

# ライフサイエンス統合データベースセンター(DBCLS)が 提案する研究データの見つけ方

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)

2022年6月6日(月)

Japan Open Science Summit 2022 E1 研究データの「新しい見つけ方」を考える

## 内容

- ライフサイエンス統合データベースセンター(DBCLS)が提案する研究データの見つけ方
  - TogoTV を使って、新たな研究データ(ベース)を発見・学習・活用する
  - 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・編集者
    - 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](#)の開発責任者
    - 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
  - [TogоАD](#)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](#)

# TogoTV を使って、新たな研究データ(ベース)を発見・学習・活用する

# 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



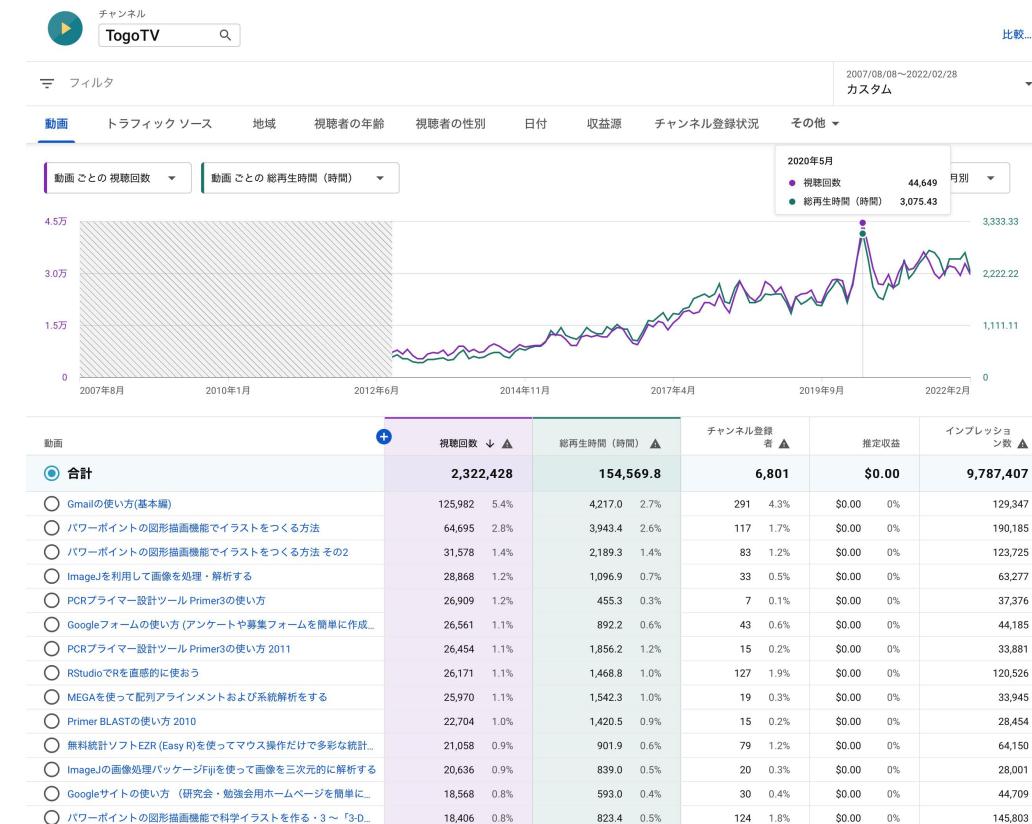
# TogoTV

- <https://togotv.dbcls.jp/> 
- 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
  - データベースやツールの動画マニュアル
  - 講演や講習会動画・資料
  - イラスト
- 2007年8月サービススタート
- 2020年11月リニューアル



# 動画マニュアル

- ウェブサイトへのアクセスの仕方から結果の解釈まで、操作の一挙手一投足がわかる
- 各動画はYouTubeに上がっており、環境に応じた解像度、倍速表示等で快適に閲覧可能
- 1,980本を超える動画が公開されており、2,300,000回以上再生(2022年2月末現在)
- コロナ禍の自粛時に過去最高の視聴数
- YouTubeのチャンネル登録をすると新規公開の通知が来て便利です



