

TogoCoord (仮)



生命科学における様々なレイヤーの配列座標の変換を目的とした取り組み

守屋勇樹¹、藤澤貴智²、山本泰智¹

1) 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)

2) 国立遺伝学研究所

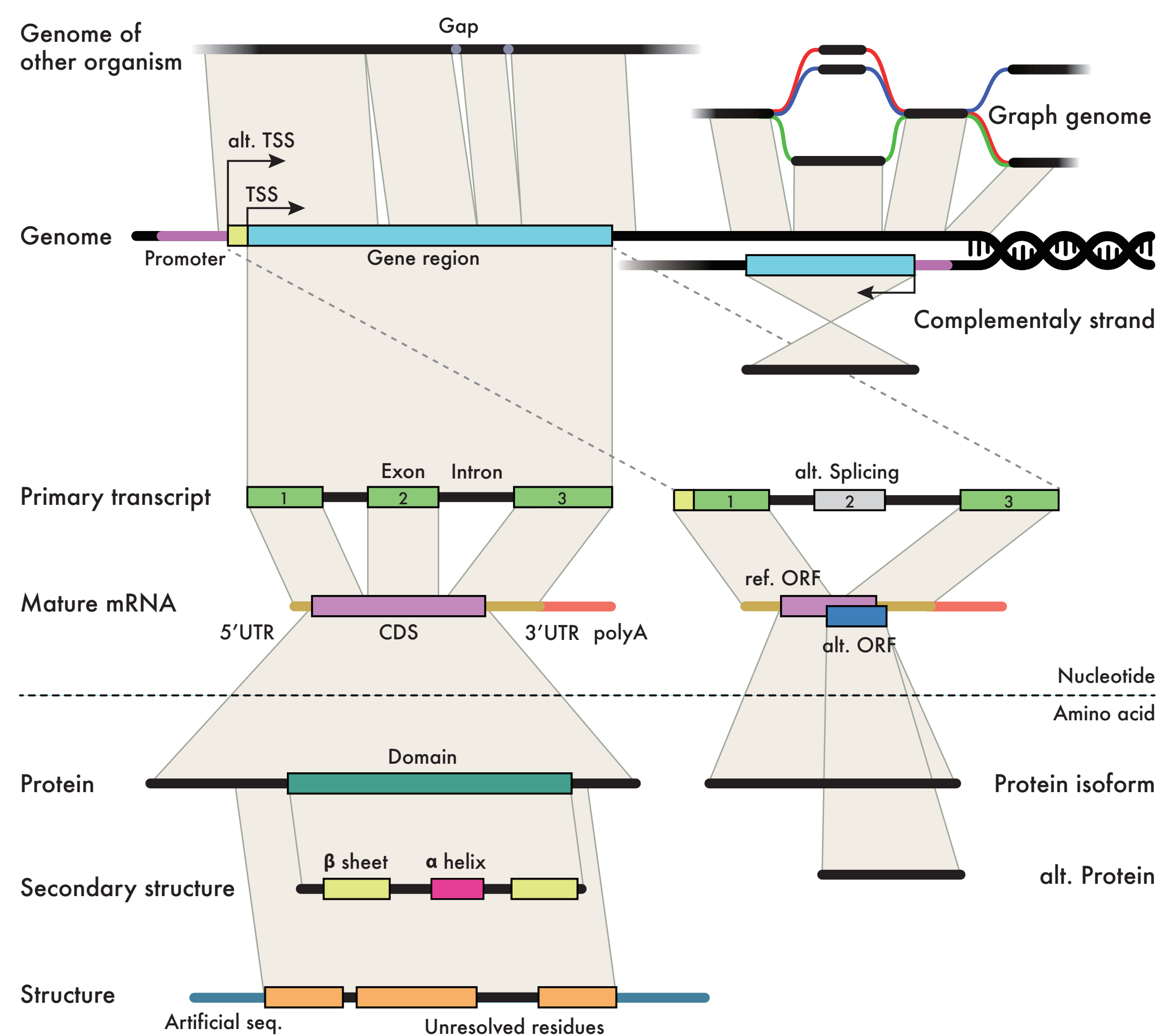
moriya@dbcls.rois.ac.jp

生命科学分野では、ゲノム、転写産物、タンパク質といった異なるレイヤーで配列情報を扱うため、複数の座標系が存在する。代表的なものとして、ゲノム座標、pre-mRNA座標、CDS座標、アミノ酸座標などが挙げられる。研究現場では、これらの座標系間の相互変換が頻繁に求められ、例えばゲノム上のバリエーションをUniProtのタンパク質配列やPDBの立体構造にマッピングする際や、生物種をまたいだオースログ間のバリエーション比較におい

