2017年6月26日
研究データ利活用協議会
公開シンポジウム

データ共有の先行事例の紹介
生命科学におけるデータ共有の歴史、現状、課題

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

生命科学分野の研究用DBの歴史と動向

- 文献データは1960年代より
- 研究データは1970年代より
- 塩基配列データバンクは日米欧の３極体制で
- タンパク質データバンクは日米欧の４センターで
- 他のオミックスデータにも共有の枠組みが拡大
  - トランスクリプトーム、プロテオーム、メタボローム、フェノーム、、、
- 分野別目的別単位でのデータ共有の枠組みも
  - 微生物、植物、実験動物、ヒト疾患、脳、、、

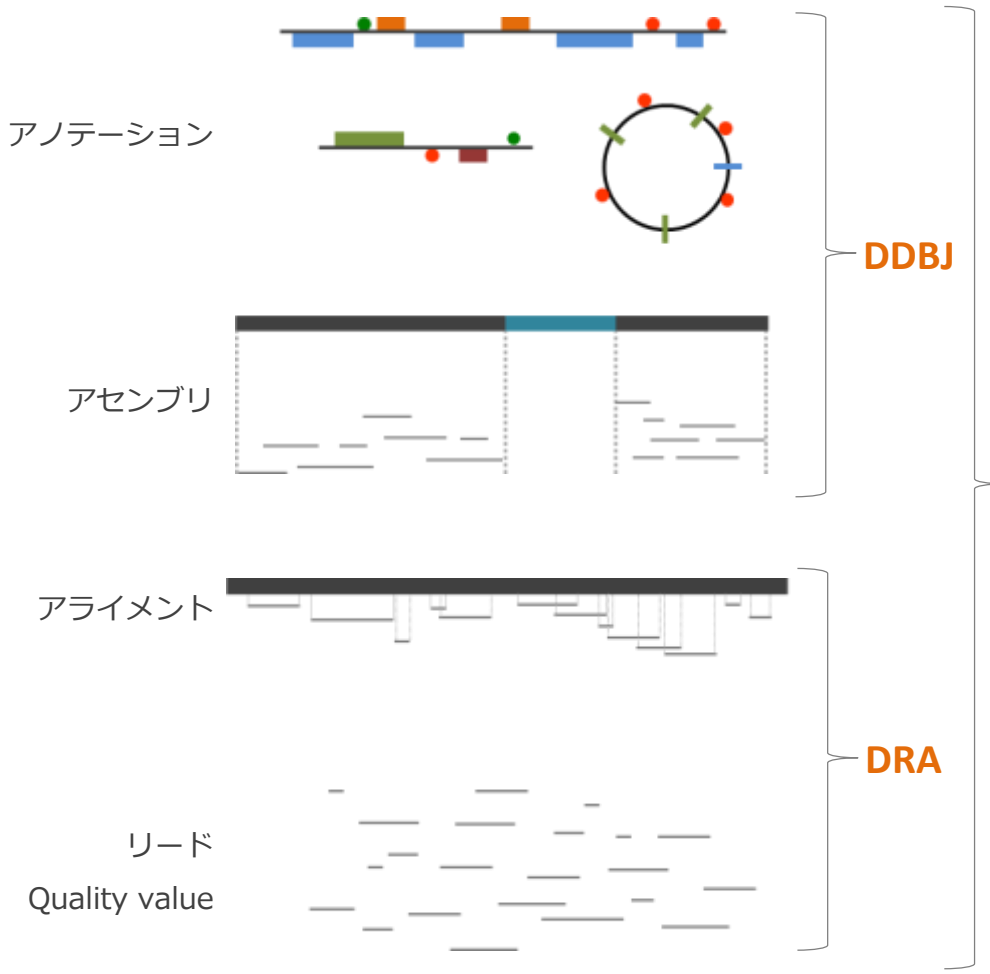
生命科学ではデータ共有がなぜ活発なのか？

- 少数の数式や法則で表現できない（データが重要）
- 研究成果の再現や検証
- データ共有による研究の促進
  - 統計解析のパワーアップ
  - 他の観点からの新発見、イノベーション促進
- 重複の排除、資金の効率化、研究不正への対応
- 資金提供機関からのデータ共有の要請
- 論文投稿時における出版社からのDB登録要請
- 受け皿としてDBセンターやアーカイブの整備
- DBは研究のインフラでありフロンティア

ゲノム関係の国内外のDBセンター

- 1980　EMBL-Bank
- 1982　LANL GenBank
- 1987　NIG DDBJ
- 1988　NIH NLM NCBI
- 1992　EMBL EBI
- 2001　JST BIRD
- 2007　ROIS DBCLS
- 2011　JST NBDC

DDBJ センターが運営するデータベース



制限公開データベース

JGA

個人レベルの遺伝型と表現型情報

NBDC

ヒトデータ審査委員会で提供と利用を審査

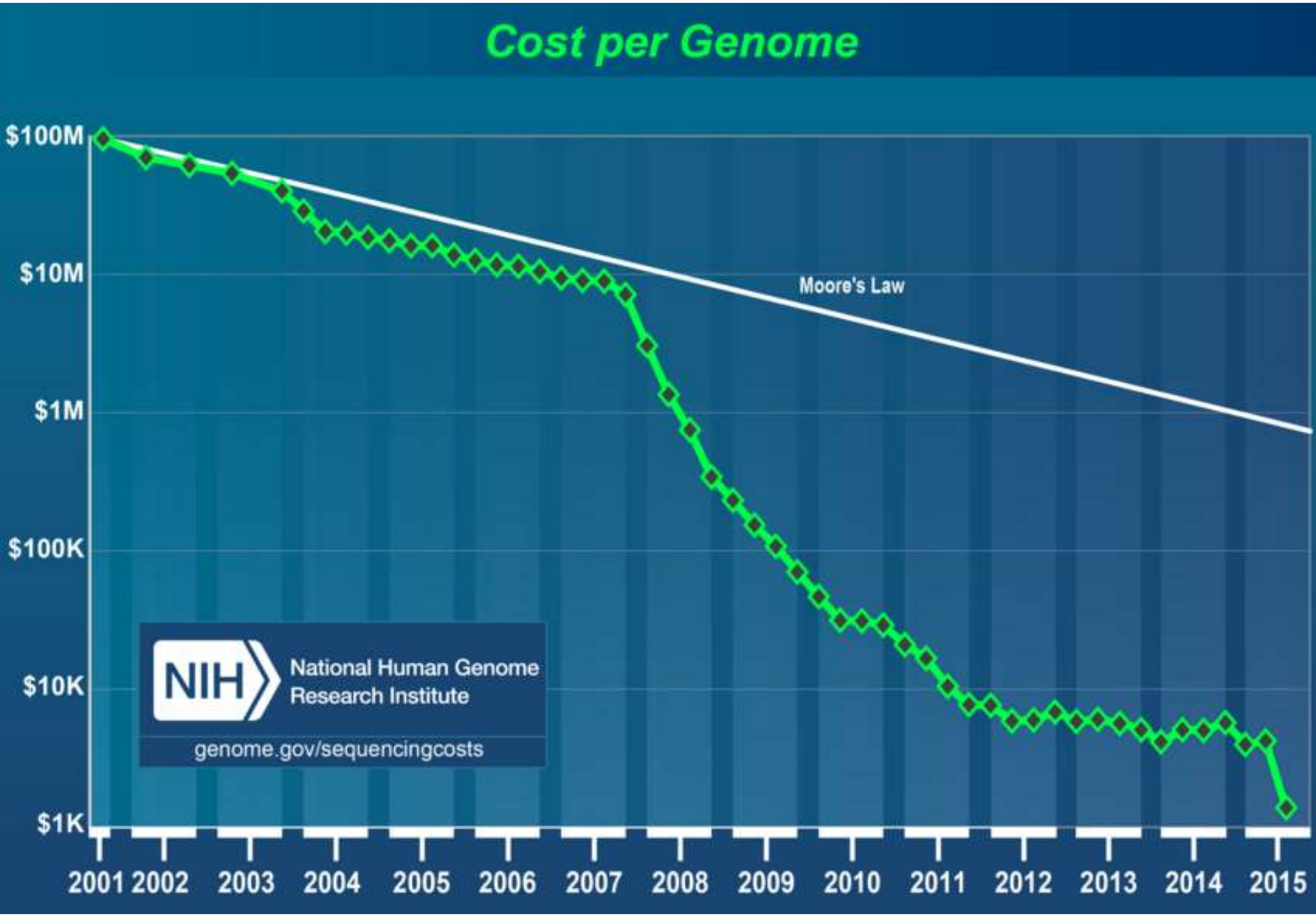
INSDC: オープンアクセスデータベース

情報の多様化と爆発

- データ爆発
  - –次世代ゲノムシークエンサーなどの計測技術の進歩
  - –ムーアの法則を凌駕
  - –10万を超えるゲノムプロジェクト進行中
  - –ゲノム以外のomicsデータや画像も急増
- 知識爆発
  - –論文数　2,700万件
  - –オープンアクセスの拡大
- データベース爆発
  - –世界　　1万から2万（日本は千）
  - –解析ツール　数千
  - –内容も非常に多様化

