

生命科学とオープンデータ

東京大学 大学院理学系研究科 生物科学専攻

科学技術振興機構(JST)
バイオサイエンスデータベースセンター(NBDC)

情報・システム研究機構(ROIS)
国立遺伝学研究所(NIG)
DDBJセンター(DDBJ)

高木利久

本日のお話

- 生命科学におけるデータ共有
– 意義、歴史、現状、課題–
- 統合データベースプロジェクト
– データ共有からオープンサイエンスへ
- 個人ゲノムデータの活用状況（海外）
- これからの展開と課題

生命科学におけるデータの共有 -意義、歴史、現状、課題-

生命科学における研究データ共有の歴史

- 文献データ(書誌情報、アブストラクト)は1960年代より
- 研究データは1970年代より
- DNAデータは日米欧3極体制で
- タンパク質立体構造データは日米欧4センターで
- 他の種類のデータにも共有の枠組みが拡大
- 分野別目的別単位でのデータ共有の枠組みも多々
 - HUPO、ADNI、GA4GH、など
- ゲノムデータの共有なしに生命科学は成立しないまでに

生命科学ではデータ共有がなぜ活発なのか？

- 少数の数式や法則で表現できない
- DBは研究のインフラでありフロンティア
- ファンディング機関からの義務付け
- 論文投稿時における出版社からの義務付け
- 研究コミュニティにおける合意
 - ヒトゲノムの公開に関するバミューダ原則
- 受け皿としてDBセンターの整備（欧米）
- データ共有による研究の促進
 - 統計解析のパワーアップ
 - 他の観点からのデータ利用、イノベーション促進
- 研究成果の再現性や検証、研究不正への対応
- データ収集の重複の排除、研究（資金）の効率化

公開の意義と公開すべきエビデンスデータの範囲

- 資金配分機関: 研究の成果を最大限活用するため
 - NIH: 資金援助した研究の結果および成果を、研究者コミュニティや一般社会が最大限利用できるようにする
 - Wellcome Trust: 資金援助した研究から出る成果の価値を最大化するため
- 出版社: 発表論文に関するエビデンスの確保
 - PLOS系列: 論文作成に関係する全データを制限なしで公開
 - Nature系列: 論文Submit時にはEditorsやReviewersが見られるように
 - Science系列: 論文中的数据を作成したデータを自由に見ることができるように

Natureの規定

Mandates for specific datasets

For the following types of data set, submission to a community-endorsed, public repository is mandatory. Accession numbers must be provided in the paper. Examples of appropriate public repositories are listed below.

Mandatory deposition	Suitable repositories
Protein sequences	Uniprot
DNA and RNA sequences	Genbank
	DNA DataBank of Japan (DDBJ)
	EMBL Nucleotide Sequence Database (ENA)
DNA and RNA sequencing data	NCBI Trace Archive
	NCBI Sequence Read Archive (SRA)
Genetic polymorphisms	dbSNP
	dbVar
	European Variation Archive (EVA)
Linked genotype and phenotype data	dbGAP
	The European Genome-phenome Archive (EGA)
Macromolecular structure	Worldwide Protein Data Bank (wwPDB)
	Biological Magnetic Resonance Data Bank (BMRB)
	Electron Microscopy Data Bank (EMDB)
Microarray data (must be MIAME compliant)	Gene Expression Omnibus (GEO)
	ArrayExpress
Crystallographic data for small molecules	Cambridge Structural Database

<http://www.nature.com/authors/policies/availability.html>

生命科学にかかわるデータの実態(1)

• データベースの数

- 世界全体:10,000から20,000
- メジャーなもの(NAR誌のDB特集収録):約1,600
- 我が国のDB数(NBDC Integbioカタログ):約1,100

• データの種類

- NAR誌での分類:15カテゴリ、40サブカテゴリ
- Integbioカタログ:生物種、対象、データ種類で分類
 - 生物種:動物、植物、原生生物、菌類、真正細菌、ウイルス
 - 対象:ゲノム、遺伝子、cDNA、多型、タンパク質、酵素、細胞
 - データ種類:配列、構造、発現、相互作用、画像、オントロジー

• ゲノムプロジェクトの数(GOLD DB):約77,000

• データDB開発国(NAR誌):約50

生命科学にかかわるデータの実態(2)

• データの量、サイズ

- 米国NCBI: SRA 3PB, dbGaP 2PB
- dbGaPには 100万人分のゲノムデータ
- 10年先には個人のゲノムデータも 10^6 から 10^9 へ
- DDBJでは年率 1.5倍程度の伸び(ムーアの法則と同程度)
- 文献(PuBMed)は2,500万件(PMCに370万のフルテキスト)

• 主要なDBセンター: NCBI, EBI, DDBJ、等

- 欧米のDBセンターは数百人規模
- 10～30 PB程度のストレージ保有
- 2020年にはゲノムだけで2EB程度必要との試算

データの利活用に関する障害

- 生命科学は研究目的、研究対象、データの種類、解釈が非常に多様
- 専門外のDBを使う必要性あり(ゲノムは生物横断的)
- DBや解析ツールの数が多すぎて使い方分からない
 - 生体内相互作用DBだけでも500以上のDB
- 注釈が信頼性のあるものとなないものが混在
- フォーマットや用語、概念がDB毎にバラバラ
 - 遺伝子の概念がDBにより違う、同じ遺伝子に多数の名前
- データの文脈依存性、曖昧性、冗長性、複雑性、誤差
- レポジトリするだけでは不十分、再利用性低い

我が国の事情

- 資金配分機関からの共有の義務化ルールなし
- プロジェクト終了すると維持管理更新されない
- データの囲い込み、データの権利関係不明
- 小規模データ多い→ビッグデータ化必要
- バイオインフォマティクス不足→競争に負ける
- 受け皿となる中核DBセンターがない（欧米は数百人規模のセンター）

我が国における統合データベースプロジェクト
データ共有からオープンサイエンスへ

我が国の生命科学DB統合プロジェクト

- データの共有、公共財化を促進し、その価値を最大化
- 内閣府CSTP主導の統合データベースプロジェクト(2006～)
 - 文科省、経産省、農水省、厚労省で実施
 - 2011年12月に四省連携のポータルサイト
- 文科省の統合データベースプロジェクト(2006～)
 - 中核センターの設立
 - 2007～ROIS ライフサイエンス統合DBセンターDBCLS
 - 2011～JST バイオサイエンスDBセンターNBDC
 - クリエイティブコモンズ(CC)ライセンスによるデータの共有
 - フォーマット、辞書、統合技術、動画教材などの開発
 - カタログ、横断検索、アーカイブの構築など種々のサービス提供
 - 研究分野ごとのデータベース統合化進行中(ファンディングによる)
 - ヒト由来データの共有・セキュリティガイドラインの作成
 - ヒトDB(オープン、制限アクセス)の構築、受入れ(DDBJと連携して)

