NBDCの提供するサービス

Catalog

- Integbioデータベースカタログ
  - データベースのデータベース

Archive

- 生命科学系データベースアーカイブ
  - データベースの安定的な長期保存

Cross search

- 生命科学データベース横断検索
  - データベース内の情報を探す

NBDC RDF Portal

- NBDC RDFポータル
  - よりデータ統合に向けた整備

TogoDX

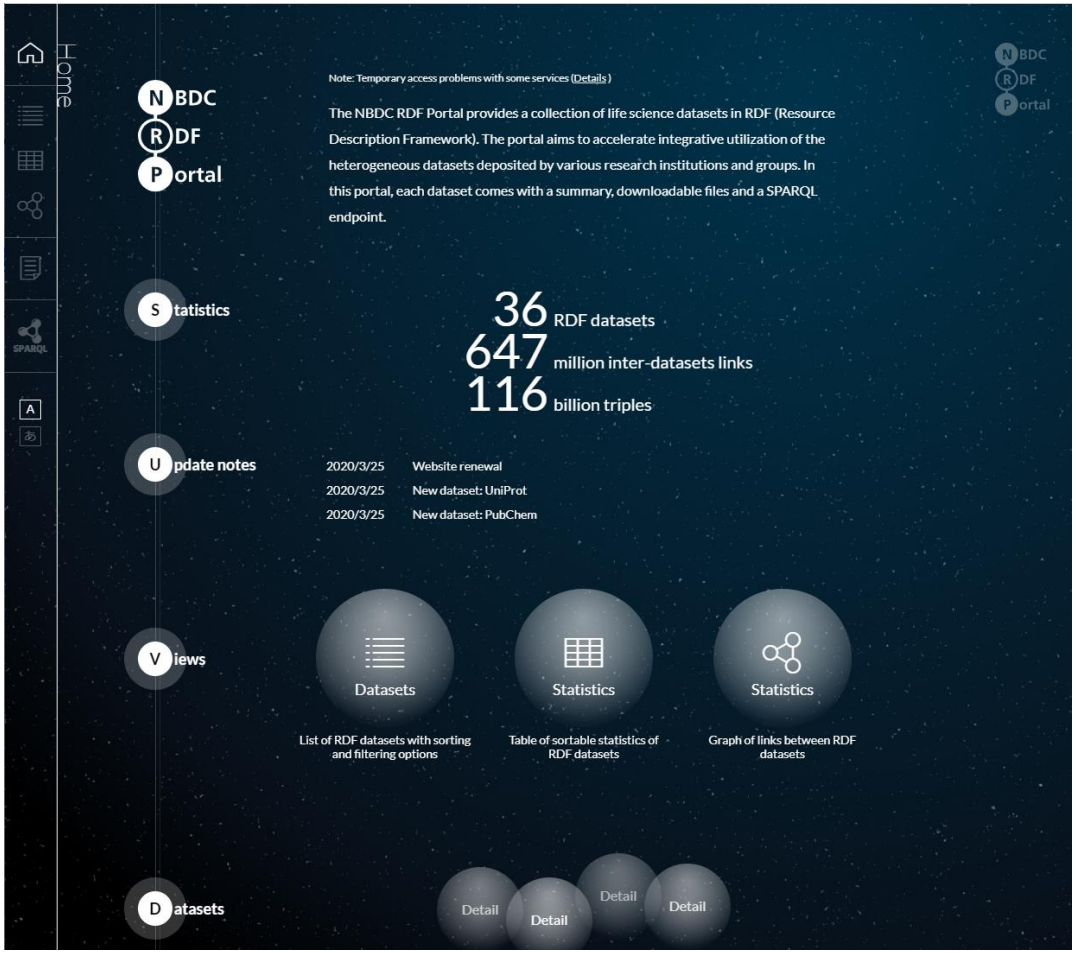
- TogoDX
  - 色んなデータベース情報を一目で分かるようにする

NBDC RDFポータル

https://integbio.jp/rdf/

「**RDFportal**」**で検索**

- 36 のRDFデータセット、約1160億トリプルを公開(2022年8月末)
- RDFデータリポジトリ
  - 「DBCLS RDF化ガイドライン*」に基づいてレビューされたデータ
  - 生命科学分野で重要なデータベースのミラーデータ
- SPARQLエンドポイント
  - RDFの検索が可能



*https://github.com/dbcls/rdfizing-db-guidelines/blob/master/dbcls-rdfizing-db-guidelines-ja.md

RDFとは？

- RDF：Resource Description Framework
  - 主語(Subject) – 述語(Predicate) – 目的語(Object)からなるデータモデル
    - 主語：モノのID (URI)
    - 述語：オントロジーで定義された属性
    - 目的語：別のモノのID (URI)または値(literal)
- 表(RDB)から考える

protein1(主語)のoriginated from(述語)は、human(目的語)である

protein	originated from	bind
protein1	human	lipid1
protein1	human	lipid2
…	…	…

分解

protein	originated from
protein1	human

protein	bind
protein1	lipid1

protein1(主語)にbind(述語)するのは、lipid1(目的語)である

データをRDF化するメリット

- すべてのデータをシンプルな構造(主語-述語-目的語=トリプル)で表現できる
  - １つのフィールドに複数の値があっても、テーブル分割の必要がない
  - データの追加時にテーブル構成を変更する必要がない
  - データの検索時にテーブル構成などを知る必要がない
- URIとオントロジー(＝概念の間の関係の体系)が使える
  - 表記のゆれがなくなる
  - モノに「意味」の情報をオントロジーの形で付加できるので、同義語や上位概念でまとめて検索できる
    （例：「哺乳類」の検索で「ヒト&マウス&ウシ&...」とする必要が無い）