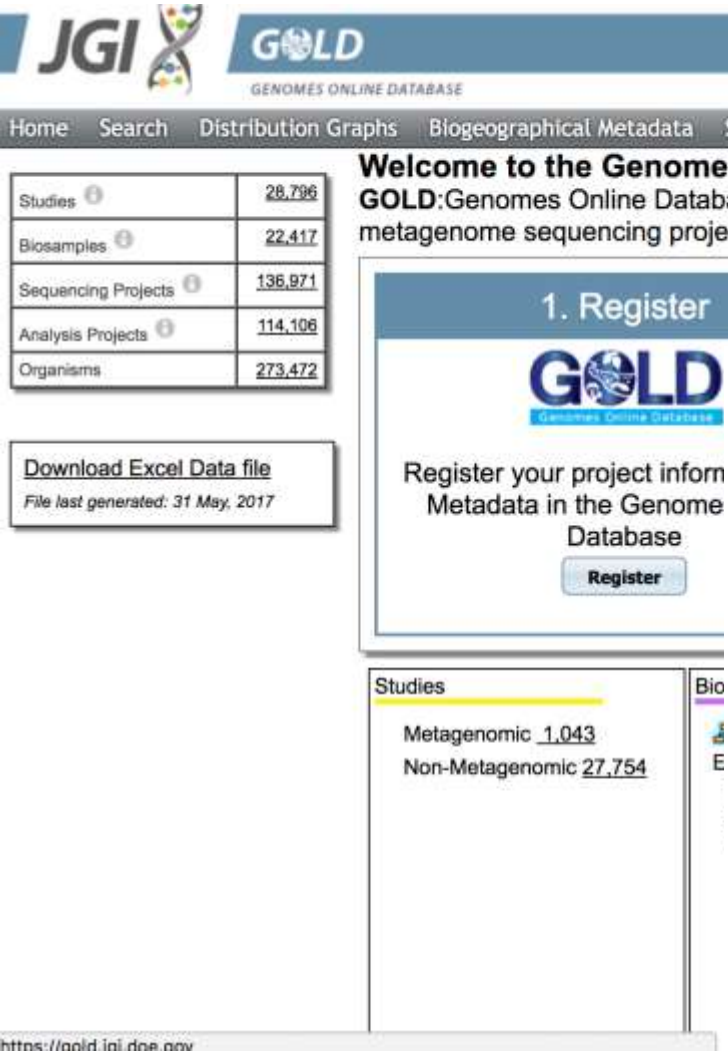
生命科学は
petaオーダーの時代に
主要DBセンターは
数十PBのデータ保有

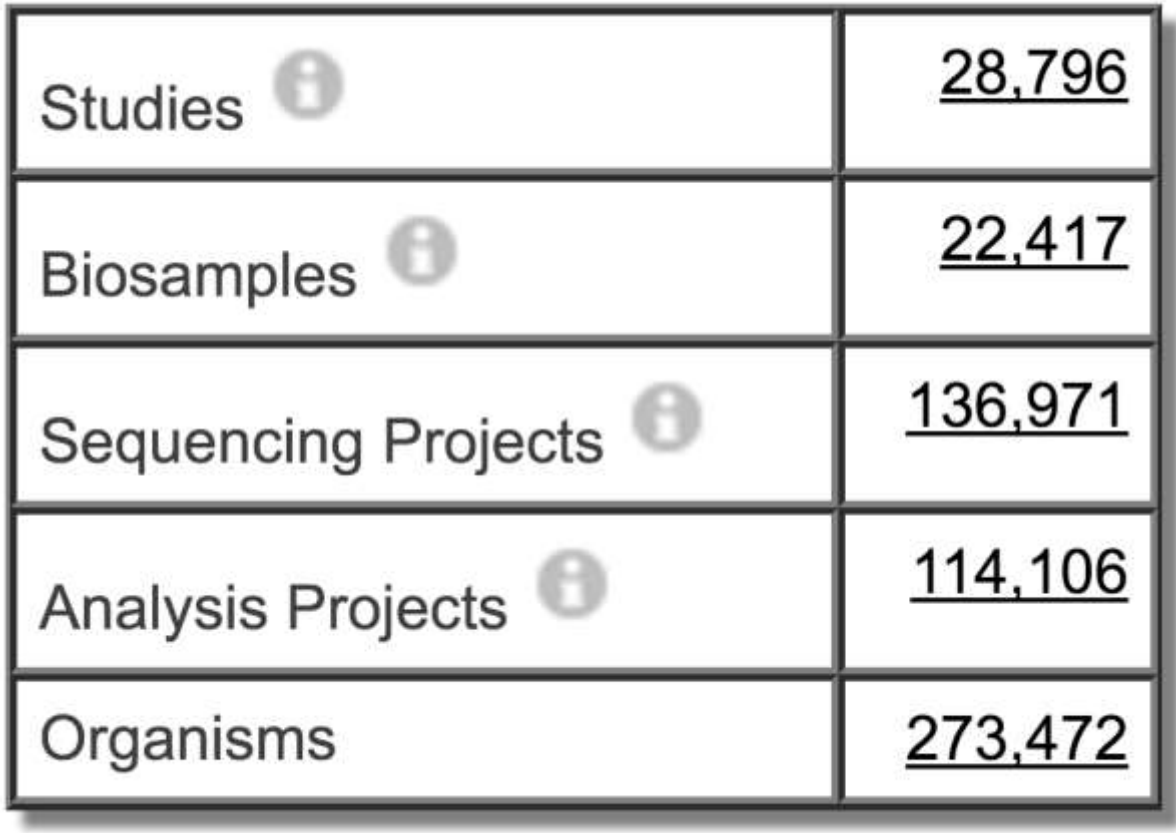
- 生命科学はビッグデータを扱う情報の学問に
  - – データ駆動型科学



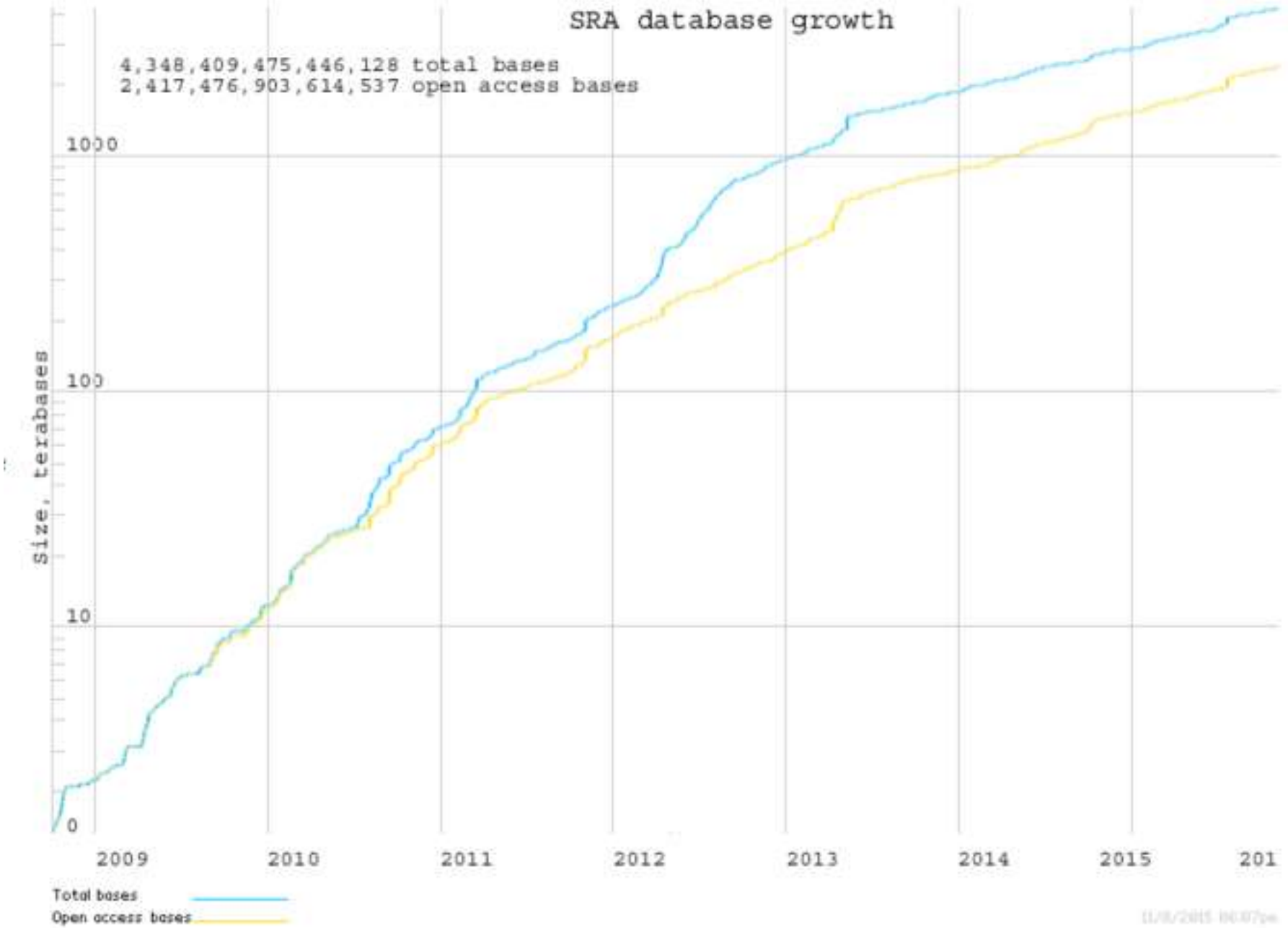
http://www.genome.gov/sequencingcosts/

様々な生物のゲノムプロジェクト





https://gold.jgi.doe.gov/



http://www.ncbi.nlm.nih.gov/Traces/sra/

生命科学におけるデータの利活用に関する障害

- 自分の専門外の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不明
  - 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いものが混在
- フォーマットや用語がバラバラ
  - 遺伝子の概念さえDBによって違う
  - 同じ遺伝子にも多数の名前あり
- データの文脈依存性、曖昧性、冗長性、複雑性
- 単純にレポジトリするだけでは再利用性低い

10年ほど前の我が国固有の事情

- 資金提供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
- 小規模プロジェクト多い→ビッグデータ化必要
- バイオインフォマティシャン不足→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）

我が国の生命科学DB統合推進事業

- 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～情報・システム研究機構ライフサイエンス統合DBセンターDBCLS
    - 2011～科学技術振興機構バイオサイエンス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）

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、さきがけ)(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。

生命科学DB統合推進事業の成果

Life Science Database Archive
LSDB Archive

- あのデータベースが、丸ごとダウンロード可能に！ -

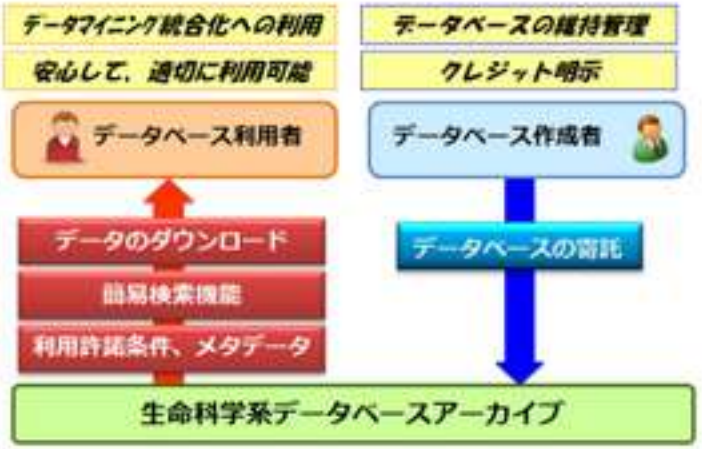
生命科学系データベース アーカイブ

ホーム | アーカイブの説明 | 寄託応募要領 | 更新履歴 | 利用状況 | ヘルプ | お問い合わせ

いくら良質なデータベースでも、説明が十分でない、利用条件が明確でない、ダウンロードできないなどの理由で十分に利用され、引用され、相応しい評価をうける機会を逃していることがあります。

