

遺伝学入門

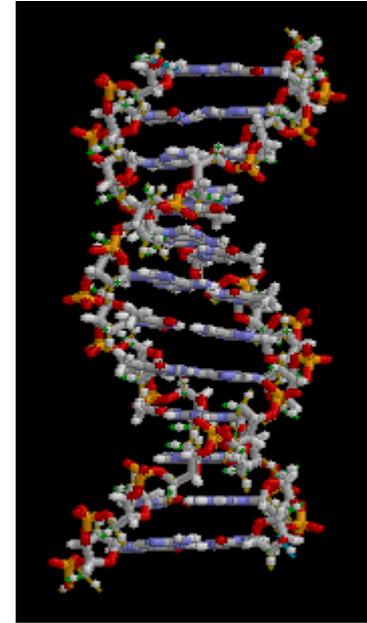
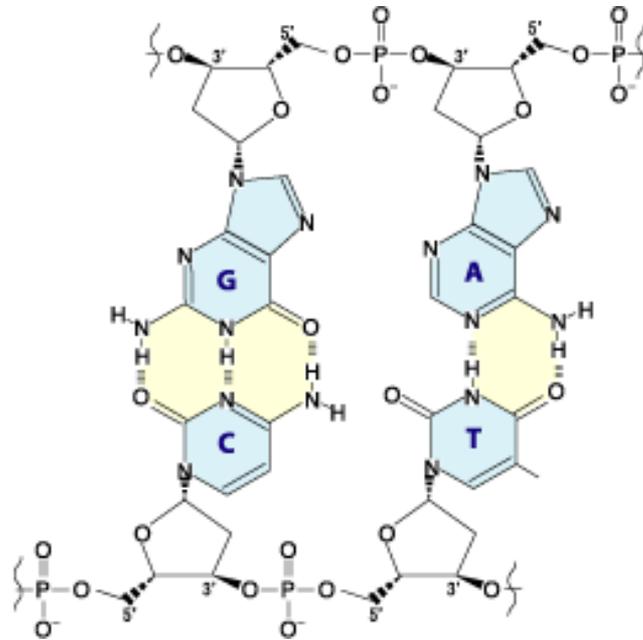
大阪大学大学院医学系研究科 遺伝統計学
東京大学大学院医学系研究科 遺伝情報学
理化学研究所生命医科学研究センター システム遺伝学チーム

<http://www.sg.med.osaka-u.ac.jp/index.html>

遺伝学入門

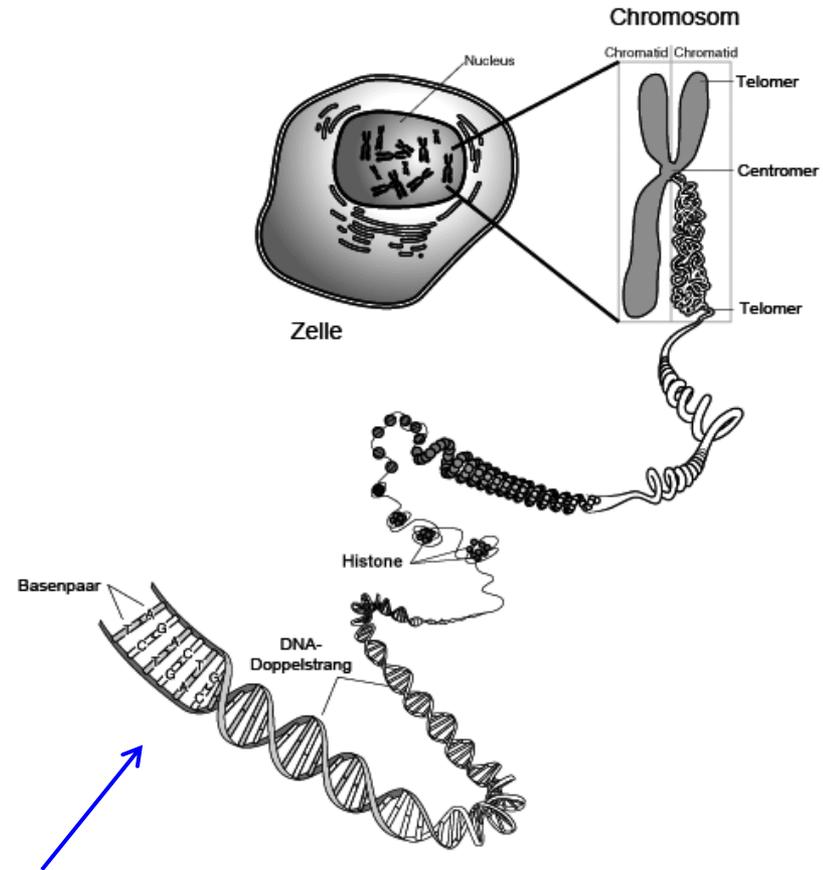
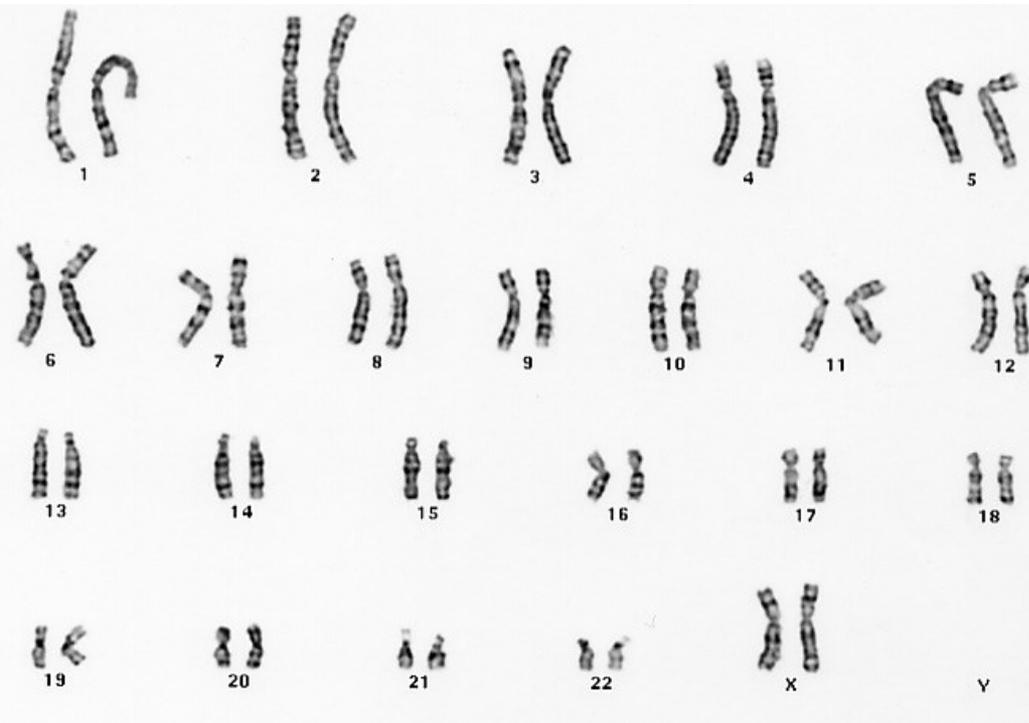
- ① **ヒトゲノム配列とその個人差**
- ② **ヒトゲノム多型を用いた疾患ゲノム解析**
- ③ **疾患病態解明への展望**
- ④ **機械学習のヒトゲノム情報への応用**
- ⑤ **ゲノム創薬への展望**

① ヒトゲノム配列とその個人差



- 生物のゲノム情報は、A(アデニン)、T(チミン)、G(グアニン)、C(シトシン)の4種類の塩基(DNA)で構成された塩基配列として、保存されています。
 - 塩基配列は、AがTと、GがCと、水素結合を介してつながることにより、相補的な2本鎖として存在します。
- (G-Cは水素結合×3、A-Tは水素結合×2なので、前者の方が結合力が強いです。)

① ヒトゲノム配列とその個人差



AGCTGATAGAT...TAACCGTATCC

- ヒトゲノム情報は、**46本の染色体**(=22*2+X/Y)に分かれています。
- 各染色体の中に、A、T、G、Cで構成された**塩基配列**が細かく折りたたまれています。

① ヒトゲノム配列とその個人差



- 2本鎖DNAの模式図として、正しいのはどれでしょうか？
- 2本鎖DNAは「右巻き」と「一回転あたり4塩基対」で構成されます。
(間違った図をよく見かけるので、注意です)

① ヒトゲノム配列とその個人差

ショウジョウバエ

8本



オオムギ

14本



ネコ

38本



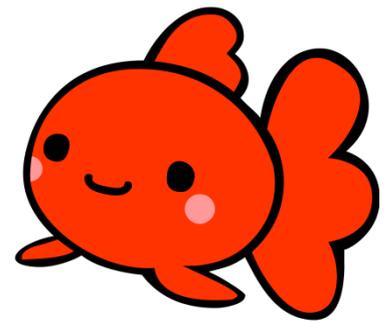
コムギ

42本



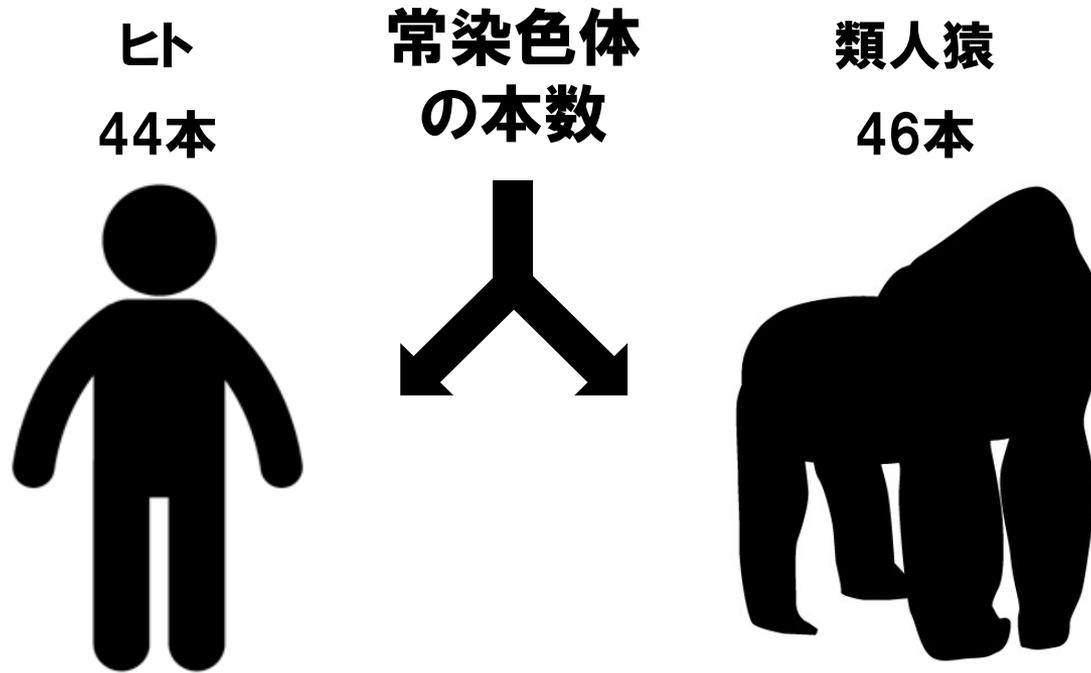
キンギョ

104本



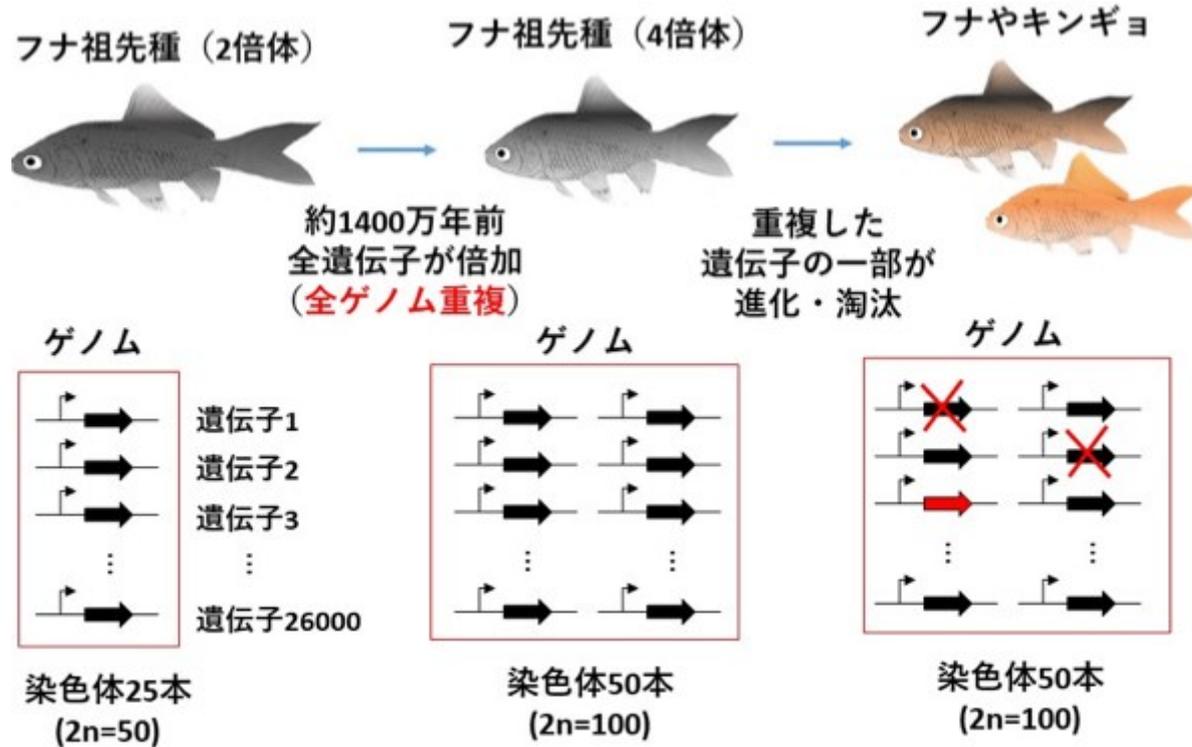
- ゲノム塩基配列や、染色体の本数は、生物種によって異なります。
- 染色体が多いほど(ゲノム配列が長いほど)高等な種、というわけでもありません。

① ヒトゲノム配列とその個人差



- ゲノム配列は安定した情報ですが、細胞分裂や、子へ受け継がれる際に、ゲノム配列の変異を生じることがあります。
- その結果、種間や個人間でのゲノム配列の違いが生じます。
- ヒトの常染色体は44本ですが、類人猿(チンパンジー・ゴリラ・オラウータン)の染色体は46本です。共通祖先からヒトへの進化の過程で、ヒトでは二つの染色体が一つに融合したのでは、と考えられています。

① ヒトゲノム配列とその個人差



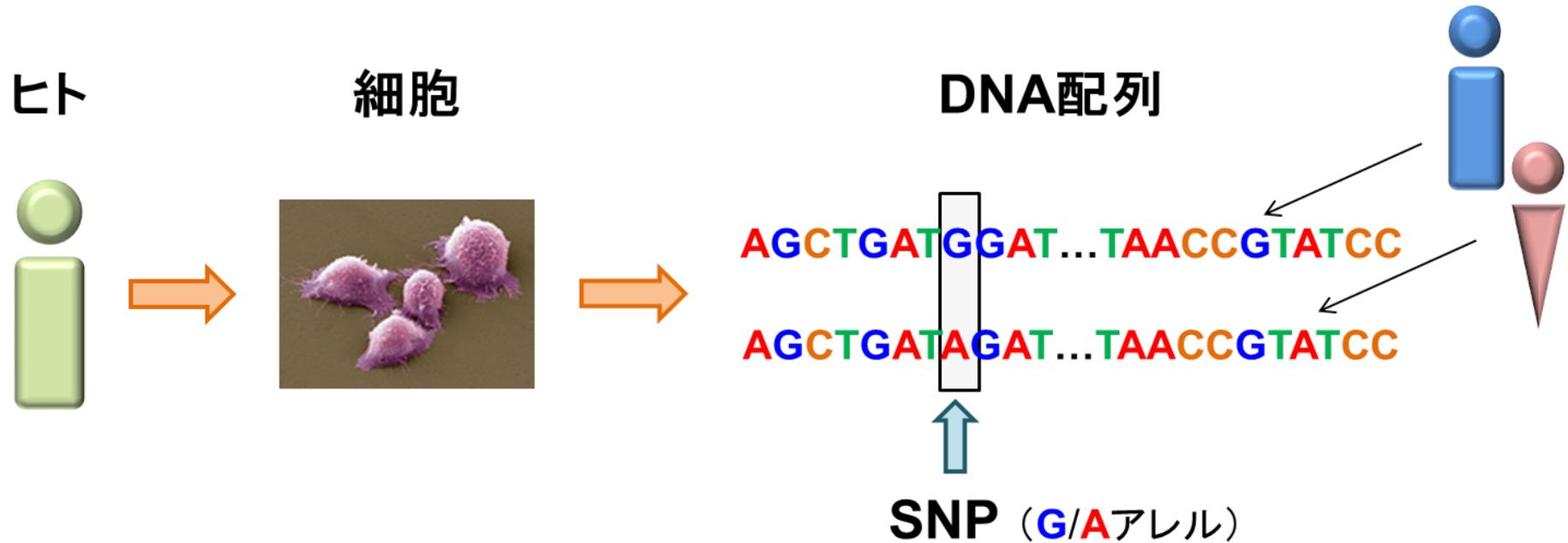
- **金魚の染色体数が多いのは、過去に全ゲノム重複(Whole Genome Duplication: WGD)と呼ばれる全遺伝子が倍加する現象が生じたからです。**
- **ゲノム長が倍だと複製コストが増加する一方、表現型の多様性の獲得や、劣性遺伝形質の出現頻度の低さなど、メリットもあるようです。**
- **非対称サブゲノム進化、という現象として知られています。**

① ヒトゲノム配列とその個人差

The screenshot shows the Science journal website interface. At the top, the Science logo is prominent, with 'AAAS' to its right. Navigation tabs include Home, News, Journals, Topics, and Careers. A search bar is located on the right. Below the navigation, a list of journal categories is provided: Science, Science Advances, Science Immunology, Science Robotics, Science Signaling, and Science Translational Medicine. The main content area features a 'SHARE' section with social media icons (Facebook, Twitter, Google+) and a 'SPECIAL REVIEWS' section for the article 'The Sequence of the Human Genome'. The article title is followed by a long list of authors. To the right of the article, there is a 'Science' journal cover image and a 'Science' sidebar with volume and issue information (Vol 291, Issue 5507, 16 February 2001) and a 'Table of Contents' link. Below the article, an 'ARTICLE TOOLS' section lists various options: Email, Print, Alerts, Citation tools, Download Powerpoint, Save to my folders, Request Permissions, and Share.

- ヒトゲノムは約30億の塩基配列で構成されています。
(父親由来、母親由来のゲノムを持つので、実際には30億×2通り)
- ヒトゲノムプロジェクトにより、ヒトゲノム配列の全容が明らかになったのは、2000年代初頭のことでした。

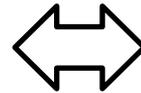
① ヒトゲノム配列とその個人差



- ヒトゲノム上の塩基配列は、ヒト個人間/個人内で少しずつ異なります。
- ヒトゲノム配列の個人差を、**多型**(polymorphism)と呼びます。
- 最も代表的なのは、一塩基単位で変異が生じる、**一塩基多型**(SNP: single nucleotide polymorphism)です。

① ヒトゲノム配列とその個人差

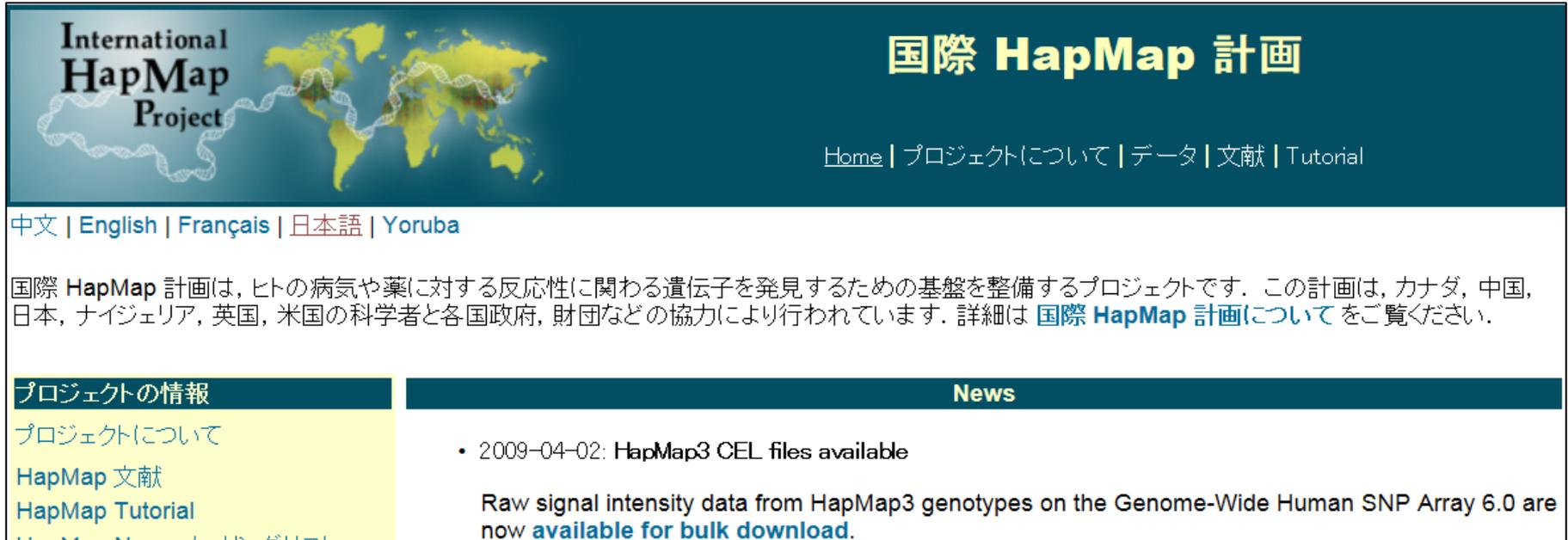
- ヒトゲノムが同じだと、見た目もそっくりです → 一卵性双生児
- ヒトゲノムが少し似ていると、見た目も少し似ています → 兄弟姉妹・親子
- ヒトゲノムが少し違うと、見た目も少し違います → 友人
- ヒトゲノムがもう少し違うと、見た目ももう少し違ってきます → 日本人と欧米人



- ゲノム配列が似ていると、外見だけでなく、**病気のかかりやすさ**や**薬の効きやすさ**も似ていることがわかっています。
- ヒトゲノム多型の情報に基づく、疾患や医療の研究が行われています。

① ヒトゲノム配列とその個人差

国際プロジェクトによるヒトゲノム個人差の解明：①



International HapMap Project

国際 HapMap 計画

Home | プロジェクトについて | データ | 文献 | Tutorial

中文 | English | Français | 日本語 | Yoruba

国際 HapMap 計画は、ヒトの病気や薬に対する反応性に関わる遺伝子を発見するための基盤を整備するプロジェクトです。この計画は、カナダ、中国、日本、ナイジェリア、英国、米国の科学者と各国政府、財団などの協力により行われています。詳細は [国際 HapMap 計画について](#) をご覧ください。

プロジェクトの情報

プロジェクトについて

HapMap 文献

HapMap Tutorial

News

- 2009-04-02: HapMap3 CEL files available

Raw signal intensity data from HapMap3 genotypes on the Genome-Wide Human SNP Array 6.0 are now [available for bulk download](#).

- 2003年に、International HapMap Projectにより、**欧米/アジア/アフリカ人270名**におけるヒトゲノム多型の概要が解読されました。
- 約200～300万のSNPが明らかになり、これまで手探りで行っていたヒト疾患感受性遺伝子の同定を網羅的に行うことが可能になりました。
- プロジェクト参加国の中で、日本は一番の貢献を果たしています。

① ヒトゲノム配列とその個人差

国際プロジェクトによるヒトゲノム個人差の解明：②

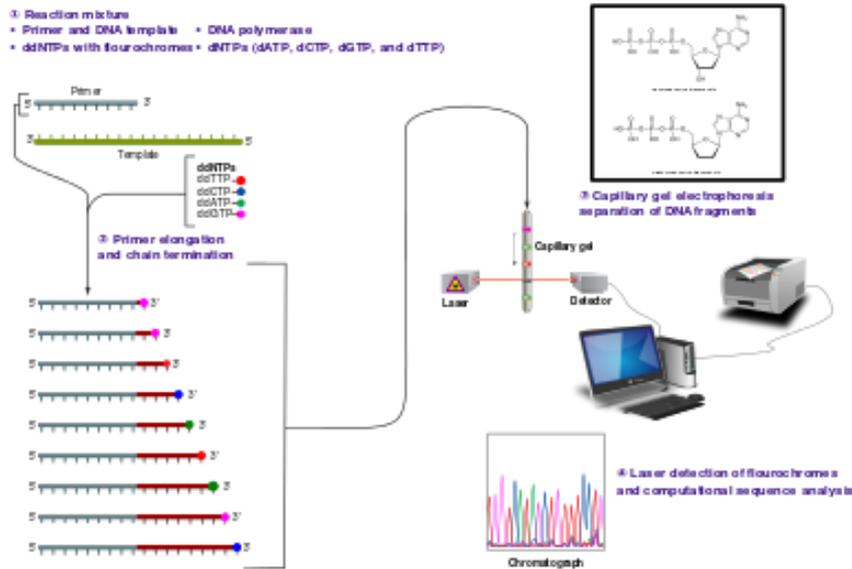


- 2010年に1000 Genomes Projectにより、ヒトゲノム多型(SNP+ α)の全容が解読されました。
- 複数集団2,500名における、約1億個のSNPが明らかになり、集団中での頻度の低いレアバリエントのカタログが公開されています。
- ヒトゲノム配列の解読には、次世代シーケンサーが使用されました。

(<https://www.internationalgenome.org/>, The 1000 Genomes Project Consortium. *Nature* 2012/2015)

① ヒトゲノム配列とその個人差

サンガー法



次世代シーケンサー



- ゲノム配列解読技術の発達が、多型情報の解読を推進しました。
- サンガー法による手作業での配列解読作業から、次世代シーケンサーによる大規模自動的な解読作業へと、効率が劇的に変化しました。
(サンガー法自体は高精度の配列解読手段として、現在でも活躍しています)

① ヒトゲノム配列とその個人差

国際プロジェクトによるヒトゲノム個人差の解明: ③

T2T Consortium

Home In the News Technology Data Publications CHM13 Cell Line Who We Are Join Us

The Telomere-to-Telomere (T2T) consortium is an open, community-based effort to generate the first complete assembly of a human genome.

Release of the first human genome assembly was a landmark achievement, and after nearly two decades of improvements, the current human reference genome (GRCh38) is the most accurate and complete vertebrate genome ever produced. However, no one chromosome has yet been finished end to end, and hundreds of gaps persist across the genome. These unresolved regions include segmental duplications, ribosomal rRNA gene arrays, and satellite arrays that harbor unexplored variation of unknown consequence. We aim to finish these remaining regions and generate the first truly complete assembly of a human genome. The ultimate goal of this effort is to drive technology to dramatically increase the throughput of complete, high quality telomere-to-telomere assemblies from diploid human genomes.

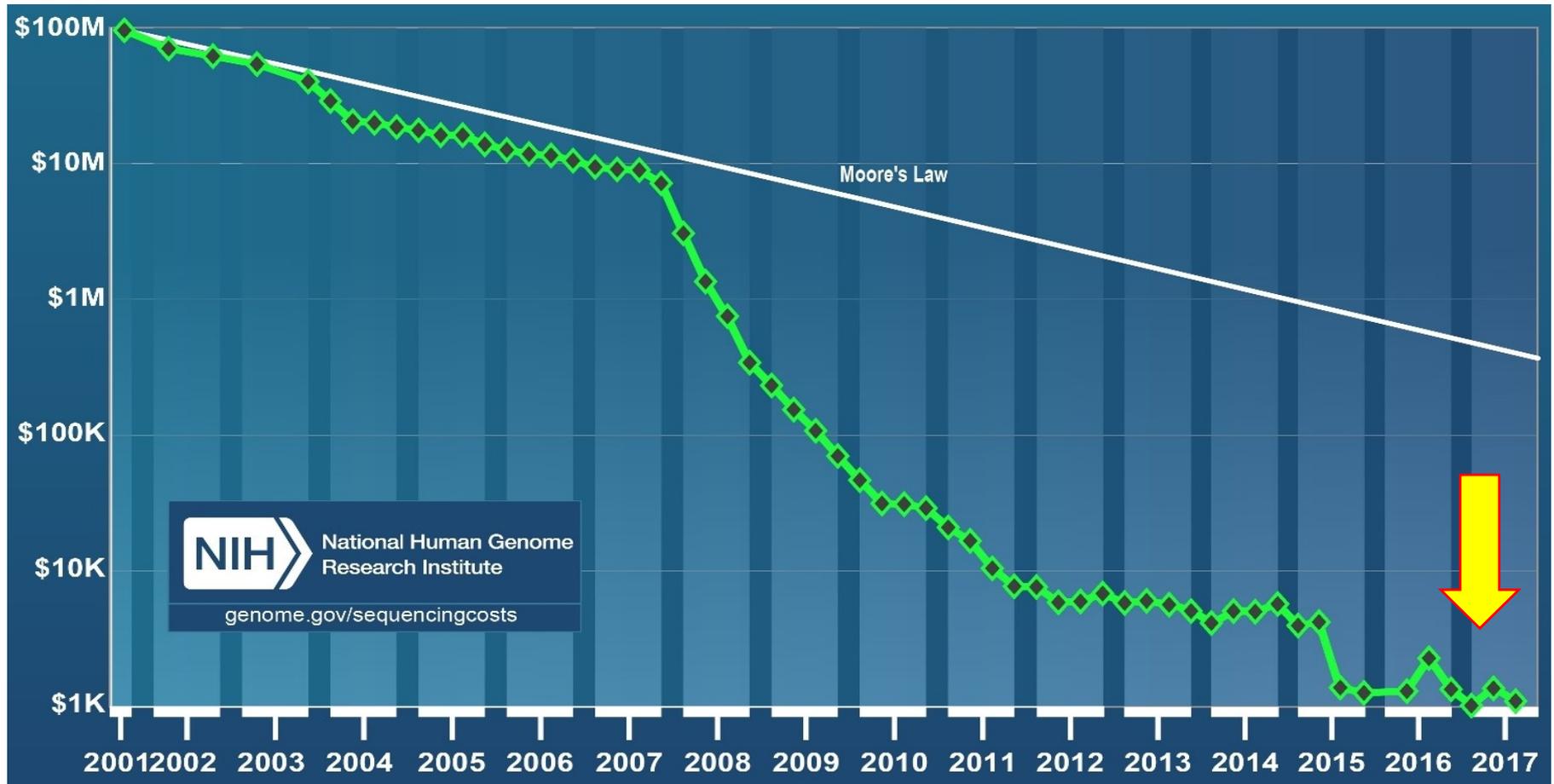
Join Us!

- **ロングリード型次世代シーケンサー**により、**難読ゲノム領域**(例:セントロメア・テロメア・ヘテロクロマチン)の配列解読が進んでいます。
- 2022年4月に、The Telomere-to-Telomere (T2T) consortiumがヒトゲノム完全解読論文を報告しています。

(<https://sites.google.com/ucsc.edu/t2tworkinggroup/home>, Nurk S et al. *Science* 2022)

① ヒトゲノム配列とその個人差

ゲノム配列情報取得コストの劇的な低下



- ヒトゲノム配列の解読コストが、**\$1,000/サンプル**が達成されました。
- 現在、国内企業の外注でも**\$600/サンプル**が達成可能です。

① ヒトゲノム配列とその個人差

ゲノム配列情報取得コストの劇的な低下

GENOMICS

Upstart DNA sequencers could be a 'game changer'

Increasing potential for fast, cheap genomes may break open biology's bottleneck and broaden clinical uses

NEWS | BIOLOGY

A \$100 genome? New DNA sequencers could be a 'game changer' for biology, medicine

"This is the year of the big shake-up."

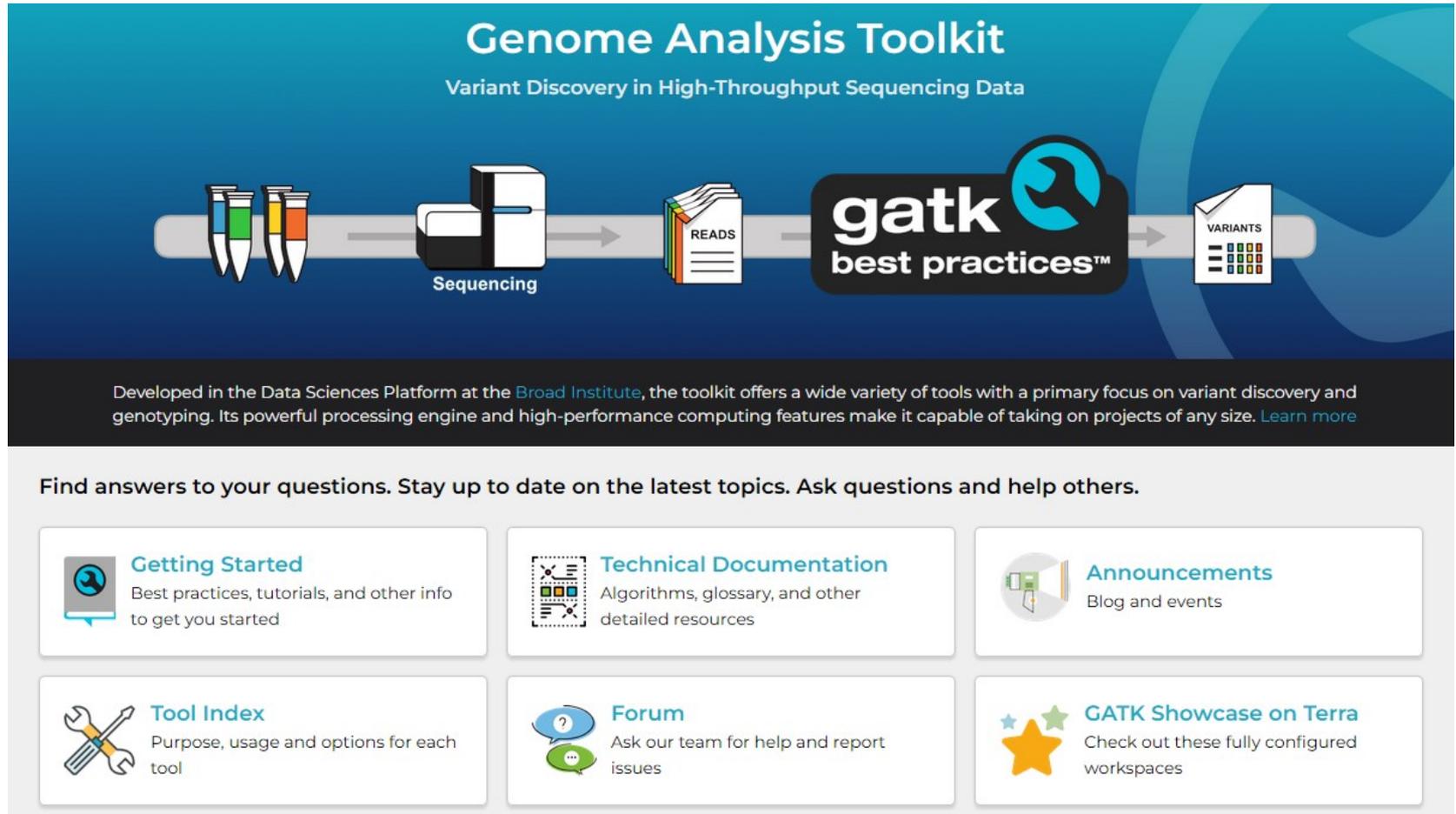
15 JUN 2022 • 1:15 PM • BY ELIZABETH PENNISI



Ultima Genomics社製
シーケンサー

- 次世代シーケンスの大半を担うイルミナ社の特許存続期間終了に伴い、更なる企業の参入とシーケンスコストの低下が見込まれています。
- **\$100/サンプル**(=100ドルゲノム)の実現も近いと言われています。

① ヒトゲノム配列とその個人差



Genome Analysis Toolkit
Variant Discovery in High-Throughput Sequencing Data

Sequencing

gatk best practices™

VARIANTS

Developed in the Data Sciences Platform at the [Broad Institute](#), the toolkit offers a wide variety of tools with a primary focus on variant discovery and genotyping. Its powerful processing engine and high-performance computing features make it capable of taking on projects of any size. [Learn more](#)

Find answers to your questions. Stay up to date on the latest topics. Ask questions and help others.