自分と他人のデータベースを簡単に統合できる
１つのクエリで複数のデータベースを一括して検索できる

２つのデータを統合

データ１

DDBJに登録された
DNA配列
AP014957.1

http://identifiers.org/insdc
/AP014957.1#sequence

taxonomy

obo:RO_0002162

taxonomy
39947

http://identifiers.org/taxonomy
/39947

表記がゆれない

データ２

taxonomy
39947

http://identifiers.org/taxonomy
/39947

ラベル

rdfs:label

Oryza sativa Japonica Group

データを統合

DDBJに登録さ
れたDNA配列
AP014957.1

の生物種は

obo:RO_0002162

Oryza sativa Japonica Group

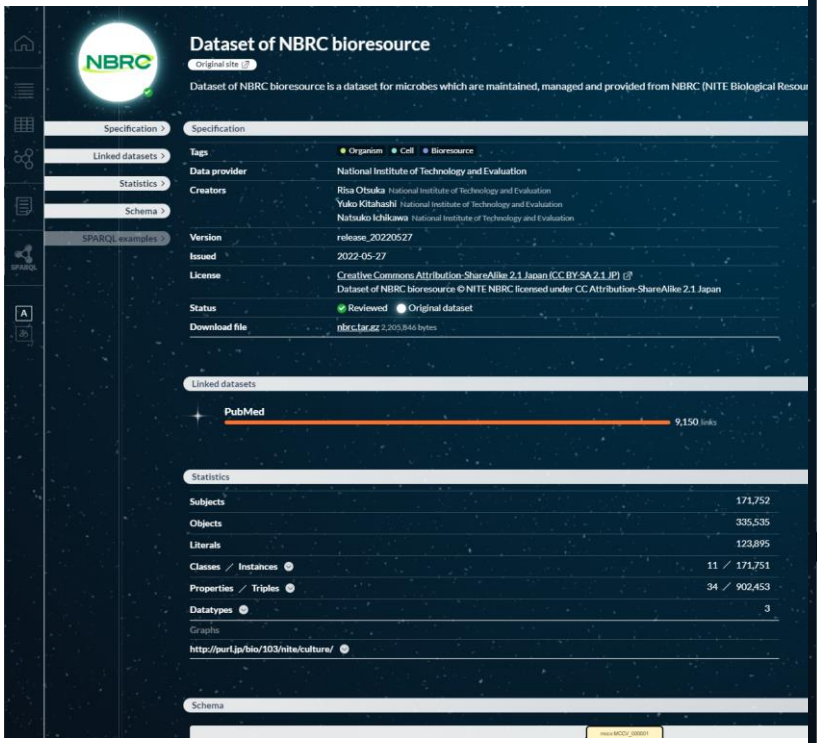
である

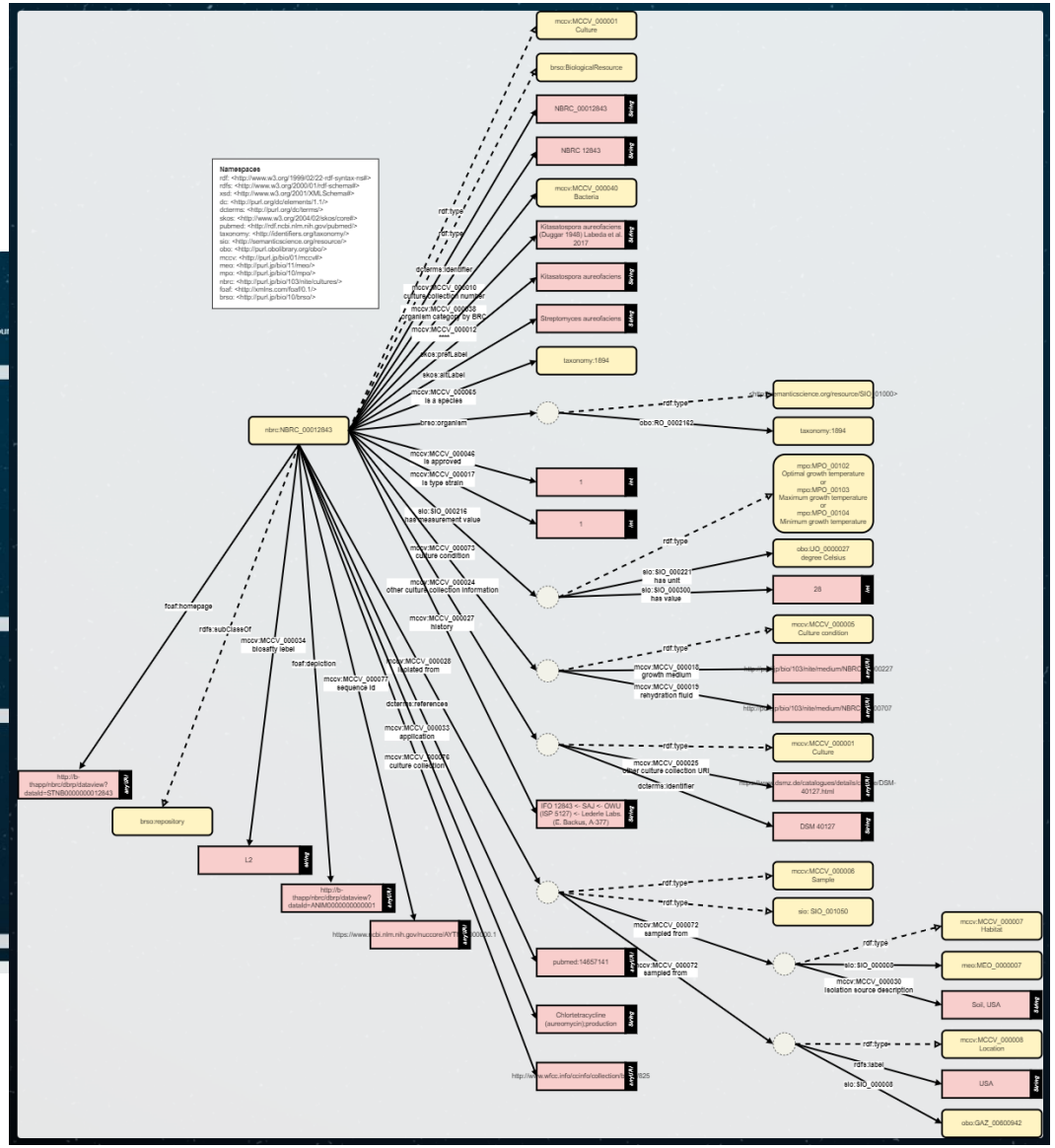
NBDC RDFポータルの特長

- データリポジトリ
  - NBDCに寄託され「DBCLS RDF化ガイドライン」に基づいてレビューされたデータと、生命科学分野で重要なデータベースのミラーデータ(https://integbio.jp/rdf/mirror/)をダウンロード可能な形で公開
  - 以下の情報を明示
    - データに関する詳細なメタデータ
    - データのスキーマ図
    - データへのサンプルクエリ
- (SPARQL)エンドポイント
  - Webインタフェース
  - ジョブ管理機能やキャッシュ機能もある(DBCLS開発のSPARQL-proxyを利用)
  - 保管するRDFデータを中心としたSPARQL検索を可能に