生命科学系データベースアーカイブは、国内のライフサイエンス研究者が生み出したデータセットをわが国の公共財としてまとめて長期間安定に維持保管し、データ説明（メタデータ）を統一して検索を容易にすると共に、利用許諾条件などの明示を行うことで、多くの人が容易にデータへアクセスしダウンロードを行えるようにするサービスです （詳細説明）。

データを長期にわたり保全し、データベース作成者のクレジットを明示する一方、公的機関や民間等様々なユーザが利用しやすい形にすることで、それぞれの研究の生命科学へのいっそうの貢献を支援します。 データベースの寄託を随時募集しています（寄託応募要領）。



http://dbarchive.biosciencedbc.jp/index.html

全メタデータをエクスポート

アーカイブデータベース一覧 (ヘルプ)

一覧内検索　　詳細検索

全 132 件 (1 件から5件)　5 件を表示

最初へ　前へ　1　2　3　4　5　…　27　次へ　最後へ

データベース	データベース運用場所	代表者	データベースカテゴリ	生物種	要約（キーワードを太字表示）	利用許諾
 **Togo Picture Gallery** ダウンロード | 簡易検索 オリジナルサイト	ライフサイエンス統合データベースセンター	小野 浩雅	画像コレクション	-	生命科学分野の**誰でも自由に閲覧・利用できる無料の画像**のコレクション。	CC 表示 詳細
 **TogoTV** ダウンロード オリジナルサイト	ライフサイエンス統合データベースセンター	小野 浩雅	動画コレクション	-	生命科学分野の有用な**データベースやツールの使い方を紹介する動画**のコレクション	CC 表示 詳細
 **抗体医薬品データベース** ダウンロード | 簡易検索 オリジナルサイト	産業技術総合研究所　創薬分子プロファイリング研究センター	福井 一彦	抗体医薬品	-	**抗体医薬品**について、承認薬や臨床試験における3段階の治験段階の情報をまとめたデータベース	CC 表示-継承 詳細
 **SKIP幹細胞情報データベース** ダウンロード | 簡易検索 オリジナルサイト	慶應義塾大学医学部臨床遺伝学センター/幹細胞情報室	小崎 健次郎	遺伝子、疾患	ヒト	各研究機関が保有する**ヒト幹細胞**サンプルに関する情報を集約したデータベース	CC 表示-継承 詳細
 **PGDBj - オルソログデータベース** ダウンロード | 簡易検索 オリジナルサイト	かずさDNA研究所	中谷 明弘	オルソログ	緑色植物、ラン藻	異なる生物種間における**アミノ酸配列の類似性**に基づいた遺伝子の**オルソログ情報**を蓄積したデータベース	CC 表示-継承 詳細

http://dbarchive.biosciencedbc.jp/index.html

NBDCにおける分野別目的別のデータベース統合

- プロテオーム統合データベースの構築
- 生命動態情報と細胞・発生画像情報の統合データベース
- エピゲノミクス統合データベースの開発と機能拡充
- ゲノム・疾患・医薬品のネットワークデータベース
- 糖鎖科学ポータルの構築
- 蛋白質構造データバンクのデータ検証高度化と統合化
- データサイエンスを加速させる微生物統合データベースの高度実用化開発
- 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ーミクスデータの統合
- 個体ゲノム時代に向けた植物ゲノム情報解析基盤の構築

