

「人工知能（AI）を用いてビッグデータから創薬を目指す」

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
（情報システム構築担当 特別補佐）

田中 博

医療ビッグデータ時代の到来

- 次世代シーケンサを始めとするBio-technologyの急速な発展によるゲノム・オミックス情報の膨大な蓄積（ゲノム・オミックス医療）
⇒医療ビッグデータ
- Mobile Healthにおける連続的生体モニタリング
⇒生体モニタ・ビッグデータ
- 生物測定法の高速化による基礎生命情報（DB）の大規模化
⇒創薬ビッグデータ

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

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**
 - **システム分子医学・Precision Medicine**
- **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
- **生命情報の大容量化 (High-Throughput Bioassay)**
 - ビッグデータ創薬

旧来のタイプの
医療データの
大容量化

新しいタイプの
医療ビッグデータ

医療の「ビッグデータ革命」

～何が新しいのか～

<目的もデータ特性も従来型と違う>

従来の生命医療情報の「ビッグデータ」

Big “Small Data” ($n \gg p$)

医療情報・疫学調査 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

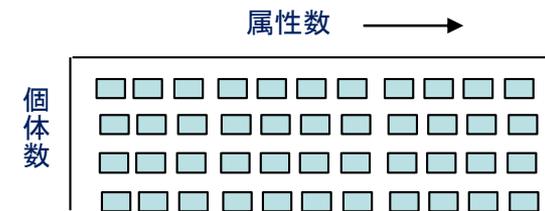
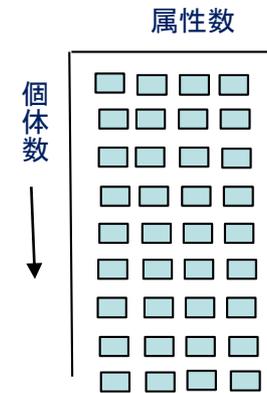
網羅的分子情報、mHealth、大規模
生体ビッグデータ

Small “Big Data” ($p \gg n$)

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
とくに多変量解析: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」の多様性を抽出する



新しいデータ科学の必要性

「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

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

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

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

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

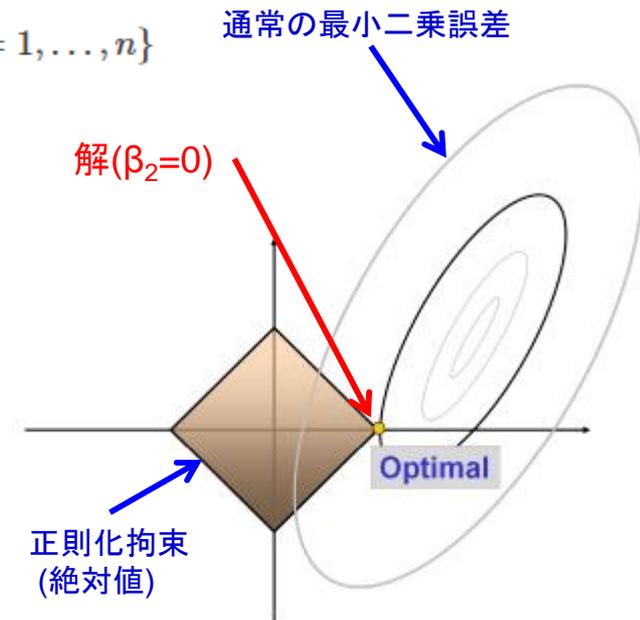
Lasso(L_1 型正則化重回帰分析)

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

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

通常の最小二乗

正則化項 (絶対値)



寄与の低い β_j は0になる \Rightarrow 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)

- 遺伝子発現プロフィールへの応用

- Biomarker (差別的発現遺伝子) 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

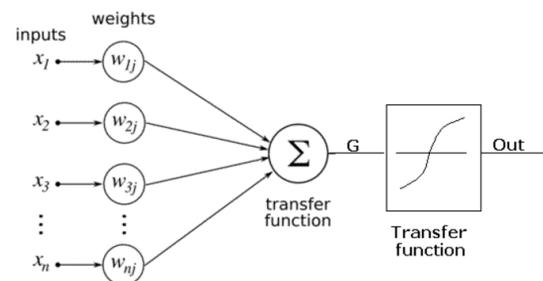
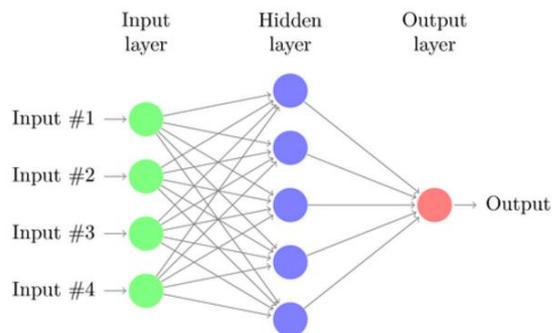
- 次を最小化

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

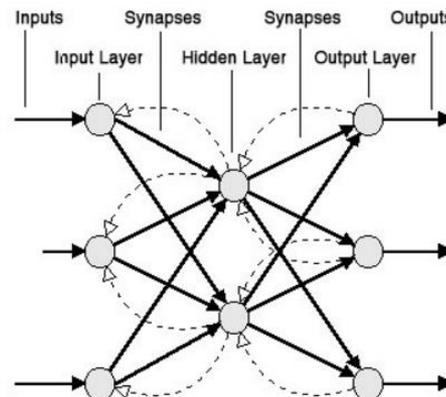
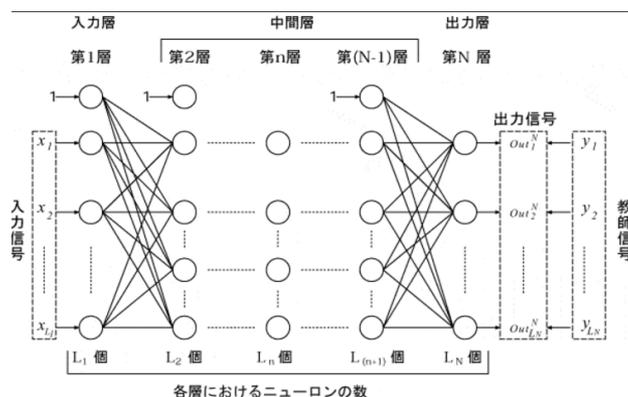
- 判別分析でも正則化により次元縮約

Deep Learningによる基軸成分の抽出

古典的Neural Network(1970年代)



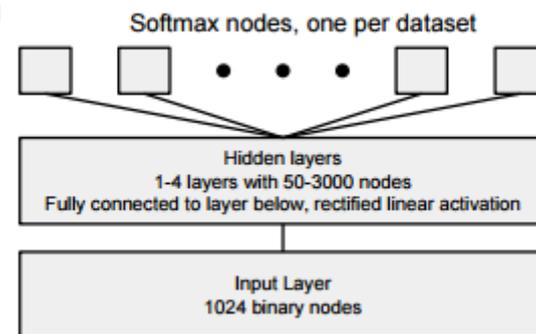
多層Neural NetworkとBack projection (1980年代)



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の変更。

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** (popular data science contest site), recently hosted the **Merck Molecular Activity Challenge** (2012).
 - to create a predictive model for the biological activity of different molecules from a collection of 15 data sets.
 - Winning model used deep learning in developing their model
- **Google in collaboration with Stanford** (2015)
 - recently released a study in collaboration with Pande Lab at Stanford University
 - demonstrating the effectiveness of deep learning as a virtual drug screening tool.
 - "Massively Multitask Networks for Drug Discovery"

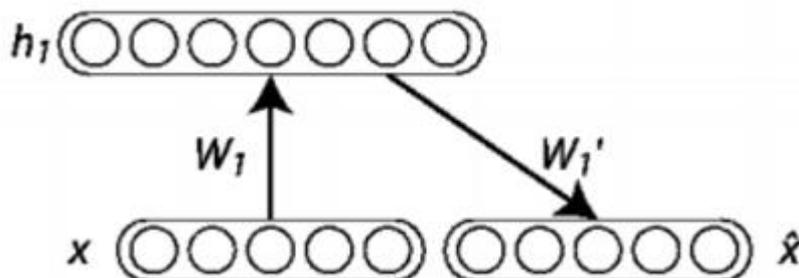


Massively Multitask Networks

Deep learning どこが新しいか

Greedy Layer-wise Training (2006, Hinton)の提案

- (1) 最初に「教師無しデータ」を利用して、各レイヤーのパラメータを一層ずつ調整。
- (2) 最初の層を学習する場合は入力を変換し逆変換をかけ元の入力と比較し一致するようにパラメータを更新。
xを入力、1層目の変換関数をf, その逆変換の関数をg ; $g(f(x))$ を計算し、**xと $g(f(x))$** が一致するようにパラメータを学習する。
パラメータが十分な数があれば元の入力をそのまま返すような関数が学習される(恒等写像)が、パラメータに**正則化**をかけて学習することにより、少ない表現力で入力の情報を表現するようにパラメータが調整される。入力情報を最も良く表現できるような関数が抽出。
基本的な特徴情報が取得される。
- (3) **autoencoder** : 変換をかけて元の信号に戻せるように学習する方法
- (4) 第一層の結合係数は**固定して**次の階層の学習に入る
- (5) 最後の層が学習できれば、最後は逆伝播で微調整する



$$\text{input: } X \quad \text{code: } h = W^T X$$

$$\text{loss: } L(X; W) = \|W h - X\|^2 + \lambda \sum_j |h_j|$$

正則化項

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回帰など優位
 - DLで構造活性相関を学習する
 - 特徴量の抽出、薬理機構への
 - 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)