# 個別動画ページ

- 見どころダイジェスト
  - YouTubeの埋め込みプレーヤなので、おなじみの操作感
  - 動画の概要を示す「見出し」から視聴箇所に移動可能
  - 動画ファイルはダウンロード可能
- 各動画には、DOI (Digital Object Identifier) が付与
  - 恒久的な URL として利用可能
  - e-learning教材として日本初



# スキル別コース

- ある目的に対して、「この順で動画をみていくと、こういうスキルを獲得できる」というような体系的な動画リスト集

The screenshot shows the TOGO TV website interface. At the top, there is a navigation bar with links for DBCLS, Research, Services, Contact, About, and a search bar. Below the navigation bar, there is a header for 'TOGO TV' with links to the homepage, information about TogoTV, video search, image search, training material search, contact, and a search bar labeled 'キーワードから動画を探す'.

The main content area is organized into several sections:

- スキル別コース (Skill-based Courses)**
  - 文章の執筆に役立つツール (Tools for improving writing skills)**: 1時間 48分. Includes videos on Google Drive, inMeXes, Overleaf, Grammarly, and diff.
  - 疾患に関する変異や遺伝子発現の情報を調べる (Investigating genetic variants and gene expression information related to diseases)**: 1時間 56分. Includes videos on TogoVar, MGEND, NCBI dbSNP, UCSC Genome Browser, and Ingenuity Variant Analysis.
  - 図表を作成する (Creating charts)**: 1時間 22分. Includes videos on Microsoft PowerPoint, Google Slides, and BioRender.
  - 公共の遺伝子発現データの検索や解析を行う (Searching and analyzing public gene expression data)**: 2時間 15分. Includes videos on NCBI GEO, AOE, and other tools.

# 公共の遺伝子発現データの検索や解析を行う

- 遺伝子発現データは、生命科学研究の中でも特に基本的で多くのDBがある
- NCBI, EBI, DDBJなどの研究センターがデータレポジトリを運営
- 個別プロジェクトからDB化され、自由に利用できるものも多い



# 疾患に関するバリアントや遺伝子発現の情報を調べる

- 遺伝統計学、臨床遺伝学の分野については、利用者の方からのリクエストも多く、ここ数年、取り上げているDB、ウェブツールが増加
- NCBI dbSNP、UCSC Genome Browser、ClinVar、gnomAD、COSMIC、GWAS Catalogなどの国際的なコンソーシアムで開発・運用されているDB
- TogoVar、MGeND、iMETHYL、jMorpなどの国内で開発・運用されているDB

病患に関するバリアントや遺伝子発現の情報を調べる 2時間21分



## 現在のラインナップ例 (今後さらに充実させていきます)

- 文献の検索や管理、情報収集に役立つツール
- ゲノムブラウザを使ってゲノム配列に関連する情報を検索・取得・可視化する
- 公共の遺伝子発現データの検索や解析を行う
- 疾患に関連する変異や遺伝子発現の情報を調べる
- 図表を作成する
- 文章の執筆に役立つツール

# 講演・講習会

- キーワードから、「講演」や「講習会」を簡単に検索可能
- 受講生の復習のみならず、初学者の学習教材として活用できます

The screenshot shows the TOGO TV website interface. At the top, there is a navigation bar with links for DBCLS, Research, Services, Contact, and About. Below the navigation is the TOGO TV logo. A search bar at the top right contains the query 'PubMed'. The main content area displays a search result page titled '「PubMed」の検索結果 57件'. The results are listed in a grid format, each entry showing a thumbnail, title, date, duration, and a brief description. The results include various lectures and workshops related to PubMed, such as 'TOGO TV 文献検索／論文執筆支援 (PubMed, Allie, Caiii, inMeXesなど)' and 'TOGO TV DBCLS で提供している 文献情報サービスほか'.