NBDCヒトデータベース／データの種類

NBDCヒトデータベース

非制限公開（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

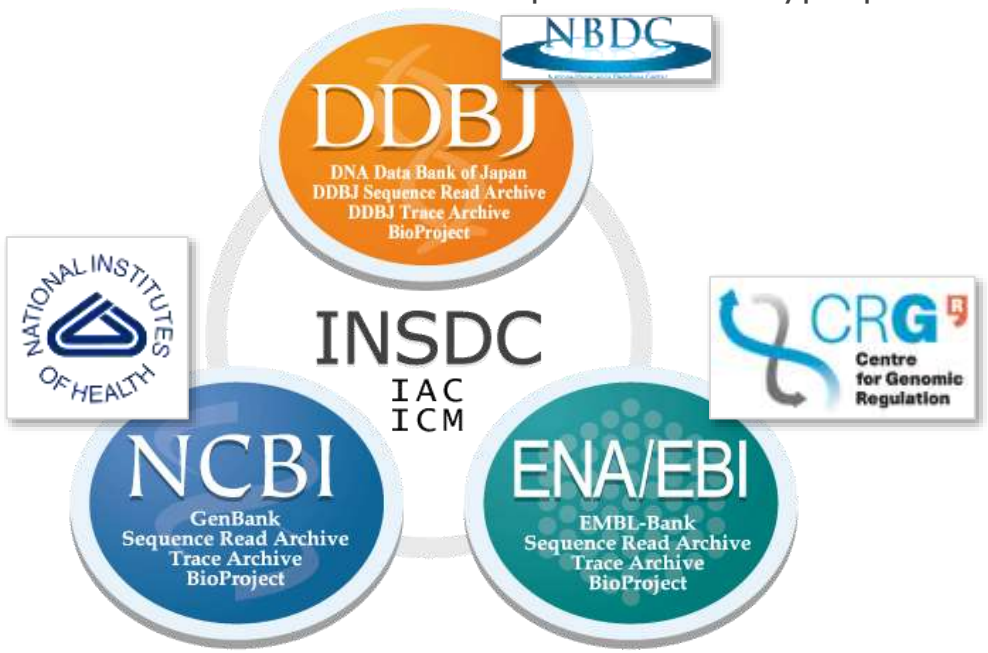
1．運用原則

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繋がらないよう、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のdbGAPで導入、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

アクセス制限（制限公開）データベース

JGA Japanese Genotype-phenotype Archive



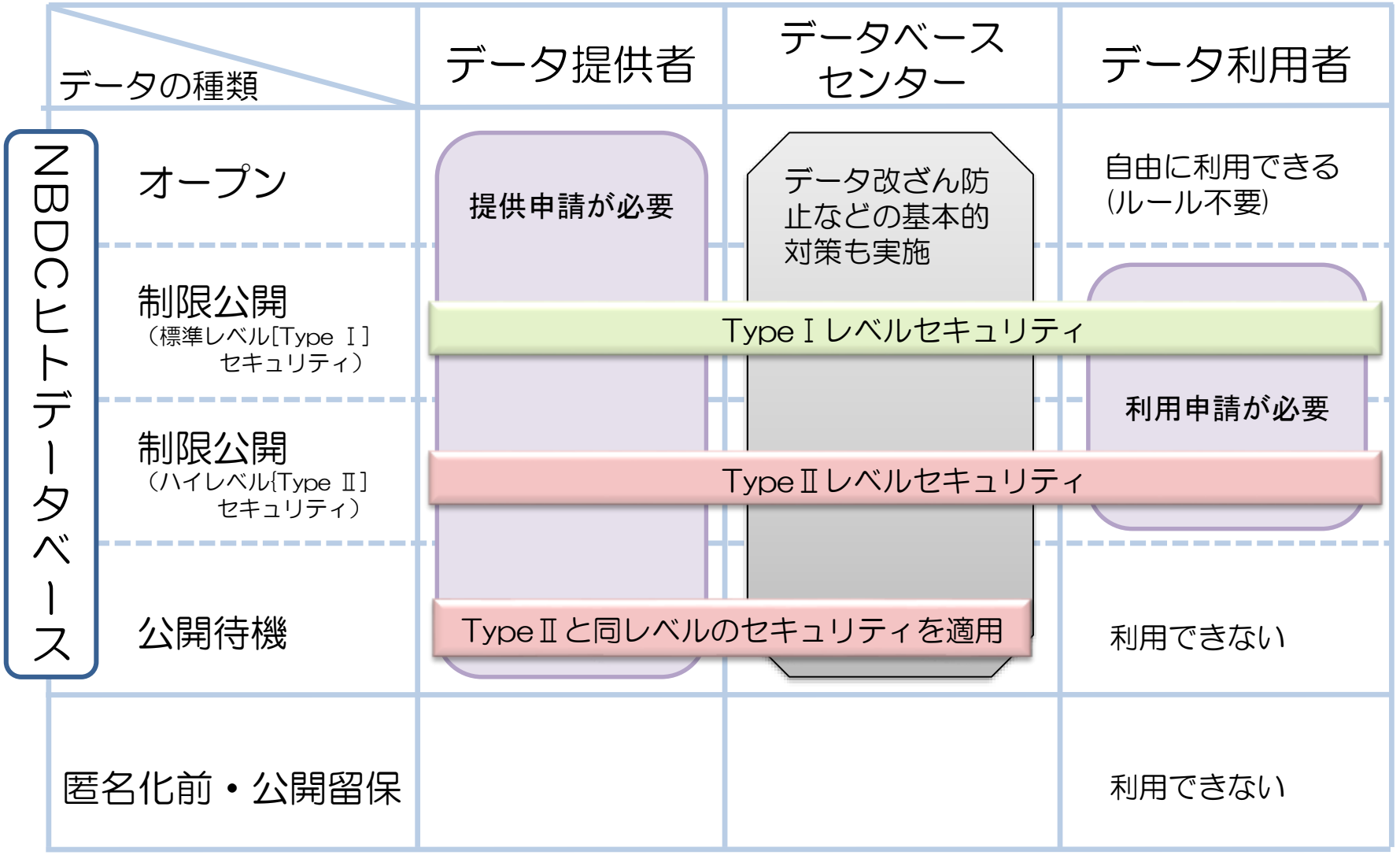
dbGaP
Database of Genotype and Phenotype

EGA
European Genome-phenome Archive

- ✓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)

NBDCヒトデータベースの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



NBDCヒトデータベース

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下さい。

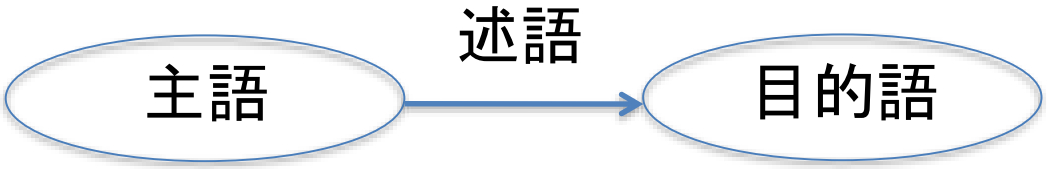
Research ID	研究題目	公開日	データの種類	研究方法	手法	参加者（対象集団）	提供者	アクセス制限
hum0001.v1	SCA31罹患患者のゲノム解析データ	v1:2013/12/01	NGS (Whole)	配列決定	Illumina	1検体	森下真一	制限(Type I)
hum0003.v1	関節リウマチ患者及び健常人における HLA領域の塩基配列比較解析	v1:2013/07/01	NGS (Target Captu					
hum0004.v1	上皮成長因子受容体遺伝子異変を有する 肺腺癌の体細胞性遺伝子変異プロファイル 解明のための分子疫学的研究	v1:2014/07/11	NGS (Exome)					
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 technoligies (Ion PGM)	17検体	宇佐美真一	オープン
hum0006.v1 (JGAS00000000004)	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）	斉藤延人	制限(Type I)

