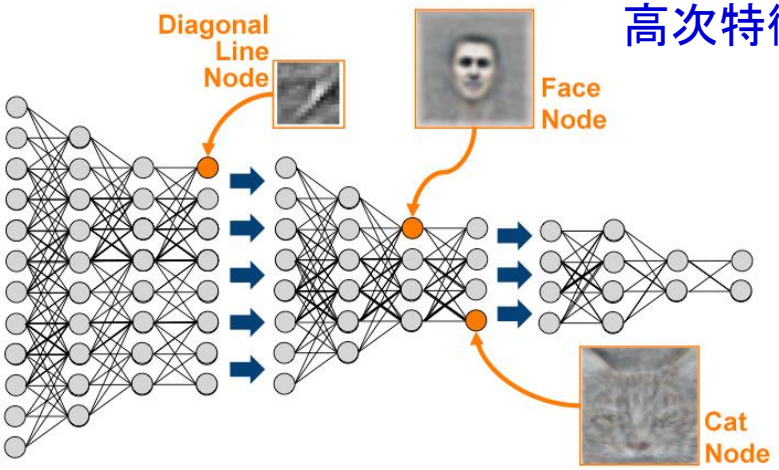
→内在的な特徴量を見出す





DLの革命点 Autoencode２

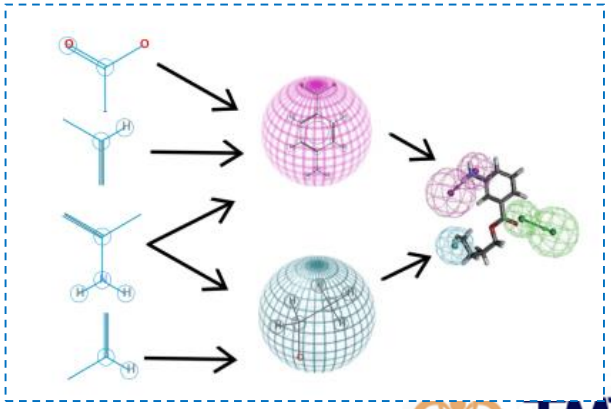
- 各層ごとに自己符号化を行うので何層でもネットを組める→Deep Learning
- 第一層で学習した特徴量を使って、つぎの階層を作るので高次の特徴量が作られる
- 特徴的表現と概念を結びつけるため「教師あり学習」が最後に必要である
- 自動特徴抽出によってこれまでの学習手法の限界を克服した（構造的理解）



高次特徴

部品的特徴

Pharmacophoreの摘出

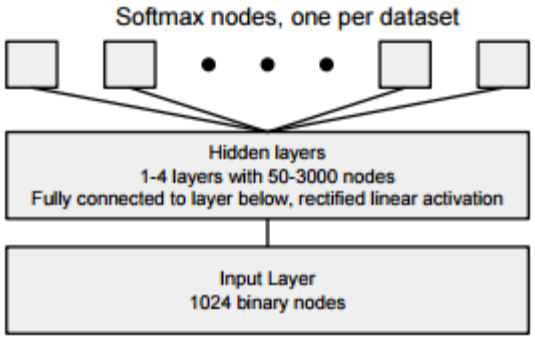


Deep learning ：創薬からの注目

- 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"



Massively Multitask Networks

Artificial Intelligenceと創薬

- 標的分子選択と妥当性検証
  - 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-nearest nb, 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

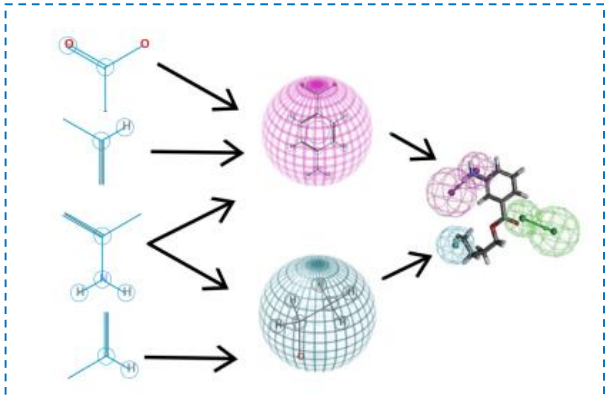


Figure Hierarchical nature of fingerprint features: by combining the ECFP features we can build reactive centers. By pooling specific reactive centers together we obtain a pharmacophore that encodes a specific pharmacological effect.

ECFP(chemical substructure: Enhanced cyan fluorescent)

DL型NNへの期待と限界

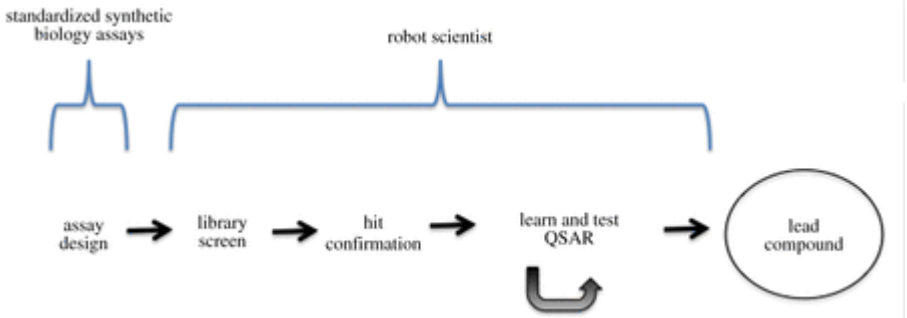
- まだ、医療・創薬の応用は緒に着いたばかりで、応用成功例は少ない
  - 画像分類・解釈と文章理解が優れているので。遺伝子発現プロファイル解析や病態推移の理解への応用が期待される
    - 例：ヒトmicrobiomeの分類・階層的表現を得た
    - ６つのがんで遺伝子発現をmiRNAとともに分類した。
    - 異なったMicroarrayを含むがん発現を分類の特徴表現を導き分類した。
    - Convolution NN を使用して画像としての遺伝子発現を分類した。
    - 遺伝子発現プロファイルの自動アノテーション
- DL型NNの限界点
  - 特徴表現を自己学習するが基本的にはBlack Box
  - 大量のデータを必要とする
  - DL型NNには数種類があり、選択問題が残る
  - 計算時間が長く、コストが大きい

TMDU

そのほかのAI創薬の話題

- Berg社のAI創薬
  - 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
  - 厳密にはAIではなくシステム薬理学・Bayes法による創薬。AI創薬と呼んでいる
- 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

