jPOSTdb: COVID-19データベースの構築

○吉沢明康[1)]、守屋勇樹[2)]、小林大樹[3)]、張智翔[4)]、奥田修二郎[5)]、田畑剛[1)6)]、河野信[2)7)]、幡野敦[3)]、高見知代[3)]、松本雅記[3)]、山ノ内祥訓[8)]、荒木令江[4)]、岩崎未央[6)]、杉山直幸[1)]、福島敦史[9)]、田中聡[10)]、五斗進[2)]、石濱泰[1)]

1)京都大学大学院薬学研究科　2)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
3)新潟大学大学院医歯学総合研究科　4)熊本大学大学院生命科学研究部　5)新潟大学医学部メディカルAIセンター　6)京都大学iPS細胞研究所
7)富山国際大学現代社会学部　8)熊本大学医学部附属病院　9)理化学研究所環境資源科学研究センター　10)Trans-IT

COVID-19とプロテオーム解析

COVID-19への対策には、ウイルス・宿主間相互作用などタンパク質レベルでの理解が必須である。

このためには、プロテオーム解析の有用性・重要性が高いが、質量分析を用いたプロテオーム解析には

- 実験系によってデータの意味が変化し得る
- 他のオミクス解析と異なった手続き（p-valueを用いない、target-decoy法という独自の方法）でFalse Discovery Rate (q-value)を計算する

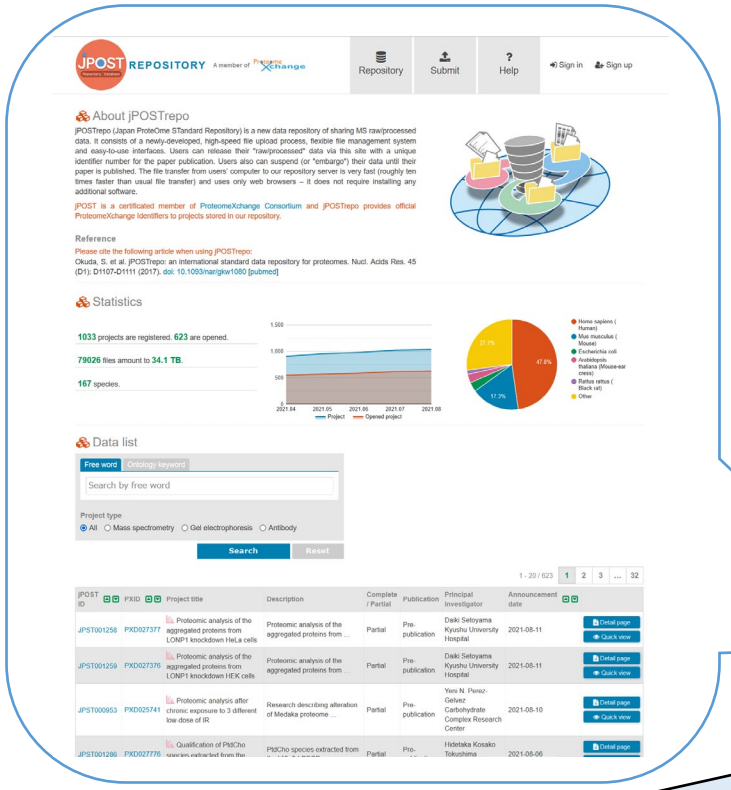
などの特徴があり、また実験系によってq-valueの基準が異なることもあるため、結果の単純な比較や他オミクスデータとの統一的な解釈が難しい。