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

公開中　52件（制限公開含む）
提供申請　123件、26万検体

RDFによるデータと知識の統合

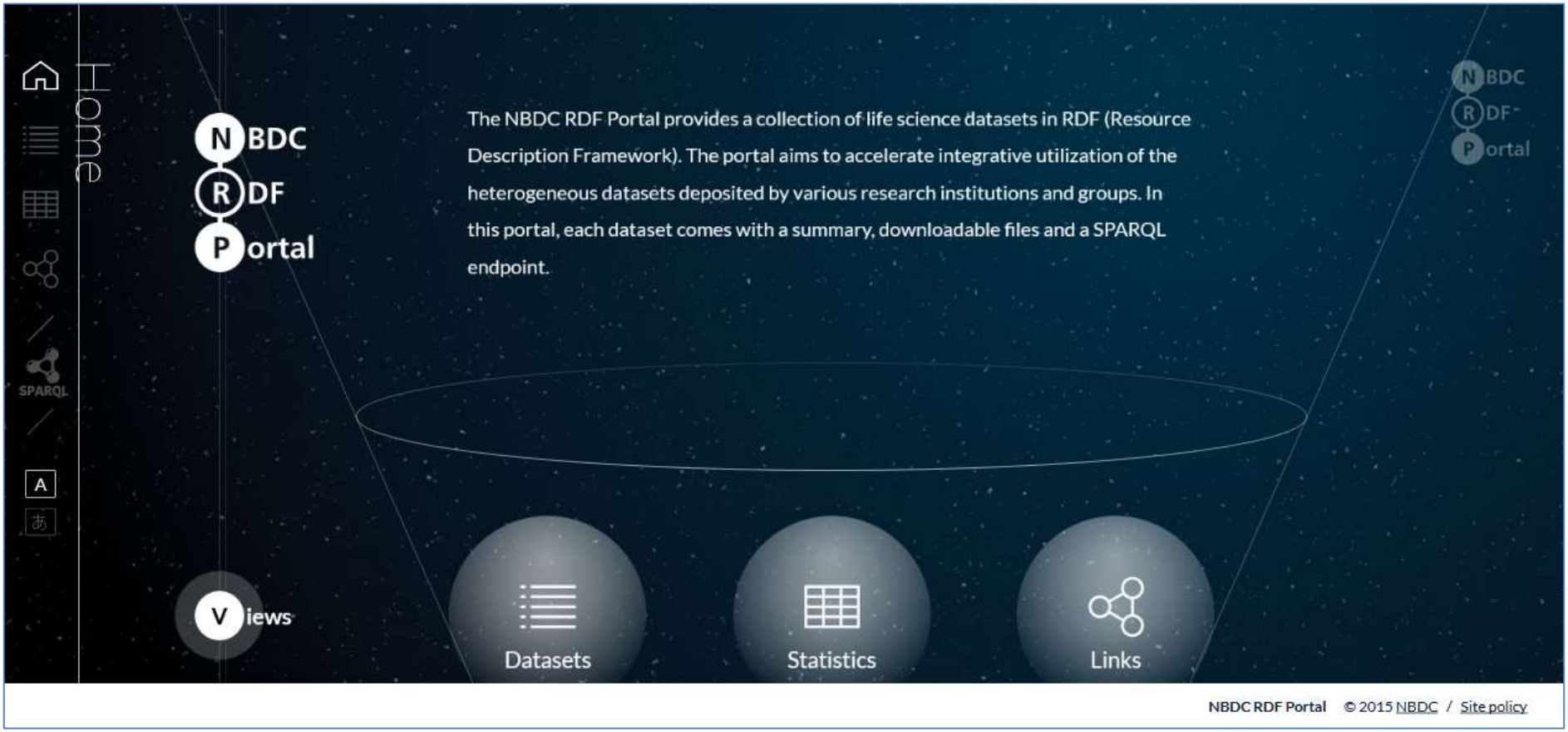
- Resource Description Frameworkの略
- （主語、述語、目的語）の3つ組（トリプル）で



- 主語、述語はURI（Uniform Resource Identifier)で

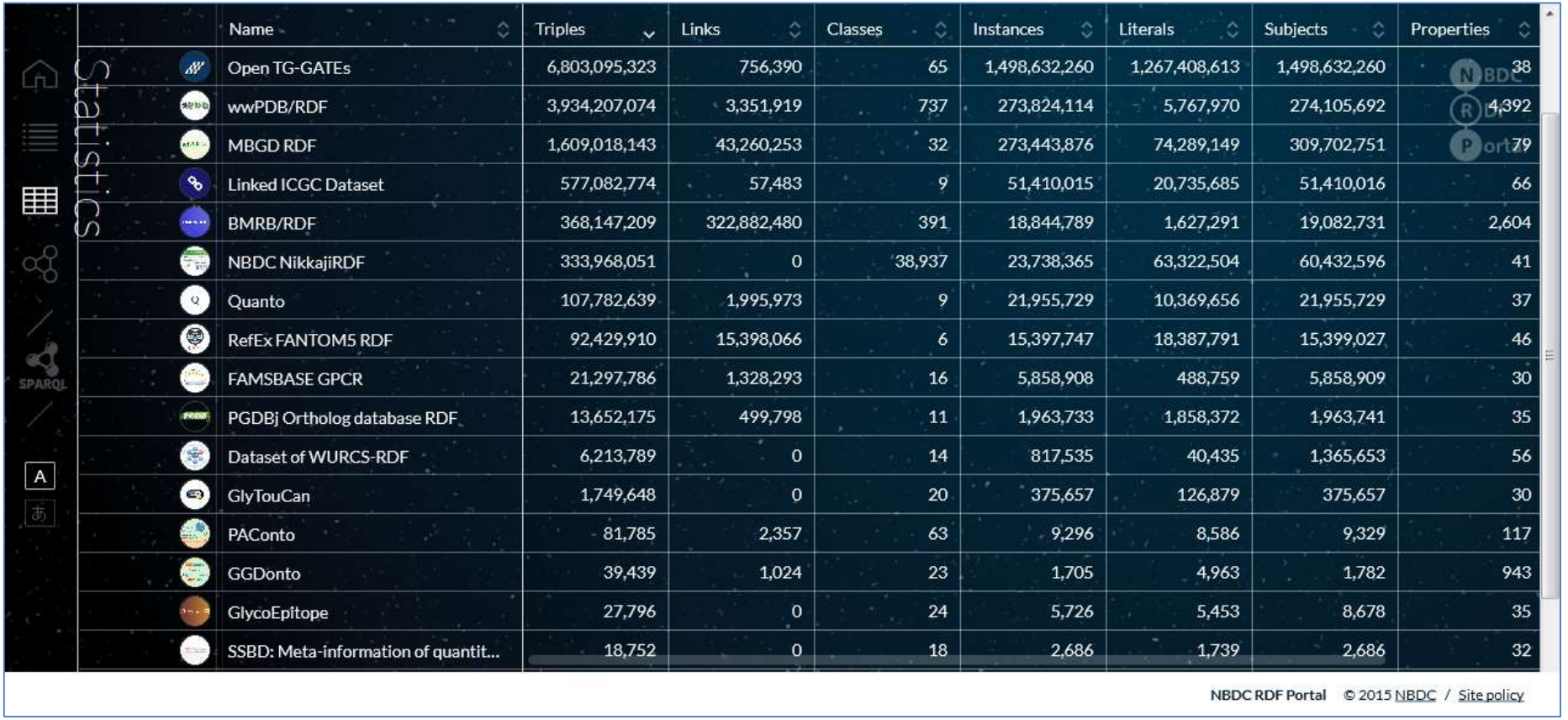
- Semantic Web
- LOD　(Linked Open Data)
- Web of Data

