ビッグデータと人工知能（AI）を用いた創薬・DR

東京医科歯科大学　医学部 臨床腫瘍学・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

本日のトピック

- 医学・ヘルスケアへのビッグデータ時代の到来
  - 医療ビッグデータの特徴
  - ビッグデータによる医療・創薬のパラダイム変化
- ビッグデータ医療・創薬の3つの流れ
  - 第１：ゲノム・オミックス医療の発展
  - 第２：Biobankとゲノムコホートの世界的興隆
  - 第３：大規模な生命情報DB/KBの出現と利用
- ビッグデータ創薬/DR
  - 大規模網羅的分子情報の知識・DBの利用
  - 1. 発現プロファイル創薬/DR
  - 2. 疾患ネットワーク創薬/DR
  - 3. Interactome 創薬/DR
  - ＜疾患-薬剤-標的分子＞の多階層ネットワーク
  - 我々の創薬/DRの研究
  - ビッグデータのclinical trialでの利用
- AI創薬/DR
  - Deep Learning 型人工知能の革命性
  - Deep LearningのAI創薬への応用
- ゲノム・オミックス医療の次世代の展開

http://ngbrc.com/concept2.html

TMDU

医学・ヘルスケア分野への
ビッグデータ時代の到来

TMDU

医療ビッグデータ時代の到来

(1) 次世代シーケンサなどによる「ゲノム/オミックス医療」による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) Biobankによるゲノム・コホート情報

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**
  - 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

新しいタイプの
医療ビッグデータ

TMDU

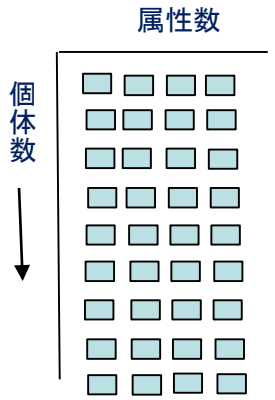
医療の「ビッグデータ革命」

～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)



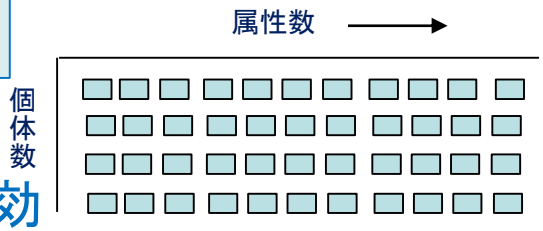
医療情報・疫学調査では 属性数：10項目程度

– 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,個別化概念にもとに破綻した
  - ＜statistical evidence based＞呪縛からの解放
    - 「標本」統計・「推測」統計学に制約されない臨床研究
  - Real World (Big) Dataの利用、からの知識生成（BD2K)
  - Register-based Randomized Clinical Trials (RRCT)
- 創薬戦略のパラダイムの転換
  - ＜ビッグデータ創薬＞: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の各ネットワーク階層間の相互写像関係

TMDU

ビッグデータ医療・創薬の3つの流れ

次世代シーケンサのインパクト

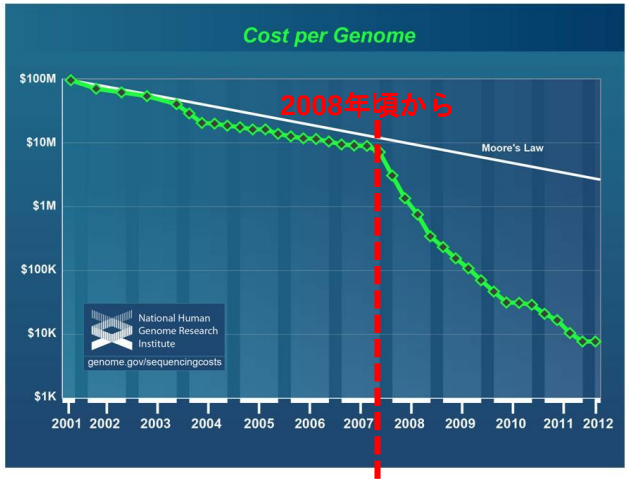
次世代シーケンサを始めとするhigh-throughput分子情報収集の急激な発展

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

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

Hiseq X システム　10台構成 (経費1/5)



DNA Sequencing Cost: the National Human Genome Research Institute

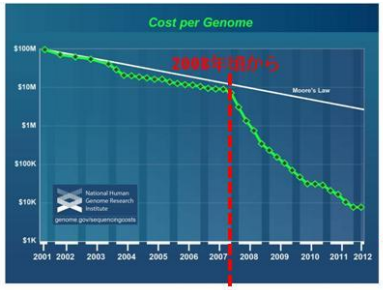
シーケンス革命　2007/8

ゲノム（配列決定）機器の進歩は、計算機のムーアの法則を越えている！

TMDU

第1の流れ

ゲノム・オミックス医療の発展



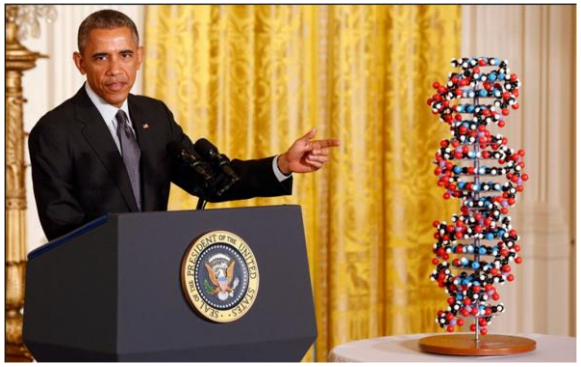
DNA Sequencing Cost: the National Human Genome Research Institute

シーケンス革命　2007/8

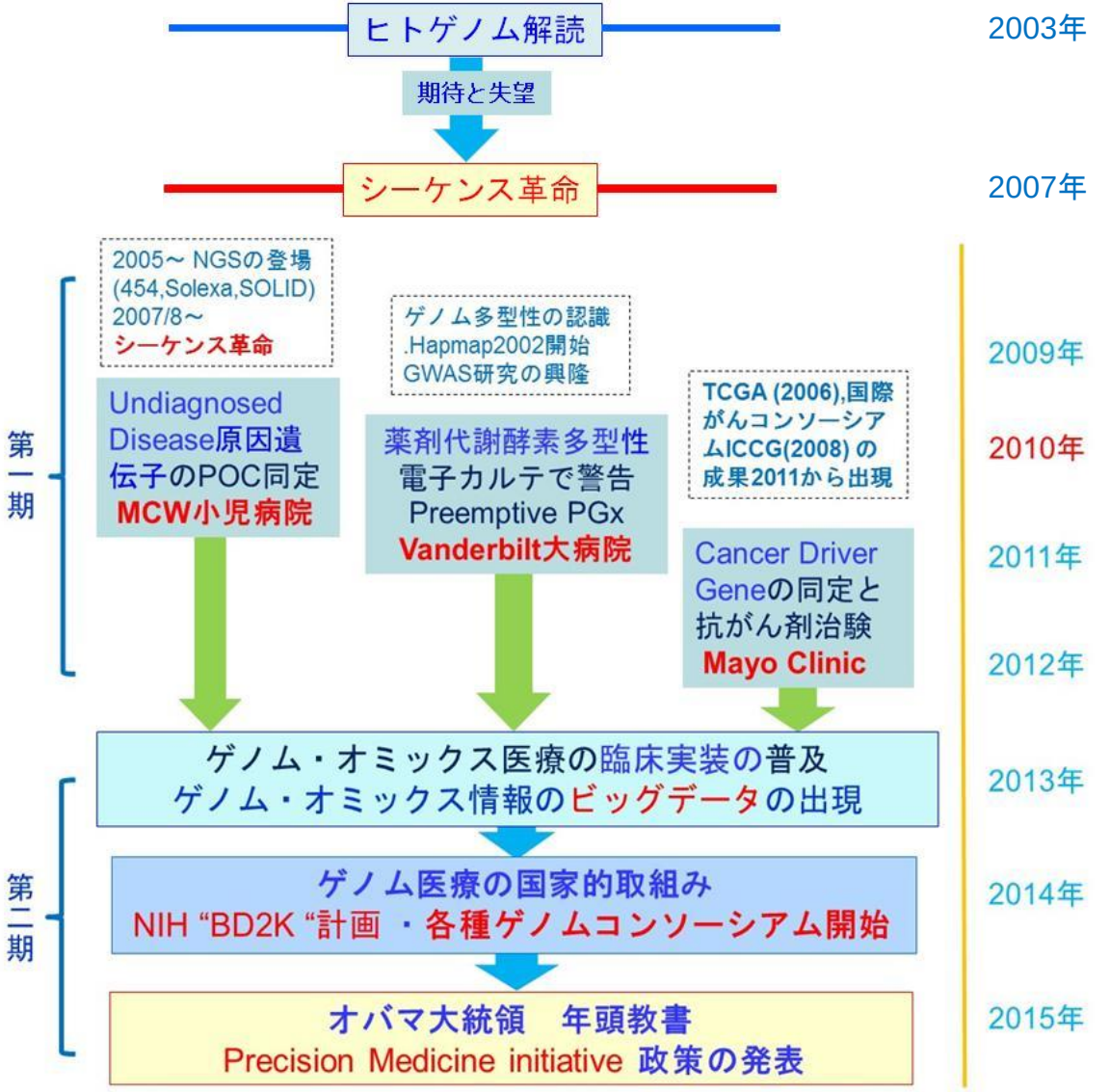
2005～ NGS 454 (LS,Roche)
2007/8～454, Solexa (Illumina), SOLiD (LT,TF)
シーケンス革命

	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



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



米国におけるゲノム医療の開始

第１世代の（生得的）ゲノム医療が中心
次の２つの潮流が同時に2010年に開始

(1) 原因不明先天的疾患(undiagnosed disease)
原因遺伝子の臨床の現場で(POC) の診断
次世代シーケンサの爆発的発展を受けて
Wisconsin 医科大学での全エキソーム解析

(2) 薬剤の代謝酵素の多型性の検査
臨床の現場で電子カルテの警告(診療支援）
Vanderbilt大学病院の先制ゲノム薬理

ゲノム医療：少数の予想される遺伝子の変異を調べる**候補遺伝子アプローチ**はすでに「遺伝子医学」で行われていた。あらかじめ候補遺伝子を決めず、網羅的でデータ駆動的なゲノム解読（**ゲノム網羅的アプローチ**）によって変異を見出す医学である

ゲノム・オミックス医療の進展とビッグ・データ

2005～ NGS登場 (454 Life sci)
2007～**シーケンス革命**

2010

第１世代

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

Early adopter 時期

Baylor**医科大学**
Mayo Clinicなど
後続病院多数

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

第２世代

国規模の計画/全国Consortium 時期

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

2015

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

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

ゲノム医療の最初の臨床実装

ゲノム医療の第1の流れ　未診断病のClinical Sequencing

Nic Volker

- Wisconsin 小児病院（全米4位）2009年、3才の男子。
- 2歳から原因不明の腸疾患で、腸のいたるところに潰瘍が発生。
- クローン病かと疑うが、**クローン病の既報の遺伝子変異なし**
- 2年間で130回の外科的切除手術を行うが再発を繰り返す。これ以上行う治療がなくなった(A. Mayer)
- Nicの全エキソンの配列を次世代シークエンサ決定
- MCWで見出された16000個のDNA配列異常を慎重に分析

XIAP(X連鎖アポトーシス阻害タンパク質遺伝
変異　TGT(cysteine)→TAT(tyrosine)（203番目）
アポトーシスの阻害因子　免疫系が腸を攻撃する自己免疫を阻害　これまでのヒトゲノム配列で見出されていない　ショウジョウバエからチンパンジー見いだせず

Medical College of Wisconsin, Human & Molecular Genetics Center
Howard Jacob
(a major mover of the whole field, Topol)

臍帯血移植(造血幹細胞移植）を実施（2010年6月）
2010年7月半ば(**42日後**）には、食事が取れるまでに回復した。
現在は普通の男子と変わらぬ健康な生活を送っている。
2010年の12月に3回連載で全米に記事・記者にピューリツア賞

TMDU

Wisconsin医科大学小児病院および Froedtert 病院のゲノム医療

- Wisconsin医科大学　Genome sequencing program
  - Nic君に続いて（翌年3月まで6例）
  - 候補選択（nomination）
    - 従来の検査・診察で診断困難な症例
  - Multidisciplinary 患者選択委員会でレビュー
  - 6-8時間のアセスメントとカウンセリング
  - 32 全ゲノム, 550 全エキソーム（2015年4月まで）
  - アメリカ病理学会（CAP)およびClinical Laboratory Improvement Amendments(CLIA:CMS) 基準：最初外注 Froedtert 病院
  - データ解析：in-houseのBIで
- Baylor医科大学病院　2番手（すでに準備?）
  - Wisconsinに続いて臨床ゲノム配列解析
  - 病院内にWhole genome laboratory 設立**(2011.Oct)**
  - In-houseでシーケンシング/変異分析
  - CAP/CLIA認証の検査室を病院内に立ち上げる。
  - 臨床分子遺伝学者によって解析・結果報告
- そのほかにWashington大学、Partnerなど多数つづく

Wisconsin
小児病院

Wisconsin 医科大学（MCW）

Baylor医科大学

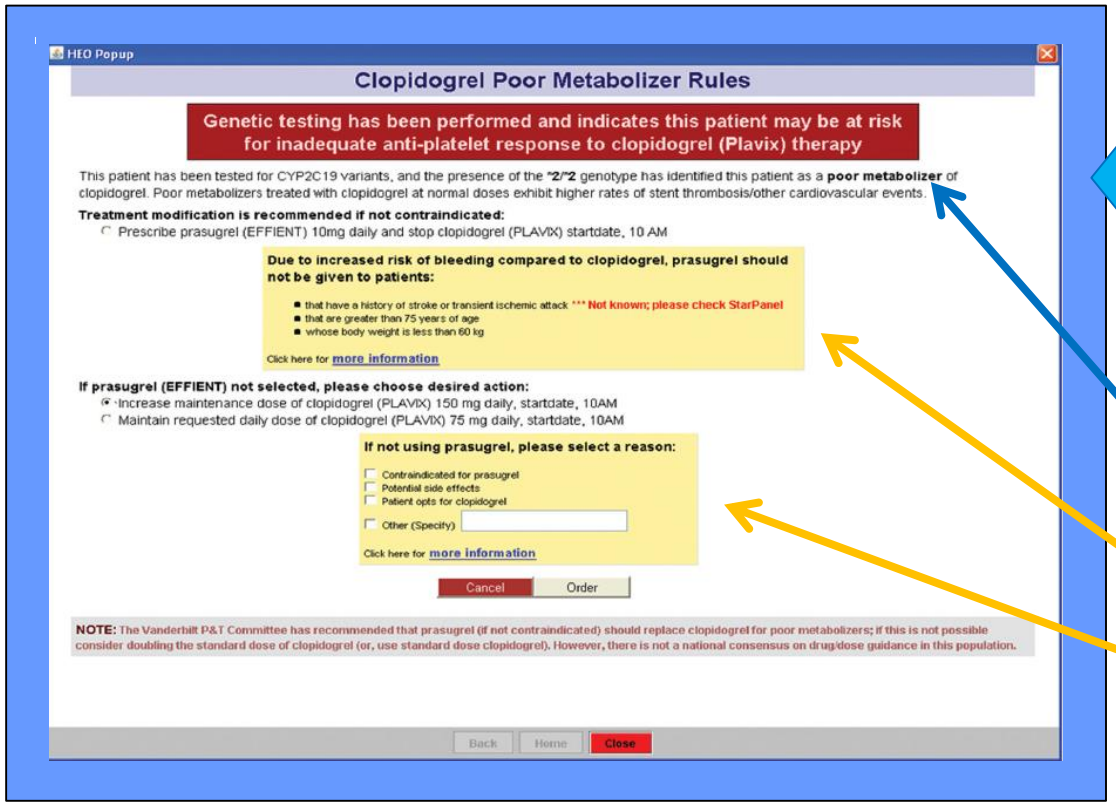
ゲノム医療の第2の流れ

薬剤代謝酵素多型性のゲノム医療
バンダービルト大学病院

■PREDICTプロジェクト

Pharmacogenomic **R**esource for
Enhanced
Decisions **i**n **C**are and **T**reatment

34項目の薬剤代謝酵素CYP多型性判定Chip
医師の処方オーダ時に警告提示（2010から）



クロピドグレル処方
電子カルテの警告画面
商品名プラビックス：抗血栓剤
ステント留置手術の後に処方

CYP2C19の多型性で*2*2の場合は
代謝機能が低いので(poor metabolizer)
血栓が凝固する
薬剤投与の応答は不十分である

この患者の場合(*2/*2)プラスグレル
(商品名エフィエント)に替えるか

分量を2倍にしろと警告している

TMDU

ゲノム医療の第３の流れ

著名ながんセンター
Dana Faber /MD Andersonなど

CSにより難治性がんのドライバー変異の同定する

組織限局的な後天的ゲノム変異のクリニカル配列解析
がんゲノムアトラス（TCGA：2006年～）および
国際がんゲノムコンソーシアム（ICGC：2008年～）
　50種のがんを500症例の全ゲノム配列解析
　2012頃から成果発表と始まった(我が国も肝臓がん)
患者個人70余の変異、全集合で3000を超える変異
がんを推進させるDriver変異と偶発的なPassenger変異

Mayo Clinic

- **全患者に全ゲノム配列解析：10万人患者（診療圏）データベース構築**
- **先制的 ゲノム薬理学（Preemptive PGx) 検査の初期の実施**
- **特別に診断する“診断オデッセイ”：Clinical Sequencing　原因不明遺伝病**

TMDU

ゲノム・オミックス医療の３つの流れ

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

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

ゲノム多型性の認識
.Hapmap2002開始
GWAS研究の興隆

薬剤代謝酵素多型性
電子カルテで警告
Preemptive PGx
Vanderbilt大病院

TCGA (2006),国際がんコンソーシアムICCG(2008) の成果2011から出現

Cancer Driver Geneの同定と抗がん剤治験
Mayo Clinic

2008年
2009年
2010年
2011年
2012年
2013年

ゲノム・オミックス医療
臨床実装(clinical implementation)

TMDU

ゲノム/オミックス医療ー米国の状況

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

- 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
  - PGx
  - がんおよび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

NHGRI Working Groupのリスト

Institution	Major Projects
MC Wisconsin	Using whole genome sequencing to establish diagnosis in patients with currently undiagnosed genetic disorders
Mount Sinai	• 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	• 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	• 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 • SLCO1B1*5 genotyping and statin adherence • Effect of genetic risk info on anxiety and adherence in T2DM

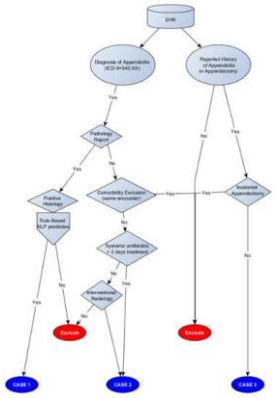
Institution	Major Projects
Alabama	Planning stages for projects in risk assessment, pharmacogenetic analysis, identification of families for further research
Baylor	Whole exome and whole genome sequencing in Mendelian disorders to improve diagnosis
Geisinger	• Selection for gastric bypass surgery vs 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	• 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	• PGx driven selection/dosing of antidepressants • CYP2C19 genotyping for antiplatelet rx post PCI
Inter-Mountain	Tumor-based screening for Lynch syndrome

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。まだゲノム医療の臨床実装は始まらず
  - 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) **ゲノム医療の臨床実装**へ舵を切る
  - MCWの臨床実装のインパクト、Vanderbiltの先制ゲノム薬理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

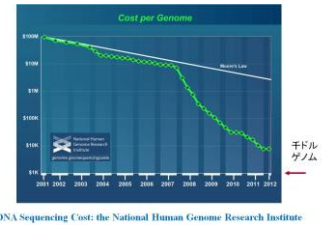
第１段階 ゲノム医療の発展

次世代シーケンシングの臨床普及(2010～)

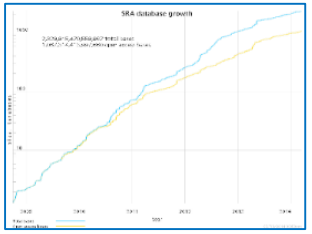
全ゲノム（x30：100Gb)・エキソーム解析（x100：6Gb）

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

電子カルテから臨床フェノタイプeMerge計画

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

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**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**としての役割が認識され、**世界的に普及**
- **疾患型BioBank（疾患ゲノムコホート）：ゲノム/オミックス個別化医療、創薬**の情報基盤：
  - 大規模調査にて、**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**
- **Population型BioBank（健常者ゲノムコホート）：個別化予防**の情報基盤：
  - 「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
  - 主に遺伝子要因情報も含めた**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。
- **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**（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）
- **米国　Precision Medicine Initiative Genome Cohort**
  - 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,1000genome,ICGC,TCGA等
  - ゲノム変異・多様体
    - dbSNP, HGMD,**Clinvar**, Clingen, OMIM,GWAS catalog, Matchmaker Exchange
    - 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

医療ビッグデータ：まとめ

- 臨床ゲノム・オミックス医療実装の進展
  - 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,疾患レジストリの拡充
  - 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の大規模化と利用
  - Clinvar, LINCS, HPRD,STRING

ビッグデータ創薬/DR

創薬を巡る状況

- 医薬品の開発費の増大
  - １医薬品を上市するのに約700億円
- 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相損耗)**
- 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
  - **ヒトのビッグデータを使う**

15~17年

		化合物数（自社）	前の段階から移行できる確率	累積成功率
2~3年	合成(抽出)化合物数	652,336		
3~5年	前臨床試験開始決定数	203		1: 3,213
3~7年	臨床試験開始数	75	1: 2.71	1: 8,698
1~2年	承認申請	26	1: 2.88	1: 25,090
	承認取得	21	1: 1.24	1: 31,064