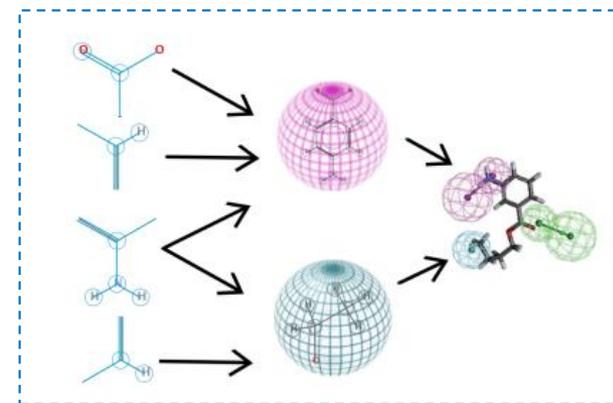
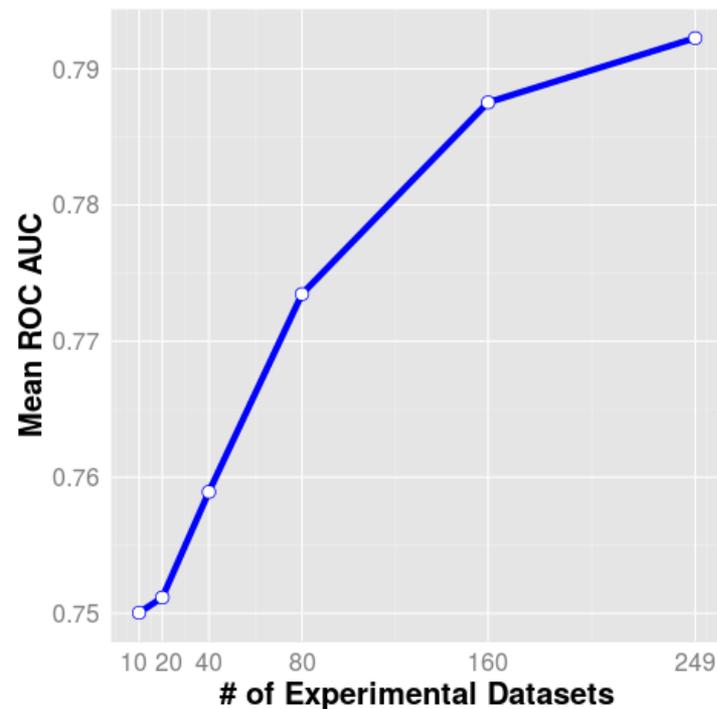


Figure . Hierarchical nature of fingerprint features: by combining the ECFP features we can build reactive centers. By pooling specific reactive centers together we obtain a pharmacophore that encodes a specific pharmacological effect.

ビッグデータのデータ原理

- スパース推定も Deep Learning も次元縮約的な特徴量を探索するビッグデータ解析のための原理に基いている。
- ただし従来型の次元縮約とは違って、縮約次元は大きく標本数 n に影響される（縮約次元は n の関数）



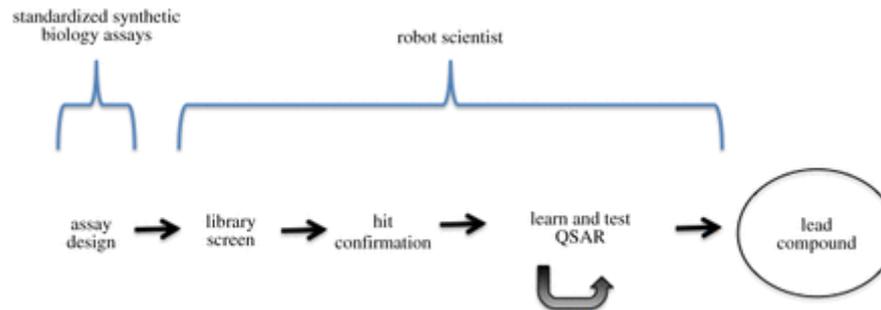
データ数が増加すると内部的な縮約次元が増加し推定精度が上昇する

そのほかのAI創薬の話題

- 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（詳細不明）
- University of Cambridge and Manchester
Artificially-intelligent Robot Scientist for new drugs
 - library screening, hit-confirmation, lead generationなどの自動化
 - Quantitative structure activity relationship (QSAR) 学習を反復する
 - Deal with 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
 - 最上位にコンセプト木（“root: assay triple screen”など）



Robot scientist Eve at work



ゲノム医療でのAI・機会学習

米国におけるゲノム・オミックス医療

2005~ NGS登場 (454 Life sci)

2007~ シーケンス革命

2010

ゲノム医療臨床実装の開始
臨床WESの最初 (MCW)
先制PGxの最初 (VU)

- MCW Nic君原因不明腸疾患 WES
XIAPの変異同定・骨髄移植
- Vanderbilt preemptive PG
(PREDICT計画) 開始

Wisconsin医科大学
臨床シーケンス初例
大きなインパクト

Baylor医科大学
Mayo Clinicなど
後続病院多数

Early adopter
時期

第1世代

2013
前後

ゲノム医療の国家的取組み
NIH "BD2K" initiative 開始
各種ゲノムコンソーシアム

ビッグ
データの
概念

NIH "Big Data to Knowledge" 計画 (2012/13)
ACGM incidental finding list 56 genes (2013)
NACHGR report "Future is here" (2013)
CPIC guideline, EGAPP guideline 2013.14

国家政策/全国Consortium
時期

第2世代

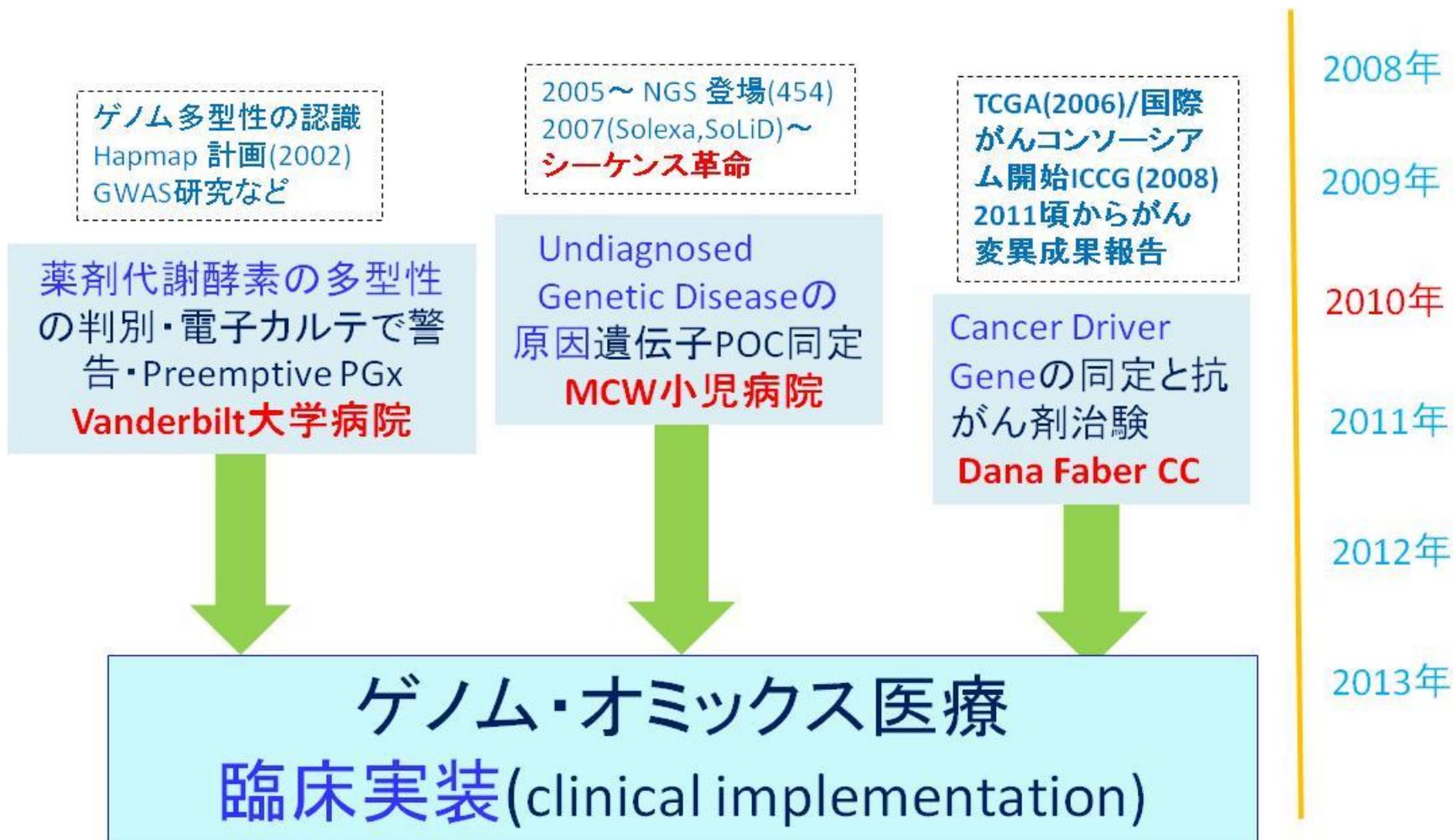
2015

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

ゲノムオミックス医療 すでに数十の医療
施設でG/O医療が病院の日常臨床実践

NIH "BD2K" COE in Data Science, DDI (2014)
ASCO "CancerLinQ", Cancer Common
"Precision Medicine (Obama)" 1 M genomic cohort

