TECHNISCHE UNIVERSITÄT MÜNCHEN
Lehrstuhl für Proteomik und Bioanalytik

Multi-omics data integration and data model optimization in ProteomicsDB

Patroklos E. Samaras

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Dmitrij Frishman

Prüfer der Dissertation:
1. Prof. Dr. Bernhard Küster
2. Prof. Dr. Julien Gagneur
3. Priv.-Doz. Dr. Martin Eisenacher

Die Dissertation wurde am 20.04.2020 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 30.09.2020 angenommen.

To Ermina
to my parents Vangelis and Ntina
and to my brothers Thanasis and Tasos

Abstract

Proteomics, transcriptomics, genomics, and phenomics are well-established scientific fields that contribute in their own way to the research of life and disease and in particular, cancer research. However, studying the data and results from only one omics-type will only reveal part of the bigger picture of what is happening inside a cell. Several methods have been developed for multi-omics data integration, however, they are either computationally expensive standalone tools or libraries that require programming expertise for their usage and the incorporation of results into further analyses. This cumulative thesis consists of two original publications that address the need for an online resource that supports scientists and clinicians with real-time multi-omics data integration, data analysis and visualization tools.

The first publication presents the initial status of ProteomicsDB, a quantitative proteomics database, and describes the extension of the platform to support new omics-types. These data required the design of new data models to solve the challenge of storing and connecting information from multiple omics types and data sources using different identifier schemes. As a result, two generic data models were implemented, one for the storage of quantitative omics expression data and one for the storage of cell viability studies. To overcome the identifier mapping challenge and to enable combined data analysis, a third model was implemented allowing inter-connections with the proteomics data already existing in the database. The data model consisted of a composition of so-called triplestores along with tables that store metadata for each identifier. The generality of this data model allowed to capture any kind of relations between identifiers, such as protein-protein interaction networks as well as pathway data.

The second publication extended the data wealth of the platform with additional proteomic, cell viability, drug-target and meltome data. In addition, the available protein properties were expanded with results from a new cellular assay containing protein turnover data. To support the interpretation of the new data types and assays, the user interface of the platform was extended with suitable visualization tools. The vast amount of data stored in ProteomicsDB enabled the development of integrative online tools, like the mRNA-guided missing value imputation method for protein expression data and the machine learning-based prediction of cell viability upon drug treatment based on protein expression data. The presented version of ProteomicsDB also enables users to upload custom expression datasets and analyse them alone or in comparison to stored data, using the analytics toolbox of the platform. Finally, all functionalities of the platform were extended to support data from any organism, transforming ProteomicsDB into a multi-omics and multi-organism resource for life science research.

Zusammenfassung

Proteomik, Transkriptomik, Genomik und Phänomik sind etablierte Forschungsbereiche, die zur Erforschung von Krankheiten und speziell zur Krebsforschung beitragen. Daten und Resultate eines einzigen Omics-Typs offenbaren jedoch nur einen Bruchteil der Vorgänge auf zellulärer Ebene. Um Daten mehrerer Omics-Typen, sogenannte Multiomik-Daten, zu integrieren, wurde eine Vielzahl von Methoden entwickelt. Diese teilen sich auf in rechenintensive Stand-alone Programme und Programmbibliotheken, welche für die Nutzung und weitere Verwertung der Ergebnisse Programmierkenntnisse voraussetzen. Diese kumulative Arbeit besteht aus zwei Veröffentlichungen, die auf den Bedarf von Wissenschaftlern und Klinikern an einer Online-Ressource eingehen, welche Echtzeit-Multiomik-Datenintegration sowie Datenanalyse- und Visualisierungstools unterstützt.

Die erste Publikation präsentiert den initialen Status von ProteomicsDB, einer quantitativen proteomischen Datenbank und beschreibt die Erweiterung der Plattform um neue Omics-Typen. Für die herausfordernde Speicherung und Verknüpfung verschiedener Omics Datentypen aus unterschiedlichen Quellen mit verschiedenen Idenfikationsbezeichnugen wurden neue Datenmodelle nötig. Als Lösung wurden zwei Datenmodelle implementiert, eines für die Speicherung quantitativer Omics-Expressionsdaten und eines für die Hinterlegung von Zellviabilitätsassays. Als Lösung für die Problematik des Abgleichs von Idenfikationsbezeichnugen wurde ein drittes Datenmodell implementiert, welches nun deren kombinierte Analyse und die Verknüpfung zu den bereits gespeicherten proteomischen Daten ermöglicht. Die verwendete Modellierung bestand aus der Kombination von sogenannten “Triplestores” und Tabellen, welche Metadaten für jeden Eintrag enthalten. Die resultierende allgemeine Anwendbarkeit erlaubt die Speicherung von Beziehungen jeglicher Art zwischen den Einträgen, wie zum Beispiel Protein-Protein-Interaktionsnetzwerke oder die Abbildung von Signalwegen.

Die zweite Publikation erweiterte die Datenmenge der Datenbank um zusätzliche proteomische Daten, Zellviabilität-, Protein-Wirkstoffinteraktions- und Meltomedaten. Die gespeicherten Proteincharakteristiken wurden um die Ergebnisse eines neuen zellulären Proteinumsatzassays erweitert. Weiterhin wurden die Visualisierungsoptionen der Plattform um zusätzliche Darstellungen für die neuen Datentypen und Assays erweitert. Die riesige in ProteomicsDB gespeicherte Datenmenge ermöglichte die Implementierung von omic-übergreifenden Onlinetools, wie eine mRNA-basierte Methode zur Imputation fehlender Proteinexpressionswerte und die von maschinellem Lernen gestützte Vorhersage der Zellviabilität nach Medikamentenbehandlung, die basierend auf Expressionsdaten von Proteinen realisiert wurde. Die beschriebene Version von ProteomicsDB ermöglicht es außerdem den Nutzern eigene Expressionsdatensätze hochzuladen und diese allein oder im Vergleich mit den hinterlegten Daten mittels der bereitgestellten Werkzeuge zu analysieren. Desweiteren wurden alle Funktionalitäten der Plattform erweitert um Daten von anderen Organismen zu unterstützen, was ProteomicsDB in eine omics- und organismus übergreifende Plattform für die Biowissenschaften transformiert.

Table of contents

Chapter 1
General Introduction

Contents

1 From genes to proteins to disease

Living organisms, prokaryotes and eukaryotes, unicellular and multicellular, are strongly dependent on the accurate and continuous flow of information that allows proper cellular function, including cell growth, replication and death. Following the central dogma of biology (1), information is encoded and decoded at different levels, starting from the "blueprint" of a cell, the genomic information that is encoded in the deoxyribonucleic acid (DNA). Information here is compressed, as a small number of genes in any organism will produce, via the transcription process, a larger number of ribonucleic acid (RNA) molecules. The human genome, for example, consisting of around 20,000 genes (2), can produce 70,000 to 100,000 RNA molecules, because of alternative splicing in a wide range of abundance values (3). This number increases, even more, when sequence mutations occur, resulting in a larger number of translated proteins. This number is not the final, though, as the number of proteoforms (4), is larger (5). Proteins and proteoforms can further undergo post-translation modifications. These modifications can happen on side chains of amino acids and play an essential role in many cellular processes, such as signalling or regulatory processes, as well as in protein homeostasis since most post-translational modification are reversible. Proteins are expressed in different subcellular organelles, in different abundances, and exhibit different degradation rates (6,7). Moreover, proteins do not act alone; they interact with each other and form complexes. Proteins and protein-complexes interact with each other as well as with RNA molecules and the surrounding environment in such a way that they ensure the normal function of the cell. The health of a cell and, subsequently, of the full organism, is highly dependent on this fine-tuned environment. Any alteration or blocking in these interactions or change in a protein's abundance can affect the normal cycle of a cell, resulting in different phenotypes (8). The state, where the regular flow of information in a cell is disturbed, causing changes in protein expression or existence is defined as a disease (9).

A disease can be caused at any level of the central biological dogma (1). A single nucleotide substitution, addition or deletion, also called a mutation in the genome sequence can have different effects. For example, it can alter a start codon on the 5' UTR of a gene, with the effect of never transcribing this gene. Alternatively, it can introduce a stop codon much earlier, altering the length of the resulting transcript and protein. It can also introduce new codons altering the normal sequence of a protein and changing its function. As proteins are directly associated with changes in the phenotype, it renders the study of the proteome of high importance with the aim to identify and cure diseases or at least try to slow the process. Current technological advances enable the full proteome acquisition from a sample. Several studies have been published exploring the full proteome of organisms to identify which genes are expressed and in which tissues and quantities (10,11). All these studies, though, suffer from the fact that not all proteins are yet detectable. Either because they measure peptides and try to infer protein existence and abundance, introducing uncertainty during the procedure, or because they are of really low abundance or length, making them harder to find. The existence of miRNA molecules in a tissue can also have an effect, as they might silence genes and consequently suppress the expression of a protein (12,13). Thus, by studying the proteome alone might provide an incomplete picture of what is happening inside a cell, tissue, or organism in the state of a disease. Proteomics can benefit from transcriptomic and genomic advances, technologies and studies to complement the space of missing protein expression. Phenomics can also contribute to the above cooperation, as changes on any level, genome, transcriptome and proteome can be correlated and associated with changes

in the phenotype of the same tissue-sample, upon exposure on the same conditions. However, each omics field has a different starting point in the research history, placing them at a different level of advancement. Every technique, though, has its benefits and should be considered equally in order to achieve a better understanding of the life inside a cell, a tissue, or even an organism. To be able to integrate data from different omics-types, there is the need to understand how each omics field is functioning and the technologies that are used.

2 Multi-omics technologies

In the past two decades, various omics technologies have contributed to detecting genes or biomarkers, with genomics and particularly transcriptomics leading the field so far. Biomarker research can further benefit by including information provided by other omics fields, such as proteomics with the study of deep proteomes, or phenomics in collaboration with drug-target interaction information. Below follows a description of the technologies that are covered in this thesis and the way that information gets generated.

