

データ(ベース)駆動型生命科学研究への招待: 統合データベースの活用事例とバイオインフォマティクスツールの使いこなし術

プラチナバイオ株式会社 事業推進部 ディレクター / 広島大学ゲノム編集イノベーションセンター バイオDX産学共創拠点 プロジェクトマネージャー
小野 浩雅

ono@pt-bio.com / onohrms@hiroshima-u.ac.jp

2024年9月30日

東京農工大学動物生化学研究室セミナー

これは、東京農工大学動物生化学研究室セミナ
ー『データ(ベース)駆動型生命科学研究への招
待: 統合データベースの活用事例とバイオイン
フォマティクスツールの使いこなし術』の講演
資料です。

- PDF のダウンロード
 - https://bit.ly/240930_TUAT
 -

© 2024 ONO Hiromasa, CC-BY-4.0 (出典明記
でご自由にお使いください)



自己紹介

- 小野 浩雅

- 統合TVの運営・編集
- 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
- 2024年4月から
 - プラチナバイオ株式会社 事業推進部 ディレクター / 広島大学ゲノム編集イノベーションセンター バイオDX产学共創拠点 プロジェクトマネージャー
 - 産業有用生物のゲノム情報の取得・目的機能に関わる遺伝子の特定からゲノム編集による機能向上まで一貫して実現できるプラットフォームをつくる

統合TV(TogoTV)

課題

- 生命科学分野では、DBやバイオインフォマティクスツールの種類や対象が日々増加している
 - 初学者にとっては、何を選ぶべきか、どう使えるのか、組み合わせて使うことができるのかが分からため利用を躊躇してしまう
 - DBやウェブツールを利活用するためのまとめた教材がほとんどない

統合TV(TogoTV)

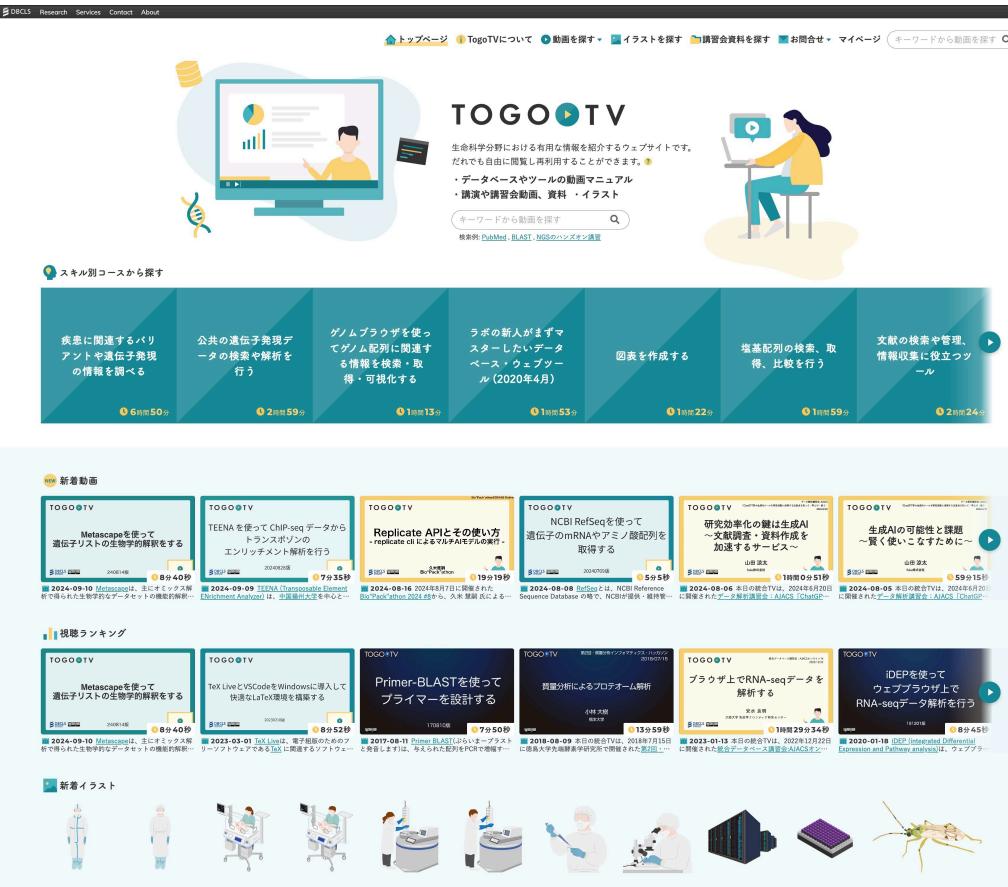
2024/07/10 東京農工大学動物生物学研究室セミナー『データ(ベース)駆動型生命科学研究への招待: 統合データベースの活用事例とバイオインフォマティクスツールの使いこなし術』

- <https://togotv.dbcls.jp/> 

○ DBCLSが運営する、生命科学分野における有用な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、イラストを独自に制作し、紹介するウェブサイト

- 実験系研究者の初学者が主要な想定利用者
- 技術習得の障壁を下げるため、自主学習、新人・後輩指導、講義・勉強会の教材として誰でも自由に使える

- 2007年8月スタート(2024年で17年)



TogoTV@YouTube

- 各動画はYouTubeで公開されており、環境に応じた解像度、倍速表示等で快適に閲覧可能
- 2,200本を超える動画が公開されており、のべ320万回(月間3万回以上)再生
(2024年8月末現在)



統計情報 (2024年 9月)

動画数

2,206 本

スキル別コース数

10 本

総再生数

320 万 (月間3万+)