ゲノム・オミックス医療の基軸



ゲノム/オミックス医療－米国の状況

現状 米国ではすでに**数十の医療施設**で
ゲノム/オミックス医療が病院の日常臨床実践

NHGRI Working Groupのリスト

- Wisconsin大学病院
 - 原因不明の遺伝疾患の診断
- Vanderbilt大学病院PREDICT計画
 - 薬剤代謝酵素の多型性
- Mayo Clinicの臨床ゲノムシーケンス
 - がんおよび非常に稀な遺伝病原因探索
 - 10万人ゲノムDB
- その他、右表にあるように多数の病院
- 分子情報と臨床情報の融合を目的として統合データベース
 - Mofit Cancer Center (Oracle HRI)
 - 製薬会社Merkと病院の契約

Institution	Major Projects
MC Wisconsin	Using whole genome sequencing to establish diagnosis in patients with currently undiagnosed genetic disorders
Mount Sinai	<ul style="list-style-type: none"> • CYP2C19 testing for antiplatelet rx post percutaneous coronary intervention • Personalized decision support for CVD risk management incorporating genetic risk info
Northwestern	Using pharmacogenomics evidence (from GWA genotyping) to guide prescriptions in primary care and assess risk for other conditions such as HFE/hemochromatosis
Cleveland Clinic	Tumor-based screening for Lynch syndrome, endometrial cancer
UCSD	<ul style="list-style-type: none"> • Screening for actionable mutations in malignant gliomas and glioblastomas for biomarker based RCTs • Targeted rx (such as RET inhibitor) of metastatic solid tumors based on tumor mutation status
Morehouse	• Exome sequencing of 1200 early onset severe African American hypertension cases and 1200 controls
Duke	<ul style="list-style-type: none"> • Computer-based family hx collection and CDS tool with 1-yr follow-up for perceptions, attitudes, behaviors related to thrombosis and breast, ovarian, and colon cancer • SLC01B1*5 genotyping and statin adherence • Effect of genetic risk info on anxiety and adherence in T2DM

Institution	Major Projects
Alabama	Planning stages for projects in risk assessment, pharmacogenetic analysis, identification of families for further research
Baylor	Whole exome and whole genome sequencing in Mendelian disorders to improve diagnosis
Geisinger	<ul style="list-style-type: none"> • Selection for gastric bypass surgery vs other wt loss means based on genetic variants predictive of long-term benefit from surgery • IL28B variants and response to hepatitis C treatment • KRAS and BRAF mutational analysis in thyroid cancer patients
Ohio State	<ul style="list-style-type: none"> • Personalized genomic med study of CHF and HTN pts randomized to genetic counseling vs usual care • CYP2C19 testing in interventional cardiovascular procedures for clopidogrel
Harvard	Whole genome sequencing with integration in EMR and CDS; pilot of 3 patients to start
U Penn	Genotyping for assessment of MI risk in Preventive Cardiology program
St. Jude's	Pre-emptive PGx genotyping in children
Vanderbilt	Pre-emptive PGx genotyping for clopidogrel, warfarin, or high-dose simvastatin
U Maryland	Develop and apply evidence-based gene/drug guidelines that allow clinicians to translate genetic test results into actionable medication prescribing decisions
Mayo	<ul style="list-style-type: none"> • PGx driven selection/dosing of antidepressants • CYP2C19 genotyping for antiplatelet rx post PCI
Inter-Mountain	Tumor-based screening for Lynch syndrome

バイオ・医療におけるBig Dataの 構成要素

- 網羅的分子情報 (Genome/Omics)
 - ゲノム・オミックス情報、multi-omics
- 臨床環境表現型 (Phenotyping)
 - eMERGE-I, PheKB, EWAS
- 知識発見システム (Data Science)
 - Data-mining, Knowledge Discovery, 人工知能

医療におけるビックデータ

ゲノム情報 + 臨床・環境情報(EMR等)

知識発見(learning) システム

米国のゲノム医療における 学習 CancerLiQ

- ASCO(米国臨床癌学会)
- The ASCO **CancerLinQ** initiative
 - “Learning Health System”あるいは”**Rapid Learning**”の概念をがん治療において実現する。
 - データがバラバラの状態を改変、集合化
 - 診療の現場(EHR)から大量の診療データを集め分析する
 - 新しい臨床治験へのガイドライン検討
 - 病院の患者によって17万人のがん症例データベースを構築。各がんについて1～2万人の症例を集める
- 学習システムを構築し、仮説を生成する治療知識を供給する。統計学習、ニューロネットを駆使して学習。

BigDataにおける学習システムの不可欠性

- 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

米国におけるゲノム医療における学習システム

米国研究施設と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、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最良の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デザインされたWatsonの試作システム
 - 脳腫瘍のglioblastoma (グリア芽細胞腫) を対象。ゲノム配列と医療情報、医学文献から個別化治療を提案
- **その他のWatsonの応用**
 - Cleveland Clinic
 - 問題解決型学習プログラム、仮説的な臨床シミュレーションの中で、Watsonを対話的に利用して症例を分析。Watsonの技術を利用した共同型の学習教育ツール
 - MDA Anderson
 - 治験に適切な患者を診療情報から選別(clinical trial)

Drug Repositioningにおける 機械学習

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 : 差別的発現遺伝子)
Signature of Differential gene Expression
- DB利用 : 発現刻印をquery, 順位尺度で類似性の高い順に化合物を提示

- 疾病特異的遺伝子発現(Disease-associated SDE)

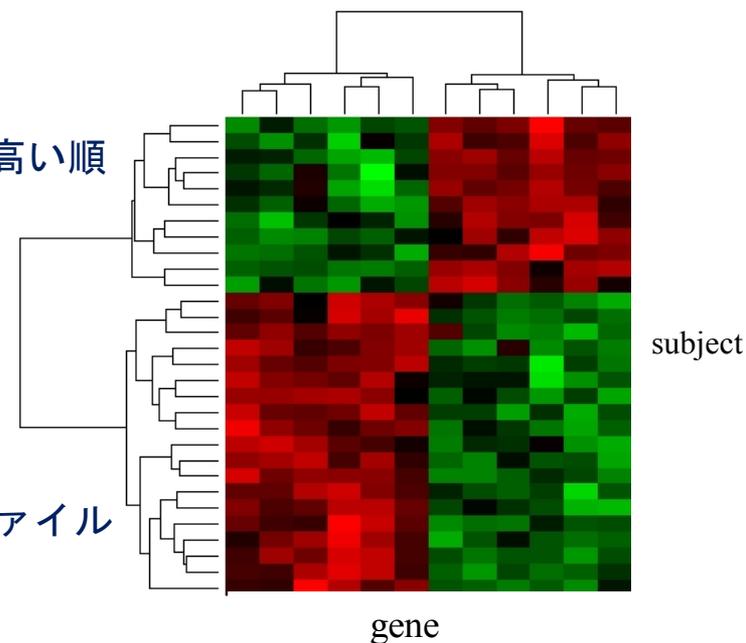
- GEO (gene expression omnibus),

- 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)。

遺伝子発現プロファイル変化

≈分子ネットワーク変化



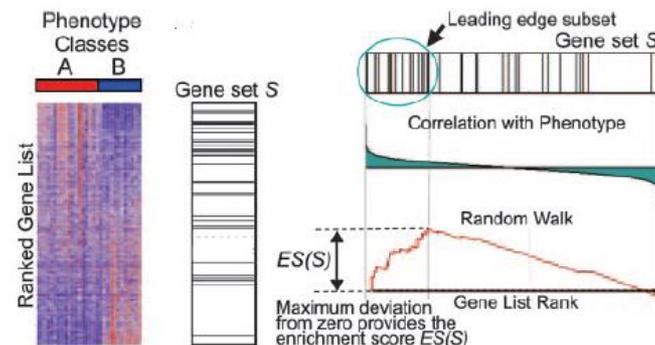
遺伝子発現プロファイルによる有効性予測

- 遺伝子発現シグネチャ逆位法 (signature reversion)

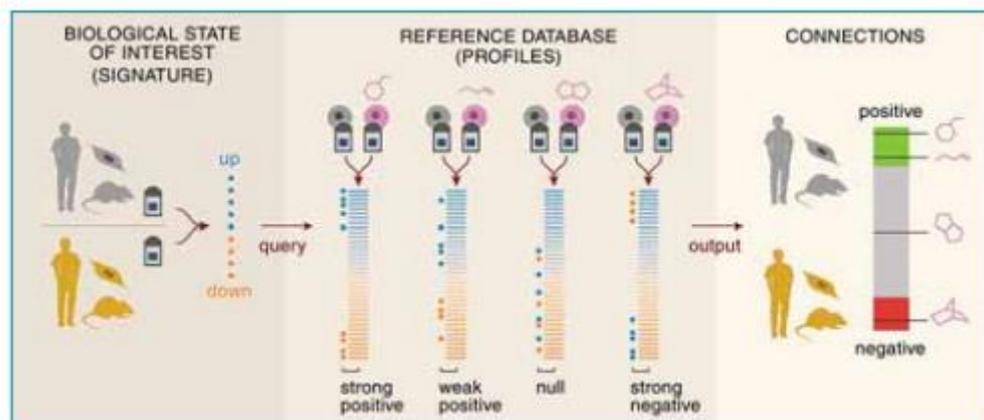
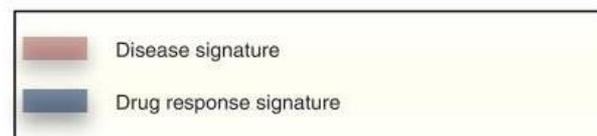
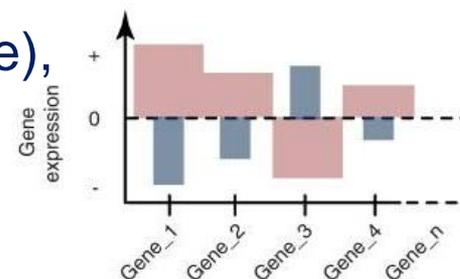
- 薬剤特異的遺伝子発現シグネチャ
- 疾患特異的遺伝子発現シグネチャ
- 有効性予測**：両者が負に相関する
- Non-parametric な相関尺度で評価