NBDC RDF Portal



- DBCLSのRDF化ガイドラインに沿う17 DBを収録
- SPARQLエンドポイントから利用可能

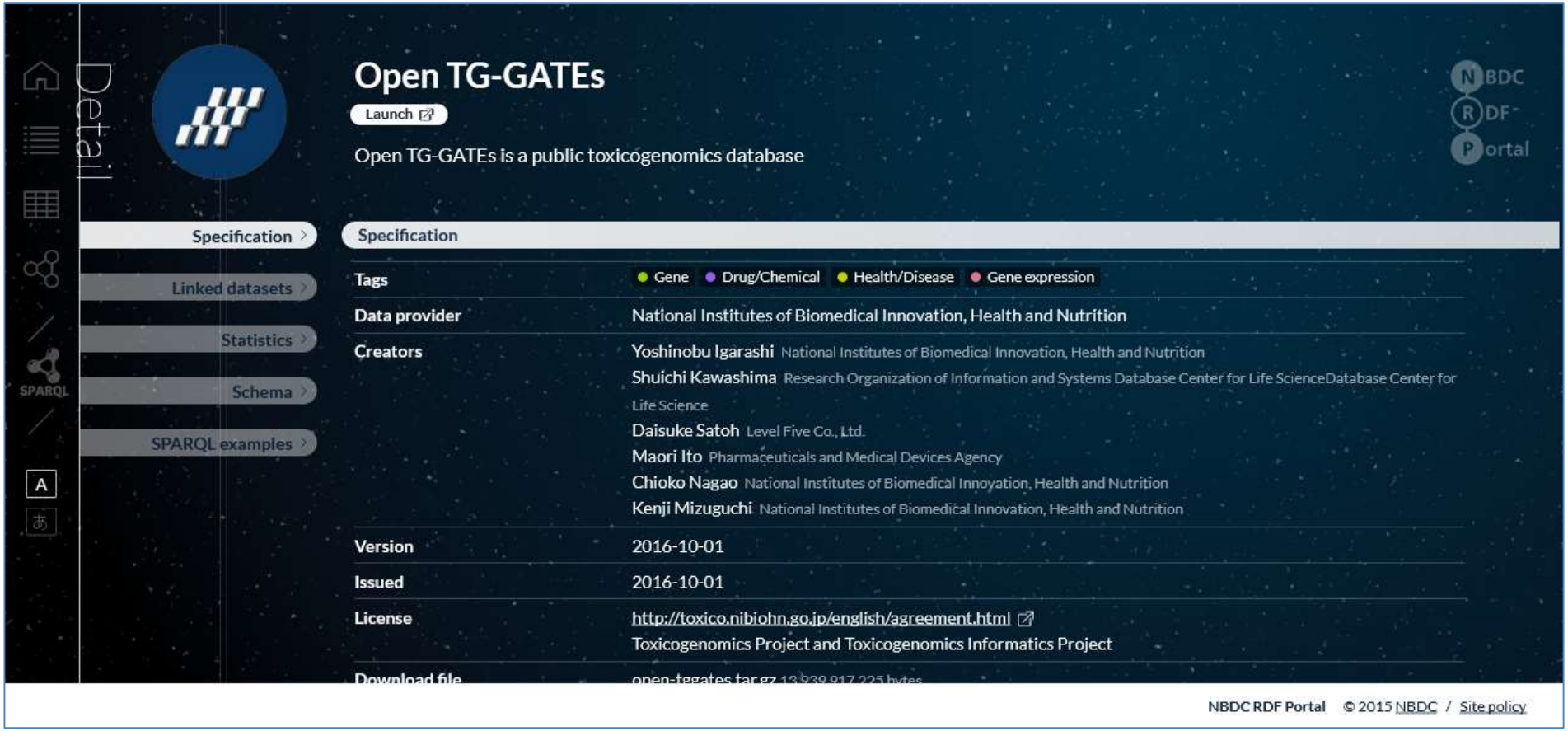
Databases in NBDC RDF Portal



Name	Triples	Links	Classes	Instances	Literals	Subjects	Properties
Open TG-GATEs	6,803,095,323	756,390	65	1,498,632,260	1,267,408,613	1,498,632,260	38
wwPDB/RDF	3,934,207,074	3,351,919	737	273,824,114	5,767,970	274,105,692	4,392
MBGD RDF	1,609,018,143	43,260,253	32	273,443,876	74,289,149	309,702,751	79
Linked ICGC Dataset	577,082,774	57,483	9	51,410,015	20,735,685	51,410,016	66
BMRB/RDF	368,147,209	322,882,480	391	18,844,789	1,627,291	19,082,731	2,604
NBDC NikkajiRDF	333,968,051	0	38,937	23,738,365	63,322,504	60,432,596	41
Quanto	107,782,639	1,995,973	9	21,955,729	10,369,656	21,955,729	37
RefEx FANTOM5 RDF	92,429,910	15,398,066	6	15,397,747	18,387,791	15,399,027	46
FAMSBASE GPCR	21,297,786	1,328,293	16	5,858,908	488,759	5,858,909	30
PGDBj Ortholog database RDF	13,652,175	499,798	11	1,963,733	1,858,372	1,963,741	35
Dataset of WURCS-RDF	6,213,789	0	14	817,535	40,435	1,365,653	56
GlyTouCan	1,749,648	0	20	375,657	126,879	375,657	30
PAConto	81,785	2,357	63	9,296	8,586	9,329	117
GGDonto	39,439	1,024	23	1,705	4,963	1,782	943
GlycoEpitope	27,796	0	24	5,726	5,453	8,678	35
SSBD: Meta-information of quantit...	18,752	0	18	2,686	1,739	2,686	32

- 現在Open TG-GATEsが最大 (70億トリプル)
- 近々DDBJ RDFの200億トリプルが公開される予定

Open TG-GATEs RDF
(170の化合物（医薬品）の毒性検査に関する情報)



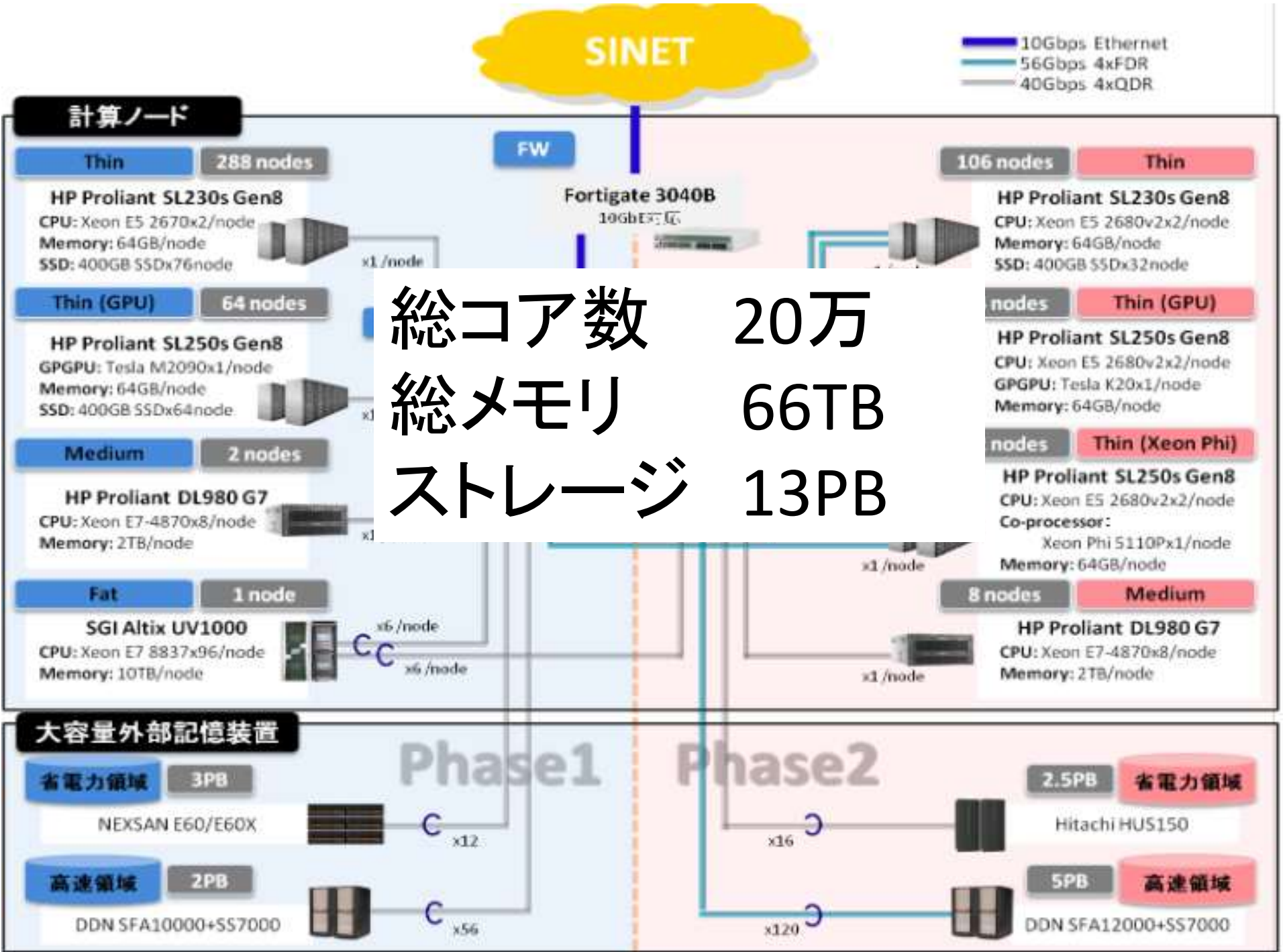
- 各RDFデータには詳細なメタデータや内部構造を示すスキーマ図が付与されている

これまでの10年を振り返って

- 当時のもくろみ
  - ITによるデータや知識の整理統合＆推論による仮設生成、質問応答
  - データ駆動型科学
- ある程度できたこと
  - FAIR
  - 生命研究者のデータ共有、DBへの理解増進
  - データシェアリングポリシー、DMP、などの導入（まだ不十分）
  - 中核センターの設置（ただし、まだ問題残っている）
- あまりできなかったこと
  - データ駆動型科学の実践
  - データ共有のインセンティブ付与
  - 人材育成
- もくろみが達成できなかった主な理由
  - ITによる統合の前にデータ共有の大きな壁
  - 生命科学データの多様性、文脈依存性、複雑性、曖昧性、など