RDFによる
すべてのDBの統合

連邦型統合DB

DB統合化基盤技術の開発

目的別統合DB

微生物DB

植物DB

ヒトDB

データバンク的DB

ゲノム

タンパク質立体構造

メタボローム

糖鎖

DBをまとめることで
使いやすく

統合化推進プログラム

個々のDBを
使いやすく

カタログ

横断検索

アーカイブ

標準化DB3
標準化DB4b

データベース

DB1a

DB1b

DB2

DB3

DB4a

DB4b

プロジェクト

P1

P2

P3

P4

測定装置

顕微鏡

MRI

chip

NGS

質量分析器

放射光施設

NBDC統合化推進プログラム

生物種メタボロームモデル・データベースの構築 有田 正規 理研 環境資源科学研究センター

ゲノムとフェノタイプ・疾患・医薬品の統合データベース 金久 實 京大化研

ゲノム・メタゲノム情報統合による微生物DBの超高度化推進 黒川 顕 東工大地球生命研究所

疾患ヒトゲノム変異の生物学的機能注釈を目指した多階層オミクスデータの統合 菅野 純夫 東大新領域

植物ゲノム情報活用のための統合研究基盤の構築 田畑 哲之 かずさDNA研究所

個別化医療に向けたヒトゲノムバリエーションデータベース 徳永 勝士 東大医

蛋白質構造データバンクの高度化と統合的運用 中村 春木 阪大蛋白研

糖鎖統合データベースおよび国際糖鎖構造リポジトリの開発 成松 久 産総研糖鎖センター

生命と環境のフェノーム統合データベース 梶屋 啓志 理研バイオリソースセンター

プロテオーム統合データベースの構築 石濱 泰 京大薬

生命動態情報と細胞・発生画像情報の統合データベース 大浪 修一 理研生命システムセンター

ヒト脳疾患画像データベース統合化研究 岩坪 威 東大医

大規模ゲノム疫学研究の統合情報基盤の構築 松田 文彦 京大医 附属ゲノム医学センター

公募要領にデータ提供協力依頼記載

- 文科省ライフ課委託プロジェクト(H20～)
- JST戦略事業(CREST、さがけ)(H23～)
- 厚労科研費(H24～)
- 文科省科研費(H25～)
- AMED-CREST, PRIME(H27～)
- 医療分野研究成果展開事業
産学連携医療イノベーション創出プログラム(H27～)
- ナショナルバイオリソースプロジェクト
「ゲノム情報等整備プログラム」(H27～)

6.8 バイオサイエンスデータベースセンターへの協力

ライフサイエンス分野の本事業実施者は、論文発表等で公表された成果に関わる生データの複製物、又は構築した公開用データベースの複製物を、バイオサイエンスデータベースセンター(※)に提供くださるようご協力をお願いします。提供された複製物は、非独占的に複製・改変その他必要な形で利用できるものとします。複製物の提供を受けた機関の求めに応じ、複製物を利用するに当たって必要となる情報の提供にもご協力をお願いすることがあります。

NBDCは、日本の生命科学研究を推進するために、データベースをつなげて使い易くします。
 そのためにNBDCや協力機関は、以下のようなサービスやウェブサイトを作成・提供しています。

生命科学全体のデータベース統合

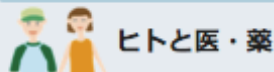
[Integbioデータベースカタログ](#)

[データベース横断検索](#) 

[生命科学系データベースアーカイブ](#)

[NBDC RDFポータル](#) **NEW !!**

分野ごとのデータベース統合



ヒトと医・薬

[NBDCヒトデータベース](#)

[ヒトゲノムバリエーションデータベース](#)

[ヒト疾患脳画像データベース](#)

[KEGG MEDICUS: 疾患・医薬品統合リソース](#)



生命を支える分子

[DDBJ: 日本DNAデータバンク](#)

[PDBj: 日本蛋白質構造データバンク](#)

[TogoProt: 蛋白質関連データベース統合検索](#)

[JCGGDB: 日本糖鎖科学統合データベース](#)

日本語や動画でわかりやすく

[新着論文レビュー / 領域融合レビュー](#)

[統合TV](#)

論文をもっと読みやすく、書きやすく

[Allie](#) / [inMeXes](#) / [TogoDoc](#)

大量の配列データを扱いやすく

[DBCLS SRA](#)

[RefEx](#) / [統合遺伝子検索 GGRNA](#)

さまざまな統合コンテンツ

[生物アイコン](#)

[生命科学系主要プロジェクト一覧](#)

[Webリソースポータルサイト](#)

[ゲノム解析ツールリンク集](#)

[HOWDY](#) / [GenLibi](#)

開発ツール

統合データベース講習会: AJACS
受入れ機関募集! 詳しくはこちら!
 募集期間: 2016年1月12日(火)~2月17日(水)



NBDCパンフレット
 (PDF: 3.22MB/
 2015/06/30更新)

新着情報



2016/02/05

[平成27年度NGSハンズオン講習会実施報告書等を公開しました。](#)

2016/02/05

[【横断検索】12件のデータベースが検索できるようになりました。](#)

2016/02/02

[【生命科学系データベースアーカイブ】「イネプロテオームデータベース」\(農研機構 作物研究所 小松節子上席研究員\)を追加しました。](#)

2016/02/02

[【NBDCヒトデータベース】オーダーメイド医療の実現プログラムからのアトピーGWAS結果をオープンデータとして公開しました\(hum0014.v](#)

4省連携によるDB統合

<http://integbio.jp/>

文科省

JSTバイオサイエンス
データベースセンター

医薬基盤・健康・栄養研究所

厚労省

The screenshot shows the homepage of the **integbio.jp** portal. The header includes the logo and the title "生命科学系データベース統合のための合同ポータルサイト". Navigation tabs include "ホーム", "統合へのステップ", "技術交流", and "イベント". The main content area is titled "4省の生命科学系データベースの統合を目指して" and contains introductory text about the project's goals and progress. A sidebar on the right offers search options: "データベースを探す (カタログ)", "データを探す (横断検索)", and "データを使う (アーカイブ)".

