

統合データベース講習会：AJACS番町3 - jPOST の使い方 -

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
守屋 勇樹 moriya@dbcls.rois.ac.jp
2019/08/07 JST

講習会の流れ

今回の講習会では、ウェブブラウザを使って jPOSTdb の操作を行うことで、データベースの使い方を練習していきます

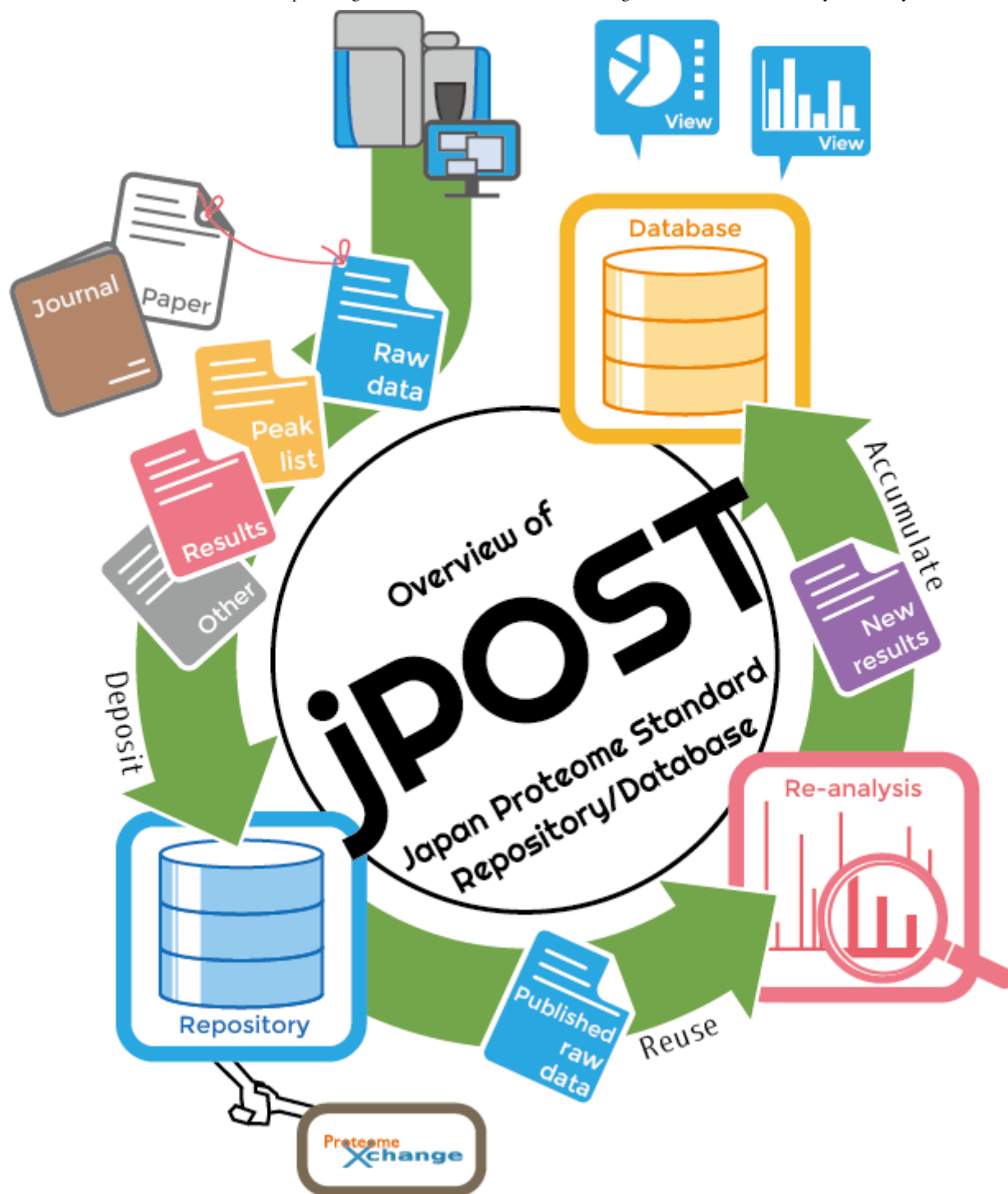
- 概要
 - jPOSTdbの概要
 - データのしぼりこみ
 - データセットページの概要
 - タンパク質ページの概要
 - Slice(データセットグループ)を作成してみる
 - 2つのSliceを比較してみる

講習に際しての注意とお願い

- みんなで同じ回線を使って同時にアクセスするとサイトにつながりにくくなることが予想されます。
 - 資料を見ながら自力で進められそうな方はどんどん先に、そうでない方は一緒にすすめていきましょう。
 - サイトの反応が悪い時はタイミングをずらして実行してみてください。
 - 反応が無いからと言って何度もクリックするとますます繋がらなくなってしまうかもしれません。おらかな気持ちで臨みましょう。
- わからないことがあったら挙手にてスタッフにお知らせください。
 - 遠慮は無用です(そのための講習会です!)。おいてけぼりは楽しくありません。
- 今回のコースでは一応Google Chromeを推奨しますが、Firefox、Safari、Edge、Operaなどのモダンブラウザでも問題ないと思います。動かなかったら今後の課題に加えますので教えて下さい。ただしInternet Explorerは古すぎるので動作を保証していません。モバイル機器での操作。動作も一部未対応です。

jPOSTdbとは

jPOSTdbは質量分析に基づくプロテオームデータを"再解析"をし、その結果を収録したデータベースです。



jPOSTrepに登録したデータを直接見れるわけではありません。

jPOSTdbの特徴

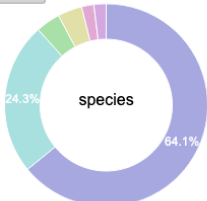
- データセット毎にプロテオームデータを俯瞰できます
 - Missing Protein探索やKEGG Pathwayへのマッピングができます
- タンパク質毎に同定されたペプチドやPTM、isoformの情報を見ることができます
- 興味のあるデータセット複数を切り出して、データを俯瞰できます (Slice)
- 簡易的なツールを用いてSlice間の比較を行うことができます

jPOSTdb webサイト

Filter ▲

Species

+



species

Reset

Dataset (103) Protein (28573)