RDFファイルのリポジトリとして

- メタデータ(下)
- スキーマ図(右)





SPARQLエンドポイントとして

SPARQL Proxy

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 + - table ?

```
# Get NBRC microbial strains sampled from soil.

PREFIX mccv: <http://purl.jp/bio/10/mccv#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX meo: <http://purl.jp/bio/11/meo/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select ?strain ?name ?meo ?meo_label
where {
?strain mccv:MCCV_000012 ?name .
?strain mccv:MCCV_000028 ?sample .
    ?sample mccv:MCCV_000072 ?habitat .
    ?habitat sio:SIO_000008 ?meo .
    ?meo rdfs:subClassOf meo:MEO_0000007 .
    ?meo rdfs:label ?meo_label
```

クエリ

Run Query (Ctrl+Enter)

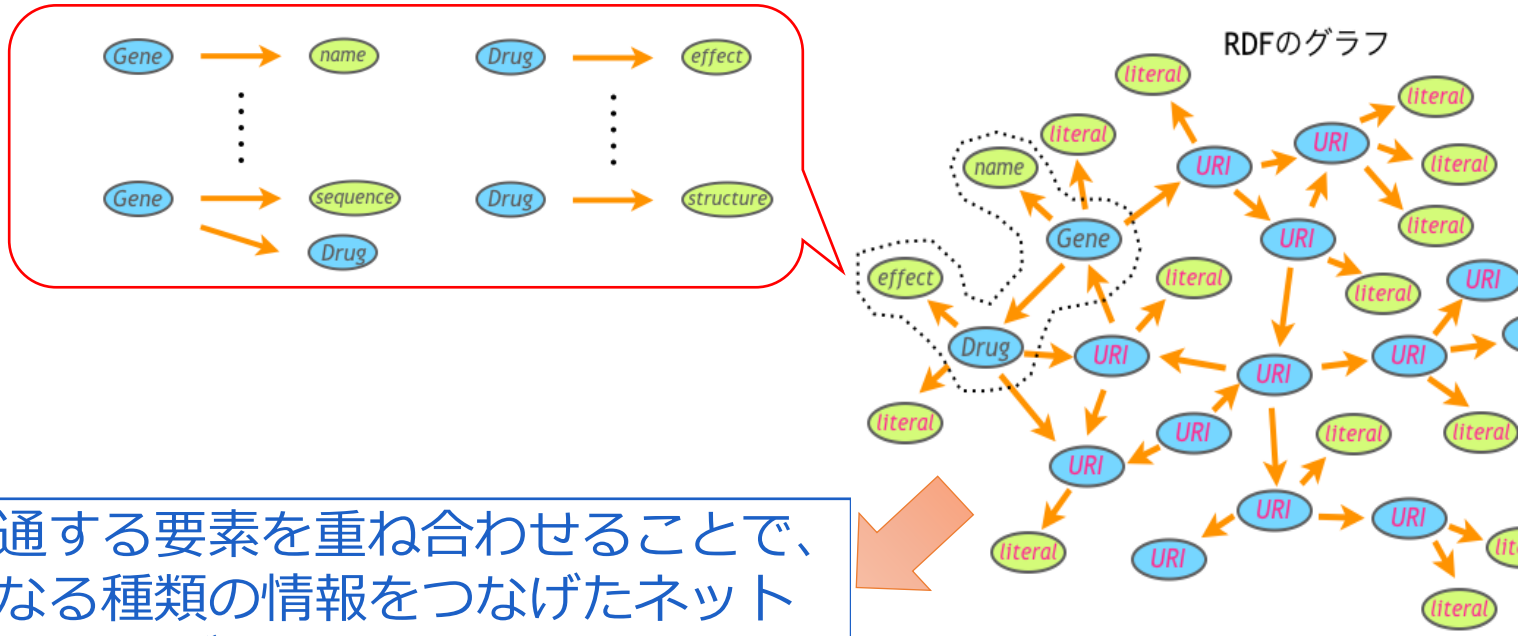
実行結果

[100 bindings. -- 1.1 sec.]

Download

strain	name	meo	meo_label
http://purl.jp/bio/103/nite/nbrc/cultures/00003776	"Streptosporangium roseum Couch 1955"	meo:MEO_0000651	"garden soil"@en
http://purl.jp/bio/103/nite/nbrc/cultures/00004402	"Aspergillus sydowii (Bainier & Sartory) Thom & Church"	meo:MEO_0000126	"forest soil"@en
http://purl.jp/bio/103/nite/nbrc/cultures/00004403	"Aspergillus sojae Sakaguchi & Yamada"	meo:MEO_0000126	"forest soil"@en
http://purl.jp/bio/103/nite/nbrc/cultures/00005422	"Clonostachys rosea (Link : Fries) Schroers et al."	meo:MEO_0000126	"forest soil"@en
http://purl.jp/bio/103/nite/nbrc/cultures/00005424	"Hyalopus nopporoensis Sasaki & Nakane"	meo:MEO_0000126	"forest soil"@en
http://purl.jp/bio/103/nite/nbrc/cultures/00005425	"Monilia geophila Oudemans"	meo:MEO_0000126	"forest soil"@en
http://purl.jp/bio/103/nite/nbrc/cultures/00005426	"Umbelopsis angularis W. Gams & M. Sugiyama"	meo:MEO_0000126	"forest soil"@en
http://purl.jp/bio/103/nite/nbrc/cultures/00005427	"Pestalotiopsis funerea (Desmazieres) Steyaert"	meo:MEO_0000126	"forest soil"@en
http://purl.jp/bio/103/nite/nbrc/cultures/00005435	"Verticillium niveostratosum Lindau"	meo:MEO_0000126	"forest soil"@en
http://purl.jp/bio/103/nite/nbrc/cultures/00005453	"Penicillium expansum Link ex S.F. Gray"	meo:MEO_0000126	"forest soil"@en
http://purl.jp/bio/103/nite/nbrc/cultures/00005764	"Penicillium jensenii Zaleski"	meo:MEO_0000126	"forest soil"@en

データベースの統合が進むと



共通する要素を重ね合わせることで、異なる種類の情報をつなげたネットワーク(=グラフ構造)を作る

グラフの部分構造を検索して抽出することで、2つの事柄等の関係性を見出す

(AJACS資料より　https://github.com/AJACS-training/AJACS67/tree/master/01_minowa)

