ProMetaDB: A unified resource of enriched proteomics metadata for easy data acquisition of proteomic datasets from different public repositories

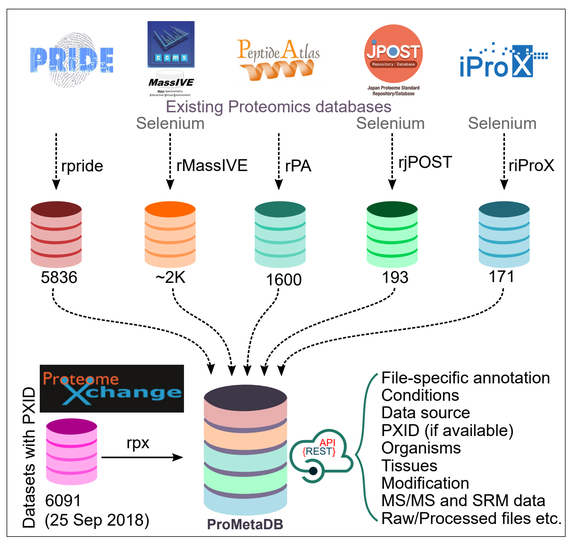
Rintu Kutum[1,2,3], Prateek Singh[1,3], Anurag Raj[1,3], Debasis Dash[1,2,3]

[1]G.N. Ramachandran Knowledge Centre for Genome Informatics, CSIR-IGIB, [2]CSIR-TRISUTRA unit, CSIR-IGIB, [3]Academy of Scientific and Innovative Research

Abstract

We are currently in the golden era of proteomics, where mass spectrometry (MS) based proteomics datasets are publically available in many repositories [1]. Meta-analysis of existing proteomics through proteogenomics approach shows great potential towards identifying novel proteins and peptides [2, 3]. In 2011, ProteomeXchange(PX) consortium has adopted a unified framework to enable rapid dissemination of proteomics datasets existing in different repositories (PRIDE, MassIVE, PeptideAtlas, jPOST and iProX) [4, 5]. PX follows the **M**inimum **I**nformation **A**bout a **P**roteomics **E**xperiment (**MIAPE**) guidelines [6] for the metadata; it provides only minimal information such as the title, description, species, modification, instrument, keywords and links to the repository etc. Here, we present **Pro**teomics **Meta**data **D**ata**b**ase (**ProMetaDB**), which contains additionally curated meta information (phenotypes, conditions/concentrations and their associated raw/processed files) for all existing proteomics studies. We believe ProMetaDB will enable researchers to perform easy and rapid acquisition of existing datasets for discovering novel proteins/peptides in this golden era of proteomics.

Data & Methodology



A Collect all existing metadata from a database through existing APIs or web scraping (selenium)

B Develop R/python packages for automatic retrieval of metadata

C Manually segregate and annotate projects from each database towards enrichment of existing metadata

D Develop database of enriched metadata for each proteomic database

E Develop ProMetaDB by merging above databases and provide web services (APIs) for easy and rapid acquisition of datasets

Results

Overview of proteomic datasets across databases

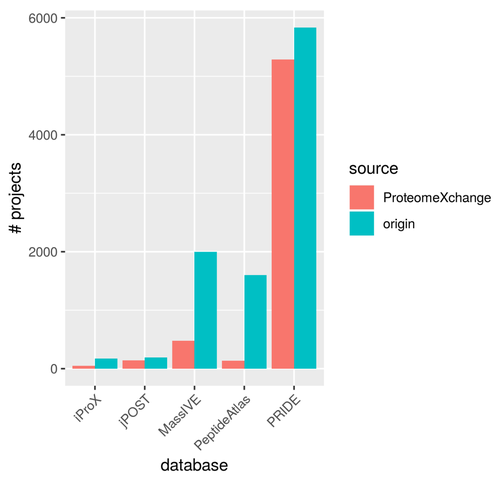


Figure 1a: Barplot showing number of proteomic projects across databases and number of projects metadata available at ProteomeXchange

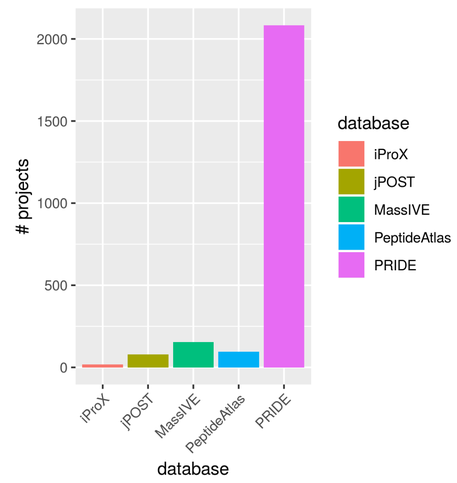


Figure 1b: Barplot showing number of proteomic projects across databases related to Human.

