バイオサイエンスにおけるデータ発見

国立研究開発法人科学技術振興機構
バイオサイエンスデータベースセンター(NBDC)
八塚　茂

CC BY

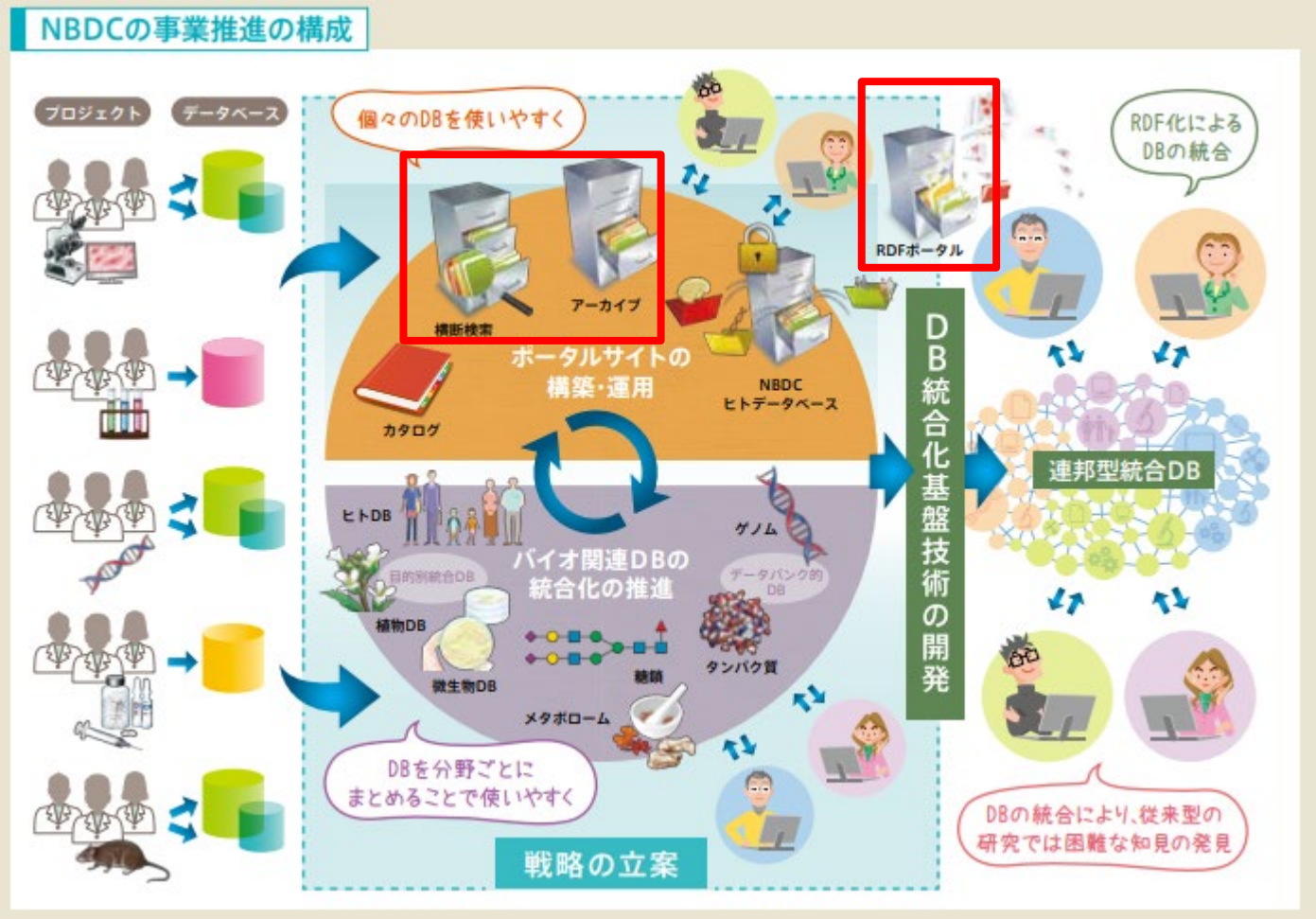
自己紹介

- 八塚　茂（やつづか　しげる）
  - 学部は人文系
  - たまたまSEになる
  - たまたまバイオ系システムの部門に配属される
    - IT&バイオバブルの栄枯盛衰を経験
  - NBDCの担当になる
    - 生命科学系データベースアーカイブの立ち上げから参画
  - 色々あってNBDCの「中の人」になる
  - NBDCの諸サービス＆ITインフラを担当