- Getting Started**
Best practices, tutorials, and other info to get you started
- Technical Documentation**
Algorithms, glossary, and other detailed resources
- Announcements**
Blog and events
- Tool Index**
Purpose, usage and options for each tool
- Forum**
Ask our team for help and report issues
- GATK Showcase on Terra**
Check out these fully configured workspaces

• 全ゲノムシーケンスの精度が高くなり、解析パイプラインの改良により、**シーケンスデータの信頼性も高くなってきました。**

• 米国Broad研究所が提供する**GATK best practices**が有名です。

(<https://gatk.broadinstitute.org/hc/en-us>)

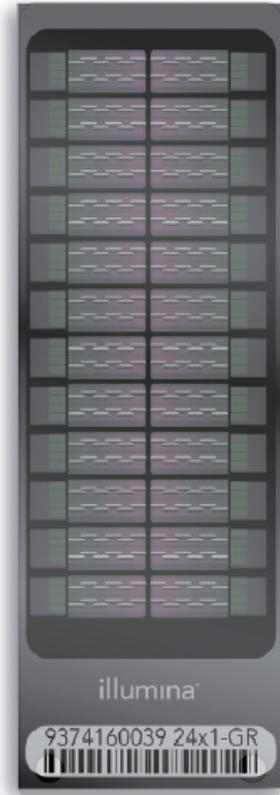
① ヒトゲノム配列とその個人差

ゲノム配列情報取得コストの劇的な低下

- インピュテーション^{*1} 精度を最大化する SNP を搭載
- 疾患研究を加速するクリニカルリサーチ SNP を搭載
- 大規模解析を支援する低価格を実現
- 日本人を含むアジア人の低頻度バリエーションまで検出 (ASA)

※1 インピュテーション:連鎖不平衡 (LD) となっている SNP 間の相関を利用することで、コンテンツが異なるマーカーセット間でのジェノタイピングデータの推測をすること。GSA、ASA は対象人種集団にて低頻度の多型も高精度にインピュテーションできる SNP を選択しています。

複数の遺伝的要因と環境要因が関与する多因子疾患は、頻度の低い多型が関与していることが多く、サンプルサイズが限られる解析での原因説明は極めて困難とされてきました。しかし近年、これまで対象群と有意差が検出できなかった疾患に対し、整備されたバイオバンクのサンプルを活用して、対象疾患のサンプル数を多く確保することで、疾患感受性座位が特定される例が報告されています。イルミナではそのような大規模アプローチに最適な、対象人種集団の異なる2つのアレイを設計しました。



Infinium Global Screening Array
24 サンプル / スライド

- ゲノム全体のSNPデータは、**\$40/サンプル**が達成されています。
- これは、**ゲノムDNA抽出費用とほぼ同じ値段**になります。

① ヒトゲノム配列とその個人差

ヒトゲノム研究の大規模化

UKバイオバンク

(<https://www.ukbiobank.ac.uk/>)

TopMedプロジェクト

(<https://topmed.nhlbi.nih.gov/>)

The screenshot shows the UK Biobank website. At the top, there are navigation links for 'Researcher log in', 'Participant log in', and 'Contact us'. Below this is a search bar and a main banner with the text 'Enabling your vision to improve public health'. The banner includes a description of the database and two buttons: 'Data Showcase' and 'Future data releases'. Below the banner is a large text block describing the UK Biobank as a large-scale biomedical database and research resource, containing in-depth genetic and health information from half a million UK participants. At the bottom, there are several smaller images and a section titled 'How GP data helps UK Biobank improve human health' with a date 'Wednesday 18th August 14:00 - 15:00'.

The screenshot shows the NHLBI Trans-Omics for Precision Medicine (TOPMed) website. At the top, there is a logo and the text 'NHLBI Trans-Omics for Precision Medicine'. Below this is a navigation menu with links for 'Centers', 'Projects/Studies', 'Working Groups', 'Data', 'Publications', 'EAP', 'ELSI', and 'Workshops'. The main content area is titled 'About TOPMed' and includes a date 'Updated 11/30/2020'. There is a 'Contents' section with a list of links: 'Overview', 'Study Characteristics', 'Study Designs', 'Participant Diversity', 'Whole Genome Sequencing', and 'Resources for the Scientific Community'. Below this is an 'Overview' section with a paragraph of text. On the right side, there is a login form with fields for 'Username' and 'Password', a 'Log in' button, and a link to 'Create new user account'.

- ヒトゲノム研究の大規模化を引き起こしています。
- UKバイオバンクは最大**50万人**のGWAS・全エクソーム・全ゲノムシーケンス(WGS)、TopMedプロジェクトは**10万人**のWGSを構築済みです。
- UKバイオバンクは、**ゲノム・表現型情報を安価に分譲**しています。

① ヒトゲノム配列とその個人差

ヒトゲノム研究の商業化

次世代シーケンサー

大学



企業

- ・次世代シーケンサー等のヒトゲノム解読機器を、研究施設だけでなく企業が保有する例が増えています。
- ・年々更新される高額機器を購入し、ペイするレベルで運用を続けるのは、**研究施設より企業の方が有利**となりつつあります。

① ヒトゲノム配列とその個人差

イルミナ社 HiSeq X Tenシリーズ(1000ドルゲノムを達成)



大規模計算機による
ビッグデータ解析



Precision Medicine
(個別化医療)



・大容量のゲノムデータが低コストで得られる時代が到来し、医療現場も巻き込む形で、ヒトゲノムを巡る研究開発が進められています。

遺伝学入門

- ① ヒトゲノム配列とその個人差
- ② ヒトゲノム多型を用いた疾患ゲノム解析
- ③ 疾患病態解明への展望
- ④ 機械学習のヒトゲノム情報への応用
- ⑤ ゲノム創薬への展望

② ヒトゲノム多型を用いた疾患ゲノム解析

ゲノム配列解読



大規模ゲノムデータ



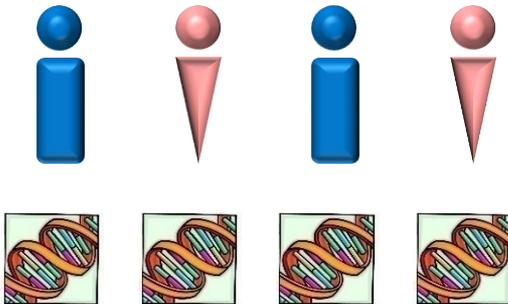
プログラミング

```
out = Log_for_NonparametricIBUmapping+n#nInput_#n
FileTools.out3File(bw[0], out);←

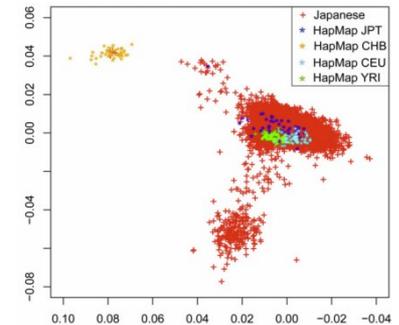
for (int i=0; i<numSNP; i++) {←
> String[] tmp = DataMap[i].split("#t");←
> SNPID[i] = tmp[1];←
> Chr[i] = Integer.parseInt(tmp[0]);←
> bp[i] = Integer.parseInt(tmp[3]);←
> Position[i] = Long.parseLong(tmp[0])*100000000
}←

for (int i=0; i<numPoint; i++) {←
> String[] tmp = DataMap[i].split("#t");←
> PointID[i] = tmp[1];←
> PointChr[i] = Integer.parseInt(tmp[0]);←
}
```

ヒトゲノムサンプル



統計解析



- **遺伝統計学**: 遺伝情報と形質情報の結びつきを統計学により評価
- ゲノムデータへの遺伝統計解析を行う、(主に)ドライの研究活動を展開

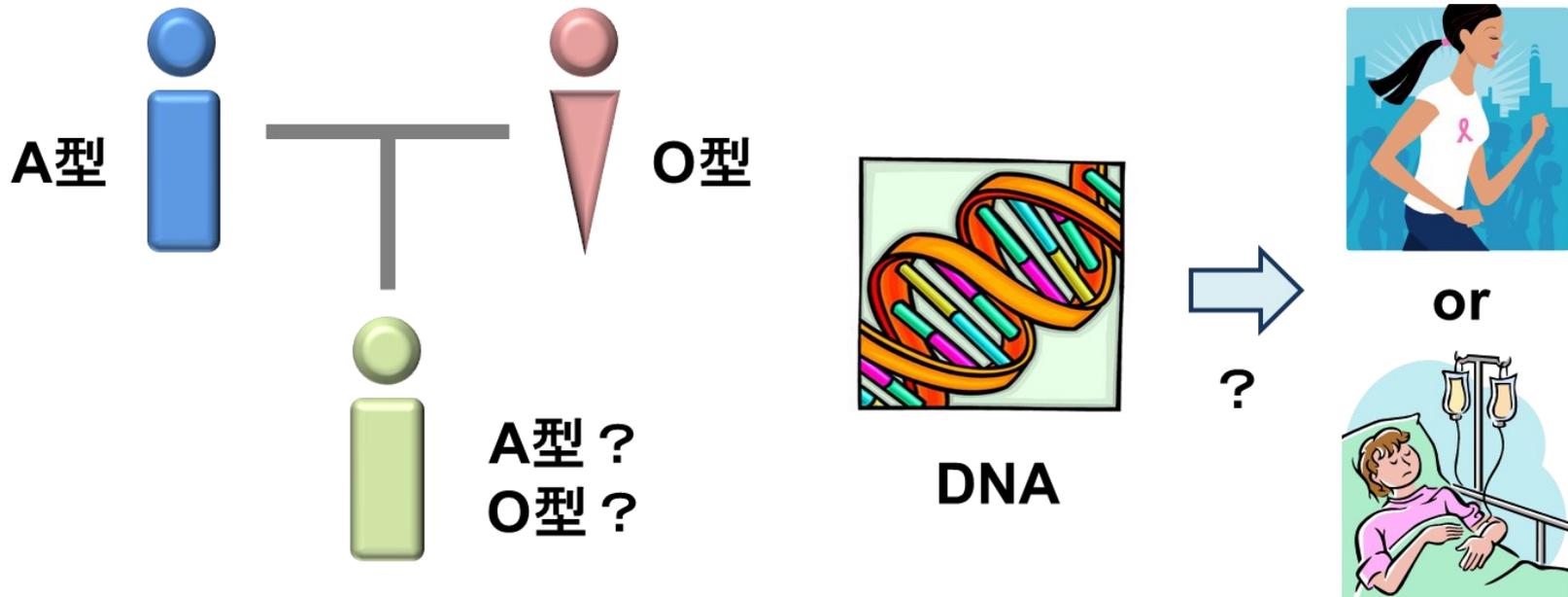
② ヒトゲノム多型を用いた疾患ゲノム解析

・**遺伝情報**・・・生物の細胞の中に、DNAとして保存されている。

ヒト個人間ではほとんど同じだが(99.9%)、少しずつ違う。

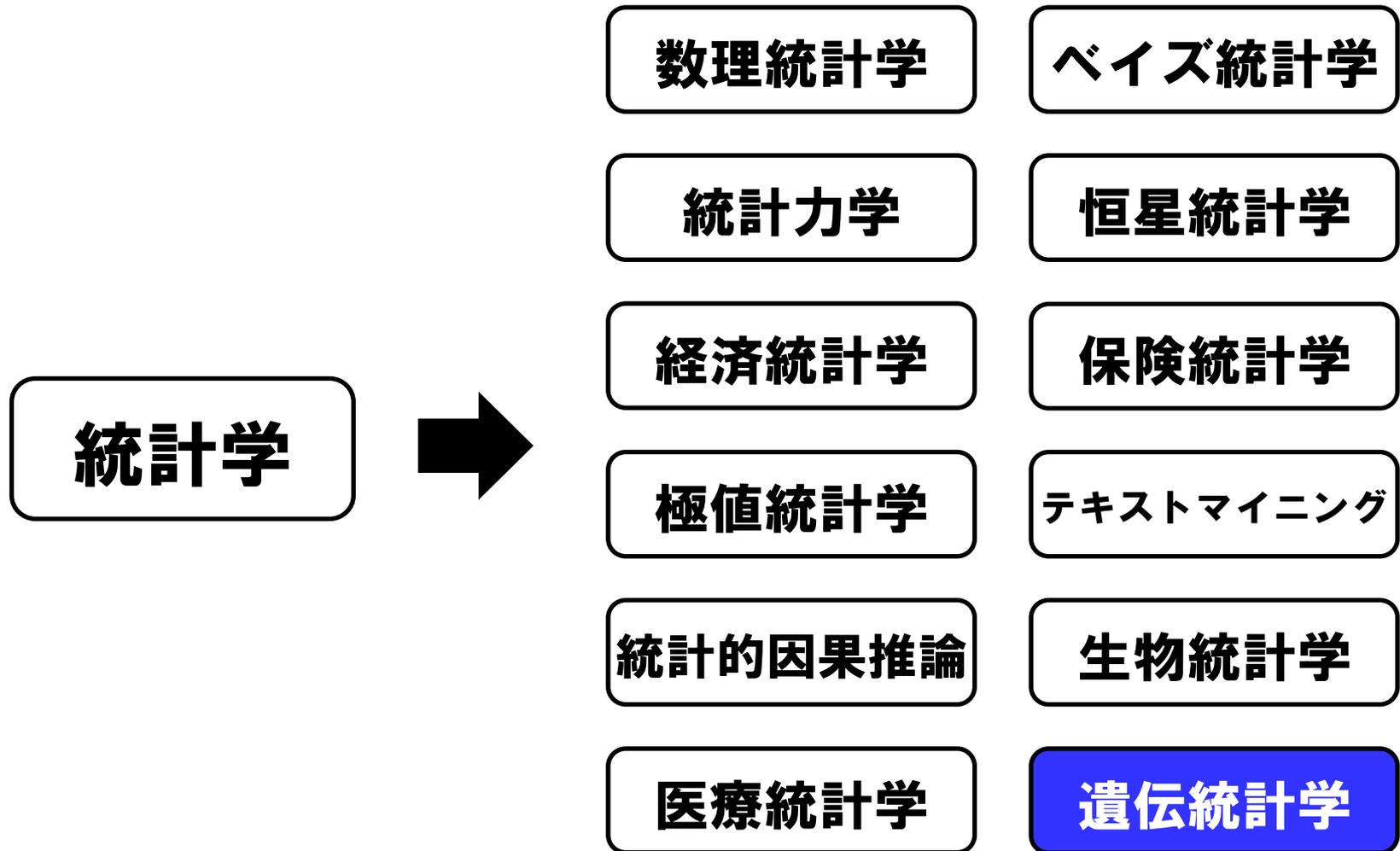
・**形質情報**・・・ヒトから得られる情報。個人間で違うものが対象となる。

(血液型、性別、身長・体重、病気 etc.)



・「**遺伝情報**」と「**形質情報**」との結びつきを、統計解析を通じて評価することで、疾患の原因となっているゲノム多型を同定することができます。

② ヒトゲノム多型を用いた疾患ゲノム解析



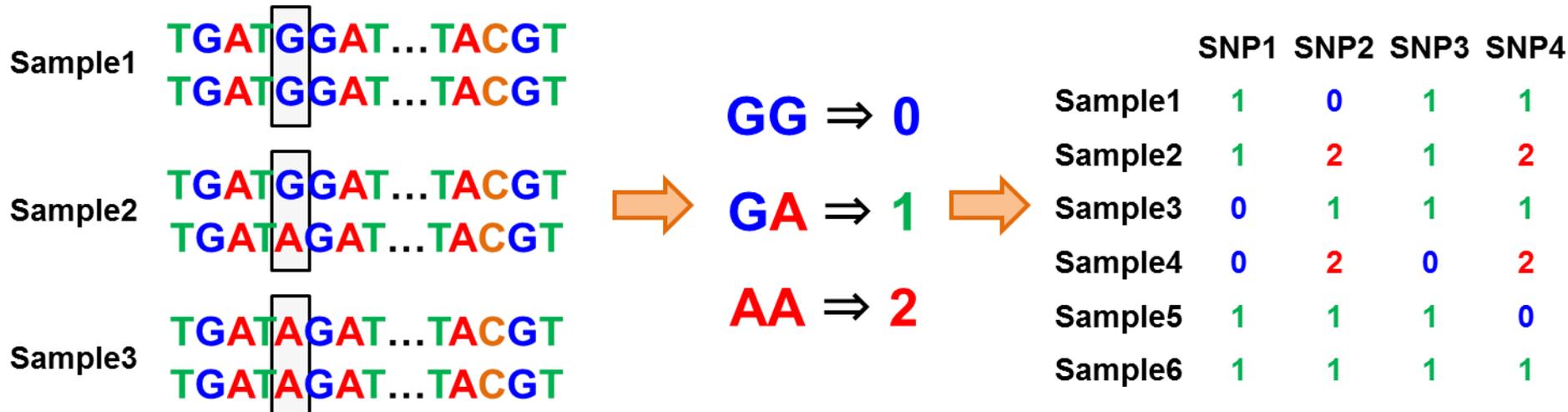
- データ解析を行う統計学には、様々な学問分野が含まれています。
- ゲノムデータの解析に特化した統計学として、**遺伝統計学**があります。

② ヒトゲノム多型を用いた疾患ゲノム解析

SNPジェノタイプデータ

0/1/2の行列

SNP



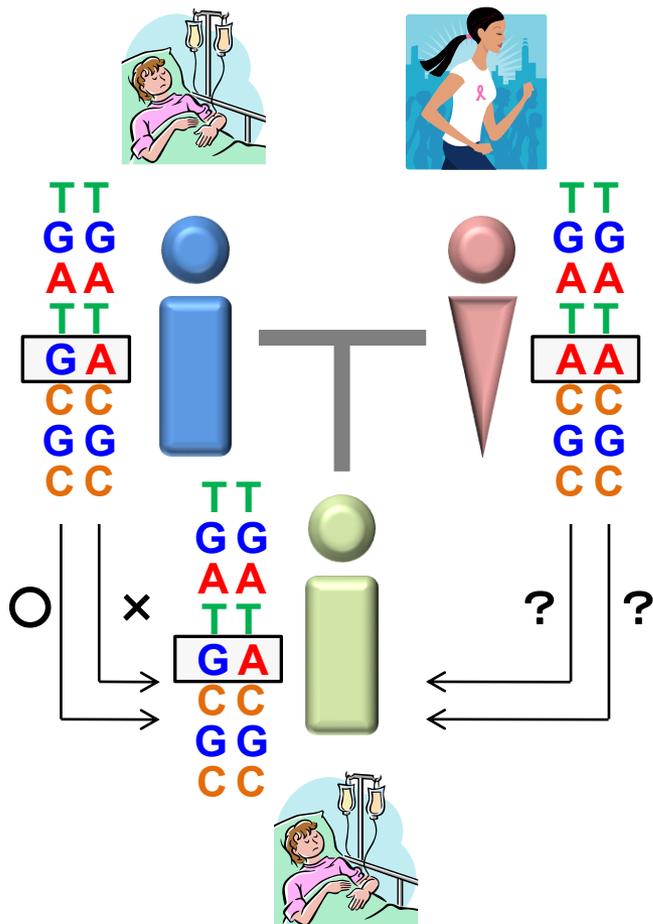
• 遺伝情報は、**デジタルな情報**として扱うことができる。

①: 個人の遺伝情報は、4文字(ATGC)×30億×2通り、で表現可能。

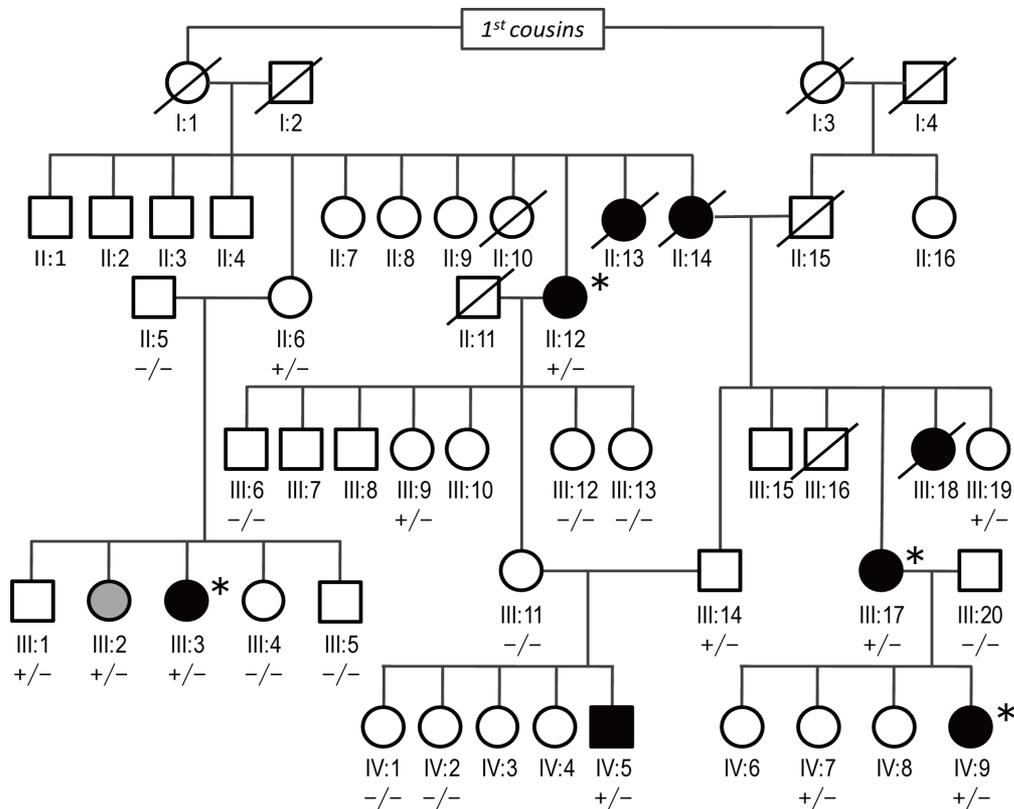
②: 個人間の違い(SNP)だけ抽出すれば、更に効率的に表現可能。

⇒ 数十万～数百万SNPの情報で、個人の遺伝情報を表現できる。²⁷

② ヒトゲノム多型を用いた疾患ゲノム解析



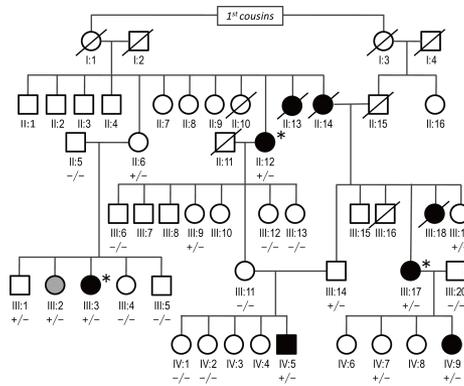
家族性関節リウマチにおける実施例



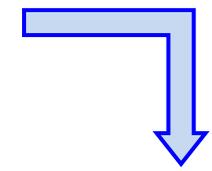
連鎖解析(Linkage analysis)

- 家系内での遺伝情報と疾患の受け継がれ方を比較することにより、疾患の原因遺伝子領域を同定する。

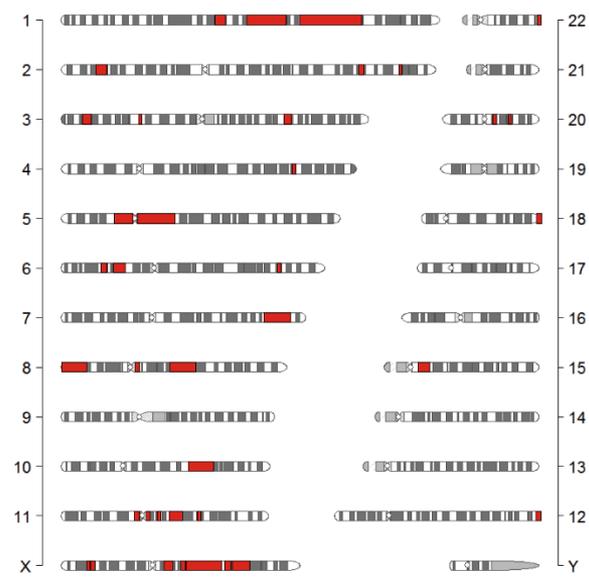
② ヒトゲノム多型を用いた疾患ゲノム解析



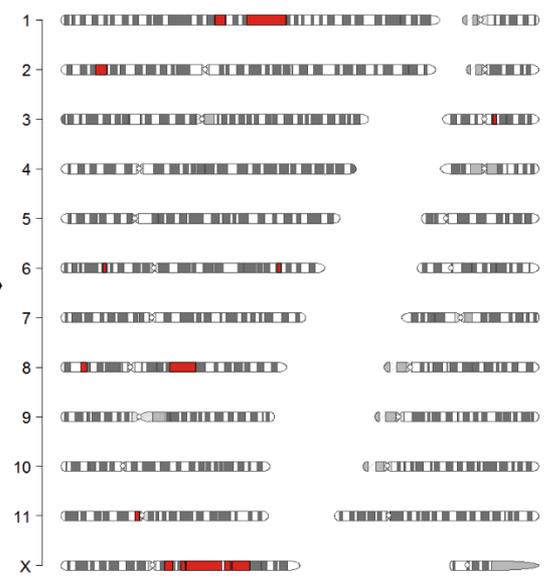
遺伝統計解析



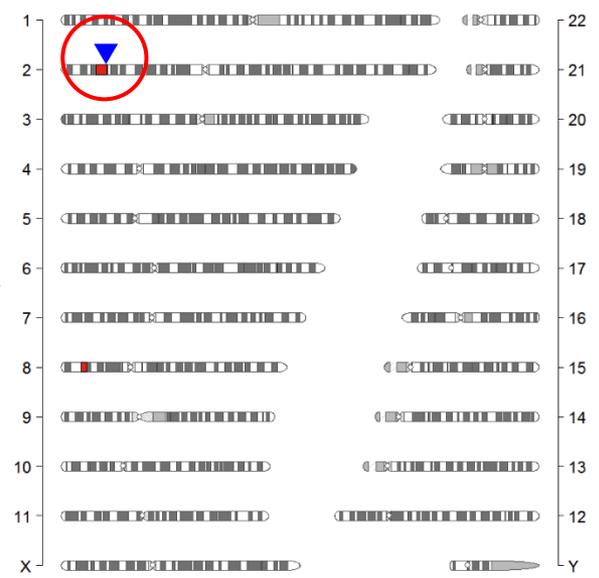
IBD stretches among 4 RA cases



IBD stretches among 5 RA cases and 1 ACPA-positive control



IBD stretches among 5 RA cases, 1 ACPA-positive control, and ≤ 9 ACPA-negative controls



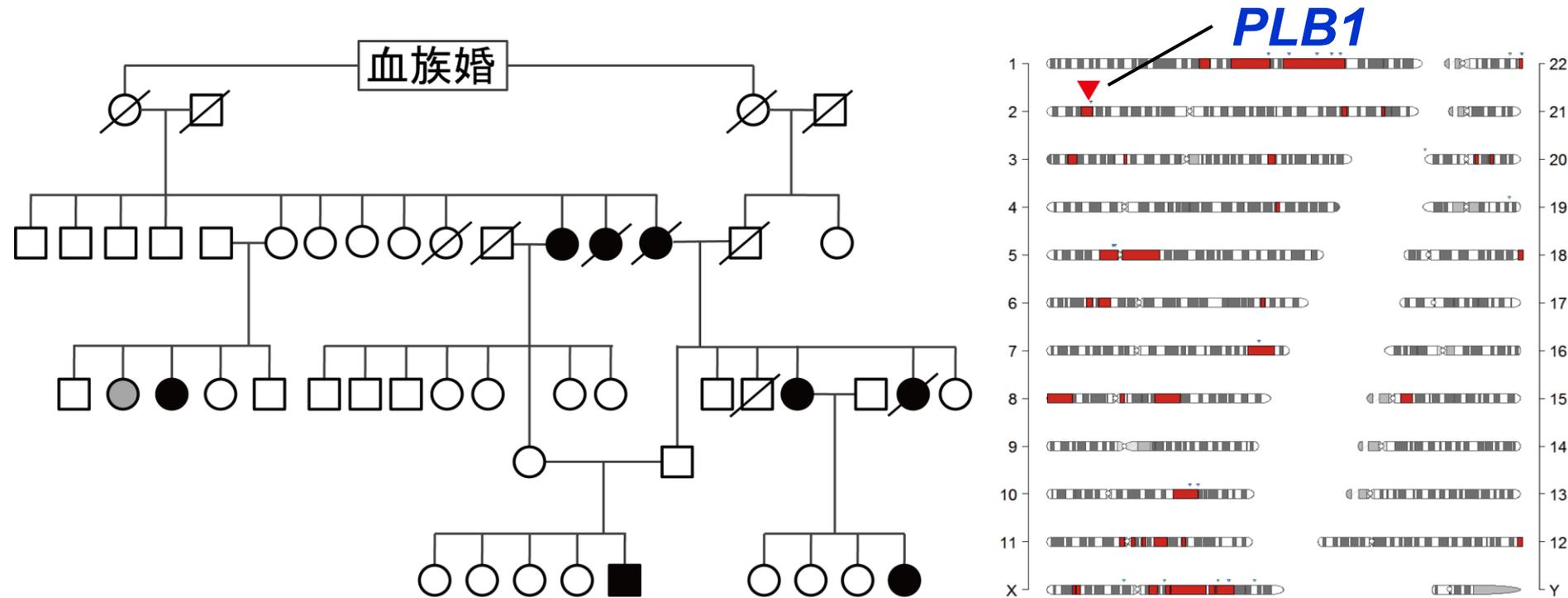
■ ... IBD stretch

▼ ... Candidate causal SNV

• 遺伝統計解析の第一目的は、疾患原因遺伝子の“住所”の同定。 30

② ヒトゲノム多型を用いた疾患ゲノム解析

自己免疫疾患家系例に対する全エクソーム解析



- 中東で報告された家系例およびケースコントロール2,000名を対象とした次世代シーケンサー解析(エクソーム解析)を通じて、関節リウマチの新規感受性遺伝子 *PLB1* を同定。

② ヒトゲノム多型を用いた疾患ゲノム解析

ABO血液型と貧血の関連

性別	基準値	血液型別ヘモグロビン平均値 (g/dl)		
		**	B*	BB
女性	11.3-15.5	12.9	13.0	13.1
男性	13.8-16.6	14.1	14.2	14.2

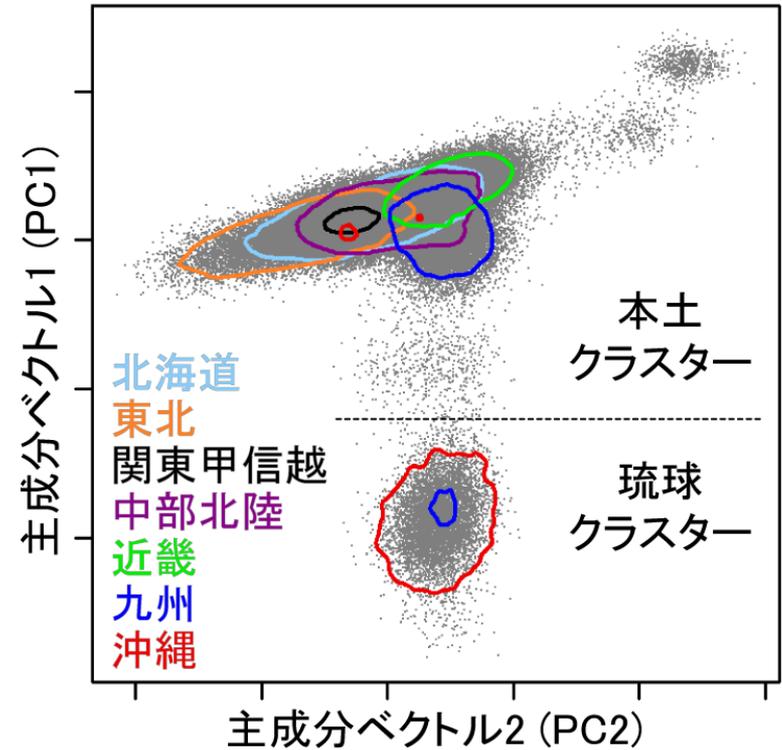
- ABO遺伝子上のSNPにより、個人の**ABO式血液型**が決定されます。
- **B型決定アレル**と、**貧血関連指標(ヘモグロビン)**との関連を同定しました。
- B型決定アレルの保有者は、非保有者と比較した場合
⇒ ヘモグロビン値が約0.15 g/dl高く、貧血リスクが7~8%低かった。

② ヒトゲノム多型を用いた疾患ゲノム解析

主成分分析による日本人集団の分類

	SNP1	SNP2	SNP3	SNP4	SNP5
Sample1	1	0	1	1	1
Sample2	1	2	1	2	1
Sample3	0	1	1	1	1
Sample4	0	2	0	2	1
Sample5	1	1	1	0	1
Sample6	1	1	1	1	1
Sample7	2	0	0	2	0
Sample8	1	1	0	1	1
Sample9	1	1	2	2	2
Sample10	0	2	1	1	2

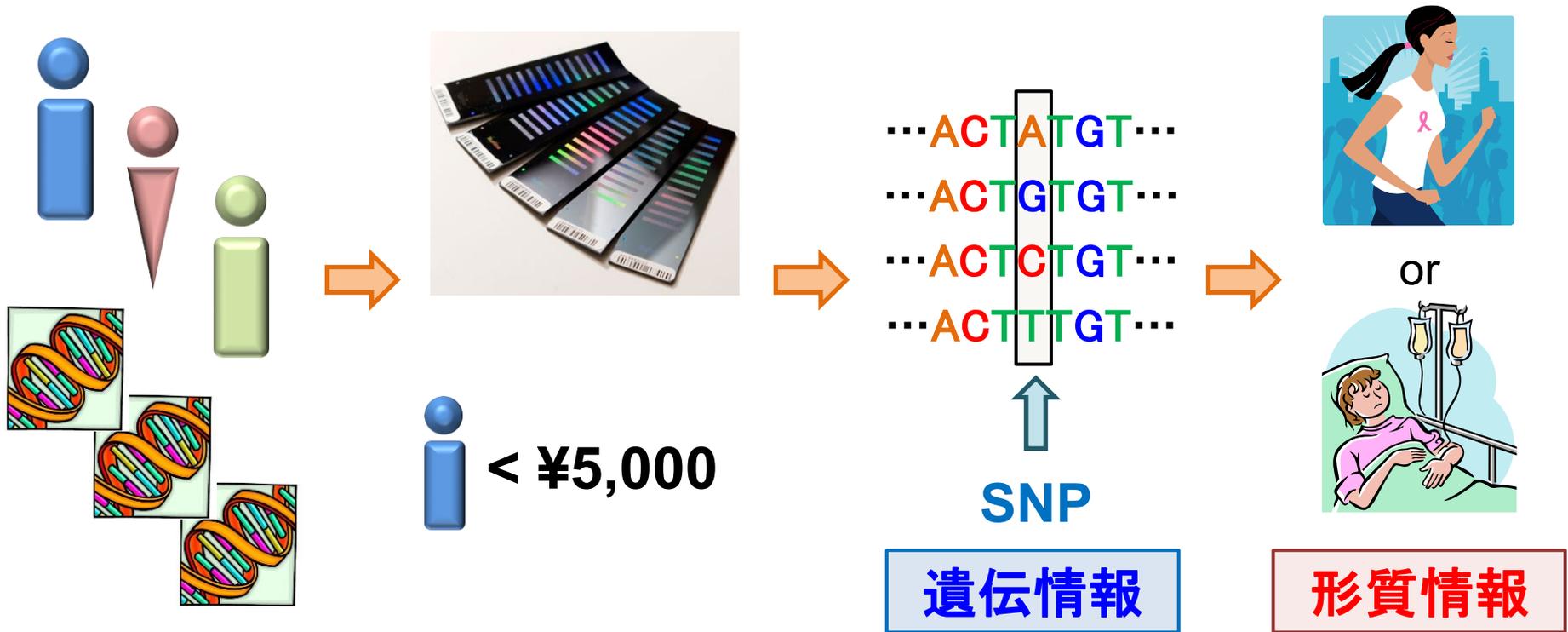
PCA
→



- 日本人集団数千人のジェノタイプデータを対象に主成分分析(PCA)を行った結果、日本人集団の遺伝的背景が、本州(本土)と沖縄(琉球)とで(少しだけ)違うことがわかった。

② ヒトゲノム多型を用いた疾患ゲノム解析

ゲノムワイド関連解析(GWAS)



- 遺伝情報と形質情報との結びつきを評価する遺伝統計学的手法。
- 数十万人を対象に、ヒトゲノム全体を網羅する数千万箇所のSNPのタイピングを実施し、対象形質との関連を評価する手法。
- 2002年に日本の理化学研究所で世界に先駆けて実施された。