製薬協データ2011

製薬協ガイド2014

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**

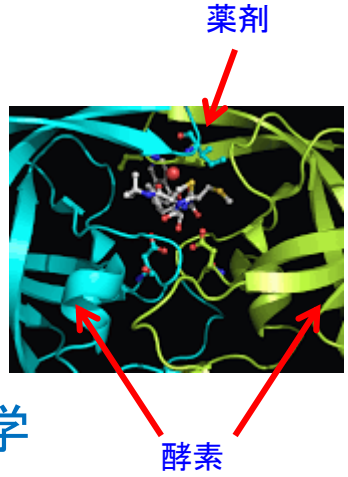
（3）**DR候補の探索**に**疾患生命情報ビッグデータ知識DB**を使用する

TMDU

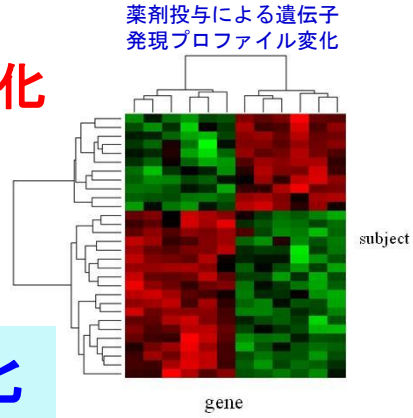
ビッグデータ創薬/DR
計算論的(*in silico*)創薬の新しい方向

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（結合ポケット）
- リガンドの分子設計・分子計算ー分子力学・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.profiling* drug design)
- 「オミックス創薬/DR, ゲノムワイド計算創薬/DR」
- **網羅的分子プロファイル・分子ネットワーク全体変化**
- **薬剤ー標的分子の結合**を取り巻く**genome-wide**な**分子環境**,「生命システムの変化」という理解
- 化合物, 標的分子, 疾患間の関係の「ビッグデータ」



まずDRでのビッグデータ戦略から創薬へ一般化

ビッグデータを医薬に生かす

創薬/DR　ビッグ網羅的分子の知識・DBの利用

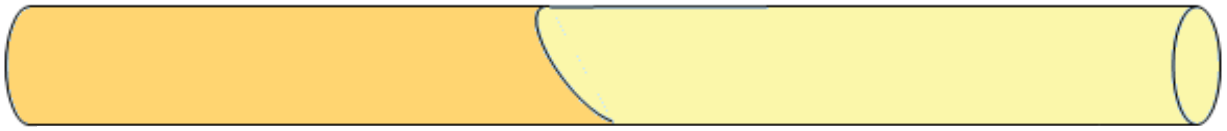
1．オミックス創薬/DR（遺伝子発現プロファイル）
2．疾患ネットワーク創薬/DR
3．Interactome 創薬/DR
4．多階層階層的ネットワーク創薬/DR

治験/調査　ビッグデータ疾患レジストリーの利用

1．疾患レジストリー準拠臨床無作為化治験
(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Biobank/Registryの利用



創薬・DR（研究開発）　治験・市販後調査（社会実装）

大規模網羅的分子情報の知識・DBの利用（創薬/DR）

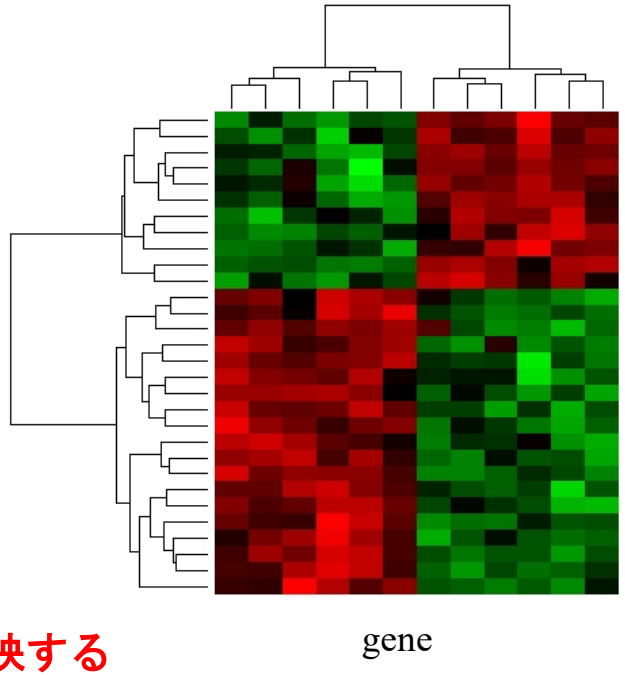
1. 発現プロファイル創薬/DR

遺伝子発現プロファイル空間を基礎にした
ビッグデータ創薬/DR
＜疾患ー薬剤＞プロファイル比較

TMDU

ビッグデータ創薬/DRの基本原理 1
発現プロファイル創薬原理

- **薬剤特異的遺伝子発現**（Drug-induced SDE)
  - CMAP(Connectivity Map)
    - 薬剤(化合物) 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**, 約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遺伝子群）
      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
    - 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

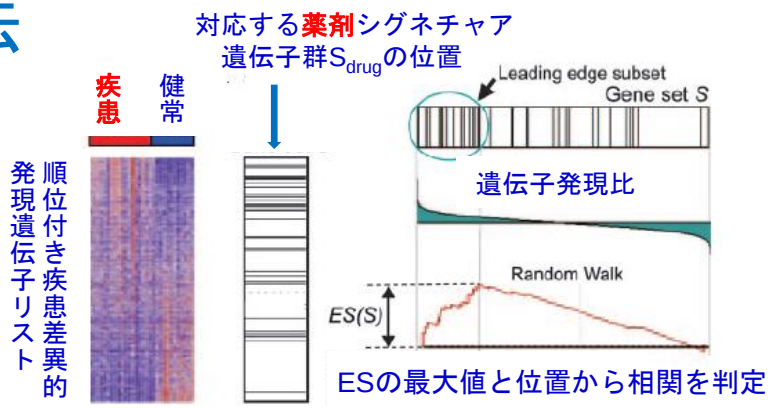


基礎には分子ネットワークの疾病/薬剤特異的変化
遺伝子発現プロファイル変化
≈　分子ネットワーク活性構造変化を反映する

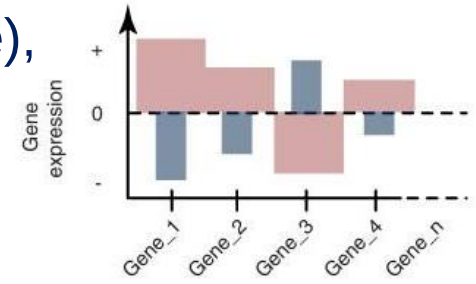
遺伝子発現プロファイルによる有効性予測

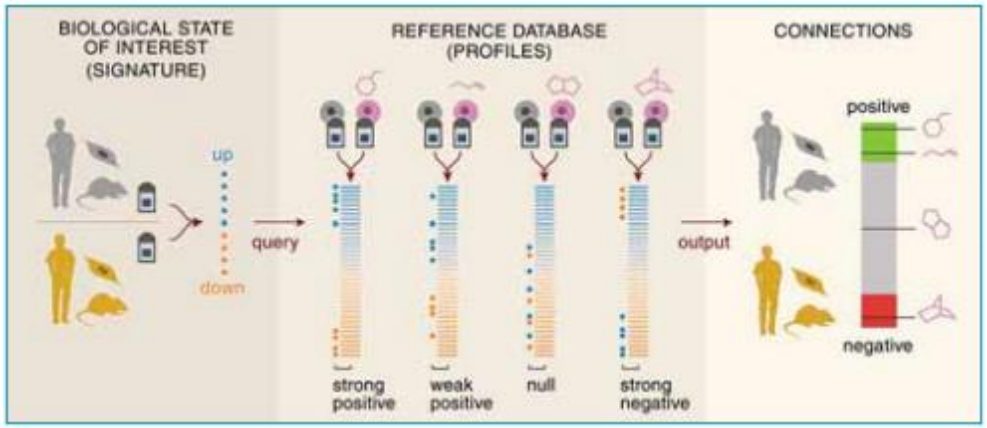
- 遺伝子発現シグネチャア逆位法 (signature reversion)
  - 薬剤特異的遺伝子発現シグネチャア
  - 疾患特異的遺伝子発現シグネチャア
  - 有効性予測：両者が負に相関する
  - Non-parametric な相関尺度で評価
    - Gene Set Enrichment Analysis（GSEA)：
    - 対照と比較して順位づけられた遺伝子リストの上位に密集しているかの尺度
  - 例：炎症性腸疾患IBDに 抗痙攣剤(topiramate), 骨格筋委縮にウルソール酸

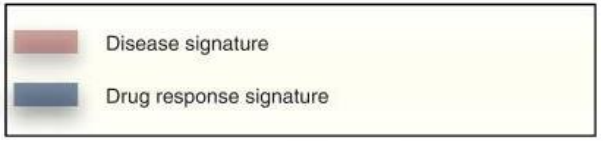
GSEA



発現比ランクの高い順から遺伝子を調べ
遺伝子リストS_{drug}中に**該当する遺伝子が存在したらES (Enrich Score)を加算**、無ければ減算



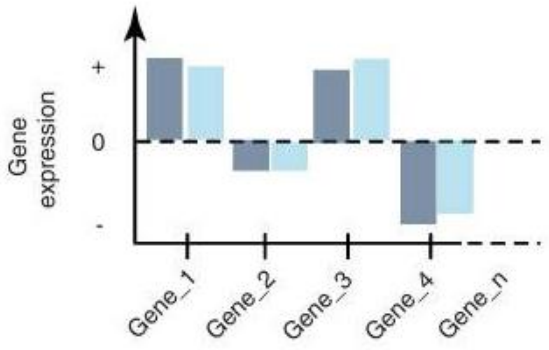




TMDU

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



発現プロファイル原理による
＜疾患－薬剤 Map＞に基づく計算DR

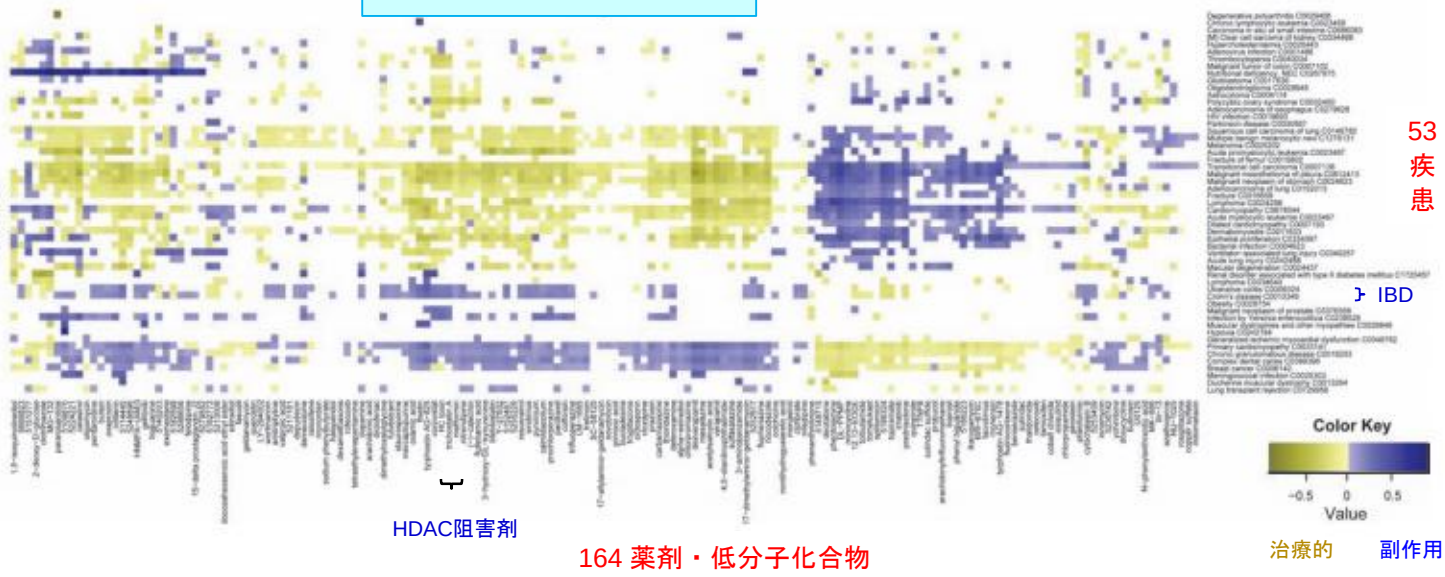
- NCBI・**GEO**から
  100疾患のシグネチャアを取得
- **c-Map**より得た**164の薬剤・化合物**
  の薬剤特異的遺伝子発現profile
  疾患-薬剤間で類似性スコアを計算
- 約16000組の疾患-薬剤間の2664組が
  有意、半数以上が治療的関連(負)あり
- 100疾患内, 53疾患有意に164薬剤と関連

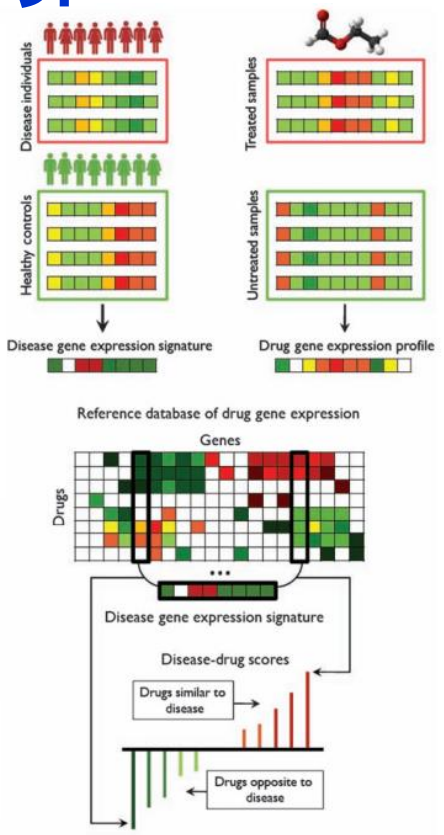
疾患-薬剤マップ

Table 1. Drugs and diseases with the most indications.

Drugs with most indications		Diseases with most indications	
Vorinostat	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)



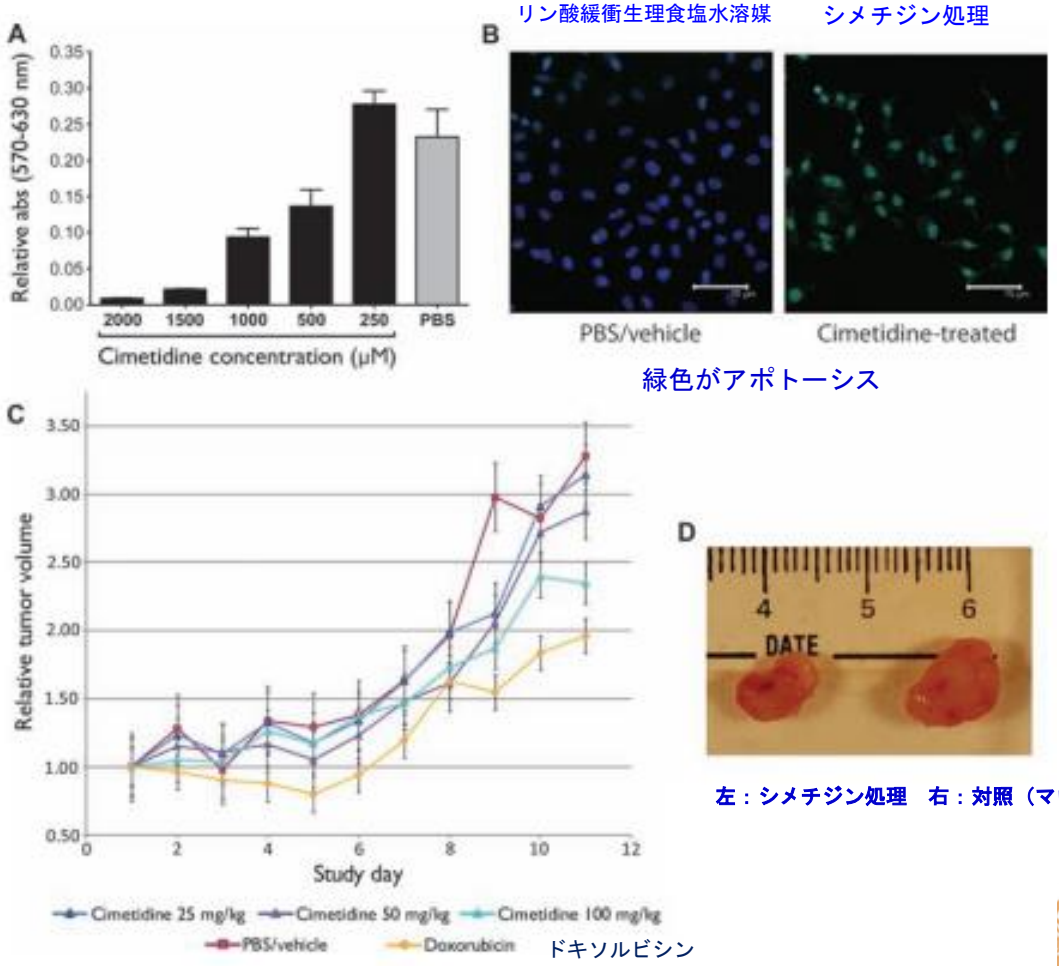


Drug group	Drugs
PI3K inhibitors	LY-294002 and wortmannin
HSP90 inhibitors	Geldanamycin, raloxifene, monorden, and sodium phenylbutyrate
HDAC inhibitors	Vorinostat, HC toxin, and trichostatin A
Salicylate anti-inflammatory agents	Sulfasalazine, mesalazine, and acetylsalicylic acid

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

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 の-0.075より高い



遺伝子発現Profiling による疾患ー薬剤ネットワーク（Hu, Agarwal）

遺伝子発現プロファイル(c-Map)での相関係数、ES指標によりネットワーク表示

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

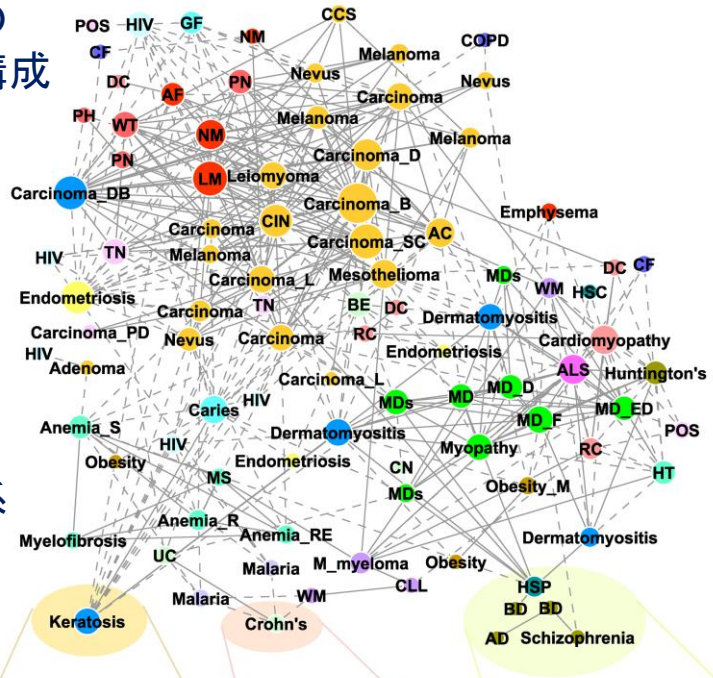
結果

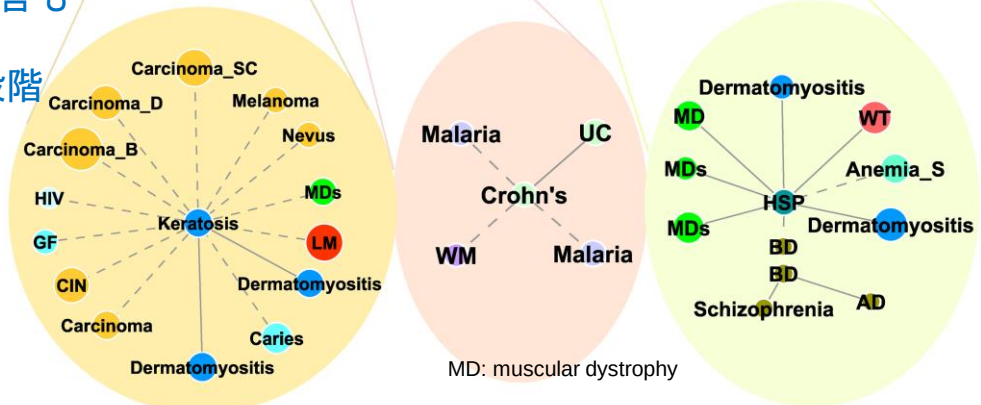
①疾患関連の60%はMeSH (既知体系）
その他は**分子レベル疾患分類学**
Transcriptomeの類似性による疾患体系

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 --精神障害も
Solar keratosis 日光性角化症
⇒ cancer(squamous) --前癌段階