図1. 4省データベース統合の段階的進展

以下は、各省で統合を進めている4つの機関それぞれのポータルサイトです。

- 文科科学省・バイオサイエンスデータベースセンター(NBDC)ポータルサイト**
- 農水省・農業生物資源研究所**
- 経産省・産総研創薬分子プロファイリング研究センター**
- 厚労省・医薬基盤・健康・栄養研究所**

農水省

農業生物資源研究所

産総研創薬分子プロファイリング
研究センター

経産省

NBDCヒトデータベース
DDBJ JGA

(Japanese Genotype-phenotype Archive)

NBDCヒトデータベース／データの種類

NBDCヒトデータベース

オープンデータ

ウェブサイト等から制限なく公開

- ・集団の統計値
- ・特定の個人由来では無い試料の解析結果

制限公開データ

(標準レベル[Type I]セキュリティ)
(ハイレベル[Type II]セキュリティ)

ヒトデータ審査委員会
(NBDC)での審査に基づき利用可能

- ・個人ごとの情報

公開待機データ

一定期間の後、制限公開データ等へ移動

匿名化

匿名化前・公開留保データ他

各プロジェクト・実施機関

ヒトデータベース基本方針

1. 運用原則

1. 『NBDCヒトデータベース』は以下の原則に基づいて運用される。

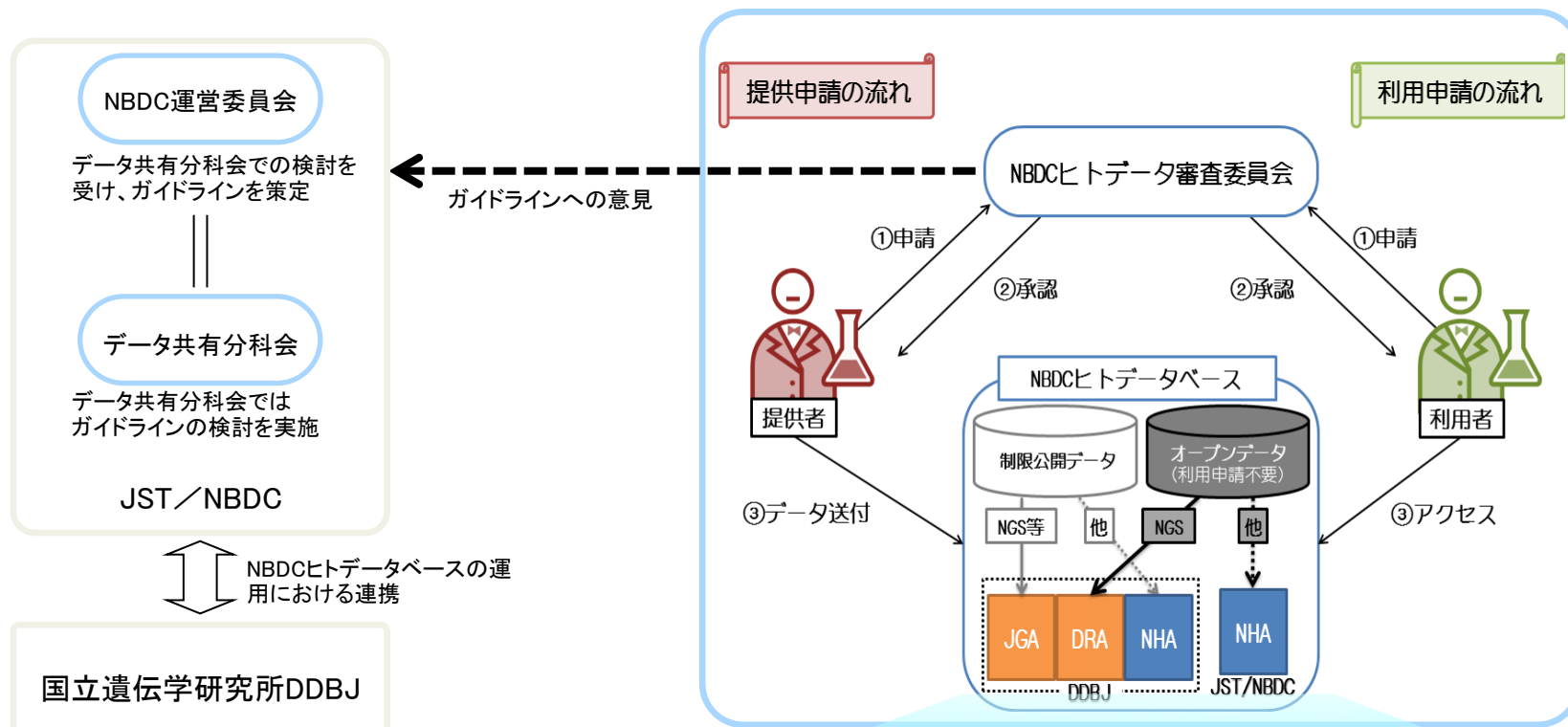
- 原則1 公的資金により産生されたヒトに関するデータをなるべく広く収集すること
- 原則2 収集したデータをなるべく広く共有できるようにすること
- 原則3 試料提供者の個人同定等につながらないよう、データの適正な管理に努めること

2. NBDCは『NBDCヒトデータベース』の運用において以下の項目を実施する。

- i. ガイドラインの整備および必要に応じた見直し
- ii. データ提供およびデータ利用申請についての審査
- iii. ウェブサイトの整備等データへのアクセス手段の維持

- インフォームドコンセントで禁止してない限り民間企業も利用可能
- 無料でデータ提供可、データ利用可
 - 今後大規模データを受入れる際はデータ提供側に課金の可能性も
 - 米国NCBIで現在検討中、欧州ではそのような動きない
- データ公開時期は提供者の意向を基本的に尊重
 - 最長でも論文出版までが通常
 - 今後NIHのガイドラインに準拠して変更する可能性も