Keyword

Page size: 10 Showing 1 to 10 of 103 entries

| Dataset ID | Project ID | Project Title | Project Date | #proteins | #spectra |
|------------------------|------------|---|--------------|-----------|----------|
| DS59_1 | JPST000059 | One-dimensional capillary liquid chromatographic sep... | 2016-07-20 | 1554 | 23407 |
| DS59_2 | JPST000059 | One-dimensional capillary liquid chromatographic sep... | 2016-07-20 | 1850 | 32343 |
| DS59_3 | JPST000059 | One-dimensional capillary liquid chromatographic sep... | 2016-07-20 | 1839 | 31708 |
| DS67_1 | JPST000067 | PC12, NF1 disease model, ITRAQ analysis | 2016-09-22 | 4125 | 53101 |
| DS81_1 | JPST000081 | Human IPS cell_201B7-P32 | 2016-07-21 | 4108 | 61097 |
| DS81_2 | JPST000081 | Human IPS cell_201B7-P32 | 2016-07-21 | 3613 | 47745 |
| DS81_3 | JPST000081 | Human IPS cell_201B7-P32 | 2016-07-21 | 3630 | 49370 |
| DS82_1 | JPST000082 | Human IPS cell_32R1-P32 | 2016-07-21 | 4657 | 78052 |
| DS82_2 | JPST000082 | Human IPS cell_32R1-P32 | 2016-07-21 | 4708 | 81255 |
| DS82_3 | JPST000082 | Human IPS cell_32R1-P32 | 2016-07-21 | 5018 | 93670 |

Page: 1

« < 1 2 3 4 5 6 7 > »

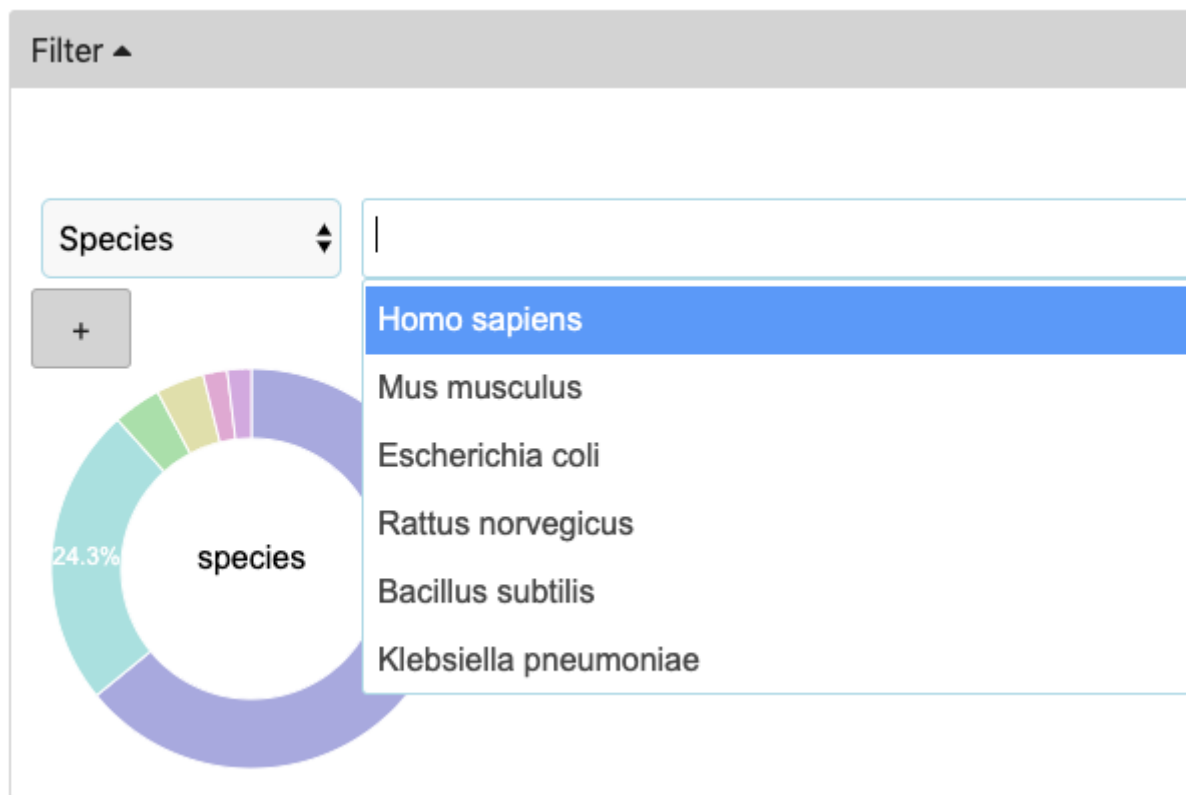
New Slice...

• Topページ

- 最初に、データベースに入っているデータセットの生物種の割合を示したチャートが出ます
 - マウスカーソルをホバーさせることで、詳細が表示されます
- 下にはデータセットのリストテーブルが表示されます
 - 2019/8/7時点で103データセットが収録されています
 - "DS59_1" というのがjPOSTdbでのデータセットのIDになっています
- 上のpie chartのある部分はデータセットの絞り込みを行うフォームになっています

データセットを絞り込んでみよう

- 例えばhumanで絞りこみをしてみましょう
 - テキストエリアをクリックすると候補が表示されます。クリックで選択されます
 - pie chartをクリックすることでも選択できます



- テーブルのリストが66データセット、14,620タンパク質に減りました
- 生物種以外に、Sample type、Cell line、Organ、Disease、Modification、Instrumentで絞り込みが行なえます
- "+" ボタンをクリックすることで、複数のカテゴリで絞り込みを行うことができます

データセットを検索してみよう

- 一旦全部の絞り込みを外しましょう
- テーブルの上にワード検索Boxがあります。キーワードに関連したデータセットが検索されます
 - データがまだ少ないので、今回は"fibroblast"で検索します

64.1%

35.7%

Dataset (9)

Protein (5711)

fibroblast

Page size: 10

| Dataset ID | Project ID | Project Title |
|------------|------------|--------------------------------|
| DS87_1 | JPST000087 | Human Fibroblast cell_aHDF1388 |
| DS87_2 | JPST000087 | Human Fibroblast cell_aHDF1388 |
| DS87_3 | JPST000087 | Human Fibroblast cell_aHDF1388 |
| DS88_1 | JPST000088 | Human Fibroblast cell_aHDF1419 |
| DS88_2 | JPST000088 | Human Fibroblast cell_aHDF1419 |