② ヒトゲノム多型を用いた疾患ゲノム解析

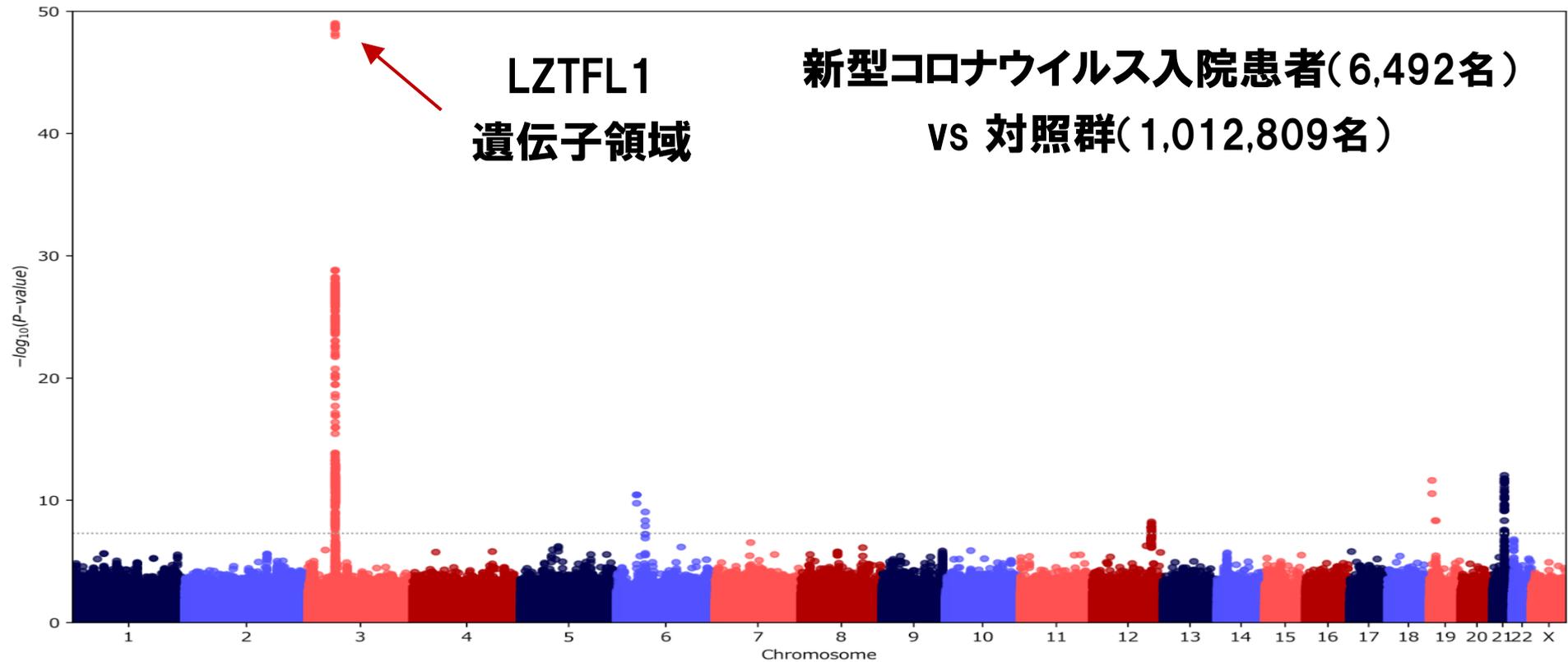
ゲノムワイド関連解析(GWAS)が同定した感受性遺伝子



- SNPタイピング技術の発達に伴い、2000年代後半以降、世界中の研究施設でゲノムワイド関連解析が実施されている。
- 2022年までに、1,000以上の形質に対し約5,000報のゲノムワイド関連解析が報告されている。

② ヒトゲノム多型を用いた疾患ゲノム解析

新型コロナウイルス重症化ゲノム解析



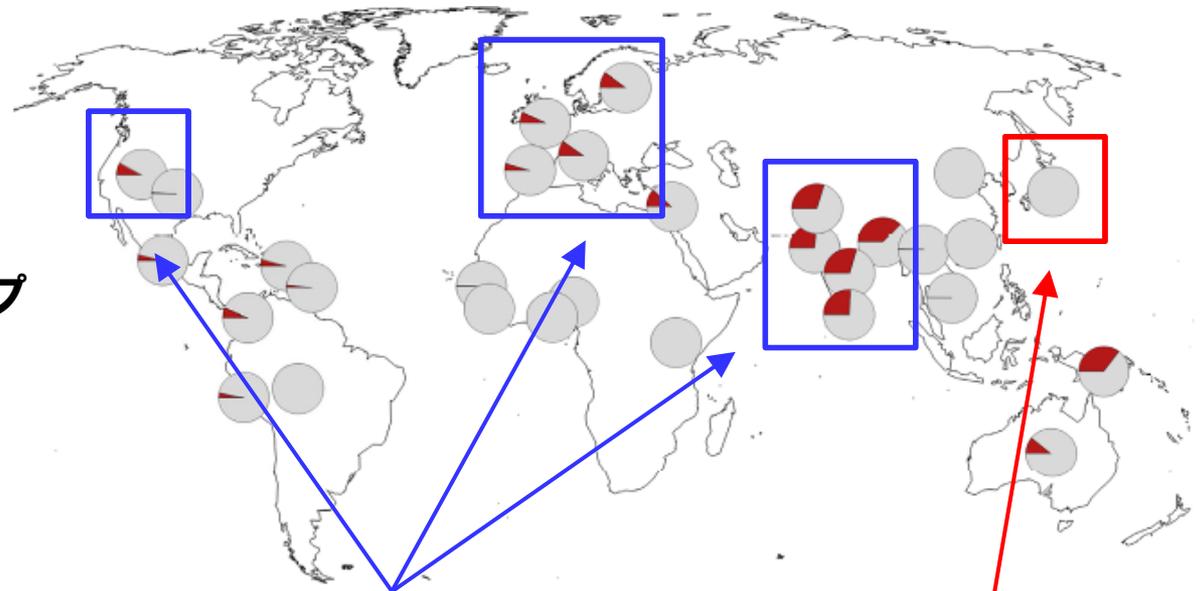
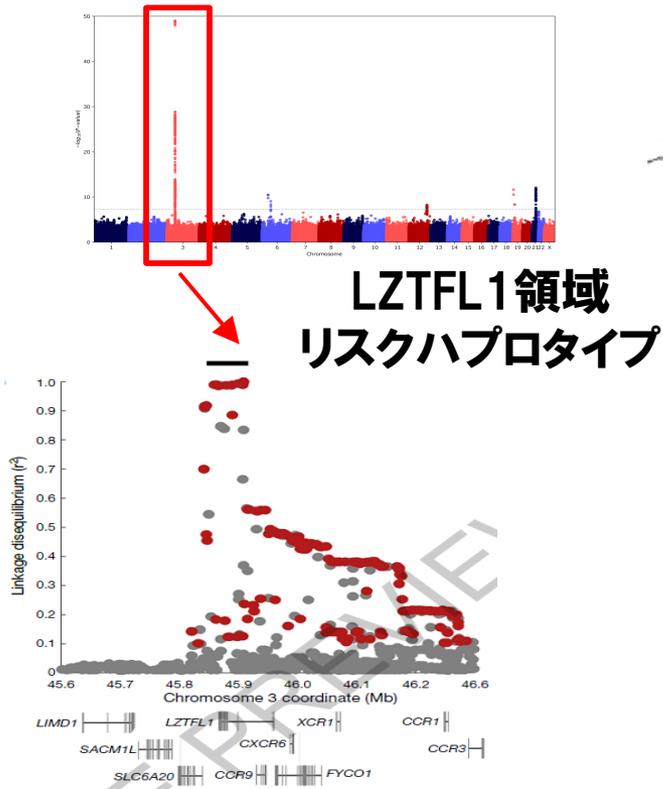
- 国際コンソーシアムCOVID-19 Host Genetics Initiativeにより、**新型コロナウイルス感染者の重症化リスクに関するGWAS**が実施された。
- 欧米人集団において、**LZTFL1遺伝子領域**が強いリスクを有することが明らかになった。

② ヒトゲノム多型を用いた疾患ゲノム解析

欧米人集団特異的COVID-19重症化リスクハプロタイプ

欧米人集団COVID-19 GWAS

COVID-19リスクハプロタイプの頻度分布



欧米人集団・中央アジア人
集団では高頻度

東アジア人集団では
ほぼ存在しない!!

- LZTFL1遺伝子領域COVID-19重症化リスクハプロタイプは、ネアンデルタール人由来であり、頻度が人類集団間で異なっていた。
- 東アジア人集団はリスクハプロタイプが存在しないことが明らかに。

② ヒトゲノム多型を用いた疾患ゲノム解析

コロナ制圧タスクフォース:オールジャパンCOVID-19研究

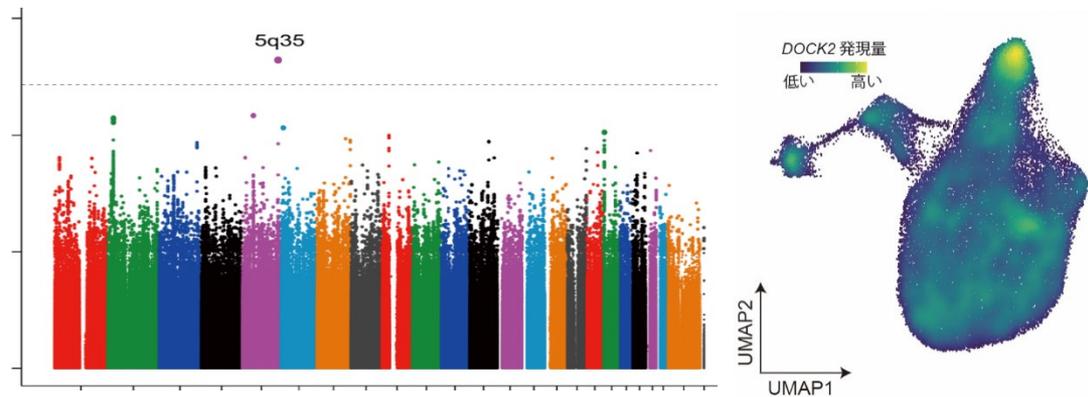
全国規模100以上の参加施設
(主催:慶応義塾大学)

コロナ制圧タスクフォース

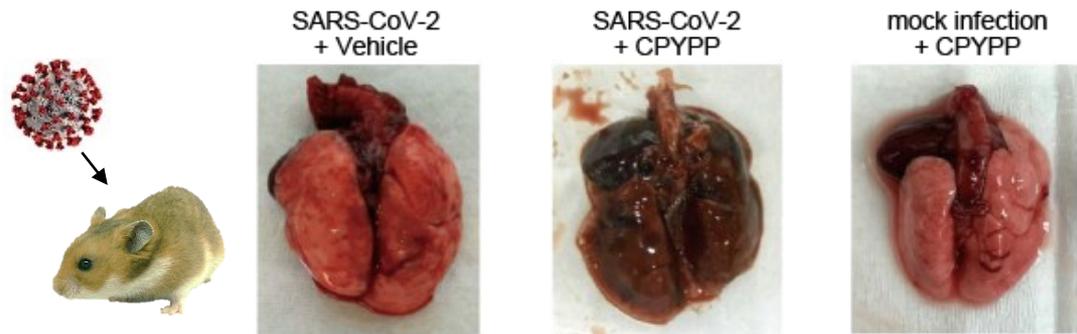


<https://www.covid19-taskforce.jp/>

GWAS・シングルセル解析:重症化遺伝子DOCK2同定



DOCK2阻害剤によるCOVID-19重症化メカニズム解明



- 日本人集団の新型コロナウイルス克服を目指し、コロナ制圧タスクフォースを結成。感染者6,000名のゲノム・RNA・血漿・臨床情報を収集。
- アジア人集団最大規模の生体レポジトリとして、ゲノム・オミクス情報に基づく予防医療の社会実装へ。

② ヒトゲノム多型を用いた疾患ゲノム解析

日経ビジネス

「ファクターXを探して」

全4回

コロナ重症化因子解明の鍵、冷蔵された患者5000人分の血液

小学校の教室の半分ほどの広さの部屋は、肌寒いほどの気温に調整されている。無人。並べられた冷蔵庫のコンプレッサーから出る低音以外は何も聞こえない。冷蔵庫の中は2つの温度帯に保たれている。一つは4度、もう一つはマイナス80度…

2021.12.24



過酷なコロナ臨床最前線、立ち上がった500人の医師

日本には帰れない——。2020年3月、米国の首都ワシントンから北西に車で30分ほどの距離にあるノース・ベセスダ。医師の南宮湖は自宅でパソコンを眺めながらため息をつき、航空会社のサイトで成田空港行きのフライトをキャンセルし…

2022.01.07



スパコンが解くパズル、コロナ重症化を招く遺伝子の特徴が見えてきた

在野の医師たちの呼びかけで新型コロナウイルスの重症化因子を探る「コロナ制圧タスクフォース」が立ち上がった。全国の医師たちの協力と献身により異例のペースで約5000人分の検体が集まった。いよいよこれを分析していく。

2022.01.14



「主役なき物語」は続く…感染症に関するアジア最大のバイオバンクへ

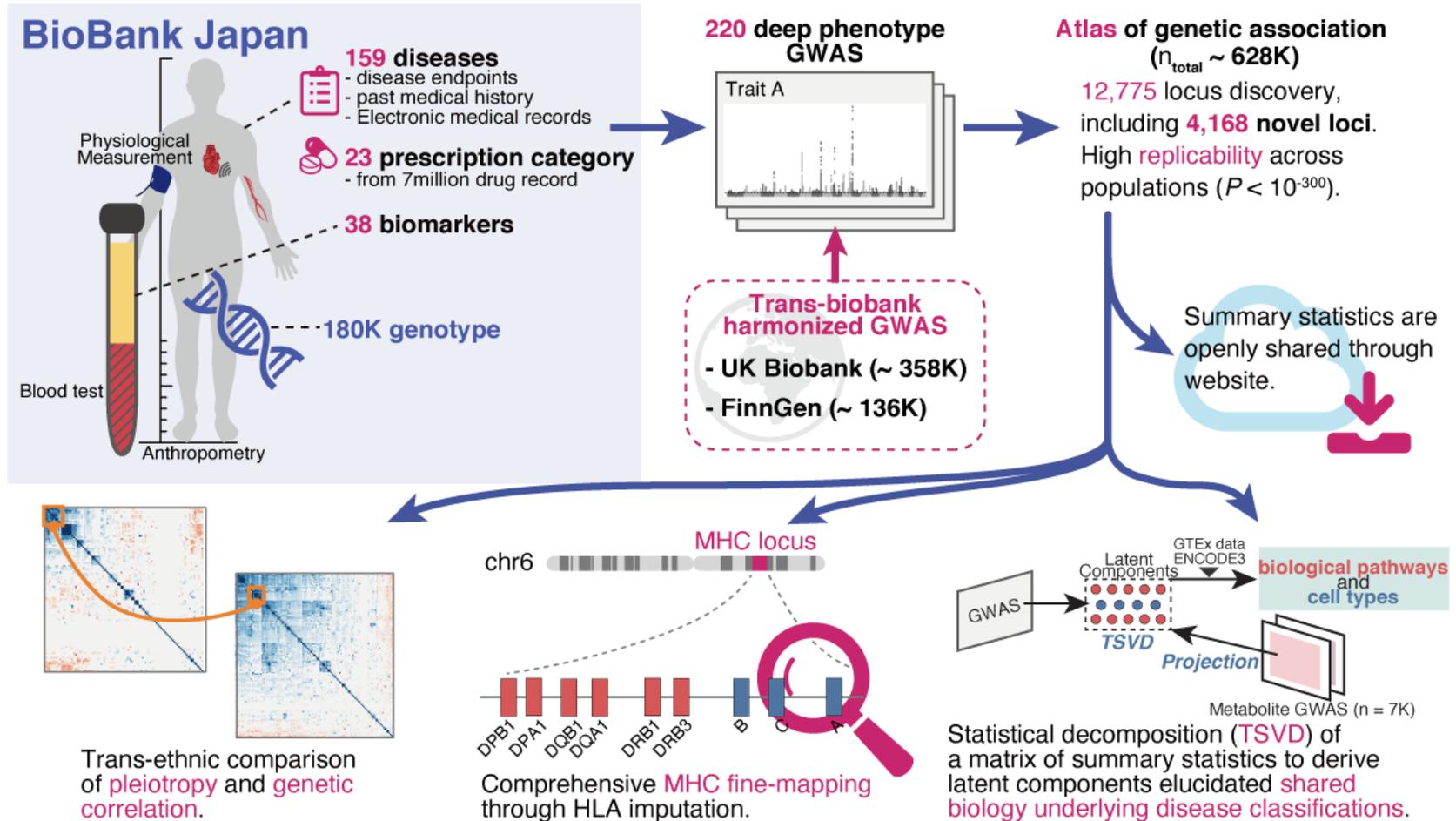
米ニューヨークにあるピュッフェの設けられたオフィスの一室、2006年。みずほコーポレート銀行（当時）の飯田浩一は、やや緊張した面持ちで朝食を口に運んでいた。向かいに座るのは、米投資銀行エパーコア・パートナーズを立ち上げた…

2022.01.21



② ヒトゲノム多型を用いた疾患ゲノム解析

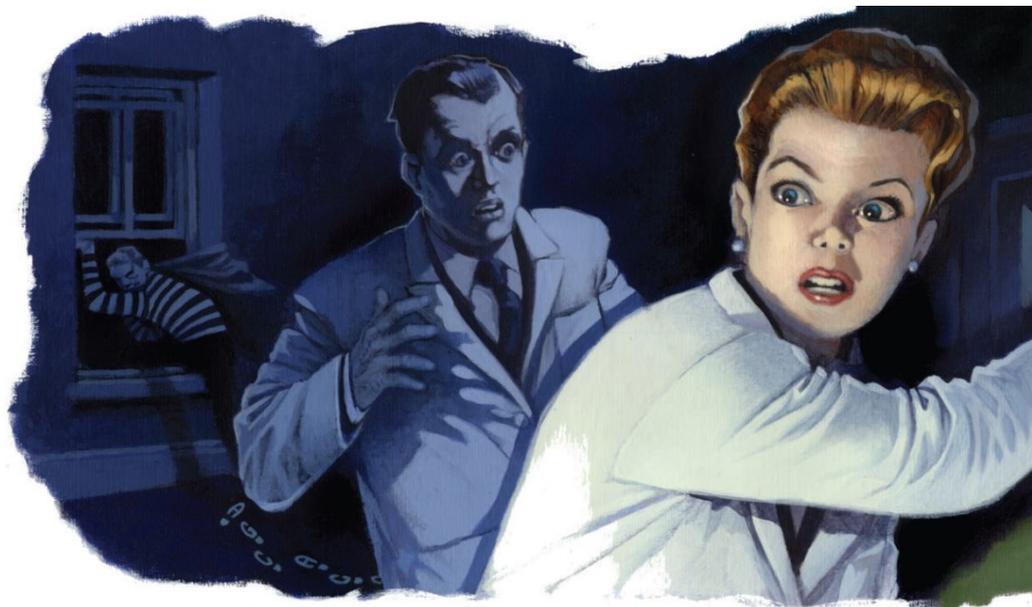
国際バイオバンク連携によるGWASメタ解析



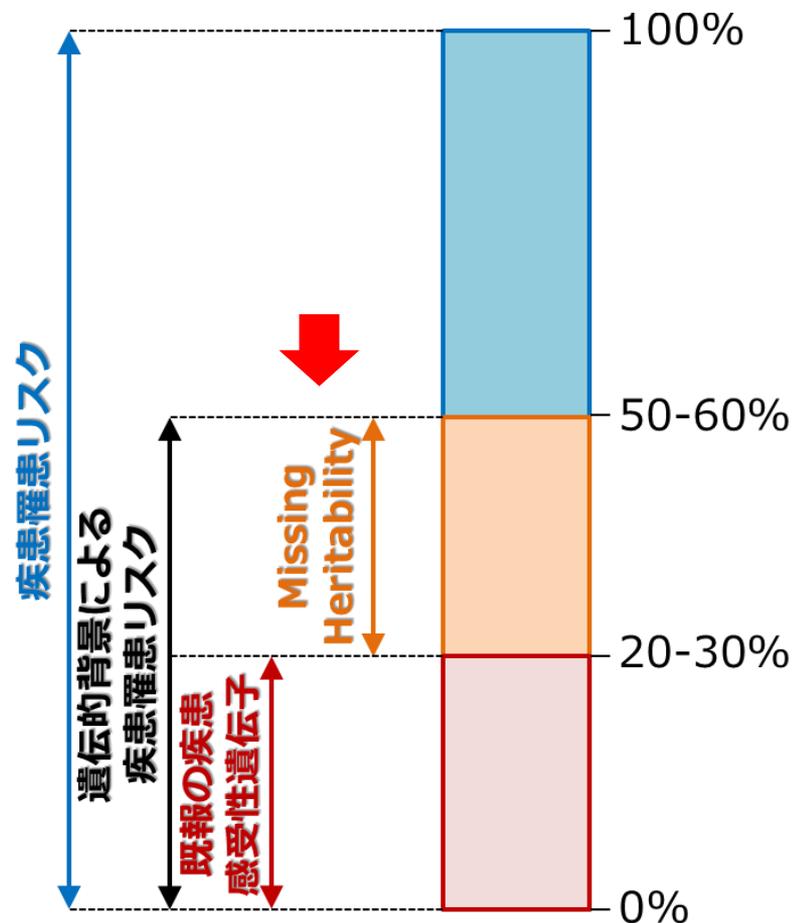
• 国際バイオバンク連携63万人・220形質のGWASメタ解析を実施。

• GWAS統計量を <https://pheweb.jp/> にて一般公開中。

② ヒトゲノム多型を用いた疾患ゲノム解析

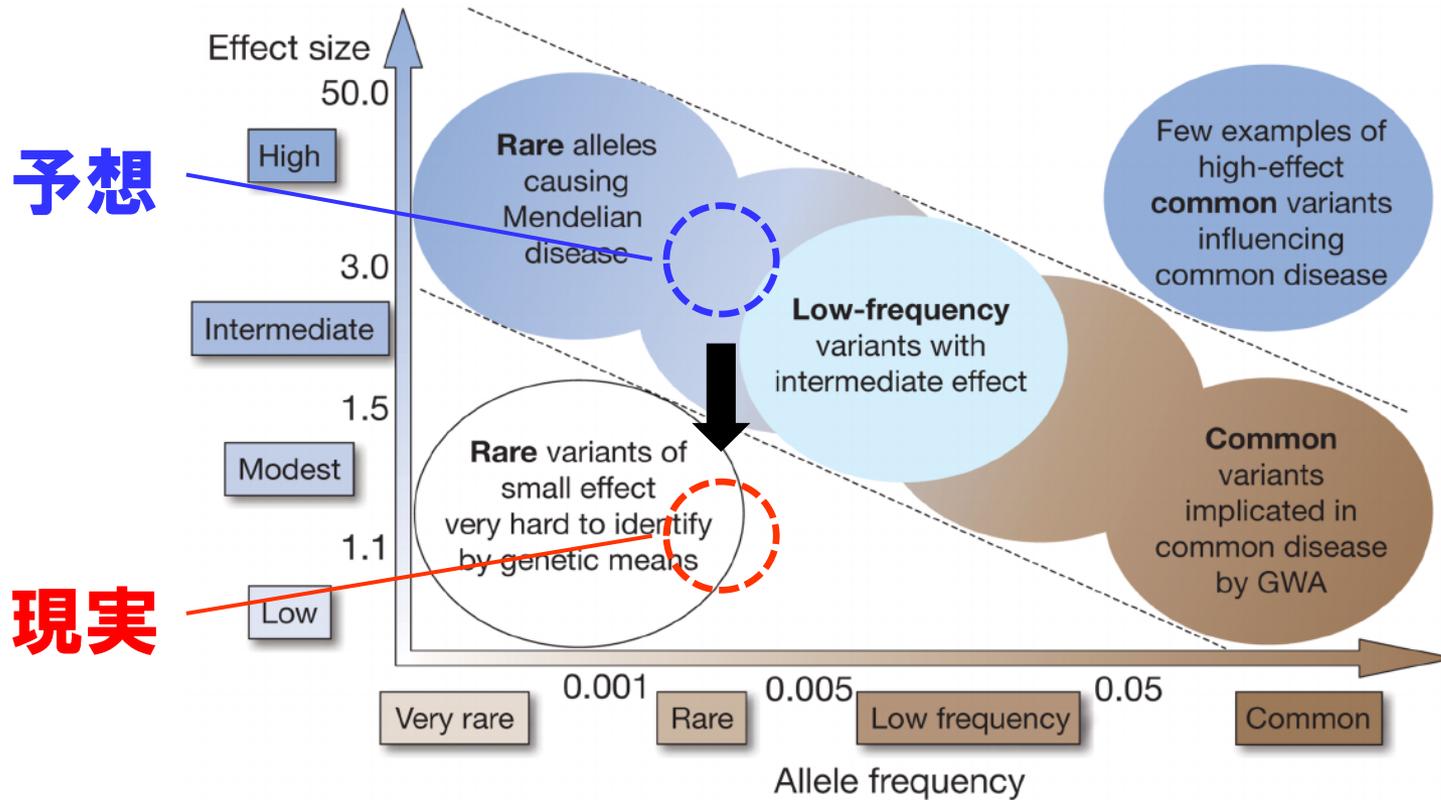


The case of the missing heritability



- GWASで同定された遺伝子変異では、疾患の遺伝的背景の一部しか説明できないことも判明し、“**missing heritability**”と呼ばれています。
- Missing heritabilityを説明するために、様々な仮説が立てられています。

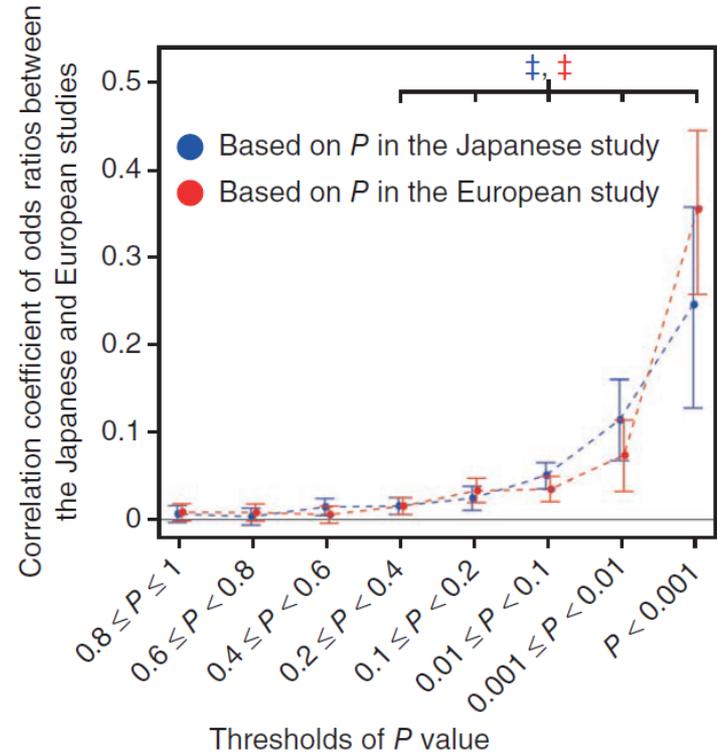
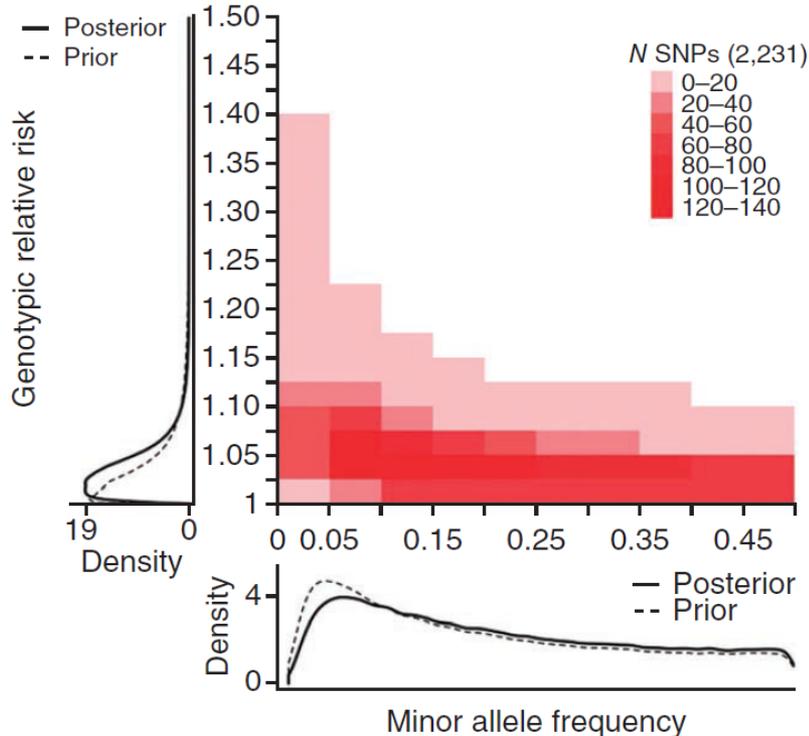
② ヒトゲノム多型を用いた疾患ゲノム解析



- 次世代シーケンサーの発達に伴い、**レアバリエント**がmissing heritabilityを説明できるのでは？、という期待が高まりました。
- しかし、レアバリエントがありふれた疾患に対して有するリスクは小さく、**missing heritability**を説明するには**不十分**であるとも考えられています。

② ヒトゲノム多型を用いた疾患ゲノム解析

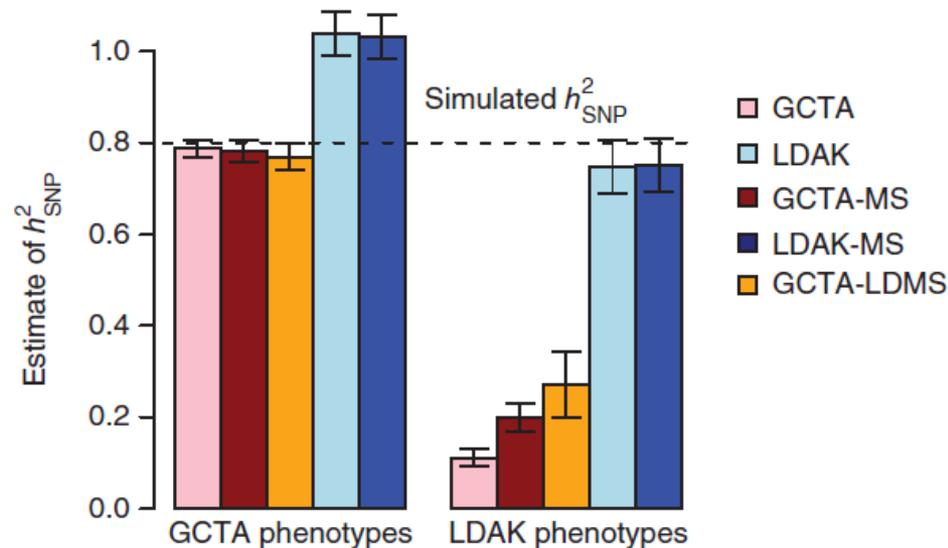
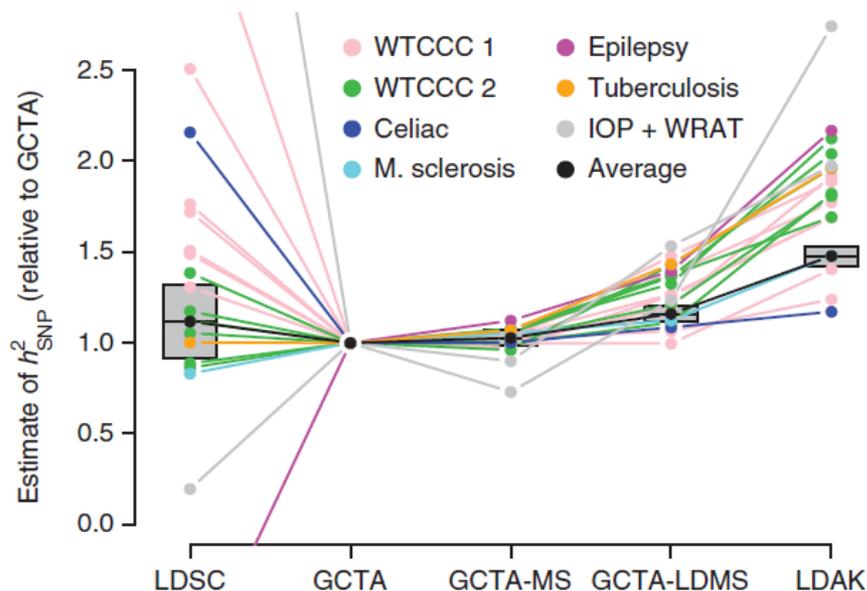
関節リウマチリスクのPolygenic modelは集団間で共有される



- ゲノム全体に存在する無数のコモンバリエントが弱い疾患リスクを有する、という”polygenic model”が提唱されています。
- Polygenic modelに基づきゲノム全体のコモンバリエントを対象とすることで、**遺伝的背景の大半を説明できる可能性が指摘されています。**

② ヒトゲノム多型を用いた疾患ゲノム解析

Polygenic modelによるheritability推定値はソフトウェアに依存



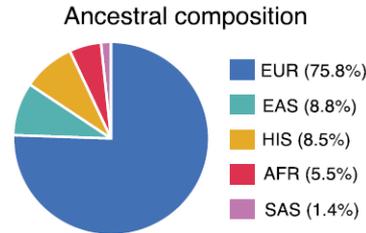
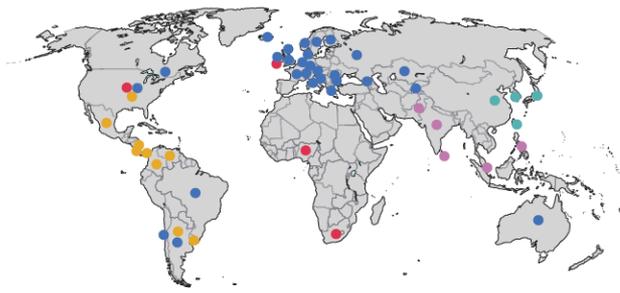
- 一方で、Polygenic modelにより説明されるheritabilityの推定値は、使用するソフトウェアに依存して大きく変化することも指摘されています。
- Missing heritabilityを巡る状況はまだ混沌としていて、全容解明には、サンプル数がまだまだ足りないと考えられています。

② ヒトゲノム多型を用いた疾患ゲノム解析

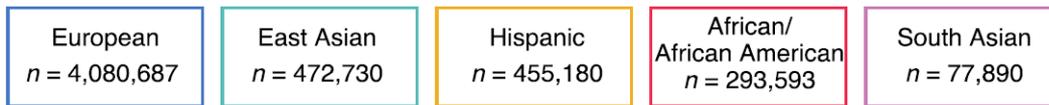
540万人GWASにより身長heritabilityの大部分が説明された

GIANTコンソーシアムによる過去最大規模のGWAS

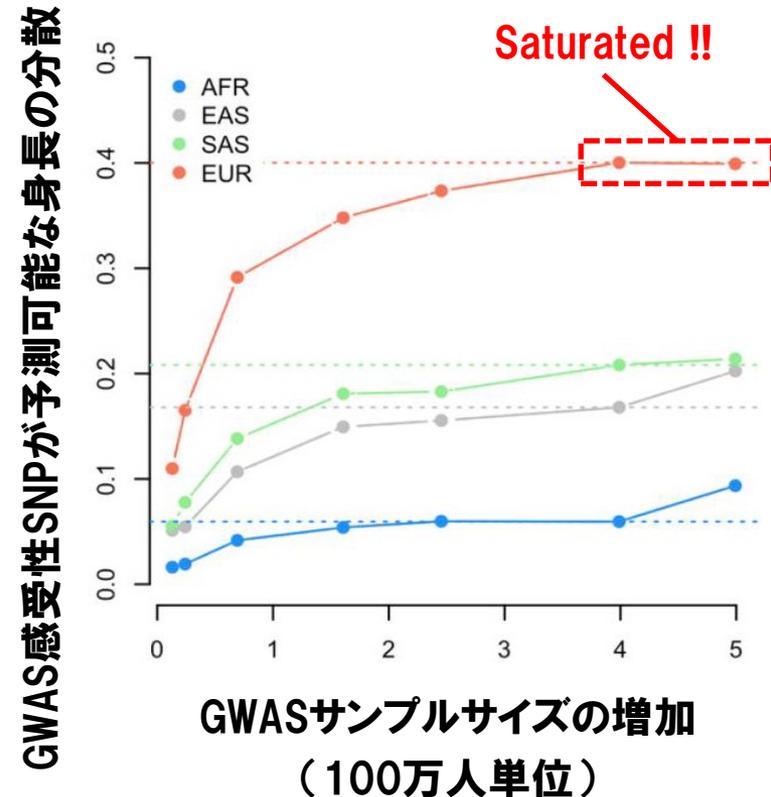
GIANT consortium: Genetic Investigation of ANthropometric Traits



Ancestry-specific meta-analysis of height



GWAS meta-analysis of height
in 281 studies



- 複数人種540万人GWASにより12,000箇所の身長感受性SNPを同定。
- Common SNPによるheritabilityの大部分が感受性SNPで説明可能であること(=saturated map of common genetic variants)が確認された。

② ヒトゲノム多型を用いた疾患ゲノム解析



長生き出来るかどうかも、ゲノムで決まるのですか？



寿命の個人差の一部は、ゲノムの個人差で説明できそうです。但し、**感受性遺伝子探しは難航**しています。

- ・私達は経験的に、ヒトの寿命には個人差があることを知っています。
- ・寿命の個人差には、性別、生活習慣、職業、地域集団など、複数の要因が関わっています。
- ・ゲノムの個人差も、寿命の個人差に影響すると考えられていますが、**寿命を決める感受性遺伝子探しは難航**しています。

② ヒトゲノム多型を用いた疾患ゲノム解析



どうして、長寿遺伝子を見つけるのは難しいのですか？



が難しいからです。

- ・長寿に関わる遺伝子を見つけるためには、長寿者のゲノム配列と比較する、対照群(コントロール群)が必要です。
- ・ゲノム以外の要因(性別、生活習慣、職業、地域集団)による影響を排除するためには、これらの背景情報が共通し、かつ長寿でない人々を集める必要があります。これはなかなか難しい作業になります。

② ヒトゲノム多型を用いた疾患ゲノム解析

ゲノムワイド関連解析による長寿者研究



SHARE



REPORT

Genetic Signatures of
Exceptional Longevity in
Humans

Paola Sebastiani^{1,*}, Nadia Solovieff¹, Annibale Puca², Stephen W.