2.1 Transcriptomics

Although genomics is an interesting field as it focuses on the structure evolution and function of genomes, it provides no information on gene expression. As described earlier, a genome contains all possible genes that exist in a cell; not all of them, though, are expressed, and if so, they are not of the same abundance. Stretches of the DNA that contain genes and encoded information are transcribed into RNA during transcription. The field that studies the function and abundance of RNA molecules is called transcriptomics and provides the first sight of gene expression. Two of the most prominent transcriptome profiling technologies are microarrays (14-17) and RNAseq (18).

2.1.1 MicroArray

A typical microarray experiment (19) is illustrated in Figure 1.1. The first step is the extraction of the total RNA from a biological sample. The mRNA is then poly-A enriched (20,21) and reverse transcribed to complementary DNA (cDNA) using fluorescently labelled nucleotides (22). Afterwards, the labelled mRNA is hybridized to a microarray. A microarray consists of probes, which are collections of oligonucleotides, complementary to the cDNA sequences that are targeted. These probes are then attached to a predefined grid. During hybridization, each probe will bind preferentially to perfectly complementary cDNA molecules. That will cause different amounts of fluorescence across regions, which can then be read with a laser scanner. The intensity of the fluorescence and the position in the grid are then associated with the cDNA sequences and their quantity in the measured samples.

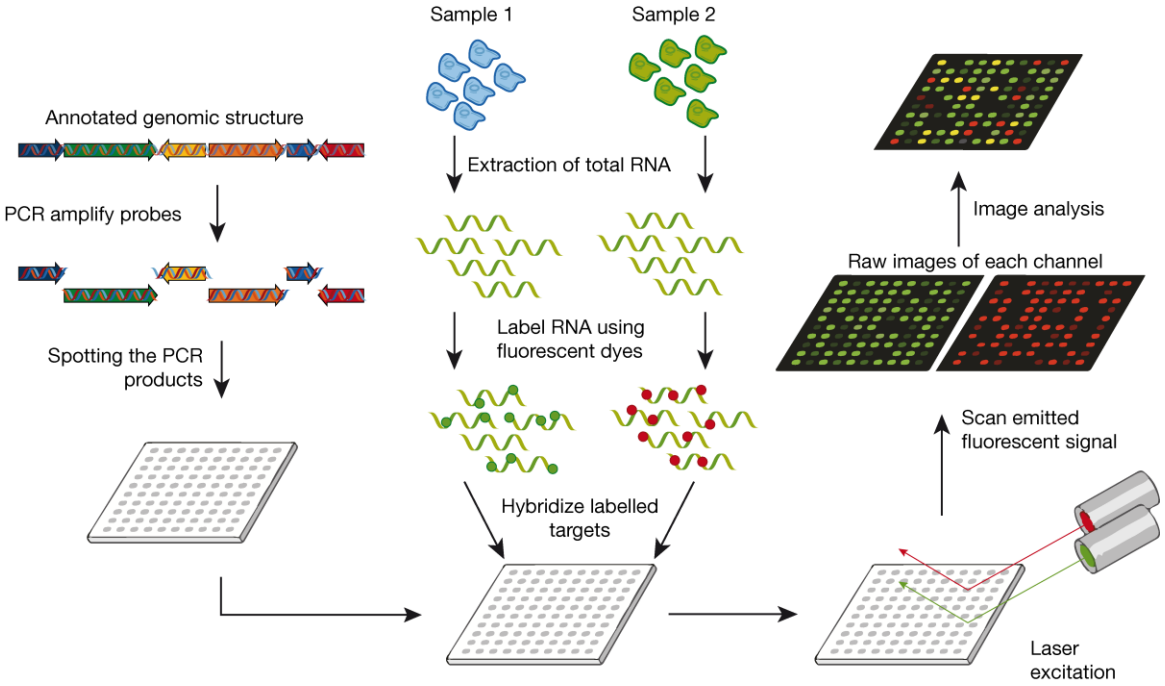

Figure 1.1 A typical MicroArray experimental workflow. Acquired and adjusted from (19).

Because of the technology and the materials that are used for the quantification of microarray data, several types of noise are introduced. This requires further data curation before any analysis can start. Data preprocessing includes but is not limited to background noise filtering, cross-hybridization adjustment in the case of non-specific binding of cDNA to the probes, as well as removal of unwanted variance, also known as batch effect (23), within or across microarray experiments. Several normalization methods have been proposed, such as the Robust Multiarray Average (RMA) normalization, which assumes that most of the transcripts do not show differential expression between conditions and normalizes all microarrays simultaneously. A modification of this method, GCRMA (24), also uses position-specific effects, as it is observed that the nucleotide sequence plays an essential role in the binding affinity of the probes to the targets. MicroArray technology offers great sensitivity and good throughput. However, it is limited to the amount of analytes on the chip. In short, MicroArray chips provide information about the quantity of a specific set of transcripts in a sample, with no sequence information, however.

2.1.2 RNAseq

The RNAseq technology compared to MicroArray chips is unbiased, as it does not require a preselection of probes. Besides, RNAseq offers not only quantification of the existing transcripts in a sample but also sequencing. Similarly to a microarray experiment, in an RNAseq experiment, the first step is the isolation of the total RNA from the biological sample and the reverse transcription to cDNA, excluding the labelling step. The cDNA is then fragmented into short nucleotide sequences. The generation of these short nucleotide sequences can also be achieved by mRNA fragmentation, followed by reverse transcription into cDNA. Both 3' and 5' ends of these fragments are ligated to short DNA adaptors, which contain functional elements for the sequencing, e.g., the primary sequencing site. The obtained cDNA library is then analyzed by an NGS platform, such as the widely used Illumina (see review (25)). The outcome of the sequencing is millions of short reads that correspond to either one or both ends of the fragments. The length of these sequences varies from tens to hundreds, depending on the technology that was used (26). The temporal and financial cost of the sequencing depends on whether there was used single-read or paired-end sequencing methods, with the second being the more expensive and time-consuming. The reads, or short sequences, are then mapped to the reference genome of the organism, and in particular, the reads are assigned to the exons that they correspond to.

The standards for quantification of the transcripts have been modified during the years, starting with RPKM and FPKM values, to RSEM, and finally, TPM values. RPKM stands for Reads Per Kilobase of transcript, per Million mapped reads and scales by transcript length taking care of the fact that longer RNA molecules generate more sequencing reads in most of the RNAseq protocols. It is calculated, as shown in Equation 1.1.

$$RPKM = \frac{numReads \times 10^3 \times 10^6}{totalNumReads \times geneLength} \quad (1.1)$$

In the case of paired-end RNAseq data, although a sequenced molecule comes from a single cDNA fragment, it can generate two reads. FPKM stands for Fragments Per Kilobase of transcript, per Million mapped reads, and uses the same formula for the calculation by replacing the number of reads with the number of fragments. RPKM and FPKM values have the disadvantage that they will not always sum up to one million, meaning that a transcript's abundance level might be affected by the average transcript's length in the measured sample. This is not the case for Transcripts Per

Million (TPM), where the sum of all transcripts should always add up to one million (Equation 1.2). There are many cases though that a single read can map to more than one gene or transcript. For these uncertain mappings, a method was proposed by (27), called RNAseq by expectation maximization (RSEM). This method uses a statistical model that derives from the sequencing process, allowing that way modelling of non-uniform read distributions.
The transcriptomics data that were used in this thesis are normalized using the TPM quantification method.

$$TPM = A \times \frac{1}{\sum A} \times 10^6, \qquad A = \frac{totalNumReads \times 10^3}{geneLength} \quad (1.2)$$

2.1.3 A reason to move beyond transcriptomics

Transcriptomics is a high-throughput and robust technology. Both methods described above provide a dense expression matrix with only a few or no missing values. However, in all cases, the actual result of these experiments is the quantification of the transcribed genes in a sample. There is no information about whether this transcript is then translated to a protein and to which abundance. Transcript existence and abundance information is still important and has many applications in diagnosis or patient classification, for example, in oncology (28). In order to find what is happening in a cell, it is necessary to dig deeper and study the existence, abundance levels, and function of proteins.

2.2 Proteomics

The entire set of proteins and proteoforms that are expressed or modified by an organism constitute the proteome. Proteomics is the field that studies the proteome, its function, and alterations under the effect of drug treatment or biological phenomena like disease, e.g., oncogenesis (29). In comparison with other omics-types, as described by (30), “the proteome is the expressed protein complement of a genome and proteomics is functional genomics at the protein level”. The accurate identification and quantification of the proteins that are expressed in a cell or tissue or, in general, in a sample across different conditions, enables the protein expression pattern recognition. By studying those patterns, modules can be identified (31) and related to phenotypes or disease states. Genomics and transcriptomics are two powerful fields with well-established technologies. In both these fields, molecule amplification is an advantage for the later identification and quantification of genes or transcripts. This step is not possible in proteomics, though, as no technique exists that would amplify proteins prior to their detection. This can cause issues, as it gets harder to detect the low abundant proteins in a sample. Various methods have been established to address the issue with the sensitivity required for the detection of such proteins, such as antibody-based affinity approaches (32). Each antibody is specific to a target. That makes antibodies a valuable tool for tracing, identifying, and quantifying the expression of certain proteins in a sample. Some of the most common assays are Western Blot (33,34) and Enzyme-Linked Immunosorbent Assays (ELISA) (35,36). Despite the sensitivity of the antibody-based assays, antibodies are cross-reactive with non-target proteins, which limits the number of proteins that antibodies are highly specific. Mass spectrometry-based approaches (37,38), are capable of identifying and confidently quantifying proteins in their whole expression dynamic range. This enables not only the detection of low abundant proteins but also full protein expression profiles comparison across multiple samples and experiments, with the ultimate goal

to study protein expression behaviour and find differences among healthy and/or disease biological samples.