て不可欠である。特に、範囲を範囲へ変換する際には、エキソン・イントロン構造による座標の不連続性や、遺伝子の逆位なども考慮しなければならず、処理は複雑になる。加えて、アイソフォーム間の差異や挿入・欠失を伴う変異の扱いも問題を複雑化させる要因であり、これまでは個別対応に依存していた。こうした課題に対処するため、汎用的かつ再利用可能な座標変換サービス、TogoCoord (仮) の取り組みを開始した。

いろいろな座標系

生命科学における様々なレイヤーの配列は、セントラルドグマ的に、または進化的に連続しており、関連性を辿ることで互いに変換することが可能なのである。



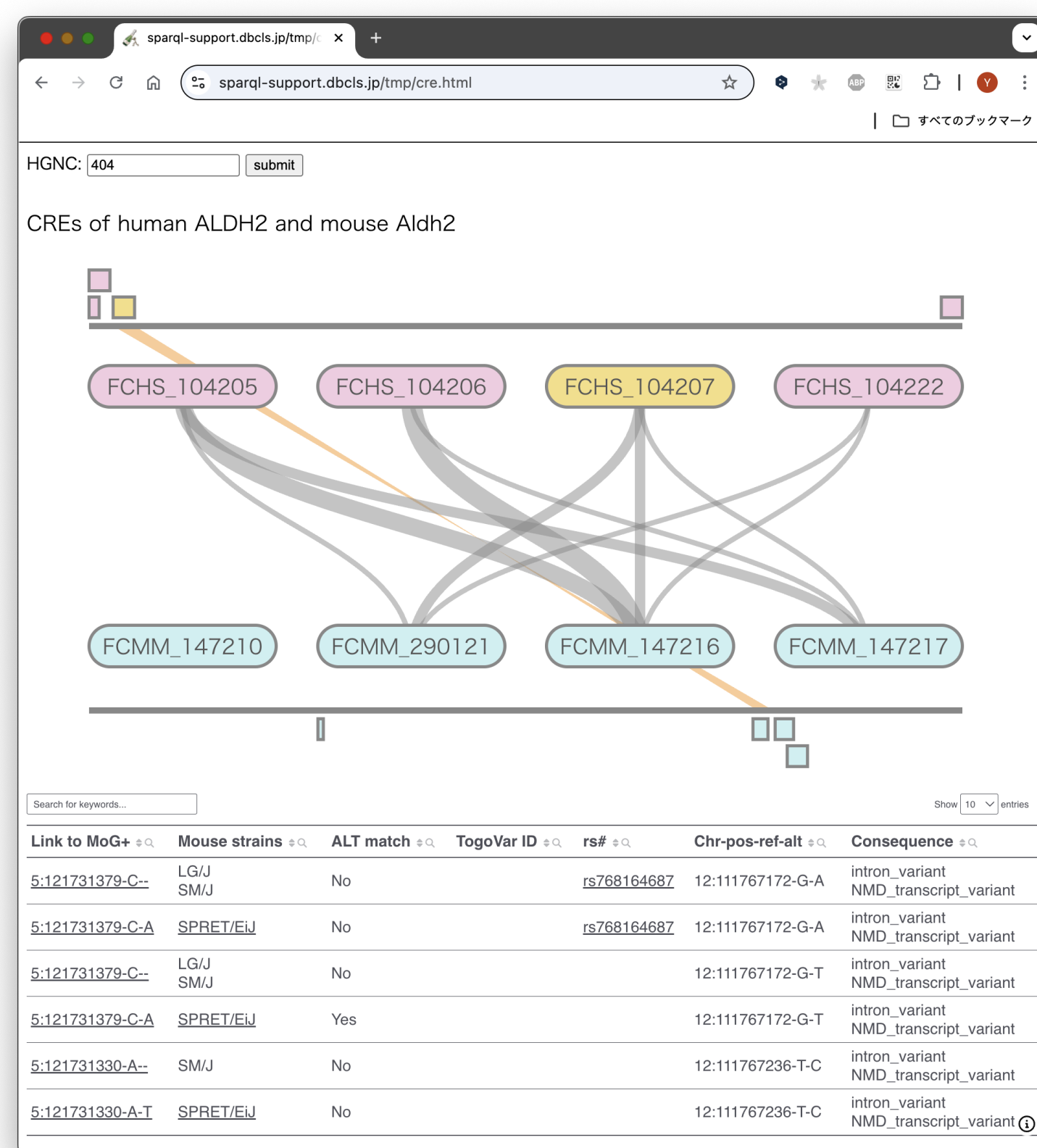
座標変換はアノテーションの伝播

異なるレイヤーの座標へと落とし込むことで、それぞれのレイヤーのアノテーションを他のレイヤーのアノテーションと比較することが可能となる。