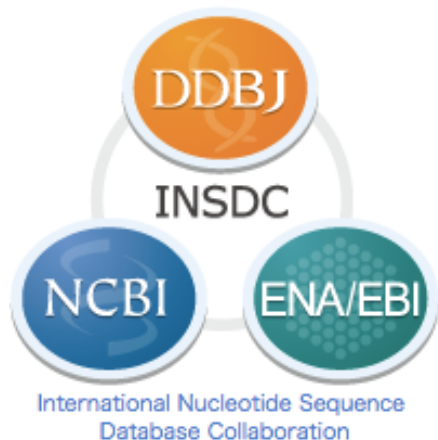
ヒトデータベース運営体制図



データベース・格納データ種類・運営者等の関係

データベース名称	JGA (制限公開)	DRA (オープン)	NHA (制限公開)	NHA (オープン)
データ種類	次世代シーケンサーデータ 発現アレイデータ 遺伝子型データ 等	次世代シーケンサーデータ	配列情報等に付随する健診・検診データ 調査票データ 画像データ 等	左記の集計・統計データ等

※DRA (DDBJ Sequence Read Archive) 、JGA (Japanese Genotype-phenotype Archive) 、NHA (NBDC Human Data Archive)



NBDCヒトデータベース／セキュリティルール

データの種類によって実施すべきセキュリティ対策を共通化

データの種類		データ提供者	データベースセンター	データ利用者
NBDCヒトデータベース	オープン	提供申請が必要	データ改ざん防止などの基本的対策も実施	自由に利用できる (ルール不要)
	制限公開 (標準レベル[Type I] セキュリティ)	Type I レベルセキュリティ		
	制限公開 (ハイレベル[Type II] セキュリティ)	Type II レベルセキュリティ		
	公開待機	Type II と同レベルのセキュリティを適用		
	匿名化前・公開留保			利用できない

NBDCヒトデータベースの現状

利用可能な研究データ一覧

データ利用方法は[こちら](#)をご覧ください。

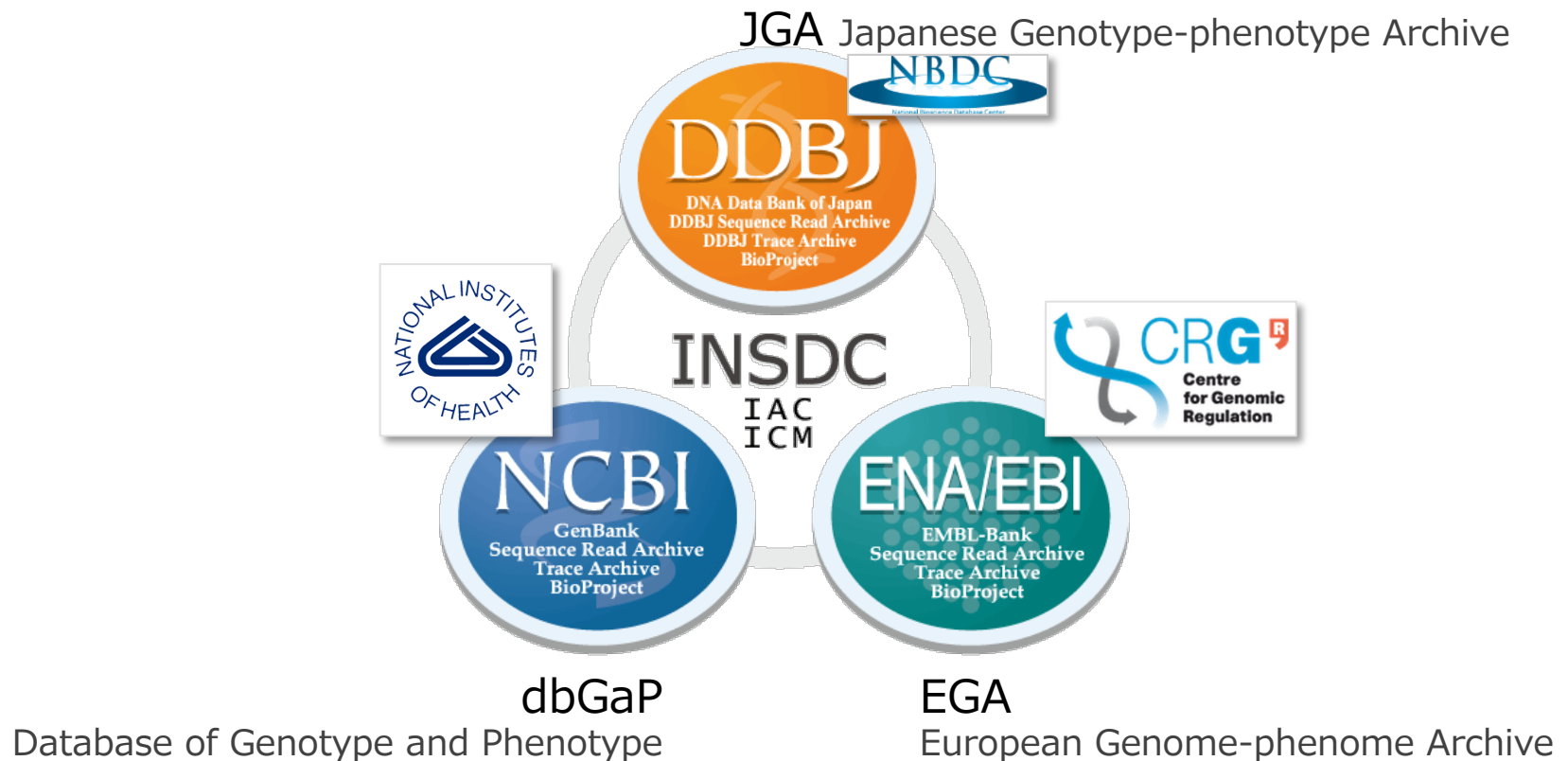
<http://humandbs.biosciencedbc.jp/>

Research ID	研究題目	公開日	データの種類	研究方法	手法	参加者 (対象集団)	提供者	アクセス制限
hum0001.v1	SCA31罹患患者のゲノム解析データ	v1:2013/12/01	NGS (Whole)	配列決定	Illumina (HiSeq2000)	1検体 (日本人)	森下真一	制限(Type I)
hum0003.v1	関節リウマチ患者及び健康人における HLA領域の塩基配列比較解析	v1:2013/07/01	NGS (Target Capture)	HLA領域 配列決定	Illumina (MiSeq)	33検体 (セルライン)	井ノ上逸朗	オープン
hum0004.v1	上皮成長因子受容体遺伝子異変を有する 肺腺癌の体細胞性遺伝子変異プロファイル 解明のための分子疫学的研究	v1:2014/07/11	NGS (Exome)	配列決定	Illumina (Genome Analyzer IIX)	97症例 (日本人)	後藤功一	制限(Type I)
hum0005.v1	難聴の遺伝性解析と臨床応用に関する研究	v1:2013/12/27	NGS (Target Capture)	9遺伝子領域 配列決定	Life technologies (Ion PGM)	17検体	宇佐美真一	オープン
hum0006.v1 (JGAS000000000004)	脳腫瘍のゲノム・遺伝子解析と その臨床病理学的意義の解明	v1:2014/01/31	NGS (Exome)	配列決定	Illumina (HiSeq 2000)	6症例23検体 (日本人)	斎藤延人	制限(Type I)