チャンネル登録数

10,200+

イラスト数

2,042 本

のべ製作者数

140+ 名

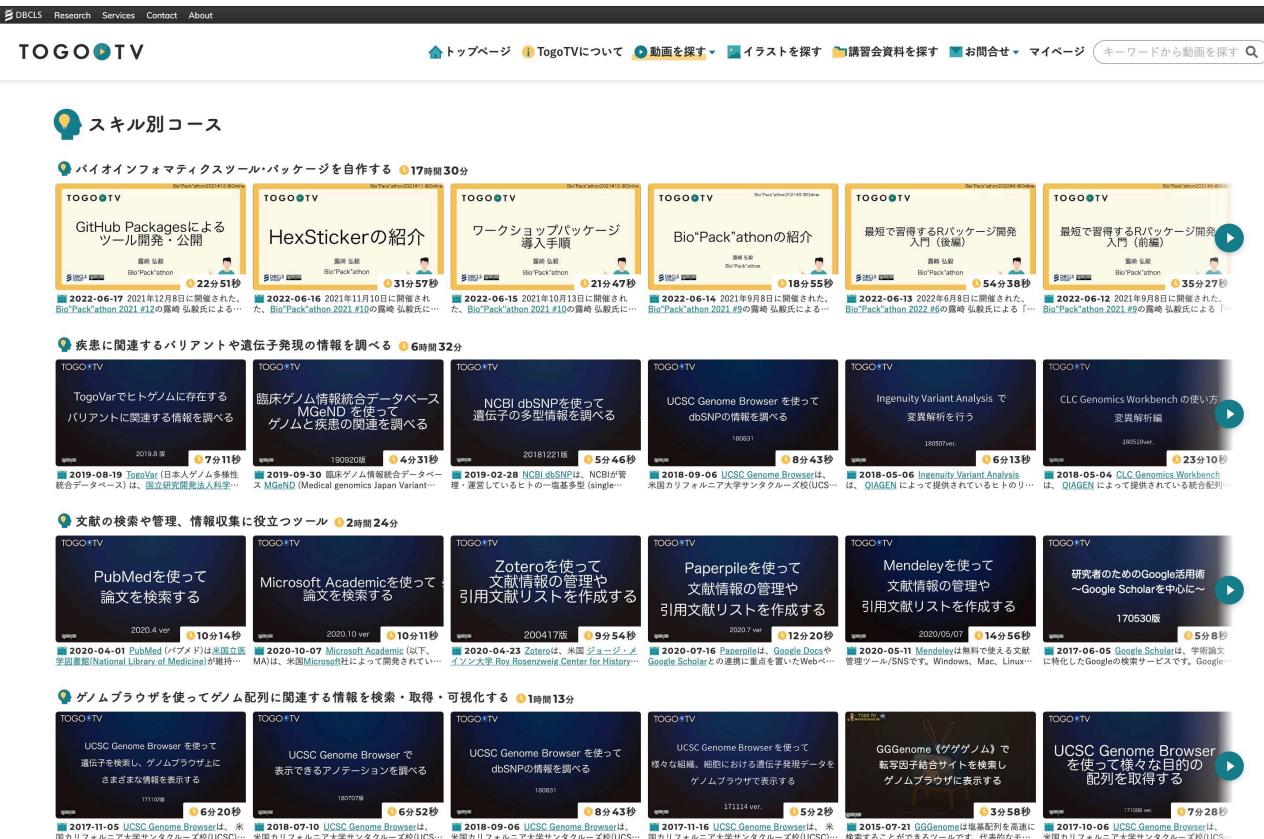
個別動画ページ

- 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
- 見どころダイジェスト
 - YouTubeの埋め込みプレイヤなので、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
- 「再生リストに保存」する機能
 - Googleアカウントでログインする
 - 利用者オリジナルの動画再生リストを作成、共有、公開することができる



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



ラボの新人がまずマスターしたいデータベース・ウェブツール(2020年4月)

- 研究室に入ってきた新人が必ず知っておくべき「論文の効率的な検索方法」、「研究発表資料の作成に必要なパワーポイントの图形描画機能」、「Google各種サービスを使って研究生活を効率化する」という3つのテーマに関する10本の動画を紹介しています
- 解説ブログもあります



文献の検索や管理、情報収集に役立つツール

- 泣く子も黙るPubMedはもちろん、ZoteroやPaperpileなどの文献情報の管理・引用文献リスト作成支援ツールの動画が人気



ウェブブラウザ上でRNA-seqデータ解析をする

- コマンド操作に苦手意識がある方でも RNA-seqデータ解析ができる時代
- さまざまなツールがあるので、自身の研究テーマに合った解析方法や図を作成するため、また操作感などが合うかなど、複数試してみるのが吉



現在のラインナップ例 (今後充実させていきます)

- ラボの新人がまずマスターしたいデータベース・ウェブツール (2020年4月)
- 文献の検索や管理、情報収集に役立つツール
- 文章の執筆に役立つツール
- 塩基配列の検索、取得、比較を行う - 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
- ウェブブラウザ上でRNA-seqデータ解析をする
- 図表を作成する
- 疾患に関連するバリエントや遺伝子発現の情報を調べる
- バイオインフォマティクスツール・パッケージを自作する

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
- サービスの開発者・研究者が、背景や周辺の基礎知識を紹介するとともにDBやウェブツールの基本的な使いこなし術や高度な組み合わせ方法などを紹介
- 講習会の復習や、他分野の研究内容を(何度も繰り返し)学習することができます。



PubMed検索のプロによる文献検索のイロハを学ぶ

- 最新の文献検索方法を知る @ AJACSオンライン16, 2023年
 - PubMedの使い方が身につき、自身に必要な文献を効率よく見つけられることを目的として、2020年にリニューアルされたPubMed (パブメド)の基本的な使い方から検索の仕組み、自分の欲しい論文を効率的に検索するにはどうすればよいかなど、PubMedを余すところなく使いこなす方法を紹介
 - PubMed以外にもある便利なサービスを一通り把握し、自身の目的に応じて使えるようにする

次世代シークエンス(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](#)

ウェブブラウザでできるRNA-seqデータ解析を実例を交えて学ぶ

- ブラウザ上でRNA-seqデータを解析する @ AJACSオンライン14, 2023年
 - バルクRNA-seq下流解析ができるようになることを目的として、ウェブブラウザを用いたバルクRNA-seq解析ツールであるiDEPの使い方をハンズオン形式で講習するとともに、類似ツールであるBioJupiesやRaNA-Seqについても紹介
 - 胸腺腫合併重症筋無力症の実データを用いた解析実例も紹介

NGS解析について、さらにもっと基礎から応用までを深く学びたい方向け(それぞれ約10-50時間程度)