jPOSTと再解析によるデータベース化

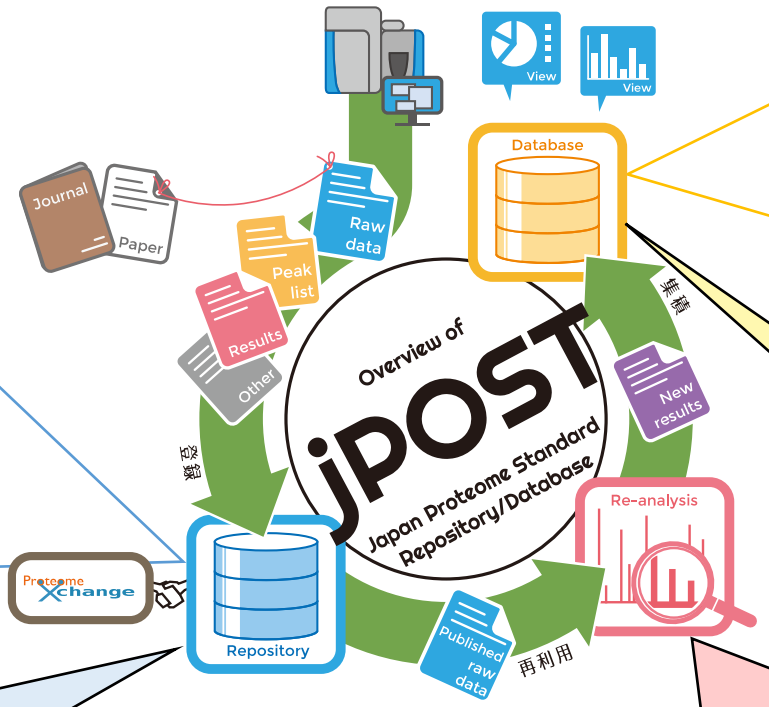
jPOSTはこのような問題に対応することを当初から念頭に置いて開発が行われてきたプロジェクトであり、リポジトリ部分 (jPOSTrepo) で収集した生データを、独自の統一的な基準に基づいて再解析し、その結果をデータベース (jPOSTdb) に収録している。

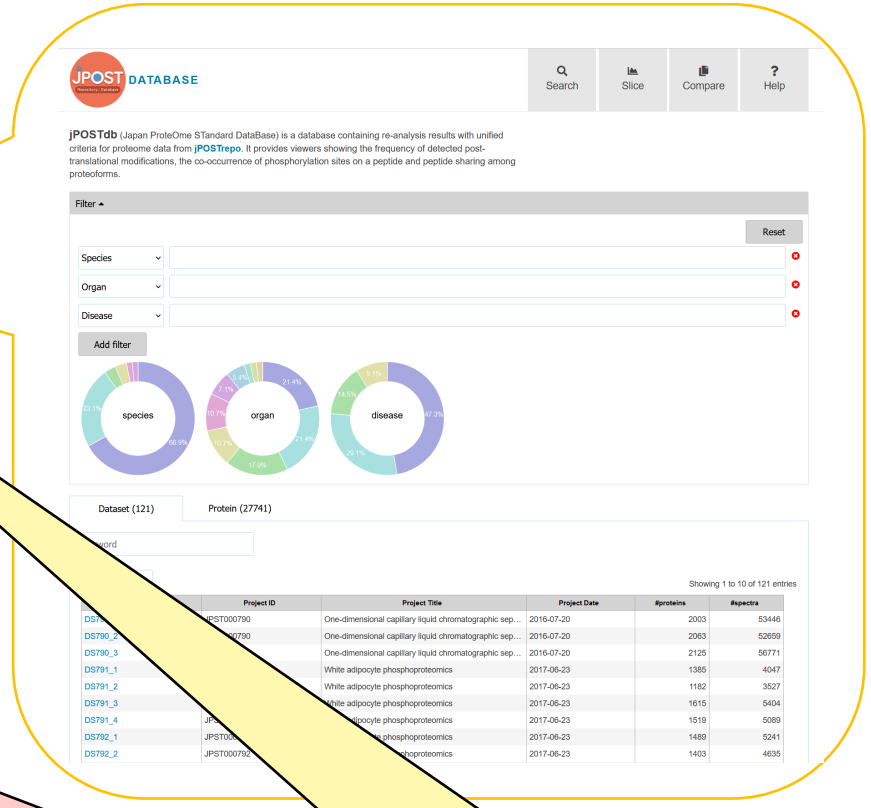
今回我々はそのリソース（再解析・jPOSTdb）を利用して、複数のCOVID-19研究の公開データを再解析した。得られた結果は、特徴を閲覧しやすい形でのデータベース化を進めている。

jPOSTdb: COVID-19データベースの構築



jPOSTの構成





今までに寄託されたデータ（2021年9月1日現在）：

- 約79,400ファイル（34.6TB）・170生物種
- 1,044プロジェクト中633プロジェクトが公開

- ［**利用方法解説文献**］Watanabe, Y., Yoshizawa, A.C., Ishihama, Y. and Okuda, S. The jPOST Repository as a Public Data Repository for Shotgun Proteomics. ***Methods in Molecular Biology***, **2259**, Chapter 20: 309-322 (2021) [PubMed: 33687724] [doi: 10.1007/978-1-0716-1178-4_20]