2.2.1 Mass spectrometry-based proteomics

There are two most dominant approaches when it comes to mass spectrometry-based proteomics. The "top-down" approach, where intact proteins are measured and the "bottom-up" approach, also referred to as shotgun proteomics, where proteins are digested into peptides using a proteolytic enzyme (e.g., Trypsin), which are then measured. Proteins are quite diverse regarding their biochemical properties, which makes the preparation of the samples as well as the data acquisition in the top-down approach challenging. In bottom-up proteomics, this complexity gets reduced as the peptides, can be further separated using online or offline liquid chromatography (LC). Here, peptides with different physical or chemical properties, such as hydrophobicity, are bound to reverse phase material and eluted at varying concentrations of an organic solvent, commonly referred to as retention time. As a previous step, peptides that bare specific modifications, for example, phosphorylated peptides, can be enriched by first passing the mixture through a titanium dioxide capillary column, also referred to as IMAC enrichment (39,40). This thesis will focus on the bottom-up approach, as all data that was used or processed in the scope of this thesis were produced by this approach.

2.2.2 Mass spectrometry

A typical mass spectrometer (MS) consists of the following three parts: the ion source or ionizer, the mass analyzer and the ion detector. The most commonly used ionization method is electrospray ionization (ESI). A fine needle, also called the emitter, is coupled online with the LC, and a high voltage is applied between the needle and the MS, which allows the peptides to be ionized at atmospheric pressure, and the ions are transferred into the high vacuum of the MS. Positive ESI spray creates peptide ions that are mostly of charge two and above. Introduction of LC additives, for example, DMSO (41), can enhance the peptide ionization. The ions in the high vacuum of the MS can be manipulated using electrical fields and separated by their mass to charge ratio (*m/z*) by the mass analyzer. Several mass analyzers exist and are commonly used.

Ion traps consist of four parallel rods-electrodes, where direct current (DC) and alternating current (AC) is applied on opposing rods. This mechanism is used to confine ions in space while they trapped in a spiral-like secular motion. Ions are sequentially ejected from the ion trap and hit a detector that records the induced current.

Quadrupoles (Q) have the same structure with ion traps, but, in contrast to the latter, a radio frequency voltage (RF) with a DC offset is applied between opposing rod pairs. That causes the ions with a specific *m/z* ratio to travel through the mass analyzer, while the rest of the ions have unstable trajectories and collide with the rods. They are usually employed as mass filters to select only one type of ion. Compared to the Ion traps, quadrupoles cannot store ions.

Time of flight (TOF) mass analyzers make use of an electric field that accelerates ions in vacuum given the same kinetic energy. Lighter ions, having a higher velocity reach the detector earlier compared to heavy ions. Therefore the *m/z* of an ion is calculated from the time it needs to reach the detector. A longer drift distance will cause better separation of the m/z ratios, leading to more precise measurements.

Fourier transform (FT) analyzers monitor the motion of the ions in a magnetic field. The ions, after excitation, orbit at their cyclotron frequency forming clusters. To determine the *m/z* of the oscillating ions, the induced image current gets recorded on two electrodes, which in turn get

Fourier transformed. Increasing the transient time, the resolution of the FT mass analyzer increases as the frequencies can be measured with high accuracy.

Orbitraps belong to the family of FT MS. They consist of 2 homocentric electrodes. Between the outer and the inner electrodes, an electric field is applied. The injected ions start to orbit steadily in elliptical trajectories around the inner electrode, while their centrifugal force balances their distance from it. This causes ions with lower *m/z* to orbit closer to the inner electrode. At the same time, ions show an axial oscillating movement across the inner electrode. The frequency of this movement is recorded and used to calculate the *m/z* ratio of the trapped ions using Fourier transformation. Orbitrap mass analyzers are nowadays the most commonly used mass spectrometers due to their high resolution and superior mass accuracy.

2.2.3 Tandem mass spectrometry

Up to this point, all mass analyzers detect and measure the *m/z* ratio of the injected peptides, which corresponds to an MS scan, also known as MS1 scan. Measuring the mass of a single ion in MS1 cannot reveal the exact peptide sequence. In order to derive sequence information, further fragmentation of a peptide into shorter parts of the initial peptide sequence is needed (42). This process is called an MS/MS scan or MS2 scan. In an MS2 scan, there are two consecutive MS stages. The MS1 scan is used to record the mass of all peptides in a certain retention time point. The mass spectrometer selects and isolates a single ion population, also called the precursor, and proceeds into further fragmentation of the selected ion population using a fragmentation method. The most commonly used fragmentation techniques are the electron-transfer dissociation (ETD) (43), the higher energy collision-induced dissociation (HCD) (44) and the collision-induced dissociation (CID) (45). The fragment ions are then recorded in an MS2 scan. In the resulted spectra, sequence information can be read as the delta (*m/z* distance) between two fragment ions.

2.2.4 Peptide identification

Although the field is named proteomics and the ultimate goal is to identify and quantify proteins, in shotgun proteomics, peptides are the ones that get measured. The spectra that are collected from the mass spectrometers are assigned to peptides, and different statistical models and strategies are applied to ensure the quality of matching a spectrum to a peptide sequence or commonly called a Peptide Spectrum Match (PSM). Afterwards, the identified peptides get assigned to their proteins of origin following several rules, which will be described later.

Database search and control for false discovery rate

Two techniques can be applied for the identification of a spectrum. *De novo* sequencing is one of them, where the observer either performs manual spectrum interpretation, calculating the distance between abundant or obvious peaks in the spectrum and assigning them to the mass of amino acids or uses appropriate software and libraries (46). This approach is sensitive to errors and becomes more complicated in the cases that one or more modifications are present in the peptide. Even solving this issue though, manual annotation of a single spectrum is time-consuming and modern mass spectrometers produce thousands of spectra per hour of measurement, rendering manual spectra interpretation not optimal for large scale studies. Another approach, called the Database search, emulates the sample preparation in the wet-lab, by applying *in-silico* digestion using the cleavage pattern of the used protease. Having the in-silico peptide sequences at hand, all possible masses can be calculated that can occur by fragmenting this peptide along its

backbone. This will generate all possible theoretical spectra of the peptides that correspond to the target organism. A theoretical spectrum consists of peaks of unknown intensity and masses that were calculated with the procedure described above (Figure 1.2). For every acquired spectrum, this database of known theoretical spectra can be filtered by mass, as this is recorded and reported back from the mass spectrometer along with the spectrum, limiting the number of candidate peptides. A comparison of the acquired spectrum to the theoretical one and implementation of similarity scores will result in a PSM, where the best matching candidate has the highest score. This approach is much faster as the database needs to be created only once at the beginning of the identification process, and by utilizing computation power, thousands of spectra can be annotated in a very short amount of time.

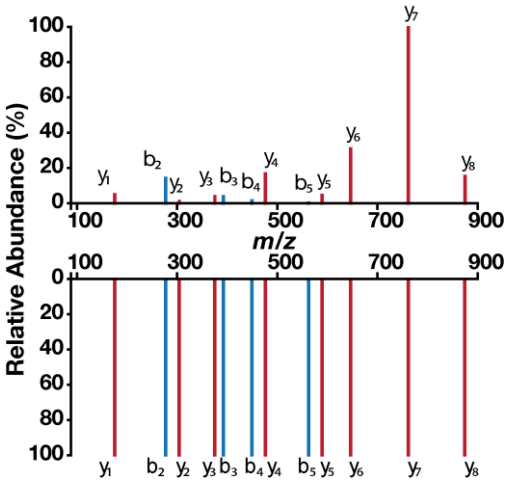

Figure 1.2 A mirrored spectrum. The top spectrum is an experimental spectrum of the peptide sequence YLDGLTAER. The bottom spectrum is the theoretical spectrum of the same peptide.

However, at this point, there are two types of errors that can occur: false negatives, which translates to a spectrum not being identified although it originates from a peptide in our sample, and false positives, which give us the wrong information of a peptide existing in our sample although it does not. The false positives or Type I errors should be avoided as they might lead to false hypotheses.

The automated matching procedure solves the issue with time. However, it offers no control over the number of false assignments or, in particular, control over the false positives. For that reason, the target-decoy approach was proposed to offer an estimation of the False Discovery Rate (FDR) in a set of PSMs. The identification procedure stays mostly the same. The database with the theoretical spectra that was created earlier is now referred to as the target space. Shuffling or reversing the peptide sequences of the target space, gives us new false theoretical spectra which constitute the decoy space. The decoy spectra share the same peptide properties (e.g., precursor mass) with the target spectra. By combining the two databases (target and decoy) and continuing with the identification process, a spectrum generating a Type I error has 50:50 chance to match to a target or a decoy theoretical spectrum. Dividing the number of decoy hits after the database search with the number of target hits (Equation 1.3) gives us the false discovery rate.

$$FDR = \frac{FP}{TP + FP} \approx \frac{decoy\ hits}{target\ hits} \quad (1.3)$$

To control for FDR and find the lowest identification score that results in the desired FDR, PSMs by descending score and the FDR is calculated in each subset from the highest score to the current one. PSMs with a score higher or equal to the lowest score that corresponds to the selected q-value cut-off are accepted.

The theoretical spectrum approach lacks information about the intensity of the fragment ions in a spectrum. That alone might lead to false matches, increasing the false discovery rate, as the same theoretical spectrum might match more than one peptide sequences. This can be solved by

substituting the theoretical spectra with experimentally acquired ones originating, for example, from synthetic peptides (47,48). However, creating custom libraries of experimental spectra for organisms where synthetic datasets do not exist yet, is time-consuming and expensive. Using these spectra from projects like ProteomeTools (47,48) as a gold standard, allowed the training of a deep neural network Prosit (49), capable of predicting peptide fragmentation accurately. As of now, Prosit can predict only unmodified peptide sequences, but there are efforts to extend the tool to all known modifications. Such a tool allows the full replacement of the theoretical spectrum databases with predicted spectra that include both mass and intensity information.