＜疾患 - 薬＞
有効性：**マラリア治療薬**
⇒ Crohn’s disease
ハンチントン病に種々の薬剤





MD: muscular dystrophy

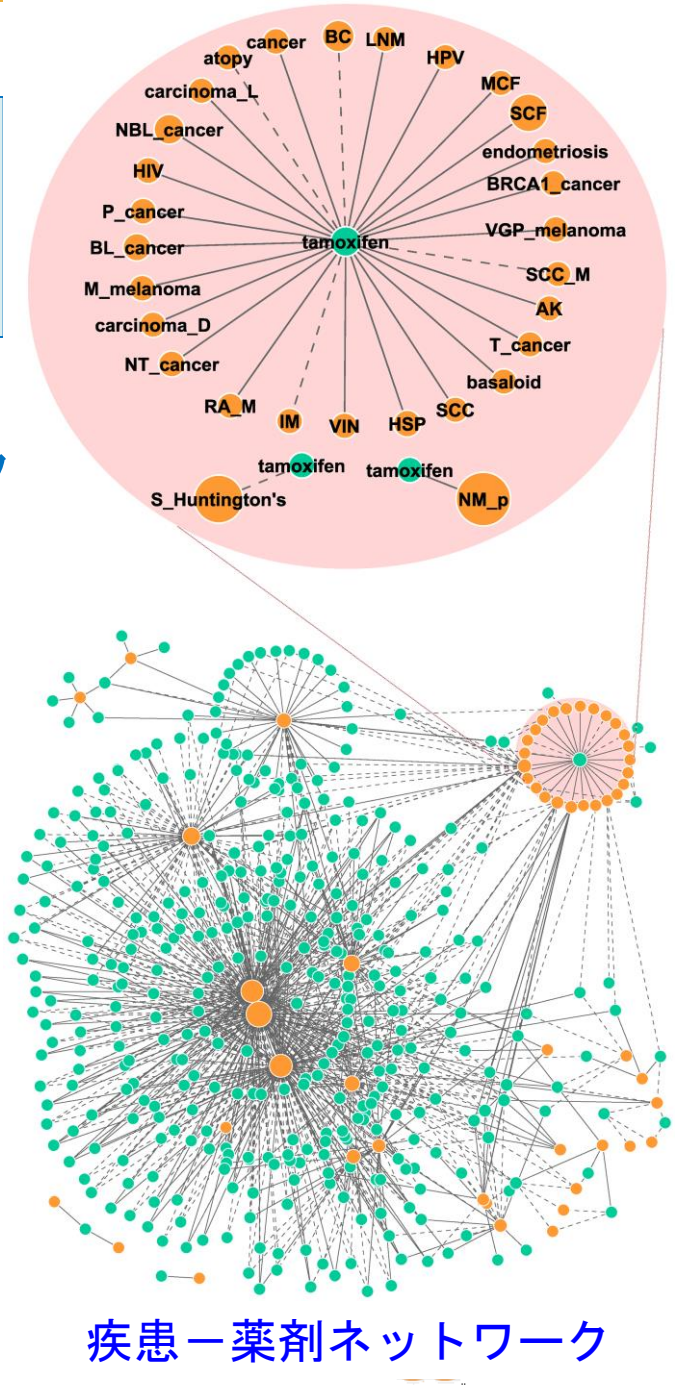
C2 virus disease
C3 parasitic diseases
C4 neoplasms
C5 musculoskeletal
C6 digestive
C7 stomatognathic
C8 respiratory tract
C10 nervous system
C12 male urogenital
C13 female urogenital
C14 cardiovascular
C15 hemic and lymphatic
C16 congenital
C17 skin/connective
C18 nutritional/metabolic
C19 endocrine system
C20 immune system
C23 pathological conds
F03 mental disorders

カラーはMeSH
同一カテゴリー
実線はMeSH内
破線はMeSH外

遺伝子発現 Profiling による Drug-Diseaseネットワーク

疾患－薬剤および薬剤－薬剤ネットワーク
（Disease-drug network:右図）
橙色節　49 疾患,　緑色節　213 薬剤
906　疾患－薬剤結合
実線　正値　破線　負値

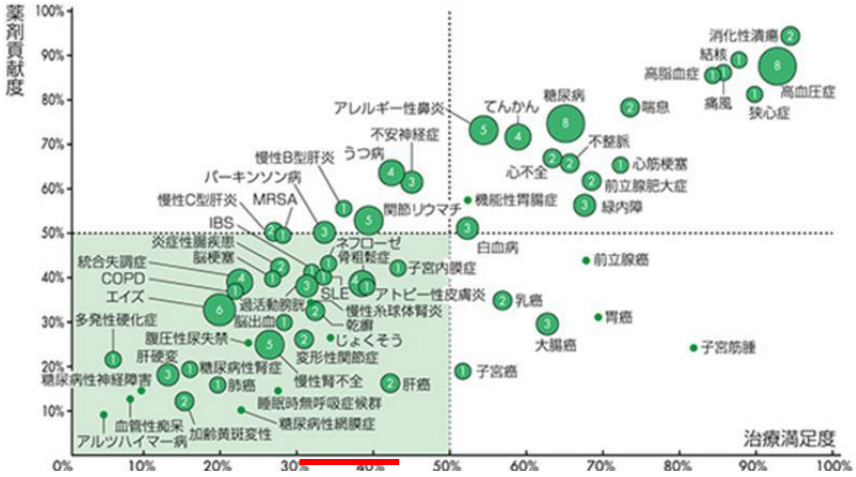
Tamoxifen (breast cancer)
有効性　負の値をもっている
⇒アトピー,
⇒マスト細胞分泌抑制、
アレルギー抑制
Hunting病に多数のDR薬
副作用　正の値をもっている
副作用の予測
⇒ 発癌性



疾患－薬剤ネットワーク

我々の研究室での成果

- 対象疾患　(Sibata et al. 2015)
  - 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), 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

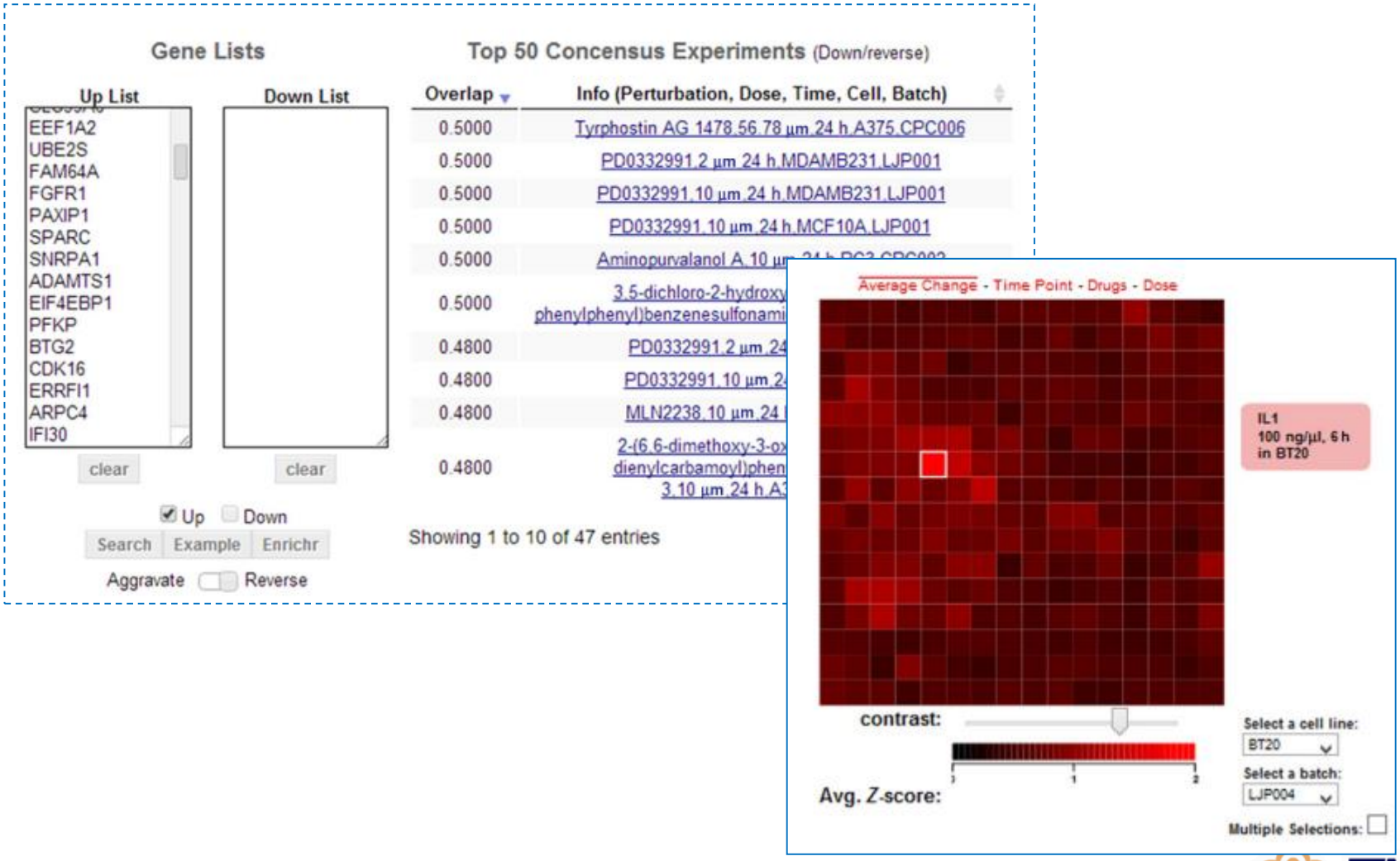
近年のビッグデータ化
LINCS

- **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
    - 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, **100万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 ---

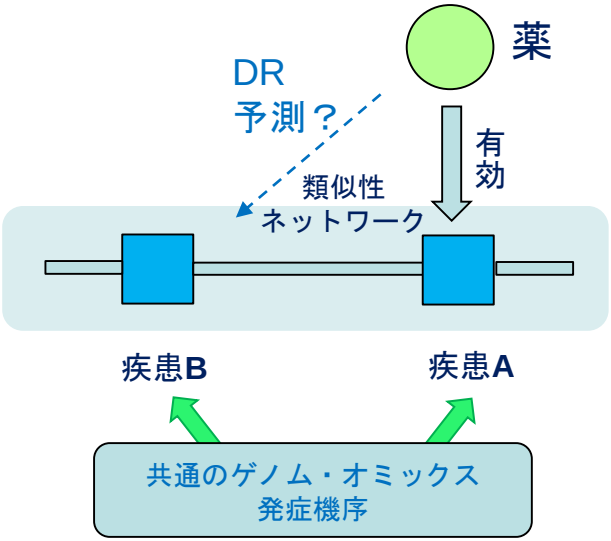


2. 疾患ネットワーク創薬/DR

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＞

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１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

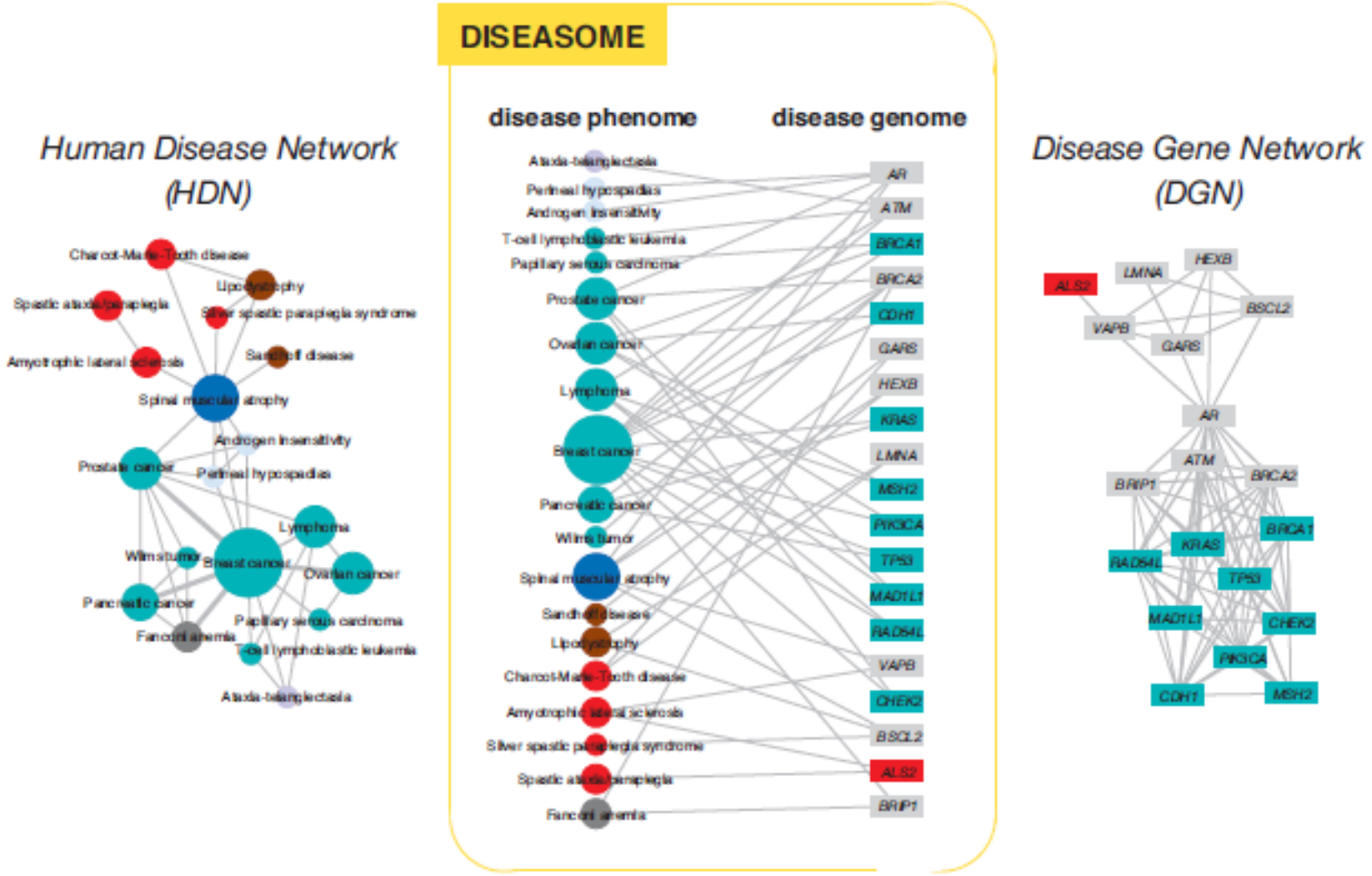
第１世代型
Diseasomeと疾患遺伝子

- **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

(Goh,Barabasi et al.)



１つ以上の疾患関連遺伝子を共有する疾患

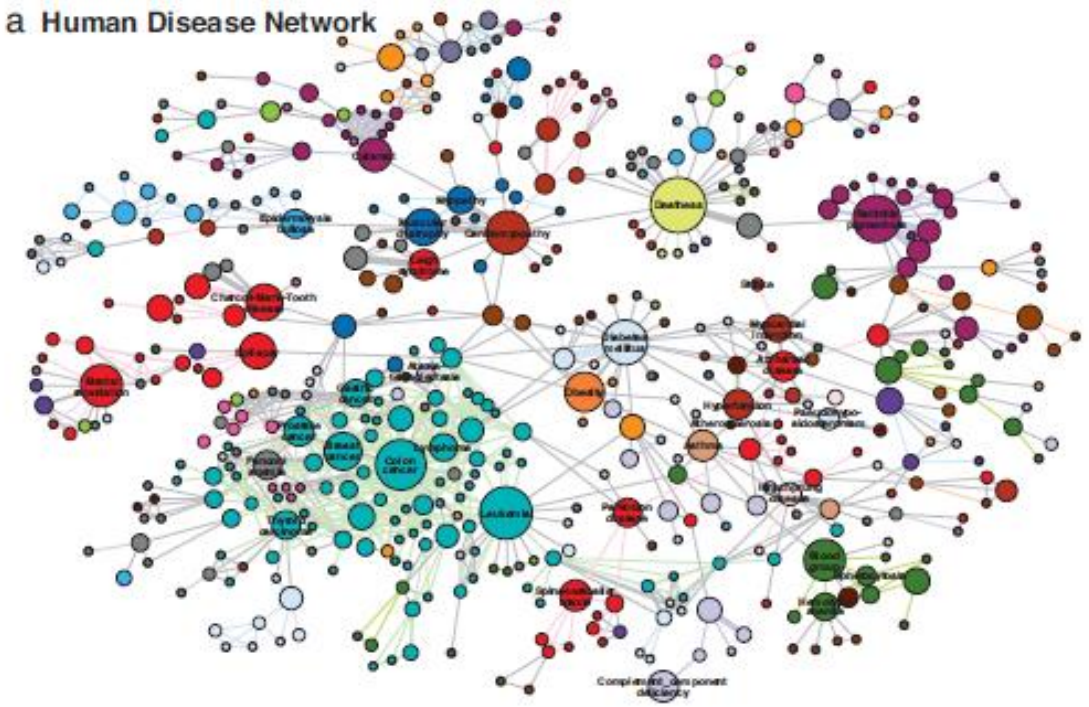
１つ以上の疾患を共有する疾患関連遺伝子

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

TMDU

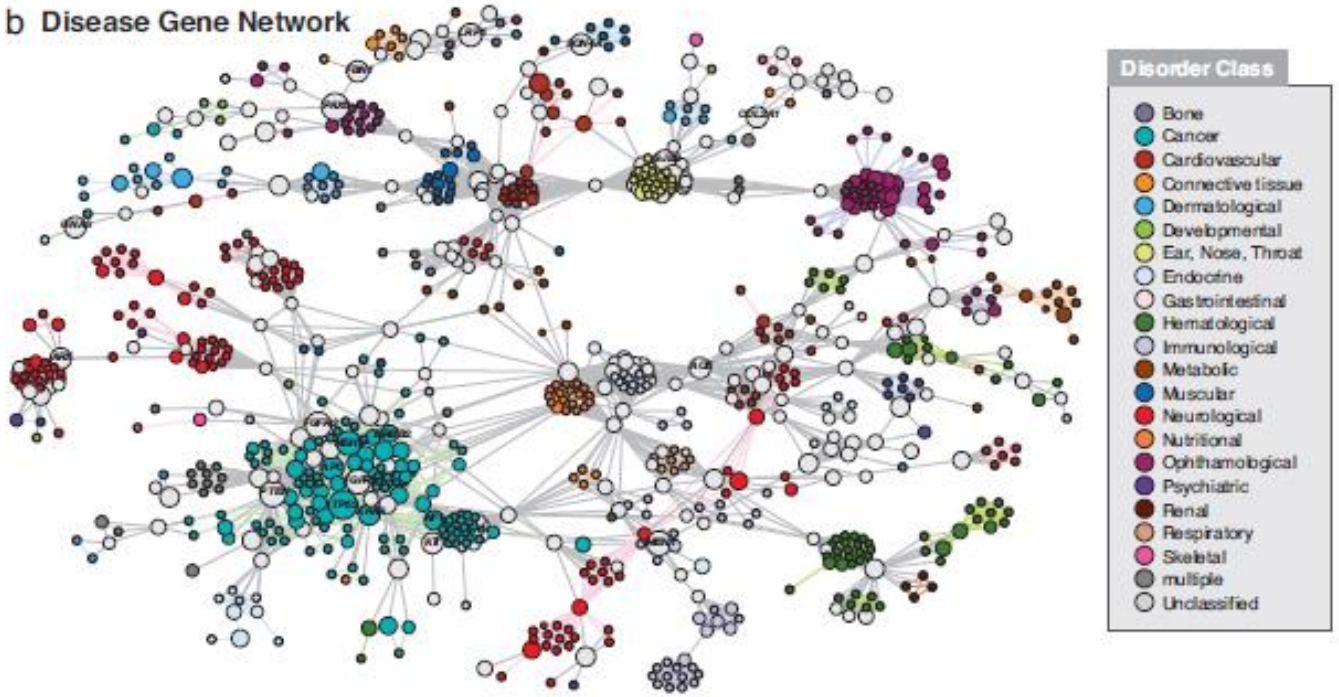
疾患
ネットワーク
(HDN)

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)

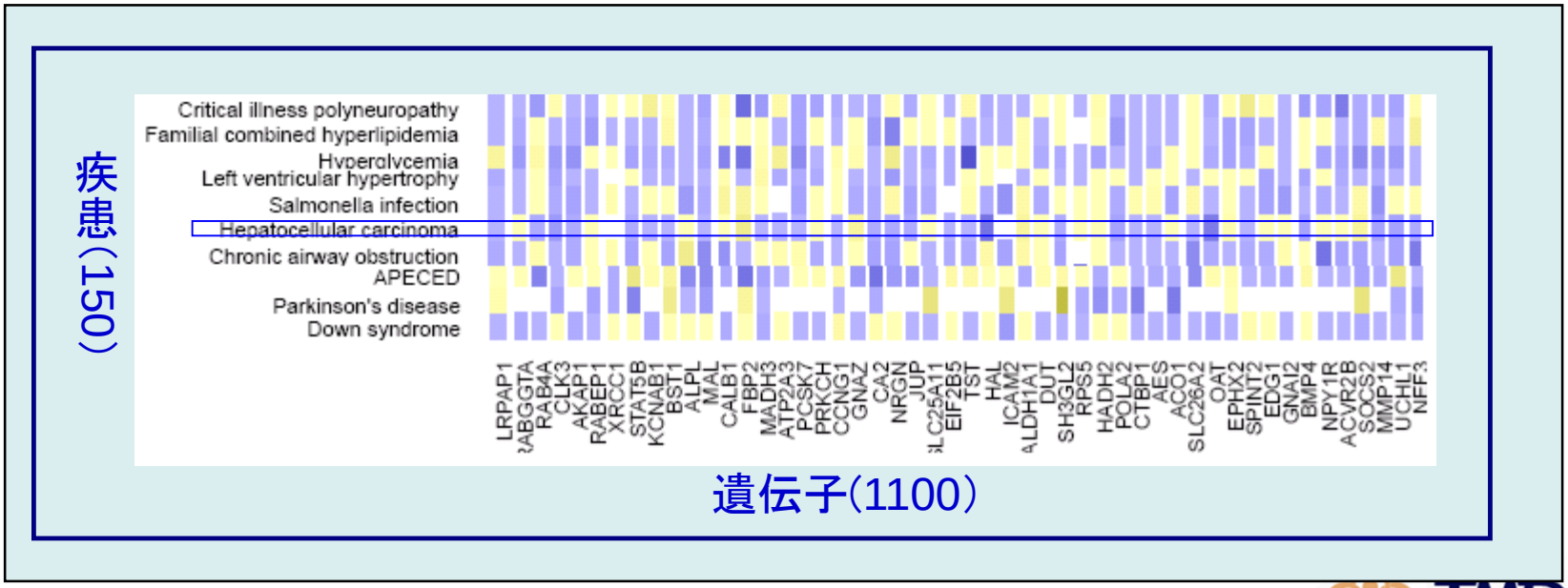
Nodeの径
その遺伝子を原因にしている疾患の数に比例
2つ以上の疾患に関与していると明灰色の遺伝子ノード