すべてのデータを統一的な方法で解析

- ピーク検出：MaxQuant
- データベース検索：MaxQuant, X!Tandem, Comet
- UniScore（旧名 jPOST score、論文執筆中）を用いて、3エンジンからの結果を統合
- q-value≦1%を満たすペプチドのみを用いる

生データのプロジェクトごとに結果を管理

- 各プロジェクトごとの結果 (Slice)
- それらの全体集合 (Globe)
- Sliceは自由にGlobeから抽出し、組み合わせて結果表示ができる
- メタデータを使ってSliceを検索

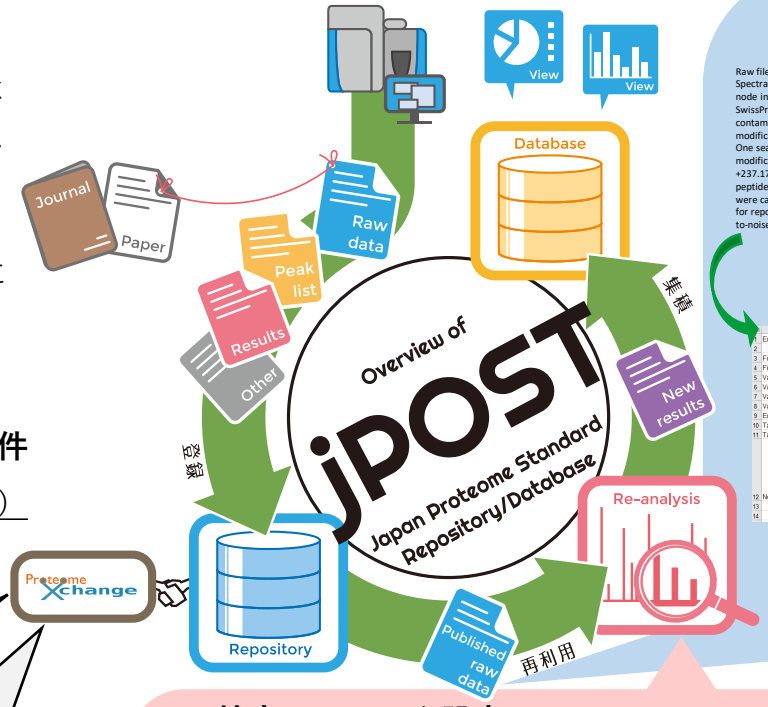
https://jpostdb.org/

jPOSTdb: COVID-19データベースの構築

COVID-19のプロテオーム解析データ

COVID-19の解析データは（現時点では）jPOSTには寄託されていないため、ProteomeXchange加盟リポジトリに寄託されたデータを用いた。

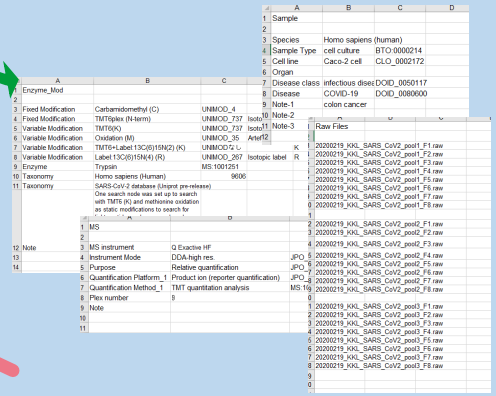
- 〔**方針**〕各プロジェクトのメタデータをjPOST形式に変換し、jPOST再解析プロトコルに流し込む
- 2020年7月時点で、生データがProteomeXchange加盟リポジトリで公開されていた研究プロジェクト：**20件**
- うち、再解析プロトコルで対応できる（解析可能な）プロジェクト：**10件**



- **メタデータの整備（例）**

（ https://www.ebi.ac.uk/pride/archive/projects/PXD017710 ）

Raw files were analyzed using Proteome Discoverer (PD) 2.4 software (ThermoFisher Scientific). Spectra were selected using default settings and database searches performed using SequestHT node in PD. Database searches were performed against trypsin digested Homo Sapiens SwissProt database, SARS-CoV-2 database (Uniprot pre-release) and FASTA files of common contaminants (`contaminants.fasta` provided with MaxQuant) for quality control. Fixed modifications were set as TMT6 at the N-terminus and carbamidomethyl at cysteine residues. One search node was set up to search with TMT6 (K) and methionine oxidation as static modifications to search for light peptides and one search node was set up with TMT6+K8 (K, +237.177), Arg10 (R, +10.008) and methionine oxidation as static modifications to identify heavy peptides. Searches were performed using Sequest HT. After search, posterior error probabilities were calculated and PSMs filtered using Percolator using default settings. Consensus Workflow for reporter ion quantification was performed with default settings, except the minimal signal-to-noise ratio was set to 5.