2.2.5 Protein inference

A protein consists of multiple peptides, but at the same time, a given peptide might be originating from multiple proteins. As in bottom-up proteomics peptides get measured and identified, it is not that straightforward to assign peptides back to their proteins of origin. Several rules and algorithms exist for protein inference. The Occam's razor rule reports the minimum set of proteins that can explain all of the observed peptides. The anti-Occam's razor reports all proteins that contain the measured peptides. Another way is to consider only the peptides that belong exclusively to one protein. The last method is the one followed for protein inference by ProteomicsDB and thus to the data that was used in this thesis.

2.2.6 Quantification

As with protein inference, protein quantification is inferred from their peptide quantification. Two approaches exist for quantification. In the first approach, label-based quantification, the peptides are labelled prior to their measurement, which introduces precise mass shifts, recognizable by the mass spectrometer. That allows measurement at the same time of multiple conditions and separate quantification in the same run (Figure 1.3). The second approach is called label free quantification and compares the results of two or more samples or conditions from separate runs. Several methods can be applied to extract the final abundance of a peptide from its acquired spectra. Spectral counting (50,51) is based on the fact that in data dependent acquisition, the most abundant peptide precursors are selected for further fragmentation and therefore trigger an MS/MS event. High abundant peptides produce stronger signal intensities and thus are selected more frequently. In intensity-based quantification(52), peptide abundance is calculated as the area of the extracted ion chromatogram (XIC) of the precursor ion over its elution profile. In that case, both the signal of the MS event and the one recorded from the MS/MS event can be used for quantification, rendering this method independent of any labelling technique.

After calculating each peptide's abundance, the estimation of protein abundance takes place. Two most commonly used methods for protein level summarization are the iBAQ (53) and the top3 (54) methods. In iBAQ, the sum of all peptides intensities that correspond to a protein is normalized to the length of that protein. Top3 uses the sum of the intensities of the three most abundant peptides of a protein.

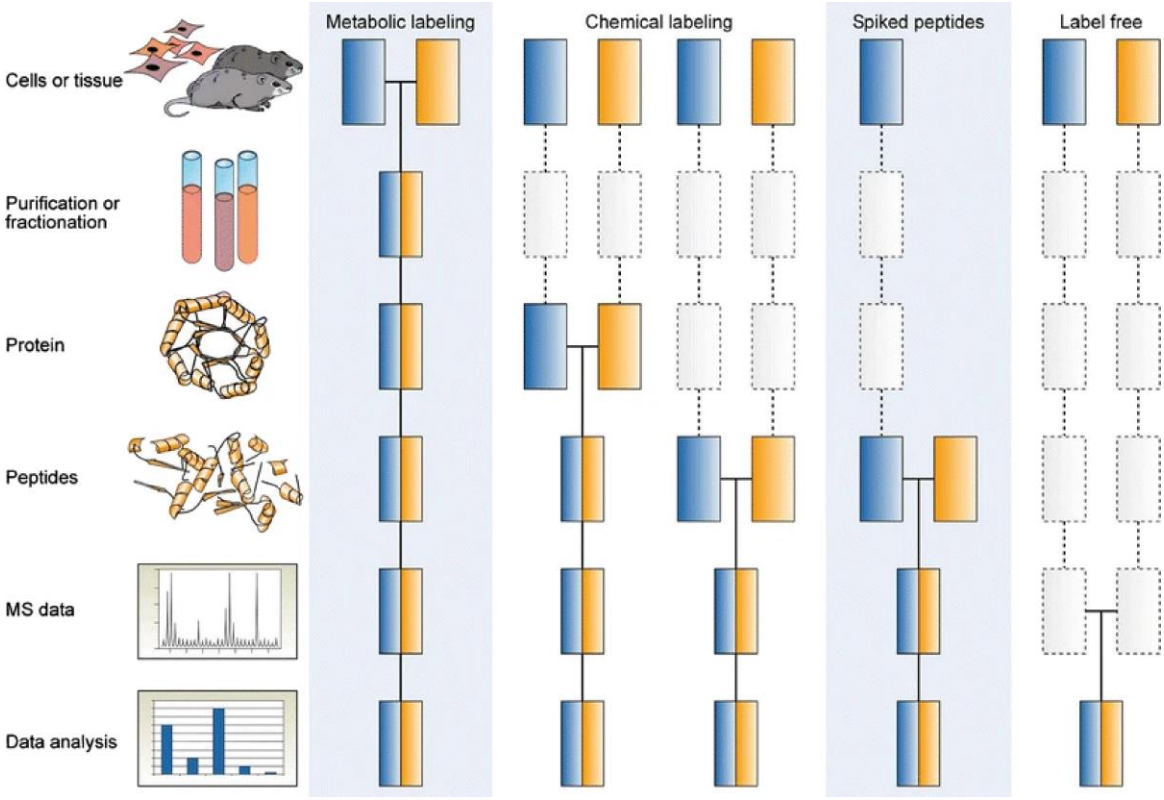

Figure 1.3 Quantitative MS workflows. Rectangles in yellow and blue represent different conditions. Horizontal lines represent the point that two samples are combined. Dashed lines indicate points at which experimental variation and thus quantification errors can occur. Figure from (55)

2.3 Phenomics

A rather interesting omics-type is phenomics, a field that is dedicated to the systematic study of phenotypes of an organism or tissue or cell line under several conditions. Recording the changes of a cell line's phenotype after exposing it to different drugs or conditions and associating these changes to changes at the cell line's proteome, can lead to the discovery of biomarkers, sets of genes or proteins that are highly associated with disease. The most dominant route to study phenotypic changes are cellular assays and in particular, cell viability assays.

2.3.1 Cell line viability assays

The most common cellular assays in the field are the cell line viability assays or also called dose-response experiments. Here different drugs, alone or in combinations, in different dosages or dosage ratios (dose) are applied on samples of the same cell line to track the effect of the selected treatment on the viability (response) of the cell line. The readout of such assays is the cell survival or death in a sample. Cell viability assays can be grouped into four main classes.

Dye exclusion assays are based on the fact that dead cells do not exclude dyes, in contrast to viable ones that do exclude them. Several dyes have been developed for this type of assays, including trypan blue, erythrosine B, eosin and Congo red with trypan blue being the most commonly used (56). Staining of the cells is a straightforward procedure; however, the full experimental procedure of a large number of samples is time-consuming (57). In this type of assays, cell death might be underestimated as there are two essential factors that can give misleading results: i) cells might undergo an early disintegration, causing them not to have been

dyed by the end of the cell culture period and ii) any surviving cells might continue to grow and proliferate during the assay. Regardless, dye exclusion assays are simple in execution, rapid and require a small number of cells.

In **colorimetric assays,** the reagents that are used, produce a colour as a response of cell viability, which allows the response measurement by a spectrophotometer. They are cheap, easy to perform and are available in commercial kits. Examples of this type of cell viability assays are the MTT assay (58), MTS assay (59,60) and XTT assay (61).

Fluorometric assays utilize fluorometers or flow cytometers to monitor the viability of the cells. They are offered as commercial kits by several companies as well as kit packages for full experimental procedures. The most common fluorometric assay is the alamarBlue (AB) assay (62). Here, the blue non-fluorescent dye resazurin enters the cells and gets reduced to resorufin with the help of different enzymes such as diaphorases (63). Contradictory, resorufin is red and highly fluorescent. Viable cells convert resazurin to resorufin constantly, increasing the total fluorescence of the medium. The ratio of the viable cells can then be measured with the use of a fluorometer.

In **luminometric assays**, the addition of the reagents produces a stable glow. This glow can be used as a signal and measured by a luminometric microplate reader (64). The assay can be performed on a 96-well plate or 384-well microtiter plate, while both cell viability and death can be recorded from the same well (65). A well-established approach is the adenosine triphosphate (ATP) assay. ATP plays a considerable role in many functions of a cell, for example, in cell signalling, biological synthesis and movement processes. Lethally damaged cells that lose their membrane integrity are not able to synthesize ATP, causing its level to drop. The enzyme luciferase can be used to detect the ATP levels of a cell in real-time through biotinylation. Biotinylation immobilizes luciferase on the cell membrane enabling the detection of the real-time release of ATP as ATP yields a luminescent signal.

2.3.2 Dose-response curves and feature extraction

In all the aforementioned assays, a single value is recorded as the viability or response of the cells after treatment with a compound at a specific dose. By repeating the experiment using different treatment doses, the results can be visualized as a dose-response curve (Figure 1.4). It is crucial at this point to include a 0 concentration treatment as the control treatment that should picture the viability of the cells under "normal" conditions. The rest of the dose-points are then normalized and compared to the control dose-point. There are several linear or sigmoidal models that can be fitted to the dose-response data, with four-parameter log-logistic regression being the most common. (Equation 1.4). The resulting model follows a sigmoidal curve indicating the overall change on the viability of the cells. In Equation 1.4 *b* is the slope of the curve, *c* is the lowest viability measurement or lower limit, *d* is the highest viability measurement, or upper bound and

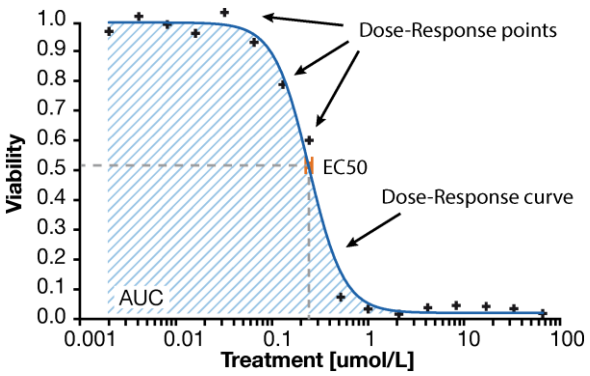

Figure 1.4 Example of a 16-dose-response curve and model attributes that can be extracted from the fitted model.

e is the effective concentration at which the viability of the sample is 50% of the maximum (control) viability.

$$f(x) = c + \frac{d - c}{1 + e^{b*log(x) - log(e)}} \quad (1.4)$$

Having the model at hand, further features can be calculated that will help to define if a cell is sensitive to a drug or not. R^2 is a value that can be exported from the model and depicts how close are the actual data points or, in this case, dose-response measurements to the fit. A value close to 1 shows a perfect fit, so further results are trustworthy, while a value close to 0 depicts a bad fit and suggests that further results should not be trusted. The area under the curve (AUC) is the integral of the resulted curve. The closer the AUC is to 1, the less effective the drug is against this cell line. There are cases that the AUC can be higher than 1, for example, when the drug has no effect on the cell line, and the cells keep proliferating. Another feature is the Relative inhibition effect (RE), which is the relative difference of the viability of the last dose to the viability of the control. A negative RE or close to 0 shows no sensitivity of the cell line to the drug, while a positive RE with a value close to 1 shows high inhibition of the cell line. Each of these attributes alone can lead to a false interpretation of the experiment's results. A typical case is an RE with a value of 1, showing high inhibition, although the EC50 value of that drug might be extremely high, causing the drug to be toxic for any cell. For that reason, they should be used in combinations, while separate thresholds can be applied to each one of them.