第２世代型

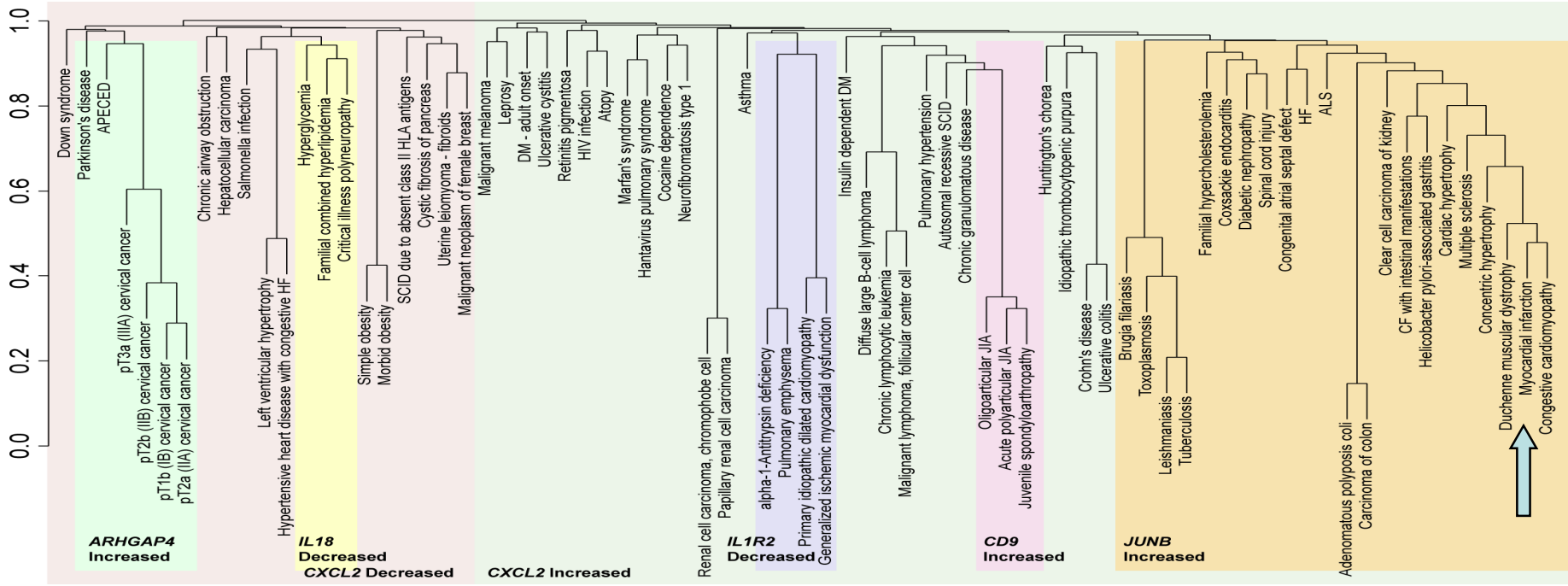
GENOMED (A.Butte et al)

- 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



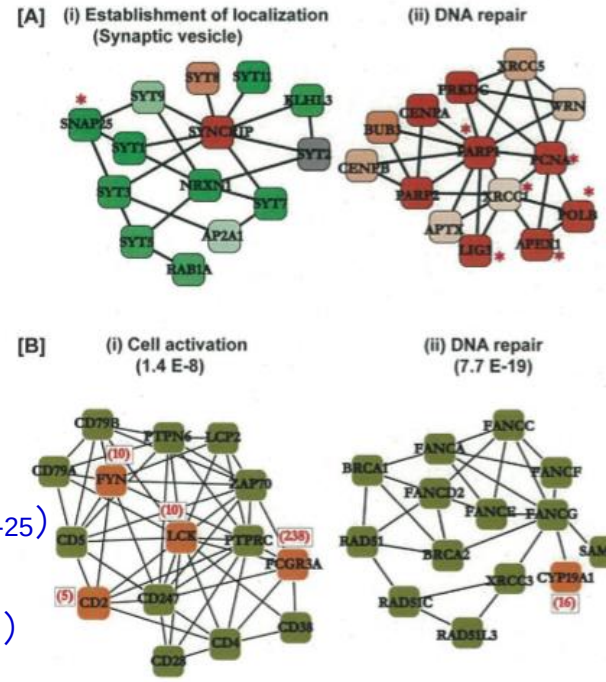
Gene-Expression Nosology of Medicine

- 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



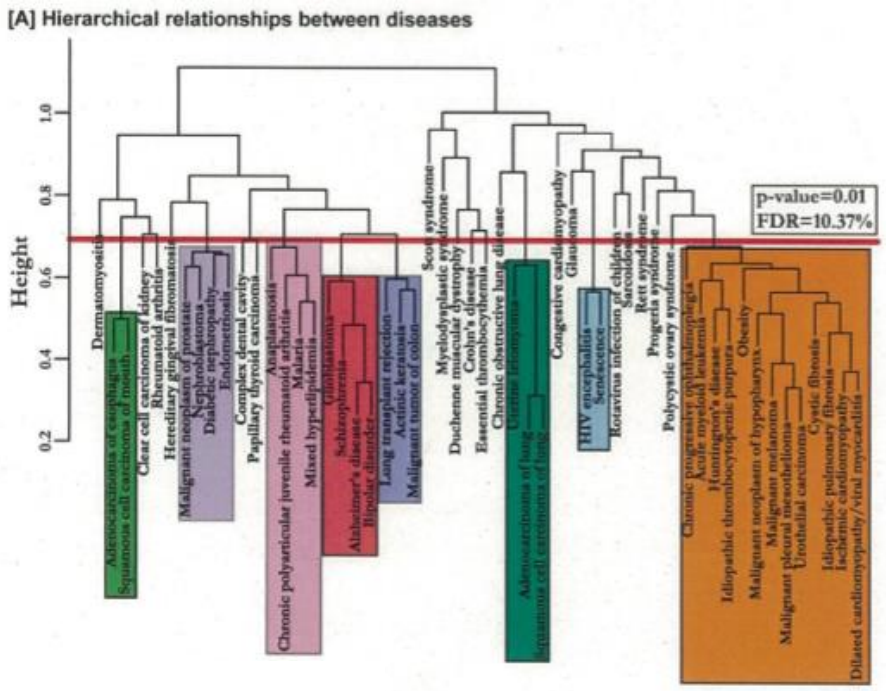
Transcriptomeの変化をPPIに投影した疾患ネットワーク（Butte)

- 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 に分解したタンパク質相互作用ネットワークの＜**機能moduleでの疾病時の平均発現変化**＞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の有意な類似性**：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pasM), 精神疾患(synapsM:SNAP-25)
- **138の有意な疾患相関**
  - 17は少なくとも1つの共通薬：14疾患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"疾患状態"モジュール」
  - 「共通疾患状態シグネチャア」薬剤標的分子に富んでいる
  - この遺伝子群を標的にする薬剤は有意に多くの他の疾患の薬剤にもなっている

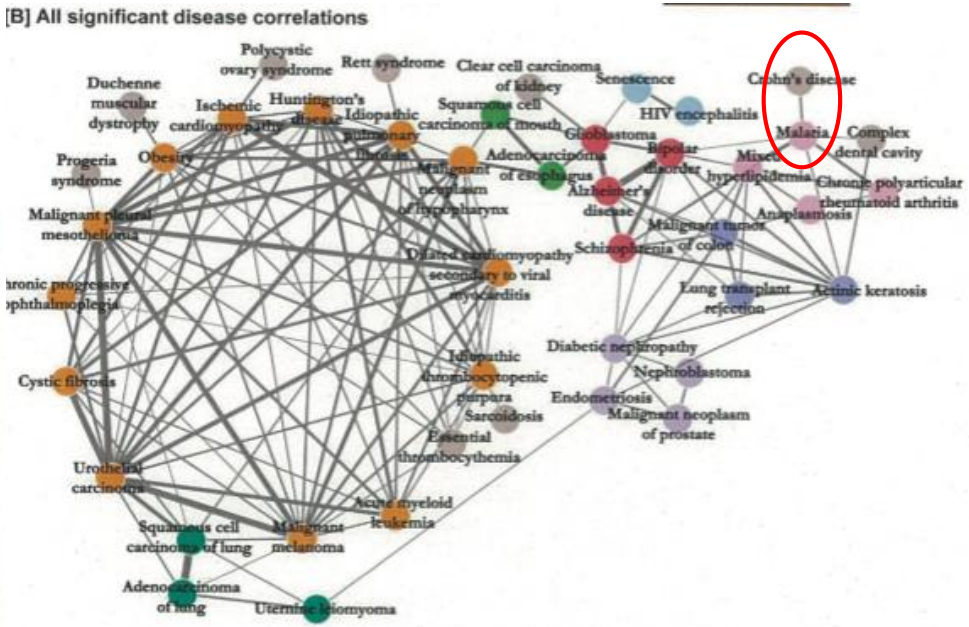


Transcriptomeの変化をPPIに投影した疾患ネットワーク（Butte)

- 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

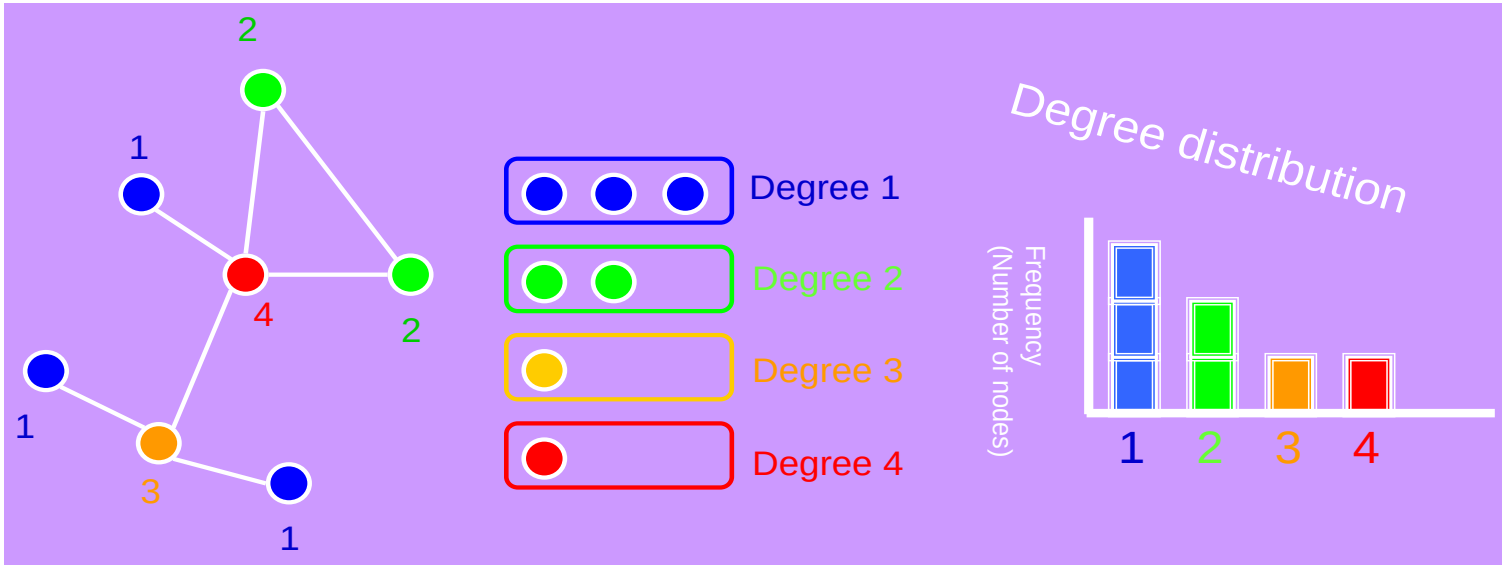
TMDU

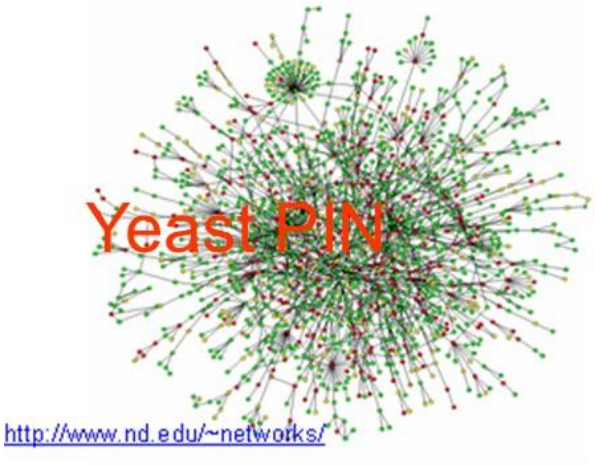
3. Interactome 創薬/DR

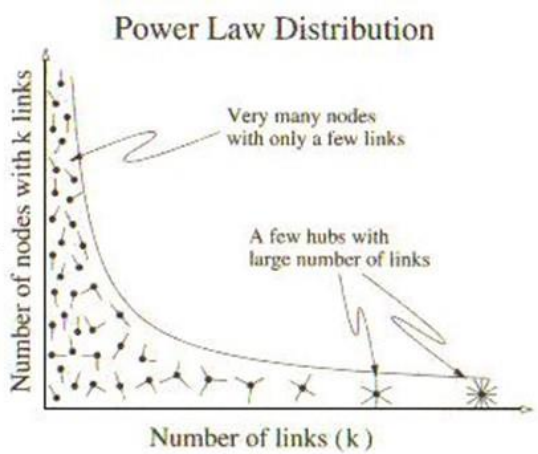
Protein間相互作用ネットワーク（PPIN）
を基礎にした
創薬/DR

＜薬剤標的分子と原因遺伝子のPPINでの距離＞

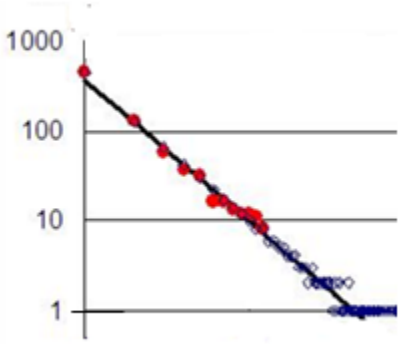
タンパク質相互作用ネットワーク(PIN)では数少ない相互作用が集中したタンパク質(hub)と相互作用が1や2の多数の末端タンパク質(branch)が存在する





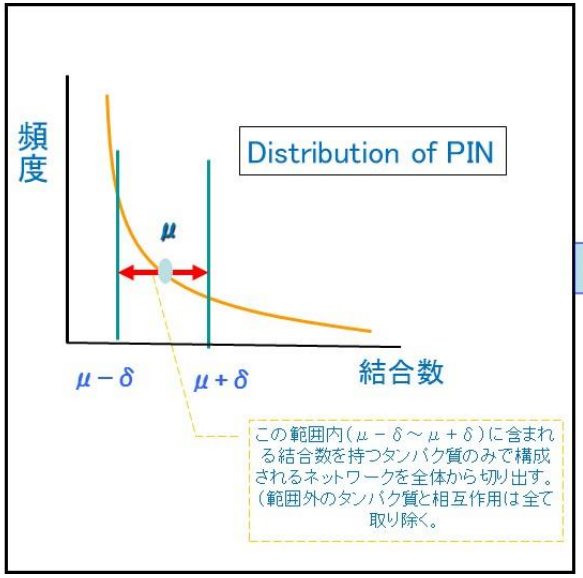


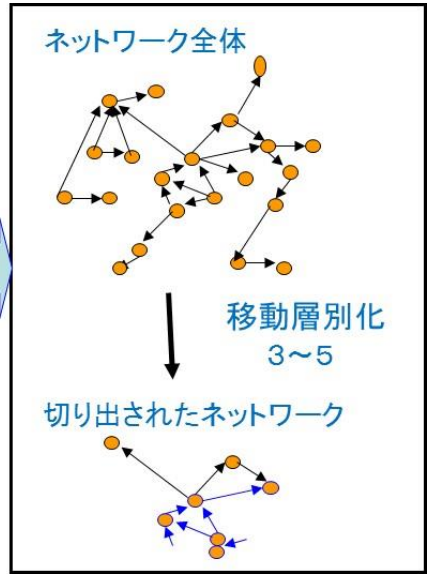
http://www.macs.hw.ac.uk/~pdw/topology/ScaleFree.html