# 塩基配列解析に関する基礎知識(遺伝子IDとそのデータベース)とゲノム編集について

- 塩基配列解析のためのデータベース・ウェブツール @ AJACSオンライン2
  - CRISPRdirectの開発者が自ら解説
    - 入力した塩基配列に対してCRISPR-Cas9システムのガイドRNAを設計することができるツール
- ウェブツールを使ってゲノム編集の標的サイトを検索する @ AJACSオンライン8
  - 適切なゲノム編集を行うための標的検索とその考え方について学びます。また、さまざまな目的に特化した標的選定やゲノム編集後の解析に活用できるツールについても紹介

## 次世代シーケンス(NGS)データ解析に必要な基礎知識とリテラシーを学ぶ

- NGSデータから新たな知識を導出するためのデータ解析リテラシー @ AJACS浜松
  - NGSデータを解析するための基礎的な考え方・知識と、データ解析プロセスをどう設計・実践していくかの技術を学びます。ソフトウェアの使い方の詳細な解説よりも、実験系研究者が独学していくために必要なことに焦点を絞っています。
  - 講義資料: [NGSデータから新たな知識を導出するためのデータ解析リテラシー](#)

# 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(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」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「前処理」が作業の8割

## 課題を解決するための取り組み、技術

- FAIR原則
  - 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

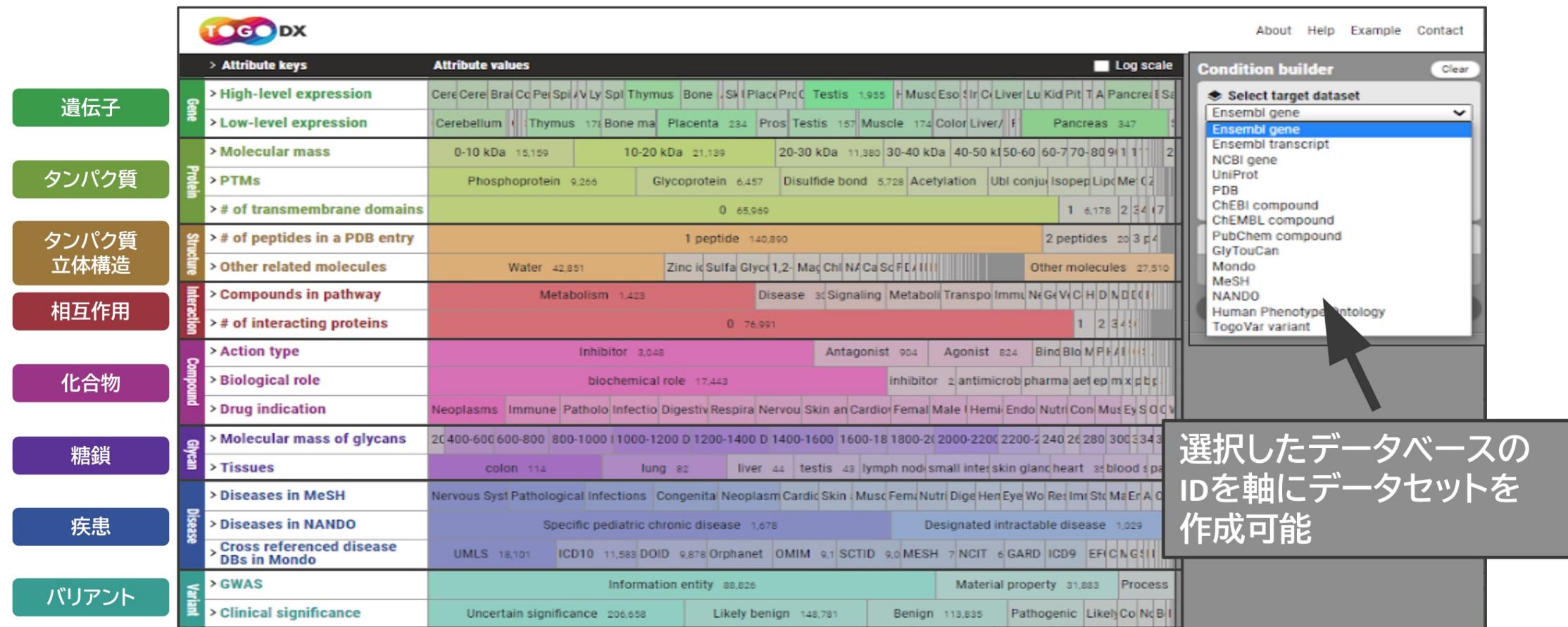
# データを統合してTogoDXというアプリケーションを作った

- 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です。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50個の attribute が利用可能

