ゲノム医療実現の基盤とバイオバンク

東京医科歯科大学　名誉教授（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　特任教授
機構長特別補佐

田中　博

わが国における

「ゲノム医療元年」

急速に展開する国際的な動向

- **米国**

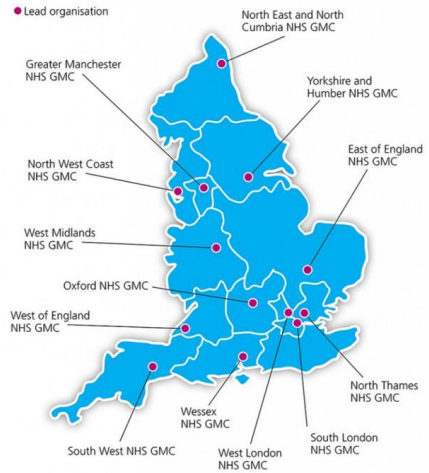
  Precision Medicine Initiative 2016～
  - オバマ大統領 一般年頭教書
  - 個別化医療というより層別化医療
  - 遺伝的素因・環境素因、mHealthも
  - 短期的：精密腫瘍学 Precision Oncology
  - 長期的：100万人ゲノム・コホート

2015年1月　オバマ大統領一般年頭教書

- **英国**

  Genomics England 2013～2017年
  - 10万人の全ゲノムがん・希少疾患・感染症
  - Genomics Expert Network for Enterprises (GENE) Consortium
    - 10万人ゲノムを使用する一年の企業トライアル
    - 13 箇所のGMC、seq解析はサンガーセンター集中

- **国際的なゲノム情報共有アライアンス**
  - **GA4GH**(Global Alliance for Genomics and Health）
  - 2013年設立(Haussler), 33ヵ国326施設、配列グラフ
  - ゲノム配列と臨床転帰の国際連携
  - Matchmaker Exchange：
    - 表現型・遺伝型の相互検索、類似疾患の原因を検索



NHS Genome MedicalCenter

いよいよ動き出した我が国の「ゲノム医療」
政府・行政

- 健康・医療戦略推進会議
  - 「ゲノム医療実現推進協議会」設置27.1
  - ゲノム医療推進方針「中間とりまとめ」
- 厚生労働省
  - 「ゲノム医療実用化研究推進事業」(AMED)26～
  - 「ゲノム医療推進本部」設置 27.9
  - 「臨床ゲノム情報統合DB事業（AMED）」
- 日本医療研究開発機構（AMED）27.4月 設置
  - 「未診断疾患イニシアチブ（IRUD）」27.10
  - 「ゲノム医療推進WG」報告　28.2
  - 「ゲノム医療実現推進プラットフォーム事業」
  - 3大バイオバンク研究基盤事業

いよいよ動き出した我が国の「ゲノム医療」
先進的医療機関

- **国立がん研究センター**
  - NCC oncopanel によるがん診断
    - ゲノム医療実用化研究推進事業（厚労、AMED）
  - SCRUM-JAPAN
    - 産学連携全国がんゲノムスクリーニング
- **静岡県立がんセンター**
  - HOPE計画
    - マルチオミックス解析によるがんの個別化医療
    - 遺伝性疾患の予防を目指す未病医学、遺伝情報結果回付
- **京大腫瘍内科**
  - OncoPrime計画
    - がんドライバー遺伝子の同定と分子標的薬選択（自由診療）
    - ４大学診療施設併設型バイオバンク
      - 京都大学、岡山大学、北海道大学、千葉大学
- **その他の医療施設**
  - 全国遺伝子医療連絡会議では国内に12の医療施設が研究予算でclinical sequence

ゲノム Healthcare実現の２つの流れ

- 臨床ゲノム医療
  - Clinical Implementation ゲノム医療の臨床実装
  - 現在、主にはClinical Sequencing at POC
  - 遺伝子変異・多型が疾患・病態に影響
  - がん、希少疾患（単因子性遺伝疾患）など
  - 米国で著明に進展
- 大規模ゲノム調査研究
  - Large-scale Genomic Study
  - 主に欧州
  - GWASからゲノム・コホート（Biobank)へ
  - 疾患ゲノム・コホート
  - Population型ゲノム・コホート

臨床ゲノム医療の流れ
―米国での進展を見る―

次世代シーケンサのインパクト

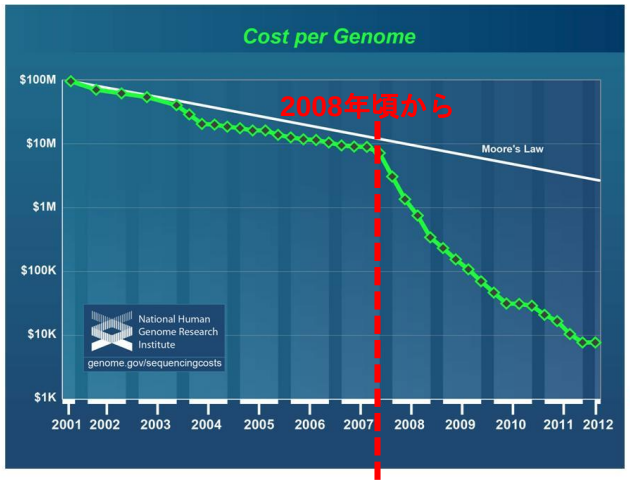
次世代シーケンサを始めとするhigh-throughput分子情報収集の急激な発展

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！

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

注：初めから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。網羅的でデータ駆動的な検査（**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

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

第１世代

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
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

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

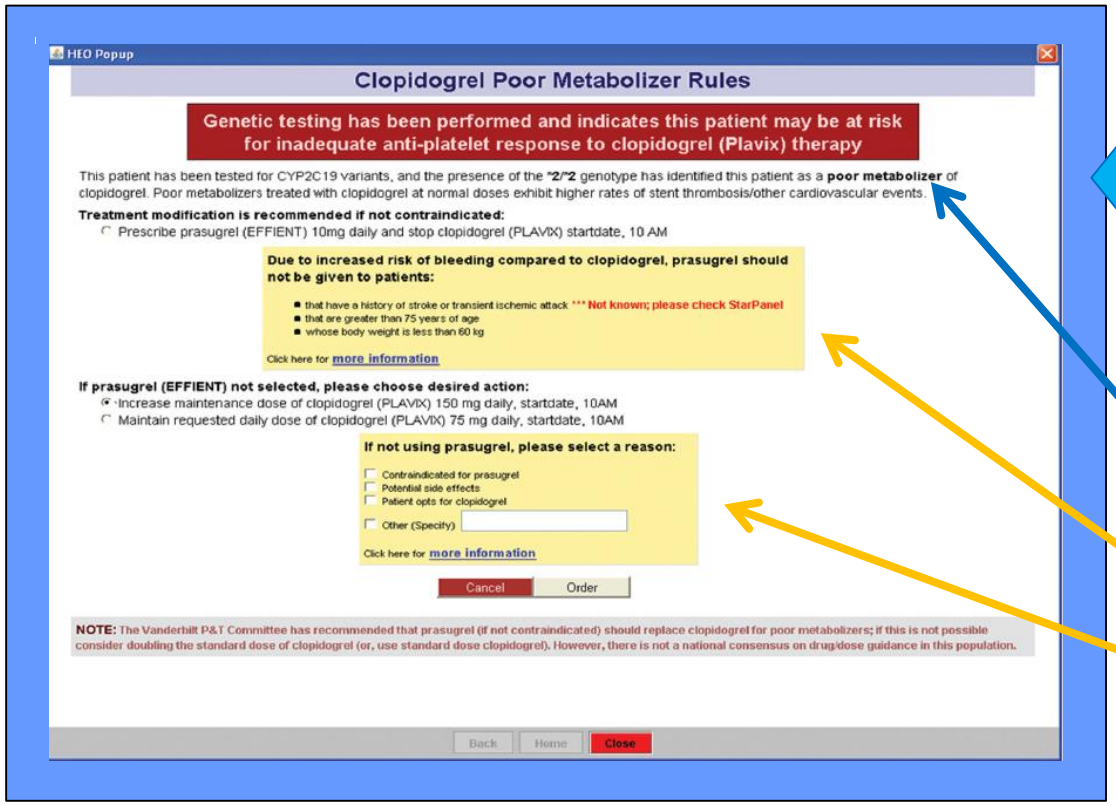
第2の流れ

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

ゲノムオミックス医療臨床実装化の第３の流れ

著名ながんセンター
Dana Faber /MD Andersonなど

第３の要素が加わる

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
国際がんゲノムコンソーシアム（ICGC：2008年から）
　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**

