Bioinformatics Analysis of Proteins and Proteomes

Md Tawhidul Islam

Master of Philosophy (Bioinformatics),
Macquarie University, Australia

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

November 2017

Department of Chemistry and Biomolecular Sciences
Macquarie University, Sydney, Australia

Dedicated to,

my late father who encouraged me for higher education.

DECLARATION

I certify that this thesis entitled “Bioinformatics Analysis of Proteins and Proteomes” is a bonafide record of research work carried out by me under the guidance of Professor Shoba Ranganathan during the years 2012-2017 for the degree of Doctor of Philosophy. The results presented in this thesis have not previously formed the basis for award of any degree, fellowship or other recognition. The particulars given in the thesis are true to the best of my knowledge.

Md Tawhidul Islam

June 2017

Acknowledgements

I am indebted to many individuals throughout my Ph.D. tenure for their assistance and support. The format of this thesis will not permit me to list every one of them. However, I like to acknowledge and thank the following key persons.

Professional

- My principal supervisor Professor Shoba Ranganathan her supervision, guidance, motivation, feedback and intellectual support.
- My co-supervisor A/Professor Joanne Jamie for her support throughout my Ph.D. tenure.
- Professor Mark Baker for the proteomics collaboration.
- My mentor Dr. Abidali Mohamedali, and my colleagues Dr. Gagan Garg, and Dr Mostafa Shaikh for their valuable guidance in various research matters.
- Macquarie University, for the award of MQRES research scholarship for pursuing my Ph.D. and PGRF funding for attending HUPO Congress 2014 conference in Madrid, Spain.
- A/Professor Tracy Rushmer and A/P Mark Molloy for their encouragement and support in extending my Ph.D. tenure to compensate the time lost due to IT infrastructure changes.
- Ms. Catherine Wong, Ms. Michelle Kang, Ms. Jane Yang, Mr. Rienzie Jayasekara, Mr. Richard Miller and Mr. Suresh Mulavineth for administrative and IT support.
- Mr. Gehendra Acharya, Mr. Glenn Satchell and Mr. Jayatheeban Soundararajan for their technical support with research cloud compute and storage support.

Personal

- I am incredibly grateful to my wife, Saima, for her patience, ongoing support, and encouragement.
- My heartfelt thanks to my mother, and all my family members for their encouragement.

Table of Contents

List of Abbreviations

API	Application programming interface
BLAST	Basic Local Alignment Search Tool
CPTAC	Clinical Proteomic Tumor Analysis Consortium
CV	Controlled vocabularies
DDA	Data-dependent acquisition
DOI	Digital Object Identifier
EBI	E European Bioinformatics Institute
FAB	Fast atom bombardment
FASTA	FAST-All
FDR	False discovery rate
FTP	File Transfer Protocol
GC	Gas chromatography
GO	Gene Ontology
GPMDB	The Global Proteome Machine Database
HAMAP	High-quality Automated and Manual Annotation of Proteins
HMM	Hidden Markov Models
HPM	Human proteome map
HPP	Human Proteome Projects
HUPO	The Human Proteome Organization
IPI	International Protein Index
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Orthology
LC	Liquid chromatography
LIT	Linear ion trap
MALDI	Matrix-assisted laser desorption/ionization
MAXQB	The MaxQuant DataBase
MeSH	Medical Subject Headings
MGF	Mascot generic format file
MOPED	Multi-Omics Profiling Expression Database
MOWSE	MOlecular Weight Search
MQRES	Macquarie University Research Excellence Scholarship
MS	Mass spectrometry
NMR	Nuclear magnetic resonance

NSF	National Science Foundation
OmicsDI	Omics Discovery Index
PASSEL	The PeptideAtlas SRM Experiment Library
PDB	Protein Data Bank
PE	Protein existence
PEFF	PSI Extended Fasta Format
PEP	Posterior error probability
PIR	Protein Information Resource
PIRSF	The Protein Information Resource SuperFamily
PMF	Peptide mass fingerprint
PRIDE	PRoteomics IDEntification
PRINTS	The PRINTS database
PROSITE	Database of protein domains, families and functional sites
PSI	Proteomics Standards Initiative
PSM	Peptide spectrum match
PTM	Post-translational modifications
PX	ProteomeXchange
PXD	ProteomeXchange Identifier
QIT	Quadrupole ion trap
RCSB	RCSB Protein Data Bank
RESID	The RESID Database
SIB	Swiss Institute of Bioinformatics
SRM	Selected reaction monitoring
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
TOF	Time of flight
TPP	Trans-Proteomics Pipeline

List of Tables

List of Figures

List of Publications included in this thesis

The following papers are presented in this thesis and are referred to from this point onwards as listed in respective sections of the thesis, with my contributions to each paper:

Publications

[1] **Islam MT,** Mohamedali A, Garg G, Khan JM, Gorse AD, Parsons J, Marshall P, Ranganathan S*, Baker MS* (2013) Unlocking the Puzzling Biology of the Black Périgord Truffle *Tuber melanosporum. J Proteome Res.* 12, 5349-56.
(i) concept: 40% (ii) data gathering: 55% (iii) data analysis: 40% (iv) development: 60% (iii) writing: 40%

[2] **Islam MT,** Garg G, Hancock WS, Risk BA, Baker MS, Ranganathan S (2014) Protannotator: A Semiautomated Pipeline for Chromosome-Wise Functional Annotation of the "Missing" Human Proteome. *J Proteome Res*. 13, 76-83.
(i) concept: 55% (ii) data gathering: 80% (iii) data analysis: 60% (iv) development: 75% (iii) writing: 45%

[3] **Islam MT,** Mohamedali A, Ranganathan S (2017) ProtAnnotator 2.0: An automated pipeline for *in silico* protein functional annotation. *Manuscript in preparation.*
Contribution to (i) concept: 60% (ii) data gathering: 90% (iii) data analysis: 60% (iv) development: 90% (iii) writing: 60%

[4] **Islam MT,** Mohamedali A, Nawar I, Baker MS, Ranganathan S (2016) A systematic bioinformatics approach to identify high quality MS data and functionally annotate proteins and proteomes. In: Mathivanan A, Keerthikumar S (Eds.) *Proteome Bioinformatics, Methods in Molecular Biology, Springer, USA,* vol. 1549, pp. 163-176.
Contribution to (i) concept: 50% (ii) data gathering: 60% (iii) data analysis: 40% (iii) writing: 45%

[5] Baker MS, Ahn SB, Mohamedali A, **Islam MT**, Cantor D, Verhaert P, Fanayan S, Sharma S, Nice EC, Connor M and Ranganathan S (2017) Accelerating the search for the missing proteins in the human proteome. *Nat. Commun*. 8, 14271. Contribution to (i) concept: 10% (review), 80% (MPP); (ii) data gathering: 100% (proteomics data), 15% (review); (iii) data analysis: 40% (proteomics data), 10% (review); (iv) development: 90% (iii) writing: 100% (MPP), review (15%)

[6] **Islam MT**, Shaikh M, Mohamedali A, Ahn SB, Baker MS, and Ranganathan S (2017) An integrated nexus for interpreting and using omics data to find human missing proteins. *Manuscript in preparation*.
Contribution to (i) concept: 60% (ii) data gathering: 70% (iii) data analysis: 50% (iv) development: 65% (iii) writing: 60%

Appendix 1

[1] **Islam MT,** Fernandes CS, Mohamedali A, Baker MS, Ranganathan S (2016) *De novo* peptide sequencing: deep mining of high-resolution mass spectrometry data. In: Mathivanan A, Keerthikumar S (Eds.) *Proteome Bioinformatics, Methods in Molecular Biology, Springer, USA,* vol. 1549, pp. 119-134.

Oral presentations based on thesis work

[1] **Islam MT** and Baker MS**.** MissingProteinPediea (MPP), a platform to uncover the human 'Missing proteins'. Bioinformatics Hub, 15th Human Proteome Organization World Congress, Taiwan, September 2016.

[2] **Islam MT**, Baker MS and Ranganathan S**.** Towards an integrated web platform to uncover the human 'Missing proteins'. Baker Lab, Australian School of Advanced Medicine (ASAM), Macquarie University, Sydney, Australia, August 2015.

Poster presentations

[1] **Islam MT**, Mohamedali A, Shaikh MAM, Ahn SB, Al-Nakli A, Baker MS, and Ranganathan S. MissingProteinPedia (MPP), a platform to uncover the human 'Missing proteins'. 15th Human Proteome Organization World Congress, Taiwan, September 2016

[2] **Islam MT,** Mohamedali A, Garg G, Baker MS**,** and Ranganathan S**.** Unlocking the Puzzling Biology of the Black Périgord Truffle *Tuber melanosporum*. 13th Human Proteome Organization World Congress, Spain, October 2014

[3] **Islam MT,** Mohamedali A, Garg G, Baker MS**,** and Ranganathan S**.** Unlocking the Puzzling Biology of the Black Périgord Truffle *Tuber melanosporum*. InCoB 2014 Conference, Australia, Aug 2014

Awards received as PhD student

[1] **Macquarie University Postgraduate Research Fund (PGRF)**, 2014 (awarded A$ 5000 for International conference travel)

[2] **Macquarie University Research Excellence Scholarship (MQRES)**, 2012-2016.

Abstract

The advancement of next-generation proteomics methodologies has led to an explosion in proteomics data. However, the analysis and interpretation of this data remains a challenge, as several proteins remain unannotated and uncharacterised for many organisms. Despite the presence of the large volume of mass spectrometry (MS) data in various datasets, over 10% human proteins are still considered ‘missing’. Bioinformatics techniques can be used to provide comprehensive annotations for entire proteomes to provide valuable information regarding putative functions of proteins that can be validated and or supplemented with experimental data.

The aims of this thesis are to tackle some of these challenges, firstly to develop a generic *in silico* bioinformatics pipeline to identify homologues and map putative functional signatures, gene ontology terms and biochemical pathways of novel organisms, or “missing’’ proteins. This pipeline was used to identify homologues for 2,587 proteins and functional annotation for 2,486 proteins from black Périgord truffle (*Tuber melanosporum Vittad*), followed by MS-based shotgun proteomics to validate 836 proteins. The same pipeline was then used to annotate the human “missing” protein sequences on each human chromosome available through the ProtAnnotator web portal, with homologues from the mammalian kingdom for 2538 (66.2%, based on September 2013 data). ProtAnnotator also functionally annotated 1945 (50.8%) “missing” human proteins. ProtAnnotator 2.0 automated the process and provides an update to the annotation of the truffle proteome.

The lack of coherency between the proteomics data submitted to various databases, processed by different search engines has limited their integration in the quest for uncovering human “missing” proteins. To this end, a scheme was worked out for comparing proteomics data from different sources, looking at proteotypicity and search engine scores, with guidelines on spectral quality analysis as well.

Finally, ProtAnnotator and the proteomics data integration strategy above were integrated, to create a novel integrated web platform (MissingProteinPedia) to define, collate and make serviceable all available data (including single proteotypic MS spectra) from various databases and web platforms for human “missing” proteins. The MissingProteinPedia (MPP) platform comprises a freely available web interface for datamining, collaboration and validation of MS and publication data. MPP permits protein-level identification of proteins that have very short tryptic peptides, such as interleukin-9, proteins traditionally known but without proteomic or antibody data as well as those that are carefully identified by our integrated computational workflow followed by expert spectral analysis.

The tools developed in this thesis provide data integration to accelerate the annotation of novel proteomes and the discovery of human missing proteins.

Chapter 1: Introduction

1.1 Overview

The genome of an organism is composed of deoxyribonucleic acid (DNA) that contains the biological instructions for life. These instructions form the blueprint for protein production that determines how the cell will function. The genetic information encoded in DNA is translated to proteins via the transcription and translation machinery [1]. These proteins then fold to acquire unique three-dimensional structures to perform specialized functions [2, 3] in the cell involving one or more proteins. Proteins are considered to be the ‘building blocks’ of the cells and play a central role (as molecular machines) to carry out tasks inside the cell [4, 5]. In other words, proteins are the parts of the engine of the cell - if the genetic coding is the roadmap, protein is the vehicle. The proteome, or the complete products of genes, is much more dynamic compared to its counterpart genomics or even transcriptomics [6], and contains specific information such as cell-type- and time-dependent expression patterns and post-translational modifications [7]. Thus, identifications of the proteins and decoding its underlying biological functions are the key to understand life within the cell, tissue or organisms.

Proteomics is the study of a proteome whereby all the protein complement of an organism, tissue or cell (or fraction) are studied in a high throughput manner [8-10]. Mass spectrometry (MS) is the most commonly used technique of choice [11] for the analysis (identification and quantitation) of proteins. The scalability, speed, ease of use and, more recently, accuracy of the data obtained from MS instruments has contributed to this rapid rise [12]. The recent publications of draft human proteomes have also assisted in bringing proteomics into the limelight [7, 13]. The use of mass spectrometry in the biological context extends not only to all corners of the research arena, but indeed to drug development, diagnostics, quality control and personalised medicine [14-16] necessitating a larger number of people having to analyse often very complex data sets. Although this is a very positive sign, it has come attached with some significant challenges, primarily around the storage and dissemination of these vast swathes of data obtained from various MS instruments [17] and concurrently the analysis and interpretation of this data [18].

For example, following the two draft releases of the human genome in 2001 [19, 20] the near-complete or 'finishing' euchromatic human genome sequence was published in 2004. It was expected to encode 20,000–25,000 protein-coding genes [21]. Since 2008, the Human

Proteome Project (HPP) has aimed to identify and functionally characterize the proteome comprehensively [22-25]. It launched the International Chromosome-centric Human Proteome Project (C-HPP) in 2012 [26] with baseline metrics [22] in a bid to accelerate the identification and annotation of the human proteome. The project validated 13,664 out of 20,128 proteins in neXtProt [27] at protein evidence level 1 (i.e. confirmed by mass spectrometry, antibody-capture, Edman sequencing, or 3D structures) [22].

Worldwide efforts were then made to identify and characterize the rest of the proteins and 1982 proteins were added to neXtProt [27] in the following year with high-quality identifications, and at the same time 638 genes were removed from the denominator as "uncertain" or "dubious" [28] resulting in a revised target of 3844 proteins (i.e. the defined "known unknown") to be identified and annotated. Two draft human proteomes [7, 13] were then published in 2014, and 17,294 proteins were identified using liquid chromatography-tandem mass spectrometry (LC-MS/MS) and in-depth bioinformatics analyses [29]. However, 64% of the peptide mass spectra (out of 25 million) did not match to any of the human proteins [30]. An advanced search identified 2,535 novel protein-coding genes from the previously defined non-protein coding regions of the DNA sequences (i.e. the "unknown unknowns"). The remaining unidentified peptides from various studies including this can provide further clues (with better algorithms and techniques) to identify 'missing' human proteins. Despite the significant progress, these draft proteomes remain incomplete and as per the HPP Mass Spectrometry (MS) Data Interpretation Guidelines 2.1 [31], with up to 15% of the predicted human proteins yet to be identified by high confidence mass spectrometry or other experimental methods [32]. This is a common phenomenon for other novel or less studied species [33, 34], and limits the capabilities in understanding proteome and utilizing such information for the improvement of human life. However, the example highlights some interesting challenges and opportunities, such as (i) technological improvements can not only help us to find the ‘known unknowns’ and but also discover ‘unknown unknowns’ (ii) New or improved methods and pipelines are required to deal with new technologies and or ‘unknown unknowns’ (e.g. identification and annotation of novel proteins) (iii) As new datasets are added to the repositories, new analysis tool or pipelines are required to predict functions more accurately from multiple data sources (discovering the ‘unknown knowns’).

This thesis demonstrates the benefit of data integration, collaboration, reusable workflows and pipelines to identify and annotate proteomes using existing knowledge. The specific aims of this thesis and how they have been addressed forms the rest of the thesis, followed by conclusions and future direction.

1.2 Mass spectrometry (MS)-based proteomics

Mass spectrometry is an analytical tool that can provide information about the qualitative and quantitative composition of both organic and inorganic analytes including their structures in complex mixtures based on mass-to-charge (m/z) ratio. It is the most commonly used technique in proteomics to calculate molecular mass to charge ratios (m/z) to identify and quantify ions [35] generated from proteins.

1.2.1 Brief history and the basics of Mass Spectrometry

1.2.1.1 History

Although the term 'proteomics' was introduced in the 1990s [10], the history of mass spectrometry goes beyond a century (see Figure 1.1). Physicist J.J. Thomson is attributed for the invention of mass spectrometry for his work on the 'negatively charged cathode ray particles' [36] and ‘mass spectrograph’ [37]. As the technology continued to develop [38, 39] in the 1940s, it was used by the physicists to discover new isotopes, purify and asses the enrichment of the fissionable ^{235}U isotope in World War II and by the chemists to monitor petroleum refinery system [40, 41]. Following the development of gas chromatography (GC)/MS [42, 43] and tandem mass spectrometry (MS/MS) [44], and the introduction of the “soft” mode chemical ionization [45] in the 1960s, mass spectrometry became the common analytical tool for the analysis of organic compounds [40]. However, due to the limitations of ionization techniques, it could not be used to analyse relatively large biomolecules, which became the catalyst for a series of improvements over the next couple of decades. These were, fast atom bombardment (FAB) [46], electrospray ionization (ESI) [47], and matrix-assisted laser desorption/ionization (MALDI) [48, 49]. It was the latter two methods that set the ground for modern mass spectrometry, and both made it possible to detect and sequence (and hence identify) polypeptides [50-52].

1.2.1.2 Basics of mass spectrometry

In the past decade, proteomics and mass spectrometry have taken tremendous strides forward, spurred on by rapid advances in technology, computing and new applications in biological and biomedical sciences.

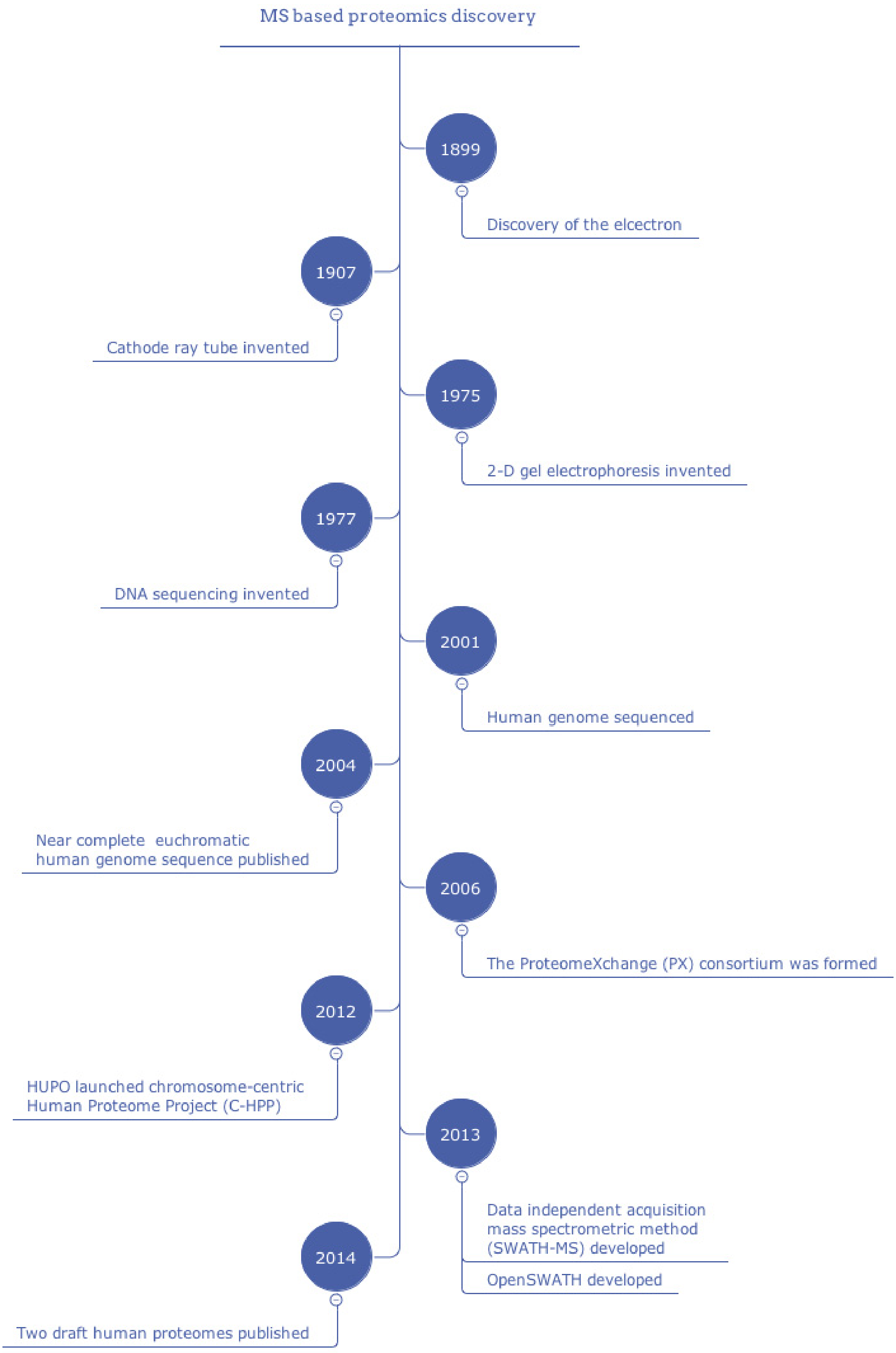

MS based proteomics discovery
1899
Discovery of the elcectron
1907
Cathode ray tube invented
1975
2-D gel electrophoresis invented
1977
DNA sequencing invented
2001
Human genome sequenced
2004
Near complete euchromatic human genome sequence published
2006
The ProteomeXchange (PX) consortium was formed
2012
HUPO launched chromosome-centric Human Proteome Project (C-HPP)
2013
Data independent acquisition mass spectrometric method (SWATH-MS) developed
OpenSWATH developed
2014
Two draft human proteomes published

Figure 1.1: A timeline showing important advances in MS based proteomics discovery over the last century.

Although a number of new high-throughput machines were developed for proteomics [53-55], the basic steps to identify proteins are as follows [40, 56]

- Prepare and isolate samples and digest using trypsin or similar enzyme
- Produce ions from the sample by adding or removing of protons or electron from analyte molecules.
- Separate various ions based on their mass-to-charge ratio (m/z), and then fragment the selected ions (in the case of hybrid systems).
- Detect fragmented ions from the previous step, measure their abundance, and convert the ions into electrical signals.
- Detect and process the transmitted signals on a computer, amplify and display them as a spectrum.

The basic components of MS spectrometer are shown in Figure 1.2.

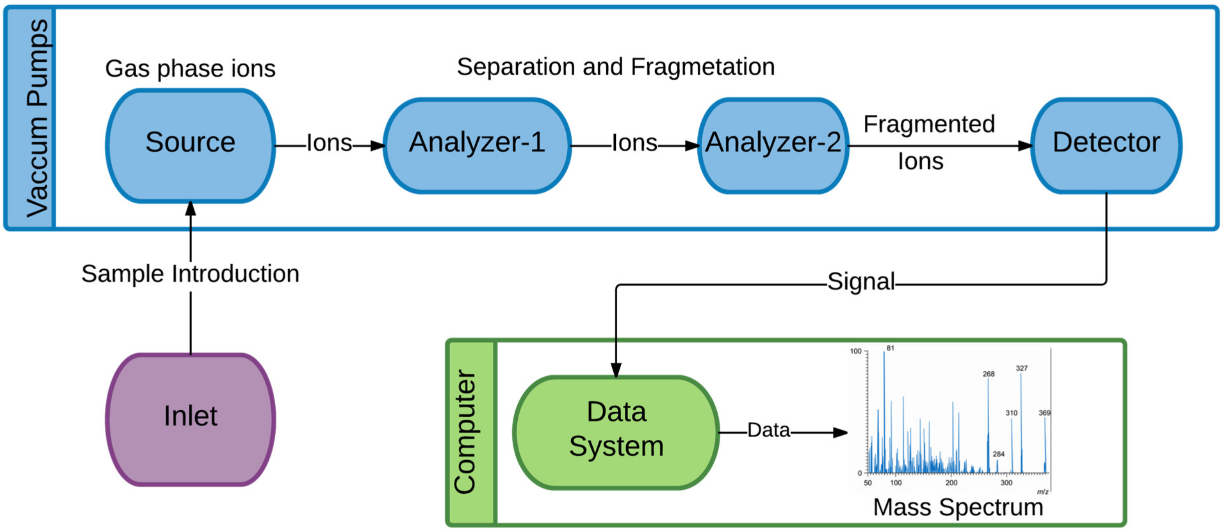


Figure 1.2: Basic components of mass spectrometer.

A basic mass spectrometer has the following components to perform the above tasks:

- *An inlet system* to introduce the sample to the ion source (e.g. gas chromatograph, direct insertion probe)
- *An ion source* to convert the substance into gas-phase ions using ionization techniques. The electron ionization [57] and chemical ionization [45] are commonly used for gas, whereas nanoelectrospray (nESI) [58] and MALDI [59] are commonly used to ionize large biomolecules in solid and liquid samples in shotgun proteomics.
- *A mass analyzer* to separate the ions into their characteristics mass components based on their mass-to-charge ratio. Commonly used mass analyser for proteomics

experiments are (i) trapping type instrument such as - quadrupole ion trap (QIT), linear ion trap (LIT), fourier transform ion cyclotron resonance (FT-ICR), and Orbitrap (ii) quadrupole (Q) (iii) time of flight (TOF) instruments [5]. A mass spectrometer can have one or more analyzers in tandem (usually called hybrid or fusion machines).

- *A detector* to detect fragmented ions from the last analyzer, then measure their abundance and convert the ions into electrical signals. These signals are then passed onto the data system.
- *A data system* to record, store and process the transmitted signals on a computer. It amplifies the signals and creates the MS spectrum.

The ion source, analyzer, and detector are controlled by a vacuum system to maintain the integrity of the sample during the transfer, and to keep the system free from gas molecules to maintain absolute paths for the ion.

1.2.2 Protein identification and characterisation techniques

Despite the rapid advances in mass spectrometry in terms of its accuracy, sensitivity, and throughput we still lack a single technique that can identify all predicted proteins produced by the genome annotation from a sample. Although a combined liquid chromatography (LC) separation and MS method (known LC-MS) can analyse thousands of proteins and peptides from a complex sample, a significant number of the peptide mass spectra remains unidentified [13, 60]. There are two major approaches for mass spectrometry-based protein deification and characterization, namely top-down and bottom-up. The bottom-up approach deals with peptide level information from a digested sample, whereas top-down approach uses intact protein molecules without the proteolytic cleavage [5, 40]. The difference between these two approaches are shown in Figure 1.3. How these two methods are applied in proteomics studies will be discussed first followed by a discussion of the pros and cons of each of these approaches.

Top-down proteomics is used to characterise intact protein molecules. The first step is the gas-phase ionization of intact proteins. Samples are then separated (and/or fractionated) into a single protein or less complex mixtures to calculate the accurate mass measurement of the intact proteins ions using high-resolution mass spectrometry and searched against a molecular mass database. The peptide mass fingerprinting is used for protein identification, or proteins can be separated further and confirmed using on-line tandem mass spectrometry.

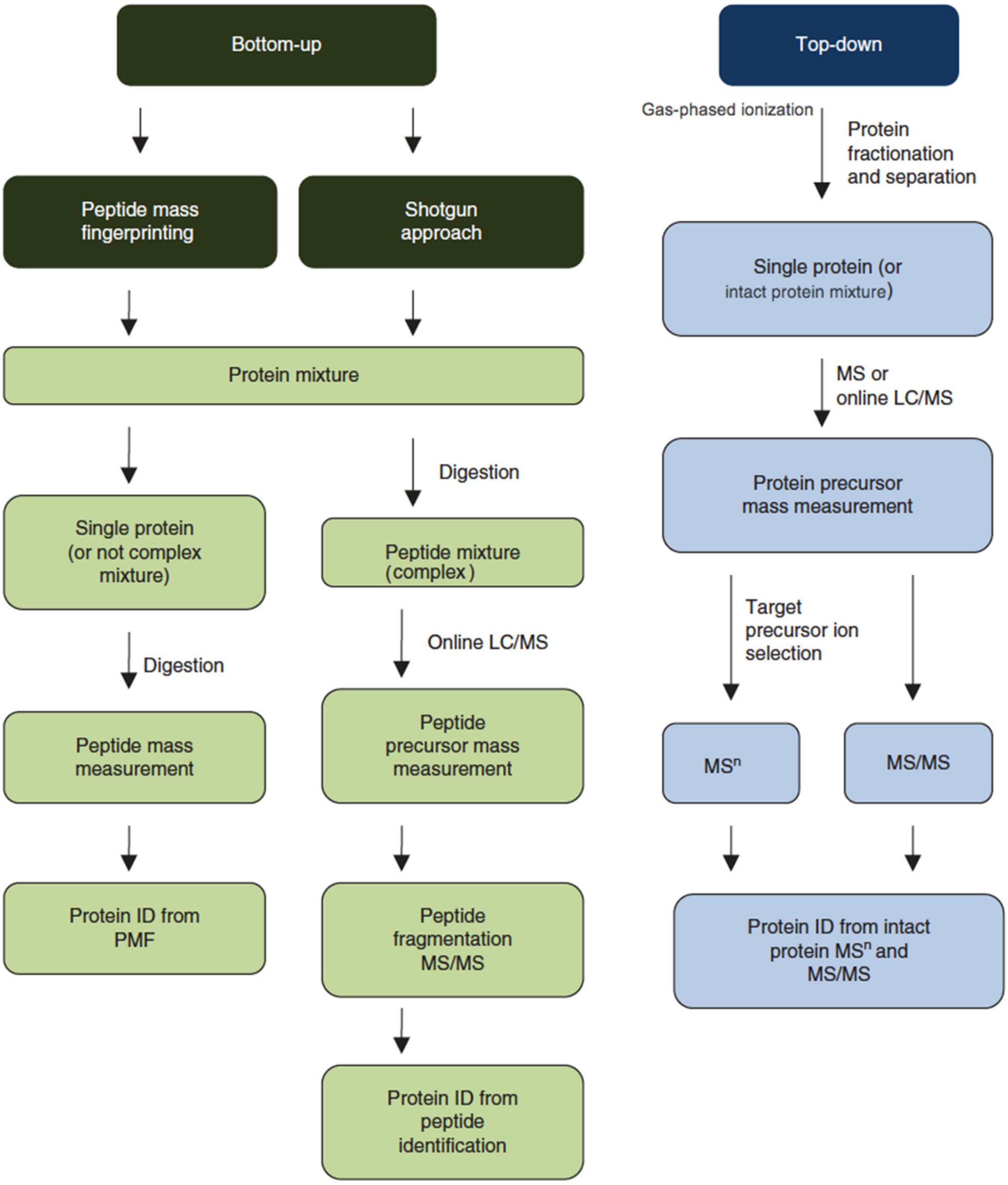

Bottom-up
Peptide mass fingerprinting
Shotgun approach
Protein mixture
Single protein (or not complex mixture)
Digestion
Peptide mixture (complex)
Digestion
Online LC/MS
Peptide mass measurement
Peptide precursor mass measurement
Protein ID from PMF
Peptide fragmentation MS/MS
Protein ID from peptide identification
Top-down
Gas-phased ionization
Protein fractionation and separation
Single protein (or intact protein mixture)
MS or online LC/MS
Protein precursor mass measurement
Target precursor ion selection
MSⁿ
MS/MS
Protein ID from intact protein MSⁿ and MS/MS

Figure 1.3: Strategies for bottom-up and top-down MS-based protein identification and characterisation. Adapted from Sokolowska, 2013.

In the bottom-up approach peptides are used as the analytes to deduce the presence of proteins from highly complex samples. The most commonly used bottom-up techniques are “sort-then-break” and “break-then-sort” approach. In the first approach, samples are separated and fractionated offline first, then digested into peptides. These peptides are then either analysed by peptide mass fingerprinting (PMF) or further separated and analysed in a tandem mass spectrometer.

However, in the "break-then-sort" method also known as “shotgun proteomics” [61], proteins are first digested into peptides using a protease (trypsin being the most commonly used [62]). These peptides are then separated using single or multi-dimensional chromatography and analysed using tandem mass spectrometer (MS/MS) [63]. The tandem MS detects single peptide ions or a set of ions for further fragmentation and finally the peptide sequences are derived from the resulting fragment ions [64, 65].

The potential advantage of top-down approach is its ability to detect protein isoform and post-translational modifications (PTM) more accurately as it analyses the whole sequence instead of a digested pepetide (i.e shotgun/bottom-up approach) [63]. It has been used to measure intact proteins up to 200 kDa using “prefolding dissociation” [66] and to identify over 1000 proteins from complex samples using a "four-dimensional separation" system [67]. However, the approach has significant limitations in protein fractionation, ionization, and fragmentation in the gas phase [62]. Proteins are less soluble under LC-MS compared smaller peptides. Complex protein samples require multiple separation techniques (e.g. combining LC with electrospray ionization (ESI)) [68]. Ionic detergents such as sodium dodecyl sulfate (SDS) are often used to solubilize large proteins. However, ESI is not compatible with SDS. It is difficult to disrupt the tertiary structure of proteins with higher molecular weight. The sensitivity of a mass spectrometer for proteins are lower than peptides[69].

On the other hand, shogun proteomics uses digested peptides (smaller than the whole sequence) therefore it is easier to ionize and the fragment ion spectra is less complex for interpretation. Advances in peptide separation techniques [70-73] make it possible to reduce complex protein into simpler fractions for analysis. Hence shotgun proteomics, despite some limitations, are widely used for proteomic analysis.

It is worth mentioning that, a complementary hybrid bottom-up and top-down approach known as middle-down proteomics is also intoroduced to analyse larger peptides (>6.3 kDa) to gain better detection of protein isoform and PTM without the complexity top-down intact protein analysis [74].

1.2.2.1 Shotgun proteomics for peptide and protein identification

As discussed in above, shotgun proteomics is preferred over other MS techniques for proteomics studies and is predominantly used in proteomics experiments. In this thesis therefore, the shotgun proteomics approach is the primary focus. The technological advancement in MS instrumentations, mass resolution [75, 76], and separation techniques such as chromatography [70, 73] to reduce complex samples have enabled protein detection with higher sensitivity. A brief overview of a generic shotgun proteomics workflow (see Figure 1.4) to identify protein and peptide will be presented below.

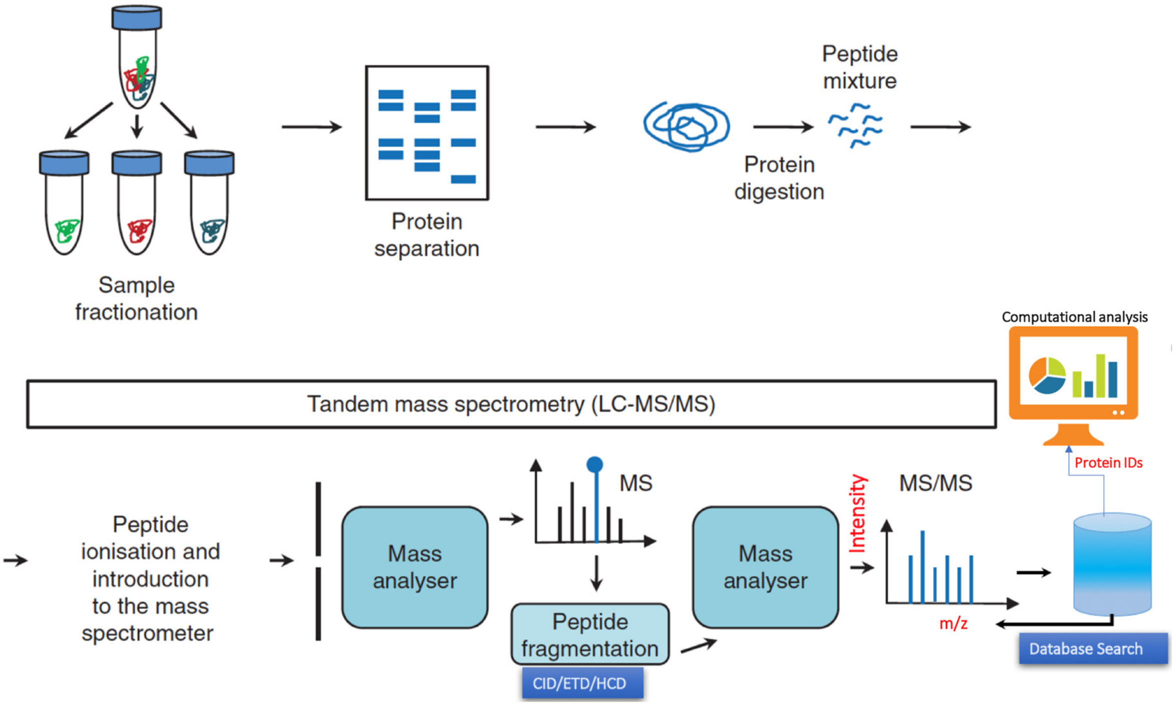


Figure 1.4: Generic shotgun proteomics workflow. Adapted from Sokolowska, 2013.

Most proteomics experiments follow a very similar procedure where the biological protein sample is purified (and/or fractionated), digested typically with trypsin and the peptides either subjected to further peptide fractionation through nano-LC followed by electrospray ionisation in tandem or subjected to MALDI (with or without peptide fractionation). The technologies used to detect the charged protein masses and the accuracy and speed of instruments differ (Ion traps, TOF and others) but the result is nearly identical where a precursor ion mass is obtained in the first MS.

The top (most abundant) precursor masses are usually then subjected to additional dissociation (CID, ETD, HCD) to break up the peptides into individual amino acids. Typically, the resulting raw data obtained from a MS instrument is analysed by searching against a database using a search engine which considers numerous variables including, intensity, error, instrument, mass modifications (in form of posttranslational modifications or introduced experimental peptide modifications as consequence of sample preparation (such as oxidation of the methionine residues) or an inherent part of the experiment (e.g. isobaric tagging). Identified proteins can be subject to further computational analysis (i.e. functional annotation). A wide range of algorithms and computational techniques are available to perform protein and peptide identification in shotgun proteomics.

1.2.3 Algorithms and computational protein and peptide identification techniques for shotgun proteomics

Over the past decade significant efforts were made to accurately identify and characterise proteins [22-25] including the release of two mass spectrometry based draft human proteomes in 2014 [7, 13]. There are two common techniques to identify proteins in shotgun/bottom-up proteomics [77], they are (i) Peptide mass fingerprint (PMF) techniques, where observed peptide masses are matched with theoretical mases of peptides from a protein mass sequence database (ii) Tandem MS based technique where the predicted peptide sequences from the experimental spectrum are either matched against theoretical fragmentation (*in silico*) of amino acids or cleaved peptides (*in silico* digestion) of a protein sequence from a protein database.

A typical Peptide mass fingerprint (PMF) workflow is to separate the proteins by 2D polyacrylamide gel electrophoresis (2D-PAGEThen use MALDI-MS to analayse and digest selected gel spots to produce one spectrum per spot. Then use MALDI-MS to digest analyse selected gel spots to generate a single spectrum per spot. Peptide masses are calculated from these spectra and compared to the masses from the digested peptide sequence (*in silico*) from a protein sequence database using various algorithms (described later in this section). This method has some advantages such as less instrumental overheads, better throughput, and its ability to eliminate the protein inference problem. However, it cannot differentiate peptides with identical masses, therefore less accurate than the tandem MS technique [77].

There are various software packages and algorithms available for protein identification using PMF (listed in Table 1.1). Aldente [78] and GPMAW [79] are a set of bioinformatics tools that allows protein identification using PMF. The MOWSE [80-82] score algorithm is based on statistical significance and is used by MASCOT [81] . In addition MASCOT also used MS-FIT [83] an extended probability-based MOWSE score. A complimentary tool, ProFound [84] is based on the Bayesian probability scoring system .

Tandem MS is one of the most commonly used shotgun proteomics methods that analyses m/z values from the MS/MS to infer proteins present in a sample by assigning peptide sequence to a spectrum either by database search or *de novo* searching method [85].

The *de novo* approach is database independent, as it uses the mass differences from MS/MS spectra to calculate the peptide sequence directly [86-88]. It is a very useful technique to identify proteins for organisms where genome sequence is not available or partially sequenced. Since this method doesn't rely on a protein database, it can potentially identify new peptides [89]. However, this method requires lot of compute resource and high-quality MS/MS spectra [90]. In most cases, MS spectra do not contain all fragment ions, and the method provides incomplete sequences and low sensitivity with high false positives [89, 91]. Hence, the database search methods preferred over *de novo* search [92].

The database search method is used more widely in shotgun proteomics [93, 94]. In this method, every experimental MS/MS spectrum from the peptide fragmentation is compared against theoretical fragmentation patterns of the peptide spectrum from the search database to identify the best possible match. Protein sequences from the search database are first digested using the input protease, then theoretical m/z spectra for the cleaved peptides are created using the monoisotopic mass of their individual amino acids.

A scoring algorithm based on the search tool used (described later in the section) is used to score the matches and the best scoring spectrum is reported as the origin of the spectrum [95]. The basic database search workflow is shown in Figure 1.5, it shows how a peptide is identified using database search method. A range of parameters and techniques are used to maintain the quality of the result [90, 96] and proteins are inferred by mapping filtered high-quality peptides to protein sequences [90, 97]. For example, spectra for the identified peptides are searched against 'decoy' protein databases [33,34] to avoid random matches. This search result is used to calculate the false discovery rate (FDR). The final peptide match results are filtered based on the set FDR value to maximise sensitivity and specificity [98-101].

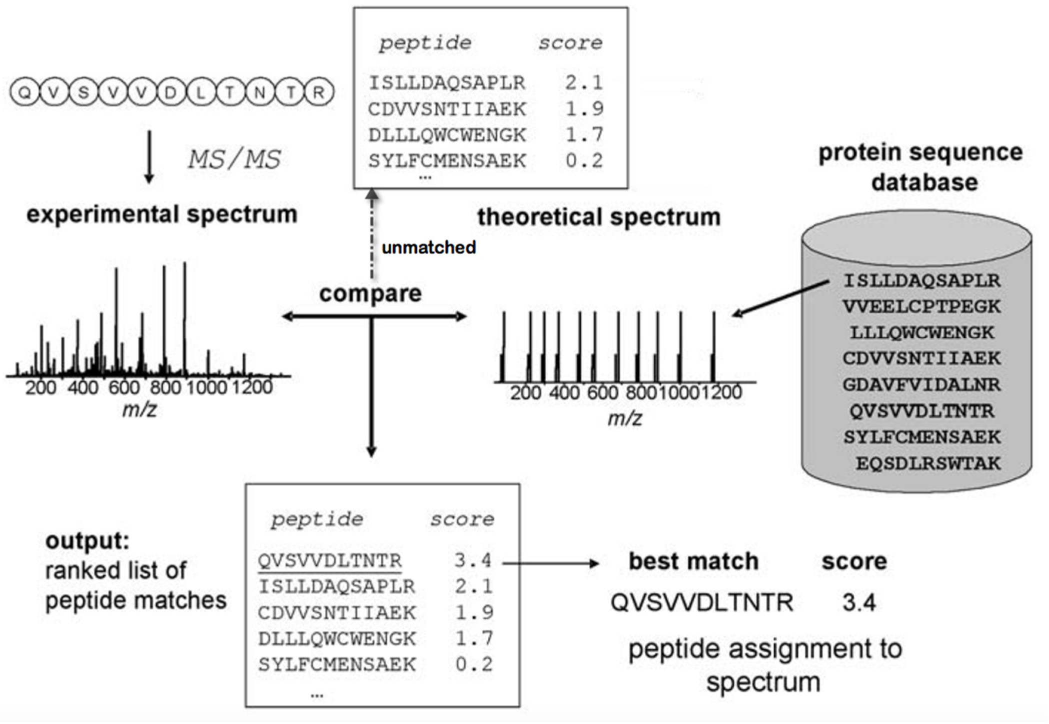


Figure 1.5: Tandem mass spectrometry (MS/MS) database searching. Adapted from *Nesvizhskii et al.*, 2007.

The scoring algorithm is one of the key components of the database search method. In general, all algorithms take the experimental spectra as input and match it against the spectral library from a protein database to assign a matching score. Then apply a statistical algorithm to infer the protein with maximum significance [98]. The key differentiators are the scoring function and the search databases. SEQUEST [93] is one of the first algorithms used for protein identification. It applies a cross-correlation (X-Corr) method to compare an experimental spectrum with the theoretical spectrum from the database. MASCOT [81] was originally built on the probability-based MOWSE score to identify proteins using peptide mass fingerprinting (PMF) though it was later extended for tandem MS identification using a proprietary algorithm. Later on X!Tandem [95], was produced as a free web based tool which calculates e-value using a hyper-geometric model for peptide spectrum match (PSM) ranking. ProbID [102] uses a Bayesian model-based probability algorithm, and OMSSA [103] uses Poisson scoring model to calculate PSM scores. MaxQuant [104] uses the integrated Andromeda [105] peptide search engine for identification. MassWiz [106] scoring function uses the number and continuity of matched fragment ions, associated intensity, and instrument specific fragmentation pattern to evaluate PSMs.

Table 1.1 Examples of software tools and databases for MS-based proteomics

Category	Tool	URL
PMF Packages	Mascot	http://www.matrixscience.com/search_form_select.html
	MS-Fit	http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msfitstandard
	ProFound	http://prowl.rockefeller.edu/prowl-cgi/profound.exe
	GPMAW	http://www.gpmaw.com/html/gpmaw.html
	Aldente	http://www.expasy.org/tools/aldente
De novo sequencing	GutenTag	https://code.google.com/archive/p/gutentag
	InsPecT	http://proteomics.ucsd.edu/software-tools/inspectms-alignment
	Lutefisk	http://www.hairyfatguy.com/lutefisk
	PEAK	http://www.bioinfor.com
	MS-Blast	http://genetics.bwh.harvard.edu/msblast/
	FASTA	http://www.ebi.ac.uk/Tools/sss/fasta/
	PepNovo	http://proteomics.ucsd.edu/software-tools/531-2
	DeNovoGUI	http://compomics.github.io/projects/denovogui.html
	pNovo+	http://pfind.ict.ac.cn/software/pNovo/index.html
	UniNovo	http://proteomics.ucsd.edu/software-tools/uninovo
	MRUniNovo	http://bioinfo.hupo.org.cn/MRUniNovo/index.php
Database search	SEQUEST	http://thermo.com
	SpectrumMill	http://www.agilent.com/home
	Mascot	http://matrixscience.com
	ProteinPilot	http://www.absciex.com
	ProteinPilot	http://www.absciex.com
	Protein-Prospector	http://prospector.ucsf.edu
	Proteinlynx GlobalServer	http://www.waters.com/waters/en_AU/ProteinLynx-Global-SERVER-

Category	Tool	URL
		%28PLGS%29/nav.htm?cid=513821&locale=en_AU
	ProbID	http://tools.proteomecenter.org/wiki/index.php?title=Software:ProbID
	X!Tandem	http://www.thegpm.org/tandem
	MS-GF+	http://proteomics.ucsd.edu/software-tools/ms-gf
	MS-GFDB	http://proteomics.ucsd.edu/software-tools/ms-gfdb
	Morpheus	https://sourceforge.net/projects/morpheus-ms
	MetaMorpheus	https://github.com/smith-chem-wisc/MetaMorpheus
	MS Amanda	http://ms.imp.ac.at/?goto=msamanda
	Andromeda	http://www.coxdocs.org/doku.php?id=maxquant:andromeda:start
	MyriMatch	https://medschool.vanderbilt.edu/msrc-bioinformatics/software
	MaxQuant	http://www.coxdocs.org/doku.php?id=maxquant:start
	OMSSA	ftp://ftp.ncbi.nlm.nih.gov/pub/lewisg/omssa
	MassWiz	https://sourceforge.net/projects/masswiz
Sequence tag and combined approach	InsPecT	http://proteomics.ucsd.edu/software-tools/inspectms-alignment/
	MSblender	http://www.marcottelab.org/index.php/MSblender
	COMPID	http://users.utu.fi/lanatr/compid.
	Popitam	https://code.google.com/archive/p/popitam
	TagRecon	https://medschool.vanderbilt.edu/msrc-bioinformatics/software
	DirectTag	https://medschool.vanderbilt.edu/msrc-bioinformatics/software
	Byonic^{TM}	http://www.proteinmetrics.com/products/byonic
Spectral library search	SpectraST	http://www.peptideatlas.org/spectrast
	X!P3	http://p3.thegpm.org/tandem/thegpm_ppp.html
	Bibliospec	skyline.gs.washington.edu

Category	Tool	URL
Multi-platform search engine and unified scoring	iProphet	https://www.systemsbiology.org/resources/software-downloads/
	Scaffold	http://www.proteomesoftware.com/products/scaffold/
	PepArML	http://peparml.sourceforge.net/
	MSblender	http://www.marcottelab.org/index.php/MSblender
	FDRAnalysis	http://code.google.com/p/web-based-multiplesearch
Spectra quality	MassWiz	https://sourceforge.net/projects/masswiz
	MaXIC-Q	http://ms.iis.sinica.edu.tw/COmics/Software_MaXIC-Q.html
	IDEAL-Q	http://ms.iis.sinica.edu.tw/COmics/Software_IDEAL-Q.html
MS data management and spectral libraries	PeptideAtlas	www.peptideatlas.org
	Proteios	www.proteios.org
	SBEAMS	http://www.sbeams.org/project_description.php
	CPAS	www.labkey.org/
	PRIDE	www.ebi.ac.uk/pride/
	MASPECTRA-S2	http://genome.tugraz.at/maspectras/maspectras_description.shtml
	Proteom-Xchange	www.proteomexchange.org
Analytics platforms and pipelines	MaxQuant	www.biochem.mpg.de/en/rd/maxquant/
	Trans-proteomic pipeline	http://tools.proteomecenter.org/wiki/index.php?title=Software:TPP
	Chorus	https://chorusproject.org/pages/index.html
	Galaxy-P	https://usegalaxyp.org/static/welcome.html
	Firmiana	http://www.firmiana.org/
	Labkey	https://www.labkey.org

Category	Tool	URL
Collection of tools and databases	OMICtools	https://omictools.com/proteomics-category
	Expasy tools	http://www.expasy.org/tools/
	EBI tools and databases	https://www.ebi.ac.uk/services
	NCBI data-bases and tools	https://www.ncbi.nlm.nih.gov/guide/data-software/
	Uniprot	www.uniprot.org
	neXtProt	www.uniprot.org, www.nextprot.org
	SPC-Prote-omics tools	http://tools.proteomecenter.org/software.php

This thesis focusses on the results from the algorithms, not detailing the efficacy or utility of each algorithm as most are proprietary, are already established in proteomics laboratories or have been extensively reviewed and therefore beyond the scope of this study.

Some hybrid software platforms have also been developed in a bid to improve the algorithms to identify more peptides. InsPecT [107] uses a hybrid tag based approach to identify best matching peptide for a spectrum. MSBlender [108] converts PSM scores from multiple search engines into a probability score for all possible matches (PSM) while considering the correlation between search scores to maximise the number of identified PSMs. COMPID [109] integrates Mascot [81] and Paragon [110] algorithms.

The availability of this wide range of free and commercially available search engine software each with their own unique algorithms and search parameters for matching spectra meant that an accurate assessment of the quality of search results can be very difficult in a highly collaborative environment. This aspect is also a significant part of determining the accuracy of an entire proteomic assessment and subsequent downstream analyses. In recent years though, there is a significant preponderance to the use of commercially available search engines primarily due to ease of use and often intuitive interface. Although proteomics data analysis can contain a significant amount of specialised data (including modifications, transitions etc.) the majority of users of proteomics technologies utilise it for identification of proteins [111] or increasingly quantitation of protein [112] upon successful identification.

Increasingly, most users, though having access to raw data, rely on search engine processed data to draw conclusions and make inferences on the nature and quality of the biological sample [112]. This could be due to several reasons ranging from lack of expertise to poor or no access to proprietary software for reanalysing or reprocessing raw MS files. In many cases, the data produced by these machines can contain noise/bias. This can lead to incorrect identification if the searches are not conducted carefully with correct search engine specific parameters. So, it is important to (a) use an optimised set of parameters (depending on the search engine) for the database search (b) have a good understanding on how to interpret the various engine scores to identify the proteins correctly. As there are no standard sets of parameters for all search engines, users must optimise the key parameters to obtain best possible results. Some of the key parameters are the enzyme, parts-per-million (ppm) value of mass tolerance, the number of miss cleavages, choice of PTMs. Some published guidelines [90] and a detailed review [92] of the parameters for various search engines are also available online. Several quality scoring methods such as false discovery rate (FDR), P value, Q-value, E value, posterior error probability (PEP) have been developed to add confidence to the statistical scores from various search engines. Many software tools like iProphet[113], Scaffold [114], PepArML [115], MSBlender [108] and FDRAnalysis [116] provide a platform to merge data from various search engines and reprocess them to give a unified score. This however comes at the cost of increased complexity, computational time requirement and ability to interpret yet another set of scores [117]. Although ideally, a reanalysis of raw data through a standard protocol would yield the best indicators of data quality, this may not be practical for all studies especially for smaller studies.

A major further compounding factor is that most freely available and proprietary search engine software do not take into consideration the quality of mass spectra. Most, if not all algorithms rely on matching theoretical spectral masses (or m/z) to those obtained from an experiment and provide a confidence score of the similarities based on a user defined error margin. This often means that poor spectra (for instance, those with a high ratio of noise to sample, low-intensity spectra, etc.) could potentially get high scores. Incorrect sequence assignments often arise due to deficiency in scoring schemas, low spectral quality, fragmentation of multiple peptide ions, presence of homologous peptides, incorrectly determined charged state or peptide mass, restricted databases for searching and variation in sequence of peptides [118]. More recently, the recognition that confidence scores and identity scores for peptide identification being insufficient have sparked interest in determining quality checks of actual raw spectra. Several methods are developed to assess the quality of the

PSMS, such as single spectrum methods, multiple spectra methods, and target-decoy analysis methods [98, 101]. A few review articles have been published to compare these methods as well defining some criteria for assessing the quality of the PSMs [96, 119, 120]. Indeed, several more sophisticated software (Manual Analysis Emulator (MAE) [121], MaXIC-Q Web[122], IDEAL-Q [123], MassWiz [106]) do exist to automate this effort but they require a re-analysis/reprocessing of the data, downloading and installing software which is not only cumbersome but often not possible for most proteomics labs. Many peer reviewed journal articles including large studies [7, 13] now add a manual spectral validation step but often refer to spectra being analysed by 'expert or experienced mass spectrometrists' without clear defining criteria. These manual validation methodologies are often subjective with criteria either undefined or varying between groups. Furthermore, such validation is not feasible for most non-specialists or educational purposes without appropriate guideline or method.

The database search method is dependent on the protein sequences available in the database. Unlike *de novo* sequencing, a protein or peptide cannot be identified if it is not available in the target database [92] even if the protein is present in the sample. The protein database comprises a set of protein coding genes of an organism that are derived using a gene prediction tool as part of the genome annotation process. The most common gene prediction techniques are *ab initio* gene prediction using a gene model and sequence similarity of genes [124, 125]. A wide range of software and tools are available to predict genes for prokaryotes and eukaryotes. These include, but are not limited to Glimmer [126], Prodigal [127], GeneMark [128], GENSCAN [129], AUGUSTUS [130] and MAKER2 [131]. With the advent of next generation sequencing technology, there has been a rapid increase in the number of completely sequenced genomes as well as the public databases that store these genomes. Despite the advances, genome annotations are not free from errors. Some of the examples include incorrect assignment of gene start site [132], translation of a gene incorrectly annotated as pseudogene [133], and unannotated new genes [134, 135]. Accurate annotation of protein-coding gene boundaries is essential for protein identification as well as downstream functional annotations. The predicted sequences generated from these annotations can then be further analysed functionally.

1.3 *In silico* protein identification and functional annotation

This section will briefly outline some of the *in silico* annotation techniques using existing information from various databases.

1.3.1 Database similarity searches

Sequence homology technique is commonly used for functional annotation. Sequences are considered homologous if they share common ancestors, therefore share similar biological functions. *In silico* sequence alignment tools such as Basic Local Alignment Search Tool (BLAST) [3, 136] can be used to determine the similarity between a query sequence and the subject sequence ignoring any evolutionary changes. The sequence identity score is a key indicator of similarity. A query sequence can be considered strongly homologous if it matches with a subject sequence with at least 50% identity with an expect value 1e-05 [137, 138]. BLAST offers specific tools for various sequence types. For example, Translated BLAST (BLASTX) to search gene sequences and Protein BLAST (BLASTP) protein sequences against a protein sequence database, and Nucleotide BLAST (BLASTN) to search gene sequences against nucleotide sequence database. Several other alignment tools offer similar functionalities such as HHblits [139], BLAT [140], and MGAlignIt [141]. Any protein sequence database can be used for the searches including some of publicly available databases such as UniProtKB [142], TrEMBL [143, 144], PDB [145, 146] (described later in section 1.4). Although almost every database provides some annotations, it is important to use one or more closely related protein sequence database to identify homologous sequences. Where possible, a reviewed protein sequence database with experimental evidence is preferred to avoid matches against unannotated or translated coding regions [138, 147]. Although the similarity searches identify similar or homologous proteins, further functional analysis can be carried out to get a better understanding of their functions.

1.3.1 Functional annotation

One of the common techniques to identify biological functions of a gene or protein is to identify the over-represented functional categories from a set of differentially expressed genes or proteins with common biological properties [148]. Gene Ontology (GO) [149, 150] provides hierarchically controlled vocabularies (CV) to classify gene function and provides a unified description of biological, cellular and molecular functions across genomes. There are three high-level categories in GO annotation - the biological process that defines a series of events, the molecular function describes activities at the molecular level, and the cellular component describes locations on the subcellular structures and macromolecular complexes. Each of these categories is then further expanded with various evidence codes. These controlled vocabularies are a good candidate for functional annotation and has been widely used by many annotation tools for example, BLAST2GO [151], PANNZER [152] , FunFHMMer [153, 154], dcGOR [155], and AgBase-Goanna [156] to annotate protein sequence. Several other tools have been developed by the members of the Gene Ontology Consortium, and are listed in the Gene Ontology website [157]. Another method is to scan protein or gene sequences against predicted protein signature models to identify domain, motifs and associated GO terms. The InterPro [158] database integrates such predicted signature models from 14 member databases including the two recently added database SFLD [159, 160] and CDD [161]. InterProScan [162] annotation package combines all protein signature method to a native form to look up corresponding InterPro and GO annotation.

Table 1.2. **List of InterPro member databases**

Database	Description	Reference
CATH-Gene3D	Predicted structural fold, protein family and domain organisation of proteins.	[163, 164]
CDD	A collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins.	[161]
MobiDB	A centralised resource of manually curated, indirect and predicted annotations of intrinsic protein disorder.	[165, 166]
HAMAP	A high-quality annotation database with manually curated profiles for protein sequence family classification and expert-curated rules for functional annotation of family members.	[167]
PANTHER	A large annotation database of protein families and function of protein-coding genes from 104 completely sequenced genome using human expertise.	[168]
Pfam	A large collection of protein families primarily based on the UniProtKB reference proteomes derived by multiple sequence alignment and hidden Markov models.	[169]
PIRSF	A comprehensive resource for comparative analysis of protein function and evolution providing multiple levels of sequence diversity from superfamilies to subfamilies reflecting the evolutionary relationship of full-length proteins and domains.	[170, 171]

Database	Description	Reference
PRINTS	A collection of groups of conserved motifs used to characterise a protein family also known as fingerprints.	[172]
ProDom	A protein domain families database produced by clustering homologous segment from the UniProt Knowledge Database.	[173]
PROSITE	An annotation database of protein domains, families and functional sites including their identification patterns and profiles.	[174]
SFLD	A hierarchical classification of functionally diverse enzymes superfamilies that relate to homologous sequence-structure to specific chemical capabilities.	[159, 160]
SMART	A resource for identification and extensive annotation of genetically mobile protein domains and the exploration of domain architectures.	[175]
SUPER-FAMILY	A library of structuraland functional annotation of genes and proteins based on a collection hidden Markov models.	[176, 177]
TIGRFAMs	A protein family database containing multiple sequence alignments, Hidden Markov Models (HMMs) for protein sequence classification.	[178]

A complete list of InterPro member databases is provided in Table 1.5. Although there are some overlaps between the databases (as well as the tools), each of them is developed for the specific purpose so the software and databases must be selected carefully to the research goal [179]. The InterProScan package can also detect pathways.

1.3.2 Pathway analysis

A biological pathway is a series of interactions between various biochemical compounds like gene, protein, protein complex, and metabolites within a cell that creates a change in the cell. Therefore, understanding which protein and genes are involved in a pathway of interest can provide vital information about protein functions and overall biology. The Kyoto Encyclopedia of Genes and Genomes (KEGG) [180-182] is an integrated database that was developed to establish links between genes and high-level functions of the cell of the organism. It comprises of 16 different databases in four categories. These categories include systems information, genomic information, chemical information and health information. KEGG LIGAND integrates all chemical information and KEGG MEDICUS acts as the reference point for all information under the health category and extends KEGG's application programming interface (API) for computational data analysis [180].

Table 1.3. List of KEGG resources.

Category	Database	Content	URL
KEGG			http://www.genome.jp/kegg/
System information	PATHWAY	KEGG pathway maps	../pathway.html
	BRITE	BRITE functional hierarchies and BRITE tables	../brite.html
	MODULE	KEGG modules	../module.html
Genomic information	ORTHOLOGY	KEGG Orthology (KO) groups	../ko.html
	GENOME	KEGG organisms (complete genomes)	../genome.html
	GENES	Gene catalogs of KEGG organisms, viruses, plasmids and addendum category	../genes.html
	SSDB	GENES sequence similarity	../ssdb/
Chemical information (KEGG LIGAND)	COMPOUND	Metabolites and other small molecules	../compound/
	GLYCAN	Glycans	../glycan/
	REACTION	Biochemical reactions	../reaction/
	RPAIR	Reactant pairs	../document/help_bget_rpair.html
	RCLASS	Reaction class	../reaction/

Category	Database	Content	URL
	ENZYME	Enzyme nomenclature	../annotation/enzyme.html
Health Information (KEGG MEDICUS)	DISEASE	Human diseases	../disease/
	DRUG	Drugs	../drug/
	DGROUP	Drug groups	../drug/
	ENVIRON	Crude drugs and health-related substances	../drug/environ.html

A complete list of KEGG resources is provided in Table 1.6. Some of the pathway mapping tools include, KEGG Orthology Based Annotation System (KOBAS) [183] which is a KEGG orthology (KO) based pathway identification web server, KEGG Automatic Annotation Server (KASS) [184] performs BLAST against KEGG GENES database for annotation, InterProScan [162] annotation also provides pathway annotation, and PathFinder [185] identifies signaling pathways in protein-protein interaction networks.

1.4 Proteomics databases

The advances in proteomics technologies complemented by the growth of the genomics technology (including the increase of genome sequences data) created an unprecedented growth the number of publicly available proteomics data repositories. The focus of these databases ranges from the mere repository of raw data to being a knowledgebase. The sharing of experimental proteomics is becoming a norm, as some scientific journals either require [186] or recommend [187, 188] researchers to submit their raw proteomics data. The shotgun proteomics experiments can produce three types of data in its life cycle. These are (a) raw data (i.e raw MS files), (b) processed results (i.e. identification and quantification information), (c) the research findings [17]. Each of these category can also contain various metadata such as quality information (QC), instrument parameters, etc. Since these datasets come from a wide range of instruments and vendors storing, analyzing and sharing these datasets without some a) common or standard data format b) overarching submission and discovery platform is a big challenge. Several data standards were developed, such as the Proteomics Standards Initiative (PSI) (led by HUPO-PSI [189]) over the years to address the first problem. For example, mzML [190] for MS data, mzIdentML [191]for MS identification, mzTab [192] for identification and quantification, mzQuantML [193]for quantification, and TraML [194] transition lists for targeted proteomics studies [195]. To address the second issue, the ProteomeXchange (PX) consortium [196] was formed in 2011 to standardise the submission process and to store and sharing data using unique identifiers (PXD number). Currently, users can submit data using any of the ProeomeXchange member repositories (PRIDE [197], PeptideAtlas [198], PASSEL [199], MassIVE [200] and jPOSTrepo [201]). Other commonly used databases are UniProt [142], neXtProt [27], GPMDB [202], ProteomicsDB [7], MaxQB [203], MOPED [204, 205], PaxDb [206], Human Proteinpedia [207], and the human proteome map (HPM) [13]. Typically, proteomics databases focus on storing data for one or more of the three categories mentioned earlier in this section. The next section will provide brief descriptions of some of the key databases.

1.4.1 ProteomeXchange (PX) consortium databases

The ProteomeXchnge [196] provides an overarching platform to submit and share data using community standard data format to its member repositories. It stores the standard experimental and technical metadata from member repositories using the PX XML format and assigns a unique and universal PXD identifier to all dataset. Depending on data and workflow, users can either make a "Complete" or "Partial" submission to any of its member repositories (refer to Figure 1.6 for more information). The key difference between the two submissions are if the protein/peptide identifications information is submitted in a standard format, the host repository can link the MS spectra directly to the submitted identification results. However, for partial submission, the unlinked search engine results are made available for download. All submission receives a PXD id, but a full submission with peak lists receives a Digital Object Identifier (DOI). Irrespective of the host repository, it provides a federated data search and discovery via the ProteomeCentral portal. All host repositories store metadata and raw data for both “complete” and “partial” submissions. All PX nodes allow users to embargo their data until publication and provide private access to the journal reviewers.

The PX host repository also have specific functions and features which will be outlined below.