文章形式のメタデータ（論文のMethodのコピーであることが多いが、COVID-19の場合は論文未発表であるケースも多い）

↓〔**変換**〕

jPOST形式メタデータ

この変換作業はマニュアルで（人手で）進める

- **COVID-19のプロテオーム解析・生データ**が公開されている（主に英PRIDE。一部は米MassIVE、中IProXにも）
- ただし**メタデータは文章記述形式**

国際標準のプロテオームデータリポジトリの連携組織（加盟6件、オブザーバー1件）
（jPOSTは2016年7月加盟）

- **検索パラメータ設定**

原則としてオリジナルと同一であるが、以下の2点は（jPOST再解析プロトコルにしたがって）変更している：

- mass tolerance: プレサーチして決定
- missed cleavage number: 1 or 2（リン酸化の場合）

- **データベース検索のための配列データベース**

SARS-CoV-2： Nextstrain (https://nextstrain.org/) 収録変異（2020年7月20日時点）をすべて反映（13,555配列）

human： Swiss-Prot + isoform (release2020_06)（42,383配列）

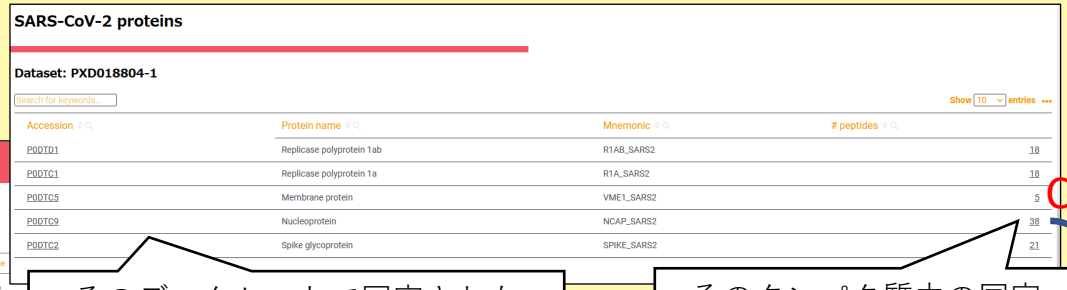
***Chlorocebus sabaeus*（ミドリザル）**： UniProt reference proteome (release2021_01)（19,229配列）

jPOSTdb: COVID-19データベースの構築

得られた結果とデータベース（開発中画面）

- q-value≦1%である配列として、**5,018,134個**のPSM (Peptide –Spectrum -Match) が得られた。
- この結果を基に、重複のないペプチドのリストを取得した。