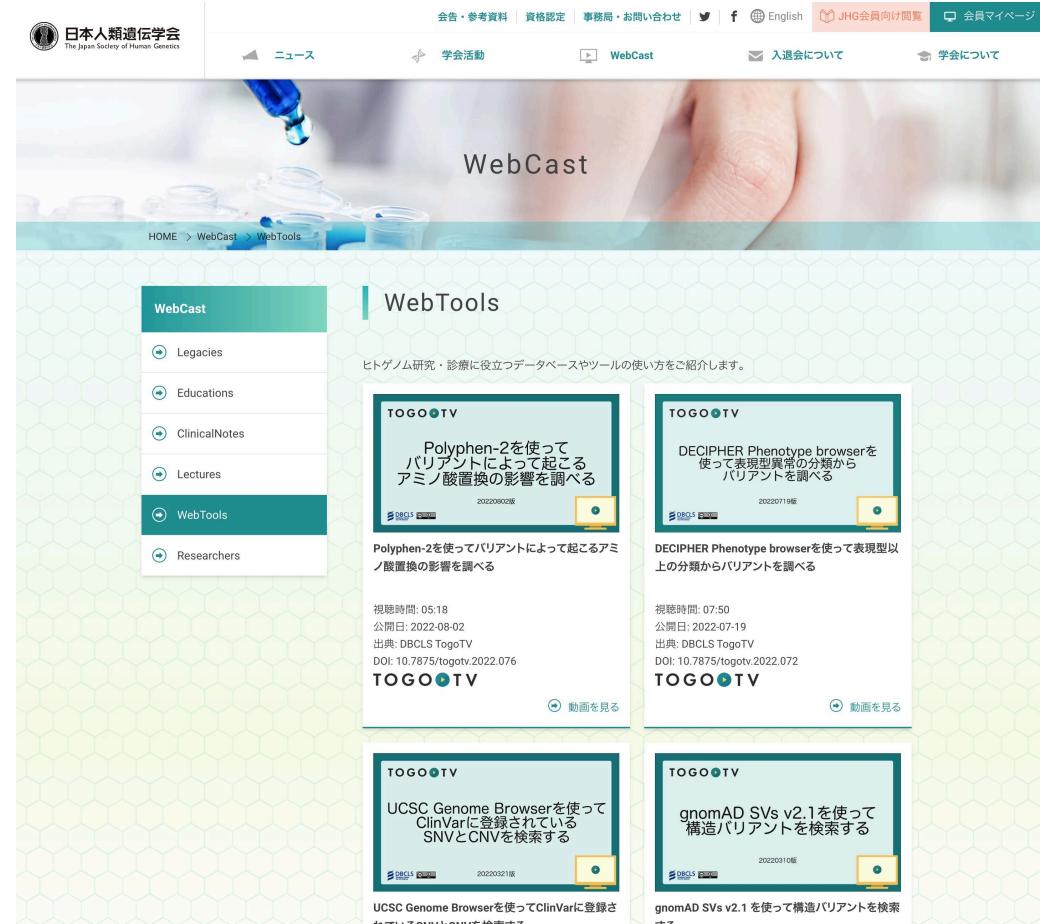
- 「バイオインフォマティクス人材育成カリキュラム (次世代シークエンサ) 速習コース(2014年8月)
- 「バイオインフォマティクス人材育成カリキュラム 次世代シークエンサ(NGS)ハンズオン講習会(2015年8月)
- NGSハンズオン講習会2016
- NGSハンズオン講習会2017
- 先進ゲノム支援(PAGS)、DDBJ、DBCLS合同情報解析講習会(2019)

ChatGPT等の生成AIツールを研究活動に活用する注意点を知って・学んで・使う

- 生成AIの可能性と課題～賢く使いこなすために～ @ データ解析講習会：AJACS, 2024年
 - 生成AI技術の基礎を振り返るとともに、生成AIツールを賢く使って研究の生産性を高めるノウハウを紹介
- 研究効率化の鍵は生成AI～文献調査・資料作成を加速するサービス～ @ データ解析講習会：AJACS, 2024年
 - 多様なシーンで研究の生産性を高めるAIサービスや自分だけのツール作りのためのノウハウを紹介

学会との連携

- 日本人類遺伝学会の教育コンテンツ配信システム **JSHG-WebCast(JWC)**
 - 統合TVと学会との連携は初めての試み
 - 統合TVで作成する動画について学会の推薦を受けた専門家のレビューを受けられる
 - 分野の専門家自身の講義・講習をコンテンツ化できる
- 日本小児遺伝学会とも連携予定



AJACSオンライン9 「疾患に関する多型データを解析する」

- 日本人類遺伝学会との連携によって、講義内容と講師を選定
 - 表現型と遺伝子型のデータを共有および比較する @ AJACSオンライン9, 2022年
 - 関心対象の表現型が解析によって明らかになった遺伝子型を説明できるか、を検証できるようになる @ 山本 俊至 先生
 - バリアントの機能を推定する @ AJACSオンライン9
 - バリアントがタンパク質の機能およびスプライシングに与える影響について *in silico* 解析を行うプログラムについて、その使い方と解析結果について理解する @ 才津 浩智 先生

イラストを探す (Togo picture gallery)

- 「イラストを探す」
- 生命科学分野のイラストが、出典明記だけで、誰でも自由に利用可能 (CC-BY-4.0)
- 研究発表のスライド・ポスター作成、資料作成等に、ぜひご活用ください
- 日本人類遺伝学会との連携はイラストでも
- 統合TVのコンテンツを利用したいのですが、著作権の扱いはどうなっていますか?
- 2024年8月末で120件の引用論文

日本分子生物学会との連携

- 2023年の第46回日本分子生物学会年会(参加者8000人以上)の参加章シール企画に公式採用されました 
- すべての素材はTogo picture galleryにあります
 - 当時なかったものは追加で制作



リクエスト募集中

- お探しの動画マニュアルや画像が見つからない場合は、[統合TV番組リクエストフォーム](#)でお気軽にリクエストしてください。
- すべて目を通しています
- 講演・講義動画やイラストの寄託も大歓迎です。

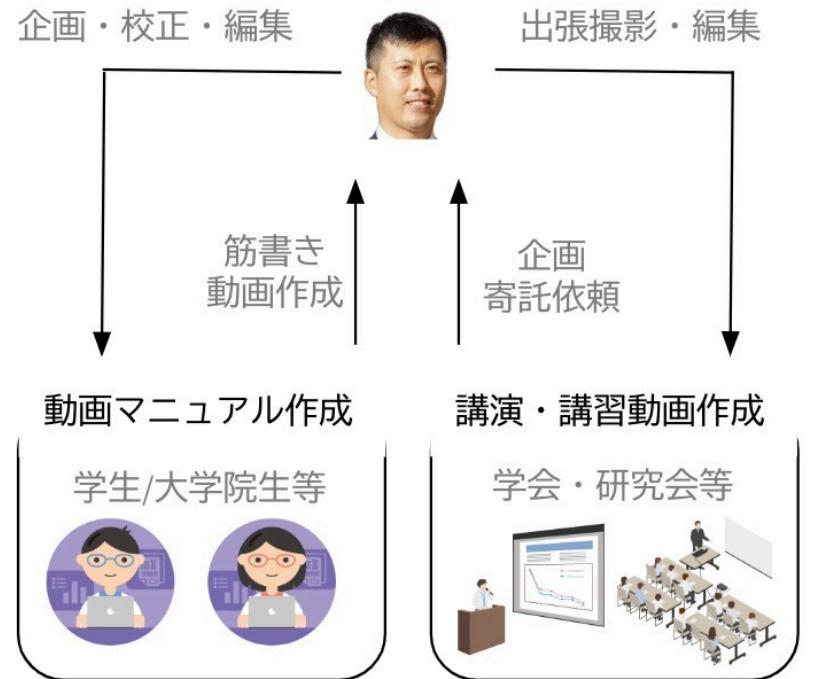
The screenshot shows the TogoTV website with a navigation bar at the top. Below the navigation, there's a main content area with a heading '番組リクエスト' (Program Request) and a sub-section '統合TV コンテンツリクエスト' (Integrated TV Content Request). Both sections contain explanatory text and a note that the content is non-public. At the bottom, there's a form field labeled 'リクエストしたいコンテンツの種類はなんですか? *' (What type of content do you want to request? *) with three options: '動画' (Video), '静止画' (Still image), and '統合TV全体や各コンテンツについてのご感想・ご意見、システム不具合、機能要望など利用者の声' (Comments, opinions, system issues, feature requests, and user voices about the entire TogoTV or its various contents).