- 2010年、長寿者のゲノム研究結果がScience誌にて報告されました。
- 長寿に関わる感受性遺伝子を100個以上も同定した研究成果で、**ついにゲノム研究で長寿の謎が解明されるのでは？**、と期待されました。

② ヒトゲノム多型を用いた疾患ゲノム解析

ゲノムワイド関連解析による長寿者研究



- この研究成果は、**残念ながら2年後に取り下げ**となりました。
- 長寿者集団と比較対照群として選んだ集団が、**別々の地域から集められていたことが原因**でした。
- つまり、長寿感受性遺伝子と考えられていたものは、**2つの地域の集団の遺伝的背景の違いを示す遺伝子**だったのです。

② ヒトゲノム多型を用いた疾患ゲノム解析



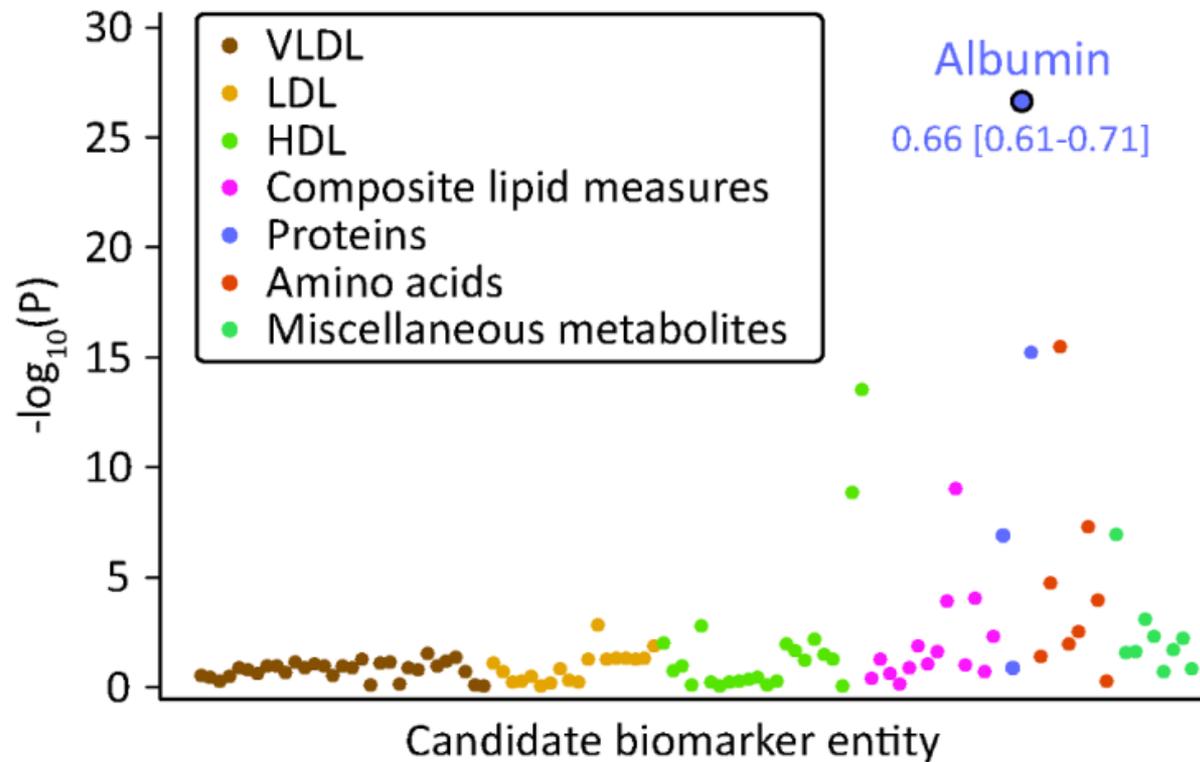
長生き出来るか、知ることはできないのでしょうか？



ゲノム情報だけで寿命を予測することは難しいですが、**あとどれくらい長生きできるかは予測可能**かもしれません。

- ・アルツハイマー病など特殊な例を除いて、ゲノム情報だけで、ヒトの寿命を予め予測することは難しいと考えられます。
- ・生活習慣や社会環境の方が、与える影響が大きいと考えられます。
- ・一方で、**あとどれくらい長生き出来るかは、現時点で取得可能な情報から予測**できるかもしれません。

② ヒトゲノム多型を用いた疾患ゲノム解析



- エストニアにおける大規模追跡研究により、**死亡の5年前から幾つかの血液検査値が特徴的な変動を生じていることが報告されました。**
- これは、死亡原因を特定しない解析の結果であり、**特定の病気の影響は小さいと考えられています。**

ヒトゲノムデータの容量ってどのくらい？

GWASヒトゲノムデータ1人分は・・・

①:70 キロバイト ← メール一通分

②:700 キロバイト

③:7 メガバイト ← 写真1枚分

④:70 メガバイト

⑤:700 メガバイト ← CD1枚分

遺伝学入門

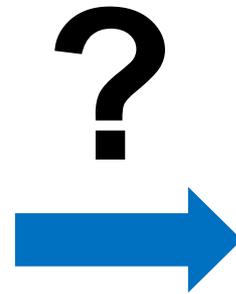
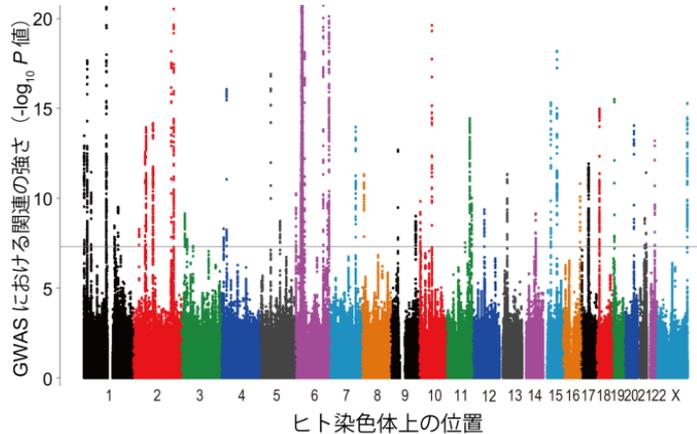
- ① ヒトゲノム配列とその個人差
- ② ヒトゲノム多型を用いた疾患ゲノム解析
- ③ 疾患病態解明への展望
- ④ 機械学習のヒトゲノム情報への応用
- ⑤ ゲノム創薬への展望

③ 疾患病態解明への展望

ゲノム解析の成果をどのように活用すれば
疾患病態・創薬・個別化医療に貢献できるのか

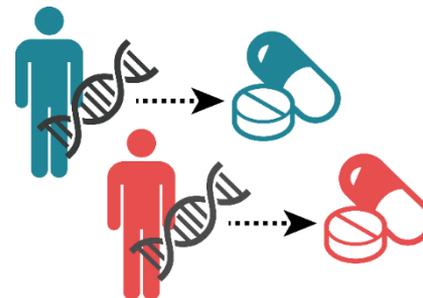
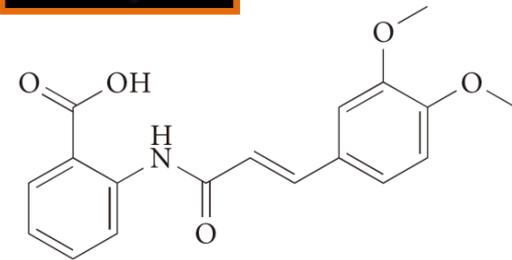
ほとんどわかっていない

大規模
ヒトゲノム解析



Disease
biology

Drug
discovery

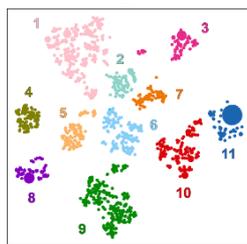
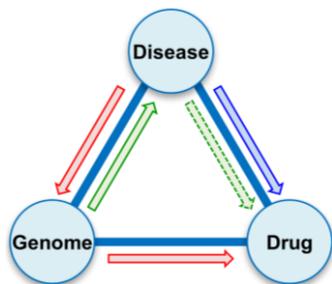
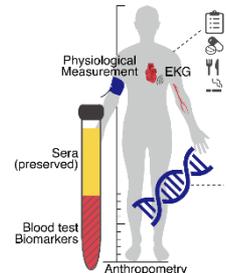
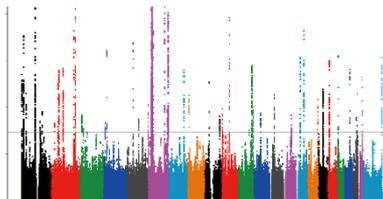


Personalized
Medicine

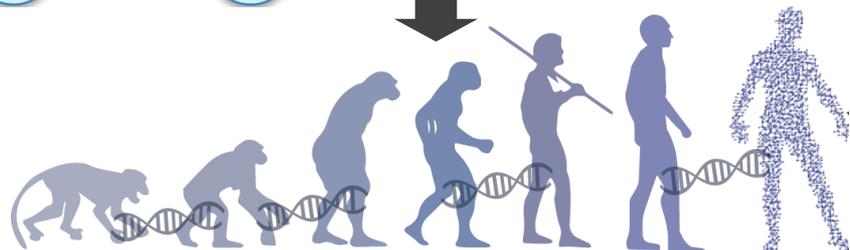
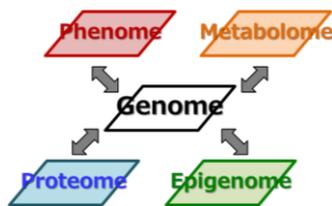
③ 疾患病態解明への展望

30億塩基対のヒトゲノム配列

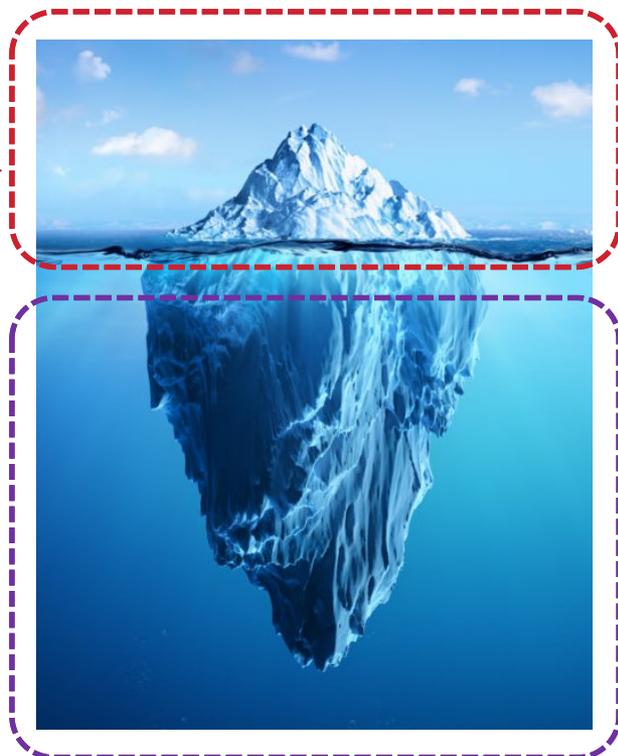
...ACTATGTAGCGTACCT...



Trans-layer omics



氷山の一角



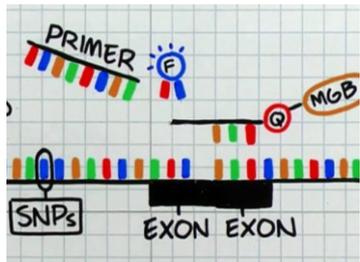
氷山の奥底は未到達

?

- 30億塩基対のヒトゲノム配列はヒトの設計図。そのメカニズムを解明すれば生命現象を再構成できるはずだが、ほとんどは**未解明**。
- 研究テーマ: 如何にヒトゲノム配列に隠された生命現象を引き出すか。

③ 疾患病態解明への展望

TaqManアッセイ



SNPマイクロアレイ



ショートリードNGS



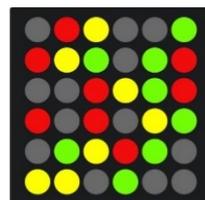
ロングリードNGS



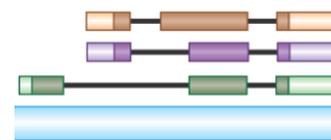
RT-PCR



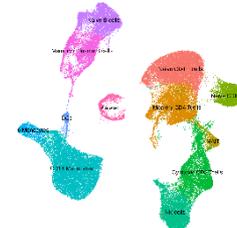
遺伝子マイクロアレイ



RNA-seq



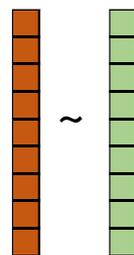
シングルセル解析



分割表検定

	AA	AG	GG
Case	360	480	160
Control	250	500	250

回帰分析



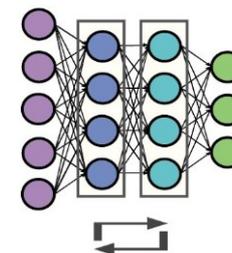
機械学習

Variants (90 K)

ACTGAACGCAAAC
CCAGTATTCTACCT
CATGACTGCAAAC
ACAGTATGCTACAT
CCTAAATGCTACCT
AAAGTCTTCTACCT
ACTATATGCTACCT

Individuals (170 K)

深層学習

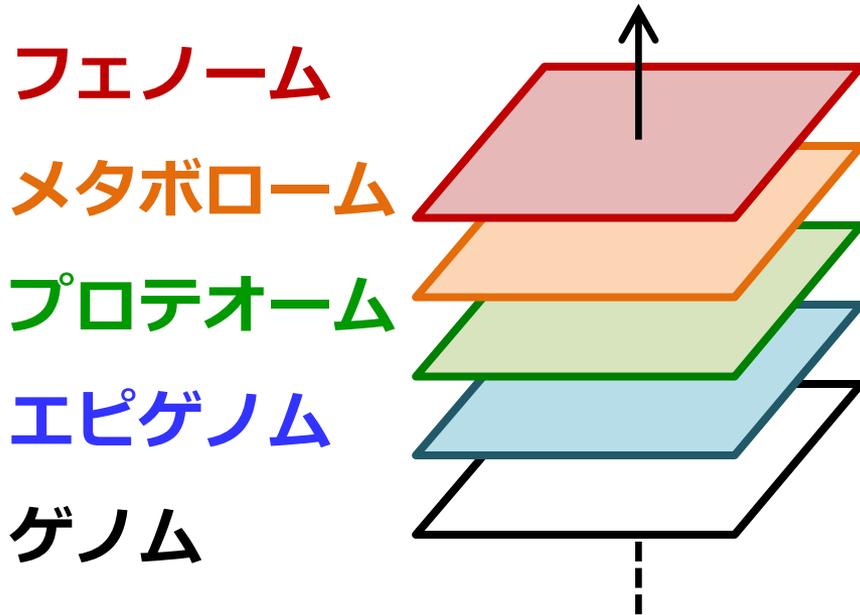


- 観測・解析技術の進歩は、常に予想を上回る速度で進む。
- Wet・Dry双方の最新解析技術の先進的導入が生命科学に不可欠。
- 実験・解析原理の正確な理解が、革新的な研究を可能にする。

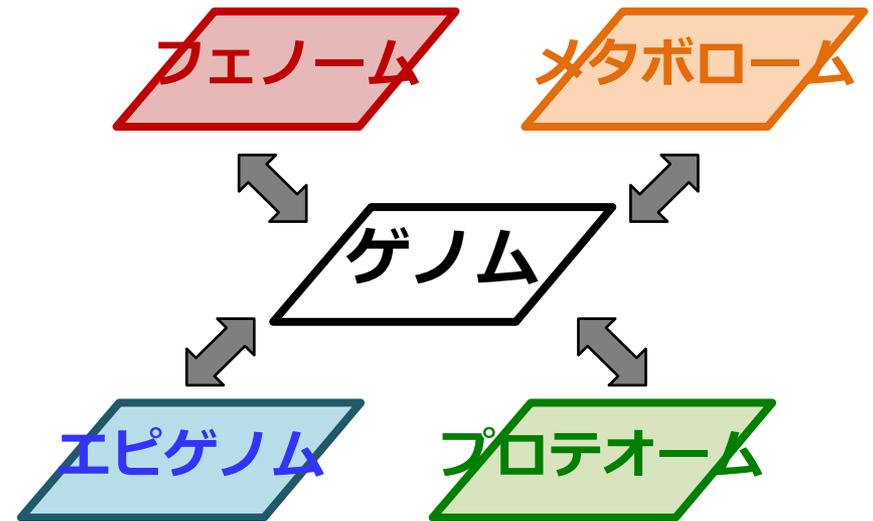
③ 疾患病態解明への展望

多層的 vs 横断的 オミクス解析

多層的オミクス解析



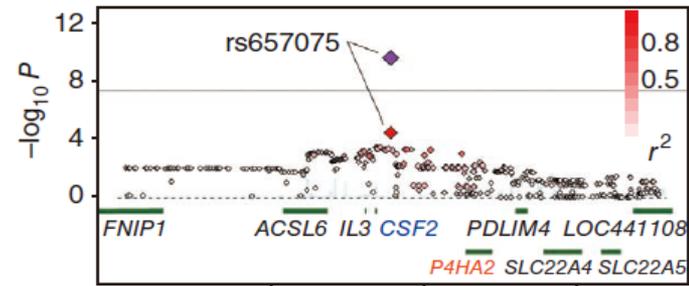
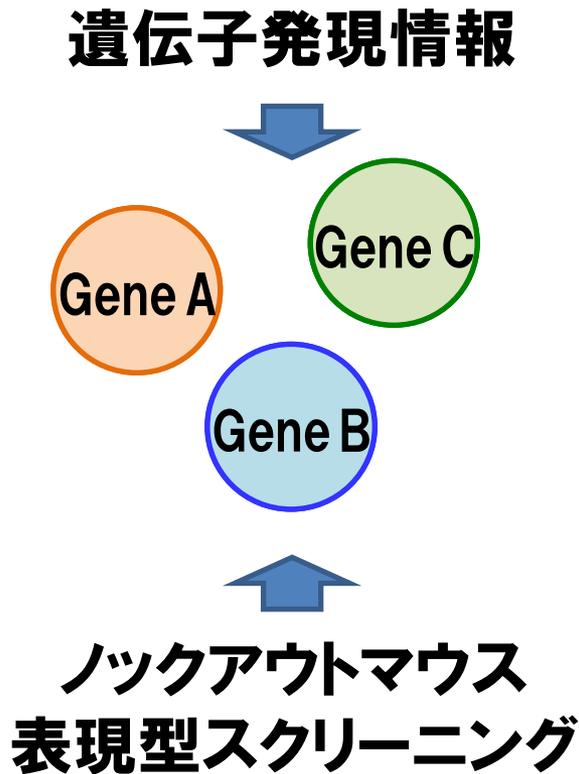
横断的オミクス解析



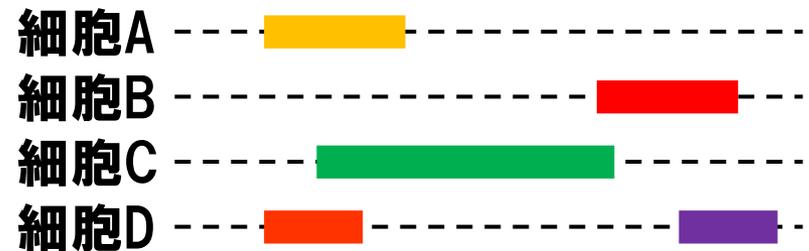
	多層的オミクス解析	横断的オミクス解析
網羅性	×	○
コスト	△	○
サンプル数	×	○
予測性	○	△

③ 疾患病態解明への展望

横断的オミクス解析の解析方法



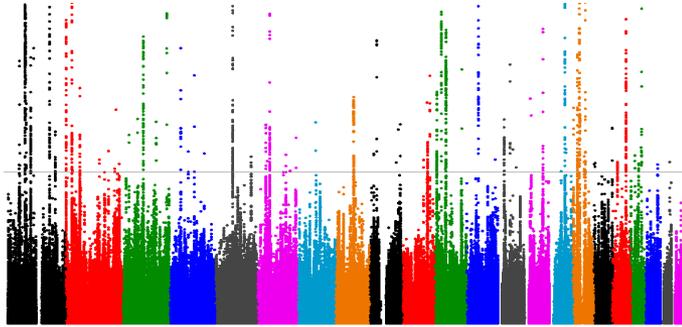
ヒトゲノム標準配列



- 異なるサンプル集団由来のオミクス情報を**共通の情報単位**に変換することで、横断的な統合が可能になる。
- 近年、**ヒトゲノム標準配列に沿ったベクトル情報**に変換する横断的オミクス解析方法の開発がさかんである。

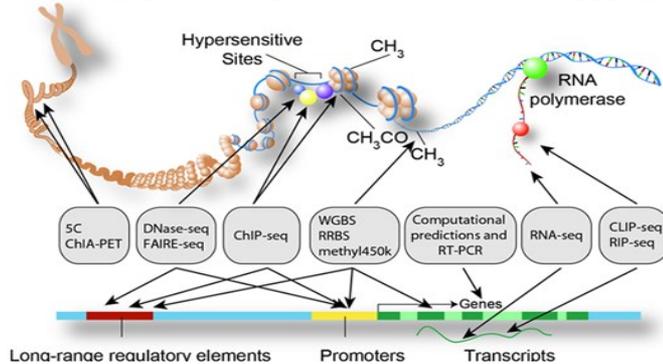
③ 疾患病態解明への展望

肥満の大規模ゲノム解析

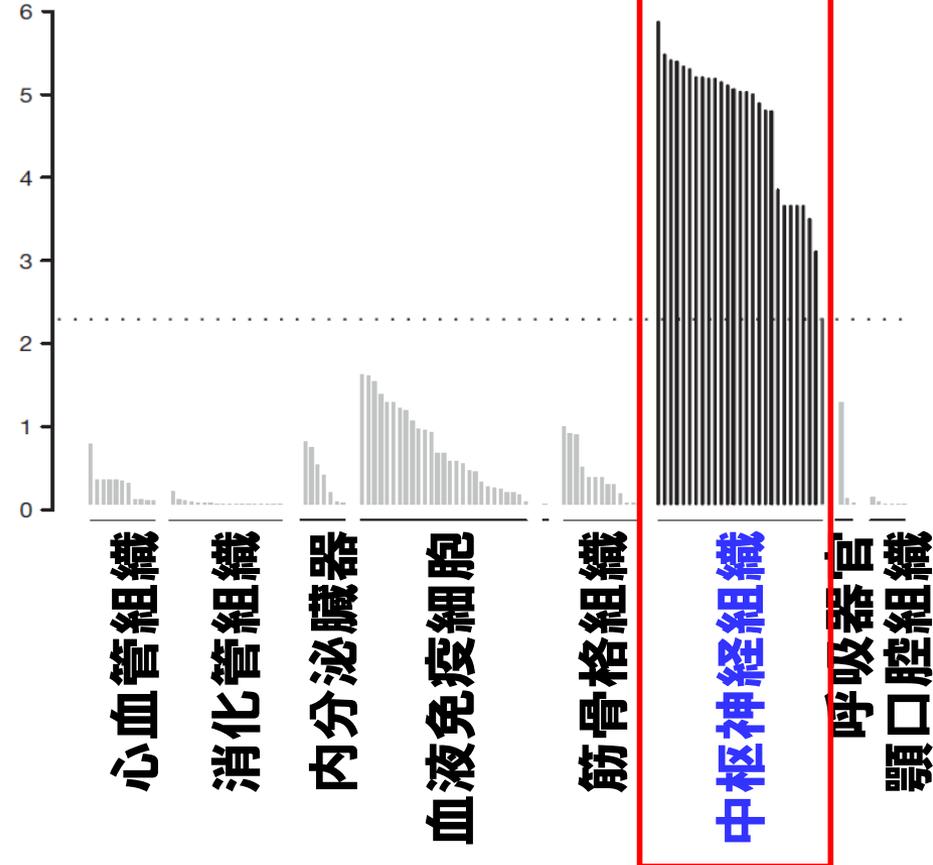


+

組織特異的エピゲノム情報



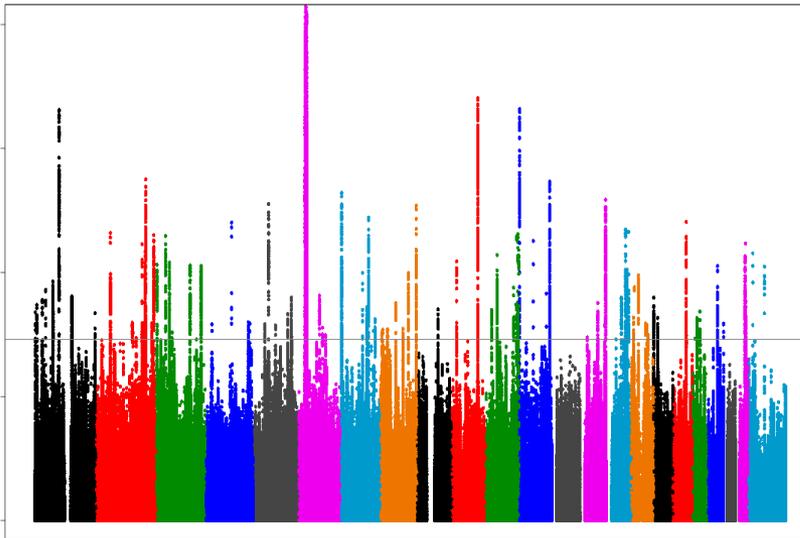
肥満感受性遺伝子の組織特異性



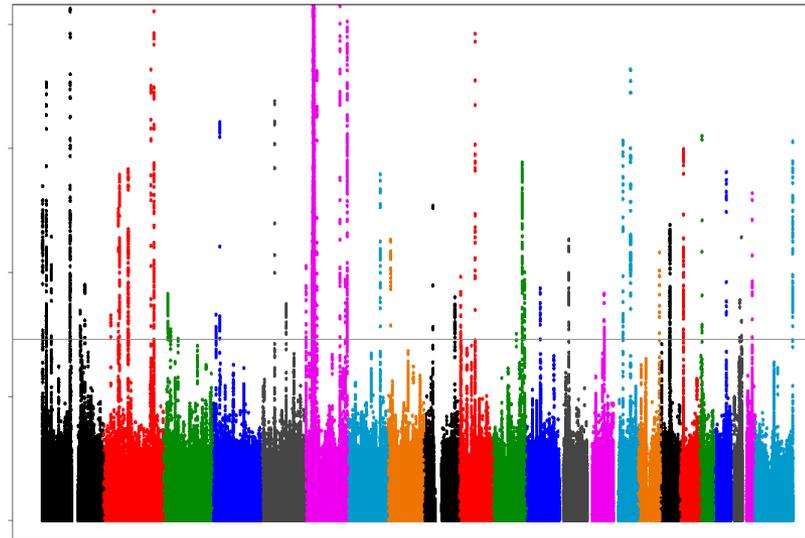
- **肥満**の大規模ゲノム解析と組織特異的エピゲノム情報(ENCODEプロジェクト)の統合により、「**脳疾患としての肥満**」が明らかになりました。

③ 疾患病態解明への展望

統合失調症のGWAS



関節リウマチのGWAS

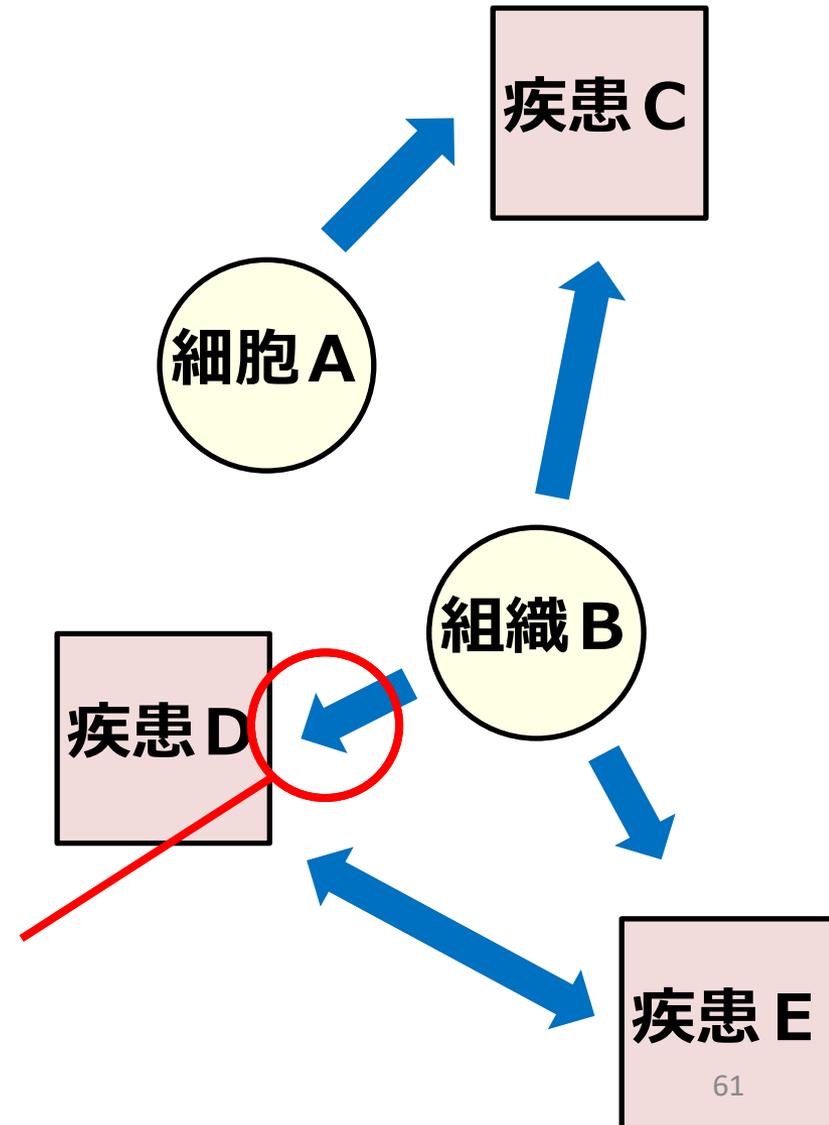
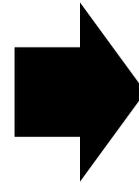
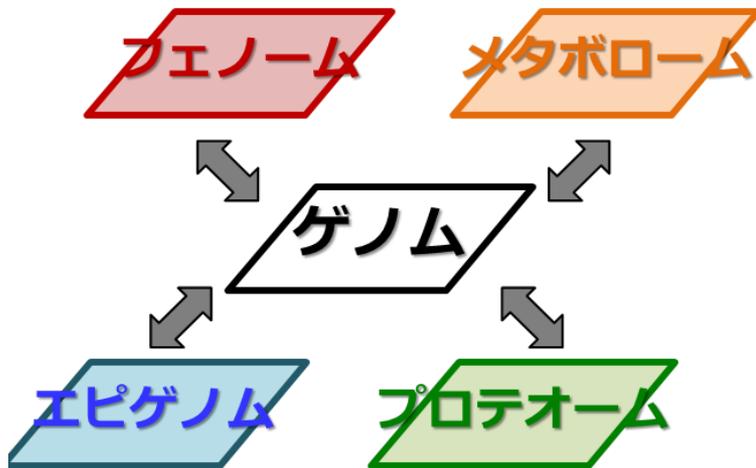


- 疫学調査により、統合失調症と関節リウマチが合併しにくい。
- もともと合併しにくいのか、治療介入の結果なのか、不明であった。
- 両疾患に対して実施されたGWASデータを照合することにより、「ゲノムレベルで合併しにくい」ことが明らかになりました。
- ヒトゲノム解析と疾患疫学との融合による、新たな発見といえます。