ゲノム・オミックス医療の３つの流れ

2005～ NGS (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

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

TCGA (2006) 国際がんコンソーシアムICCG 2008開始の成果 2011から出現 シーケンス革命

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

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

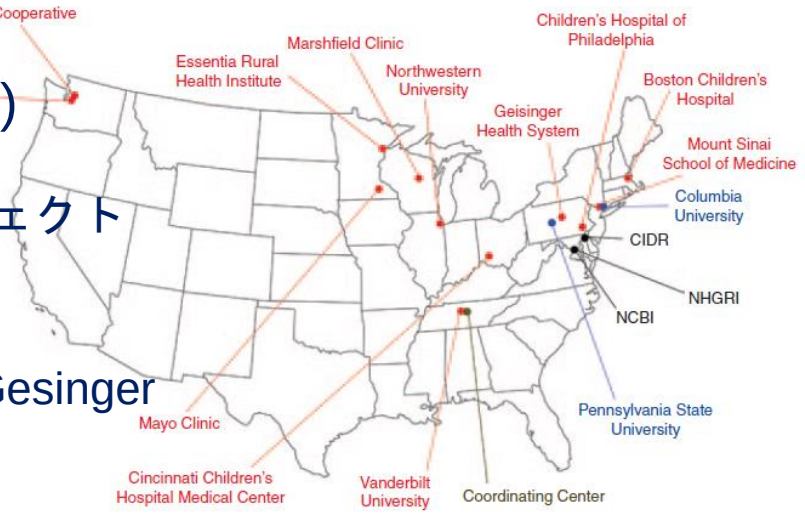
Institution	Major Projects
Alabama	Planning stages for projects in risk assessment, pharmacogenetic analysis, identification of families for further research
Baylor	Whole exome and whole genome sequencing in Mendelian disorders to improve diagnosis
Geisinger	• Selection for gastric bypass surgery vs 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

臨床表現型 *eMERGE*プロジェクト

electronic MEdical Record + GEnome (NIH grand)

電子カルテからphenotyping

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



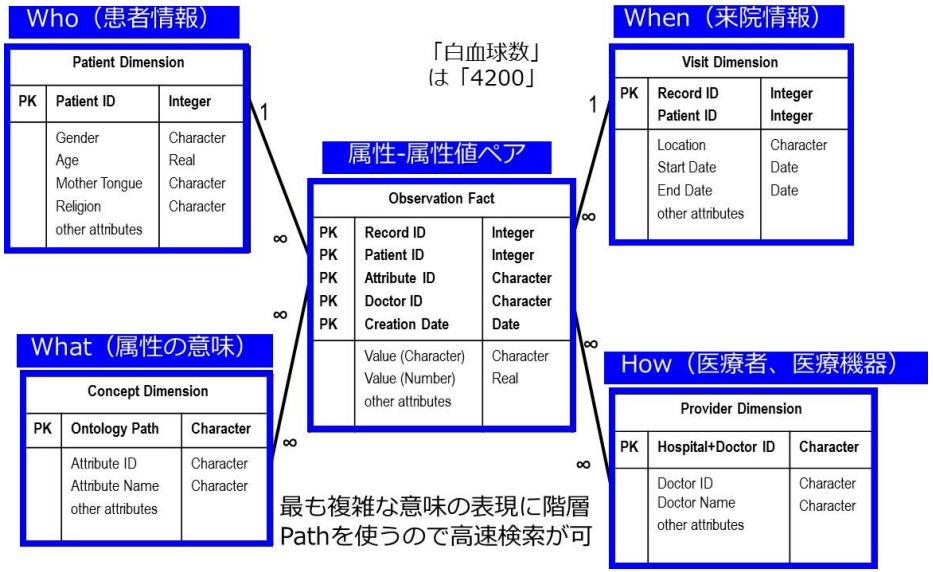
臨床データの表現型形式化（Phenotyping）の問題

i2b2 (Informatics for Integrating Biology and the Bedside)

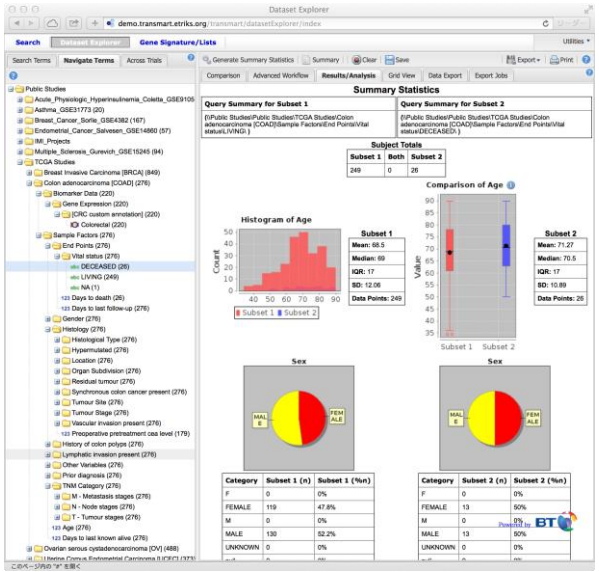
- 格納すべきあらゆる情報を
  主語（subject）述語（predicate）目的語（object）のトリプレットで形式化、
- オントロジーとの組み合わせで検索可能とする、特徴的な設計
- Star Schema:データベーススキーマの１つ、その中心に位置するobservation_factテーブルに集約される。

tranSMART – トランスレーショナル生物医学研究のプラットホーム

- tranSMART Foundationにより開発されているオープンソース(GPL3)のプラットフォーム：データマート方式
- 転帰（outcome）などにより集団を抽出し、ヒートマップ, 相関解析, クラスタ分析, 主成分分析, 生存時間分析などの解析が可能 (IMI: Innovative Med. Initiative)

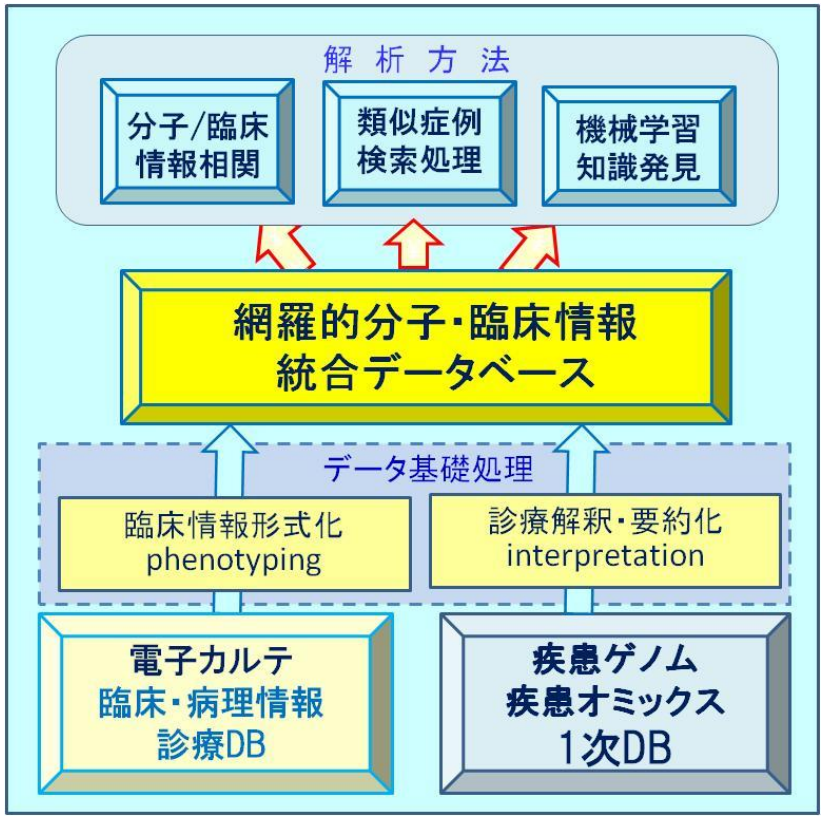


Node	Description
Biomarker Data	Measurements of biomarkers such as RBM antigens, gene expressions, antibodies and antigens in ELISA tests, and SNPs.
Clinical Data	Primary and secondary endpoints, and other measurements from the study.
Samples and Timepoints	Tested samples (such as tissue or blood) and time periods when the samples were taken.
Scheduled Visits	Periodic stages of the trial during which patients are seen.
Design Factors	Compounds involved in the study, dosages, and regularity with which the compounds were administered. **Note:** With clinical trials, this node is typically named Treatment Groups.
Sample Factors	Patient information, such as demographics and medical history.