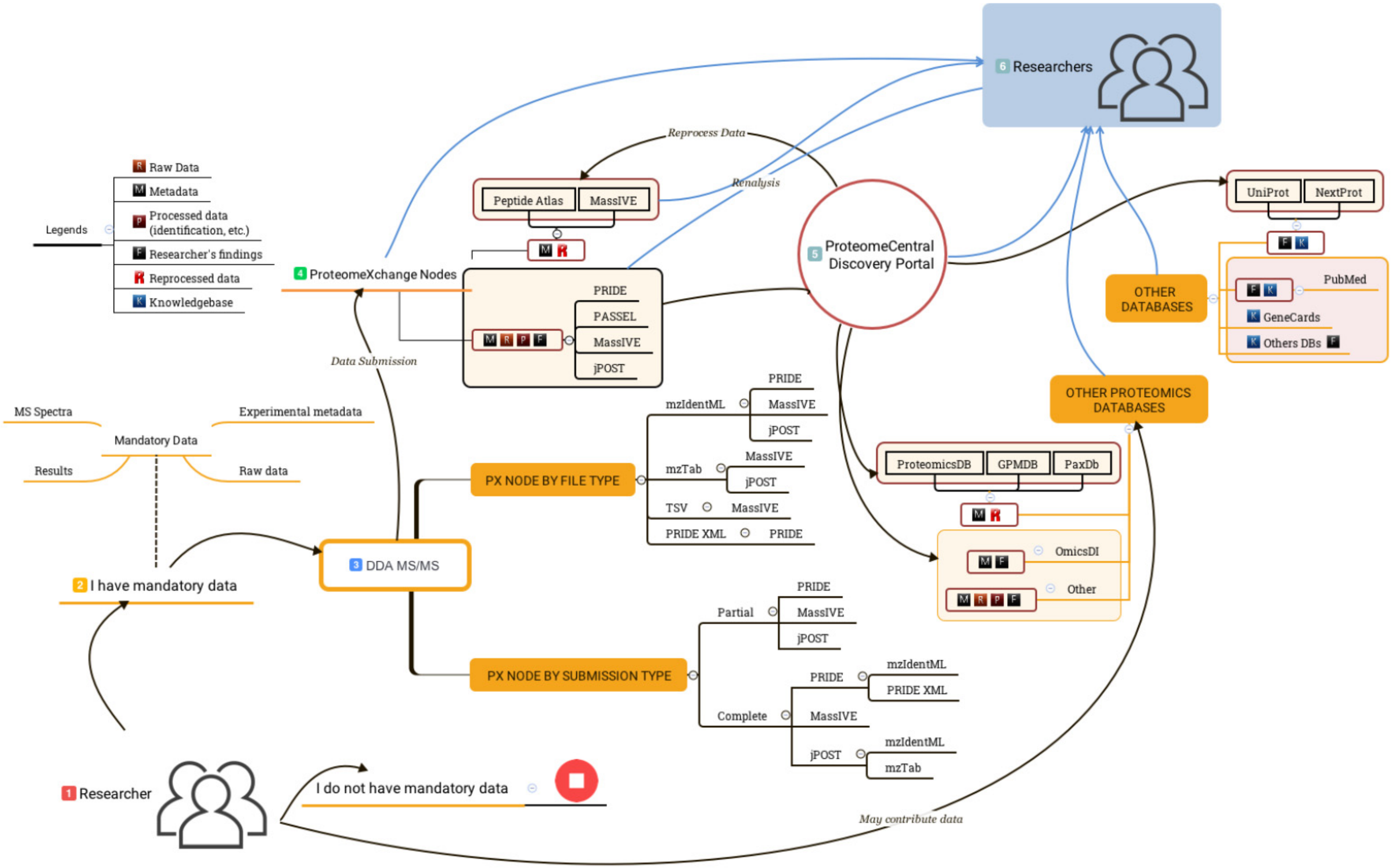


Figure 1.6: Current state of proteomics databases and a summary of submission guidelines for various ProteomeXchange (PX) repositories, depending on the data and submission types for DDA MS/MS experiments.

1.4.1.1 PRIDE

PRoteomics IDEntification (PRIDE) [197] is one of the most matured, comprehensive, and founding node of the PX consortium. It stores both raw and processed data (i.e. identification, PTMs, expression). In addition to the standard ProteomeXchange submission platform, PRIDE also provides a high-speed command line submission tool for larger datasets [208]. For the "complete" submission, PRIDE requires users to submit identification data using PRIDE XML, or PSI mzIdentML (with the corresponding peak lists) format. It provides an open source tool [209] to convert data from the non-standard format. The "partial" submission system doesn't have this requirement and is suitable other MS techniques such as SWATH -MS (DIA), and search engine results that are not supported by any conversion tool. PRIDE provides a web-based data discovery portal [210] [PRIDE Archive] to search and download data. Users can search data and access data at various level: project, protein, peptide, assay. It provides File Transfer Protocol (FTP) and command line access to the data. Complete submissions are processed in a way that end users can download the files, and inspect the spectra using PRIDE Inspector tool [211]. It also provides RESTful web services [212] for programmatic access to the data. In addition to this, PRIDE also provides reprocessed "spectrum-clustering" data via the PRIDE cluster web interface [213].

1.4.1.2 PeptideAtlas

PeptideAtlas [198] is a widely used curated data repository for protein expression data for shotgun proteomics. Data can be submitted to one of the PX nodes or via the PeptideAtlas online submission portal [214]. PeptideAtlas then re-analyses the raw LC–MS/MS spectra and organises them into various builds (data from single proteome or subproteome). The reanalysis pipeline first searches the spectra against a comprehensive proteomics database (UniProtKB/Swiss-Prot [142], Ensembl [215] and International Protein Index (IPI) [216]), then processed using Trans-Proteomics Pipeline (TPP) [217]. The TPP pipeline leverages a range of protein and peptide identification tools – PeptideProphet [218], InterProphet [113], ProteinProphet [219] to assign probability of the identification. MAYU [220] was developed to control the protein FDR when various databases are combined for identification. Finally, the proteins are annotated with supporting information such as genome mappings proteotypicity, etc. to make the final build of a release. The reanalysed datasets are assigned with RPXD identifiers to link them with the original PX dataset. The PeptideAtlas webserver acts as the discovery portal. It also provides a "Chromosome Explorer" [221] to provide a chromosome centric view of the information, and a customised dedicated discovery portal [222] to assist the HUPO - CHPP ‘missing’ protein identification initiative [23].

1.4.1.3 PASSEL

PASSEL [199] stores targeted proteomics quantification data. It is the Selected reaction monitoring (SRM) experiment library (both experimental and raw data) of the PeptideAtlas. Users can submit data using the online data submission portal [223]. Users are required to provide transition list for SRM data in mProphet [224] format as PASSEL reprocesses all data using mQuest/mProphet to maintain consistency throughout. The PASSEL reprocessing workflow also incorporates the measured transitions into the SRMAtlas [225]. A submission template along with supported data formats are also available online [214], and each submission also receives a 'PASS' identification number. These submitted files along with the reprocessed data are made available to the wider scientific community portal via the PASSEL data discovery portal [226]. The portal provides two discovery services to access SRM experimental data (raw and reprocessed) and transition results. "PASSEL Experiments" service allows users to search experimental and raw data using available filters for a single experiment [227]. The "PASSEL Data" service [228] also known as "Transition Group Browser" lets users identify the best performing assays from the entire database instead of a single experiment. Users can apply filters like protein, peptide, and various MS parameters to find SRM transitions and chromatograms from available experiments [229]. It also provides links to links to any available PeptideAtlas [198] and SRMAtlas [225] data.

1.4.1.4 MassIVE

MassIVE [200] is another node of ProteomeXchange that allows users to submit and access MS experimental data and associated files. Unlike other repositories, MassIVE provides a social or community science platform that lets users add new data, or reanalyze existing data and comment on new or renalysed datasets to enrich the original submission. The data submission is a two-step process; all registered users are assigned with an FTP account via the ProteoSAFe web interface [230]. It recommends users use their preferred FTP client to upload data. MassiVE dataset files are organised into ten categories, such as license file, spectrum files, etc. [231]. Hence users are recommended to name and organise datasets into various folders according to the main dataset categories during the upload process. Once the datasets are uploaded, users can log into the ProteoSAFe web interface [230] to link their files to specific groups and invoke the submission workflow for these files. Users are also required to submit necessary metadata for each dataset via the ProteoSAFe web page. It is recommended to provide raw spectrum and search engine files for all submissions however, a complete submission must include peak lists and result files. A detailed submission guideline along with the requirements for complete submission is available at the ProteoSAFe web page [231].

Once the data is submitted, users can either make it public or provide password protected access to other users. MassIVE ProteoSAFe [230] provides two discovery options i) browse [232] ii) search [233]. The browse page allows users to sort and filter data from a tabular web page, and search page provides a range of search criteria to find a dataset. Users can then invoke renalysis workflow for selected datasets. MassIVE onsite renalysis workflows use wide a range of bioinformatics tools [234] to allow users to reanalyse their own or other data. Users can then comment on or submit the reanalysis to enrich the original submission.

1.4.1.5 jPOSTrepo

jPOSTrepo [201] is the most recent member of ProteomeXchange consortium that enables users to share and reuse data generated from various proteomics project. It provides an easy-to-use flexible file management platform with high-speed file uploading mechanism. The jPOSTrepo submission workflow [235] requires users to create "preset" as the first step to document experimental procedures and wet lab protocols. Currently, the workflow offers four "presets." These are 'Sample', 'Fractionation', 'Enzyme/Modifications', and 'MS mode'. In the next step, users can create a 'profile' using a combination of the four presets and link it to their raw proteomics data. The same 'profile' can be attached to multiple files using submission portal. The platform supports raw MS files as well as the peak list files. Upon the meta-data profile linkage, users can upload the files from their computer using the web browser. The upload protocol uses a parallel transfer protocol to chunk individual files to maximise bandwidth over high latency connections. All complete submissions are subject to a validation process. After a successful validation, the workflow allows the user to "lock" the submission process to obtain a PX and jPOST identifier. The jPOSTRepo data discovery portal [236] lets users search public projects by ontology terms and controlled vocabulary (CVs). Users can also search a project by principal investigator names listed in presets and projects [201]. The discovery portal also offers a "quick" and "detail" view options. The quick view allows users to view high-level information on the same page, and users can also select the detailed view mode for more information.

1.4.2 Other Proteomics Databases

In this section, some of the proteomics databases will be discussed that are not part of the ProteomeXchange consortium but provide valuable proteomics experimental data to the community. Some of these databases reuse data from ProteomeXchange repositories and provide further information.

1.4.2.1 GPMDB

GPMDB [202] is a widely used database platform to analyse and validate tandem MS data using the X!Tandem [95] search engine. Users can go to the submission page [237] directly or navigate the page through the boutique proteomes option to select a species-specific database. Users can then upload a file to run X!Tandem [95] search using the Ensembl [215] database. Currently, the platform supports common, mzXML, mzData, DTA, PKL and MGF files. The identification information is stored in XML files, and these files are then indexed in a MySQL database [17]. Upon a successful search, users can opt in to submit the data to the GPMDB repository. The platform also allows both restricted and anonymous contributions. Additionally, the search portal also offers two other algorithms X!P3 [238] and X!Hunter [239] to analyse data. The X!P3 algorithm uses a list of a frequently detected proteotypic peptides for identification whereas the X!Hunter takes the experimental spectra and compares it with the consensus mass spectra from the GPMDB. The GPMDB data discovery portal also provides both a keyword and ontology-based search engine. Users can leverage a range data visualization tools (e.g. gene view, protein and observed peptide sequence view, and X!Tandem view) to view selected dataset as well as annotated spectra.

1.4.2.2 ProteomicsDB

ProteomicsDB [7] is an MS-based protein expression database of the draft human proteome. It stores protein and peptide identification and quantification information. It is underpinned by the SAP HANA platform and offers high-throughput in-memory data analysis and visualisation capabilities. In May 2017, the repository contained protein identification information covering 80% of the human proteome (15721 of 19629 proteins) from 78 projects comprising 418 experiments [240]. Although the platform used to offer various ways to upload data to the repository in the past, has discontinued the data repository service from 25 January 2017 and encouraging the users to submit or resubmit their data to one of the ProteomeXchange consortia. The ProteomicsDB team will continue to support the central part of the database. The site aims to link all published datasets to the ProteomeXchange dataset to maintain consistency. The data discovery portal allows users to browse human proteome and associated information (e.g. function, expression, identification). It also provides users an option to browse proteins by chromosome. Once a protein is selected, users are directed to the protein view. The protein view page offers a range of information tabs (e.g. summary, FDR estimation, proteotypicity, expression, biochemical assays, etc.). The expression tab of the protein view also allows users to visualise expressions across the complete human body

1.4.2.3 HPM

Like ProteomicsDB [7], HPM [13] is another database that was developed as a result of draft human proteome release. The database contains protein identification and expression information for 84% (17, 294 protein) of the protein coding genes of the near complete human genome. The information provided in the database was derived from 30 histologically normal human tissue samples from fetal tissues/adult tissues/hematopoietic cells using high-resolution MS (high-high mode). The MS/MS data was then searched against the human RefSeq database using MASCOT [81], and SEQUEST [93], search engines, and identification results stored in a MySQL database for public sharing [13]. The HPM database doesn't allow users to contribute their data. However, using the public data discovery portal users can search and visualise protein expressions and identification evidence based on protein, specific pathway, and gene family. Users can also restrict the search results to specific tissue or cell. Once a protein is selected, users can navigate to the peptide tab to view identified sequences, associated modification, mass-to-charge ratio, charge and visualise the best available high-resolution MS spectrum. The MS-based proteomics data is available via PRIDE, a member of the ProteomeXchange consortium and the PXD identifier for the dataset is PXD000561.

1.4.2.4 MaxQB

MaxQB [203] is a repository for high-resolution MS-based proteomics experiments. It stores protein and peptide identification and quantification information along with their spectra (high or low resolution) and allows joint analysis and comparison across project data. The project specific cutoff scores are adjusted to maintain the database-wide false discovery rate. Although only the Mann group contributes these datasets, the research community can access the public datasets using the data discovery portal. The identification workflow is tightly integrated with MaxQuant, and the private data submitters are prompted with an option to submit data at the end of the analysis. MaxQB also offers a manual private submission mechanism to the Mann group users. The MaxQB data discovery portal provides a basic and an advanced search function to query the database using search terms like protein or gene name, organism, etc. to view protein and peptide identification information. The portal also provides protein expression information within a proteome. It uses spectral count based iBAQ algorithm [241] to estimate protein quantification.

1.4.2.5 MOPED

MOPED [204, 205] is a multi-omics expression data repository from several model organisms including human. It provides protein-level expression, quantitative data from standard analysis and links genes, proteins, pathways, and external data sources. The platform is no longer maintained since 1 October 2015 and provided on an 'as is' basis to the community. Although the database is not updated, all datasets submitted to MOPED included a minimum set of standard metadata. The MOPED pipeline [242] renalysed MS data using public repository using the SPIRE [243] environment. MOPED discovery portal offers seven options to search and visualise and compare data. These are (i) protein absolute expression tab; to compare protein concentrations within and across various expression (ii and iii) relative gene/protein expression tab; to explore ratios of gene/protein concentration in comparative experiments iv) pathways tab; to show expression across experiments, conditions, tissues, and localizations v) experiment tab; to browse recently added experiments vi) disease tab; shows the gene-disease co-appearance in published article vii) visualisation tab; allows users drill down proteins by organism, tissue, localisation, and cell and display the absolute or relative matrix of identified proteins based on the condition.

1.4.2.6 PaxDb

PaxDb [206] is a comprehensive meta-resource containing whole genome protein abundance information across organisms and tissues. It aims to collect proteome-wide protein abundance information, irrespective of the underlying measurement technique. It uses information from publicly available repositories (mostly from PRIDE and PeptideAtlas) and maps them onto a common namespace and reprocesses the MS data using a standardized spectral counting pipeline. This pipeline uses the PeptideAtlas scoring and cutoffs to count the number of identification of peptides across the whole PeptideAtlas build. Each protein abundance data set is then remapped to a reference model organism genome/proteome sequence database that is imported from the STRING database [244]. For MS/MS data the pipeline remaps each peptide to the corresponding protein sequence from the database. The protein abundance values are converted for each dataset into protein abundance estimates using "parts per million" (ppm) value. The PaxDb discovery portal allows users to query single protein across all organism and multiple proteins for a single organism. For single protein search, the results page shows the abundance information from all available datasets, abundance (in ppm), and the ranking. The abundance information can be filtered further to narrow down the results. Users can also view the abundances in other organisms from this page as well the interaction network provided by STRING. For multiple protein search, related datasets and abundances are listed in a tabular form. Users can download the results as a tab-delimited file or click on individual proteins to navigate to the single protein view. The download data page allows users to download all datasets or per-species abundance file in a compressed format.

1.4.2.7 Human ProteinPedia

Human Proteinpedia [207] is a public web platform for sharing and integration of human protein expression and annotation information from multiple experimental platforms. It derives data from various experimental platforms, which include co-immunoprecipitation and mass spectrometry-based protein-protein interaction or western blotting based protein-protein interaction, fluorescence based experiments, immunohistochemistry, MS analysis data, protein and peptide microarray, western blotting and yeast two-hybrid-based protein-protein interaction. Besides experimental platforms, its annotation platform covers a diverse range annotation features. These include post-translational modifications, subcellular localization, protein-protein interactions, enzyme substrates, and tissue, cell line, and disease tissue expression. Registered users can log into the submission portal to submit annotation for any of the above mention features. Annotations must be submitted using the defined format [245]. Users can also leverage the 'batch upload' feature to provide large datasets. Unlike other proteomics databases, Human Proteinpedia doesn't mandate users to submit all raw data. However, it doesn't accept any *in silico* annotation and recommends users to provide raw data where possible. The database enriches human protein reference data (HPRD) [246] using these community annotations. The data discovery portal allows users to search and access data using three different ways. These are i) using gene symbol, protein name or protein accession numbers ii) by datasets and iii) by experimental platform. It also provides a data download page to browse and download any public dataset.

1.4.2.8 UniProt

UniProt is one of the most comprehensive and widely used protein sequence and annotation databases. UniProt provides four different databases to addresses different scientific use cases. These resources are developed and maintained by the UniProt Consortium. The consortium was formed in 2002 to combine the databases, resources, and expertise of its member institutes - the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). Swiss-Prot [247-249] and TrEMBL [247] were developed and maintained by EMBL-EBI and SIB jointly. The PIR managed the PIR-PSD and related databases [143, 144, 250].

The UniProt Knowledgebase (UniProtKB) is the central repository that provides high-quality sequences for many species. It has two components - i) UniProtKB/Swiss-Prot contains non-redundant protein sequence data that are manually annotated by the UniProt curators. The information is curated by various computational analysis as well as extracting information from scientific literature. (ii) UniProtKB/TrEMBL contains protein sequences that are computationally analysed and subject to automatic classification and annotation. These records are manually annotated and integrated to UniProtKB/Swiss based on 'defined annotation priorities' [247]. The UniProt Archive (UniPrac) acts as a non-redundant unannotated archive of publicly available protein sequence databases. UniParc merges all sequences with 100% identity match over their entire length from various databases regardless of their species. It cross-references each sequence to their source including all versions of each sequence to avoid redundancy during the data retrieval [251]. The UniProt Reference Clusters (UniRef) provides clustered set of UniProtKB sequences, selected UniPrac records as well as Ensembl [215] protein translations from selected species. It contains three databases - UniRef100, UniRef90 and UniRef50, and are produced by clustering various sequence records by percentage of sequence identity [252]. The UniProt Metagenomic and Environmental Sequences DB (UniMes) includes predicted metagenomic sequences derived from and environmental samples. The dataset is further enriched by InterPro annotation [162] to show protein families, domains and functional sites [253].

The UniProt resources also use information from previously mentioned publicly available proteomics databases to enrich the annotation. The UniProtKB protein existence [254] annotation uses protein/peptide identification information from publicly available databases and literature. It captures post-translational modifications (PTMs) from these databases. The UniProtKB also provides a direct link to some of the databases in the protein annotation page such as PaxDb, PRIDE, PeptideAtlas, MaxQB, iPTMnet [255] and PhosphoSitePlus [256].

1.4.2.9 neXtProt

neXtProt [27] is a web-based protein knowledgebase that aims to provide updated information on human proteins. It takes the annotated information from UniProtKB and compliments it further by offering manual curation, developing tools for various research use cases, and providing use case centric views of the data. Users can download the annotations from the neXtProt FTP site as XML or PSI Extended Fasta Format (PEFF) file. It also sources or integrates data from a wide range of other databases including previously mentioned PeptideAtlas and SRMAtlas to capture various information such as protein expression, micro-array, and cDNA, subcellular localization, information from MS experiments data, Gene Ontology (GO), etc. More recently it started to capture information on the phenotypic effect of genetic variations from the scientific literature and provides a 'phenotypic view' to interrogate any possible phenotypic impact of genetic variation [27]. It also provides a web tool 'unicity checker' [257] to identify proteotypic peptide from a protein sequence. Like UniProtKB, it also provides protein existence level (PE) annotation and considered to be the primary repository of the C-HPP initiative. The neXtProt PE level 2 to 4 are classified as the 'missing proteins.' The platform also provides a periodic release of the PE status to facilitate the C-HPP project [25, 26, 28].

1.4.2.10 GeneCardsSuites

The GeneCardsSuites [258] provides a suite of databases that include gene-centric annotations and information that are automatically mined from 150 databases[259]. Among them, the GeneCards [258] is a wiki-like service that provides annotation information from these data sources in a single web page per gene (also known as cards) or provides a link the resource if the data is not captured. Users can use the GeneALaCart the web-based service generate an annotation file for a list of genes. The PathCards [260] contains integrated information of human pathways. The MalaCards [261] is a database of the human maladies, and the GeneLoc [262] uses the computation algorithm to provide an integrated map of human chromosomes. The GeneAnalytics tool [263] can analyse data from GeneCards and MalaCards to provide an interactive view of highly enriched annotations such as gene associations to disorders, pathways, GO terms, expression, and compounds.

1.4.2.11 PubMed

PubMed [264] is an online service that provides free access to bibliographic information, abstract, and some full articles (via PubMed Central) from MEDLINE, PreMEDLINE, and various life science journals. It also provides publisher supplied citations, and all indexed or archived journals are subject to a quality assurance check. It also contains links to full texts. PubMed also provides programmatic access to its database. PubMed can be queried using Medical Subject Headings (MeSH) terms as well as keywords such as title, author name, journal name, etc.

In addition to the standard search, PubMed provides purpose-built clinical query service [265] to search literature related to medical genetics, disease, and treatment, etc. and its LinkOut service [266] gives access to biological databases, research tools, etc.

Table 1.4. List of proteomics databases

Database	Category[1]	URL	Accept Public data[2]
ProteomeXchange	Raw and processed data	http://www.proteomexchange.org	Yes
PRIDE	Raw and processed data	https://www.ebi.ac.uk/pride	Yes
MassIVE	Raw data	https://massive.ucsd.edu	Yes
jPOST	Raw and processed data	https://repository.jpostdb.org	Yes
PASSEL	Raw and processed data	http://www.peptideatlas.org/passel	Yes
PeptideAtlas	(Re)Processed data	http://www.peptideatlas.org	Yes
GPMDB	Processed data	http://gpmdb.thegpm.org	Yes
ProteomicsDB	Processed data	https://www.proteomicsdb.org	No
HPM	Processed data	http://www.humanproteomemap.org	No
MaxQB	Processed data	http://maxqb.biochem.mpg.de/mxdb	No
MOPED	Processed data	https://www.proteinspire.org/	No
PaxDb	(Re)Processed data	http://pax-db.org	No
Human ProteinPedia	Processed data	http://www.humanproteinpedia.org/	Yes
UniProt	Knowledgebase	http://www.uniprot.org	Yes
neXtProt	Knowledgebase	https://www.nextprot.org/	No
GeneCards	Knowledgebase	http://www.genecards.org	No
PubMed	Knowledgebase	https://www.ncbi.nlm.nih.gov/pubmed/	No
PDB	Knowledgebase	https://www.wwpdb.org/	Yes

[1] Category – Raw data = raw proteomics data, process data = peptide identification and quantification data, knowledgebase = scientific knowledges or research outputs

[2] Does the repository allow end users to upload or contribute data directly?

1.4.2.12 Protein Data Bank (PDB)

PDB is the most widely used and largest global dataset for experimentally derived three-dimensional (3D) structures of proteins, DNA and RNA molecules [145, 267, 268]. The dataset provides community supplied annotated and validated primary, secondary and tertiary structural data obtained by X-ray crystallography or Nuclear magnetic resonance (NMR) spectroscopy. The platform is managed by the Worldwide Protein Data Bank Organization [267] with three regional data centers or nodes. These are RCSB Protein Data Bank (RCSB PDB) [145, 146], Protein Data Bank Japan or (PDBj) [269] and Protein Data Bank in Europe (PDBe) [270]. Among them, the RCSB PDB provides a broad range of analysis, visualisation and validation tools for research and education purposees. The 'NGL [271] viewer' enables users to view the structure from the web interface without installing any software. The 'Structural View of Biology' integrates PDB data with other primary and community-derived data resources to decipher the biological processes and mechanisms. It also provides a mapping of the Protein Modification (PMS) to the RESID database [272]. The PDB resources have been used to create many other databases [273]. The PDB sequences are useful resources for database similarity searches for *in silico* annotation, as proteins can be considered homologous with structural identity match 25% or more [33, 274].

1.4.3 Database challenges

In the previous section, I have highlighted a large number proteomics centric databases and their contribution to the community. The growth of proteomics data is on the rise, and the community is more openly sharing data [275]. The ProteomeXchange initiative has significantly eased the raw data storage problem. However, the opportunity to make the best use of these datasets are not free from challenges. Although all repositories provide their own data discovery portal, finding all datasets related to a topic from various databases is far from trivial. Each of these databases has their own implementation platform, controlled vocabularies, and they are purpose built for specific types of data. Proteomics experiments are more complex as they come from various instruments and experimental conditions, data format and quality metrics. So, finding all datasets related to a topic of interest is not enough, one must know how to integrate or harmonise data from heterogeneous sources and platforms. For example, a user may be interested in the search for all information about some human 'missing' proteins.

To get the most out the vast public knowledge, the user needs first to identify all the databases that have some information (including how to search them in each database). Then the user needs to understand the quality metrics for each of the datasets to filter and extract all information in human/machine readable formats, which often requires multi-domain skills as well as storage and compute resources, and finally, the user needs to run further analysis to combine data from various sources (more compute, and complex multi-domain skills). This is even more complex for multi-omics research. For example, data from a proteogenomics study may have been partially submitted to a proteomics database (proteomics data) and genomics database (for genomic data and findings). So, linking information from various sources and the ability to integrate and interpret data from diverse biological domains pose significant challenges. Although some of the database platforms support basic data sharing, they do not provide enhanced collaboration facilities between research groups to conduct complex studies.

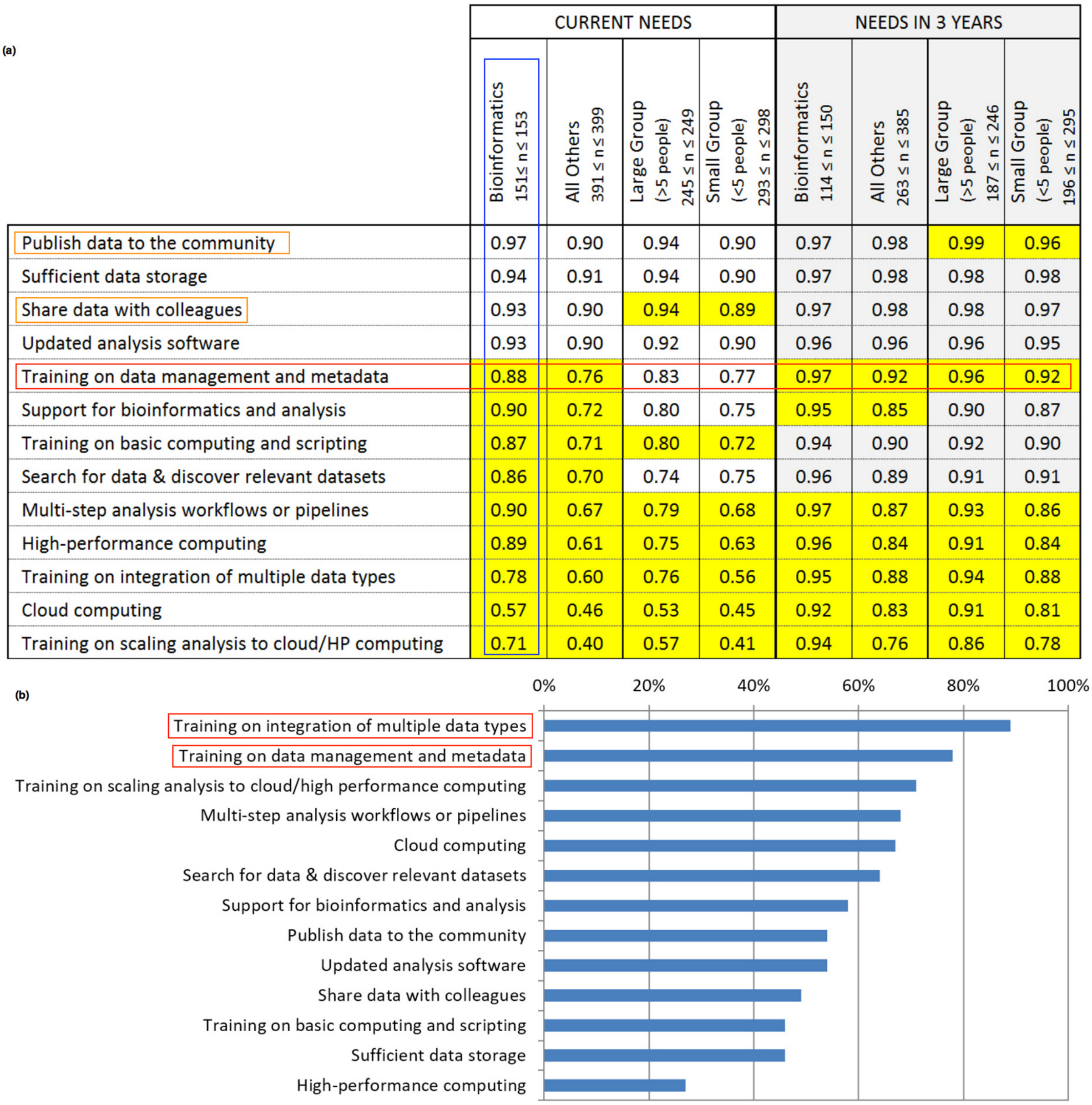
Several databases such as UniProt, neXtProt and GeneCards act as knowledgebases (instead of raw data provider), and ProteomeXchange offers some common standards for storing and sharing data, the generalised nature of these platforms often does not meet the needs of domain-specific research. On the other hand, platforms like Chorus and Firmiana offer data storage, sharing, and analysis facility under a single platform. However, they work in silos and do not provide integration with the public knowledgebase. Hence, data discovery and integration of data from various sources continue to challenge the scientific community (including the proteome bioinformatics community).

The National Science Foundation (NSF) surveyed 704 Biological Sciences Directorate principle investigators (BIO PIs) in 2016 to identify the current and future priorities to enable big data research. The survey report (preprint version - February 2017) [276] showed over 95% PIs are either sharing data with the community or will continue to share data in future. 90% of them indicated that they are currently or will be analysing big data. However, 89% of them reported that they need training on integration of multiple data types and 78% indicated the need for training in data management and metadata (see Figure 1.7 (a) and 1.7 (b)). All survey participants are considered "competitive researchers" in their field as each of them has secured at least one peer-reviewed grant. Hence, the outcome of the study further highlights the need for integrating data from heterogeneous experiments and computing platform.

A recent article published around the same time by Martens et al. discussed some data use and reuse examples along with potential opportunities and challenges. The article called this the 'Golden Age' to work publicly shared proteomics resources. However, it concurs that integrating data generated from the multi-omics study is a challenging task and so far, have been achieved by two large consortia with their own data repository (The Clinical Proteomic Tumor Analysis Consortium (CPTAC) [277, 278]) or organism specific resources (*'Saccharomyces'* Genome database [279]) [275].

Omics Discovery Index (OmicsDI) [280] has been developed recently to index multi-omics datasets. It links and indexes 11 data repositories with host repository supplied shared metadata. OmicsDI offers three types of metadata submission options- (i) mandatory, (ii) recommended, and (iii) additional fields. The mandatory information contains basic collection administration level information. The recommended fields aim to collect basic sample protocol, omics type, tissue and instrument details. The additional fields aim to capture PTM, chromatographic protocol, quantification method, taxonomy and protein/metabolite identifier. While keeping the mandatory data simple makes is easier to link the database, the OmicsDI system will potentially index large number of datasets without the protein identifier (and other key information). While it is an excellent starting point to identify datasets from various sources, researchers may not be able to find all datasets by searching them with their protein identifier (even if the datasets are already indexed) and will eventually have to go to the source repository to (a) perform another search to ensure a comprehensive coverage of the dataset (b) collect additional information (e.g. search engine, identification etc.).

One solution is to create new platforms or extend some of the current platforms to integrate data from various sources in a reusable format. As the community matures, it will drive new metadata standards, and science will demand new types data or metadata for new discoveries. Therefore, the platform needs to be able to adapt to community needs (not the opposite). At the same time, the focus should be complementing any existing efforts rather than duplicating resources or data, and increase collaboration where possible (especially for targeted studies).



(a)

	CURRENT NEEDS				NEEDS IN 3 YEARS			
	Bioinformatics 151≤ n≤ 153	All Others 391 ≤ n≤ 399	Large Group (>5 people) 245 ≤ n ≤ 249	Small Group (<5 people) 293 ≤ n ≤ 298	Bioinformatics 114 ≤ n ≤ 150	All Others 263 ≤ n ≤ 385	Large Group (>5 people) 187 ≤ n ≤ 246	Small Group (<5 people) 196 ≤ n ≤ 295
Publish data to the community	0.97	0.90	0.94	0.90	0.97	0.98	0.99	0.96
Sufficient data storage	0.94	0.91	0.94	0.90	0.97	0.98	0.98	0.98
Share data with colleagues	0.93	0.90	0.94	0.89	0.97	0.98	0.98	0.97
Updated analysis software	0.93	0.90	0.92	0.90	0.96	0.96	0.96	0.95
Training on data management and metadata	0.88	0.76	0.83	0.77	0.97	0.92	0.96	0.92
Support for bioinformatics and analysis	0.90	0.72	0.80	0.75	0.95	0.85	0.90	0.87
Training on basic computing and scripting	0.87	0.71	0.80	0.72	0.94	0.90	0.92	0.90
Search for data & discover relevant datasets	0.86	0.70	0.74	0.75	0.96	0.89	0.91	0.91
Multi-step analysis workflows or pipelines	0.90	0.67	0.79	0.68	0.97	0.87	0.93	0.86
High-performance computing	0.89	0.61	0.75	0.63	0.96	0.84	0.91	0.84
Training on integration of multiple data types	0.78	0.60	0.76	0.56	0.95	0.88	0.94	0.88
Cloud computing	0.57	0.46	0.53	0.45	0.92	0.83	0.91	0.81
Training on scaling analysis to cloud/HP computing	0.71	0.40	0.57	0.41	0.94	0.76	0.86	0.78

(b)
0%
20%
40%
60%
80%
100%
Training on integration of multiple data types
Training on data management and metadata
Training on scaling analysis to cloud/high performance computing
Multi-step analysis workflows or pipelines
Cloud computing
Search for data & discover relevant datasets
Support for bioinformatics and analysis
Publish data to the community
Updated analysis software
Share data with colleagues
Training on basic computing and scripting
Sufficient data storage
High-performance computing

Figure 1.7 (a) Percentage of affirmative responses, current and future data needs of BIO PIs. Yellow highlighting indicates a statistically significant chi square result between groups (bioinformaticians versus others; large research groups versus small)**. (b) Unmet data needs of BIO PIs.** Percent responding negatively (318 ≤ n ≤510)**.** Source *Barone et al., 2017*

1.5 Aims and objectives

Identifications of proteins and decoding their underlying biological functions are the key to understanding life within the cell, tissue or organisms. Bioinformatics techniques can be used to find putative functions of proteins using publicly available databases. In-house experimental or publicly available proteomics data can be used to supplement the annotation. Mass Spectrometry techniques have been widely used for proteomics studies. Many free and commercially available search engine software and algorithms is available to analyse data. However, the interpretation and quality assessment of MS search engine results and mass spectra remain a challenge. Some guidelines are required for non-specialist users to assist them in interpreting such data. While there is a rapid growth in the number proteomics databases (including the volume of data) accessible to the community, the lack of integration between these heterogeneous datasets creates further challenges to analyse and interpret information to study any proteome comprehensively. An overarching platform can be developed to i) integrate data or metadata from various sources ii) provide a collaborative platform to share data and knowledge between research groups to analyse complex data or facilitate large studies. Specific aims are listed below, with **6** publications presented in this thesis:

Aim 1

Develop a generic *in silico* bioinformatics pipeline to identify homologues and map putative functional signatures, gene ontology terms and biochemical pathways of relatively less studied or novel organisms, or “missing’’ proteins using existing resources and share the annotations with the scientific community.

Objectives:

1. Develop a generic *in silico* bioinformatics pipeline to functionally annotate the relatively less studied black Périgord truffle proteome from the 2010 *T. melanosporum* genome comprising 12,771 putative non-redundant proteins. Conduct a shotgun proteomics study (using a combined 1D PAGE and high accuracy LC-MS/MS) to validate and supplement the annotation (Publication 1).
2. Demonstrate the application of the bioinformatics pipeline to annotate the human “missing” protein sequences for each human chromosome and develop a web portal to share the annotations with the scientific community (Publication 2).
3. Develop a fully automated, robust and generic functional annotation platform to annotate any given proteome (Publication 3).

Aim 2

Develop protocols for functional annotations using existing knowledgebase and guidelines to complement the annotations using publicly available MS datasets derived from various MS instruments and search engines.

Objective

4. Develop *in silico* function annotation protocol, and guidelines for comparing proteomics data from different sources, search engine scores, identifying proteotypicity as well as guidelines on spectral quality analysis (Publication 4).

Aim 3

Develop an integrated data capture, deposition and sharing and collaboration platform for a domain specific research need using existing and community contributed datasets to fast track the research process.

Objectives

5. Apply the guidelines from objective #2 and extend the previously developed web portal to provide a single platform to automatically capture and integrate available information about human ‘missing’ proteins from various databases, including a data deposition and secure sharing interface to capture unpublished, draft or laboratory data via citizen science contribution (Publications 5, 6).

Chapter 2: Methods and applications

Methods and applications that were developed and used in this study are summarised in Table 2.1. The related publications have also been listed and included in the relevant chapter.

Table 2.1: Methods, applications and publications

Methods/Applications	**Chapter**	**Thesis Publication(s)**
Unlocking the Puzzling Biology of the Black Périgord Truffle	3	1
Protannotator: A Semiautomated Pipeline for Chromosome-Wise Functional Annotation of the "Missing" Human Proteome	4	2
ProtAnnotator 2.0: An automated pipeline for *in silico* protein functional annotation	5	3
A systematic bioinformatics approach to identify high quality MS data and functionally annotate proteins and proteomes	6	4
MissingProteinPedia for accelerating the search for the missing proteins in the human proteome	7	5,6

Chapter 3: Unlocking the Puzzling Biology of the Black Périgord Truffle

3.1 Summary

Proteins are responsible for almost every activity and processes within the cell, hence understanding the functions of protein is the key to uncovering the biology of any organism. The advances in next generation sequencing technology and the reduced cost of the genome sequencing are the catalysts for the rapid growth in genome sequences. Protein sequences are mostly predicted, from genome annotation phase. Only 1% of these proteins have experimental functional annotations [281]. The databases are growing at a much faster rate than our biological understanding of these sequenced organisms. Hence, there is a clear need for accurate and high-throughput *in silico* annotation methods for functional annotation of proteins, for novel and less studied proteins.

In this chapter, I carried out functional annotations of publicly available black Périgord truffle (*Tuber melanosporum Vittad*) proteome sequences [34], a highly prized and relatively less studied organism. Only 14 out of 12,771 *T. melanosporum* proteins sequences had been reviewed and manually annotated in UniProt, with no experimental evidence. I developed an *in silico* functional annotation technique using proteins sequences from existing knowledge base (UniProt) to identify homologous, and functionally annotate proteins based on GO, pathways, and protein domain mapping. Details of the annotation pipeline, tools, and proteomics validation are described in publication 1.

3.2 Publication 1

Reprinted with permission from: Mohammad Tawhidul Islam, Abidali Mohamedali, Gagan Garg, et al (2013) Unlocking the Puzzling Biology of the Black Périgord Truffle Tuber melanosporum. *Journal of Proteome Research*. vol. 12, no. 12, pp. 5349-5356.

Journal of proteome research

Article

pubs.acs.org/jpr

Unlocking the Puzzling Biology of the Black Périgord Truffle *Tuber melanosporum*

Mohammad Tawhidul Islam,[†,‡,#] Abidali Mohamedali,[†,#] Gagan Garg,[†,‡,#] Javed Mohammed Khan,[†,‡] Alain-Dominique Gorse,[§] Jeremy Parsons,[§] Peter Marshall,[∥] Shoba Ranganathan,[*,†,‡,⊥] and Mark S. Baker[*,†]

[†]Department of Chemistry and Biomolecular Sciences, Macquarie University, NSW 2109, Australia

[‡]ARC Centre of Excellence in Bioinformatics, Macquarie University, NSW 2109, Australia

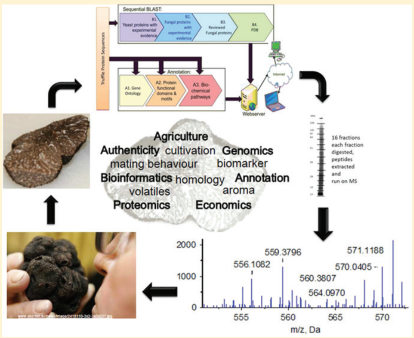
[§]QFAB, The University of Queensland, Queensland Bioscience Precinct, QLD 4072, Australia

[∥]Terra Preta Truffles, 389 Sawyers Ridge Road, Reidsdale, NSW 2622, Australia

[⊥]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 117597 Singapore

S *Supporting Information*

ABSTRACT: The black Périgord truffle (*Tuber melanosporum* Vittad.) is a highly prized food today, with its unique scent (i.e., perfume) and texture. Despite these attributes, it remains relatively poorly studied, lacking "omics" information to characterize its biology and biochemistry, especially changes associated with freshness and the proteins/metabolites responsible for its organoleptic properties. In this study, we have functionally annotated the truffle proteome from the 2010 *T. melanosporum* genome comprising 12 771 putative nonredundant proteins. Using sequential BLAST search strategies, we identified homologues for 2587 proteins with 2486 (96.0%) fungal homologues (available from http://biolinfo.org/protannotator/blacktruffle.php). A combined 1D PAGE and high-accuracy LC−MS/MS proteomic study was employed to validate the results of the functional annotation and identified 836 (6.5%) proteins, of which 47.5% (i.e., 397) were present in our bioinformatics studies. Our study, functionally annotating 6487 black Périgord truffle proteins and confirming 836 by proteomic experiments, is by far the most comprehensive study to date contributing significantly to the scientific community. This study has resulted in the functional characterization of novel proteins to increase our biological understanding of this organism and to uncover potential biomarkers of authenticity, freshness, and perfume maturation.


Sequential BLAST:
B1. Yeast proteins with experimental evidence
B2. Fungal proteins with experimental evidence
B3. Reviewed Fungal proteins
B4. PDB
Truffle Protein Sequences
Annotation:
A1. Gene Ontology
A2. Protein functional domains & motifs
A3. Bio-chemical pathways
Webserver
Internet
Agriculture
Authenticity
cultivation
Genomics
mating behaviour
biomarker
Bioinformatics
homology
Annotation
volatiles
aroma
Proteomics
Economics
16 fractions each fraction digested, peptides extracted and run on MS
2000
1000
0
559.3796
556.1082
571.1188
570.0405
560.3807
564.0970
555
560
565
570
m/z, Da

KEYWORDS: *black truffle, organoleptic, proteome, functional annotation, fungal proteomics*

INTRODUCTION

Truffles are fungi that produce subterranean fruiting bodies through the establishment of an ectomycorrhizal symbiotic relationship with the roots of host plants,[1] usually in a mutualistic fashion utilizing animals in their lifecycle to distribute spores. Among the different indigenous truffle species described, many have very pronounced organoleptic properties that are capable of attracting animals (including man) to the fruiting body. Collectively, these organoleptic properties also have accorded some truffle species their high economic importance.[2] The fruiting body of the black Périgord truffle (*Tuber melanosporum* Vittad) is one of the most prized delicacies in any gourmet food repertoire as evidenced by the exorbitant prices they fetch in world markets (≥$2,000 USD/kg).[3] This rare 'black diamond' of the kitchen has long intrigued distinguished chefs and biologists alike, due to its combination of smooth texture, pungent odor/perfume, and musty earthy flavor. In addition, its unique and often cryptic symbiotic relationship with oak and hazelnut trees has thwarted numerous efforts at routine cultivation.[4] In the past decade, the harvest of the black Périgord truffle has plummeted in Europe due to the effects of climate change, and loss of suitable arable land, encroachment of introduced species, and other factors.[5] This scarcity, coupled to increased awareness and demand for truffles has led to increasing prices and hence the inevitable replacement by similar black truffles (e.g., the Chinese black truffle *Tuber himalayensis*[6] or members of the *Tuber indicum* group[7]) into significant markets. For these reasons, a more focused study on the ecology, biology, and behavior of the black Périgord truffle has become increasingly necessary. To this end, the 2010 publication of the genome[4] of the black Périgord truffle has resulted in an understanding, on the genomic level, of its carbohydrate metabolism,[8] transcription,[9] mating behavior,[10] volatiles that produce aroma,[11] and other aspects of transcriptional and genomic control. Despite this,

Received: June 30, 2013
Published: October 22, 2013

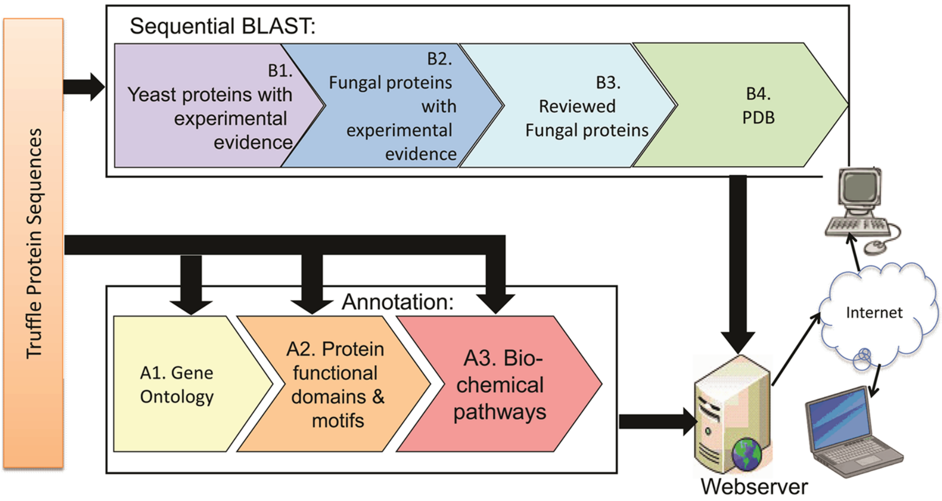

Truffle Protein Sequences
Sequential BLAST:
B1. Yeast proteins with experimental evidence
B2. Fungal proteins with experimental evidence
B3. Reviewed Fungal proteins
B4. PDB
Annotation:
A1. Gene Ontology
A2. Protein functional domains & motifs
A3. Bio-chemical pathways
Internet
Webserver

Figure 1. Summary of the pipeline used to annotate the truffle proteome. Proteins were passed through a series of databases to determine homology (sequential BLAST) as well as databases to confer annotations based on gene ontology, protein functional domains, and motifs as well as biochemical pathways.

only 14 proteins have "reviewed" annotations in UniProt[12] (v 2013_05): "reviewed" data sets contain proteins entries that are all manually annotated and reviewed in UniProtKB/Swiss-Prot database. Clearly, the full utility of the published genome in terms of the proteome remains to be realized.

To have a comprehensive understanding of the biology of this organism, it is imperative that a systems biology approach (bioinformatics and proteomics) is employed. Bioinformatics can provide compressive annotations for entire proteomes, providing valuable information regarding putative functions of proteins. Fungal proteomics over the past years has led to numerous advancements in biological understanding of the unique behaviors of filamentous fungi that have resulted in commercial gain,[13] control of pathogenic fungi,[14] and the discovery of biomarkers of freshness and authenticity[15] as well as ecology.[4] The differences between the unique biochemistries of different economically important truffle species (e.g., white truffle *Tuber magnatum* Pico, summer truffle *Tuber aestivum*, and winter truffle *Tuber brumale*) are yet to be determined comprehensively. A recent study has attempted to utilize proteomics to decipher differences in the growing conditions of the *Tuber* family with the successful identification of only 17 proteins from *T. magnatum* matched to the *T. melanosporum* gene database.[16] It is beyond doubt that the proteome of the fruiting body of the black Périgord truffle would yield a significant and useful data set for understanding this and related organisms with bioinformatics studies accelerating and guiding the complex proteomics studies.

In this study, we applied a "sequential BLAST" method previously adopted to functionally annotate the "missing" proteins of Human Chromosome 7.[17] In this approach, data similarity searching was carried out sequentially against carefully selected reference databases instead of repeated searches against the same database, as implemented in PSI-BLAST, a variant of the standard BLAST search engine,[18] to overcome the limitation of matches to predominantly "hypothetical" proteins (11 870), as reported in the Supporting Information (TuberGM_annot.xls) of the *T. melanosporum* genome.[4] This study is a combined bioinformatics and proteomics study to characterize the truffle proteome. Putative biological functions of the truffle proteome are assigned primarily by identifying homologues from well-annotated experimentally validated yeast and fungal proteins. We have also used bioinformatics analyses to ascribe functional annotations in terms of protein domains, gene ontology, and biochemical pathways. We have then attempted to validate the proteome using shotgun proteomic analysis of the fruiting body, which has provided a list of potential proteins involved in the production of the truffles' aroma profile.

■ MATERIALS AND METHODS

1. Bioinformatics Analysis

Data Sources. Black Périgord truffle (*T. melanosporum* Vittad.) protein sequences were extracted from the MycorWeb database [http://mycor.nancy.inra.fr/IMGC/TuberGenome/download.php?select=fast][4] in FASTA (a special file format generated by the 'FAST-All' software package) format.[4] Of the 12 826 coding DNA sequences (CDS, predicted Genoscope gene models) obtained, the truffle proteome comprised 12 771 nonredundant proteins after removing duplicate entries, although the *T. melanosporum* genome publication[4] reported only 7496 as "protein coding genes." Reviewed proteins with protein level experimental evidence are the best source of functional information, followed by reviewed proteins. Since yeast (*Saccharomyces cerevisiae*) is the most studied fungal species, our first choice for seeking truffle homologues was the set of reviewed yeast proteins with experimental evidence, followed by fungal proteins with experimental evidence and then reviewed fungal proteins. To carry out the functional characterization of the *T. melanosporum* proteins, we downloaded and set up local databases for BLAST[18] similarity searches. These databases included yeast proteins with experimental evidence (7503 sequences), fungal proteins with experimental evidence (9450 sequences), and reviewed fungal proteins (31 031 sequences). In addition, another search was conducted against all Protein Data Bank (PDB) proteins (236 604 sequences) to assign homologues from proteins with structure, as 3D structures are known to be evolutionarily conserved even under very low amino acid sequence similarity.[17] Protein data sets for the sequential BLAST searches were extracted from the UniProt/SwissProt database (v 2013_05, release 01-05-2013).[12]

Database Similarity Searches. Database similarity searches for the truffle proteome were conducted using BLASTP.[18] A match was deemed a strong indicator of homology if the query sequence matched a database sequence with high significance (i.e., very low E value: $< 1 \times 10^{-5}$) and sequence identity of at least 50%. Sequential BLASTP runs were performed against the four data sets (including PDB) described above (using default parameters) for mapping a known protein sequence against a database of protein sequences. Sequences that did not have any match from the first run were passed to the next round of BLASTP to search the second data set, then the third and the fourth. As reviewed sequences with protein level experimental evidence are considered the most reliable source of homologues, these were used as the first database for BLASTP similarity searches. These proteins were also subjected to further *in silico* analyses, as described later.

Functional Annotation. Protein functional annotation in terms of protein domains, motifs, and signatures provides vital clues to biological function for experimental validation. InterProScan[19] comprises 14 programs for matching a query sequence against 13 protein domain and functional site databases and represents the most comprehensive protein functional annotation software currently available. All black Périgord truffle proteins were initially characterized through InterProScan[19] domain/motif analyses. InterProScan also provides gene ontology (GO) annotations. Pathway mapping for all of these proteins was carried out using KOBAS (KEGG Orthology-Based Annotation System, KOBAS-2.0).[20] All results from domain/motif analyses, GO annotation, and KEGG pathway mapping were used for preliminary functional annotation of these proteins.

The different bioinformatics analysis steps used for annotating the truffle proteome were integrated into a pipeline, illustrated in Figure 1.

Proteomics Studies. Proteomic studies were carried out to validate our bioinformatics approach. Freshly harvested Australian black Périgord truffles (*T. melanosporum* Vittad.) were kindly donated by Terra Preta Truffles (NSW, Australia). Truffles were stored on ice overnight for transport to the laboratory, and the best sample was selected as “representative” of the mature fruiting body. Approximately 50 mg of the inner tissue of the selected sample was freeze-crushed in liquid N_2, and the resulting powder was dissolved in 1 mL of 4× LDS buffer in the presence of both protease and phosphatase inhibitors. The sample was probe-sonicated (3 × 10 pulses, output 3) (Branson sonifier 450) until the solution was homogeneous and was centrifuged at 10 000*g* for 10 min to remove insoluble particulate matter, acetone-precipitated overnight at −80 °C, and centrifuged for 20 min at 10 000*g* . The pellet was then resuspended in 4× LDS buffer and protein-quantified using a BCA assay (Thermo-Pierce, Rockford, IL) according to the manufacturer’s instructions.

1D Gel Electrophoresis and Slice-and-Dice Proteomics. Resuspended protein (100 μg) was run on precast 4−12% linear gradient SDS polyacrylamide gel (Invitrogen, USA) under reducing conditions as per the manufacturer’s instructions. The gel was then fixed in 40% ethanol (v/v), 10% acetic acid (v/v) for 2 h and stained overnight with Flamingo Pink (BioRad, Hercules, USA) and imaged on the Typhoon Trio Variable Mode Laser Imager (GE Healthcare, Uppsala, Sweden) with photomultiplier tube (PMT) voltage set to 5 V below saturation of the most intense spot. The entire gel lane was divided into 16 fractions, digested using trypsin, and extracted using standard procedures described elsewhere.[21]

LC Coupled to Mass Spectrometry. The digested peptides (10 μL) were injected onto a peptide chromatography trap (Michrome peptide Captrap) on an Eksigent Ultranano LC system for preconcentration. Desalting is a standard procedure prior to LC as salt usually produces aberrant peaks on the spectrum, depending on the conductance values of the solution. Peptides were desalted using 0.1% formic acid, 2% ACN, at 5 μL/min for 10 min. The peptide trap was then switched online with an analytical column (SGE ProteCol C18, 300 Å, 3 μm, 150 μm × 10 cm). Peptides were eluted from the column using a linear solvent gradient, consisting of 0.1% formic acid as mobile phase A and 90% ACN/0.1% formic acid as mobile phase B, at 600 nL/min, starting from 2% B and going to 40% B over 140 min. After peptide elution, the column was cleaned with 80% buffer B for 19 min and then equilibrated with buffer A for 15 min before the next sample was injected. The reverse-phase nanoLC eluent was subject to positive ion nanoflow electrospray analyses in an information-dependent acquisition mode (IDA) on a Triple TOF 5600 (ABSciex, Toronto, Canada) at 15 kV acceleration voltage. MS data were collected using an ion spray voltage of 2.4 kV, curtain gas of 20 PSI, nebulizer gas of 15 PSI, and an interface heater temperature of 150 °C. In IDA mode, a TOFMS survey scan was acquired (m/z 350−1500, 0.25 s), with the 15 most intense multiply charged ions (counts >150) in the survey scan sequentially subjected to MS/MS analysis. MS/MS spectra were accumulated for 50 ms in the mass range m/z 100−1500 with the total cycle time 1.05 s, with a mass accuracy 1 ppm.

Database Searching of Proteomic Data. The experimental nanoLC−ESI−MS/MS data were submitted to Mascot after raw files were converted to .mgf format and searched against the *T. melanosporum* database (called here BT_Prot) which was derived from the 12 771 nonredundant sequences of the truffle proteome. The 16 fractions were processed individually with output files for each fraction, then merged, and a nonredundant output file was generated for protein identifications with $\log_e$ scores < 1. Search parameters included MS and MS/MS tolerance of ±100 ppm and ±0.2 Da, respectively. Carbamidomethyl was considered a fixed modification. In addition, variable modifications of methionine, threonine, and deamidation of asparagine and glutamine were also considered. Additional searching was performed against the decoy database in Mascot to evaluate false discovery rates (FDRs). Peptide FDR of a list is 2 × total numbers of peptides representing reversed protein hits in the list/total number of peptides representing all proteins in the list × 100. Protein FDR was calculated for each list of proteins using number of reversed protein hits in the list/total number of proteins in the list × 100.[22]

■ RESULTS AND DISCUSSION

Bioinformatics Analysis

The “sequential BLAST” approach we used involved repeated similarity searching against different select databases (Figure 1). Reviewed sequences (from UniProt) with protein level experimental evidence were used as the first database for BLASTP similarity searches. To identify the optimal sequence identity for this study, we ran our workflow with very high sequence identity cutoff for BLAST, then reduced it by 5% on each run and compared the results against the 836 proteins that were identified by proteomics in a single preliminary shotgun approach.

At 50% sequence identity we were able to find homologues for 2486 proteins (19.5% of 12 771). Reducing the threshold to a sequence identity ≤50% yielded 7447 (58.31%) out of 12 771 proteins. The black Périgord truffle is a relatively under-studied organism with unique features and biochemistry. It was therefore expected that most of the black Périgord truffle that is homologous to other sequences will have a lower coverage (i.e., sequence identity). Although sequence identity ≤50% yielded many more protein matches, to retain high-quality results, we have only considered results with sequence identity ≥50% for this study.

Sequential-BLAST Similarity Search

In the first round of our sequential-BLAST approach, we assessed the 12 771 proteins against yeast protein sequences with experimental protein evidence. Of these 1794 (14.0%) black Périgord truffle proteins showed significant matches, with 3 hits having ≥99% sequence identity, 11 hits with 90−95% sequence identity, and 8 hits with 85−90% sequence identity. The coverage ranged from 50 to 99.2% with *E* values of 8.00 × 10^{-6} to 0. No significant matches with coverage >50% were reported for the remaining sequences (Supporting Information, Supplementary Table S1). The second BLAST search against fungal protein sequences with experimental protein evidence for the remaining 10 977 black Périgord truffle proteins showed matches for 109 sequences (0.85%) with 50−89.6% sequence coverage and *E* values of 2.00 × 10^{-6} to 0, with two hits having ≥85% sequence identity (Supporting Information, Supplementary Table S2). The third BLAST search against reviewed fungal protein sequences for the remaining 10 868 proteins yielded significant results for 583 sequences (4.6%). For these matches, the coverage ranged from 50 to 100%, with *E* values ranging from 9.00 × 10^{-6} to 0. Of these, seven had 100% sequence identity, one with 95.5% sequence identity and 10 with sequence identities between 85 and 90% (Supporting Information, Supplementary Table S3). The remainder of these proteins (10 285) were matched against solved protein structures from the PDB. 101 proteins showed matches with coverage ranging from 50 to 80% with *E* values of 9.00 × 10^{-20} to 0 (Supporting Information, Supplementary Table S4). Since structures of homologous proteins show functional conservation up to sequence identities as low as 25%,[23] the knowledge of homologous structural information for these truffle proteins provides important functional clues.

The results clearly indicate that the black truffle's unique biology in the context of its evolution shows only a distant relationship to any other commonly studied fungus (such as yeast). The proteins that matched with very high similarity (identity >70%) were proteins known to be evolutionarily conserved[24] in most eukaryotic organisms and accounted for a very small proportion of the total protein complement (552 proteins). A significant proportion of the unique biology of this organism thus lies in the 10 184 putative proteins that had very low similarity to any known proteins

Functional Annotation

InterProScan for the 12 771 *T. melanosporum* proteome provided annotations for 6487 proteins (50.8%) with 1369 unique GO annotations (Supporting Information, Supplementary Table S5), while 1309 genes were manually curated by the genome consortium but are not publicly available for comparison.[4] Our analysis on GO biological processes revealed that the majority of the proteins were involved in dUTP metabolic processes (674), oxidation−reduction process (482), metabolic process (276), translation (184), and protein phosphorylation (184). A similar analysis on GO molecular function revealed that ATP binding (763), hydrolase activity (414), and catalytic activity (335) were the most common annotations. Protein domain and family mapping provided InterPro domains for 946, family for 910, active sites for 46, conserved sites for 162, and repeats for 18 proteins compared with only protein family annotations provided for the genome.[4]

Analysis by KEGG pathways revealed a large proportion of proteins identified from metabolic pathways (961), while proteins involved in the production of secondary metabolites were also seen (239 proteins) (Supporting Information, Supplementary Table S6). The top KEGG pathways are listed in Table 1. These findings suggest that the truffle is very metabolically active, possessing a cohort of biochemical and enzymatic activity that may explain to some degree its ability to produce over 90 volatiles[25] that modulate its flavor profile as well as its complex lifecycle.

Table 1. Top 5 KEGG Pathways for Bioinformatics and Proteomics Analyses

proteins from bioinformatics analysis		proteins from MS/proteomics analysis	
description	total match	description	total match
metabolic pathways	961	metabolic pathways	158
pyrimidine metabolism	399	biosynthesis of secondary metabolites	75
biosynthesis of secondary metabolites	239	ribosome	50
cell cycle - yeast	161	glycolysis/gluconeogenesis	21
meiosis - yeast	96	protein processing in endoplasmic reticulum	19

Mass Spectrometric Evidence for *T. melanosporum* Proteins

Mass spectrometry analyses of the *T. melanosporum* proteome identified 836 proteins (Supporting Information, Supplementary Table S7) that were assigned to the BT_Prot database (Mascot version 2.3.0). Analyses of the 836 proteins showed that 91% were identified by peptides ranging between 1 and 100 per protein, with the other 9% between 101 and 1010 hits (results not shown). A protein was positively identified if it had a minimum of one unique peptide with at least 99% confidence.

Numerous mass spectra (65 260) were not assigned to any protein as the stringency of the cutoff scores for accepting a peptide match was set quite high (99%) and the peptide tolerance window for experimental results was set low (±0.2 Da); this meant that numerous spectra did not pass high stringency tests. This meant that the data used were of sufficient quality to justify accuracy of the results. The spectra obtained were matched against the predicted translations of open reading frames from the genome. Undoubtedly, there would be a degree of mismatch between experimental and predicted results. There is a possibility that the unmatched peptides may have come from cross-contamination from peptides of other organisms/species.[26] The truffle, being in a symbiotic relationship, may indeed be sharing a significant amount of its structure with its host (hazelnut and oak trees). Additional investigations into potentially finding proteins in the unmatched peptide data were considered beyond the scope of this study, and future studies are planned to investigate this further.

The discovery of a total of 836 proteins (from a potential total of 12 771) by proteomics in a single preliminary shotgun approach in this study was higher than expected, as previous studies performed in unrelated fungal species identify a far lower proportion of nonredundant matches.[27−29] The proteomic coverage could still be increased using more sophisticated separation technologies such as PROOF fractionation,[30] which involves passing the sample through a tandem cation and anion exchange column, followed by peptide IPG-IEF prefractionation prior to MS analysis, which has been shown to significantly increase proteome coverage.[31] A more sophisticated information-independent SWATH-type analysis[32] could also be employed to ensure that any future findings can be quantitatively analyzed.

Comparison between Bioinformatics Analysis and Mass Spectrometry Evidence

Comparing the proteomics findings with our bioinformatics studies, in the first round of our BLASTP similarity search of these 836 proteins against reviewed yeast proteins with experimental evidence, 315 proteins (37.7%) were found to have significant matches (with 1 hit having ≥99% identity and 7 hits with ≥90% identity). In the second round of our BLASTP similarity search (against reviewed fungal proteins with experimental evidence), 18 proteins (2.15%) were found to have matches, with hits ≥85% identity. The third round of BLASTP similarity search (against reviewed fungal proteins) resulted in 55 proteins (6.58%; including one match with 100% identity and two matches with ≥90% identity). Nine proteins matched 3D structures in the PDB. Yeast is one of the most widely studied organisms/fungi over the years, and not surprisingly >50% of all matches found were to yeast proteins with experimental evidence. Of the *T. melanosporum* proteins identified by our proteomics study, 439 (52.5%) had no matches in our bioinformatics results with sequence identity >50%. The results have been summarized in Table 2, with details of the BLASTP matches for these 836 proteins provided as Supporting Information (Table S8).

Furthermore, we compared our findings with the 14 UniProt proteins identified in *T. melanosporum* (strain Mel28) that have been previously reviewed but not verified experimentally (Supporting Information, Supplementary Table S9). Out of these 14 reviewed proteins, three were identified by mass spectrometry with protein coverage ranging from 16 to 40% (Supporting Information, Supplementary Table S10). These proteins include a probable amino/metallo-peptidase (Gene ID: AMPP1; UniProt ID: D5GAC6), a nuclear- and cytoplasm-residing dioxygenase involved in L-methionine salvage (Gene ID: MTND; UniProt ID: D5GE59), and an integral membrane catalytic subunit of a signal peptidase complex found in the ER membrane (Gene ID: SEC11; UniProt ID: D5GNC3) that was deduced to be involved in proteolysis and signal peptide processing. In addition, of the 17 proteins reported in the recent proteomics study of *T. magnatum*,[16] we have identified *T. melanosporum* homologues to 12 proteins, of which three were found only by bioinformatics analysis and one uniquely in our proteomics data.

Table 2. Comparative Summary of Bioinformatics Analysis and Mass Spectrometry Evidences

description	bioinformatics analysis	proteins from mass spectra (validated by bioinformatics)
total number of proteins	12 771	836
reviewed yeast protein sequences with experimental protein evidence	1794	315
reviewed fungal protein sequences with experimental protein evidence	109	18
reviewed fungal protein sequences	583	55
protein data bank (PDB)	101	9
functional annotation (GO, InterPro, KEGG)	2486	731

Overall, at least 105 of the 836 proteins could not be assigned any putative function. These proteins could be used as putative candidate markers of truffles. Alternatively, a proportion of these might be species-specific biomarkers for the black Périgord truffle (*T. melanosporum*). Further proteomics studies using single or multiple reaction monitoring experiments (SRM/MRM)[33] can quantitatively be used to study this cohort of proteins to discern markers of authenticity, a study beyond the scope of this report.