ビッグデータDBを利用したDR/創薬

計算論的創薬
computational drug discovery

これまでの計算論的創薬
- 分子(構造)中心（molecular-structure oriented）

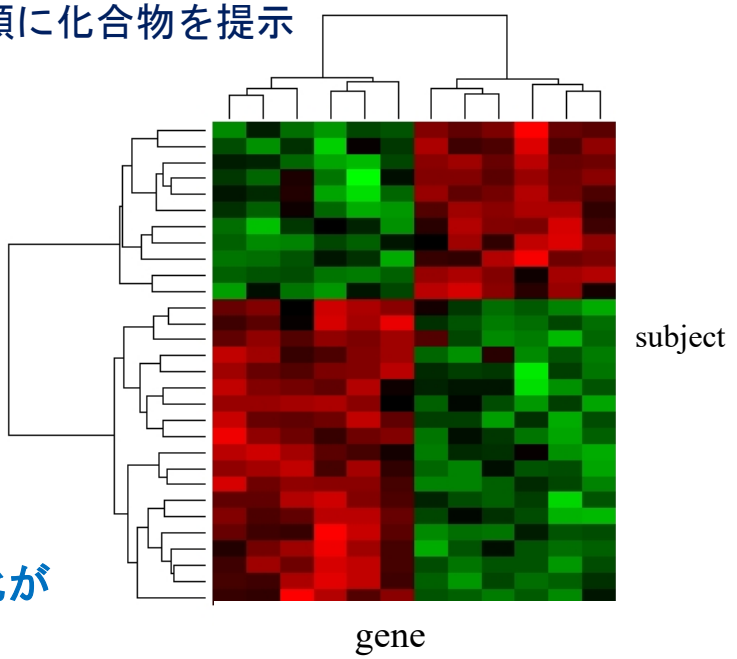
分子構造解析・分子設計（*in silico* drug design)
- Structure-based rational drug design
- 標的分子の分子構造解析
  - 薬剤（リガンド）との結合構造の分子構造解析
- リガンドの分子設計ー分子力学・量子化学
- リード化合物の構造最適化
- 定量的構造活性相関(QSAR)の利用

新しい計算論的創薬のアプローチ(*mol. profile* drug design)
- オミックス創薬・システム創薬
- 網羅的分子プロファイル・分子ネットワーク変化中心
- **薬剤ー標的分子の結合**を取り巻く**genom-wideな分子環境**
- 標的分子より「疾患システム」という対象の把握
- 化合物ー疾患の反応関係の「ビッグデータ」利用

TMDU

オミックス創薬の原理

- 薬剤特異的遺伝子発現（Drug-induced SDE)
  - CMAP：Connectivity Map
    - 薬剤投与による遺伝子発現プロファイルの変化
    - 米国 Broad Institute,1309化合物, MCF7,PC5など5 がんセルライン, 7000 遺伝子発現プロファイル
    - Signature (遺伝子発現刻印：差異的発現遺伝子の代表的集合）
      Signature of Differential gene Expression
    - DB利用：SDEをquery, 順位尺度で類似性の高い順に化合物を提示
    - 最近LINCS　100万サンプルへ拡張
- 疾病特異的遺伝子発現
  (Disease-associated SDE)
  - GEO（gene expression omnibus),
    - 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
    - ArrayExpressもEBIが作成、サンプル数同程度



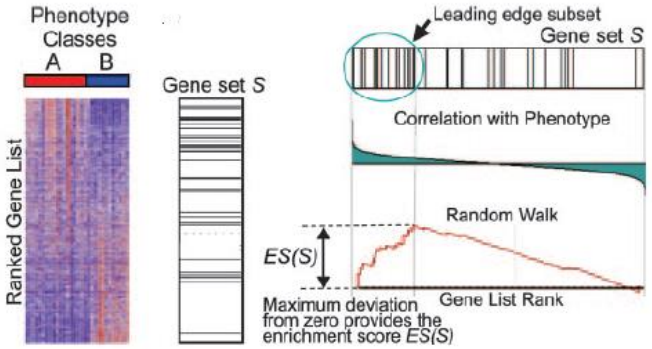
本来は、分子ネットワークの疾病/薬剤特異的変化が基本（第３世代網羅的医学）。

遺伝子発現プロファイル変化
≈　分子ネットワーク変化

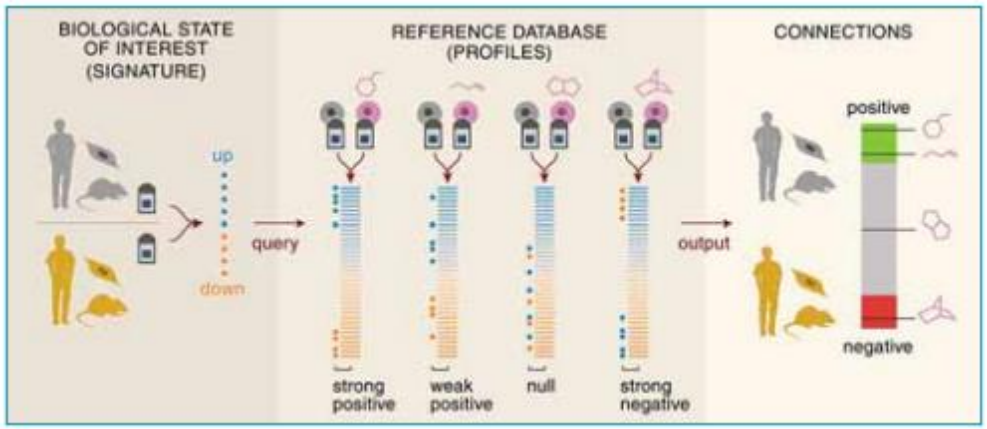
TMDU

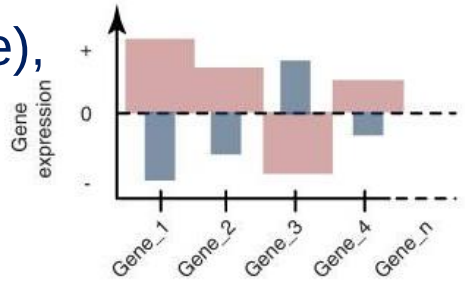
遺伝子発現プロファイルによる有効性予測

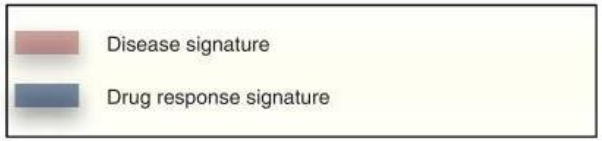
- 遺伝子発現シグネチャア逆位法 (signature reversion)
  - 薬剤特異的遺伝子発現シグネチャア
  - 疾患特異的遺伝子発現シグネチャア
  - 有効性予測：両者が負に相関する
  - Non-parametric な相関尺度で評価
    - Gene Set Enrichment Analysis（GSEA)：ES score
    - 対照と比較して順位づけられた遺伝子リストの上位に密集しているかの尺度
  - 例：炎症性腸疾患IBDに 抗痙攣剤(topiramate), 骨格筋委縮にウルソール酸