また、アノテーションの少ない生物においても、他の生物との座標変換を経ることにより、より多くのアノテーション情報を利用できる。

ヒトとマウスのゲノム座標

座標変換と共有オースログ転写因子から、ヒトとマウスの対応するCREs上のVariantを比較することで、実験マウスの探索を促進。(Fanta.bio, MoG+, TogoVar 連携)



UCSCの配布しているchain fileを利用した liftover による座標変換

これらの座標変換は統一的な仕組みが提供されているわけではなく、個々の実装において苦勞することも多い。TogoCoord のような座標に関わる包括的なサービスがあれば車輪の再開発を防ぐことが期待できる。

Location ID と faldo JSON-LD

<https://github.com/dbclj/universal-public-genome>

GCA000000000-J000000:467

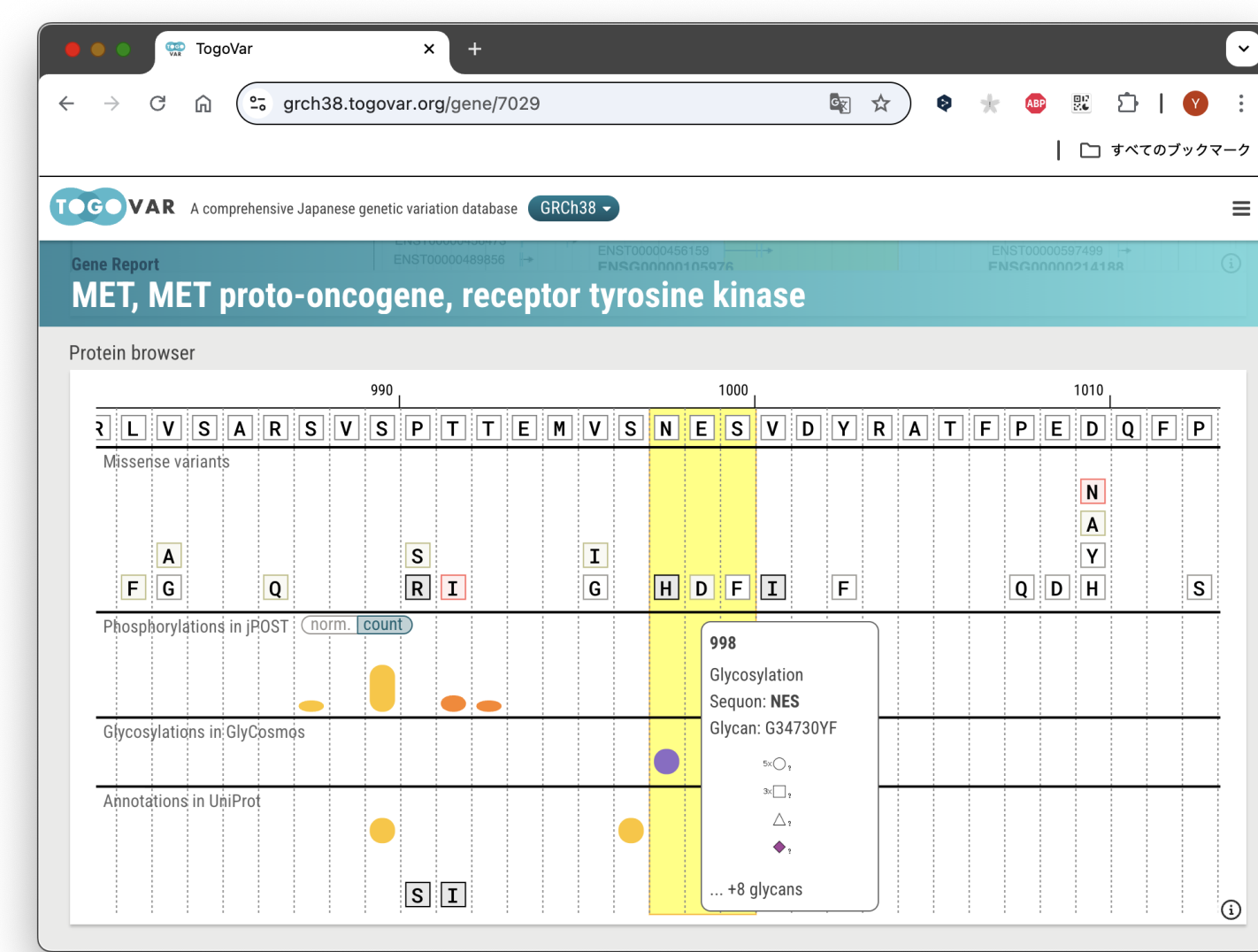
```
{
  "@context": "http://example.org/context/faldo.jsonld",
  "id": "http://example.org/GCA000000000-J000000-location",
  "faldo:location": {
    "type": "faldo:ExactPosition",
    "faldo:position": 467,
    "faldo:reference": "insdc:J000000"
  }
}
```

