2015年10月5日
トーゴーの日シンポジウム2015

オープンライフサイエンスと統合データベース

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)

高木利久

© 2015　高木利久（バイオサイエンスデータベースセンター）licensed under CC表示2.1日本

トーゴーの日シンポジウム

- 第1回
  - 日時：2009年10月5日
  - 場所：東京大学弥生講堂一条ホール
  - 主催：ROISライフサイエンス統合データベースセンター(DBCLS)
- 第0回
  - 日時：2009年6月12日
  - 場所：東京大学武田先端知ビル
  - 主催：ROISライフサイエンス統合データベースセンター(DBCLS)
- 第２回以降
  - 日時：10月５日に開催
- 第3回以降
  - 主催：JSTバイオサイエンスデータベースセンター(NBDC)

- 今年が7回目（8回目）

科学における情報の上手な権利化と共有化

2009年10月5日　東京大学弥生講堂一条ホール

- 「閉会の挨拶」 堀田 凱樹(情報･システム研究機構)
- 「開会挨拶･開催趣旨説明」 高木 利久(DBCLS)
- 「デジタル科学への最後のステップ」 大久保 公策（国立遺伝学研究所)
- “Copyright in the Digital Age and Its Impact on Scientific Data Sharing” Lawrence Lessig　(Harvard Law School, Professor)

http://togotv.dbcls.jp/ja/20100512.html

- 「デジタル時代の著作権とイノベーション」 中山 信弘(明治大学)
- 「公表したものは共有財産」 長尾 真　(国立国会図書館長)
- 「情報の共有と独占について」 末吉 亙(末吉綜合法律事務所)
- 「OECD データ･アクセス･ガイドラインおよび欧州におけるデータ･アクセス推進政策動向」 深作 裕喜子 (INNOVMOND)
- 「パネル討論」（大久保 公策、高木 利久、末吉 亙、長洲 毅志（エーザイ株式会社）、永山 裕二（文化庁）、深作 裕喜子）

ライフサイエンスを取り巻く最近のキーワード

オープンサイエンス

第4の科学

データサイエンス

ビッグデータ

マルチオミックス

クラウド

統合データベース

AMED

ゲノムコホート

ゲノム医療

植物の分子デザイン

人工知能(AI)

IBMワトソン

ディープラーニング

データの産出、共有、統合、解析の流れ

健康長寿社会　予防医学　ゲノム医療　高効率高機能農業

予測、設計

モデル構築解析（仮説駆動型）　二つのアプローチの融合　オミックス解析（データ駆動型）

数理モデル　システムバイオロジー　シミュレーション　人工知能　バイオインフォマティクス　機械学習　統計解析

バイオインフォ人材育成

産学連携パートナーシップ

国際連携、国際共同研究

DBセンター、バイオインフォセンター（のアライアンス）によるデータと知識の共有・統合・解析プラットフォーム

オントロジー　RDF　国際標準化　テキストマイニング　著作権、CC　ELSI　暗号化匿名化　クラウド

高速ネットワークによるスパコン、大容量ストレッジの連携によるデータ科学推進ネットワーク

特許　報告書　データベース　論文

生体情報　臨床情報　医薬品情報　生活情報　環境情報　育種情報　農業情報

病院　研究機関　大学　民間企業　ベンチャー　センター　プロジェクト　コホート　コンソーシアム

ウエアラブルセンサー　生体センサー　環境センサー　IoT 次世代顕微鏡　NGS　質量分析器　MRI　飛行ドローン

iPS技術　ゲノム編集技術

ヒト　バイオバンク　モデル生物　バイオリソース　環境（腸内　土壌　空気中　海洋）　化合物　微生物　植物

ライフ分野の研究データ共有の歴史と動向

- 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(INSDC)は日米欧の３極体制で
- タンパク質データバンク(wwPDB)は日米欧の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

データ共有がなぜ進められて来たのか？

- 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
- 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

データの利用に関する障害

- 自分の専門外のDBを使う必要性あり
- 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、誤差
- 単にレポジトリするだけでは再利用性低い

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
- データの囲い込み（公開されないデータも多数）
- データの権利関係不明
- バイオインフォマティシャンの不足→競争に負ける
- 受け皿となる中核DBセンターがない
- プロジェクト終了するとデータの維持管理更新されない