2.4 Multi-omics data integration

As shown above, each omics technology comes with its advantages and disadvantages. Each one alone reveals an incomplete picture of what is happening inside a cell in the state of a disease. Integrating information from many different omics technologies can help in this cause. In the last years, different multi-omics data integration tools have been developed either as standalone applications or as programming language-specific packages and libraries, e.g. R packages.

In general, the developed tools can be grouped into three main categories based on their applications; biomarker prediction, disease classification and subtyping and tools that provide insights into the biology behind a disease.

An example of a biomarker prediction tool based on multi-omics data integration is MOFA (66). MOFA is a computational method for the discovery of principal sources of variation in multi-omics datasets. It infers hidden factors that capture technical and biological sources of variability using a Bayesian framework. It was validated using multi-omics data from 200 samples with chronic lymphocytic leukaemia (CLL), including gene expression profiling, DNA methylation and drug response data using 63 drugs.

A multi-omics data integration tool that is used for disease subtyping is moCluster (67), which identifies patterns across multi-omics datasets. Using sparse consensus Principal Component Analysis (PCA), it identifies latent variables that are then clustered using traditional clustering methods, such as K-means. It was used in the analysis of 83 colorectal cancer cell lines across gene and protein expression data as well as DNA methylation data and identified four integrative subtypes, two of them not having been discovered in studies previous to the moCluster publication.

Multiple co-inertia analysis (MCIA) (68), is another integration approach that contributes to disease subtyping while it also provides insight into the disease biology. This approach normalizes

the given set of features (e.g. genes and proteins) using a covariance optimization criterion and projects the different datasets into the same space. Visual observation of the sample space using different axes of MCIA can reveal disease subtypes. In the tested case, the first MCIA axis separated the samples to proliferative and immunoreactive, while the second axis separated the samples with differentiated subtype from the ones with mesenchymal subtype (68). Taking a look at the feature space of the same datasets and applying ingenuity pathway analysis (IPA) on cell line-specific features resulted in revealing significantly enriched canonical pathways that were relevant to the cell line, such as melanoma development signalling pathway in melanoma cells. Therefore, the sample space of MCIA can help in disease subtyping, while the feature space in deriving insights into the disease biology.

Multi-omics data integration can also have applications on the exploration of other omics-types. Transcriptomics expression data can be used for missing value imputation in proteomics (69), replacing standard imputation methods like the replacement of missing values using either a constant value or by random sampling from a distribution close to the detection limit. However, all these tools and methods exist as standalone packages or are described in specific studies, not being offered in any platform for online usage across public datasets, requiring manual data manipulation and programming skills for their usage.

3 Bioinformatics

Modern sequencing techniques, together with high-throughput technologies, increase the amount of produced data by gigabytes (GB) or terabytes (TB) on a daily basis. Right from the beginning, there was a need for public databases for storage and access to the produced and published biological data. The exponential increase in the amount of data that are uploaded in public databases (Figure 1.5) makes it clear that there has to be a robust, well-defined infrastructure to handle the data growth that is foreseeable approaching. This data amount and variety transform biological databases into gold mines for data mining and knowledge discovery applications.

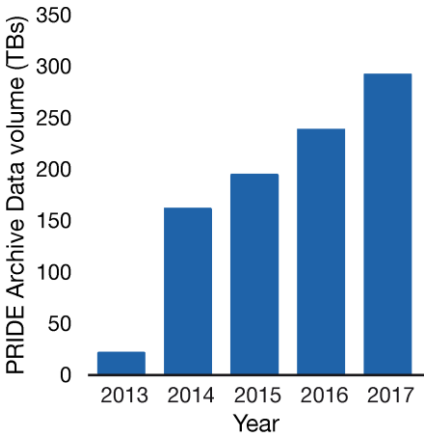

Figure 1.5 Histogram showing the annual growth of data in PRIDE Archive. Figure adjusted from (70)

Biological databases can be categorized either by the origin of the stored data or by the type of data they store. Using the first categorization approach, biological databases can be divided into:

- **Primary/archival** databases, that store raw experimental data, such as GenBank (70), EMBL-bank (71) and PRIDE (72), and
- **Secondary/curated** databases, which rely on data stored in primary databases but store extra information and metadata like sequence motifs, gene- transcript- and protein-sequences, gene locations in chromosomes, protein-protein interactions even evolutional relations between species that are not always experimentally supported. Such examples are UniProt (73) and PROSITE (74).

Based on the second way of categorization, databases can be divided as follows:

- Nucleotide sequence databases, such as NCBI, and EBI,
- Protein sequence databases, such as UniProt and PROSITE,
- Genomic databases, such as Ensembl as well as databases that are organism specific, e.g. FlyBase (75) and MGI Mouse Genome (76),
- Structural databases, storing protein structures and structure predictions like PDB (77), SCOP (Structural Classification of Proteins) (78) and CATH (Protein Structure Classification) (79)
- RNA databases like Rfam (80), mirBase (81), and
- Microarray databases, such as ArrayExpress (82).

This chapter constitutes a review of the most popular and relevant to the study databases and data repositories.

3.1 Public repositories (archival databases)

Over the last decade, data sharing in the scientific community got more and more attention. Starting with genomic and transcriptomic studies, journals and scientific publishers in general, oblige submitting authors to provide along with their study, the corresponding raw and/or processed datasets. The same applies nowadays for the proteomics data and is expanded over the years to every other omics-type that exists. The scope of this enforcement is not only for the scientific community to be able to reproduce the published results but to enable the data

reusability and repurposing. Some of the most popular omics data repositories are described in this chapter.

3.1.1 Gene Expression Omnibus – GEO

Although the name suggests that these repository stores genomics data, GEO stores and shares the biggest collection of genomic and transcriptomic numerical data. Their database organizes the data in several layers, some of them stored as provided by the users and some of them after reprocessing from data curators.

Data submitters upload a text file describing the sequencing platform that was used to produce the data and in case of array-based platforms, a data table describing the array template. This file is stored as a GPL record under a unique identifier. As a next step, submitters upload files describing the biological samples, the protocols followed, and the conditions applied to them, as well as their abundance measurements. Each sample gets a unique identifier and is stored as a GSM entry. The last file that submitters have to upload is a record that summarizes an experiment. This record is stored and gets a unique GSE identifier. Data curators at the GEO process this entry and reassemble it into GEO dataset records (GDS). Experiments or samples that belong to the same GDS are biologically and statistically comparable. Both GDS and GSE entries are query-able through the GEO website. However, only GDS entries constitute the basis of the GEO's visualization and analysis tools. It is worth to mention at this point that not all uploaded data are suitable to assemble a GDS entry, meaning that not all data are used in the visualization and analysis tools.

Although, as described, a considerable amount of metadata accompanies the raw or processed uploaded data, there is no controlled vocabulary that should be followed and not all requested information is mandatory to complete. The renders the automatic retrieval and reprocessing of separate datasets or samples a rather tricky task, as each sample has to be curated separately and brought by the user to a generic format so that it can be compared to others.

3.1.2 PRIDE

The Proteomics Identifications Database (PRIDE) (72) it the largest repository in the world of mass spectrometry-based proteomics. It was set up in 2004 at the European Bioinformatics Institute (EBI). Submitters have to create a user account, request a unique identifier and upload all the raw files used in their "to be published" study, as well as a document in a specific format providing information about the study that is submitted and a description of each raw file that is uploaded. Each uploaded dataset is assigned with a unique ProteomeXchange accession to be reported in the corresponding manuscript. Upload of identification and/or quantification results is also possible but not mandatory. Lack of precise experimental and meta-data annotation makes reprocessing of these raw files also challenging. Moreover, except for the submitted files, the

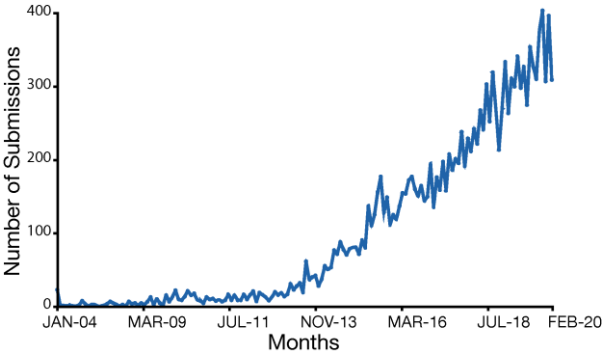

Figure 1.6 Number of submissions per month on PRIDE. Figure acquired and modified from https://www.ebi.ac.uk/pride/statisticsdetails on 27.02.2020

protein identifications or expression values regarding a specific sample measurement are not queryable.