GCA000000000-J000000:complement(join(2691..4571,4918..5163))

```
{
  "@context": "http://example.org/context/faldo.jsonld",
  "id": "http://example.org/GCA000000000-J000000-location",
  "faldo:location": {
    "type": "faldo:ListOfRegions",
    "faldo:strand": "faldo:NegativeStrand",
    "faldo:member": [
      {
        "type": "faldo:Region",
        "faldo:begin": {
          "type": "faldo:ExactPosition",
          "faldo:reference": "insdc:J000000",
          "faldo:position": 2691,
          "faldo:end": {
            "type": "faldo:ExactPosition",
            "faldo:reference": "insdc:J000000",
            "faldo:position": 4571,
            "faldo:order": 1
          }
        },
        "type": "faldo:Region",
        "faldo:begin": {
          "type": "faldo:ExactPosition",
          "faldo:reference": "insdc:J000000",
          "faldo:position": 4918,
          "faldo:end": {
            "type": "faldo:ExactPosition",
            "faldo:reference": "insdc:J000000",
            "faldo:position": 5163,
            "faldo:order": 2
          }
        }
      }
    ]
  }
}
```

ゲノム座標とアミノ酸座標

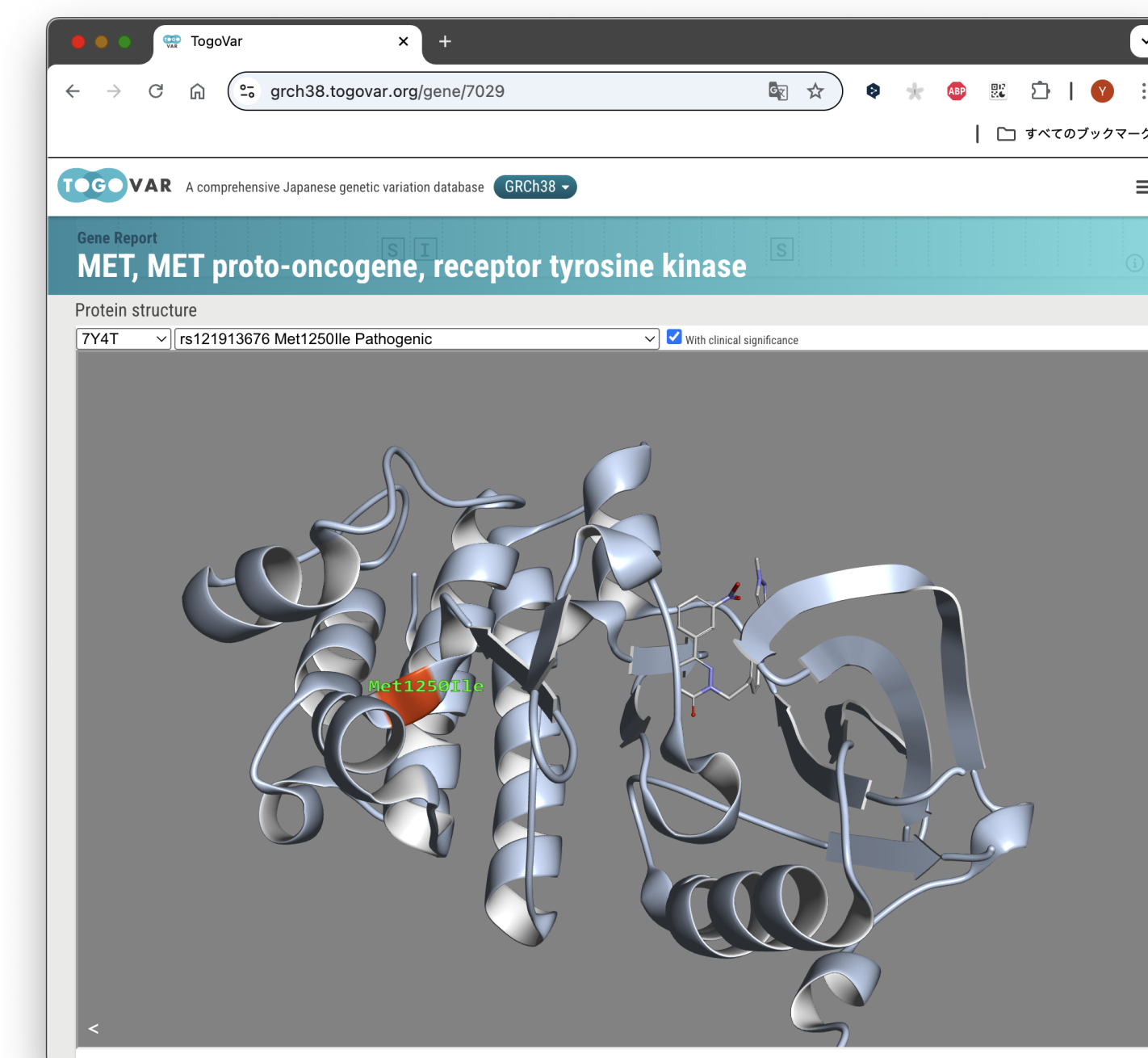
Variant が引き起こすアミノ酸変異と、元のタンパク質配列におけるリン酸化や糖鎖修飾への影響の可能性を可視化 (TogoVar)。