2024/04/10 東京農工大学動物生化学研究室セミナー『データベース駆動型生命科学研究への招待: 統合データベースの活用事例とバイオインフォマティクスツールの使いこなし術』

統合TVコンテンツを作つてみたい方も募集中

- ・ 動画やイラストを制作してくれる方を随時募集中です!
- ・ オンラインで完結する作成環境を整備しており、遠隔地でもOKです。
- ・ 「本業優先」でノルマ無し
- ・ 謝金あります 💰
- ・ 動画制作の実際: 「[画面録画/編集ソフトウェア Camtasia 2019 を使って統合TVの動画を作成する](#)」
- ・ 静止画制作の実際: 「[ライフサイエンス分野のイラスト集を作る仕事@AJACS本郷](#)」

コンテンツ制作体制の概要



企画立案・校正・コミュニケーション

ClickUp 進捗管理

Camtasia® 動画撮影・編集

Google Sheets GitHub コンテンツ管理
ウェブサイト構築

参考・関連書籍1

- 実験医学増刊 Vol.40 No.17 バイオDBとウェブツール ラボで使える最新70選 知る・学ぶ・使う、バイオDX時代の羅針盤



参考・関連書籍2

- 生命科学研究のためのデジタルツール入門
第2版 - 結果に差がつく使いこなし術 -



生命科学分野におけるデータベースの統合とその利活用およびデータ駆動型研究を行うためのバイオインフォマティクスツール

- 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 @ DBCLS
 - **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

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汗
 - 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、相互連携性) 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
 - バイオインフォマティクスの出番
 - 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

- BioHackathon (バイオハッカソン)
 - 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
 - 生成AIとの相性が良さそう？

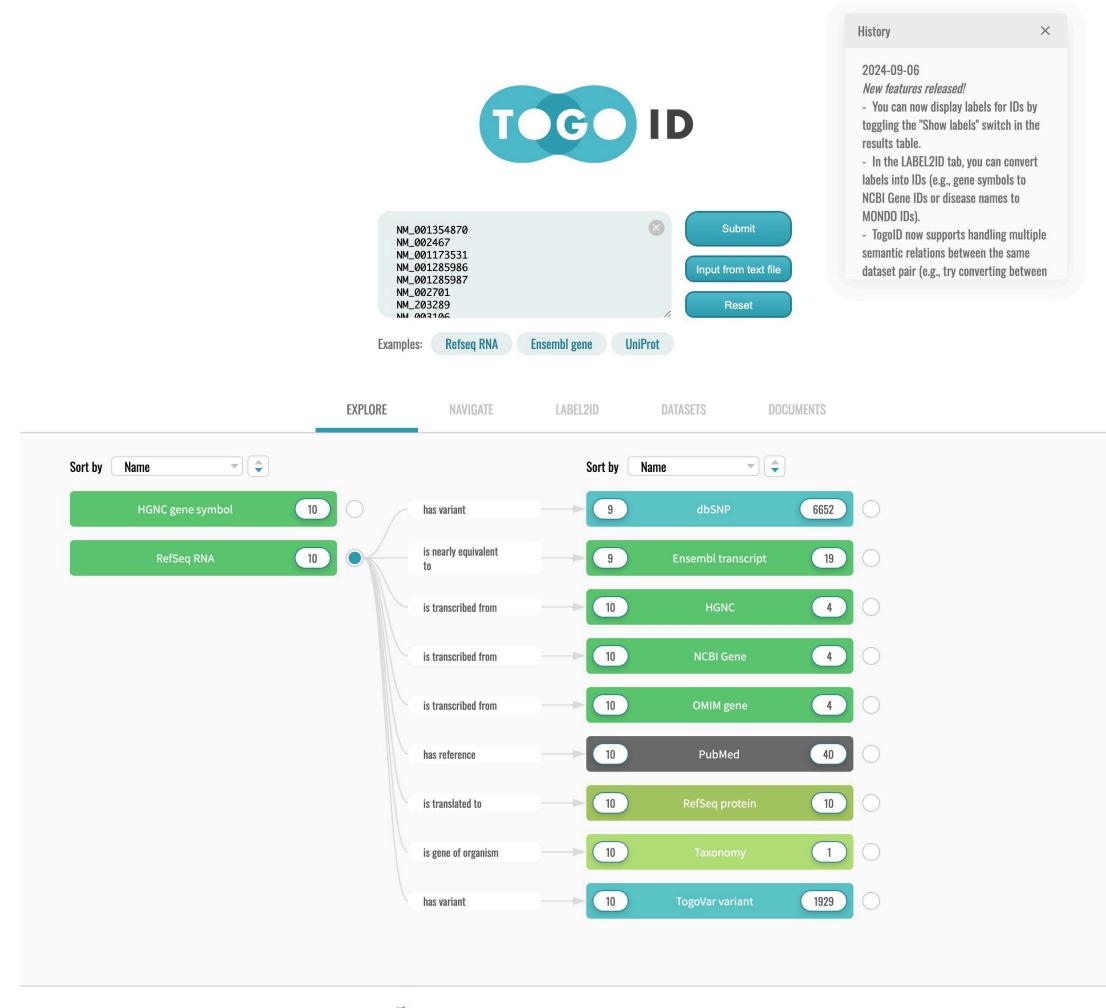
TogolD を使って生命科学系データベース間のつながりを探索的に確認しながらID変換を行う