GSEA



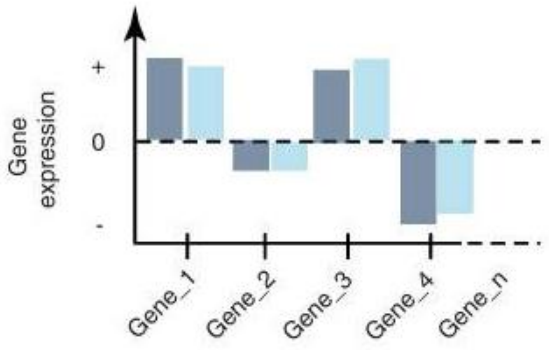




TMDU

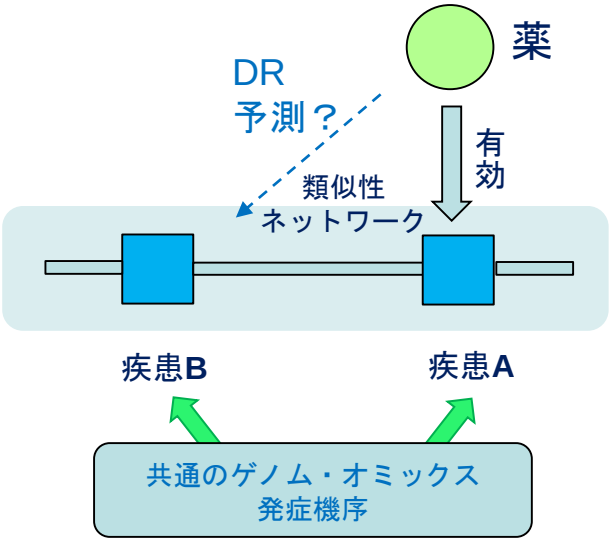
遺伝子発現プロファイルによる毒性予測

- 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類似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



疾患ネットワークに基づいた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による疾患分類
  - 発症機構類似性を基準に疾患ネットワーク
  - ゲノム・オミックス医学の疾病概念が基礎



TMDU

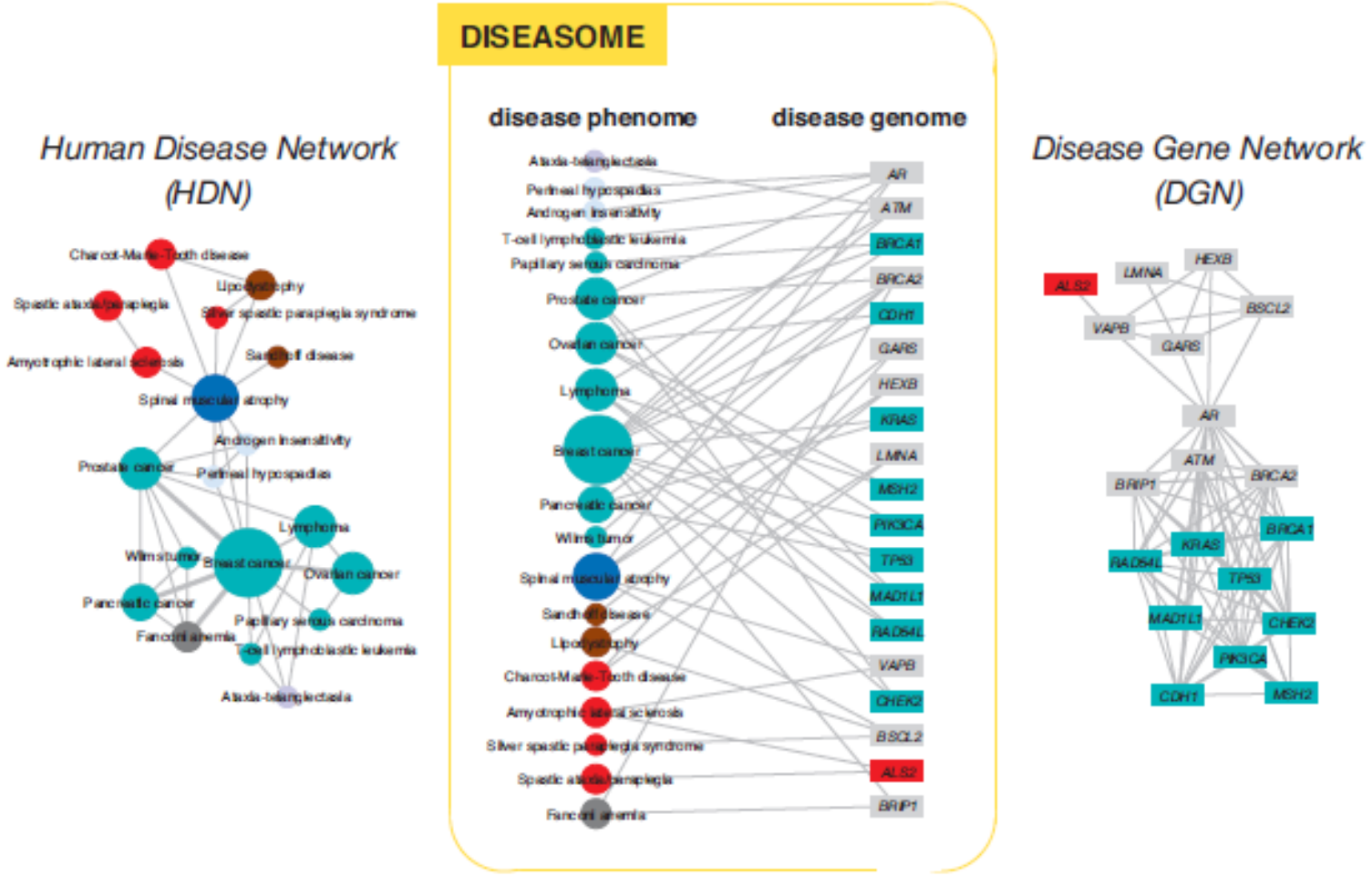
第１世代型
Diseasomeと疾患遺伝子

- 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を持つ

TMDU

疾患ネットワーク Diseasome

(Goh,Barabasi et al.)