臨床ゲノム医療の統合情報基盤

統合臨床オミックス・データベース
(integrated Clinical Omics Database)



電子カルテから入力された臨床・病理的診療情報と疾患ゲノム・オミックスから入力された分子情報は

1. データ基礎処理部

電子カルテから必要な情報をphenotyping して所定の形式に
分子情報はゲノムはvariant call、オミックス情報はsignature情報を中心にする

2. 統合データベース本体

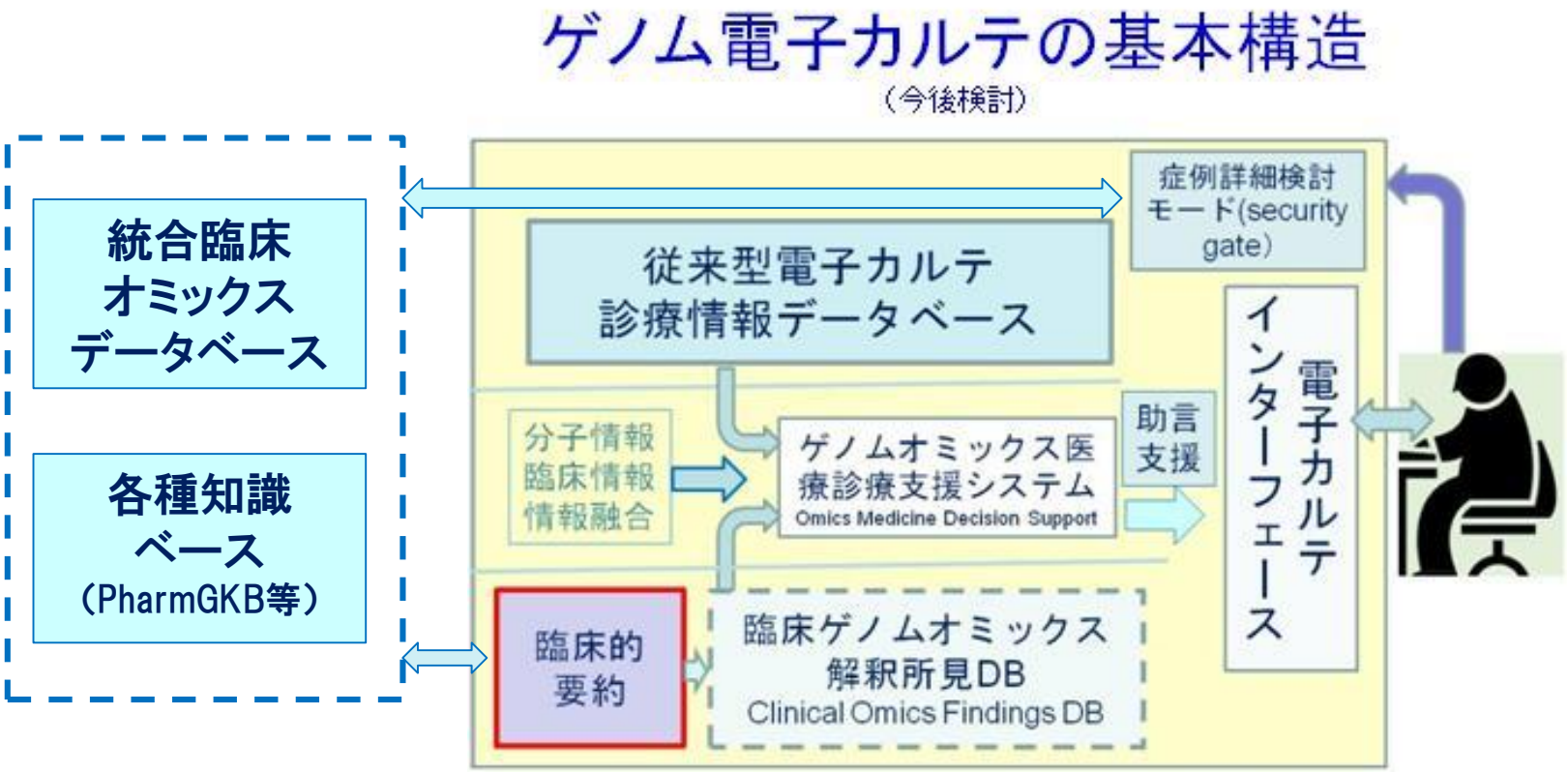
どのようなデータ形式か検討の必要
RDF化やi2b2方式など

3. データ解析部

分子・臨床情報相関解析、類似症例検索、機械学習の各システムを開発

病院内のゲノムオミックス支援電子カルテシステム

統合臨床オミックスデータベースを情報基盤にして
診療のゲノム・オミックス医療を実践する電子カルテEHR。
Geisinger Hospital などで実践



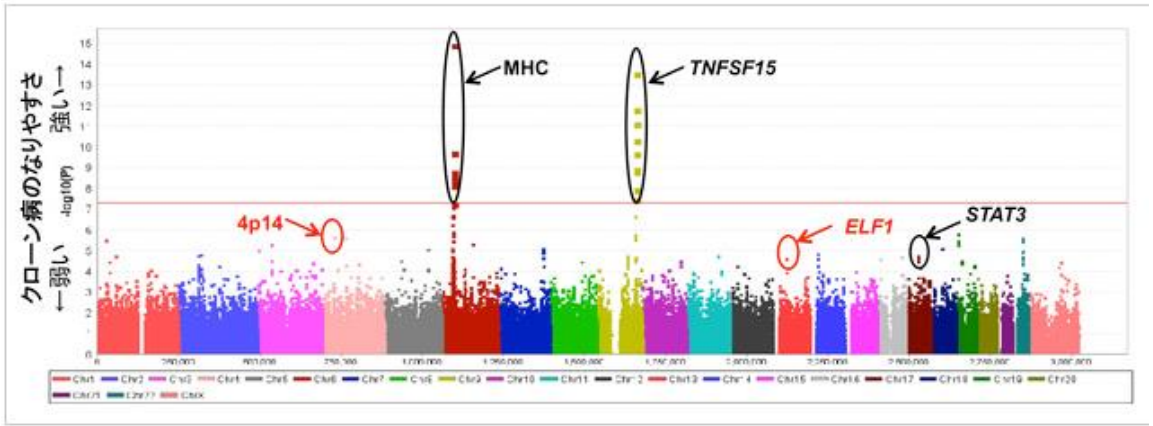
大規模ゲノム調査研究の流れ

ゲノムワイド関連解析
Genome-wide association study; GWAS

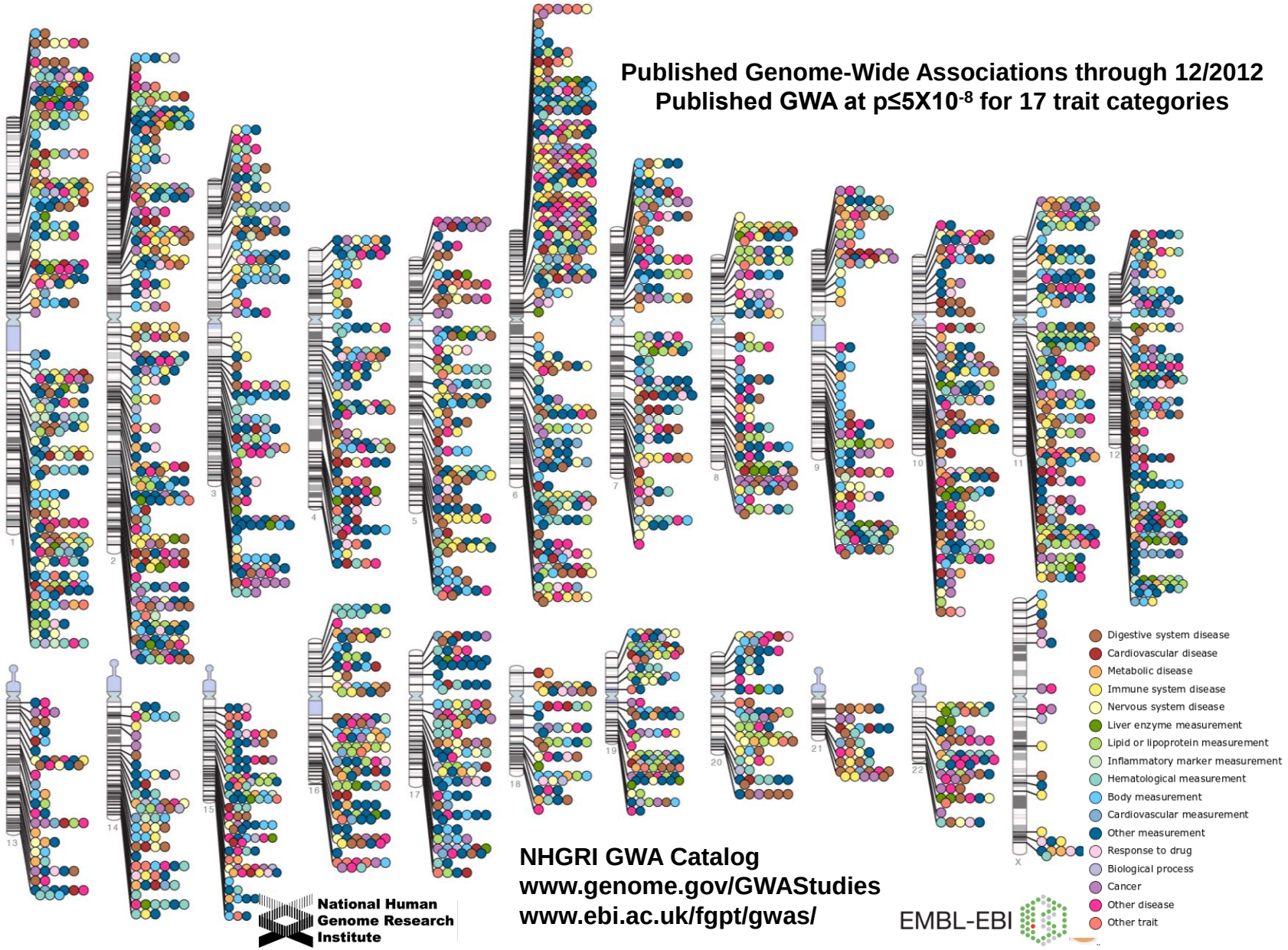
- ゲノム全体を網羅する一塩基多型情報と、疾患の有無や量的形質などの表現型情報との関連を統計的に調べる遺伝統計学の一手法
- ある疾患の患者(case）とその疾患に罹患していない健常者（control)との間で、～100万箇所の多型（主にSNP)の頻度の分布(差異)を調べ、有意な統計的連関があるかどうか統計的に検定し、疾患関連遺伝子を見出す。多重比較補正　Bonferroni補正