「バイオサイエンス研究データ流通業」

JST | NBDC

NBDCについて



NBDCパンフレット
https://biosciencedbc.jp/about-us/files/NBDC_pamphlet202006.pdf

バイオサイエンスにおけるデータとは？

- 分類はいろいろ
  - オミクスによる分類
  - データ形式による分類
  - 生物種による分類
  - ・・・

セントラル
ドグマ

オミクスによる分類

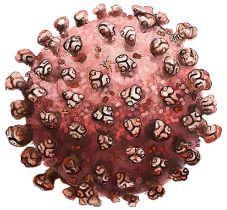
- Genomics
  - 遺伝子やそれを構成するDNAに関するデータ
- Transcriptomics
  - mRNAなど転写産物に関するデータ
- Proteomics
  - タンパク質に関するデータ
- Glycomics
  - 糖鎖に関するデータ
- Metabolomics
  - 代謝物に関するデータ
- 他にもあります・・・

JST | NBDC

データ内容による分類

- 配列（シーケンス）関連データ
  - DNAやアミノ酸の配列
  - クオリティやゲノム・タンパク質上での位置情報も含まれる
- 構造データ
  - タンパク質などの構造（2次元・3次元）
- アノテーションデータ
  - 遺伝子、タンパク質、その他のものの説明
- 文献データ
  - 論文そのものや書誌情報
- 画像データ（動画も）
  - 顕微鏡写真など
- 他にも様々な種類のデータが・・・

生物種による分類

- ヒト
- マウス
- 植物
- 細菌
- ウィルス 
- 他にも多数・・・

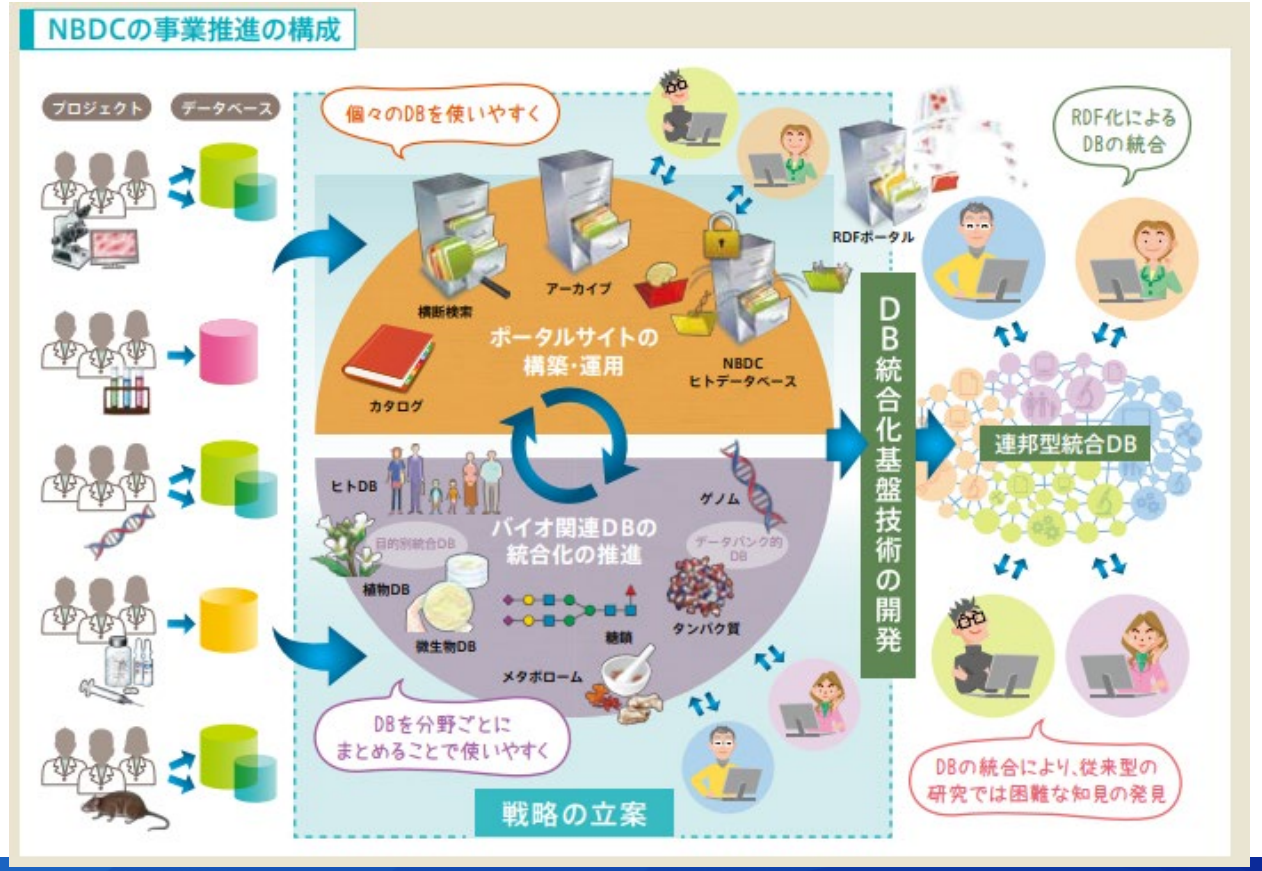
本スライドの画像は全て
© 2016 DBCLS TogoTV / CC-BY-4.0