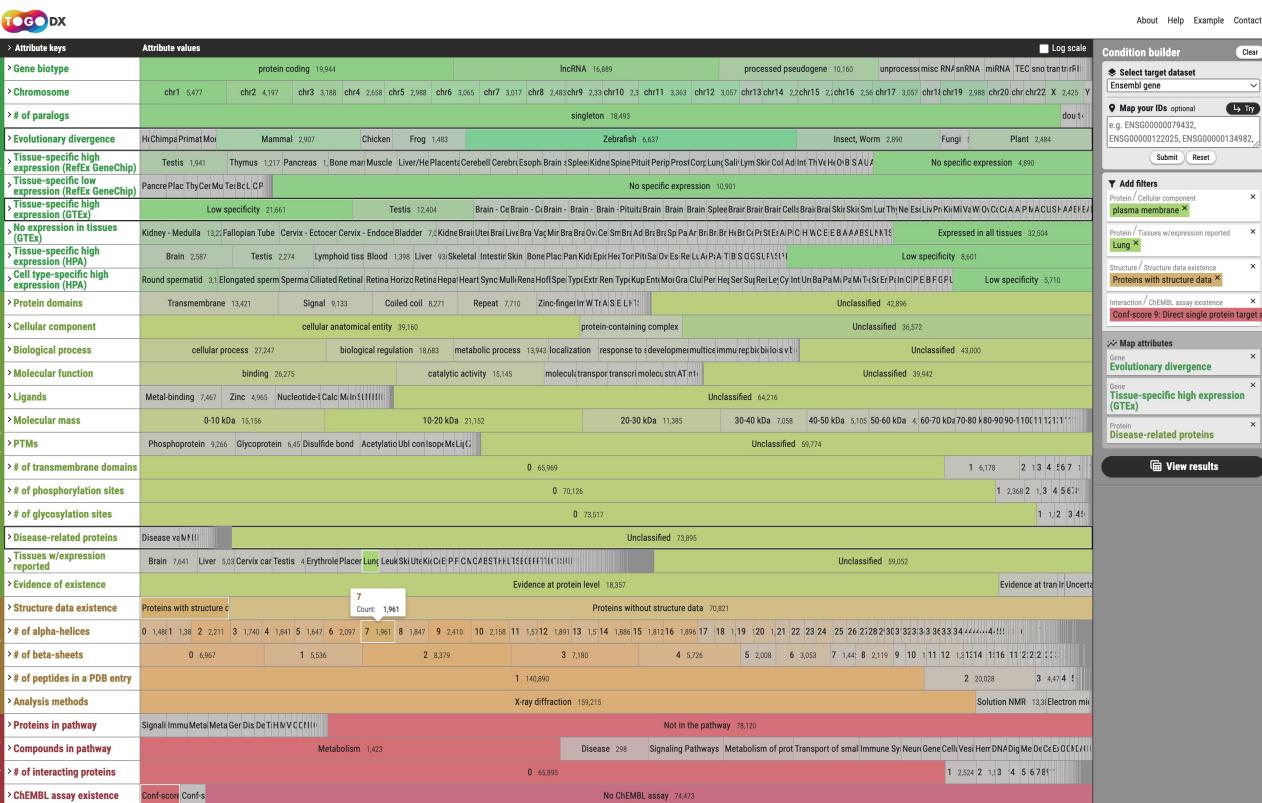
# TogoDX で統合されたデータを俯瞰する



# 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

○ Hits/All  
 Hits/All (%)  
 Hits only

UniProt	uniprot Tissues w/expression reported	uniprot Cellular component	uniprot Structure data existence	uniprot ChEMBL assay existence
000206 TLR4_HUMAN	000206 Lung	000206 plasma membrane	000206 Proteins with structure data	000206 Conf-score 9: Direct single protein target assigned
000443 P3C2A_HUMAN	000443 Lung	000443 plasma membrane	000443 Proteins with structure data	000443 Conf-score 9: Direct single protein target assigned
014788 TNF11_HUMAN	014788 Lung	014788 plasma membrane	014788 Proteins with structure data	014788 Conf-score 9: Direct single protein target assigned
014924 RGS12_HUMAN	014924 Lung	014924 plasma membrane	014924 Proteins with structure data	014924 Conf-score 9: Direct single protein target assigned