- 9つのデータセットが引っかけり、5,711タンパク質が収録されていました

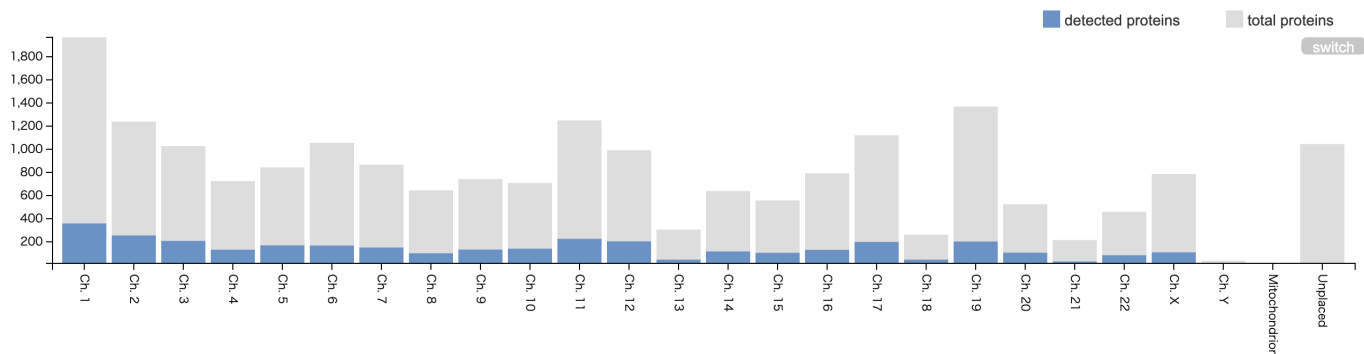
データセットの中身を見ていこう

- テーブルの "DS87_1" をクリックしてデータセットDS87_1の中身を表示しましょう

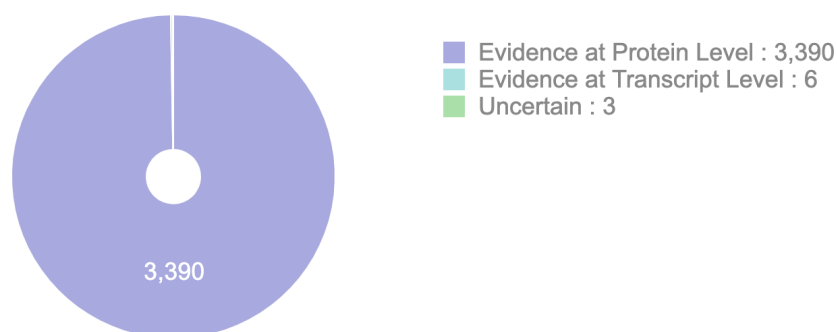
Dataset: DS87_1

| | | |
|---------------------|---|--------------|
| Project ID | JPST000087 | 3 datasets |
| Project title | Human Fibroblast cell_aHDF1388-P9 | |
| Project description | Proteome analyses of human fibroblast cell line (aHDF1388-P9) were carried out using a two-meter monolithic silica C18 capillary columns without prefractionation. Tryptic peptides from 4 μg cell lysates were directly injected onto the 100 μm i.d. monolithic silica-C18 column and an 8-h gradient was applied at 500 nL/min. Triplicate measurements were conducted and the merged results were searched against UniProt DB to identify peptides and proteins. The obtained results were filtered at 1% FDR thresholds by searching against the randomized decoy DB both at peptide and protein levels. This dataset with the PX complete submission format will be further used to generate a customized DB in this jPOST project. | |
| Sample | species: | Homo sapiens |
| | sample type: | Cell Line |
| | cell line: | aHDF1388-P9 |
| Statistics | # proteins: | 3,439 |
| | # peptides: | 23,628 |
| | # spectra: | 57,242 |
| Download | DS87_1.tar | |

- データセットのメタデータがまとめられたテーブルが表示されます
- Project ID、Project title、Project descriptionはリポジトリに登録したときのものが表示されます
- Sampleの情報、統計情報、データをまとめてDLするリンクがあります

Chromosome Info.