データはどこにあるか？

- バイオサイエンス分野では、大量のデータがデータベース・リポジトリで公開されている
  - データベース：主に運営者が作成したデータを公開
  - リポジトリ：運営者以外の研究者からデータを収集して公開
    - INSDC(International Nucleotide Sequence Database Collaboration)
      - NCBI（米国）
      - EBI（欧州）
      - DDBJ（日本・国立遺伝学研究所）
    - wwPDB(Worldwide Protein Data Bank)：タンパク質構造の世界的リポジトリネットワーク(since 1971!)
      - PDBj（日本・大阪大学蛋白質研究所）など
    - Uniprot：アミノ酸情報の大規模リポジトリ
    - jPOST（日本）：計測機器からの生データを中心としたプロテオームリポジトリ
    - …
    - 生命科学系データベースアーカイブ：上記のリポジトリでは扱いきれないデータを収集
- 論文にはこれらのデータベース・リポジトリにおけるデータのPID（アクセション番号）を記載することが当たり前になっている

公開データが増えたのはいいけれど・・・

- データベース・リポジトリが乱立
  - どこに・どんなデータがあるかわからなくなる
  - 予算がなくなったらデータも廃棄？
- データ（とメタデータ）を様々な観点から「統合」する必要がある



NBDCパンフレット
https://biosciencedbc.jp/about-us/files/NBDC_pamphlet202006.pdf

データの発見方法（その１）

- まずはデータベース・リポジトリを探す
  - Integbioデータベースカタログ
    - https://integbio.jp/dbcatalog/?lang=ja

データの発見方法（その１）

- 各データベース・リポジトリで検索する
  - DDBJで遺伝子情報をキーワードを使って検索する例
    http://ddbj.nig.ac.jp/arsa/

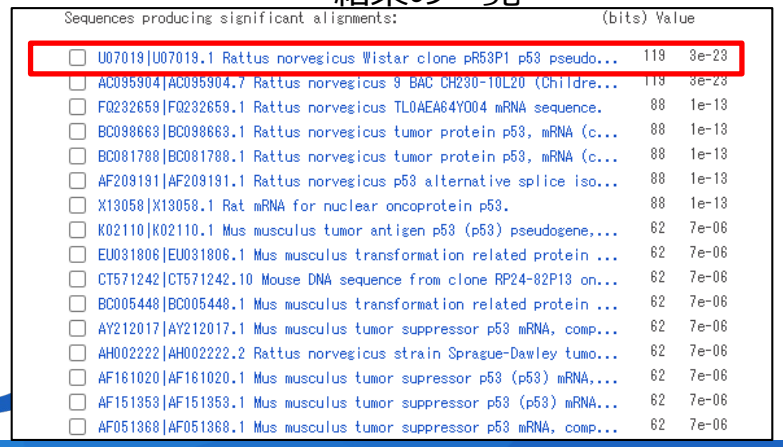
遺伝子情報

```
LOCUS       RNU07019                1273 bp    DNA     linear   ROD 02-FEB-1996
DEFINITION  Rattus norvegicus Wistar clone pR53P1 p53 pseudogene.
ACCESSION   U07019
VERSION     U07019.1
KEYWORDS    .
SOURCE      Rattus norvegicus (Norway rat)
  ORGANISM  Rattus norvegicus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha;
            Muroidea; Muridae; Murinae; Rattus.
REFERENCE   1  (bases 1 to 1273)
  AUTHORS   Lin,Y. and Chan,S.H.
  TITLE     Cloning and characterization of two processed p53 pseudogenes from
            the rat genome
  JOURNAL   Gene 156 (2), 183-189 (1995)
   PUBMED   7758955
REFERENCE   2  (bases 1 to 1273)
  AUTHORS   Lin,Y.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1994) Yue Lin, Department of Microbiology,
            National University of Singapore, Lower Kent Ridge Road, Singapore
            0511, Singapore
FEATURES             Location/Qualifiers
     source          1..1273
                     /organism="Rattus norvegicus"
                     /mol_type="genomic DNA"
                     /strain="Wistar"
                     /db_xref="taxon:10116"
                     /clone="pR53P1"
                     /sex="male"
                     /cell_type="hepatocyte"
                     /tissue_type="liver"
     exon            <1..14
                     /note="corresponds to a part of exon 1 of the rat p53
                     gene"
                     /pseudo
     exon            15..97
                     /note="corresponds to exon 2 of the rat p53 gene"
                     /pseudo
     gene            24..1245
                     /gene="p53"
     CDS             24..1245
                     /gene="p53"
                     /pseudo
```

