

# 生命科学分野におけるデータ(ベース)の統合と、データ駆動型研究を行うためのリソース開発

大学共同利用機関法人 情報・システム研究機構

データサイエンス共同利用基盤施設

ライフサイエンス統合データベースセンター (DBCLS)

小野 浩雅

[hono@dbcls.rois.ac.jp](mailto:hono@dbcls.rois.ac.jp)

2023年6月19日(月)

Japan Open Science Summit 2023 D1 研究基盤プラットフォーム再考 - FAIRの「I」

## 内容

- 生命科学分野におけるデータ(ベース)の統合と、データ駆動型研究を行うためのリソース開発
  - **TogoID** を使って生命科学系データベース間のつながりを探索的に確認しながらID変換を行う
  - **TogoDX/human** を使って統合されたデータを探索・俯瞰・抽出する

# 自己紹介

- 小野 浩雅
  - [TogoTV](#)の運営・編集
    - 生命科学分野の有用なDBやウェブツールの活用法を動画で紹介するウェブサイト
  - [RefEx](#)の開発
    - 遺伝子発現解析の基準となる各遺伝子の遺伝子発現量を簡単に検索、閲覧できるウェブツール
  - [TogoID](#)の開発
    - 生命科学系データベース間のつながりを探索的に確認しながらID変換を行うことができるウェブツール
  - [Twitter@h\\_ono](#)

# 生命科学研究はデータベース作り

- さまざまな実験で得られたデータは、論文投稿時などに公共データベース上に登録し、その後誰でも参照可能になるようにすることが義務付けられていることが多い
- 公共データベースには多種多様なデータが日々大量に登録、蓄積され続けている
- データをうまく活用すれば、多くのメリット(がありそうなことは皆感じている)
  - 予備実験をせずに済む
  - 自分の実験結果を支持する知見が得られる
  - 多角的な視点からの新たな仮説生成
- 似たようなものがいくつもありどれを使ってよいかわからない
  - 使えそうなものが見つかっても、実際に使うのは大変/使えない場合も多い

# データ(ベース)を統合的に組み合わせて、データ駆動型研究を行う

- 生命科学の目的は様々な要素が相互作用している複雑なシステムの理解
  - 多種多様なデータの「統合」(÷相互運用性) が必須
- 課題
  - データベースごとに異なるインターフェース
  - データベース間を繋ぐリンク情報の欠如
  - データベースごとに異なる出力形式
- 横断的にDBを使うには手間がかかりすぎる
  - 個別にデータベースを解析して組み合わせるための「前処理」が作業の8割

# 課題を解決するための取り組み

- BioHackathon
  - 生命科学分野のデータベース統合の技術基盤の確立を目的として、年1回日本各地で開催している国際開発者会議
  - BioHackathon 2015@長崎 にてFAIR原則の内容に関する議論が行われた
    - Wilkinson MD et. al.: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data., [doi:10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18) (2016)



## FAIR原則と知識グラフ

- Findable、Accessible、**Interoperable**、Reusable
- それぞれのデータが使いやすくなった（ちょっとずつなってる）
  - 幅広いデータが統合できる時代になった
- 知識グラフによる生命科学分野のデータベース統合
  - 複数のデータセットが共通のURIで連結される
  - 各データとそのつながりの意味が表現できる、すなわち、知識が表現できる

TogolD を使って生命科学系データベース間のつながりを  
探索的に確認しながらID変換を行う

# 生命科学分野におけるID変換の必要性

- 様々なデータベース (DB) を活用するには異なる ID 間のリンクが重要
  - 使いたい解析ツールが手元にある ID では使えない
  - 等価なものに対する ID 間で変換したい
    - 例: NCBI Gene ID  Ensembl Gene ID
  - 関連する情報を取得したい
    - 遺伝子が関与する疾患、化合物が関与するパスウェイ etc.
- 既存の ID 変換サービスの問題点
  - 対象としている DB の範囲が限られる
  - 大元の DB の更新に追従していない
  - プログラムから利用できるAPI が提供されていない