NBDCの提供するサービス

Catalog

- Integbioデータベースカタログ
  - データベースのデータベース

Archive

- 生命科学系データベースアーカイブ
  - データベースの安定的な長期保存

Cross search

- 生命科学データベース横断検索
  - データベース内の情報を探す

NBDC RDF Portal

- NBDC RDFポータル
  - よりデータ統合に向けた整備

TogoDX

- TogoDX
  - 色んなデータベース情報を一目で分かるようにする

生命科学のデータベースをまとめて俯瞰・探索できるサイト

https://togodx.dbcls.jp/human/

「togodx」で検索

遺伝子

タンパク質

タンパク質
立体構造

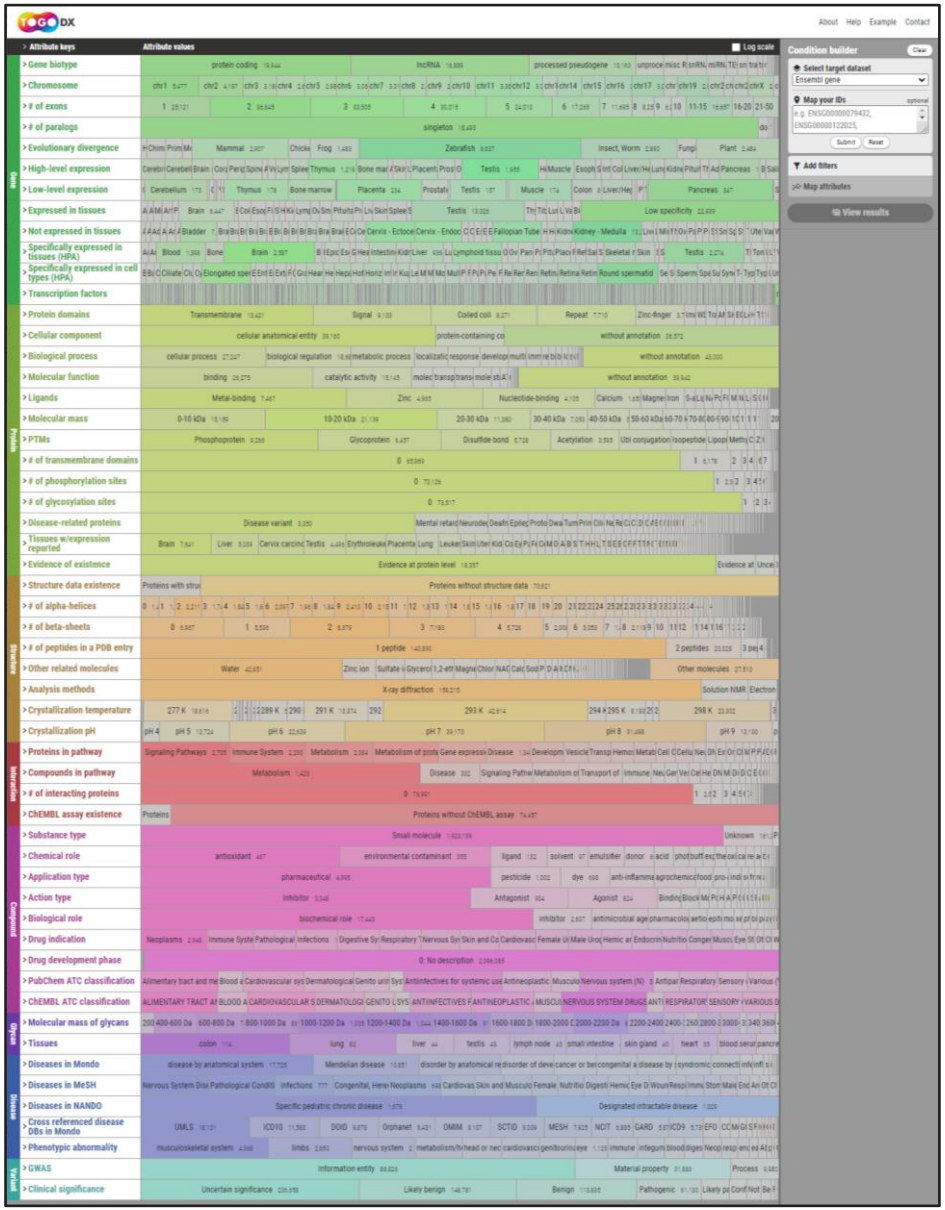
相互作用

化合物

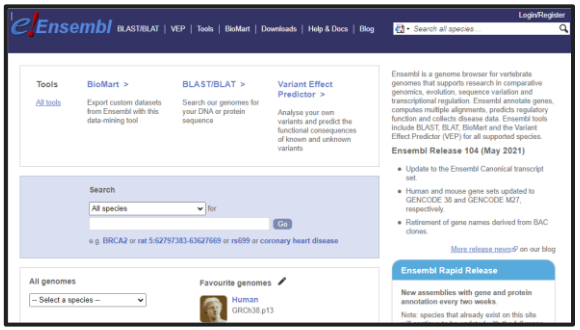
糖鎖

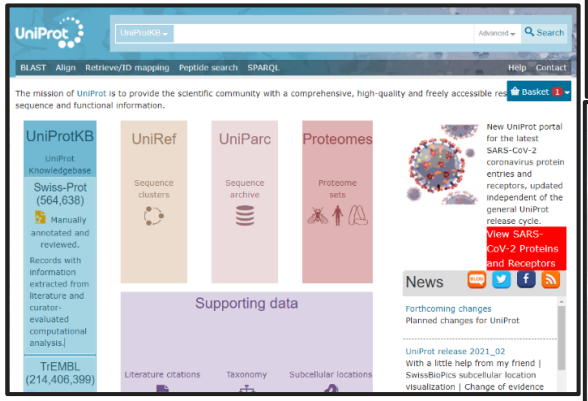
疾患

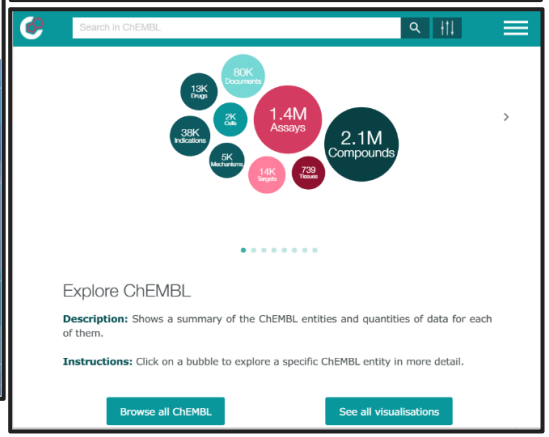
バリアント

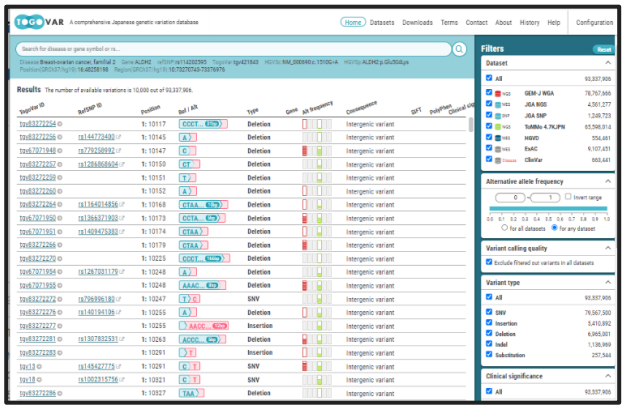


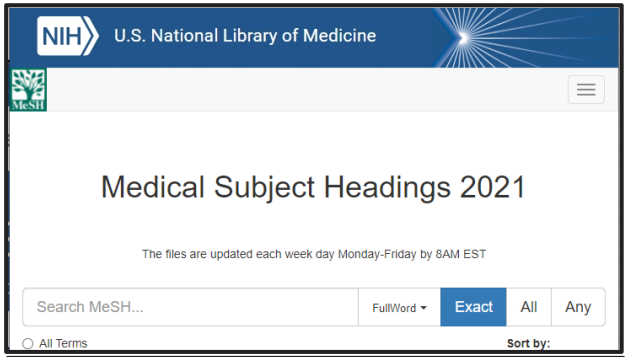
TogoDXで検索できるサイト

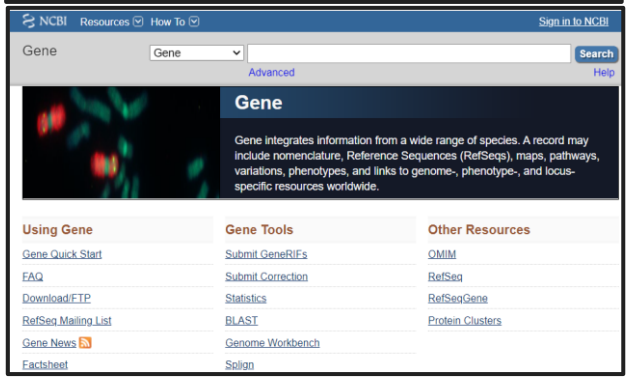






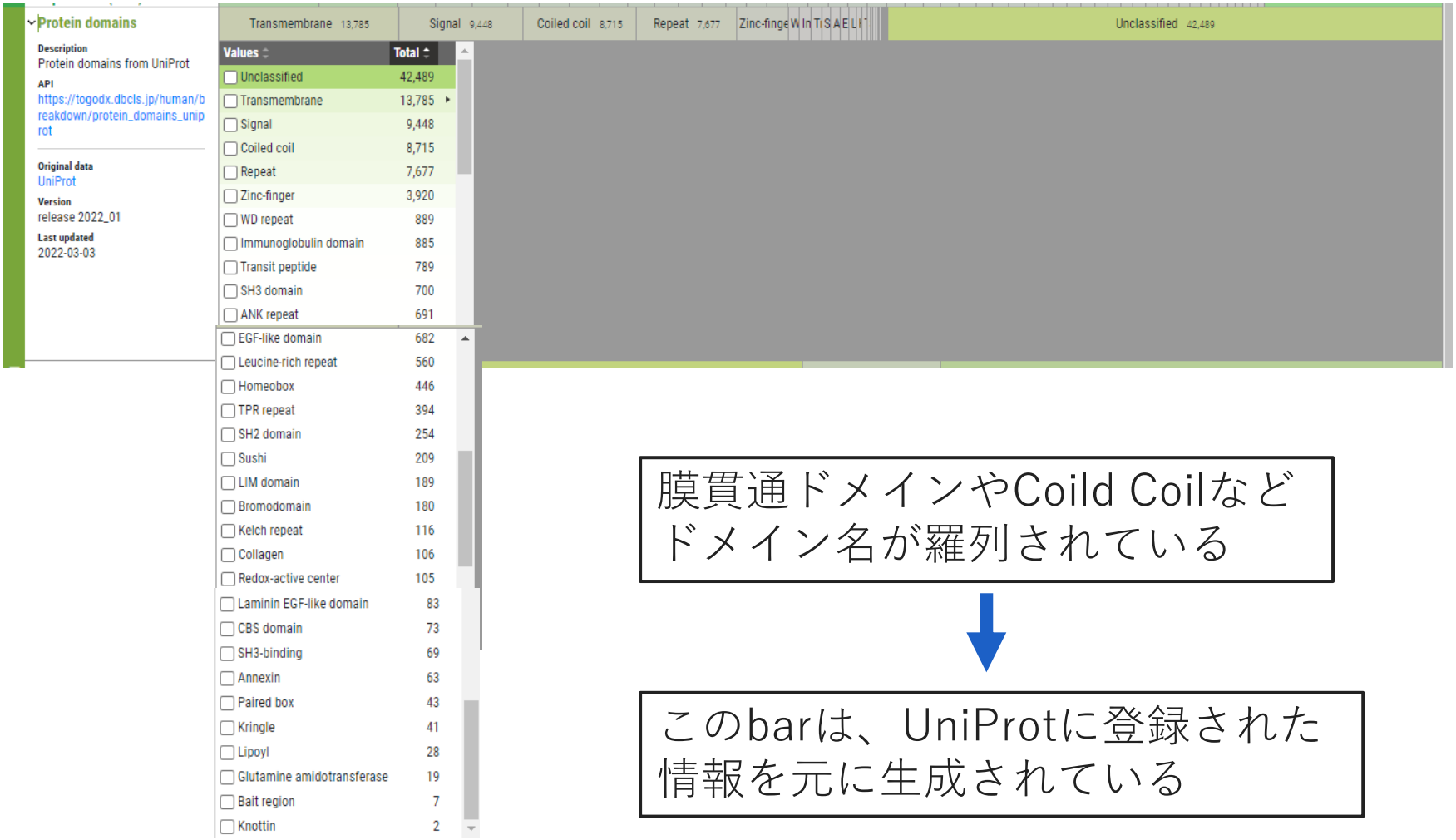






などなど.....