- 同定されたタンパク質の染色体毎の数を表示しています
 - UniProt IDベースのカウントになるので、遺伝子数とは厳密には異なります

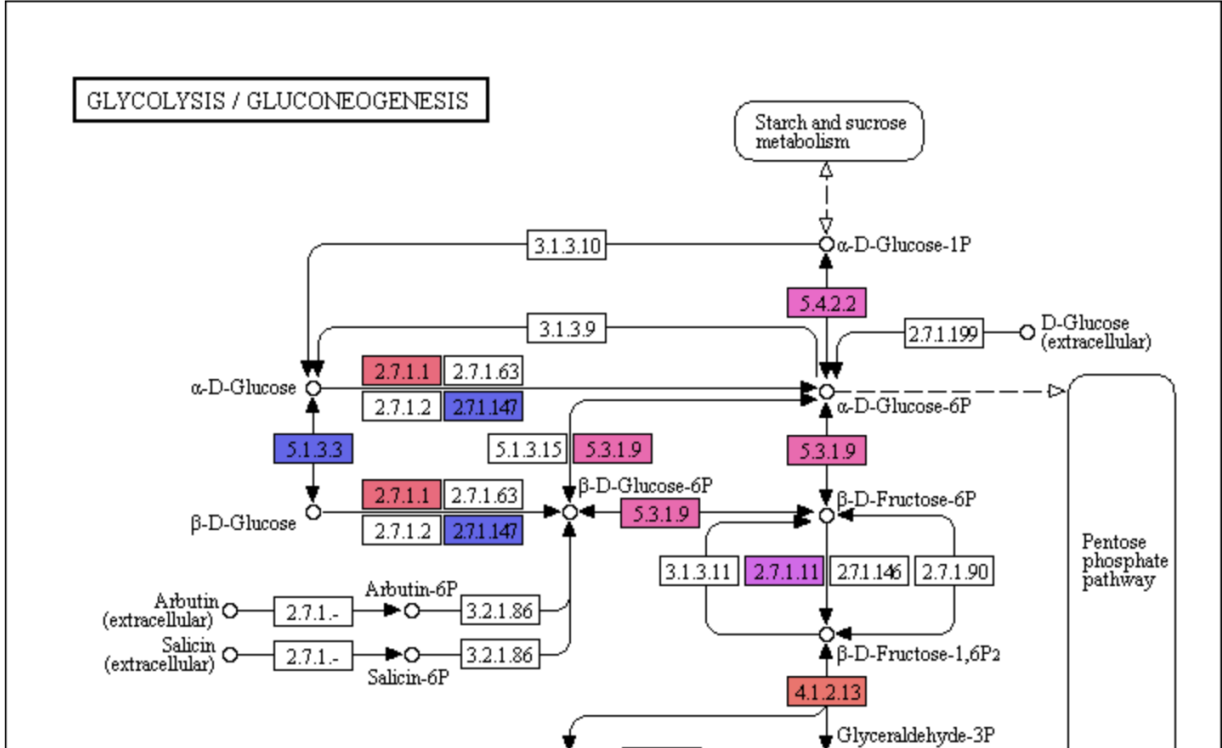
Protein Existence

- Protein existenceのレベルの内訳を表示しています
 - ヒトはneXtProtに基づいており、それ以外の生物種はUniProtのデータに基づいています
 - それぞれのタンパク質のリストを見ることができます

KEGG Pathway Mapping

2,849 proteins with KO (KEGG Orthology)
2,592 KOs

- データセットのタンパク質をKEGG Pathwayにマッピングすることができます



・色はスペクトルカウントに基づいた色になっており、青が少なく、赤が多いことを意味しています

Protein (3430)

Peptide (23586)

Page size: 10

Showing 1 to 10 of 3430 entries

| Protein Name | Accession | ID | Length | Sequence |
|---|-----------|-------------|--------|--|
| 14-3-3 protein beta/alpha | P31946 | 1433B_HUMAN | 246 | MTMDKSELVQKAKLAEQAERYDDMAAMKAVTEQ... |
| 14-3-3 protein epsilon | P62258 | 1433E_HUMAN | 255 | MDDREDLVYQAKLAEQAERYDEMVESMKKVAGMD... |
| 14-3-3 protein eta | Q04917 | 1433F_HUMAN | 246 | MGDREQLLQRLARLAEQAERYDDMASAMKAVTELN... |
| 14-3-3 protein gamma | P61981 | 1433G_HUMAN | 247 | MVDREQLVQKARLAEQAERYDDMAAMKNVTELN... |
| 14-3-3 protein theta | P27348 | 1433T_HUMAN | 245 | MEKTELIQKAKLAEQAERYDDMATCMKAVTEQGA... |
| 14-3-3 protein zeta/delta | P63104 | 1433Z_HUMAN | 245 | MDKNELVQKAKLAEQAERYDDMAACMKSVTEQGA... |
| HLA class I histocompatibility antigen, A-2 alpha chain | P01892 | 1A02_HUMAN | 365 | MAVMAPRTLVLSSGALALTQTWAGSHSMRYFFTSV... |
| HLA class I histocompatibility antigen, B-49 alpha chain | P30487 | 1B49_HUMAN | 362 | MRVTAPRTLVLSSGALALTETWAGSHSMRYFHTAM... |
| HLA class I histocompatibility antigen, Cw-3 alpha chain | P04222 | 1C03_HUMAN | 366 | MRVMAPRTLILSSGALALTETWAGSHSMRYFYTAV... |
| HLA class I histocompatibility antigen, Cw-18 alpha chain | Q29865 | 1C18_HUMAN | 366 | MRVMAPRALILSSGALALTETWAGSHSMRYFDTA... |

Page: 1

<< < 1 2 3 4 5 6 7 > >>

タンパク質をみましょう

- ・ ページ一番上の"Search"からTopの検索画面に戻れます
- ・ テーブルを"Dataset"から"Protein"に切り替えます
- ・ 表示されるデータの都合が良いので、今回は"BCLF1"で検索します

Dataset (103)

Protein (2)

BCLF1

Page size: 10

| Protein Name | Accession | ID |
|---|-----------|-------------|
| Bcl-2-associated transcription factor 1 | Q9NYF8 | BCLF1_HUMAN |
| Bcl-2-associated transcription factor 1 | Q8K019 | BCLF1_MOUSE |

Page: 1

- ヒトとマウスでタンパク質が引っかかりました
- Protein nameをクリックするとタンパク質ページに以降します。今回はヒトのBCLF1を見てみましょう

Protein: Q9NYF8

Protein Name

Bcl-2-associated transcription factor 1

Protein ID

BCLF1_HUMAN

Gene Name

BCLAF1

Accession

Q9NYF8

Length

920 aa

Sequence

MGRNSRSRSHSRKSRSSQSSSRSRSHSRKKRYSSRSRSTYSRSRSDRMYSRDYRRDYRNRRGMRRPYGYRGRGRGYQGGGGRYHRGGYRPVWNRHRSRPRRGRSRSPKRRSVSSQRS
QEQTKKAEGEPQEESPLKSKSQEEPDTFEHDPSEIDFNKSSATSGDIWPLGLSAYDNSPRSPHSPPIATPPSQSSSCDAPMLSTVHSAKNTPSQHSHSIQHSPERSGSGSVGNGSSRYSP
PDGGDQETAKTGKFLKRFTDEESRVFLDRGNTRDKEASKEKGSEKGRAEGEWEDQEALDYFSDKESGKQKFNDSEGGDTEETEDYRQFRKSVLADQGKSFATASHRNTEEEGLKYKSKVSLKGN
NSERITVKKETQSPQVKSEKLDLFDYSPPLHKNLDAKSTFREESPLRIKMIASDSHRPEVKLKMAPVPLDDSNRPASLTKDRLLASTLVHSHVKKKEQEFRSIFDHIKLPQASKSTSESFIQH
RQKSPETIHRRIDISPSTLRKHTRLAGEERVFKENQKGDKKLRCDSDLRHDIDRRRKERSKERGDSKGSRESSGSRKQEKTPKDYKEYKSYKDDSKHKREQDHSRSSSSASPSPPSSREEKES
GRARGVFAGTNTGPNNSTTFQKRPKEEWDPEYTPKSKKYFLHDDRDDGVDYWAKRGRGRGTQGRGRGRFNFKKSGSSPKWTHDKYQGDGIVEDEEETMENNEEKKDRRKEEKE