Functional annotation for the 836 proteins (Supporting Information, Supplementary Table S11) provided GO annotations for 698, InterPro domains for 768, and enzyme codes (ECs) for 225. The ECs were related to 90 corresponding KEGG biochemical pathways. Our analysis of the GO terms for the 836 proteins revealed that the majority were involved in binding (436), catalytic activity (350), localization within cells (288), and cell organelles (151) or with taking part in metabolic biological processes (416) or cellular (232) processes (Figure 2). Analysis by KEGG pathways revealed a large proportion of proteins identified from metabolic pathways (158), while proteins involved in the production of secondary metabolites (75 proteins) were also noted (Table 1). In all, 731 (87.4%) of the 836 proteins were successfully annotated with either GO, InterPro domains, or KEGG pathways. However, with the currently available biological knowledge, 105 proteins could not be annotated at all. These findings reflect those obtained from the bioinformatics studies and suggest that the proteomic assessment was representative of the currently annotated data.

Proteins from Truffle That Confer Aroma

The enzymes that catalyze the production of volatiles that confer the unique aroma of the black truffle were analyzed from the pathway relating to the production of secondary metabolites (including the sulfur and methane metabolism pathways) in the KEGG database. A comprehensive analysis of the functionally annotated proteins found by proteomics and bioinformatics revealed that the proportion of proteins involved in the production of secondary metabolites in the fruiting body of the black truffle (from proteomics) (9.6%) was similar to that found *in silico* (9%). This list of proteins involved in secondary metabolism was matched against enzymes known to be involved in pathways that produce volatiles found in truffles from previous biochemical studies.[25,34−37] A total of nine proteins were identified (Table 3).

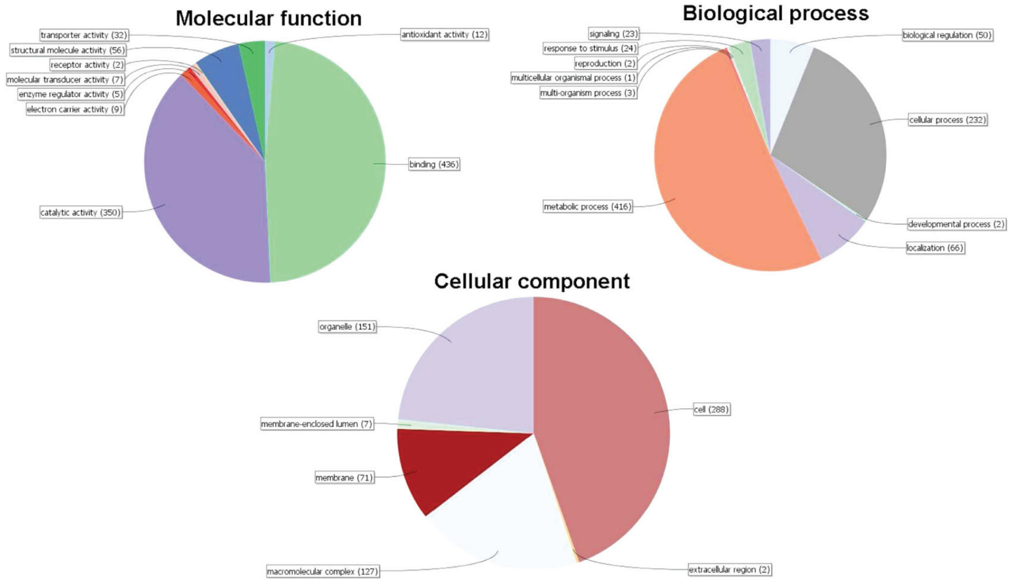


Figure 2. Pictorial representation of GO distributions for the 836 *T. melanosporum* proteins. Pie charts depicting the distributions are shown with respective molecular functions, cellular components, and biological processes labeled and number of proteins involved shown in parentheses.

Table 3. List of Previously[23,32−35] Identified Volatiles and the Enzymes Involved in Their Synthesis Matched to Proteins Obtained from the *T. melanosporum* Proteins Annotated with Bioinformatics and Substantiated with Proteomic (MS) Evidence (indicated with *)

volatile	KEGG compound code	enzymes	truffle protein	annotation
acetaldehyde	C00084	ribose-5-phosphate transaldolase, or fluorothreonine transaldolase	GSTUMT00000035001*	transaldolase
acetaldehyde	C00084	aldehyde dehydrogenase	GSTUMT00003865001*	aldehyde/histidinol dehydrogenase
acetaldehyde	C00084	alcohol dehydrogenase	GSTUMT00006862001*,[a]	alcohol dehydrogenase
anisole and methoxybenzene	C01403	phenol *O*-methyltransferase	GSTUMT00004489001*	sterol methyltransferase
2-methyl butanal	C02223	branched-chain-2-oxoacid decarboxylase	GSTUMT00001753001*, GSTUMT00000274001, GSTUMT00003265001	2-oxoacid dehydrogenase acyltransferase
propanal	C00479	propanediol dehydratase	GSTUMT00010247001	dihydroxy-acid/6-phosphogluconate dehydratase
phenylacetaldehyde	C00601	amine oxidase (pyridoxal containing)	GSTUMT00008176001	pyridoxal phosphate-dependent transferase

[a] Proteins identified as associated with volatiles in the genome publication.[4]

The organoleptic properties, particularly the unique aroma, of the truffle are arguably its most valuable asset, not only biologically in attracting animals for sporulation[38] but economically for gastronomists and food lovers. It has been previously shown from the analysis of the genome that the truffle possesses much of the machinery required for synthesizing its aroma. In this preliminary study, for the first time, we were able to potentially identify nine proteins responsible for part of the aroma profile of truffles, although more biochemical analyses need to be carried out to confirm the findings. Of these, only one (GSTUMT00006862001) has been identified as involved in secondary metabolism by the 2010 genome publication.[4]

Two compounds, DMS and 2-methylbutanal, when mixed in the right proportions, mimic the aroma of the black Périgord truffle, *T. melanosporum*.[39] The latter mixture has been used for standardly by the food industry to imitate black truffle aroma. The interesting discovery from bioinformatics with proteomics evidence of the enzyme potentially responsible for the metabolism of 2-methyl butanal validates the approaches taken.

A large proportion of the proteins annotated form secondary metabolite pathways that have proteomic as well as bioinformatics evidence did not match to known pathways of volatile synthesis. This, compounded by the fact that the enzymes involved in the production of some volatiles remain to be discovered, suggests that a more comprehensive biochemical study of the enzyme components of the *T. melanosporum* is warranted. It is hardly surprising that such low numbers of enzymes were shown to be involved in the production of over 90 volatiles considering over 70% of the proteome is yet to be annotated. The potential to discover novel enzymes that could be of economic, medicinal, or other uses remains a tantalizing possibility.

■ CONCLUSIONS

Only 14 black truffle proteins have been reviewed in UniProt. The remainder, although recently annotated, are not yet reviewed and await curation, while 1309 genes were manually curated by the truffle genome consortium.[4] We have provided high-quality bioinformatics annotations for 2587 sequences and proteomic evidence of 836 truffle proteins. Using selected high-quality protein databases for similarity searches using BLAST sequentially, we identified homologues with experimental evidence for 14.9% of the black Périgord truffle proteome, with a further 4.6% mapping to reviewed fungal proteins and another 0.8% mapping to protein structures, totalling 2587 proteins (20.2% of the *T. melanosporum* 12 771 proteins). The acquisition of functional experimental evidence of these proteins is quite possible, as most matches were to a well-characterized fungus, *S. cerevisiae*. Additionally, using a suite of bioinformatics tools, we have assigned putative biological functions in terms of gene ontology, biochemical pathway, and domain/motif signatures for 2486 of these 2587 sequences. Using a proteomics approach, we have provided proteomic evidence of 836 proteins, none of which have been reported experimentally to date, including three of the 14 UniProt reviewed proteins that lacked proteomics evidence. Using a combination of computational strategies, we were able to identify nine proteins responsible for part of the aroma profile of truffles and for the first time suggest a potential enzymatic pathway for the production of one of the primary volatiles in black truffle. Approximately 20% of the 12 771 proteins have been assigned putative biological functionality, providing valuable clues for experimental validation and future work. We have described a generic framework that is validated by our proteomics studies and can be used to annotate the proteome of any novel organism.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Table S1: BLAST hits showing high % identity against reviewed yeast proteins with experimental evidence. Hits sorted by % identity and only the hits with ≥50% sequence identities are shown. **Table S2**: BLAST hits showing high % identity against reviewed fungal proteins with experimental evidence. Hits sorted by % identity and only the hits with ≥50% sequence identities are shown. **Table S3**: BLAST hits showing high % identity against reviewed fungal proteins. Hits sorted by % identity and only the hits with ≥50% sequence identities are shown. **Table S4**: BLAST hits showing high % identity against PDB proteins. Hits sorted by % identity and only the hits with ≥50% sequence identities are shown. **Table S5**: InterPro and GO annotations of *T. melanosporum* proteins. **Table S6**: KEGG Pathways for *T. melanosporum* proteins. **Table S7**: List of proteins identified by mass spectrometry. **Table S8**: BLAST hits of proteins found by mass spectrometry. Hits sorted by % identity and only the hits with ≥50% sequence identities are shown. **Table S9**: Summary of UniProt reviewed proteins for *T. melanosporum* (strain Mel28). **Table S10**: Summary of proteins found by mass spectrometry and reviewed in UniProt (2013_05) . **Table S11**: Functional annotations of the 836 *T. melanosporum* proteins. Hits sorted by #GOs, GO, InterProScan, and KEGG Enzyme Codes. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*Tel: +61 2 9850 8211. Fax: +61 2 9850 8313. E-mail: shoba.ranganathan@mq.edu.au.
*E-mail: mark.baker@mq.edu.au.

Author Contributions

#M.T.I., A.M., and G.G. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We sincerely acknowledge the contributions of Dr. Alison Kan, who assisted in the proteomics analysis reported in this paper. Proteomics analysis was carried out at the Australian Proteome Analysis Facility, a Major National Research Facility of Australia. Financial support was provided by a Macquarie University Research Excellence Scholarship (MQRES) to MTI in the Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, Australia.

■ REFERENCES

(1) Trappe, J. M.; Claridge, A. W. The hidden life of truffles. *Sci. Am.* **2010**, *302* (4), 78−82, 84.

(2) Maser, C.; Claridge, A. W.; Trappe, J. M. *Trees, Truffles, and Beasts: How Forests Function*; Rutgers University Press: New Brunswick, NJ, 2008; p xvi, 280 p, 8 p of plates.

(3) Bohannon, J. Genetics. Rooting around the truffle genome. *Science* **2009**, *323* (5917), 1006−7.

(4) Martin, F.; Kohler, A.; Murat, C.; Balestrini, R.; Coutinho, P. M.; Jaillon, O.; Montanini, B.; Morin, E.; Noel, B.; Percudani, R.; Porcel, B.; Rubini, A.; Amicucci, A.; Amselem, J.; Anthouard, V.; Arcioni, S.; Artiguenave, F.; Aury, J. M.; Ballario, P.; Bolchi, A.; Brenna, A.; Brun, A.; Buee, M.; Cantarel, B.; Chevalier, G.; Couloux, A.; Da Silva, C.; Denoeud, F.; Duplessis, S.; Ghignone, S.; Hilselberger, B.; Iotti, M.; Marcais, B.; Mello, A.; Miranda, M.; Pacioni, G.; Quesneville, H.; Riccioni, C.; Ruotolo, R.; Splivallo, R.; Stocchi, V.; Tisserant, E.; Viscomi, A. R.; Zambonelli, A.; Zampieri, E.; Henrissat, B.; Lebrun, M. H.; Paolocci, F.; Bonfante, P.; Ottonello, S.; Wincker, P. Perigord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. *Nature* **2010**, *464* (7291), 1033−8.

(5) Martin, F. Unearthing the truffle genome. *New Phytol.* **2011**, *189* (3), 645−6.

(6) Paolocci, F.; Rubini, A.; Granetti, B.; Arcioni, S. Typing *Tuber melanosporum* and Chinese black truffle species by molecular markers. *FEMS Microbiol. Lett.* **1997**, *153* (2), 255−60.

(7) Murat, C.; Martin, F. Sex and truffles: first evidence of Perigord black truffle outcrosses. *New Phytol.* **2008**, *180* (2), 260−3.

(8) Ceccaroli, P.; Buffalini, M.; Saltarelli, R.; Barbieri, E.; Polidori, E.; Ottonello, S.; Kohler, A.; Tisserant, E.; Martin, F.; Stocchi, V. Genomic profiling of carbohydrate metabolism in the ectomycorrhizal fungus *Tuber melanosporum*. *New Phytol.* **2011**, *189* (3), 751−64.

(9) Montanini, B.; Levati, E.; Bolchi, A.; Kohler, A.; Morin, E.; Tisserant, E.; Martin, F.; Ottonello, S. Genome-wide search and functional identification of transcription factors in the mycorrhizal fungus *Tuber melanosporum*. *New Phytol.* **2011**, *189* (3), 736−50.

(10) Rubini, A.; Belfiori, B.; Riccioni, C.; Arcioni, S.; Martin, F.; Paolocci, F. *Tuber melanosporum*: mating type distribution in a natural plantation and dynamics of strains of different mating types on the roots of nursery-inoculated host plants. *New Phytol.* **2011**, *189* (3), 723−35.

(11) Li, Y. Y.; Wang, G.; Li, H. M.; Zhong, J. J.; Tang, Y. J. Volatile organic compounds from a *Tuber melanosporum* fermentation system. *Food Chem.* **2012**, *135* (4), 2628−37.

(12) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2012**, *40*, (Database issue), D71-5.

(13) Doyle, S. Fungal proteomics: from identification to function. *FEMS Microbiol. Lett.* **2011**, *321* (1), 1−9.

(14) Kniemeyer, O.; Schmidt, A. D.; Vodisch, M.; Wartenberg, D.; Brakhage, A. A. Identification of virulence determinants of the human

pathogenic fungi *Aspergillus fumigatus* and *Candida albicans* by proteomics. *Int. J. Med. Microbiol.* **2011**, *301* (5), 368−77.

(15) Vincent, D.; Balesdent, M. H.; Gibon, J.; Claverol, S.; Lapaillerie, D.; Lomenech, A. M.; Blaise, F.; Rouxel, T.; Martin, F.; Bonneu, M.; Amselem, J.; Dominguez, V.; Howlett, B. J.; Wincker, P.; Joets, J.; Lebrun, M. H.; Plomion, C. Hunting down fungal secretomes using liquid-phase IEF prior to high resolution 2-DE. *Electrophoresis* **2009**, *30* (23), 4118−36.

(16) Vita, F.; Lucarotti, V.; Alpi, E.; Balestrini, R.; Mello, A.; Bachi, A.; Alessio, M.; Alpi, A. Proteins from *Tuber magnatum Pico* fruiting bodies naturally grown in different areas of Italy. *Proteome Sci.* **2013**, *11* (1), 7.

(17) Ranganathan, S.; Khan, J. M.; Garg, G.; Baker, M. S. Functional annotation of the human chromosome 7 ″missing″ proteins: a bioinformatics approach. *J. Proteome Res.* **2013**, *12* (6), 2504−10.

(18) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389−402.

(19) Quevillon, E.; Silventoinen, V.; Pillai, S.; Harte, N.; Mulder, N.; Apweiler, R.; Lopez, R. InterProScan: protein domains identifier. *Nucleic Acids Res.* **2005**, *33*, W116−20.

(20) Xie, C.; Mao, X.; Huang, J.; Ding, Y.; Wu, J.; Dong, S.; Kong, L.; Gao, G.; Li, C. Y.; Wei, L. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* **2011**, *39*, W316−22.

(21) Rosenfeld, J.; Capdevielle, J.; Guillemot, J. C.; Ferrara, P. In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis. *Anal. Biochem.* **1992**, *203* (1), 173−9.

(22) Mirzaei, M.; Pascovici, D.; Atwell, B. J.; Haynes, P. A. Differential regulation of aquaporins, small GTPases and V-ATPases proteins in rice leaves subjected to drought stress and recovery. *Proteomics* **2012**, *12* (6), 864−877.

(23) Martin, A. C.; MacArthur, M. W.; Thornton, J. M. Assessment of comparative modeling in CASP2. *Proteins* **1997**, No. Suppl 1, 14−28.

(24) Basu, M. K.; aaCarmel, L.; Rogozin, I. B.; Koonin, E. V. Evolution of protein domain promiscuity in eukaryotes. *Genome Res.* **2008**, *18* (3), 449−61.

(25) Splivallo, R.; Ottonello, S.; Mello, A.; Karlovsky, P. Truffle volatiles: from chemical ecology to aroma biosynthesis. *New Phytol.* **2011**, *189* (3), 688−99.

(26) Ding, Q.; Xiao, L.; Xiong, S.; Jia, Y.; Que, H.; Guo, Y.; Liu, S. Unmatched masses in peptide mass fingerprints caused by cross-contamination: An updated statistical result. *Proteomics* **2003**, *3*, 1313−1317.

(27) Wang, M.; Gu, B.; Huang, J.; Jiang, S.; Chen, Y.; Yin, Y.; Pan, Y.; Yu, G.; Li, Y.; Wong, B. H.; Liang, Y.; Sun, H. Transcriptome and proteome exploration to provide a resource for the study of *Agrocybe aegerita*. *PLoS One* **2013**, *8* (2), e56686.

(28) Horie, K.; Rakwal, R.; Hirano, M.; Shibato, J.; Nam, H. W.; Kim, Y. S.; Kouzuma, Y.; Agrawal, G. K.; Masuo, Y.; Yonekura, M. Proteomics of two cultivated mushrooms *Sparassis crispa* and *Hericium erinaceum* provides insight into their numerous functional protein components and diversity. *J. Proteome Res.* **2008**, *7* (5), 1819−35.

(29) Andersson, K. M.; Meerupati, T.; Levander, F.; Friman, E.; Ahren, D.; Tunlid, A. Characterization of the proteome of the nematode-trapping cells of the fungus *Monacrosporium haptotylum*. *Appl. Environ. Microbiol.* **2013**, *79*, 4993−5004.

(30) Tan, S. H.; Mohamedali, A.; Kapur, A.; Baker, M. S. Ultradepletion of Human Plasma using Chicken Antibodies: A Proof of Concept Study. *J. Proteome Res.* **2013**, *12* (6), 2399−413.

(31) McQuade, L. R.; Schmidt, U.; Pascovici, D.; Stojanov, T.; Baker, M. S. Improved membrane proteomics coverage of human embryonic stem cells by peptide IPG-IEF. *J. Proteome Res.* **2009**, *8* (12), 5642−9.

(32) Gillet, L. C.; Navarro, P.; Tate, S.; Rost, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics* **2012**, *11* (6), O111 016717.

(33) Boja, E. S.; Rodriguez, H. Mass spectrometry-based targeted quantitative proteomics: achieving sensitive and reproducible detection of proteins. *Proteomics* **2012**, *12* (8), 1093−110.

(34) Cullere, L.; Ferreira, V.; Venturini, M. E.; Marco, P.; Blanco, D. Potential aromatic compounds as markers to differentiate between *Tuber melanosporum* and *Tuber indicum* truffles. *Food Chem.* **2013**, *141* (1), 105−10.

(35) Davoli, P.; Bellesia, F.; Pinetti, A. Comments on Truffle aroma analysis by headspace solid phase microextraction [Is butylated hydroxytoluene (BHT) a ″natural″ volatile constituent of truffles?]. *J. Agric. Food Chem.* **2003**, *51* (15), 4483 author reply 4484..

(36) Diaz, P.; Ibanez, E.; Senorans, F. J.; Reglero, G. Truffle aroma characterization by headspace solid-phase microextraction. *J. Chromatogr., A* **2003**, *1017* (1−2), 207−14.

(37) Diaz, P.; Senorans, F. J.; Reglero, G.; Ibanez, E. Truffle aroma analysis by headspace solid phase microextraction. *J. Agric. Food Chem.* **2002**, *50* (22), 6468−72.

(38) Linde, C. C.; Selmes, H. Genetic Diversity and Mating Type Distribution of *Tuber melanosporum* and Their Significance to Truffle Cultivation in Artificially Planted Truffieres in Australia. *Appl. Environ. Microbiol.* **2012**, *78* (18), 6534−9.

(39) Cullere, L.; Ferreira, V.; Chevret, B.; Venturini, M. E.; Sanchez-Gimeno, A. C.; Blanco, D. Characterisation of aroma active compounds in black truffles (*Tuber melanosporum*) and summer truffles (*Tuber aestivum*) by gas chromatography-olfactometry. *Food Chem.* **2010**, *122* (1), 300−306.

3.3 Conclusions

In this study, using the sequential similarity search technique with high-quality protein sequences databases we identified homologous for 2486 proteins (UniProt databases). Additionally, we identified structural similarity to 101 proteins with the Protein Data Bank (PDB) sequences. Our approach identified functional annotations for 96% of these proteins. The shotgun proteomics identified 836 proteins, and 47% of these proteins were also identified by the *in silico* approach. Furthermore, a deeper computational analysis on the functional annotation provided by our approach revealed nine proteins, responsible for the aroma profile and a potential enzymatic pathway to produce one of the primary volatiles of the black truffle. This demonstrates that our approach (validated and complemented by the proteomics study) can functionally annotate proteins with high confidence which can lead to biological understandings of organisms. Our approach described here provides a generic functional annotation framework and that can be applied to other novel or less studied organisms. To demonstrate the reusability and generic nature of the framework, we applied this technique to functionally annotate the human 'missing' proteins (described in Chapter 4).

Chapter 4: Protannotator: A Semiautomated Pipeline for Chromosome-Wise Functional Annotation of the "Missing" Human Proteome

4.1 Summary

In this chapter, we show the application of the generic annotation strategy outlined in the earlier section. The Human Proteome Project (HUPO) launched the Chromosome-centric HPP (C-HPP) project [23] with 25 member institutes [26] across the world to accurately identify and characterise all human proteins (chromosome by chromosome) using strict baseline metrics [282]. On behalf of C-HPPP, neXtProt classifies all human proteins to five protein existence (PE) levels based on the information available according to its guideline. Proteins identified by mass spectrometry, immunohistochemistry, 3D structure, and or amino acid sequencing are classified as PE1, whereas proteins identified without protein expression evidence, but detected by transcript expression, homology, and predicted gene models are categorised as PE2, PE3, and PE4 respectively. Proteins with protein existence level PE2-PE4 are classified as 'missing' protein. neXtProt reported 19% of the human proteins as 'missing' (current at the time of the study, September 2013 release). In other words, very little information is available for these proteins. Identifying homologous proteins and functional annotations from the most up to date information can not only uncover the unknown biology of these proteins but also accelerate the identification of these proteins.

4.2 Publication 2

Mohammad T. Islam, Gagan Garg, William S. Hancock, Brian A. Risk, Mark S. Baker, and Shoba Ranganathan (2014) Protannotator: A Semiautomated Pipeline for Chromosome-Wise Functional Annotation of the “Missing” Human Proteome. *Journal of Proteome Research*, vol. 13, no.1, pp. 76-83. DOI: 10.1021/pr400794x

Journal of proteome research

Article

pubs.acs.org/jpr

Protannotator: A Semiautomated Pipeline for Chromosome-Wise Functional Annotation of the "Missing" Human Proteome

Mohammad T. Islam,[†,‡,#] Gagan Garg,[†,‡,#] William S. Hancock,[§] Brian A. Risk,[‖] Mark S. Baker,[†] and Shoba Ranganathan*,[†,‡,⊥]

[†]Department of Chemistry and Biomolecular Sciences and [‡]ARC Centre of Excellence in Bioinformatics, Macquarie University, Sydney, NSW 2109, Australia

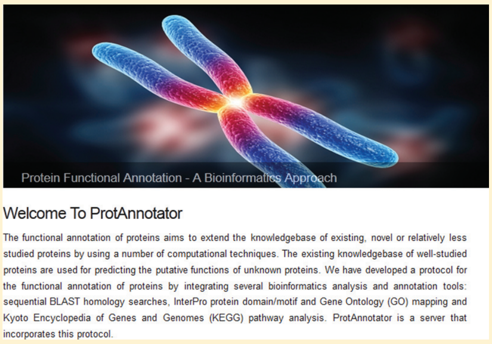
[§]Barnett Institute, Northeastern University, 140 The Fenway, Boston, Massachusetts 02115, United States

[‖]College of Arts and Sciences, Boise State University, 1910 University Drive, Boise, Idaho 83725, United States

[⊥]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 14 Medical Drive, 117599 Singapore

S *Supporting Information*

ABSTRACT: The chromosome-centric human proteome project (C-HPP) aims to define the complete set of proteins encoded in each human chromosome. The neXtProt database (September 2013) lists 20 128 proteins for the human proteome, of which 3831 human proteins (~19%) are considered "missing" according to the standard metrics table (released September 27, 2013). In support of the C-HPP initiative, we have extended the annotation strategy developed for human chromosome 7 "missing" proteins into a semiautomated pipeline to functionally annotate the "missing" human proteome. This pipeline integrates a suite of bioinformatics analysis and annotation software tools to identify homologues and map putative functional signatures, gene ontology, and biochemical pathways. From sequential BLAST searches, we have primarily identified homologues from reviewed nonhuman mammalian proteins with protein evidence for 1271 (33.2%) "missing" proteins, followed by 703 (18.4%) homologues from reviewed nonhuman mammalian proteins and subsequently 564 (14.7%) homologues from reviewed human proteins. Functional annotations for 1945 (50.8%) "missing" proteins were also determined. To accelerate the identification of "missing" proteins from proteomics studies, we generated proteotypic peptides *in silico*. Matching these proteotypic peptides to ENCODE proteogenomic data resulted in proteomic evidence for 107 (2.8%) of the 3831 "missing proteins, while evidence from a recent membrane proteomic study supported the existence for another 15 "missing" proteins. The chromosome-wise functional annotation of all "missing" proteins is freely available to the scientific community through our web server (http://biolinfo.org/protannotator).



Protein Functional Annotation - A Bioinformatics Approach

Welcome To ProtAnnotator

The functional annotation of proteins aims to extend the knowledgebase of existing, novel or relatively less studied proteins by using a number of computational techniques. The existing knowledgebase of well-studied proteins are used for predicting the putative functions of unknown proteins. We have developed a protocol for the functional annotation of proteins by integrating several bioinformatics analysis and annotation tools: sequential BLAST homology searches, InterPro protein domain/motif and Gene Ontology (GO) mapping and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis. ProtAnnotator is a server that incorporates this protocol.

KEYWORDS: *Human Proteome Project, human chromosome, missing proteins, sequential BLAST, functional annotation, proteotypic peptides, proteogenomics*

INTRODUCTION

The interpretation of the human genome depends on detailed annotation, usually at the nucleotide level, the protein level, and the process level,[1] for which the functional annotation of proteins is crucial at the process level. Since 2008, the Human Proteome Organization (HUPO) has pursued the comprehensive identification and functional characterization of the human proteome via the Human Proteome Project (HPP),[2] of which the chromosome-centric HPP (C-HPP) approach seeks to catalog the human proteome on the basis of chromosomes.[3−5] The International Chromosome-centric Human Proteome Project (C-HPP), launched in 2012, marks the first step toward the genome-wide chromosome by chromosome characterization of the human proteome.[6] Such an approach would address a key aim of the human genome project, viz. personalized medicine, by providing sensitive and highly specific protein biomarkers for early onset diagnosis, prognosis and treatment of several diseases, providing clinical and translational proteomic solutions.[7]

The three pillars of HPP are mass spectrometric proteomics, antibody/affinity capturing agents, and a knowledgebase,[2] embodied by the neXtProt database,[8] where detailed information on the human proteome is collated, curated, and organized for rapid access of information on a query protein. Our group carried out the functional annotation of missing

Special Issue: Chromosome-centric Human Proteome Project

Received: August 1, 2013
Published: December 6, 2013

proteins of human chromosome 7 (hChr7),[9] developing a sequential BLAST homology search approach, with the neXtProt[8] data available at the time. A list of missing proteins was also compiled and investigated for chromosome 17,[10] where proteins having no proteomic identifications from sources including PeptideAtlas[11] or GPM (http://www.thegpm.org/) were considered "missing." Currently, for the entire human genome, neXtProt (as of September 2013) lists 20 128 proteins, by extending their data sources to incorporate all peptides from PeptideAtlas Human builds (August 2013) as "GOLD" (i.e., <1% error) as well as 20 other studies (unpublished data). Thus, 3831 proteins (~19%; excluding unmapped and redundant sequences) are considered "missing," based on the currently available C-HPP standard metrics table, developed with neXtProt data (http://www.c-hpp.org/gnuboard4/bbs/board.php?bo_table=public).

The UniProtKB/Swiss-Prot database[12] (release 2013_10) with >541 000 entries of reviewed and annotated proteins serves as the highest quality database for bioinformatics studies. Homologous sequences often display identical or similar biological functions. The biological knowledge available in this database can be mined by similarity searches to identify if any of these missing proteins are homologous to similar proteins in higher mammals or other species. BLAST[13,14] programs are widely used for sequence similarity searches, using the default nonredundant (NR) data sets, which include putative, unannotated, or translated coding regions. Thus, a similarity search against NR data sets may result in matches to large numbers of unreviewed or unannotated proteins. Previously, we have annotated less studied organisms such as helminth parasites and fungal pathogens[15−19] by combining similarity searches and functional annotations including gene ontology (GO), biochemical pathways, and functional domains and motifs. Following a targeted BLAST approach, labeled "sequential BLAST search" from our previous hChr7 "missing" protein annotation,[9] where we have run repeated BLAST searches against selected databases providing high-quality reviewed annotations, we now present a semiautomated pipeline for the annotation of "missing" proteins. This is a generic approach and can be adopted for the annotation of any novel proteome, for example, black Périgord truffle (*Tuber melanosporum*).[20] Using this approach, we have annotated the entire set of "missing" proteins in the human proteome. Out of 3831 "missing" proteins, 1271 sequences (33.2%) were homologous to nonhuman reviewed mammalian proteins with proteomic evidence, while 703 proteins (18.4%) had nonhuman reviewed mammalian homologues. 1945 (50.8%) of the "missing" proteins were assigned putative GO and domain/motif annotations, using strict parameters (detailed in Materials and Methods), while 1250 (32.7%) "missing" proteins were mapped to biochemical pathways. We have also generated proteotypic peptides to facilitate proteomic identification of the "missing" proteins. These proteotypic peptides enabled us to garner proteomic evidence for 107 "missing" proteins, using proteogenomic data accurately matching the peptides from the ENCODE project.[21−23] Also, a recent in-depth proteomic study of breast cancer tissues by Muraoka et al.[24] has reported 851 membrane proteins that currently lack evidence by mass spectrometry in the neXtProt database. From this study, we have identified 15 additional "missing" proteins, which together with the ENCODE proteogenomic data have provided proteomic evidence for 122 "missing" proteins. The annotated data for the human proteome have been compiled into a database, which is freely available to the scientific community.

■ MATERIALS AND METHODS

1. Data Sources

Chromosome reports for each human chromosome were downloaded from the neXtProt database[8] (release September 2013). From these reports, sequences for "missing" proteins were extracted in FASTA format. A number of protein data sets were downloaded from UniProtKB/Swiss-Prot database[12] to our local Linux server for database similarity search. These include nonhuman reviewed mammalian proteins with experimental evidence (14 910 sequences), nonhuman reviewed mammalian proteins (45 926 sequences), human-reviewed proteins (23 515 sequences), and Protein Data Bank (PDB)[25] proteins (260 382 sequences), as in our hChr7 study.[9] We used the PDB to obtain possible matches against proteins with known structures from nonmammalian organisms. Verification data sets used for this study comprise the set of all mammalian proteins (1 155 455 sequences) and the nonmammalian protein set with protein evidence (70 830 sequences).

2. Database Similarity Search

Database similarity search technique is used to identify if a novel sequence is homologous to sequences that are already available in existing databases. BLAST[13,14] is A widely used tool for sequence similarity search. A query sequence is considered to be strongly homologous if it matches against a subject sequence with high significance (*E* value: 1×10^{-5}, compared with 1×10^{-3})[26] and sequence identity of at least 50%.[27]

We ran BLASTP searches sequentially against the data sets previously described using default parameters with a minimum *E* value of 1×10^{-5}. Those sequences that yielded no matches against the first data set were matched against the next data set, then the third, and so on. Missing proteins were also functionally annotated based on GO, pathways, and protein domain mapping. This is described in detail in Section 5 (Protannotator bioinformatics pipeline).

3. Functional Annotation of Missing Proteins

"Missing" proteins are provided putative functional annotation by mapping to protein domain, motif, and families. Functional annotations were further strengthened by assigning GO terms to the proteins. InterProScan[28] is widely used for protein functional annotation. It scans the protein sequences using the different protein signature recognition methods (Hidden Markov Model and BLAST) in its 13 protein domain and functional site databases combined in InterPro[29] database.

To obtain the best results, we ran InterProScan with default programs, described in detail in our previous paper,[9] while KOBAS[30] (KEGG Orthology-Based Annotation System, KOBAS-2.0) results provided pathway mapping. The program identifies statistically significantly enriched pathways by first mapping the proteins to genes in KEGG GENES based on BLAST searches followed by mapping against the whole human genome as the background. These programs have been successfully employed for the comprehensive annotation of novel and uncharacterized sequences[15−18] and in recent genome projects of less studied organisms.[31,32]

4. *In Silico* Tryptic Digestion

Protein Digestion Simulator[33] was used with default parameters (fragment mass range of 400−6000 Da; pI range of 0−14; mass

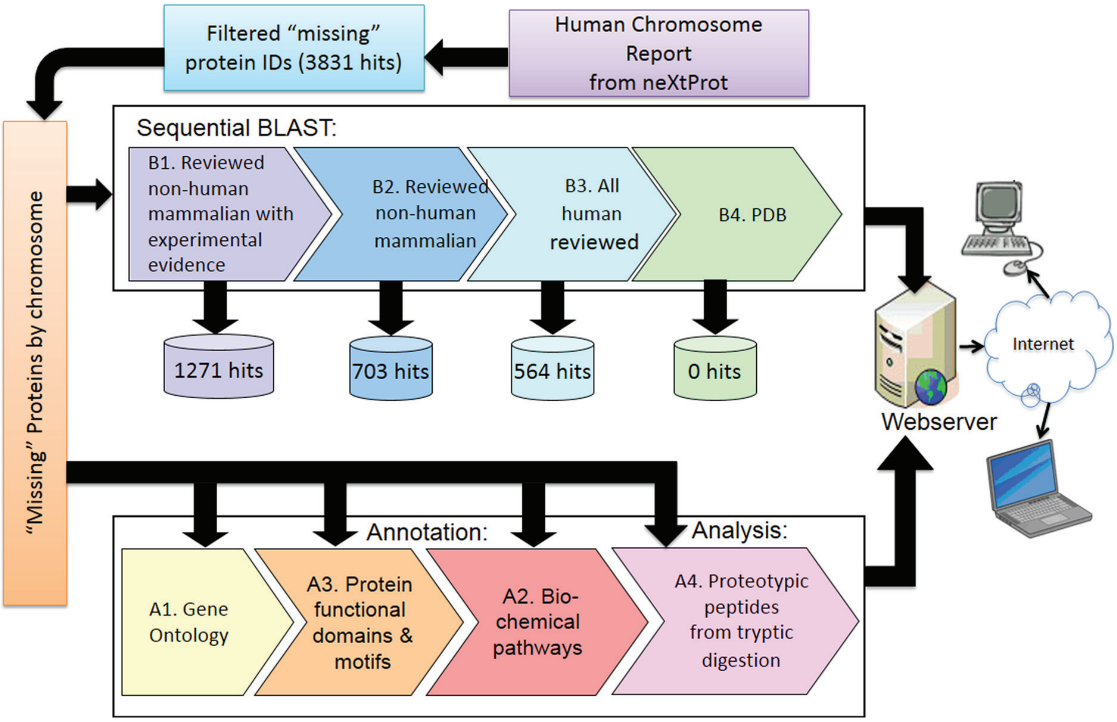

Filtered "missing" protein IDs (3831 hits)
Human Chromosome Report from neXtProt
"Missing" Proteins by chromosome
Sequential BLAST:
B1. Reviewed non-human mammalian with experimental evidence
B2. Reviewed non-human mammalian
B3. All human reviewed
B4. PDB
1271 hits
703 hits
564 hits
0 hits
Webserver
Internet
Annotation:
Analysis:
A1. Gene Ontology
A3. Protein functional domains & motifs
A2. Bio-chemical pathways
A4. Proteotypic peptides from tryptic digestion

Figure 1. Top-level architecture of the pipeline for annotating human "missing" proteins. Proteins were passed through a series of databases to determine homology (sequential BLAST) as well as annotation databases based on GO, protein functional domains, motifs, and biochemical pathways.

tolerance of 5 ppm; Hopp and Woods[34] hydrophobicity mode) to computationally digest the "missing proteins" with trypsin, to identify proteotypic peptide sequences. Input sequences were validated, and duplicates were removed. These peptide sequences were then matched against ENCODE proteogenomic data,[20] generated on the basis of peptide spectrum scoring system[23] using the Peppy software.[22]

5. Protannotator Bioinformatics Pipeline

We have developed a semiautomated pipeline, called Protannotator, for the "missing" human proteome annotation based on the workflow reported in our previous hChr7 study.[9] All programs and tools used in this study were installed on a Linux cluster running on Ubuntu server operating system. The data are served using the Apache webserver with a PHP front end. The different components of the workflow system are linked using Perl, Python, and bash shell scripts into a workflow. The top-level architecture of the pipeline is illustrated in Figure 1.

In the first step, Protannotator extracted the proteome details of each chromosome from neXtProt, available as chromosome reports. Human proteins were then sorted based on the availability of protein evidence. The system then identified proteins characterized as "missing" proteins based on accession numbers provided for protein evidence level 2–4 (consistent with the recent C-HPP standard metrics table) and extracted their sequences in FASTA format from UniProt using Linux's wget utility.

The protein sequence files thus extracted were processed by the Database Similarity Search (DSS) module of the pipeline. DSS uses a series of sequential BLAST searches to identify high-quality matches. The "missing" proteins are first searched against reviewed nonhuman mammalian proteins with experimental evidence, with the unmatched sequences then searched against nonhuman reviewed mammalian proteins. The unmatched sequences from the second search are then searched against all human reviewed proteins and finally PDB proteins. "Missing" proteins were also searched for sequence similarity against all mammalian proteins and all nonmammalian proteins with proteomic evidence database, for verification.

The Protannotator system then employs InterProScan to characterize "missing" proteins with high-quality annotations based on protein functional domains and motifs along with GO terms. Pathway mapping was then carried out using KOBAS. InterProScan results were processed using IPRStats[35] for compiling statistics from InterProScan results as well as visualization of the output information.

All annotation information was then uploaded to a static webpage for the scientific community to view or download, by chromosome, permitting different C-HPP research groups across the globe to search the information on "missing" proteins for their respective chromosomes.

■ RESULTS AND DISCUSSION

All 20 128 human proteins were sorted based on the availability of protein evidence. 3831 proteins (~19%) were identified as "missing" based on protein evidence level 2–4 (consistent with the recent C-HPP standard metrics table and available as Supporting Information: Table S1). The number of "missing" proteins across the human proteome is steadily decreasing due to the large-scale proteomic effort across the globe. In our previous study of hChr7, 170 proteins were reported as "missing" as compared with 186 "missing" proteins in the current study.

1. Sequential-BLAST Similarity Search

The first sequential-BLAST run, against reviewed nonhuman mammalian protein sequences with proteomic evidence, resulted in 1271 “missing” proteins (33.2%) with significant matches, with >50% identity and E values of 0 to 1×10^{-05} (available from Supporting Information: Table S1). All top hits were selected, and matches with sequence identity ≥50% are considered as significant for this study. The remaining 2560 proteins were then searched against nonhuman reviewed mammalian protein sequences, using BLAST, with significant results for 703 sequences (18.4% of the 3831 “missing” proteins, available from Supporting Information: Table S2). For these matches, E values ranged from 0 to 2.00×10^{-6}. The third BLAST search against reviewed human proteins reported matches for 564 sequences (14.7% of the 3831 “missing” proteins, Supporting Information: Table S3), with E values ranging from 0 to 2.00×10^{-6}. In the final round of BLAST search, the remaining 1857 sequences were searched against PDB to check for similarity against sequences of known protein structures, with no match identified. Mapping all the “missing” proteins to the validation databases (results not shown) comprising all mammalian proteins and all nonmammalian proteins with experimental protein evidence also yielded a null result. The results from the sequential BLAST searches are shown in Table 1. Compared with the 127 “missing” proteins annotated using the sequential BLAST strategy in our previous hChr7,[9] 134 have been annotated in the current study, possibly as a consequence of the larger BLAST search databases in the current analysis and also due to the increased number of “missing” proteins, according to C-HPP standard metrics table.

Table 1. Sequential BLAST Matches for Human “Missing” Proteins

chromosome	number of missing proteins	reviewed mammalian proteins with experimental evidence	reviewed mammalian proteins	reviewed human proteins
Chr1	412	79	89	55
Chr2	190	55	51	25
Chr3	179	50	43	30
Chr4	130	53	28	10
Chr5	160	69	28	14
Chr6	176	52	31	28
Chr7	186	78	35	21
Chr8	108	36	26	22
Chr9	162	45	35	32
Chr10	152	53	21	41
Chr11	354	89	47	25
Chr12	169	60	35	17
Chr13[a]	53	17	12	13
Chr14	105	22	17	10
Chr15	111	44	22	11
Chr16	139	49	24	31
Chr17	191	50	46	35
Chr18	44	19	6	8
Chr19	369	230	30	25
Chr20	96	24	27	23
Chr21	59	9	8	27
Chr22	87	29	14	20
ChrX	182	52	28	40
ChrY[a]	17	7	0	1

[a] neXtProt Chr13 has one more protein and ChrY one less protein that the C-HPP standard metrics table.

The organism-wise distribution of the first two rounds of BLAST matches is shown in Figure 2, with the largest number of homologues in mouse, followed by rat and cow. Primates are not well-represented in this study, unlike in our previous report on hChr7.[9]

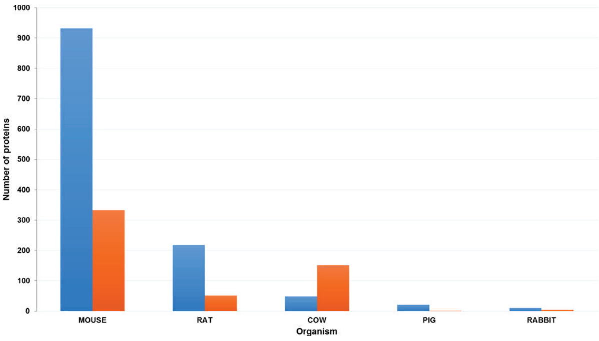


Figure 2. Significant BLAST hits grouped by organism. Blue bars represent the outcome of the first round of sequential BLAST against reviewed nonhuman mammalian proteins with experimental evidence, while the red bars represent the second round of BLAST against reviewed nonhuman mammalian proteins.

2. Functional Annotation

Functional annotation was carried out for all 3831 “missing” proteins, unlike the novel “missing” proteins alone in our previous study of hChr7.[9] InterProScan annotated 1945 “missing” proteins with GO annotation results in mapping of missing proteins to 2269 biological process (BP), 2059 cellular component (CC), and 3731 molecular function (MF) terms. Several GO terms such as *protein binding* and *membrane* reported in recent hChr7-centric proteomic analysis of human colon carcinoma cell lines[36] were also found among the hChr7 “missing” proteins. These 1945 “missing” proteins also mapped to 3019 domains, 4783 families, 162 repeats, 82 conserved sites, 9 binding sites, and 4 active sites. Recently published annotations of male specific chromosome Y proteins[37] were also reflected in chromosome Y “missing” protein mapping. DAZ proteins bind RNA in germ cells and are involved in primordial germ cell population maintenance.[38] The top 15 InterPro codes identified for the human “missing” proteins are shown in Table 2.

“Missing” proteins were also mapped to KEGG biochemical pathways using KOBAS, with 642 proteins annotated with pathway information. *Olfactory receptor* (IPR000725), the second InterPro hit, was listed as the top hit in the KEGG pathway mapping (*Olfactory transduction*). Out of 366 proteins mapped to the *Olfactory receptor family*, 360 were also mapped to the *G-protein-coupled receptor family* (IPR000276). Olfactory receptors were also reported recently in p13.2 and p13.3 regions of chromosome 17.[10] These receptors are associated with the biological process of *G-protein-coupled receptor signaling pathway* and the molecular function of *olfactory receptor activity*. G-protein-coupled receptor signaling pathway and olfactory receptor activity have been reported in genes clustered largely in a localized domain of chromosome 11.[39] *Zinc finger domains* (IPR013087, IPR001841, and IPR007087) comprise another important protein domain, in which zinc plays a structural role for the stability of the small domain. These protein domains are structurally diverse and are present among proteins responsible for a broad range of cellular functions, such as replication and

Table 2. Top 15 InterProScan Hits for Human "Missing" Proteins

InterPro code	description	number of missing proteins mapped	chromosome(s) mapped
IPR000276	G protein-coupled receptor, rhodopsin-like (family)	445	Chr10
IPR000725	olfactory receptor (family)	366	Chr10
IPR013087	zinc finger C2H2-type/integrase DNA-binding domain	129	Chr6
IPR001909	Krueppel-associated box (domain)	80	Chr6
IPR009057	homeodomain-like (domain)	68	Chr3
IPR001356	homeobox domain	62	Chr3
IPR002110	ankyrin repeat (repeat)	51	Chr8, Chr22
IPR007087	zinc finger, C2H2 (domain)	48	Chr7
IPR002126	cadherin (domain)	45	Chr4
IPR015919	cadherin-like (domain)	45	Chr4
IPR002494	high sulfur keratin associated protein (family)	32	Chr16
IPR001841	zinc finger, RING-type (domain)	31	Chr10
IPR015943	WD40/YVTN repeat-like-containing domain	29	Chr19
IPR011598	Myc-type, basic helix−loop−helix (bHLH) domain	29	Chr19
IPR011992	EF-hand domain pair (domain)	25	Chr19

repair, transcription and translation, metabolism and signaling, cell proliferation, and apoptosis.[40] The *homeobox* domain noted in our results (Table 2) was also recently reported in q21.32 region of chromosome 17.[10]

The sensory system was the main category of KEGG biochemical pathways reports for the "missing" proteins, with 390 proteins mapped to *Olfactory transduction* (372) and *Taste transduction* (18) pathways. Another 13 human "missing" proteins were mapped to pathways involved in Huntington's disease (HD), a neurodegenerative genetic disorder. This result indicates their involvement in human diseases, which needs further proteomic investigation. The top 10 KEGG pathway mappings are shown in Table 3. Recently, chromosome 19 genes/proteins have been related to 80 human diseases.[41] We mapped missing proteins from chromosome 19 to Alzheimer's, Parkinson's, and Huntington's diseases. The details of InterProScan and KEGG mapping are documented in the Supporting Information: Tables S4 and S5.

Table 3. Top Ten KEGG Pathways for Human "Missing" Proteins

pathway description	no. of proteins
olfactory transduction	372
neuroactive ligand−receptor interaction	70
metabolic pathways	66
taste transduction	18
GABAergic synapse	17
glutamatergic synapse	16
calcium signaling pathway	16
natural killer cell mediated cytotoxicity	14
retrograde endocannabinoid signaling	14
antigen processing and presentation	13
Huntington's disease	13

The functional annotation of the 3831 "missing" proteins is summarized in Figure 3, with 608 (15.9%) proteins having GO, InterPro domains as well as biochemical pathway annotations, 1337 (34.9%) proteins have InterPro domains and GO annotations alone, while 642 (16.8%) proteins have only KEGG biochemical pathway annotations. 1244 (32.5%) proteins could not be assigned any functional annotation with the currently available biological knowledge and may be considered novel.

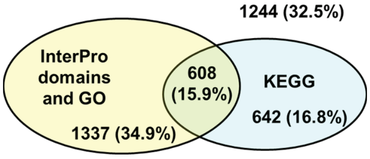


Figure 3. Summary of functional annotations for the "missing" proteins. Gene ontology (GO), functional domains/motif (InterPro) annotations were obtained for 1945 (50.8%) proteins. KEGG pathways (KEGG) annotations were obtained for 1250 (32.6%) proteins. 608 proteins had GO, Interpro domains, and KEGG pathway annotations. 1244 proteins remain unannotated.

3. *In Silico* Tryptic Digestion and ENCODE Proteogenomic Data

The Protein Digestion Simulator[28] was used to generate *in silico* proteotypic peptides, with trypsin selected as the proteolytic enzyme. Monoisotopic masses, pI, and hydrophobicity values for the tryptic peptides were computed (results not shown). These digested peptides were matched against the high-quality proteogenomic peptide data (58 601 records, based on the criteria described elsewhere[23] using the Peppy software[22]) from the ENCODE project[21] for proteomic evidence for the entire set of "missing" proteins. We found 245 peptides that matched the ENCODE data, with 1−44 peptides per protein. We have used the criteria of at least one or more peptide matching accurately (i.e., 100% identity) to the proteogenomic peptides, as we are matching protein sequences, to emulate the false positive discovery rate of Risk et al.,[22] who have set the threshold at >1 peptide matching to six-frame translations of genomic DNA. The peptides provide proteomic evidence of 107 "missing" proteins (with at least one peptide) for review and integration into the neXtProt chromosome summary lists. These peptides were found in 571 locations, with 316 in the positive orientation and 255 in the reverse orientation, that is, coded by the complementary strand. The genomic locations for two proteins (NX_Q9Y5G0 and NX_Q9Y5G1) on hChr5 could not be determined from neXtProt. The mapping results have been summarized in Table 4, and details of the mapping as well as the mapping regions are documented in Supporting Information: Table S6.

We have validated the ENCODE data mapping with neXtProt assigned genomic coordinates for each protein. Of the 571 locations, 202 matched the genomic coordinates

Table 4. Summary of ENCODE Data Mapping of Peptides from Human “Missing” Proteins

chromosome	number of proteins	number of peptides	positive	negative	number of peptides on both strands	total matched with neXtProt coding region	total unmatched with nextProt coding region
Chr1	7	17	17	18	2	17	18
Chr2	7	27	66	57	21	27	96
Chr3	2	2	1	1	0	2	0
Chr4	3	3	2	1	0	3	0
Chr5[a]	17	29	20	7	0	27	0
Chr6	9	44	72	55	24	3	124
Chr7	9	16	15	17	7	16	16
Chr8	1	1	1	1	1	1	1
Chr9	6	13	15	26	8	13	28
Chr10	5	8	21	3	2	8	16
Chr11	3	4	3	2	1	4	1
Chr12	5	7	4	5	0	7	2
Chr13	3	4	1	3	0	4	0
Chr14	1	1	1	0	0	1	0
Chr15	1	2	2	0	0	2	0
Chr16	3	27	50	27	26	27	50
Chr17	5	8	11	8	4	10	9
Chr18	1	1	0	1	0	1	0
Chr19	6	11	4	10	1	12	2
Chr20	1	1	1	0	0	1	0
Chr21	0	0	0	0	0	0	0
Chr22	3	4	0	1	0	1	0
ChrX	8	14	8	12	1	14	6
ChrY	1	1	1	0	0	1	0
Total	107	245	316	255	98	202	369

[a] For NX_Q9Y5G0 and NX_Q9Y5G1, the genomic location on chromosome 5 is not available in the neXtProt database, although ENCODE proteogenomic data mapped to these proteins.

assigned by neXtProt for 107 proteins, as two proteins (NX_Q9Y5G0, NX_Q9Y5G1) on Chr5 without genomic coordinates in neXtProt were excluded. 98 peptides were found in both orientations, that is, the coding regions covering one ENCODE peptide as well as at least one reverse peptide, requiring further experimental validation. These peptides can be used for the synthesis of antibodies and for future in vitro studies that could lead to proteomic identification of the proteins.

4. Membrane Proteomics “Missing” Protein List Comparison

We have compared the 3831 “missing” proteins with the 851 “missing” proteins identified by Muraoka et al.[24] and found that a total of 17 proteins have been provided proteomic evidence from this study. The chromosome-wise results are presented in Table 5.

Furthermore, we have summarized the proteomic evidence from the ENCODE project[21] and the membrane proteomic study of Muraoka et al.[24] (detailed in Supporting Information: Table S7). Two “missing” proteins (NX_P0CK97 and NX_Q9H0R5) have proteomic evidence from both experimental studies, with 105 “missing proteins uniquely supported by ENCODE data and 15 proteins by membrane proteomic data alone. In all, 122 (3.2%) “missing” proteins now have proteomic evidence

Table 5. Summary of Membrane Proteomic Evidence for Human “Missing” Proteins

chromosome	number of proteins	chromosome	number of proteins
Chr1	3	Chr11	2
Chr6	1	Chr12	3
Chr7	2	Chr16	1
Chr9	1	Chr17	1
Chr10	2	Chr19	1

■ CONCLUSIONS

We have compiled the chromosome-wise set proteins from the human proteome for 3831 “missing” proteins, as listed in the C-HPP standard metrics table, for *in silico* analysis and annotation. Using selected high-quality protein databases, similarity searches running BLAST sequentially identified homologues with experimental evidence for 33.2% of the “missing” proteins, with another 18.4% mapping to reviewed nonhuman mammalian proteins. As our study has used existing information to identify homologous proteins, further experimental work is required to confirm the existence of the proteins that are not identified by Protannotator, which is outside the scope of the work. However, with homologues identified from higher mammals, these proteins have a high probability of acquiring experimental evidence in the near future. Using a suite of bioinformatics tools, we have assigned putative biological functions in terms of GO and domain/motif signatures for 1945 (50.8%) and biochemical pathways for 1250 (32.6%) of the “missing” sequences. Despite the current level of biological knowledge in the databases, 1244 sequences (32.5%) remain unannotated by our sequential BLAST and computational annotation strategy.

By using a combination of computational tools, close to 50% of “missing” proteins in the human genome have been assigned putative biological functionality, providing valuable clues for experimental validation assays. *In silico* tryptic digestion generated proteotypic peptides with which we were able to ascribe proteomic evidence for 107 (2.8%), thereby linking genomics and proteomics via bioinformatics. Additionally, proteomic evidence for another 15 "missing" proteins was provided by the recent membrane protein study of Muraoka et al.,[24] bringing the total of neXtProt "missing" proteins with proteomic evidence to 122. Our results, available freely through Protannotator, will benefit proteomic identification of the human "missing" proteome. The computational approach we have described is generic and can be used to annotate the proteome of any novel organism, such as the black Périgord truffle.[20] We plan to further automate the system (wherever possible) and provide updated information via the Protannotator web portal, to track proteomic or bioinformatics evidence for the unannotated set of "missing" proteins.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Table S1: Significant BLAST hits against nonhuman reviewed mammalian proteins with experimental evidence. Table S2: Significant BLAST hits against nonhuman reviewed mammalian proteins. Table S3: Significant BLAST hits against human reviewed proteins. Table S4: Functional annotations of human missing proteins using InterProScan mapping. Table S5: KEGG pathway mapping of "missing" proteins. Table S6: ENCODE proteogenomic peptides mapping to the human "missing" proteins. Table S7: Summary of proteomic evidence for human "missing" proteins from ENCODE proteogenomic and membrane proteomic data. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: shoba.ranganathan@mq.edu.au. Phone: +612-9850-6262. Fax: +612-9850-8313.

Author Contributions

#M.T.I. and G.G. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Financial support was provided by a Macquarie University Research Excellence Scholarship (MQRES) to M.T.I. in the Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, Australia

■ ABBREVIATIONS

C-HPP, Chromosome-centric Human Proteome Project; Chr, chromosome; EC, enzyme code; GO, gene ontology; KOBAS, KEGG Orthology-Based Annotation System; KEGG, Kyoto Encyclopedia of Genes and Genomes; NET, normalized elution time; NR, nonredundant; SCX, strong cation exchange

■ REFERENCES

(1) Stein, L. Genome annotation: from sequence to biology. *Nat. Rev. Genet.* **2001**, *2*, 493−503.

(2) Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K.; Beretta, L.; Bergeron, J.; Borchers, C. H.; Corthals, G. L.; Costello, C. E.; Deutsch, E. W.; Domon, B.; Hancock, W.; He, F.; Hochstrasser, D.; Marko-Varga, G.; Salekdeh, G. H.; Sechi, S.; Snyder, M.; Srivastava, S.; Uhlen, M.; Wu, C. H.; Yamamoto, T.; Paik, Y. K.; Omenn, G. S. The human proteome project: current state and future direction. *Mol Cell Proteomics* **2011**, *10* (7), M111 009993.

(3) Hancock, W.; Omenn, G.; Legrain, P.; Paik, Y. K. Proteomics, human proteome project, and chromosomes. *J. Proteome Res.* **2011**, *10*, 210.

(4) Paik, Y. K.; Jeong, S. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Cho, S. Y.; Lee, H. J.; Na, K.; Choi, E. Y.; Yan, F.; Zhang, F.; Zhang, Y.; Snyder, M.; Cheng, Y.; Chen, R.; Marko-Varga, G.; Deutsch, E. W.; Kim, H.; Kwon, J. Y.; Aebersold, R.; Bairoch, A.; Taylor, A. D.; Kim, K. Y.; Lee, E. Y.; Hochstrasser, D.; Legrain, P.; Hancock, W. S. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* **2012**, *30* (3), 221−3.

(5) Paik, Y. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Marko-Varga, G.; Aebersold, R.; Bairoch, A.; Yamamoto, T.; Legrain, P.; Lee, H. J.; Na, K.; Jeong, S. K.; He, F.; Binz, P. A.; Nishimura, T.; Keown, P.; Baker, M. S.; Yoo, J. S.; Garin, J.; Archakov, A.; Bergeron, J.; Salekdeh, G. H.; Hancock, W. S. Standard guidelines for the chromosome-centric human proteome project. *J. Proteome Res.* **2012**, *11* (4), 2005−13.

(6) Marko-Varga, G.; Omenn, G. S.; Paik, Y. K.; Hancock, W. S. A first step toward completion of a genome-wide characterization of the human proteome. *J. Proteome Res.* **2013**, *12* (1), 1−5.

(7) Baker, M. S. Building the 'practical' human proteome project - the next big thing in basic and clinical proteomics. *Curr. Opin. Mol. Ther.* **2009**, *11* (6), 600−2.

(8) Gaudet, P.; Argoud-Puy, G.; Cusin, I.; Duek, P.; Evalet, O.; Gateau, A.; Gleizes, A.; Pereira, M.; Zahn-Zabal, M.; Zwahlen, C.; Bairoch, A.; Lane, L. neXtProt: organizing protein knowledge in the context of human proteome projects. *J. Proteome Res.* **2013**, *12* (1), 293−8.

(9) Ranganathan, S.; Khan, J. M.; Garg, G.; Baker, M. S. Functional Annotation of the Human Chromosome 7 "Missing" Proteins: A Bioinformatics Approach. *J. Proteome Res.* **2013**, 2504−10.

(10) Liu, S.; Im, H.; Bairoch, A.; Cristofanilli, M.; Chen, R.; Deutsch, E. W.; Dalton, S.; Fenyo, D.; Fanayan, S.; Gates, C.; Gaudet, P.; Hincapie, M.; Hanash, S.; Kim, H.; Jeong, S. K.; Lundberg, E.; Mias, G.; Menon, R.; Mu, Z.; Nice, E.; Paik, Y. K.; Uhlen, M.; Wells, L.; Wu, S. L.; Yan, F.; Zhang, F.; Zhang, Y.; Snyder, M.; Omenn, G. S.; Beavis, R. C.; Hancock, W. S. A chromosome-centric human proteome project (C-HPP) to characterize the sets of proteins encoded in chromosome 17. *J. Proteome Res.* **2013**, *12* (1), 45−57.

(11) Farrah, T.; Deutsch, E. W.; Hoopmann, M. R.; Hallows, J. L.; Sun, Z.; Huang, C. Y.; Moritz, R. L. The state of the human proteome in 2012 as viewed through PeptideAtlas. *J. Proteome Res.* **2013**, *12* (1), 162−71.

(12) UniProt Consortium.. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2012**, *40* (Database issue), D71−75.

(13) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403−410.

(14) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389−402.

(15) Nagaraj, S. H.; Gasser, R. B.; Ranganathan, S. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief Bioinf.* **2007**, *8* (1), 6−21.

(16) Ranganathan, S.; Menon, R.; Gasser, R. B. Advanced *in silico* analysis of expressed sequence tag (EST) data for parasitic nematodes of major socio-economic importance–fundamental insights toward biotechnological outcomes. *Biotechnol. Adv.* **2009**, *27* (4), 439−448.

(17) Garg, G.; Ranganathan, S. *In silico* secretome analysis approach using next generation sequencing transcriptomic data. *BMC Genomics* **2011**, *12* (Suppl 3), S14.

(18) Menon, R.; Garg, G.; Gasser, R. B.; Ranganathan, S. TranSeqAnnotator: large-scale analysis of transcriptomic data. *BMC Bioinf.* **2012**, *13* (Suppl17), S24.

(19) Garg, G.; Ranganathan, S. High-throughput functional annotation and data mining of fungal genomes to identify therapeutic targets. In *Laboratory Protocols in Fungal Biology: Current Methods in Fungal Biology*, Gupta, V. K.; Tuohy, M. G.; Ayyachamy, M.; Turner, K. M.; O'Donovan, A., Eds.; Springer: New York, 2013.

(20) Islam, M. T.; Mohamedali, A.; Garg, G.; Khan, J. M.; Gorse, A. D.; Parsons, J.; Marshall, P.; Ranganathan, S.; Baker, M. S. Unlocking the puzzling biology of the black Périgord truffle *Tuber melanosporum*. *J. Proteome Res.* **2013**, *12* (12), 5349−5356.

(21) Khatun, J.; Yu, Y.; Wrobel, J. A.; Risk, B. A.; Gunawardena, H. P.; Secrest, A.; Spitzer, W. J.; Xie, L.; Wang, L.; Chen, X.; Giddings, M. C. Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions. *BMC Genomics* **2013**, *14*, 141.

(22) Risk, B. A.; Spitzer, W. J.; Giddings, M. C. Peppy: proteogenomic search software. *J. Proteome Res.* **2013**, *12* (6), 3019−25.

(23) Risk, B. A.; Edwards, N. J.; Giddings, M. C. A Peptide-Spectrum Scoring System Based on Ion Alignment, Intensity, and Pair Probabilities. *J. Proteome Res.* **2013**, *12* (9), 4240−4247.

(24) Muraoka, S.; Kume, H.; Adachi, J.; Shiromizu, T.; Watanabe, S.; Masuda, T.; Ishihama, Y.; Tomonaga, T. In-depth membrane proteomic study of breast cancer tissues for the generation of a chromosome-based protein list. *J. Proteome Res.* **2013**, *12* (1), 208−13.

(25) Protein Data Bank (PDB): http://www.rcsb.org/pdb/.

(26) Boekhorst, J.; Snel, B. Identification of homologs in insignificant blast hits by exploiting extrinsic gene properties. *BMC Bioinf.* **2007**, *8*, 356.

(27) Sander, C.; Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **1991**, *9* (1), 56−68.

(28) Quevillon, E.; Silventoinen, V.; Pillai, S.; Harte, N.; Mulder, N.; Apweiler, R.; Lopez, R. InterProScan: protein domains identifier. *Nucleic Acids Res.* **2005**, *33* (Web Server issue), W116−20.

(29) Hunter, S.; Jones, P.; Mitchell, A.; Apweiler, R.; Attwood, T. K.; Bateman, A.; Bernard, T.; Binns, D.; Bork, P.; Burge, S.; de Castro, E.; Coggill, P.; Corbett, M.; Das, U.; Daugherty, L.; Duquenne, L.; Finn, R. D.; Fraser, M.; Gough, J.; Haft, D.; Hulo, N.; Kahn, D.; Kelly, E.; Letunic, I.; Lonsdale, D.; Lopez, R.; Madera, M.; Maslen, J.; McAnulla, C.; McDowall, J.; McMenamin, C.; Mi, H.; Mutowo-Muellenet, P.; Mulder, N.; Natale, D.; Orengo, C.; Pesseat, S.; Punta, M.; Quinn, A. F.; Rivoire, C.; Sangrador-Vegas, A.; Selengut, J. D.; Sigrist, C. J. A.; Scheremetjew, M.; Tate, J.; Thimmajanarthanan, M.; Thomas, P. D.; Wu, C. H.; Yeats, C.; Yong, S.-Y. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* **2012**, *40* (D1), D306−D312.

(30) Xie, C.; Mao, X.; Huang, J.; Ding, Y.; Wu, J.; Dong, S.; Kong, L.; Gao, G.; Li, C. Y.; Wei, L. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* **2011**, *39* (Web Server issue), W316−22.

(31) Young, N. D.; Jex, A. R.; Li, B.; Liu, S.; Yang, L.; Xiong, Z.; Li, Y.; Cantacessi, C.; Hall, R. S.; Xu, X.; Chen, F.; Wu, X.; Zerlotini, A.; Oliveira, G.; Hofmann, A.; Zhang, G.; Fang, X.; Kang, Y.; Campbell, B. E.; Loukas, A.; Ranganathan, S.; Rollinson, D.; Rinaldi, G.; Brindley, P. J.; Yang, H.; Wang, J.; Wang, J.; Gasser, R. B. Whole-genome sequence of Schistosoma haematobium. *Nat. Genet.* **2012**, *44* (2), 221−5.

(32) Jex, A. R.; Liu, S.; Li, B.; Young, N. D.; Hall, R. S.; Li, Y.; Yang, L.; Zeng, N.; Xu, X.; Xiong, Z.; Chen, F.; Wu, X.; Zhang, G.; Fang, X.; Kang, Y.; Anderson, G. A.; Harris, T. W.; Campbell, B. E.; Vlaminck, J.; Wang, T.; Cantacessi, C.; Schwarz, E. M.; Ranganathan, S.; Geldhof, P.; Nejsum, P.; Sternberg, P. W.; Yang, H.; Wang, J.; Wang, J.; Gasser, R. B. Ascaris suum draft genome. *Nature* **2011**, *479* (7374), 529−33.

(33) Protein Digestion Simulator: http://omics.pnl.gov/.

(34) Hopp, T. P.; Woods, K. R. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 3824−8.

(35) Kelly, R.; Vincent, D.; Friedberg, I. IPRStats: visualization of the functional potential of an InterProScan run. *BMC Bioinf.* **2010**, *11* (Suppl 12), S13.