生命科学分野におけるID変換の必要性

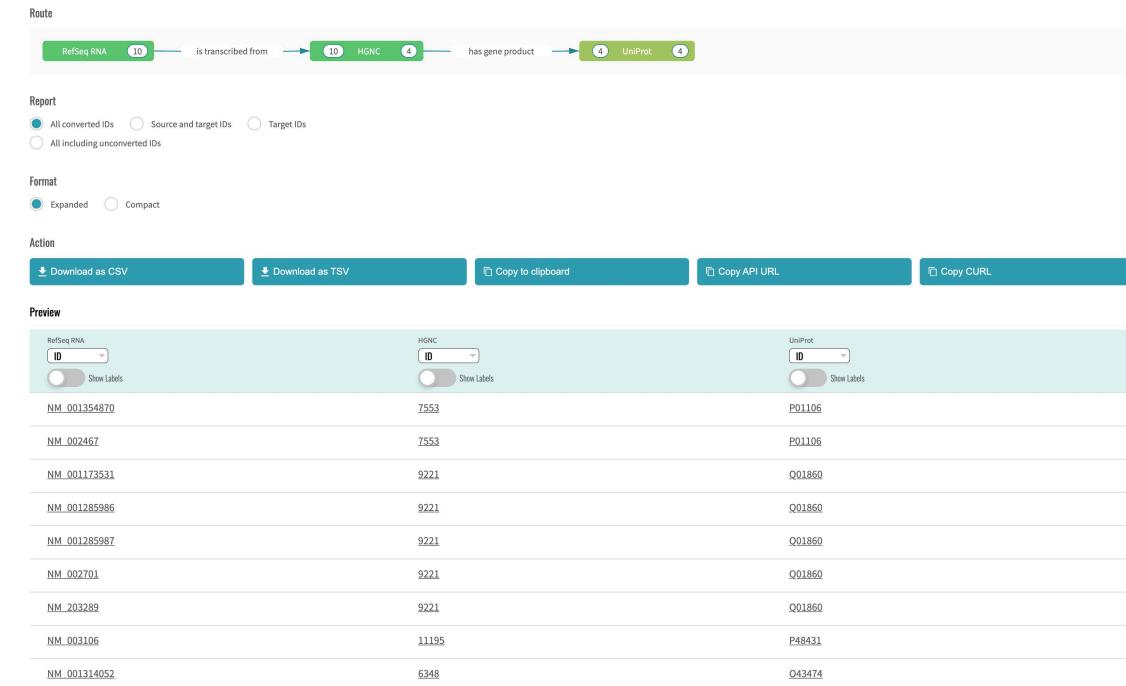
- 様々なデータベース (DB) を活用するには異なる ID 間のリンクが重要
 - 使いたい解析ツールが手元にある ID では使えない
 - 等価なものに対する ID 間で変換したい
 - 例: NCBI Gene ID  Ensembl Gene ID
 - 関連する情報を取得したい
 - 遺伝子が関与する疾患、化合物が関与するパスウェイ etc.
- 既存の ID 変換サービスの問題点
 - 対象としている DB の範囲が限られる
 - 大元の DB の更新に追従していない
 - プログラムから利用できるAPI が提供されていない

TogolD

- 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)



- 73のデータベースに由来する 104 のデータセットのペアを収載 (2024年7月末現在)
 - 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



Ver.2.0公開!(2024年9月)

1. ラベルとIDの相互変換機能の追加 (LABEL2ID)