公開データ31件(制限なし9件)
公開待機データ・データ登録中23件
合計 延べにして4万人程度

大規模な国のプロジェクトと連携
・東北メディカル・メガバンク機構
・次世代がん研究
・オーダーメイド医療プログラム

アクセス制限データベース



- ✓ JGA と EGA は SRA をベースにしたデータモデルを使用
- ✓ dbGaP と EGA は概要情報を交換 (JGA も参画予定)

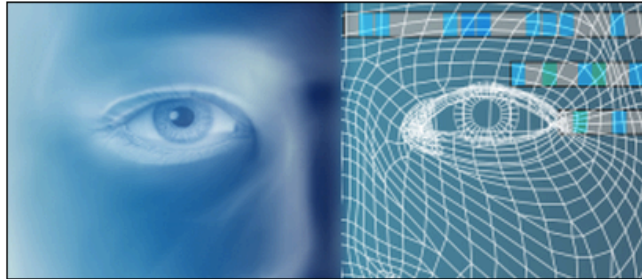
欧米の個人ゲノムデータベースの (米国dbGaP、欧州EGA) 活用状況

dbGaP@米国NIH NLM NCBI

NCBI Resources ▾ How To ▾ Sign in to NCBI

dbGaP dbGaP ▾ Search

[Limits](#) [Advanced](#)



dbGaP

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype.

Getting Started

[dbGaP Tutorial](#)

[Overview](#)

[FAQ](#)

[How to Submit](#)

[Browse Top Level Studies](#)

Access dbGaP Data

[Collections](#)

[Apply for Controlled Access Data](#)

[Public Data via ftp Download](#)

[Association Results Browser](#)

[Phenotype-Genotype Integrator](#)

Important Links

[Summary Statistics](#)

[dbGaP RSS Feed](#) 

[Code of Conduct](#)

[Security Procedures](#)

[Contact Us](#)

dbGaP provides two levels of access - [open](#) and [controlled](#) - in order to allow broad release of non-sensitive data, while providing oversight and investigator accountability for sensitive data sets involving personal health information. Summaries of studies and the contents of measured variables as well as original study document text are generally available to the public, while access to individual-level data including phenotypic data tables and genotypes require varying levels of authorization.

<http://www.ncbi.nlm.nih.gov/gap>

NIH Genomic Data Sharing Policy

U.S. Department of Health & Human Services

www.hhs.gov

www.nih.gov

NIH Genomic Data Sharing (GDS)

Google™ Custom Search Search

Home

Policy

Policy Oversight

Researchers

Institutions & IRBs

Data Repositories

FAQs

Related Resources

Subscribe to the GDS LISTSERV

Policy

NIH Genomic re... provides op... To promote appropriate... Genomic D... <http://grant...> <https://fede...>

NIH-funded investigators are expected to submit **human genomic data** from studies that fall within the scope of these Policies to the NIH database of Genotypes and Phenotypes (**dbGaP**) or other relevant NIH-designated data repositories such as Gene Expression Omnibus (**GEO**), Sequence Read Archive (**SRA**), or **the Cancer Genomics Hub**.