Tissue specific metadata enrichment of PRIDE database

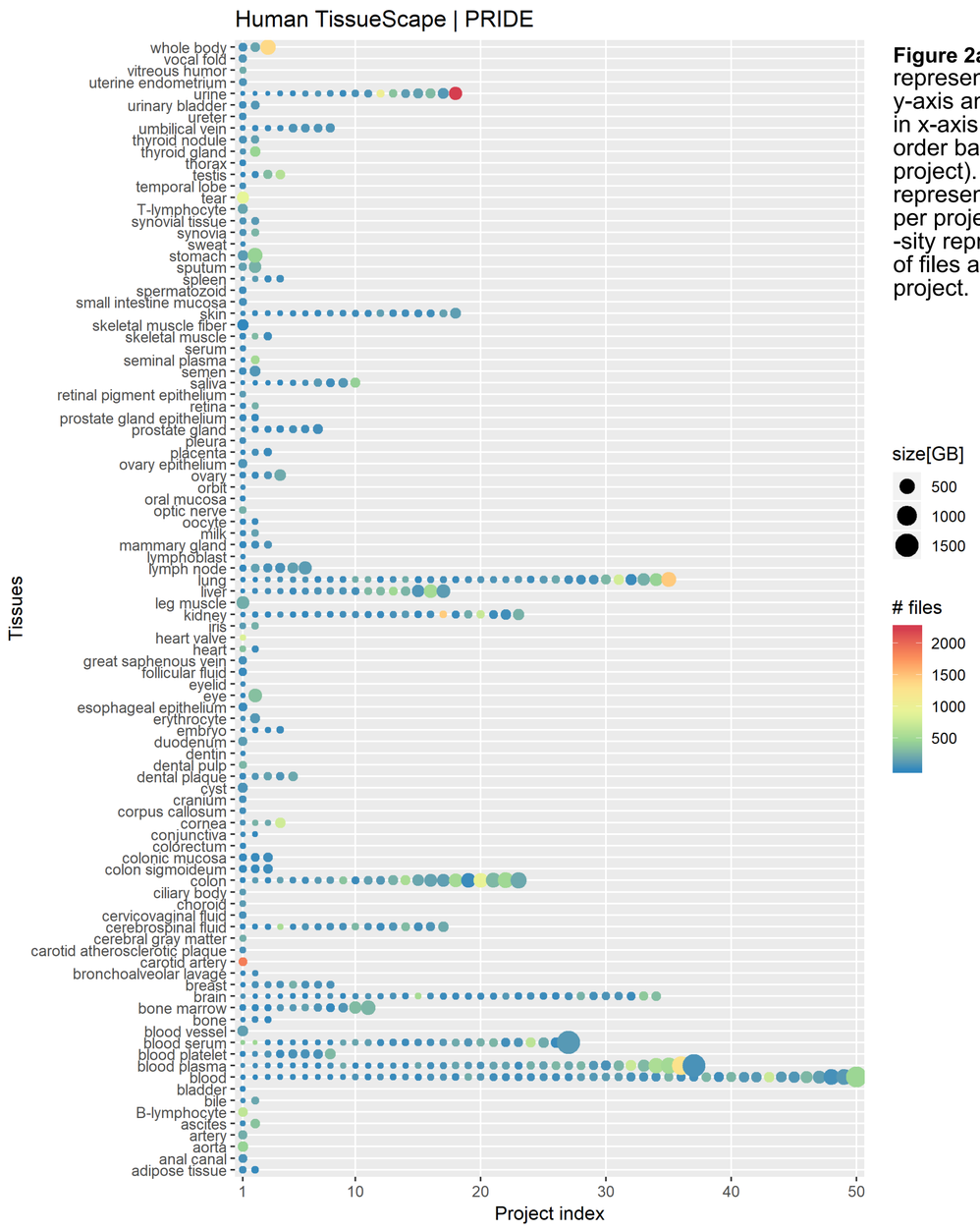


Figure 2a: Human TissueScape representing tissue tags in y-axis and projects indeces in x-axis (sorted in ascending order based total size of the project). The bubble size represents the total size(GB) per project. The color inten-sity represents total number of files associated within a project.