１つ以上の疾患関連遺伝子を共有する疾患

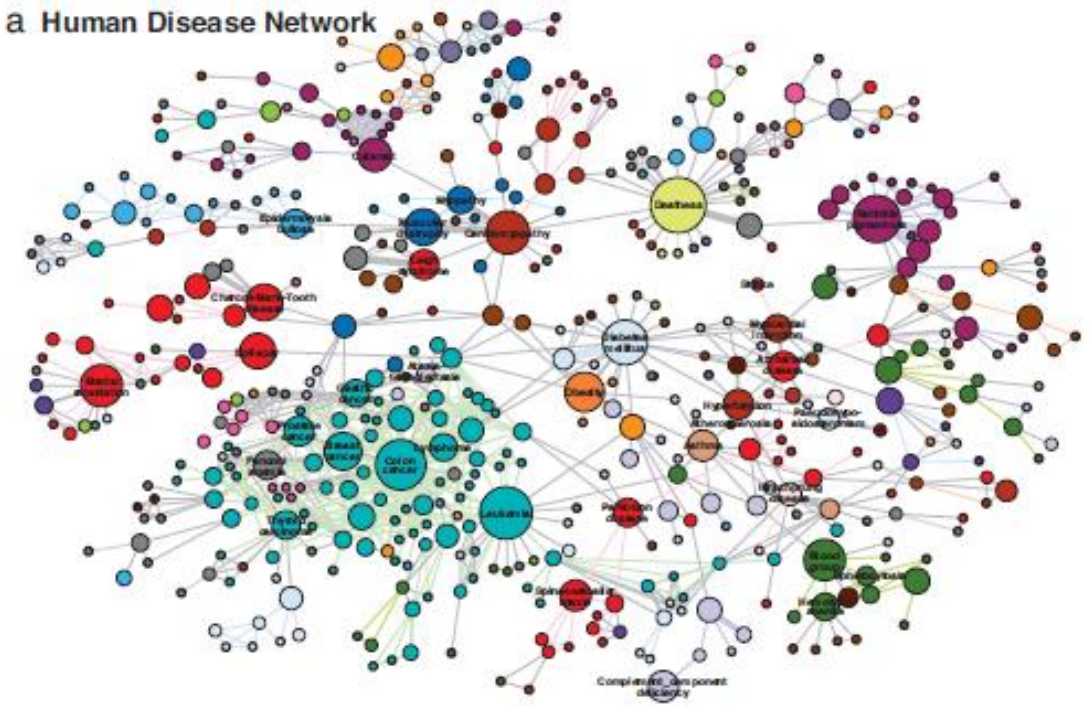
１つ以上の疾患を共有する疾患関連遺伝子

Kwang-Il Goh*, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-Laszlo Barabasi The human disease network　PNAS2007

TMDU

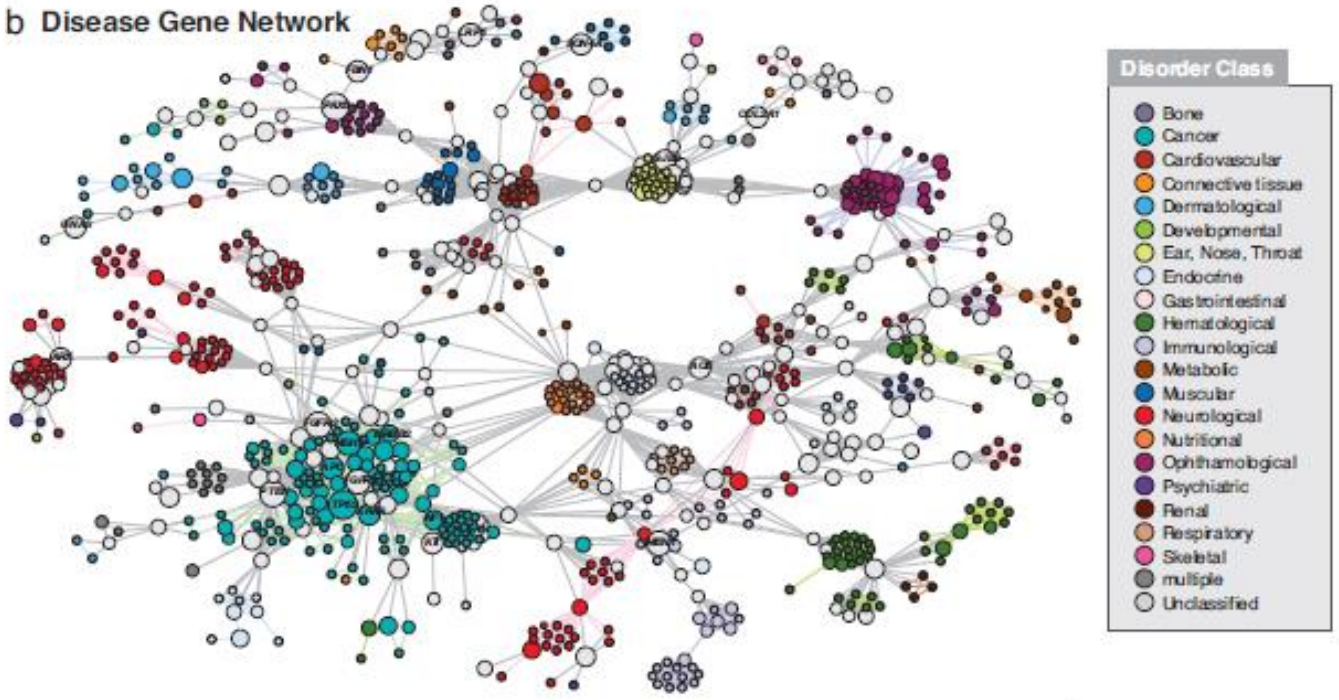
疾患
ネットワーク
(HDN)

Nodeの径
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)

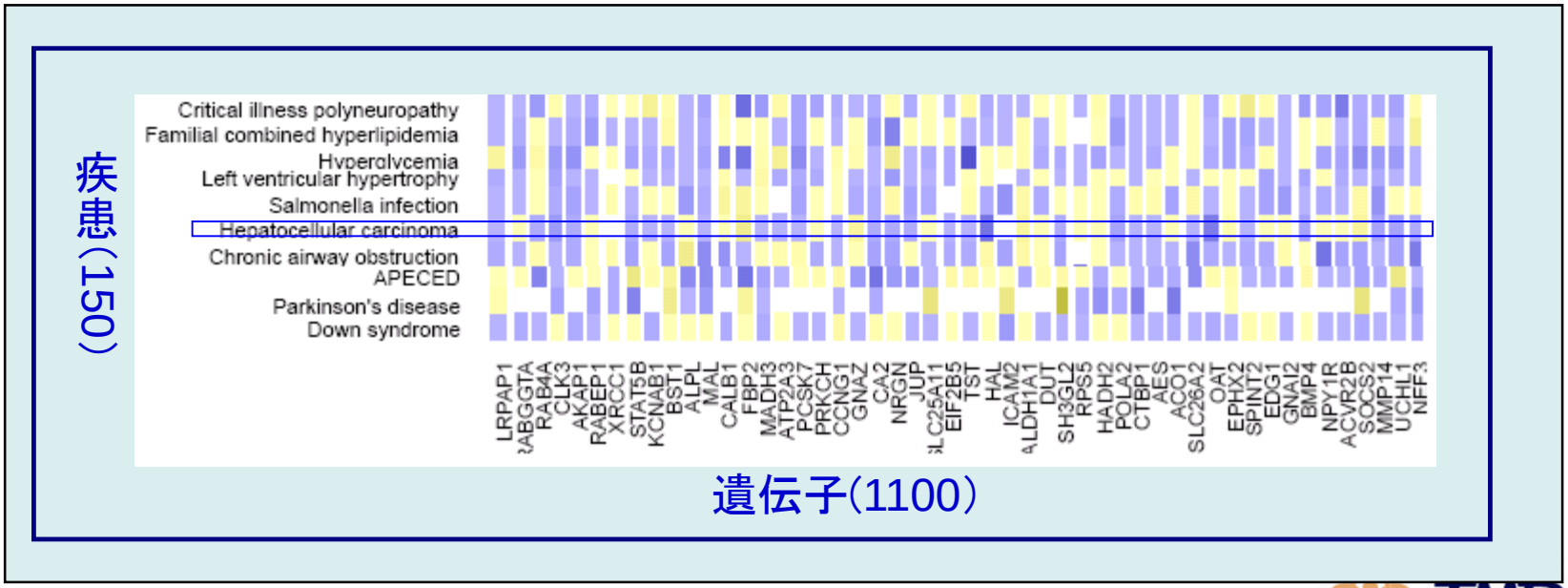
Nodeの径
その遺伝子を原因にしている疾患の数に比例
2つ以上の疾患に関与していると明灰色の遺伝子ノード