	AA	Aa	aa
case			
control			

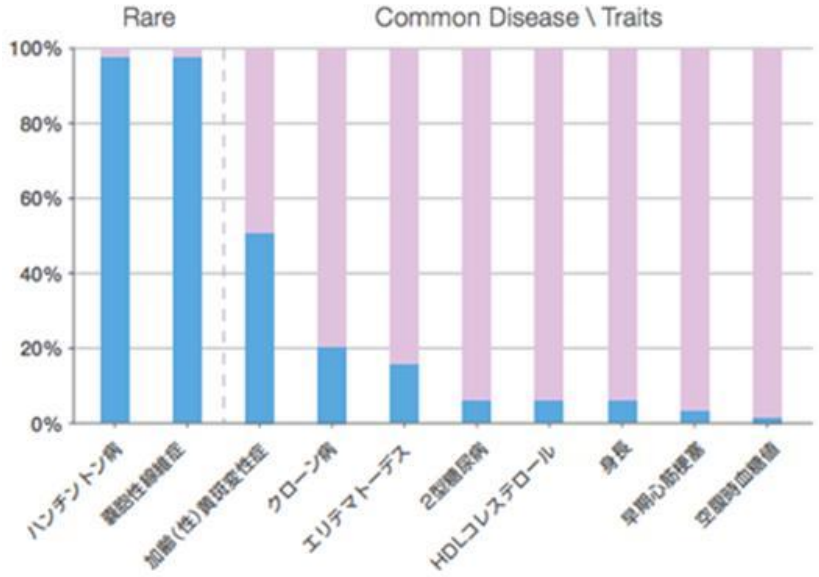
-log p 値のマンハッタンプロット

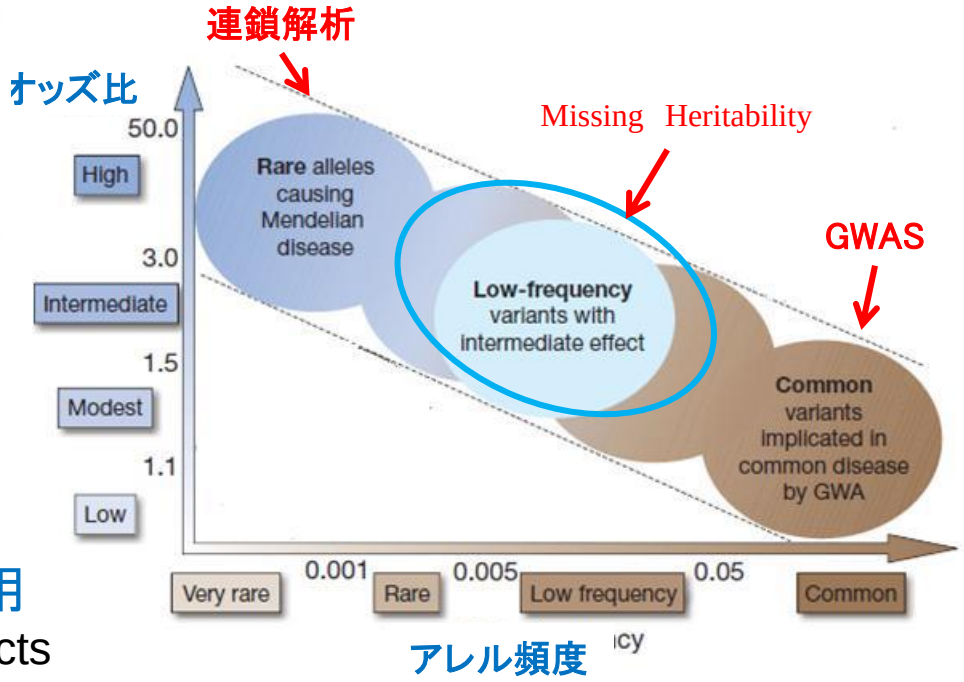


GWASで同定された関連遺伝子のマップ



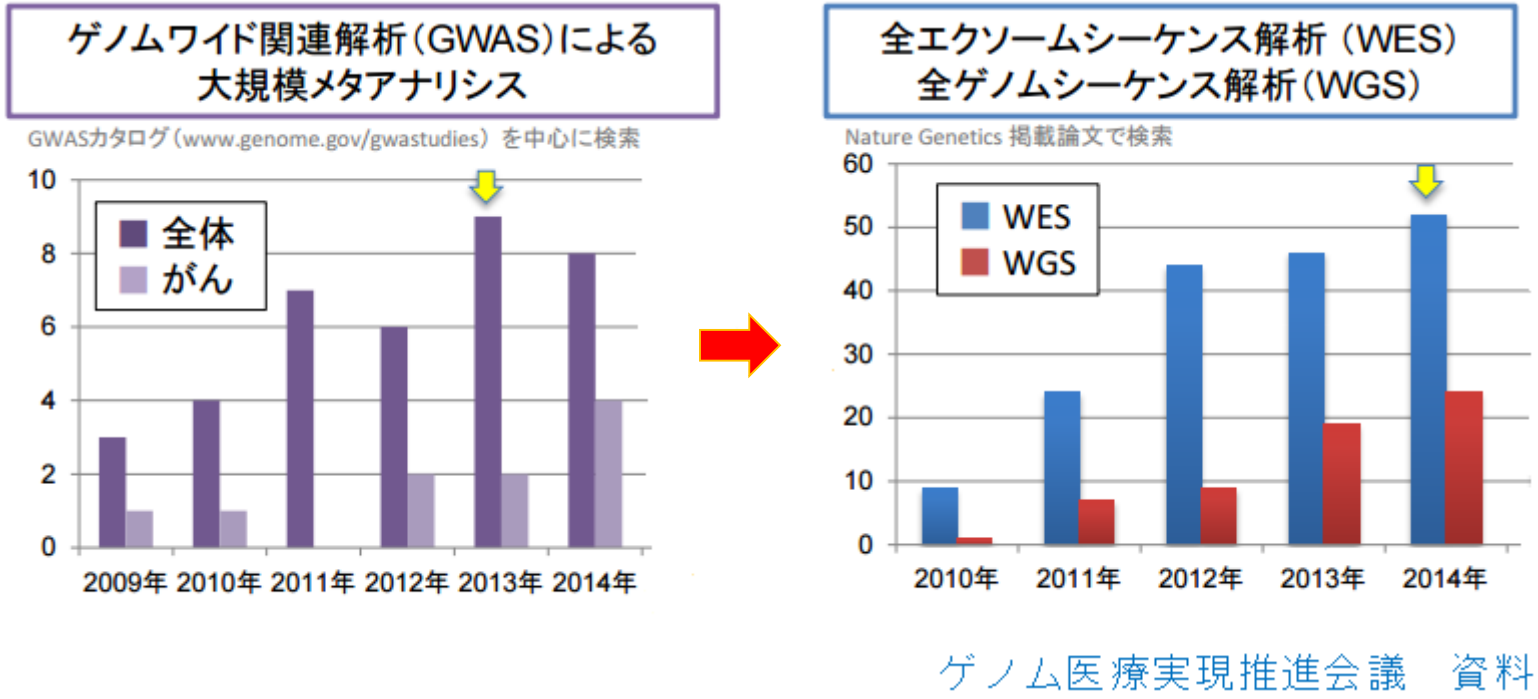
単遺伝子的アプローチでは未知な部分が多すぎる





- 遺伝継承性の20%～30%
- 我々の見解
  - 遺伝子間（Gene-gene）相互作用
    - Pathway-integrated polygenic effects
  - 遺伝子環境 相互作用（Gene-Environment）
  - **相互作用を１項目のみで評価**
    他の相互作用項の効果で相殺