Except for the data archive, PRIDE offers since 2015 a new repository for spectral libraries, called PRIDE Peptidome. Uploaded studies that include peptide identification information are used to extract spectra with peptide identifications. The extracted spectra with or without identifications are clustered into clusters. Low-quality clusters are then filtered out. The remaining clusters provide consensus spectra and peptide identifications. The final clusters are stored in PRIDE Cluster (83) and are accessible either via File Transfer Protocol (FTP) or API. The same cluster can also be used to generate spectral libraries for different organisms, and they are available for download and usage in spectrum search engines/tools. PRIDE also offers a collection of offline software tools that users can download and use on their personal computers, like:

- ms-data-core-api (84) that can be used for reading any proteomics data format to be later used in any application.
- PIA: Protein inference toolbox (85), which include algorithms for mass spectrometry-based protein inference and identification.
- PRIDE Mod (84), a modification library for the retrieval of protein modification information, by providing an identifier from preselected databases (e.g. UniMod)

By the time of writing of this thesis, PRIDE still remains the primary repository for the storage of mass spectrometry raw files, with a continuing growth on the number of datasets that are submitted per month (Figure 1.6).

3.1.3 MassIVE

The NIH-funded Center for Computational Mass Spectrometry developed a community resource for the storage and sharing of mass spectrometry data in the scientific community. Users have to first to register for an account in order to be able to access their private FTP directory and upload their raw files. It is suggested to pre-organize the files per dataset or experiment to make the rest of the data submission process easier. After the successful upload of the data, the user can select and trigger one of the platform's workflows for each dataset separately. To do so, the user must provide meta-data concerning the origin of the raw files with respect to the species, the instrument that was used, any allowed post-translational modifications (PTM) as well as a short description of the corresponding dataset or study. Each dataset can be accompanied by a set of keywords that will enable the indexing and accurate searching of the corresponding dataset. Datasets can also be assigned a ProteomeXchange accession so that they comply with standard publication requirements. If users also upload identification result files, they must map them to the corresponding raw files manually. There is also the option to let MassIVE run its own workflow and produce the corresponding identification results, also applying global FDR cutoffs. Peptide and protein identification results are query-able in their website, however not easily accessible via an application programming interface (API) calls. MassIVE also offers tools for the visualization of spectra as well as protein coverage plots.

MassiVE is another powerful repository, but its usage is limited to the storage and retrieval of raw files, spectra and peptide identification when available. A drawback, in this case, is that MassIVE stores no information on experimental designs so far and provides no quantification on the stored samples.

3.2 Biological databases (curated databases)

There is a clear line that separates public repositories from biological databases. It is the raw data along with any annotation that is provided and stored in the repositories, against the knowledge that is extracted from these raw data, proper data annotation and global comparisons among the resulting information that is stored in the databases. The pioneer databases or also referred to as knowledge bases (KB) in the omics world are described in this chapter.

3.2.1 Bgee

Bgee (86) is a database for the storage, sharing and comparison of gene expression data produced by different platforms (e.g. RNAseq, Affymetrix) originating from different species. The database stores expression data, originating only from "normal" healthy tissues, reprocessed by Bgee data curators. The reason behind the reprocessing of publicly available data is to provide a comparable reference of gene expression datasets. Another feature of the platform is the hosting of analysis or post-processing tools, such as call of presence or absence of gene expression, differential gene expression analysis (over- and under-expression) as well as information about gene orthology and homology between organs and species. That enables intra-species gene expression comparisons. The data curators of Bgee selected and downloaded raw data files from GEO and ArrayExpress. The raw data originate from healthy tissue samples from various organisms. Each dataset was reprocessed using the same pipeline to ensure maximum quality and later aggregated into organism-specific datasets. Reprocessed data concerning a specific organism, are provided in compressed comma separated format files, in the 'Species' page. Here, general information is provided for the selected organism as well as the gene expression calls files, containing records of unique combinations of a gene, tissue or anatomical entity and developmental stage of it, with the reported presence or absence of gene expression.

Anatomical ontologies of several species are aligned using ontology alignment methods (87). Via this alignment, homology relationships were designed between these ontologies. As a next step, developmental ontologies were mapped to the aligned anatomical ones. Using the provided annotation of the raw transcriptomics data and the by manual annotation where it was missing, Bgee integrates heterogeneous gene expression data on the new ontology. The anatomical and developmental ontologies of the different species can be accessed and visualized as tree structures in the web interface of Bgee.

Their 'Gene Search' functionality allows users to query genes of interest and locate them in the available organisms. By selecting a gene/organism entry on the returned list, users get transferred in a gene and organism-specific page. Here general information about the selected gene is displayed as well as gene expression values across different tissues with the possible expansion to tissue developmental stages. The gene expression values are normalized between 0 and 100, producing that way the visualized expression scores. Finally, cross-references to other resources are provided.

Bgee is accessible via a web interface but also offers programmatic access via an R package (BgeeDB) that provides full access to annotations, quantitative transcriptomics data as well as the gene expression calls. Last but not least, the R package includes functions to perform a GO-like enrichment analysis, where genes are mapped to anatomical entities based on their expression patterns.

3.2.2 UniProt

The number of known and annotated human protein-coding genes is not constant, and every day new scientific discoveries change our view on that. The Universal Protein Resource (UniProt) is a

comprehensive and long-term supported resource for protein sequence and annotation data. As of today, UniProt hosts and provides access to protein sequences for more than 84000 species that have sequenced genomes (84,387, release 2018_07). Most of these proteomes are results of the translation of genomes that are submitted in ENA, DDBJ (88) and GenBank. These sequences are then further enriched with metadata and annotations coming from Vectorbase (89), Ensembl (3) and WormBase ParaSite (90). A challenge that UniProt faces is the continuing growth of sequenced genomes, which most of the times results in the availability of sequences of very similar strains. In 2015 they introduced a redundancy removal process that identifies and removes identical proteomes of a species before they get imported into the UniProt knowledgebase (UniProtKB). The remaining proteomes are clustered using a Reference Proteome set (about 9% of all proteomes), which are manually selected by the scientific community with the goal of providing best proteome annotation per cluster.
UniProtKB is one of the primary sources for annotations of the human proteome and offers daily snapshots. The UniProtKB database is composed of two resources: TrEMBL contains automatically annotated gene and protein sequences and is based on snapshots of the human genome, while Swiss-Prot is a manually annotated and reviewed database, records of which are extracted from literature and curator-evaluated computational analyses. UniProt is the gold-standard protein-sequence database in mass spectrometry-based proteomics and is used to identify proteins by common database-centric search approaches.

3.2.3 ProteomicsDB

ProteomicsDB is an online resource that was initially developed for hosting the first mass spectrometry-based draft of the human proteome (10). A detailed description of the database, as well as the online platform and its capabilities, are described in the two manuscripts in the Appendix. The first manuscript provides a thorough view over the initial implementation as well as the first expansions and addition of new analysis tools, while the second is focused more on the integration of data originating from different omics types.
ProteomicsDB is, by the time of writing of this thesis, the only online resource that fills in the gap between experimental raw data and curated data, by storing and providing access to fully annotated experimental designs on protein expression data. Each project in ProteomicsDB is linked to a scientific publication and provides some minimal description of it. A PubMed id is provided as a cross-reference to the initial manuscript as well as a ProteomXchange identifier that links to the underlying raw data. Each project is a collection of different experiments in one study. ProteomicsDB enforces this separation between experiments, to ensure the best annotation of their experimental design. Each experiment has a unique internal identifier and is described by a name and a short description. A mandatory field in an experiment entry is the definition of the scope of the experiment, for example, full proteome, kinobead assay, thermal shift assay, or protein turnover assay. Every experiment includes a set of samples. Each sample also gets a unique identifier and is described by rich metadata, including:

- the tissue and species of origin of the corresponding sample,
- any kind of sample treatment (Temperature or inhibitor treatment) along with treatment details (e.g. time course) and the treatment agent of p
- the labelling method that was applied (e.g. TMT, SILAC) that was applied if so,
- the protease that was used for the digestion of the proteins in a sample and the digestion method (e.g. in-gel, in-solution),

- the online liquid chromatography system along with the mass spectrometer, the mass detector and resolution in every MS level and
- the acquisition mode (e.g. DDA, DIA, SRM)

Each experiment also includes a list of the raw files that belong to it. As the last step, each raw file is mapped to the corresponding sample. In the case of TMT-plex multiple files are assigned to one sample and vice versa.

Having all the above in place, ProteomicsDB can store full experimental designs by defining samples as biological or technical replicates and assigning them to treatments and conditions. An experimental design includes a full overview of a dose-, time- or temperature-dependent wet-lab experiment. Columns represent the different sample replicates and rows the drugs and concentrations, or temperatures or durations that were applied to those replicates. Each stored experimental design can be later visualized as a curve in the biochemical assay tab. All the above constitute the metadata around the actual data that is served via the platform.

ProteomicsDB plays another important role in these gaps between repositories and databases. As discussed earlier, there are repositories that store raw files and maybe identifications if not identification information only. Nevertheless, no other platform exists that connects metadata annotation with peptide and protein identification and quantification, with ProteomicsDB being the pioneer. ProteomicsDB remains the main point of reference when researchers want to compare protein expression across samples, identify selective compounds and design combination treatments as well as explore the evidence behind each stored peptide identification and its quality, by comparing each experimental spectrum to matching reference spectra from synthetic peptides or to predicted reference spectra.