ライフサイエンスに係るデータの実態（1）

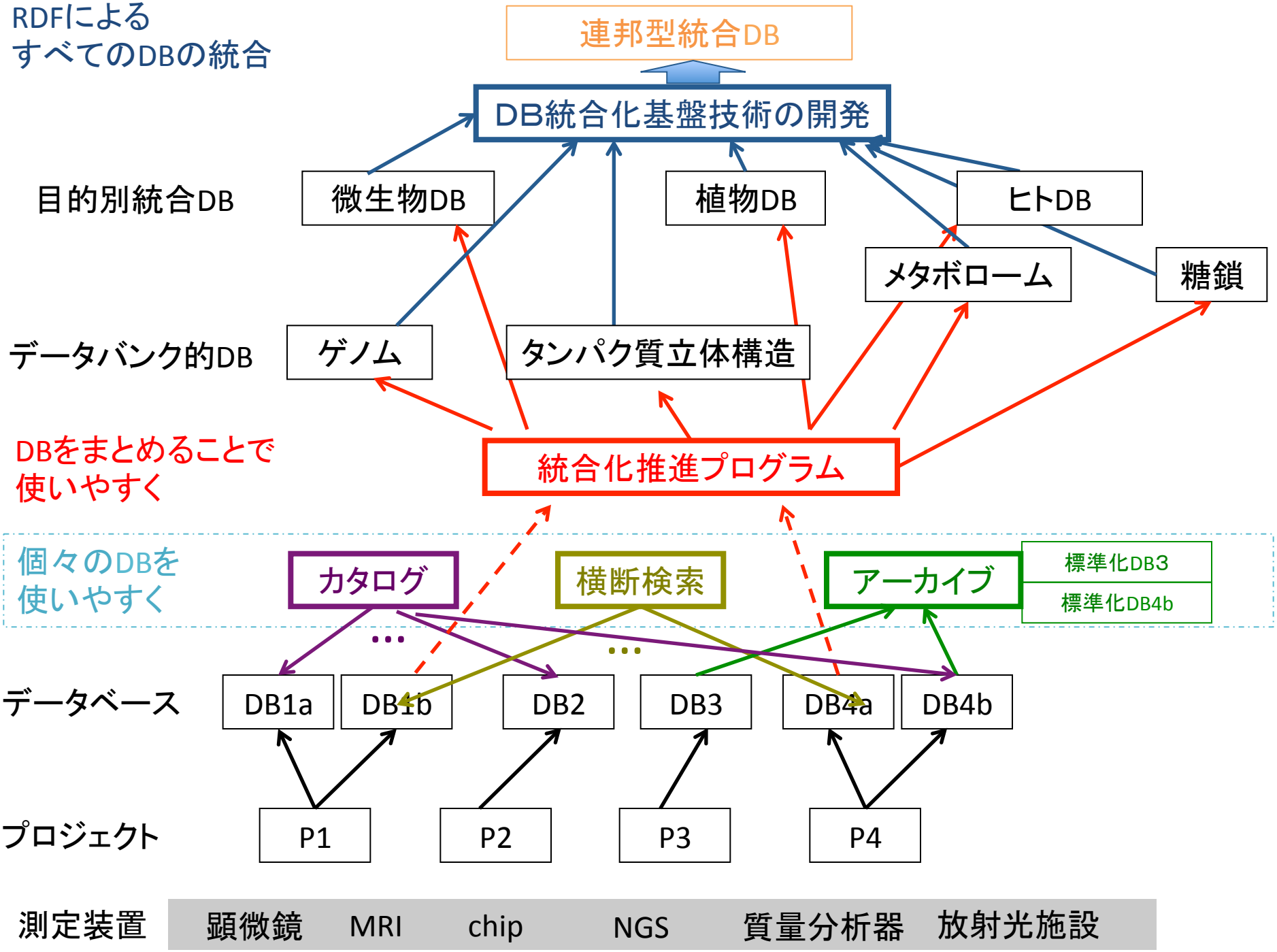
- データベースの数
  - 世界全体:10,000から20,000
  - メジャーなもの（NAR誌のDB特集収録）:約1,600
  - 我が国のDB数(NBDC Integbioカタログ）:約1,000
- データの種類
  - NAR誌での分類：15カテゴリ、40サブカテゴリ
  - Integbioカタログ:生物種、対象、データ種類で分類
    - 生物種：動物、植物、原生生物、菌類、真正細菌、ウイルス
    - 対象：ゲノム、遺伝子、cDNA、多型、タンパク質、酵素、細胞
    - データ種類：配列、構造、発現、相互作用、画像、オントロジー
- ゲノムプロジェクトの数(JGI GOLD)：約7万
- データDB開発国(NAR誌）：約50