GWASからBiobankへ



ゲノム医療実現推進会議　資料

GWASの研究は峠を超えた。これから関心は全ゲノムのコホート研究（前向き）へ移っている

Biobankとゲノムコホート

- **バイオバンクの目的･機能の変化**
  - 従来は再生医療ための生体標本や臨床研究の資料保存、
  - ゲノム/オミックス個別化医療、創薬の情報基盤

  ① 疾患型BioBank：
  - 疾患罹患患者の網羅的分子情報（ゲノムなど）と
  - 臨床表現型（臨床検査、画像、処方歴、病態経過、転帰など）の収集。
  - 目的：**個別化医療**の層別化パターンの網羅的摘出、疾患ゲノムコホート

  ② Population型BioBank：
  - 「健常者」前向きコホート。調査開始時の網羅的分子情報　と生活環境情報（exposome）を集めて、長期間追跡するゲノム・コホート
  - 目的　**個別化予防**の情報基盤　疾患発症リスク＝遺伝子要因 x 環境要因
- **欧州の代表的なBiobank**
  - **UK biobank**
    - 50万人の健常者。40～69歳（2006-2010, 62Mポンド), 追加2011-16, 25Mポンド
    - 健診データ（血液・尿・唾液サンプル、生活情報）を集め、健康医療状況を追跡）
  - **Genomics England**,
    - 2013開始、2017年までに 10万人のゲノム 配列収集。
    - 最初の対象は稀少疾患（患者・家族）、がん患者、最初はEnglandのみ
  - **BBMRI**（Biobank/Biomole. Res. Infra.)
    - 250以上の欧州各国のBioBankを統合

国内の主なバイオバンク・ゲノムコホートの状況

我が国における主なバイオバンク・ゲノムコホートを対象者、規模、目的で大別

大規模／多目的

健常者

患者

小規模／特定目的（特定疾患）

東北メディカル・メガバンク
東北大学・岩手医科大学
2013年～/被災住民8万人＋3世代7万人

JPHC(多目的コホート)
国立がん研究センター
1990,93,2011(NEXT）～地域住民/10万人

J-MICC(多施設共同コホート)
愛知がんセンター等、多施設共同研究
2005年～/10万人、愛知がんセ・名大

山形分子疫学コホート　2002年/9100人

久山コホート
追跡率99%・剖検率80%
50年の歴史/8000人九州大学

ながはま0次予防コホート
京都大学　2002年～/9100人

鶴岡メタボロームコホート
鶴岡市・慶応大学　2013年～/10000人

バイオバンク・ジャパン
（BBJ）/東大医科研・理研
2003年～/20万人（第3期～）
47疾患・12医療機関

ナショナルセンターバンク（NCBN）2012～
- 国立がん研究センター
- 国立循環器病研究センター
- 国立精神・神経医療研究センター
- 国立長寿医療研究センター
- 国立国際医療研究センター
- 国立成育医療研究センター

難病バンク
2009年～医薬基盤研究所（40種以上分譲）

※内閣官房 健康・医療戦略室作成資料より抜粋

Biobank/ゲノムコホートへの期待

- 疾患型バイオバンク/ゲノムコホート
  - 個別化医療パターンの網羅的摘出
  - 病院ゲノム・オミックス医療DBと相互補完
  - 疾患時間経過とゲノム・オミックス疾患機序の追跡
- 健常者（population型)コホート

(1) 前向きコホート: 発症要因同定「個別化予防]

疾患発症相対リスク　相互作用を評価

＝遺伝子要因 x 環境生活習慣要因

(2)「健康から疾患発症に至る過程」を多数収集

「先制医療受攻状態」(vulnerable period)同定

⇒ 先制医療創薬の開発

⇒ QOLにも医療経済的にも有効な政策

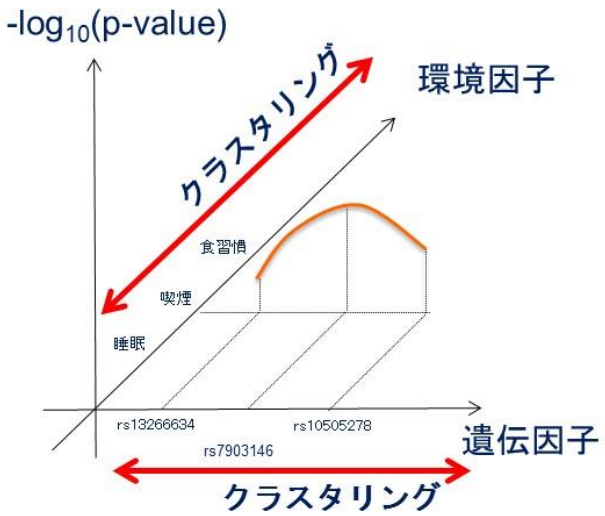
個別化予防：特異的な遺伝子・環境相互作用

Idiosyncratic Effect of Combination of GxE factors

- 遺伝的素因と環境の相互作用
- 相互作用の特異的組合せ効果
  - ハワイの白人、日系人と結腸がん発生
  - **相対リスクの乗算ではない。**
  - Idiosyncratic Effect

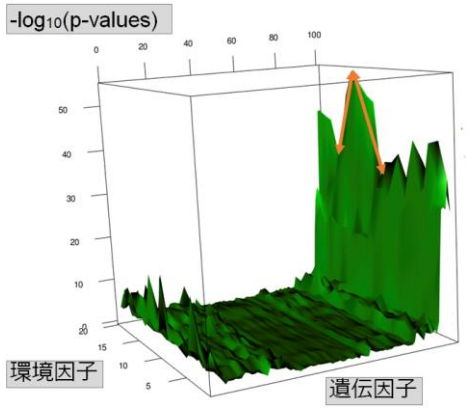
HCA(ヘテロサイクリックアミン, 肉を高温で焼いた時に生成される発癌物質

		CYP1A2 Phenotype ≦Median		CYP1A2 Phenotype ＞Median	
		Likes rare/medium meat	Likes well-done meat	Likes rare/medium meat	Likes well done meat
Non-Smoker	NAT2 Slow	1	1.9	0.9	1.2
	NAT2 Rapid	0.9	0.8	0.8	1.3
Ever-Smoker	NAT2 Slow	1	0.9	1.3	0.6
	NAT2 Rapid	1.2	1.3	0.9	8.8