(36) Fanayan, S.; Smith, J. T.; Sethi, M. K.; Cantor, D.; Goode, R.; Simpson, R. J.; Baker, M. S.; Hancock, W. S.; Nice, E. Chromosome 7-centric analysis of proteomics data from a panel of human colon carcinoma cell lines. *J. Proteome Res.* **2013**, *12* (1), 89−96.

(37) Jangravi, Z.; Alikhani, M.; Arefnezhad, B.; Sharifi Tabar, M.; Taleahmad, S.; Karamzadeh, R.; Jadaliha, M.; Mousavi, S. A.; Ahmadi Rastegar, D.; Parsamatin, P.; Vakilian, H.; Mirshahvaladi, S.; Sabbaghian, M.; Mohseni Meybodi, A.; Mirzaei, M.; Shahhoseini, M.; Ebrahimi, M.; Piryaei, A.; Moosavi-Movahedi, A. A.; Haynes, P. A.; Goodchild, A. K.; Nasr-Esfahani, M. H.; Jabbari, E.; Baharvand, H.; Sedighi Gilani, M. A.; Gourabi, H.; Salekdeh, G. H. A fresh look at the male-specific region of the human Y chromosome. *J. Proteome Res.* **2013**, *12* (1), 6−22.

(38) Reynolds, N.; Cooke, H. Role of the DAZ genes in male fertility. *Reprod. Biomed. Online* **2005**, *10*, 72.

(39) Kwon, K. H.; Kim, J. Y.; Kim, S. Y.; Min, H. K.; Lee, H. J.; Ji, I. J.; Kang, T.; Park, G. W.; An, H. J.; Lee, B.; Ravid, R.; Ferrer, I.; Chung, C. K.; Paik, Y. K.; Hancock, W. S.; Park, Y. M.; Yoo, J. S. Chromosome 11-centric human proteome analysis of human brain hippocampus tissue. *J. Proteome Res.* **2013**, *12* (1), 97−105.

(40) Krishna, S. S.; Majumdar, I.; Grishin, N. V. Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res.* **2003**, Jan 15; *31*(2):532-50.

(41) Nilsson, C. L.; Berven, F.; Selheim, F.; Liu, H.; Moskal, J. R.; Kroes, R. A.; Sulman, E. P.; Conrad, C. A.; Lang, F. F.; Andren, P. E.; Nilsson, A.; Carlsohn, E.; Lilja, H.; Malm, J.; Fenyo, D.; Subramaniyam, D.; Wang, X.; Gonzales-Gonzales, M.; Dasilva, N.; Diez, P.; Fuentes, M.; Vegvari, A.; Sjodin, K.; Welinder, C.; Laurell, T.; Fehniger, T. E.; Lindberg, H.; Rezeli, M.; Edula, G.; Hober, S.; Marko-Varga, G. Chromosome 19 annotations with disease speciation: a first report from the Global Research Consortium. *J. Proteome Res.* **2013**, *12* (1), 135−50.

4.3 Conclusions

We devised a set of sequential blast search databases (from the mammalian kingdom, described in the publication) tailored for human proteome based on our annotation framework. We developed ProtAnnotator, a semi-automated pipeline using these datasets to perform the functional annotation for these sequences. We then obtained the missing protein sequences for all chromosomes from UniProt and processed them through this pipeline. Our workflow identified homologues for 66.2% and functional annotation for 50.8% of these human 'missing' proteins. As the C-HPP Project is a worldwide collaboration, we developed a web portal to share the updated annotation results freely to the community. However, neXtProt updates the PE status with frequent releases, meaning an automated pipeline is needed to (i) annotate any newly listed proteins (ii) provide up to date annotations in line with database and software update as these can lead to new annotations and enhance our knowledge about the proteins. Besides the C-HPP initiative, an automated platform is also necessary to annotate vast swathes of newly sequenced or novel genomes or targeted annotation studies.

Chapter 5: ProtAnnotator 2.0: An automated pipeline for *in silico* protein functional annotation

5.1 Summary

As discussed in the previous chapter, the human missing protein sequences are regularly updated with proteins being added or removed from the list. So, there is a clear need for an automated annotation system to provide updated annotation information to the C-HPP community. At the same time, the low cost, but high throughput next generation sequencing technology is allowing the community to publish new genomes at a rapid rate. That makes it practically impossible to annotate these proteomes experimentally. High accuracy and high throughput annotation strategies are needed to keep up with this influx of the new genomes.

In this study, we extended our ProtAnnotator platform to incorporate a more generic, easy to use automated annotation platform. The cloud-based platform offers end users to execute the previously mentioned workflows (Chapter 3 and Chapter 4) by simply uploading a sequence file in FAST-All (FASTA) format. Most importantly, it offers users an authenticated secure portal to create their own annotation pipeline using their own reference sequence databases to match their experiments. It provides an industry standard high-speed data transfer mechanism to upload data into the environment with latest software and databases to perform the annotation (detailed in Publication 3).

5.2 Publication 3

ProtAnnotator 2.0: An automated pipeline for *in silico* protein functional annotation

Mohammad T. Islam[1]*, Abidali Mohamedali*[1]*, Shoba Ranganathan*[1*]

[1]Department of Chemistry and Biomolecular Sciences, Faculty of Science and engineering, Macquarie University, Sydney, Australia.

[*]Corresponding author

KEYWORDS: Functional annotation, black Périgord truffle, Human Proteome Project missing proteins, sequential BLAST, proteotypic peptides, proteogenomics

ABSTRACT

The functional annotation of a proteome uncovers biological functions of all its proteins. The rapid growth in protein sequence databases provides an excellent opportunity to predict putative functions of the unknown or relatively less studied proteins by mining the knowledge from well-studied proteins from these databases. Re-annotation has been recognized to be efficacious but tedious. Various constantly evolving computation tools and databases are available for annotation but have not been cohesively used in an automated fashion to streamline annotation. Here we present ProtAnnotator 2.0, an updated version of our previously published semi-automated *in silico* functional annotation that automates the entire process under a single web platform. The platform is underpinned by several freely available annotation tools and allows users to design custom pipelines and import their sequences databases for more accurate and targeted studies according to their needs.

INTRODUCTION

Annotation of proteins or genes is a critical step in elucidating the biological relevance of any biomolecule. This annotation can be at a nucleotide, protein or process level, with the process or functional annotation usually the most useful [1]. Experimental validation and annotation of gene function accounts for less than 1% of annotated proteins [2] and therefore there exist over 126 methods of predicting functional annotations *in silico* [3]. This is particularly relevant in the context of next generation sequencing where multiple new eukaryotic species are being sequenced at an increasing frequency [4] with increasingly distant phylogenetic divergence. This rapid rise in full genome sequencing has posed numerous challenges aside from the data challenges with annotation of proteins using older methodologies becoming limited.

The most common bioinformatics based annotation methodologies rely on evaluating sequence or structural homology search to a single database [5] with the premise that proteins with high sequence similarity will have similar functions. This approach is sufficient where genomes are phylogenetically close, or abundant experimental data is available. Often though this approach in itself is practically relatively limited for numerous reasons and hence, a significant development was, in addition to sequence homology, a search against protein domain family resources (such as Pfam [2, 5]. Although this approach circumvented many of the pitfalls of new genome annotation, some problems such as bias in functional annotations and mis-annotations still remain [6].

In proteomics, the *in silico* representation of proteins is almost wholly reliant on some form of genomic database or knowledgebase such as UniProt [7]. One of the standard methods is to align query sequences against a target sequence database to identify similar sequences. Although the UniProt non-redundant (NR) database has been widely used for this purpose, it is important to use targeted high quality (such as experimentally validated sequences, reviewed sequences) and or closely related database for several reasons. Searching against NR database may lead to matches against un-reviewed or unannotated proteins [8] that may not be useful for downstream annotations. Experimentally validated and or reviewed protein sequences from closely related species yield more accurately identified homologues for novel and/or less studied or complex proteomes. For example, the less studied Truffle fruiting body is believed to contain bacteria, yeasts, and filamentous fungi [9]. So, a targeted annotation strategy is required to understand its complex biology. Although most of the sequences are annotated in the public databases, as new genome and proteomes are published there is a need to re-annotate proteins for specific studies. The updates and new developments of methods, database, software, and tools necessitate on demand annotation platform, as re-annotation of the entire repository may not always be practical.

In a quest to annotate the 'missing' proteins from the human proteome, we had developed a semi-automated protein annotation pipeline, which integrated not only sequence homology (sequential BLAST approach) and protein family resources, but also gene ontology and biochemical pathways [8, 10, 11]. This novel approach tapped into novel datasets and annotating proteins based on the best available evidence.

One of the most significant outcomes was the uncovering of a range of 'missing' proteins [12] as part of HUPO's human proteome project (HPP) objective of characterizing the functions of all proteins in the human [13]. A wider ranging and interesting use of the methodology was the ability to annotate the proteins derived from the fruiting body of the black Périgord truffle [10]. This application and the clear gap in the functional annotation of proteins from other species led us to develop an updated version of the first iteration of the tool. Here we present a fully automated version 2.0 of ProtAnnotator (ProtAnnotator 2.0), the key features of which include the ability of a user to input their sequences to annotate proteins based on original pipeline published, and to seamlessly develop their own unique pipelines based on user-supplied databases. This automation will not only assist in efficient annotation of new genomes, it will significantly assist in re-annotation efforts. We demonstrate here the re-annotation of our original truffle data to arrive at significantly different results than just a few years ago.

MATERIALS AND METHODS

1. Overview

ProtAnnotator 2.0 is a web platform with integrated annotation pipelines based on our previously published method and workflows [8, 10, 11]. The core of this is the annotation workflow. The platform offers two types of workflows, public and private (i.e. authenticated) workflows. The public workflows are our previously published workflows [8, 10, 11] the 'sequential BLAST' databases for these workflows are fixed (periodically updated) and users can upload their protein sequences in FASTA format to run any of the available workflows without an account.

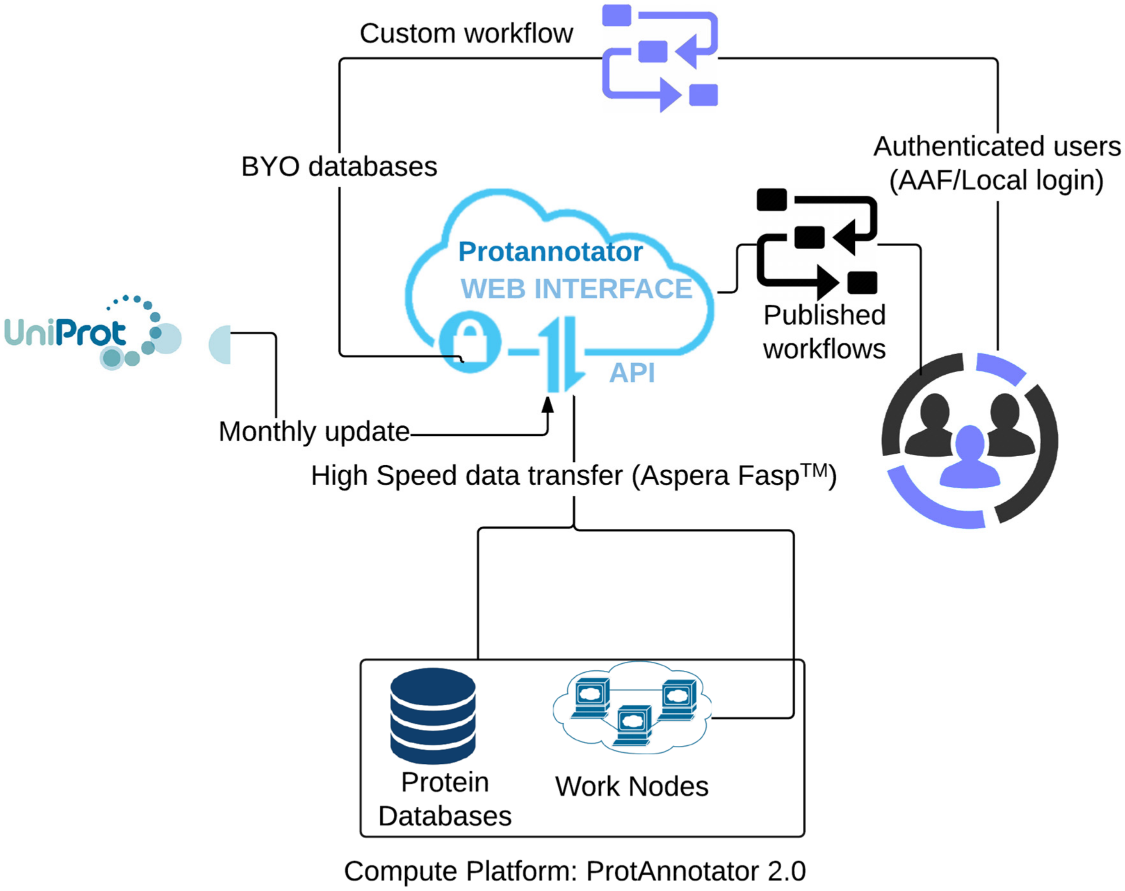


Figure 1. Top-level architecture of ProtAnnotator 2.0.

These workflows are provided on an *as is* basis, the search databases are updated monthly but it does not allow users to bring their own search databases. The private workflows are dynamic, and allow researchers to bring their own (BYO) databases and build custom sequential BLAST pipelines. Users are required to follow a simple registration process to be able to run the dynamic private workflows. Both public and private workflows offer users to identify domain/motif analyses, GO annotation and pathway mapping using the InterProScan [14] annotation tool. The basic flow (as depicted in Figure 1) is, the user selects a workflow published or private (upon successful registration), the user then uploads the query sequence file and builds the sequential BLAST [15] pipeline by uploading database files for each BLAST run (for private BYO database workflow only, see Figure 2), then selects the annotation options and submits the job. The user can visit the public portal to run a published workflow by adding the contact details and uploading a sequence file. The BLAST databases for published workflows are updated monthly. To design a custom workflow, the user needs to login to the portal, then create a workflow, and upload all BLAST databases. Files are uploaded using Aspera Fasp^{TM} protocol. A work node then processes jobs.

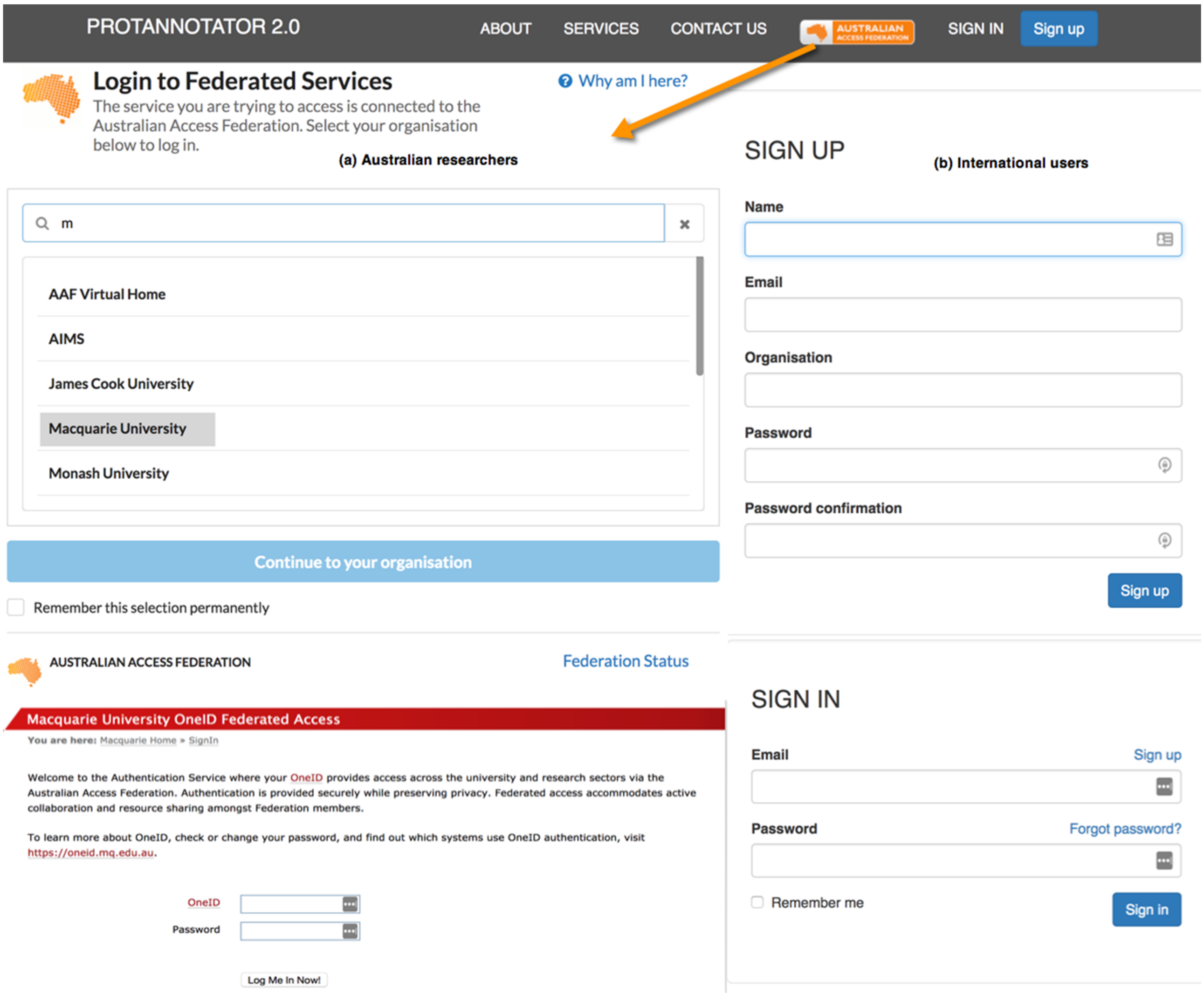
Once the job is submitted, the user receives a confirmation email with the job ID and selected parameters with the job being sent to a queue. An available ProtAnnotator work node then picks up the job, processes it and upon completion, sends an email to the user to allow result download over an Aspera connection.

2. Cloud platform

The web interface is developed using the Ruby on Rails open source software. The annotation pipeline along with all pre and post-processing steps are developed in Python, which integrates the BLAST sequential alignment tool, and InterProScan annotation tool to create cohesive annotation pipeline. The web platform, software tool, databases and associated compute work nodes are hosted on an OpenStack research cloud platform [16]. The webserver is build on Ubuntu (16.04 LTS x86_64) and the work nodes are built on (CentOS 6.7 (Final) x86_64) Linux operating system with open-source Sun Grid Engine (SGE) scheduler to manage the compute resources efficiently. The pipeline uses the Distributed Resource Management Application API (DRMAA) [17] bindings for Python to interact with SGE scheduler to process jobs. Each work node is built with 16 CPUs and 64 gigabytes of memory. Processed results are kept for two weeks after the job is completed and raw data files are deleted after successful completion of a job to maximise the utilization of the resource. The system is designed in a way that additional work nodes can be added to manage excessive workloads without code modifications.

3. Authentication and authorization

All users must be registered and authorised by an administrator to run the BYO database workflows. The ProtAnnotator 2.0 system is designed to provide simple access to the researchers as well as minimizing the number of accounts managed by the system (where possible). As such it offers two types of authentication model for end users (see Figure 2). The system is integrated with the Australian Access Federation (AAF)[18], a leading identity broker that offers federated identity management service for the education and research services. The researchers from all Australian universities and leading research organization can log into the system using their institutional login accounts. The users are provisioned in the ProtAnnotator system upon first successful login (Just-in-Time Provisioning). The international users are required to register using the sign-up service of ProtAnnotator. In both cases, users must be authorised by an administrator to access and create any workflow. When a user logs in for the first time (using AAF) or sign-up using the form, an email is sent to the administrator to approve the request. Protannotator applies a role based authorisation model, and workflows are associated with roles. An administrator can assign a role using the admin interface. Once an appropriate role is assigned to the user, the system sends an email to the user. The user can then log into the system to create and execute workflows.



The service you are trying to access is connected to the Australian Access Federation. Select your organisation below to log in.

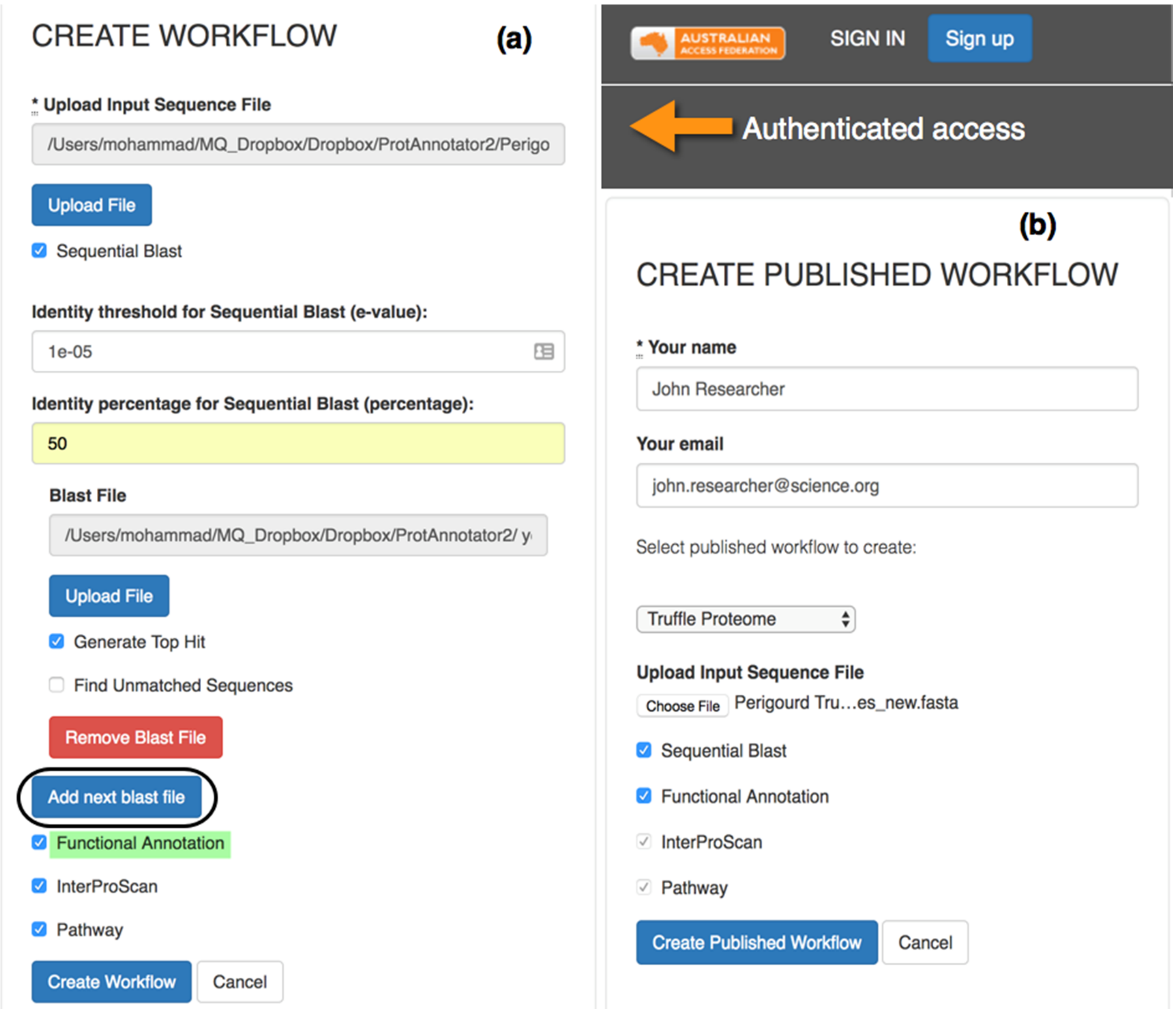
Welcome to the Authentication Service where your OneID provides access across the university and research sectors via the Australian Access Federation. Authentication is provided securely while preserving privacy. Federated access accommodates active collaboration and resource sharing amongst Federation members.

To learn more about OneID, check or change your password, and find out which systems use OneID authentication, visit https://oneid.mq.edu.au.

Figure 2. Proannotator2 authentication methods. (a) Australian researchers first select the AAF button, then select their institution and use their university credential to access the system. **(b)** International users first sign up using the form then uses the non-AAF log in option to access the system.

3. Workflow and pipeline

The basic concept of the workflow is to offer a sequential sequence similarity search against a set of user-defined databases as well as identifying domain/motif analyses, GO annotation and pathway mapping for the same. The query sequences are searched against the first database (user defined), then the unmatched sequences are used as the input for the next round of similarity search and so on. The pipeline currently uses BLAST (version 2.6.0+) for sequence similarity search. End users can select the E-value and percentage identity as the cut-off value for the sequential BLAST (default value is 1e-05, and 50% respectively).



CREATE WORKFLOW

* Upload Input Sequence File

/Users/mohammad/MQ_Dropbox/Dropbox/ProtAnnotator2/Perigo

☑ Sequential Blast

Identity threshold for Sequential Blast (e-value):

1e-05

Identity percentage for Sequential Blast (percentage):

50

Blast File

/Users/mohammad/MQ_Dropbox/Dropbox/ProtAnnotator2/ y

☑ Generate Top Hit

☐ Find Unmatched Sequences

☑ Functional Annotation

☑ InterProScan

☑ Pathway

CREATE PUBLISHED WORKFLOW

* Your name

John Researcher

Your email

john.researcher@science.org

Select published workflow to create:

Truffle Proteome

Upload Input Sequence File

Choose File Perigourd Tru…es_new.fasta

☑ Sequential Blast

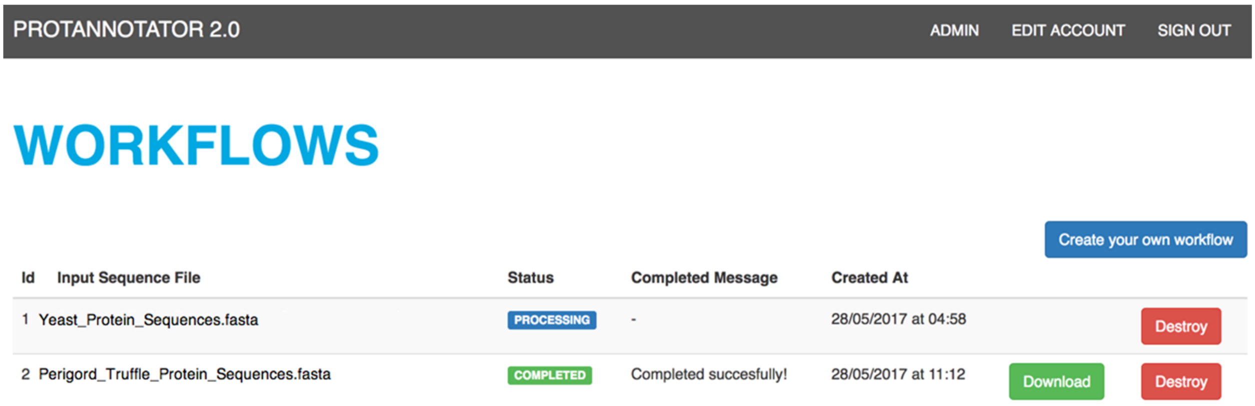
☑ Functional Annotation

✓ InterProScan

✓ Pathway

Figure 3. User interface for public (published) and private (BYO database) workflows. (a) Published workflow- Users can select a published workflow and simply upload a protein sequence file for annotation. **(b) Private (BYO database) workflow** – users need to log into the system, and then create a workflow. Multiple search database files can be uploaded using the *Add next BLAST file* option.

InterProScan (InterPro 63.0 and InterProScan 5.24-63.0) is used for the functional annotation and pathway analysis (cross-links to KEGG, MetaCyc, Reactome, and UniPathway) of the proteins. The registered users can also log into the system to check the status of the BYO workflows and download results (Figure 4).



WORKFLOWS

Id	Input Sequence File	Status	Completed Message	Created At		
1	Yeast_Protein_Sequences.fasta	PROCESSING	-	28/05/2017 at 04:58		Destroy
2	Perigord_Truffle_Protein_Sequences.fasta	COMPLETED	Completed succesfully!	28/05/2017 at 11:12	Download	Destroy

Figure 4. ProtAnnotator 2.0 job status and download page for authenticated users

5. Data sources

We previously applied our semi-automated ProtAnnotator pipeline to annotate human ‘missing’ proteins[8] and Black Périgord truffle proteins[10]. Since the ‘missing’ proteins sequences are regularly updated [19] we have used the static 12,771 non-redundant Black Périgord truffle (*T. melanosporum Vittad*) protein sequences from our previous proteomics and annotation study to demonstrate the platform and its applicability. These sequences were downloaded from the MycorWeb database [http://mycor.nancy.inra.fr/index.html] in FASTA format. We also download the following databases for sequence similarity search, reviewed yeast proteins with experimental protein evidence (7,740 sequences, previously 7,503), reviewed fungal proteins with experimental protein evidence (9,634 sequences, previously 9,450), reviewed fungal proteins (32,674 sequences, previously 31,031) and Protein Data Bank (PDB) protein (402,312 sequences, previously 236,604) sequences.

RESULTS AND DISCUSSION

To demonstrate the applicability of the automated ProtAnnotator 2.0 platform, we have reanalyzed the sequences we had previously reported using the newly updated platform against updated blast databases mentioned in the data source section. As shown in the previous section, the recent versions of all of the BLAST databases have more sequences than previously reported. The underlying tools for the pipeline are also updated. Our previous analysis pipeline used BLAST (+ 2.2.27) and InterproScan 4.8.

One of the biggest challenges for any cloud platform is the speed of the data transfer; users often spend more time moving data between local computers and the cloud instance than the actual time to process the jobs. Many standard transfer protocols such as FTP, GridFTP, RSYNC, WebDAVs, and HTTPS either do not provide resume transfer option or offers slow transfer speed especially with high latency (long distance transfer). So if a transfer is interrupted for any reason, users often need to transfer the whole file again. The sequences files, especially for the BYO database workflow, can be hundreds of gigabytes. To address this issue, ProtAnnotator is integrated with the Aspera FASP^{TM} transfer protocol using Aspera Connect Server API [20]. The Aspera FASP transfer protocol is hundreds of times faster than FTP and HTTP protocol [20]. Its client side connect plugin (free to users) provides options to end users to select their own bandwidth to transfer data. The ProtAnnotator cloud platform is hosted in a research cloud infrastructure[16] that utilises Australia's Academic and Research Network (AARNet) [21] with a 10Gbps connection. ProtAnnotator users are required to install the Aspera Connect Plug-in (freely available at http:// downloads.asperasoft.com/connect2/) for the first time. If the plug-in is not installed, the web interface will prompt the user to download the plugin with a link.

Despite the increase in the number of sequences in target databases, the sequential BLAST workflow identified homologous for 2,468 proteins (see Supplementary Table S1) as opposed to 2,486 proteins from our previous study (see Table 1). The ProtAnnotator 2.0 workflow identified homologues for 111 proteins that were not identified by our previous study. However, homologues for 129 previously identified were eliminated by the new analysis due to low percentage identity (refer to Supplementary Table S4). We used updated protein sequence databases for the similarity search, so the results are expected to vary. Besides, our previous study was conducted using BLAST+ 2.2.27. Since then BLAST tool underwent numerous improvements. In the latest iteration (version 2.6.0) the gapped alignment starting point was changed to minimize sub-optimal alignments [22]. Both of these factors may have contributed to the variation of the percentage identity scores for the 129 proteins. However, the pipeline identified a significantly large number of similarities (1,393 proteins) with the PDB sequences (see Supplementary Table S2) compared to our previous analysis of 101 proteins (see Table 1). The number of PDB sequences has almost doubled since our last study, so the increased identification was expected.

In our previous study, the annotation pipeline identified functional annotations and pathways for 20% of the black Périgord truffle sequences, however the updated pipeline identified at least one functional annotation or pathway for 82% proteins. This significant increase in annotations suggests that reference databases are constantly evolving and growing, it is imperative that before embarking on novel studies or interpreting results, that sequences be run through such pipelines and updated.

Table 1 Significant BLAST hits from the sequential BLAST pipeline. A comparative summary of previous bioinformatics analysis and recent bioinformatics analysis using ProtAnnotator 2.0

Description	Previously published results	ProtAnnotator 2.0 results with updated tools and databases
Total Number of proteins	12,771	12,771
Reviewed yeast protein sequences with experimental protein evidence	1,794	1,753
Reviewed fungal protein sequences with experimental protein evidence	109	131
Reviewed fungal protein sequences	583	584
Protein Data Bank (PDB)	101	1,393
Functional Annotation (GO, InterPro, Pathway)	2,587	10,511

Until now although recognized as a significant need [23], this has largely been a cumbersome process [23, 24] which meant that most scientists likely rely on previously annotated genomes. The efficacy of the disparate results obtained from re-annotation using updated databases and a more efficient methodology (keeping the same principles) is not more stark than as demonstrated in Table 2. The top 5 major pathways identified do not even resemble the previous results (with many pathways previously identified showing now much lower in the current analysis) and therefore the possibility of understanding and deciphering the biology and biochemistry of the organism is greatly improved.

The huge disparity could be that we used InterProScan 4.8 for our previous study. The software has been updated, and several new InterPro databases have also been added to InterProScan version 5 [14] . In addition to KEGG, the InterProScan version 5 also provides pathway identification from MetaCyc, Reactome, and UniPathway all of which contributed to a significantly high number of functional annotations compared to our previous study.

Table 2- Top five pathways identified by previous pipeline and current analysis using the updated platform.

Previously published pathways		**Pathways from ProtAnnotator 2.0 analyses**	
#Term	**Total Match**	**#Term**	**Total Match**
Metabolic pathways	961	Major pathway of rRNA processing in the nucleolus and cytosol	444
Pyrimidine metabolism	399	SRP-dependent cotranslational protein targeting to membrane	361
Biosynthesis of secondary metabolites	239	L13a-mediated translational silencing of Ceruloplasmin expression	336
Cell cycle - yeast	161	Formation of a pool of free 40S subunits	318
Meiosis - yeast	96	Selenocysteine synthesis	302

The efficacy of re-annotation of protein function was highlighted even further when analyzing the results from the GO analysis (Figure 5). We used REVIGO [25] treemap to identify the representative non-redundant subset of GO terms to show the differences in representative GO terms for the two analyses (Figure 5) and the differences between the top 5 pathways are shown in Table 2. We note that the RISC complex for example is highly overrepresented by our new annotation results. It is almost as if two separate organisms are being studied. It is clear that the basic biochemical processes in the *T. melanosporum* revolve around core cellular processes unlike our previous conclusions that suggest a greater preponderance on the processes that produce secondary metabolites. An in-depth analysis of the newly uncovered biology of the black truffle from this analysis, although beyond the scope of this study, needs to be carried out, especially in light of the new evidence. Indeed, a deeper analysis and appreciation of the role of complex biochemical pathways [26], biological functions to understand the enigmatic life cycle of this fungus its complex relationship not only with it symbiont host but also to other microorganisms [27] and the biology of its aroma [28] can greatly assist in not only developing cultivation programs [29] but in harnessing the organism's hereto unknown processes.

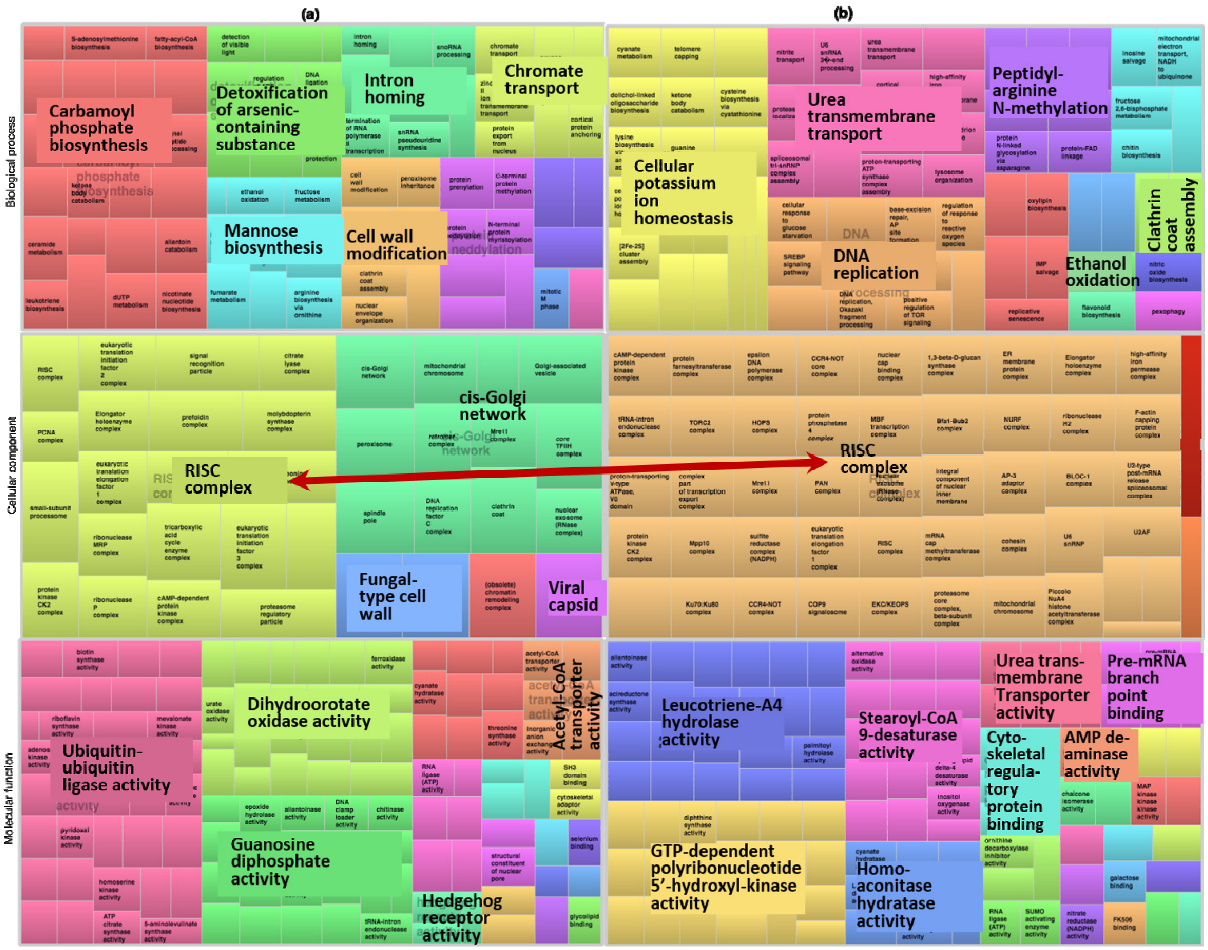


Figure 5. Treemap of the representative non-redundant GO terms, (a) results from the previous study, (b) annotations from the ProtAnnotator 2.0 pipeline

It is evident from the results that, the new analysis can provide us further insight into understanding the biology of this organism that was not possible from the previous tools and or databases.

CONCLUSIONS

The ProtAnnotator automated annotation platform is generic, and its custom workflow and 'bring your own database' feature can be used to annotate proteins from any novel or less studied organism. The reanalysis of the black Périgord truffle data using the platform demonstrates that the platform is capable of providing updated insights into protein functions and can be tailored for targeted annotation studies.

AVAILABILITY

The web platform can be accessed at https://protannotator.biolinfo.org

ACKNOWLEDGMENT

Financial support was provided by a Macquarie University Research Excellence Scholarship (MQRES) to MTI in the Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, Australia. We would like to thank Jayatheeban Soundararajan for his assistance in the web development of ProtAnnotator. Digital storage and computing was provided by Intersect Australia Space and Time, with acknowledgement to the Australian Government NCRIS Research Data Services (RDS) and National eResearch Collaboration Tools and Resources (Nectar).

Supporting Information. Tables S1-S4 as described in the main manuscript are provided as MS Excel files (*.xls). (provided on CD attached with the thesis). The Table descriptions are as follows:

Table S1: Significant BLAST hits from the sequential BLAST pipeline

Table S2: Significant BLAST hits from the similarity searches against the PDB database.

Table S3: Functional annotations and pathway mapping of black Périgord truffle proteins using InterProScan mapping.

Table S4: BLAST hits from the sequential similarity searches that were identified by the previous study but eliminated by ProtAnnotator 2.0 due to low percentage identify (<50%)

REFERENCES

1. Mudge JM, Harrow J: **The state of play in higher eukaryote gene annotation**. *Nat Rev Genet* 2016, **17**(12):758-772.

2. Das S, Orengo CA: **Protein function annotation using protein domain family resources**. *Methods* 2016, **93**:24-34.

3. Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor KM, Ben-Hur A *et al*: **An expanded evaluation of protein function prediction methods shows an improvement in accuracy**. *Genome Biol* 2016, **17**(1):184.

4. Schmidt B, Hildebrandt A: **Next-generation sequencing: big data meets high performance computing**. *Drug Discov Today* 2017, **22**(4):712-717.

5. Addou S, Rentzsch R, Lee D, Orengo CA: **Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer**. *J Mol Biol* 2009, **387**(2):416-430.

6. Schnoes AM, Ream DC, Thorman AW, Babbitt PC, Friedberg I: **Biases in the experimental annotations of protein function and their effect on our understanding of protein function space**. *PLoS Comput Biol* 2013, **9**(5):e1003063.

7. The UniProt C: **UniProt: the universal protein knowledgebase**. *Nucleic Acids Res* 2017, **45**(D1):D158-D169.

8. Islam MT, Garg G, Hancock WS, Risk BA, Baker MS, Ranganathan S: **Protannotator: a semiautomated pipeline for chromosome-wise functional annotation of the "missing" human proteome**. *J Proteome Res* 2014, **13**(1):76-83.

9. Benucci GM, Bonito GM: **The Truffle Microbiome: Species and Geography Effects on Bacteria Associated with Fruiting Bodies of Hypogeous Pezizales**. *Microb Ecol* 2016, **72**(1):4-8.

10. Islam MT, Mohamedali A, Garg G, Khan JM, Gorse AD, Parsons J, Marshall P, Ranganathan S, Baker MS: **Unlocking the puzzling biology of the black Perigord truffle Tuber melanosporum**. *J Proteome Res* 2013, **12**(12):5349-5356.

11. Ranganathan S, Khan JM, Garg G, Baker MS: **Functional annotation of the human chromosome 7 "missing" proteins: a bioinformatics approach**. *J Proteome Res* 2013, **12**(6):2504-2510.

12. Baker MS, Ahn SB, Mohamedali A, Islam MT, Cantor D, Verhaert PD, Fanayan S, Sharma S, Nice EC, Connor M *et al*: **Accelerating the search for the missing proteins in the human proteome**. *Nat Commun* 2017, **8**:14271.

13. Omenn GS: **Advances of the HUPO Human Proteome Project with broad applications for life sciences research**. *Expert Rev Proteomics* 2017, **14**(2):109-111.

14. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G *et al*: **InterProScan 5: genome-scale protein function classification**. *Bioinformatics* 2014, **30**(9):1236-1240.

15. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

16. **The National eResearch Collaboration Tools and Resources (Nectar) Cloud** [https://nectar.org.au/research-cloud]

17. Troger P, Rajic H, Haas A, Domagalski P: **Standardization of an API for Distributed Resource Management Systems**. 2007:619-626.

18. **Australian Access Federation** [https://aaf.edu.au/]

19. Omenn GS, Lane L, Lundberg EK, Beavis RC, Overall CM, Deutsch EW: **Metrics for the Human Proteome Project 2016: Progress on Identifying and Characterizing the Human Proteome, Including Post-Translational Modifications**. *J Proteome Res* 2016, **15**(11):3951-3960.

20. **Aspera High-Speed File Transfer Software** [http://asperasoft.com/]

21. **AARNet** [https://www.aarnet.edu.au/]

22. **BLAST+ Release Notes** [https://www.ncbi.nlm.nih.gov/books/NBK131777/]

23. Chowdhary N, Selvaraj A, KrishnaKumaar L, Kumar GR: **Genome Wide Re-Annotation of Caldicellulosiruptor saccharolyticus with New Insights into Genes Involved in Biomass Degradation and Hydrogen Production**. *PLoS One* 2015, **10**(7):e0133183.

24. Ouzounis CA, Karp PD: **The past, present and future of genome-wide re-annotation**. *Genome Biol* 2002, **3**(2):COMMENT2001.

25. Supek F, Bosnjak M, Skunca N, Smuc T: **REVIGO summarizes and visualizes long lists of gene ontology terms**. *PLoS One* 2011, **6**(7):e21800.

26. Splivallo R: **Biological Significance of Truffle Secondary Metabolites**. In: *Secondary Metabolites in Soil Ecology.* Edited by Karlovsky P. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008: 141-165.

27. Vahdatzadeh M, Deveau A, Splivallo R: **The Role of the Microbiome of Truffles in Aroma Formation: a Meta-Analysis Approach**. *Appl Environ Microbiol* 2015, **81**(20):6946-6952.

28. Pacioni G, Rapino C, Zarivi O, Falconi A, Leonardi M, Battista N, Colafarina S, Sergi M, Bonfigli A, Miranda M *et al*: **Truffles contain endocannabinoid metabolic enzymes and anandamide**. *Phytochemistry* 2015, **110**:104-110.

29. Iotti M, Piattoni F, Leonardi P, Hall IR, Zambonelli A: **First evidence for truffle production from plants inoculated with mycelial pure cultures**. *Mycorrhiza* 2016, **26**(7):793-798.

5.3 Conclusions

The reanalysis of the black perigord truffle data using this newly developed platform shows our platform can leverage updated knowledge base and database to provide more accurate as well as new annotations of the query proteins. For example, the platform identified homologous for 111 new proteins (with E-value 1e-05, and with at least 50% identity) compared to our previous study. At the same time, another 129 proteins that were identified previously (with the same criteria) were assigned a lower identity and removed from the annotation list. It also identified structural similarities for 1,393 proteins with PDB sequences as opposed to 101 proteins from our previous study. The platform also identified functional annotation for a significantly large number of proteins (82%) compared to our previous study (20%). While the automated platform is an information technology enabler to empower the researchers, users need to select their search databases carefully to match their studies to get the best output from this pipeline. Besides, many proteomics databases are available (as discussed in Chapter 1) that can be utilised to complement these annotations with experimental evidence (especially for the human missing protein identification). However, the datasets within each of these proteomics databases are generated using various MS platforms and search engines with various platform and search engine specific parameters. This coupled with the lack of integration between different databases, create further challenges for the researchers to analyse or interpret data. A guideline is needed for researchers to identify protein sequence and proteomics database as well as how to interpret data from heterogeneous MS data sources to annotate and complement the annotations accurately.

Chapter 6: A systematic bioinformatics approach to identify high quality MS data and functionally annotate proteins and proteomes

6.1 Summary

In the previous chapter, we presented the ProtAnnotator 2.0 automated platform to use existing knowledge to identify functional annotations of proteins. A large number of proteomics databases (described in section 1.4) are available to uncover the biology of a target proteome. A set of carefully selected high-quality protein sequence databases (closely related, experimentally validated and or reviewed) can be used to significantly increase the quality of the similarity search. The processed MS results or raw MS data (after reprocessing) from various proteomics databases can then be used to complement the annotation by providing identification or experimental validation. The reuse of such databases is crucial for large collaborative or community efforts such as the 'missing' protein identification initiative of the C-HPP consortium to fast-track the identification process without duplicating the experiment or efforts. However, the consolidated analysis and interpretation of proteomics data from various sources is a complicated task as the datasets come from a wide range of MS platform, search engine software and platform-specific parameters. Scientists often rely on the results from the automatic search engine of the MS platform (without checking the quality of the MS spectra) to identify proteins that can lead to incorrect identifications. Although some automated software and tools are available to automate the quality check (described in section 1.2.3), they are compute intensive and or rely on reprocessing data that may not always be practical.

6.2 Publication 4

Pages 83-96 of this thesis have been removed as they contain published material under copyright. Removed contents published as:

Islam M.T., Mohamedali A., Ahn S.B., Nawar I., Baker M.S., Ranganathan S. (2017) A Systematic Bioinformatics Approach to Identify High Quality Mass Spectrometry Data and Functionally Annotate Proteins and Proteomes. In: Keerthikumar S., Mathivanan S. (eds) Proteome Bioinformatics. Methods in Molecular Biology, vol 1549. Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-6740-7_13

6.3 Conclusions

In this study, we developed a simple protocol for the scientific community to functionally annotate protein using our previously described method as well as an MS evidence workflow to identify and mine publicly available (or individual lab) MS data to complement the annotation. We used walk-through examples from the missing human proteins to demonstrate the annotation protocol and the MS evidence workflow. The MS evidence workflow provides a guideline to interpret cross-platform MS search results to identify high-quality peptides, check their proteotypicity and finally check the quality of the MS spectra to detect and annotate a protein accurately. This combined annotation and evidence workflow can be used in any proteomics laboratory with minimal computational resources to annotate and interpret data. The automated annotation platform (underpinned by the annotation protocol) described in the previous chapter can be integrated with the MS evidence workflow to develop a cohesive identification and annotation pipeline to capture information from cross-platform sources to provide annotation and evidence for larger studies, such as the missing human protein identification project with high stringency identification and evidence criteria.

Chapter 7: Accelerating the search for the missing proteins in the human proteome

7.1 Summary

In the previous chapter, we discussed that our automated protein functional annotation platform can be integrated with the MS evidence workflow to build an automated protein identification and characterisation (both public and private source) pipeline to underpin extensive studies. In Chapter 1, we provided a comprehensive literature review on proteomics knowledge bases, tools, and algorithms. We then demonstrated that these databases could indeed be used for protein characterisation and identification (Chapter 3-6) of the novel, unknown and 'missing' proteomes. However, despite the advances in proteomics technologies and considerable growth of proteomics knowledge base, the questions that baffle us are, why are ~10% of the human proteins are still considered missing and what can be done to accelerate the identifications to uncover unknown human biology eventually? In this study, we comprehensively review the current HPP metrics and approaches to identify the challenges and possible solutions in finding these missing proteins.

7.2 Publication 5

nature COMMUNICATIONS

REVIEW

Received 16 May 2016 | Accepted 6 Dec 2016 | Published 24 Jan 2017 DOI: 10.1038/ncomms14271 OPEN

Accelerating the search for the missing proteins in the human proteome

Mark S. Baker[1], Seong Beom Ahn[1], Abidali Mohamedali[1,2], Mohammad T. Islam[2], David Cantor[1], Peter D. Verhaert[3], Susan Fanayan[1], Samridhi Sharma[1], Edouard C. Nice[4], Mark Connor[1] & Shoba Ranganathan[2]

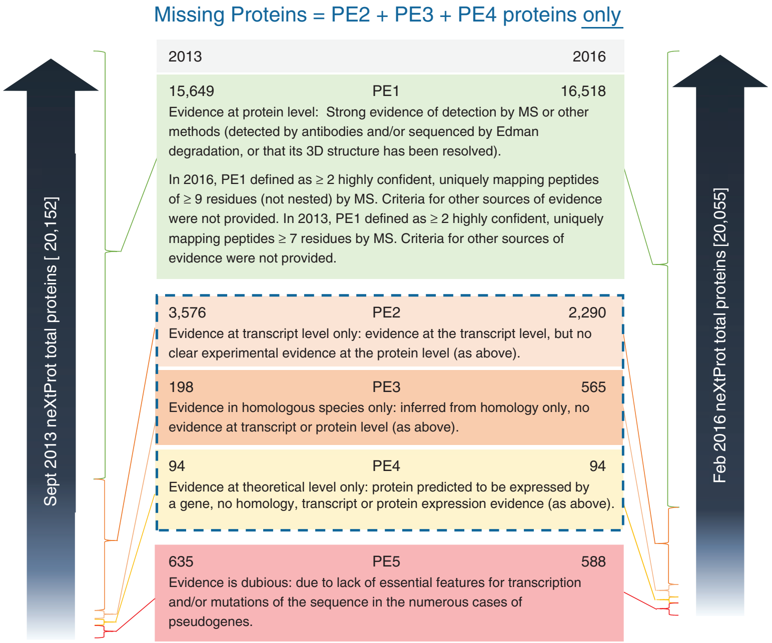
The Human Proteome Project (HPP) aims to discover high-stringency data for all proteins encoded by the human genome. Currently, ~18% of the proteins in the human proteome (the missing proteins) do not have high-stringency evidence (for example, mass spectrometry) confirming their existence, while much additional information is available about many of these missing proteins. Here, we present MissingProteinPedia as a community resource to accelerate the discovery and understanding of these missing proteins.

The Human Proteome Project (HPP) supports defining what it is to be human in molecular terms. It strives to 'know thyself' by finding high-stringency evidence for the ~20,000 proteins encoded by the human genome. Here, we focus on what has been termed the human proteome's 'missing proteins', discuss what renders them currently unobservable using high-stringency proteomic approaches, and outline a road-map that aims to accelerate the HPP. We review milestones and the progress of this global scientific effort to accurately identify and understand the biology of genome-coded human proteins. We focus on what has been achieved to date and we identify some areas where progress may be made. We provide a comprehensive survey of the characteristics of the so-called 'missing proteins', a term initially coined by Hancock and colleagues defined in Box 1 (refs 1,2), and we emphasize why they may be difficult to detect using mass spectrometry (MS) and/or validated antibody (Abs) techniques. Our re-analysis of publicly available MS data for the largest family of missing proteins (olfactory receptors), viewed in conjunction with other specific missing protein examples reveals a need for the community to capture as much complementary evidence as possible about missing proteins, in addition to high-stringency MS data. With this aim, we launch MissingProteinPedia (http://www.missingproteins.org), a community biological database that is complementary to the high-stringency HPP methodologies currently underway. MissingProteinPedia is a low-stringency communal database that will increase our understanding of the spatiotemporal biology of missing proteins, and accelerate their discovery by high-stringency MS.

[1] Department of Biomedical Sciences, Faculty of Medicine & Health Sciences, Macquarie University, New South Wales 2109, Australia. [2] Department of Chemistry & Biomolecular Sciences, Macquarie University, New South Wales 2109, Australia. [3] Department of Biology, Antwerp University, Antwerpen 2020, Belgium. [4] Department of Biochemistry and Molecular Biology, Monash University, Victoria 3800, Australia. Correspondence and requests for materials should be addressed to M.B. (email: mark.baker@mq.edu.au).

Box 1 | neXtProt protein evidence (PE) definitions and 2013→2016 PE data comparison.

neXtProt assigns every one of the 20,055 human proteome proteins as either PE1-5, using evolving communal metrics that have become stricter to improve identification confidence. HPP PE status from 2013→2016 is shown below (with protein numbers indicated top left and right of boxes).


Missing Proteins = PE2 + PE3 + PE4 proteins only
Sept 2013 neXtProt total proteins [20,152]
Feb 2016 neXtProt total proteins [20,055]
2013
2016

15,649 PE1 16,518
Evidence at protein level: Strong evidence of detection by MS or other methods (detected by antibodies and/or sequenced by Edman degradation, or that its 3D structure has been resolved).

In 2016, PE1 defined as ≥ 2 highly confident, uniquely mapping peptides of ≥ 9 residues (not nested) by MS. Criteria for other sources of evidence were not provided. In 2013, PE1 defined as ≥ 2 highly confident, uniquely mapping peptides ≥ 7 residues by MS. Criteria for other sources of evidence were not provided.

3,576 PE2 2,290
Evidence at transcript level only: evidence at the transcript level, but no clear experimental evidence at the protein level (as above).

198 PE3 565
Evidence in homologous species only: inferred from homology only, no evidence at transcript or protein level (as above).

94 PE4 94
Evidence at theoretical level only: protein predicted to be expressed by a gene, no homology, transcript or protein expression evidence (as above).

635 PE5 588
Evidence is dubious: due to lack of essential features for transcription and/or mutations of the sequence in the numerous cases of pseudogenes.

The PE2-4 proteins are now considered as the missing proteins since insufficient evidence has been produced as per the HPP metrics. The criteria for categorizing PE status, using data other than MS, remains to be communally defined. neXtProt protein data are constantly updated, so PE numbers vary with each version release.

Human Proteome Project (HPP) goals and progress

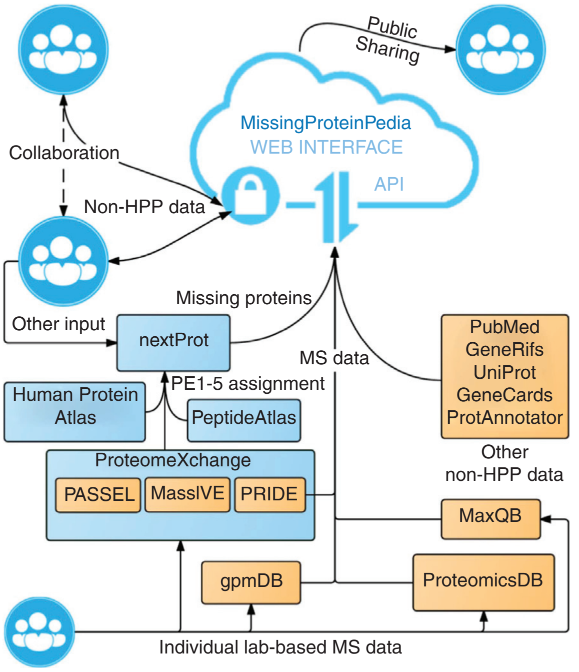
Science is rapidly becoming a global endeavour, with high-quality curation and annotation of data becoming the responsibility of the whole scientific community. Despite the Delphic maxim 'know thyself' being inscribed on the forecourt of the Temple of Apollo in ancient Greece during the sixth century BC, we still do not have a comprehensive description of what it means to be human in strictly molecular terms (that is, genome + epigenome + transcriptome + proteome + peptidome + metabolome). In 2010, the Human Proteome Organization (HUPO) formally initiated a flagship project called the Human Proteome Project (HPP). This ambitious project contributes to humans knowing themselves by collecting credible, high-stringency MS and other evidence for the ~20,000 or so proteins coded by human genes. The long-term aims of HPP are twofold. First, it aims to complete the protein 'parts list' of *Homo sapiens* by identifying and characterizing at least one protein product and as many post-translational modifications, single amino acid polymorphisms and splice variant isoforms as possible for each protein-coding gene. Second, it aims to transform proteomics so it becomes complementary to genomics across clinical, biomedical and life sciences, through technological advances and creation of knowledgebases for the identification, quantitation and characterization of the functionally networked human proteome.

In order to ensure all encoded proteins would be revealed and that all important biology and diseases would be represented, the HPP was amalgamated under two distinct but overlapping streams called the chromosome-centric (C-HPP) and Biology/Disease (B/D-HPP) Human Proteome Projects[3]. These are underpinned by three resource pillars; (i) MS, (ii) Affinity Reagents (for example, Abs), and (iii) a Knowledgebase. In addition to re-analysing and reporting HPP data, a number of complementary groups (PeptideAtlas; http://www.peptideatlas.org, neXtProt; http://www.neXtProt.org, GPMDB; http://www.gpmdb.org and Human Protein Atlas (HPA); http://www.proteinatlas.org) work cooperatively to provide annual HPP updates, present chromosome-by-chromosome tabulations, evolve high-stringency HPP data analysis metrics[4,5], and supply HPP data deposition guidelines for all researchers[6]. Critically, the HPP consortium encourages concurrent raw data deposition through standardized MS portals (for example, ProteomeXchange; shown as a schema in Box 2). The HPP also undertakes critical, annual re-analyses and reporting of the growing MS dataset with accompanying metadata using community-approved, high-stringency metrics.

The desire to build a reproducible, definable, metrics-driven, annotated HPP of the highest quality necessitated the imposition of terms defining the categories of evidence obtained. To enable this, it was communally agreed that the protein-centric knowledge platform neXtProt[7,8] would classify HPP proteins by protein existence (PE), based on partial/complete Edman sequencing, identification by MS, 3D structure (X-ray/NMR), good quality protein–protein interaction data and/or detection of a protein by validated Abs (for example, in the HPA[9]). Metrics, guidelines and/or PE categories have been agreed on and revised through community forums, facilitated by HUPO. Since the HPP was launched in 2010, we have learned many lessons. The importance of 'speaking the same language' with regard to MS analysis metrics and data submission guidelines has been prominent. Kim *et al.*[10] and Wilhelm *et al.*[11] proposed drafts of the human

Box 2 | Integration of MissingProteinPedia with HPP.

The *MissingProteinPedia* is a publicly available protein data and information sharing web system that aims to collate any relevant data pertaining to any PE2-4 protein. At its core is a flexible schema-based database-driven web system allowing captures of all PE2-4 protein PubMed data, based upon gene and protein including synonyms. The database also allows unpublished, preliminary or proprietary data (for example, antibody, MS, cell biology and genetic studies) to be shared with collaborators via a protected interface.



Schema 1: The *MissingProteinPedia* collates and displays protein information from existing databases using various web services and application programming interfaces. Furthermore, the web interface allows researchers to collaborate and share data not available through other databases. The schema includes the recent illustration of the high-stringency HPP metrics engine[9].

MissingProteinPedia facilitates HPP cross-disciplinary collaboration by providing a complementary, unfiltered, lower stringency perspective to both the HPP metrics and guidelines approaches, enabling community evaluation and scrutiny. *MissingProteinPedia* incorporates text mining technology to fetch and search accumulated UniProt, GeneCards, GeneRifs, PubMed and ProtAnnotator PE2-4 data. In addition, *MissingProteinPedia* summarize publicly available MS data from PRIDE, GPMDB, ProteomicsDB and MaxQB for relevant PE2-4 proteins. It also allows community users to annotate data and administrators to curate information before web publication.

proteome in 2014. These studies challenged the imposition of communal metrics, including previously agreed consensus regarding protein target-decoy false discovery rates (FDRs) and requisite minimum proteotypic peptide length (≥7 amino acids in 2014). The term proteotypic in this context refers to a human peptide sequence of any length found by MS that is uniquely derived from a single known human protein expressed by the genome. The term is often used interchangeably with the commonly used terms, uniquely expressed and unitypic. In the HPP, proteotypic peptides (that is, two proteotypic peptides of suitable length) are employed to identify the expression of a human protein by MS methods. Discussion around the impact of single amino acid variation on application of the term proteotypic are currently underway.

Conclusions from both the human proteome drafts[10,11] were considered contentious[12,13] because they chose to interrogate MS findings using different metrics to those established by the HPP after communal agreement. Because of debate around these publications, large-scale heterogeneous datasets were recognized as raising questions related to assumptions around FDR protocols[12]. Encouragingly, positive, collaborative, communal efforts (for example, revised data deposition guidelines and clear diagrammatic representations of data re-analysis workflows and metrics) are underway and will resolve many of the issues raised. In response, the HPP Knowledgebase pillar proposed more rigorous metrics for substantiating claims of the identification of previously unobserved proteins (that is, PE2-5 proteins; Box 1). It has been proposed that datasets should be culled at 1% protein FDR with additional estimates of peptide and peptide spectral match (PSM) level FDRs and notification of the numbers of proteins, peptides and spectra passing/failing these thresholds. In late 2015, PeptideAtlas proposed increasing the minimum thresholds to two proteotypic peptides of ≥9 amino acids with raw spectra to be made publicly available (downgrading 432 previously validated PE1 proteins)[4]. Some exceptions included predicted proteins that are unable to be cleaved to form at least two tryptic proteotypic peptides of required length[4]. While neXtProt initially retained less stringent criteria thresholds of two proteotypic peptides of ≥7 amino acids or one proteotypic peptide of ≥9 amino acids (that is, with consequent downgrading of 20 PE1 proteins), in February 2016 they aligned with the more stringent PeptideAtlas metrics. These developments were incorporated into both the 2016 HPP metrics and HPP guidelines for data submission that have been recently

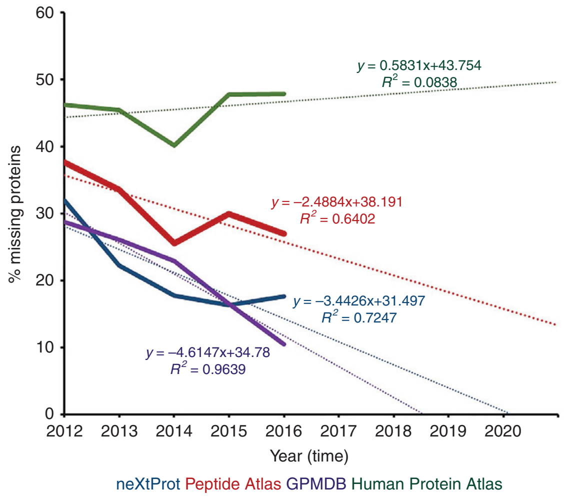


Figure 1 | Extrapolation of linear best-fit rate equations demonstrates the rate at which various HPP input databases and GPMDB are currently 'finding' PE2-4 proteins. Data required for this analysis (2012 → 2014) was extracted from Omenn *et al.*[4] , with additional (2015 and 2016) statistics obtained from neXtProt, Peptide Atlas and GPMDB. Note: GPMDB data are not currently captured by neXtProt as part of the data input into the HPP (see Box 2), but GPMDB plays a role in defining annual HPP metrics.

published[5,6]. It should be noted that while the observation of two ≥9 amino acid proteotypic peptides by highly accurate MS dramatically reduces statistical uncertainty, it does not make the putative identification of any protein unequivocal.

What is known about missing proteins

On behalf of the HPP, neXtProt curates, integrates and computes PE (PE1-5) scores based on experimental information from multiple types of enquiry (see Box 1). In this review, we focus solely on those proteins that are classified as being either PE2 (evidence only at transcript level), PE3 (inferred from homology) or PE4 (proteins inferred to exist). These three PE groups have been collectively and colloquially defined as the HPP 'missing proteins'[1,2] (Box 1), although a recent study erroneously mentions missing proteins include PE5s[14], which are highly unlikely to be translated. Definitions for PE1-5 (ref. 4) proteins are released by neXtProt before annual HUPO Congresses.

The HPP endorses open, community-wide use of standardized re-analysis pipelines, with attention to the evolving HPP guidelines for researcher data submission[6] and metrics used for global concatenated communal data re-analyses[4,5]. It also encourages confirmation of novel findings with advanced MS methods (for example, selected reaction monitoring (SRM) and data-independent acquisition, including new methods such as SWATH-MS[15]). This process implies that PE2-4 proteins need to be re-classified regularly (that is, upgraded or downgraded) after agreed, metrics-driven, communal re-analysis, preferably with publication of the rationale for their re-assignment. This high-stringency approach is crucial for quality assurance and is favoured over any individual laboratory MS data analysis, that can result in potentially contestable claims that regularly arise for 'finding' suites (sometimes hundreds) of PE2-4 missing proteins.