3.3 Community standards and the need for them

Genomics and transcriptomics are well-established fields, where scientists have deployed standards that are respected and followed by researchers of these fields (91). Especially in RNAseq experiments, there is a standard pipeline that almost everyone follows. After the experimental procedure that was described above, there are standard steps that are followed and lead to differential expression analysis. These standards include but are not limited to:

- Quality control. FastQC (92) and FaQCs (93), are two of the most common software used for sequence quality analysis
- Data (read count) output format. The General Feature Format (GFF) is used here to store genes or transcripts and read counts
- Trimming and filtering for bad quality reads. Trimmomatic (94) and FASTX-Toolkit (95) are usually used to remove reads and bases of low quality as well as trim adaptor-sequences. Nowadays, this step can be avoided, as output from Illumina contains high quality reads.
- Alignment. It can be done with multiple tools. The golden standard today is STAR (96) alignment, a transcriptome specialised tool.
- Counting and Quantification. STAR can also provide counts but is not always the best tool for that task. Most commonly used software for quantification are featureCounts, which is available as an R package, Kallisto (97) and Salmon (98).
- Differential Expression (DE). Plenty of tools have been developed and used during the years. The ones that are most commonly used (as reviewed by (99,100)) are DESeq (101) and DESeq2 (102).

The proteomics field, being younger than the other two, has only a few well-defined standards. Each vendor's mass spectrometer outputs raw files in its own data format. Each proteomics laboratory follows a different workflow to prepare their samples for measurement, while the details of the workflow or sample preparation, most of the times, does not accompany the published raw files. Raw files that are deposited on publically available repositories are usually not accompanied by the identification or quantification results and not mapped to the measured samples. Moreover, identification and quantification are handled by tens of different software solutions, all following different assumptions or being non-deterministic, providing different results in the same version but also across versions of the same software. In 2002, the HUPO Proteomics Standard Initiative (HUPO-PSI) (103) was founded. HUPO-PSI defines community standards and open protocols for data representation in proteomics to enable data exchange, verification and comparison among laboratories. They define standards on different levels, such as Mass Spectrometry (PSI-MSS), Proteomics Informatics (PSI-PI) and Quality Control (PSI-QC). In Mass spectrometry, a default file format is proposed, mzML, for the representation of raw mass spectrometer output. mzML is a merged version of 2 preexisting formats, the mzData and the mzXML and is conducted using the XML markup language. A second initiative here is the definition of Controlled Vocabularies for use and description of Mass Spectrometers as wells as Protein identification and quantification software. The PSI-PI group introduced two file formats for the representation of protein identification and quantification results, named mzIdentML and mzQuantML respectively. The PSI-QC group charter is a multidisciplinary team and introduces the qcML file format. qcML files are designed for the exchange of QC metrics derived from mass spectrometry results. However, these standards are not mandatory to follow for publication and so far not so many laboratories follow them. Due to this, the data analysis workflows also differ from laboratory to laboratory and following different workflows using different software can lead to different results. There is an incrementing need in the proteomics field for common and standardized data processing and analysis options. In particular, online platforms should, as a first step, define and use a standard processing pipeline for all the proteomic data that they store and analyze. That would already enable the cross-platform data comparisons, but would also allow the easy integration of data originating from other platforms.

4 Databases and database management systems

All repositories that were described so far store a large amount of data. Storing all this data as files of the file system is not worthy. With the exception being the raw data files that are stored in repositories that are not queried or processed to provide information to the user, all the rest of the data and metadata that they store need to be organized in databases. Databases (DB) are collections of elements relative to each other, which are structured and stored in an appropriate way. A Database management system (DBMS), is a software that implements all functions that need to be supported, such as selections, insertions, deletions, contemporary access and security. A DBMS is capable of handling multiple DBs at the same time in the same machine.

4.1 Database models

Over the last 50 years, many different database models have been proposed and implemented. In 1970, Edgar Frank Codd proposed the **relational data model** (104). The model's simple and comprehensive architecture led to its integration to many high-load applications. One of the model's main advantages is that it can be described in a mathematical way using either Set Theory or Predicate Logic. The main points of the relational data model are:

- The support of the data independence, so that changes in the structure and organization of the DB will not affect the connected application,
- The avoidance of redundancy, which is the case when duplicated data is stored in several parts of the database.

The relational model defines the rules for database normalization, also called normal forms (NF). Each form is named with an incrementing integer and has the property that it should respect the previous form. The First normal form (1NF) sets the rule that information is stored in a table where each column contains atomic values. In 1NF, there are no repeating groups of columns. The second normal form (2NF) requires the table to be in 1NF, and all table's columns should be dependent on the tables primary key. The third normal form (3NF) requires the table in 2NF, and there should be no column with transitive dependency on the primary key. The 2NF and 3NF are concerned with functional dependencies; in other words, constraints between two sets of columns in a table. The fourth normal form (4NF) is concerned with multivalued dependencies. A multivalued dependency requires a table with at least three columns (e.g. X, Y and Z), where for a value of column X there are well-defined sets of values for the other two columns y and Z respectively while these two sets are independent to each other. It is quite rare that a table in the 4NF will not comply with the next NF, the fifth normal form (5NF), also known as project-join normal form (PJNF). In the 5NF, a table is in the state that it cannot be further loss-less decomposed in 2 or more tables. It is quite rare that a table in the 4NF does not comply with the 5NF. Finally, the sixth normal model requires every table to consist of the primary key and maximum one more column. It is rarely followed as it proposes an extreme table decomposition that increases query complexity. Data stored in a relational database can be queried by using the Structured Query Language (SQL).

A different way of storing data is to not organize them in tables known as relations in the relational model but to store them as entities that have relations to each other in the form of triples. A triple is a data entity consisting of a subject, a predicate and an object, as defined from the Resource Description Framework (RDF) (105). The database that stores triples is called a triplestore or RDF-store, and it can be queried using semantic queries. A popular query language for RDF-stores is SPARQL. The RDF data model represents named properties and property values. It consists of the

three following object types: Resources, Properties and Statements. A **resource** is anything that is described by an RDF expression. It can be a web page, a part of a webpage, a collection of webpages (website) or even something not accessible via web, like a printed book. A resource is always named by a Universal Resource Identifier (URI). A **property** is a characteristic or relation or attribute that is used to describe a resource. Each property defines the resource types it can describe, the allowed values as a form of enumerations and relationships with other properties. A resource along with a named property and the value of this property constitute an RDF **statement**, with the resource being the subject, the property being the predicate and the value being the object of the aforementioned triple. The object in a statement can have to types of values: another resource or a literal value such as a string or an integer. The definition of an RDF schema is written in the Extensible Markup Language (XML). The RDF data model is also called a semantic data model, as it captures the meaning of an application environment, something that is not possible in a relational data model. As a result, RDF data models are the standard in describing ontologies. For better exploitation of the capabilities of such a data representation, more tools have been developed on top of that model. The Web Ontology Language (OWL and OWL2), is a semantic web language for the detailed description of entities relations between entities and groups of entities. It is a computational logic-based language so that computer programs can exploit the knowledge that is expressed in OWL. OWL follows the same syntax and principles with RDF, while at the same time, it expands the RDF properties' **attributes**.

As the triples can also be described as rules, rule-based inference engines have been developed, also known as reasoners. As stated in their name, they exploit the information described in properties, like in case of a property being bi-directional or transitive. As an example, having the triplets <10><greater_than><5>, and <5><greater_than><3>, if the predicate <greater_than> was defined as a transitive property, the reasoner would create the inferred triple <10><greater_than><3>. Of course, this is an easy example that even pure SQL is able to resolve. It gets more complicated when a property defined as the reverse of another (like greater/smaller or parent_of/child_of) and becomes even more complicated when these attributes are combined. OWL same as RDF can be queried using SPARQL. RDF and OWL are well-defined standards by the World Wide Web Consortium (W3C).

Semantic data models opened the way for the creation of several others. Graph data models are an extension that includes though more structures than a single triplestore. Here resources are depicted as nodes and properties as edges. All attributes of a resource are stored as metadata of a node, as well as the attributes of a property as metadata of that edge. Graph databases exploit this data representation by implementing algorithms originating from the field of graph theory, enabling the fast calculation of shortest paths, nearest neighbours and other algorithms. A common language to query data from graph databases is the so-called GraphQL.

Each of the aforementioned data models is better suitable with specific use-cases that should be considered when storing biological data. Protein or transcript expression data are usually organized by sample, tissue and experiment all grouped in a broader project topic. This type of data is often accompanied by metadata like the description of the sample or the scope of an experiment and the protocols that were followed. This use-case needs a relational data model as it describes one-to-many relations between different tables. This model supports the continuous addition of new data that correspond to these tables and can be easily expanded to include new tables with relations to the existing ones. Data retrieval is also trivial as collecting information for a sample of interest or all samples that a protein is expressed above a defined level, needs to

access only specific tables. However, the relational data model is not the best choice for protein-protein or protein-drug interaction networks, especially when this data needs to be treated as a network or a graph. In this case, the graph data model is the first choice as it is optimized for the storage of relations between different entities and includes algorithms for the in-depth exploration of any graph. It is not optimized though for data aggregation, at least not in the same way as a relational data model. As a result, a combination of data models can provide better solutions than the usage of a single data model for the storage of different kinds of data.

4.2 SAP HANA

SAP HANA is an in-memory relational database management system (RDBMS) developed by SAP SE. It bundles together three different layers (Figure 1.7).

The calculation layer holds the data, database and DBMS all under the umbrella of a service named index server. The index server is responsible for the process of incoming queries with the help of a query plan optimizer and the execution engine. Data can be stored in row- as well as column-oriented tables. In the second option, HANA utilizes several compression strategies for every column. As a first step, each column uses a dictionary, which maps all distinct values of a column to number. Finally, run length-, prefix- or sparse- encoding encryption algorithms are directly applied when possible. The in-memory storage removes the overhead of querying and retrieval of data from the secondary memory, the hard disk. That allows the full utilization of the processor and main memory which enables the real-time processing of vast amounts of data. With the level of compression that was described earlier, it is computationally hard to perform insertions on the fly. For that reason, HANA uses secondary storage called the Delta storage. That allows the efficient merging of big chunks of imported data to the actual compressed storage in the hard disk. The delta storage exists only in main memory, which means that any change in it has to be written back to the disk. This operation is performed by using delta logs. In the case of write-failure, the database can recover by using the delta logs and replicate the last changes.