各barについて (例:Protein domain)

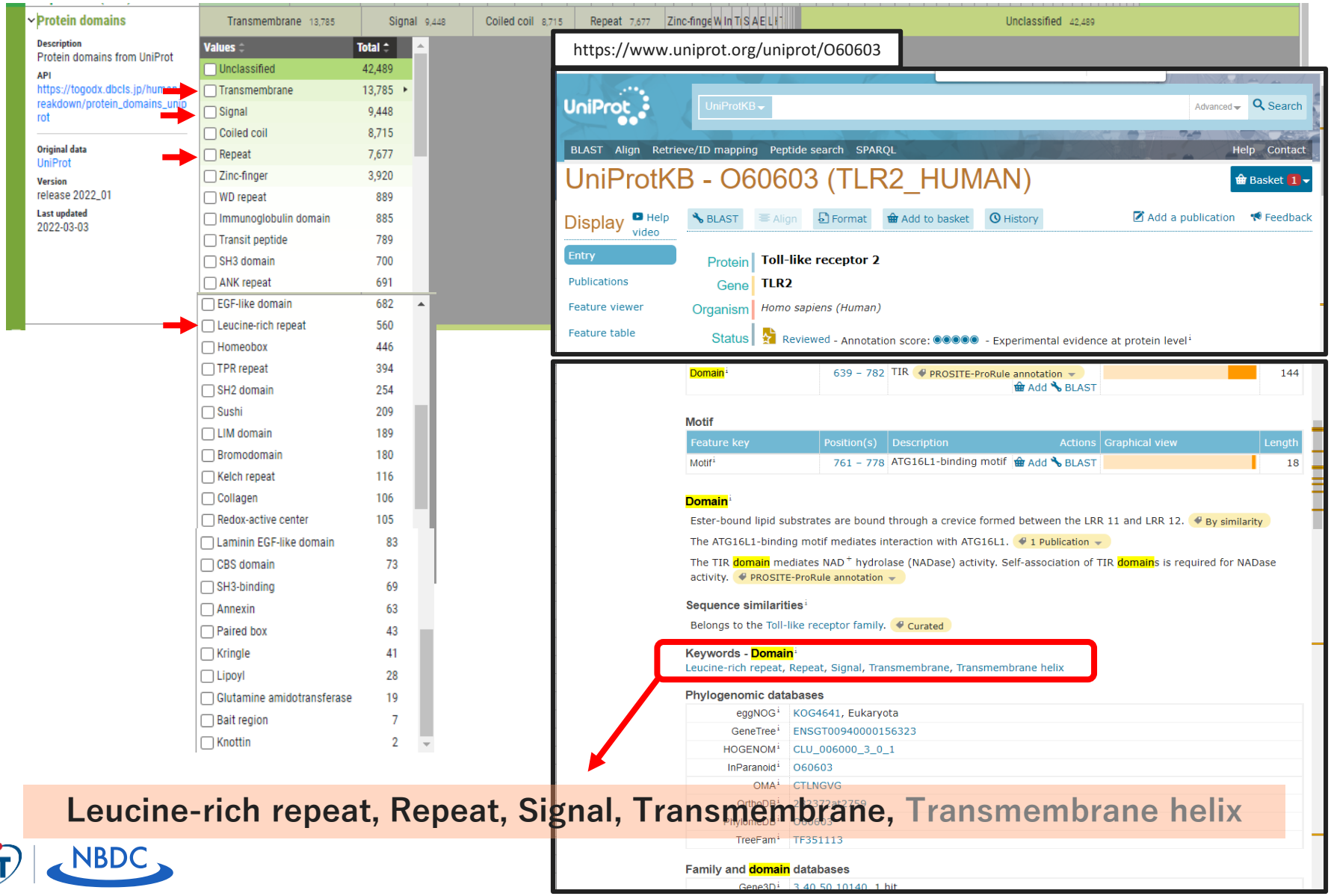


膜貫通ドメインやCoild Coilなど
ドメイン名が羅列されている

↓

このbarは、UniProtに登録された
情報を元に生成されている

各barについて (例:Protein domain)



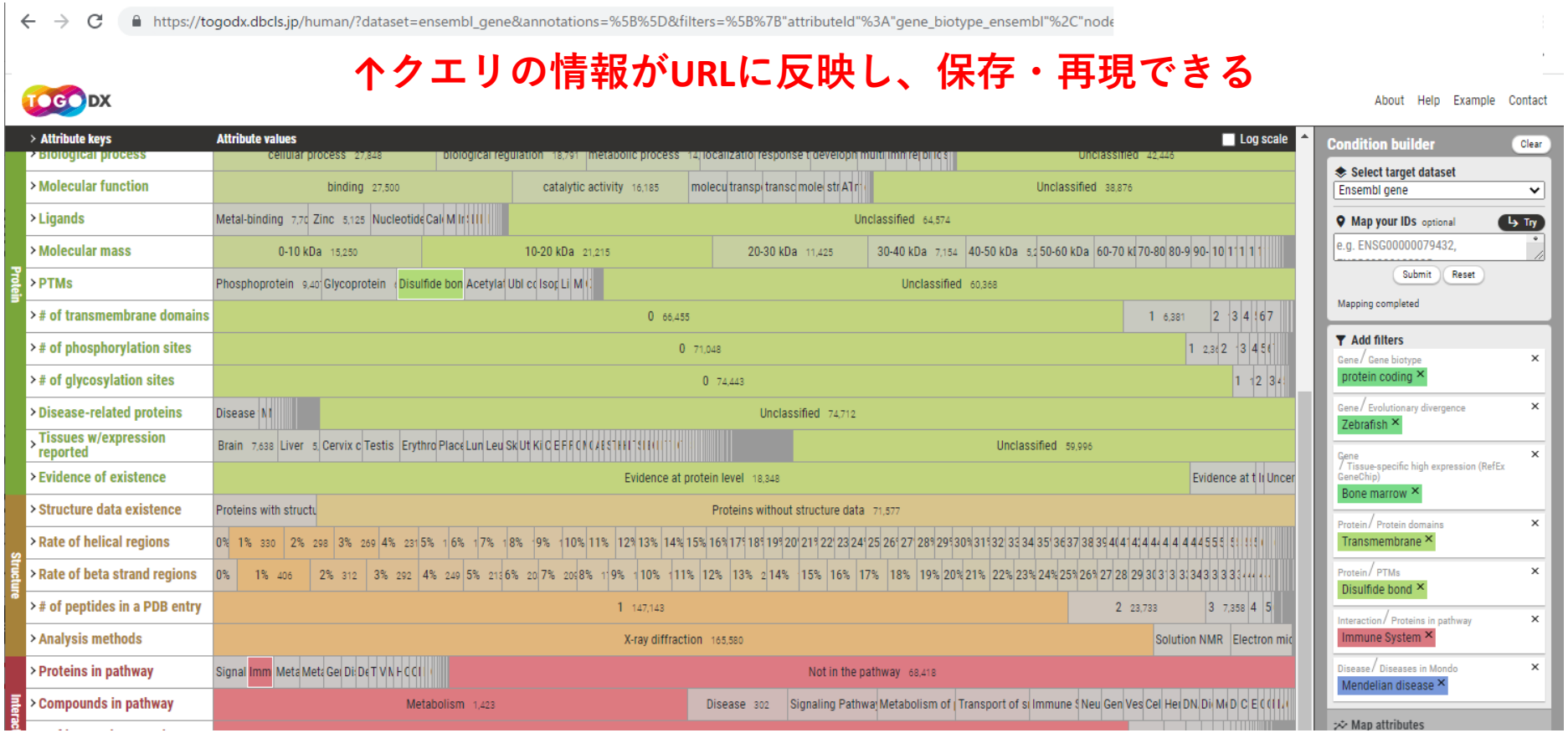
Leucine-rich repeat, Repeat, Signal, Transmembrane, Transmembrane helix