③ 疾患病態解明への展望

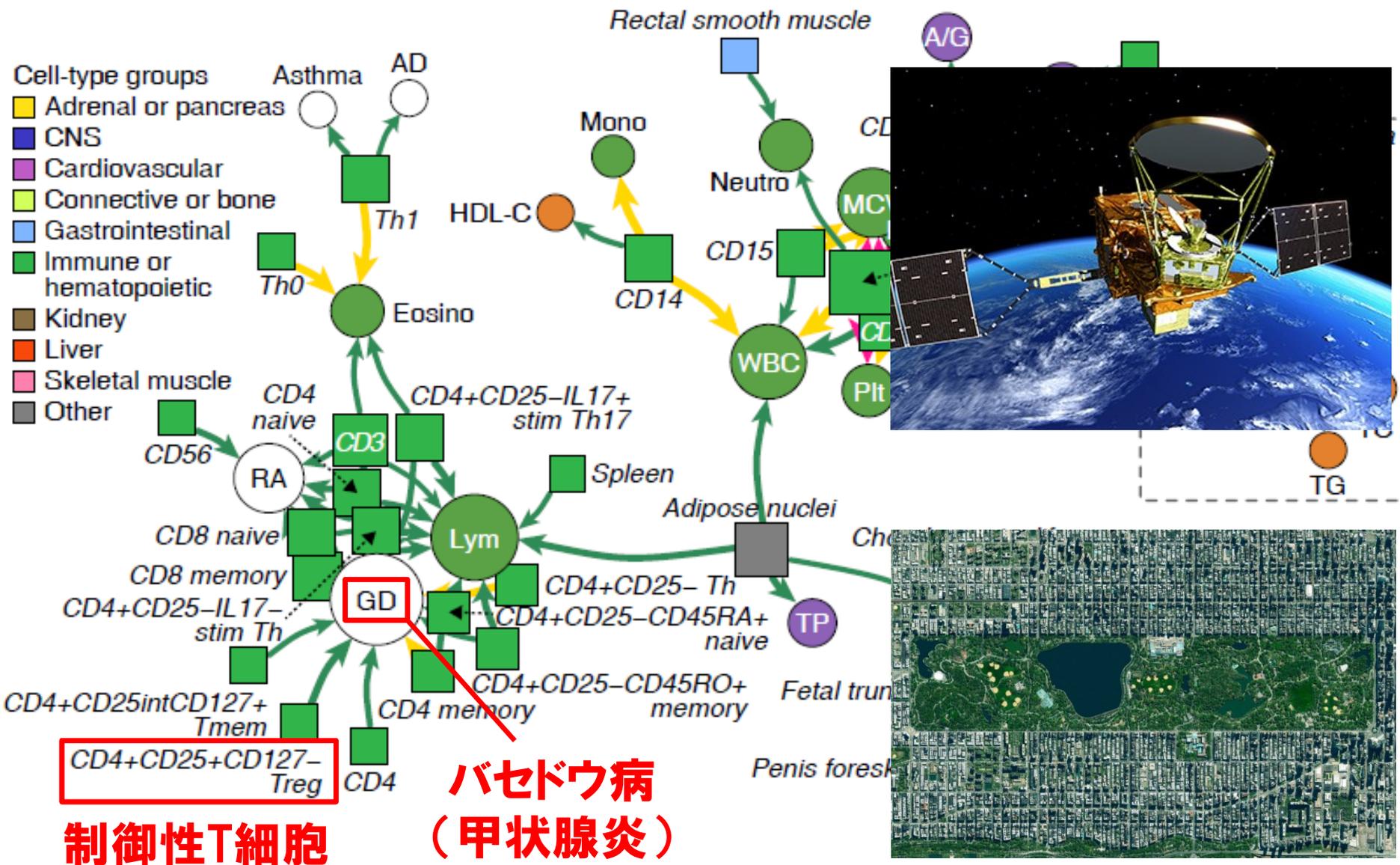
横断的オミクス解析による疾患ネットワーク解明

横断的オミクス解析



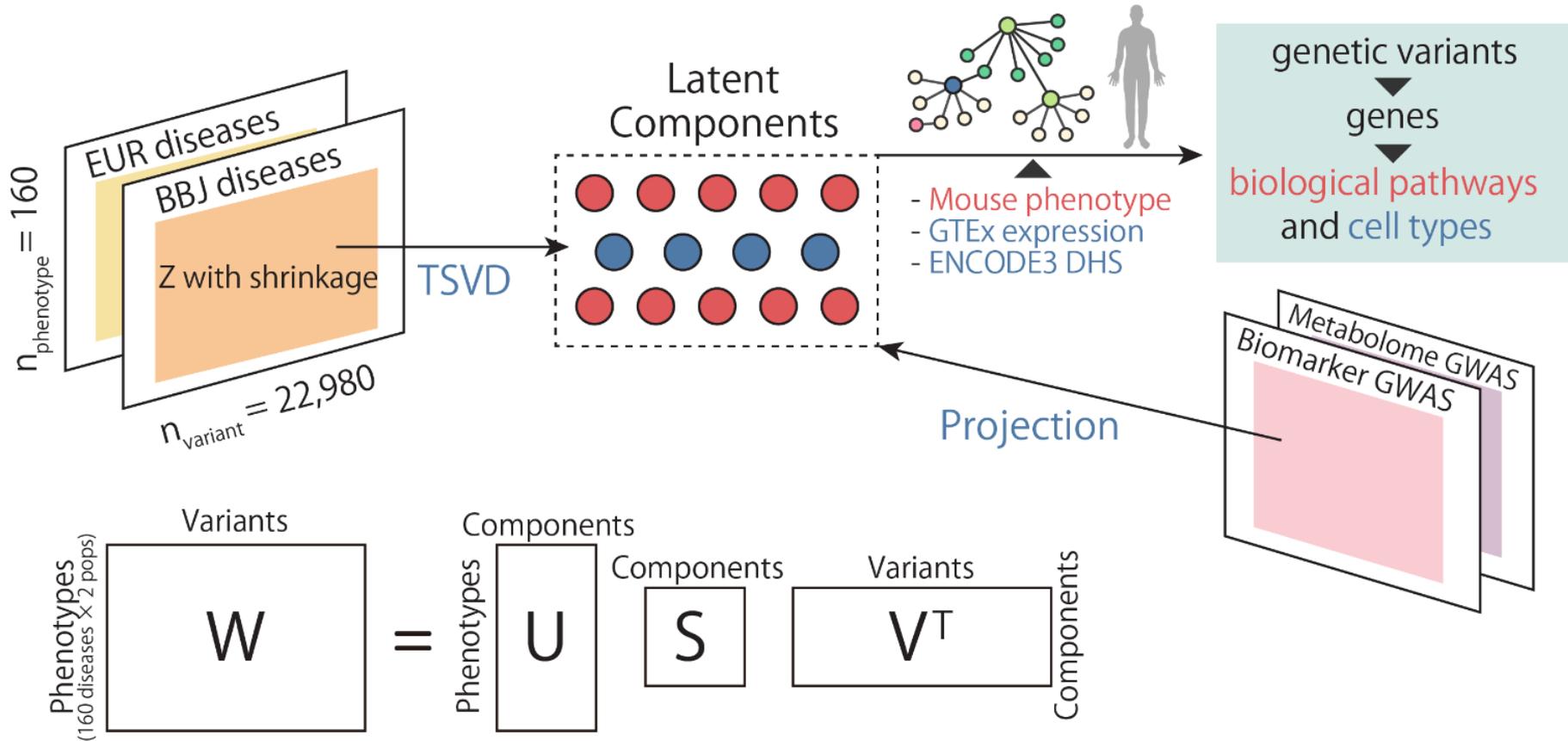
ゲノム情報の活用により
因果関係を反映した
ネットワーク構築が可能

③ 疾患病態解明への展望



③ 疾患病態解明への展望

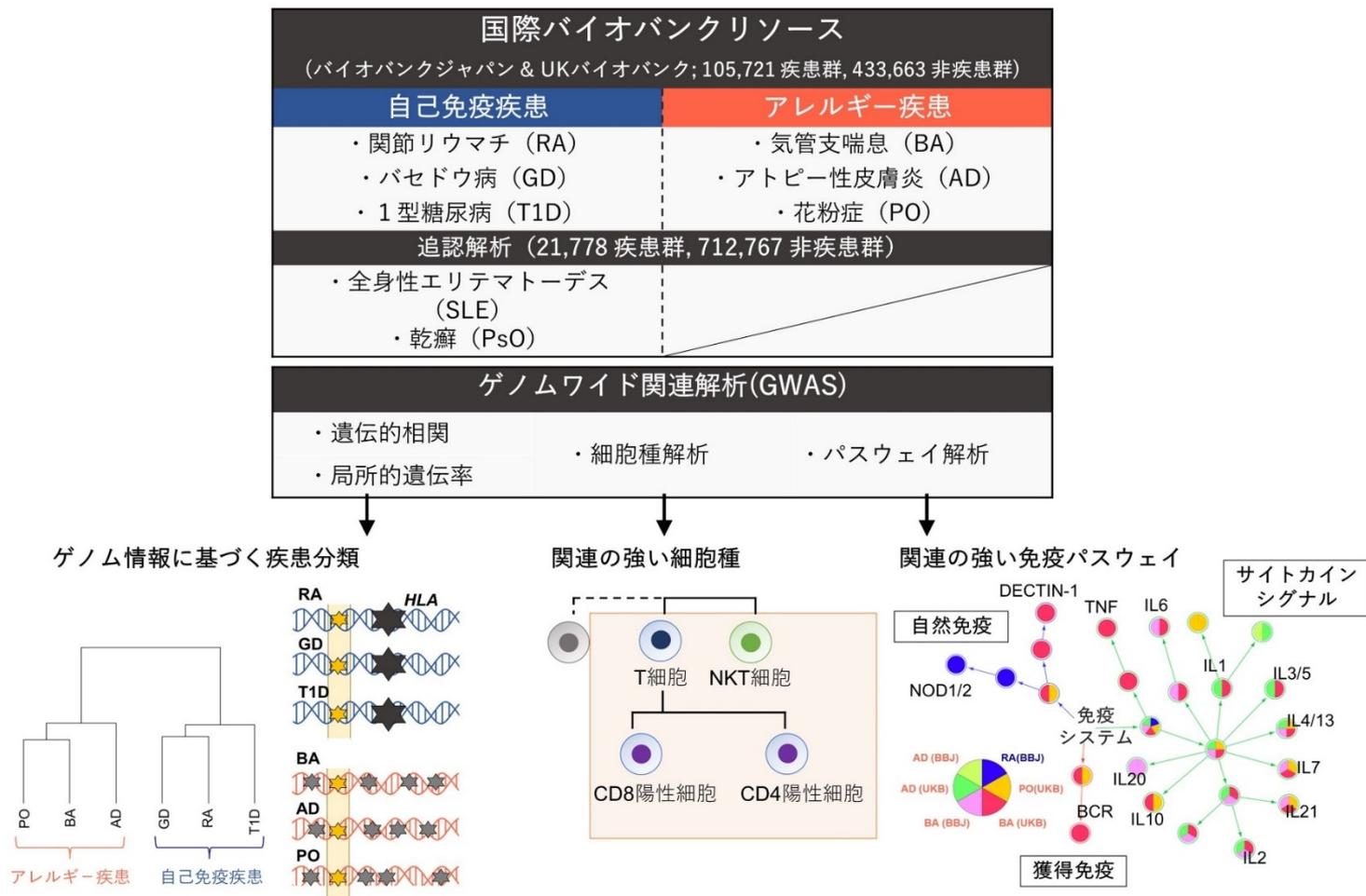
複数疾患の遺伝的背景の網羅的比較



- 多数のGWAS統計量への特異値分解(tSVD)の適用により、**関節リウマチとSLE、喘息・アレルギー性鼻炎・アトピー性皮膚炎**など、複数の形質に共通した感受性遺伝子・パスウェイを抽出可能になった。

③ 疾患病態解明への展望

複数疾患の遺伝的背景の網羅的比較

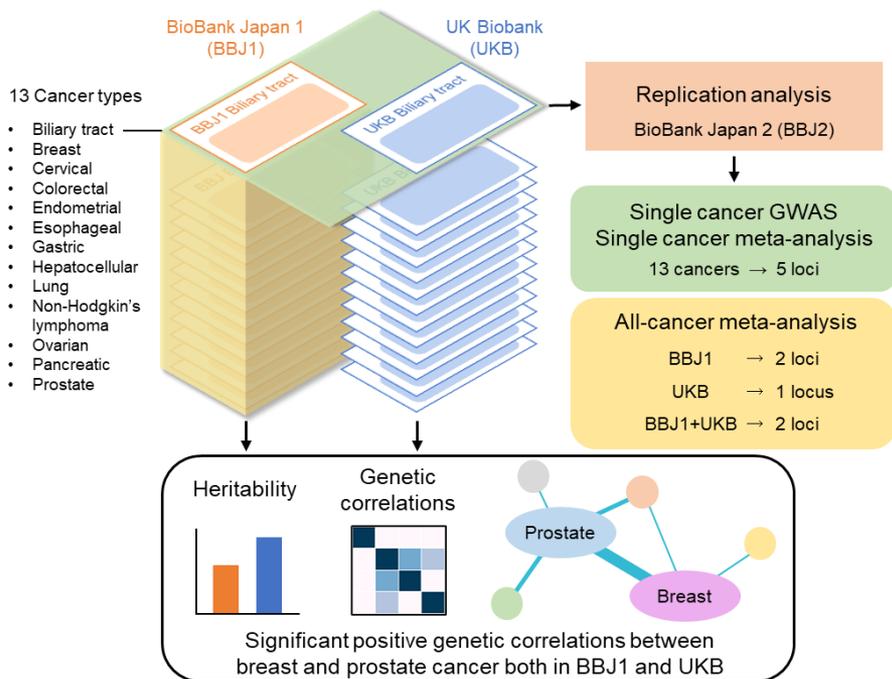


- 自己免疫疾患(関節リウマチ・バセドウ病・1型糖尿病)、アレルギー疾患(喘息・アトピー性皮膚炎・花粉症)を統合するGWASにより、共通した遺伝的背景や細胞組織特異性が明らかに。

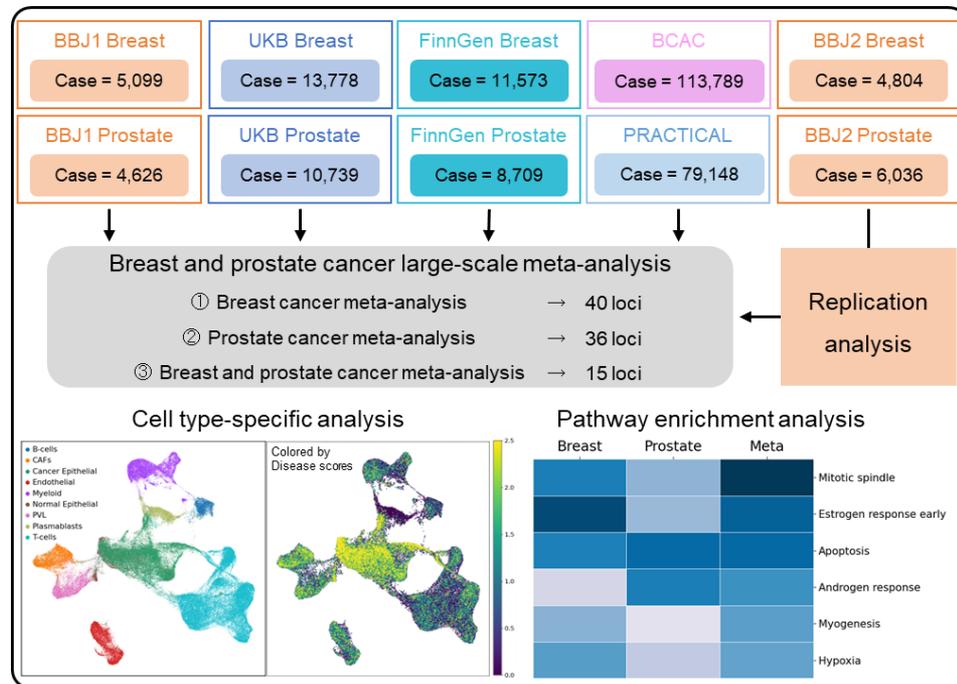
③ 疾患病態解明への展望

がん横断的な遺伝的背景の比較

13種類のがん横断的GWAS解析



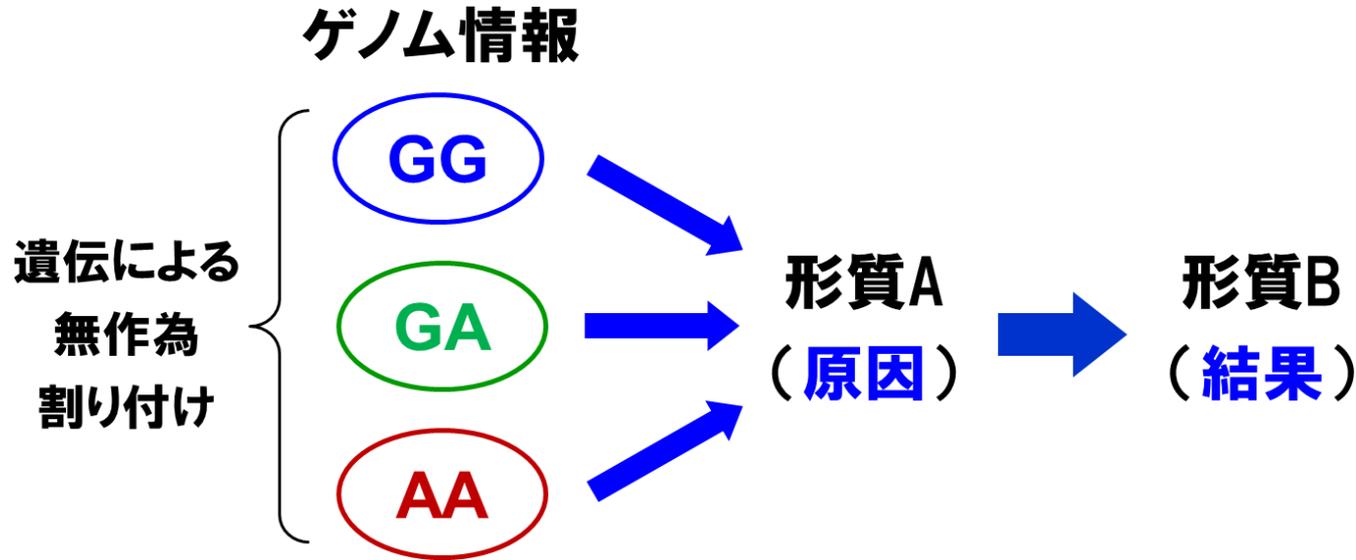
前立腺がん・乳がん に共通した遺伝的背景



- 13種類のがんを横断的に統合するGWAS解析を、人種集団横断的に実施。複数のがん種に共通してリスクを有する遺伝子領域を同定。
- 前立腺がんと乳がん で共通した遺伝的背景を同定。両者を統合するGWASで、感受性遺伝子領域と細胞組織特異性を同定。

③ 疾患病態解明への展望

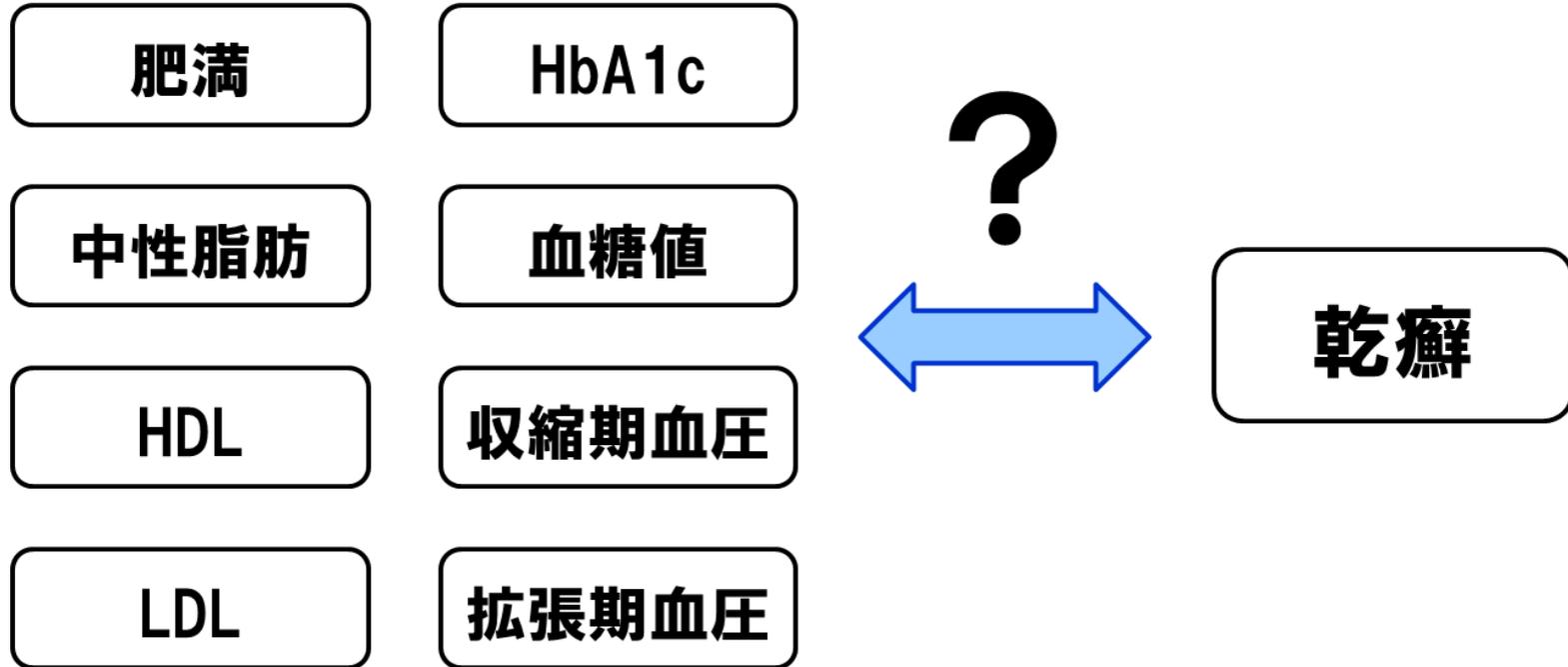
MRによる複数形質の遺伝的背景の因果関係の解明



- 形質間の遺伝的な相関関係と、因果関係は異なります。
- 複数の形質間の遺伝的背景の因果関係に着目した手法として、**Mendelian Randomization (MR)**が挙げられます。
- 遺伝情報が無作為抽出で親から子に伝わること(Mendelの法則)を、無作為化割り付け(Randomization)として活用する解析手法です。
- 脂質や炎症マーカーと心筋梗塞の因果関係の研究が有名です。

③ 疾患病態解明への展望

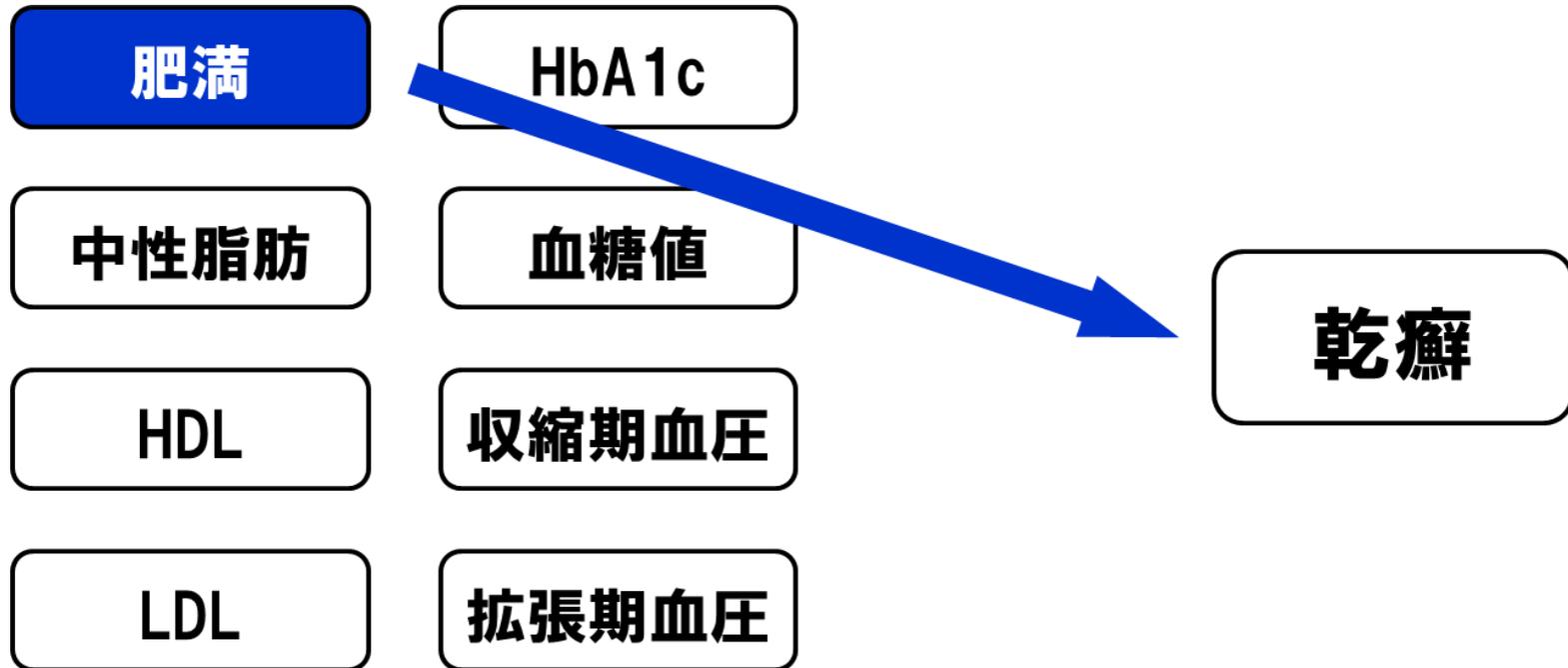
MRによる複数形質の遺伝的背景の因果関係の解明



- **乾癬罹患者ではメタボリック関連指標が悪化する傾向にあります。**
- **メタボリック指標が乾癬発症リスクを高めるのか、乾癬発症がメタボリック指標の悪化を引き起こすのか、因果関係の議論が続いていました。**
- **どのメタボリック指標が最も乾癬と関わりが強いのか、不明でした。**

③ 疾患病態解明への展望

MRによる複数形質の遺伝的背景の因果関係の解明

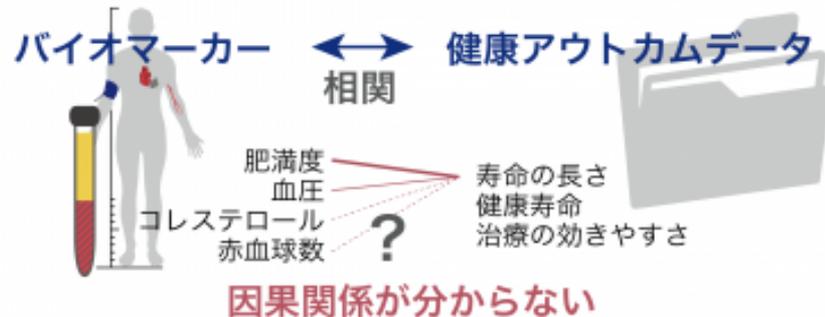


- 乾癬およびメタボリック関連指標のゲノムワイド関連解析結果を、欧米人集団および日本人集団で収集し、**大規模MR解析**を実施しました。
- **肥満が乾癬の原因**となることを明らかになりました。
- **肥満改善や肥満治療薬が乾癬治療標的**となる可能性が示されました。

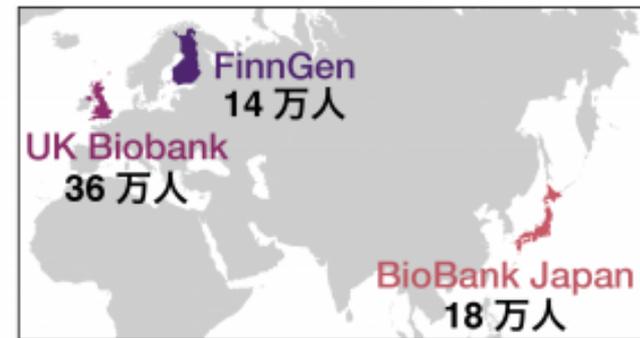
③ 疾患病態解明への展望

Polygenic risk scoreを活用した健康寿命バイオマーカー探索

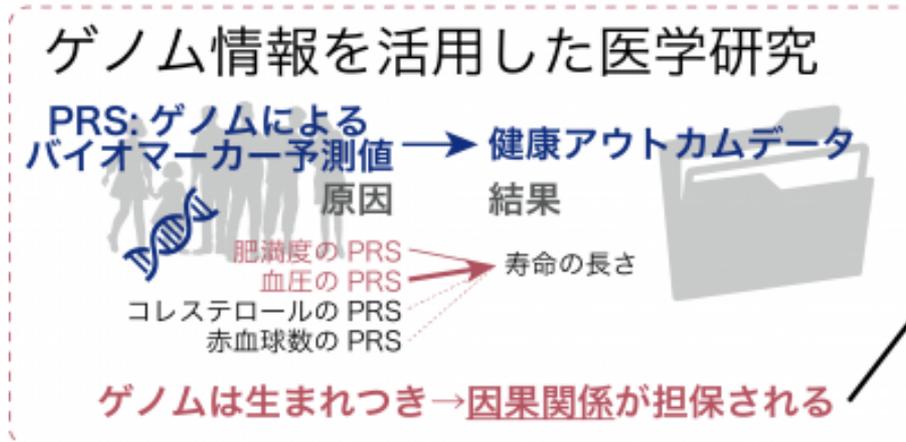
これまでの医学研究



世界3カ国 70万人のゲノムデータに応用



治療により健康アウトカムの改善が期待できる
バイオマーカーを特定

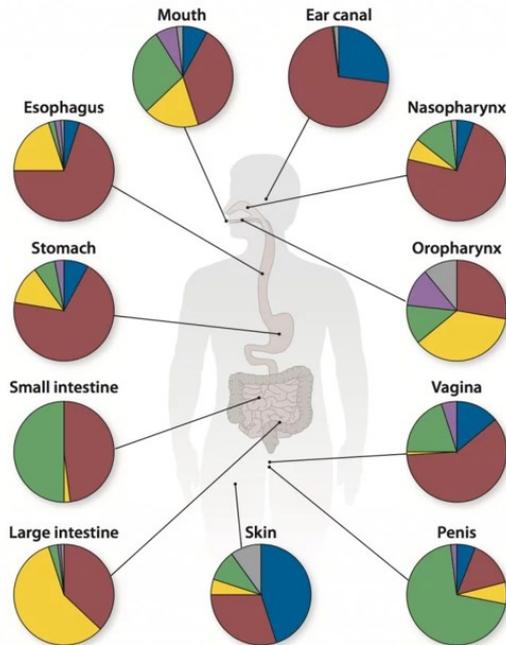


- 国際バイオバンク67万人のpolygenic risk score (PRS)解析を実施。健康の指標バイオマーカーPRSと寿命との因果関係を横断的に調べる手法を開発。
- 高血圧・肥満が健康寿命を短縮し、その治療が健康寿命延伸に貢献する可能性を示した。

③ 疾患病態解明への展望

- 宿主であるヒトや動物と共生関係にある多種多様な微生物の集まり。
- 代謝物の産生や病原体の排除、免疫の賦活化など、宿主と様々な相互作用を持ち、「**第二の臓器**」と呼ばれる。
- 微生物叢の個人差は腸管疾患に限らず、2型糖尿病や心血管障害、炎症性腸疾患、がんなど**多くの疾患の病因**に関与している。

様々な部位に共生する微生物叢



微生物叢のもたらす肥満

無菌のモデルマウス



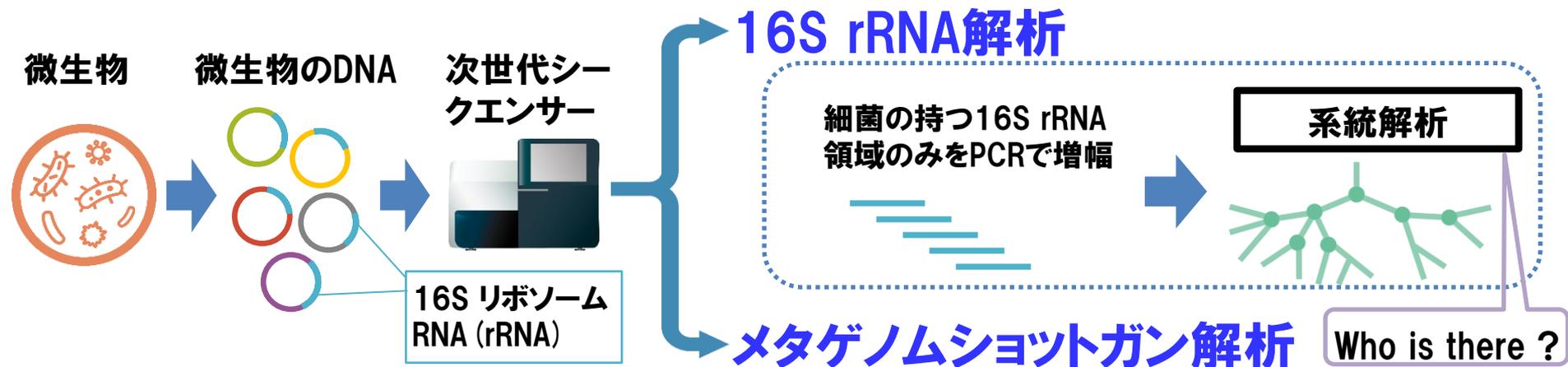
肥満マウスの腸内細菌叢を投与

痩せ型マウスの腸内細菌叢を投与

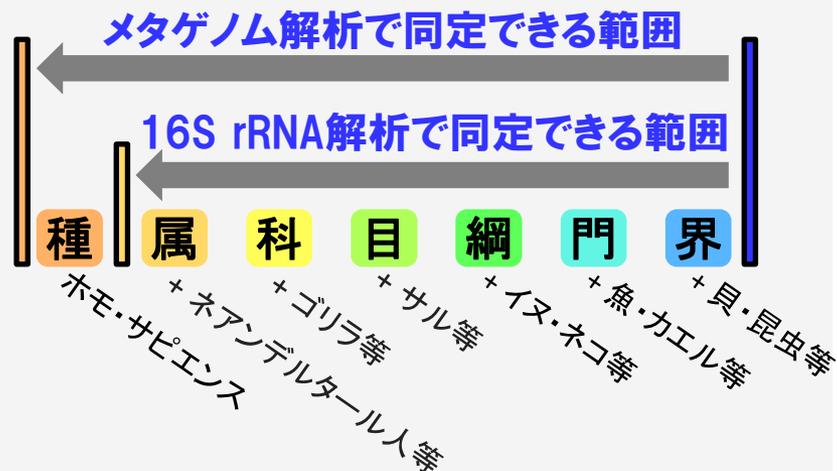


③ 疾患病態解明への展望

16S rRNA解析とメタゲノムショットガン解析



生物の階級・各解析の対象範囲



ショットガン シーケンシング

- 全DNAを断片化
- 細菌に加えて真菌、古細菌、ウイルスを対象

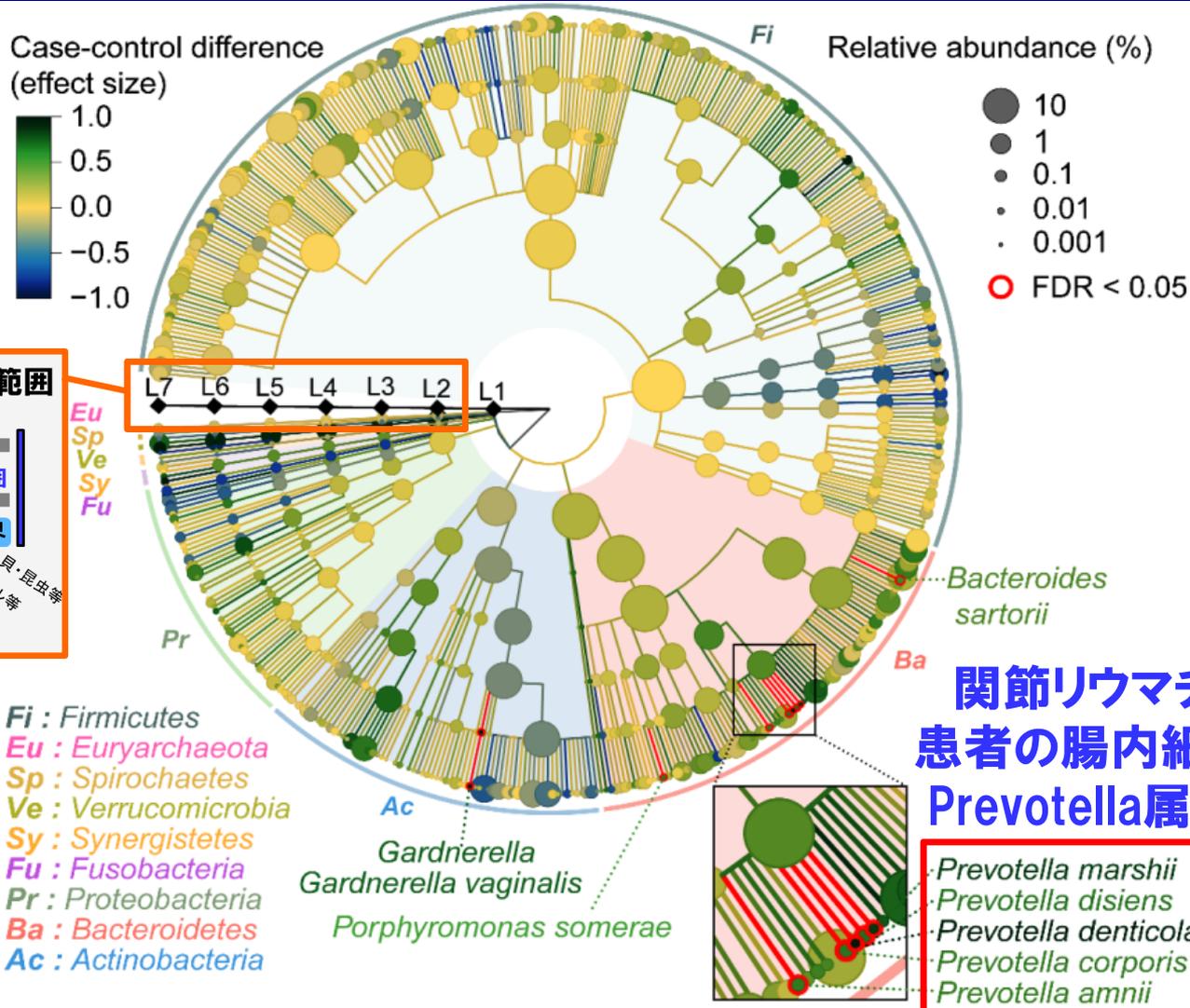


What are they doing ?