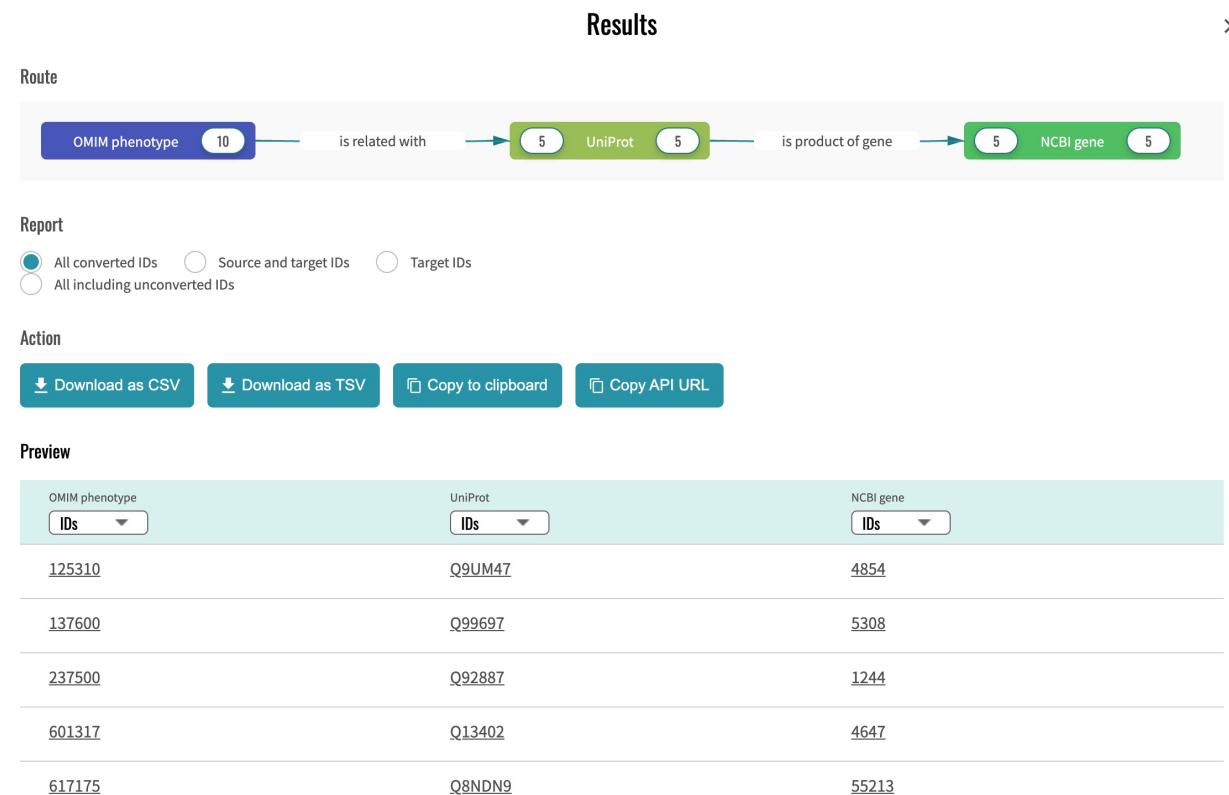
# TogolD

<https://togoid/dbcls.jp/>

- 生命科学系データベース間のつながりを探索的に確認しながらID変換を行うウェブアプリケーション
- Shuya Ikeda, Hiromasa Ono et. al.: TogolD: an exploratory ID converter to bridge biological datasets, Bioinformatics, doi:[10.1093/bioinformatics/btac491](https://doi.org/10.1093/bioinformatics/btac491) (2022)



- 64のデータベースに由来する 89 のデータセットのペアを収載
  - 遺伝子から化合物、疾患等までを網羅
  - 毎週の定期更新
- 生物学的意味を持つID間の対応関係を独自のオントロジーとして整備
- GitHubレポジトリは公開
  - 誰でも自由に参照したり新規データセットペアを提案できる