遺伝因子・環境因子相互作用の同定

シミュレーション実験

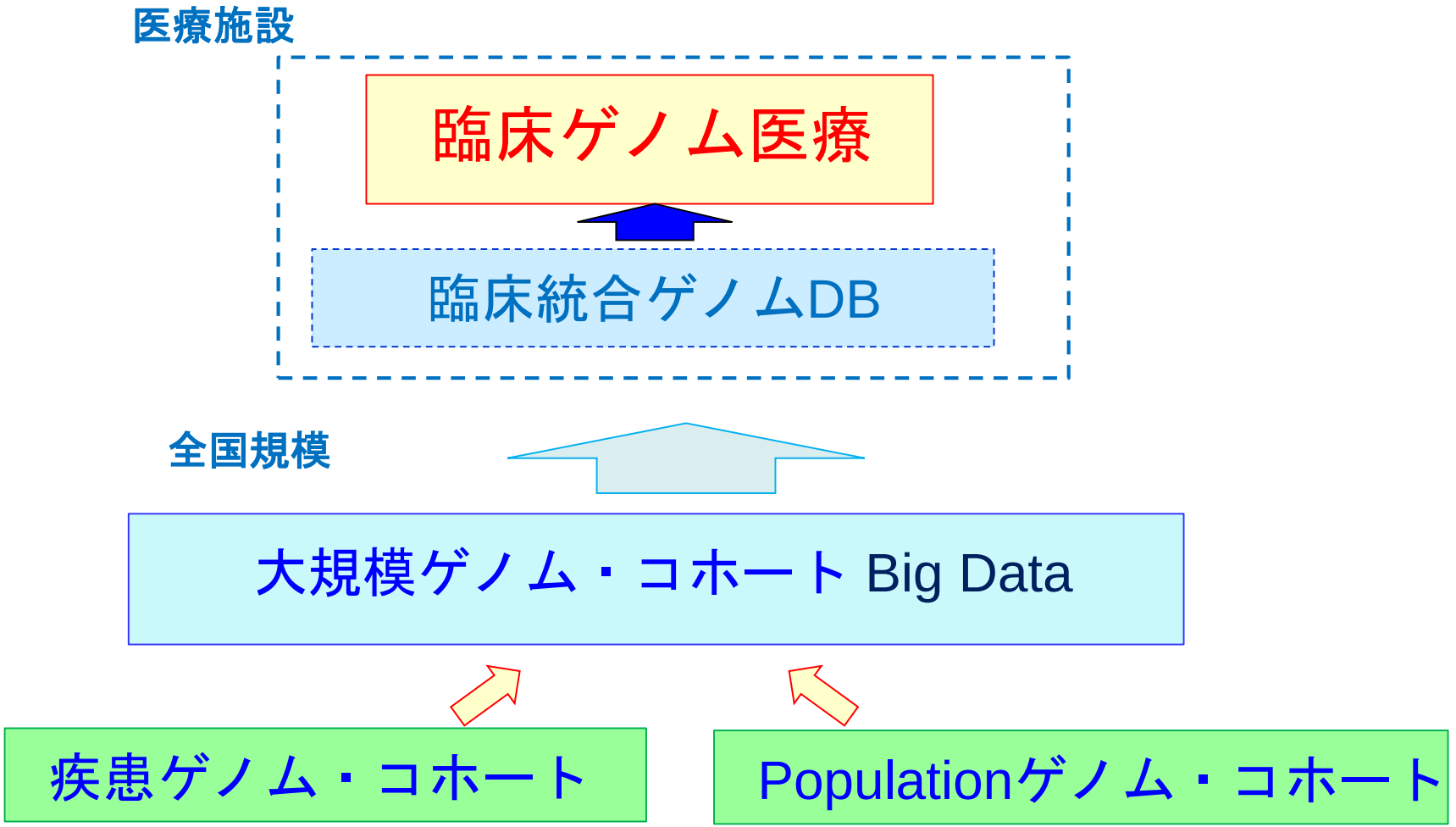


L. Le Marchand, JH. Hankin, LR. Wilkens, et alCombined Effects of Well-done Red Meat, Smoking, and Rapid N-Acetyltransferase 2 and CYP1A2 Phenotypes in Increasing Colorectal Cancer Risk, Cancer Epidemiol. Biomarkers Prev 2001;10:1259-1266

将来

ゲノム・オミックス医療の大規模ビッグデータの形成と知識発見

2つの流れは将来融合して ゲノム医療を支える



米国でのゲノム医療の推移
第２世代化

ゲノムビッグデータ時代の到来（米国）

ゲノム医療の実践

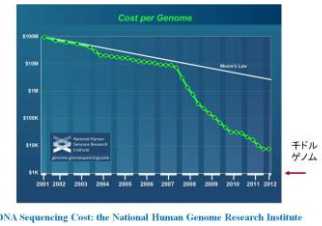
第１段階 ゲノム医療の発展

次世代シーケンシングの臨床普及(2010～)

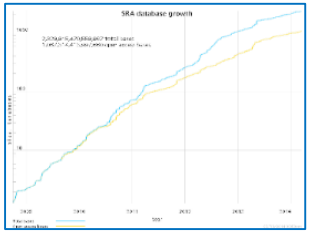
全ゲノム（x30：100Gb)・エキソーム解析（x100：6Gb）

米国では数十の著名病院で実施

ゲノム・オミックス情報の蓄積



2000兆塩基 (2 Pb)
が登録(NCBI:SRA)



医療ビッグデータ

第２段階 医療ビッグデータ時代

医療情報との統合

電子カルテからの
臨床フェノタイプ

ゲノム・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

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

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

第１世代

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

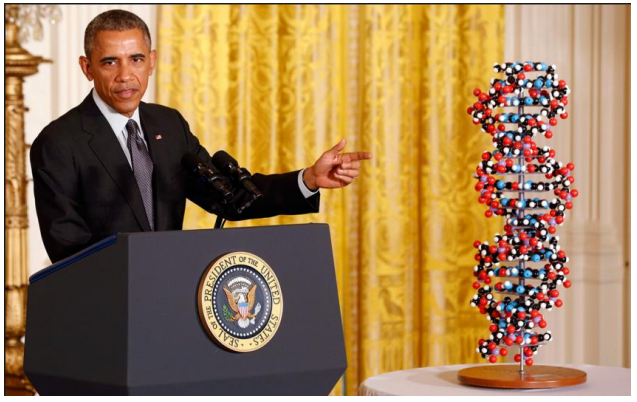
国家戦略としての「医療ビッグデータ」

NIH「ビッグデータから知識へ」計画

- BD2K: ”Big Data to Knowledge” Initiative 開始
  - ゲノム・オミックス医療の普及により、臨床シーケンス情報蓄積の大量化蓄積に対応して政策開始
  - 研究費の配分2013年に提案。計画実施は2014年から
  - Francis Collins長官談「NIH全規模での優先計画」
- NIH：BD2Kの2014年から助成
  - 医療におけるデータ科学の全米COE創設
    - ピッツバーグ, UCSC, ハーバード,**コロンビア大学、イリノイ大学など11施設　32Mドル**
  - Data Scientist 人材養成への予算措置
  - データ発見索引 DDI (Data Discovery Index) Consortium
    - データベースカタログの発展・Pub MEDのDB版
    - UCSDに委託： BioCADDIEを中心にDDI開発の準備を担当
- 米国はすでに戦略的に対応している。わが国は？

Precision Medicineとは何か

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



2015年1月　オバマ大統領一般年頭教書

趣旨：基本は、個別化医療 Personalized Medicine の概念と変わらないが、目指していたのは診断/ 治療の個人化ではなく層別化であることを明確化

概念の拡張：Personalized Medicine提唱時から10数年

医療ビッグデータ時代の到来による個別化医療の拡張

（１）遺伝素因 X 環境(生活習慣)要因のスキーマ重視

SNPや変異（Genome)だけでなく環境・生活習慣要因(Exposome)の重視、疾患発症は2要因の相互作用と明快に強調。臨床表現型（Clinical Phenome)も疾患発症後には不可欠。３つの成因の重視

（２）日常生理モニタリング情報の包摂

モバイルヘルス(mHealth)・ wearable sensor大量継続情報収集の重視

（３）ゲノムコホート・Biobankの重視

Precision Medicineを実現基盤ゲノムコホート/Biobankの重視。
Real world dataの重視

ゲノム医療ビッグデータのための医療人工知能

Sparse ModelingとDeep Learningに共通するもの

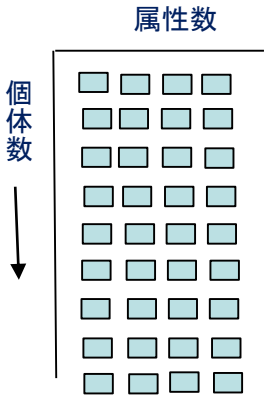
ゲノム医療の「ビッグデータ革命」

～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)



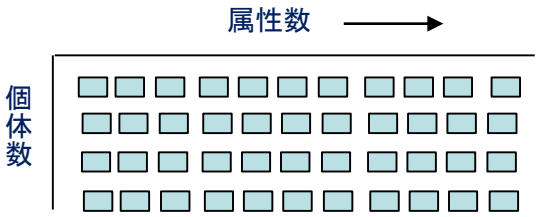
医療情報・疫学調査では 属性数：10項目程度

– 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

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」の多様性を摘出

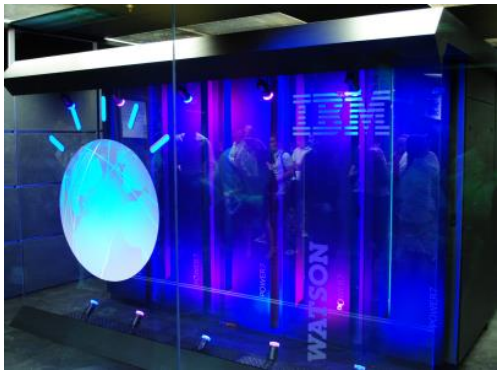
新しいデータ科学の必要性

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、
  - 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治療知識を統計学習、
  - ニューロネット、機械を駆使して知識抽出。