- 環境中に含まれる全微生物のゲノム情報の集合に対する網羅的な解析。
- 細菌のゲノムの一部を対象とする16S rRNA解析に代わる新しい手法。

③ 疾患病態解明への展望



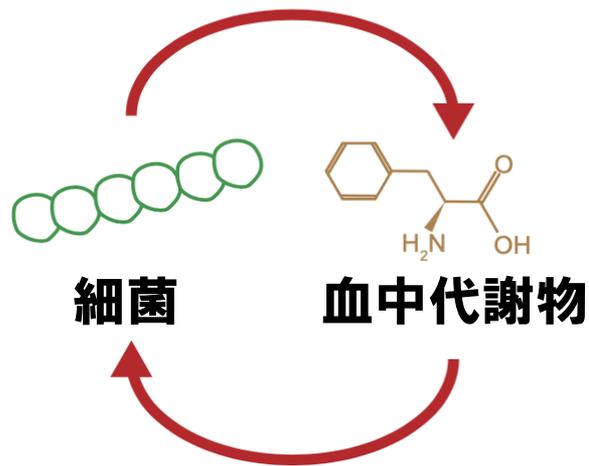
**関節リウマチ(RA)
患者の腸内細菌叢で
Prevotella属が増加**

微生物叢ショットガンシーケンスに基づくメタゲノムワイド関連解析により、関節リウマチに特異的な菌種・代謝物・パスウェイを同定。

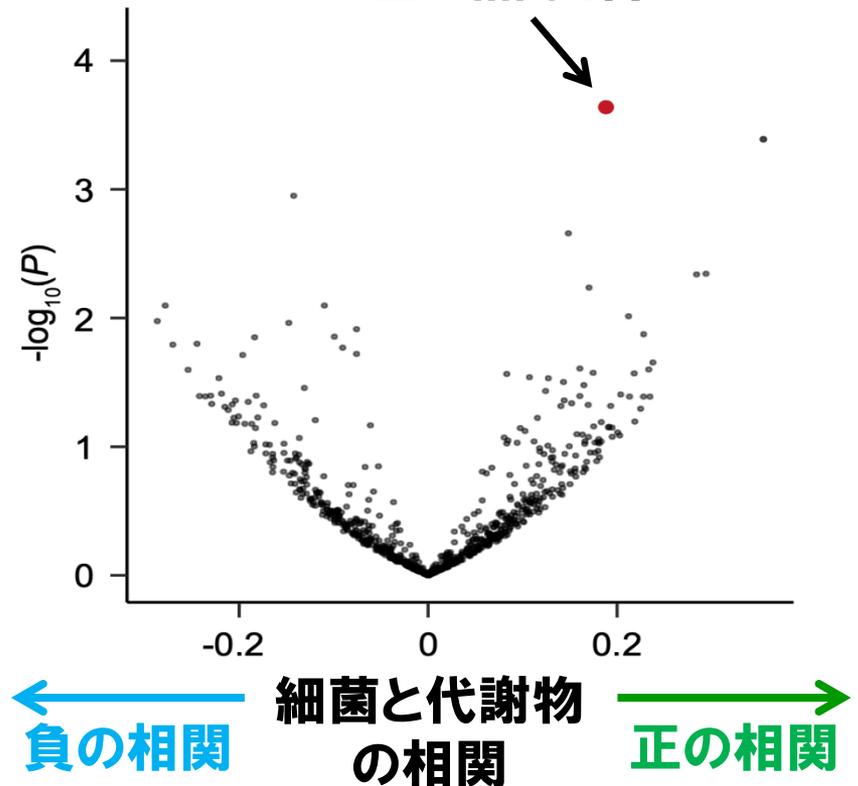
③ 疾患病態解明への展望

メタゲノム-メタボローム関連解析

SLE患者・対照群(103人)



Streptococcus intermediusと
アシルカルニチン(18:1)は
正の相関を持つ

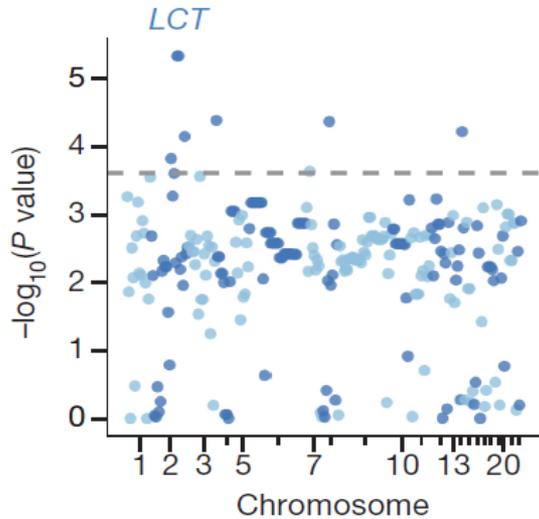


- 全身性エリテマトーデス(SLE)患者の腸内細菌叢でStreptococcus属が増加し血中代謝物アシルカルニチン量と正の相関を有することを同定した。

③ 疾患病態解明への展望

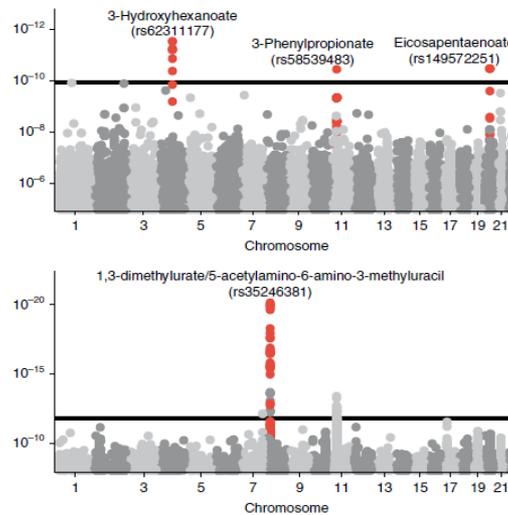
日本人集団メタゲノム示準データベース構築

メタゲノム⇔ヒトゲノム



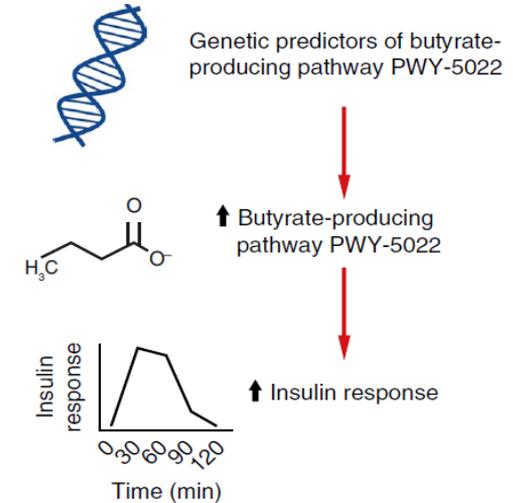
(Rothschild D et al. *Nature* 2018)

メタゲノム⇔メタボローム



(Zierer J et al. *Nat Genet* 2018)

メタゲノム⇔ヒトゲノム ⇔メタボローム

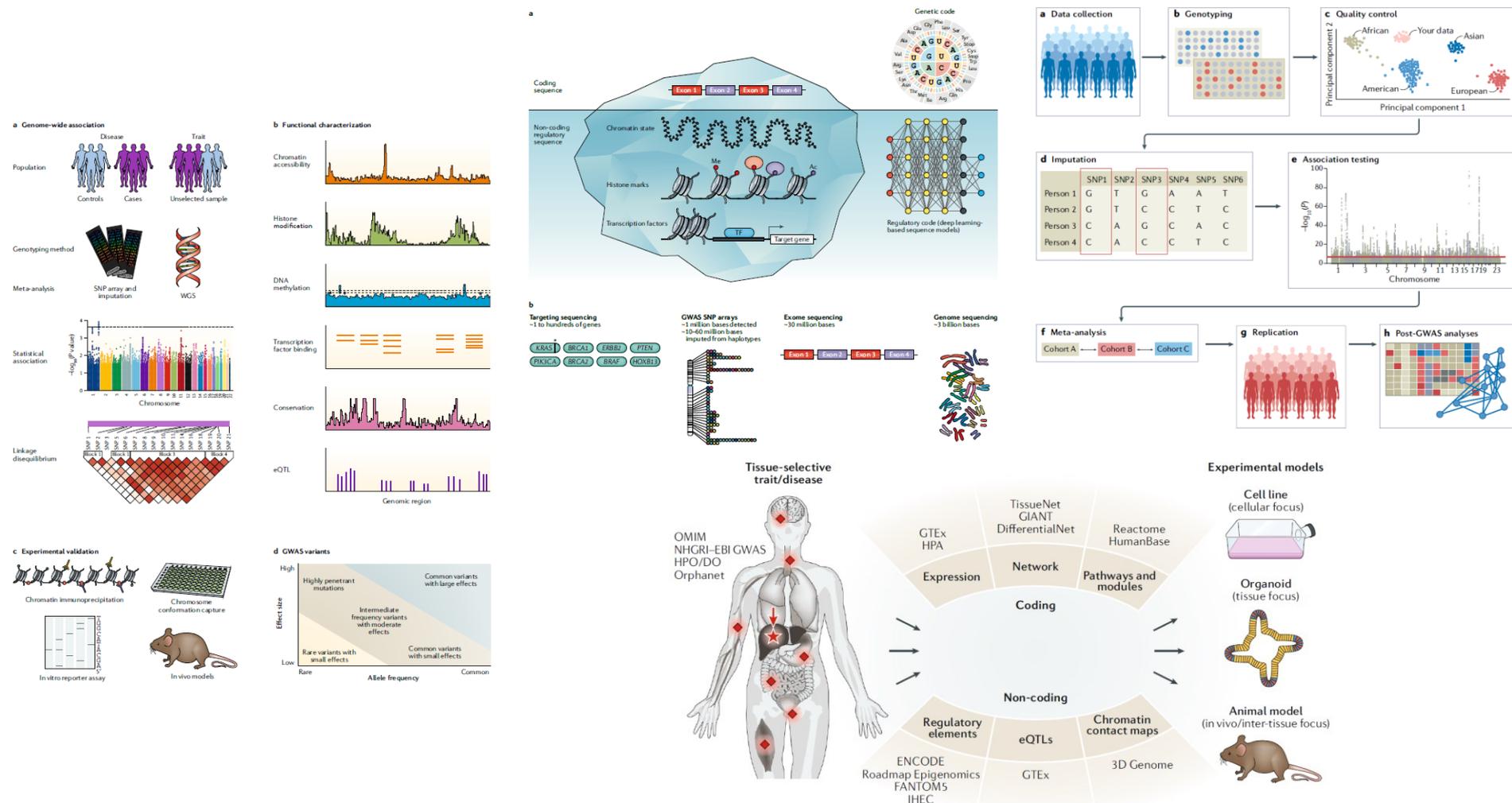


(Sanna S et al. *Nat Genet* 2019)

- 微生物叢と宿主の相互作用の理解には、各人種集団に特異的な、ヒトゲノム・メタゲノム・メタボロームの多様性のつながりの解明が不可欠。
- ヒト疾患の発症予測や病態解明においても有用と期待される。
(例: ヒトゲノム個人差 → 糞便微生物叢由来の短鎖脂肪酸 → インスリン抵抗性)
- 日本人集団においても、ヒトゲノム・メタゲノム・メタボローム多様性を網羅した示準データベース構築が不可欠。

③ 疾患病態解明への展望

• GWAS・オミクス解析の現状の把握には、総説論文が参考になります。



(Pasaniuc B et al. *Nat Rev Genet* 2017, Tam V et al. *Nat Rev Genet* 2019, Hekselman I et al. *Nat Rev Genet* 2020, Wang AK et al. *Nat Rev Genet* 2021, Uffelmann E et al. *Nat Rev Methods Primers* 2021)

遺伝学入門

- ① ヒトゲノム配列とその個人差
- ② ヒトゲノム多型を用いた疾患ゲノム解析
- ③ 疾患病態解明への展望
- ④ 機械学習のヒトゲノム情報への応用
- ⑤ ゲノム創薬への展望

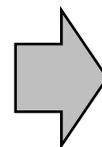
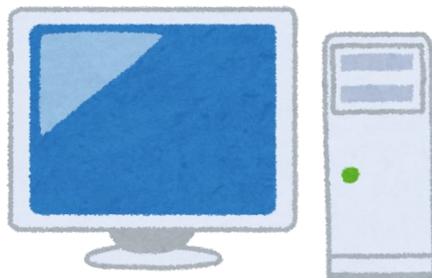
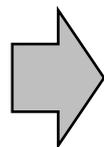
④ 機械学習のヒトゲノム情報への応用

機械学習とは何か？

入力データ

学習 & タスク

出力データ



- 与えられた**入力データ**に基づき**学習**し、**学習結果**に基づき設定された**タスク**をこなし、**出力データ**を返却する作業を、**機械学習**といいます。
- 機械学習には、小規模～大規模、シンプル～複雑、単数～複数、様々な学習内容が含まれます。

④ 機械学習のヒトゲノム情報への応用

人工知能・機械学習・ディープラーニングの包含関係

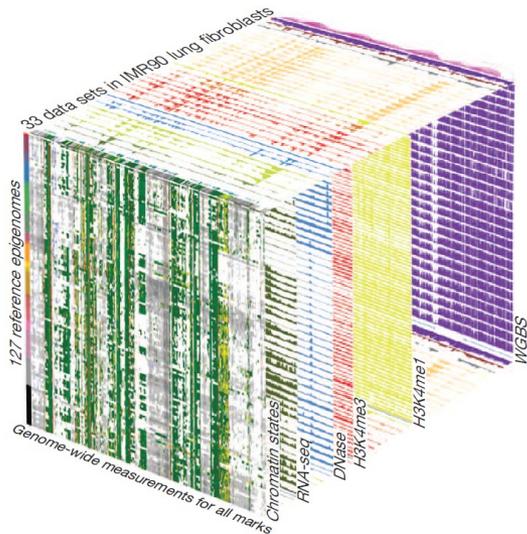


- 昨今注目されている**人工知能**([Artificial Intelligence; AI](#))や**機械学習**([Machine Learning; ML](#))もゲノム解析に有用です。
- 人工知能は、人間の知的活動をコンピューターで実現する技術です。
- 機械学習は、人工知能の実践的な解析技術の一つで、与えられたデータを学習して特徴を把握し、判断や予測を行うための手法になります。
- **ディープラーニング**は機械学習のアルゴリズムの一つで、脳の神経回路を模倣して、複数のデータ処理層を組み合わせていることが特徴です。

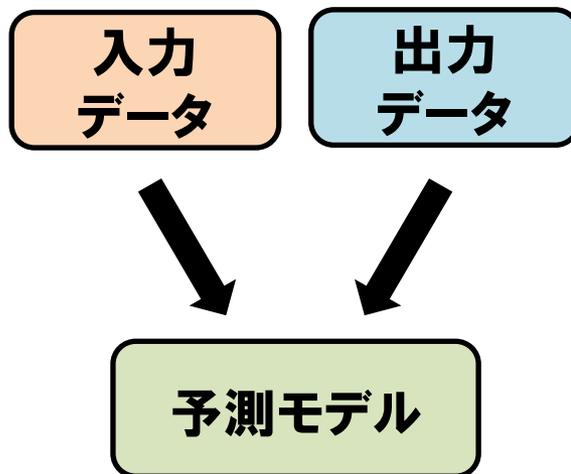
④ 機械学習のヒトゲノム情報への応用

機械学習における教師ありと教師なしの違い

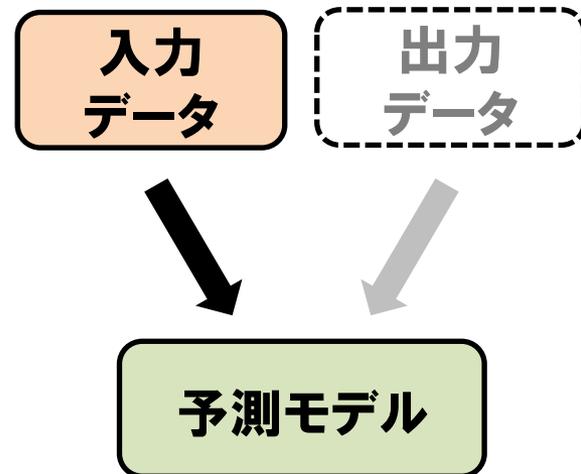
高次元データ



教師あり機械学習



教師なし機械学習

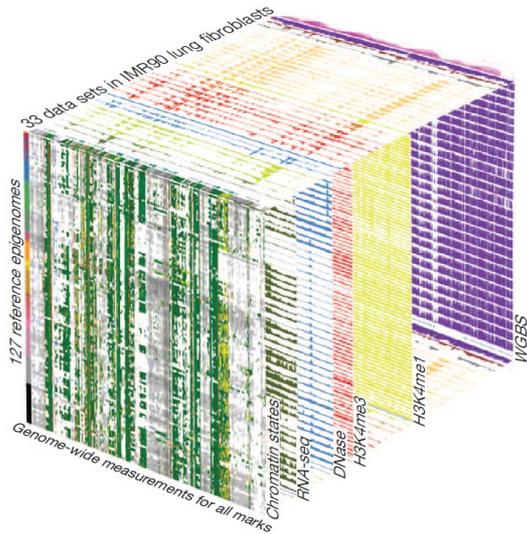


- 機械学習は大きく分けて**教師あり**と**教師なし**に分類されます。
- **教師あり**: 予め入力データと出力データ(=正解ラベル)が与えられた状態で、予測モデルを開発する手法です。
- **教師なし**: 入力データのみから、特徴量を捉えて、分類を行う手法です。
- ゲノムデータ解析の分野では、教師無し機械学習を適用することで、サンプルや遺伝子領域の分類を行うことがよくあります。

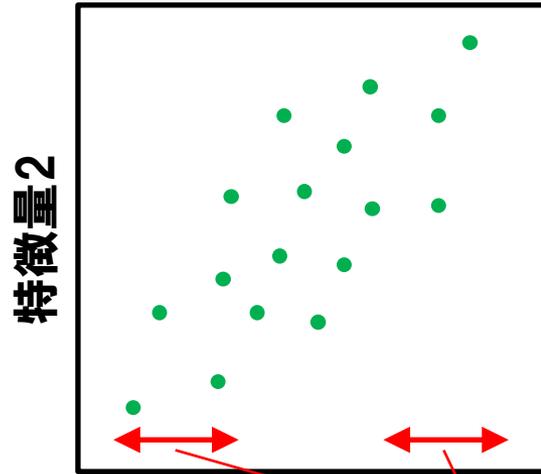
④ 機械学習のヒトゲノム情報への応用

機械学習における線形と非線形の違い

高次元データ



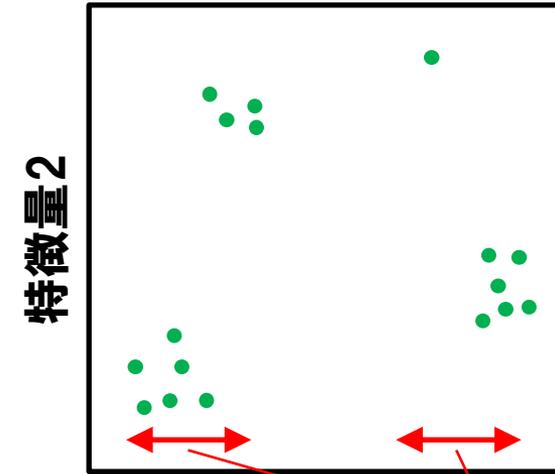
線形機械学習



特徴量1

同じ距離

非線形機械学習

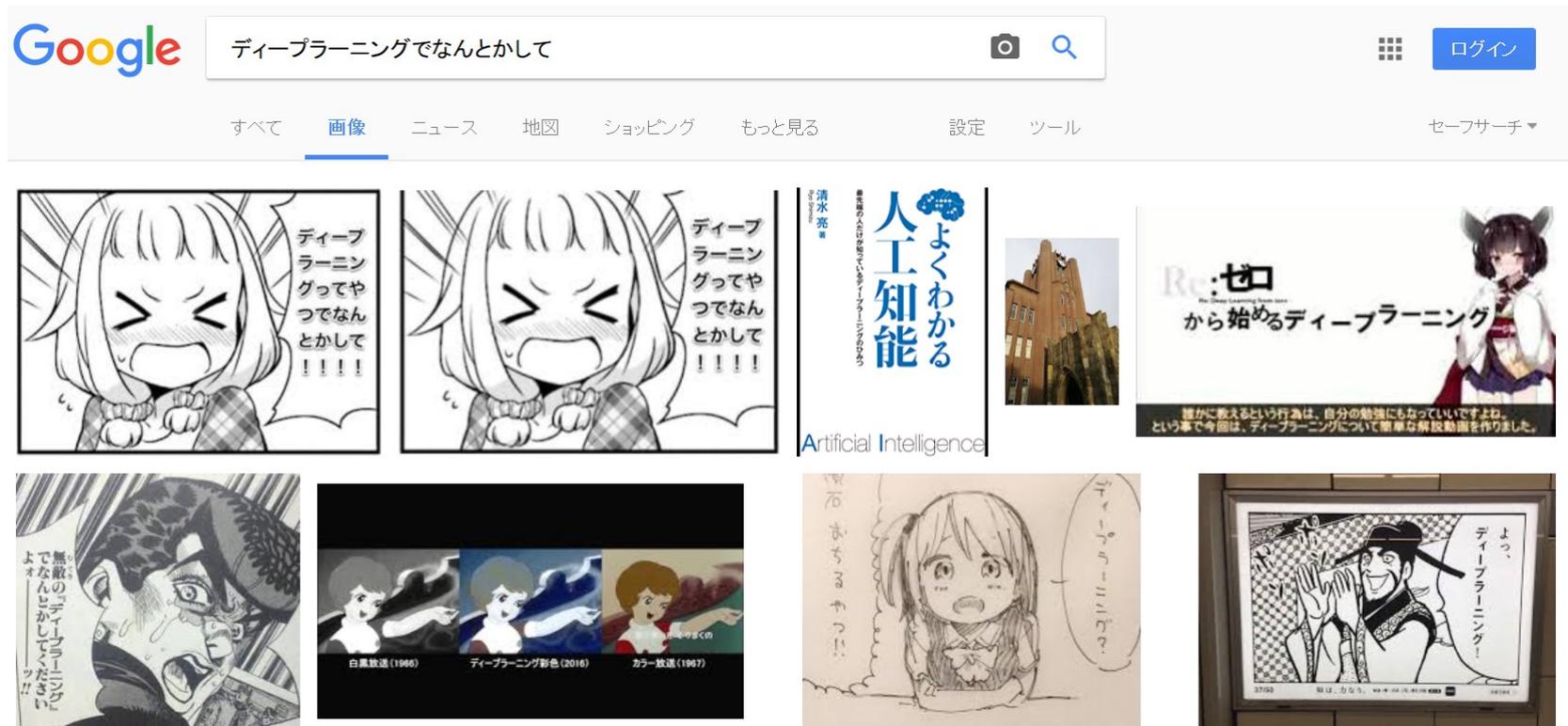


特徴量1

同じ距離でない

- 機械学習の計算方法は大きく分けて**線形**と**非線形**に分類され、高次元データから得られた各サンプルの**特徴量**の値の意味づけが異なります。
- **線形**: 特徴量の差とサンプル間の距離が比例関係にある。特徴量を**共変量**や異なる**機械学習の入力情報**として使用しやすいです。
- **非線形**: 特徴量の差とサンプル間の距離が比例しない。**サンプルを分離**する際に有用です。

④ 機械学習のヒトゲノム情報への応用



- 最近、人工知能が世間を賑わせています。
- GPU技術を活用したディープラーニングの応用と、対象となるデータ容量の膨大な増加がブームの立役者となっています。
- ゲノム情報にディープラーニングを適用する研究も進んでいます。
- そのためには、ゲノムデータ行列の特徴を知ることが重要となりそうです。

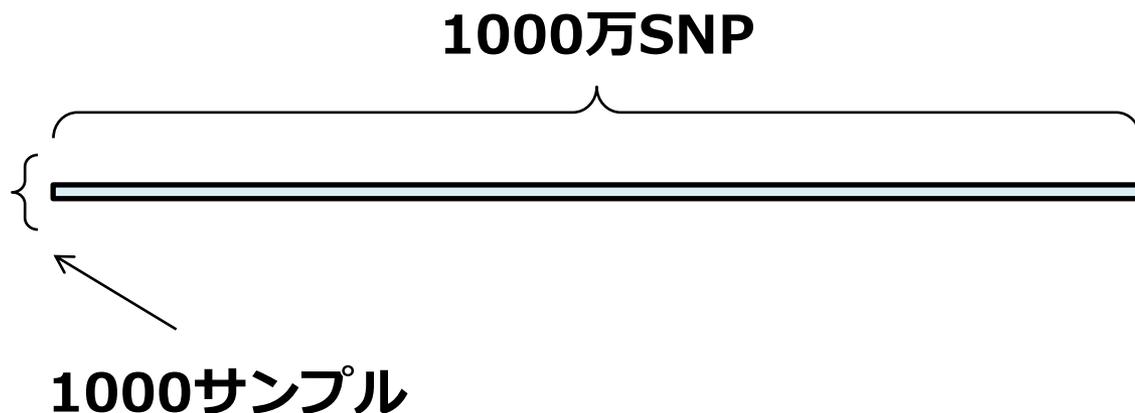
④ 機械学習のヒトゲノム情報への応用

ゲノムデータ行列の特徴: その①

通常の行列

01201201201201201
01201201201201201
01201201201201201
⋮
01201201201201201
01201201201201201
01201201201201201

ヒトゲノムデータ行列



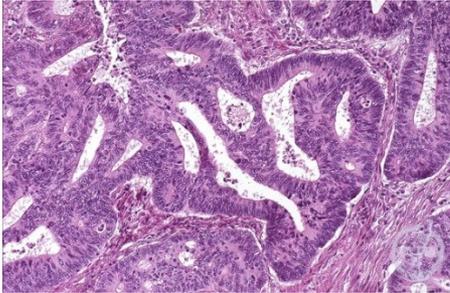
- ゲノムデータ行列は**行と列の長さが著しく違う**、という特徴があります。
- $p \gg n$ 問題や、機械学習における過学習を引き起こす原因となります。
- リッジ回帰やLASSOでも**完全には解決できず、今後の課題**です。

④ 機械学習のヒトゲノム情報への応用

ゲノムデータ行列の特徴: その②



画像データ



ゲノムデータ

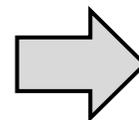
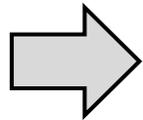
	SNP1	SNP2	SNP3	SNP4
Sample1	1	0	1	1
Sample2	1	2	1	2
Sample3	0	1	1	1
Sample4	0	2	0	2
Sample5	1	1	1	0
Sample6	1	1	1	1

- **ディープラーニング**: 複雑系を比較的単純なデータに写像する作業。
- **主成分分析**: 単純なデータから、データを代表する特徴量を抽出。
- ゲノムデータは、**比較的単純なデータ**に分類されます。

④ 機械学習のヒトゲノム情報への応用

ディープ
ラーニング

	SNP1	SNP2	SNP3	SNP4
Sample1	1	0	1	1
Sample2	1	2	1	2
Sample3	0	1	1	1
Sample4	0	2	0	2
Sample5	1	1	1	0
Sample6	1	1	1	1



or

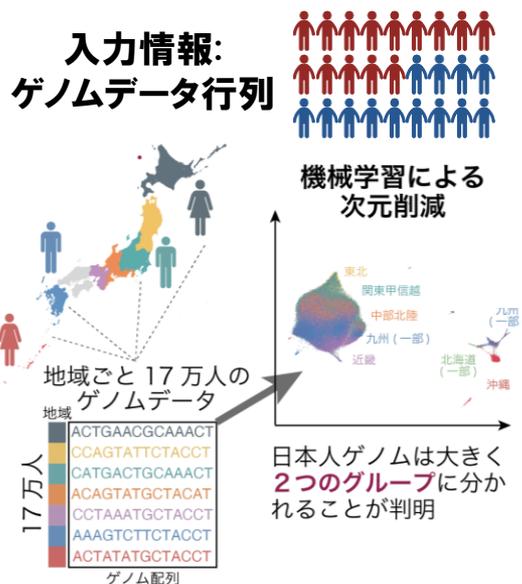


- ディープラーニングは、**解析の途中経過がブラックボックス**です。
- どんなに正しい予測でも、**ブラックボックスから得られた結果を医療に使ってもいいのでしょうか？**
- 予測が万一間違った場合、**人工知能は責任を取れるのでしょうか？**
- ゲノム情報への人工知能への適応は、**技術面以外の課題**も多そうです。

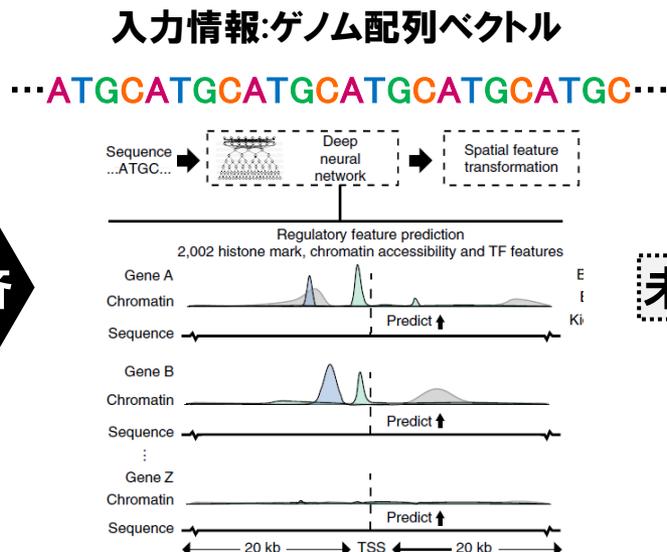
④ 機械学習のヒトゲノム情報への応用

- 人工知能技術(AI)における**深層学習の技術変革**が著しく、従来は困難であった問題の解決を生命科学の広い分野で達成しています。
- ヒトゲノム解析では、**集団ゲノムデータへの機械学習** → **参照ゲノム配列への深層学習**へと進化するも、**集団ゲノムデータへの深層学習**が課題です。
- 集団ゲノムデータ行列への深層学習の適用は、高精度予測能を要する個別化医療実装に必須と期待されるが、フレームワークの構築は道半ばです。

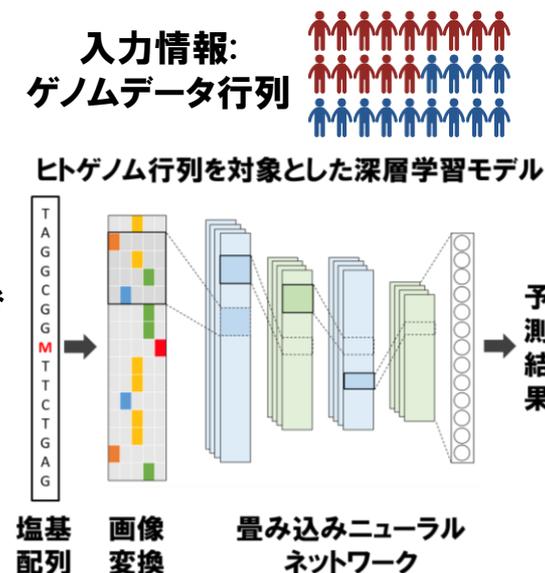
集団ゲノムデータ行列への教師なし機械学習



参照ゲノム配列ベクトルへの教師あり深層学習



集団ゲノムデータ行列への教師あり深層学習



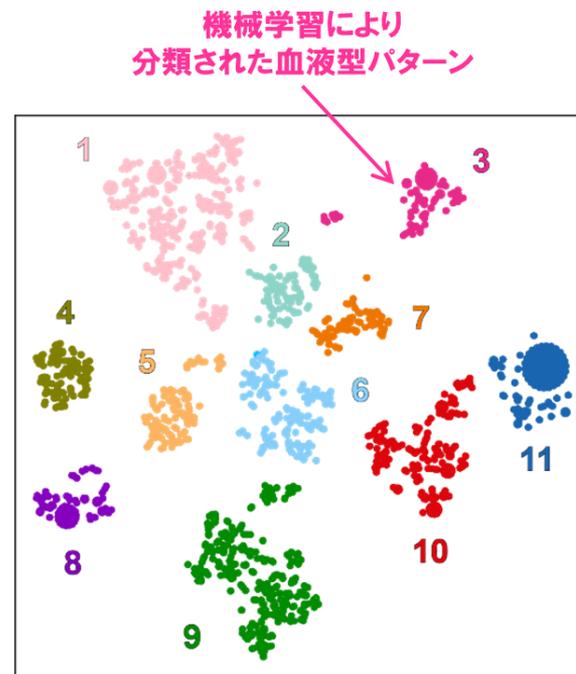
④ 機械学習のヒトゲノム情報への応用

ゲノムデータ行列に対する**非線形機械学習**の適用

次世代シーケンス技術
によるHLA遺伝子解読



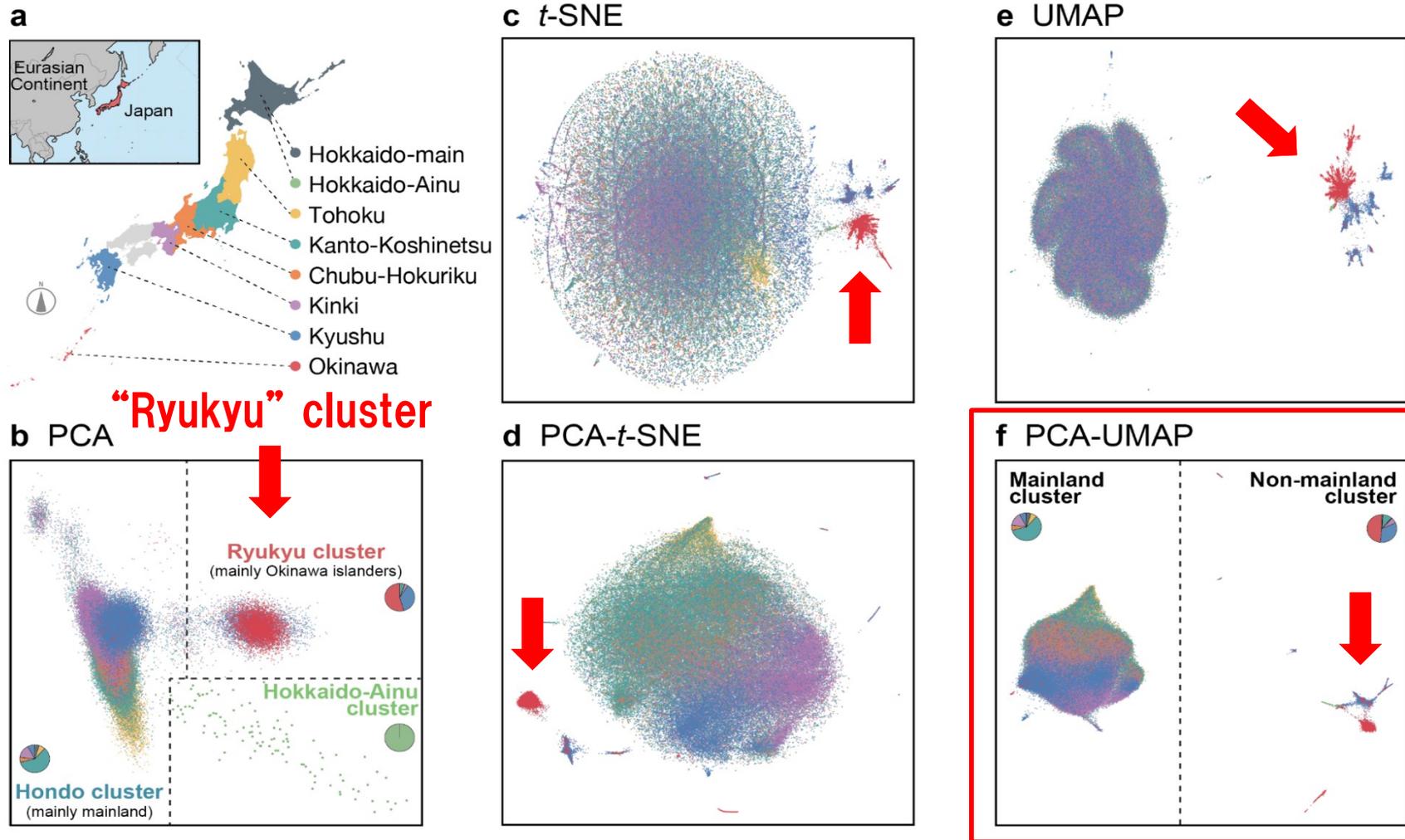
機械学習による日本人集団の
白血球の血液型分類



- ゲノムデータ行列に対する**非線形機械学習手法の適用**も進んでいます。
- HLA遺伝子型への**非線形機械学習の適用**により、日本人集団の白血球の血液型のパターン分類が可能になりました。