- 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

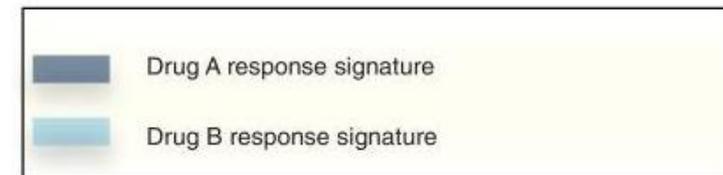
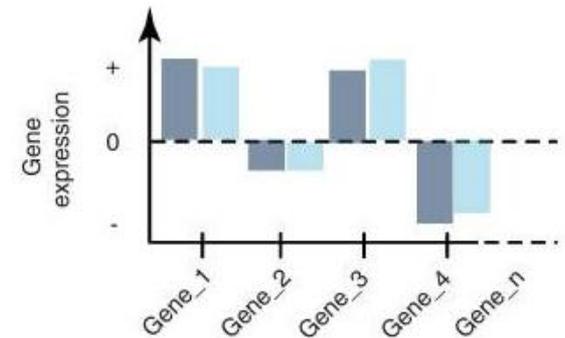


GSEA



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－薬剤間
 - － 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
- 薬剤－疾患間
 - － 薬剤特異的シグネチャと
 - － 疾患特異的シグネチャが
 - － ノンパラメトリック相関 正
 - － **毒性・副作用の予測**



Drug Repositioning

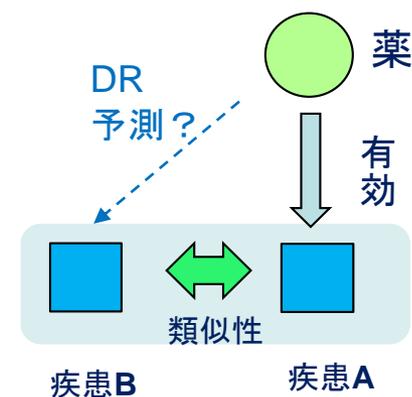
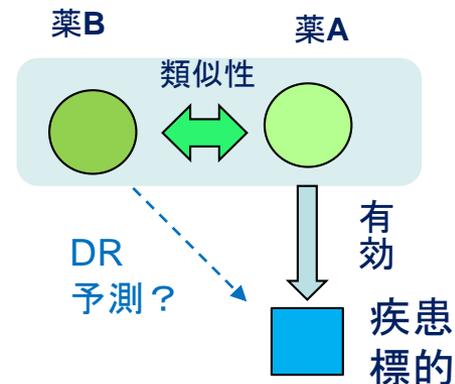
ヒトでの安全性と体内動態が十分に分かっている
既承認薬の標的分子や作用パスウェイなどを、体系的・論理的・網羅的に解析することにより新しい
薬理効果を発見し、その薬を別の疾患治療薬として
開発する創薬戦略

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**

合理的DR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



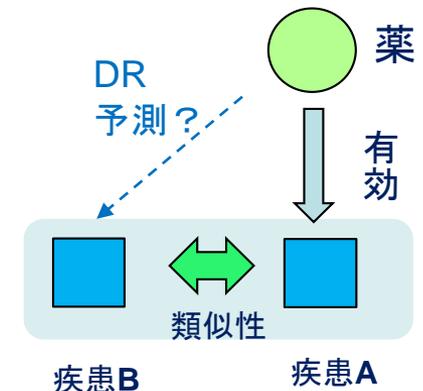
疾患ネットワークによるDR

第1段階 疾患ネットワークの構築

- 疾患のゲノム・オミックス機序の類似性をネットワークで表したもの

第2段階 DR候補予測

- 既存の<疾患-薬剤>を決め
近傍の疾患をDR候補とする

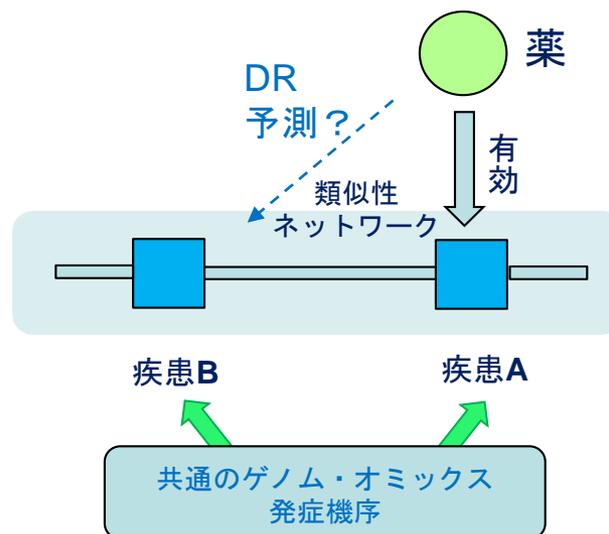


第3段階 有効性・毒性の確認

- シグネチャア逆位法、連座法で有効性、毒性を予測

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年に亘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



疾患の成立機序における主要機序

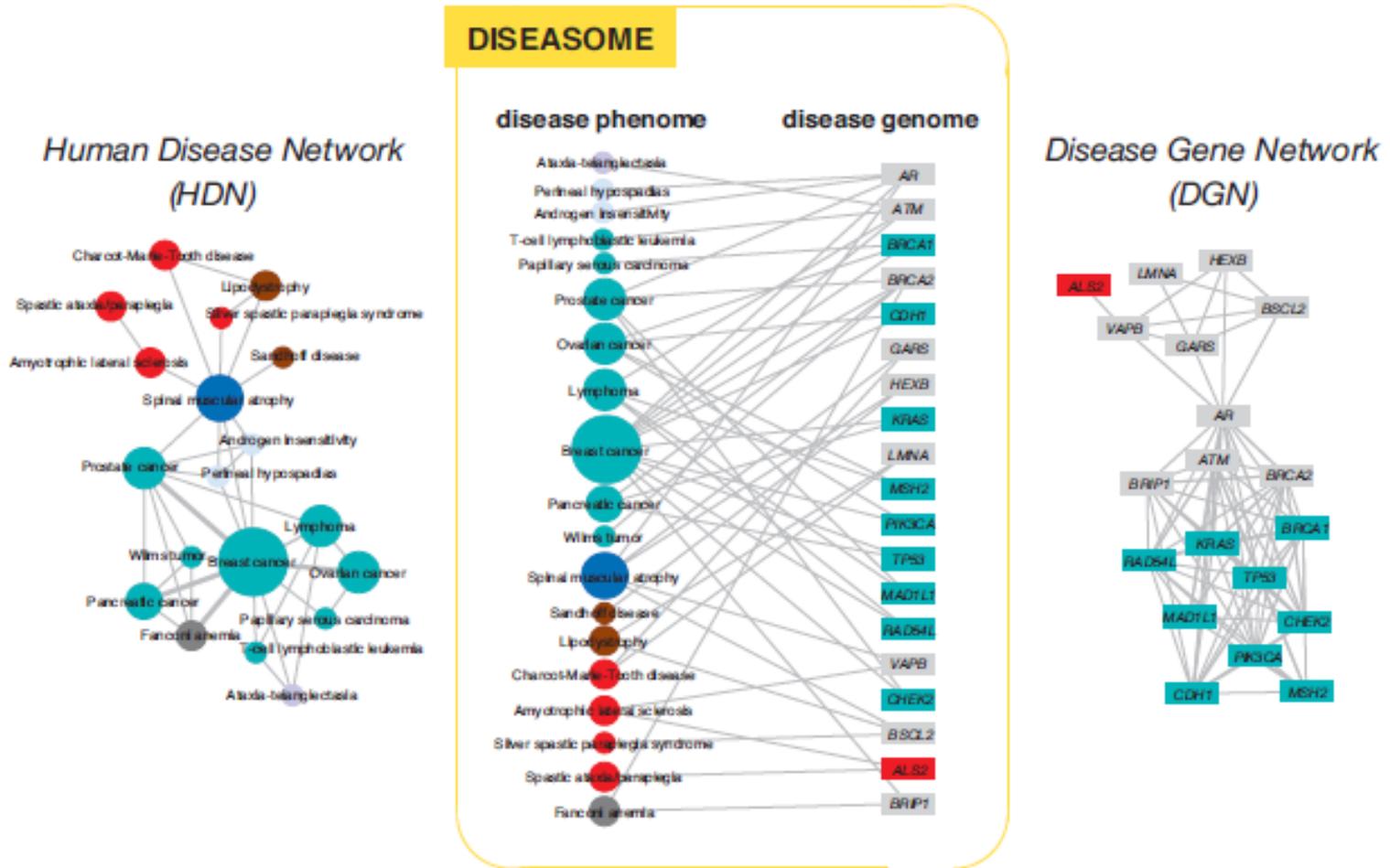
- 疾患関連遺伝子型（第一世代型）
 - 原因遺伝子、疾患感受性遺伝子の変異・多型が主要発症機序
- 疾患オミックス型（第2世代型）
 - 疾患オミックスプロファイルの変容が主要発症機序
 - Transdisease omics
- 疾患分子ネットワーク型（第3世代型）
 - 分子ネットワークの歪みが主要発症機序
 - がんなどで遺伝子型（肺腺がん等）でない通常のがん