**TogoDX/human を使って統合されたデータを探索・俯瞰・抽出する**

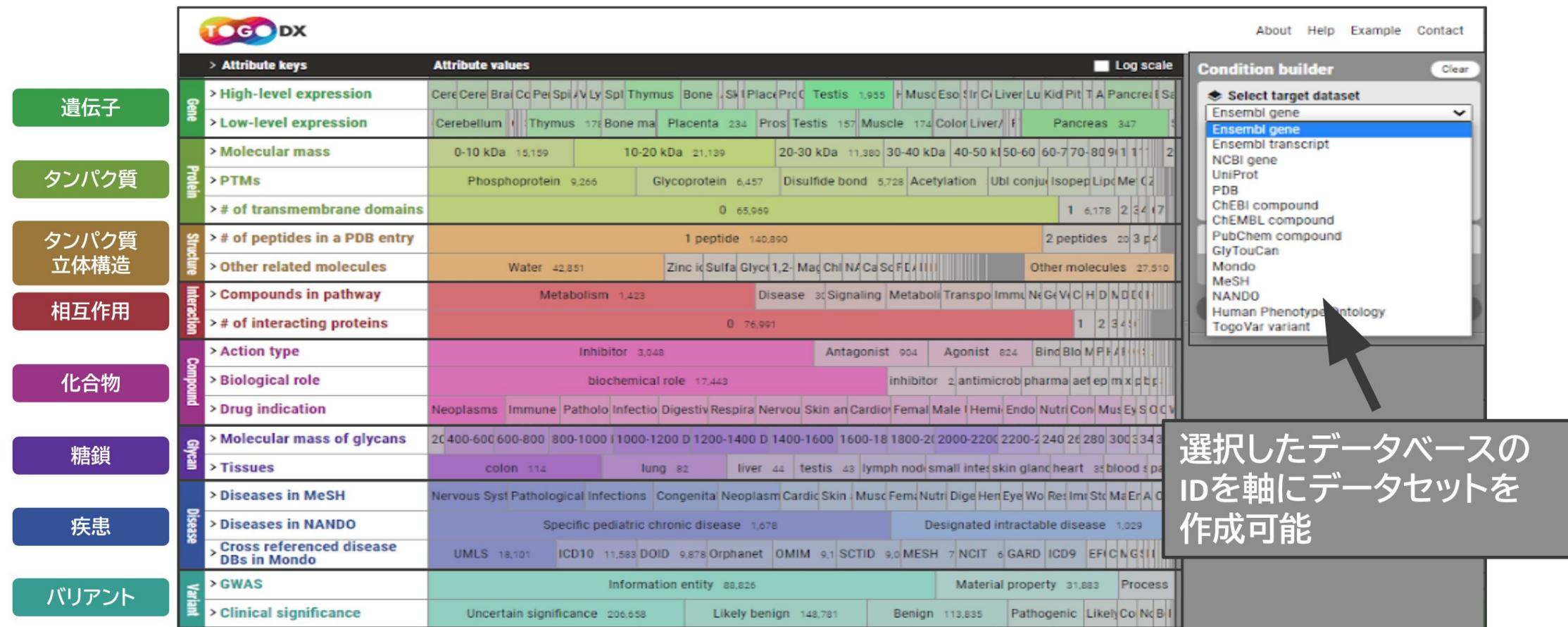
# データを統合してTogoDXというアプリケーションを作った

- データは統合できるが、それをどう理解・探索・解析するか？
- 統合されたデータに適したインターフェースが必要
  - ヒトデータはデータ量も膨大で多岐にわたっている
  - 誰も本当に統合・俯瞰したことはないんじゃないか
    - それができる TogoDX/human を作った
    - データを入れ替えれば、マウスや〇〇など他のテーマでも流用できる

## TogoDX/human <https://togodx.dbcls.jp/human/>

- 国内外のデータベースから収集・統合した、ヒトに関する遺伝子、タンパク質、化合物、疾患などの情報をワンストップで探索することができるサービス
- TogoDX(Data eXplorer) は、生命科学分野における様々なデータベースを統合的に探索し、俯瞰するためのフレームワーク
  - 膨大な情報を多様な属性 (attribute) によって柔軟に絞り込み、必要な情報を抽出できる新しい仕組み
- TogoDX/humanでは、21個のデータベースに由来する64個の attribute が利用可能