It should be stressed that the PE2-4 proteins only represent a list of proteins currently not fulfilling HPP metrics, and that these lists have evolved since the launch of the HPP. Recent HPP questions involved issues around assessing MS quality, validating automated findings and considering potential alternative protein assignments for specific PSMs. Due to the evolution of HPP data submission guidelines and data re-analysis metrics, we have a higher baseline of proteins at PE2-4 levels from which ongoing discovery and transition to PE1 status continues. Current metrics for a protein to be PE1 are based on statistical calculations minimizing the risk that any peptide can be randomly mapped to multiple genes products.

Of the 20,055 currently allocated proteins in the human proteome (neXtProt 12 February 2016), only 16,518 were PE1, with a further 588 considered at best to be hypothetical (PE5). This means that at present 2,949 proteins are PE2-4; composed of 2,290 PE2 (transcript only), 565 PE3 (inferred from homology) and 94 PE4 (predicted). While only 2,949 PE2-4 proteins remain to be confirmed by high-stringency HPP metrics, our current approach takes little account of the potential goldmine of valid data available from other sectors of the scientific community. We argue that collectively alternative sources of complementary data provide clues that may facilitate the discovery of additional PE2-4 proteins by subsequent HPP MS metrics. Recognizing this fact, we acted upon comments made by researchers outside the proteomics community who argued that in order to be functional or biologically relevant a protein did not need to be reduced to any statistically required number of proteotypic peptides of any predefined length. As an example, they noted the many highly bioactive secretory peptides, such as neuropeptides, which were crucial to human biology. Several of these peptides are very short (<9 amino acids) secreted proteoforms that perform essential functions as intercellular signals. However, such peptides do not fall within the currently accepted thresholds in bottom-up HPP MS experiments. These constraints (that is, two uniquely mapping proteotypic peptides at least nine amino acids long) preclude discovery and annotation of these peptides as PE1, as well as incorporation into high-stringency datasets. Thus, short peptide proteoforms, such as the orexigenic neuropeptide QRFP, continue to be 'missing' in HPP databases, annotated as known only at the transcript level (https://www.nextprot.org/entry/NX_P83859/sequence). Similar arguments have been made about proteins unable to be cleaved by trypsin to produce two uniquely mapping proteotypic peptides of at least nine amino acids.

Analysis of Box 1 data reveals significant HPP progress. Over the period 2013–16, PE1 assignments have increased by 5% from 15,649 to 16,518 (78 → 82% of the estimated human proteome), with 1,079 PE2-5 entries re-assigned as PE1. This has occurred despite deliberate efforts to increase stringent MS metrics, leading to 432 PE1 proteins being downgraded to PE2-5 proteins. Interestingly, the data demonstrate that 22 new PE1 proteins were listed, which were previously not present at any PE level (for example, UMAD1, SULT1A4, MYH16).

Unfortunately, as can occur when 'big data' is not endorsed through annual community jamboree/forums, experimental evidences and detailed rationales for such re-classifications are not currently made public nor are they easily accessible to non-experts. We therefore encourage establishing annual PE annotation/assignment jamborees, analogous to how the human genome project dealt with similar challenges.

Applying best fit linear extrapolations to all available PE re-assignment data[5] (Fig. 1), it appears that with current neXtProt high-stringency metrics, the HPP will likely reach completion of ≥95% parts list coverage (PE1 status) near the close of the current decade (that is, 2020). As the final arbitrators

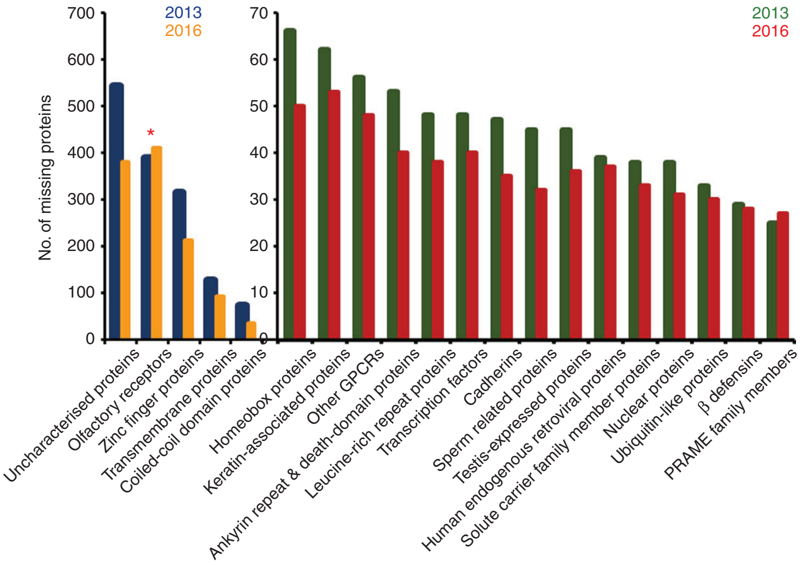


Figure 2 | Top 20 missing protein families to determine protein families enriched in the February 2016 neXtProt PE2-4 report list. According to these data, olfactory receptors (ORs; marked with a red asterisk *) represent the largest family of PE2-4 proteins. The olfactory receptors also show the largest increase between 2013 and 2016 (that is, 15% in 2016 from 10% in 2013) when compared to the other families. The scale '0–70' represents a magnified axis scale for protein descriptors having <70 missing proteins. Blue and green colours represent PE2-4 proteins from 2013 whereas orange and red colours represent 2016 missing proteins.

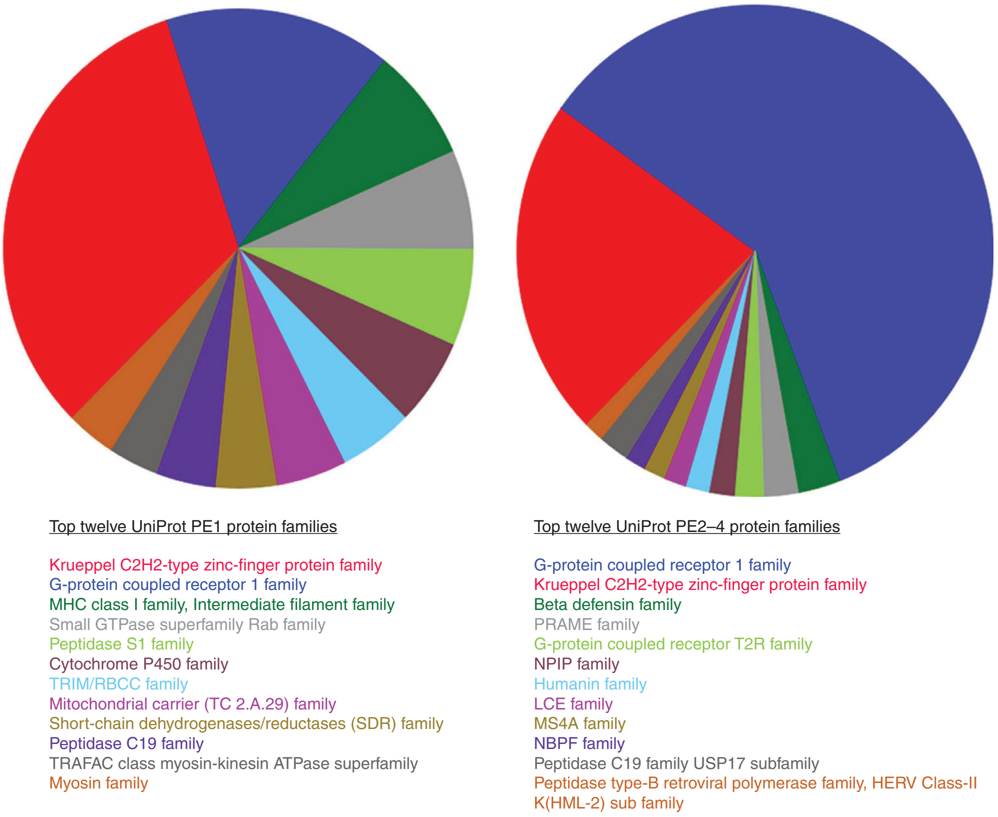


Figure 3 | Most prolific PE1 and 12 PE2-4 UniProt protein families represented in the HPP neXtProt February 2016 release. The most represented PE1 families (left hand side) are the Kreuppel zinc-finger protein family followed by the G-protein coupled receptor 1 family. These two families are also at the top of the PE2-4 category (right hand side) with the order reversed.

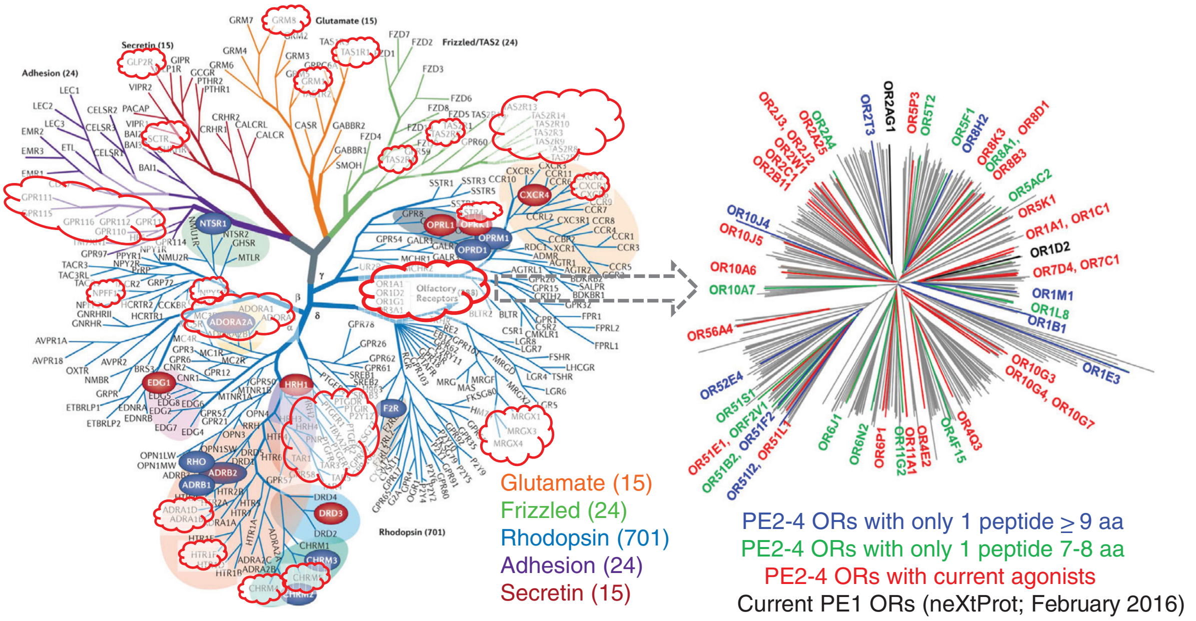


Figure 4 | Phylogenetic analysis of PE distribution across GPCRs and olfactory receptors. In this composite figure, GPCR (left) family branches (largest 'receptor' subset of all human and the PE2-4 proteins) are shown in an unrooted phylogenetic tree from Panther analyses with PE2-4 GPCRs highlighted inside red clouds, and an unrooted GCPR subset phylogenetic tree showing olfactory receptors (right) was produced using iTOP[56], from neXtProt February 2016 PE1 olfactory receptors or best available, manually validated proteotypic MS evidence for olfactory receptor was retrieved. olfactory receptors with functional activity (known agonists) are shown in red in the left figure, as from Mainland *et al.*[16]. GPCR figure modified with permission from Macmillan Publishers Ltd: *Nature Reviews. Drug Discovery*, Stevens *et al.*[57] copyright 2013.

of HPP PE1 calls, the statistical analysis of neXtProt is particularly telling, with a recent lag/hiatus evident. Equally, extrapolating PeptideAtlas data alone suggests 95% completion somewhere around 2030–40.

Orthogonal efforts to find missing proteins

A major outcome from the C-HPP effort to date has been that researchers have been made to consider possible reasons why PE2-4 proteins have not been found by MS, Ab-based or other methods. This has now inspired the development of novel strategies to find the PE2-4 proteins, or understand why they are missing. Some approaches, envisaged to date, include subcellular enrichment of families, groups, clades or classes (for example, membrane proteins); more extensive protein and peptide fractionation before MS; increased MS accuracy, sensitivity and throughput; more reliable, specific and accurately validated Ab technologies, which are currently underway with collaborative efforts by the HPP Ab technology pillar; scrutiny of proteins not amenable to tryptic digestion, those failing to yield 'flying' tryptic peptides or those outside observable mass range detection settings[14]; analysis of cross-linked or otherwise insoluble proteins; examination of rare human tissues/cells under differing spatiotemporal conditions or differentiation states; exposure of tissues to pathophysiological and/or environmental cues, and finally; broadening the capture of data from solely MS and Ab-based data streams.

Bioinformatics efforts to understand missing proteins

Given the current scientific and protein informatics data detailed in Supplementary Table 1 and with a view to finding more PE2-4 proteins, we additionally undertook bioinformatics analyses of all PE2-4 proteins according to their families, sub-families, clades, groups, ontologies, pathways and networks. Figures 2–4 summarize these analyses with increasing depth across neXtProt descriptors (Fig. 2), comparison of protein biologies between PE1 and PE2-4 (Fig. 3), and PE2-4 G protein-coupled receptor (GPCR) family (Fig. 4, left) and OR* (Fig. 4, right) clade phylogenetic tree analyses, focussing on the most populous protein families from Figs 2 and 3.

Analyses of major descriptors (that is, protein subfamilies, classes, domain-type) for neXtProt 2016 PE2-4s indicated that five groups of proteins were highly represented. The PE2-4 groups with greater than 50 members in decreasing order are: olfactory receptors (red * in Fig. 2), zinc finger proteins, non-GPCR transmembrane proteins, coil-coil domain proteins and homeobox proteins (Fig. 2). Encouragingly, our analysis demonstrates a decrease in the percentage of HPP PE2-4 proteins assigned as 'uncharacterized' by neXtProt over the 2013–16 period. These data also demonstrate the substantial success made across all major (that is, the top 20) protein groups, with the sole exception of the enigmatic olfactory receptors. In agreement with these data, Panther Protein Class analysis of 2,491 classifiable genes confirmed the major PE2-4 protein types were: receptors (PC00197), transcription factors (PC00218), transferases (PC00220), transporters (PC00227), membrane traffic proteins (PC00150), enzyme modulators (PC00095) and signalling molecules (PC00207), with other groups represented at low percentages.

Analysis of the top 12 UniProt families found in the 2016 PE2-4 and the PE1 lists (Fig. 3) demonstrates a highly significant enrichment of GPCR type 1 family missing proteins, and a reduction in the % of zinc finger proteins in the PE2-4 proteins list. Furthermore, we note that when the highest 12 families are examined in the PE2-4 list, the vast majority of those families' members are found to be 'missing', with relatively few PE1 representatives. Only three families (that is, Kruppel C2H2-type zinc finger, GPCR type 1 and Peptidase C19 protein families) were common to both the major PE1 and the major PE2-4 families. Interestingly, PE1 assignments account for only 22% of all GPCR type 1 proteins while it accounts for 59% of the Kruppel zinc finger proteins. If one considers only the PE2-4 'missing' proteins, GPCR type 1 members represent 25% and zinc finger family members 9%. On a family-by-family basis, apart from Kruppel zinc finger (34%) and peptidase C19 (31%) proteins, the remainder of the top 12 families are noticeably composed of missing proteins (that is, range 50–95% of the total family membership). This implies that when a major family is 'missing' by current HPP metrics, extremely limited high-stringency MS knowledge exists for any member of that protein family (for example, of 22 known PRAME proteins 19, 86% are assigned as PE2-4 and re-analysis of olfactory receptor MS data summarized in Supplementary Table 2 shows all (100%) are currently missing).

The olfactory receptor family missing proteins

Subsequently, we examined the largest PE2-4 family, namely human GPCRs (shown in dark blue in Fig. 3). These are responsible for cellular responses to everything from protons and photons to hormones of >30 kd, metals, nutrients, small molecules including volatiles and neurotransmitters through many of our major senses (that is, sight, olfaction and taste). GPCRs also are the most important pharmaceutical drug target and largest family (>800) in the human proteome, as well as the largest membrane receptor family. They instigate signalling through nucleotide exchange involving heterotrimeric G-proteins and can be classified into five major families and subdivided into subfamilies based on sequence homology, to (1) rhodopsin (class A), (2) secretin, (3) adhesion (class B), (4) glutamate (class C), and (5) Frizzled/taste receptor 2 (TAS2). Phylogenetic analysis of GPCR PE2-4 proteins demonstrates that although singleton representatives and a few clusters are distributed across all five major subfamily branches/classes (Fig. 4), by far the highest proportion of missing proteins ($n=400$; ~15% of all human PE2-4 proteins) emanate from the rhodopsin branch of the unrooted GPCR phylogenetic tree where the olfactory receptors reside. Note that family members with determined crystal structures are highlighted on the phylogenetic tree in coloured ovals (including ADORA2A, which has been recently re-classified by neXtProt as PE1).

Discovering functionality of the complete missing human olfactory receptor repertoire has proved difficult with only 49/~400 human olfactory receptors having known ligands before the recent studies of Mainland *et al.*[16]. Using high-throughput screens of human olfactory receptors against 73 potential ligands they identified agonists for 27 receptors (coloured red in Fig. 4, right), including 18 that were previously orphan receptors. Their dataset addressed a bottleneck in research around functionality of human olfactory receptors by showing how physical olfaction stimuli can signal post-receptor activation. Correlating odorant ligands to olfactory receptors provides a valuable database, identifying functional olfactory receptors with potential to be strategically targeted through proteomic approaches and subsequent conversion to PE1 proteins.

The recent studies by Kim *et al*,[10] and Wilhelm *et al.*[11] generated intense interest in MS evidence for the expression of the chemosensory olfactory receptor family, as they claimed to have 'unearthed' a surprisingly high number of 108 and 200 PE2-4 olfactory receptors, respectively. Of the human genome's 480 olfactory receptor genes in the latest version of neXtProt, 12 are considered hypothetical or putative (PE5). The remaining 468 olfactory receptor genes code for 411 unique proteins, with only two classified as PE1, and the remaining 409 classified as PE2-4. The claims for finding missing olfactory receptors by the draft human proteome papers above were rapidly critiqued by Ezkurdia *et al.*[12] and Deutsch *et al*,[13] on the basis of marginal spectral quality, deficiency of stringent protein/peptide 1% FDR criteria, use of short peptides, and erroneous or potentially ambiguous peptide identification, with the suggestion that these claims represent 'the cream of false positives'. Collectively, these errors led Ezkurdia *et al.*[12] and Deutsch *et al.*[13] to conclude that there was little evidence for even a single olfactory receptor (including the two listed in previous releases of PeptideAtlas). Incidentally, 10 olfactory receptors were considered 'found' by Choong *et al.*[17] in the 2015 release of neXtProt with MS and Ab evidence. However, this evidence was considered insufficient for all these 10 olfactory receptors, suggesting that currently known olfactory receptor proteins may not possess sufficiently documented protein evidence in neXtProt.

From the amazing repertoire of 411 unique olfactory receptor proteins, only two are currently considered PE1 in the neXtProt 2016 release (namely, OR2AG1 and OR1D2; coloured black in Fig. 4, right). For OR1D2, no MS or Ab evidence is available, with three publications cited as functional evidence. For OR2AG1, neXtProt reports a single peptide 7 amino acids long, with no Ab evidence and functional evidence from two publications[18,19]. One of these studies[18] equally reports function for another olfactory receptor, namely OR1F12 but this remains classified by neXtProt as PE4, whose status is based upon sequence homology. Thus, it appears that both these PE1 olfactory receptor proteins do not actually conform to HPP MS-based metrics and require

Box 3 | Accelerating discovery of the complete human proteome.

We recognize the tremendous achievement the Human Proteome Project has made since its 2010 launch by making available high-quality, communal MS (and other) data for ~82% of the human proteome (February 2016).

To accelerate discovery of the 15% of the human proteome defined as the missing PE2-4 proteins, we recommend and encourage the following:

1. All proteomics practitioners, human researchers and human biology/medicine journals renew their efforts to observe current high-stringency HPP re-analysis *metrics* and researcher data submission *guidelines*.
2. All MS data should be incorporated into a single database (for example ProteomeXchange), including MS databases not currently captured, where data are provided transparently for any claim for a current PE2-4 protein.
3. The HPP should communally develop metrics and guidelines for processes by which they deal with all non-MS data sources. In particular, transparency around how protein evidence scoring for non-MS data needs to be communally accepted and reported.
4. An annual jamboree to evaluate and approve both MS and non-MS protein evidence reclassification proposals.
5. All possible biological data concerning the PE2-4 missing proteins to be comprehensively captured in *Missing-ProteinPedia*.

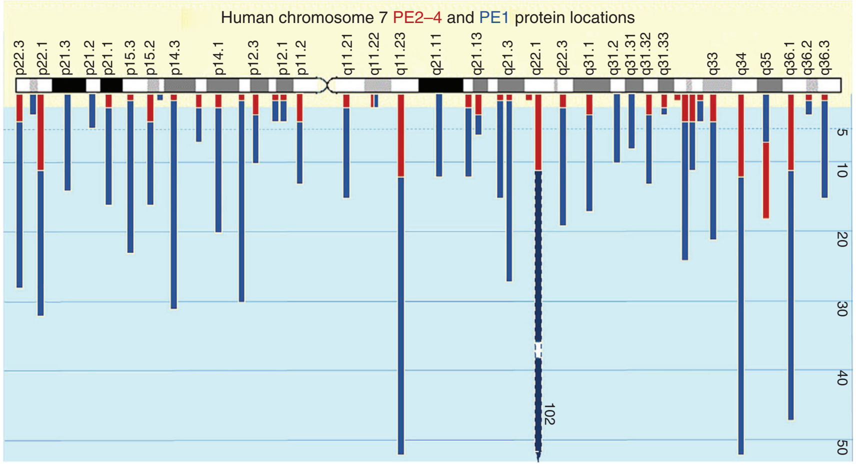


Figure 5 | Positional mapping of the PE1 (757) and PE2-4 (139) proteins along human Chr 7. The data show random distribution of both along the complete length of human Chr 7. However, Giemsa banding patterns of light (GC-rich) and dark (GC-poor) bands are shown that debatably correspond to regions of gene density from light (higher gene density) to dark (lower gene density)[58].

closer community examination (Box 3), as does the way we consider functional/biological data as evidence for PE.

Olfactory receptors are involved under most physiological situations with odour recognition but have recently been shown to be expressed in multiple epithelial tissues with many potential chemosensory roles[20–22]. Criticisms of olfactory receptor restriction to nasal epithelial tissue are ill-advised[11] and appear erroneous[12,20–22]. Given these data and the comprehensive olfactory receptor functional studies conducted by Mainland *et al.*[16], we believe that a systematic capture of non-MS data and a communal re-assessment of all olfactory receptor PE assignments would be timely. To bring additional perspective to the olfactory receptor mêlée and to emphasize the challenges we face in finding the missing olfactory receptors by high-stringency MS, we undertook an analysis of all currently available raw olfactory receptor spectra from public repositories. This re-analysis reinforces that the best available MS data fail to provide high-stringency PE1 level proof for any GPCR olfactory receptor members using current metrics (Supplementary Fig. 1 and Supplementary Table 2). Despite 2,361 manuscripts revealed by an ‘olfactory receptor and human’ PubMed keyword search, only piecemeal MS evidence for any human olfactory receptor is currently available.

To verify the *status quo*, we trawled public MS proteomic repositories (including GPMDB, PRIDE, ProteomicsDB, MAXQB and Human ProteinPedia), and aggregated 122,717 peptide MS entries (PSMs of length ≥7 aa), including many with multiple PE2-4 olfactory receptor observations. This collective dataset was processed through a semi-automated workflow (Supplementary Fig. 1), including manual spectral validation to filter reliable peptide assignments, with consideration of leucine/isoleucine ambiguity and BLAST analysis to account for possible single amino acid variations coding for peptides, as detailed elsewhere[23]. Briefly, the data (using Batch Peptide Match) identified 4,751 proteotypic olfactory receptor peptides (3.9%), following removal of non-proteotypic and decoy peptides. Of the proteotypic peptides, only 286 (6%) were tagged with a high search engine confidence value score by either SEQUEST, Mascot or MaxQuant. Finally, manual spectral validation (taking into consideration, noise, error rates (to matched peptide sequence), the run of B and Y singly charged ions, unassigned peaks and relative intensity of the spectrum) allowed us to sift out 64 high quality spectra for 24 peptides. As two overlapping peptides could be merged for a single olfactory receptor, this culminated in 23 unique olfactory receptor peptides. In summary, this analysis provided MS evidence for 23 of 409 missing olfactory receptors (5.6%).

The best available MS evidence for these 23 olfactory receptors is shown in Supplementary Table 2, and it includes peptides from GPMDB (1 green, 1 yellow and 5 red peptides), PRIDE (10 peptides) and ProteomicsDB (7 peptides). It should be noted that 14 PSMs represent a single 7–8 amino acid peptide, while 9 possess a single PSM of >9 amino acids. Proteins derived from matches were cross-referenced against HPA with no (zero) olfactory receptors found in the current (May 2016) high confidence HPA premium dataset. In addition, 13 peptides (Supplementary Table 2) were found to have complete or partial matches with 14 SRM peptides listed in the current version of SRMAtlas.

In summary, we demonstrate that many missing PE2-4 olfactory receptors possess single high-confidence PSM evidence, although best available MS spectra are insufficient to meet current HPP metrics. These could be considered as PE2-4 proteins ‘waiting in the wings’, requiring confirmatory proteotypic PSM identifications at the required length to reach high-stringency requirements.

Chromosome 7 example missing proteins

Under the C-HPP, the proteomic information found across chromosomes 1-22, X, Y and mitochondrial DNA are being studied by country-based or regional cluster teams. Australia and New Zealand undertook analysis of the proteins coded by human chromosome 7 (Chr 7)[24,25]. As part of our ongoing efforts, we demonstrate that current PE2-4 proteins are located across the length of the long and short arms, approximately equally dispersed across the length of Chr 7 (Fig. 5). This holds true for the majority (but not all) chromosomes examined to date. At one chromosomal location, namely 7q35, a significantly greater number of PE2-4 proteins (18/25) were found than PE1 proteins (7/25). Interestingly, however, when Giemsa (that is, reported relative gene richness) staining patterns along Chr 7 were compared for PE2-4 and PE1 distribution, we observed that 56% PE2-4s emanate from high gene density Chr 7 regions, 12% from moderate, 25% from low-moderate and only 1.5% from regions of low gene density. PE1 proteins generally distribute across Chr 7 locations with PE2-4 proteins, with few regions (only p22.2, p21.3, p21.2, p15.1, q21.11, q31.2 and q31.31) not having both PE classifications represented. Chr 7 PE2-4 proteins do not emanate from gene-poor regions and hence it is reasonable to suspect that other factors (for example, low spatiotemporal expression) are more likely to explain why they have not been found by high-stringency MS to date. These observations need to be replicated for all chromosomes by other C-HPP teams.

Of the 134 Chr 7 PE2-4 proteins, 27 are known to be GPCRs. The majority of these encode olfactory (15) or taste-related (six) receptors, with only four 'orphan' GPCRs and two well-described GPCRs (5-HT_{5A} and mGluR8). There are many reasons why these proteins may still be considered missing. First, they all have restricted anatomical expression. In particular, the receptors for odours and ingested chemicals, which are likely expressed in only a few cells in specific regions of the body. Further, many missing proteins may be localized to a few discrete cells and/or difficult to access cellular compartments, like axon terminals, inner/outer hair cells (OHCs) or cilia on olfactory sensory neurones. Second, receptor expression may be extremely low even where they are physiologically active. Finally, it is possible that gene products are not translated/transcribed under normal physiological situations, or indeed at all. Their absence from proteomic databases suggests they are not highly abundant but it does not mean they are not important or not expressed. Indeed, a cursory examination of Chr 7 PE2-4 GPCR proteins reveals many non-proteomic studies show these GPCRs represent a very active part of the human proteome. Using the BPS/IUPHAR Concise Guide to Pharmacology (http://www.guidetopharmacology.org/index.jsp)[26] as a starting point for analysis, we provide some examples. First, HTR5A is part of the large family of receptors for the neurotransmitter serotonin (5-HT). When expressed, 5-HT_{5A} receptors stimulate G protein activity resulting in inhibition of adenylyl cyclase[27], indicating it is a functional GPCR. mRNA for 5HT_{5A} receptor has been detected in the human brain by *in situ* hybridization[28] and PCR[29]. However, our search shows no reports of protein localization by immunohistochemistry or identification by western blot in any human tissue. Mice with a 5-HT_{5A} receptor deletion have altered behaviour and a distinct response to the serotonin receptor ligand LSD[30], indicating the protein is functional. It is likely that low levels of protein and restricted anatomical localization preclude identification of 5-HT_{5A} receptors by MS.

A second receptor we considered is GRM8 (metabotropic glutamate receptor 8, $mGlu_8$), which is part of the large family of receptors for the prominent neurotransmitter glutamate. In a heterologous expression system, activation of $mGlu_8$ receptors results in inhibition of adenylyl cyclase[31], indicating it is a functional GPCR. *In situ* hybridization reveals discrete but low levels of mRNA in human brain[32,33], while $mGlu_8$ mRNA has been reported in cancer cell lines[34], hippocampal cells[35], astrocytes[36] and in patient tissue in epilepsy or multiple sclerosis. Murine deletion of $mGlu_8$ affects hippocampal synaptic transmission[37], suggesting function under physiological conditions. Low levels and restricted anatomical localization may preclude identification of $mGlu_8$ receptors by MS, although the receptor is also large and has a complex genetic structure, which probably leads to alternatively splice transcripts, and potentially several protein species[33,38].

Finally, GPR22 (Probable G-protein coupled receptor 22) is a class A GPCR, with mRNA expressed in human heart and brain[39–42]. Interestingly, GPR22 has an unusually AT-rich mRNA, and only when enrichment is artificially rectified by introduction of G-C bases can signalling be restored in heterologous expression systems (Gi/o-mediated stimulation of G protein activity and constitutive inhibition of AC activity[41]). No ligand has been identified for GPR22, and GPR22 knockouts seem physiologically unremarkable. However, GPR22 mRNA is significantly reduced by aortic banding, a procedure that mimics cardiac hypertrophy produced by high blood pressure, and in GPR22 knockouts heart failure follows more rapidly than in wild type animals, implying a role for responses to cardiac stress[41]. There is no peer-reviewed report of GPR22 immunoreactivity in human tissues, although several corporate sites show neurons and other cells displaying putative GPR22 immunoreactivity. Sera from mice immunized against a human GPR22 peptide label cells in rat heart, although staining suggests GPR22 is restricted to subsets of myocytes[41]. The lack of an identified ligand for GPR22 has dampened enthusiasm for further pursuing functional studies through conventional biochemistry, and coupled with lack of neuronal phenotype in GPR22 null mice, it is not surprising no further attention has been paid to it. Unlike 5HT_{5A} and $mGlu_8$ receptors, which likely have roles in normal physiology (even if understudied), there is little evidence to speak for or against function of GPR22, despite mRNA being detected by multiple investigators. However, for even the most obscure (non-olfactory) PE2-4 GPCRs, some evidence exists, suggesting that they are expressed in some tissues under certain conditions.

While we can learn much from an analysis of the Chr 7 PE2-4 GPCR proteins, the reasons for other proteins apparently 'falling through the cracks' and having PE2-4 assignments may be legion. Below, we examine two current PE2-4 examples that appear to have strong biological non-HPP evidence that, combined with the olfactory receptor data above, argue for a broader, community-based, open data base strategy. We propose that opening up the HPP to consider other sources of data might concomitantly accelerate re-classification of PE2-4 proteins to PE1 status through the existing high-stringency HPP workflow.

In an orthogonal approach to understand the Chr 7 PE2-4 proteins, an example was randomly selected. Prestin (gene name *SLC26A5*) retrieved 91 peer-reviewed PubMed manuscripts, with the oldest in 2000 entitled 'Prestin is the motor protein of cochlear outer hair cells'[43], while another was a recent review of structural and functional properties[44]. Antibodypedia unearthed 83 anti-prestin Abs from 15 different vendors (http://www.antibodypedia.com/explore/prestin). Though not listed on the Therapeutic Target database, Drugbank or Binding DB, prestin's substrates are listed as Cl^- and $HCO^-{}_3$ by the IUPHAR-DB (pharmacological targets) database[45]. Additionally, the Human Gene Mutation Database lists two prestin missense/nonsense mutations that produce deafness/autism phenotypes (CM075015 and CM124551), with one splice-variant linked with deafness (CS030995). Furthermore, the gene is known to have 15 transcripts. Equally, 12 patients with overlapping copy number variants are listed in DECIPHER: Database of Genomic variants and phenotype in Humans Using Ensembl Resources. Additionally, zebrafish studies captured in ZFIN include several CRISPR targeting agents (http://zfin.org/ZDB-GENE-030131-1566) directed against prestin. In conclusion, this randomly selected Chr 7 PE2-4 protein shows there is copious public functional evidence at the protein level available, despite there being zero high-stringency MS or acceptable Ab evidence.

Particular physiological, cell and molecular factors make prestin intractable to being found by MS. First, it is a bullet-shaped membrane protein that is localized only on the OHCs of

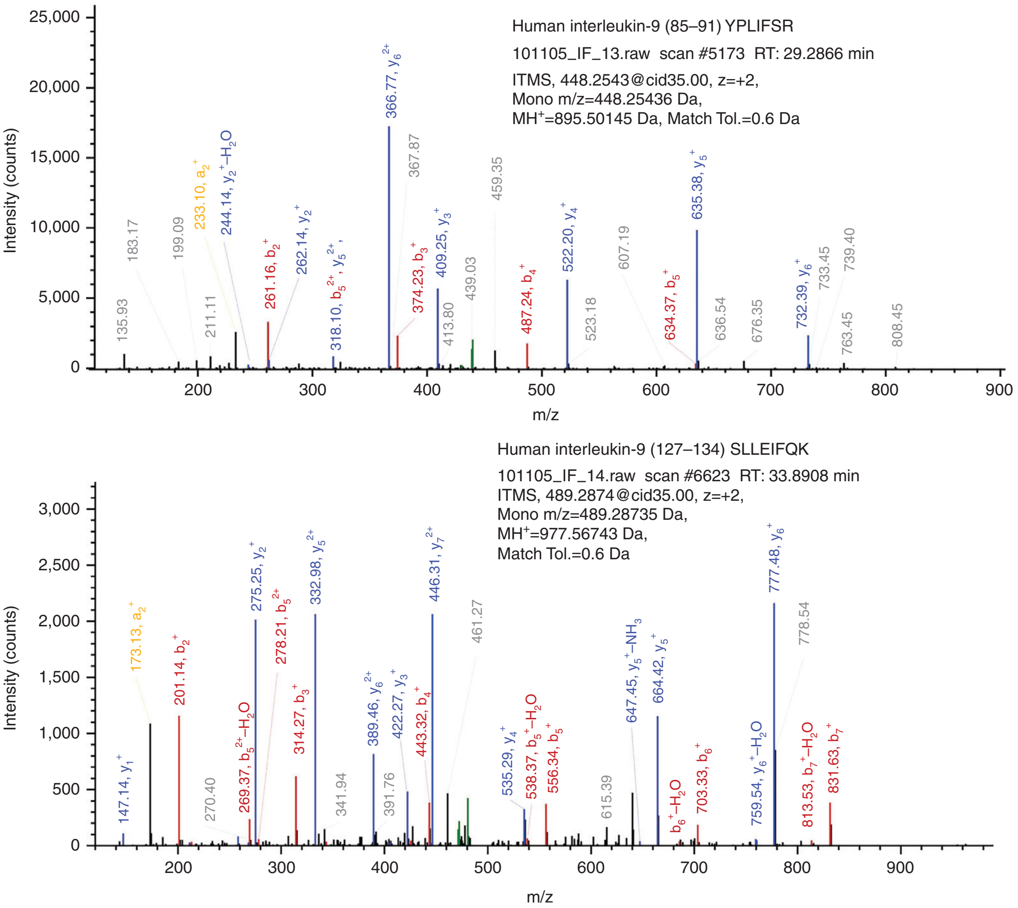


Figure 6 | Fragmentation spectra of two IL-9 proteotypic peptides detected in the secretome of activated T-cells. Although not yet observed in any publicly available MS databases, both of these peptides are predicted to be proteotypic by neXtProt Unicity checker (https://search.nextprot.org/viewers/unicity-checker/app/index.html).

the mammalian inner ear[46]. This presents three challenges; highly specific tissue of origin, low copy number and membrane localization. OHCs are relatively few in number and are in the minority of the cells of the cochlea[47], requiring specialized techniques such as laser capture microdissection to capture cells from very thin tissue sections. Each cochlear microdissection performed by Anderson *et al.*[47] found only 200–300 OHCs per human being, far below the number required for routine proteomic analysis, let alone those involving OHC plasma membrane preparations. Equally, we know that membrane proteins are notoriously resistant to purification and identification by traditional techniques; requiring specialized enrichment strategies due to low copy number per cell, high-hydrophobicity and potential shielding of tryptic cleavage sites by either co-localized membrane proteins or the lipid bilayer itself. It is understandable why prestin is currently a PE2 (transcript evidence only) protein, even though 10 synthetic 10-28mer proteotypic peptides have been reported in neXtProt[45], but no endogenous peptides have yet been captured experimentally by MS.

Interleukin 9 an example missing protein

A number of small biologically active secretory proteins risk being overlooked primarily because of their typical low abundance *in vivo* (in particular relative to the extremely high level of extracellular 'background' proteins), in combination with a specific spatiotemporal expression/secretion profile, a very limited number of predicted potential proteotypic peptides and a relatively high ratio of post-translationally modified residues. One obvious example is the MS detection of interleukin-9 (IL-9) in secretome analysis of post-activation primary cultured T cells. Previous studies of the secretome of cells *ex vivo* had never identified IL-9, as they typically involve only short culture times. To facilitate secretome analysis, typical studies analyse cells grown in serum-free media, inevitably generating considerable cellular stress (with many stress- and apoptosis-related proteins detected). When we analysed cells grown for several days in the presence of foetal bovine serum (described in Supplementary Note 1), a very high percentage (≈95%) of detected tryptic peptides from the conditioned media proteins are evidently of bovine serum origin. After exclusion of bovine proteins and human T cell secretory proteins released from control 'resting' (non-activated) cells, many other secretory proteins (for example, missing interleukins) are now exclusively detected from activated cells. Among these is the 125 amino acid residue, currently PE2 protein, IL-9. MS analyses reveal that IL-9 generates two proteotypic peptides of 7 and 8 residues, respectively (Fig. 6). Subsequent deposition of this and similar data into ProteomeXchange with annual communal re-analysis with stringent criteria will result in the re-classification of IL-9 as PE1. Similar discoveries accompanied with appropriate MS data deposition are expected to result in the re-classification of PE2-4 missing proteins that are unable to generate any proteotypic peptides acceptable to the HPP metrics, yielding a dramatic increase in the rates of discovery of missing proteins.

Complementary efforts to characterize missing proteins

At present, there are also unrelated efforts (for example, Antibodypedia) to capture standardized, non-HPA affinity reagent data. Abs represent the main thrust one of the three pillars of the HPP initiative, and Ab-based techniques (for example Ab-enrichment, immunohistochemistry, western blot) support the search for the PE2-4 missing proteins[48]. However, issues around validity of Ab data have recently been raised across many forums, including this journal[49]. Key problems revolve around selectivity, acceptability and suitability for a given specific application. To facilitate resolving these issues, efforts are being made (for example, Antibodypedia, HPA) to collect, in searchable databases, detailed information concerning Ab validation and their use, and in some cases, literature performance review. Clearly, careful validation of all Abs is mandatory to allow researchers to make informed choices about suitable reagents with the knowledge that they are specific, selective, fit-for-purpose and reproducible in the context for which they are required[50]. Such validation should include western blot, immunohistochemistry, immunofluorescence, flow cytometry and microarrays, and ideally also Surface Plasmon Resonance data with detailed kinetic information. Where possible, the use of gene knockout/gene silencing (for example RNAi, CRISPR/Cas9) to confirm specificity has also been proposed[51]. Both polyclonal Abs (ideally affinity-purified) and monoclonal Abs have their advantages and disadvantages in the search for the PE2-4 missing proteins. Multiple epitopes, accessible by polyclonal Abs, can facilitate targeting specific proteins in complexes where some epitopes may be masked. They do, however, often have higher non-specific background and cannot be replaced once stocks are depleted. Monoclonal Abs, by contrast, are a renewable resource and typically have high affinity, high specificity and reduced non-specific binding[52], while binding only a single epitope. Furthermore, monoclonal Ab libraries against target proteins can be readily generated[53]. For the missing proteins, a further dilemma is how to obtain an appropriate antigen for immunization. A potentially generic approach is the use of a proteospecific recombinant protein fragment and Protein Epitope Signature Tags (PrESTs)[54]. In a recent study, this approach has successfully generated a panel of monoclonal Abs and affinity purified polyclonal Abs against a number of targets, including some missing proteins[48].

MissingProteinPedia

The availability of large volumes of published, peer-reviewed, credible scientific data for PE2-4 proteins outside of high-stringency PE1 MS and Ab-based evidence (for example, IL-9 and prestin) struck us as a resource we could further exploit. Given the need to accelerate the HPP, we contend that the acquisition of such additional data streams concerning the biology of all PE2-4 proteins is self-evident. This has inspired us to explore, create and launch a communal database called MissingProteinPedia. This database assembles in one repository the vast amounts of publicly available, complementary data about all the current PE2-4 proteins that sit outside of the well-justified, high-stringency HPP pipeline. We contend that the knowledge captured by MissingProteinPedia will accelerate the communal HPP effort, as we seek strategies to allow the generation of high confidence MS evidence for as many PE2-4 proteins as possible. In addition, by providing an assembly of all available biological clues in one repository about every single current PE2-4 protein, it is likely that the MissingProteinPedia database may assist C-HPP chromosomal teams that have accepted the ‘Top 50 Missing Protein Marathon Challenge’ launched recently at the 15th HUPO 2016 World Congress in Taipei to successfully identify an additional 50 PE2-4 proteins per chromosome to those already found by high stringency methods.

MissingProteinPedia is an open, comprehensive, communal, evidence-based, searchable and sortable (by chromosome, tissue and keywords) community knowledgebase, addressing the HPP’s PE2-4 proteins. The launch of MissingProteinPedia aims to capture the broadest level of scientific data necessary to increase the rate at which PE2-4 proteins are validated. MissingProteinPedia represents a new community-based proteomics tool, analogous to human genome annotation jamborees[55], where open big data contributions are invited from the broader scientific community regarding evidence for the existence of any missing protein. Unlike the high-stringency HPP data re-analysis, MissingProteinPedia makes no attempt to edit or judge the quality of submitted data, rather utilizing data to expose hidden possibilities not deposited into the current HUPO-accredited databases, including legacy lab books, unpublished works and data found in commercial/protected environments. It is anticipated that MissingProteinPedia collation will reveal clues that will contribute to an acceleration of high quality MS and qualified Ab data that allow confirmation beyond reasonable doubt of many of the current PE2-4 missing proteins. We believe MissingProteinPedia can cooperate and be easily integrated with high-stringency HPP data re-analysis, assisting the completion of the first phase of the HPP on schedule.

In summary, MissingProteinPedia aims to define, summarize and discuss all available data (including single proteotypic MS spectra) for the so-called missing proteins, emphasizing why they may be currently difficult to observe/find, using standard proteomics MS and Ab-based techniques.

Conclusions and the way forward

The HPP was launched in 2010 and since then has grown organically with a general initial phase aimed at providing knowledge about the human proteome parts list. Progress has entailed the formation of a two-pronged strategy (C-HPP and B/D-HPP) culminating in the creation of guidelines and repositories (for example, ProteomeXchange) for MS and Ab-based (for example, HPA) data deposition; metrics for communal, annual MS re-analysis (for example, PeptideAtlas); categorization of the ~20,000 basal components of the human proteome into PE levels (PE1-5; neXtProt); and forums for discussion and communication between research teams (for example, annual HUPO Congresses and HHP workshops).

The controversial release of the two draft human proteome papers[10,11] has compelled researchers to recognize that the HPP is still in its infancy and much remains to be done. This is especially so with regard to the absence of a universally agreed long-term strategy for piloting the project into the future the capture of high-stringency data from all potential MS and Ab sources, capture of the breadth of other scientific human protein data to searchable knowledgebases, and finally the dissemination of the impact and success of the HPP to the public.

Of 20,055 human proteins (neXtProt, February 2016), 16,518 are PE1 (known), a further 2,949 are currently PE2-4 proteins (missing), while 588 PE5 proteins are considered only to be hypothetical. Current PE1-5 assignment strategies do not take into account all other alternative data streams available from the broader scientific community, preferentially relying on high-stringency MS data.

Analysis undertaken herein demonstrates that the rate of progress of the HPP in finding PE1 proteins needs to be accelerated in order to meet proposed HPP decadal plans. To hasten the progress of the current high-stringency HPP engine, we propose to capture other credible scientific data focussing on the PE2-4 missing proteins. This complementary engine is called the MissingProteinPedia and provides clues in the search for missing proteins, learning more about proteins that fall through the cracks of current data re-analysis. It is our hope that the communal MissingProteinPedia tool will allow researchers to better understand where, how, when and why PE2-4 proteins can be found. Capture of high-stringency data will populate the pool of PE1 proteins more readily and efficiently, building our knowledge of what it is to be human in strictly molecular terms.

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD005656.

References

1. Paik, Y. K. *et al.* The Chromosome-centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* **30,** 221–223 (2012).
   Aims to define full set of human proteins encoded by ~20,300 genes, chromosome-by-chromosome including tissue localization, isoforms and PTMs using MS and Abs. First coined term 'missing proteins'.
2. Paik, Y. K. *et al.* Standard guidelines for the Chromosome-centric Human Proteome Project. *J. Proteome Res.* **11,** 2005–2013 (2012).
3. Legrain, P. *et al.* The Human Proteome Project: current state and future direction. *Mol. Cell Proteomics* **10,** M111.009993 (2011).
4. Omenn, G. S. *et al.* Metrics for the Human Proteome Project 2015: progress on the Human Proteome and Guidelines for High-confidence Protein Identification. *J. Proteome Res.* **14,** 3452–3460 (2015).
5. Omenn, G. S. *et al.* Metrics for the Human Proteome Project 2016: progress on identifying and characterizing the human proteome, including post-translational modifications. *J. Proteome Res.* **15,** 3951–3960 (2016).
   Update on HPP annual communal data re-analyses that adopted higher stringency MS metrics for protein evidence (PE1 = two unitypic peptides > 9 residues). HPP (neXtProt version 2016-02) has 16,518 PE1 proteins, with 2,949 PE2-4 missing proteins and 485 reclassified by higher stringency HPP Guidelines v2.0 to reduce false positives.
6. Deutsch, E. W. *et al.* Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1. *J. Proteome Res.* **15,** 3961–3970 (2016).
7. Gaudet, P. *et al.* neXtProt: organizing protein knowledge in the context of human proteome projects. *J. Proteome Res.* **12,** 293–298 (2013).
8. Lane, L. *et al.* neXtProt: a knowledge platform for human proteins. *Nucleic Acids Res.* **40,** D76–D83 (2012).
   Describes neXtProt the human protein-centric knowledge platform that supports and reports the HPP.
9. Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347,** 1260419 (2015).
10. Kim, M. S. *et al.* A draft map of the human proteome. *Nature* **509,** 575–581 (2014).
11. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–587 (2014).
12. Ezkurdia, I., Vazquez, J., Valencia, A. & Tress, M. Analyzing the first drafts of the human proteome. *J. Proteome Res.* **13,** 3854–3855 (2014).
13. Deutsch, E. W. *et al.* State of the human proteome in 2014/2015 as viewed through PeptideAtlas: enhancing accuracy and coverage through the AtlasProphet. *J. Proteome Res.* **14,** 3461–3473 (2015).
14. Elguoshy, A. *et al.* Why are they missing?: bioinformatics characterization of missing human proteins. *J. Proteomics* **149,** 7–14 (2016).
    Recent physicochemical analysis of missing proteins, erroneously including PE5 along with the current PE2-4 missing protein definition. Claim 24% PE2-4 proteins possess hydrophobic transmembrane domains and a significant number do not generate suitable unitypic tryptic peptides.
15. Gillet, L. C. *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell Proteomics* **11,** O111.016717 (2012).
16. Mainland, J. D. *et al.* The missense of smell: functional variability in the human odorant receptor repertoire. *Nat. Neurosci.* **17,** 114–120 (2014).
17. Choong, W. K. *et al.* Informatics view on the challenges of identifying missing proteins from shotgun proteomics. *J. Proteome Res.* **14,** 5396–5407 (2015).
18. Neuhaus, E. M., Mashukova, A., Zhang, W., Barbour, J. & Hatt, H. A specific heat shock protein enhances the expression of mammalian olfactory receptor proteins. *Chem. Senses* **31,** 445–452 (2006).
19. Mashukova, A., Spehr, M., Hatt, H. & Neuhaus, E. M. Beta-arrestin2-mediated internalization of mammalian odorant receptors. *J. Neurosci.* **26,** 9902–9912 (2006).
20. Kang, N. & Koo, J. Olfactory receptors in non-chemosensory tissues. *BMB Rep.* **45,** 612–622 (2012).
21. Flegel, C., Manteniotis, S., Osthold, S., Hatt, H. & Gisselmann, G. Expression profile of ectopic olfactory receptors determined by deep sequencing. *PLoS ONE* **8,** e55368 (2013).
22. Ferrer, I. *et al.* Olfactory receptors in non-chemosensory organs: the nervous system in health and disease. *Front Aging Neurosci.* **8,** 163 (2016).
23. Islam, M. T. *et al.* A systematic bioinformatics approach to identify high quality MS data and functionally annotate proteins and proteomes. *Methods Mol. Biol.* **1549,** 163–176 (2016).
    A simple and intuitive MS evidence workflow for verifying peptides from proteins, along with in silico functional annotation from ProtAnnotator that is integrated into MissingProteinPedia.
24. Ranganathan, S., Khan, J. M., Garg, G. & Baker, M. S. Functional annotation of the human chromosome 7 'missing' proteins: a bioinformatics approach. *J. Proteome Res.* **12,** 2504–2510 (2013).
25. Islam, M. T. *et al.* Protannotator: a semiautomated pipeline for chromosome-wise functional annotation of the 'missing' human proteome. *J. Proteome Res.* **13,** 76–83 (2014).
26. Alexander, S. P. *et al.* The Concise Guide to PHARMACOLOGY 2015/16: Overview. *Br. J. Pharmacol.* **172,** 5729–5743 (2015).
27. Hurley, P. T. *et al.* Functional coupling of a recombinant human 5-HT5A receptor to G-proteins in HEK-293 cells. *Br. J. Pharmacol.* **124,** 1238–1244 (1998).
28. Pasqualetti, M. *et al.* Distribution of the 5-HT5A serotonin receptor mRNA in the human brain. *Brain Res. Mol. Brain Res.* **56,** 1–8 (1998).
29. Rees, S. *et al.* Cloning and characterisation of the human 5-HT5A serotonin receptor. *FEBS Lett.* **355,** 242–246 (1994).
30. Grailhe, R. *et al.* Increased exploratory activity and altered response to LSD in mice lacking the 5-HT(5A) receptor. *Neuron* **22,** 581–591 (1999).
31. Wu, S. *et al.* Group III human metabotropic glutamate receptors 4, 7 and 8: molecular cloning, functional expression, and comparison of pharmacological properties in RGT cells. *Brain Res. Mol. Brain Res.* **53,** 88–97 (1998).
32. Berthele, A. *et al.* Expression of metabotropic glutamate receptor subtype mRNA (mGluR1-8) in human cerebellum. *Neuroreport* **10,** 3861–3867 (1999).
33. Malherbe, P. *et al.* Cloning and functional expression of alternative spliced variants of the human metabotropic glutamate receptor 8. *Brain Res. Mol. Brain Res.* **67,** 201–210 (1999).
34. Stepulak, A. *et al.* Expression of glutamate receptor subunits in human cancers. *Histochem. Cell Biol.* **132,** 435–445 (2009).
35. Tang, F. R. & Lee, W. L. Expression of the group II and III metabotropic glutamate receptors in the hippocampus of patients with mesial temporal lobe epilepsy. *J. Neurocytol.* **30,** 137–143 (2001).
36. Geurts, J. J. *et al.* Expression patterns of Group III metabotropic glutamate receptors mGluR4 and mGluR8 in multiple sclerosis lesions. *J. Neuroimmunol.* **158,** 182–190 (2005).
37. Zhai, J. *et al.* Modulation of lateral perforant path excitatory responses by metabotropic glutamate 8 (mGlu8) receptors. *Neuropharmacology* **43,** 223–230 (2002).
38. Scherer, S. W., Soder, S., Duvoisin, R. M., Huizenga, J. J. & Tsui, L. C. The human metabotropic glutamate receptor 8 (GRM8) gene: a disproportionately large gene located at 7q31.3-q32.1. *Genomics* **44,** 232–236 (1997).
39. O'Dowd, B. F. *et al.* Cloning and chromosomal mapping of four putative novel human G-protein-coupled receptor genes. *Gene* **187,** 75–81 (1997).
40. Lee, J., Hever, A., Willhite, D., Zlotnik, A. & Hevezi, P. Effects of RNA degradation on gene expression analysis of human postmortem tissues. *Faseb J.* **19,** 1356–1358 (2005).
41. Adams, J. W. *et al.* Myocardial expression, signaling, and function of GPR22: a protective role for an orphan G protein-coupled receptor. *Am. J. Physiol. Heart Circ. Physiol.* **295,** H509–H521 (2008).
42. Raine, E. V. *et al.* Gene expression analysis reveals HBP1 as a key target for the osteoarthritis susceptibility locus that maps to chromosome 7q22. *Ann. Rheum. Dis.* **71,** 2020–2027 (2012).
43. Zheng, J. *et al.* Prestin is the motor protein of cochlear outer hair cells. *Nature* **405,** 149–155 (2000).

44. He, D. Z., Lovas, S., Ai, Y., Li, Y. & Beisel, K. W. Prestin at year 14: progress and prospect. *Hear. Res.* **311,** 25–35 (2014).
45. Mistrik, P., Daudet, N., Morandell, K. & Ashmore, J. F. Mammalian prestin is a weak Cl(−)/HCO(3)(−) electrogenic antiporter. *J. Physiol.* **590,** 5597–5610 (2012).
46. Mio, K. *et al.* The motor protein prestin is a bullet-shaped molecule with inner cavities. *J. Biol. Chem.* **283,** 1137–1145 (2008).
47. Anderson, C. T. & Zheng, J. Isolation of outer hair cells from the cochlear sensory epithelium in whole-mount preparation using laser capture microdissection. *J. Neurosci. Methods* **162,** 229–236 (2007).
48. Horvatovich, P. *et al.* Quest for missing proteins: update 2015 on Chromosome-Centric Human Proteome Project. *J. Proteome Res.* **14,** 3415–3431 (2015).
49. Baker, M. Antibody anarchy: a call to order. *Nature* **527,** 545–551 (2015).
50. Bordeaux, J. *et al.* Antibody validation. *Biotechniques* **48,** 197–209 (2010).
51. Barrangou, R. *et al.* Advances in CRISPR-Cas9 genome engineering: lessons learned from RNA interference. *Nucleic Acids Res.* **43,** 3407–3419 (2015).
52. Colwill, K. & Graslund, S. A roadmap to generate renewable protein binders to the human proteome. *Nat. Methods* **8,** 551–558 (2011).
53. Layton, D., Laverty, C. & Nice, E. C. Design and operation of an automated high-throughput monoclonal antibody facility. *Biophys. Rev.* **5,** 47–55 (2012).
54. Larsson, K. *et al.* Multiplexed PrEST immunization for high-throughput affinity proteomics. *J. Immunol. Methods* **315,** 110–120 (2006).
55. Thiele, I. & Palsson, B. Ø. Reconstruction annotation jamborees: a community approach to systems biology. *Mol. Syst. Biol.* **6,** 361–361 (2010).
56. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44,** W242–W245 (2016).
57. Stevens, R. C. *et al.* The GPCR Network: a large-scale collaboration to determine human GPCR structure and function. *Nat. Rev. Drug Discov.* **12,** 25–34 (2013).
58. Niimura, Y. & Gojobori, T. *In silico* chromosome staining: reconstruction of Giemsa bands from the whole human genome sequence. *Proc. Natl Acad. Sci. USA* **99,** 797–802 (2002).

Acknowledgements

We thank Dr M Shaikh for assistance with web engineering the MissingProteinPedia web interface. MTI acknowledges the award of a Macquarie University Research Excellence Scholarship (MQRES). We thank Professor G Omenn, Dr L Lane, Dr E Deutsch, Dr H Cheruku, Ms I Nawar and the HUPO community for helpful discussions during the course of this work. Digital storage and computing was provided by Intersect Australia Space and Time, with acknowledgement to the Australian Government NCRIS Research Data Services (RDS) and National eResearch Collaboration Tools and Resources (Nectar).

Author contributions

M.S.B. conceived MissingProteinPedia. M.S.B., S.R. and E.C.N. planned this study. S.R. and M.S.B. named MissingProteinPedia. S.R. and M.T.I. organized all necessary MissingProteinPedia compute resources. M.T.I. designed, developed and implemented the MissingProteinPedia database, all automated workflows and the community web portal. S.B.A., A.M., M.T.I., M.S.B. and S.R. assembled, interrogated spectra manually, re-analysed and reported olfactory receptor data. E.C.N. coordinated the Australia–New Zealand Chr7 initiative. P.V. contributed analysis of small peptides and new IL-9 MS data. D.C. contributed to the prestin analysis. M.C. provided missing protein pharmacological data review. S.B.A., A.M., M.T.I., D.C., S.S., S.F., S.R. and M.S.B. produced graphics, formatting and referencing. All authors contributed to the writing/reviewing of each version of this manuscript.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Baker, M. S. *et al.* Accelerating the search for the missing proteins in the human proteome. *Nat. Commun.* **8,** 14271 doi: 10.1038/ncomms14271 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

7.3 Publication 6

An integrated nexus for interpreting and using omics data to find human missing proteins

Mohammad T Islam[‡], Mostafa Shaikh[‡], Abidali Mohamedali[‡] , Seong Beom Ahn[†], Mark S. Baker[†] and Shoba Ranganathan[‡,*]

[†]Department of Biomedical Sciences, Faculty of Medicine & Health Sciences and
[‡]Department of Chemistry & Biomolecular Sciences, Macquarie University, NSW, 2109, Australia.

We present ‘Missing ProteinPedia’ web-based protein knowledgebase, a purpose-built community-driven integrated nexus. It offers automatic data capture from various public databases, and user driven data deposition, validation, secure sharing and community annotation or curation service to capture all available evidence for human ‘missing proteins’. We demonstrate that the platform can complement existing C-HPP ‘missing protein’ identification efforts by connecting information from various public and community sources, with sufficient data currently available to even find substantive protein level evidence.

The international Chromosome-centric Human Proteome Project (C-HPP) aims to identify and characterise human proteome with strict baseline metrics[1,2]. Proteins are classified into five protein existence (PE) levels based on any credible experimental information collected by the neXtProt classification workflow, starting with known proteins (PE1) to hypothetical proteins (PE5). Proteins classified as PE2-PE4 (with evidence only at transcript level, inferred from homology and proteins inferred to exist, respectively) are considered the 'missing proteins' [2] a term originally coined by Hancock and colleagues.

The exponential growth of scientific data in the past decade has been unprecedented and extremely challenging not only in terms of accessibility but more so in interpretation. Several high quality genomic and proteomic databases exist but with little or no significant cross integration to allow any meaningful comprehensive interpretation of these vast datasets. This fact is especially stark when, despite the volume of data available worldwide, over 2500 human proteins (~10%) are still considered ‘missing’. The ProteomeXchange initiative[3] to standardize the submission and dissemination of data has been very effective and now consists of 5 member repositories. At present PeptideAtlas and GPMDB reanalyse all major MS datasets from the ProteomeXchange member repositories using the highly stringent guidelines metrics[1]. However, a significant number of datasets are available[4] with great potential[5], and if studied together with other complementary evidences, can increase our understanding of spatiotemporal biology[6] to identify human 'missing' proteins. However, identifying the relevant datasets from various databases as well as accessing and analysing the data from different platforms, experimental protocols, metadata standards, and omics platform is a highly challenging task[5,7]. The recently developed Omics Discovery Index[8] (OmicsDI) provides an index of datasets from 11 repositories. It also maintains a minimum metadata requirement and provides a template for recommended and additional metadata submission for flexibility. This is indeed an excellent resource to identify related datasets, and enables users to shortlist these datasets. However, users are still required to access and analyse data from the source repositories directly. Besides, the protein and metabolite identifiers (i.e. a list of identified proteins per dataset) are non-mandatory additional fields of the metadata template, meaning OmicsDI may contain datasets that are related to some ‘missing’ proteins but without specific protein identifiers, making it very difficult for users to find them. Often repositories have to make the tough decision to balance capturing detailed information and making submission easier for the users. However, it is crucial to identify as many datasets as possible and disseminate them in their most reusable/interpretable format to the community to accelerate the discovery process. While some databases provide basic collaboration features, they rarely support integration with other databases. As the C-HPP ‘missing’ protein identification is a worldwide initiative, advanced scientific collaboration and database crosstalk are key to its success.

‘Missing ProteinPedia (MPP)’ is a web platform (missingproteins.org) that is designed to harmonise very disparate data sets from various disciplines and databases by automatic data integration and community contribution to provide a single interface for users to find information about ‘missing’ proteins. It provides a public discovery portal and an authenticated data submission portal. It is designed as a multi-community based platform (Figure 1) where each research group has access to their dedicated portal, and the community administrators manage user registration and authorisation processes. Available authorisation levels are *admin* or *read and write* or *read only*. This authorisation layer is implemented to empower the community and is opaque to the MPP system administrators. Community users can log in to the portal and contribute information/evidence (e.g. putative/known functions and inhibitor studies) in free text format for any of the information blocks (see supplementary data for the available data blocks). Each block of information can be kept private or shared with other research group or made open to the wider community using the publish data option. This gives the flexibility to keep data private until its ready to be shared (e.g. journal publication, etc.). MPP provides separate user interface according to the permission level (see supplementary file for more information). The interface also offers an approval before publishing feature to allow lab heads to curate the data before it is shared with the wider community via the public portal. In a highly collaborative environment, it is important to acknowledge authors' contributions to maintain the credibility of the information as well as incentivising the author[9]. To make the citation easier, we implemented a metadata snippet with predicted search function that allows users to cite any work using the "Crossref metadata search API" within MPP (shown in supplementary figures). Users can just perform a free text search, and select one or more references and cite them using few clicks. If the article or dataset is not indexed in CrossRef, users can add the manually.

MPP captures data from various MS specific databases, protein knowledgebases and other complimentary sources. At present, it integrates information from four proteomics databases to capture peptide identification information for all available proteins. The databases are the PRoteomics IDEntifications (PRIDE), the Global Proteome Machine Database (GPMDB), ProteomicsDB, and the MaxQuant DataBase (MaxQB). The MPP then applies our previously published guidelines[10] to check the proteotypicity of each peptide, and filter and sort the peptides based on their search engine scores and provides a community annotation interface for manual spectral quality analysis to validate or rank spectra via a collaborative community effort. It allows three annotations per peptide spectrum to provide the best available identification information from each of these databases. The peptide identification and annotation results can be viewed from the public interface of MPP. In addition, it provides an interface to assist researchers mine their own MS search results to identify any missing proteins in their data, and rank them. Researchers can submit a list of identified protein IDs (instead of full search engine results for confidentiality), and MPP then searches for possible 'missing' proteins and provides the user with options to contribute their data. MPP stores relevant results in reusable formats and doesn't replicate or store any raw data.

Currently MPP captures non-MS information from GeneCards[11], UniProt[12], GeneRIF[13], and functional annotations from ProtAnnotator[14]. We chose to use information from GeneCards as it sources information from 150 databases. MPP collects relevant publications from PubMed and also indexes the MeSH terms for all publications. It provides an interactive browser for users to search, sort and filter publications using MeSH terms and various controlled vocabularies. The PubMed annotation interface allows users to annotate citations based on their relevance to human missing proteins. Expert annotations and verified citations using this interface are given a higher rank. MPP also indexes and provides an interactive browser for all relevant datasets for each protein that are indexed in OmicsDI. In order to accommodate diverse metadata needs, MPP has been developed with a flexible metadata platform. Community administrators can design a new information block or metadata form and submit it to the MPP administrator for approval. Unlike neXtProt, MissingProteinPedia is a low-stringency public database that aims to collect information that is not deposited in C-HPP accredited databases and lets the community judge the quality of the data. MPP is designed to be a community-centric portal. Besides community contributed data and annotation, it also allows an administrator to create a community science campaign and tag proteins under that campaign.

Based on an analysis of PubMed references in MPP, we noticed that a 2015 paper (Chick et al.[15]) mapped to several missing proteins. Furthermore, as required as evidence by C-HPP, the data accruing from that publication had been deposited in ProteomeXchange (http://www.ebi.ac.uk/pride/archive/projects/PXD001468). We then applied our guidelines[10] to check the proteotypicity of each peptide, then to filter and sort the peptides based on their search engine scores followed by manual validation of the spectra to provide the best available identification information for each missing protein. High quality MS results were identified for 18 proteins (listed in Table 1), with two entries (NX_A6NJT0 and NX_Q9Y2G7) having MS evidence got PE1 status compliant with HPP high-stringency metrics (≥ 2 uniquely mapping peptides ≥ 9 residues). Ten more proteins have 1 uniquely mapping peptide ≥ 9 residues. Of these, NX_O15391 has two overlapping peptides of length 9 and 8 residues, which have been merged in Table 1. The spectra for peptides ≥ 9 residues is available in the supplementary data file. Another six have 1 or more uniquely mapping 7-8 residue peptides. Nine of these peptides are supported by identical SRM peptides (some as short as 8 residues) while another two peptides have partial overlap with SRM data. In summary, two proteins currently qualify for PE1 status according, while 16 others have some high quality MS evidence, from a single ProteomeXchange entry.

We demonstrate that high-quality data on missing proteins exists in publicly available databases and combining the information captured or contributed via this platform with the information available in other proteomics databases may give clues to discovering them. Furthermore, scientists can use their own datasets in MPP and demonstrate that either by using very high resolution MS, or analysing difficult to access tissues (e.g. ovary, brain, etc.) to accurately identify missing proteins. This platform focuses the collaborative efforts of the proteomics community to complement the existing C-HPP effort to accelerate the discovery of missing proteins.