Results cont'd

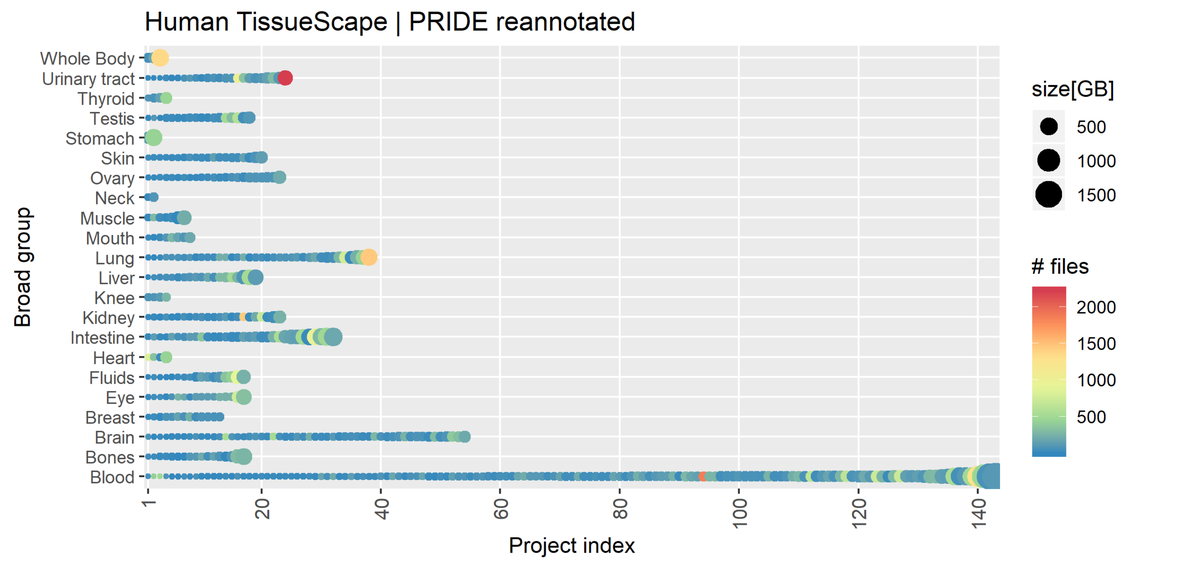


Figure 2b: Human TissueScape after reannotating tissue tags into broad group shown in y-axis and x-axis represents the iproject index arranged in ascending order based on total size of the project. The bubble size depicts the total size(GB) of the project and color gradient represents the total number of files per project,

Enrichment of metadata in MassIVE database | Human

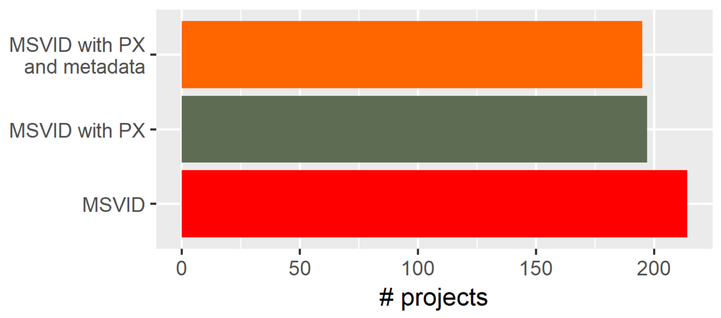


Figure 3a: Barplot of completed MassIVE datasets (red) for human. MassIVE ID (MSVID) without PX ID is shown in middle and 195 MSVIDs (orange) with complete metadata and PXID were used for further reannotation of metadata.