ヒトデータは 品質管理が済み次第dbGaPに登録。
非公開期間は最大6か月。エンバーゴは設定できない。

<https://gds.nih.gov/>

dbGaPの登録と利用

登録Study数636

年間利用申請 >1万件

データ利用の論文数 >920

データサイズ 2PB

dbSNP/dbVar など他の NCBI
リソースと有機的に連携

EGAの登録と利用

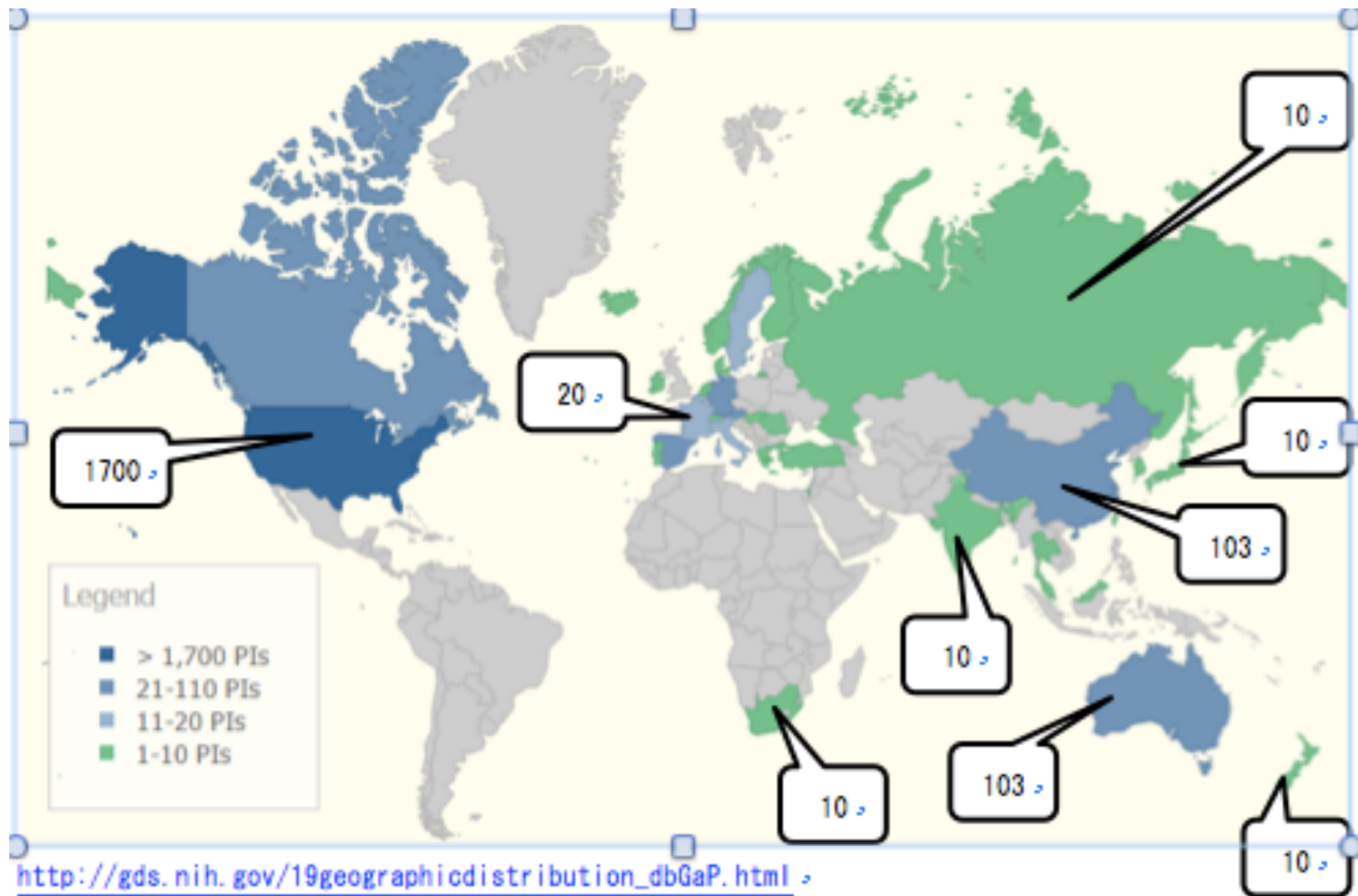
登録Study数830

年間利用申請 >5000件

データサイズ 1.6PB

EBI変異DB、UK10Kなどと連携

dbGaPの利用の地理的分布



これからの展開と課題

国際的なゲノムデータの共有の枠組み GA4GH

Global Alliance for Genomics and Health (GA4GH) とは

- 様々な機関が集結し、参加者の自主性を尊重しながらも、ゲノムデータと臨床情報を最大限に統合・活用し、ヒトの健康増進を図る。
- 最大限に活用するために、ゲノムデータと臨床情報の信頼のおける、自発的かつ安全な共有を可能にする国際的に共通の枠組みを創る。
- 共通の枠組みとは、倫理・法律上の規約、データ共有のためのソフトウェアやデータフォーマット、情報セキュリティ規約などを指す。

GA4GH 加入組織

- 387組織(38カ国) ※2016年2月現在
 - NIH, ELIXIR, Google, Amazon, Illuminaなど、研究機関に限らず、医療機関、IT企業等も参加。
- 日本からは11組織がメンバーに
 - ライフサイエンス統合データベースセンター (DBCLS)
 - エーザイ株式会社
 - Genomedia株式会社
 - 日本医療政策機構(HGPI)
 - 科学技術振興機構 バイオサイエンスデータベースセンター (NBDC)
 - 日本人類遺伝学会(JSHG)
 - 国立がん研究センター(NCC)
 - 国立遺伝学研究所 DDBJセンター
 - 大阪大学大学院 医学系研究科・医学部
 - 理化学研究所
 - 株式会社理研ジェネシス

GA4GH 運営体制

- 運営委員会の下に4つのワーキンググループを設置

1. Clinical Working Group

- Phenotypeデータの統一フォーマットの開発(オントロジー)やゲノムデータとのリンク付け方法の確立を目的としている。

2. Data Working Group

- データ形式、クラウド環境における安全な保管、ゲノム情報を共有するためのアプリケーション・プログラミング・インターフェース(API)の開発、データを使いやすくするためのアプリケーション開発といった、技術開発を実施している。

3. Regulatory and Ethics Working Group

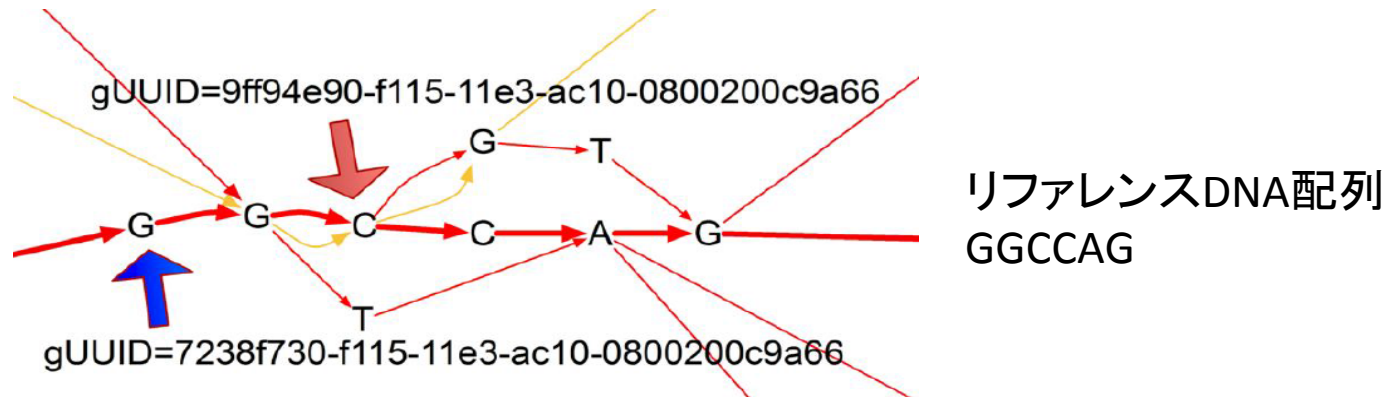
- 国際ガイドラインや倫理的な枠組みを作成し、ゲノムデータ・臨床情報の信頼のおける共有を世界規模で活性化させることを目的としている。

4. Security Working Group

- データセキュリティ、アクセス制御、監査機能、プライバシー保護について検討している。

Reference Graph

Data Working GroupのReference Variation Task Teamでの活動



<https://genomicsandhealth.org/files/public/6-Beacon-HausslerGA4GHLeiden.pdf>

- DNA配列をグラフ(各塩基を節、隣接する塩基を枝で接続)で表現
- リファレンスDNA配列を1本の経路で表現
- リファレンスに対する変異をリファレンス配列の経路から分岐した経路で表現