Acknowledgements

MTI acknowledges the award of a Macquarie University Research Excellence Scholarship (MQRES). Digital storage and computing was provided by Intersect Australia Space and Time, with acknowledgement to the Australian Government NCRIS Research Data Services (RDS) and National eResearch Collaboration Tools and Resources (Nectar). Joel Chick and Mehdi Mirzaei.

Authors' Contributions

SR and MTI. organized all necessary MissingProteinPedia compute resources. MTI. designed and implemented all automated workflows, and pipelines and the MissingProteinPedia database. MTI and MS developed the MissingProteinPedia database, web platform, and data visualisations. All authors assembled, interrogated spectra manually, re-analysed and reported MS missing protein data, produced graphics, formatting and referencing. All authors contributed to the writing the manuscript.

Declaration of competing financial interests

All authors declare they have no competing or other interests that might be perceived to influence the results and/or discussion reported in this article.

References

1 Omenn, G. S. *et al.* Metrics for the Human Proteome Project 2016: Progress on Identifying and Characterizing the Human Proteome, Including Post-Translational Modifications. *J Proteome Res* **15**, 3951-3960, doi:10.1021/acs.jproteome.6b00511 (2016).

2 Deutsch, E. W. *et al.* Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1. *J Proteome Res* **15**, 3961-3970, doi:10.1021/acs.jproteome.6b00392 (2016).

3 Deutsch, E. W. *et al.* The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res* **45**, D1100-D1106, doi:10.1093/nar/gkw936 (2017).

4 Perez-Riverol, Y., Alpi, E., Wang, R., Hermjakob, H. & Vizcaino, J. A. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. *Proteomics* **15**, 930-949, doi:10.1002/pmic.201400302 (2015).

5 Martens, L. & Vizcaino, J. A. A Golden Age for Working with Public Proteomics Data. *Trends Biochem Sci* **42**, 333-341, doi:10.1016/j.tibs.2017.01.001 (2017).

6 Baker, M. S. *et al.* Accelerating the search for the missing proteins in the human proteome. *Nat Commun* **8**, 14271, doi:10.1038/ncomms14271 (2017).

7 Barone, L., Williams, J. & Micklos, D. Unmet Needs for Analyzing Biological Big Data: A Survey of 704 NSF Principal Investigators. *bioRxiv*, doi:10.1101/108555 (2017).

8 Perez-Riverol, Y. *et al.* Discovering and linking public omics data sets using the Omics Discovery Index. *Nat Biotechnol* **35**, 406-409, doi:10.1038/nbt.3790 (2017).

9 Credit where credit is overdue. *Nat Biotechnol* **27**, 579, doi:10.1038/nbt0709-579 (2009).

10 Islam, M. T. *et al.* A Systematic Bioinformatics Approach to Identify High Quality Mass Spectrometry Data and Functionally Annotate Proteins and Proteomes. *Methods Mol Biol* **1549**, 163-176, doi:10.1007/978-1-4939-6740-7_13 (2017).

11 *GeneCards®: The Human Gene Database*, <http://www.genecards.org> (

12 The UniProt, C. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**, D158-D169, doi:10.1093/nar/gkw1099 (2017).

13 Mitchell, J. A. *et al.* Gene indexing: characterization and analysis of NLM's GeneRIFs. *AMIA Annu Symp Proc*, 460-464 (2003).

14 Islam, M. T. *et al.* Protannotator: a semiautomated pipeline for chromosome-wise functional annotation of the "missing" human proteome. *J Proteome Res* **13**, 76-83, doi:10.1021/pr400794x (2014).

15 Chick, J. M. *et al.* A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat Biotechnol* **33**, 743-749, doi:10.1038/nbt.3267 (2015).

Table 1: Currently assigned PE2-4 missing proteins with manually-curated, best available proteotypic peptide MS spectra.

Details of neXtProt identifier (ID), chromosome (Chr) number, peptide identified, with its length and start and end positions on protein sequence, number of MS observations (Obs.) and SRM peptide positions(s) if any, at the peptide position.

#	neXtProt ID	Chr	Gene Name	Peptides identified	Peptide length	Peptide position	Obs.	SRM peptide position(s)
Proteins that **currently qualify** as PE1 proteins according to neXtProt Feb 2017 HPP high-stringency metrics (≥ 2 uniquely mapping peptides ≥ 9 residues)								
1	NX_A6NJT	7	UNCX	DAASCGPGAAVAA	16	306-331	7	306-321; 329-344
				GGAGLEPAPK	10	398-407	2	-
				TNFTGWQLEELEK	13	110-122	2	110-122
2	NX_Q9Y2G	19	ZFP30	ECGQAFLCSTGLR	13	302-314	1	302-314
				IFTCGSDLR	9	222-230	3	-
Proteins that **previously qualified** as PE1 proteins according to 2015 HPP neXtProt metrics (≥ 2 uniquely mapping peptides ≥ 7 residues or 1 uniquely mapping peptide ≥ 9 residues)								
1	NX_O15391	9	YY2	KFAQSTNLK	9	351-359	2	352-359; 351-359
2	NX_Q8IUZ	1	LRRC49	LLNFQHNFITR	11	116-126	1	116-126
3	NX_Q7Z6R	6	TFAP2D	LGLNLPAGR	9	272-280	2	-
4	NX_Q12999	2	TSPAN31	SQSPTCQMCGEK	12	152-163	6	152-163
5	NX_Q8NEK	19	ZNF548	LVCSMNVGNSLAK	13	504-516	2	504-516
6	NX_Q8N8J	19	ZNF615	LYTCSECGK	9	399-407	2	-
7	NX_Q5T5D	1	ZNF684	SYTVENAYECSECG	15	152-166	1	-
8	NX_Q9UC0	22	ZNF70	GLEQMAVIYK	10	40-49	1	40-49
9	NX_Q03936	7	ZNF92	GGYNGLNQCLTTT	16	128-143	1	128-143
10	NX_A6NGE	X	DCAF8L1	FVYEACGAR	9	172-180	2	-
Proteins that **did not previously and do not currently qualify** as PE1 proteins according to 2016 HPP neXtProt high-stringency metrics (i.e., only 1 uniquely mapping 7-8 residue peptide)								
1	NX_Q9C09	3	DCLK3	LADFGLAK	8	497-504	2	-
				LLVVDPK	7	593-599	6	-
2	NX_P0C091	4	FREM3	SLWLDPLR	8	76-83	1	-
3	NX_P50391	10	NPY4R	ALEFLADK	8	191-198	2	191-198
4	NX_A6NKT	2	RGPD3	ILQNYDNK	8	1384-1391	4	1384-1391
5	NX_P17025	X	ZNF182	STLIIHQR	8	387-394	1	-
6	NX_Q3MJ6	6	ZSCAN23	SLIQVLGK	8	208-215	1	-

* overlapping peptides with overlap underlined

Figures

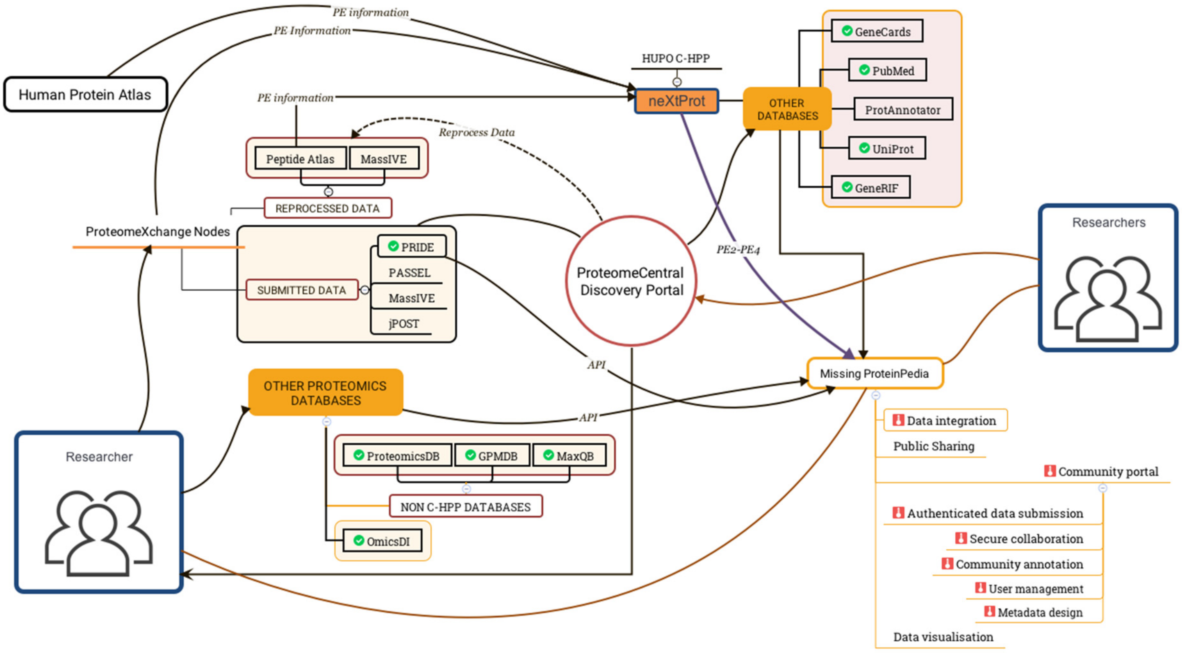


Figure 1: The MissingProteinPedia (MPP) collates and displays protein information from existing databases using various web services. The MPP web interface allows researchers deposit data, annotate information and to collaborate and share data that are not available C-HPP endorsed databases.

An integrated nexus for interpreting and using omics data to find human missing proteins

Mohammad T Islam[‡], Mostafa Shaikh[‡], Abidali Mohamedali[‡] , Seong Beom Ahn[†], Mark S. Baker[†] and Shoba Ranganathan[‡,*]

[†]Department of Biomedical Sciences, Faculty of Medicine & Health Sciences and

[‡]Department of Chemistry & Biomolecular Sciences, Macquarie University, NSW, 2109, Australia.

Supplementary information

Table of Contents

1 MissingProteinPedia data and metadata

MissingPeoteinPedia (MPP) is designed to automatically capture both MS and non-MS information related to human missing proteins from various public databases. In addition, it also allows end users to contribute both MS identification and non-MS (orthogonal) information about missing proteins. MPP does not store any raw data, it only aims to store processed results and metadata related to missing proteins (see Table S1). MPP strongly recommends users to submit raw data directly into the relevant community standard repositories such as ProteomeXchange [1]. At present MPP data capture services are configure to mine (i) orthogonal information from GeneCards [2], UniProt[3], GeneRIF [4] and PubMed [5] including MeSH terms pertaining to each publications; (ii) MS identification information from the PRoteomics IDEntifications (PRIDE) [6], the Global Proteome Machine Database (GPMDB)[7], ProteomicsDB[8], and the MaxQuant DataBase (MaxQB) [9]; (iii) and indexes multi-omics metadata related to the datasets from eleven different repositories using the Omics Discovery Index [10] service, these datasets are PRIDE, PeptideAtlas [11], MassIVE [12] and GPMDB for proteomics datasets; MetaboLights [13], Global Natural Products Social Molecular Networking (GNPS) [14], the Metabolomics Workbench [15]; ArrayExpress [16] and Expression Atlas [17] for transcriptomics datasets, and Metabolome Express[18] for metabolomics datasets; and European Genome-Phenome Archive (EGA) [19] for genomics and phenotypic data.

Table S1: List of information sources for MissingProteinPedia (MPP)

Source	[i]Data type	Omics type	Update Frequency[ii]
PRIDE	MS identification, metadata	Proteomics	Monthly
GPMDB	MS identification	Proteomics	Monthly
ProteomicsDB	MS identification	Proteomics	Monthly
MaxQB	MS identification	Proteomics	Monthly
PeptideAtlas	Metadata	Proteomics	Monthly
MassIVE	Metadata	Proteomics	Monthly

[i] Data type is the type of information stored in MPP. MS identification includes protein and peptide identification data including MS search engine results. Metadata includes information about available datasets, including the experimental conditions, tissue/cell type, related citation etc.

[ii] At present MPP is capturing the data monthly to match various update schedules for each of these databases. We aim to provide more regular updates for some of the databases in near future.

Source	[iii]Data type	Omics type	Update Frequency[iv]
MetaboLights	Metadata	Metabolomics	Monthly
GNPS	Metadata	Metabolomics	Monthly
Metabolomics Workbench	Metadata	Metabolomics	Monthly
Metabolome Express	Metadata	Metabolomics	Monthly
EGA	Metadata	Genomics	Monthly
Expression Atlas	Metadata	Transcriptomics	Monthly
ArrayExpress	Metadata	Transcriptomics, genomics, metabolomics	Monthly
UniProt	Metadata	Orthogonal	Fortnightly
GeneCards	Metadata	Orthogonal	Quarterly[v]
GeneRIF	Metadata	Orthogonal	
PubMed	Metadata	Orthogonal	Fortnightly

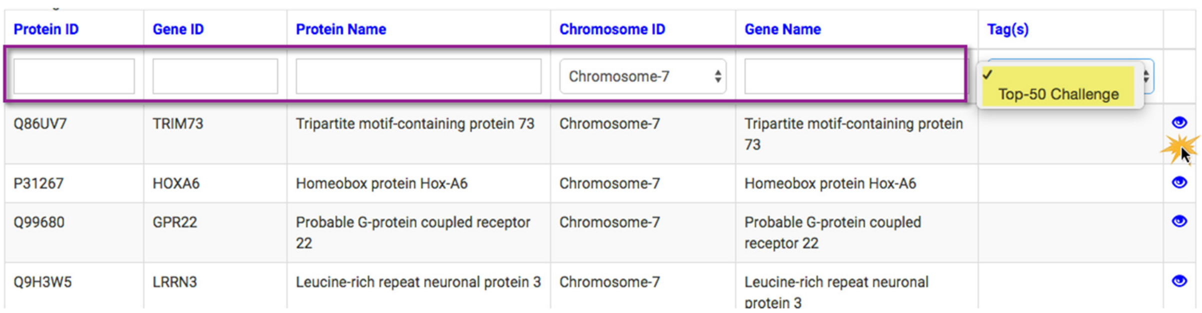
[iii]Data type is the type of information stored in MPP. MS identification includes protein and peptide identification data including MS search engine results. Metadata includes information about available datasets, including the experimental conditions, tissue/cell type, related citation etc.

[iv] At present MPP is capturing the data monthly to match various update schedules for each of these databases. We aim to provide more regular updates for some of the databases in near future.

[v] Aligned with neXtProt missing protein release

2 Data discovery and visualization

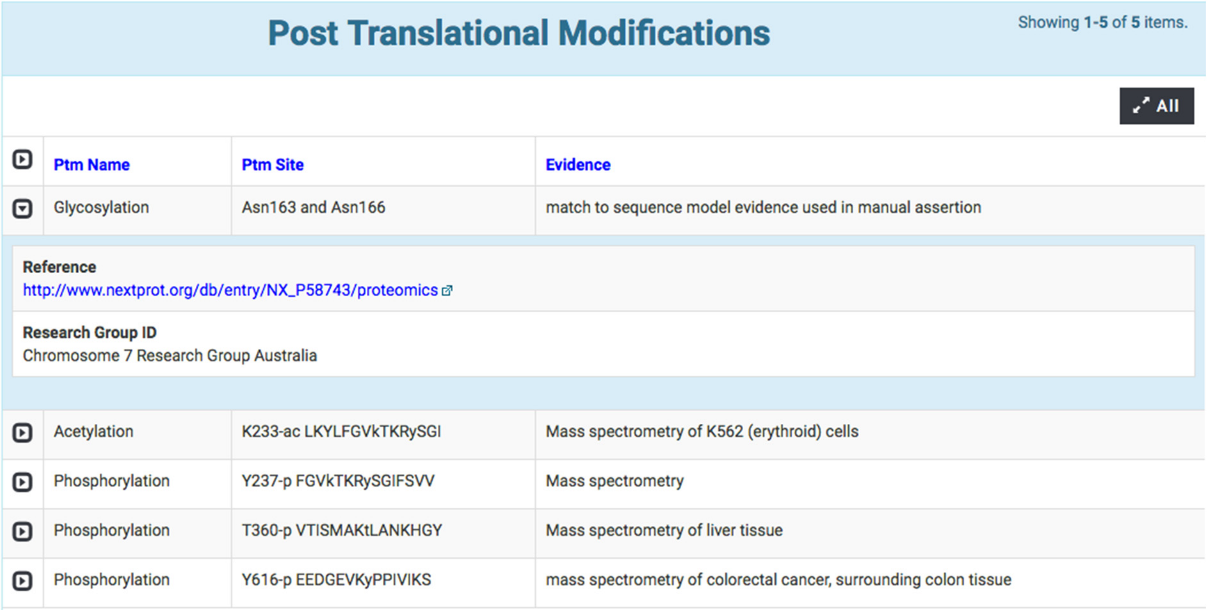
MPP offers a public and private data discovery service via the missingproteins.org web portal. It offers a simple filter, search and browse option to view protein data. Users can filter proteins by chromosome, or any tag (i.e. community campaign, or other tags) then search for proteins using the available controlled vocabularies (see Figure S1). Users need to login to view any privately shared data (see section 3.2).



Protein ID	Gene ID	Protein Name	Chromosome ID	Gene Name	Tag(s)	
			Chromosome-7		✓ Top-50 Challenge	
Q86UV7	TRIM73	Tripartite motif-containing protein 73	Chromosome-7	Tripartite motif-containing protein 73		👁
P31267	HOXA6	Homeobox protein Hox-A6	Chromosome-7	Homeobox protein Hox-A6		👁
Q99680	GPR22	Probable G-protein coupled receptor 22	Chromosome-7	Probable G-protein coupled receptor 22		👁
Q9H3W5	LRRN3	Leucine-rich repeat neuronal protein 3	Chromosome-7	Leucine-rich repeat neuronal protein 3		👁

Figure S1 MissingProteinPedia data discovery options

MPP offers a mix of tabular view and interactive visualisation options to present complex data or metadata to the users. See Figure S2 for an example of a tabular view, and Section 2.1 for some examples of interactive data visualisations



Post Translational Modifications

Showing 1-5 of 5 items.

	Ptm Name	Ptm Site	Evidence
	Glycosylation	Asn163 and Asn166	match to sequence model evidence used in manual assertion

Reference
http://www.nextprot.org/db/entry/NX_P58743/proteomics

Research Group ID
Chromosome 7 Research Group Australia

	Ptm Name	Ptm Site	Evidence
	Acetylation	K233-ac LKYLFGVkTKRySGI	Mass spectrometry of K562 (erythroid) cells
	Phosphorylation	Y237-p FGVkTKRySGIFSVV	Mass spectrometry
	Phosphorylation	T360-p VTISMAKtLANKHGY	Mass spectrometry of liver tissue
	Phosphorylation	Y616-p EEDGEVKyPPIVIKS	mass spectrometry of colorectal cancer, surrounding colon tissue

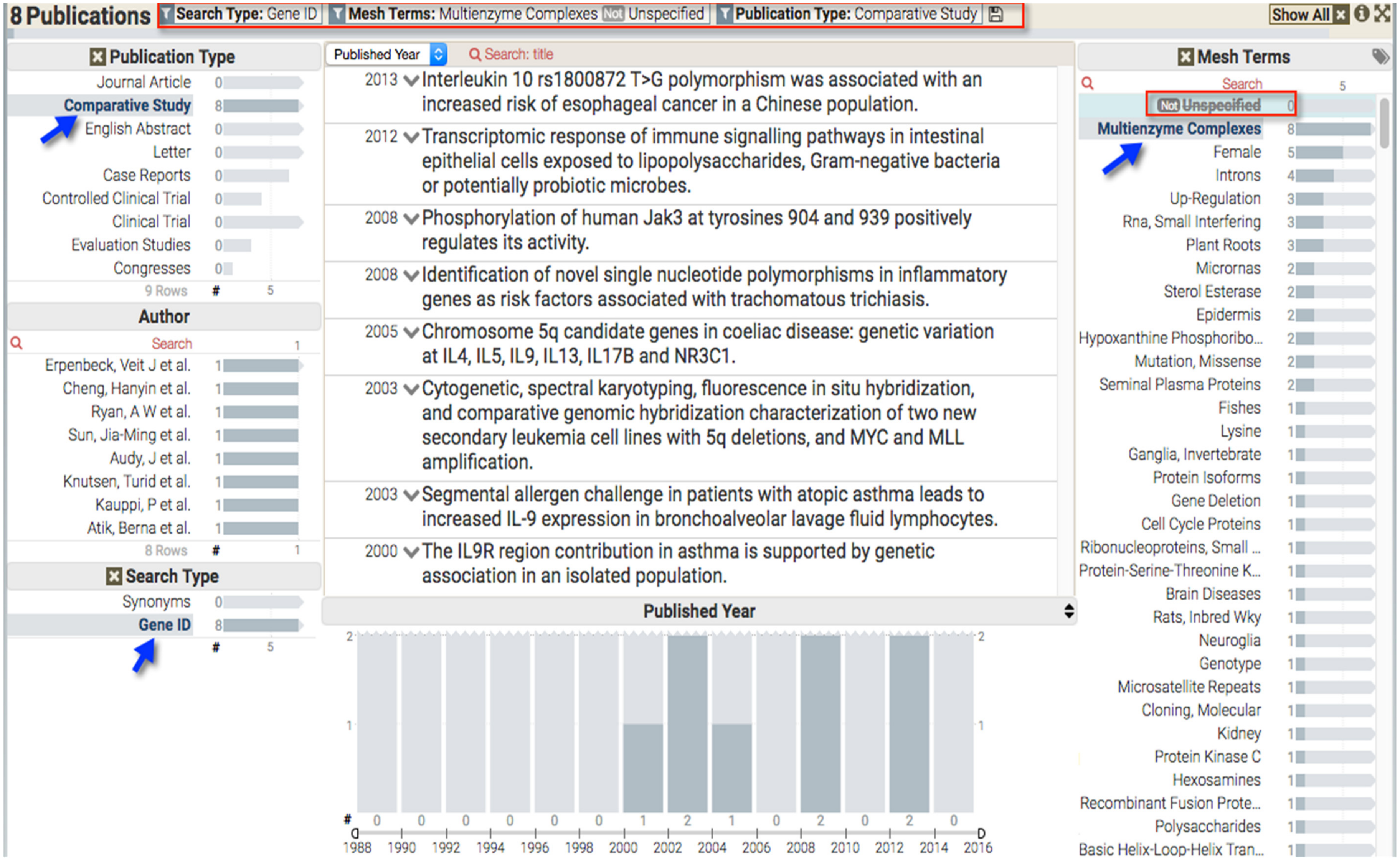
.

Figure S2: Tabular view of an information block

Data visualisation

2.1.1 Orthogonal information and metadata browser

MPP offers several data visualisation options to the end users to easily view and interpret the available data. The interactive PubMed (Figure S3) Citation and multi-omics metadata browser (Figure S4) were developed using Keshif browser [20] underpinned by a Data-Driven document (D3) [21] JavaScript library. These browsers allow users to interact with data, use complex filter to view the specific entries for individual protein entries.



8 Publications | Search Type: Gene ID | Mesh Terms: Multienzyme Complexes Not Unspecified | Publication Type: Comparative Study

Publication Type

Journal Article	0
Comparative Study	8
English Abstract	0
Letter	0
Case Reports	0
Controlled Clinical Trial	0
Clinical Trial	0
Evaluation Studies	0
Congresses	0
9 Rows	#

Author

Erpenbeck, Veit J et al.	1
Cheng, Hanyin et al.	1
Ryan, A W et al.	1
Sun, Jia-Ming et al.	1
Audy, J et al.	1
Knutsen, Turid et al.	1
Kauppi, P et al.	1
Atik, Berna et al.	1
8 Rows	#

Search Type

Synonyms	0
Gene ID	8

Published Year | Search: title

2013 Interleukin 10 rs1800872 T>G polymorphism was associated with an increased risk of esophageal cancer in a Chinese population.

2012 Transcriptomic response of immune signalling pathways in intestinal epithelial cells exposed to lipopolysaccharides, Gram-negative bacteria or potentially probiotic microbes.

2008 Phosphorylation of human Jak3 at tyrosines 904 and 939 positively regulates its activity.

2008 Identification of novel single nucleotide polymorphisms in inflammatory genes as risk factors associated with trachomatous trichiasis.

2005 Chromosome 5q candidate genes in coeliac disease: genetic variation at IL4, IL5, IL9, IL13, IL17B and NR3C1.

2003 Cytogenetic, spectral karyotyping, fluorescence in situ hybridization, and comparative genomic hybridization characterization of two new secondary leukemia cell lines with 5q deletions, and MYC and MLL amplification.

2003 Segmental allergen challenge in patients with atopic asthma leads to increased IL-9 expression in bronchoalveolar lavage fluid lymphocytes.

2000 The IL9R region contribution in asthma is supported by genetic association in an isolated population.

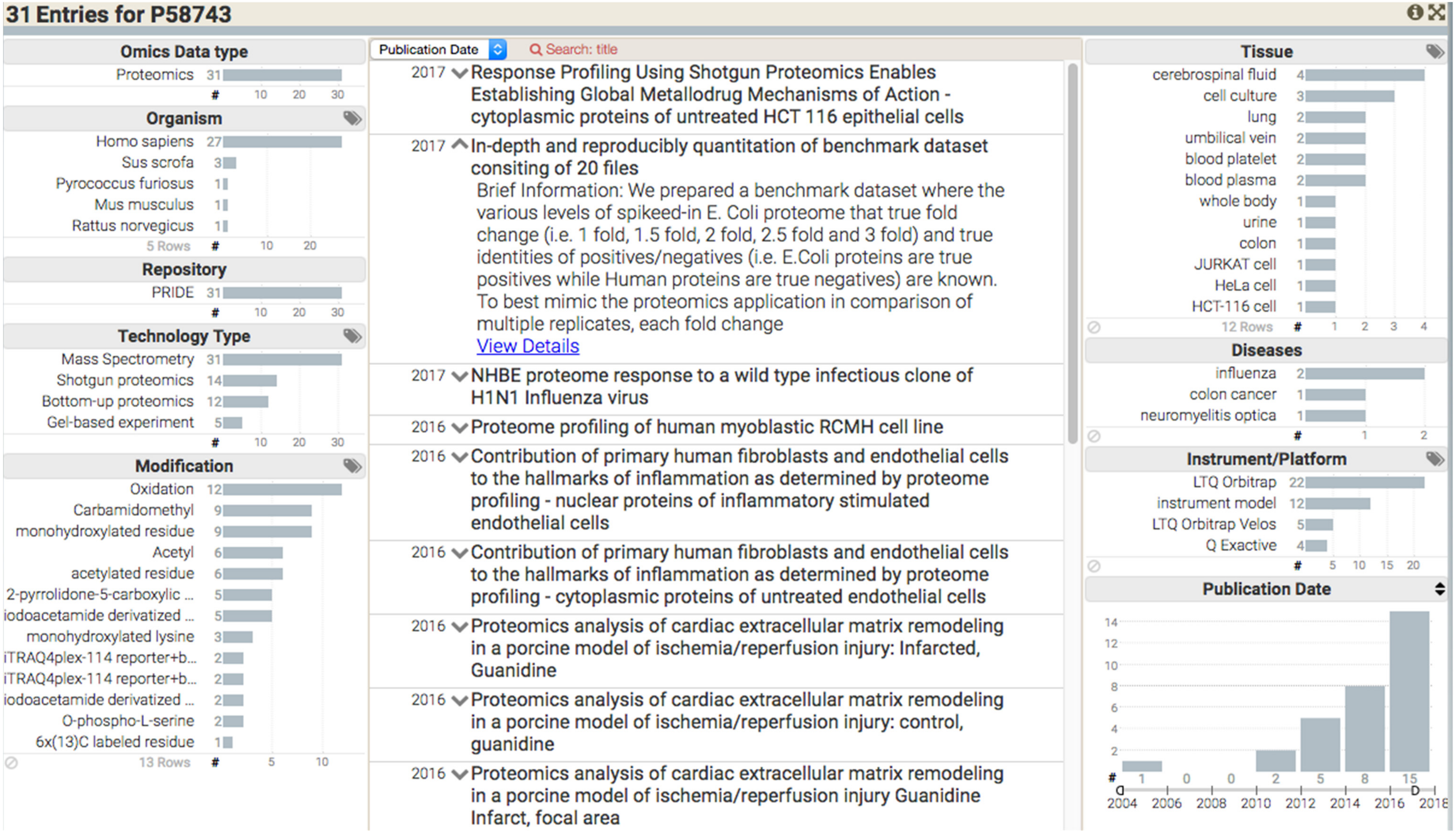
Published Year

#	1988	1990	1992	1994	1996	1998	2000	2002	2004	2006	2008	2010	2012	2014
	0	0	0	0	0	0	1	2	1	0	2	0	2	0

Mesh Terms

Not Unspecified	0
Multienzyme Complexes	8
Female	5
Introns	4
Up-Regulation	3
Rna, Small Interfering	3
Plant Roots	3
Micrornas	2
Sterol Esterase	2
Epidermis	2
Hypoxanthine Phosphoribo...	2
Mutation, Missense	2
Seminal Plasma Proteins	2
Fishes	1
Lysine	1
Ganglia, Invertebrate	1
Protein Isoforms	1
Gene Deletion	1
Cell Cycle Proteins	1
Ribonucleoproteins, Small ...	1
Protein-Serine-Threonine K...	1
Brain Diseases	1
Rats, Inbred Wky	1
Neuroglia	1
Genotype	1
Microsatellite Repeats	1
Cloning, Molecular	1
Kidney	1
Protein Kinase C	1
Hexosamines	1
Recombinant Fusion Prote...	1
Polysaccharides	1
Basic Helix-Loop-Helix Tran...	1

Figure S3: Interactive browser for PubMed citation for a selected protein. The highlighted **sections demonstrate the different filters that are available**



31 Entries for P58743

Omics Data type

Proteomics	31

Organism

Homo sapiens	27
Sus scrofa	3
Pyrococcus furiosus	1
Mus musculus	1
Rattus norvegicus	1

5 Rows

Repository

PRIDE	31

Technology Type

Mass Spectrometry	31
Shotgun proteomics	14
Bottom-up proteomics	12
Gel-based experiment	5

Modification

Oxidation	12
Carbamidomethyl	9
monohydroxylated residue	9
Acetyl	6
acetylated residue	6
2-pyrrolidone-5-carboxylic ...	5
iodoacetamide derivatized ...	5
monohydroxylated lysine	3
iTRAQ4plex-114 reporter+b...	2
iTRAQ4plex-114 reporter+b...	2
iodoacetamide derivatized ...	2
O-phospho-L-serine	2
6x(13)C labeled residue	1

13 Rows

Publication Date | Search: title

2017 Response Profiling Using Shotgun Proteomics Enables Establishing Global Metallodrug Mechanisms of Action - cytoplasmic proteins of untreated HCT 116 epithelial cells

2017 In-depth and reproducibly quantitation of benchmark dataset consiting of 20 files

Brief Information: We prepared a benchmark dataset where the various levels of spikeed-in E. Coli proteome that true fold change (i.e. 1 fold, 1.5 fold, 2 fold, 2.5 fold and 3 fold) and true identities of positives/negatives (i.e. E.Coli proteins are true positives while Human proteins are true negatives) are known. To best mimic the proteomics application in comparison of multiple replicates, each fold change

View Details

2017 NHBE proteome response to a wild type infectious clone of H1N1 Influenza virus

2016 Proteome profiling of human myoblastic RCMH cell line

2016 Contribution of primary human fibroblasts and endothelial cells to the hallmarks of inflammation as determined by proteome profiling - nuclear proteins of inflammatory stimulated endothelial cells

2016 Contribution of primary human fibroblasts and endothelial cells to the hallmarks of inflammation as determined by proteome profiling - cytoplasmic proteins of untreated endothelial cells

2016 Proteomics analysis of cardiac extracellular matrix remodeling in a porcine model of ischemia/reperfusion injury: Infarcted, Guanidine

2016 Proteomics analysis of cardiac extracellular matrix remodeling in a porcine model of ischemia/reperfusion injury: control, guanidine

2016 Proteomics analysis of cardiac extracellular matrix remodeling in a porcine model of ischemia/reperfusion injury Guanidine Infarct, focal area

Tissue

cerebrospinal fluid	4
cell culture	3
lung	2
umbilical vein	2
blood platelet	2
blood plasma	2
whole body	1
urine	1
colon	1
JURKAT cell	1
HeLa cell	1
HCT-116 cell	1

12 Rows

Diseases

influenza	2
colon cancer	1
neuromyelitis optica	1

Instrument/Platform

LTQ Orbitrap	22
instrument model	12
LTQ Orbitrap Velos	5
Q Exactive	4

Publication Date

Figure S4: Interactive browser for multi-omics metadata from various repositories for a selected Protein

2.1.2 MS peptide identification browser

Each protein can have hundreds of peptide evidence entries from various databases (see Figure S5). It is extremely challenging for users to scroll through these entries in a tabular view.

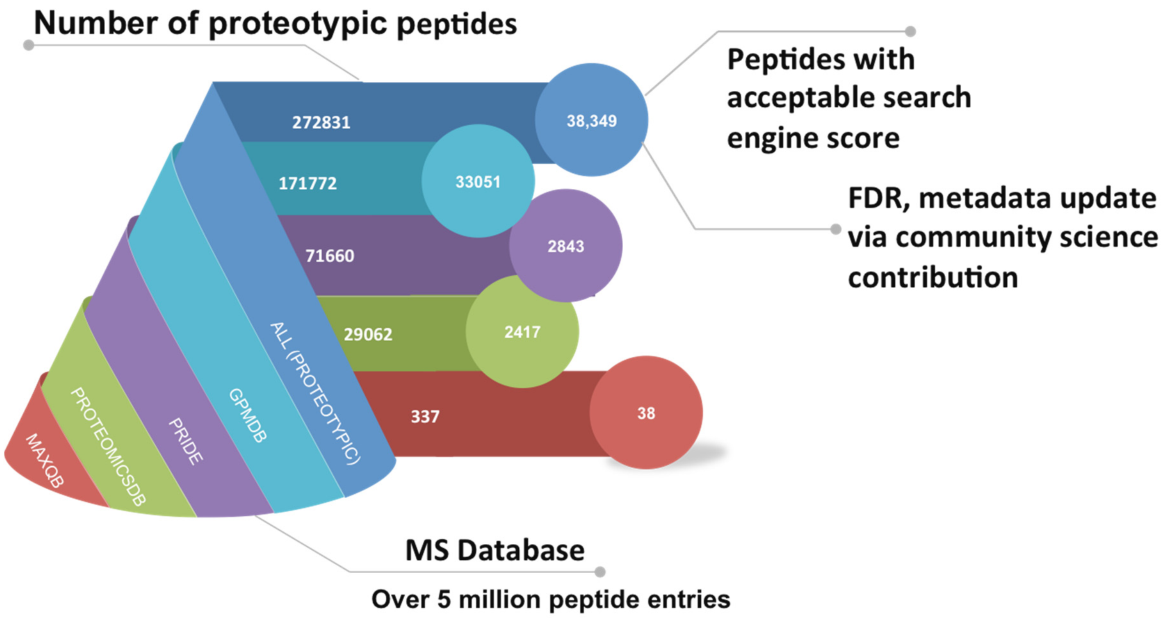


Figure S5 Number of peptide identification entries in MPP at the launch (only proteotypic peptides are shown here) from each of the MS data repositories (in different colour). All collected peptides are filtered first for proteotypicity then, for acceptable search scores, then open for the community to annotate and score the spectra.

MPP combines peptide entries for each protein from all available databases and provides a tabular and an interactive weighted tree view of the aggregated data. The weighted tree graph is created using D3 JavaScript library. A user selects a protein and an interactive weithed graph is generated dynamically. The graph provides an easy to navigate interface to view all available proteins by their proteotypicty, followed by the quality of the search engine score based on our previously published method [22]. The spectra for the shortlisted peptides with high search engine confidence scores can be further validated and annotated by multiple community members or users (refer to section 4.1).

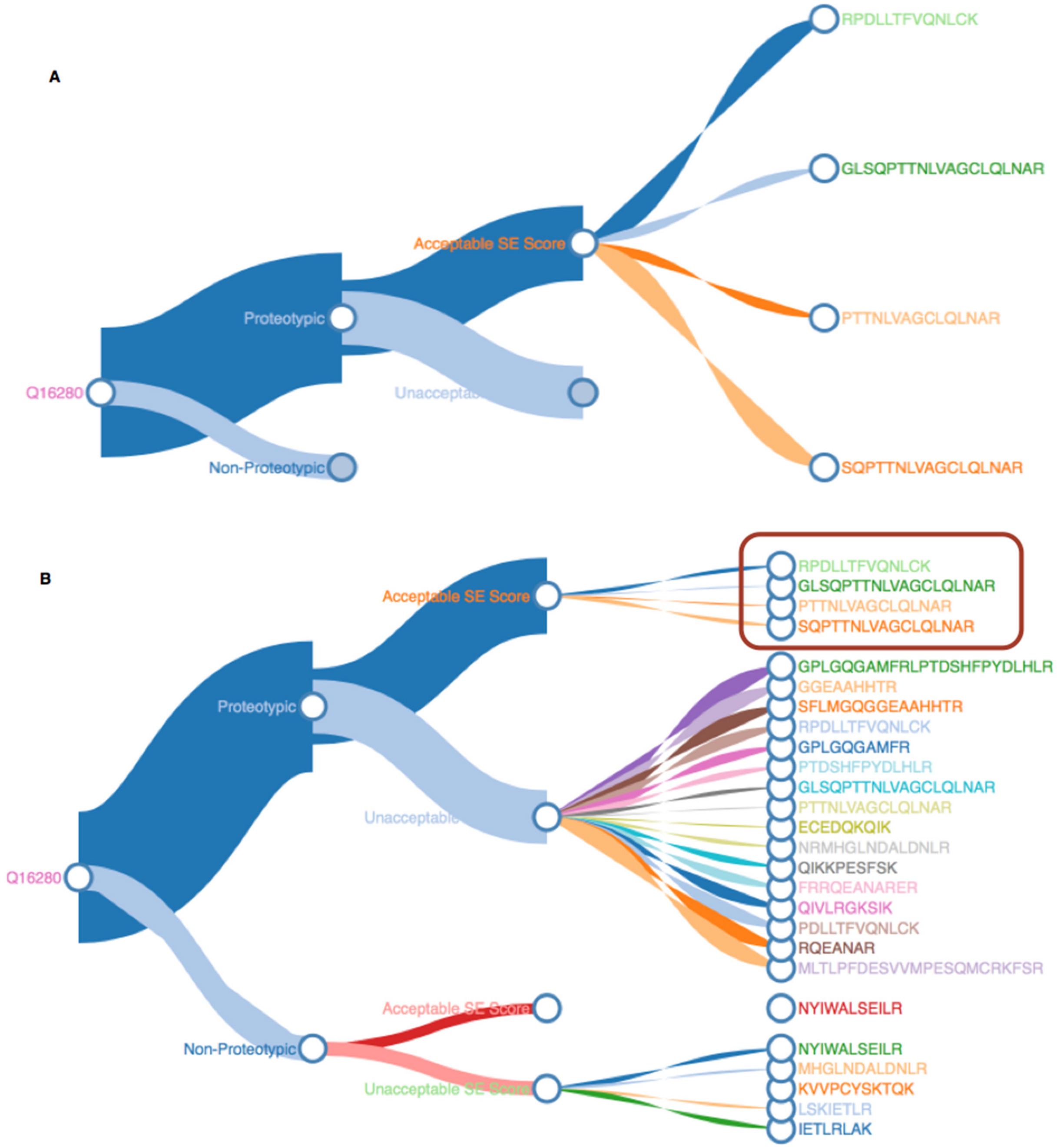
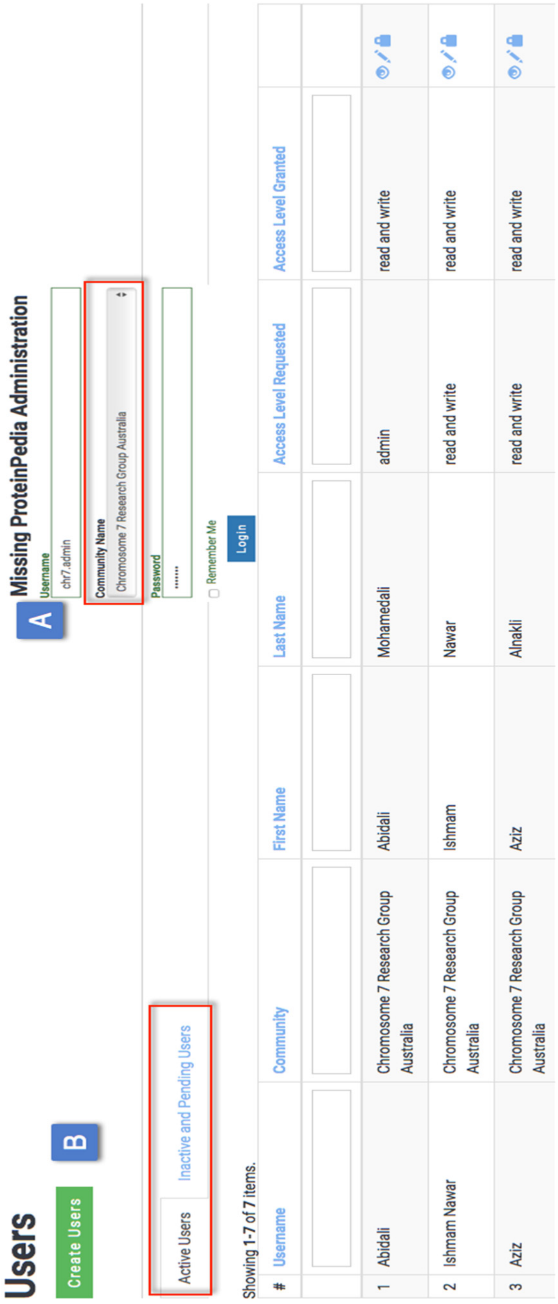


Figure S6: Interactive weighted tree graph view of MS peptide identification information- I mouse click on each node reveals more of the tree

3 User management, data deposition and sharing

3.1 Authentication, authorization

MPP's authentication and authorisation model are based around the research community. Each research community is assigned to one or more administrators by a system administrator. The community administrator is responsible for user management of the respective community and this is opaque to the system administrator (see Figure S7).



Missing ProteinPedia Administration

Username: chr7.admin

Community Name: Chromosome 7 Research Group Australia

Password: •••••••

☐ Remember Me

Login

Users

Create Users

Active Users | Inactive and Pending Users

Showing 1-7 of 7 items.

#	Username	Community	First Name	Last Name	Access Level Requested	Access Level Granted	
1	Abidali	Chromosome 7 Research Group Australia	Abidali	Mohamedali	admin	read and write	
2	Ishmam Nawar	Chromosome 7 Research Group Australia	Ishmam	Nawar	read and write	read and write	
3	Aziz	Chromosome 7 Research Group Australia	Aziz	Alnakli	read and write	read and write	

Figure S7: User management dashboard for a community administrator

MPP currently permits the following roles: admin, read only, read and write, and super admin.

3.2 Data deposition and Sharing

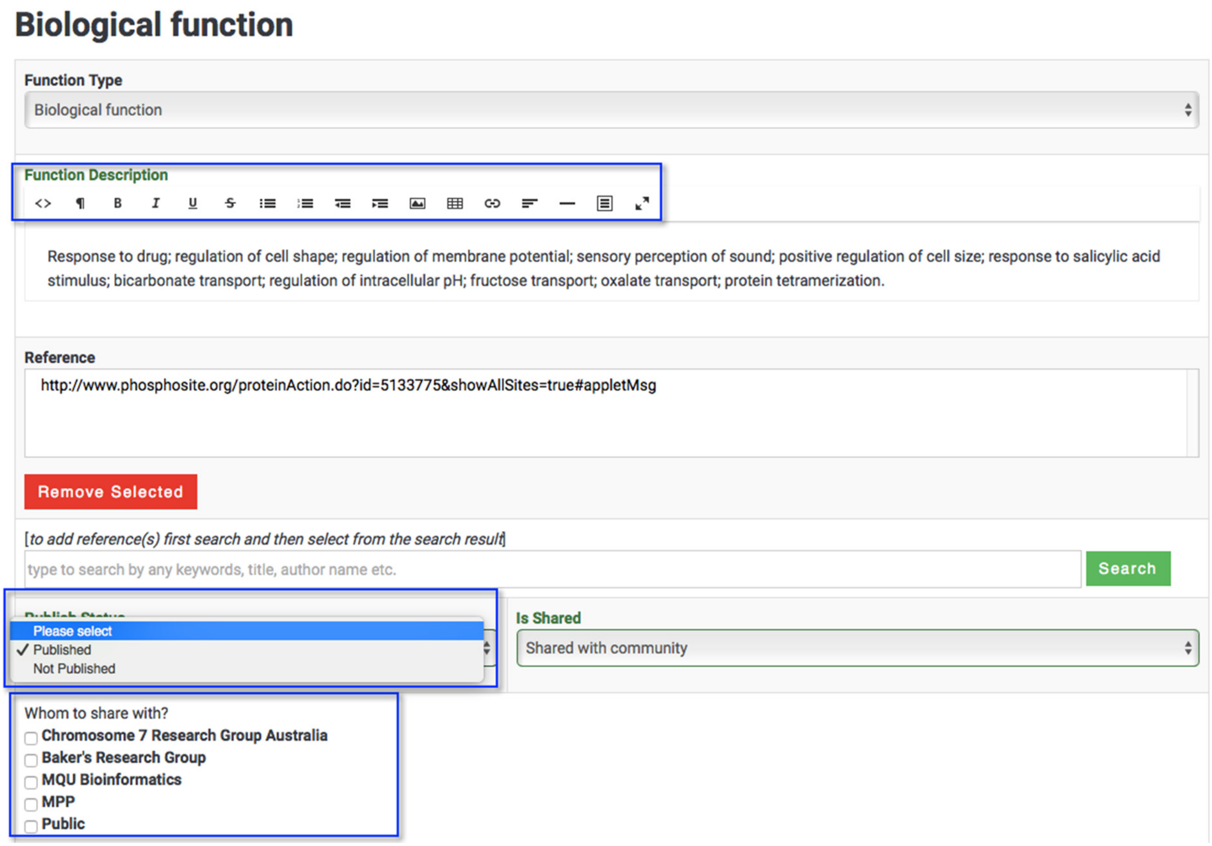
MPP allows secure data deposition and collaboration. The sharing permission applies to each information block which means each information blocks (Table S2) can be shared between different users within or outside the community. Users can enter data using rich text editor (Figure S8), add citations for the added information (Figure S9), then save the data as private or share it with another community. Users can leave the information with “Not Published” status to allow a senior researchers or administrators to approve the change before publishing it.

Table S2: Available information blocks (MS or orthogonal) and the data entry options

Information block	Auto capture	Manual entries	Community annotation
Basic protein informatio (gene name, alternate name, descriptions etc)	Yes	Yes	
Gene Reference Into Function (GeneRIF)	Yes		
Relevant citations within the PubMed literature	Yes	Yes	Yes
Putative/known Functions (generic, biological, molecular, primary secondaty, etc.	Yes	Yes	
Localisation		Yes	
Homologues, Orthologues, Paralogues and Family		Yes	
Sequence Similarity and Functional Annotation	Yes		
Post Translational Modifications		Yes	
Protein Protein Interactions		Yes	
Best Available Mass Spectra without FDR (PRIDE, GPMDB,ProteomicsDB and MaxQB)	Yes		Yes
Lab specific MS data		Yes	Yes (if shared)
Structural Studies		Yes	
Disease Databases		Yes	
Behavioural Studies		Yes	
Chemical Proteomics		Yes	
Knockout Databases		Yes	
Drug Studies		Yes	

3.2.1 Data citation

It is important to credit the original author for the information added in the system. MPP is integrated with "Crossref metadata search API"[23] to make this easier for the end users. A reference snippet is available on each information block for the user to simply use the predictive search function to search for an article, then select and cite it (see Figure S9). If the article or dataset is not indexed in CrossRef, users can add the manually via the interface.



Biological function

Function Type
Biological function

Function Description

Response to drug; regulation of cell shape; regulation of membrane potential; sensory perception of sound; positive regulation of cell size; response to salicylic acid stimulus; bicarbonate transport; regulation of intracellular pH; fructose transport; oxalate transport; protein tetramerization.

Reference
http://www.phosphosite.org/proteinAction.do?id=5133775&showAllSites=true#appletMsg

Remove Selected

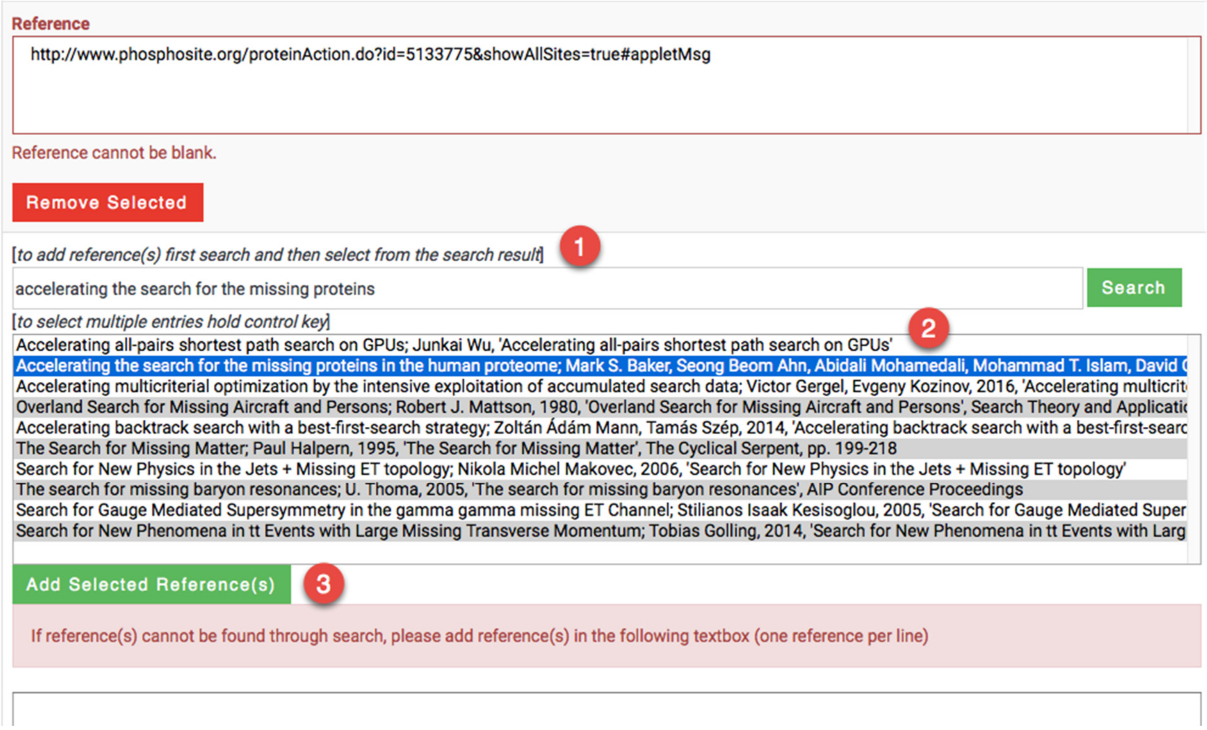
[*to add reference(s) first search and then select from the search result*]
type to search by any keywords, title, author name etc.
Search

Publish Status
Please select
✓ Published
Not Published

Is Shared
Shared with community

Whom to share with?
□ Chromosome 7 Research Group Australia
□ Baker's Research Group
□ MQU Bioinformatics
□ MPP
□ Public

Figure S8 Sample rich text editor data entry form



Reference
http://www.phosphosite.org/proteinAction.do?id=5133775&showAllSites=true#appletMsg

Reference cannot be blank.

Remove Selected

[*to add reference(s) first search and then select from the search result*]
1
accelerating the search for the missing proteins
Search

[*to select multiple entries hold control key*]
2
Accelerating all-pairs shortest path search on GPUs; Junkai Wu, 'Accelerating all-pairs shortest path search on GPUs'
Accelerating the search for the missing proteins in the human proteome; Mark S. Baker, Seong Beom Ahn, Abidali Mohamedali, Mohammad T. Islam, David C
Accelerating multicriterial optimization by the intensive exploitation of accumulated search data; Victor Gergel, Evgeny Kozinov, 2016, 'Accelerating multicrit
Overland Search for Missing Aircraft and Persons; Robert J. Mattson, 1980, 'Overland Search for Missing Aircraft and Persons', Search Theory and Applicati
Accelerating backtrack search with a best-first-search strategy; Zoltán Ádám Mann, Tamás Szép, 2014, 'Accelerating backtrack search with a best-first-searc
The Search for Missing Matter; Paul Halpern, 1995, 'The Search for Missing Matter', The Cyclical Serpent, pp. 199-218
Search for New Physics in the Jets + Missing ET topology; Nikola Michel Makovec, 2006, 'Search for New Physics in the Jets + Missing ET topology'
The search for missing baryon resonances; U. Thoma, 2005, 'The search for missing baryon resonances', AIP Conference Proceedings
Search for Gauge Mediated Supersymmetry in the gamma gamma missing ET Channel; Stilianos Isaak Kesisoglou, 2005, 'Search for Gauge Mediated Super
Search for New Phenomena in tt Events with Large Missing Transverse Momentum; Tobias Golling, 2014, 'Search for New Phenomena in tt Events with Larg

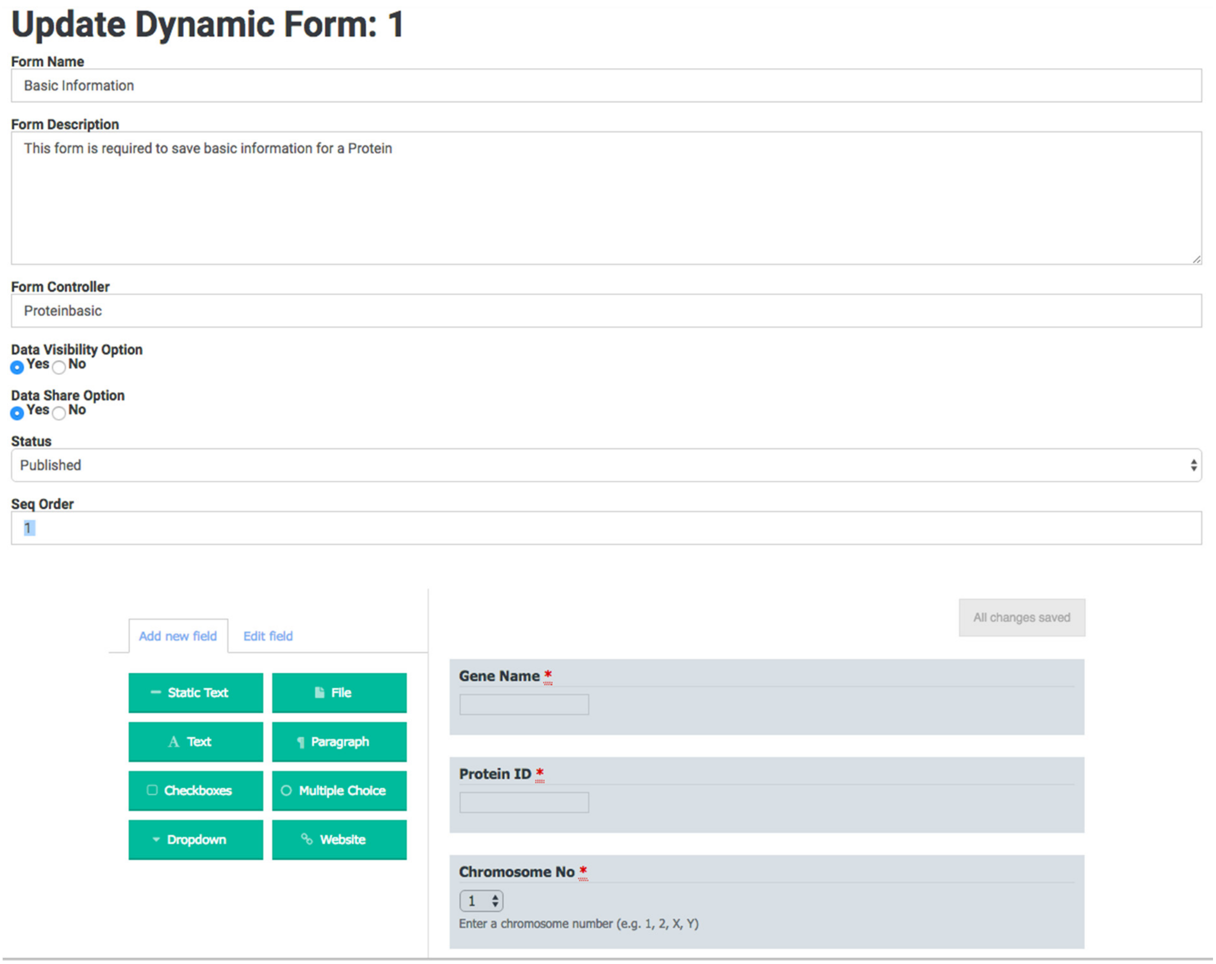
Add Selected Reference(s)
3

If reference(s) cannot be found through search, please add reference(s) in the following textbox (one reference per line)

Figure S9: Adding a reference using cross ref metadata search in 3 steps. Step 1 is to search an article, step 2 is selecting the relevant one (or multiple) from the list and finally step 3 is adding relevant articles**.**

3.3 Flexible metadata schema

MPP allows a wide range of metadata entry with flexible form and information blocks. The MPP aims to collate as much information as possible. Hence it has been developed with a flexible metadata platform to allow the addition of new information blocks. Community administrators can design a new information block using the dynamic form designer within MPP (Figure S10) and send it to the MPP Administrator for approval and implementation.



Update Dynamic Form: 1

Form Name: Basic Information

Form Description: This form is required to save basic information for a Protein

Form Controller: Proteinbasic

Data Visibility Option: ● Yes ○ No

Data Share Option: ● Yes ○ No

Status: Published

Seq Order: 1

Gene Name *

Protein ID *

Chromosome No *: 1
Enter a chromosome number (e.g. 1, 2, X, Y)

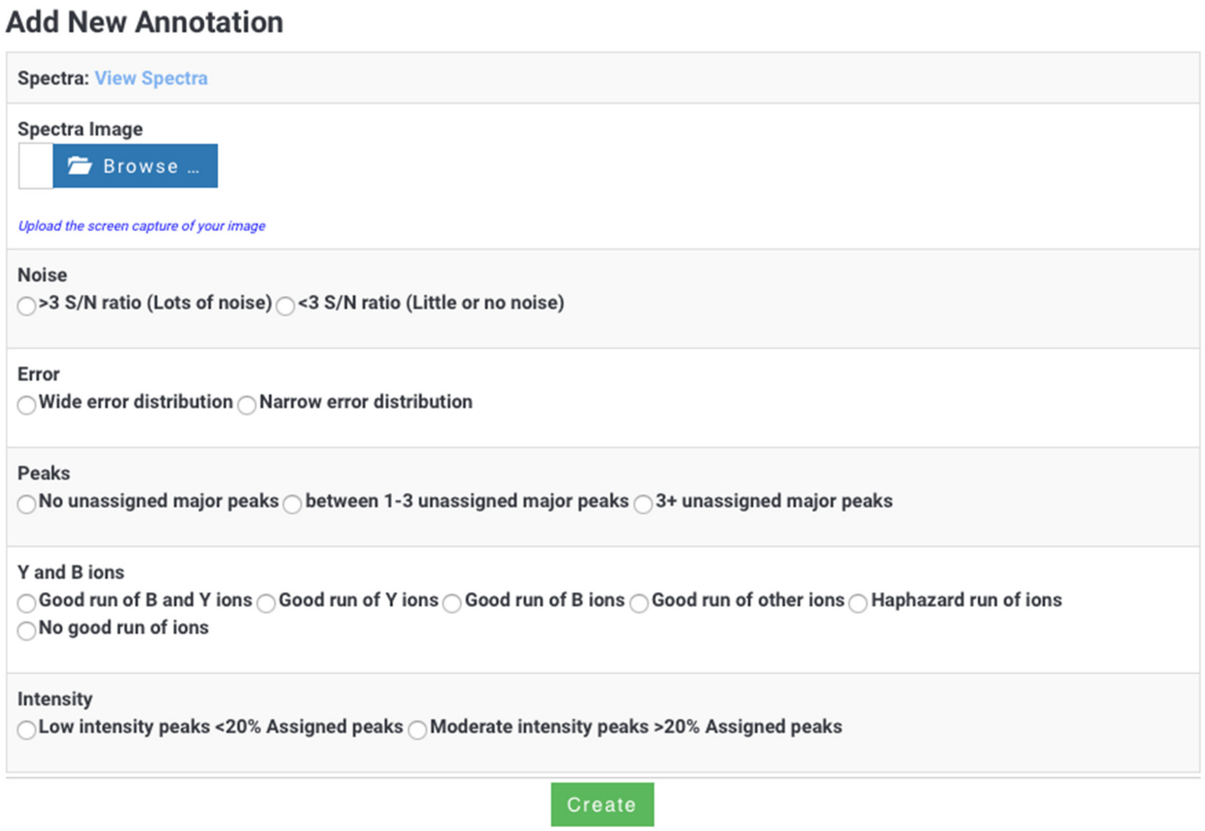
Figure S10: Dynamic metadata form designer

4 Community science and collaboration

MPP aims to provide all available information about a protein in a non-judgemental way. It doesn't apply metrics to the data. Instead, it intends to equip the scientific community with relevant information and tools to allow them to make their own judgment or enrich and annotate the data in a collaborative environment.

4.1 MS peptide identification spectra annotation

MPP sorts all peptide identification data from various databases and provides an interface to filter MS peptide identification data (section 2.1.2, Figure S6), then annotate them using the spectra annotation interface (Figure S11). It allows three curators to annotate a single spectrum and sorts the results based on a simple algorithm where if one annotators feedback is vastly differently form the other 2, the average of the other 2 is taken, whereas if the feedback is widely different for all three, the spectrum is referred to the administrator to arbitrate. The annotated spectra can be viewed via both public and private interface (Figure S12).



Add New Annotation

Spectra: View Spectra

Spectra Image

Upload the screen capture of your image

Noise

○>3 S/N ratio (Lots of noise) ○<3 S/N ratio (Little or no noise)

Error

○Wide error distribution ○Narrow error distribution

Peaks

○No unassigned major peaks ○between 1-3 unassigned major peaks ○3+ unassigned major peaks

Y and B ions

○Good run of B and Y ions ○Good run of Y ions ○Good run of B ions ○Good run of other ions ○Haphazard run of ions
○No good run of ions

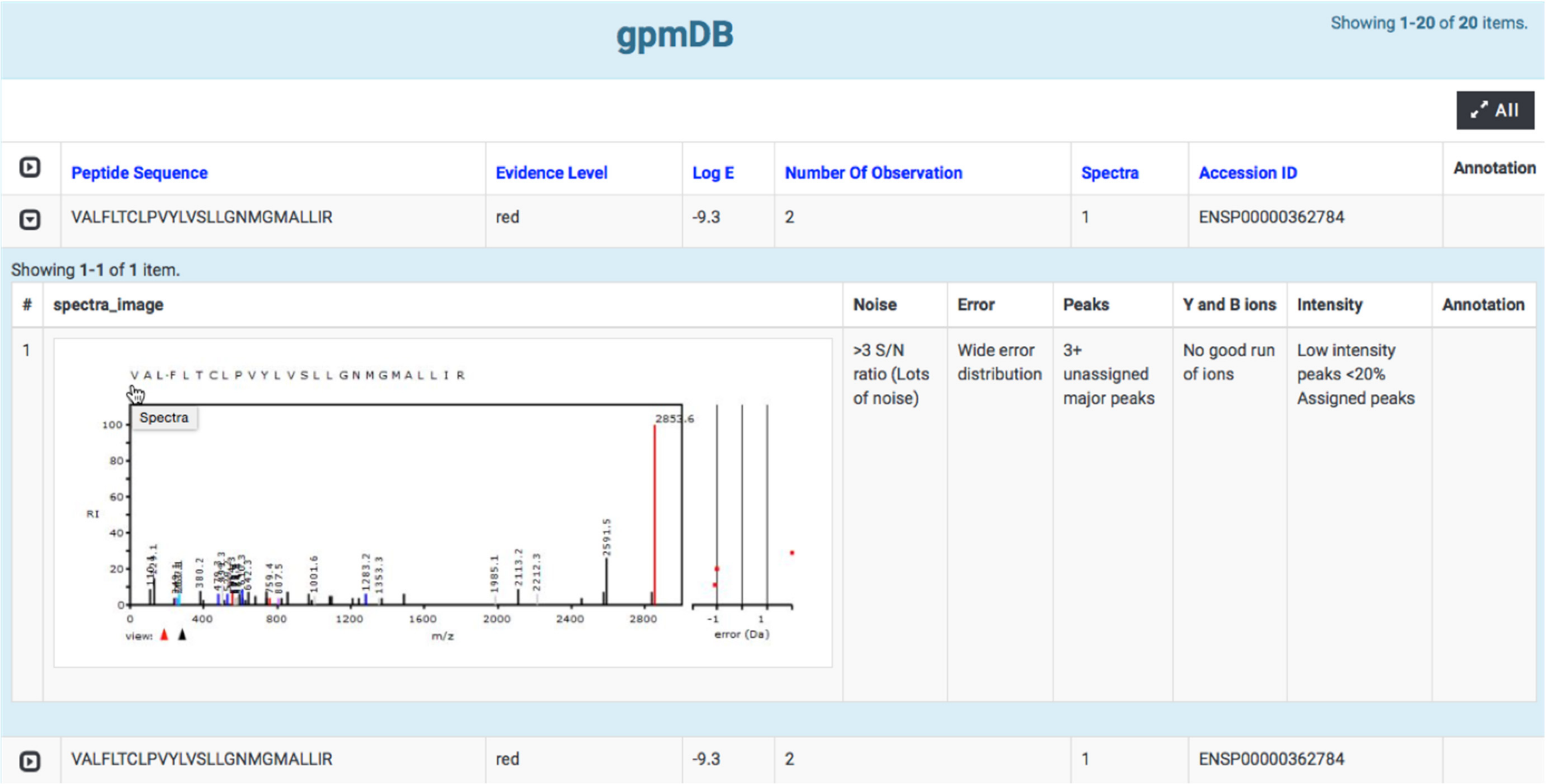
Intensity

○Low intensity peaks <20% Assigned peaks ○Moderate intensity peaks >20% Assigned peaks

Figure S11: MS spectra annotation interface

4.2 PubMed citation annotation

MPP mines relevant citations for a proteins using the PubMed API using gene name, protein name, and synonyms. It then filters out any that are incorrectly returned by the API (e.g. keyword matches) and ranks the citation based on the frequency of the keywords (whole words) used for the search. The user can view the abstract on screen, then select the annotate option to annotate a citation. (Figure S13).



gpmDB

Showing 1-20 of 20 items.