爆発するデータ、知識への対応

データ共有のコストは誰が負担？
何をDBとして残すべきか？
スパコン（ストレージ、計算パワー）などの基盤整備
様々な技術の開発

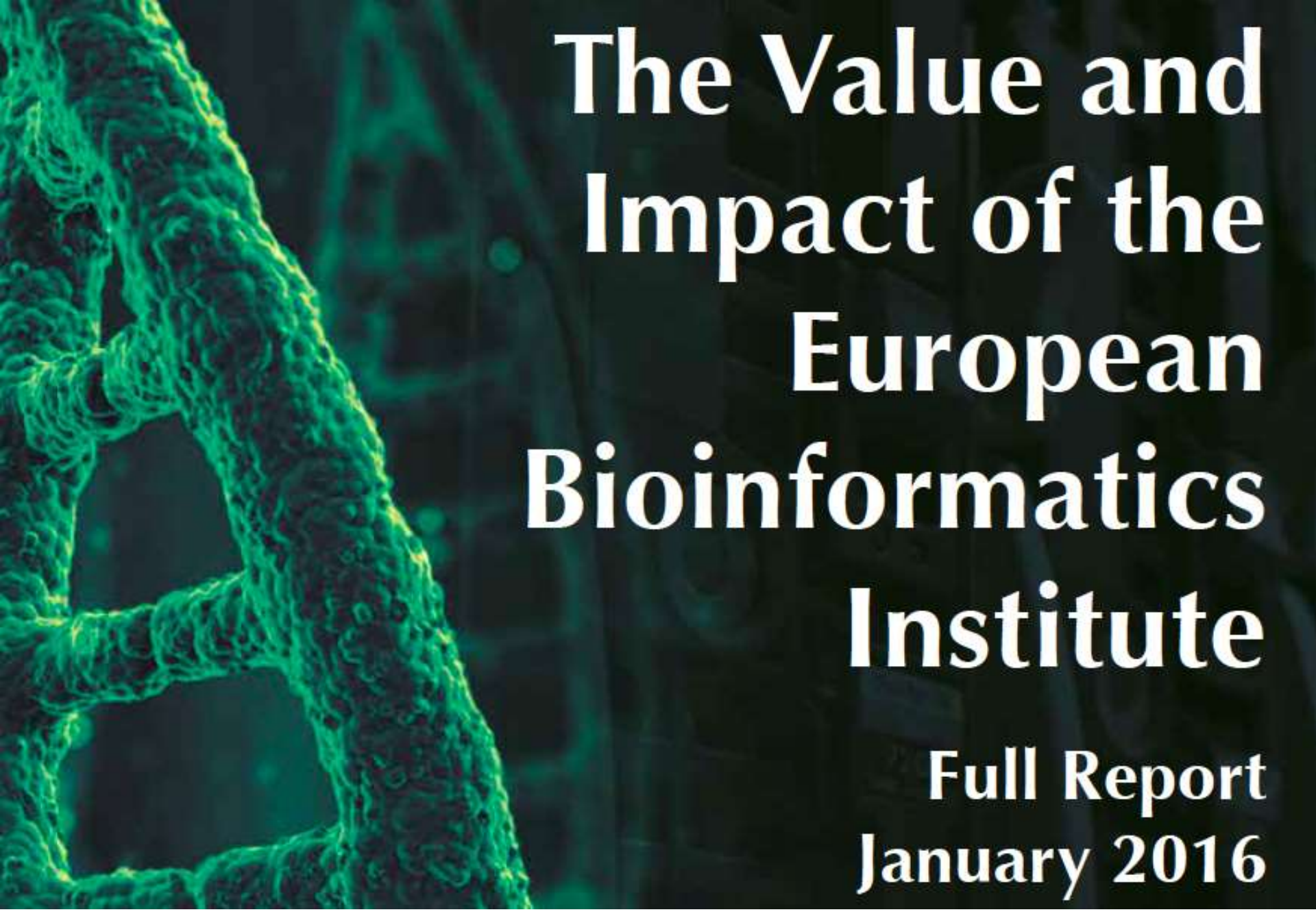

SINET
10Gbps Ethernet
56Gbps 4xFDR
40Gbps 4xQDR
計算ノード
FW
Fortigate 3040B
Thin
288 nodes
HP Proliant SL230s Gen8
CPU: Xeon E5 2670x2/node
Memory: 64GB/node
SSD: 400GB SSDx76node
x1 /node
Thin (GPU)
64 nodes
HP Proliant SL250s Gen8
GPGPU: Tesla M2090x1/node
Memory: 64GB/node
SSD: 400GB SSDx64node
Medium
2 nodes
HP Proliant DL980 G7
CPU: Xeon E7-4870x8/node
Memory: 2TB/node
Fat
1 node
SGI Altix UV1000
CPU: Xeon E7 8837x96/node
Memory: 10TB/node
x6 /node
x6 /node
106 nodes
Thin
HP Proliant SL230s Gen8
CPU: Xeon E5 2680v2x2/node
Memory: 64GB/node
SSD: 400GB SSDx32node
nodes
Thin (GPU)
HP Proliant SL250s Gen8
CPU: Xeon E5 2680v2x2/node
GPGPU: Tesla K20x1/node
Memory: 64GB/node
nodes
Thin (Xeon Phi)
HP Proliant SL250s Gen8
CPU: Xeon E5 2680v2x2/node
Co-processor:
Xeon Phi 5110Px1/node
Memory: 64GB/node
x1 /node
8 nodes
Medium
HP Proliant DL980 G7
CPU: Xeon E7-4870x8/node
Memory: 2TB/node
x1 /node
総コア数 20万
総メモリ 66TB
ストレージ 13PB
大容量外部記憶装置
Phase1
Phase2
省電力領域
3PB
NEXSAN E60/E60X
x12
高速領域
2PB
DDN SFA10000+SS7000
x56
x16
x120
2.5PB
省電力領域
Hitachi HUS150
5PB
高速領域
DDN SFA12000+SS7000

遺伝研スパコン利用機関（抜粋）

国内250以上の大学、研究機関,2500名の研究者が利用

持続可能な体制や基盤の構築

オープンサイエンス、データ共有の効能の評価
プロジェクトと連動した予算の確保



https://beagrie.com/static/resource/EBI-impact-report.pdf

我が国におけるDBセンターの連携・一元化

（国際的な）ヒト由来データの共有

マシンリーダブルコンセント、フォーマット、研究者認証、アクセス権限、
セキュリティ、ELSI（個人情報保護法など）、
マイクロアトリビューション

Global Alliance
for Genomics & Health

Become a Member

ABOUT GLOBAL ALLIANCE | WORKING GROUPS | PRODUCTS & PROJECTS | MEMBERSHIP | NEWS & EVENTS | CONTACT US

Home → About the Global Alliance

About the Global Alliance

- Mission & Founding Principles
- How We Work
- Governance
- History
- Key Documents

About the Global Alliance

The **Global Alliance for Genomics and Health** (Global Alliance) was formed to help accelerate the potential of genomic medicine to advance human health. It brings together over 400 leading institutions working in healthcare, research, disease advocacy, life science, and information technology. The partners in the Global Alliance are working together to create a common framework of harmonized approaches to enable the responsible, voluntary, and secure sharing of genomic and clinical data.

http://genomicsandhealth.org/about-global-alliance

GA4GH 加入組織

- 411組織（41カ国）
  - 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12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
  - 株式会社テンクー

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。

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

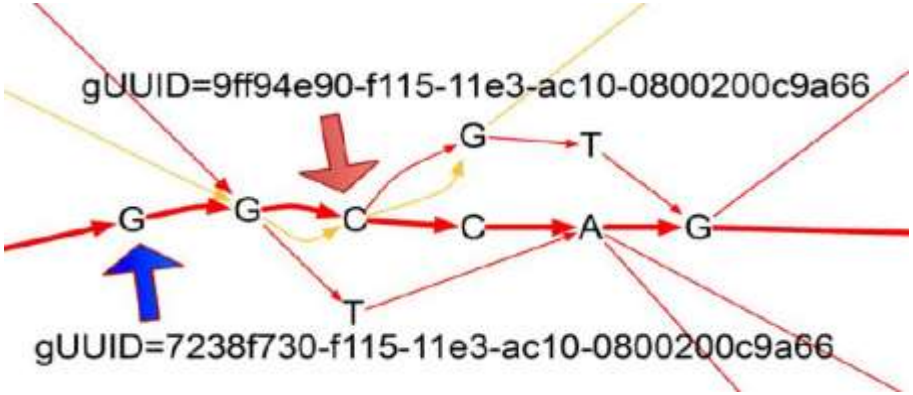
乳ガンやその他のガンの遺伝要因の理解を深めるために、世界中からガンに関与する遺伝子多型データを共有するための試み。まずは乳ガンのデータ共有を進めている。