データの発見方法（その１）

- 各データベース・リポジトリで検索する
  - DDBJで遺伝子情報をDNA配列の類似により検索する例
    http://ddbj.nig.ac.jp/blast/

結果の一覧



遺伝子情報

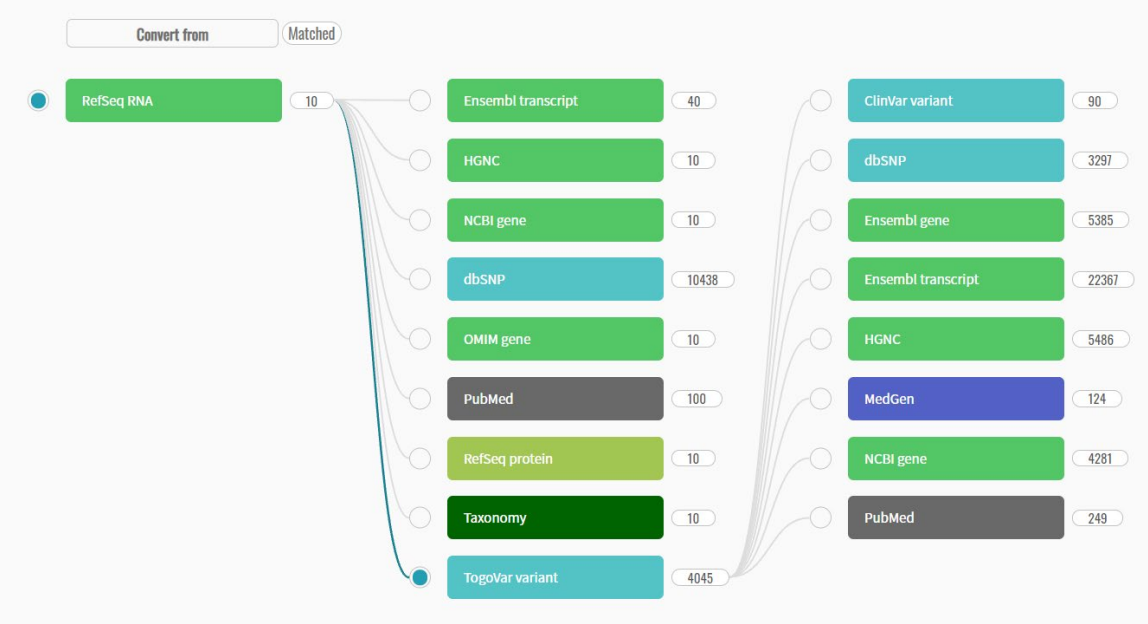
```
LOCUS       RNU07019                1273 bp    DNA     linear   ROD 02-FEB-1996
DEFINITION  Rattus norvegicus Wistar clone pR53P1 p53 pseudogene.
ACCESSION   U07019
VERSION     U07019.1
KEYWORDS    .
SOURCE      Rattus norvegicus (Norway rat)
  ORGANISM  Rattus norvegicus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha;
            Muroidea; Muridae; Murinae; Rattus.
REFERENCE   1  (bases 1 to 1273)
  AUTHORS   Lin,Y. and Chan,S.H.
  TITLE     Cloning and characterization of two processed p53 pseudogenes from
            the rat genome
  JOURNAL   Gene 156 (2), 183-189 (1995)
   PUBMED   7758955
REFERENCE   2  (bases 1 to 1273)
  AUTHORS   Lin,Y.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1994) Yue Lin, Department of Microbiology,
            National University of Singapore, Lower Kent Ridge Road, Singapore
            0511, Singapore
FEATURES             Location/Qualifiers
     source          1..1273
                     /organism="Rattus norvegicus"
                     /mol_type="genomic DNA"
                     /strain="Wistar"
                     /db_xref="taxon:10116"
                     /clone="pR53P1"
                     /sex="male"
                     /cell_type="hepatocyte"
                     /tissue_type="liver"
     exon            <1..14
                     /note="corresponds to a part of exon 1 of the rat p53
                     gene"
                     /pseudo
     exon            15..97
                     /note="corresponds to exon 2 of the rat p53 gene"
                     /pseudo
     gene            24..1245
                     /gene="p53"
     CDS             24..1245
                     /gene="p53"
                     /pseudo
```