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

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

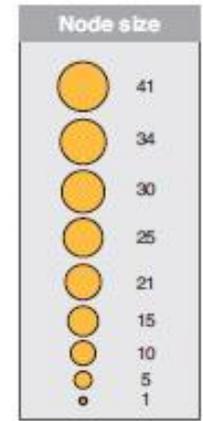
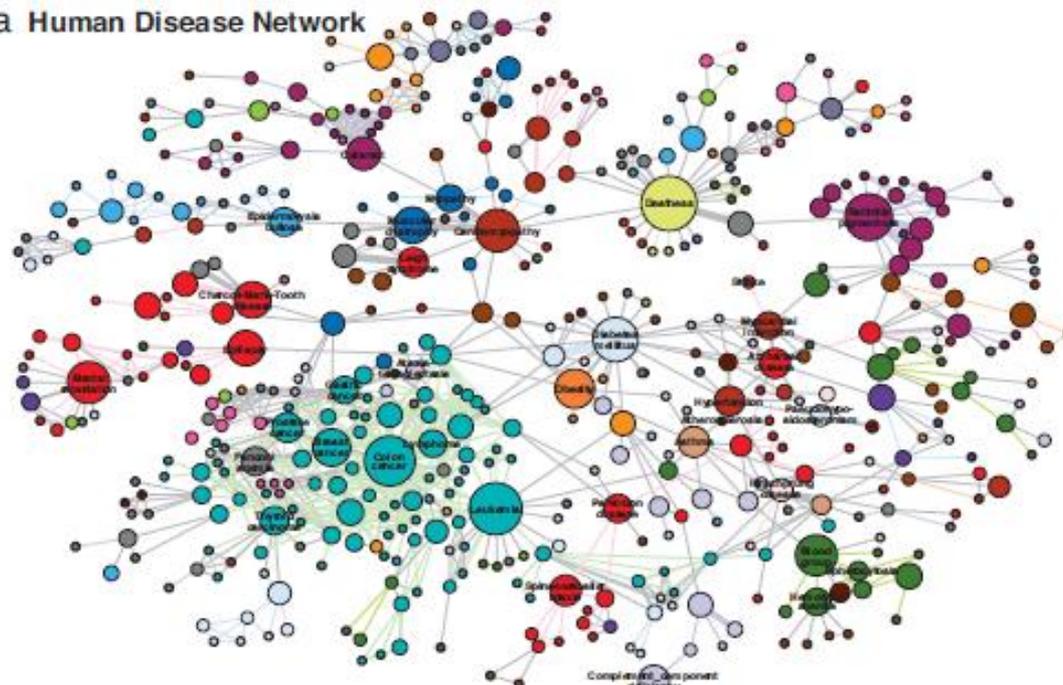
Kwang-Il Goh*, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-Laszlo Barabasi The human disease network PNAS2007

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

a Human Disease Network

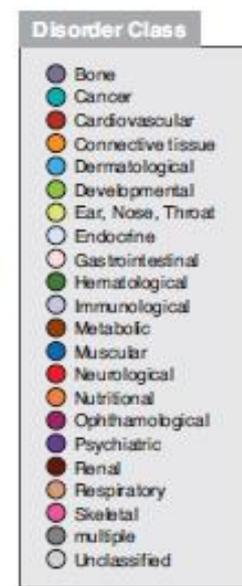
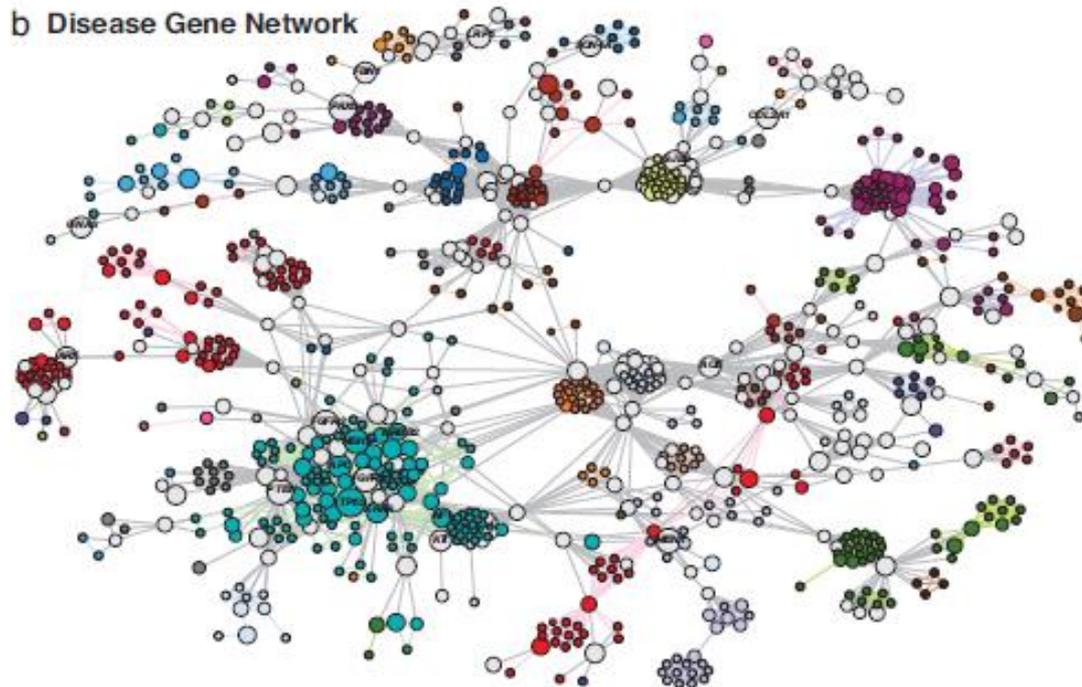


疾患遺伝子 ネットワーク (DGN)