データセット（プロジェクト）のリスト



そのデータセットで同定されたSARS-CoV-2のタンパク質のリスト

そのタンパク質中の同定されたペプチドの個数

タンパク質配列に同定ペプチドをマッピング

変異 (SAAV) マッピング

同定ペプチド情報

- jPOSTdbの既存の機能を用いてウイルスタンパク質の視点から情報を要約した。
- 宿主タンパク質との相互作用情報などの表示機能を追加中。

jPOSTdb: COVID-19データベースの構築

収録データの内容の例（再解析の効果）

PXD018357（ウイルス感染ヒト細胞のリン酸化プロテオーム）の再解析結果を、オリジナル論文[1]の解析結果と比較した。

解析条件

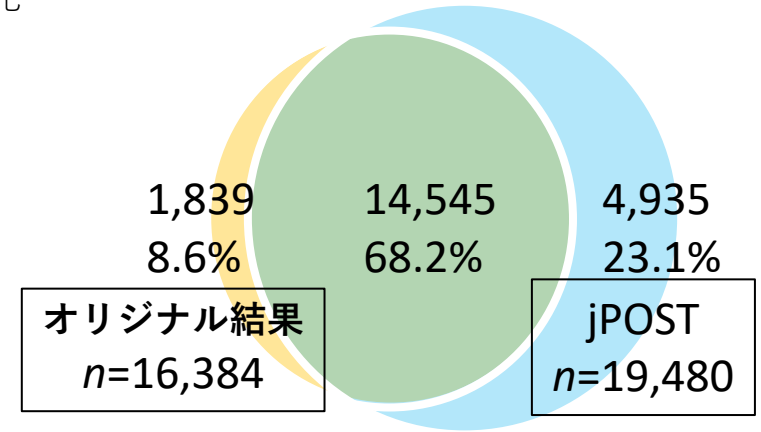
実験条件：
- 試料 … ヒト培養細胞 (Caco-2) ・SARS-CoV-2
- 酵素消化 … Trypsin+Lys-C

データベース検索条件（オリジナル論文）：
- 酵素条件 … Trypsin
- Missed cleavage number (mc) <=2
- PSMの結果に対してq-value<=1%

データベース検索条件（jPOST再解析）：
- 酵素条件 … Trypsin/P
- mc <=2
- PSM (q-value<=1%)からUniqueなペプチドのリストを作成、再度q-value<=1%

（**次ページ**）同定ペプチド上のリン酸化サイトの存在をPhosPep Analyzer [2] で検証し、その結果をオリジナルと比較した。
（オリジナルの結果は、公開データを基に再描画したもの）

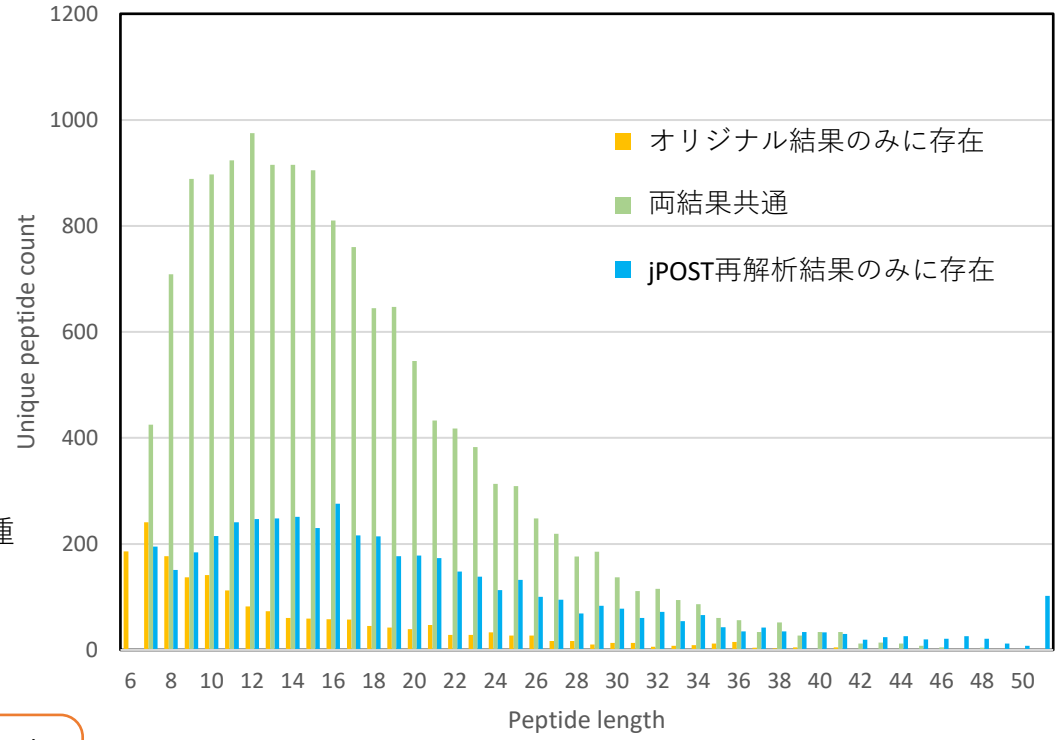
(A) 両解析結果のペプチド総数比較



(A) 同定されたペプチドのアミノ酸配列と修飾の種類・個数ごとにカウント。
　なおjPOSTでは「修飾同一種類個数・サイト違い」の結果も得ている（*n*=27,171）。