データの発見方法（その２）

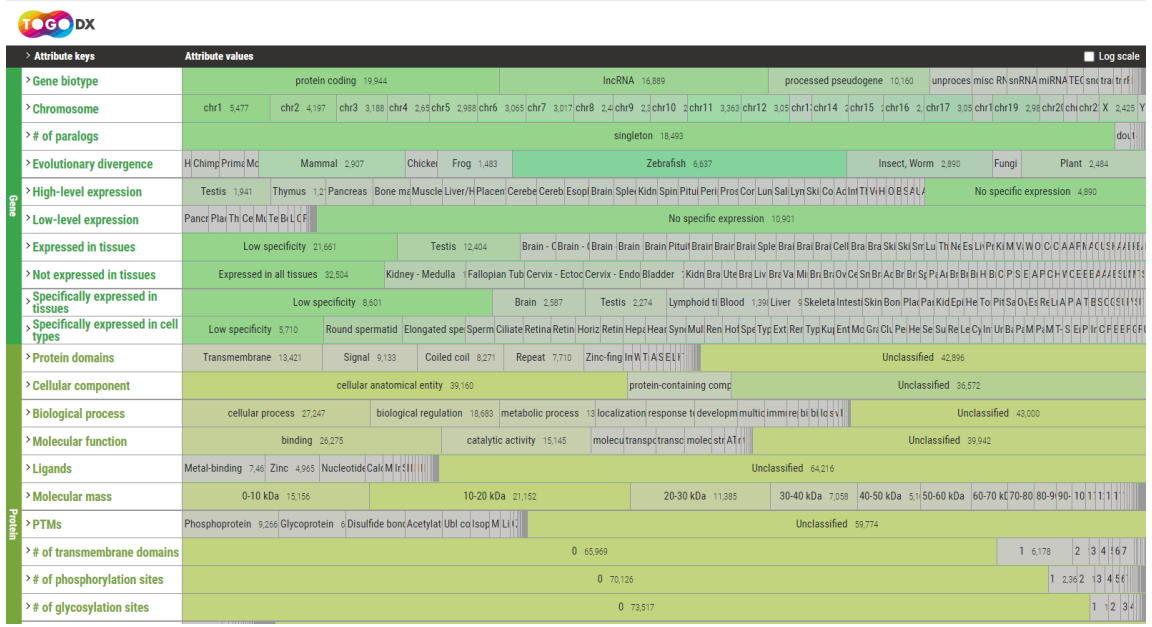
- 複数のデータベース・リポジトリを一括検索
  - 生命科学データベース横断検索
    - https://dbsearch.biosciencedbc.jp/
    - 700以上のデータベースからキーワードを一括して検索できる

データの発見方法（その３）

- データIDの関係をたどる
  - TogoID
    https://togoid.dbcls.jp/



- 各データベース・リポジトリのデータ内訳やそれらの関係を探る
  - TogoDX/human
    https://togodx.dbcls.jp/human/



講習会もあります

統合データベース講習会：AJACS　https://biosciencedbc.jp/event/ajacs/

「統合データベース講習会：AJACS（あじゃっくす）」は、生命科学系のデータベースやツールの使い方、データベースを統合する活動を紹介する講習会です。データ解析の“入り口”を学び、研究に役立つデータベースを知ることができます。これまでに全国各地やオンラインにて80回以上開催しています。受講は無料です。お気軽にご参加ください。

一覧

2021年度

2022年1月20日　タンパク質のデータベースを活用する（AJACSオンライン10）
オンライン

2021年12月16日　疾患に関する多型データを解析する（AJACSオンライン９）
オンライン

2021年8月19日　遺伝子発現データを解釈しゲノム編集に応用する（AJACSオンライン８）
オンライン

2021年7月15日　BLAST検索とゲノムブラウザ、NGSデータベースを活用する（AJACSオンライン７）

JST | NBDC

データ発見を促すために

- データをただ公開するだけでは不十分
  - FAIR原則
- データベース・リポジトリの運営には多様な専門家の協業が不可欠
  - 企画
    - 仕様
    - メタデータ
  - 運用管理
  - ライセンス・契約
  - アノテーション/キュレーション
    - 形式のチェック（データ間の整合性、用語・表記などの統一）
    - 内容のレビュー
  - データの加工
  - システムインフラの維持

赤字のタスクはあまり認知されていないが、運営には不可欠