All

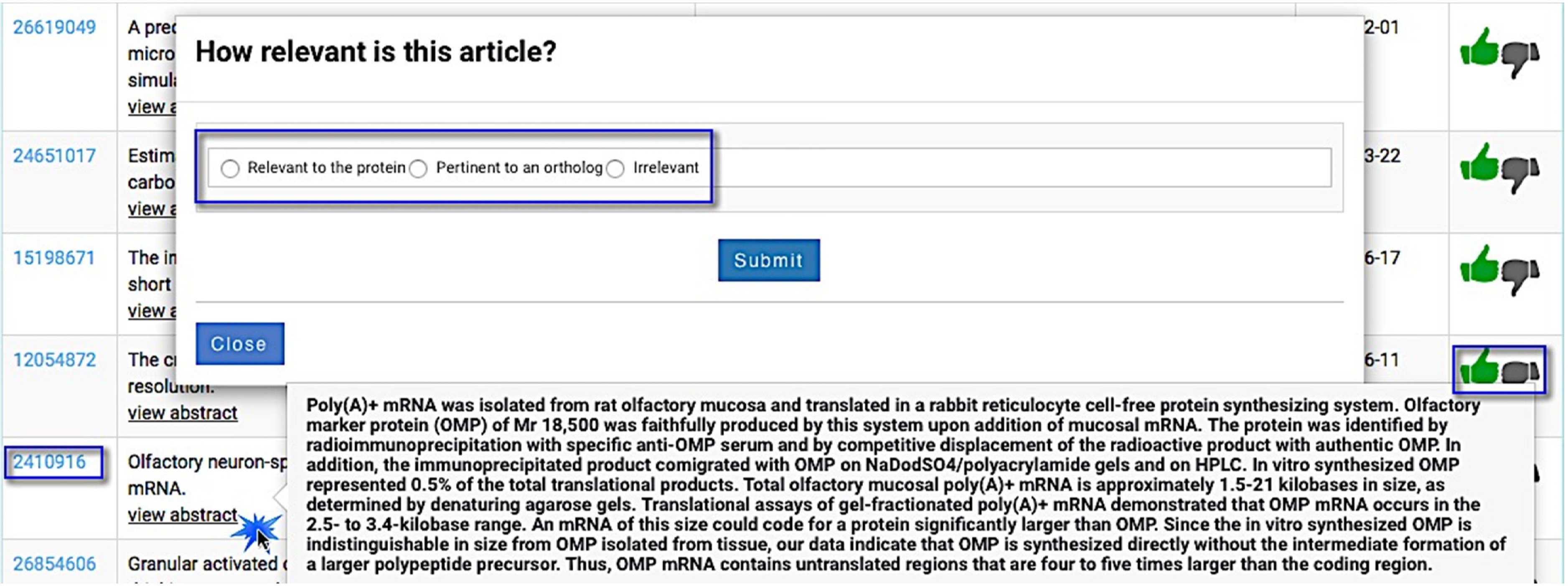
Peptide Sequence	Evidence Level	Log E	Number Of Observation	Spectra	Accession ID	Annotation
VALFLTCLPVYLVSLLGNMGMALLIR	red	-9.3	2	1	ENSP00000362784	

Showing 1-1 of 1 item.

#	spectra_image	Noise	Error	Peaks	Y and B ions	Intensity	Annotation
1		>3 S/N ratio (Lots of noise)	Wide error distribution	3+ unassigned major peaks	No good run of ions	Low intensity peaks <20% Assigned peaks	

Peptide Sequence	Evidence Level	Log E	Number Of Observation	Spectra	Accession ID	Annotation
VALFLTCLPVYLVSLLGNMGMALLIR	red	-9.3	2	1	ENSP00000362784	

Figure S12: Public view of a peptide identification entry community annotated spectra



How relevant is this article?

○ Relevant to the protein ○ Pertinent to an ortholog ○ Irrelevant

Submit

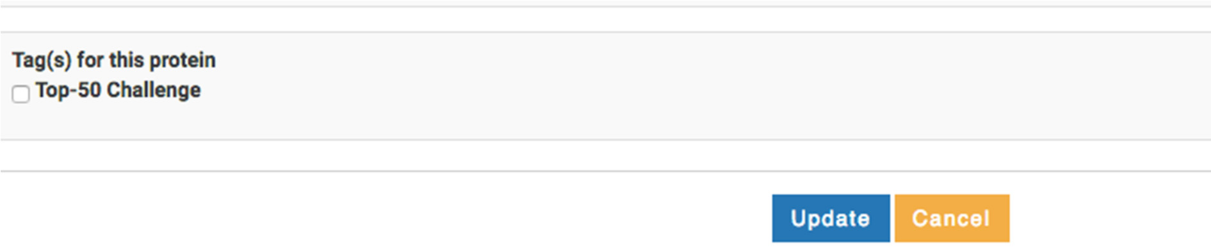
Close

Poly(A)+ mRNA was isolated from rat olfactory mucosa and translated in a rabbit reticulocyte cell-free protein synthesizing system. Olfactory marker protein (OMP) of Mr 18,500 was faithfully produced by this system upon addition of mucosal mRNA. The protein was identified by radioimmunoprecipitation with specific anti-OMP serum and by competitive displacement of the radioactive product with authentic OMP. In addition, the immunoprecipitated product comigrated with OMP on NaDodSO4/polyacrylamide gels and on HPLC. In vitro synthesized OMP represented 0.5% of the total translational products. Total olfactory mucosal poly(A)+ mRNA is approximately 1.5-21 kilobases in size, as determined by denaturing agarose gels. Translational assays of gel-fractionated poly(A)+ mRNA demonstrated that OMP mRNA occurs in the 2.5- to 3.4-kilobase range. An mRNA of this size could code for a protein significantly larger than OMP. Since the in vitro synthesized OMP is indistinguishable in size from OMP isolated from tissue, our data indicate that OMP is synthesized directly without the intermediate formation of a larger polypeptide precursor. Thus, OMP mRNA contains untranslated regions that are four to five times larger than the coding region.

Figure 13: Community annotation interface for PubMed Citation

4.3 Community campaign and protein tagging

A tagging system is created within MPP that allows the system administrator to create a new campaign (such as the "Top 50 Missing Protein Challenge"), and then tag proteins for the specified campaign. This allows the contributing members to track the progress of the campaign. The tagging can also be used for other generic purposes.



Tag(s) for this protein

□ Top-50 Challenge

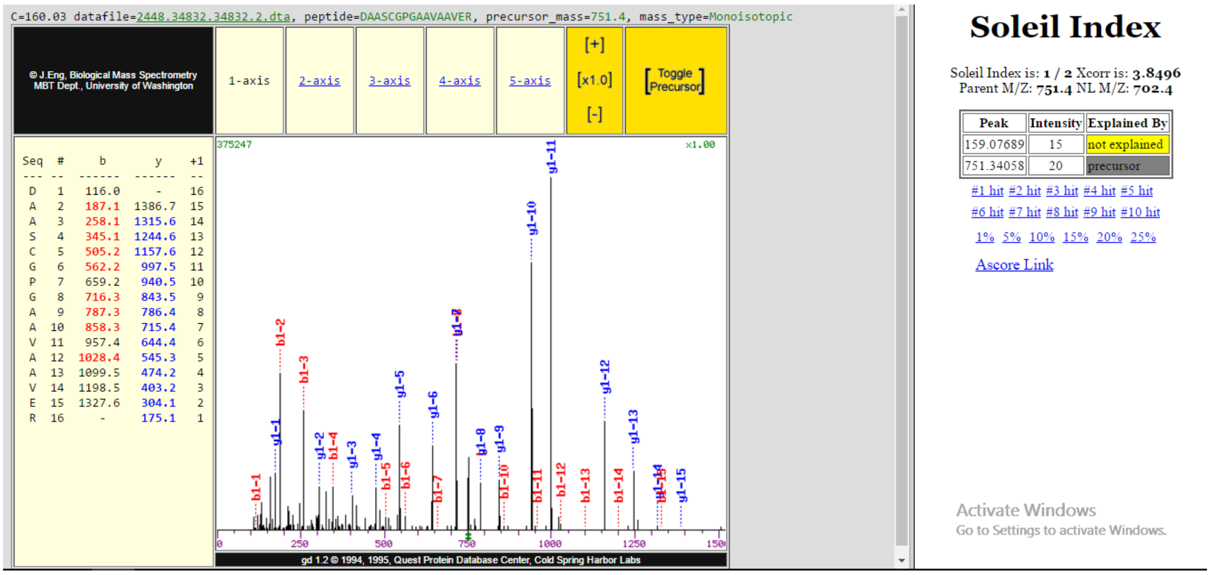
Figure S14: MPP tag a protein for a specific community campaign

5 Best available spectra for the identified MS peptides

5.1 Representative Spectra

1. A6NJT0

1.1 - DAASCGPGAAVAAVER



```
C=160.03 datafile=2448.34832.34832.2.dta, peptide=DAASCGPGAAVAAVER, precursor_mass=751.4, mass_type=Monoisotopic
```

© J.Eng, Biological Mass Spectrometry MBT Dept., University of Washington

```
Seq  #     b       y    +1
--- --  ------  ------  --
 D   1   116.0     -    16
 A   2   187.1  1386.7  15
 A   3   258.1  1315.6  14
 S   4   345.1  1244.6  13
 C   5   505.2  1157.6  12
 G   6   562.2   997.5  11
 P   7   659.2   940.5  10
 G   8   716.3   843.5   9
 A   9   787.3   786.4   8
 A  10   858.3   715.4   7
 V  11   957.4   644.4   6
 A  12  1028.4   545.3   5
 A  13  1099.5   474.2   4
 V  14  1198.5   403.2   3
 E  15  1327.6   304.1   2
 R  16     -     175.1   1
```

gd 1.2 © 1994, 1995, Quest Protein Database Center, Cold Spring Harbor Labs

Soleil Index

Soleil Index is: **1 / 2** Xcorr is: **3.8496**
Parent M/Z: **751.4** NL M/Z: **702.4**

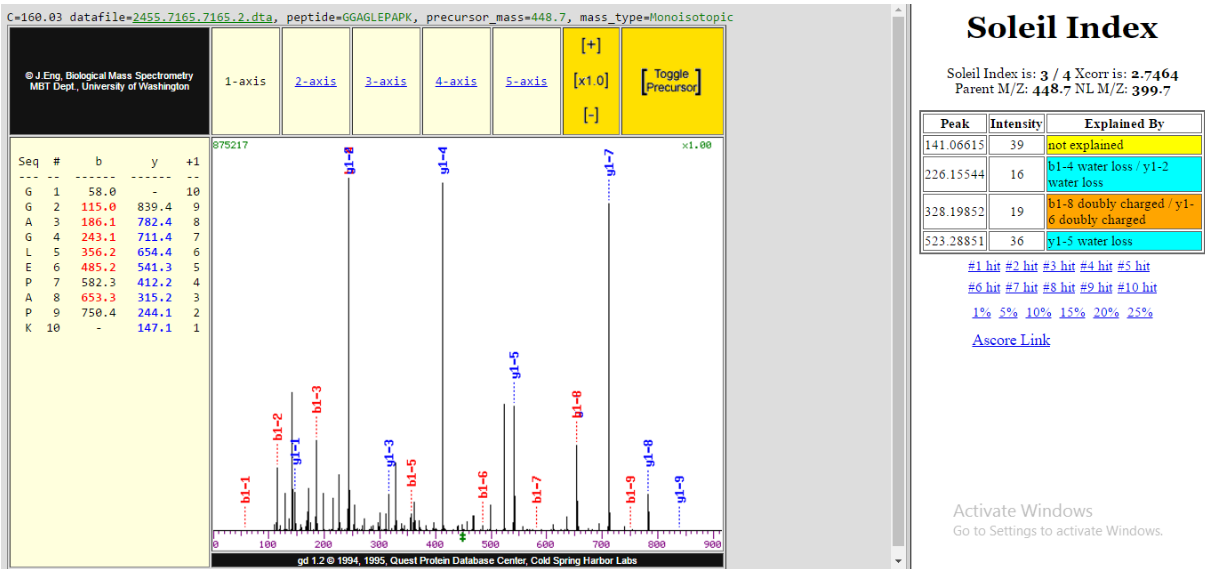
Peak	Intensity	Explained By
159.07689	15	not explained
751.34058	20	precursor

#1 hit #2 hit #3 hit #4 hit #5 hit
#6 hit #7 hit #8 hit #9 hit #10 hit

1% 5% 10% 15% 20% 25%

Ascore Link

1.2- GGAGLEPAPK



C=160.03 datafile=2455.7165.7165.2.dta, peptide=GGAGLEPAPK, precursor_mass=448.7, mass_type=Monoisotopic

© J.Eng, Biological Mass Spectrometry
MBT Dept., University of Washington

```
Seq  #     b       y     +1
--- --  ------  ------   --
 G   1    58.0      -    10
 G   2   115.0   839.4    9
 A   3   186.1   782.4    8
 G   4   243.1   711.4    7
 L   5   356.2   654.4    6
 E   6   485.2   541.3    5
 P   7   582.3   412.2    4
 A   8   653.3   315.2    3
 P   9   750.4   244.1    2
 K  10      -    147.1    1
```

gd 1.2 © 1994, 1995, Quest Protein Database Center, Cold Spring Harbor Labs

Soleil Index

Soleil Index is: **3 / 4** Xcorr is: **2.7464**
Parent M/Z: **448.7** NL M/Z: **399.7**

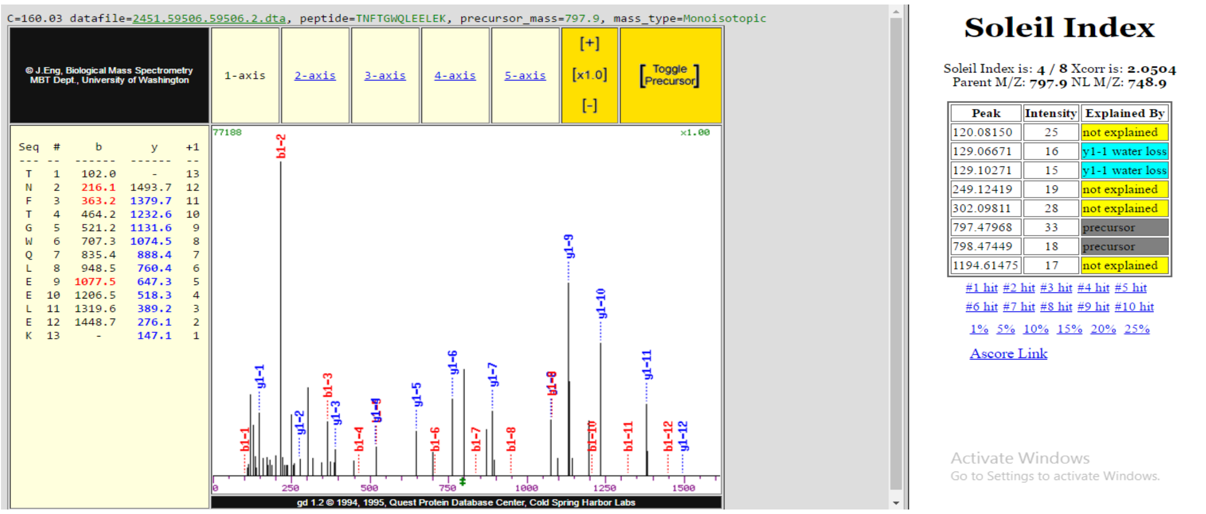
Peak	Intensity	Explained By
141.06615	39	not explained
226.15544	16	b1-4 water loss / y1-2 water loss
328.19852	19	b1-8 doubly charged / y1-6 doubly charged
523.28851	36	y1-5 water loss

#1 hit #2 hit #3 hit #4 hit #5 hit
#6 hit #7 hit #8 hit #9 hit #10 hit

1% 5% 10% 15% 20% 25%

Ascore Link

1.3- TNFTGWQLEELEK



C=160.03 datafile=2451.59506.59506.2.dta, peptide=TNFTGWQLEELEK, precursor_mass=797.9, mass_type=Monoisotopic

© J.Eng, Biological Mass Spectrometry
MBT Dept., University of Washington

```
Seq  #     b       y     +1
--- --  ------  ------   --
 T   1   102.0      -    13
 N   2   216.1  1493.7   12
 F   3   363.2  1379.7   11
 T   4   464.2  1232.6   10
 G   5   521.2  1131.6    9
 W   6   707.3  1074.5    8
 Q   7   835.4   888.4    7
 L   8   948.5   760.4    6
 E   9  1077.5   647.3    5
 E  10  1206.5   518.3    4
 L  11  1319.6   389.2    3
 E  12  1448.7   276.1    2
 K  13      -    147.1    1
```

gd 1.2 © 1994, 1995, Quest Protein Database Center, Cold Spring Harbor Labs

Soleil Index

Soleil Index is: **4 / 8** Xcorr is: **2.0504**
Parent M/Z: **797.9** NL M/Z: **748.9**

Peak	Intensity	Explained By
120.08150	25	not explained
129.06671	16	y1-1 water loss
129.10271	15	y1-1 water loss
249.12419	19	not explained
302.09811	28	not explained
797.47968	33	precursor
798.47449	18	precursor
1194.61475	17	not explained

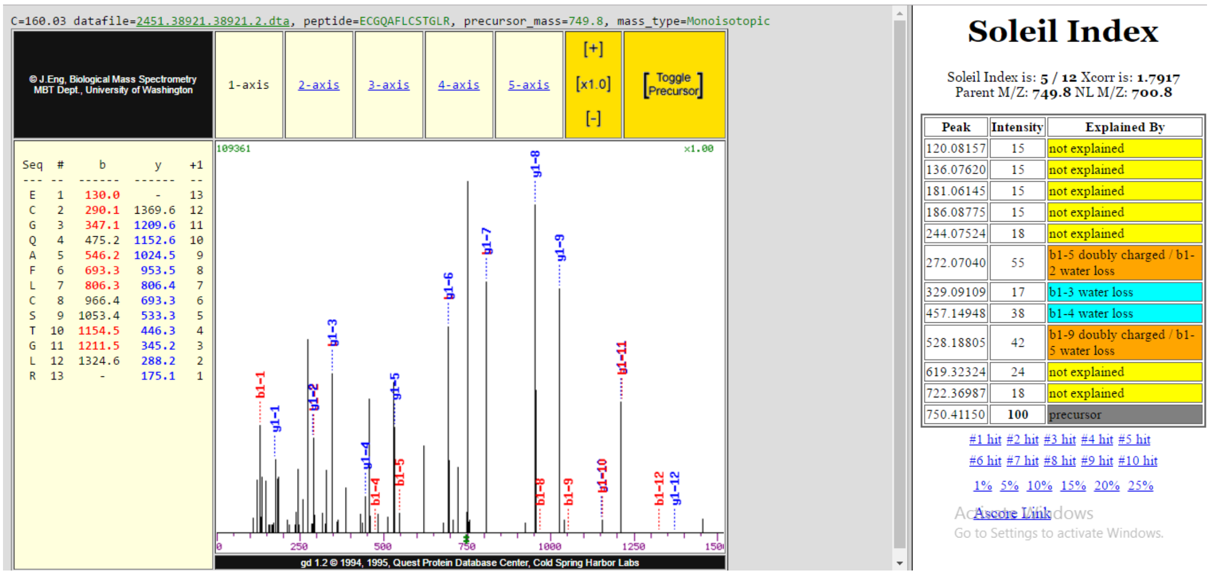
#1 hit #2 hit #3 hit #4 hit #5 hit
#6 hit #7 hit #8 hit #9 hit #10 hit

1% 5% 10% 15% 20% 25%

Ascore Link

2. Q9Y2G7

2.1-ECGQAFLCSTGLR



C=160.03 datafile=2451.38921.38921.2.dta, peptide=ECGQAFLCSTGLR, precursor_mass=749.8, mass_type=Monoisotopic

© J.Eng, Biological Mass Spectrometry
MBT Dept., University of Washington

```
Seq  #     b       y     +1
--- --  ------  ------   --
 E   1   130.0      -    13
 C   2   290.1  1369.6   12
 G   3   347.1  1209.6   11
 Q   4   475.2  1152.6   10
 A   5   546.2  1024.5    9
 F   6   693.3   953.5    8
 L   7   806.3   806.4    7
 C   8   966.4   693.3    6
 S   9  1053.4   533.3    5
 T  10  1154.5   446.3    4
 G  11  1211.5   345.2    3
 L  12  1324.6   288.2    2
 R  13      -    175.1    1
```

gd 1.2 © 1994, 1995, Quest Protein Database Center, Cold Spring Harbor Labs

Soleil Index

Soleil Index is: **5 / 12** Xcorr is: **1.7917**
Parent M/Z: **749.8** NL M/Z: **700.8**

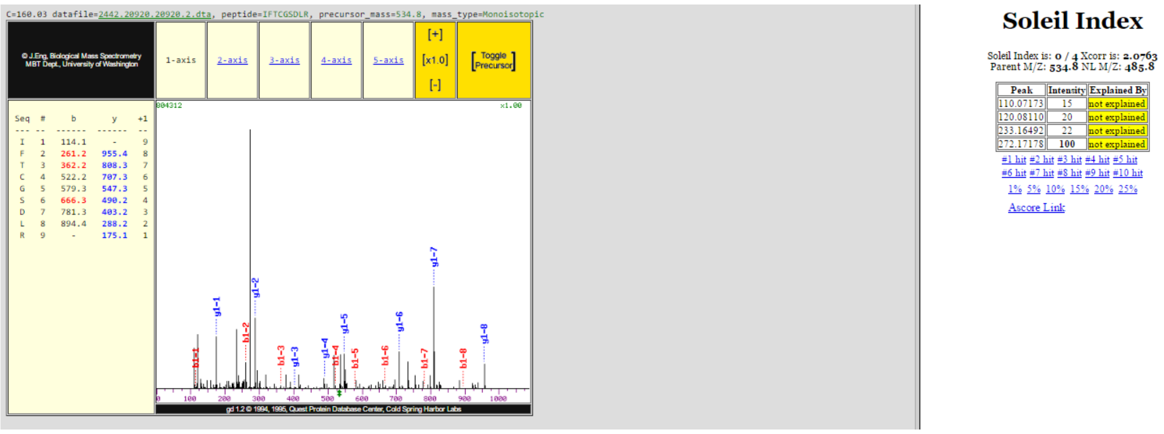
Peak	Intensity	Explained By
120.08157	15	not explained
136.07620	15	not explained
181.06145	15	not explained
186.08775	15	not explained
244.07524	18	not explained
272.07040	55	b1-5 doubly charged / b1-2 water loss
329.09109	17	b1-3 water loss
457.14948	38	b1-4 water loss
528.18805	42	b1-9 doubly charged / b1-5 water loss
619.32324	24	not explained
722.36987	18	not explained
750.41150	**100**	precursor

#1 hit #2 hit #3 hit #4 hit #5 hit
#6 hit #7 hit #8 hit #9 hit #10 hit

1% 5% 10% 15% 20% 25%

Ascore Link

2.2 - IFTCGSDLR



C=160.03 datafile=2442.20920.20920.2.dta, peptide=IFTCGSDLR, precursor_mass=534.8, mass_type=Monoisotopic

```
Seq  #     b       y     +1
--- --  ------  ------   --
 I   1   114.1     -      9
 F   2   261.2   955.4    8
 T   3   362.2   808.3    7
 C   4   522.2   707.3    6
 G   5   579.3   547.3    5
 S   6   666.3   490.2    4
 D   7   781.3   403.2    3
 L   8   894.4   288.2    2
 R   9     -     175.1    1
```

Soleil Index

Soleil Index is: **0 / 4** Xcorr is: **2.0763**
Parent M/Z: **534.8** NL M/Z: **485.8**

Peak	Intensity	Explained By
110.07173	15	not explained
120.08110	20	not explained
233.16492	22	not explained
272.17178	**100**	not explained

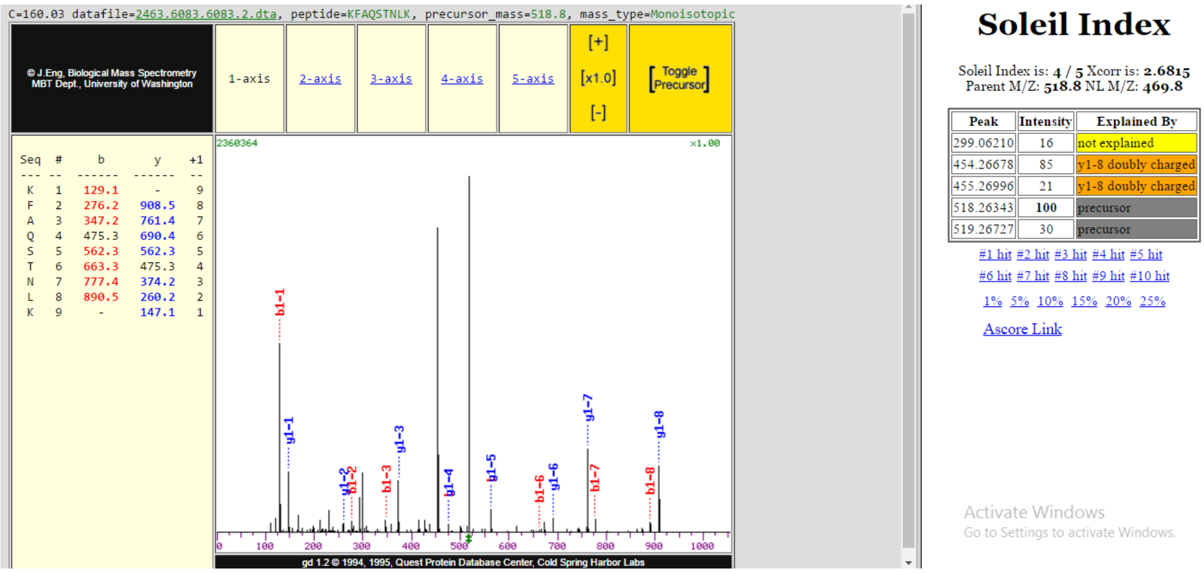
#1 hit #2 hit #3 hit #4 hit #5 hit
#6 hit #7 hit #8 hit #9 hit #10 hit

1% 5% 10% 15% 20% 25%

Ascore Link

3. O15391

3.1- KFAQSTNLK



C=160.03 datafile=2463.6083.6083.2.dta, peptide=KFAQSTNLK, precursor_mass=518.8, mass_type=Monoisotopic

© J.Eng, Biological Mass Spectrometry
MBT Dept., University of Washington
1-axis
2-axis
3-axis
4-axis
5-axis
[+]
[x1.0]
[-]
Toggle Precursor
2360364
x1.00
b1-1
y1-1
y1-2
b1-2
b1-3
y1-3
y1-4
y1-5
b1-6
y1-6
y1-7
b1-7
b1-8
y1-8
0
100
200
300
400
500
600
700
800
900
1000
gd 1.2 © 1994, 1995, Quest Protein Database Center, Cold Spring Harbor Labs
Activate Windows
Go to Settings to activate Windows.

```
Seq  #     b       y     +1
--- --  ------  ------   --
 K   1   129.1     -      9
 F   2   276.2   908.5    8
 A   3   347.2   761.4    7
 Q   4   475.3   690.4    6
 S   5   562.3   562.3    5
 T   6   663.3   475.3    4
 N   7   777.4   374.2    3
 L   8   890.5   260.2    2
 K   9     -     147.1    1
```

Soleil Index

Soleil Index is: **4 / 5** Xcorr is: **2.6815**
Parent M/Z: **518.8** NL M/Z: **469.8**

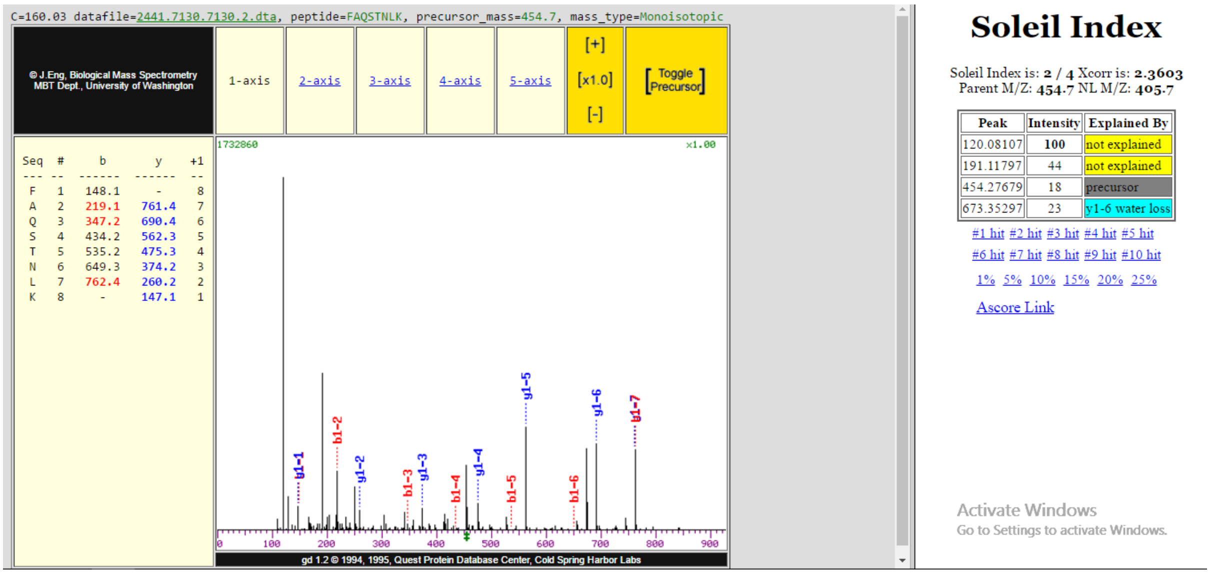
Peak	Intensity	Explained By
299.06210	16	not explained
454.26678	85	y1-8 doubly charged
455.26996	21	y1-8 doubly charged
518.26343	**100**	precursor
519.26727	30	precursor

#1 hit #2 hit #3 hit #4 hit #5 hit
#6 hit #7 hit #8 hit #9 hit #10 hit

1% 5% 10% 15% 20% 25%

Ascore Link

3.2- FAQSTNLK



C=160.03 datafile=2441.7130.7130.2.dta, peptide=FAQSTNLK, precursor_mass=454.7, mass_type=Monoisotopic

© J.Eng, Biological Mass Spectrometry
MBT Dept., University of Washington
1-axis
2-axis
3-axis
4-axis
5-axis
[+]
[x1.0]
[-]
Toggle Precursor
1732860
x1.00
y1-1
b1-2
y1-2
b1-3
y1-3
b1-4
y1-4
y1-5
b1-5
b1-6
y1-6
y1-7
0
100
200
300
400
500
600
700
800
900
gd 1.2 © 1994, 1995, Quest Protein Database Center, Cold Spring Harbor Labs
Activate Windows
Go to Settings to activate Windows.

```
Seq  #     b       y     +1
--- --  ------  ------   --
 F   1   148.1     -      8
 A   2   219.1   761.4    7
 Q   3   347.2   690.4    6
 S   4   434.2   562.3    5
 T   5   535.2   475.3    4
 N   6   649.3   374.2    3
 L   7   762.4   260.2    2
 K   8     -     147.1    1
```

Soleil Index

Soleil Index is: **2 / 4** Xcorr is: **2.3603**
Parent M/Z: **454.7** NL M/Z: **405.7**

Peak	Intensity	Explained By
120.08107	**100**	not explained
191.11797	44	not explained
454.27679	18	precursor
673.35297	23	y1-6 water loss

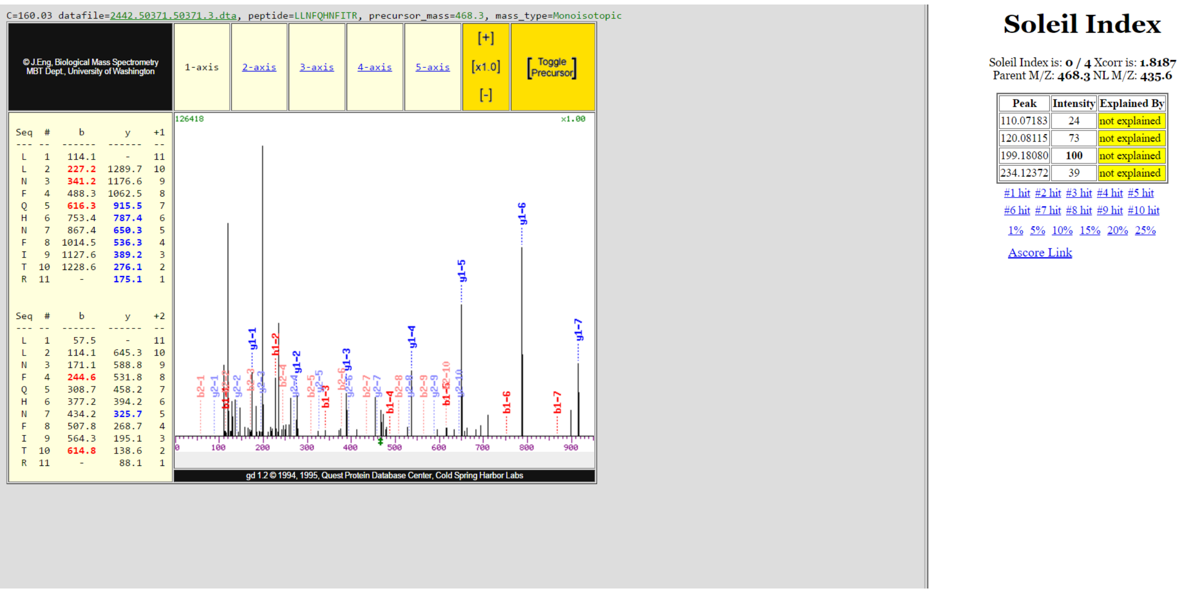
#1 hit #2 hit #3 hit #4 hit #5 hit
#6 hit #7 hit #8 hit #9 hit #10 hit

1% 5% 10% 15% 20% 25%

Ascore Link

4. Q8IUZ0

4- LLNFQHNFITR



C=160.03 datafile=2442.50371.50371.3.dta, peptide=LLNFQHNFITR, precursor_mass=468.3, mass_type=Monoisotopic

```
Seq  #     b       y     +1
--- --  ------  ------  --
 L   1   114.1     -    11
 L   2   227.2  1289.7  10
 N   3   341.2  1176.6   9
 F   4   488.3  1062.5   8
 Q   5   616.3   915.5   7
 H   6   753.4   787.4   6
 N   7   867.4   650.3   5
 F   8  1014.5   536.3   4
 I   9  1127.6   389.2   3
 T  10  1228.6   276.1   2
 R  11     -     175.1   1


Seq  #     b       y     +2
--- --  ------  ------  --
 L   1    57.5     -    11
 L   2   114.1   645.3  10
 N   3   171.1   588.8   9
 F   4   244.6   531.8   8
 Q   5   308.7   458.2   7
 H   6   377.2   394.2   6
 N   7   434.2   325.7   5
 F   8   507.8   268.7   4
 I   9   564.3   195.1   3
 T  10   614.8   138.6   2
 R  11     -      88.1   1
```

gd 1.2 © 1994, 1995, Quest Protein Database Center, Cold Spring Harbor Labs

Soleil Index

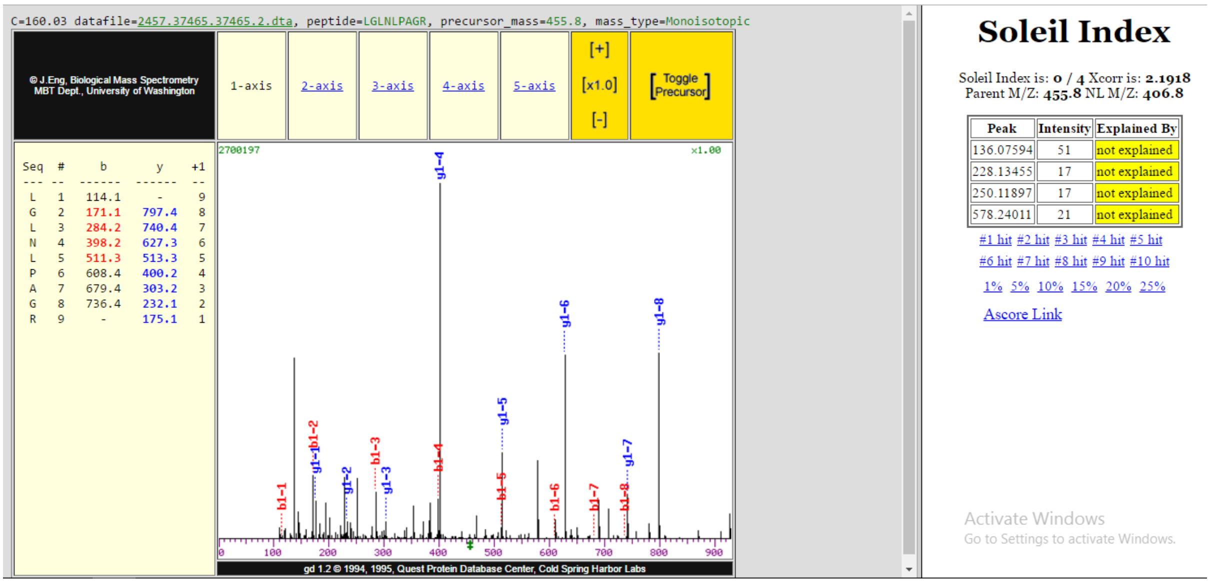
Soleil Index is: **0 / 4** Xcorr is: **1.8187**
Parent M/Z: **468.3** NL M/Z: **435.6**

Peak	Intensity	Explained By
110.07183	24	not explained
120.08115	73	not explained
199.18080	**100**	not explained
234.12372	39	not explained

#1 hit #2 hit #3 hit #4 hit #5 hit
#6 hit #7 hit #8 hit #9 hit #10 hit
1% 5% 10% 15% 20% 25%
Ascore Link

5. Q7Z6R9

5- LGLNLPAGR



C=160.03 datafile=2457.37465.37465.2.dta, peptide=LGLNLPAGR, precursor_mass=455.8, mass_type=Monoisotopic

```
Seq  #     b       y     +1
--- --  ------  ------  --
 L   1   114.1     -     9
 G   2   171.1   797.4   8
 L   3   284.2   740.4   7
 N   4   398.2   627.3   6
 L   5   511.3   513.3   5
 P   6   608.4   400.2   4
 A   7   679.4   303.2   3
 G   8   736.4   232.1   2
 R   9     -     175.1   1
```

gd 1.2 © 1994, 1995, Quest Protein Database Center, Cold Spring Harbor Labs

Soleil Index

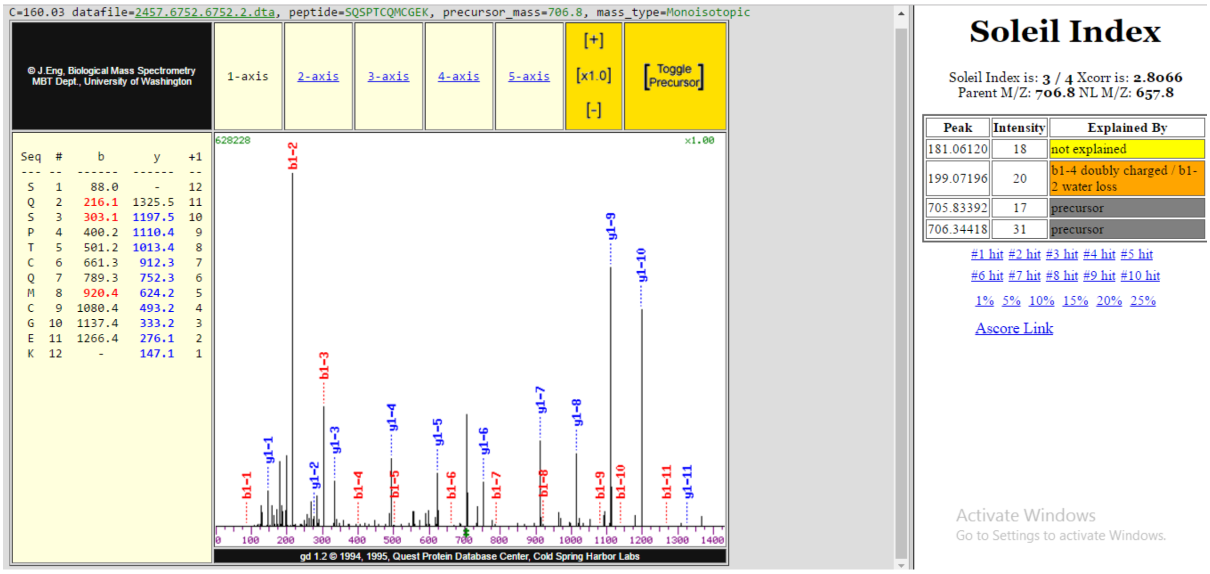
Soleil Index is: **0 / 4** Xcorr is: **2.1918**
Parent M/Z: **455.8** NL M/Z: **406.8**

Peak	Intensity	Explained By
136.07594	51	not explained
228.13455	17	not explained
250.11897	17	not explained
578.24011	21	not explained

#1 hit #2 hit #3 hit #4 hit #5 hit
#6 hit #7 hit #8 hit #9 hit #10 hit
1% 5% 10% 15% 20% 25%
Ascore Link

6. Q12999

6- SQSPTCQMCGEK



C=160.03 datafile=2457.6752.6752.2.dta, peptide=SQSPTCQMCGEK, precursor_mass=706.8, mass_type=Monoisotopic

© J.Eng, Biological Mass Spectrometry MBT Dept., University of Washington

```
Seq  #     b       y    +1
--- --  ------  ------  --
 S   1    88.0     -    12
 Q   2   216.1  1325.5  11
 S   3   303.1  1197.5  10
 P   4   400.2  1110.4   9
 T   5   501.2  1013.4   8
 C   6   661.3   912.3   7
 Q   7   789.3   752.3   6
 M   8   920.4   624.2   5
 C   9  1080.4   493.2   4
 G  10  1137.4   333.2   3
 E  11  1266.4   276.1   2
 K  12     -     147.1   1
```

Soleil Index

Soleil Index is: **3 / 4** Xcorr is: **2.8066**
Parent M/Z: **706.8** NL M/Z: **657.8**

Peak	Intensity	Explained By
181.06120	18	not explained
199.07196	20	b1-4 doubly charged / b1-2 water loss
705.83392	17	precursor
706.34418	31	precursor

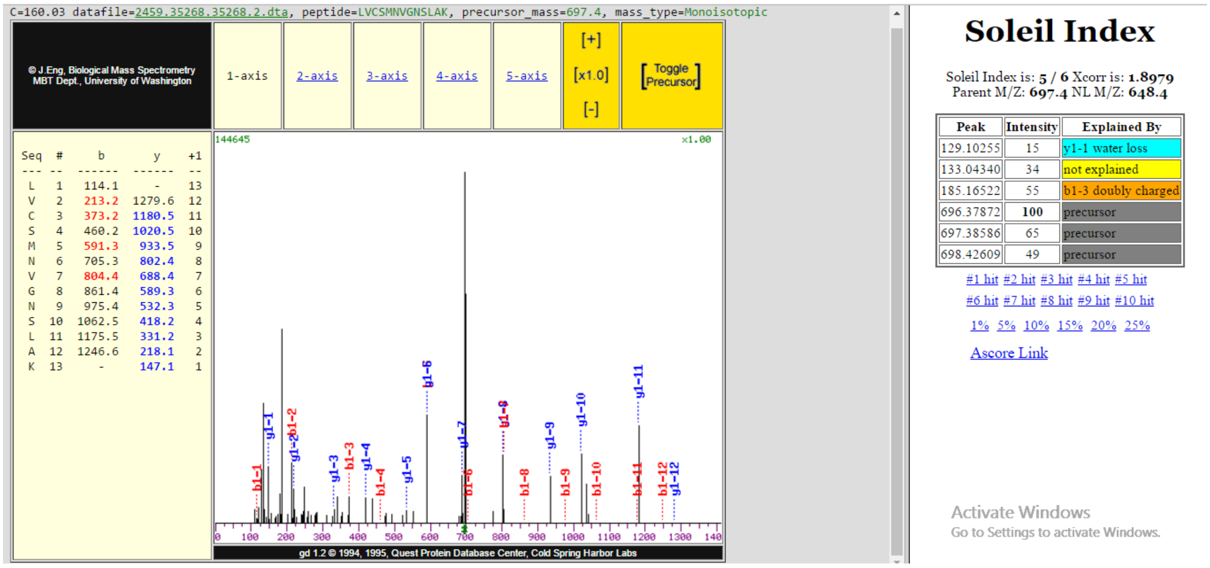
#1 hit #2 hit #3 hit #4 hit #5 hit
#6 hit #7 hit #8 hit #9 hit #10 hit

1% 5% 10% 15% 20% 25%

Ascore Link

7. Q8NEK5

7- LVCSMNVGNSLAK



C=160.03 datafile=2459.35268.35268.2.dta, peptide=LVCSMNVGNSLAK, precursor_mass=697.4, mass_type=Monoisotopic

© J.Eng, Biological Mass Spectrometry MBT Dept., University of Washington

```
Seq  #     b       y    +1
--- --  ------  ------  --
 L   1   114.1     -    13
 V   2   213.2  1279.6  12
 C   3   373.2  1180.5  11
 S   4   460.2  1020.5  10
 M   5   591.3   933.5   9
 N   6   705.3   802.4   8
 V   7   804.4   688.4   7
 G   8   861.4   589.3   6
 N   9   975.4   532.3   5
 S  10  1062.5   418.2   4
 L  11  1175.5   331.2   3
 A  12  1246.6   218.1   2
 K  13     -     147.1   1
```

Soleil Index

Soleil Index is: **5 / 6** Xcorr is: **1.8979**
Parent M/Z: **697.4** NL M/Z: **648.4**

Peak	Intensity	Explained By
129.10255	15	y1-1 water loss
133.04340	34	not explained
185.16522	55	b1-3 doubly charged
696.37872	**100**	precursor
697.38586	65	precursor
698.42609	49	precursor

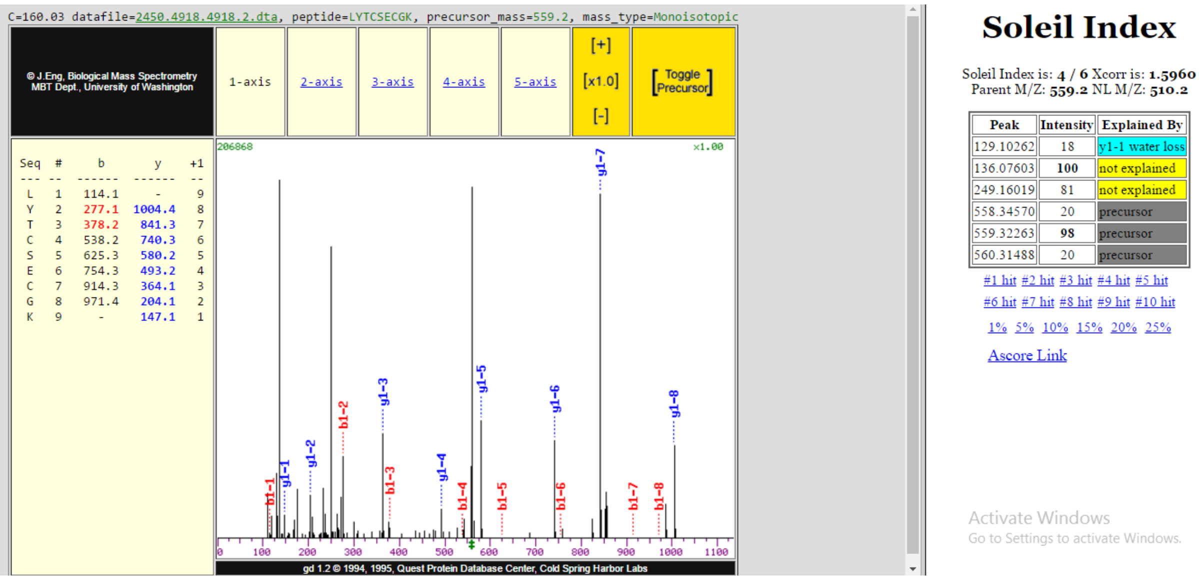
#1 hit #2 hit #3 hit #4 hit #5 hit
#6 hit #7 hit #8 hit #9 hit #10 hit

1% 5% 10% 15% 20% 25%

Ascore Link

8. Q8N8J6

8- LYTCSECGG



C=160.03 datafile=2450.4918.4918.2.dta, peptide=LYTCSECGK, precursor_mass=559.2, mass_type=Monoisotopic

```
Seq  #    b       y     +1
---  --  ------  ------  --
 L   1   114.1     -     9
 Y   2   277.1  1004.4   8
 T   3   378.2   841.3   7
 C   4   538.2   740.3   6
 S   5   625.3   580.2   5
 E   6   754.3   493.2   4
 C   7   914.3   364.1   3
 G   8   971.4   204.1   2
 K   9     -     147.1   1
```

Soleil Index

Soleil Index is: **4 / 6** Xcorr is: **1.5960**
Parent M/Z: **559.2** NL M/Z: **510.2**

Peak	Intensity	Explained By
129.10262	18	y1-1 water loss
136.07603	**100**	not explained
249.16019	81	not explained
558.34570	20	precursor
559.32263	**98**	precursor
560.31488	20	precursor

#1 hit #2 hit #3 hit #4 hit #5 hit
#6 hit #7 hit #8 hit #9 hit #10 hit

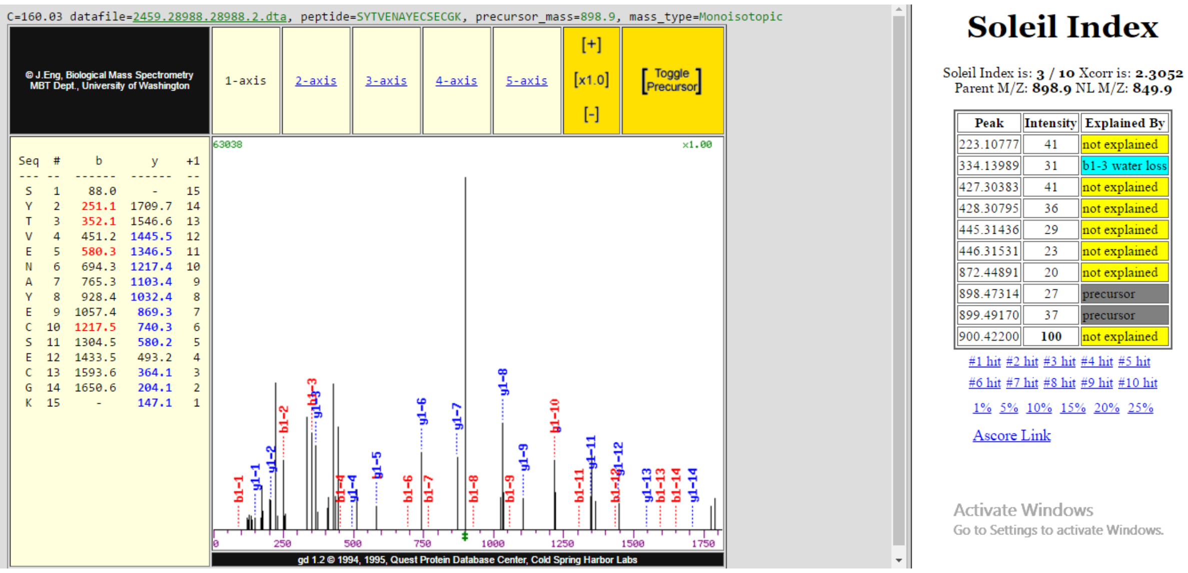
1% 5% 10% 15% 20% 25%

Ascore Link

Activate Windows
Go to Settings to activate Windows.

9. Q5T5D7

9- SYTVENAYECSECGK



C=160.03 datafile=2459.28988.28988.2.dta, peptide=SYTVENAYECSECGK, precursor_mass=898.9, mass_type=Monoisotopic

```
Seq  #    b       y     +1
---  --  ------  ------  --
 S   1    88.0     -     15
 Y   2   251.1  1709.7   14
 T   3   352.1  1546.6   13
 V   4   451.2  1445.5   12
 E   5   580.3  1346.5   11
 N   6   694.3  1217.4   10
 A   7   765.3  1103.4    9
 Y   8   928.4  1032.4    8
 E   9  1057.4   869.3    7
 C  10  1217.5   740.3    6
 S  11  1304.5   580.2    5
 E  12  1433.5   493.2    4
 C  13  1593.6   364.1    3
 G  14  1650.6   204.1    2
 K  15     -     147.1    1
```

Soleil Index

Soleil Index is: **3 / 10** Xcorr is: **2.3052**
Parent M/Z: **898.9** NL M/Z: **849.9**

Peak	Intensity	Explained By
223.10777	41	not explained
334.13989	31	b1-3 water loss
427.30383	41	not explained
428.30795	36	not explained
445.31436	29	not explained
446.31531	23	not explained
872.44891	20	not explained
898.47314	27	precursor
899.49170	37	precursor
900.42200	**100**	not explained

#1 hit #2 hit #3 hit #4 hit #5 hit
#6 hit #7 hit #8 hit #9 hit #10 hit

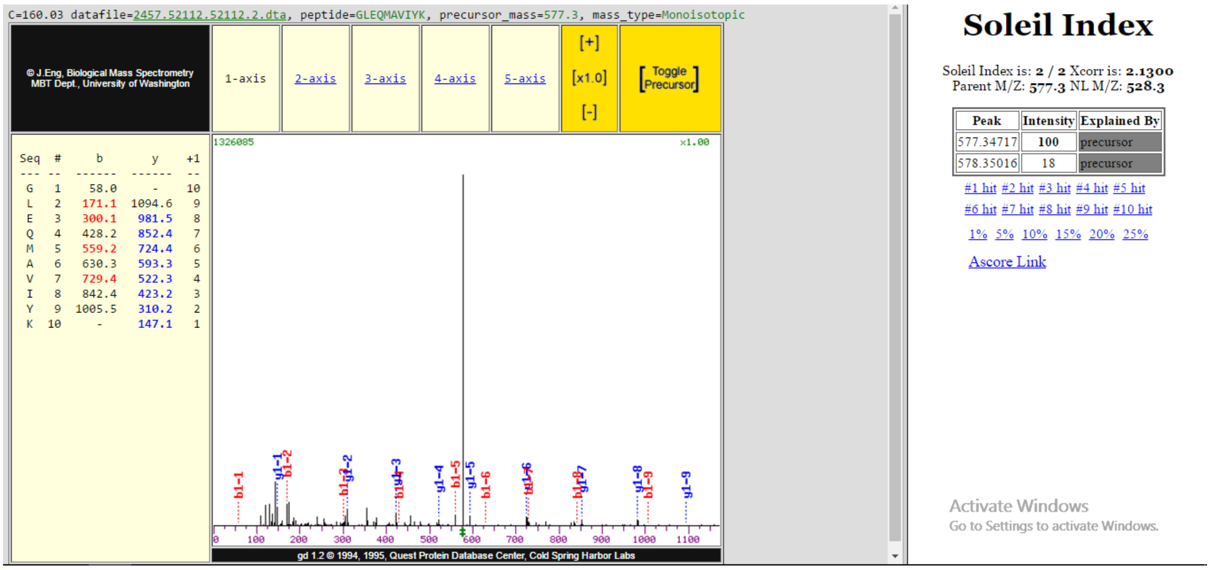
1% 5% 10% 15% 20% 25%

Ascore Link

Activate Windows
Go to Settings to activate Windows.

10. Q9UC06

10- GLEQMAVIYK



C=160.03 datafile=2457.52112.52112.2.dta, peptide=GLEQMAVIYK, precursor_mass=577.3, mass_type=Monoisotopic

```
Seq  #     b       y     +1
--- --  ------  ------   --
 G   1    58.0      -    10
 L   2   171.1  1094.6    9
 E   3   300.1   981.5    8
 Q   4   428.2   852.4    7
 M   5   559.2   724.4    6
 A   6   630.3   593.3    5
 V   7   729.4   522.3    4
 I   8   842.4   423.2    3
 Y   9  1005.5   310.2    2
 K  10      -    147.1    1
```

Soleil Index

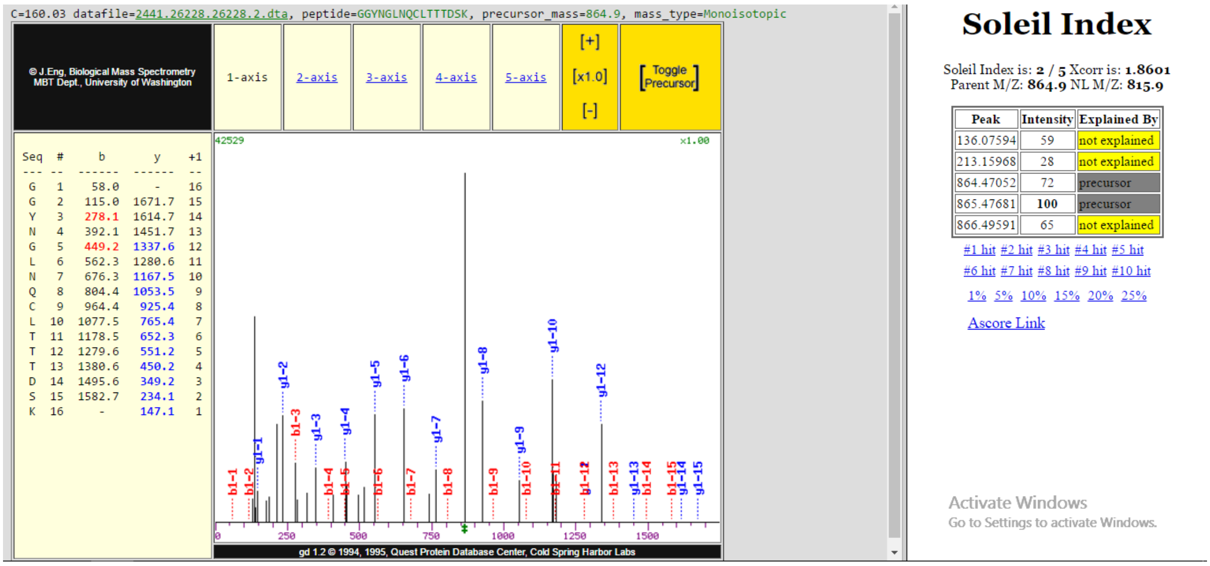
Soleil Index is: **2 / 2** Xcorr is: **2.1300**
Parent M/Z: **577.3** NL M/Z: **528.3**

Peak	Intensity	Explained By
577.34717	**100**	precursor
578.35016	18	precursor

© J.Eng, Biological Mass Spectrometry MBT Dept., University of Washington
1-axis
2-axis
3-axis
4-axis
5-axis
[+]
[x1.0]
[-]
Toggle Precursor
1326085
x1.00
b1-1
y1-1
b1-2
b1-3
y1-2
b1-4
y1-3
y1-4
b1-5
y1-5
b1-6
y1-6
b1-7
b1-8
y1-7
y1-8
b1-9
y1-9
0
100
200
300
400
500
600
700
800
900
1000
1100
gd 1.2 © 1994, 1995, Quest Protein Database Center, Cold Spring Harbor Labs
#1 hit #2 hit #3 hit #4 hit #5 hit
#6 hit #7 hit #8 hit #9 hit #10 hit
1% 5% 10% 15% 20% 25%
Ascore Link
Activate Windows
Go to Settings to activate Windows.

11. Q03936

11- GGYNGLNQCLTTTDSK



C=160.03 datafile=2441.26228.26228.2.dta, peptide=GGYNGLNQCLTTTDSK, precursor_mass=864.9, mass_type=Monoisotopic

```
Seq  #     b       y     +1
--- --  ------  ------   --
 G   1    58.0      -    16
 G   2   115.0  1671.7   15
 Y   3   278.1  1614.7   14
 N   4   392.1  1451.7   13
 G   5   449.2  1337.6   12
 L   6   562.3  1280.6   11
 N   7   676.3  1167.5   10
 Q   8   804.4  1053.5    9
 C   9   964.4   925.4    8
 L  10  1077.5   765.4    7
 T  11  1178.5   652.3    6
 T  12  1279.6   551.2    5
 T  13  1380.6   450.2    4
 D  14  1495.6   349.2    3
 S  15  1582.7   234.1    2
 K  16      -    147.1    1
```

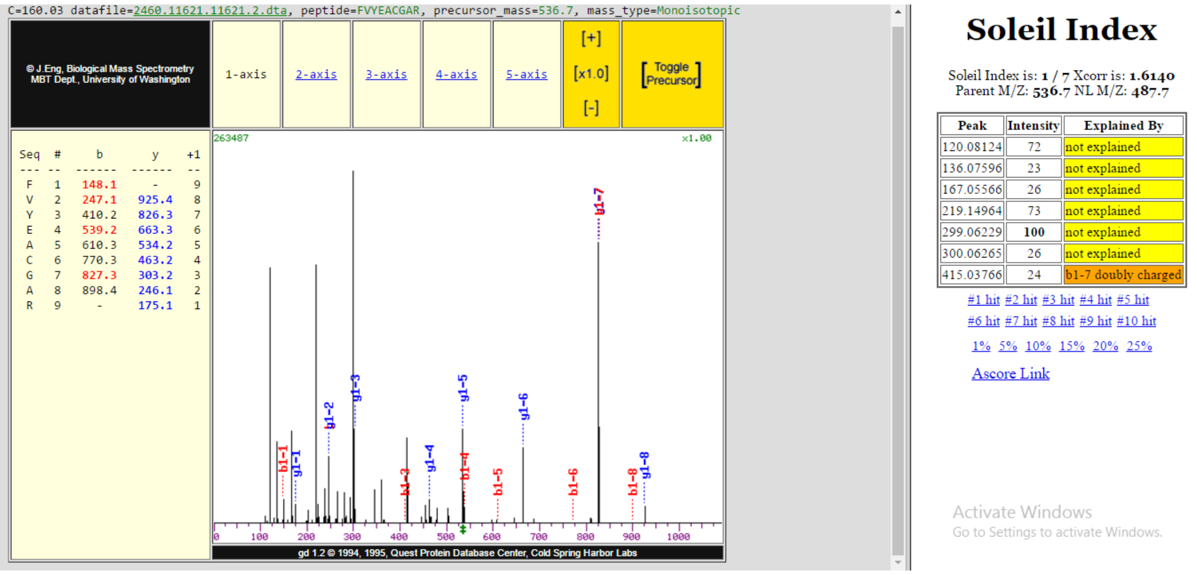
Soleil Index

Soleil Index is: **2 / 5** Xcorr is: **1.8601**
Parent M/Z: **864.9** NL M/Z: **815.9**

Peak	Intensity	Explained By
136.07594	59	not explained
213.15968	28	not explained
864.47052	72	precursor
865.47681	**100**	precursor
866.49591	65	not explained

12. A6NGE4

12- FVYEACGAR



C=160.03 datafile=2460.11621.11621.2.dta, peptide=FVYEACGAR, precursor_mass=536.7, mass_type=Monoisotopic

```
Seq  #     b       y    +1
---  --  ------  ------  --
 F   1   148.1     -     9
 V   2   247.1   925.4   8
 Y   3   410.2   826.3   7
 E   4   539.2   663.3   6
 A   5   610.3   534.2   5
 C   6   770.3   463.2   4
 G   7   827.3   303.2   3
 A   8   898.4   246.1   2
 R   9     -     175.1   1
```

Soleil Index

Soleil Index is: **1 / 7** Xcorr is: **1.6140**
Parent M/Z: **536.7** NL M/Z: **487.7**

Peak	Intensity	Explained By
120.08124	72	not explained
136.07596	23	not explained
167.05566	26	not explained
219.14964	73	not explained
299.06229	**100**	not explained
300.06265	26	not explained
415.03766	24	b1-7 doubly charged

#1 hit #2 hit #3 hit #4 hit #5 hit
#6 hit #7 hit #8 hit #9 hit #10 hit

1% 5% 10% 15% 20% 25%

Ascore Link

References

1. 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 Res* 2017, **45**(D1):D1100-D1106.
2. **GeneCards®: The Human Gene Database** [http://www.genecards.org/]
3. The UniProt C: **UniProt: the universal protein knowledgebase**. *Nucleic Acids Res* 2017, **45**(D1):D158-D169.
4. Mitchell JA, Aronson AR, Mork JG, Folk LC, Humphrey SM, Ward JM: **Gene indexing: characterization and analysis of NLM's GeneRIFs**. *AMIA Annu Symp Proc* 2003:460-464.
5. **PubMed Clinical Queries** [https://www.ncbi.nlm.nih.gov/entrez/query/static/clinical.shtml]
6. Vizcaino JA, Csordas A, del-Toro N, Dianes JA, Griss J, Lavidas I, Mayer G, Perez-Riverol Y, Reisinger F, Ternent T *et al*: **2016 update of the PRIDE database and its related tools**. *Nucleic Acids Res* 2016, **44**(D1):D447-456.
7. Craig R, Cortens JP, Beavis RC: **Open source system for analyzing, validating, and storing protein identification data**. *J Proteome Res* 2004, **3**(6):1234-1242.
8. Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H: **Mass-spectrometry-based draft of the human proteome**. *Nature* 2014, **509**(7502):582-587.
9. Schaab C, Geiger T, Stoehr G, Cox J, Mann M: **Analysis of high accuracy, quantitative proteomics data in the MaxQB database**. *Mol Cell Proteomics* 2012, **11**(3):M111 014068.