多様性を持つゲノム配列の集合をグラフで表現することにより、ゲノム配列の既知のあらゆる変異を表現でき、既存の文字列表現での不完全性、矛盾を解消することを目指す。

実証プロジェクト



Beacon Project

遺伝情報を国際的に共有するオープンウェブサービス。分散しているゲノムデータを検索しやすくすることを目的としており、現時点では、指定した条件を満たすデータを含むデータベースを示す(2015/6現在、252 Datasetsが検索対象,)。



BRCA Challenge

乳ガンやその他のガンの遺伝要因の理解を深めるために、世界中からガンに関与する遺伝子多型データを共有するための試み。まずは乳ガンのデータ共有を進めている。

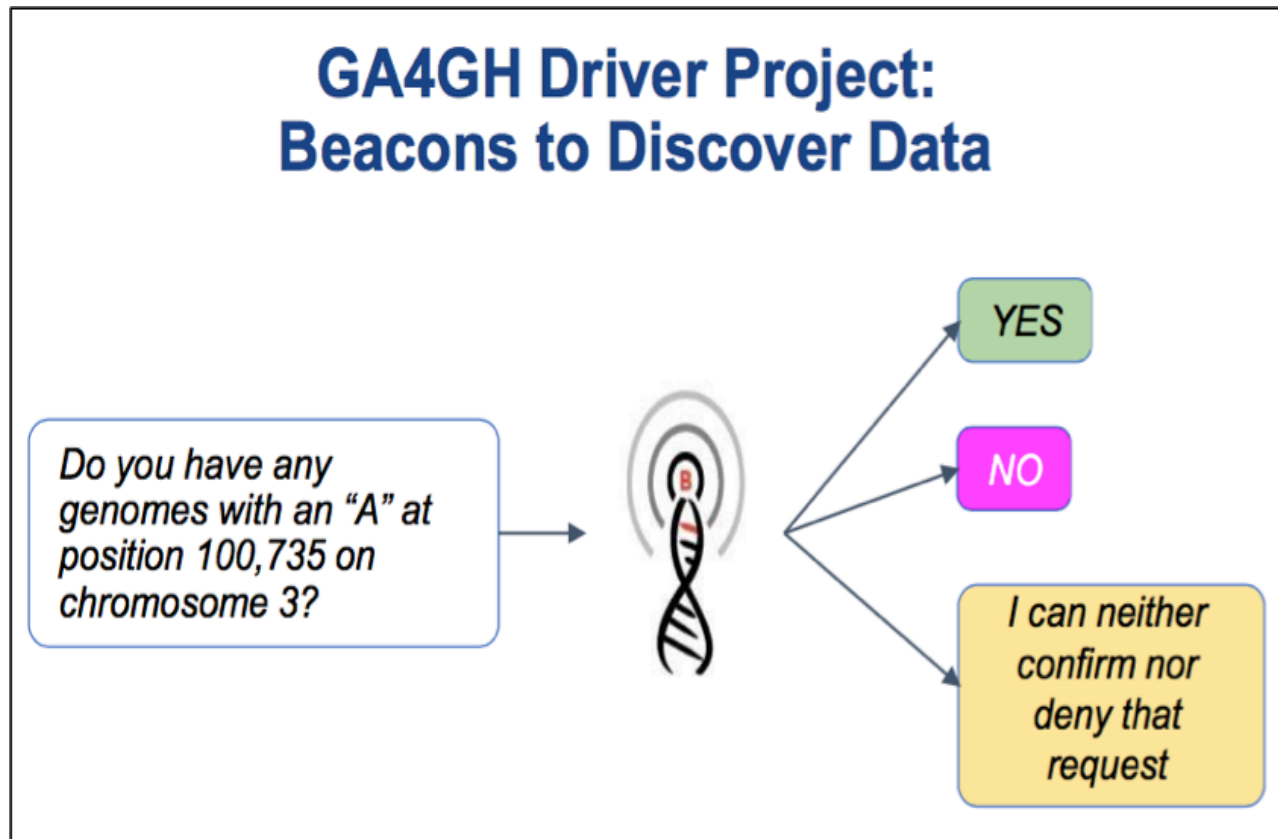


Matchmaker Exchange

類似の表現型情報や遺伝子型情報を共有することで、希少疾患や診断未確定疾病の理解を深めるためのデータベース連邦型ネットワークシステム。

Beacon

検索対象DBが条件(ゲノム上の特定の位置の塩基が、指定した塩基か否か)を満たすゲノムデータ(頻度だけでなく個人ゲノム)を持っているかをyes/noで返す。分散しているゲノムデータを検索しやすくする。