014936 CSKP_HUMAN	014936 Lung	014936 plasma membrane	014936 Proteins with structure data	014936 Conf-score 9: Direct single protein target assigned
015492 RGS16_HUMAN	015492 Lung	015492 plasma membrane	015492 Proteins with structure data	015492 Conf-score 9: Direct single protein target assigned
043318 M3K7_HUMAN	043318 Lung	043318 plasma membrane	043318 Proteins with structure data	043318 Conf-score 9: Direct single protein target assigned
043570 CAH12_HUMAN	043570 Lung	043570 plasma membrane	043570 Proteins with structure data	043570 Conf-score 9: Direct single protein target assigned
075084 FZD7_HUMAN	075084 Lung	075084 plasma membrane	075084 Proteins with structure data	075084 Conf-score 9: Direct single protein target assigned
075762 TRPA1_HUMAN	075762 Lung	075762 plasma membrane	075762 Proteins with structure data	075762 Conf-score 9: Direct single protein target assigned

← Return

Saved conditions

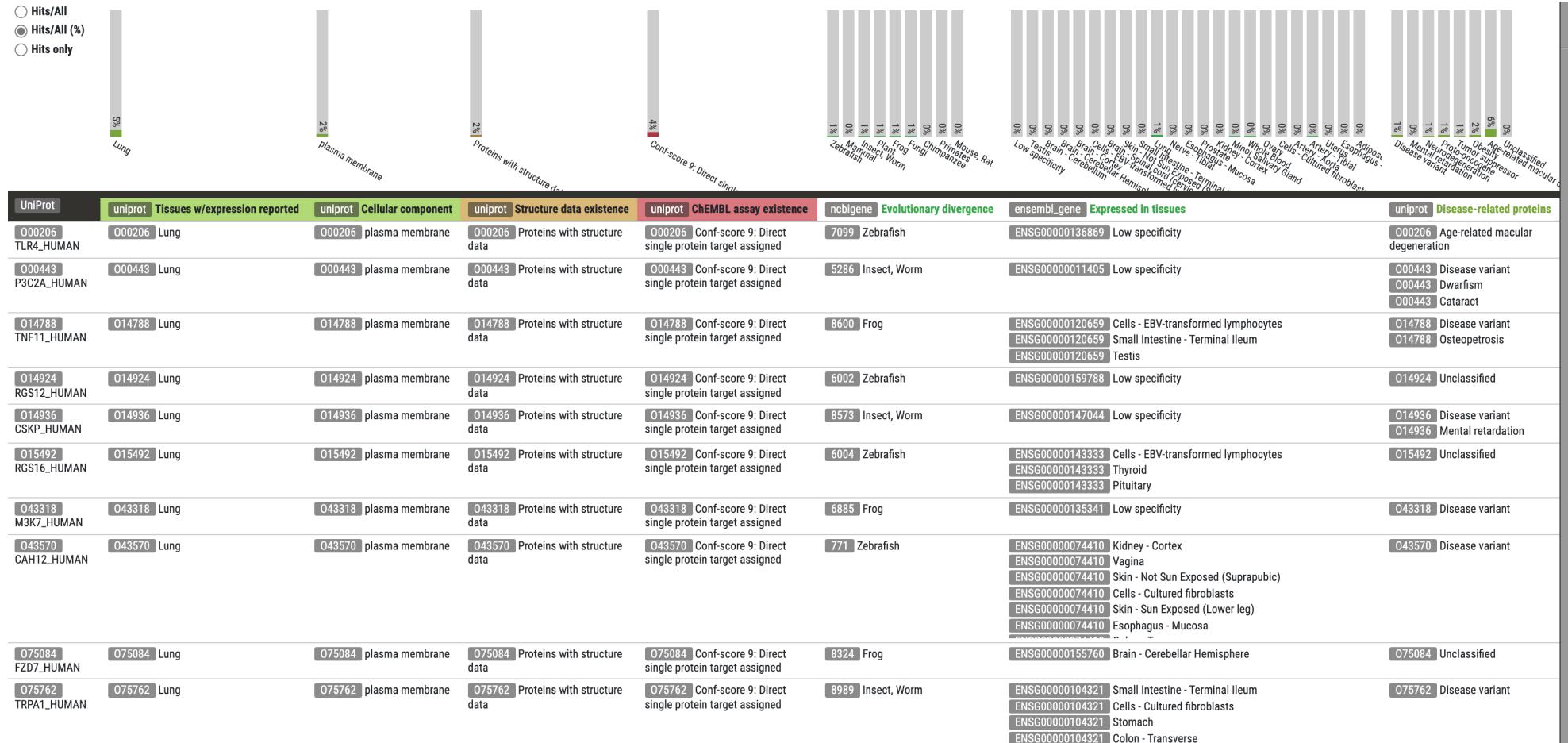
U... Tissues ... Cellu... Structu... ChEMBL... X