第２世代型

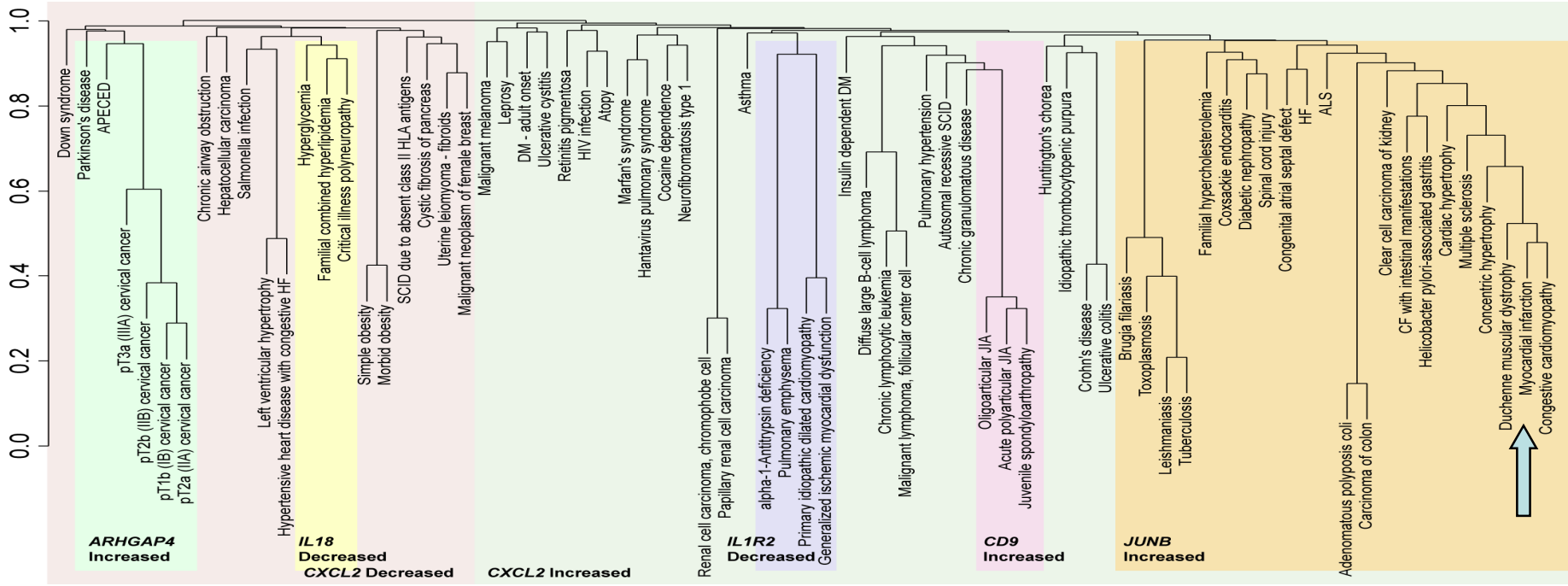
GENOMED (A.Butte et al)

- 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

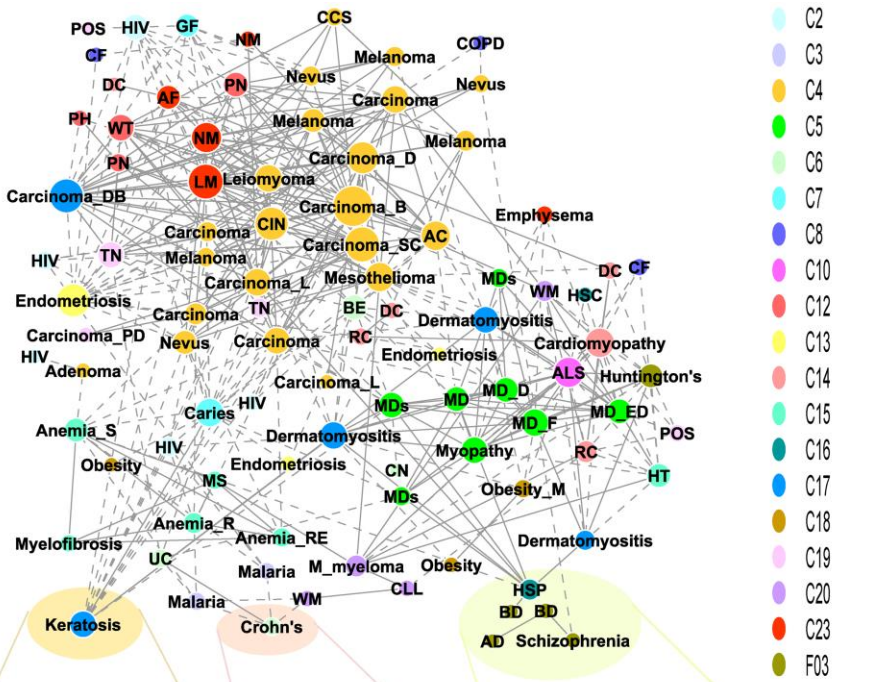


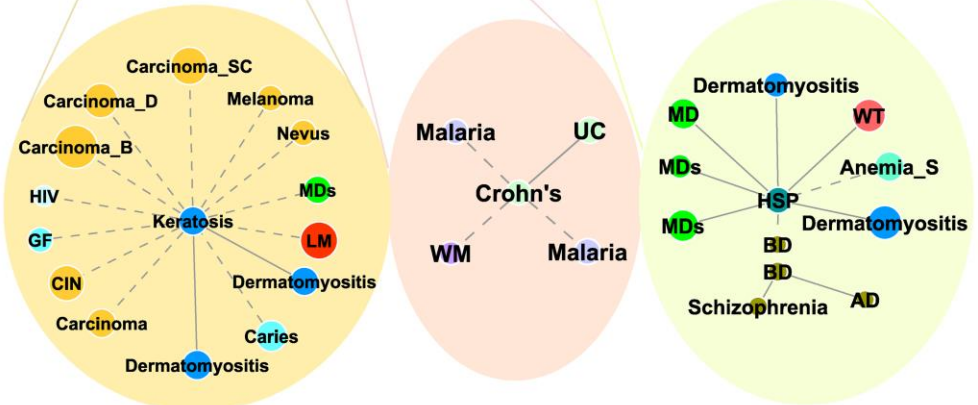
Transcriptional Profiling による疾患ネットワーク

Hu, Agarwal 遺伝子発現プロファイルとGSEA関連尺度によるリンク

疾患（disease-disease）645 nodes
疾患- 薬（disease-drug）5008 pairs

Solar keratosis 日光性角化症
⇒ cancer(squamous)
Crohn’s disease
⇒　マラリア
Hereditary Spastic Paraplegia
（遺伝性痙攣性対麻痺）
⇒bipolar双極性うつ病