実際に使ってみる「複数のbarで絞りこむ」

1, "Select target category" (IDリストの軸)を選択
　⇒Ensembl geneを選択
2, “Attribute values”を選択(**異なるbarは”and”、同じbarは”or”**)
　⇒Gene/protein coding; Gene/Zebrafish; Gene/Bone marrow; Protein/Transmembrane;
　　Disulfide bond; Immune System; Mendelian disease

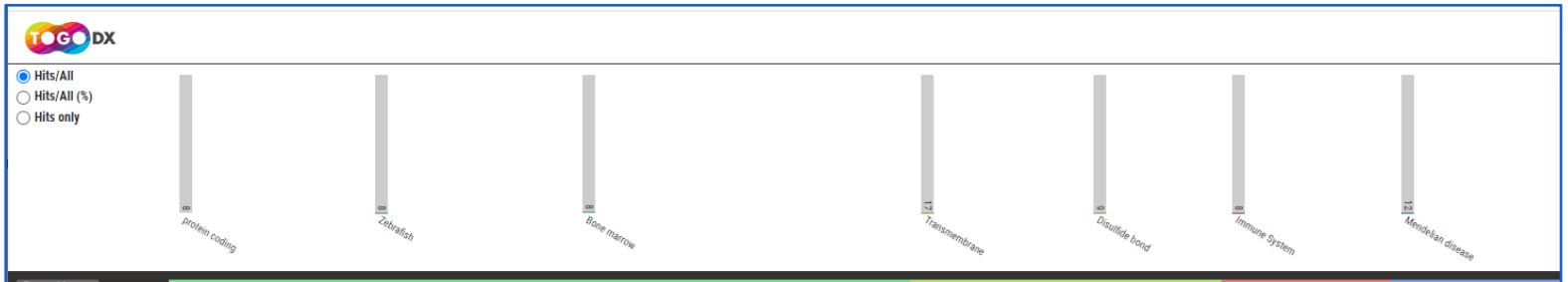
書き下すと、、、

「**遺伝子に紐づくエントリーで** ●タンパク質をコード ●Zebrafishと分岐した時代から存在 ●骨髄で高発現 ●膜貫通ドメインを持つ ●ジルスルフィド結合を持つ ●免疫システムに関与 ●メンデル遺伝病に関与**を満たすものは？**」