再解析によって、長い（＝信頼性の高い）ペプチドが、より多数同定された。

(B) 両解析結果のペプチド長ごとの比較



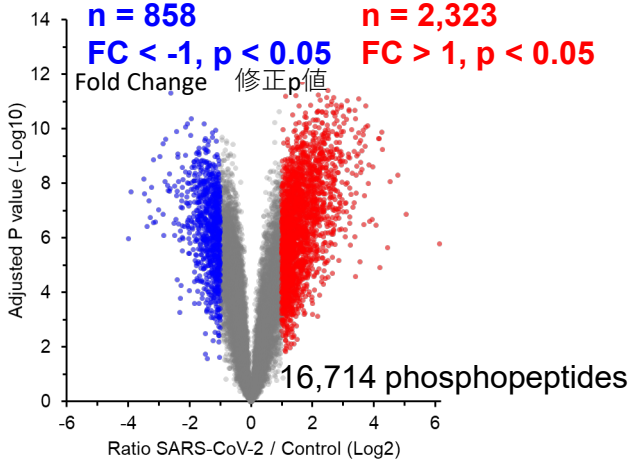
(B) (A)と同じ条件で、アミノ酸配列の長さごとにカウント。UniScoreは短いペプチドには不利に働くので、結果の個数が減る。特に配列長≦6の結果はすべて除外されている。

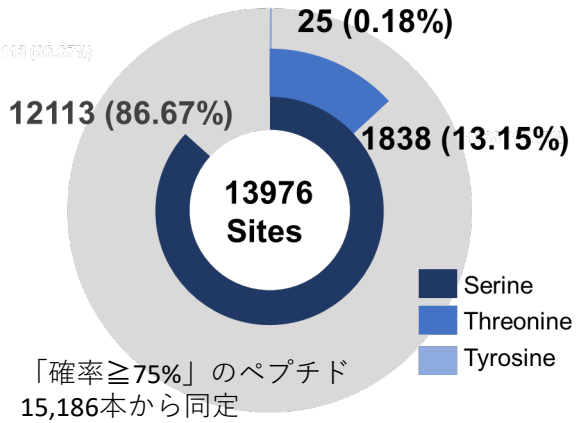
文献：
[1] Klann *et al.*, *Molecular Cell* **80**, 164 (2020)
[2] Nakagami *et al*. *Plant Physiol*. **153**, 1161 (2010)