④ 機械学習のヒトゲノム情報への応用

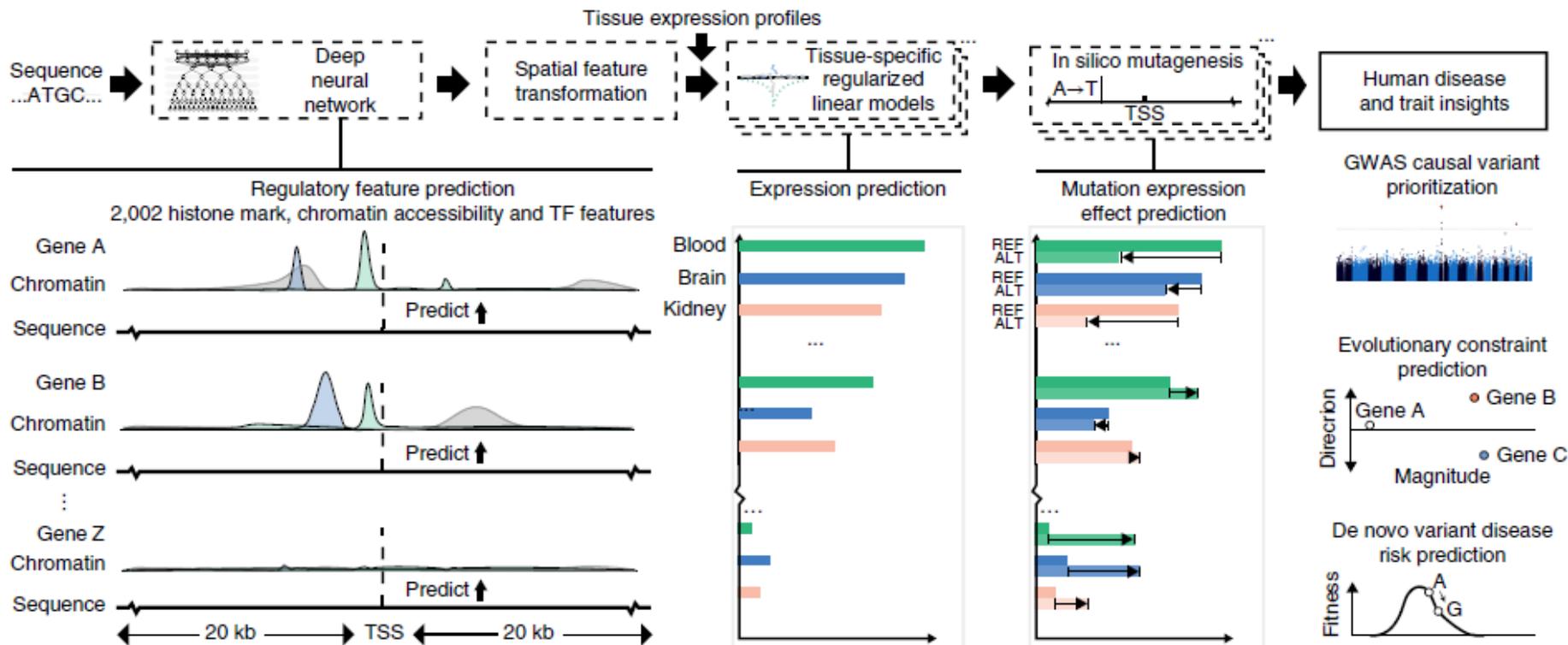
ヒト集団ゲノムデータ行列への教師なし深層学習の適用



• 線形-非線形のハイブリッド型機械学習により、日本人集団の詳細な構造化の解明(“Hondo”, “Ryukyu”, “Hokkaido-Ainu”)が可能になりました。

④ 機械学習のヒトゲノム情報への応用

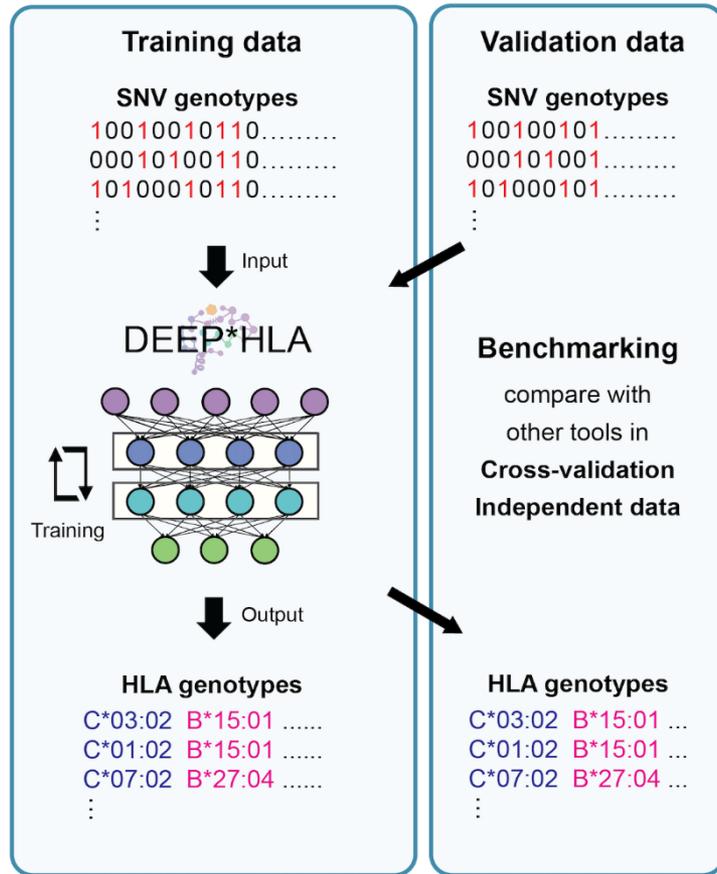
ヒトゲノム配列ベクトルへの教師あり深層学習の適用



- 30億のヒトゲノム配列そのものをディープラーニングの対象とすることで、ヒトゲノム配列だけからエピゲノム修飾や遺伝子発現制御を予測し、ヒト疾患に(真の)関連を持つ遺伝子変異を絞り込むことが可能となり、注目を集めています。

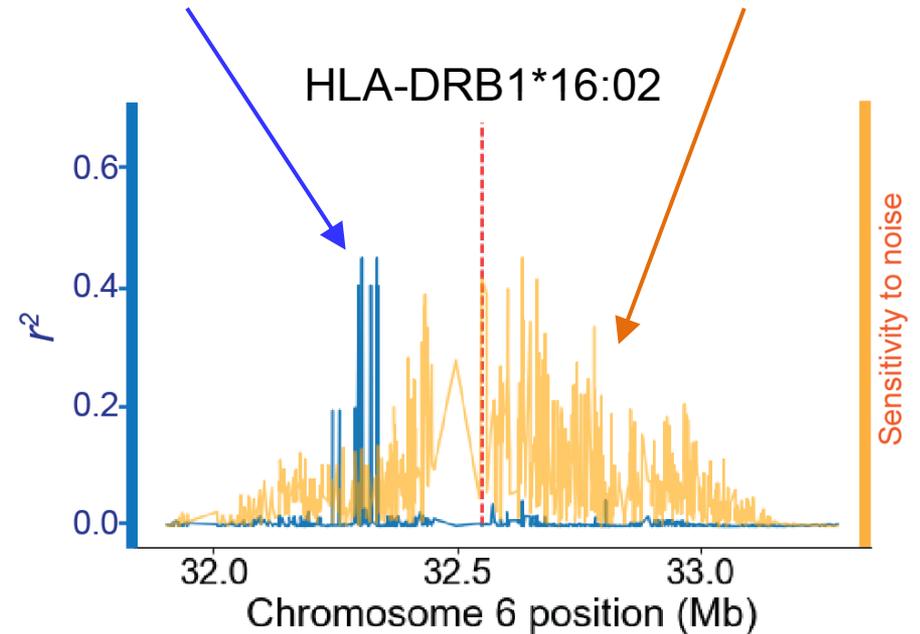
④ 機械学習のヒトゲノム情報への応用

ヒト集団ゲノムデータ行列への教師あり深層学習の適用



既存の機械学習が
参考にする箇所

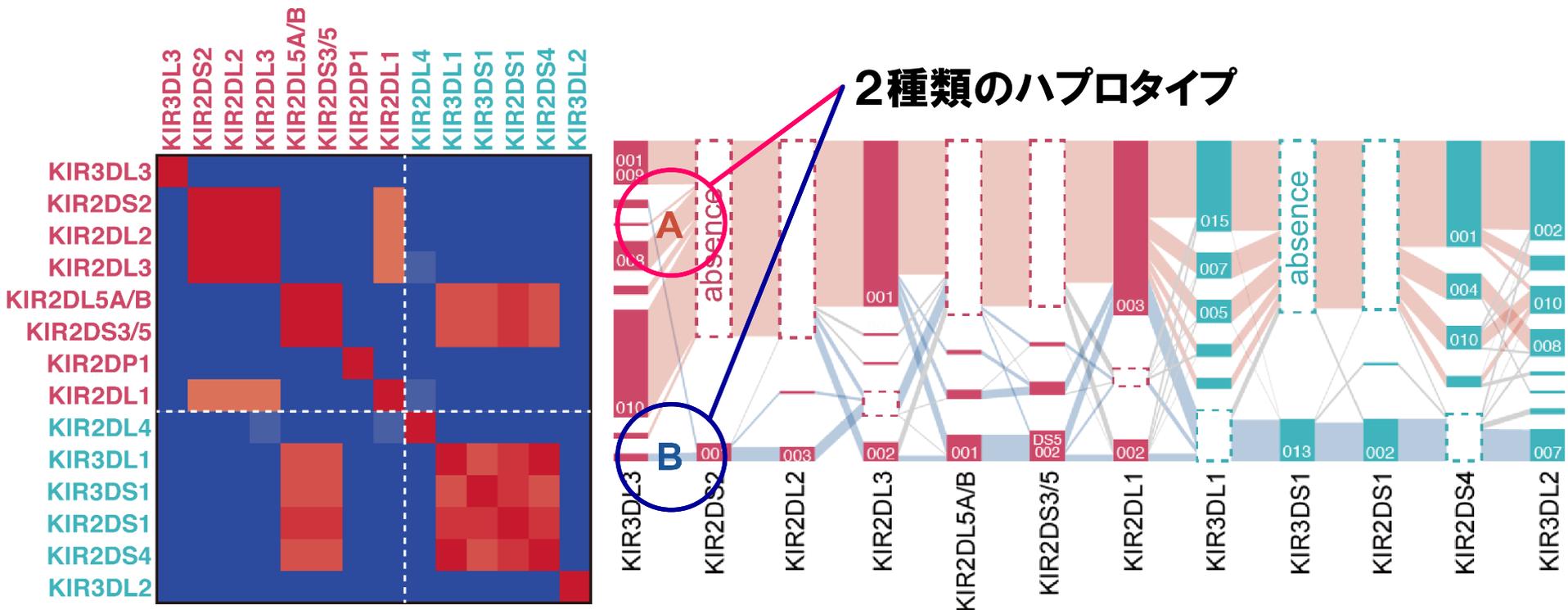
深層学習が
参考にする箇所



- 深層学習によるHLA imputation法(DEEP*HLA)を開発。MHC領域内のヒト集団ゲノム行列を画像変換することで深層学習の適用を可能にした。
- 従来の機械学習(例:マルコフ連鎖)と比較して、**希なHLA遺伝子型の推定精度が改善**。

④ 機械学習のヒトゲノム情報への応用

KIR imputation法の日本人集団への実装

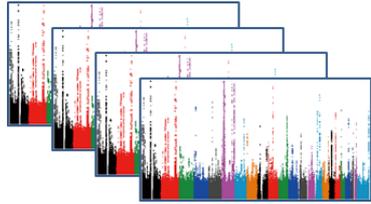


- 日本人集団1,150名を対象に、次世代シーケンサーを用いた**全KIR関連遺伝子型**を詳細に決定。
- KIR遺伝子間のLD構造・ハプロタイプ構造を解明し、**A/Bの2種類のハプロタイプの組み合わせ**で説明されることが明らかとなった。
- フェノムワイド関連解析(PheWAS)を70以上のヒト形質に対して実施するも、予想に反し**有意な関連は認められなかった**。

④ 機械学習のヒトゲノム情報への応用

入力情報

疾患ゲノム情報

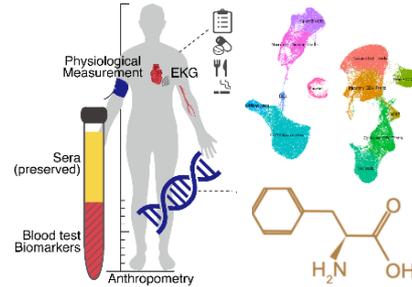


$$PRS = \sum_i \beta_i x_i$$



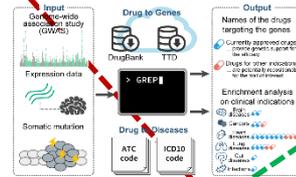
出力情報

形質情報



遺伝子変異 機能注釈

...AGCTCCATGCATGTCATGC...
 ...T...
 ...T...
 ...A...



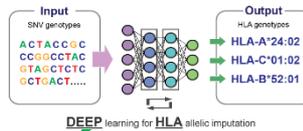
個人ゲノム

DRB1*01:01
 DRB1*01:04
 DRB1*04:05
 DRB1*09:01
 ...

集団ゲノム情報

ACTGAACGCAA	ACT
CCAGTATTCTAC	CT
CATGACTGCAA	ACT
ACAGTATGCTAC	AT
CCTAAATGCTAC	CT
AAAGTCTTCTAC	CT
ACTATATGCTAC	CT

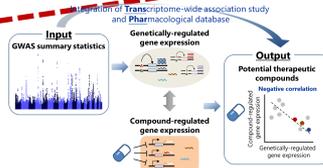
DEEP+HLA



創薬標的



Trans-Phar

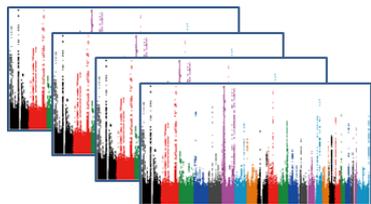


•「Uni-modal」から「Multi-modal」機械学習への転換による「ヒト人生」の予測。

④ 機械学習のヒトゲノム情報への応用

入力情報

疾患ゲノム情報

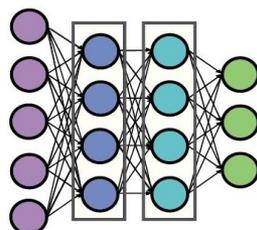
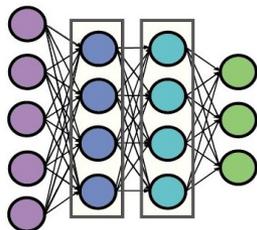


遺伝子変異 機能注釈

...AGCTCCATGCATGTCATGC...
...T...
...T...
...A...

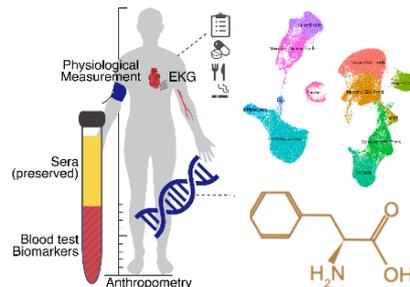
集団ゲノム情報

ACTGAACGCAA	ACT
CCAGTATTCTA	CC
CATGACTGC	CAT
ACAGTATGCT	AC
CCTAAATGCT	CCT
AAAGTCTTCT	AA
ACTATATGCT	ACT



出力情報

形質情報



個人ゲノム

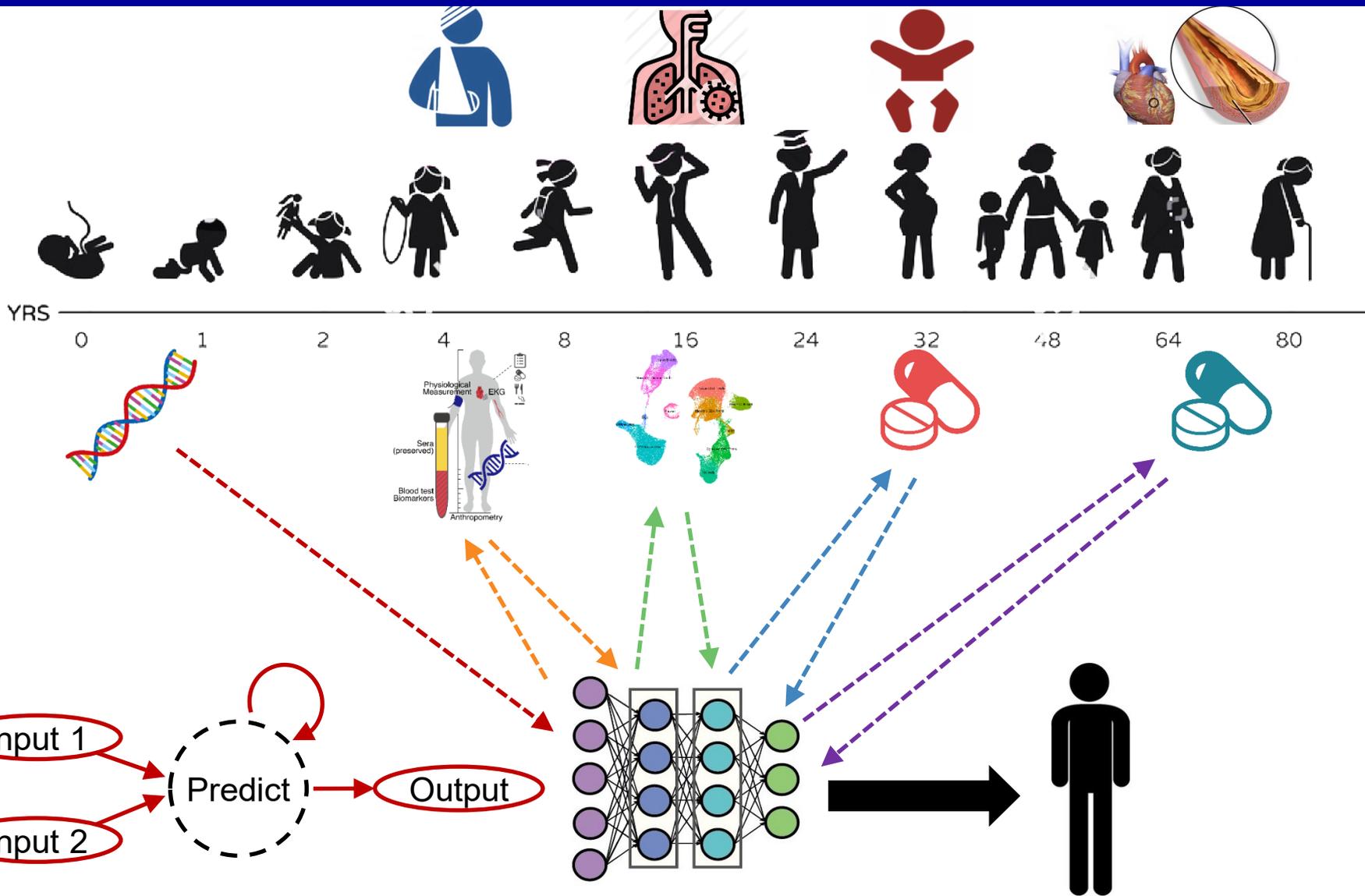
DRB1*01:01
DRB1*01:04
DRB1*04:05
DRB1*09:01
⋮

創薬標的



• 「Uni-modal」から「Multi-modal」機械学習への転換による「ヒト人生」の予測。

④ 機械学習のヒトゲノム情報への応用



•「Uni-directional」から「bi-directional & feedback-looping」機械学習への転換による「ヒト人生」の予測。

遺伝学入門

- ① ヒトゲノム配列とその個人差
- ② ヒトゲノム多型を用いた疾患ゲノム解析
- ③ 疾患病態解明への展望
- ④ 機械学習のヒトゲノム情報への応用
- ⑤ ゲノム創薬への展望

⑤ ゲノム創薬への展望

十年～数十年の歳月

ヒット化合物



第I相臨床試験

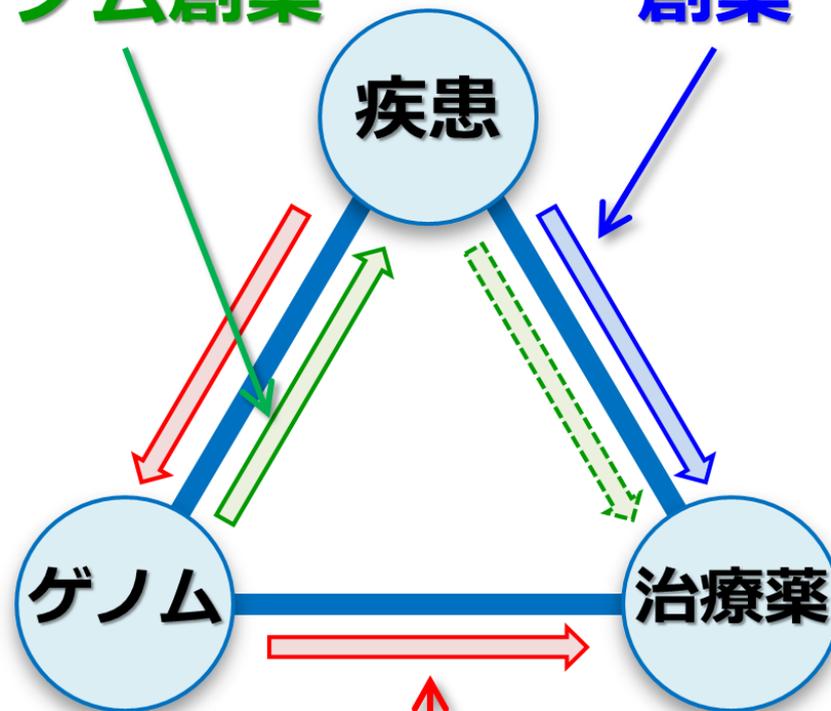


上市



これまでのゲノム創薬

これまでの創薬



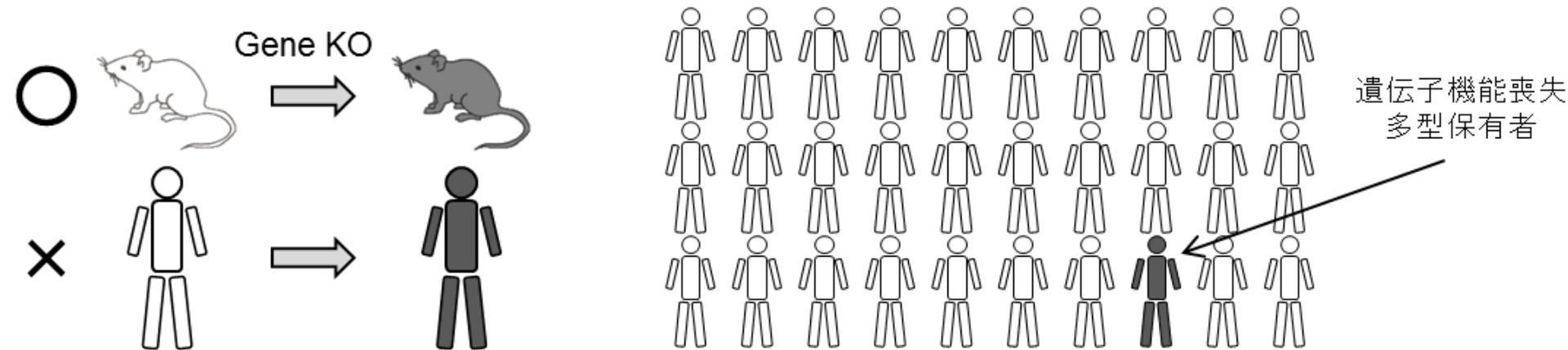
これからのゲノム創薬

⑤ ゲノム創薬への展望

ヒトにおける「バーチャル」な介入実験としての遺伝的多様性の活用

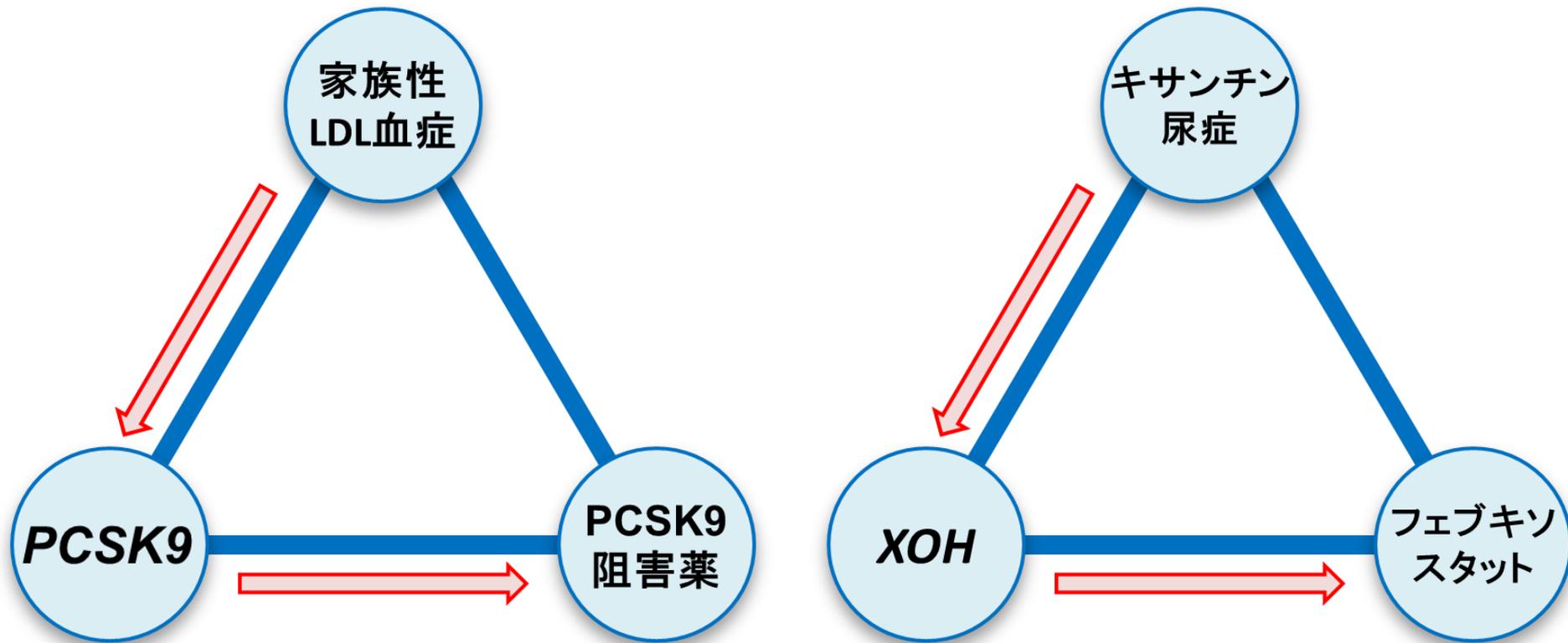
倫理的観点によりヒトにおける
介入実験は不可能

ヒト集団の遺伝的背景の多様性に基づく
「バーチャル」な介入実験結果の観測は可能



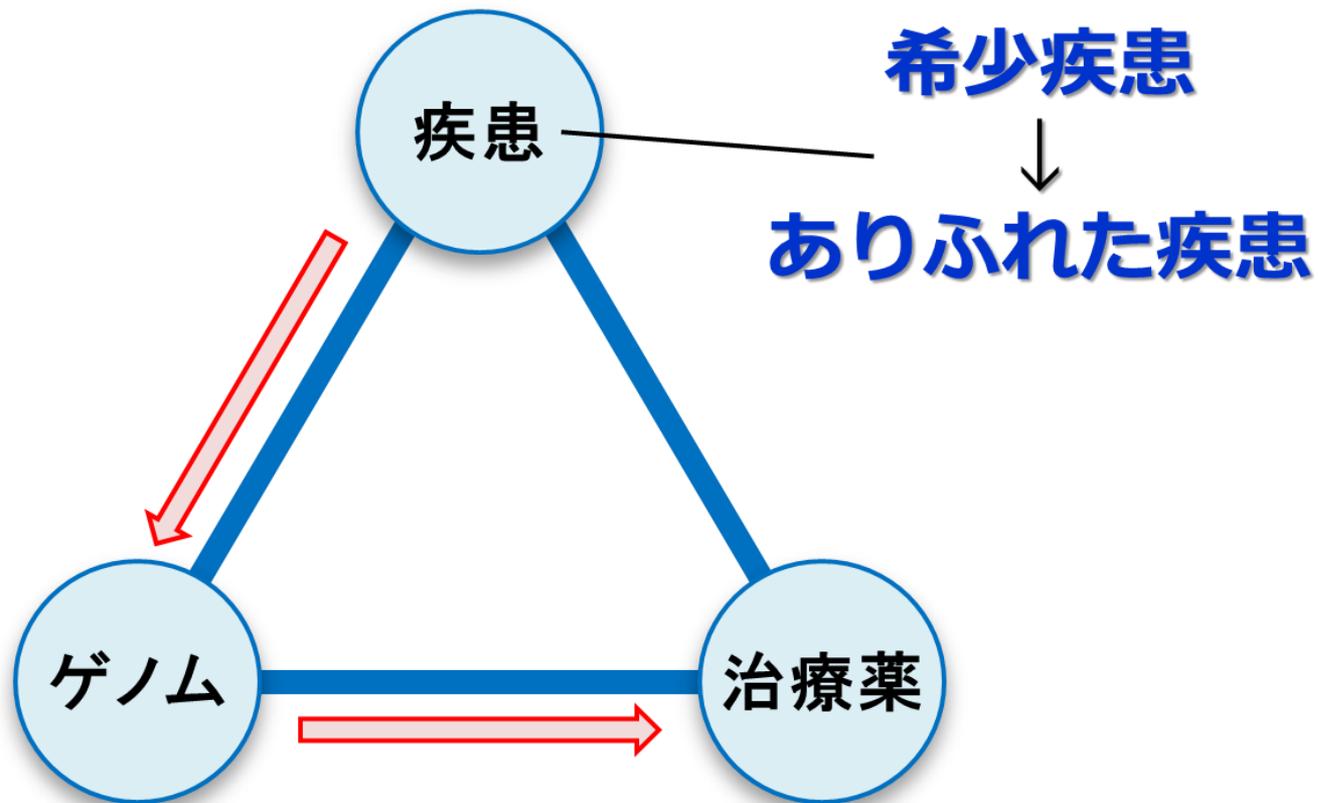
- 創薬の対象はマウスではなくヒト、故に **ヒト疾患サンプル由来の情報を創薬の過程に導入する必要があります。**
- ヒトゲノム解析を通じてヒト集団の遺伝的多様性を扱うことにより、倫理的に不可能な介入実験を「バーチャル」に観測可能に。
- 「**ヒト集団こそがヒトの真のモデル生物**」という意見が提唱されています。

⑤ ゲノム創薬への展望



- ゲノム創薬の成功例として、PCSK9阻害剤(高脂血症)やフェブキシソスタット(痛風)、SGLT2阻害剤(2型糖尿病)、が知られています。
- 遺伝子機能を修飾した際のヒト形質への影響が、ゲノム解析を通じて予め判明していて、そこから阻害薬の開発が促進された、と解釈できます。

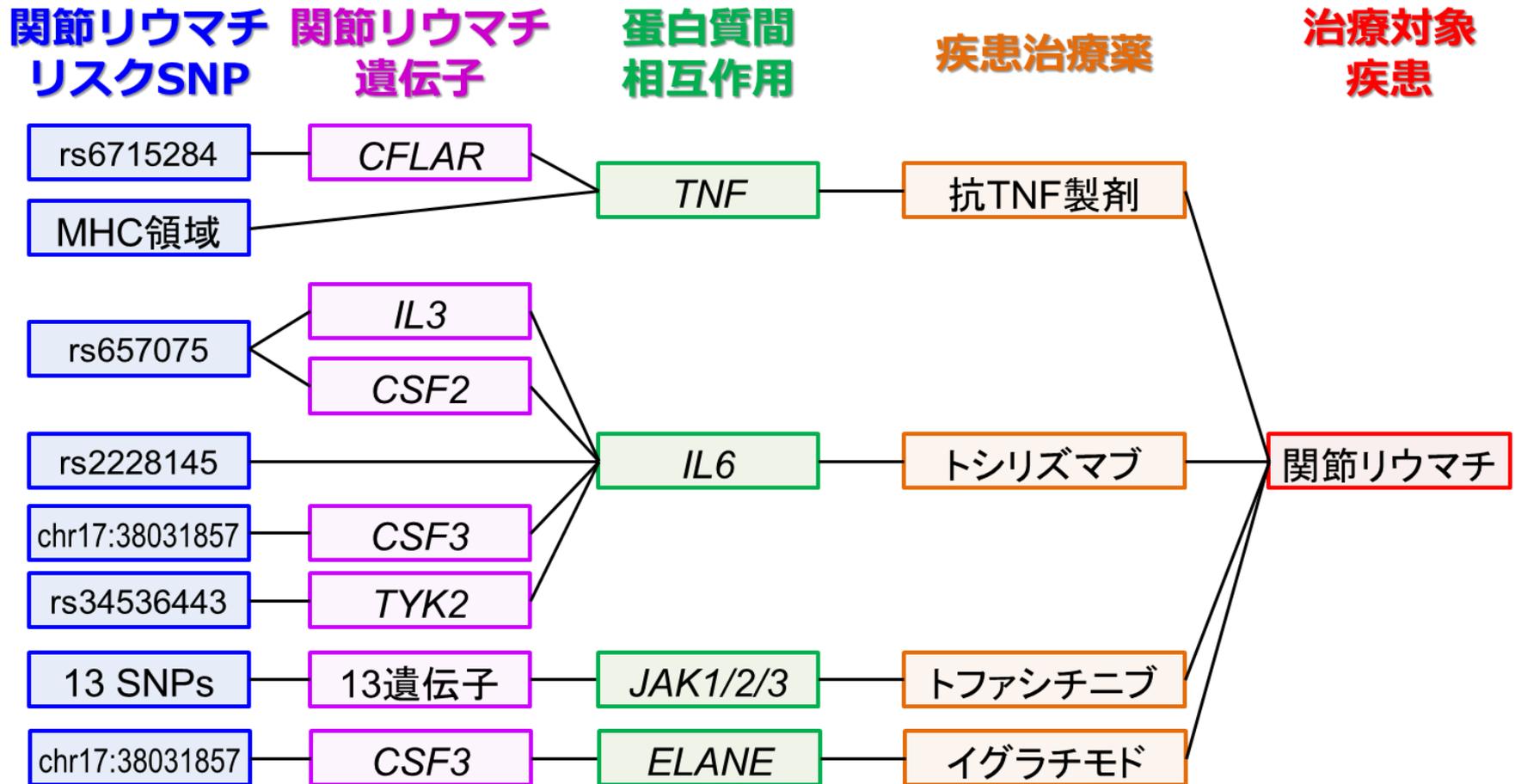
⑤ ゲノム創薬への展望



- メンデル型希少疾患だけではソースとして不十分。
- 同じ枠組みで、より多くの疾患に対するヒトゲノム解析結果を創薬に役立てることができるのでは？
- これが、これからのゲノム創薬の課題と考えられます。