Ensembl Variant Effect Predictor (VEP) によるアノテーションに基づいた座標変換

ゲノム座標と構造座標

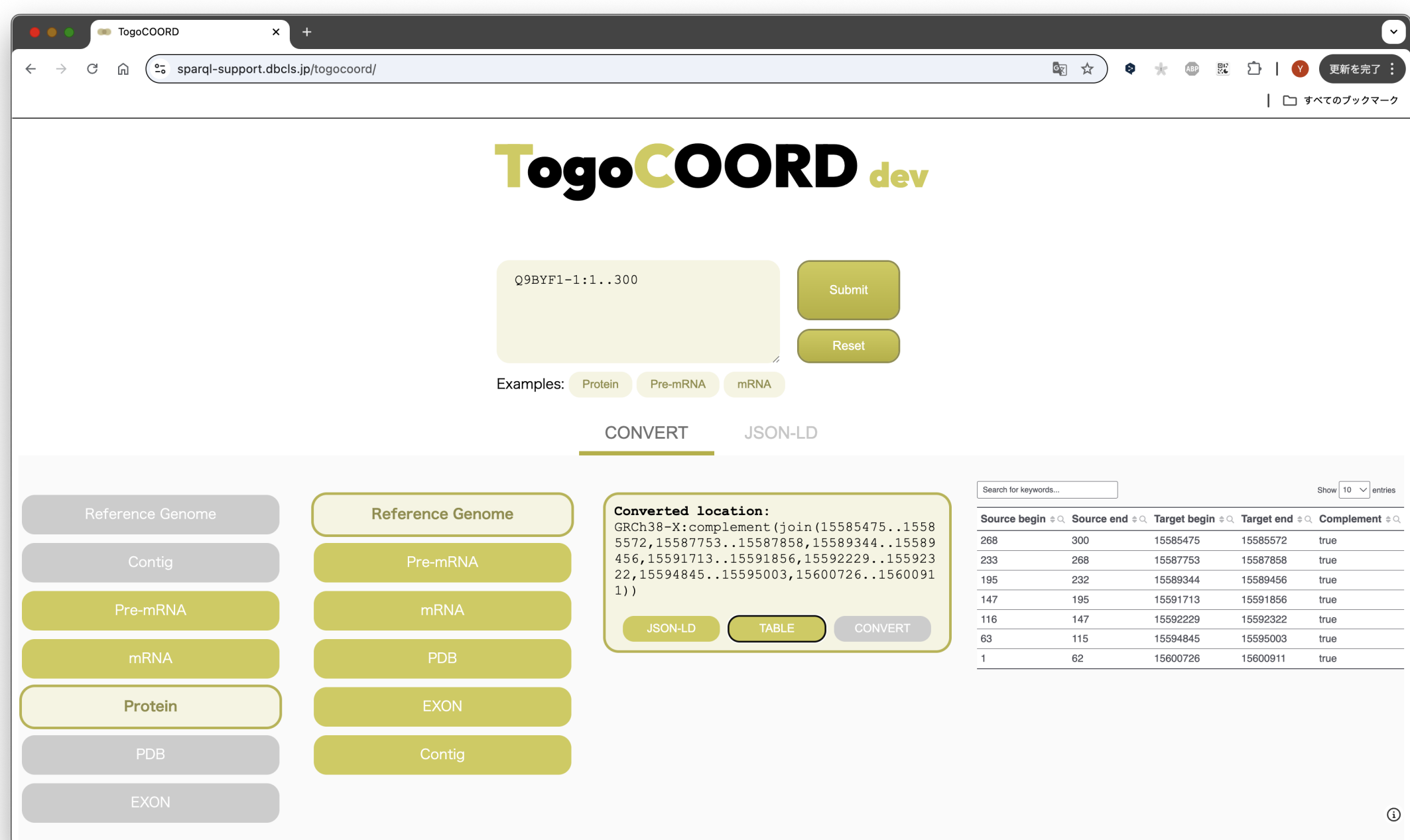
PDB へのマッピングにより Variant がタンパク質表面か内部かを判別。(TogoVar)



PDB の UniProt アラインメントに基づいて、信頼性の低いアミノ酸残基を避けて座標変換

実装とUI

座標間の対応を取るのためのデータや API を収集し、任意のレイヤーの任意のポジション (アノテーションの座標) を連続的に変換できるシステム構築を目的とし、そのため、INSDC Feature Table の座標表記を元にした、任意のポジションの組み合わせを ID として表現できる location ID を利用することとした。



タンパク質の範囲をゲノムに変換した例

展望

開発版ではヒトの一部のレイヤーのみを対象とし、実装を行ったが、対象とする生物種によっては座標対応のためのデータ取得やサービスが困難である。そのため、ゲノム座標を中心とした最低限の座標変換の標準化と、アノテーションの豊富な生物への LiftOver を目的とした開発を進める。



DBCLS
Database Center
for Life Science

Database Center for Life Science,
Joint Support-Center for Data Science Research, Research Organization of Information and Systems
Univ. of Tokyo Kashiwanoha-campus Station Satellite 6F, 178-4-4, Wakashiba, Kashiwa, Chiba 277-0871, JAPAN

Licensed under CC-BY 4.0 ©2023 Yuki Moriya (DBCLS)