jPOSTdb: COVID-19データベースの構築

(C) 発現量に変化のあったリン酸化ペプチドと同定されたリン酸化サイトの個数比較

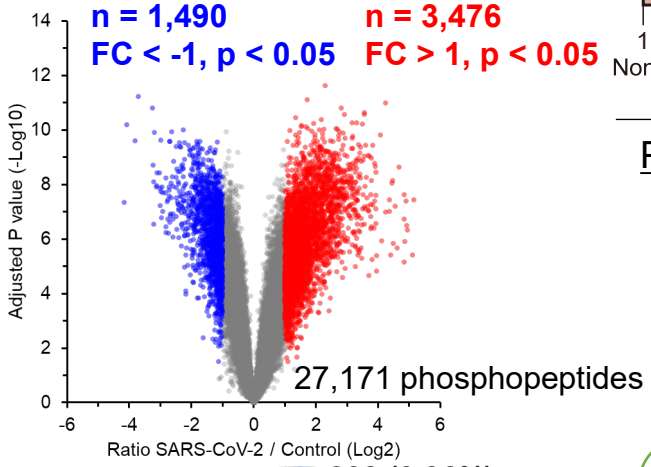
オリジナル (Mol Cell) の結果

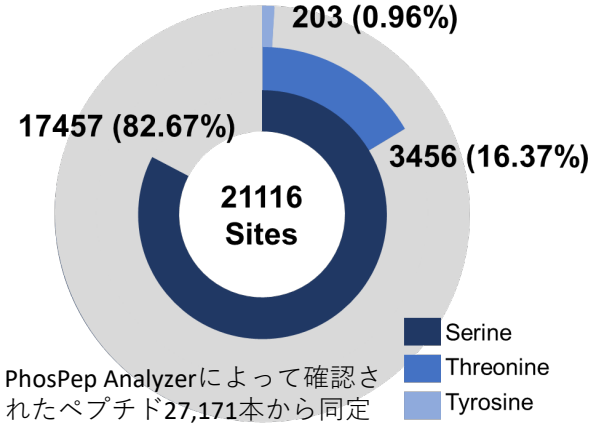




「確率≧75%」のペプチド15,186本から同定

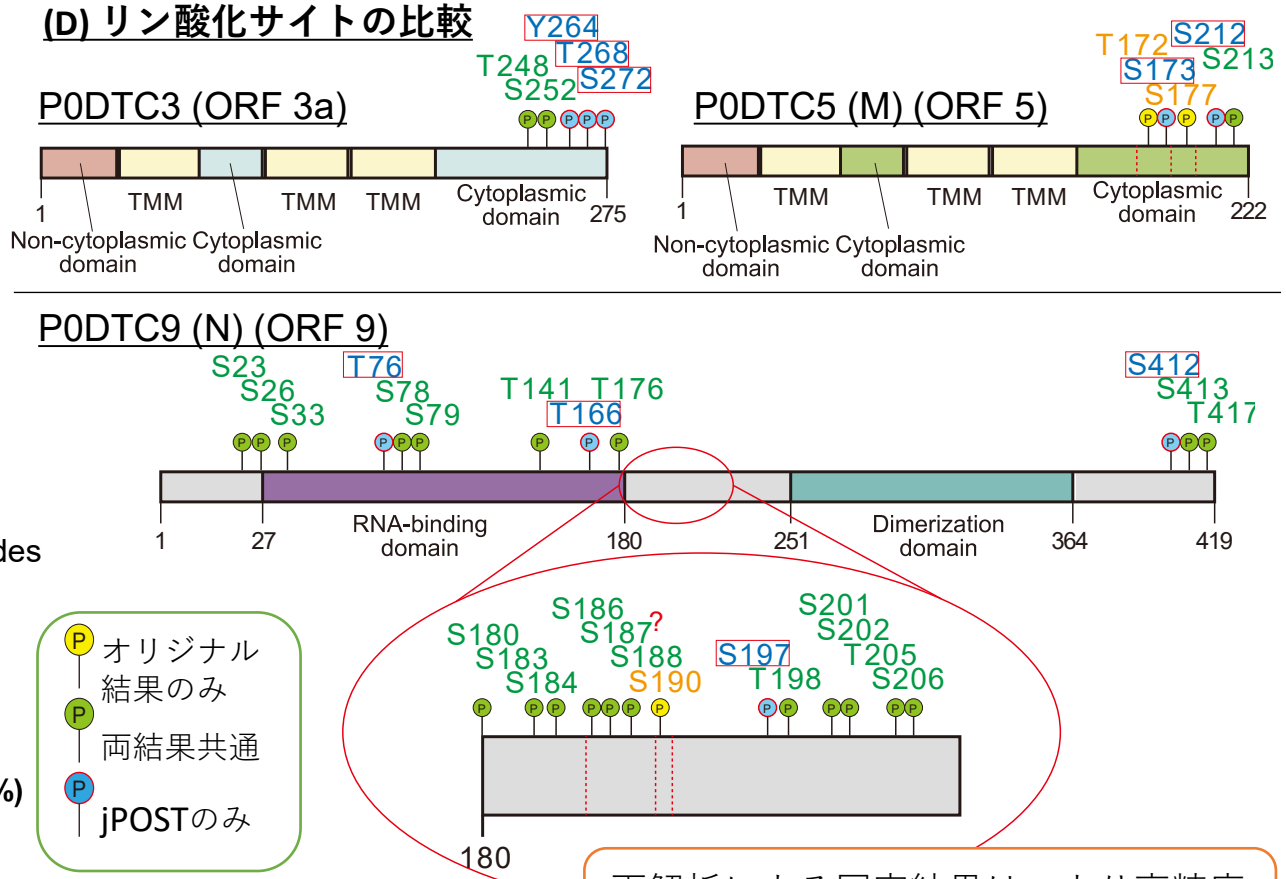
jPOST再解析の結果





PhosPep Analyzerによって確認されたペプチド27,171本から同定

(D) リン酸化サイトの比較



- P オリジナル結果のみ
- P 両結果共通
- P jPOSTのみ

? … オリジナル論文には記載なし、同補足データには記載あり

再解析による同定結果は、より高精度と考えられる。

本研究・開発は科学技術振興機構(JST)・バイオサイエンスデータベースセンター(NBDC)による統合化推進プログラム予算によって実施した。