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

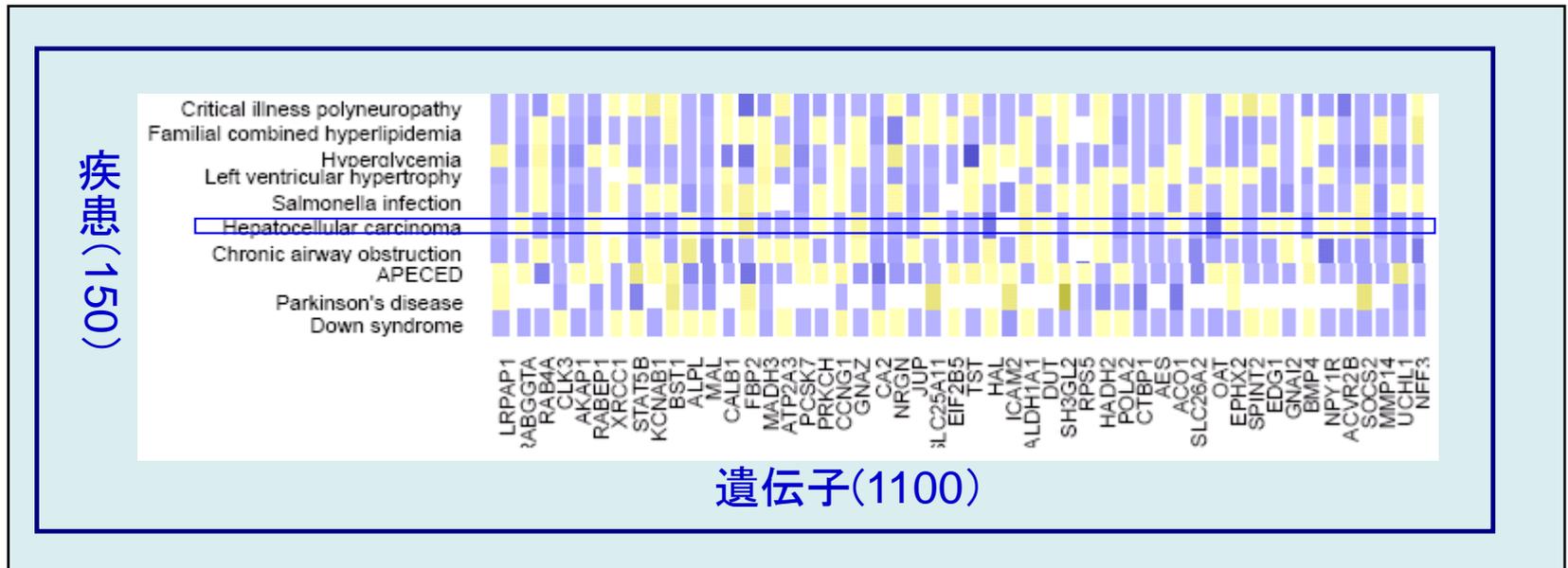
b Disease Gene Network



第2世代型

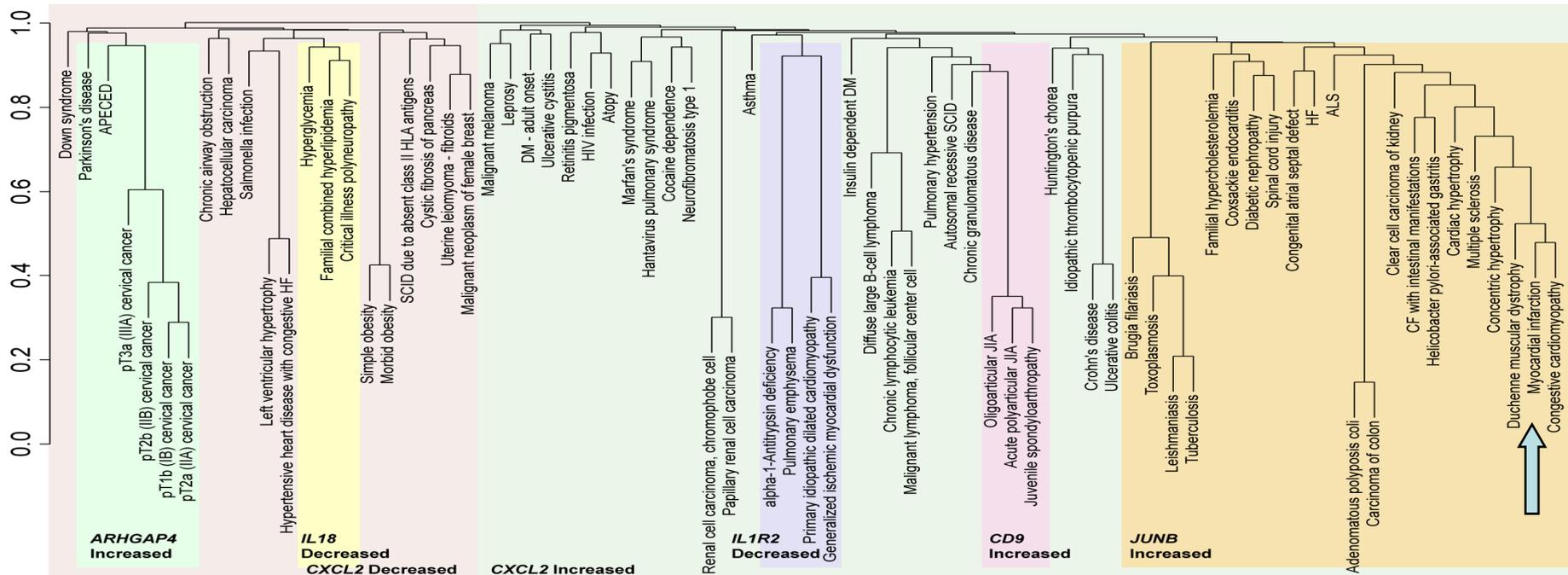
GENOMED (A. Butte et al)

- 遺伝子発現DBのGEO (Gene Expression Omnibus) 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に近い



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)

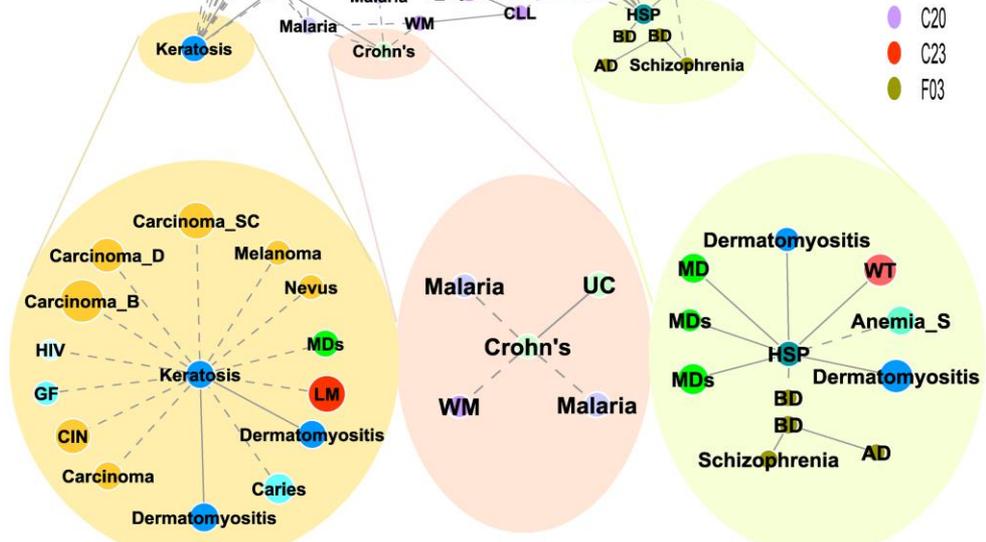
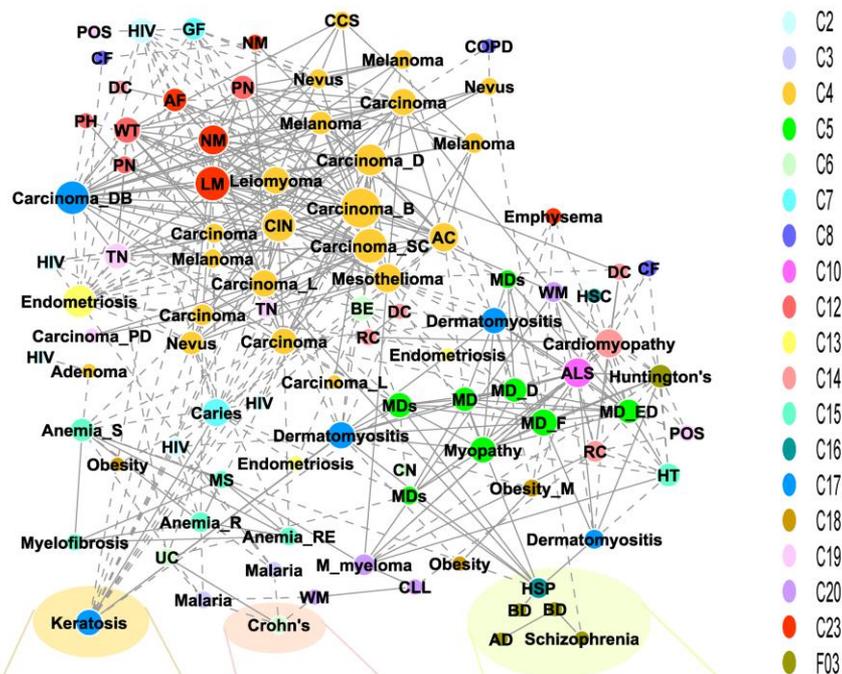
Crohn's disease

⇒ 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

Transcriptional Profiling による 疾患ネットワーク

Disease-drug network

橙色 医薬品
緑色 病気

Tamoxifen (breast cancer)

(負の値をもっている)

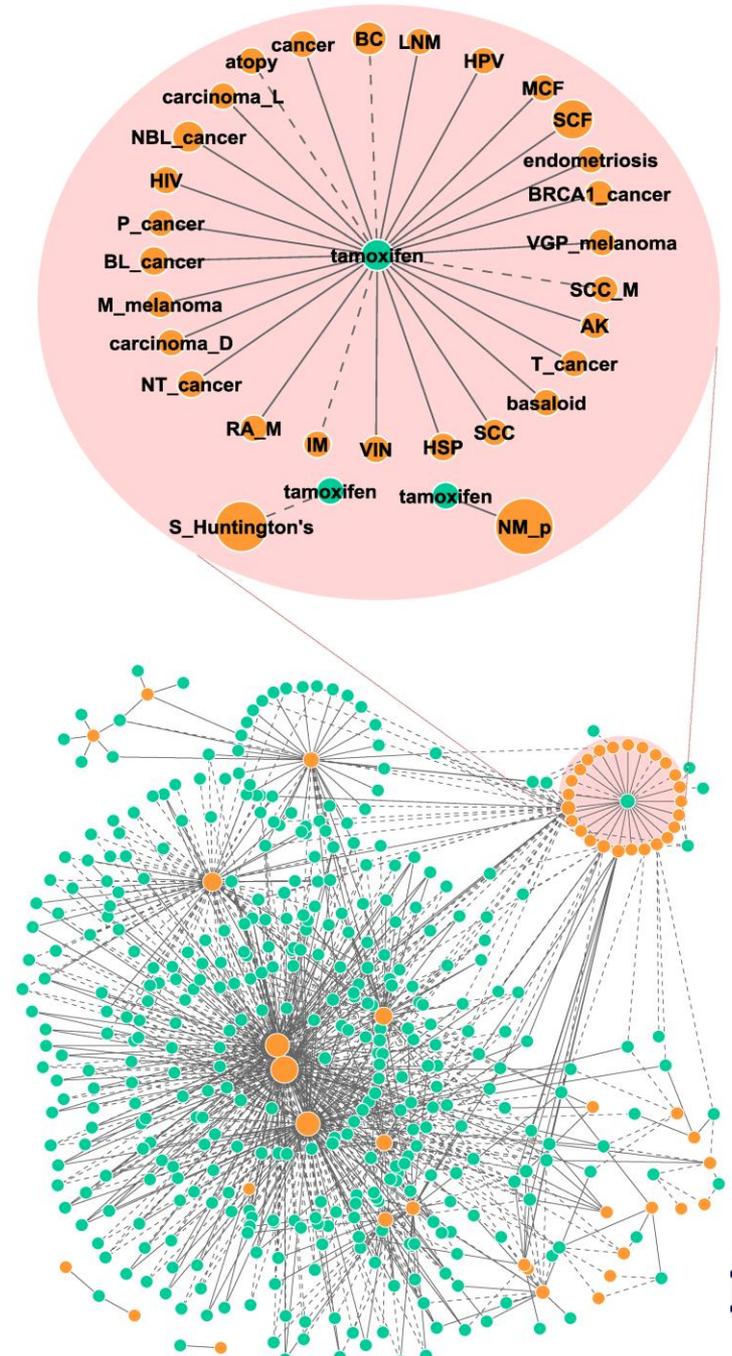
⇒ アトピー,

⇒ マスト細胞分泌抑制、
アレルギー抑制

(正の値をもっている)