ライフサイエンスに係るデータの実態（2）

- データのサイズ
  - 米国NCBI: SRA 4PB, dbGaP 2PB
  - 年率 1.5倍程度の伸び（ムーアの法則と同じ程度）
  - 文献(PuBMed)は2,500万件
- 主要なDBセンター：NCBI, EBI, DDBJ、等
  - ゲノムのデータが中心
  - 欧米のセンターは数百人規模
  - 20～30peta程度のストレージ保有
  - 2020年にはゲノムだけで2EB程度必要との試算あり

我が国のライフサイエンス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）


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
メタボローム
糖鎖
データバンク的DB
ゲノム
タンパク質立体構造
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/DBCLS共同研究の４本柱

1. RDF統合化のための基盤技術開発

高度なRDF技術の実用化

Keyword RDFデータストア、分散検索、セキュリティ、個人ゲノム

2. 統合化支援

統合化推進プログラム等のRDF化をサポート。RDF基盤データの整備

Keyword RDF化、オントロジー開発、国際連携、ハッカソン、文献アノテーション

3. エンドユーザ向けデータベース利用技術開発

大規模RDFデータを活用したアプリケーション

Keyword NGSメタデータ、遺伝子発現リファレンス、大規模塩基配列検索、質問応答、自然言語処理、日本語コンテンツ、学習動画コンテンツ

4. 既存・新規サービス/データベースの運用と拡張

公開サービスやデータベースの持続的な運用

Keyword DBサーバー運用、負荷分散、クラウド、データ更新・バージョン管理

5. 広報活動

公開サービスの広報、教育活動等

Keyword 講習会開催、学会展示会

これからの課題

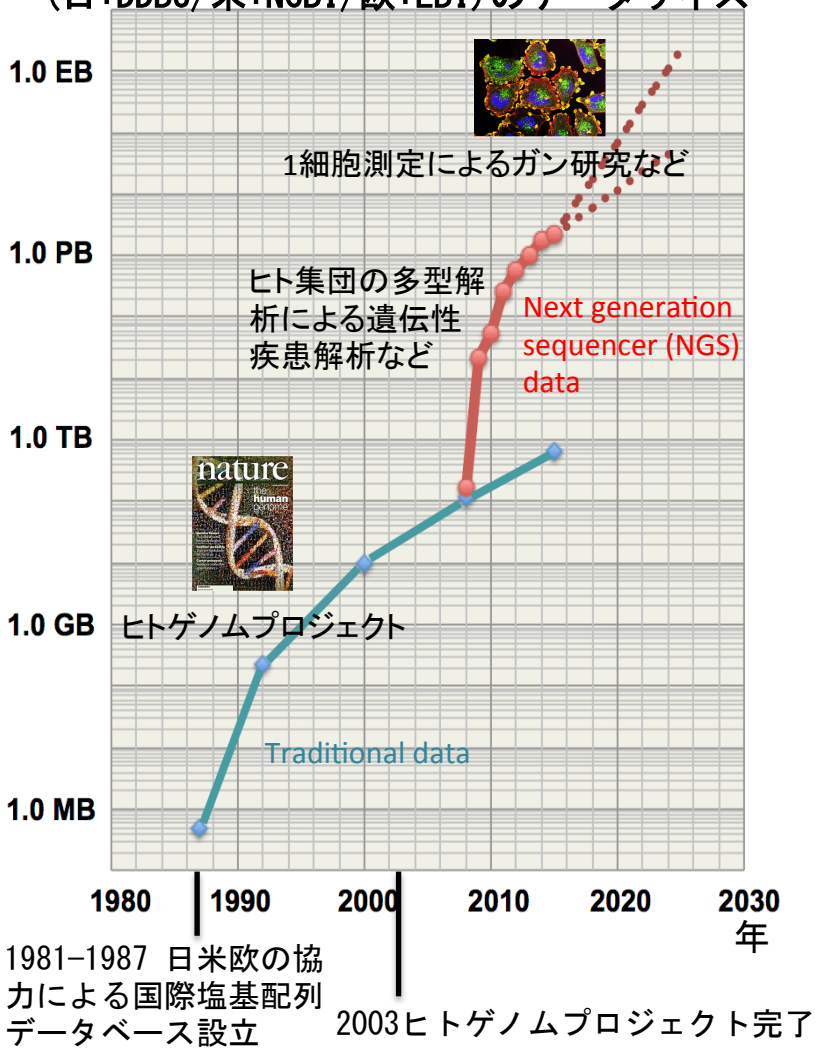
- データ量の増大への対応
- AMEDとの連携、協力
- 応用事例（成功事例）の積み上げ
- DBを活用したデータ解析
- 国際連携
- 人材育成への貢献
- DBセンターのアライアンス、統合