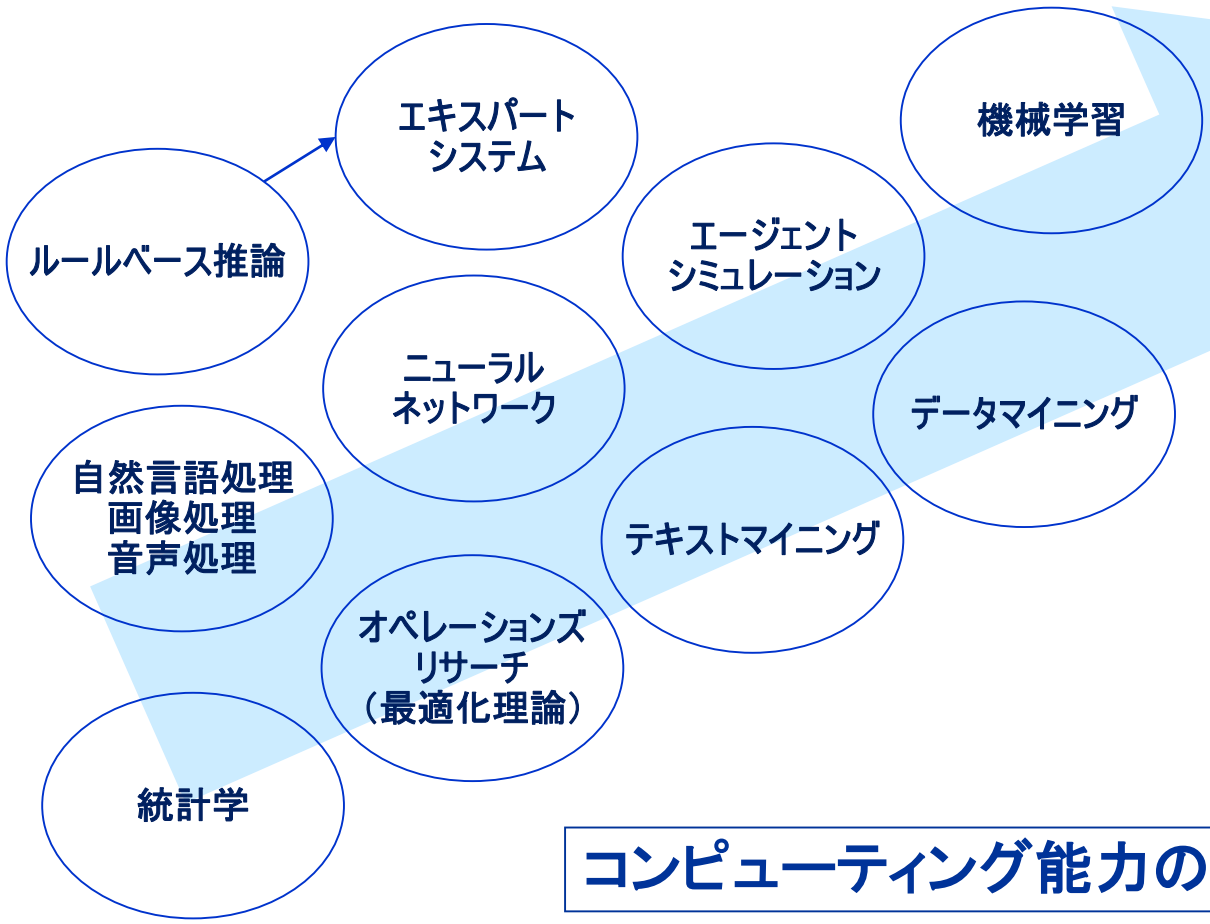
  BigDataにおけるLearning systemの不可欠性
  - 2013年に、CancerLinQのプロトタイプを完成、10万人以上の乳がんを蓄積、完全規模へ継続構築中
- IBM Watsonのがんセンターへの普及
  - Memorial Sloan-Kettering Cancer がんせんたー
  - Watsonを母体にThe Oncology Expert Adviser software 開発
  - New York ゲノムセンターrとグリア芽細胞腫の治療方針
- Google X project, “Human Longevity Inc.”など人工知能の利用

IBM Watson Learning Big Data



人工知能への期待

人工知能（AI）の分野



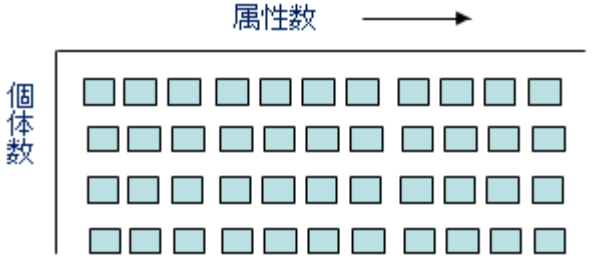
- 機械学習技術に基づくWeb検索は実用化され、有用性が確認されている
- クレジットカードの不正検知や銀行の信用業務に機械学習は利用
- プロセス制御にも機械学習の応用が進出している

TMDU

「ビッグデータ」のData 縮約原理

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

ビッグデータ方法論的スパース仮説

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）

ビッグデータ次元縮約

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)
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

ビッグデータ解析に向けた 2つの機械学習・AI方法の適用

- 数理的なデータマイニング
- 探索的な統計的データ処理の枠内での次元縮約
  ⇒スパースモデリング(疎性モデル）による
  データ行列の強制的「次元落ち」(L_1正則化)
- ニューロネットワーク：
- Deep Learning 特徴量抽出による次元縮約
  ⇒Deep LearningのAutoEncode機能
  を用いた実質的な独立次元の抽出に
  基いた解析・予測

数理的な機械学習の次元縮約
スパースモデルによる次元抑制

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

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

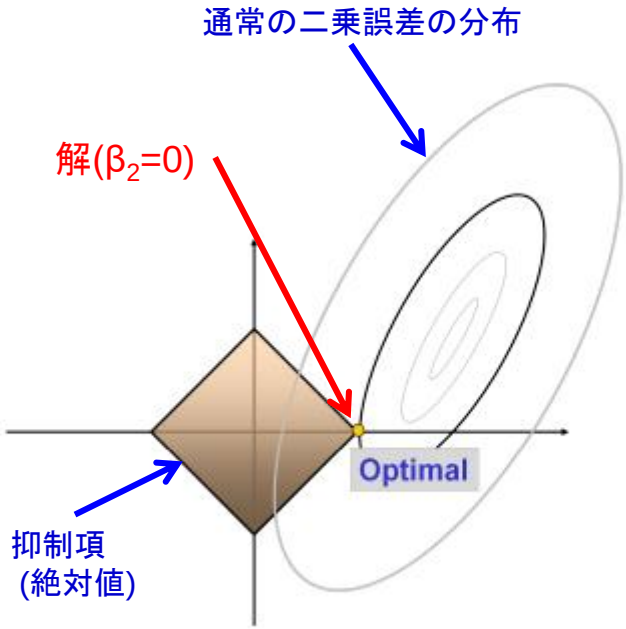
目的変数　　説明変数

Lasso(L_1型正則化重回帰分析)

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

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

通常の二乗誤差項　　次元抑制項（正則化項）



この和を最小にする係数 βj を求める

寄与の低い係数 βj は0になる⇒変数選択と次元落ちが同時に達成できる

様々なスパースモデルの利用

- GWASへの応用

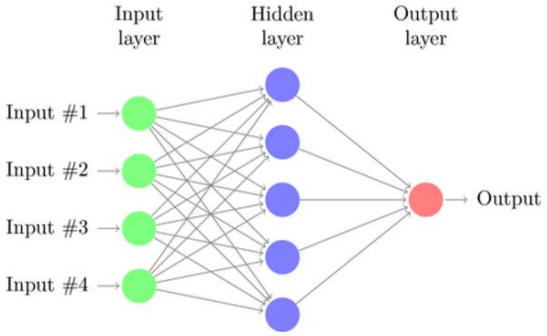
  GWASにおけるgene-gene interactionの取り込み(主効果と相互作用）
  - Correlated SNPs (Ayers and Cordell, 2010)
  - 検出力がさらに増加し、 false-discovery rate (FDR) が低くなった（He and Lin, 2011）
  - Pathwayに含まれているSNP間だけ相互作用を認める（Lu, Latourelle, 2013)
- 遺伝子発現プロファイルへの応用
  - Biomarker（差異的発現遺伝子）が明確化
- 主成分分析にスパース正則化
  - 主成分の解釈が容易になる
  - 次を最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

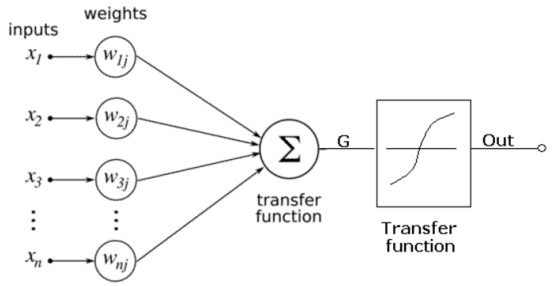
$$Q_\lambda(v_1, X) = \frac{1}{2}\text{trace}[(X - z_1 v_1^T)^T (X - z_1 v_1^T)] + \sum_{j=1}^{p} p_\lambda(|v_{1j}|),$$