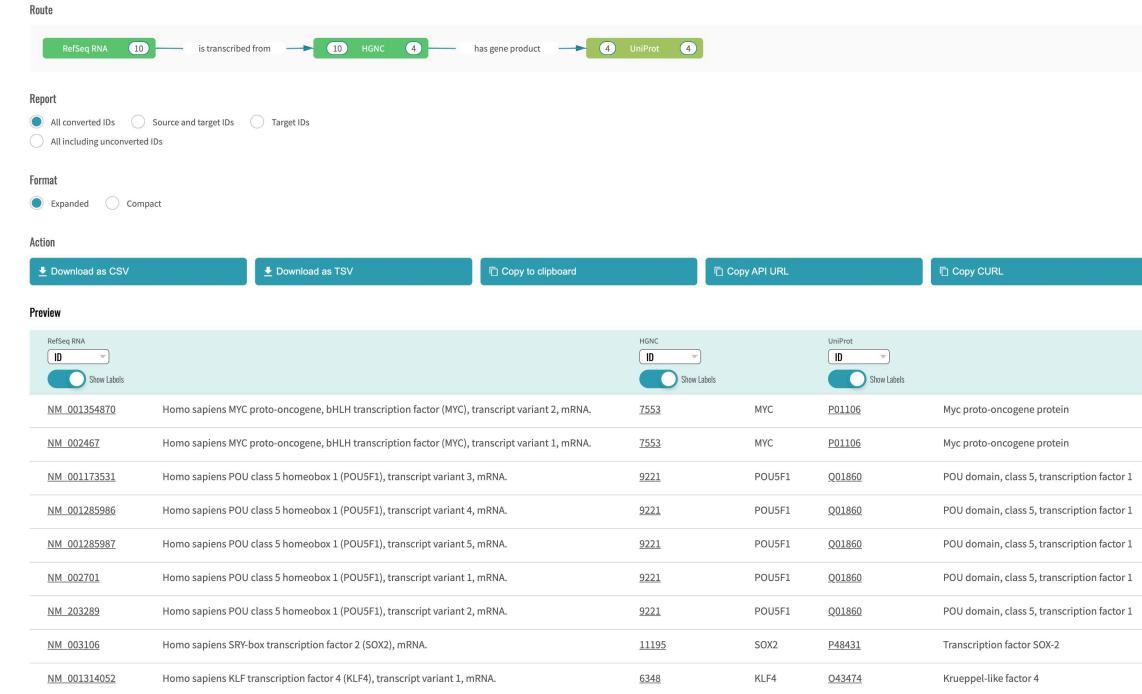
- 遺伝子シンボルや疾患名などのラベルをIDに変換可能に

2. IDに対応するラベルの表示機能

- 変換後のIDが示す内容を理解しやすく

3. 複数の意味的関係への対応

- 例：タンパク質と糖鎖の関係
 - タンパク質が酵素として糖鎖を処理
 - タンパク質が糖鎖によって修飾される



TogoDX/human を使って統合されたデータを探索・俯瞰・抽出する

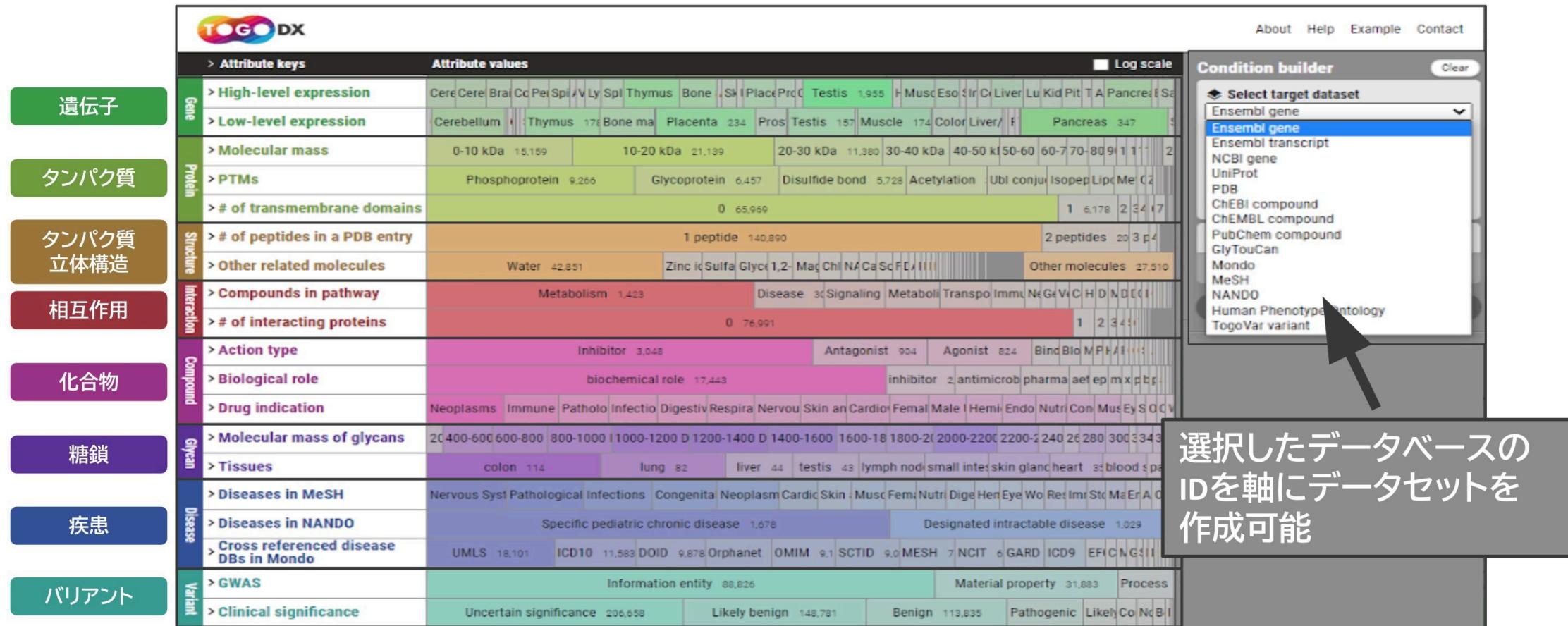
データを統合してTogoDXというアプリケーションを作った

- 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、マウス版や異なる他のテーマでも流用できる

TogoDX/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、20個のデータベースに由来する65個の attribute が利用可能