実際に使ってみる「複数のbarで絞りこむ」

結果



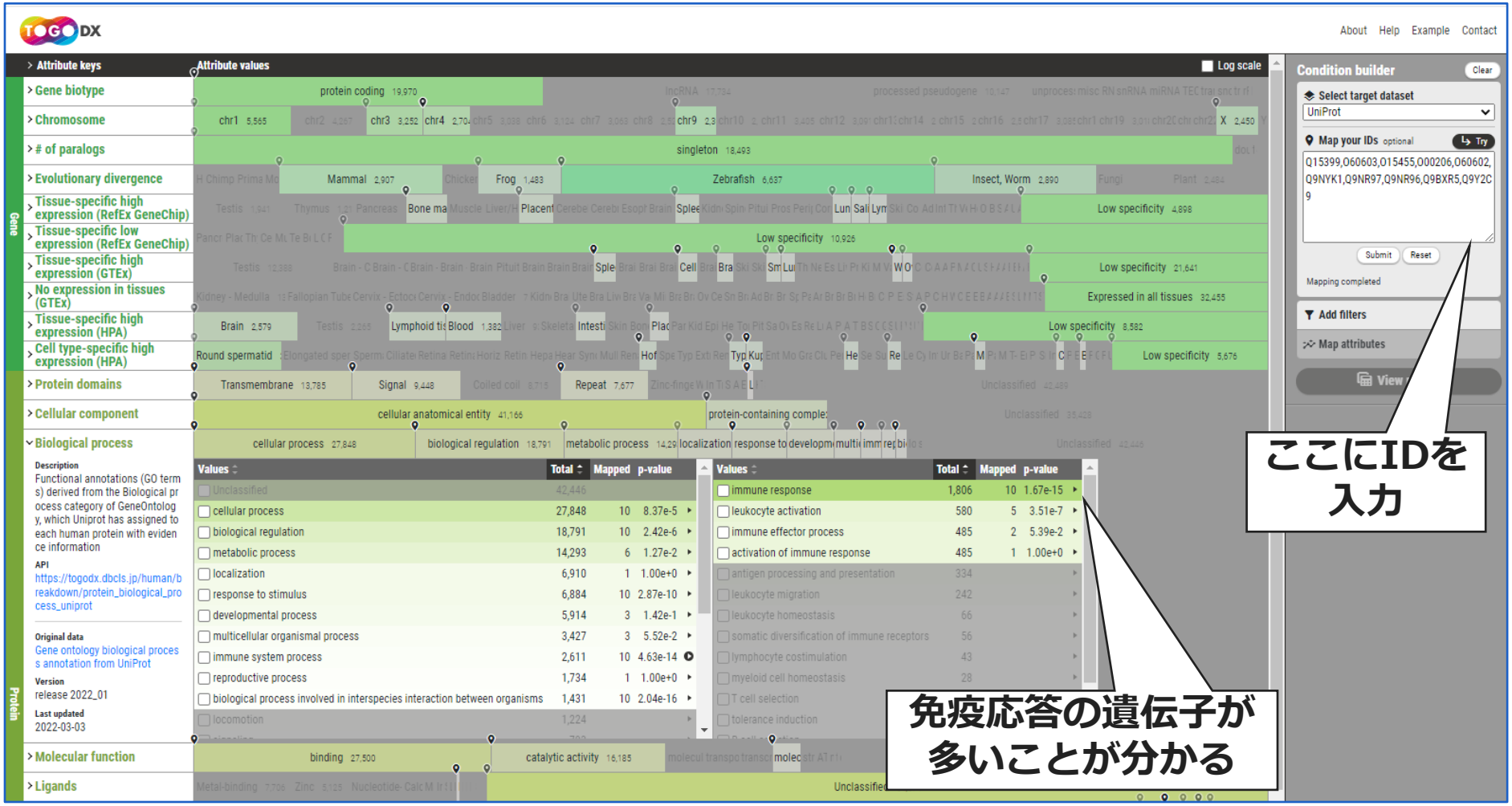
Ensembl gene	Ensembl gene Gene biotype	NCBI gene Evolutionary divergence	NCBI gene Tissue-specific high expression (RefEx GeneChip)	UniProt Protein domains	UniProt PTMs	UniProt Proteins in pathway	Mondo Diseases in Mondo
ENSG00000259207 ITGB3	ENSG00000259207 protein coding	3690 Zebrafish	3690 Bone marrow	I3L4X8 Transmembrane P05106 Transmembrane	P05106 Disulfide bond	P05106 Immune System	0030996 Mendelian disease 0031009 Mendelian disease
ENSG00000180871 CXCR2	ENSG00000180871 protein coding	3579 Zebrafish	3579 Bone marrow	C9J1J7 Transmembrane C9J2F9 Transmembrane C9JG19 Transmembrane C9JW47 Transmembrane P25025 Transmembrane	P25025 Disulfide bond	P25025 Immune System	0030374 Mendelian disease
ENSG00000177663 IL17RA	ENSG00000177663 protein coding	23765 Zebrafish	23765 Bone marrow	Q96F46 Transmembrane	Q96F46 Disulfide bond	Q96F46 Immune System	0013500 Mendelian disease
ENSG00000005961 ITGA2B	ENSG00000005961 protein coding	3674 Zebrafish	3674 Bone marrow	A0A3B3IU79 Transmembrane P08514 Transmembrane	P08514 Disulfide bond	P08514 Immune System	0008552 Mendelian disease 0031332 Mendelian disease
ENSG00000147168 IL2RG	ENSG00000147168 protein coding	3561 Zebrafish	3561 Bone marrow	P31785 Transmembrane	P31785 Disulfide bond	P31785 Immune System	0010315 Mendelian disease 0010730 Mendelian disease
ENSG00000119535 CSF3R	ENSG00000119535 protein coding	1441 Zebrafish	1441 Bone marrow	H0YE86 Transmembrane Q99062 Transmembrane	Q99062 Disulfide bond	Q99062 Immune System	0008092 Mendelian disease 0014865 Mendelian disease
ENSG00000160712 IL6R	ENSG00000160712 protein coding	3570 Zebrafish	3570 Bone marrow	H0Y949 Transmembrane P08887 Transmembrane	P08887 Disulfide bond	P08887 Immune System	0030069 Mendelian disease
ENSG00000105369 CD79A	ENSG00000105369 protein coding	973 Zebrafish	973 Bone marrow	M0QX61 Transmembrane P11912 Transmembrane	M0QX61 Disulfide bond P11912 Disulfide bond	P11912 Immune System	0013288 Mendelian disease

Ensambleの遺伝子のなかで、、、
- ●タンパク質をコード
- ●Zebrafishと分岐した時代から存在
- ●骨髄で高発現
- ●膜貫通ドメインを持つ
- ●ジルスルフィド結合を持つ
- ●免疫システムに関与
- ●メンデル遺伝病

この条件を満たす遺伝子は、8つ
- ・ITGB3：integrin subunit beta 3
- ・CXCR2：ケモカイン受容体
- ・IL17RA：interleukin 17 receptor A
- ・ITGA2B：integrin subunit alpha 2b
- ・IL2RG：interleukin 2 receprot受容体γ鎖
- ・CSF3R：血管新生因子レセプター
- ・IL6R： interleukin 6 receptor
- ・CD79A： B細胞マーカーの一つ

実際に使ってみる「手持ちのIDを調べる」

自然免疫に関するToll like receptor (TLR1-10)までのUniProt IDをMap your IDsに渡す



どんなIDセットが入力されているのかが分かる

説明動画もあります!!

TogoDX/Humanの使い方

目次

- 画面構成
  - Explorer画面
    - 属性詳細
  - Results画面
- 操作方法
  - Add filtersを用いた探索方法
  - Map Attributesを用いた探索方法
  - Map your IDsを用いた探索方法
- その他
  - ページヘッダ
  - 初期化
  - IDのリンク関係

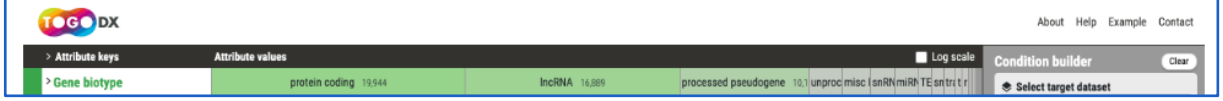
是非一度アクセス してください！

動画マニュアル

TogoDX/human 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

画面構成

Explorer画面



NBDCの提供するサービス

Catalog

- Integbioデータベースカタログ
  - データベースのデータベース

Archive

- 生命科学系データベースアーカイブ
  - データベースの安定的な長期保存

Cross search

- 生命科学データベース横断検索
  - データベース内の情報を探す

NBDC RDF Portal

- NBDC RDFポータル
  - よりデータ統合に向けた整備

TogoDX

- TogoDX
  - 色んなデータベース情報を一目で分かるようにする

つかわれてこそ価値が生まれますので
ぜひ一度使ってみて下さい