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。

Reference Graph

Data Working GroupのReference Variation Task Teamでの活動



リファレンスDNA配列
GGCCAG

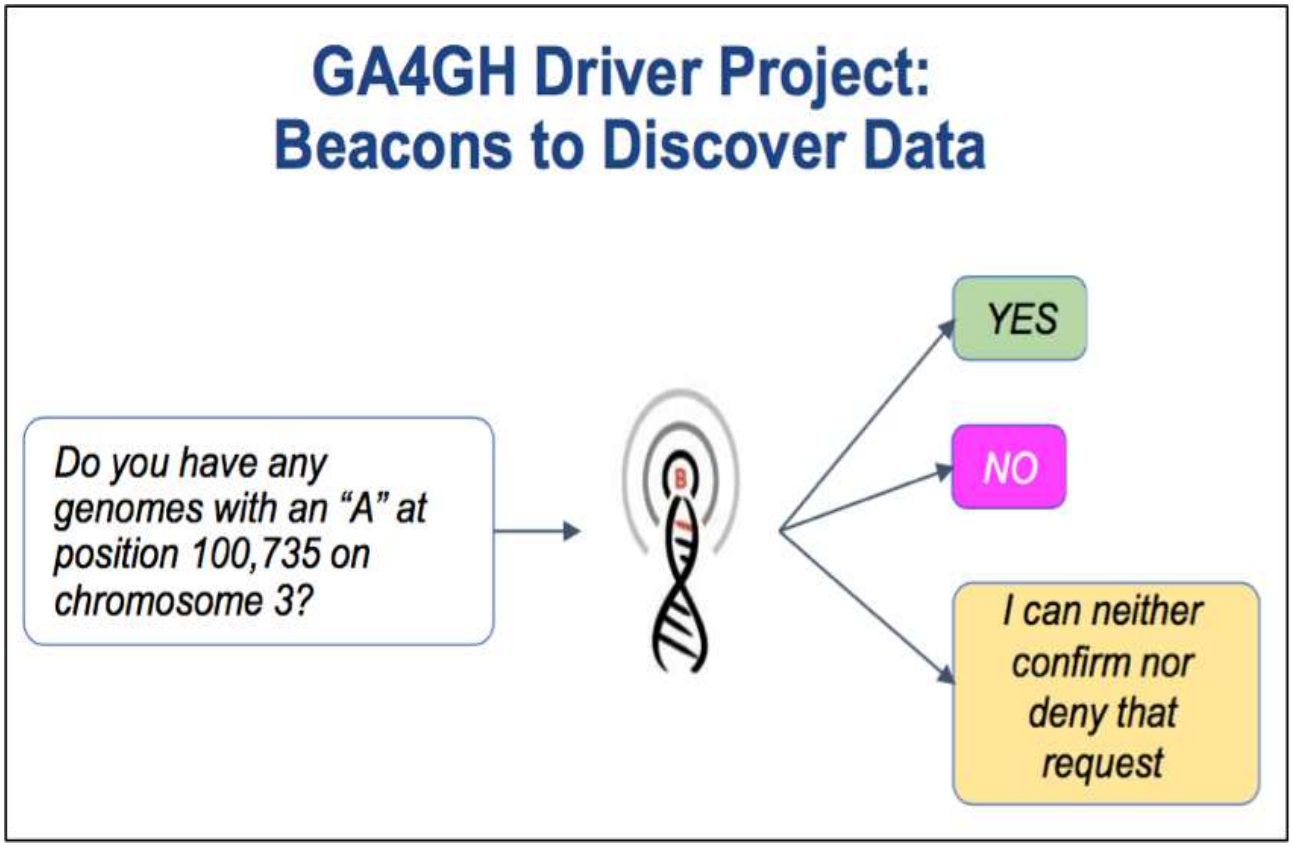
https://genomicsandhealth.org/files/public/6-Beacon-HausslerGA4GHLeiden.pdf

- 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。

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。

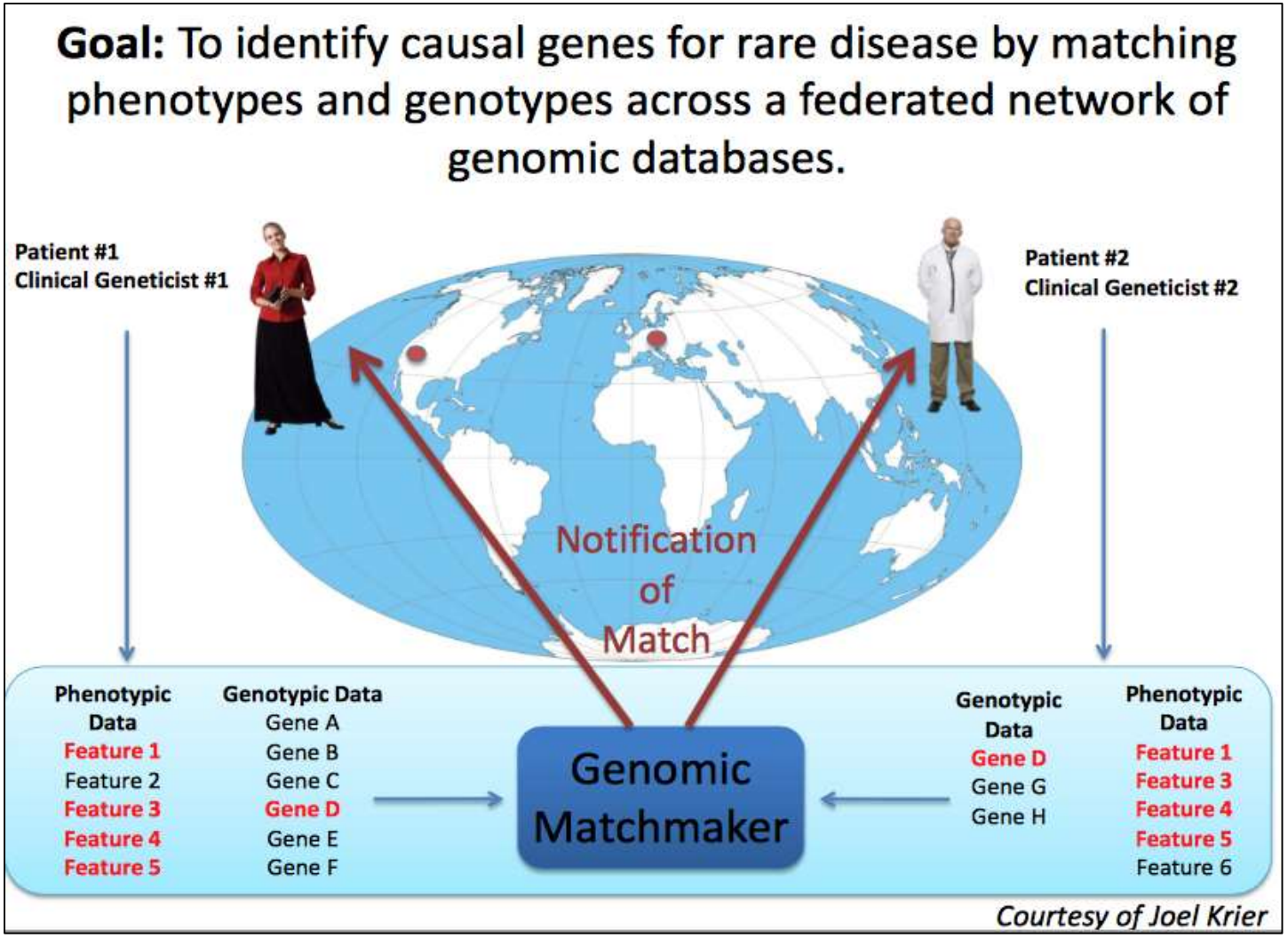


https://genomicsandhealth.org/files/public/6-Beacon-HausslerGA4GHLeiden.pdf

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を持つ希少疾患者を探す仕組み



https://genomicsandhealth.org/files/public/8Plenary2Presentation-MatchmakerExchange-HeidiRehm.pdf

人材の育成・発掘

教育体制、ポジション、キャリアパス

なぜ人材育成はうまく行かなかった？

- 人材は育ってきたが需要の拡大の方が大きかった！
- なぜ需要を見越して対応できなかったのか？
  - 教育体制の問題
  - 受け皿（産業界、研究機関、大学）の問題
  - 制度面の問題、データ囲い込み問題
  - ポジション、キャリアパスの問題
  - 評価の問題
- 人材の分類とそれに合わせた対応策必要
- 参入障壁の解消（良いデータベースの開発）