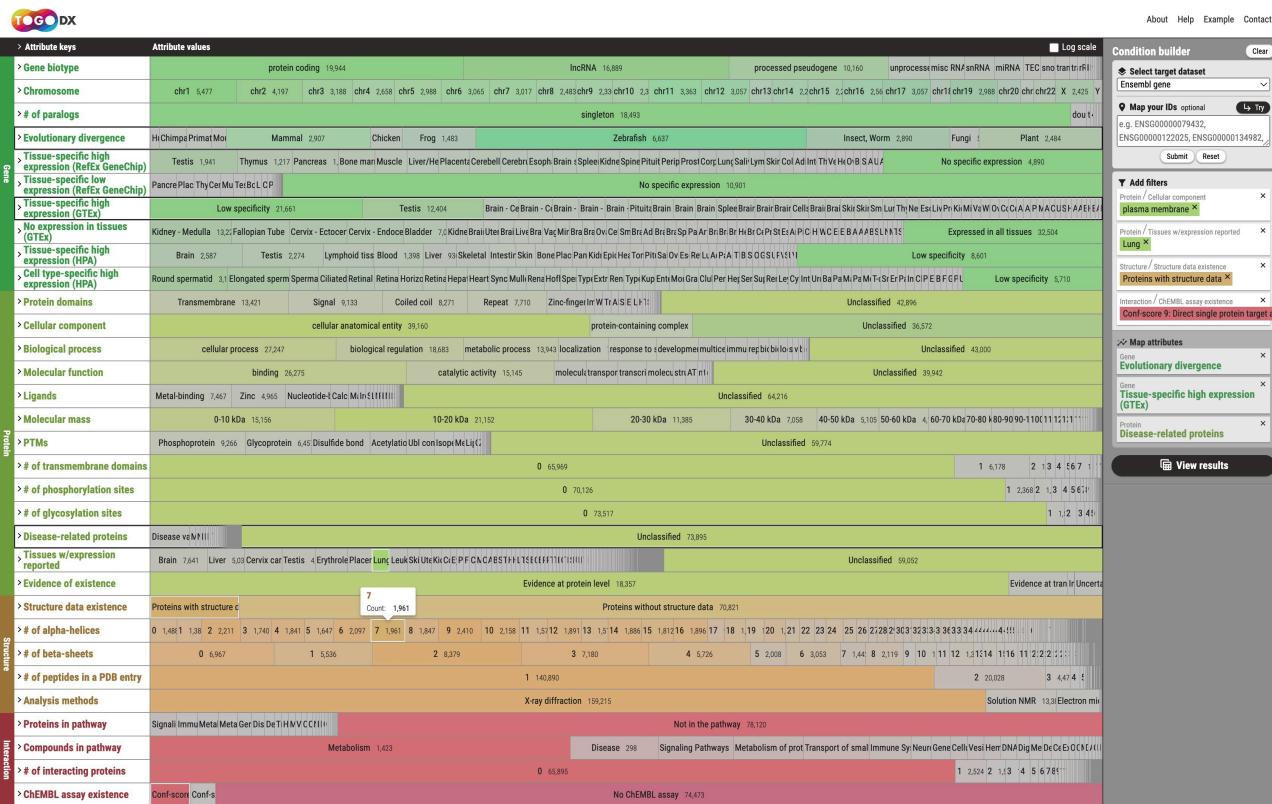
TogoDX で統合されたデータを俯瞰する



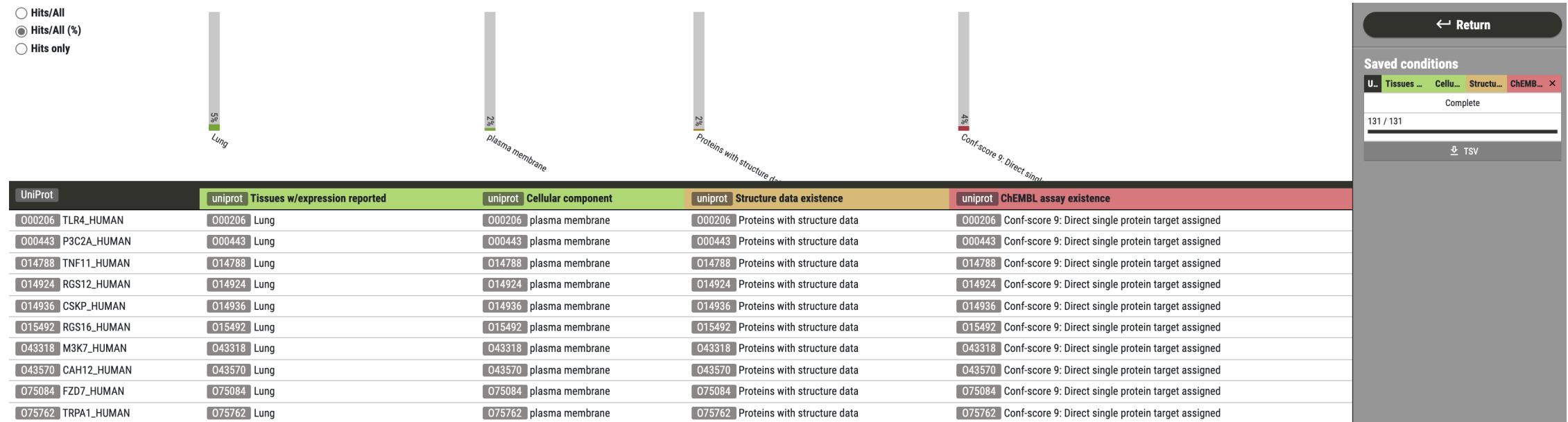
探索例

- 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

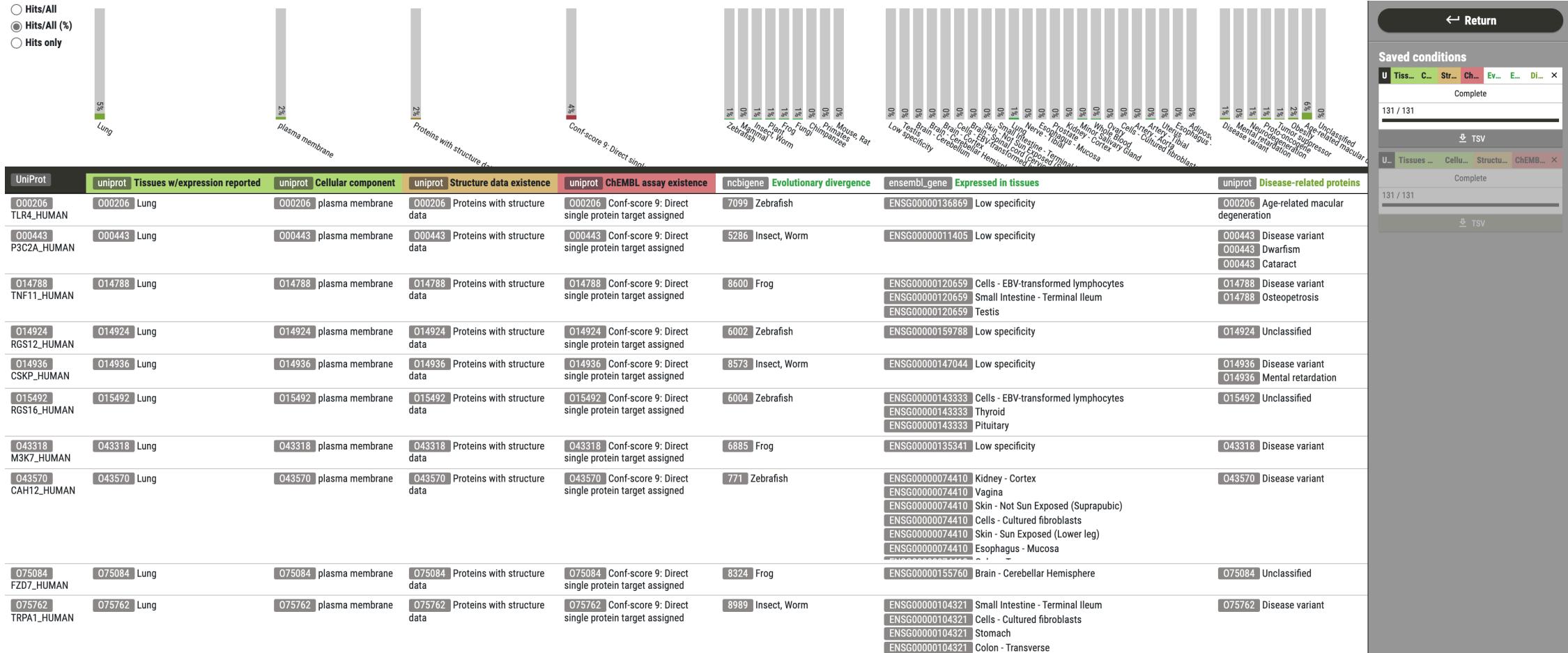
ヒトのタンパク質の一覧をワンストップで取得することができる💡



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



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

The screenshot displays the TogoDX/human interface, which is a bioinformatics tool for searching and comparing gene data across multiple species. The interface consists of two main panels, each showing a grid of gene information.

Left Panel:

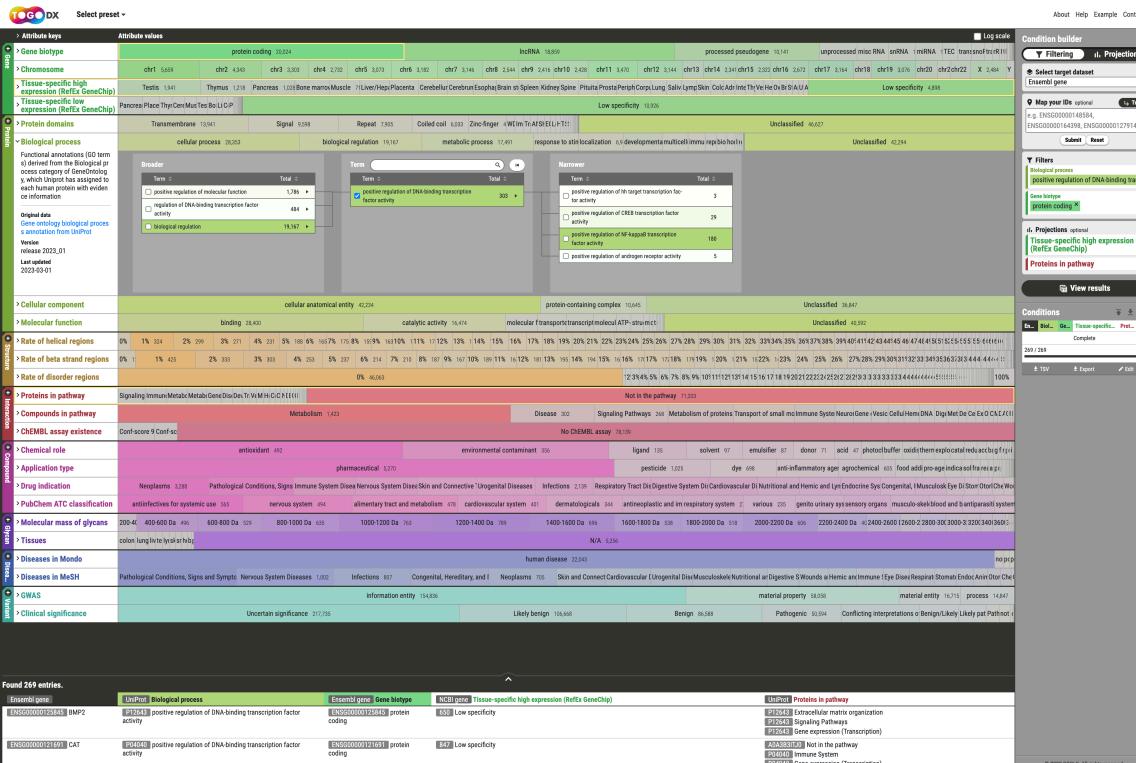
- Header:** Shows the TogoDX logo and navigation links (About, Help, Example, Contact).
- Search Bar:** Includes fields for "Selected target dataset" (set to "ExPozit"), "Map your IDs" (with a dropdown menu showing "1-10" and "P01486, P33223, P21359-1, P01111, P02484"), and "Submit" and "Reset" buttons.
- Condition builder:** A sidebar with checkboxes for "Gene", "Chromosome", "D of paralogues", "Evolutionary divergence", "Tissue-specific high expression (GeneChip)", "Tissue-specific low expression (GeneChip)", "Tissue-specific high expression (GTEX)", "Tissue-specific high expression (HTRA)", "Cell type-specific high expression (GTEX)", "Protein domains", "Cellular component", "Biological process", "Molecular function", "Ligands", "Molecular mass", "PTMs", "Transmembrane domain", "Phosphorylation sites", "Glycosylation sites", "Disease-related proteins", "Tissues expression reported", "Evidence of existence", "Structure data existence", "Proteins in pathway", "Components in pathway", "Chemical role", "Application type", "Action type", "Biological role", "Drug indication", "Drug development phase", "PubChem ATC classification", "CHEMBL ATC classification", "Molecular mass of glycans", "Tissues", "Subsumption", "Diseases in Mondo", "Diseases in Mesh", "Diseases in MANDO", "Cross-referenced diseases in Mondo", "Phenotypic abnormality", "GWAS", and "Clinical significance".
- Grid:** The main area displays a grid of gene entries. Each row contains a gene ID (e.g., dr1, dr2, dr3, etc.), chromosome number, name, and various expression and function details. The grid is color-coded by tissue type (e.g., Brain, Liver, Heart, Lung, Skin, etc.).

Right Panel:

- Header:** Similar to the left panel, with the TogoDX logo and navigation links.
- Search Bar:** Shows the same search parameters as the left panel.
- Condition builder:** A sidebar with checkboxes for "Gene", "Chromosome", "D of paralogues", "Evolutionary divergence", "Tissue-specific high expression (GeneChip)", "Tissue-specific low expression (GeneChip)", "Tissue-specific high expression (GTEX)", "Tissue-specific high expression (HTRA)", "Cell type-specific high expression (GTEX)", "Protein domains", "Cellular component", "Biological process", "Molecular function", "Ligands", "Molecular mass", "PTMs", "Transmembrane domain", "Phosphorylation sites", "Glycosylation sites", "Disease-related proteins", "Tissues expression reported", "Evidence of existence", "Structure data existence", "Proteins in pathway", "Components in pathway", "Chemical role", "Application type", "Action type", "Biological role", "Drug indication", "Drug development phase", "PubChem ATC classification", "CHEMBL ATC classification", "Molecular mass of glycans", "Tissues", "Subsumption", "Diseases in Mondo", "Diseases in Mesh", "Diseases in MANDO", "Cross-referenced diseases in Mondo", "Phenotypic abnormality", "GWAS", and "Clinical significance".
- Grid:** The main area displays a grid of gene entries, similar to the left panel, showing various gene properties and their expression patterns across different tissues.

Ver.1.2 公開!(2023年9月)

- UI更新・機能追加
 - 表示するAttributeを取捨選択できる機能およびPresetの新設
 - 深い階層のTermを検索可能なオントロジーブラウザ
 - Filterを組み合わせた探索結果をプレビューできる機能
 - JSONファイルによる探索条件の保存・共有機能



動画マニュアル @ 統合TV

- TogoDX/Human v1.2を使ってヒトのデータベースを統合的に探索、俯瞰、抽出する
(基本操作編)
- TogoDX/Human v1.2を使ってヒトのデータベースを統合的に探索、俯瞰、抽出する
(一般疾患編)
 - 生活習慣病の一つである**2型糖尿病**を例に、2型糖尿病と関連が既知のTCF7L2の特徴をもとに、複数のデータベースの情報を組み合わせて他の関連遺伝子候補を探索する方法
- TogoDX/Human v1.2を使ってヒトのデータベースを統合的に探索、俯瞰、抽出する
(希少疾患編)
 - 希少疾患である**鰓耳腎症候群**を例に、鰓耳腎症候群と関連が既知の遺伝子群の特徴をもとに、複数のデータベースの情報を組み合わせて他の関連遺伝子候補を探索する方法

TogoID & TogoDX/human 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きている
- 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)仕組みができるばかりつつある

全体のまとめ

- 生命科学分野におけるデータ駆動型研究の重要性
 - 日々増加・進化するDBやツールを効果的に活用する能力が必須
 - 統合TVなどのリソースを活用し、常に最新の知識・スキルを習得
 - 正面からしか見られなかったものが横や後ろやナナメから見ることができる
のがデータ駆動型研究のいいところ
- 次世代の研究者へのメッセージ
 - 「バイオインフォマティクス」も顕微鏡や実験試薬などと同じ「道具(ツール)」
 - 便利な「道具」を知って、その使い方が分かれば、あとはみなさん自身の情報分析力と想像力の勝負