Complete

131 / 131

↓ TSV

# 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

Complete

131 / 131

TSV

U Tiss... C... Str... Ch... Ev... E... Di... x

Complete

131 / 131

TSV

# TogoDXでIDリストをマッピングする

The screenshot displays the TogoDX web application interface, which allows users to search for biological entities using various identifiers and filters.

**Left Panel (Search Results 1):**

- Attribute keys:** Gene, biotype, Chromosome, # of paralogs, Evolutionary divergence, Tissue-specific low expression (BlastGenome), Tissue-specific high expression (BlastGenome), No expression in tissues (GTEX), Tissue-specific high expression (HMP), Cell type-specific high expression (HMP).
- Attribute values:** protein coding: 1544, Chr1: 5477, Chr2: 4,157, Chr3: 3,188, Chr4: 2,058, Chr5: 2,188, Chr6: 3,045, Chr7: 3,017, Chr8: 2,480, Chr9: 1,231, Chr10: 3,031, Chr11: 3,034, Chr12: 3,030, Chr13: 2,041, Chr14: 1,232, Chr15: 2,046, Chr17: 1,031, Chr18: 2,049, Chr20: 2,042, X: 3,425, Y: 1,643, Z: 1,643.
- Condition builder:** Select target dataset: UniProt, Map your ID: optional, P01115, P064841, Mapping combined.
- Saved conditions:** View results, Filter, Tissues, Bio, Disease, Plant.

**Right Panel (Search Results 2):**

- Attribute keys:** Gene, biotype, Chromosome, # of paralogs, Evolutionary divergence, Tissue-specific low expression (BlastGenome), Tissue-specific high expression (BlastGenome), No expression in tissues (GTEX), Tissue-specific high expression (HMP), Cell type-specific high expression (HMP), Protein domains, Cellular component, Biological process, Molecular function, Ligands, Molecular mass, PTMs, # of transmembrane domains, # of phosphorylation sites, # of glycosylation sites, Disease-related proteins, Tissues & expression, Evidence of existence.
- Attribute values:** protein coding: 1544, Chr1: 5477, Chr2: 4,157, Chr3: 3,188, Chr4: 2,058, Chr5: 2,188, Chr6: 3,045, Chr7: 3,017, Chr8: 2,480, Chr9: 1,231, Chr10: 3,031, Chr11: 3,034, Chr12: 3,030, Chr13: 2,041, Chr14: 1,232, Chr15: 2,046, Chr17: 1,031, Chr18: 2,049, Chr20: 2,042, X: 3,425, Y: 1,643, Z: 1,643.
- Condition builder:** Select target dataset: UniProt, Map your ID: optional, Q97272, Q9743, Q9751, Q9734, Q9745, Mapping combined.
- Saved conditions:** View results, Filter, Tissues, Bio, Disease, Plant.

## まとめ

- TogoTV を使って、新たな研究データ(ベース)を発見・学習・活用する
  - バイオインフォマティクスツールやデータベースの動画マニュアルを上手に活用して、自身の研究に関連する「新しい研究データ」を発見する
- TogoDX/human を使って統合されたデータを探索・俯瞰・抽出する
  - 高度に統合されたデータベースを探索・俯瞰し、「新しい研究データ」につながる知識を抽出できる仕組みが出来上がりつつある