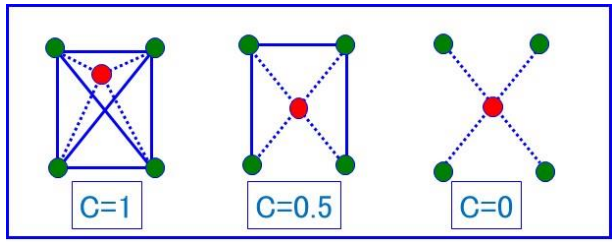
Log-log変換で直線

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



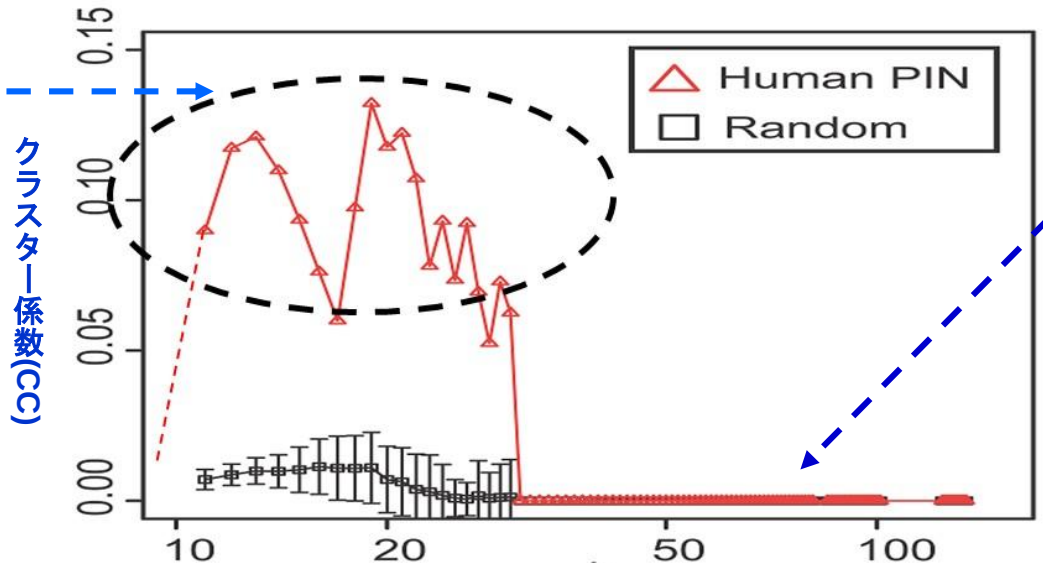


クラスター係数



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

中程度の結合数（7～42）を持つタンパク質は多数の密なモジュールを構成

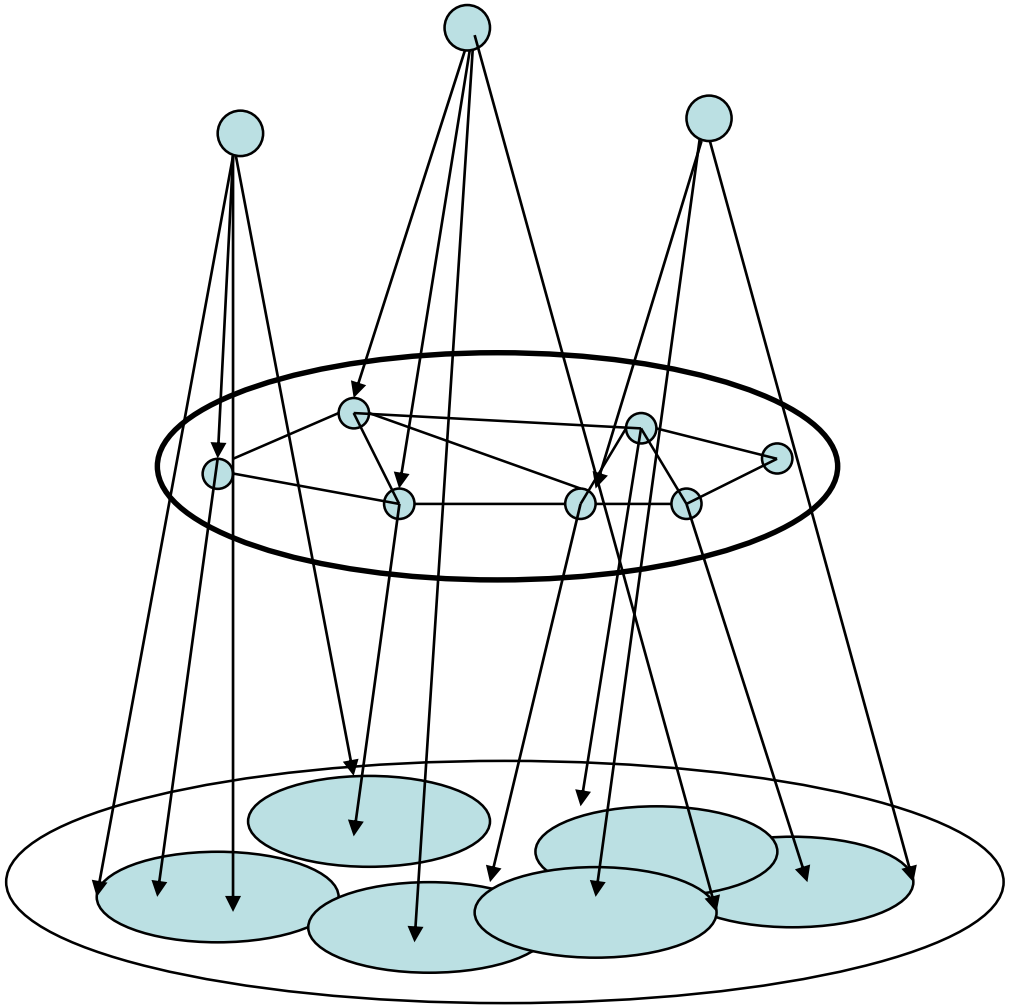


高い結合次数を持つノード（スーパーハブ）はお互いに密に結合しない

TMDU

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



高次層
高結合数分子
ハブ分子
度数 ＞31 ヒト
＞39 酵母

中間層
中結合数分子
バックボーン分子
度数 6 ～30 ヒト
6 ～38 酵母

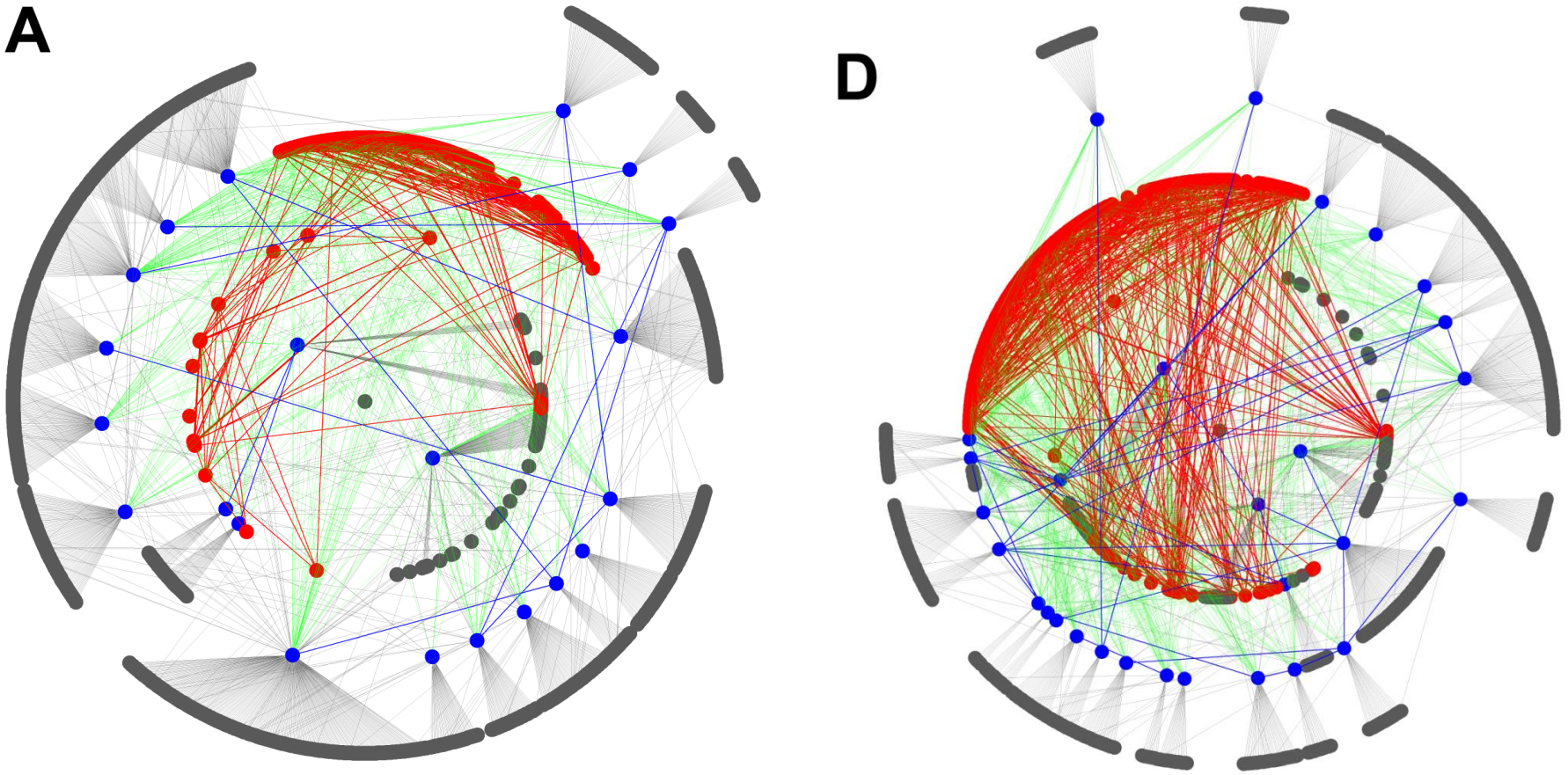
低次層
低結合数分子
ブランチ分子
度数 <6

TMDU

InteractomeのCloud Topology
（3環トポロジー）

Yeast

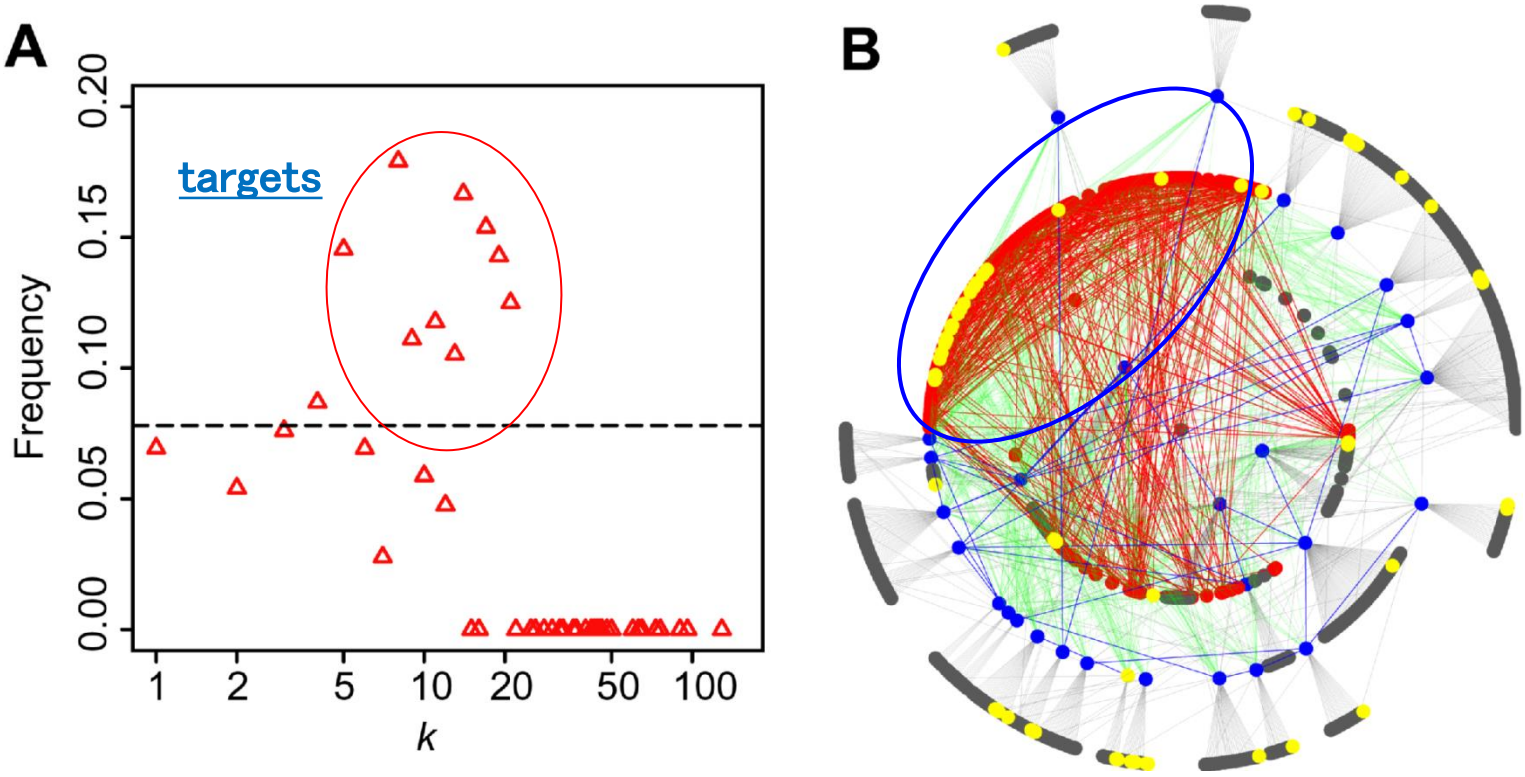
Human



Middle-degree ノードは PPI backbone を形成する.

灰色, 赤, 青は、それぞれ低層、中層、高層のdegreeのノードをそれぞれ表す。

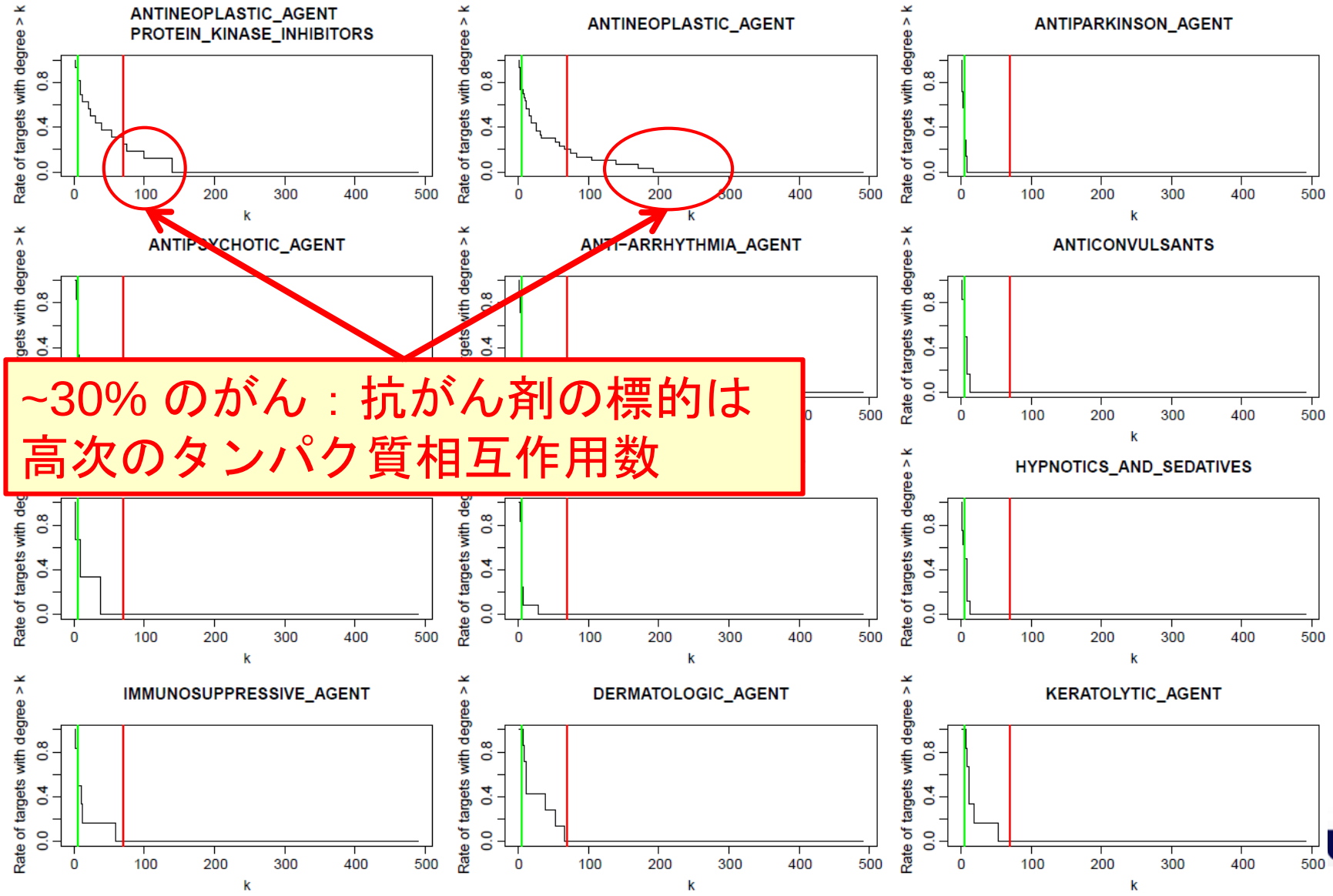
薬剤標的分子と結合度数



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

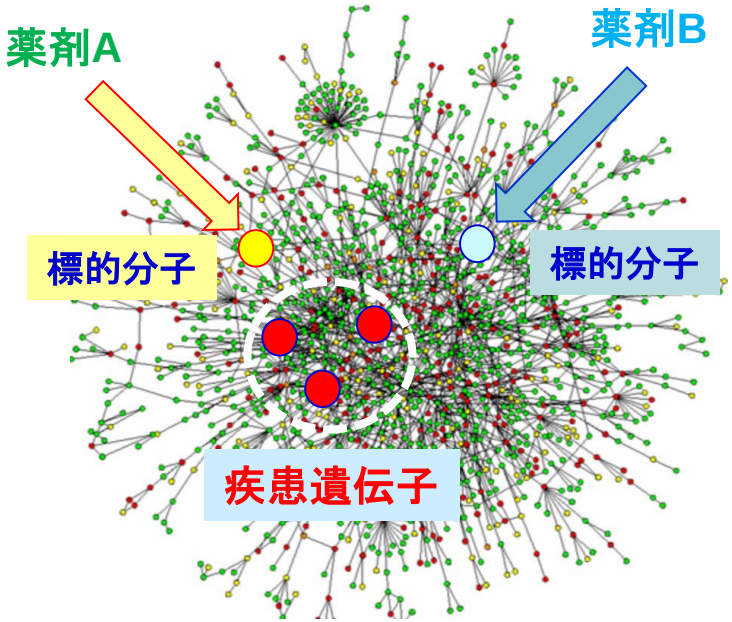
がんの疾患遺伝子は高次結合ハブのタンパク質が多い



~30% のがん：抗がん剤の標的は
高次のタンパク質相互作用数

Interactome（PPIN）創薬/DR

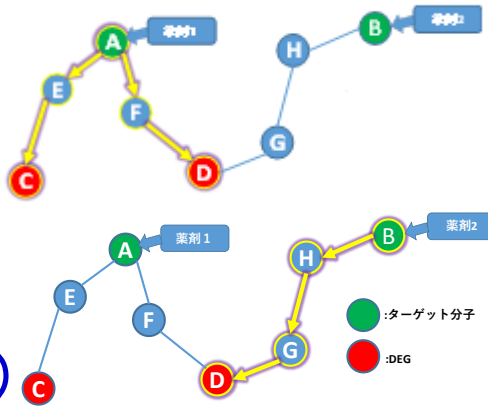
- **タンパク質相互作用ネットワーク（PPIN）空間での創薬/DR戦略**
- Interactomeのネットワーク場を基礎にして距離（類似性）を検討
- **薬　剤**：**薬剤の標的分子**（タンパク質）によって　PPI場と繋がる
- **疾　患**：**疾患特異的発現遺伝子**を**タンパク質**へ翻訳、PPI場と繋がる
- PPIN場内での**薬剤と疾患の「代理人(疾患遺伝子）」**の**距離・親近性**を基準に、薬理作用のインパクト力を評価――標的分子の概念が入る
- **random walking**で総合的な近さの評価を行う



タンパク質相互作用
ネットワーク
PPIN: interactome

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（Human Protein Reference Database）
    - 37,070 PPI, 9465 タンパク質
  - STRINGS（Search Tool for the Retrieval of INteracting Genes/proteins)
    - 184 M PPI, 9,643,763タンパク質 --- 個々に計算



- **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）
- **疾患遺伝子の差異的遺伝子発現データ（DEG）**
  - 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,FDR<0.01, **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

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

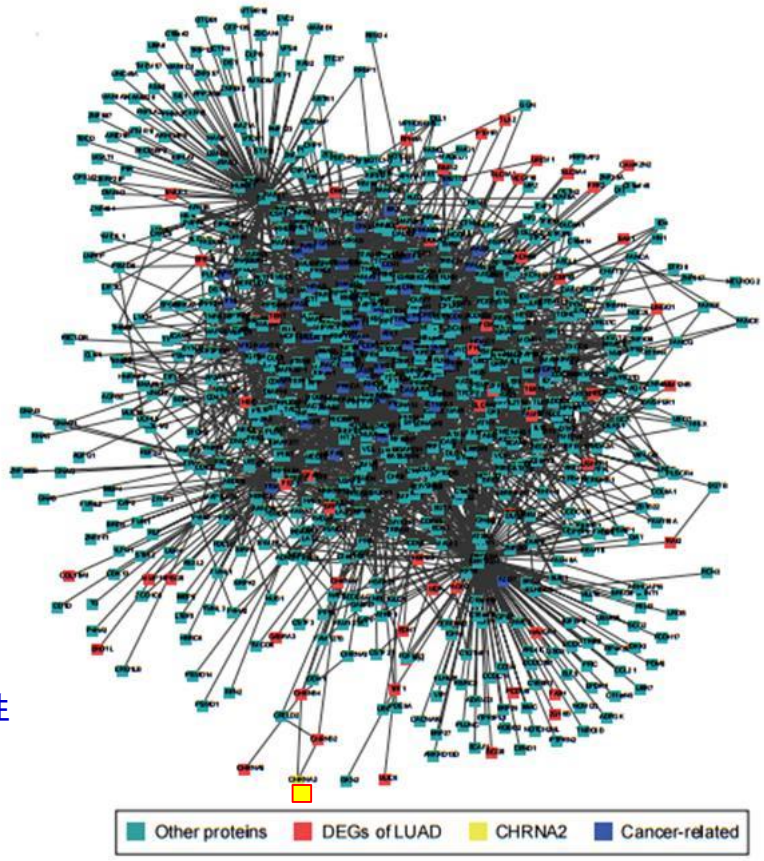
$\mathbf{P}^{s}$：時点sでの各ノードでの滞在確率　$\mathbf{M}$ ：各ノードへの遷移確率　γ：再出発確率

TMDU

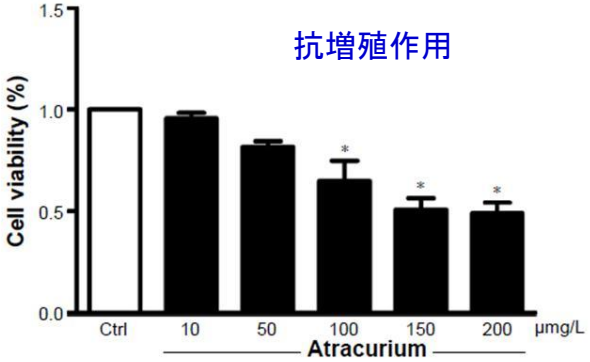
Interactome DR 結果の検証

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

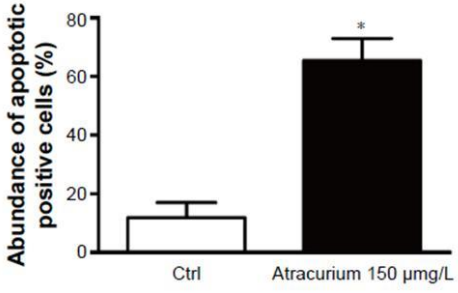
- **HPRD**と**STRINGS**の両方のPPINのランダム歩行でtop5%で共通な**145薬剤**を同定
- 最高スコアを挙げた**Atractium**を選択
- 薬剤標的は**CHRNA2**(Cholinergic Receptor Nicotinic Alpha 2）**でアポトーシス経路である**
- **培養細胞A549**（ヒト肺胞基底上皮腺癌）胞）の抗増殖作用を確認