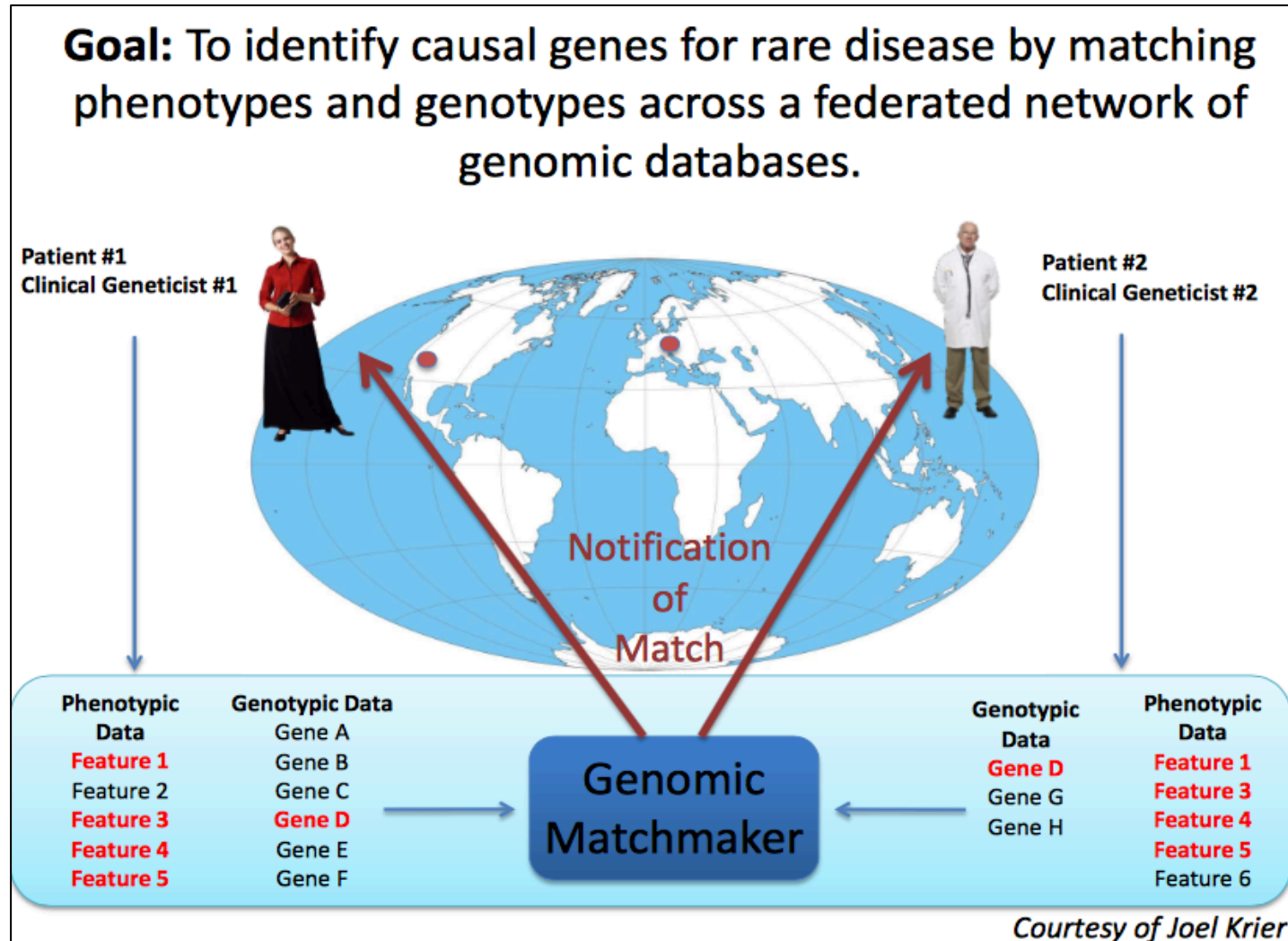


<https://genomicsandhealth.org/files/public/6-Beacon-HausslerGA4GHLeiden.pdf>

プログラムをダウンロードでき、誰でもBeaconを公開できる。

Matchmaker Exchange

分散するDBから類似のphenotype/genotypeを持つ希少疾患を探す仕組み



これからの課題

- RDFによる分野を超えたデータ統合とエンドユーザ向けのインタフェース構築
- オープンサイエンスへの対応
 - データ共有ポリシー、インセンティブ、評価の仕組み
- データ産出プロジェクト開始時からの支援、連携
 - 研究申請時にデータ管理プランの提出
- イノベーションの促進支援
 - 例:ゲノム医療プラットフォーム事業
 - 例:CREST 植物頑健性
- 個人情報保護法改正への対応
 - 個人識別符号、要配慮個人情報
- スパコンの活用、クラウドの活用
- データ量、データ種類の増大への対応、持続可能な体制構築
 - データ共有のためのコスト負担モデルの構築
- データベースセンターの統合
 - 我が国には4つ、5つ程度の小規模なセンターや拠点あり

RDFとは？

- RDF: Resource Description Framework
- 主語 (S)・述語 (P)・目的語 (O) の三つ組で表現



- リテラル以外はURIs (URLs)を使うことでリソースを一意に同定可能
- オントロジーを利用することで推論が可能に
- W3Cによって標準化されている
- 標準化された検索用言語 SPARQL
- 分散検索が可能
- 三つ組をつなげていくことで、Linked Open Data networkとしてさまざまなデータがつながる

NBDC RDFポータルサイト

NBDC
RDF
Portal

The NBDC RDF Portal provides a collection of life science datasets in RDF (Resource Description Framework). The portal aims to accelerate integrative utilization of the heterogeneous datasets deposited by various research institutions and groups. In this portal, each dataset comes with a summary, downloadable files and a SPARQL endpoint.

Views



Datasets

List of RDF datasets with sorting and filtering options



Statistics

Table of sortable statistics of RDF datasets



Links

Graph of links between RDF datasets

<http://integbio.jp/rdf/>

改正個人情報保護法における ゲノムデータ、ゲノム情報の扱い

改正個人情報保護法におけるゲノムデータ等の取扱いについて

(意見とりまとめ)

平成 28 年 1 月 22 日

ゲノム情報を用いた医療等の実用化推進タスクフォース

- 社会通念上、「ゲノムデータ」は「個人識別符号」に位置づけられるもの
- 「ゲノム情報」は配慮を要すべき情報に該当する場合がある
- 具体的な解釈は今後、個人情報保護委員会で

計算ノード

Thin 288 nodes

HP Proliant SL230s Gen8

CPU: Xeon E5 2670x2/node

Memory: 64GB/node

SSD: 400GB SSDx76/node

x1/node

Thin (GPU) 64 nodes

HP Proliant SL250s Gen8

GPGPU: Tesla M2090x1/node

Memory: 64GB/node

SSD: 400GB SSDx64/node

x1/node

Medium 2 nodes

HP Proliant DL980 G7

CPU: Xeon E7-4870x8/node

Memory: 2TB/node

x1/node

Fat 1 node

SGI Altix UV1000

CPU: Xeon E7 8837x96/node

Memory: 10TB/node

FW

Fortigate 3040B

10GbE対応

InfiniBandスイッチ

x4

Mellanox
GD 4700 x1
GD 4036 x9
GD 4036E x6

Mellanox
SX 6536 x1
SX 6036 x1

106 nodes Thin

HP Proliant SL230s Gen8

CPU: Xeon E5 2680v2x2/node

Memory: 64GB/node

SSD: 400GB SSDx32/node

x1/node

64 nodes Thin (GPU)

HP Proliant SL250s Gen8

CPU: Xeon E5 2680v2x2/node

GPGPU: Tesla K20x1/node

Memory: 64GB/node

x1/node

32 nodes Thin (Xeon Phi)

HP Proliant SL250s Gen8

CPU: Xeon E5 2680v2x2/node

Co-processor:

Xeon Phi 5110Px1/node

Memory: 64GB/node

x1/node

nodes Medium

HP Proliant DL980 G7

CPU: Xeon E7-4870x8/node

Memory: 2TB/node

DDBJスパコン
520機関、2500名の利用者

大容量外部記憶装置

省電力領域 3PB

NEXSAN E60/E60X

x12

高速領域 2PB

DDN SFA10000+SS7000

x56

Phase1

Phase2

2.5PB 省電力領域

Hitachi HUS150

x16

5PB 高速領域

DDN SFA12000+SS7000

x120