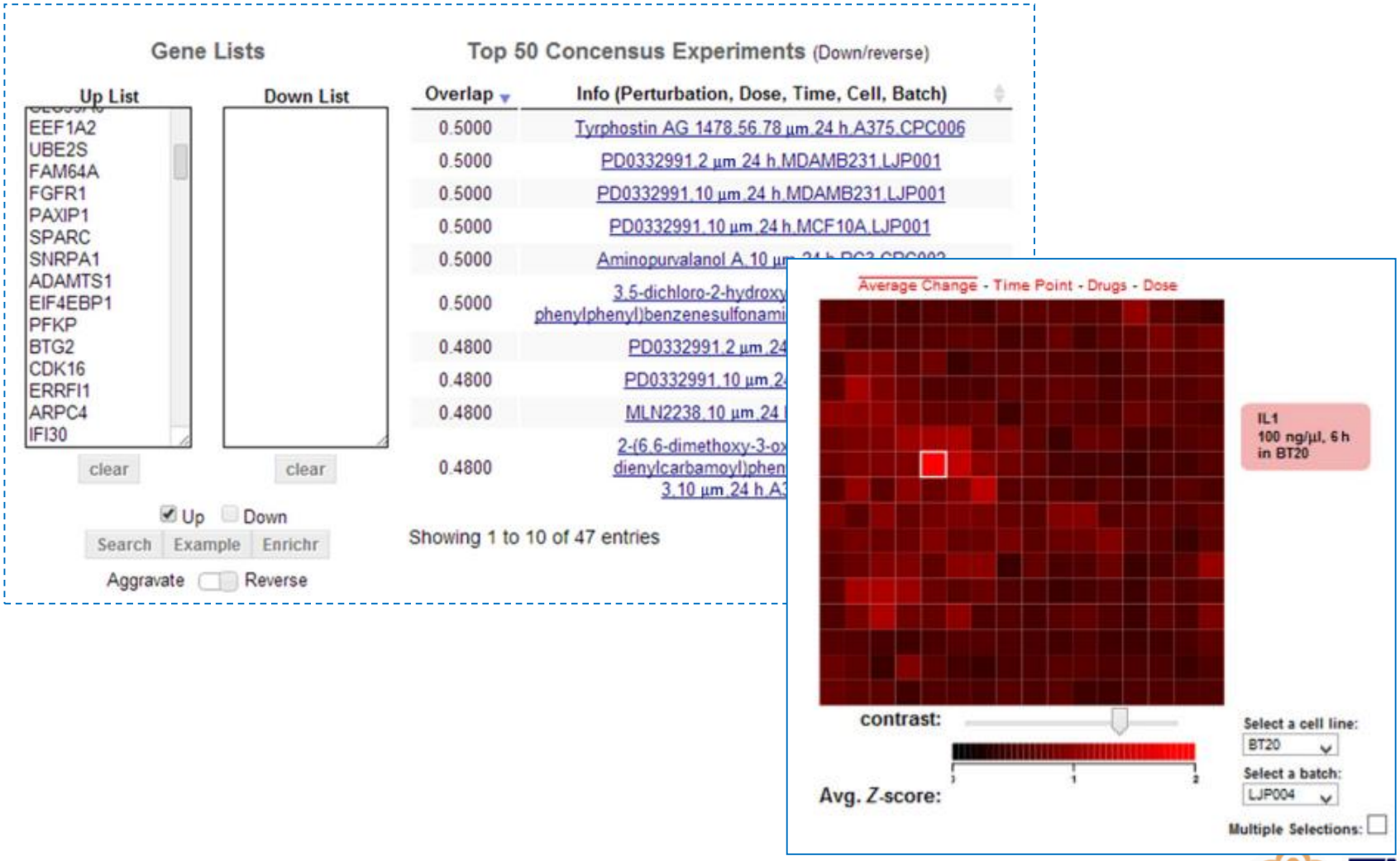
カラーはMeSH
同一カテゴリー

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）を与えて、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
    - 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(RNAj, 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