Location

Chromosome 6

Statistics

peptides:

92

spectra:

4,069

unique peptides (UniProt entry level):

54

unique peptides (gene name level):

89

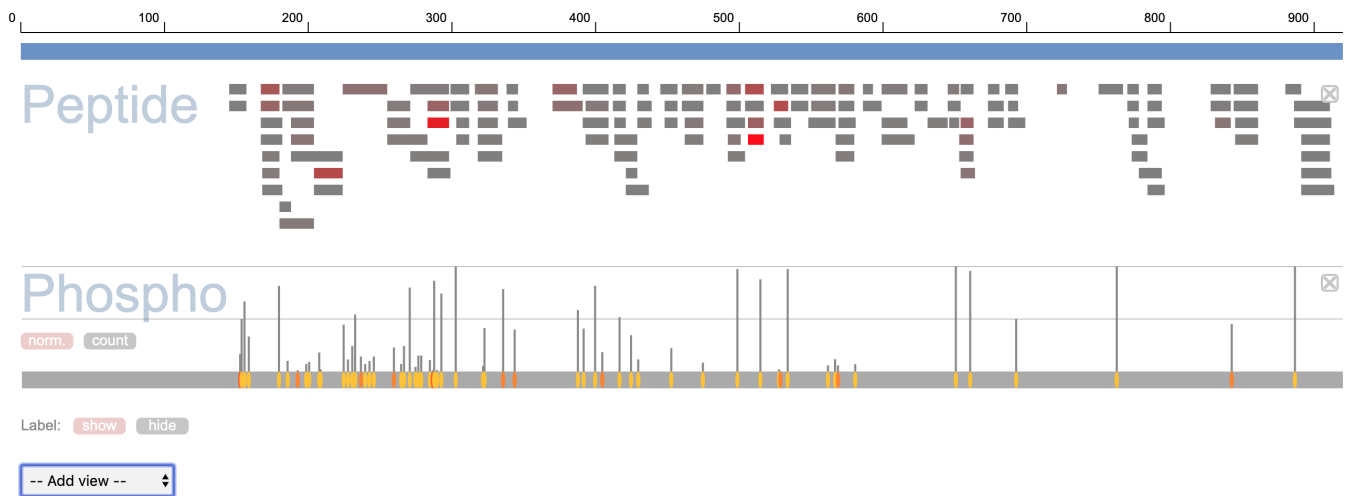
for Absolute Quantification

GeneID: 9774

search in iMPAQT

- タンパク質の情報テーブルを表示しています
- ヒトのタンパク質の場合iMPAQT へのリンクがあります

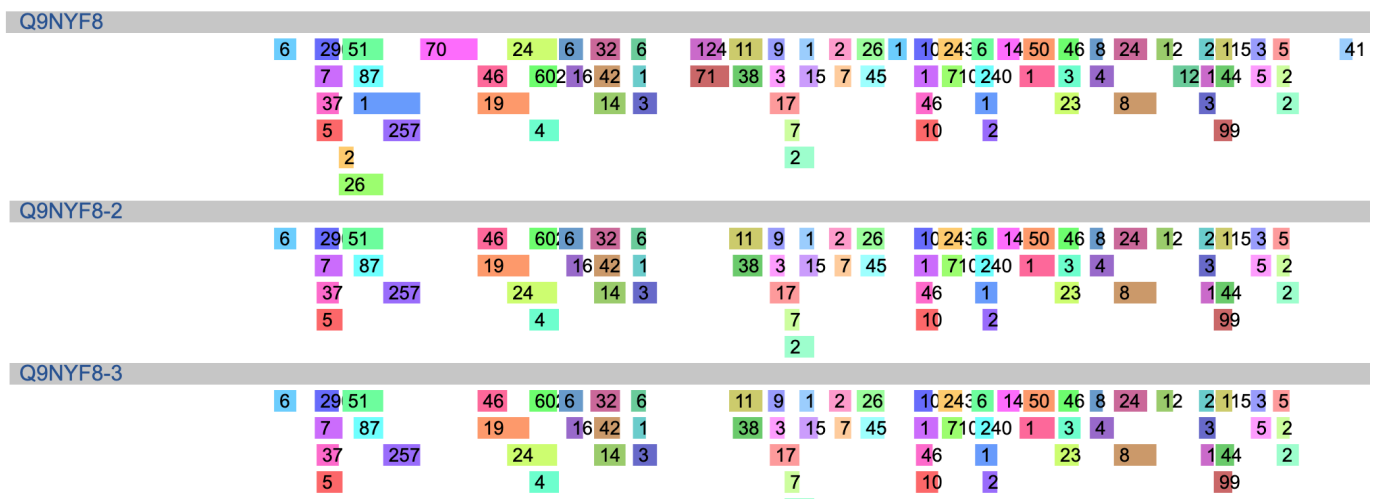
Protein Browser



- タンパク質ブラウザを表示しています
 - スクロール操作で拡大縮小できます
- プルダウンメニューから幼児する情報を追加できます
 - Peptide: 検出されたペプチドのマッピング (Default)
 - 赤は同定数の多いペプチドを示しています
 - Phospho site: リン酸化部位
 - 軸の長さは検出された量を、そのポジションをカバーしているペプチドの数で補正した割合を示しています
 - "count" ボタンをクリックすると実数に変わります
 - P-site linkage: リン酸化部位のペプチド上での共起情報
 - UniProt annotation: UniProtに収録されている既知の修飾情報

Peptide Sharing

shared in isoforms ▾



- isoform間、タンパク質間のペプチドの共有情報を表示しています
 - 同じ色のBoxが同一のペプチドになります

Sliceを作ってみよう

- 2019/8/2https://raw.githubusercontent.com/AJACS-training/AJACS77/master/05_kobayashi_moriya/README.md
- Sliceとは、データベースからユーザの興味のあるデータセットを切り出した、データベースのサブセットです
 - iPS細胞のデータセットでSliceを作ってみよう
 - Topの検索ページで"iPS cell"でデータセットを検索します



Dataset (15)