抗増殖作用



アポトーシス活動性



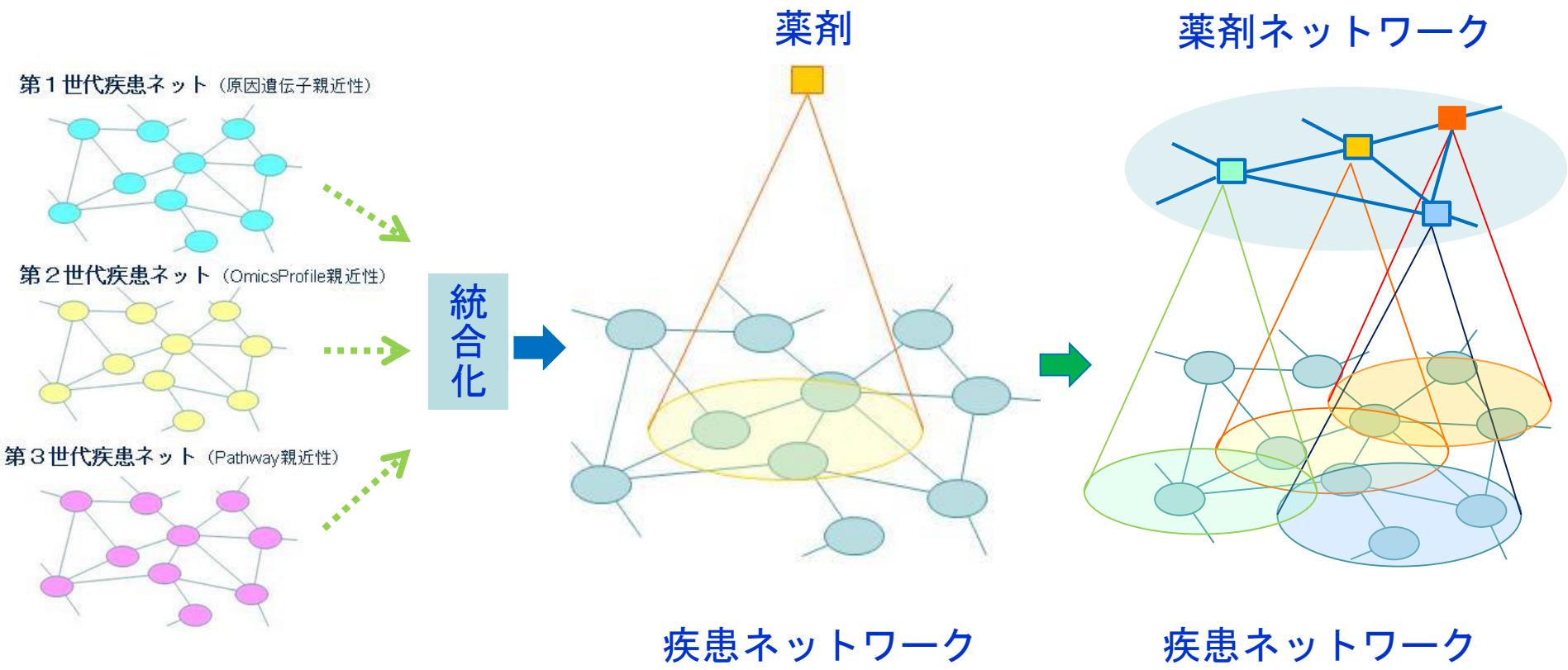
TMDU

＜疾患-薬剤-標的分子＞の多階層ネットワークによる

ビッグデータ創薬/DR

疾患ネットワークとDrug Projection Map

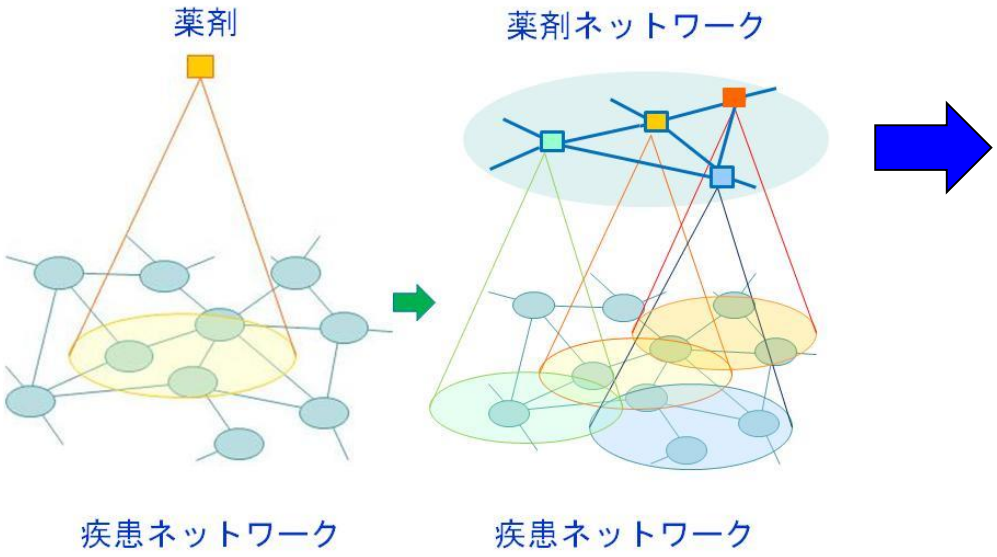
薬剤/疾患ネットワークの構築

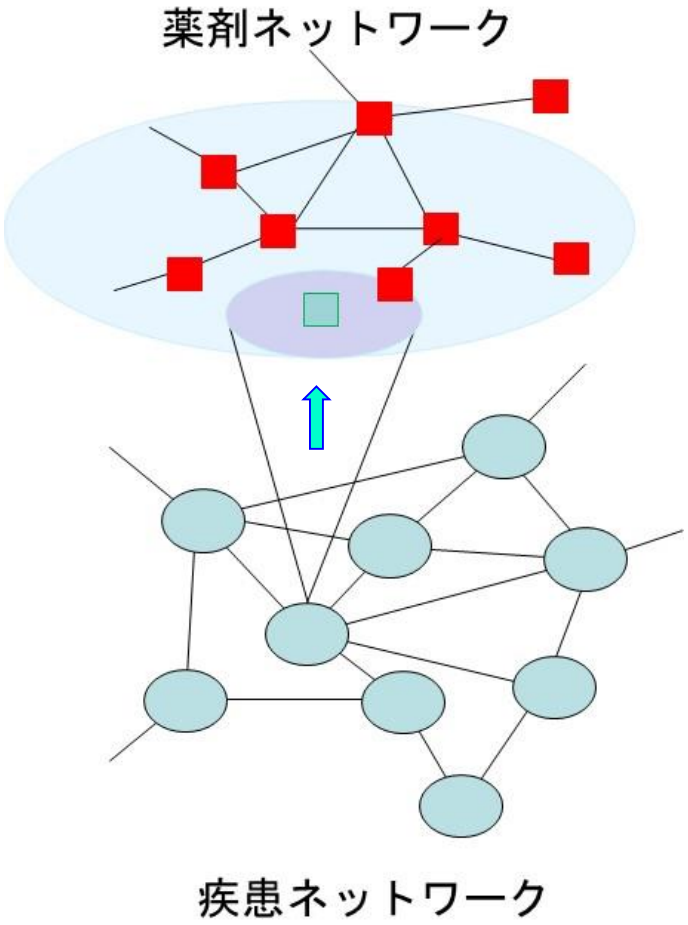


DRの方法論から創薬方法論へ

- 疾患ネットワークの十全な形成
  - 様々なオミックスの**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**有効性**は証明
- 創薬への発展
  - **薬剤・化合物階層のネットワーク**は既に確立
  - cMapでは不十分・LINCS(2014)出現
  - 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
- 疾患ネットワークと薬剤ネットワーク間写像
  - **Dual Network-based Drug Discovery**

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





Heterogeneous Network 創薬/DR

(Wang et al. 2014)

- **標的分子情報をDR計算に入れる**
  - 標的探索とDRを同時に進める
- **３層ネットワーク構成**
  - 疾患-疾患 (d_i),薬剤-薬剤(r_i),標的-標的 (t_i)
  - 疾患-薬剤間の距離を2つのレベル内での距離から計算
- 各ネットワークで距離定義
  - 疾患：表現型⇒MeSHの共通項数
  - 薬剤：化学構造⇒Tanimotoスコア
  - 標的：Protein配列類似性⇒Smith-Waterman法
  - 疾患-薬剤：過去研究、薬剤-標的ネット：Drugbankより
  - 失われた疾患-病気 Edge復元
- **結合係数W(*i*,*j*)更新法**

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

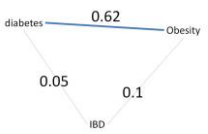
$$w(d,t)=\sum_{r_i\in R}\sum_{r_j\in R} w(d,r_i)\times w(r_i,r_j)\times w(r_j,t)$$

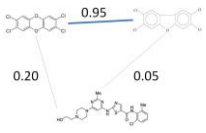
$$w(d,r)=\sum_{t_i\in T}\sum_{t_j\in T} w(d,t_i)\times w(t_i,t_j)\times w(t_j,r)$$

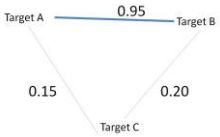
結合係数更新のマトリックス表示

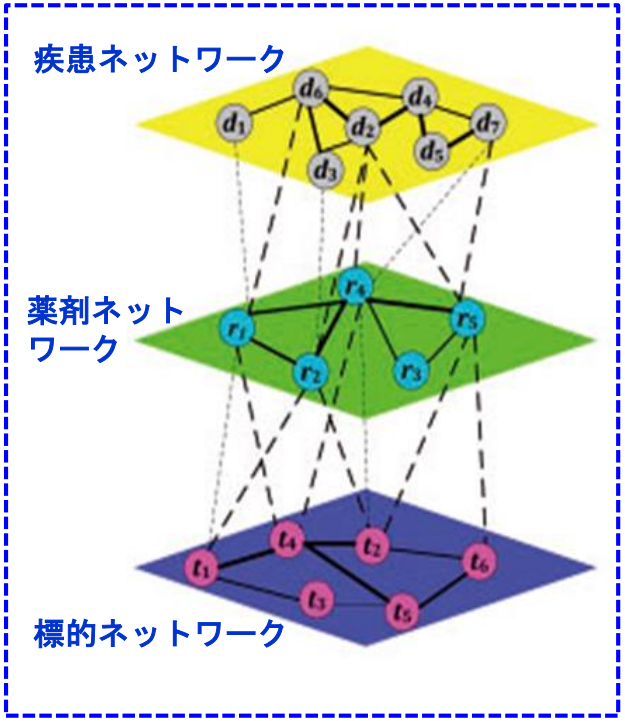
$$W_{dr}^{k+1}=\alpha W_{dr}^{k}\times(W_{rr}\times W_{rt}^{k}\times W_{tt}\times W_{rt}^{kT})+(1-\alpha)W_{dr}^{0}$$

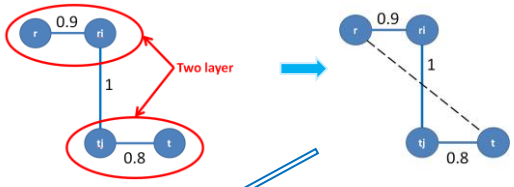
$$W_{rt}^{k+1}=\alpha(W_{dr}^{kT}\times W_{dd}\times W_{dr}^{k}\times W_{rr})\times W_{rt}^{k}+(1-\alpha)W_{rt}^{0}$$

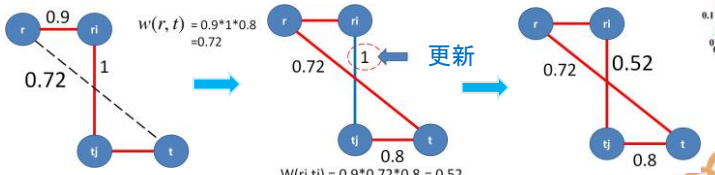


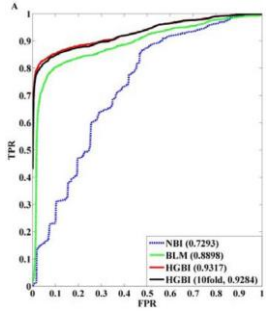












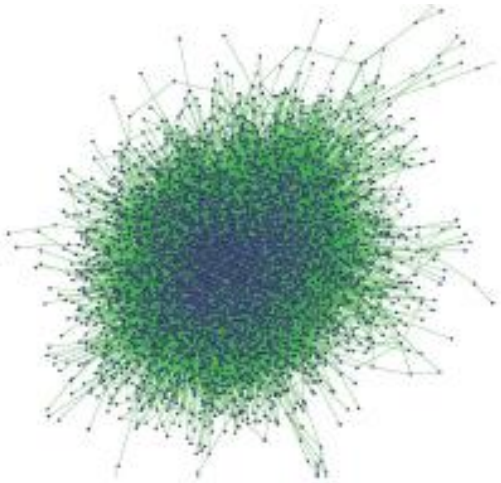
従来の方法より
DR推定精度高い
ROC曲線

我々の創薬/DRの研究
network-guided interactome DR

機械学習とネットワーク解析を組み合わせたDR(drug repositioning）

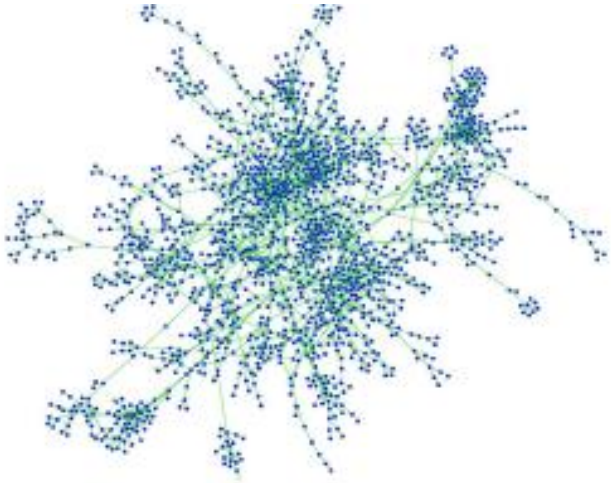
- **Step 1：ネットワーク解析を行い、機械学習モデルを生成するための特徴量を抽出する。**
- **Step 2：機械学習を用いて対象疾患（疾患Ａ）に対する新規標的分子を予測する**
- **step 3：予測された新規標的分子をベースにして、この疾患Ａに対する新しいrepositionableなdrugを推測する。**

使用したデータベース

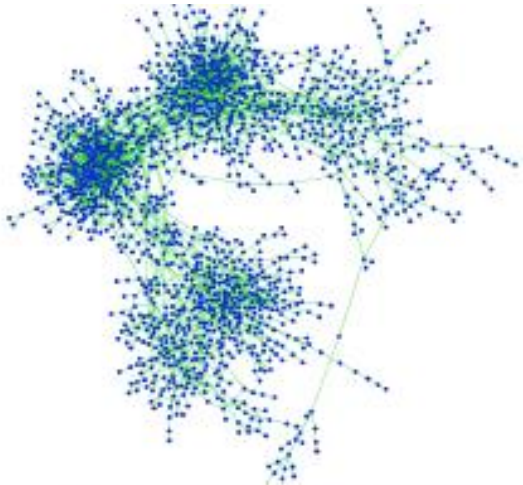


Directed protein-protein interaction (signaling) network

Vinayagam A et al. (2011) A directed protein interaction network for investigating intracellular signal transduction. *Sci Signal* 4(189):rs8



Synthetic dosage lethal interaction network



Synthetic lethal interaction network

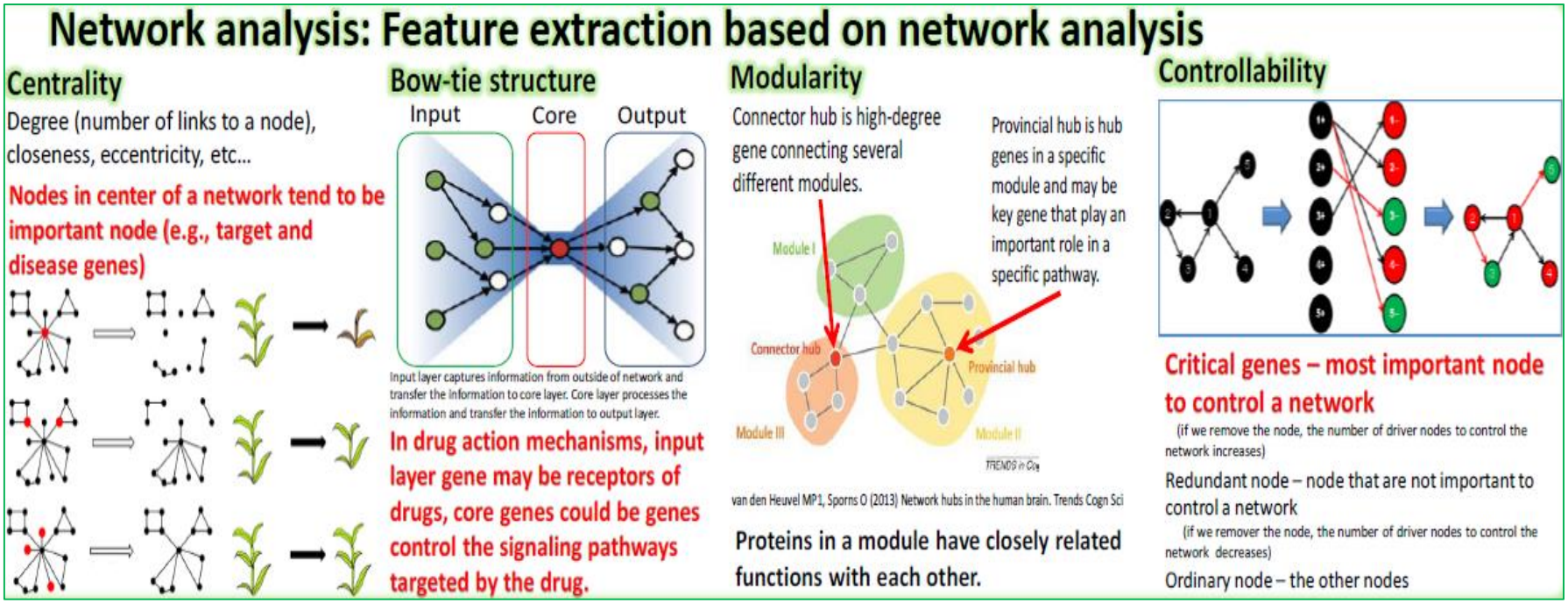
Jerby-Arnon L et al (2014) Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. *Cell* 2014 158(5):1199-209

３種のネットワークと、薬剤と標的分子の情報を、訓練データ・試験データの構築に用いた。

TMDU

Step 1：Network解析による各遺伝子の特徴量の抽出

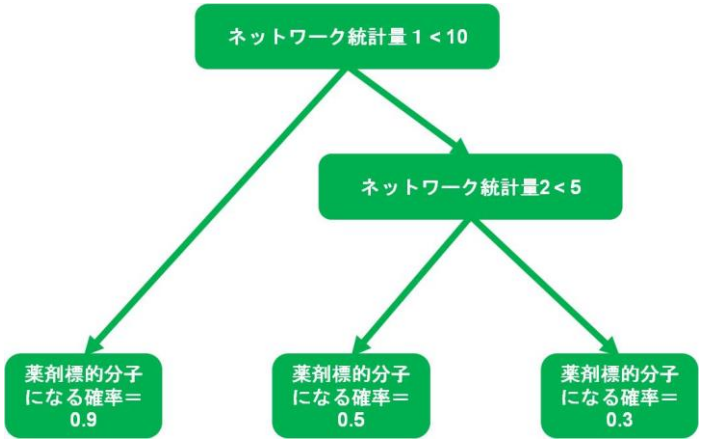
21種のネットワーク統計量（下図）を、各遺伝子に対して、各ネットワークを解析して算出した合計63種のネットワーク統計量を各遺伝子の特徴量として予測モデルの構築に用いた。