- 判別分析でも正則化により次元縮約

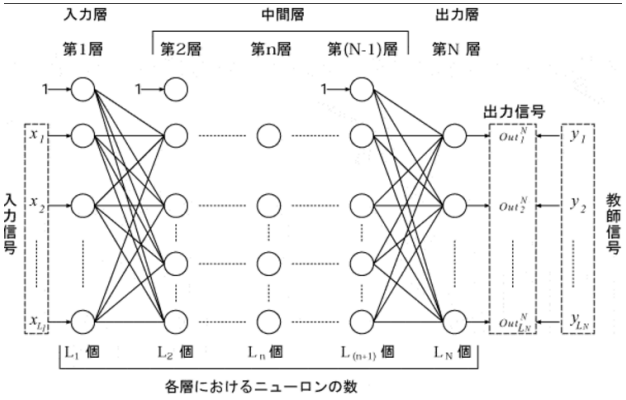
Deep Learningによる基軸成分の抽出

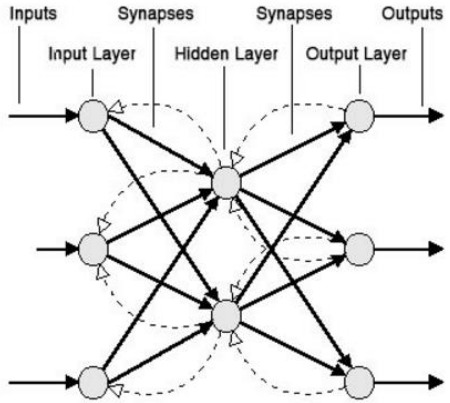
古典的Neural Network(1970年代　Perceptron）





多層Neural NetworkとBack projection (1980年代）



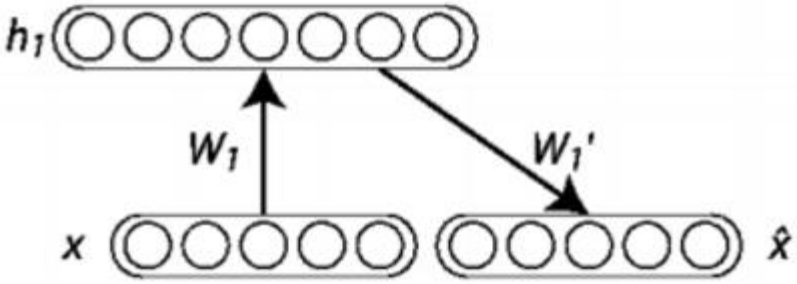


Back Propagation (1986　Rumelhart)
望ましい出力との誤差を教師信号として与える事により、次第に結合係数を変化させ、最終的に正しい出力が得られるようにする。
結合係数を変える事を学習と呼ぶ。
この学習方法には、最急降下法（勾配法）が使われる。出力層へ寄与の高いノードの重みが変更。

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



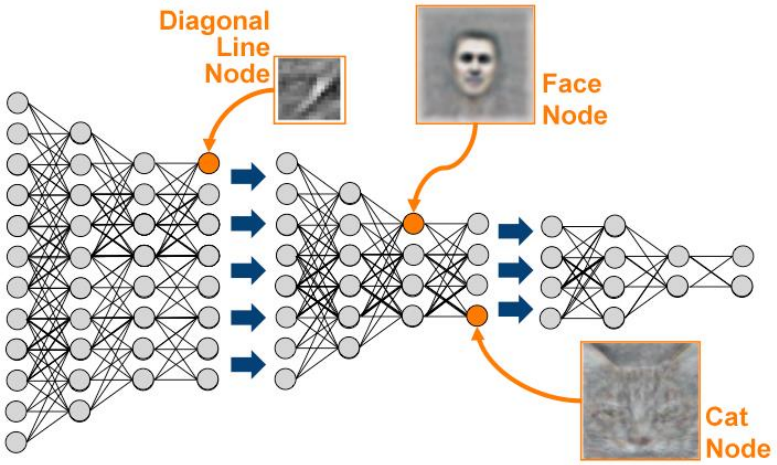
正則化項

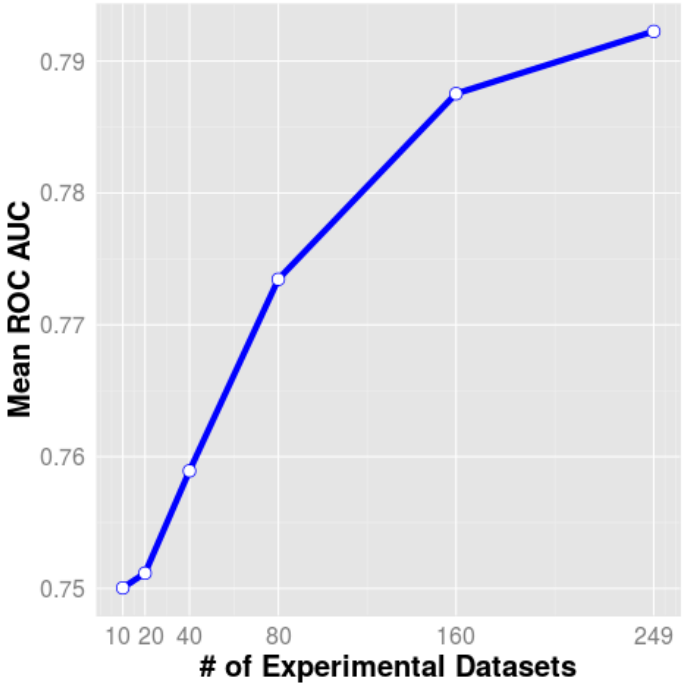
input: X code: $h = W^T X$

loss: $L(X;W) = \|Wh - X\|^2 + \lambda \sum_j |h_j|$

スパース仮説

- スパースモデルも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仮説的原理に基いている。
- 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）データを多数集めれば集めるほど認識が深まる。
- いつまでのOpen性(open-endedness)がある。





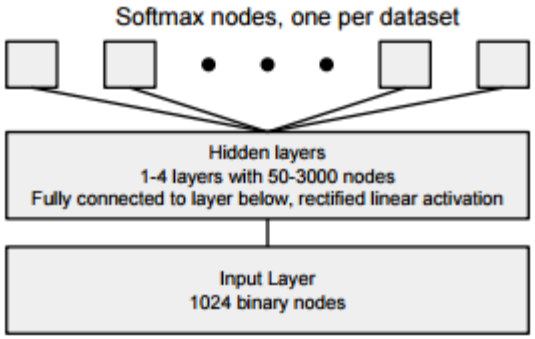
データ数が増加すると内部的な縮約次元が増加し推定精度が上昇する

医療の「ビッグデータ」革命は どんな既存のパラダイムに挑戦しているか

- Population medicineのパラダイム転換
  - <One size fits for all>のPopulation医療はもはや成り立たない
  - 個別化医療 “Personalized (Precision) medicine”
    - 個別化医療を実現するために＜個別化・層別化パターン＞を網羅的に調べる：どこまでの粒度で個別化・層別化すればよいか
- Clinical research（臨床研究）のパラダイム転換
  - 臨床研究を科学にする従来の基準RCTは,個別化概念に破綻した
  - <statistical evidence based>呪縛からの解放
  - 「標本」統計・「推測」統計学に限定されない臨床研究
  - Real World Data:ビッグデータ知識生成（BD2K)
- 創薬の戦略パラダイムの転換
  - ビッグデータ創薬の可能性
  - 創薬・育薬のReal World Dataの利用

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　人工知能の利用
- 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 Standford (2015)
  - Stanford 第学の 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

AI（人工知能）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
    - Deep Learningで構造活性相関を学習する
    - Ligand-based 標的予測,7種の予測法とAUC比較
    - Deep learningは、 SVM, k-近隣法, logistic回帰などより優位
    - 特徴量の抽出、薬理機構の理解達成
  - リード化合物 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)

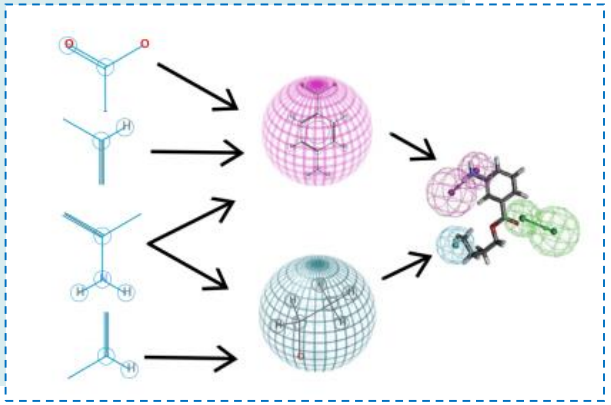


Figure Hierarchical nature of fingerprint features: by combining the ECFP features we can build reactive centers. By pooling specific reactive centers together we obtain a pharmacophore that encodes a specific pharmacological effect.

ご清聴ありがとうございました