Protein (6904)

iPS cell

Page size: 10

| Dataset ID | Project ID | Project Title |
|------------|------------|--------------------------|
| DS81_1 | JPST000081 | Human iPS cell_201B7-P32 |
| DS81_2 | JPST000081 | Human iPS cell_201B7-P32 |
| DS81_3 | JPST000081 | Human iPS cell_201B7-P32 |
| DS82_1 | JPST000082 | Human iPS cell_32R1-P32 |

- テーブルの下にある"New Slice..."ボタンをクリックします

New Slice

Slice1

Description

Filter ▲

Species

× Homo sapiens

+

iPS cell

Page size: 10

Showing 1 to 10 of 15 entries (filtered from 103 entries)

| | Dataset ID | Project ID | Project Title |
|-------------------------------------|------------|------------|--------------------------|
| <input checked="" type="checkbox"/> | DS81_1 | JPST000081 | Human iPS cell_201B7-P32 |
| <input checked="" type="checkbox"/> | DS81_2 | JPST000081 | Human iPS cell_201B7-P32 |
| <input checked="" type="checkbox"/> | DS81_3 | JPST000081 | Human iPS cell_201B7-P32 |
| <input type="checkbox"/> | DS82_1 | JPST000082 | Human iPS cell_32R1-P32 |
| <input type="checkbox"/> | DS82_2 | JPST000082 | Human iPS cell_32R1-P32 |
| <input type="checkbox"/> | DS82_3 | JPST000082 | Human iPS cell_32R1-P32 |
| <input type="checkbox"/> | DS83_1 | JPST000083 | Human iPS cell_414C2-P43 |
| <input type="checkbox"/> | DS83_2 | JPST000083 | Human iPS cell_414C2-P43 |
| <input type="checkbox"/> | DS83_3 | JPST000083 | Human iPS cell_414C2-P43 |
| <input type="checkbox"/> | DS85_1 | JPST000085 | Human iPS cell_585A1-P55 |

Page: 1

«

<

1

2

>

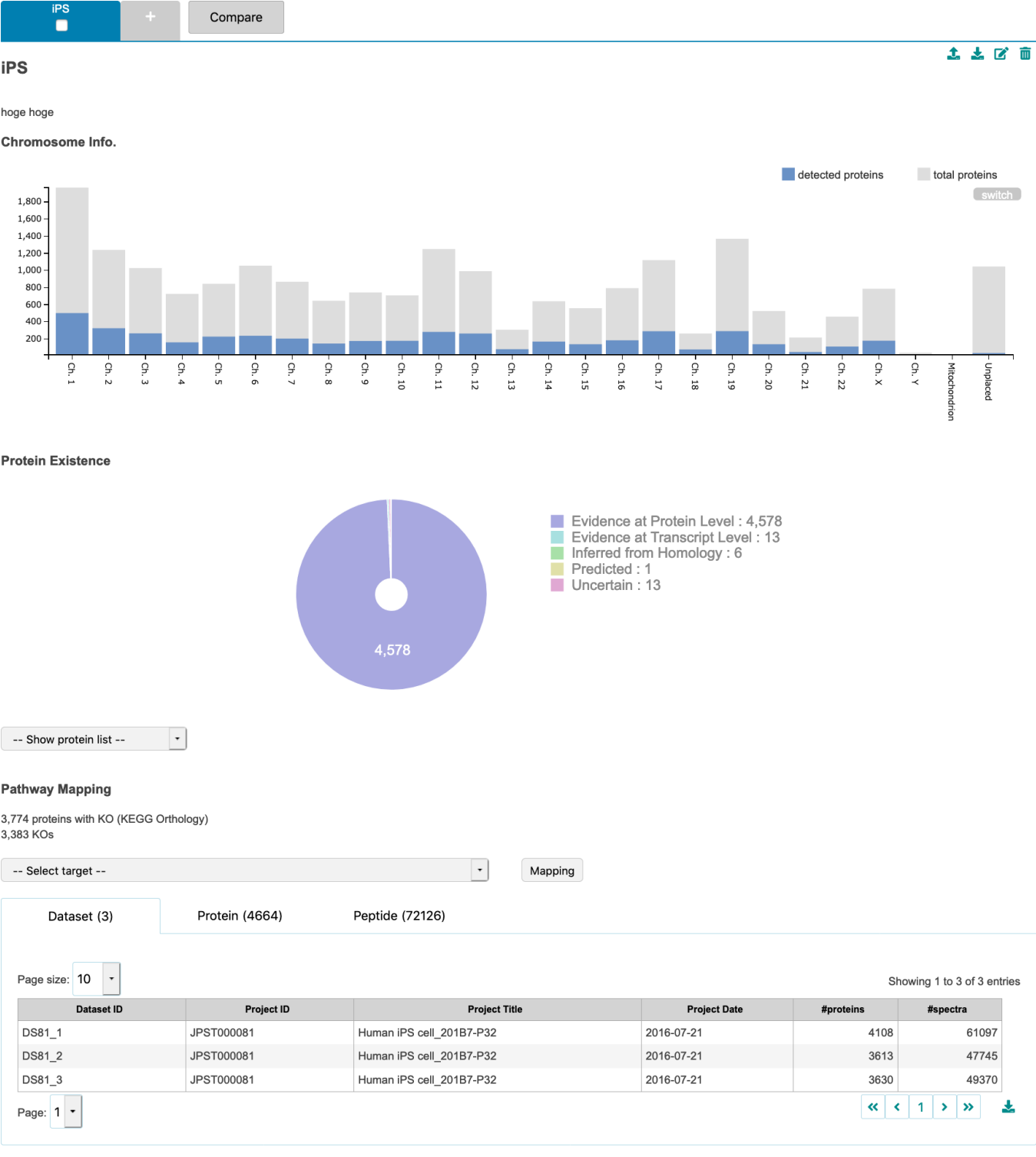
»

↓

Add

Cancel


- Slice作成画面に移動します
 - この画面でも絞り込み検索が可能です
 - わかりやすい名前をつけ、必要なら説明を入れましょう
- 今回はテーブルの上の3つDS81_1, DS81_2, DS81_3でSliceを作ります
 - 3つのチェックボックスにチェックを入れます
- "Add"ボタンをクリックすると、Sliceが作成されます



- 選択したデータセットをまとめた情報が表示されます
 - 見方はデータセット毎のページと同様です

Slice同士を比較してみよう

- jPOSTdbではSlice同士を比較する簡易な解析機能がついています
- もう一つ"fibroblast"でSliceを作成します
 - Sliceページのタブに並んでる"+"マークからも新規Sliceが作成できます