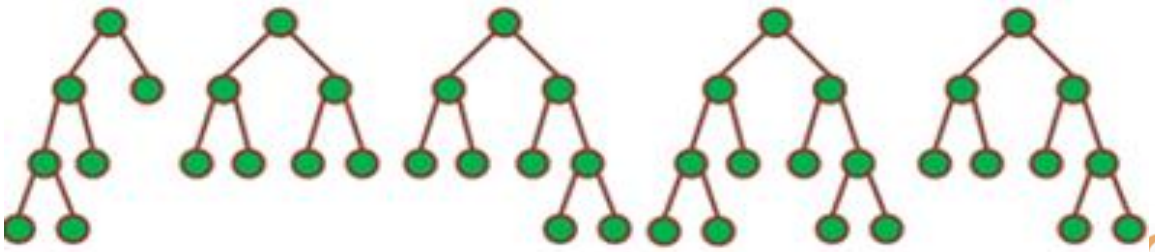
Step 2 機械学習モデルによる決定木

Random Forest

63 個の特徴量からランダムに、いくつかの統計量を選んで、ルールベースの決定木を多数作成

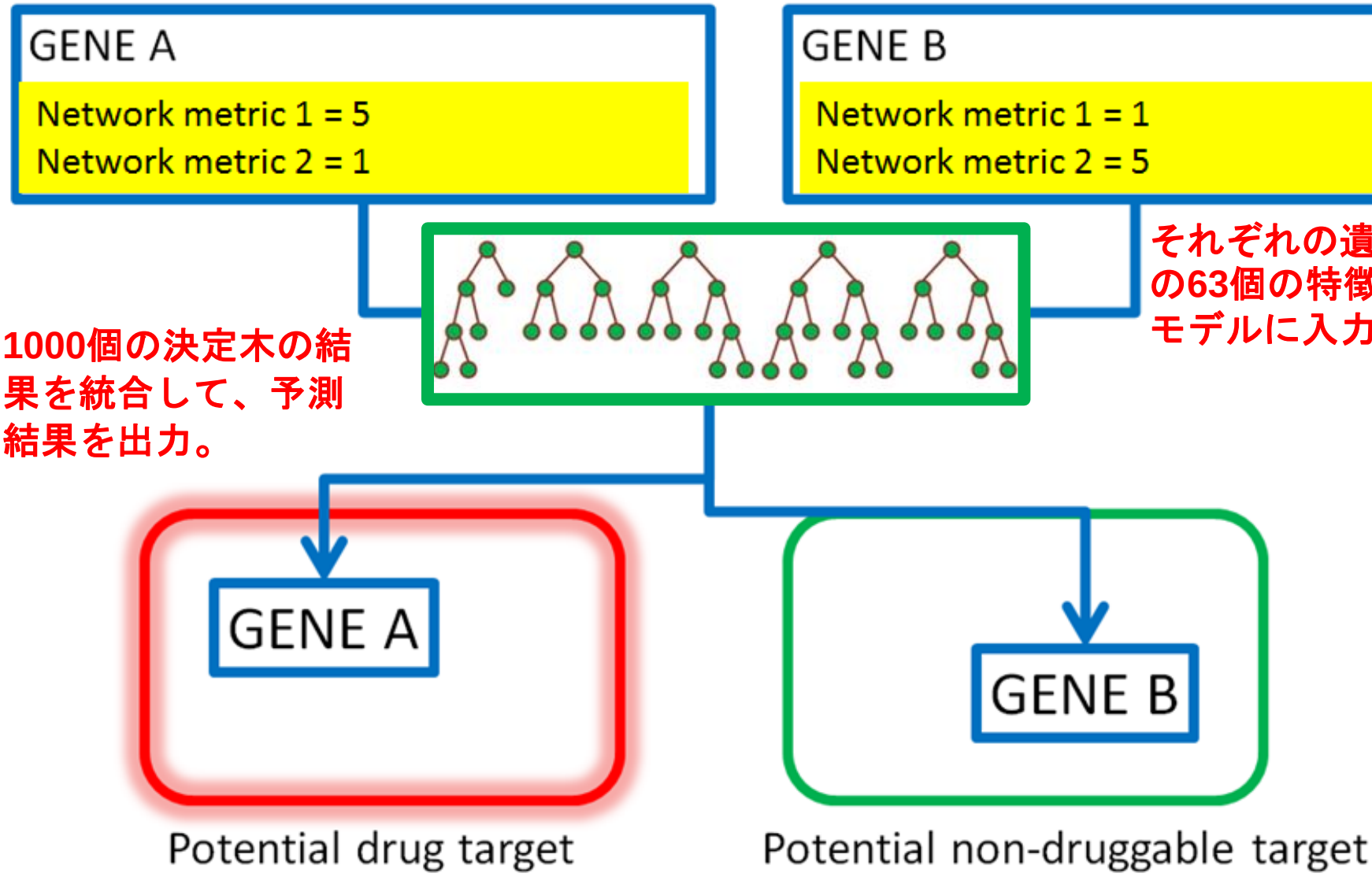


何度もランダムに特徴量を選んで、選んだ特徴量を用いて、多数の決定木モデルを作成する。今回は１０００個の決定木モデルを作成した。



TMDU

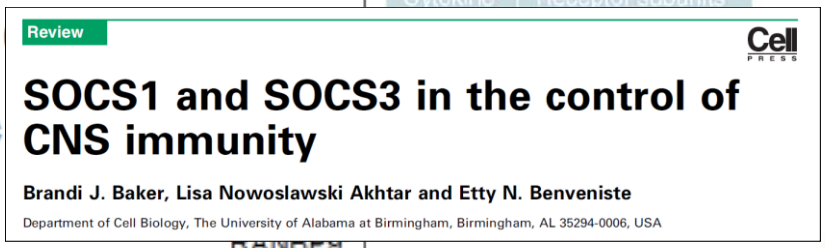
機械学習モデル　Random Forest



＜疾患- 標的分子＞予測結果

Target disease	Top 25 genes for potential novel drug target for a target disease
Anti-Alzheimer's disease	***SOCS1***; 1A; HSPG2; PARD6A; BLNK; PRKCI; YES1; PPP3CA; MMP11; SPTAN1; PTPRC; FBLN1; RPLP1; FAM107A; TRADD; NR4A1; LAT; NF2; PRKCE; KIT; NID1
Anti-Anxiety	***S1PR1; CNR1; MTNR1A; CCL4; F2; TEC; L8; CRHR1; AGTR2; OPRD1; IL8RA; RNF43; RHO***; SP6; RAB13; DRD4; IL8RB; MMP9; MMP2; OPR 1; IL1B; GNAS; S1PR3; KIT; GRM2
Anti-Rheumatoid	***SLC22A5; GRASP; KIT; SLC22A4; CFH; HSP90AA1; UBB; DHRS3; SCTR; ADORA1; N***; C6orf47; NR4A1
Anti-Breast Cancer	***SHC1; NFKB1; RELA; ID2; RAC1; SRC; RUNX; HDAC2; I***
Anti-Colorectal	***HC1; GHR; HRAS; PLCG1; CK; LYN***
Anti-Pancreatic	***H2; DSCC1; TAF9; AIF1; L2; NSMCE4A; ASPH***;
Anti-Melanoma	RAP; HMGB2; FEN1; NA; MUC1; LMNB1; GDE1; HNRNPA PSEN1; P

SOCS1はJAK/STAT pathwayを解してcytokine responseを変動させ、central nerve systemの炎症を制御


Review

Cell PRESS

SOCS1 and SOCS3 in the control of CNS immunity

Brandi J. Baker, Lisa Nowoslawski Akhtar and Etty N. Benveniste

Department of Cell Biology, The University of Alabama at Birmingham, Birmingham, AL 35294-0006, USA

Trends in Immunology Vol.30 No.8

しかし、SOCS1は上流の遺伝子なので、この下流の遺伝子をh標的にした方が、長期投与には良いかもしれない。

TMDU

疾患-標的分子リンクの同定よりDRへ

機械学習で予測された、新規標的の情報(disease A とtargetの情報,標的がdisease Aの新規標的分子、青いリンク)を、既知のdrug-target-disease interaction networkをマップし薬剤の新しい適用疾患（赤リンク）を予測

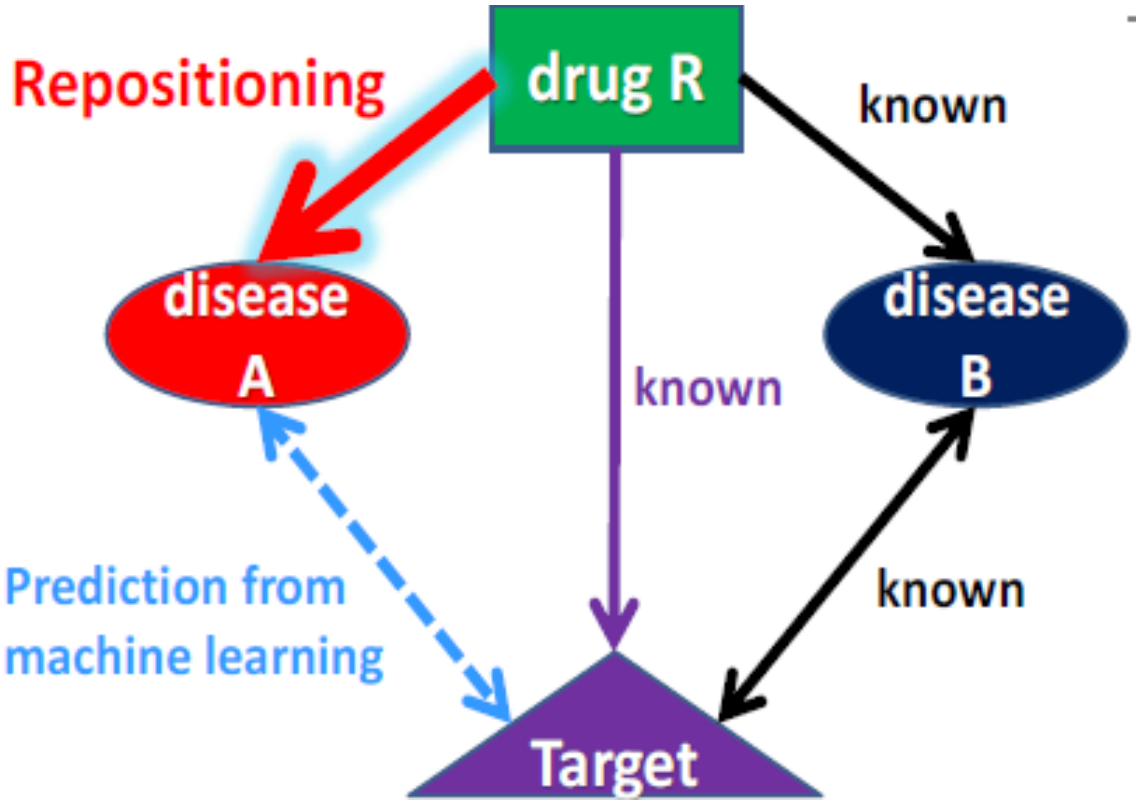


Table 3. Predicted repositionable drug candidates

Target disease	Candidate repositionable drug
Anti-Alzheimer's disease	Imatinib; Marimasta[t]; Nilotinib; Regorafenib; Sorafenib; Tamoxifen; Urokinase
Anti-Anxiety	3-Methylfentanyl; Agome[p]atine; Amitriptyline; Amoxapine; Antihemophilic Factor; Apomorphine; Aripiprazole; Bromocripti[n]e; Buprenorphine; Butorphanol; Cabergoline; Canakinumab; Captopril; Chlorpromazine; Coagula[t]ion Factor IX; Codeine; Dextromethorphan; Dextropropoxyphene; Dopamine; Drotrecogin alfa; Ethylmo[r]phine; Etorphine; Fentanyl; Halothane; Hirulog; Hydrocodone; Hydromorphone; Imatinib; Ketamine; Ketobemidone; L-DOPA; Lepirudin; Levallorphan; Levorphanol; Lisuride; Loperamide; Lo[x]apine; Marimastat; Melatonin; Menadione; Methadone; Methadyl Acetate; Methotrimeprazine; Mino[c]ycline; Morphine; Naloxone; Naltrexone; Nilotinib; Olanzapine; Ondansetron; Oxycodone; Oxymorpho[n]e; Paliperidone; Pergolide; Pethidine; Pramipexole; Promazine; Propiomazine; Quetiapine; Regorafenib; Remifentanil; Remoxipride; Risperidone; Ropinirole; Rotigotine; Sorafenib; Sufentanil; Suramin; Thi[e]thylperazine; Ziprasidone
Anti-Rheumatoid	Acetylcholine; Adenosine; Amiloride; Aminohippurate; Aminophylline; Amphetamine; Ampicillin; Azidocillin; Benzylpenicil[l]in; Cefalotin; Cefdinir; Cefixime; Cephalexin; Choline; Cimetidine; Clonidine; Cyclacillin; Desipramine; Diphenhydramine; Dopamine; Dyphylline; Enprofylline; Epinephrine; Furosemide; Grepafloxa[c]in; Histamine Phosphate; Imatinib; Imipramine; Insulin, isophane; Ipratropium bromide; L-Arginine; L-C[a]rnitine; Levofloxacin; Lidocaine; Liothyronine; Lomefloxacin; Mepyramine; Methamphetamine; Niac[i]n; Nicotine; Nilotinib; Norepinephrine; Norfloxacin; Ofloxacin; Oxtriphylline; Pentoxifylline; Probene[c]id; Procainamide; Quinidine; Quinine; Regorafenib; Rifabutin; Secretin; Sorafenib; Spermine; Te[o]adenoson; Testosterone; Tetraethylammonium; Theophylline; Thiamine;

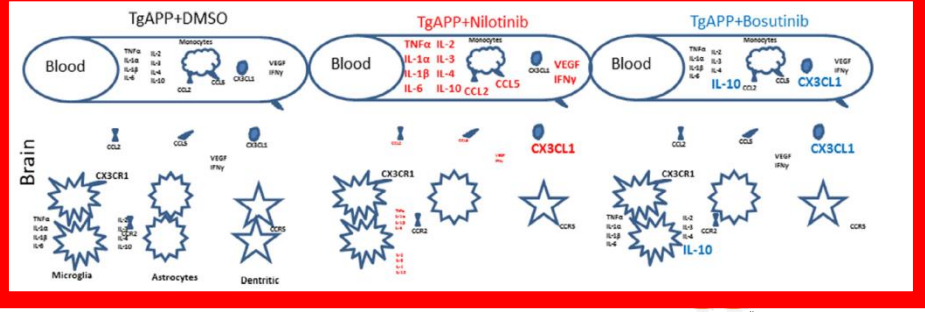
Neuroscience 304 (2015) 316–327

NILOTINIB AND BOSUTINIB MODULATE PRE-PLAQUE ALTERATIONS OF BLOOD IMMUNE MARKERS AND NEURO-INFLAMMATION IN ALZHEIMER'S DISEASE MODELS

I. LONSKAYA,[a] M. L. HEBRON,[a] S. T. SELBY,[a] R. S. TURNER[b] AND C. E.-H. MOUSSA[a]*

[a] *Department of Neurology, Laboratory for Dementia and Parkinsonism, Georgetown University Medical Center, Washington D.C. 20007, USA*

[b] *Department of Neurology, Memory Disorders Program, Georgetown University Medical Center, Washington D.C. 20007, USA*



ビッグデータの
clinical trialでの利用

―RCT, EBMからの呪縛の解放―

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網羅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**
- **我が国での推進**
  - 電子カルテの整備が遅れているので、急にRWDの創薬利用は困難
    - 米国はHITECH法のお陰でmeaningful use 基準で90%の普及
  - まずは、**疾患レジストリー、疾患型Biobank**からはじめる
  - **On demand型のリソース提供**は始まっている（病院併設型Biobank)

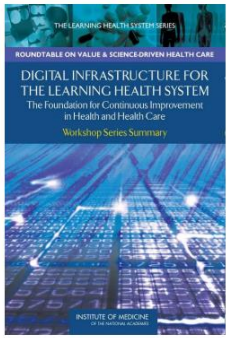
「学習する医療システム」
Learning Health System

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

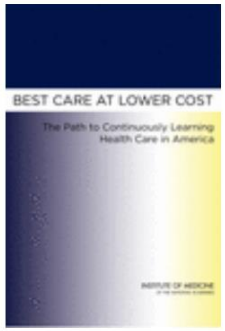
- IOM “Clinical Data as a Basic Staple of Health Learning”
- 医療システムのデジタル化（IT化）は必然の傾向である
- **「ルーチンの医療活動から集められたデータ（定式化臨床研究と違って）がLHSを支える鍵である」**
- データを共有することによって学習して医療システムを改善
- **RCTは「黄金基準」であるが、通常の医療システムの外で**実施されている。医療が実際対象とする患者集団を代表しているのか。
- **RCTは時間が掛かり費用もかかる**
- 有効な知識の蓄積の速度が加速する

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

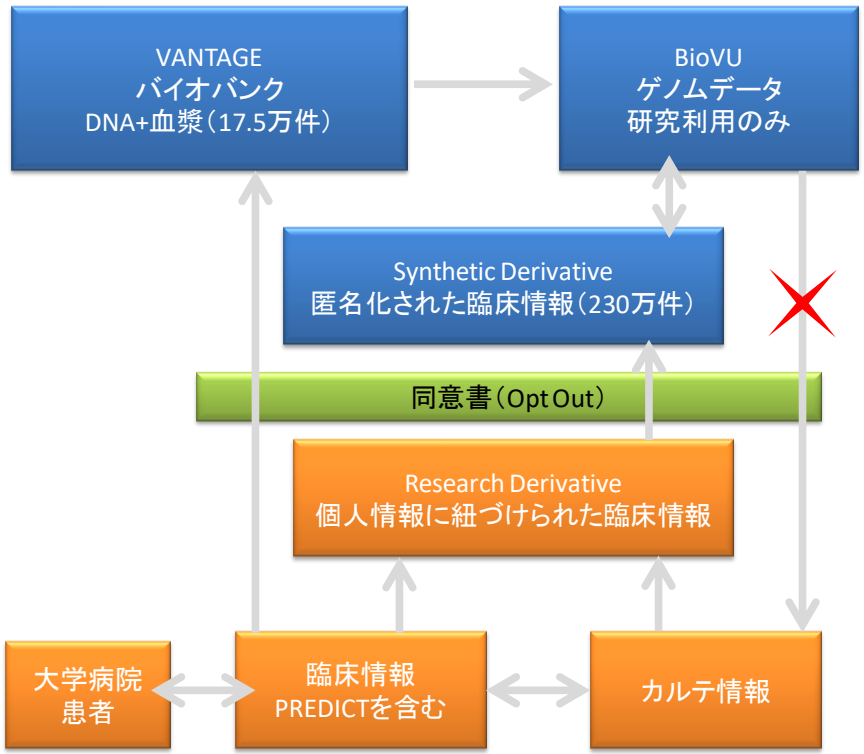
Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care



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

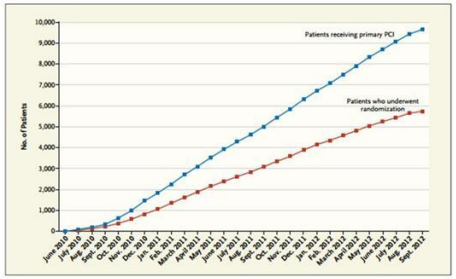
クロビドグレル（抗血栓剤）の遺伝子多型関してABCB1, CYP2C19、さらにPON1の多型が知られていたが、ヒトを対象とした臨床実験の報告はなかった。SDから循環器疾患でclopidogrelの投与歴の対象者（ケース群）およびコントロール群を選出。BuiVUから遺伝型を決定する。この条件に合致するケース群は255件。解析の結果、CYP2C19*2とABCB1の関与は有意。PON1は非有意が判明した。

TMDU

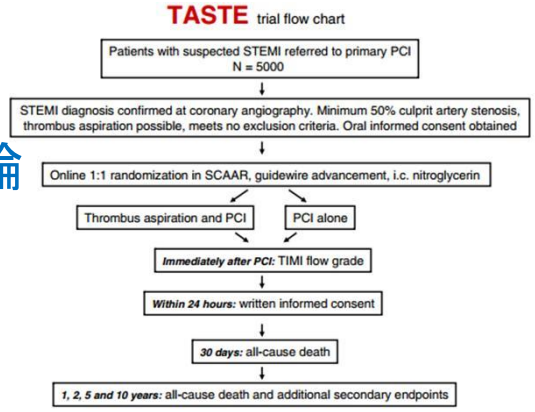
Biobank (registry) 準拠の創薬過程・治験

- スウェーデンの**TASTE**(ST segment-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治験によって相反する結論
- 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.



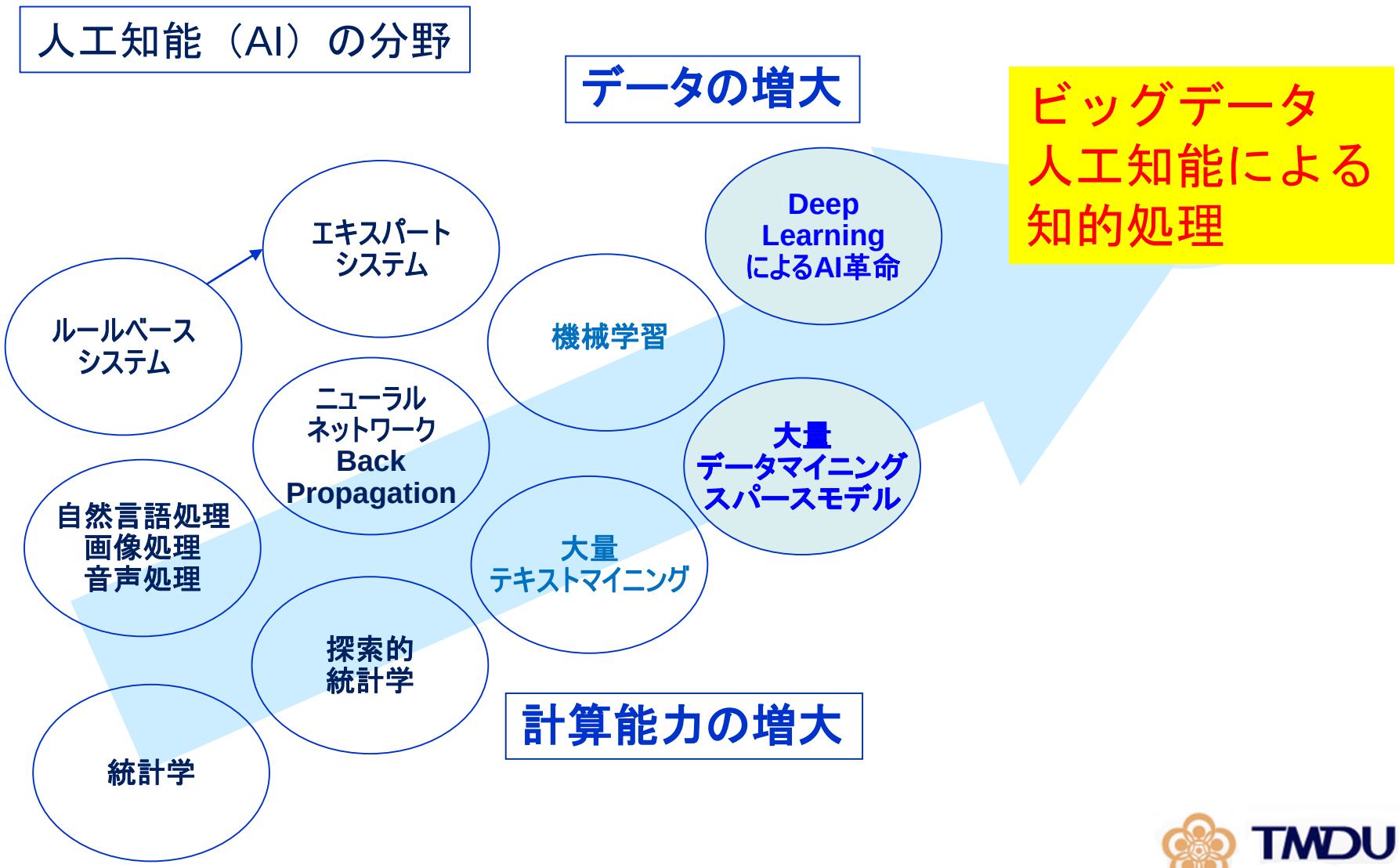
RRCTの特徴

- 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**
  - **迅速な治験参加者登録**
  - **非登録患者の調整**
  - **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

ビッグデータ研究とRCTの融合: 将来の治験方式

AI創薬/DR

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

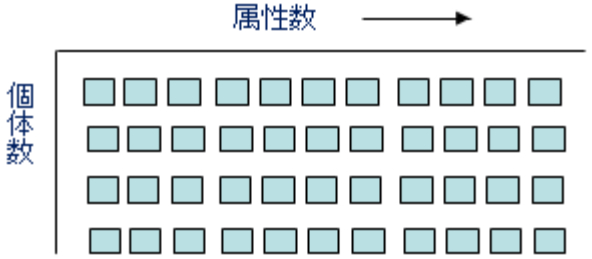
2000

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 $\{(y_i, \mathbf{x}_i); i = 1, \ldots, n\}$

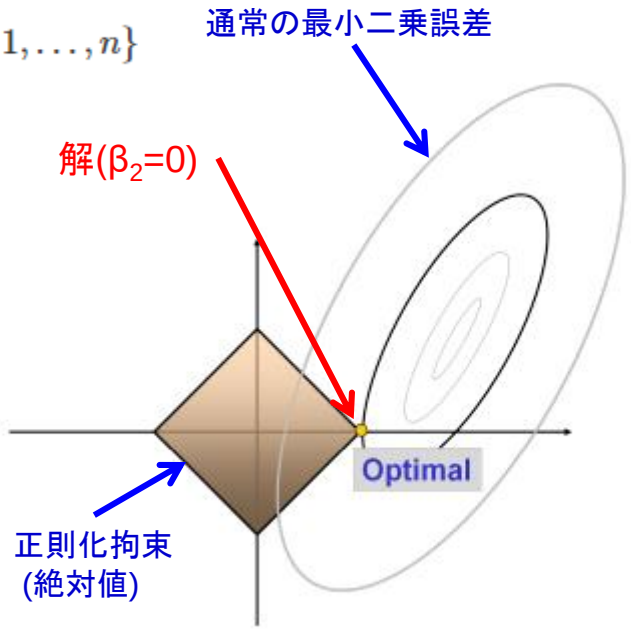
$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

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.”