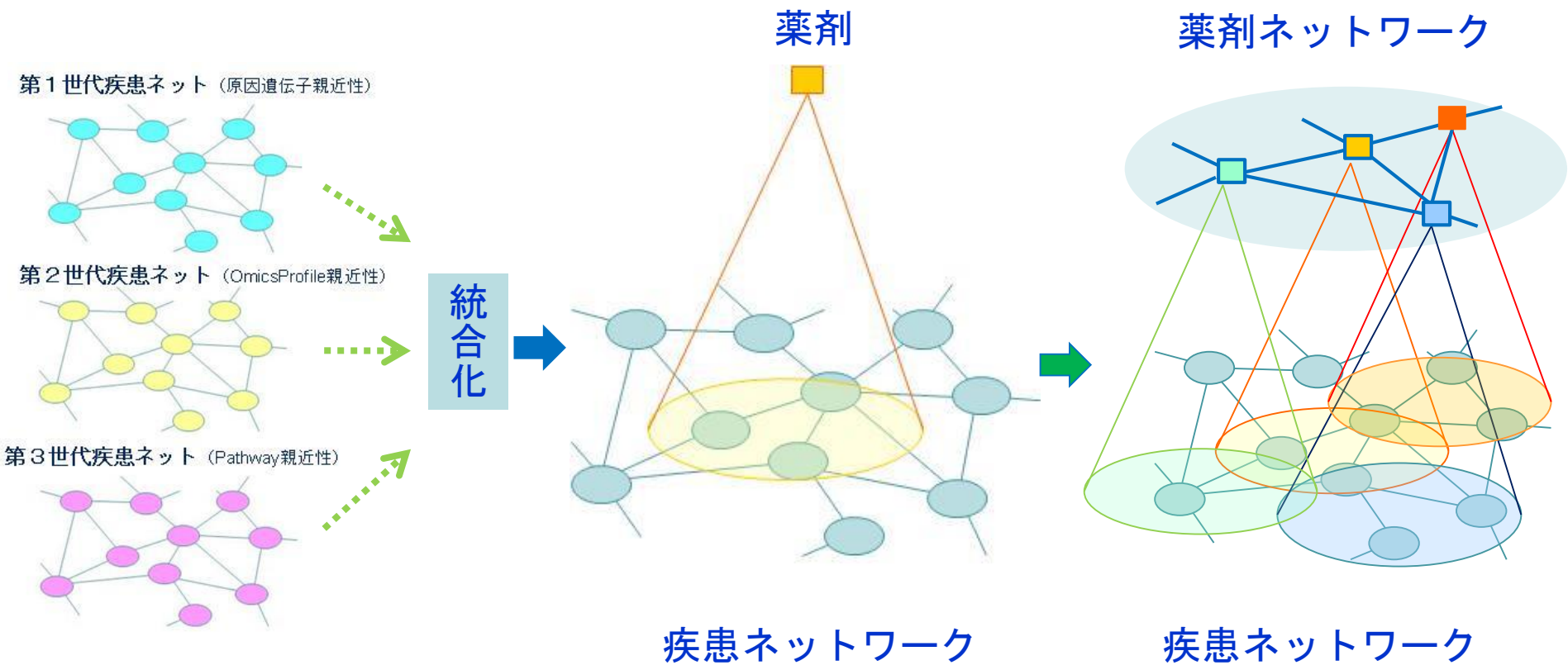
LINCSの問合せ画面

--- LINCS Canvas Browser ---



疾患ネットワークとDrug Projection Map

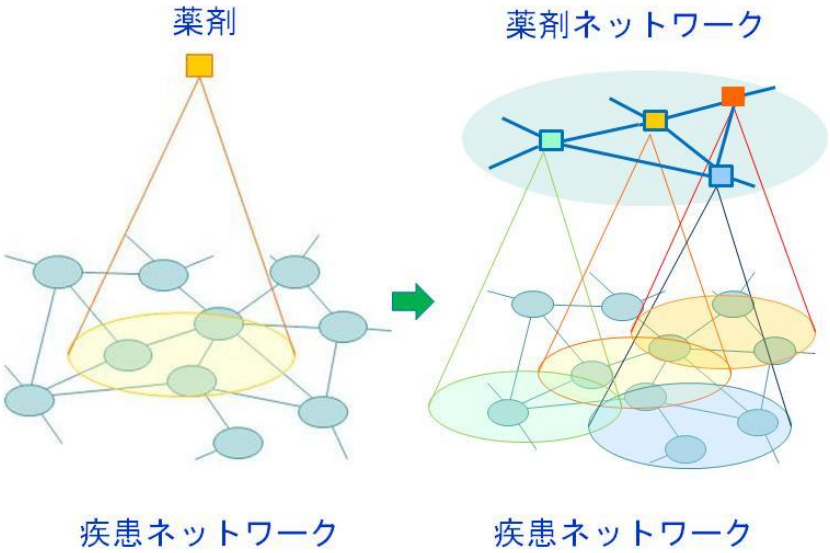
DR informaticsの構築

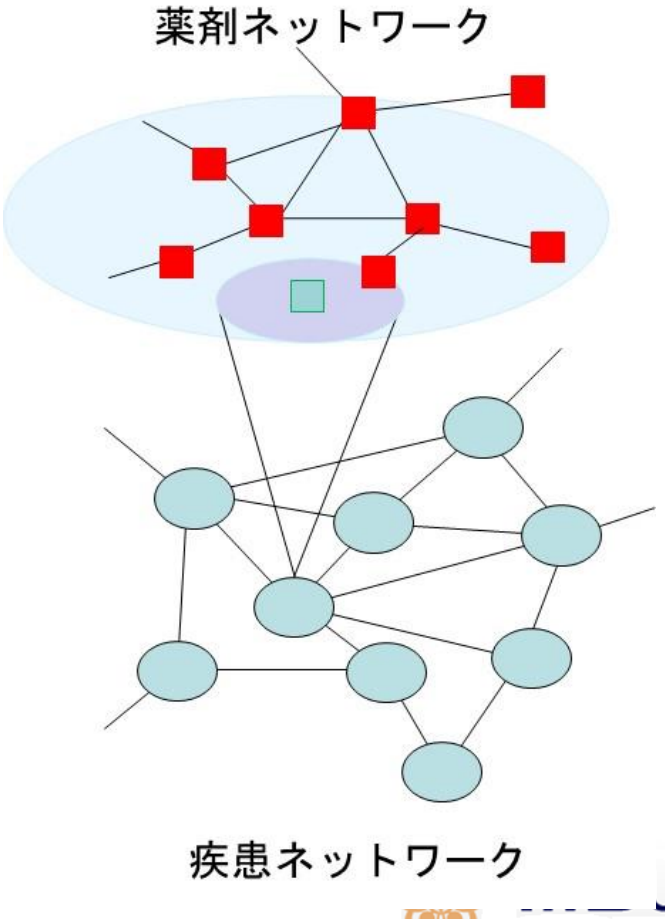


DRの方法論から創薬方法論へ

- 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)
  - 疾患ネットワークがの確立が重要
  - 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
  - Dual Topology-based Drug Disccovery

疾病から薬剤ネットワークへの逆投影
Double Topolpgy 双投影　創薬方法論





Real-World- (Big) Dataを用いた創薬/DR戦略

―RCT, EBMからの呪縛の解放―

TMDU

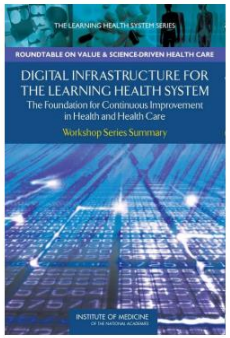
「学習する医療システム」
Learning Health System

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

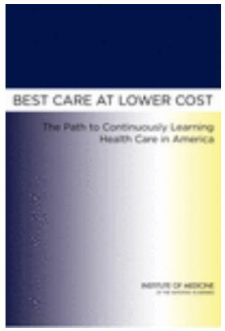
- IOM “Clinical Data as a Basic Staple of Health Learning”
- 医療システムのデジタル化（IT化）は必然の傾向である
- 「ルーチンの医療活動から集められたデータ（形式的臨床研究と違って）がLHSを支える鍵である」
- データを共有することによって学習して医療システムを改善
- RCTは「黄金基準」であるが、通常の医療システムの外で実施されている。医療が実際対象とする患者集団を代表しているのか。
- RCTは時間が掛かり費用もかかる
- 有効な知識の蓄積の速度が加速する