**JPOST DATABASE**
Repository Database

IPS

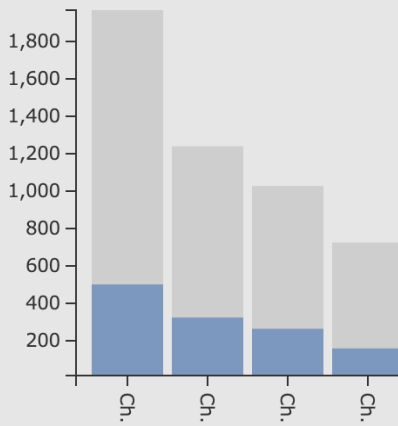
+

Compare

IPS

hoge hoge

Chromosome Info.



| Chromosome | Protein Existence |
|------------|-------------------|
| Ch. 1 | ~1,800 |
| Ch. 2 | ~1,200 |
| Ch. 3 | ~1,000 |
| Ch. 4 | ~700 |

Protein Existence

New Slice

fibroblast

hoge fuga piyo

Filter ▲

+

fibroblast

Page size: 10

Showing

| <input type="checkbox"/> | Dataset ID | Project ID | |
|-------------------------------------|------------|------------|---|
| <input checked="" type="checkbox"/> | DS87_1 | JPST000087 | ト |
| <input checked="" type="checkbox"/> | DS87_2 | JPST000087 | ト |
| <input checked="" type="checkbox"/> | DS87_3 | JPST000087 | ト |
| <input type="checkbox"/> | DS88_1 | JPST000088 | ト |
| <input type="checkbox"/> | DS88_2 | JPST000088 | ト |

- 2つのSliceを比較します
 - タブにあるチェックボックスをチェックして"Compare"ボタンをクリックします



IPS

fibroblast

+

Compare

fibroblast

hoge fuga piyo

- 比較ページに移動します

Slice1: IPS

Slice2: fibroblast

Statistics

| | slice: IPS | | share | slice: fibroblast | |
|------------|------------|--------|--------|-------------------|--------|
| | total | unique | | total | unique |
| # datasets | 3 | | | 3 | |
| # proteins | 4,684 | 1,450 | 3,234 | 4,618 | 1,384 |
| # peptides | 33,837 | 13,661 | 20,176 | 37,078 | 16,902 |

Differential Expression Analysis

- The quantification is based on spectral counting.
- Some methods need at least 2 datasets in either slice.

-- Select method --

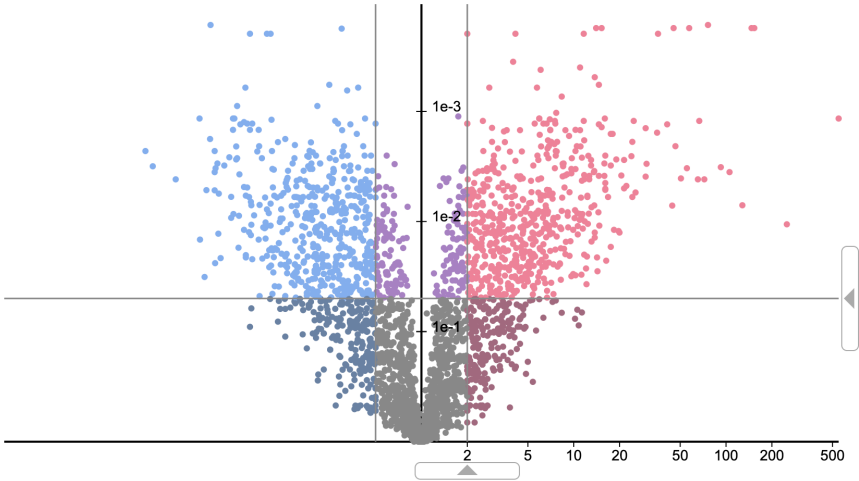
- 統計情報が表示されます
- また、簡易的な変動解析をすることができます
 - 現時点では定量データをあまり収録していないので、スペクトルカウントを補正して半定量をしています
 - スペクトルカウントはタンパク質に紐付いたスペクトルの数のことです
 - 今回は"Empirical Bayes estimation"を使います
 - Wilcoxon rank sum testはサンプル数が多いときによく使われます
 - Empirical Bayes estimationは比較的少ないサンプル数でも比較が可能です
 - Fold change of the meanは統計処理はしないので1サンプルでも比べられます

Differential Expression Analysis

- The quantification is based on spectral counting.
- Some methods need at least 2 datasets in either slice.

-- Select method --

Fold change >= 2
p-value <= 5e-2
Proteins: 1103
both



Page Size: 15

search protein

Showing 1 to 15 of 1103 entries

| | Protein name | Accession | ID | Fold change | log(fc) | p-value |
|--|--|-----------|-------------|-------------|---------|----------|
| | Collagen alpha-3(VI) chain | P12111 | CO6A3_HUMAN | 547 | 9.1 | 1.16e-03 |
| | Peripherin | P41219 | PERL_HUMAN | 251 | 7.97 | 1.06e-02 |
| | Neuroblast differentiation-associated protein AHNK | Q09666 | AHNK_HUMAN | 153 | 7.26 | 1.75e-04 |
| | Vimentin | P08670 | VIME_HUMAN | 147 | 7.2 | 1.75e-04 |
| | Calpain-2 catalytic subunit | P17655 | CAN2_HUMAN | 128 | 7 | 7.12e-03 |