Detail metadata enrichment for 195 MassIVE datasets

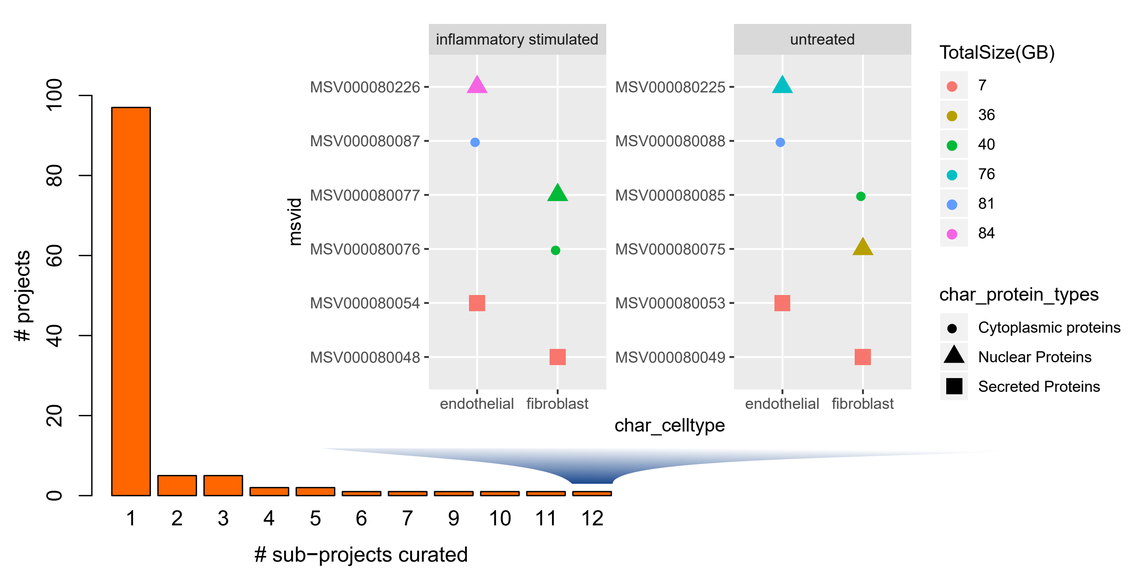


Figure 3b: Barplot derived from 195 MassIVE projects showing number of projects (y-axis) along with number of curated sub-projects (x-axis). The curation is performed based on similarity of project titles and common submitter name (main figure). The subfigure represents the manually curated metadata (char_celltype, char_protein_types and char_conditions) along with 12 MassIVE projects (sub-projects) that can be merged into single project.

Conclusions

- Acquired metadata from MassIVE, jPOST and iProX through web scraping (selenium)
- Developed an R package called **rpride** for automatic retrieval of metadata from PRIDE database.
- Manually reannotated metadata for human tissue-specific proteomics datasets from PRIDE database
- Manually annotated and enriched the existing metadata such as cell type, conditions (stimulated/unstimulated) etc., for 195 MassIVE projects that corresponding to 119 unique projects

Future directions

- To develop rMassIVE, rPA, rjPOST and riProX *R packages* for automatic retrieval of available metadata.
- To manually annotate to enrich the existing metadata for remaining projects in each proteomic database
- To develop enriched metadata database for each proteomic database
- To develop ProMetaDB and provide web serverics (API) for easy and rapid acquisition of proteomics datasets

References

1. Martens L, Vizcaino JA: **A Golden Age for Working with Public Proteomics Data**. *Trends in biochemical sciences* 2017, 42(5):333-341.
2. McAfee A, Foster LJ: **Proteogenomics: Recycling Public Data to Improve Genome Annotations**. *Methods in enzymology* 2017, 585:217-243.
3. Barbieri R, Guryev V, Brandsma CA, Suits F, Bischoff R, Horvatovich P: **Proteogenomics: Key Driver for Clinical Discovery and Personalized Medicine.***Advances in experimental medicine and biology* 2016, 926:21-47.
4. Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, Dianes JA, Sun Z, Farrah T, Bandeira N et al: **ProteomeXchange provides globally coordinated proteomics data submission and dissemination**. *Nature biotechnology* 2014, 32(3):223-226.
5. Deutsch EW, Csordas A, Sun Z, Jarnuczak A, Perez-Riverol Y, Ternent T, Campbell DS, Bernal-Llinares M, Okuda S, Kawano S et al: **The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition.***Nucleic acids research* 2017, 45(D1):D1100-D1106.
6. Martinez-Bartolome S, Binz PA, Albar JP: **The Minimal Information about a Proteomics Experiment (MIAPE) from the Proteomics Standards Initiative**. *Methods Mol Biol* 2014, 1072:765-780.

Acknowledgement

RK and DD would like to acknowledge CSIR (MLP0901) and DBT (GAP0134) for funding. PS and AR would like to acknowledge DBT-BINC and DST-INSPIRE for the fellowship respectively. All authors would like to thank HPC computing facility at CSIR-IGIB