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

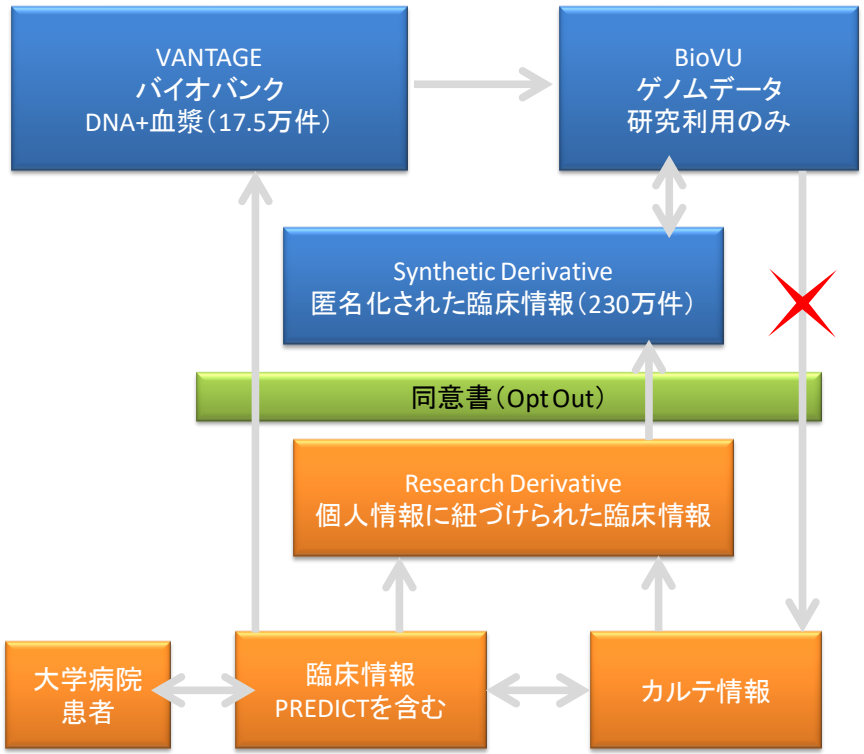
Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care



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「相転移」
  - 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準拠の創薬・治験

- 疾患レジストリ―/疾患型バイオバンク準拠型ランダマイズ治験
  - スウェーデンのTASTE(ST-Elevation MI in Scand.)
  - Registry-based randomized clinical trial
  - 疾患レジストリ―の登録患者から治験に適した治験対象者を選び
  - 選んだ集団で治験薬・対照薬をランダマイズして割付ける(50$)
  - 治験のエンドポイントは疾患レジストリーの追跡で観測される
  - 観測研究であるPopulation 型コホートでは困難か

ゲノム・オミックス医療の次世代の展開

第２世代のゲノム医療へ

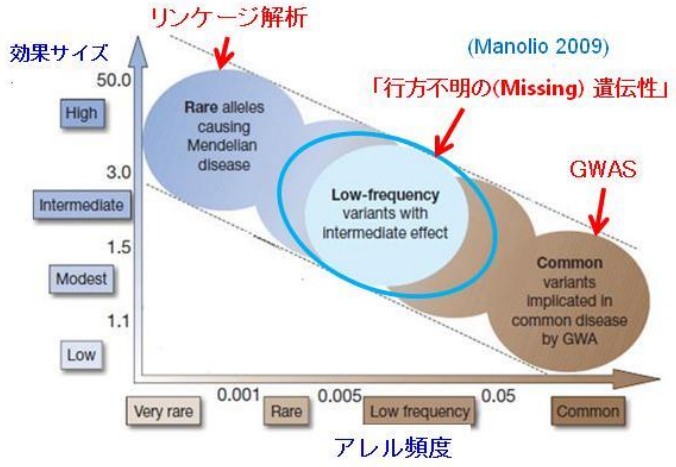
成功した臨床実装

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



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

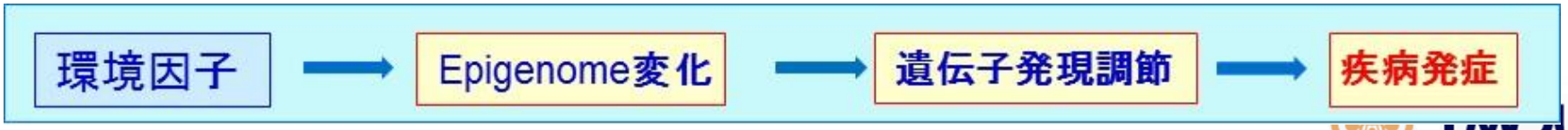
GWASでSNPの相対リスクが低い(1.1～1.3)理由： **GxE組合せ特異的効果を環境要因の全てに亘って平均しているからである**

発達プログラム説 DOHaD

Developmental Origin of Health and Disease)

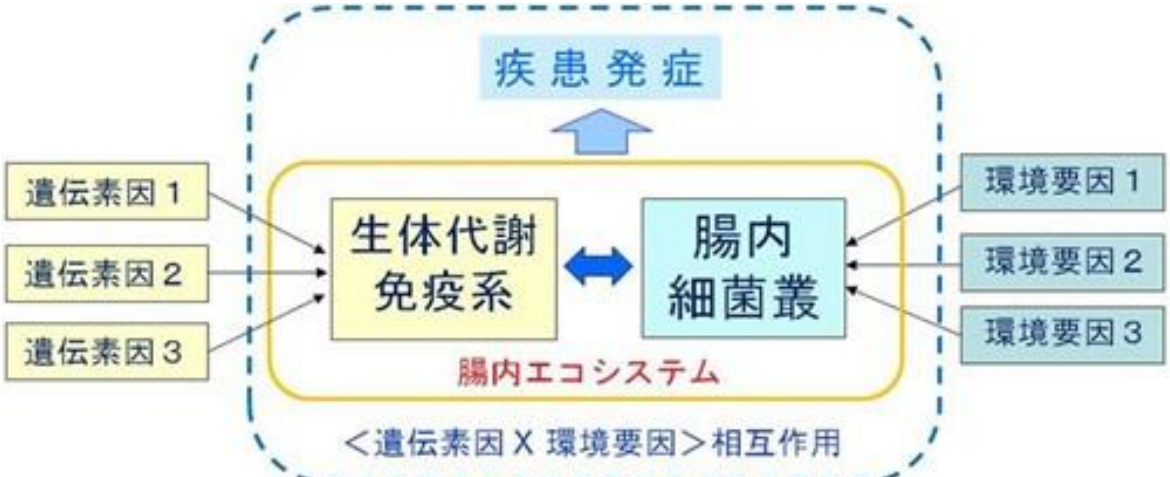
オランダ
飢饉 (1944)

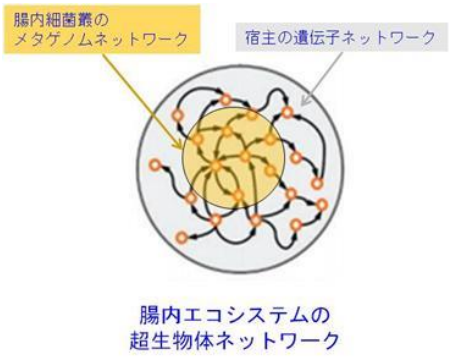
- オランダ飢饉
  - 第2次大戦末期、ナチスの封鎖、約半年間酷い飢饉
  - 飢饉の期間に胎児、戦後30年
  - 成人期:肥満,糖尿病,心筋梗塞,統合失調
- Baker仮説：英国心筋梗塞増加
- エピジェネティク機構
  - 過度な低栄養：肝臓のPPARα/γ（倹約遺伝子）メチル化低下・遺伝子発現がオン
  - エピジェネティック変化は**可変**：短期的変化、**長期的「記憶」**次の世代も



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

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
  - エピゲノム機構、メタゲノム機構
  - ＜疾患スプラ・ゲノム機構＞

TMDU

ご清聴ありがとうございました