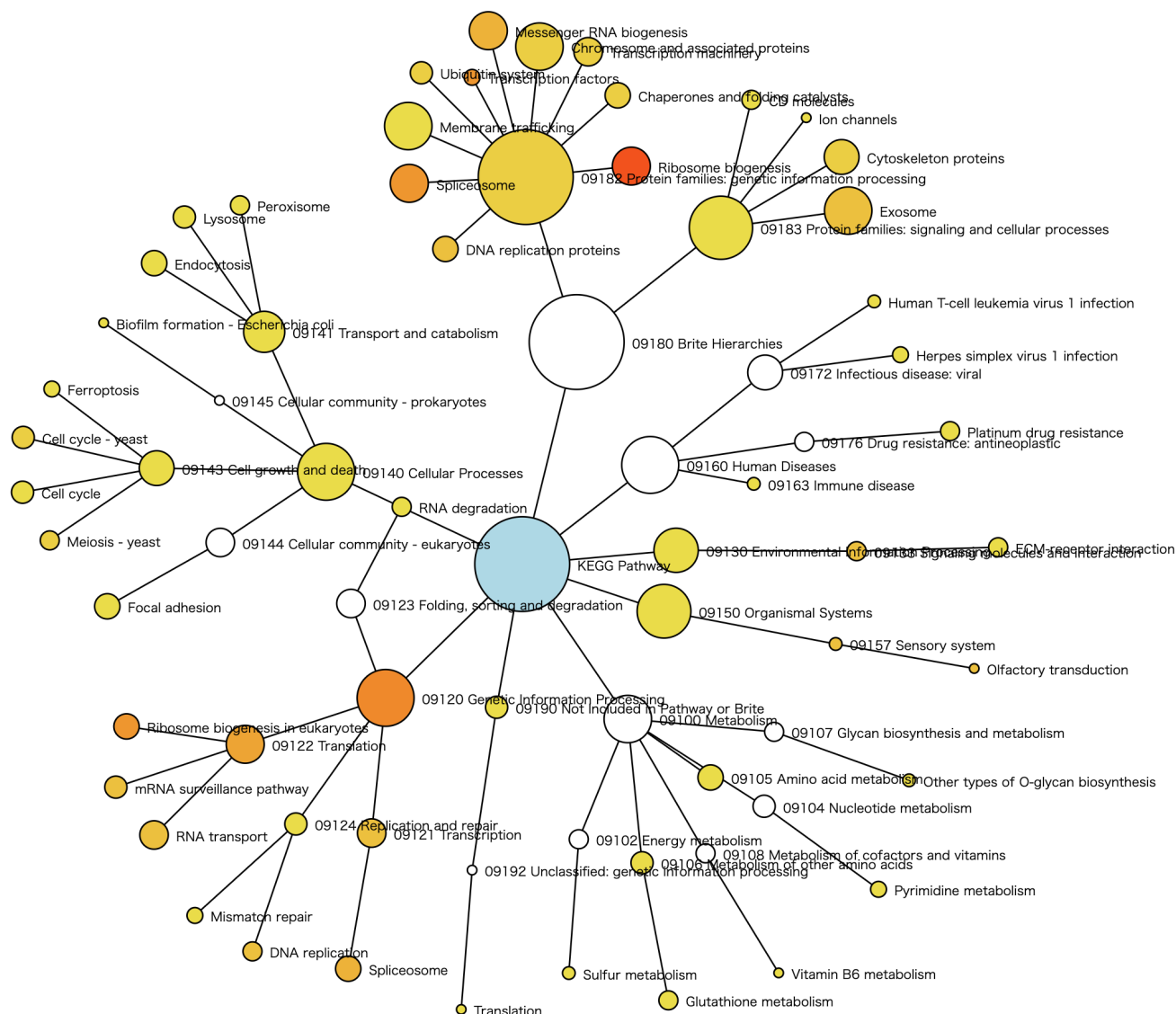
- Volcanoプロットが表示されます
 - 各プロットがタンパク質になります
 - 横軸がFold changeを示し、縦軸はp-valueになっています
 - 左（青）が発現量が低下したタンパク質、右（赤）が上昇したタンパク質を示しています
 - 青と赤で示されたプロットが有意な変動が見られたタンパク質ということになり、下のテーブルにリストが表示されています
 - Thresholdはユーザがバーを移動させることで変えられます
 - プロット左にある"both"ボタンをクリックすることで上昇したものだけ、下降したものだけを表示することもできます
 - Fold change of the meanの場合はヒストグラム風のプロットになります

Enrichment Analysis

- Protein set enrichment analysis for selected proteins in the upper plot and table.

-- Select target --

- 選択した有意な変動の見られたタンパク質リストを用いて、エンリッチメント解析を行うことができます
 - KEGG Pathway, GO, ChIP-Atlasなどを用いて解析ができます
 - 今回は"KEGG Pathway"のカテゴリでエンリッチメント解析をしてみましょう



- ネットワークグラフが表示されます
 - 各ノードがカテゴリを示しています
 - ノードのサイズがカテゴリのタンパク質数を示しています
 - 白いノードはサイズの上限を設定しているので、色付きノードより大きく表示されません
 - 黄色からオレンジ色のノードはエンリッチしているカテゴリを意味しています
 - 色の濃いノードがとりエンリッチしています
- jPOSTdbはChIP-Atlasと連携しており、ヒトをはじめ、いくつかの生物種ではjPOSTdbで変動解析した結果をChIP-Atlasの解析フォームを用いてエンリッチメント解析ができます

ChIP-Atlas - Enrichment Analysis

Tutorial movie ▾

Analyze your data with public ChIP-seq data.

H. sapiens M. musculus R. norvegicus D. melanogaster C. elegans S. cerevisiae

1. Antigen Class

All antigens (43086)
DNase-seq (1632)
Histone (10578)
RNA polymerase (1532)
TFs and others (9868)
Input control (5080)
Unclassified (10053)
No description (4343)

2. Cell type Class

All cell types (43086)
Adipocyte (324)
Blood (10674)
Bone (819)
Breast (5050)
Cardiovascular (1136)
Digestive tract (2970)
Epidermis (1244)
Skeletal muscle (684)

3. Threshold for Significance ⓘ

50
100
200
500

4. Select your data

☐ Genomic regions (BED) or sequence motif ⓘ
☒ Gene list (Gene symbols) ⓘ

COL6A3
PRPH
AHNAK
VIM
CAPN2
COL6A2
EHD2
ANXA1

ファイルを選択 選択されていません
Choose local file [Try with example](#)

5. Select dataset to be compared

☐ Refseq coding genes (excluding user data) ⓘ
☒ Gene list (Gene symbols) ⓘ

KRT18
TOP2A
MSH6
IGF2BP1
TOP2B
AFDN
PSIP1
HMGB3

ファイルを選択 選択されていません
Choose local file [Try with example](#)

6. Describe datasets

User data title ⓘ
My data

Compared data title ⓘ
Control

Project title ⓘ
My project

Distance range from TSS ⓘ
- 50 bp ≤ TSS ≤ + 50 bp

submit

Estimated run time: -