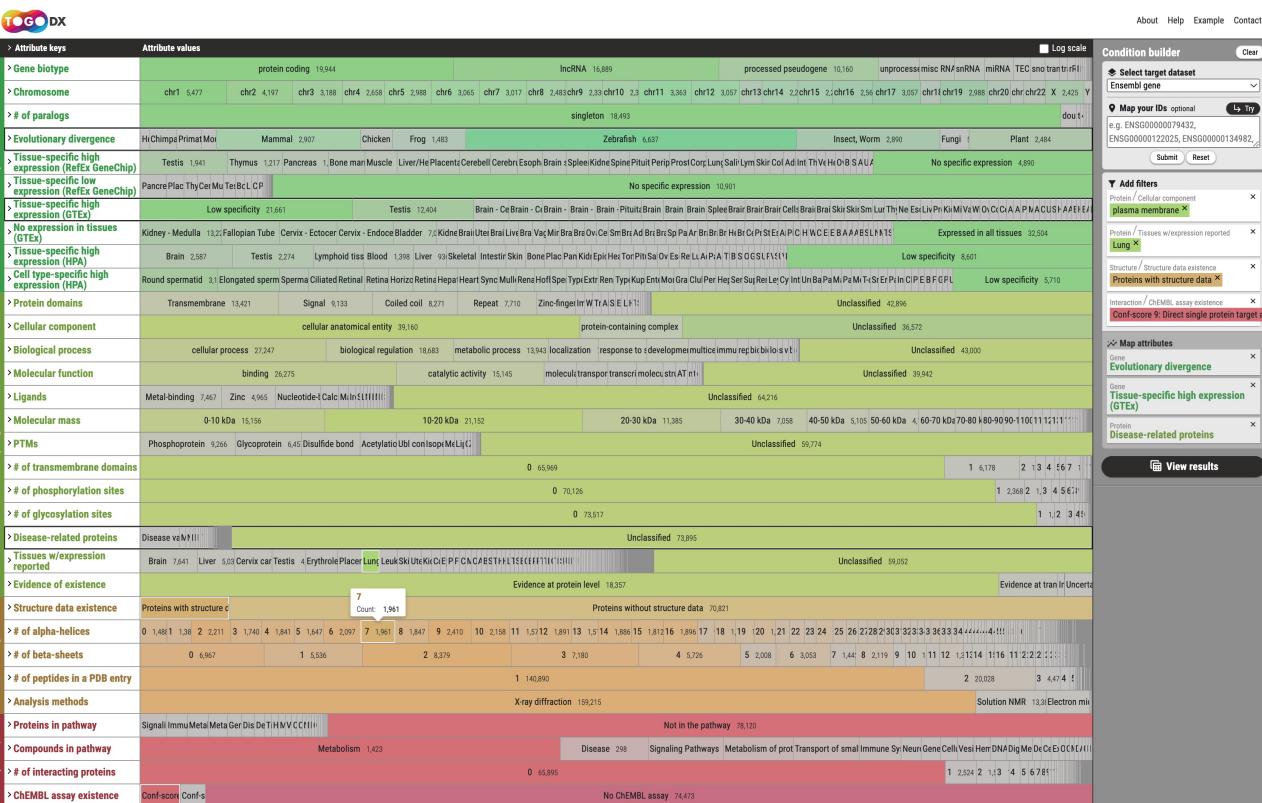
# TogoDX で統合されたデータを俯瞰する



# 検索例

- 肺でタンパク質として発現が確認され、
- 細胞膜表面に局在し、
- タンパク質立体構造が明らかになっており、
- 対応する医薬品が開発されている

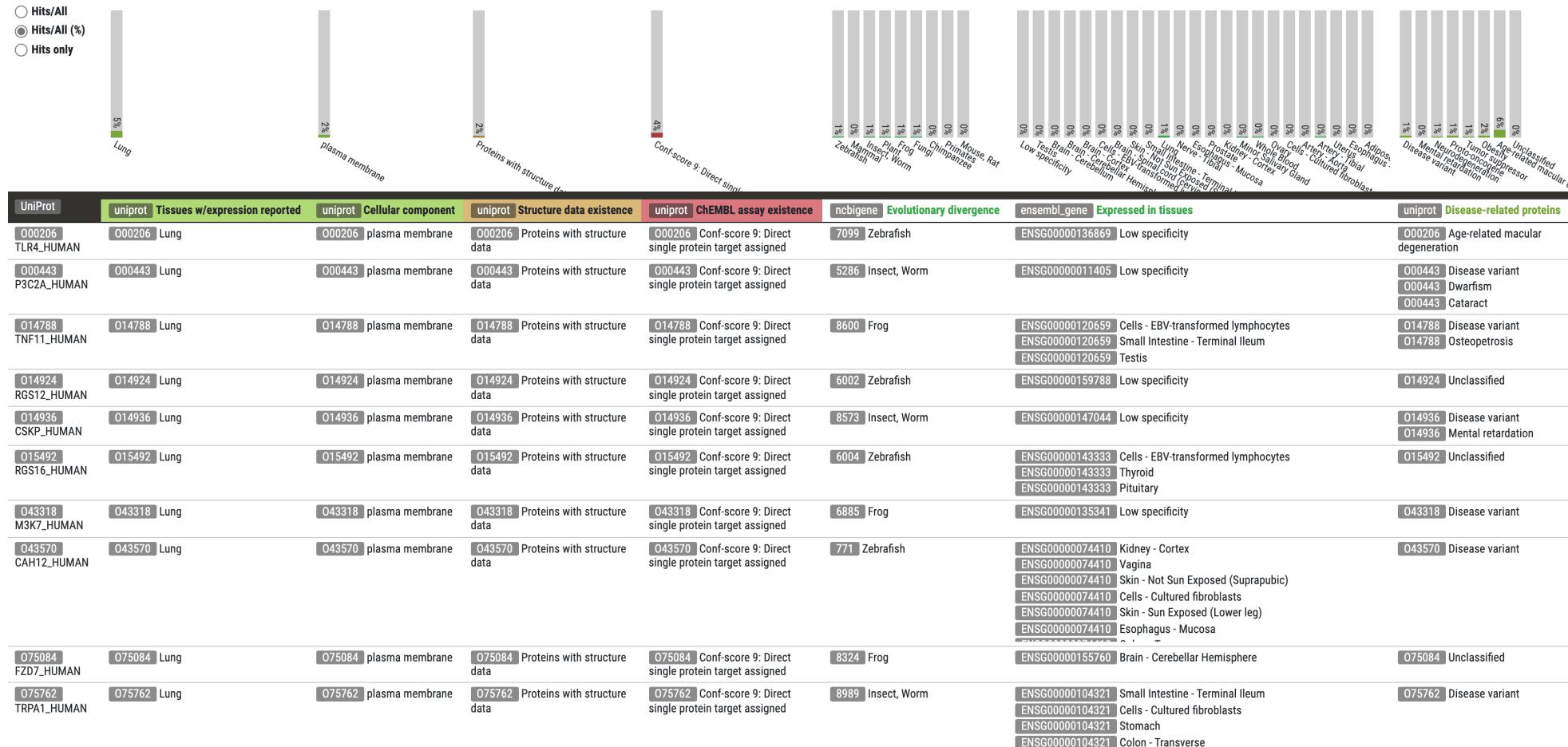
ヒトのタンパク質の一覧を取得する



# 選択した条件を全て満たすIDのリストを抽出できる



# 抽出したリストにおける他の属性の分布を調べる



← Return

Saved conditions

U	Tiss...	C...	Str...	Ch...	Ev...	E...	Di...	X
131 / 131	Complete							
	TSV							
U...	Tissues ...	Cellu...	Structu...	ChEMB...				X
131 / 131	Complete							
	TSV							

# 自分の持つIDリストの偏りをTogoDX/humanのデータにマッピングして調べる

Two screenshots of the TogoDX/human interface showing the mapping of a local ID list against the Human dataset.

**Screenshot 1 (Left):** Shows the search results for a query related to "proteins coding 15944". The results are filtered by "Selected target dataset: UniProt" and "Map year: 14-19". The results include various gene and protein entries, such as "protein coding 15944", "chromosome 1", "dof of paralogues", and "Evolutionary divergence". A "View results" button is present at the bottom right.

**Screenshot 2 (Right):** Shows the search results for a query related to "proteins coding 15944". The results are also filtered by "Selected target dataset: UniProt" and "Map year: 14-19". The results are identical to the first screenshot, showing the same gene and protein entries. A "View results" button is also present at the bottom right.

## まとめ

- TogoID を使って生命科学系データベース間のつながりを探索的に確認しながらID変換を行う
  - **多種多様なIDを統一的に利用できるよう整備することで、生命科学データの「相互運用性」を高めるよう取り組んでいる**
- TogoDX/human を使って統合されたデータを探索・俯瞰・抽出する
  - 「相互運用性」を高めることによって**高度に統合されたデータベースを探索・俯瞰することで新たな知識を抽出できる(データ駆動型生命科学研究の)仕組み**ができることがあるつつある