遺伝研スパコン

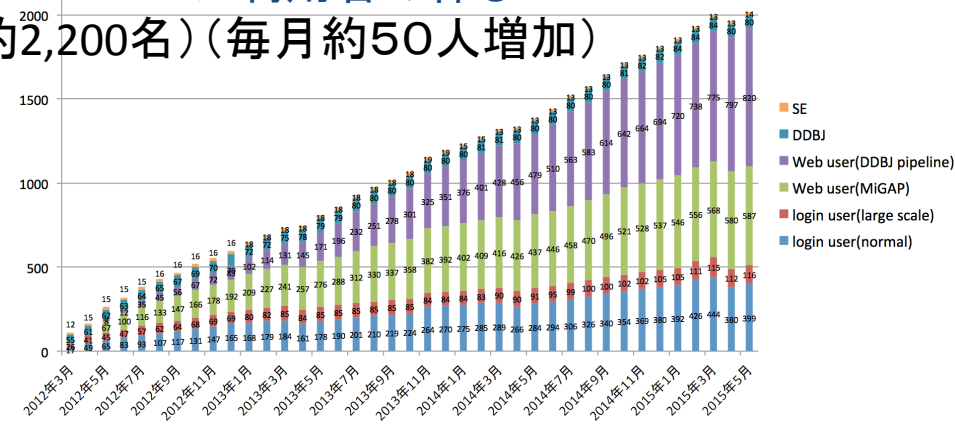
- コア数　11,240 コア
- 演算性能　373 TFLOPS　総メモリ　65.4 TB
- 高速ストレージ　7 PB　バックアップ　5.5 PB

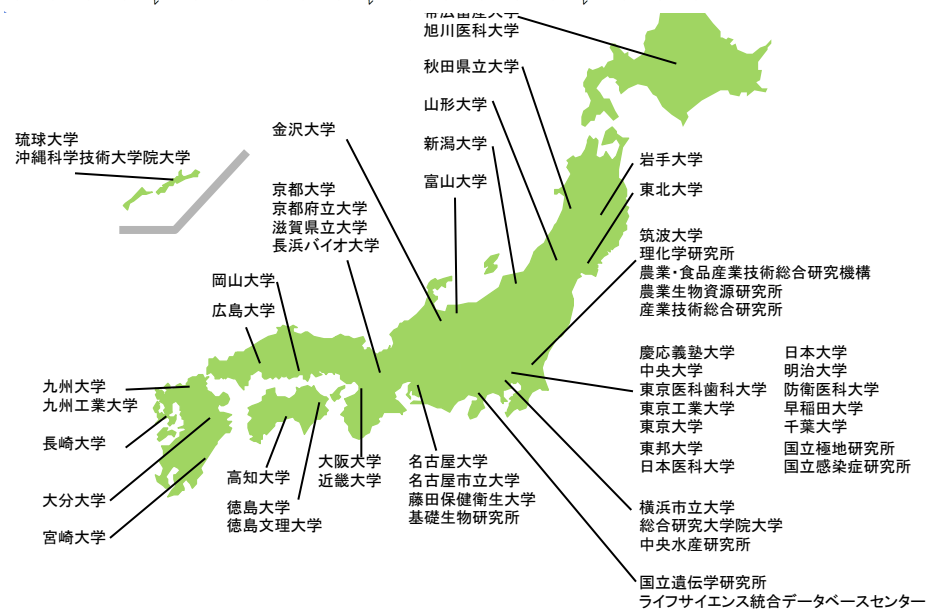
•データ量の爆発的伸び

国際塩基配列データベース
（日:DDBJ/米:NCBI/欧:EBI）のデータサイズ



•DDBJスパコン利用者の伸び
（約2,200名）（毎月約50人増加）





DDBJ利用機関（520機関）の分布

GA4GH(Global Alliance for Genomics and Health)
組織メンバー

347組織（34カ国）

NIH, ELIXIR, Google, Amazon, Illumina、など

日本は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

Global Alliance for Genomics & Health

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（252のデータセットが提供されている）。

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

似たような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。

今回のトーゴーの日シンポジウム

- 基調講演、招待講演
  - オープンサイエンス
  - 人工知能
  - データサイエンス
  - AMED
- 統合DBの応用・活用事例
- H27年度からの統合化推進新規プログラム
- 統合プロジェクトの成果はポスター発表で