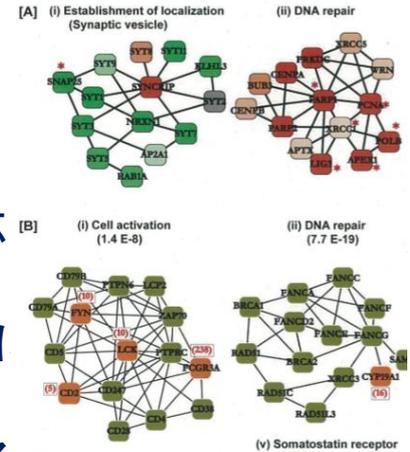
副作用の予測

⇒ carcinogenic

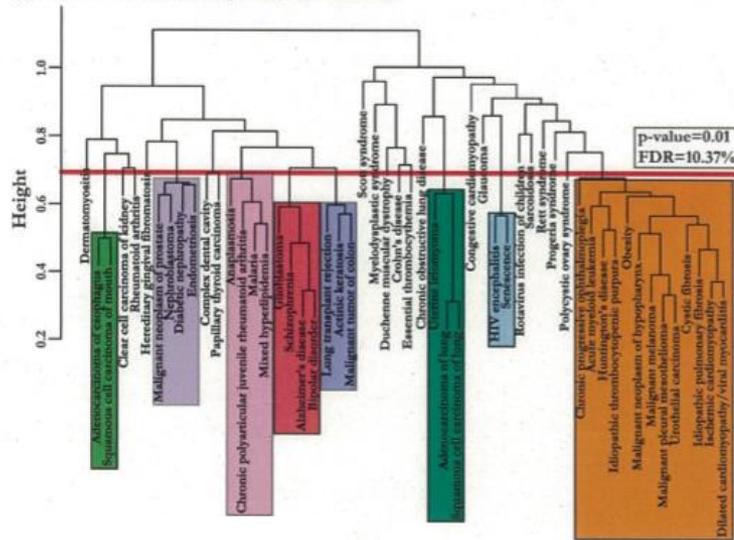


Transcriptomeの変化をPPIに投影した疾患ネットワーク (Butte)

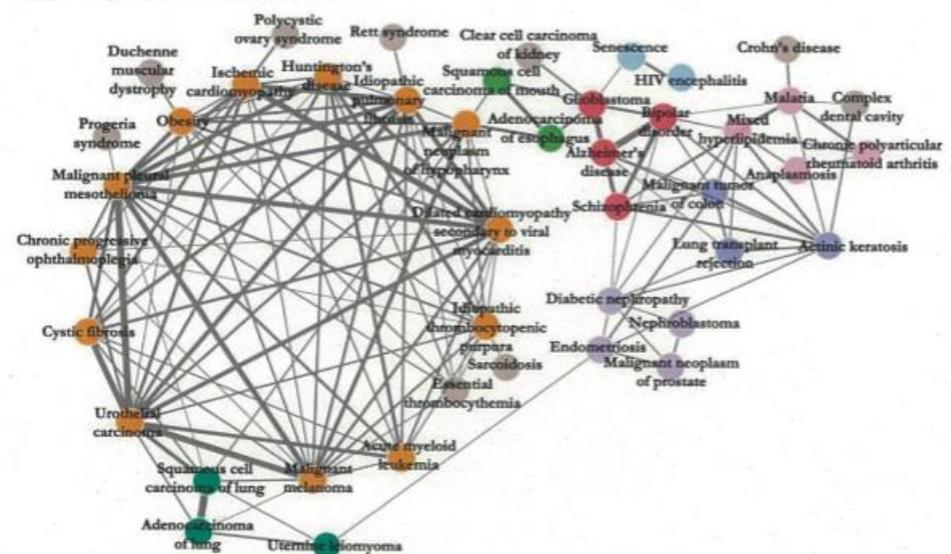
- 遺伝子発現プロファイルを直接使うのではなく4620Moduleに分解したPPIネットワークでの疾患での平均発現変化をつかう
 - 遺伝子発現プロファイルより疾患によって変化するmoduleを調べる 病気に対するPPIの応マラリアとクローン病
 - moduleの遺伝子の変化を平均して遺伝子の代わりmoduleの発現平均スコアを用いる
- 疾患の大半を占める<共通疾患状態シグネチャ>



[A] Hierarchical relationships between diseases

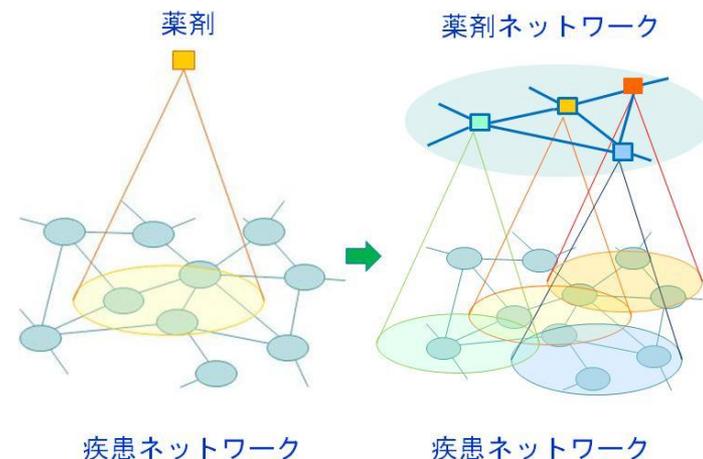


[B] All significant disease correlations



創薬への機械学習・AIの可能性

- 疾患ネットワークと薬剤（Chemical）ネットワークの2つのネットワークにおける双方向projection関係
- 既知の双対写像の関係から、未知の関係の学習
 - ビッグデータのデータ原理
 - 双対写像の有効次元縮約
 - Deep learning
 - スパース推定
 - 糖尿病性網膜症に抗炎症剤



ご清聴ありがとうございました



「学習する医療システム」 Learning Health System

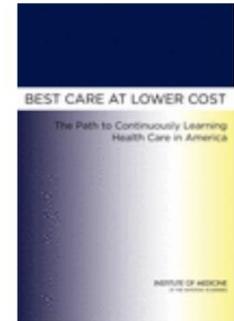
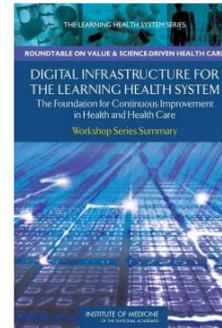
新しい生物医学知識が臨床実践に給されるまで17年
臨床データを用いて医療を実施しながら医療を改善

- 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

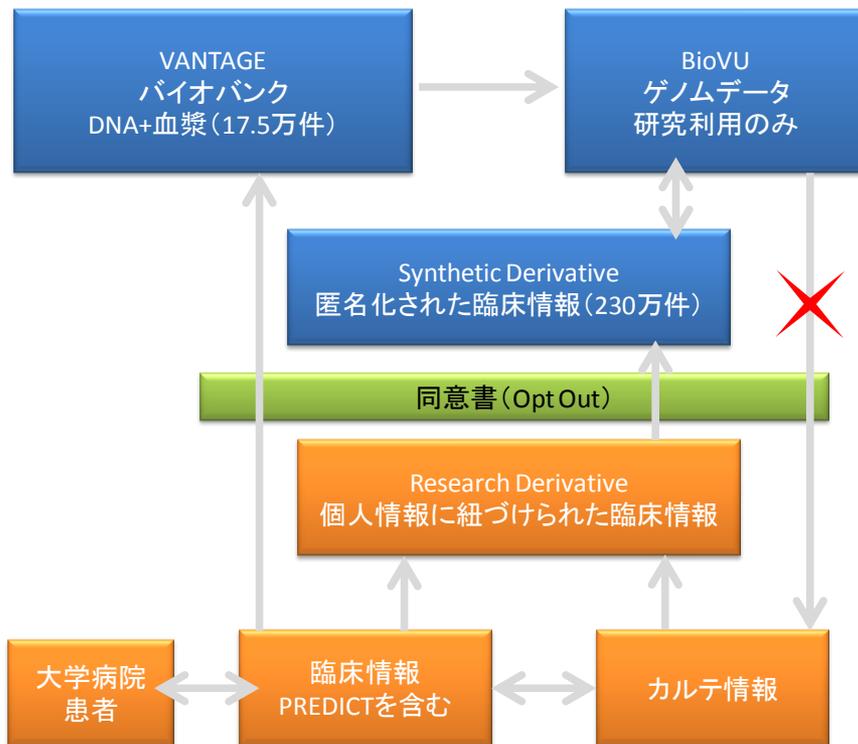
Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care

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）の遺伝子多型に関して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。BioVU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。

新しい臨床研究のパラダイム Real World BD2K

- 大半のRCTは医療の外に人工的な環境
- 「標本」からの「推測」の概念
- 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」
- しかし、新たなパラダイムを構築するまで時間・研究が必要
- 我が国の戦略 段階的移行 BioBankの利用段階



Digital Medicine

Precision Medicine

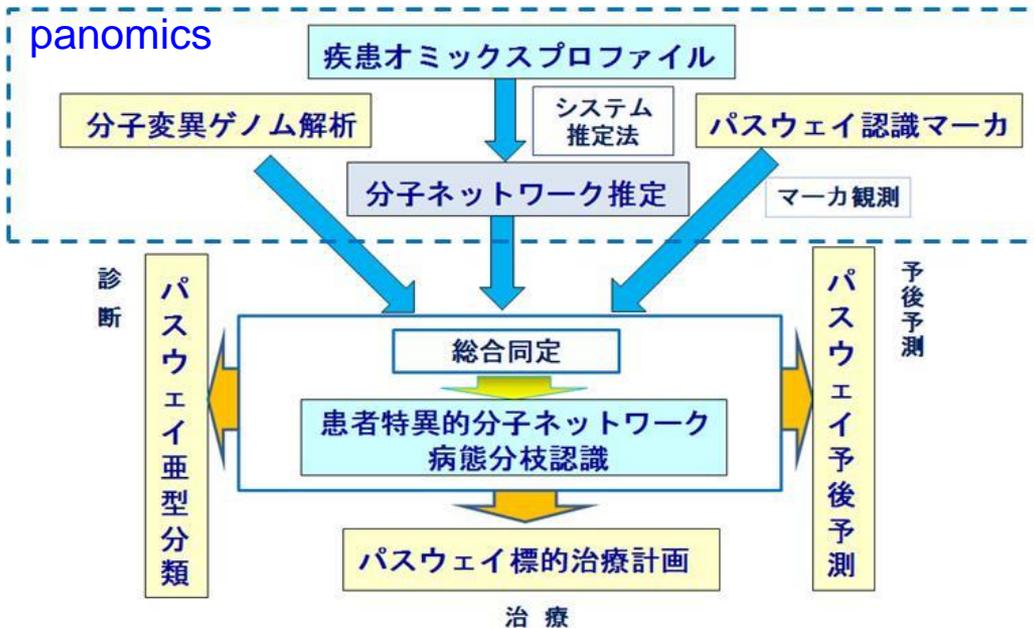
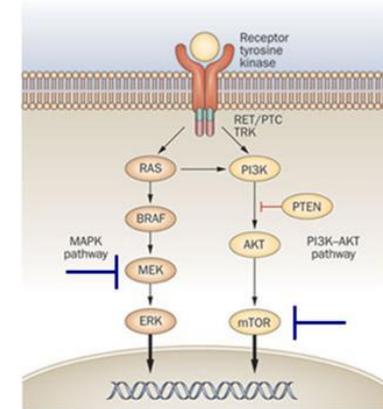
層別化パターンの詳細化・パスウェイ同定
⇒ 医療の的確性の増大