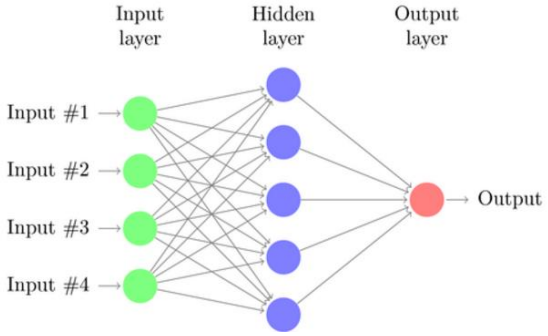
IBM Watson Learning Big Data

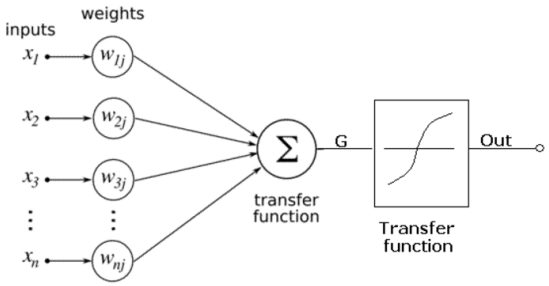
TMDU

Deep Learning 型人工知能の革命性

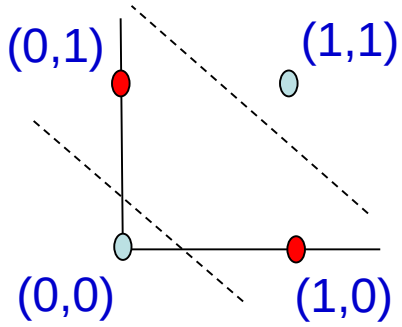
従来のニューロネットワーク

古典的Neural Network・パーセプトロン(1970年代）



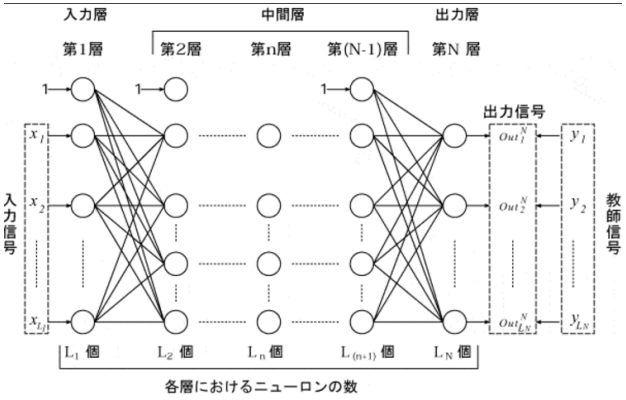


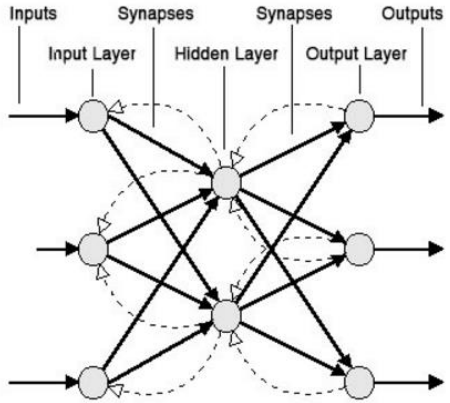
XOR



線形分離できない

多層Neural NetworkとBack projection (1980年代）





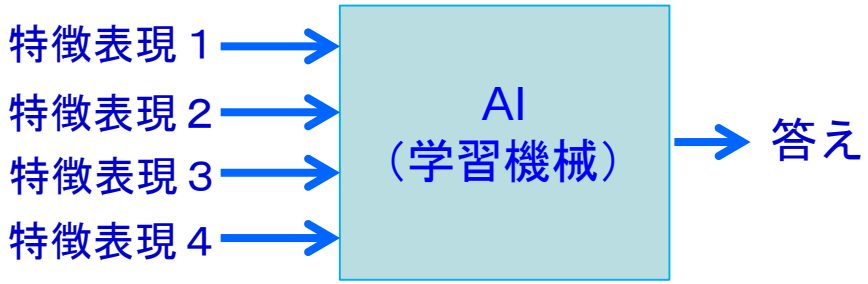
Back Propagation (1986　Rumelhart)
望ましい出力との誤差を教師信号として与える事により、次第に結合係数を変化させ、最終的に正しい出力が得られるようにする。
結合係数を変える事を学習と呼ぶ。
この学習方法には、最急降下法（勾配法）が使われる。出力層へ寄与の高いノードの重みが変更。

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

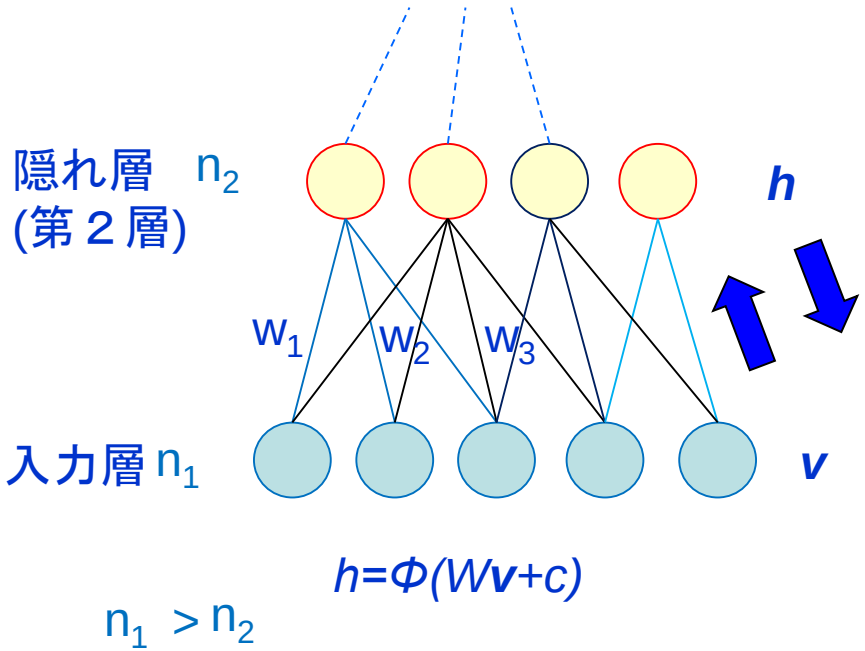
- DLは、まずは対象の固有の構造を記述する特徴表現や対象の高次特徴量を自ら学ぶ「教師なし学習」を行う
- 「内在的な特徴表現の学習」を自動的に行う
  - 自己符号化（Autoencoder）
  - 制限ボルツマンマシン
- 最終層で、人間の概念との相同をとるため「教師あり学習」

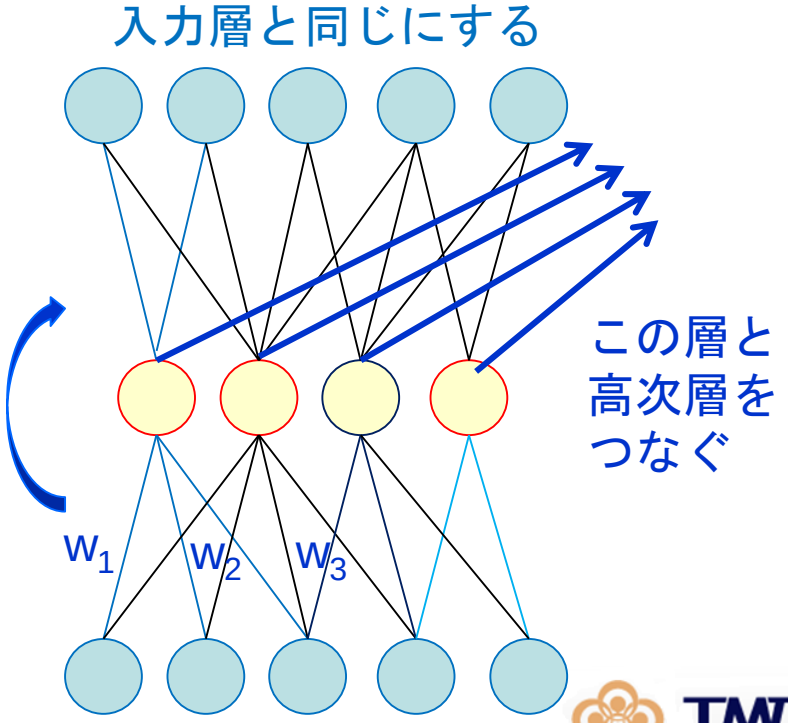
DLの革命点 Autoencode 1

- 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

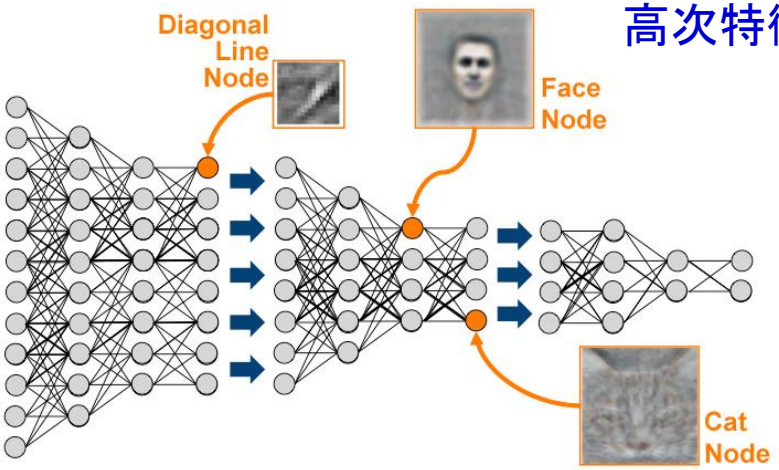
→内在的な特徴量を見出す





DLの革命点 Autoencode２

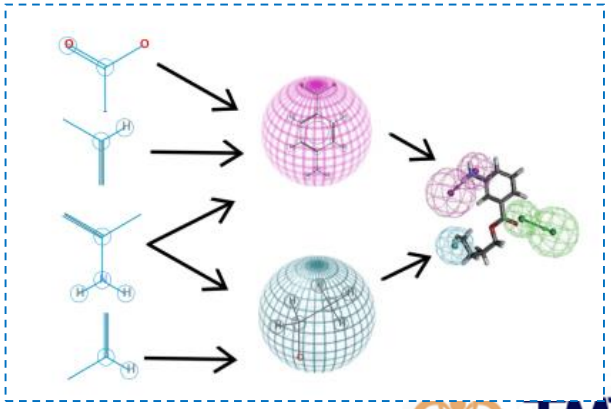
- 各層ごとに自己符号化を行うので何層でもネットを組める→Deep Learning
- 第一層で学習した特徴量を使って、つぎの階層を作るので高次の特徴量が作られる
- 特徴的表現と概念を結びつけるため「教師あり学習」が最後に必要である
- 自動特徴抽出によってこれまでの学習手法の限界を克服した（構造的理解）



高次特徴

部品的特徴

Pharmacophoreの摘出



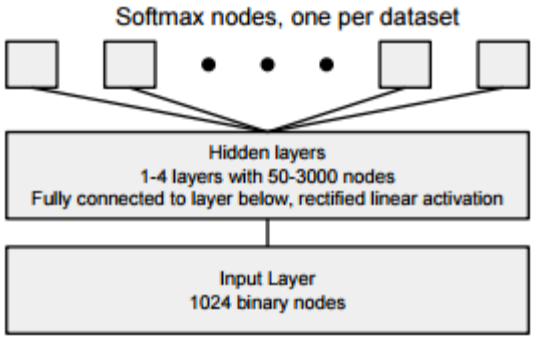
Deep Learningの創薬へ応用

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

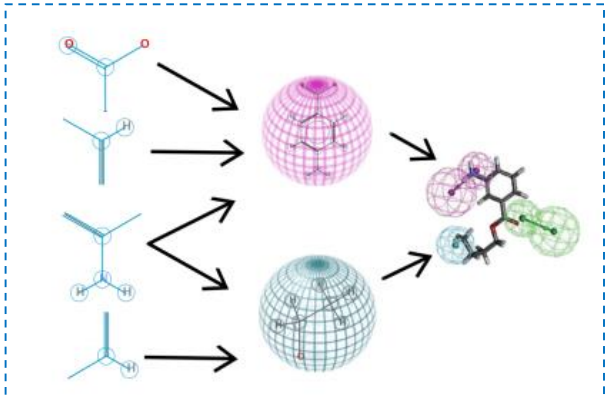


Figure Hierarchical nature of fingerprint features: by combining the ECFP features we can build reactive centers. By pooling specific reactive centers together we obtain a pharmacophore that encodes a specific pharmacological effect.

ECFP(chemical substructure: Enhanced cyan fluorescent)

DL型NNへの期待と困難点

- まだ、医療・創薬のDL応用は緒に着いたばかりで、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 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
- DL型NNの困難点
  - 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、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

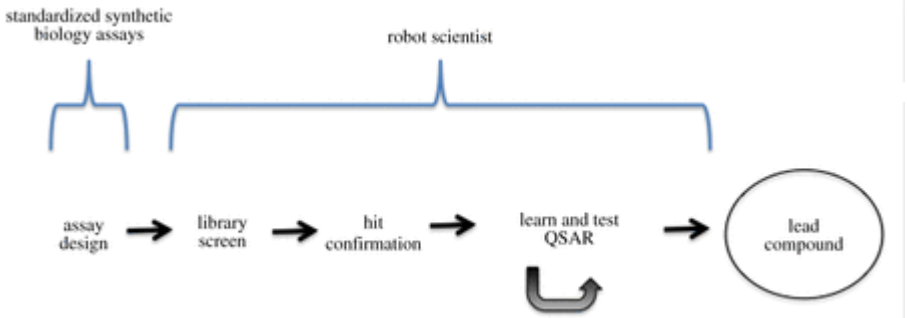
AI創薬/DRの方向性

- 我々の研究の「ネットワーク解析と機械学習に基づいたDR」
  - 機械学習の時点で、DLも試みたが特徴量抽出後であったためか、精度はRandom Forrestの方が高かった
  - ネットワーク特徴量を与えるのではなく、ネットワーク総体の情報を与えてStacked Autoencoder で特徴量を学習させる
- AI創薬の基本方針
  - 創薬・DRが根拠とするネットワーク/空間をDeep Learningで学習する戦略が期待される

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

ゲノム・オミックス医療の次世代の展開とlife-long healthcareへ

第２世代のゲノム医療へ

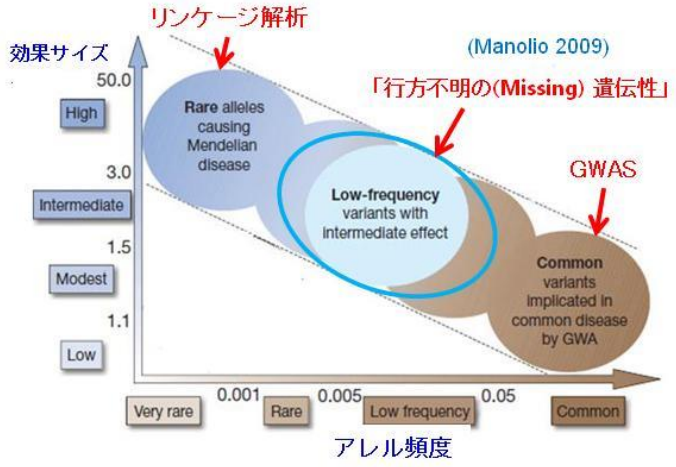
成功した臨床実装

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

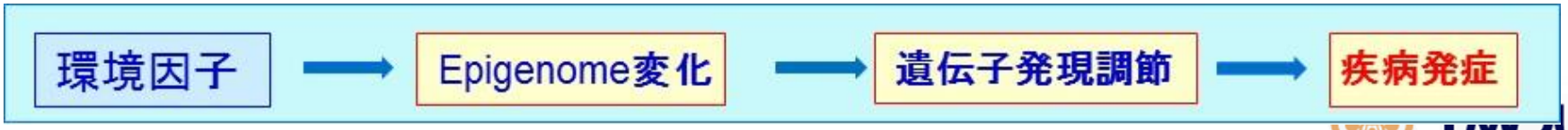
GWASでSNPの相対リスクが低い(1.1～1.3)理由： **GxE組合せ特異的効果を環境要因の全てに亘って平均しているからである**

発達プログラム説 DOHaD

Developmental Origin of Health and Disease)

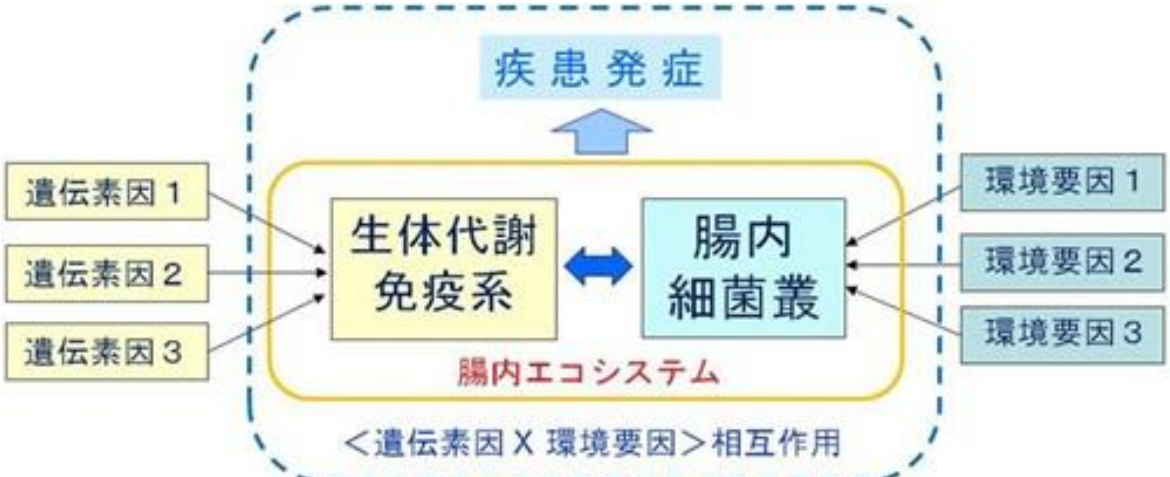
オランダ
飢饉 (1944)

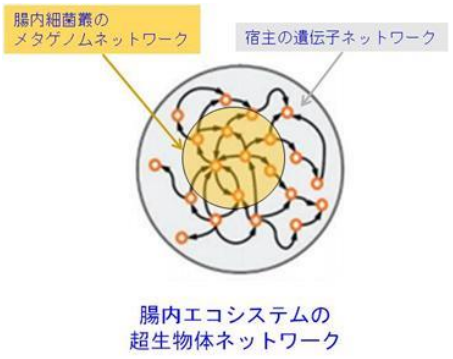
- オランダ飢饉
  - 第2次大戦末期、ナチスの封鎖、約半年間酷い飢饉
  - 飢饉の期間に胎児、戦後30年
  - 成人期:肥満,糖尿病,心筋梗塞,統合失調
- Baker仮説：英国心筋梗塞増加
- エピジェネティク機構
  - 過度な低栄養：肝臓のPPARα/γ（倹約遺伝子）メチル化低下・遺伝子発現がオン
  - エピジェネティック変化は**可変**：短期的変化、**長期的「記憶」**次の世代も



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、個別化治療に取組む
  製薬産業もlife-long healthcareの産業へ

- ゲノム・オミックス情報と医療・健康

- 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
  - ＜トランス・ゲノム機構＞へ

TMDU

ご清聴ありがとうございました