10. Perez-Riverol Y, Bai M, da Veiga Leprevost F, Squizzato S, Park YM, Haug K, Carroll AJ, Spalding D, Paschall J, Wang M *et al*: **Discovering and linking public omics data sets using the Omics Discovery Index**. *Nature Biotechnology* 2017, **35**(5):406-409.
11. Deutsch EW, Lam H, Aebersold R: **PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows**. *EMBO Rep* 2008, **9**(5):429-434.
12. **UCSD/CCMS - MassIVE Datasets - Mass Spectrometry Repository Dataset List** [http://massive.ucsd.edu/ProteoSAFe/datasets.jsp]
13. Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M, Mahendraker T, Williams M, Neumann S, Rocca-Serra P *et al*: **MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data**. *Nucleic Acids Res* 2013, **41**(Database issue):D781-786.
14. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T *et al*: **Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking**. *Nat Biotechnol* 2016, **34**(8):828-837.
15. Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, Edison A, Fiehn O, Higashi R, Nair KS *et al*: **Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools**. *Nucleic Acids Res* 2016, **44**(D1):D463-470.
16. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T *et al*: **ArrayExpress update--simplifying data submissions**. *Nucleic Acids Res* 2015, **43**(Database issue):D1113-1116.
17. Kapushesky M, Adamusiak T, Burdett T, Culhane A, Farne A, Filippov A, Holloway E, Klebanov A, Kryvych N, Kurbatova N *et al*: **Gene Expression Atlas update--a value-added database of microarray and sequencing-based functional genomics experiments**. *Nucleic Acids Res* 2012, **40**(Database issue):D1077-1081.
18. Carroll AJ, Badger MR, Harvey Millar A: **The MetabolomeExpress Project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets**. *BMC Bioinformatics* 2010, **11**:376.
19. Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Ur-Rehman S, Saunders G, Kandasamy J, Caccamo M, Leinonen R *et al*: **The European Genome-phenome Archive of human data consented for biomedical research**. *Nat Genet* 2015, **47**(7):692-695.
20. Yalçın MA, Elmqvist N, Bederson BB: **Keshif: Out-of-the-Box Visual and Interactive Data Exploration Environment**.
21. Bostock M, Ogievetsky V, Heer J: **D3: Data-Driven Documents**. *IEEE Transactions on Visualization and Computer Graphics* 2011, **17**(12):2301-2309.
22. Islam MT, Mohamedali A, Ahn SB, Nawar I, Baker MS, Ranganathan S: **A Systematic Bioinformatics Approach to Identify High Quality Mass Spectrometry Data and Functionally Annotate Proteins and Proteomes**. *Methods Mol Biol* 2017, **1549**:163-176.
23. **Crossref Metadata Search** [http://search.crossref.org/]

7.4 Conclusions

We extended our bioinformatics pipeline to incorporate the MS evidence workflow from the previous Chapter. The extended pipeline was used to programmatically capture peptide identification data from a mix of HUPO and non-HUPO preferred databases (GPMDB, PRIDE, ProteomicsDB, MAXQB) for all olfactory receptors family proteins (largest protein family with 409 missing proteins). We also collected data from Human ProteinPedia manually. The pipeline captured 122,717 peptide MS entries with at least seven amino acids length. The proteotypicity checker identified and retained 4,751 peptides as proteotypic peptides. The pipeline then compared the search engine confidence scores and retained 6% (286) of these entries having acceptable search engine ratings. After manually analysing the spectra for these 286 peptides using the guideline described in the previous chapter, we identified MS evidences for 23 olfactory proteins (see publication 5). The extended pipeline demonstrated many single high-confidence peptide MS evidence are available in the public domain. Although best available peptide MS and the associated MS spectra are insufficient to meet current HPP metric, storing such evidence in a public repository and frequently updating the repository automatically including any community contributed data can lead us to find the missing pieces of the puzzle. However, there are some proteins when cleaved by trypsin may not produce two uniquely mapping proteotypic peptides of at least nine amino acids. Many biologically important highly bioactive secretory peptides are very short secreted proteoforms. These peptide proteoforms will remain undetected as per the current HPP guidelines [283]. By using the bottom up MS data for interleukin-9 (IL-9) from our collaborators lab, we demonstrated that IL-9 indeed contains two proteotypic peptides of 7 and 8 residues with high-quality MS spectra (see Publication 5), and our approach can lead to the identification of other proteins with short peptides.

We then applied an orthogonal approach to identify complementary evidence from various databases using a randomly selected protein (Prestin/SLC26A5) from chromosome 7 missing proteins. We identified 91 peer-reviewed manuscripts in PubMed for Prestin and its alternate names. One study determined Prestin is expressed in outer hair cells and classified it as the 'motor protein of the cochlear outer hair cell' [284]. The search also picked up a recent review article of Prestin's structural and functional properties [285]. Some other evidence includes 83 anti-prestin Abs in Antibodypedia [286], two missense/nonsense mutations responsible for deafness/autism in Human Gene Mutation

Database [287] and more (at the time of the study). Details of other complementary findings are available in Publication 5. A recent search (4 June 2017) on the Antibodypedia reveals 95 antibodies from 18 providers for Prestin [288]. This highlight, although some missing proteins do not contain high-stringency MS or acceptable Ab evidence, various public and communal databases contain protein level functional evidence that needs to be captured (on a regular basis) and considered to identify and characterise missing proteins.

It was evident from our study that a public platform capable of capturing mainstream MS/AB evidence as well as other orthogonal information from various sources and providing deep analyses of such information can accelerate the identification of missing proteins and uncover the hidden biology. Therefore, I integrated ProtAnnotator 2.0 and the extended proteomics data integration strategy used in this study to create a novel integrated web platform (MissingProteinPedia) to mine various MS/AB and other orthogonal information from different databases and web platforms for human "missing" proteins. The key components for the integrated MissingProteinPedia (MPP) web platform are, a web interface for users to mine and rank MS data from individual labs, an automatic data capture tool to collect orthogonal information from various databases, and an MS data capture service to collate and integrate mass spectrometry data from different databases with a community annotation module to validate shortlisted peptide spectra via collaborative community effort. At the launch, the platform captured over 5 million peptide identification records from different databases. Of these, 272, 831 peptides were proteotypic (at least 7 aa long), 38% of which contained acceptable search engine scores that are ready for the scientific community to annotate or validate via the community annotation module. The purpose-built community platform offers a single public data discovery platform as well as a secure sharing platform where the data, as well as the user authentication and authorisations are entirely controlled by the community administrators. Some other feature includes interactive PubMed data browser, interactive MS peptide quality browser, semi-automated dynamic form builder, etc. (see publication 6 for other features). Following the launch at 2016 HUPO World Congress, the platform was presented and discussed with the community at the Bioinformatics Hub at the HUPO 2016 World Congress [289, 290] to receive feedback. Some of the MPP features resulted from this consultation include PubMed annotation module, filtered and or ranked view of MS peptide information. To demonstrate the applicability of the platform, we used MS data from our collaborator's lab and identified MS evidence for 9 proteins of which 2 comply with HPP guidelines and can be considered candidates for PE1 status (see publication 6).

Chapter 8: Conclusions and future directions

8.1 Summary, Significance and contributions

Proteins are responsible for almost every action within the cell, hence understanding the protein function is the key to uncovering the biology of every living organism. Naturally, the identification of a protein becomes the precursor for its functional study. The advancement of next-generation sequencing and high-throughput proteomics technology enable us to conduct large-scale protein prediction and confirm their identification via experiments respectively. Despite the progress in both techniques, a significant number of proteins remain unidentified by experimental method, and 1% of the protein functions are experimentally validated [281]. The unprecedented growth in genome sequences means it is almost impossible for experimental proteomics strategies to catch up with the backlog of identification and characterization. Hence, there is an opportunity for improved computational methods to close this gap.

In chapter one, I carried out a detailed review to obtain a better understanding of the end-to-end workflow of protein identification and characterisation to understand the gaps and opportunities. At first, all available proteomics techniques, algorithms, and associated tools were studied. A brief review of genome annotation (gene and protein prediction) was conducted, followed by a detailed study of the *in silico* functional annotation strategies and as well as current proteomics databases. It became apparent that one of the solutions to close the gap of identification and characterisation is, in fact, lies within the problem. In other words, knowledgebase created by the vast growth of sequencing and technologies can be used to identify and or characterise proteins. Based on this literature review, the objective for this thesis was formed, with results presented as publications (listed in Chapter 2). The review provides a baseline for similar studies.

A novel *in silico* sequence homology and functional annotation strategy was developed for novel or less studied proteomes (Chapter 3). The black Périgord truffle (*Tuber melanosporum Vittad*), one of the highly prized but less known proteomes concerning its biological functions. The genome was sequenced in 2010 [34] with only 14 reviewed proteins with no experimental evidence in UniProt database, became an excellent candidate to apply our methodology. Fresh fruiting bodies were also collected from our collaborator's truffle field in Australia to validate and complement our method using a shotgun proteomics experiment. The bioinformatics approach successfully identified homologous for 2486 proteins (UniProt databases), structural similarity of 101 proteins with the Protein Data Bank (PDB) sequences, and functional annotations for 96% of these proteins. The shotgun proteomics identified 836 proteins, 47% of which were also detected by the bioinformatics method. The computational analysis on the functional annotation provided by our approach identified nine proteins, responsible for the aroma profile and a potential enzymatic pathway to produce one of the primary volatiles of the black truffle. Given that the UniProt database contains only 14 reviewed proteins to date, this study made a significant contribution in identifying black truffle proteins and uncovering its biology.

To determine the generic nature of this approach as well as assisting the C-HPP initiative to identify and characterise the missing human proteins, a semi-automated pipeline was developed using the previously mentioned annotation strategy. The pipeline was used to annotate human missing proteins (3831 sequences) for all chromosomes (Chapter 4) for the first time. It identified homologues from the mammalian kingdom for 66.2% of the missing human proteins and functional annotations for 50.8% of the missing human proteins. The ProtAnnotaor web portal was also created to regularly update and share the annotation results with the scientific community. This demonstrates that the annotation strategy can identify functional annotation for any novel or less studied organism. However, another challenge and opportunity of the technological advancement were the protein databases get regularly updated with new information means repeated analyses of the annotation can provide annotations for previously unannotated proteins and or updated annotation for previously annotated proteins. Our study in chapter 3 and 4 demonstrated that sequential blast using high-quality protein sequence datasets for closely related organism provides high-quality annotations. That means the target blast databases will vary for different species. On the other hand, neXtProt regularly updates the missing protein list, which means the human protein annotation needs a regular update too. Hence, the ProtAnnotator pipeline was extended and developed to be an entirely automated, generic cloud-based protein functional annotation platform for functional annotation (Chapter 5). The platform allows users to customise their annotation workflow using their own databases. The platform was used to reanalyse the data from black Périgord truffle study. The platform identified homologous for 111 new proteins, and structural similarity with PDB sequences for 1,393 proteins compared to 101 proteins from the previous study. The platform culled

129 previously identified homologous as they were below the percentage identity threshold based on the new analysis (with updated tools and databases). It identified functional annotation for 82% of the black Périgord truffle proteins compared 20% proteins from the earlier study. This demonstrates that the new platform underpinned by updated tools and database can provide more accurate annotation, as well as the need for renalyses of protein annotations to detect new and accurate annotation. The automated ProtAnnotaror 2.0 cloud-based annotation platform is available for the community to annotate any proteomes functionally using their own datasets. The platform also underpins the functional annotation knowledgebase of the MissingProteinPedia [283].

However, the quality of the annotation relies on end users selecting closely related, high-quality target sequence databases. Moreover, coupling the bioinformatics approach with experimental proteomics (i.e. pan 'omics' approach) can complement the annotation considerably. A broad range of proteomics techniques, platform, tools, and algorithms are available for proteomics study (described in Chapter 1), and most of the proteomics databases offer minimum to no integration between cross-platform information. Hence, it is very difficult for general users to analyse or interpret proteomics data. To address this issue, a protocol for functional annotation and guideline to analyse and interpret proteomics data was developed using walkthrough examples (Chapter 6) that can not only be useful for general users, but also for the advanced users to develop high-throughput pipelines for large scale studies.

The automated platform (Chapter 5), and the protocol and guidelines for MS evidence workflow (Chapter 6) were used to develop an automated pipeline to capture data for the missing protein review study (Chapter 7). The study highlighted the need to capture MS evidence from both HPP and non-HPP databases, orthogonal information from the public and private domain, as well as increased community collaboration to accelerate the identification of human missing proteins. Hence, an integrated web platform was developed to mine both MS and orthogonal information from various sources on a regular basis and provide a purpose-built community portal for secure collaboration. The study itself provided best available MS (39) and orthogonal information (2) for 41 missing proteins during the implementation phase, of which 2 proteins met the high stringency HPP PE1 (protein level 1) criteria for them to be classified as identified. The platform also provides a range of community science features to fast track the missing protein identification and characterisation. The broad range of data compiled by the MPP will contribute to the validation or identification of the missing proteins. The platform can be integrated with the existing high-stringency HPP data reanalysis as a complementary resource.

To summarise, all high-throughput protein annotation and charecterisation methods, protocols, and platforms developed in this thesis can be used individually or collectively to study proteomes and their functions.

8.2 Innovations

The highlights of the thesis are the development of the *in silico* functional annotation, MS evidence workflow, and development of the integrated MissingProteinPedia web platform to uncover the human missing proteins. The generic functional annotation strategy can be used to annotate proteins from any novel or less studied proteomes. The MS evidence workflow provides a guideline to interpret and analyse proteomics data from multiple platforms. The integrated MissingProteinPedia web platform is the first of its kind that not only integrates data from various sources but also conducts in-depth analyses of the data, and facilitates community science and collaboration. Although the platform described here is used for human missing proteins, the platform can be used for similar studies or other proteomes.

8.4 Future directions

The literature review (Chapter 1) of the proteomics method was limited to Data-dependent acquisition (DDA). The Data-independent acquisition (DIA), and post-translational modifications were excluded from the review to limit the scope of this thesis. This section can be extended in future. Subsequently, Chapter 6 can be extended with additional protocols and or guidelines to include analysis and interpretation data generated by the DIA method. To this end, we have started exploring the use of *de novo* sequencing approaches for peptide identification (Appendix 1: Publication 7).

The MissingProteinPedia (MPP) is designed in a way that it can be extended in many different directions. At present, it captures data from four proteomics databases. It can vertically extend to capture data from various other databases described in Chapter 1. The platform currently integrates with the recently published (May 2017) Omics Discovery

Index (OmicsDI) service that aims to index datasets from multi-omics platforms. At present, MPP captures the index and available metadata for all human missing proteins and provides a per protein-based interactive browser to navigate the metadata. OmicsDI currently indexed 48,213 proteomics datasets (human studies) that are related to human missing proteins. MPP can be extended to identify any new datasets indexed by the OmicsDI and then capture the peptide identification information directly from the source database (as OmicsDI does not store this information). Although OmicsDI captures multi-omics datasets, at present it does not have any multi-omics dataset indexed for missing proteins. However, the current MPP integration allows us to identify any new multi-omics (as they are added) databases and incorporate them into the data capture service. The MPP analysis platform can then be extended horizontally to include multi-omics data.

References

1. Bruce Alberts AJ, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter: **Molecular Biology of the Cell**, 5 edn. New York: Garland Science; 2007.
2. Pauling L, Corey RB, Branson HR: **The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain**. *Proc Natl Acad Sci U S A* 1951, **37**(4):205-211.
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.
4. White DM: **Hasn't Evolution Been Proven True? | Answers in Genesis.** In: *The New Answers Book 1.* Edited by Ham K. USA: New Leaf Publishing Group, Inc.; 2008: 283-295
5. Sokolowska I, Wetie AGN, Woods AG, Darie CC: **Applications of Mass Spectrometry in Proteomics**. *Australian Journal of Chemistry* 2013, **66**(7):721-733.
6. Doytchinova IA, Taylor P, Flower DR: **Proteomics in Vaccinology and Immunobiology: An Informatics Perspective of the Immunone**. *J Biomed Biotechnol* 2003, **2003**(5):267-290.
7. Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H: **Mass-spectrometry-based draft of the human proteome**. *Nature* 2014, **509**(7502):582-587.
8. Tyers M, Mann M: **From genomics to proteomics**. *Nature* 2003, **422**(6928):193-197.
9. Pandey A, Mann M: **Proteomics to study genes and genomes**. *Nature* 2000, **405**(6788):837-846.
10. Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez JC, Yan JX, Gooley AA, Hughes G, Humphery-Smith I *et al*: **From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis**. *Biotechnology (N Y)* 1996, **14**(1):61-65.
11. Gstaiger M, Aebersold R: **Applying mass spectrometry-based proteomics to genetics, genomics and network biology**. *Nat Rev Genet* 2009, **10**(9):617-627.
12. Zhang G, Annan RS, Carr SA, Neubert TA: **Overview of Peptide and protein analysis by mass spectrometry**. *Current protocols in molecular biology / edited by Frederick M Ausubel [et al]* 2014, **108**:10 21 11-10 21 30.
13. Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S: **A draft map of the human proteome**. *Nature* 2014, **509**(7502):575-581.

13. Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S: **A draft map of the human proteome**. *Nature* 2014, **509**(7502):575-581.

14. Vargas AJ, Harris CC: **Biomarker development in the precision medicine era: lung cancer as a case study**. *Nat Rev Cancer* 2016, **16**(8):525-537.

15. Adaway JE, Keevil BG, Owen LJ: **Liquid chromatography tandem mass spectrometry in the clinical laboratory**. *Ann Clin Biochem* 2015, **52**(Pt 1):18-38.

16. Chen R, Snyder M: **Promise of personalized omics to precision medicine**. *Wiley Interdiscip Rev Syst Biol Med* 2013, **5**(1):73-82.

17. Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcaino JA: **Making proteomics data accessible and reusable: Current state of proteomics databases and repositories**. *Proteomics* 2014.

18. Wetie AG, Shipp DA, Darie CC: **Bottlenecks in proteomics**. *Advances in experimental medicine and biology* 2014, **806**:581-593.

19. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.

20. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome**. *Science* 2001, **291**(5507):1304-1351.

21. International Human Genome Sequencing C: **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**(7011):931-945.

22. Paik YK, Omenn GS, Uhlen M, Hanash S, Marko-Varga G, Aebersold R, Bairoch A, Yamamoto T, Legrain P, Lee HJ *et al*: **Standard guidelines for the chromosome-centric human proteome project**. *J Proteome Res* 2012, **11**(4):2005-2013.

23. Paik YK, Jeong SK, Omenn GS, Uhlen M, Hanash S, Cho SY, Lee HJ, Na K, Choi EY, Yan F *et al*: **The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome**. *Nat Biotechnol* 2012, **30**(3):221-223.

24. Legrain P, Aebersold R, Archakov A, Bairoch A, Bala K, Beretta L, Bergeron J, Borchers CH, Corthals GL, Costello CE *et al*: **The human proteome project: current state and future direction**. *Mol Cell Proteomics* 2011, **10**(7):M111 009993.

25. Hancock W, Omenn G, Legrain P, Paik YK: **Proteomics, human proteome project, and chromosomes**. *J Proteome Res* 2011, **10**(1):210.

26. Marko-Varga G, Omenn GS, Paik YK, Hancock WS: **A first step toward completion of a genome-wide characterization of the human proteome**. *J Proteome Res* 2013, **12**(1):1-5.

27. Gaudet P, Michel PA, Zahn-Zabal M, Britan A, Cusin I, Domagalski M, Duek PD, Gateau A, Gleizes A, Hinard V *et al*: **The neXtProt knowledgebase on human proteins: 2017 update**. *Nucleic Acids Res* 2017, **45**(D1):D177-D182.

28. Lane L, Bairoch A, Beavis RC, Deutsch EW, Gaudet P, Lundberg E, Omenn GS: **Metrics for the Human Proteome Project 2013-2014 and strategies for finding missing proteins**. *J Proteome Res* 2014, **13**(1):15-20.

29. Chang Z, Gu L: **Is the mission to identify all the human proteins achievable?--Commenting on the human proteome draft maps**. *Sci China Life Sci* 2014, **57**(10):1039-1040.

30. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S *et al*: **A draft map of the human proteome**. *Nature* 2014, **509**(7502):575-581.

31. Deutsch EW, Overall CM, Van Eyk JE, Baker MS, Paik YK, Weintraub ST, Lane L, Martens L, Vandenbrouck Y, Kusebauch U *et al*: **Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1**. *J Proteome Res* 2016, **15**(11):3961-3970.

32. Paik YK, Overall CM, Deutsch EW, Hancock WS, Omenn GS: **Progress in the Chromosome-Centric Human Proteome Project as Highlighted in the Annual Special Issue IV**. *J Proteome Res* 2016, **15**(11):3945-3950.

33. Islam MT, Mohamedali A, Garg G, Khan JM, Gorse AD, Parsons J, Marshall P, Ranganathan S, Baker MS: **Unlocking the puzzling biology of the black Perigord truffle Tuber melanosporum**. *J Proteome Res* 2013, **12**(12):5349-5356.

34. Martin F, Kohler A, Murat C, Balestrini R, Coutinho PM, Jaillon O, Montanini B, Morin E, Noel B, Percudani R *et al*: **Perigord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis**. *Nature* 2010, **464**(7291):1033-1038.

35. Aebersold R, Mann M: **Mass spectrometry-based proteomics**. *Nature* 2003, **422**(6928):198-207.

36. Thomson JJ: *Philos Mag* 1897, **44**:293-316.

37. Thomson JJ: **On Rays of Positive Electricity,**. *Philosophical Magazine Series 6* 1907, **13**(77):561.

38. Nier AO: *Rev Sci Instrum* 1940, **11**:212.

39. Aston FW: *Philos Mag* 1920, **39**:449-455.

40. Dass C: **Fundamentals of Contemporary Mass Spectrometry**. 2007.

41. Yates Iii JR: **A century of mass spectrometry: from atoms to proteomes**. *Nat Meth* 2011, **8**(8):633-637.

42. Ryhage R: **MS as a detector for GC**. *Anal Chem* 1964, **36**:759–764.

43. Watson JTaB, K. : **High-resolution MS of GC effluents**. *Anal Chem* 1965, **37**(844–851).

44. Jennings KR: **Collision-induced decompositions of aromatic molecular ions**. *Int J Mass Spectrom Ion Phys* 1968, **1**:227-235.

45. Munson MSBaF, F. H.: **Chemical ionization mass spectrometry, I: General introduction**. *J Am Chem Soc* 1966, **88**:2621–2630.

46. Barber M, Bordoli,R. S., Sedgwick, R. D. and Tyler, A. N. : **Fast atom bombardment of solids (F.A.B.): a new ion source for mass spectrometry**. *J Chem Soc Chem Commun* 1981:325–327.

47. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM: **Electrospray ionization for mass spectrometry of large biomolecules**. *Science* 1989, **246**(4926):64-71.

48. Karas M, Hillenkamp F: **Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons**. *Analytical Chemistry* 1988, **60**(20):2299-2301.

49. Tanaka K, Waki H, Ido Y, Akita S, Yoshida Y, Yoshida T, Matsuo T: **Protein and polymer analyses up tom/z 100 000 by laser ionization time-of-flight mass spectrometry**. *Rapid Communications in Mass Spectrometry* 1988, **2**(8):151-153.

50. Barber M, Bordoli RS, Sedgwick RD, Tyler AN: **Fast atom bombardment of solids as an ion source in mass spectrometry**. *Nature* 1981, **293**(5830):270-275.

51. Chowdhury SK, Katta V, Chait BT: **Electrospray ionization mass spectrometric peptide mapping: a rapid, sensitive technique for protein structure analysis**. *Biochem Biophys Res Commun* 1990, **167**(2):686-692.

52. Domon B, Aebersold R: **Mass spectrometry and protein analysis**. *Science* 2006, **312**(5771):212-217.

53. Chalmers MJ, Gaskell SJ: **Advances in mass spectrometry for proteome analysis**. *Curr Opin Biotechnol* 2000, **11**(4):384-390.

54. Naylor S, Kumar R: **Emerging role of mass spectrometry in structural and functional proteomics**. *Adv Protein Chem* 2003, **65**:217-248.

55. Reinders J, Lewandrowski U, Moebius J, Wagner Y, Sickmann A: **Challenges in mass spectrometry-based proteomics**. *Proteomics* 2004, **4**(12):3686-3703.

56. de Hoffmann E, Stroobant V: **Mass Spectrometry: Principles and Applications, 3rd Edition**: Wiley-Interscience; 2007.

57. Bleakney W: **A New Method of Positive Ray Analysis and Its Application to the Measurement of Ionization Potentials in Mercury Vapor**. *Physical Review* 1929, **34**(1):157-160.

58. Wilm M, Mann M: **Analytical properties of the nanoelectrospray ion source**. *Anal Chem* 1996, **68**(1):1-8.

59. Hillenkamp F, Karas M: **Mass spectrometry of peptides and proteins by matrix-assisted ultraviolet laser desorption/ionization**. *Methods Enzymol* 1990, **193**:280-295.

60. Griss J, Perez-Riverol Y, Lewis S, Tabb DL, Dianes JA, Del-Toro N, Rurik M, Walzer MW, Kohlbacher O, Hermjakob H *et al*: **Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets**. *Nat Methods* 2016, **13**(8):651-656.

61. McDonald WH, Yates JR, 3rd: **Shotgun proteomics: integrating technologies to answer biological questions**. *Curr Opin Mol Ther* 2003, **5**(3):302-309.

62. Zhang Y, Fonslow BR, Shan B, Baek MC, Yates JR, 3rd: **Protein analysis by shotgun/bottom-up proteomics**. *Chem Rev* 2013, **113**(4):2343-2394.

63. Han X, Aslanian A, Yates JR: **Mass spectrometry for proteomics**. *Current Opinion in Chemical Biology* 2008, **12**(5):483-490.

64. Hunt DF, Shabanowitz J, Yates JR, Zhu NZ, Russell DH, Castro ME: **Tandem quadrupole Fourier-transform mass spectrometry of oligopeptides and small proteins**. *Proceedings of the National Academy of Sciences* 1987, **84**(3):620-623.

65. Geiger T, Cox J, Mann M: **Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation**. *Mol Cell Proteomics* 2010, **9**(10):2252-2261.

66. Han X, Jin M, Breuker K, McLafferty FW: **Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons**. *Science* 2006, **314**(5796):109-112.

67. Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, Tipton JD, Vellaichamy A, Kellie JF, Li M *et al*: **Mapping intact protein isoforms in discovery mode using top-down proteomics**. *Nature* 2011, **480**(7376):254-258.

68. Catherman AD, Skinner OS, Kelleher NL: **Top Down proteomics: facts and perspectives**. *Biochem Biophys Res Commun* 2014, **445**(4):683-693.

69. **Proteomics: top-down or bottom-up?** [http://www.news-medical.net/whitepaper/20160112/Proteomics-top-down-or-bottom-up.aspx]

70. Lee WC, Lee KH: **Applications of affinity chromatography in proteomics**. *Anal Biochem* 2004, **324**(1):1-10.

71. Livesay EA, Tang K, Taylor BK, Buschbach MA, Hopkins DF, LaMarche BL, Zhao R, Shen Y, Orton DJ, Moore RJ *et al*: **Fully automated four-column capillary LC-MS system for maximizing throughput in proteomic analyses**. *Anal Chem* 2008, **80**(1):294-302.

72. Motoyama A, Xu T, Ruse CI, Wohlschlegel JA, Yates JR, 3rd: **Anion and cation mixed-bed ion exchange for enhanced multidimensional separations of peptides and phosphopeptides**. *Anal Chem* 2007, **79**(10):3623-3634.

73. Shi Y, Xiang R, Horvath C, Wilkins JA: **The role of liquid chromatography in proteomics**. *J Chromatogr A* 2004, **1053**(1-2):27-36.

74. Wu C, Tran JC, Zamdborg L, Durbin KR, Li M, Ahlf DR, Early BP, Thomas PM, Sweedler JV, Kelleher NL: **A protease for 'middle-down' proteomics**. *Nat Methods* 2012, **9**(8):822-824.

75. Michalski A, Damoc E, Hauschild JP, Lange O, Wieghaus A, Makarov A, Nagaraj N, Cox J, Mann M, Horning S: **Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer**. *Mol Cell Proteomics* 2011, **10**(9):M111 011015.

76. Olsen JV, de Godoy LM, Li G, Macek B, Mortensen P, Pesch R, Makarov A, Lange O, Horning S, Mann M: **Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap**. *Mol Cell Proteomics* 2005, **4**(12):2010-2021.

77. Szabo Z, Janaky T: **Challenges and developments in protein identification using mass spectrometry**. *TrAC Trends in Analytical Chemistry* 2015, **69**:76-87.

78. Tuloup M, Hernandez C, Coro I, Hoogland C, Binz P, Appel R: **Aldente and BioGraph: An improved peptide mass fingerprinting protein identification environment**. In: *Proceedings of the Swiss Proteomics Society 2003 Congress: Understanding Biological Systems through Proteomics: 2003*. 174-176.

79. Peri S, Steen H, Pandey A: **GPMAW &#x2013; a software tool for analyzing proteins and peptides**. *Trends in Biochemical Sciences*, **26**(11):687-689.

80. Pappin DJ, Hojrup P, Bleasby AJ: **Rapid identification of proteins by peptide-mass fingerprinting**. *Curr Biol* 1993, **3**(6):327-332.

81. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data**. *Electrophoresis* 1999, **20**(18):3551-3567.

82. Song Z, Chen L, Ganapathy A, Wan XF, Brechenmacher L, Tao N, Emerich D, Stacey G, Xu D: **Development and assessment of scoring functions for protein identification using PMF data**. *Electrophoresis* 2007, **28**(5):864-870.

83. **ProteinProspector** [http://prospector.ucsf.edu/prospector/]

84. Zhang W, Chait BT: **ProFound: an expert system for protein identification using mass spectrometric peptide mapping information**. *Anal Chem* 2000, **72**(11):2482-2489.

85. Steen H, Mann M: **The ABC's (and XYZ's) of peptide sequencing**. *Nat Rev Mol Cell Biol* 2004, **5**(9):699-711.

86. Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA: ***De novo* peptide sequencing via tandem mass spectrometry**. *J Comput Biol* 1999, **6**(3-4):327-342.

87. Kim S, Gupta N, Bandeira N, Pevzner PA: **Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra**. *Mol Cell Proteomics* 2009, **8**(1):53-69.

88. Taylor JA, Johnson RS: **Sequence database searches via de novo peptide sequencing by tandem mass spectrometry**. *Rapid Commun Mass Spectrom* 1997, **11**(9):1067-1075.

89. Frank A, Pevzner P: **PepNovo: de novo peptide sequencing via probabilistic network modeling**. *Anal Chem* 2005, **77**(4):964-973.

90. Nesvizhskii AI: **Protein identification by tandem mass spectrometry and sequence database searching**. *Methods Mol Biol* 2007, **367**:87-119.

91. Frank AM, Savitski MM, Nielsen ML, Zubarev RA, Pevzner PA: **De novo peptide sequencing and identification with precision mass spectrometry**. *J Proteome Res* 2007, **6**(1):114-123.

92. Eng JK, Searle BC, Clauser KR, Tabb DL: **A face in the crowd: recognizing peptides through database search**. *Mol Cell Proteomics* 2011, **10**(11):R111 009522.

93. Eng JK, McCormack AL, Yates JR: **An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database**. *J Am Soc Mass Spectrom* 1994, **5**(11):976-989.

94. Yates JR, 3rd, Eng JK, McCormack AL, Schieltz D: **Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database**. *Anal Chem* 1995, **67**(8):1426-1436.

95. Craig R, Beavis RC: **TANDEM: matching proteins with tandem mass spectra**. *Bioinformatics* 2004, **20**(9):1466-1467.

96. Granholm V, Kall L: **Quality assessments of peptide-spectrum matches in shotgun proteomics**. *Proteomics* 2011, **11**(6):1086-1093.

97. Huang T, Wang J, Yu W, He Z: **Protein inference: a review**. *Brief Bioinform* 2012, **13**(5):586-614.

98. Elias JE, Gygi SP: **Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry**. *Nat Methods* 2007, **4**(3):207-214.

99. Kall L, Storey JD, MacCoss MJ, Noble WS: **Assigning significance to peptides identified by tandem mass spectrometry using decoy databases**. *J Proteome Res* 2008, **7**(1):29-34.

100. Yadav AK, Kumar D, Dash D: **Learning from decoys to improve the sensitivity and specificity of proteomics database search results**. *PLoS One* 2012, **7**(11):e50651.

101. Elias JE, Gygi SP: **Target-decoy search strategy for mass spectrometry-based proteomics**. *Methods Mol Biol* 2010, **604**:55-71.

102. Zhang N, Aebersold R, Schwikowski B: **ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data**. *Proteomics* 2002, **2**(10):1406-1412.

103. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH: **Open mass spectrometry search algorithm**. *J Proteome Res* 2004, **3**(5):958-964.

104. Cox J, Mann M: **MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification**. *Nat Biotechnol* 2008, **26**(12):1367-1372.

105. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M: **Andromeda: a peptide search engine integrated into the MaxQuant environment**. *J Proteome Res* 2011, **10**(4):1794-1805.

106. Yadav AK, Kumar D, Dash D: **MassWiz: a novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry**. *J Proteome Res* 2011, **10**(5):2154-2160.

107. Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V: **InsPecT: identification of posttranslationally modified peptides from tandem mass spectra**. *Anal Chem* 2005, **77**(14):4626-4639.

108. Kwon T, Choi H, Vogel C, Nesvizhskii AI, Marcotte EM: **MSblender: A probabilistic approach for integrating peptide identifications from multiple database search engines**. *Journal of proteome research* 2011, **10**(7):2949-2958.

109. Lietzen N, Natri L, Nevalainen OS, Salmi J, Nyman TA: **Compid: a new software tool to integrate and compare MS/MS based protein identification results from Mascot and Paragon**. *J Proteome Res* 2010, **9**(12):6795-6800.

110. Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP, Hunter CL, Nuwaysir LM, Schaeffer DA: **The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra**. *Mol Cell Proteomics* 2007, **6**(9):1638-1655.

111. Hamdan MH, Righetti PG: **Proteomics Today: Protein Assessment and Biomarkers Using Mass Spectrometry, 2D Electrophoresis,and Microarray Technology**: Wiley; 2005.

112. Carnielli CM, Winck FV, Paes Leme AF: **Functional annotation and biological interpretation of proteomics data**. *Biochimica et biophysica acta* 2014, **1854**(1):46-54.

113. Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, Tasman N, Mendoza L, Moritz RL, Aebersold R, Nesvizhskii AI: **iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates**. *Molecular & Cellular Proteomics* 2011, **10**(12):M111. 007690.

114. Searle BC: **Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies**. *Proteomics* 2010, **10**(6):1265-1269.

115. Edwards N, Wu X, Tseng C-W: **An unsupervised, model-free, machine-learning combiner for peptide identifications from tandem mass spectra**. *Clinical Proteomics* 2009, **5**(1):23-36.

116. Wedge DC, Krishna R, Blackhurst P, Siepen JA, Jones AR, Hubbard SJ: **FDRAnalysis: a tool for the integrated analysis of tandem mass spectrometry identification results from multiple search engines**. *Journal of proteome research* 2011, **10**(4):2088-2094.

117. Shteynberg D, Nesvizhskii AI, Moritz RL, Deutsch EW: **Combining results of multiple search engines in proteomics**. *Molecular & Cellular Proteomics* 2013, **12**(9):2383-2393.

118. Matthiesen R: **Mass spectrometry data analysis in proteomics**, vol. 1: Springer; 2007.

119. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR, 3rd: **Direct analysis of protein complexes using mass spectrometry**. *Nat Biotechnol* 1999, **17**(7):676-682.

120. Yu K, Sabelli A, DeKeukelaere L, Park R, Sindi S, Gatsonis CA, Salomon A: **Integrated platform for manual and high-throughput statistical validation of tandem mass spectra**. *Proteomics* 2009, **9**(11):3115-3125.

121. Sun S, Meyer-Arendt K, Eichelberger B, Brown R, Yen CY, Old WM, Pierce K, Cios KJ, Ahn NG, Resing KA: **Improved validation of peptide MS/MS assignments using spectral intensity prediction**. *Mol Cell Proteomics* 2007, **6**(1):1-17.

122. Tsou CC, Tsui YH, Yian YH, Chen YJ, Yang HY, Yu CY, Lynn KS, Chen YJ, Sung TY, Hsu WL: **MaXIC-Q Web: a fully automated web service using statistical and computational methods for protein quantitation based on stable isotope labeling and LC-MS**. *Nucleic Acids Res* 2009, **37**(Web Server issue):W661-669.

123. Tsou CC, Tsai CF, Tsui YH, Sudhir PR, Wang YT, Chen YJ, Chen JY, Sung TY, Hsu WL: **IDEAL-Q, an automated tool for label-free quantitation analysis using an efficient peptide alignment approach and spectral data validation**. *Mol Cell Proteomics* 2010, **9**(1):131-144.

124. Yandell M, Ence D: **A beginner's guide to eukaryotic genome annotation**. *Nat Rev Genet* 2012, **13**(5):329-342.

125. Richardson EJ, Watson M: **The automatic annotation of bacterial genomes**. *Brief Bioinform* 2013, **14**(1):1-12.

126. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL: **Improved microbial gene identification with GLIMMER**. *Nucleic Acids Res* 1999, **27**(23):4636-4641.

127. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification**. *BMC Bioinformatics* 2010, **11**:119.

128. Lukashin AV, Borodovsky M: **GeneMark.hmm: new solutions for gene finding**. *Nucleic Acids Res* 1998, **26**(4):1107-1115.

129. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA**. *J Mol Biol* 1997, **268**(1):78-94.

130. Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel**. *Bioinformatics* 2003, **19 Suppl 2**:ii215-225.

131. Holt C, Yandell M: **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects**. *BMC Bioinformatics* 2011, **12**:491.

132. Aivaliotis M, Gevaert K, Falb M, Tebbe A, Konstantinidis K, Bisle B, Klein C, Martens L, Staes A, Timmerman E *et al*: **Large-scale identification of N-terminal peptides in the halophilic archaea Halobacterium salinarum and Natronomonas pharaonis**. *J Proteome Res* 2007, **6**(6):2195-2204.

133. Lamontagne J, Beland M, Forest A, Cote-Martin A, Nassif N, Tomaki F, Moriyon I, Moreno E, Paramithiotis E: **Proteomics-based confirmation of protein expression and correction of annotation errors in the Brucella abortus genome**. *BMC Genomics* 2010, **11**:300.

134. Kuster B, Mortensen P, Andersen JS, Mann M: **Mass spectrometry allows direct identification of proteins in large genomes**. *Proteomics* 2001, **1**(5):641-650.

135. Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Briggs SP: **Discovery and revision of Arabidopsis genes by proteogenomics**. *Proc Natl Acad Sci U S A* 2008, **105**(52):21034-21038.

136. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

137. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment**. *Proteins* 1991, **9**(1):56-68.

138. Islam MT, Garg G, Hancock WS, Risk BA, Baker MS, Ranganathan S: **Protannotator: a semiautomated pipeline for chromosome-wise functional annotation of the "missing" human proteome**. *J Proteome Res* 2014, **13**(1):76-83.

139. Remmert M, Biegert A, Hauser A, Soding J: **HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment**. *Nat Methods* 2011, **9**(2):173-175.

140. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res* 2002, **12**(4):656-664.

141. Lee BT, Tan TW, Ranganathan S: **MGAlignIt: A web service for the alignment of mRNA/EST and genomic sequences**. *Nucleic Acids Res* 2003, **31**(13):3533-3536.

142. The UniProt C: **UniProt: the universal protein knowledgebase**. *Nucleic Acids Res* 2017, **45**(D1):D158-D169.

143. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I *et al*: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003**. *Nucleic Acids Res* 2003, **31**(1):365-370.

144. O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, Apweiler R: **High-quality protein knowledge resource: SWISS-PROT and TrEMBL**. *Brief Bioinform* 2002, **3**(3):275-284.

145. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic acids research* 2000, **28**(1):235-242.

146. Rose PW, Prlic A, Altunkaya A, Bi C, Bradley AR, Christie CH, Costanzo LD, Duarte JM, Dutta S, Feng Z *et al*: **The RCSB protein data bank: integrative view of protein, gene and 3D structural information**. *Nucleic Acids Res* 2017, **45**(D1):D271-D281.

147. Ranganathan S, Khan JM, Garg G, Baker MS: **Functional annotation of the human chromosome 7 "missing" proteins: a bioinformatics approach**. *J Proteome Res* 2013, **12**(6):2504-2510.

148. Young MD, Wakefield MJ, Smyth GK, Oshlack A: **Gene ontology analysis for RNA-seq: accounting for selection bias**. *Genome Biol* 2010, **11**(2):R14.

149. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.

150. Gene Ontology C: **Gene Ontology Consortium: going forward**. *Nucleic Acids Res* 2015, **43**(Database issue):D1049-1056.

151. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**(18):3674-3676.

152. Koskinen P, Toronen P, Nokso-Koivisto J, Holm L: **PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment**. *Bioinformatics* 2015, **31**(10):1544-1552.

153. Das S, Lee D, Sillitoe I, Dawson NL, Lees JG, Orengo CA: **Functional classification of CATH superfamilies: a domain-based approach for protein function annotation**. *Bioinformatics* 2015, **31**(21):3460-3467.

154. Das S, Sillitoe I, Lee D, Lees JG, Dawson NL, Ward J, Orengo CA: **CATH FunFHMMer web server: protein functional annotations using functional family assignments**. *Nucleic Acids Res* 2015, **43**(W1):W148-153.

155. Fang H: **dcGOR: an R package for analysing ontologies and protein domain annotations**. *PLoS Comput Biol* 2014, **10**(10):e1003929.

156. McCarthy FM, Wang N, Magee GB, Nanduri B, Lawrence ML, Camon EB, Barrell DG, Hill DP, Dolan ME, Williams WP *et al*: **AgBase: a functional genomics resource for agriculture**. *BMC Genomics* 2006, **7**:229.

157. [ftp://ftp.geneontology.org/pub/go/www/GO.tools.annotation.shtml]

158. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztanyi Z, El-Gebali S, Fraser M *et al*: **InterPro in 2017-beyond protein family and domain annotations**. *Nucleic Acids Res* 2017, **45**(D1):D190-D199.

159. Akiva E, Brown S, Almonacid DE, Barber AE, 2nd, Custer AF, Hicks MA, Huang CC, Lauck F, Mashiyama ST, Meng EC *et al*: **The Structure-Function Linkage Database**. *Nucleic Acids Res* 2014, **42**(Database issue):D521-530.

160. Holliday GL, Brown SD, Akiva E, Mischel D, Hicks MA, Morris JH, Huang CC, Meng EC, Pegg SC, Ferrin TE *et al*: **Biocuration in the structure-function linkage database: the anatomy of a superfamily**. *Database (Oxford)* 2017, **2017**(1).

161. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR *et al*: **CDD/SPARCLE: functional classification of proteins via subfamily domain architectures**. *Nucleic Acids Res* 2017, **45**(D1):D200-D203.

162. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G *et al*: **InterProScan 5: genome-scale protein function classification**. *Bioinformatics* 2014, **30**(9):1236-1240.

163. Dawson NL, Sillitoe I, Lees JG, Lam SD, Orengo CA: **CATH-Gene3D: Generation of the Resource and Its Use in Obtaining Structural and Functional Annotations for Protein Sequences**. *Methods Mol Biol* 2017, **1558**:79-110.

164. Lam SD, Dawson NL, Das S, Sillitoe I, Ashford P, Lee D, Lehtinen S, Orengo CA, Lees JG: **Gene3D: expanding the utility of domain assignments**. *Nucleic Acids Res* 2016, **44**(D1):D404-409.

165. Necci M, Piovesan D, Dosztanyi Z, Tosatto SCE: **MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins**. *Bioinformatics* 2017, **33**(9):1402-1404.

166. Di Domenico T, Walsh I, Martin AJ, Tosatto SC: **MobiDB: a comprehensive database of intrinsic protein disorder annotations**. *Bioinformatics* 2012, **28**(15):2080-2081.

167. Pedruzzi I, Rivoire C, Auchincloss AH, Coudert E, Keller G, de Castro E, Baratin D, Cuche BA, Bougueleret L, Poux S *et al*: **HAMAP in 2015: updates to the protein family classification and annotation system**. *Nucleic Acids Res* 2015, **43**(Database issue):D1064-1070.

168. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD: **PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements**. *Nucleic Acids Res* 2017, **45**(D1):D183-D189.

169. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A *et al*: **The Pfam protein families database: towards a more sustainable future**. *Nucleic Acids Res* 2016, **44**(D1):D279-285.

170. Nikolskaya AN, Arighi CN, Huang H, Barker WC, Wu CH: **PIRSF family classification system for protein functional and evolutionary analysis**. *Evol Bioinform Online* 2007, **2**:197-209.

171. Wu CH, Nikolskaya A, Huang H, Yeh LS, Natale DA, Vinayaka CR, Hu ZZ, Mazumder R, Kumar S, Kourtesis P *et al*: **PIRSF: family classification system at the Protein Information Resource**. *Nucleic Acids Res* 2004, **32**(Database issue):D112-114.

172. Attwood TK, Coletta A, Muirhead G, Pavlopoulou A, Philippou PB, Popov I, Roma-Mateo C, Theodosiou A, Mitchell AL: **The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012**. *Database (Oxford)* 2012, **2012**:bas019.

173. Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, Kahn D: **The ProDom database of protein domain families: more emphasis on 3D**. *Nucleic Acids Res* 2005, **33**(Database issue):D212-215.

174. Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I: **New and continuing developments at PROSITE**. *Nucleic Acids Res* 2013, **41**(Database issue):D344-347.

175. Letunic I, Doerks T, Bork P: **SMART: recent updates, new developments and status in 2015**. *Nucleic Acids Res* 2015, **43**(Database issue):D257-260.

176. Oates ME, Stahlhacke J, Vavoulis DV, Smithers B, Rackham OJ, Sardar AJ, Zaucha J, Thurlby N, Fang H, Gough J: **The SUPERFAMILY 1.75 database in 2014: a doubling of data**. *Nucleic Acids Res* 2015, **43**(Database issue):D227-233.

177. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J: **SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny**. *Nucleic Acids Res* 2009, **37**(Database issue):D380-386.

178. Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E: **TIGRFAMs and Genome Properties in 2013**. *Nucleic Acids Res* 2013, **41**(Database issue):D387-395.

179. Blake JA: **Ten quick tips for using the gene ontology**. *PLoS Comput Biol* 2013, **9**(11):e1003343.

180. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M: **KEGG as a reference resource for gene and protein annotation**. *Nucleic Acids Res* 2016, **44**(D1):D457-462.

181. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.

182. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K: **KEGG: new perspectives on genomes, pathways, diseases and drugs**. *Nucleic Acids Res* 2017, **45**(D1):D353-D361.

183. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L: **KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases**. *Nucleic Acids Res* 2011, **39**(Web Server issue):W316-322.

184. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: **KAAS: an automatic genome annotation and pathway reconstruction server**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W182-185.

185. Bebek G, Yang J: **PathFinder: mining signal transduction pathway segments from protein-protein interaction networks**. *BMC Bioinformatics* 2007, **8**:335.

186. Burlingame A, Carr SA, Bradshaw RA, Chalkley RJ: **On Credibility, Clarity, and Compliance**. *Mol Cell Proteomics* 2015, **14**(7):1731-1733.

187. **Democratizing proteomics data**. *Nature Biotechnology* 2007, **25**(3):262-262.

188. **Thou shalt share your data**. *Nature Methods* 2008, **5**(3):209-209.

189. Kaiser J: **Proteomics. Public-private group maps out initiatives**. *Science* 2002, **296**(5569):827.

190. Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Rompp A, Neumann S, Pizarro AD *et al*: **mzML--a community standard for mass spectrometry data**. *Mol Cell Proteomics* 2011, **10**(1):R110 000133.

191. Jones AR, Eisenacher M, Mayer G, Kohlbacher O, Siepen J, Hubbard SJ, Selley JN, Searle BC, Shofstahl J, Seymour SL *et al*: **The mzIdentML data standard for mass spectrometry-based proteomics results**. *Mol Cell Proteomics* 2012, **11**(7):M111 014381.

192. Griss J, Jones AR, Sachsenberg T, Walzer M, Gatto L, Hartler J, Thallinger GG, Salek RM, Steinbeck C, Neuhauser N *et al*: **The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience**. *Mol Cell Proteomics* 2014, **13**(10):2765-2775.

193. Walzer M, Qi D, Mayer G, Uszkoreit J, Eisenacher M, Sachsenberg T, Gonzalez-Galarza FF, Fan J, Bessant C, Deutsch EW *et al*: **The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics**. *Mol Cell Proteomics* 2013, **12**(8):2332-2340.

194. Deutsch EW, Chambers M, Neumann S, Levander F, Binz PA, Shofstahl J, Campbell DS, Mendoza L, Ovelleiro D, Helsens K *et al*: **TraML--a standard format for exchange of selected reaction monitoring transition lists**. *Mol Cell Proteomics* 2012, **11**(4):R111 015040.

195. Vaudel M, Verheggen K, Csordas A, Raeder H, Berven FS, Martens L, Vizcaino JA, Barsnes H: **Exploring the potential of public proteomics data**. *Proteomics* 2016, **16**(2):214-225.

196. 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 Res* 2017, **45**(D1):D1100-D1106.

197. Vizcaino JA, Csordas A, del-Toro N, Dianes JA, Griss J, Lavidas I, Mayer G, Perez-Riverol Y, Reisinger F, Ternent T *et al*: **2016 update of the PRIDE database and its related tools**. *Nucleic Acids Res* 2016, **44**(D1):D447-456.

198. Deutsch EW, Lam H, Aebersold R: **PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows**. *EMBO Rep* 2008, **9**(5):429-434.

199. Farrah T, Deutsch EW, Kreisberg R, Sun Z, Campbell DS, Mendoza L, Kusebauch U, Brusniak MY, Huttenhain R, Schiess R *et al*: **PASSEL: the PeptideAtlas SRMexperiment library**. *Proteomics* 2012, **12**(8):1170-1175.

200. **Welcome to MassIVE** [http://massive.ucsd.edu]

201. Okuda S, Watanabe Y, Moriya Y, Kawano S, Yamamoto T, Matsumoto M, Takami T, Kobayashi D, Araki N, Yoshizawa AC *et al*: **jPOSTrepo: an international standard data repository for proteomes**. *Nucleic Acids Res* 2017, **45**(D1):D1107-D1111.

202. Craig R, Cortens JP, Beavis RC: **Open source system for analyzing, validating, and storing protein identification data**. *J Proteome Res* 2004, **3**(6):1234-1242.

203. Schaab C, Geiger T, Stoehr G, Cox J, Mann M: **Analysis of high accuracy, quantitative proteomics data in the MaxQB database**. *Mol Cell Proteomics* 2012, **11**(3):M111 014068.

204. Montague E, Janko I, Stanberry L, Lee E, Choiniere J, Anderson N, Stewart E, Broomall W, Higdon R, Kolker N *et al*: **Beyond protein expression, MOPED goes multi-omics**. *Nucleic Acids Res* 2015, **43**(Database issue):D1145-1151.

205. Montague E, Stanberry L, Higdon R, Janko I, Lee E, Anderson N, Choiniere J, Stewart E, Yandl G, Broomall W *et al*: **MOPED 2.5--an integrated multi-omics resource: multi-omics profiling expression database now includes transcriptomics data**. *OMICS* 2014, **18**(6):335-343.

206. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C: **Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines**. *Proteomics* 2015, **15**(18):3163-3168.

207. Kandasamy K, Keerthikumar S, Goel R, Mathivanan S, Patankar N, Shafreen B, Renuse S, Pawar H, Ramachandra YL, Acharya PK *et al*: **Human Proteinpedia: a unified discovery resource for proteomics research**. *Nucleic Acids Res* 2009, **37**(Database issue):D773-781.

208. Ternent T, Csordas A, Qi D, Gomez-Baena G, Beynon RJ, Jones AR, Hermjakob H, Vizcaino JA: **How to submit MS proteomics data to ProteomeXchange via the PRIDE database**. *Proteomics* 2014, **14**(20):2233-2241.

209. Cote RG, Griss J, Dianes JA, Wang R, Wright JC, van den Toorn HW, van Breukelen B, Heck AJ, Hulstaert N, Martens L *et al*: **The PRoteomics IDEntification (PRIDE) Converter 2 framework: an improved suite of tools to

facilitate data submission to the PRIDE database and the ProteomeXchange consortium. *Mol Cell Proteomics* 2012, **11**(12):1682-1689.

210. **PRIDE Archive** [https://www.ebi.ac.uk/pride/archive/]

211. Wang R, Fabregat A, Rios D, Ovelleiro D, Foster JM, Cote RG, Griss J, Csordas A, Perez-Riverol Y, Reisinger F *et al*: **PRIDE Inspector: a tool to visualize and validate MS proteomics data**. *Nat Biotechnol* 2012, **30**(2):135-137.

212. Reisinger F, del-Toro N, Ternent T, Hermjakob H, Vizcaino JA: **Introducing the PRIDE Archive RESTful web services**. *Nucleic Acids Res* 2015, **43**(W1):W599-604.

213. Griss J, Foster JM, Hermjakob H, Vizcaino JA: **PRIDE Cluster: building a consensus of proteomics data**. *Nat Methods* 2013, **10**(2):95-96.

214. **ISB Data Server** [http://www.peptideatlas.org/upload/]

215. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L *et al*: **Ensembl 2016**. *Nucleic Acids Res* 2016, **44**(D1):D710-716.

216. Griss J, Martin M, O'Donovan C, Apweiler R, Hermjakob H, Vizcaino JA: **Consequences of the discontinuation of the International Protein Index (IPI) database and its substitution by the UniProtKB "complete proteome" sets**. *Proteomics* 2011, **11**(22):4434-4438.

217. Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, Sun Z, Nilsson E, Pratt B, Prazen B *et al*: **A guided tour of the Trans-Proteomic Pipeline**. *Proteomics* 2010, **10**(6):1150-1159.

218. Keller A, Nesvizhskii AI, Kolker E, Aebersold R: **Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search**. *Anal Chem* 2002, **74**(20):5383-5392.

219. Nesvizhskii AI, Keller A, Kolker E, Aebersold R: **A statistical model for identifying proteins by tandem mass spectrometry**. *Anal Chem* 2003, **75**(17):4646-4658.

220. Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, Buhmann JM, Hengartner MO, Aebersold R: **Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry**. *Mol Cell Proteomics* 2009, **8**(11):2405-2417.

221. **PeptideAtlas** [http://www.peptideatlas.org/peptideatlasExplorer/]

222. **PeptideAtlas and the Chromosome-Centric Human Proteome Project** [http://www.peptideatlas.org/hupo/c-hpp/]

223. **ISB Data Server** [http://www.peptideatlas.org/submit]

224. Reiter L, Rinner O, Picotti P, Huttenhain R, Beck M, Brusniak MY, Hengartner MO, Aebersold R: **mProphet: automated data processing and statistical validation for large-scale SRM experiments**. *Nat Methods* 2011, **8**(5):430-435.

225. Kusebauch U, Campbell DS, Deutsch EW, Chu CS, Spicer DA, Brusniak MY, Slagel J, Sun Z, Stevens J, Grimes B *et al*: **Human SRMAtlas: A Resource of Targeted Assays to Quantify the Complete Human Proteome**. *Cell* 2016, **166**(3):766-778.

226. **PASSEL** [http://www.peptideatlas.org/passel]

227. **SRM Experiments Available in PASSEL** [https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/GetSELExperiments]

228. **PeptideAtlas** [https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/GetSELTransitions]

229. Kusebauch U, Deutsch EW, Campbell DS, Sun Z, Farrah T, Moritz RL: **Using PeptideAtlas, SRMAtlas, and PASSEL: Comprehensive Resources for Discovery and Targeted Proteomics**. *Curr Protoc Bioinformatics* 2014, **46**:13 25 11-28.

230. **CCMS ProteoSAFe Workflow Input Form** [URLhttp://massive.ucsd.edu/ProteoSAFe]

231. **MassIVE Dataset Submission - Confluence** [https://bix-lab.ucsd.edu/display/PS/MassIVE+Dataset+Submission -MassIVEDatasetSubmission-1UploadDatatoMassIVE]

232. **UCSD/CCMS - MassIVE Datasets - Mass Spectrometry Repository Dataset List** [http://massive.ucsd.edu/ProteoSAFe/datasets.jsp]

233. **MassIVE Search** [http://massive.ucsd.edu/ProteoSAFe/massive_search.jsp]

234. **CSE Bioinformatics Group** [http://proteomics.ucsd.edu/Software]

235. **jPOSTrepo (Japan ProteOme STandard Repository)** [https://repository.jpostdb.org/submit]

236. **jPOSTrepo (Japan ProteOme STandard Repository)** [https://repository.jpostdb.org]

237. **GPM Cyclone, simple search form** [http://h003.thegpm.org/tandem/thegpm_tandem.html]

238. Craig R, Cortens JP, Beavis RC: **The use of proteotypic peptide libraries for protein identification**. *Rapid Commun Mass Spectrom* 2005, **19**(13):1844-1850.

239. Craig R, Cortens JC, Fenyo D, Beavis RC: **Using annotated peptide mass spectrum libraries for protein identification**. *J Proteome Res* 2006, **5**(8):1843-1849.

240. **Proteomics DB** [https://www.proteomicsdb.org/ - overview]

241. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M: **Global quantification of mammalian gene expression control**. *Nature* 2011, **473**(7347):337-342.

242. Higdon R, Stewart E, Stanberry L, Haynes W, Choiniere J, Montague E, Anderson N, Yandl G, Janko I, Broomall W *et al*: **MOPED enables discoveries through consistently processed proteomics data**. *J Proteome Res* 2014, **13**(1):107-113.

243. Kolker E, Higdon R, Morgan P, Sedensky M, Welch D, Bauman A, Stewart E, Haynes W, Broomall W, Kolker N: **SPIRE: Systematic protein investigative research environment**. *J Proteomics* 2011, **75**(1):122-126.

244. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP *et al*: **STRING v10: protein-protein interaction networks, integrated over the tree of life**. *Nucleic Acids Res* 2015, **43**(Database issue):D447-452.

245. Muthusamy B, Thomas JK, Prasad TS, Pandey A: **Access guide to human proteinpedia**. *Curr Protoc Bioinformatics* 2013, **Chapter 1**:Unit 1 21.

246. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A *et al*: **Human Protein Reference Database--2009 update**. *Nucleic Acids Res* 2009, **37**(Database issue):D767-772.

247. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000**. *Nucleic Acids Res* 2000, **28**(1):45-48.

248. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence data bank and its new supplement TREMBL**. *Nucleic Acids Res* 1996, **24**(1):21-25.

249. Bairoch A: **Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times!** *Bioinformatics* 2000, **16**(1):48-64.

250. Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE *et al*: **The Protein Information Resource**. *Nucleic Acids Res* 2003, **31**(1):345-347.

251. Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, Apweiler R: **UniProt archive**. *Bioinformatics* 2004, **20**(17):3236-3237.

252. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH: **UniRef: comprehensive and non-redundant UniProt reference clusters**. *Bioinformatics* 2007, **23**(10):1282-1288.

253. **The universal protein resource (UniProt)**. *Nucleic Acids Res* 2008, **36**(Database issue):D190-195.

254. **European Bioinformatics Institute** [http://www.uniprot.org/help/protein_existence]

255. Ross KE, Huang H, Ren J, Arighi CN, Li G, Tudor CO, Lv M, Lee JY, Chen SC, Vijay-Shanker K *et al*: **iPTMnet: Integrative Bioinformatics for Studying PTM Networks**. *Methods Mol Biol* 2017, **1558**:333-353.

256. Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V, Sullivan M: **PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse**. *Nucleic Acids Res* 2012, **40**(Database issue):D261-270.

257. Schaeffer M, Gateau A, Teixeira D, Michel PA, Zahn-Zabal M, Lane L: **The neXtProt peptide uniqueness checker: a tool for the proteomics community**. *Bioinformatics* 2017.

258. **GeneCards®: The Human Gene Database** [http://www.genecards.org/]

259. **GeneCards Sources** [http://www.genecards.org/Guide/Sources]

260. Belinky F, Nativ N, Stelzer G, Zimmerman S, Iny Stein T, Safran M, Lancet D: **PathCards: multi-source consolidation of human biological pathways**. *Database (Oxford)* 2015, **2015**.

261. Rappaport N, Twik M, Plaschkes I, Nudel R, Iny Stein T, Levitt J, Gershoni M, Morrey CP, Safran M, Lancet D: **MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search**. *Nucleic Acids Res* 2017, **45**(D1):D877-D887.

262. Rosen N, Chalifa-Caspi V, Shmueli O, Adato A, Lapidot M, Stampnitzky J, Safran M, Lancet D: **GeneLoc: exon-based integration of human genome maps**. *Bioinformatics* 2003, **19 Suppl 1**:i222-224.

263. Ben-Ari Fuchs S, Lieder I, Stelzer G, Mazor Y, Buzhor E, Kaplan S, Bogoch Y, Plaschkes I, Shitrit A, Rappaport N *et al*: **GeneAnalytics: An Integrative Gene Set Analysis Tool for Next Generation Sequencing, RNAseq and Microarray Data**. *OMICS* 2016, **20**(3):139-151.

264. **National Center for Biotechnology Information** [https://www.ncbi.nlm.nih.gov/pubmed/]

265. **PubMed Clinical Queries** [https://www.ncbi.nlm.nih.gov/entrez/query/static/clinical.shtml]

266. **LinkOut: General Information** [https://www.ncbi.nlm.nih.gov/entrez/linkout/]

267. Berman H, Henrick K, Nakamura H: **Announcing the worldwide Protein Data Bank**. *Nature Structural Biology* 2003, **10**(12):980-980.

268. Berman HM: **The Protein Data Bank: a historical perspective**. *Acta Crystallogr A* 2008, **64**(Pt 1):88-95.

269. Kinjo AR, Suzuki H, Yamashita R, Ikegawa Y, Kudou T, Igarashi R, Kengaku Y, Cho H, Standley DM, Nakagawa A *et al*: **Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format**. *Nucleic Acids Res* 2012, **40**(Database issue):D453-460.

270. Velankar S, van Ginkel G, Alhroub Y, Battle GM, Berrisford JM, Conroy MJ, Dana JM, Gore SP, Gutmanas A, Haslam P *et al*: **PDBe: improved accessibility of macromolecular structure data from PDB and EMDB**. *Nucleic Acids Res* 2016, **44**(D1):D385-395.

271. Rose AS, Hildebrand PW: **NGL Viewer: a web application for molecular visualization**. *Nucleic Acids Res* 2015, **43**(W1):W576-579.

272. Garavelli JS: **The RESID database of protein structure modifications: 2000 update**. *Nucleic Acids Res* 2000, **28**(1):209-211.

273. Touw WG, Baakman C, Black J, te Beek TA, Krieger E, Joosten RP, Vriend G: **A series of PDB-related databanks for everyday needs**. *Nucleic Acids Res* 2015, **43**(Database issue):D364-368.

274. Cardozo T, Batalov S, Abagyan R: **Estimating local backbone structural deviation in homology models**. *Comput Chem* 2000, **24**(1):13-31.

275. Martens L, Vizcaino JA: **A Golden Age for Working with Public Proteomics Data**. *Trends Biochem Sci* 2017, **42**(5):333-341.

276. Barone L, Williams J, Micklos D: **Unmet Needs for Analyzing Biological Big Data: A Survey of 704 NSF Principal Investigators**. *bioRxiv* 2017.

277. Edwards NJ, Oberti M, Thangudu RR, Cai S, McGarvey PB, Jacob S, Madhavan S, Ketchum KA: **The CPTAC Data Portal: A Resource for Cancer Proteomics Research**. *J Proteome Res* 2015, **14**(6):2707-2713.

278. Huang KL, Li S, Mertins P, Cao S, Gunawardena HP, Ruggles KV, Mani DR, Clauser KR, Tanioka M, Usary J *et al*: **Proteogenomic integration reveals therapeutic targets in breast cancer xenografts**. *Nat Commun* 2017, **8**:14864.

279. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR *et al*: **Saccharomyces Genome Database: the genomics resource of budding yeast**. *Nucleic Acids Res* 2012, **40**(Database issue):D700-705.

280. Perez-Riverol Y, Bai M, da Veiga Leprevost F, Squizzato S, Park YM, Haug K, Carroll AJ, Spalding D, Paschall J, Wang M *et al*: **Discovering and linking public omics data sets using the Omics Discovery Index**. *Nature Biotechnology* 2017, **35**(5):406-409.

281. Das S, Orengo CA: **Protein function annotation using protein domain family resources**. *Methods* 2016, **93**:24-34.

282. Omenn GS, Lane L, Lundberg EK, Beavis RC, Overall CM, Deutsch EW: **Metrics for the Human Proteome Project 2016: Progress on Identifying and Characterizing the Human Proteome, Including Post-Translational Modifications**. *J Proteome Res* 2016, **15**(11):3951-3960.

283. Baker MS, Ahn SB, Mohamedali A, Islam MT, Cantor D, Verhaert PD, Fanayan S, Sharma S, Nice EC, Connor M *et al*: **Accelerating the search for the missing proteins in the human proteome**. *Nat Commun* 2017, **8**:14271.

284. Zheng J, Shen W, He DZ, Long KB, Madison LD, Dallos P: **Prestin is the motor protein of cochlear outer hair cells**. *Nature* 2000, **405**(6783):149-155.

285. He DZ, Lovas S, Ai Y, Li Y, Beisel KW: **Prestin at year 14: progress and prospect**. *Hear Res* 2014, **311**:25-35.

286. Uhlen M, Bandrowski A, Carr S, Edwards A, Ellenberg J, Lundberg E, Rimm DL, Rodriguez H, Hiltke T, Snyder M *et al*: **A proposal for validation of antibodies**. *Nat Methods* 2016, **13**(10):823-827.

287. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN: **The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine**. *Hum Genet* 2014, **133**(1):1-9.

288. **Antibodypedia - Explore prestin** [http://www.antibodypedia.com/explore/prestin]

289. **HUPO - Bioinformatics Hub at the HUPO 2016** [https://www.hupo.org/News/4398286]

290. **CompMS/Overview** [https://github.com/CompMS/Overview/wiki/HUPO-2016]

Appendix 1

Publication 7

Pages 181-196 of this thesis have been removed as they contain published material under copyright. Removed contents published as:

Islam M.T., Mohamedali A., Fernandes C.S., Baker M.S., Ranganathan S. (2017) De Novo Peptide Sequencing: Deep Mining of High-Resolution Mass Spectrometry Data. In: Keerthikumar S., Mathivanan S. (eds) *Proteome Bioinformatics. Methods in Molecular Biology*, vol 1549. Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-6740-7_10