Participatory Medicine

生涯的時間軸での医療・ケア
⇒ Intelligent Proactive Consumerの増大

Precision 医学とは分子システム医学のことである

- 個別化・分子標的創薬
 - 疾患（がん）を分子変異で層別化
 - バイオマーカ（分子変異 genomic biomarker)
 - 基本概念 **Oncogene addiction**
- システム分子医学的疾患認識へ（分子システム）
 - 「がんはパスウェイの病気である」
 - **Pathway Addiction**
 - Panomicsより患者特異的パスウェイ分枝を決定
 - 疾患のシステムの患者特異性



1. 疾患オミックスプロファイルから
 → **患者特異的分子ネットワーク(個別化医療)の**
 調節不全分枝 同定
 Dysregulated pathway/subnetwork の同定

2. パンオミックスによる臨床的実践の戦略

- 遺伝子発現プロファイル** 推定法による分子ネットワークの同定 (80%)
- 次世代シーケンシング** 転写因子や信号パスウェイスイッチ分子の変異
- リン酸化プロテオーム** パスウェイバイオマーカ
リン酸化状況の認識

Precision 医学とは分子システム医学のことである

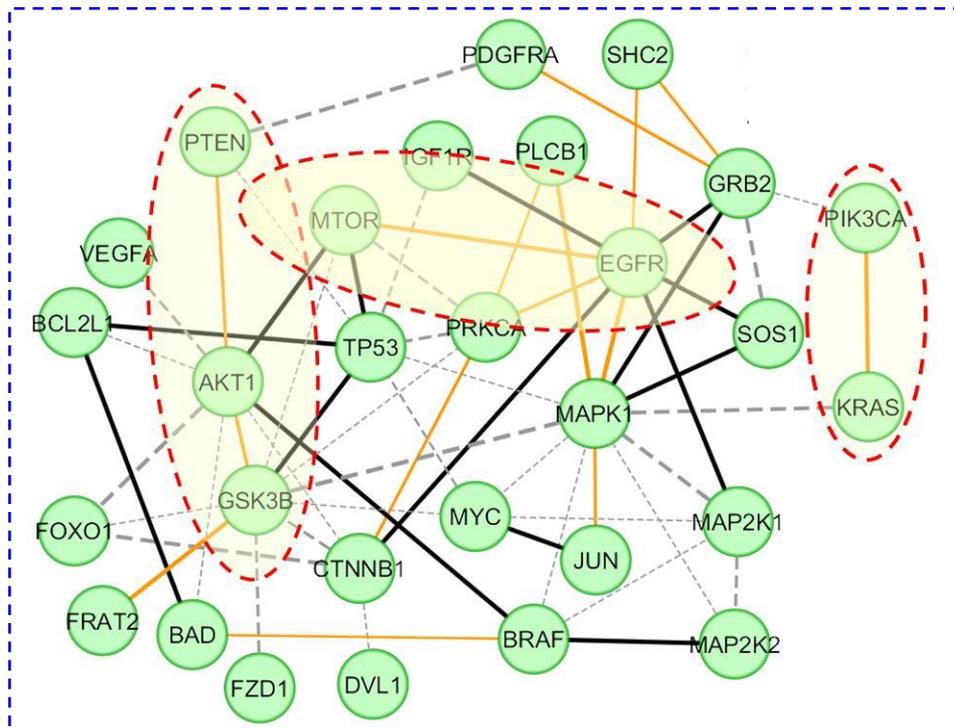
個人のpanomics1例だけではネットワークは同定できない
臨床表現型と合わせて層別化する。類似症例検索より推測

疾患ネットワーク同定方法論

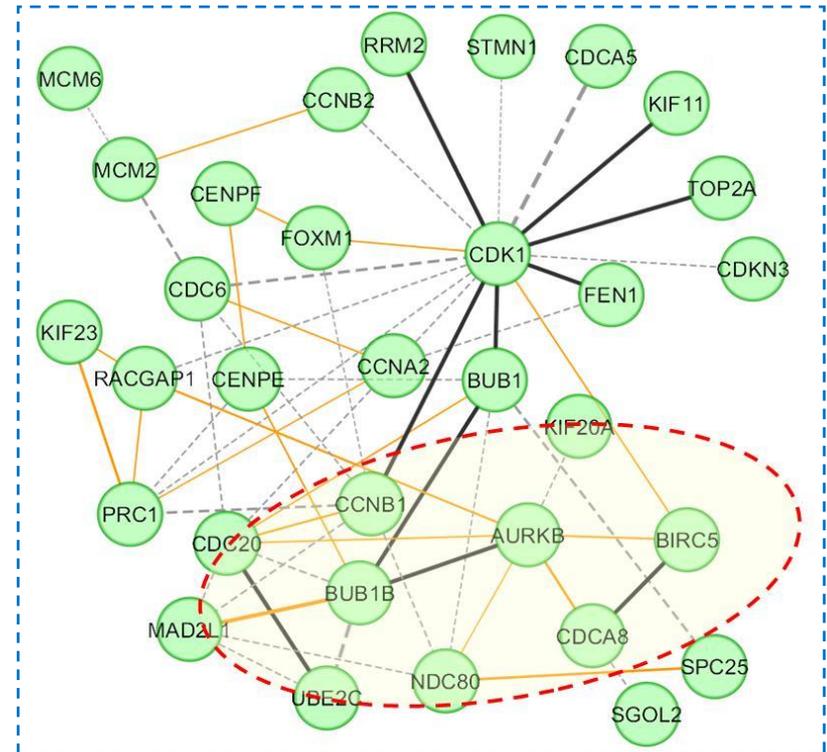
東京医科歯科大学病院 肝胆膵外科にて肝胆膵外科で手術を受けた40症例

— 予後良好群20例 vs. 予後不良群20例

細胞増殖分子ネットワーク推定
differential branches 増殖系の過剰活性

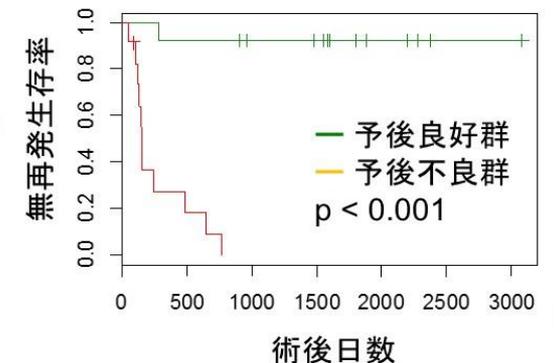


細胞周期分子ネットワーク推定
differential branches 結果 AurkBを起点としたNW



差異的発現分枝

— 予後不良群に特徴的
— 両者共通
..... 既知PPIあり

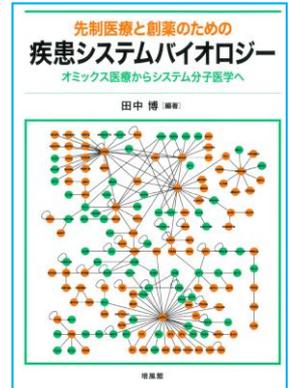
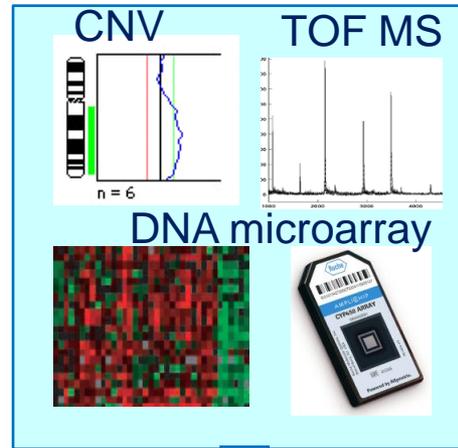


Precision Medicineとは 分子システム医学のことである

断層像
再構成技術



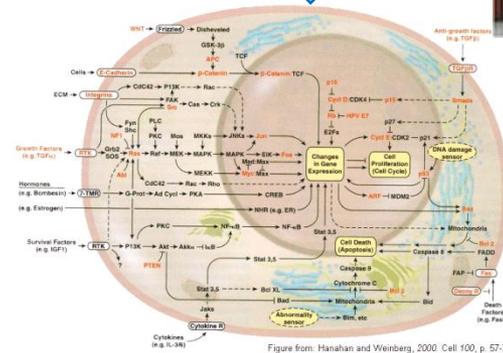
疾患
パンオミクス
プロフィール



スーパー
コンピュータ



逆計算



合理的な
診断・治療
予後予測

患者特異的分子ネットワーク
調節不全分枝同定

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、臨床表現型ネットワークと分子ネットワークを繋げる機構**
 - 遺伝子を通して疾患間を移動できる
 - Systems pathobiology, nosology, personalized medicine