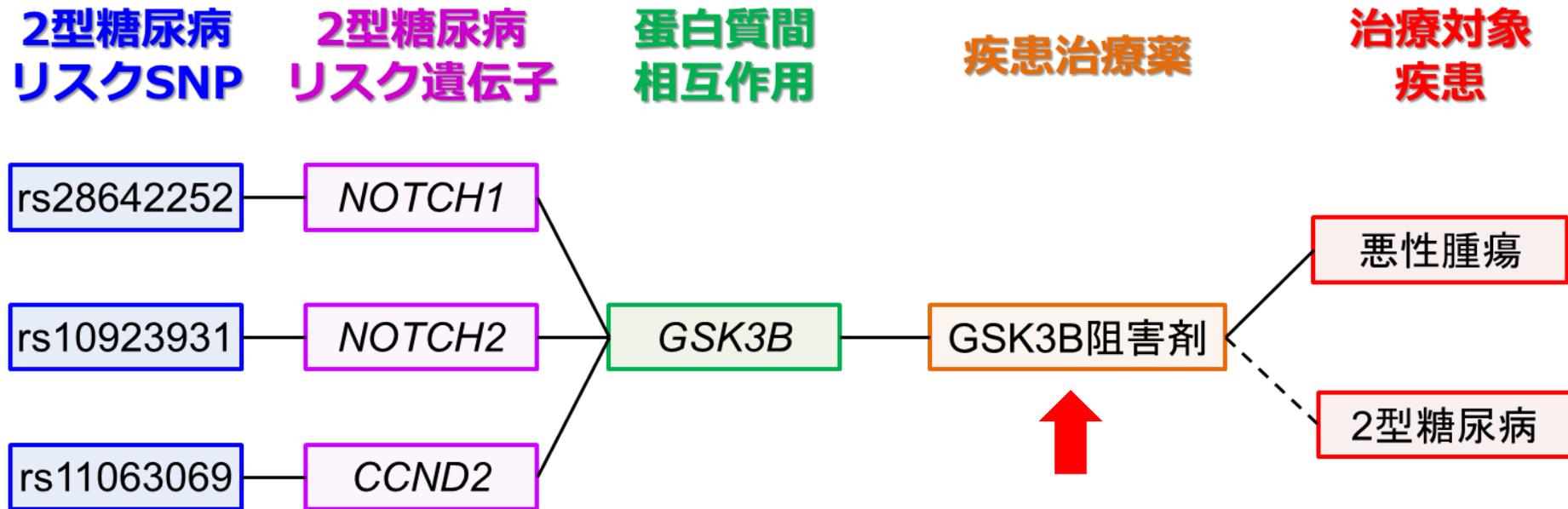
⑤ ゲノム創薬への展望

関節リウマチ感受性遺伝子と治療薬標的遺伝子



• 関節リウマチ感受性遺伝子は、蛋白質相互作用ネットワークを介して、治療薬標的遺伝子とつながっていることが明らかになりました。

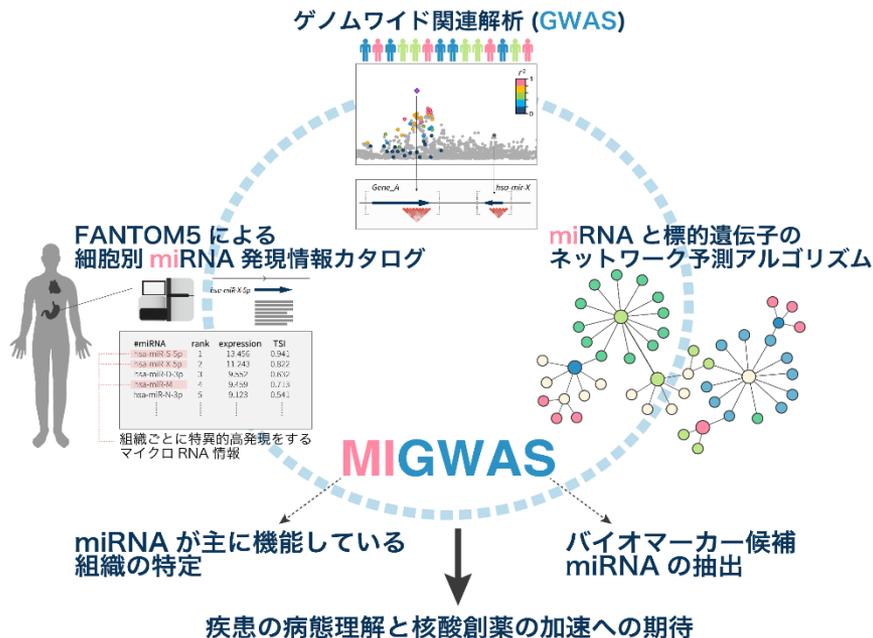
⑤ ゲノム創薬への展望



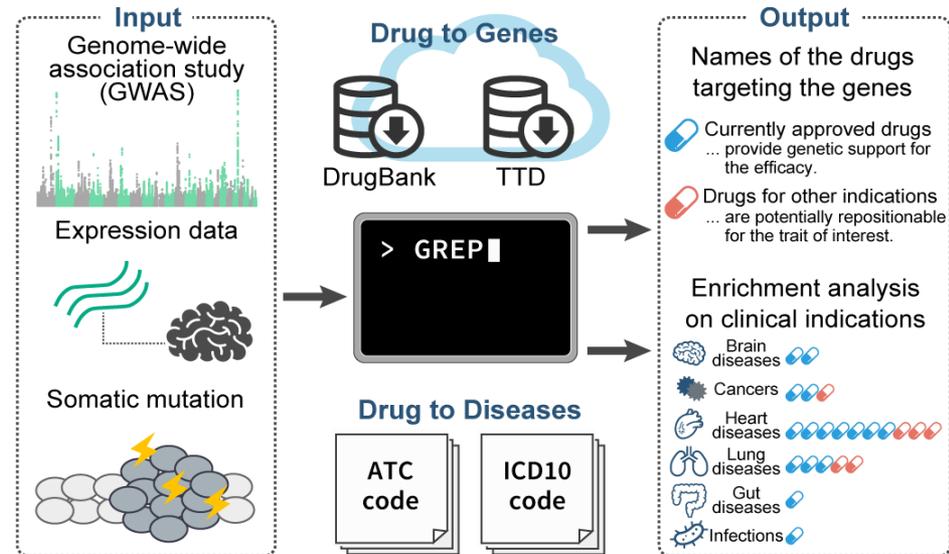
- 既存の治療薬の他の疾患への適用拡大(ドラッグ・リポジショニング)が、新規創薬のコスト改善に貢献すると注目されています。
- 疾患感受性遺伝子を直接標的とした治療薬に注目することで、ドラッグ・リポジショニング候補探索を効率化できることを報告。
- GSK3B阻害剤(2型糖尿病)やCDK4/6阻害剤(関節リウマチ)を同定。

⑤ ゲノム創薬への展望

MIGWAS: miRNA-target gene networks enrichment on GWAS



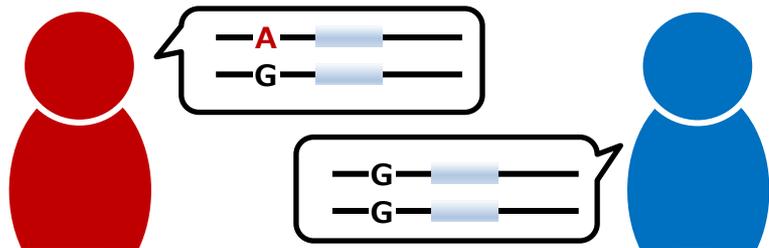
GREP: Genome for REPositioning drugs



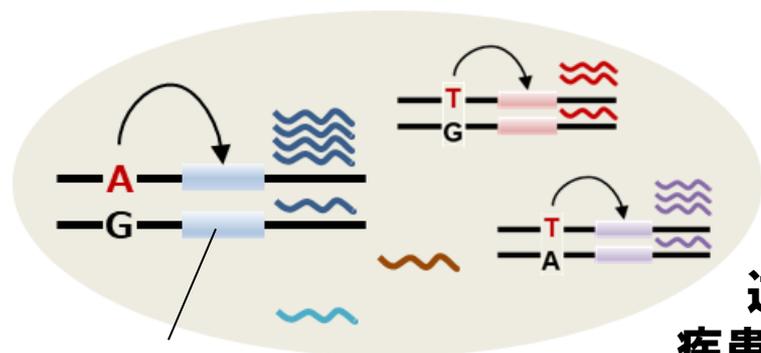
- 疾患感受性遺伝子を直接標的とした治療薬に注目することで、**ドラッグ・リポジショニング**候補探索の効率化が可能。
- **MIGWAS**: 組織特異的miRNA発現情報に基づく核酸ゲノム創薬ソフト
- **GREP**: WHO ATC分類を活用した自動化リポジショニングソフト

⑤ ゲノム創薬への展望

疾患リスクを有する
ゲノム上の変異(遺伝子多型)

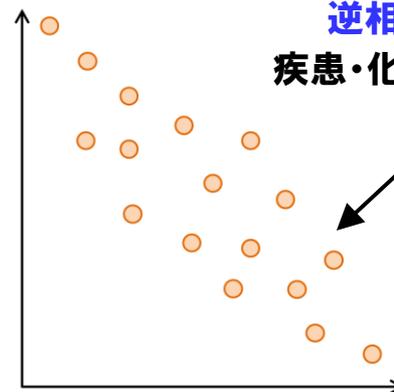


疾患リスク遺伝子多型からケース・
コントロール群間の遺伝子発現量変化を推定
(Transcriptome-Wide Association Study: TWAS)



遺伝子

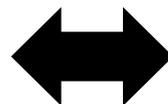
ケース・コントロール群間の
遺伝子発現変化量



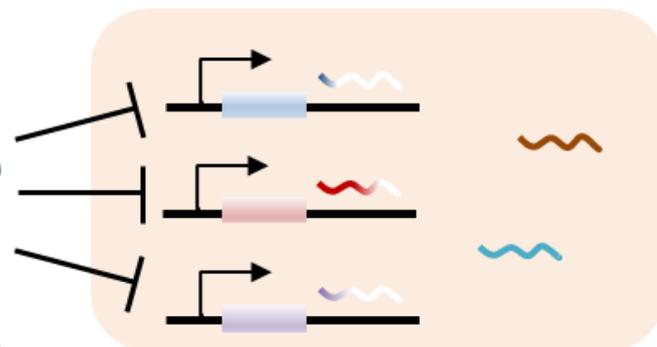
遺伝子発現変化量が
逆相関するような
疾患・化合物ペアを同定

化合物投与時の細胞組織特異的
遺伝子発現変化量

20,000化合物の細胞添加時の
遺伝子発現量変化データベース



逆相関解析で
疾患治療に効果的な
薬剤候補を同定



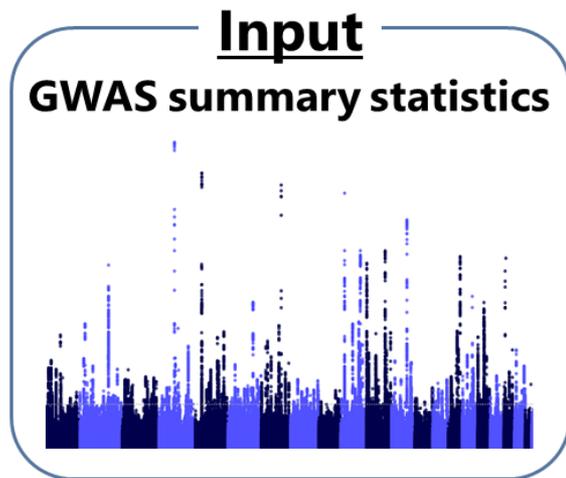
⑤ ゲノム創薬への展望

Trans-Phar: TWASを用いたインシリコ・ドラッグスクリーニング

Trans-Phar

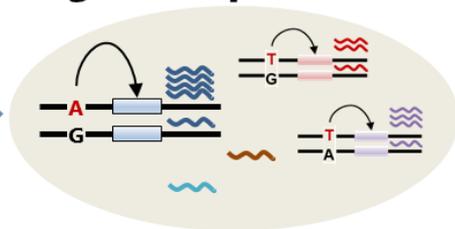
Integration of **Transcriptome**-wide association study and **Pharmacological** database

LINCS
Connectivity map
(L1000)

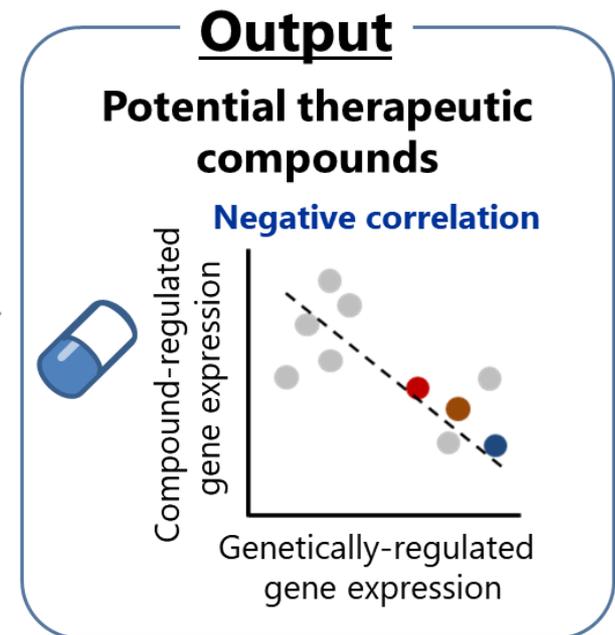
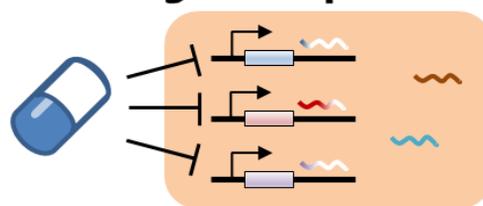


Cell type-specific
TWAS

Genetically-regulated
gene expression



Compound-regulated
gene expression

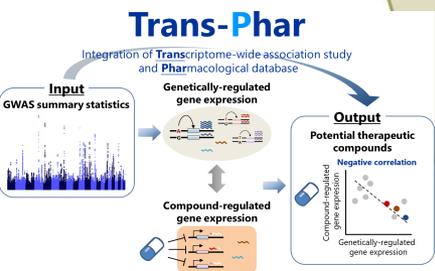
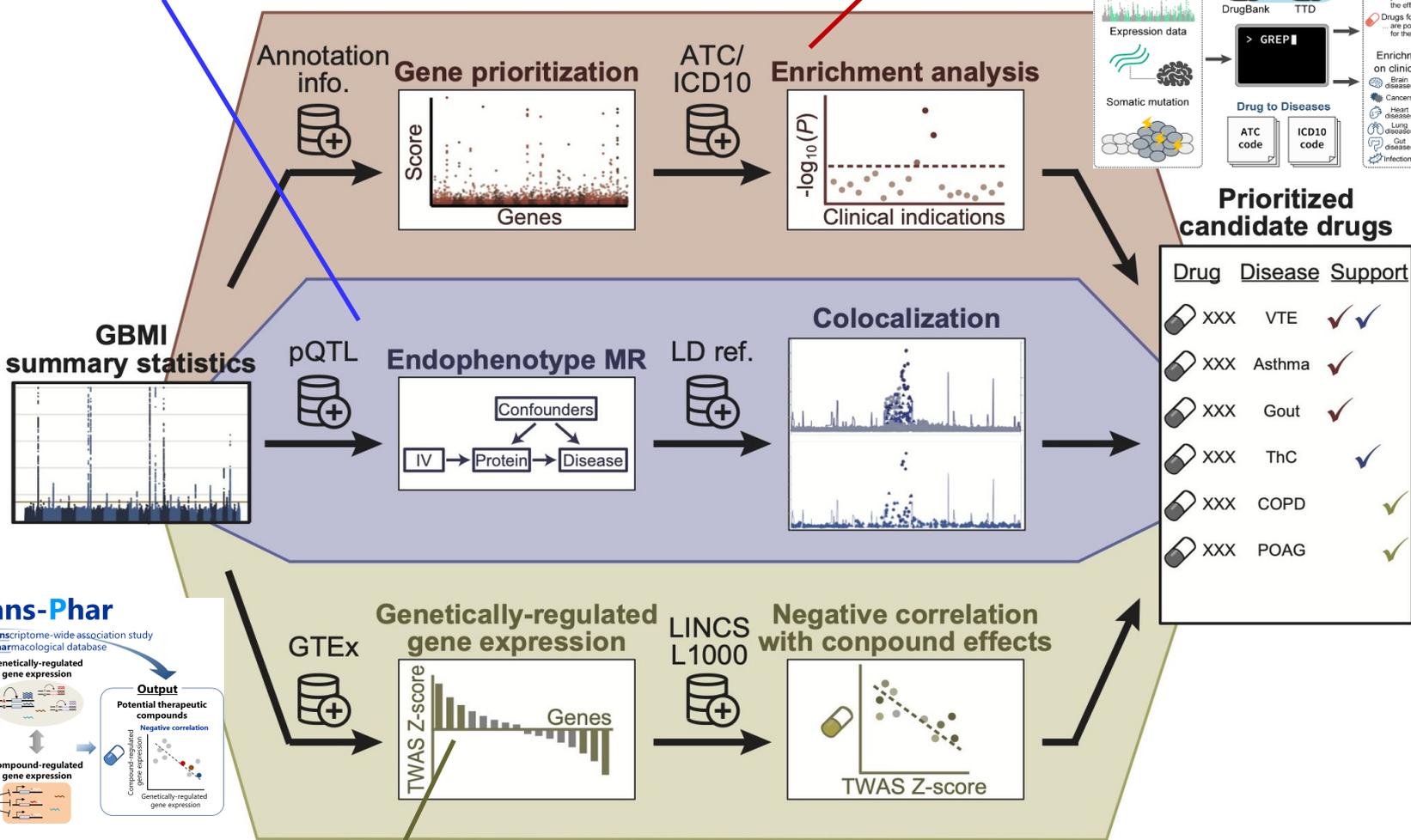
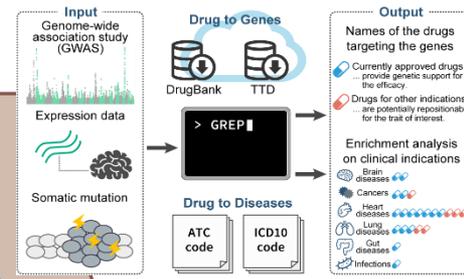


⑤ ゲノム創薬への展望

バイオバンク連携時代のゲノム創薬ベストプラクティスの提唱

中間形質QTL解析 & メンデルランダム化解析

疾患感受性遺伝子-創薬標的遺伝子
オーバーラップ解析



⑤ ゲノム創薬への展望

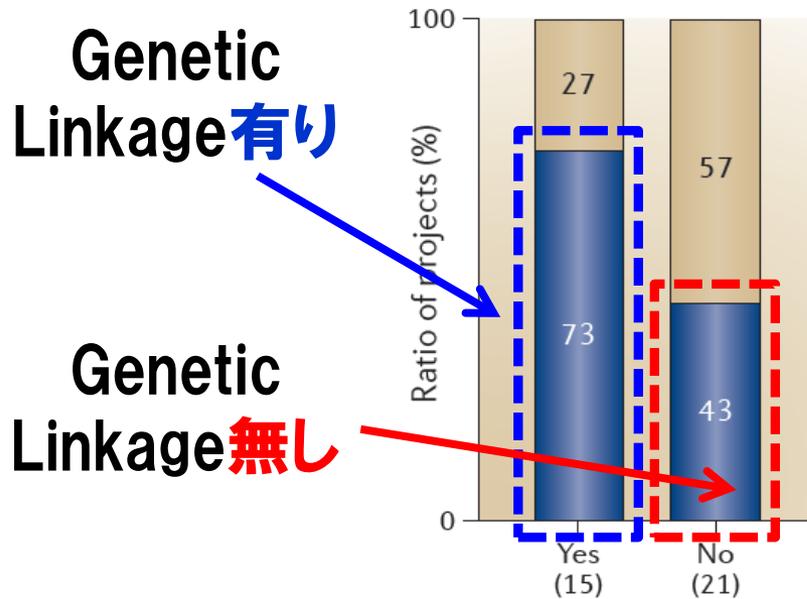
Table 2 | Selected bioinformatic tools that facilitate GWAS-informed drug repurposing

Software	Description	Code or website link	Refs
Open Targets Genetics	Single GWAS variant–gene drug repurposing	https://genetics.opentargets.org/ https://github.com/opentargets	111
Priority index	Single GWAS variant–gene drug repurposing integrated with network-based prioritization using genes from GWAS as seed genes	http://pi.well.ox.ac.uk:3010/	112
GREP	Identifies drug classes enriched for genes prioritized from GWAS or pathways revealed by gene-set association	https://github.com/saorisakaue/GREP	113
Drug Targetor	Ranks drug repurposing candidates based on gene-set association of genes targeted by a compound and the direction of effect from a TWAS of its target genes relative to the drug's mechanism of action	https://drugtargetor.com/index_v1.21.html	114
Trans-Phar	Signature matching using TWAS results and drug perturbagen expression data	https://github.com/konumat/Trans-Phar	115
PS4DR	Signature matching integrating GWAS with disease-associated and drug-induced expression profiles	https://github.com/ps4dr/ps4dr	116
FUSION	TWAS software — results derived can inform drug repurposing opportunities	https://github.com/gusevlab/fusion_twas	46
S-PrediXcan	TWAS software — results derived can inform drug repurposing opportunities	https://github.com/hakyimlab/PrediXcan	45
MAGMA	Gene-based and gene-set association — results derived can inform drug repurposing opportunities	https://ctg.cncr.nl/software/magma	76
TwoSampleMR	Mendelian randomization methods — results derived can inform drug repurposing opportunities	https://github.com/MRCIEU/TwoSampleMR	132

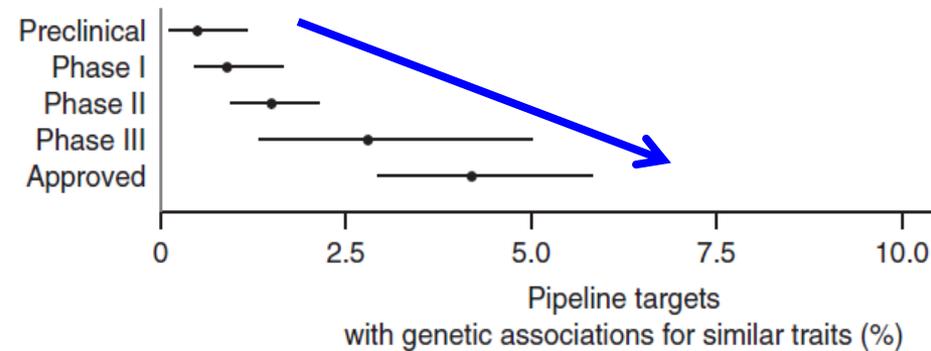
・疾患ゲノム情報から新規創薬標的・化合物スクリーニングを可能にするゲノム創薬ソフトウェアの開発が進んでいます。

⑤ ゲノム創薬への展望

「Genetic Linkage」 (by AstraZeneca)



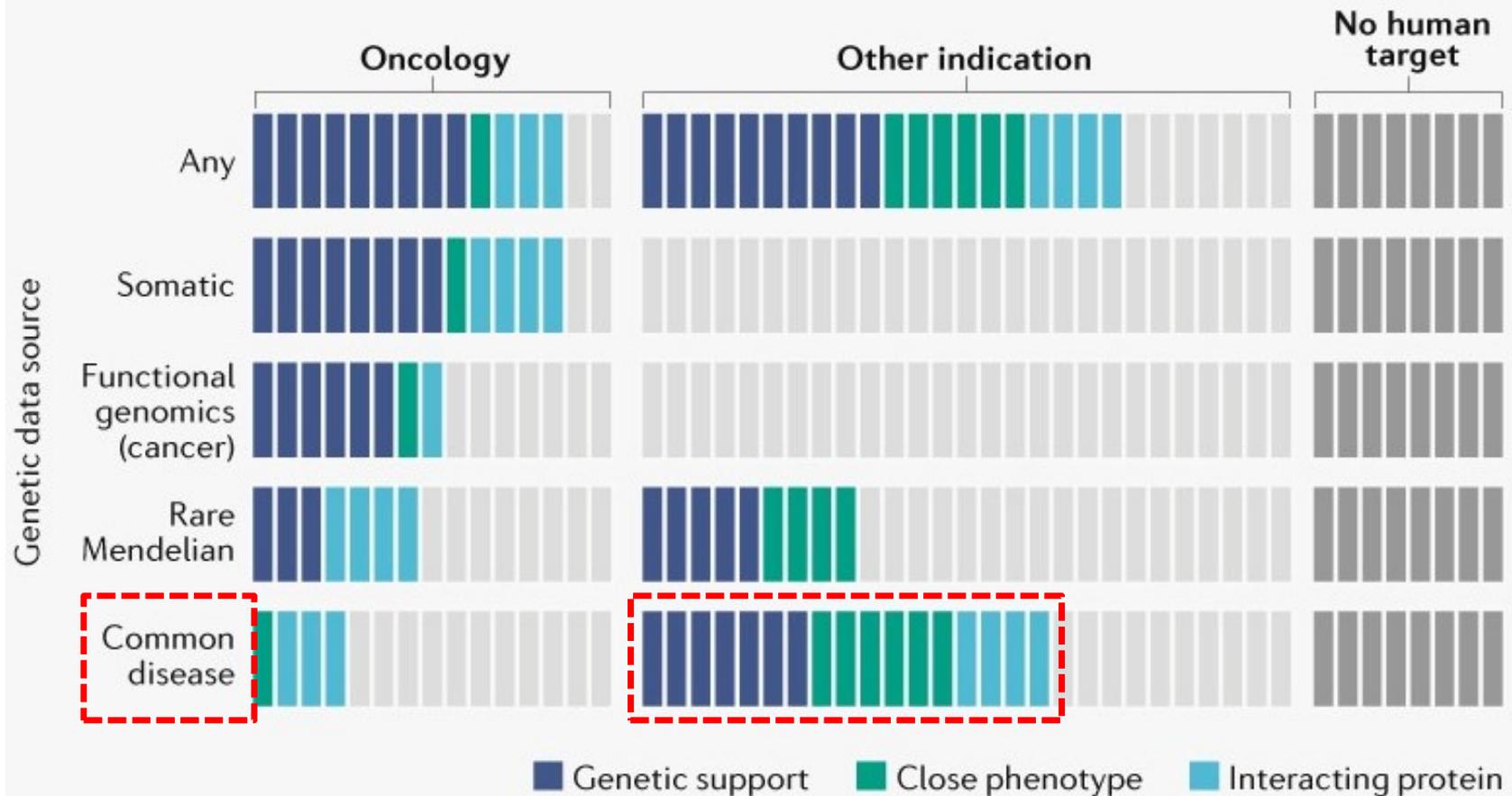
「Genetic Evidence」 (by GlaxoSmithKline)



創薬ステージが進むにつれ
Genetic Evidenceが増す

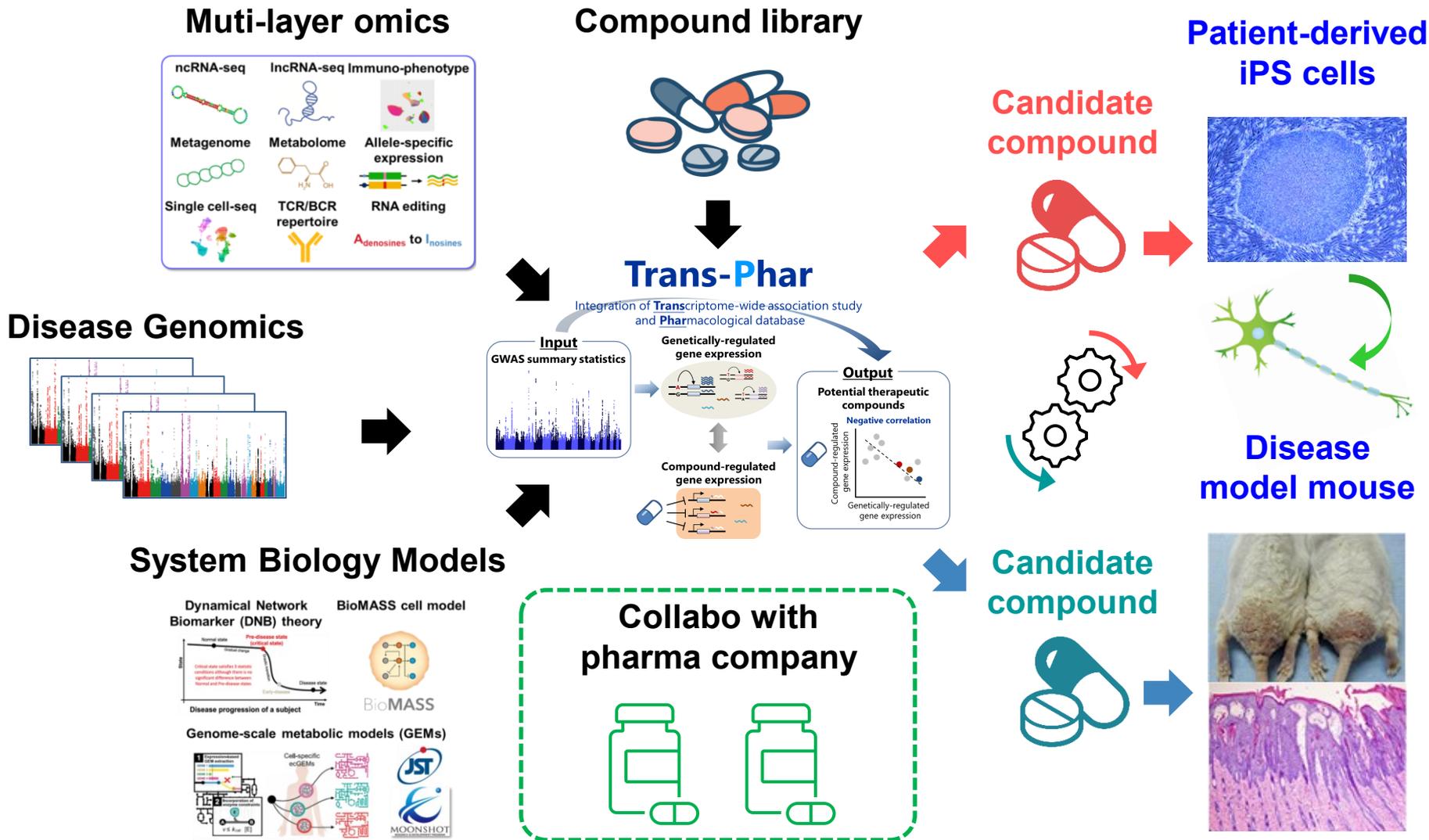
- 実際に、海外の製薬企業により、疾患ゲノム情報が創薬プロセスを効率化することが報告されています(AstraZeneca社、GlaxoSmithKline社)。
- 長く、成功率の低い創薬開発の枠組みに、どうやって疾患ゲノム情報を組み込んでいくか、に注目が集まっています。

⑤ ゲノム創薬への展望



- 2021年に米国食品医薬品局(FDA)に承認された50の治療薬のうち、**2/3の治療薬は疾患ゲノム情報の根拠がある**ものでした。
- ありふれた疾患(common disease)のゲノム解析に基づく治療薬の割合が高く、ゲノムワイド関連解析の活用の重要性が示唆されました。

⑤ ゲノム創薬への展望



• これからのゲノム創薬は、疾患オミクス情報に基づくdry研究と、iPS細胞(+遺伝情報)・モデル生物等を用いたwet研究の融合と期待されます。

⑤ ゲノム創薬への展望



疾患発症リスクの大きい遺伝子ほど、創薬標的として有望ですか？



リスクの大きさやP値の小ささは、対象集団で、該当遺伝子のどの箇所に、どの程度の頻度のSNPが存在するか、に依存するので、一概には言えません。

- ・リスクの大きさやP値の小ささは、対象集団で、遺伝子のどの箇所に、どんな頻度のSNPが存在するか、に依存します。
- ・有望な創薬標的遺伝子でも、該当集団で、近傍にある程度の頻度の機能的SNPが存在しなければ、関連解析で同定することはできません。
- ・多彩な人種でゲノム解析を実施することが、有望な創薬標的の効率的な同定につながると期待されます。

⑤ ゲノム創薬への展望

海外の製薬会社における取り組み

- 海外の製薬会社では、**ゲノム創薬への注目/投資**が高まっています。
- **優秀なゲノム研究者をアカデミアから引き抜き**、ゲノム創薬研究だけでなく、**治験対象薬物の決定プロセスにも関与**させています。
- 逆に、製薬企業で活躍したゲノム研究者が、**アカデミアに引き抜かれていく例も多くあり**、アカデミア⇔企業の距離が近いといえます。
- **臨床試験実施時にゲノムを収集し**、データベース化しています。
- 国際的なゲノム研究プロジェクトに、**自社のゲノムデータを用いた参画**を行い、最新の解析結果や、解析方法について**情報収集**を行っています。
- **バイオバンクに出資し**、**大規模ゲノム・表現型データへのアクセス**を確保する動きも活発化しています。

⑤ ゲノム創薬への展望

フィンランドバイオバンク(FinnGen)パートナー企業



(<https://www.finngen.fi/en/>)

- フィンランドのバイオバンク事業：FinnGenが注目を集めています。
- 運営費用として製薬企業から出資を募り、**出資企業を”Partners”**と位置付け、一般公開に先立ちゲノム情報・解析結果へのアクセスを提供。
- 「**いずれは公開される情報にいち早くアクセスする目的**」で、多額の出資を行うことが、これらの企業の目的に合致していた、という解釈も。

⑤ ゲノム創薬への展望

UKバイオバンク45万人全エクソーム解析論文

Article

Exome sequencing and analysis of 454,787 UK Biobank participants

Joshua D. Backman¹, Alexander H. Li¹, Anthony Marcketta¹, Dylan Sun¹, Joelle Mbatchou¹, Michael D. Kessler¹, Christian Benner¹, Daren Liu¹, Adam E. Locke¹, Suganthi Balasubramanian¹, Ashish Yadav¹, Nilanjana Banerjee¹, Christopher E. Gillies¹,

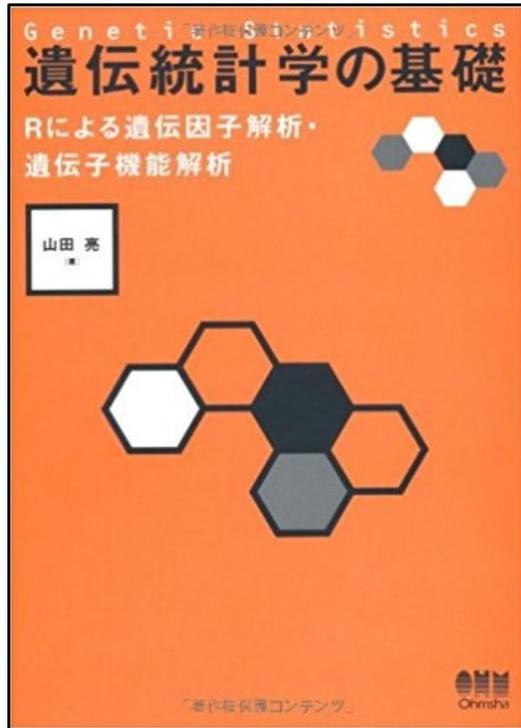
1. Regeneron Genetics Center

- 各国のバイオバンクサンプルのシーケンス解析を格安で受注し、**データアクセス権を製薬企業向けに売るビジネスモデル**、も登場しました。
- UKバイオバンクの全エクソーム解析論文を執筆したのは、ゲノムベンチャー企業Regeneronでした。

終わりに

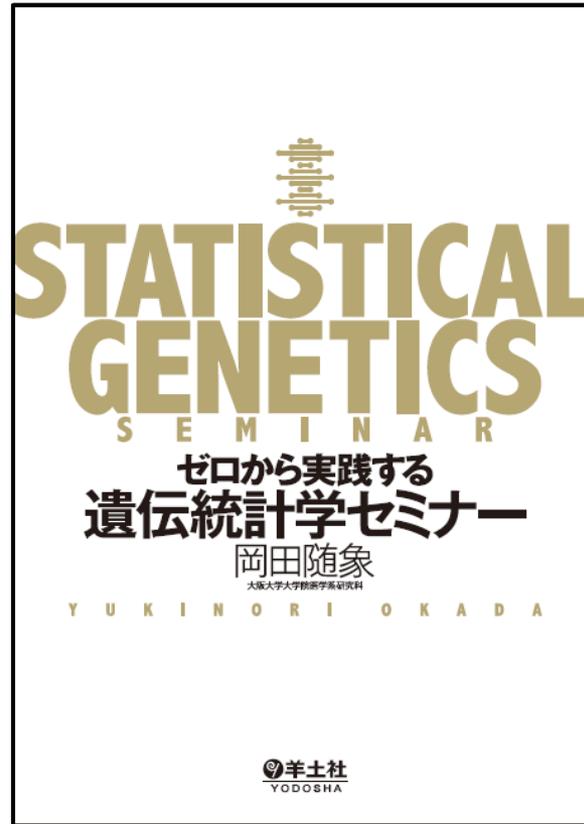
- ヒトゲノム研究や、遺伝統計学の現状を、簡単になぞってみました。
- ヒトゲノムデータ解読コストの低下に伴い、もの凄い勢いでゲノムデータが作成され、アカデミアだけでなく社会に蓄積されていっています。
- 遺伝統計解析により、ヒト疾患ゲノム研究がゲノム創薬に貢献できる可能性が示され、今後の更なる発展が期待されています。
- ヒトゲノム研究の分野は、今後5年、10年、20年と、想像もつかない速度で発展していくものと思われます。
- ヒトゲノムと直接関わらない研究分野においても、ゲノム研究の趨勢やリソースとしての価値を理解することが、重要になると考えられます。

終わりに



- 「遺伝統計学のお薦めの教科書はありませんか？」と、よく訊かれます。
- 国内の教科書としては、上記3冊がお薦めです。
- ゲノムデータ解析手法は、進歩が早いので、最新の知見を得るには、論文や総説に目を通す必要があります。

終わりに



amazon.com
販売サイト
QRコード



- 「遺伝統計学・夏の学校」の講義演習内容が、羊土社さんから書籍化されました。
- 当研究室のWebサイトで公開している講義演習内容とほぼ同一の内容になりますが、書店で見かけたら是非手に取ってみて下さい。

終わりに

次代を担う若手研究者の人材育成



- 生命科学の幅広い研究分野で、情報解析技術が必要な時代に。
- 大阪大学・東京大学医学部の授業で、ゲノムデータ解析演習を開始。