The control flow layer includes all functionalities regarding the backend webserver. It deploys backend endpoints close to the database with intermediate security layers so that only the proper database users with the appropriate privileges have the right to expose data. It includes the integrated extended application services (XS engine) and an HTTP server (web dispatcher). The side by side existence of the two services enables the use of ODATA services (xsodata in SAP HANA), server-side JavaScript modules (XSJS) that are handled and executed by the XS engine and finally the frontend framework.

Modern SAP HANA installations come with two different versions of the extended application services: the classic (XSC) service and the advanced (XSA) service. In XSC, all backend and frontend procedures and files are organized in so-called Delivery Units (DU). Each DU is fully compatible and track-able with version control systems, like GIT, and can be deployed by the in-HANA repository system called REGI. The frontend in DUs is part of the SAPUI5/OPEUI5 framework. XSA renames and reorganizes the DUs into multi-target applications (MTA), which includes more strict protocols regarding the security and access levels of each user to the separate layers of the backend and frontend. XSC can serve many DUs at once without further configuration. Contradictory, XSA is based on the logic that MTAs should be fully isolated applications, a fact that makes the transition from the one engine to the other quite challenging. In XSA each MTA deploys tables, procedures and data in its own container (HDI). An MTA belongs to a single user by default, and only this user has access to the corresponding HDI container. The same applies to *.xsjs

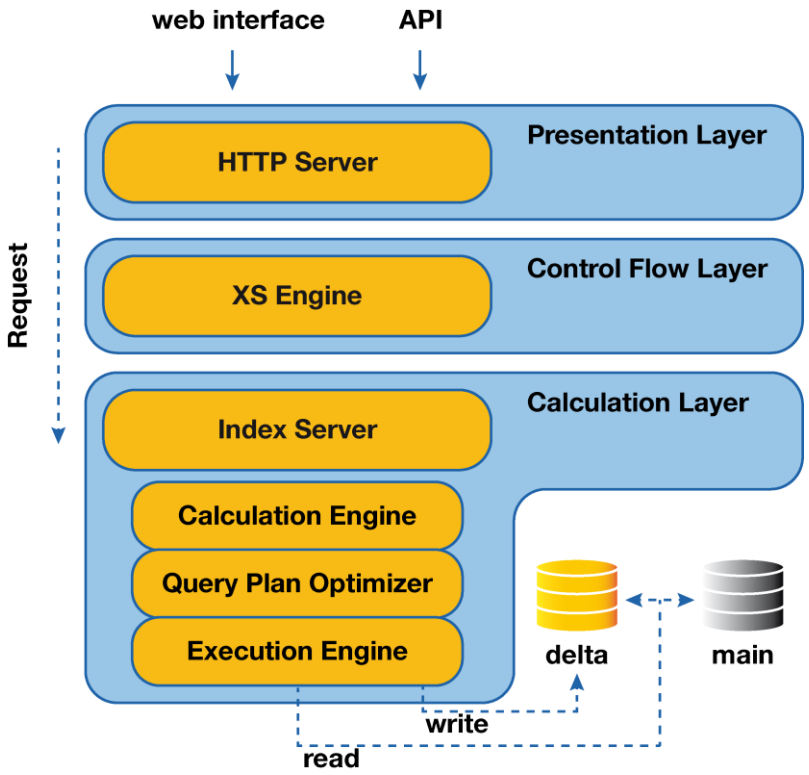

Figure 1.7 SAP HANA schematic overview. SAP HANA is a single system, which bundles together the three depicted layers.

endpoints, as in XSA they are stored and served in their defined HDI container, making it extremely hard for MTA applications to reuse existing endpoints from one another. In most cases, a total reimplementation of an XSC DU is needed in order to be compatible with XSA.

SAP HANA supports several standards for data access, such as Java Database Connectivity (JDBC) and Open Database Connectivity (ODBC) adapters. Moreover, it includes several adapters for connection and integration of external services (e.g. R server) or data sources (e.g. remote connections via JDBC to other databases) or C++ libraries via the application function library (AFL). AFL, together with an adapter that implements the Google Remote Procedure Calls (GRPC) enables the use of external Machine Learning and Deep Learning models. All these services can be implemented directly in HANA in the form of procedures. Although HANA has an embedded Graph Engine, with the purpose of emulating a graph database, its performance is not comparable with other Graph Databases.

This work is based on HANA 2.X that enables runtime on IBM Power infrastructure and utilizes the XSC engine with a focus on functionalities that can be easily migrated to the XSA in the future.

4.3 Software communication standards

Most of the web platforms that are described in this chapter use either a web interface for the access and query of the raw or processed data, or an FTP connection for direct download of the available files. This type of interfaces is necessary for human intelligence as they provide the data in a more extended way and organized in meaningful for the human mind tabs and webpages. Even the data requests executed on the website to the server, with the purpose of data visualization is many times organized in a human-readable way. In these years that computer technology thrives, machine learning and deep learning tools evolve, and data mining and

knowledge discovery are getting more important every day, the primary way of communication between a software program and a resource or online database is via APIs.

4.3.1 Application programming interface – API

APIs, in computer science, are interfaces or protocols for the communication between different software programs or parts of them. The purpose of an API is to simplify software implementation and maintenance. There are different forms of API specifications, including specifications for data structures, variables or remotes calls. Along with the API specification, documentation is provided usually describing its usage and facilitating implementation. Another specification of an API is the one that describes an interface between a server and a client or more commonly called backend and frontend, respectively. In that specification, the frontend makes a request to the server using a defined format and expects a response in a predefined format. In some cases, instead of the return of a response, a request might initiate predefined actions. The frontend requests use standard protocols such as the Hypertext Transfer Protocol (HTTP), which are then called HTTP requests. The response message of the server is in a structured format, most commonly using XML or JavaScript Object Notation (JSON). This specification is called a Web API. A Web API can consist of multiple APIs from different servers, and in that case, it is called a mashup. In this case, especially in the social media space, content that is created in one place can be shared and published dynamically in multiple web locations.

APIs can be released in the following policies: private, partner or public APIs. A private API is used only internally in a company or organization. A partner API is used from specific partner applications, allowing direct access to certain actions of the main application. That way, the main application can also track the usage of their API and exercise quality control. Finally, a public PAI is open for usage to the public. A common issue with public APIs is their interface stability. Changes that are applied to the API should be documented and early announced as it may break compatibility with client applications that exploit its functionality. A best practice is declaring older versions of API endpoints as deprecated and release of the modified endpoints as newer versions. Deprecation informs other developers of future removal of the specific endpoint so that they can slowly migrate to newer ones.

Although most of the afore-mentioned online resources and databases offer API endpoints that are open to the public, most of them do not follow the same response structure. An example is the retrieval of spectra that can be represented in different formats.

4.3.2 F.A.I.R. principles

APIs that serve the same type of data but in different formats, introduce difficulties in the implementation of both mashups and applications that need to extract knowledge from these APIs. Missing documentation of other APIs makes it impossible for applications to interpret the returned data, especially in cases that no metadata is served.

In 2016, Wilkinson et al. came with a well-defined set of principles as guidelines that should be followed by resources that want to enhance their data reusability, known as F.A.I.R principles (106). The word FAIR is formed by the initial letters of the words describing the four main principles: Findability, Accessibility, Interoperability and Reusability. Previous attempts to define such principles focused on the human scholar. FAIR principles focus on making the data interpretable and automatically findable by the machines as well as enhancing data reusability by the individuals. The first principle, findability, describes the way data and metadata should be served and organized so that machine and human can find and identify the uniqueness of each

retrieved entity. Data and metadata should be described by globally unique identifiers (e.g. DOI). That renders data and metadata indexable and consequently searchable by search engines. The second principle describes the ways that data should be accessed. It requires a well-described and open communication protocol that in case of limited access, it allows for an authentication/authorization procedure. The third principle enforces the usage of controlled vocabularies that are either described from the primary resource or defined in public ontologies. That way, machine and human can avoid mistakes of mapping terms to meanings. Interoperability also demands the use and sharing of qualified references and cross-references to other resources. That way, data can be integrated among resources and interoperable with workflows and applications regarding storage, processing, and analysis. Finally, the fourth principle is also the ultimate goal of the FAIR principles, reusability of the stored data. It requires data and metadata to be richly described, accompanied by a plethora of relevant attributes. If possible, data should follow community standards.

There are already many resources that are considered FAIR, based on the way they serve their content. UniProt is a nice example as all entries are identified with unique and stable Universal Resource Location (URL) links. Following these links, content can be served in a plethora of formats, including a web page, RDF and plain-txt, covering that way the 'F' and 'A' principles. Each record in UniProt is accompanied by rich metadata that is served either in HTML format (human-readable) or RDF (machine-readable), while in the case of RDF they make use of controlled vocabularies. That way, UniProt also respects the third 'I' principle. Moreover, each record in UniProt offers different cross-references providing links to external resources, enabling that way strong and rich citations. All these links are machine-readable when accessing a record-webpage in its RDF format. That enables data reusability ('R' principle) (107).

The process of making a resource FAIR is called FAIRification (Figure 1.8) and consists of the following steps:

1. Retrieval of the nonFAIR data.
2. Analysis of the retrieved data and definition of concepts that could describe them. At this step definition of possible relations between the data is also necessary as it will lead easier to the completion of step 3.
3. Definition of a semantic model that describes the meaning of the several entities and relations, in an accurate but also computer-actionable way. Existing semantic models can help speed up the FAIRification process.
4. Make data linkable by applying the semantic model of step 3 on the nonFAIR data.
5. Licensing, as the absence of a license, might be a driving factor in people not reusing the supplied data, regardless of them being open data.
6. Definition of rich metadata, as these will enhance the information that is encoded, and
7. Publication of the FAIR resource. With the last step, metadata of the resource will be indexed by search engines.

GO FAIR (https://www.go-fair.org) is a detailed resource providing guides and examples that help resource in the process of getting FAIR.

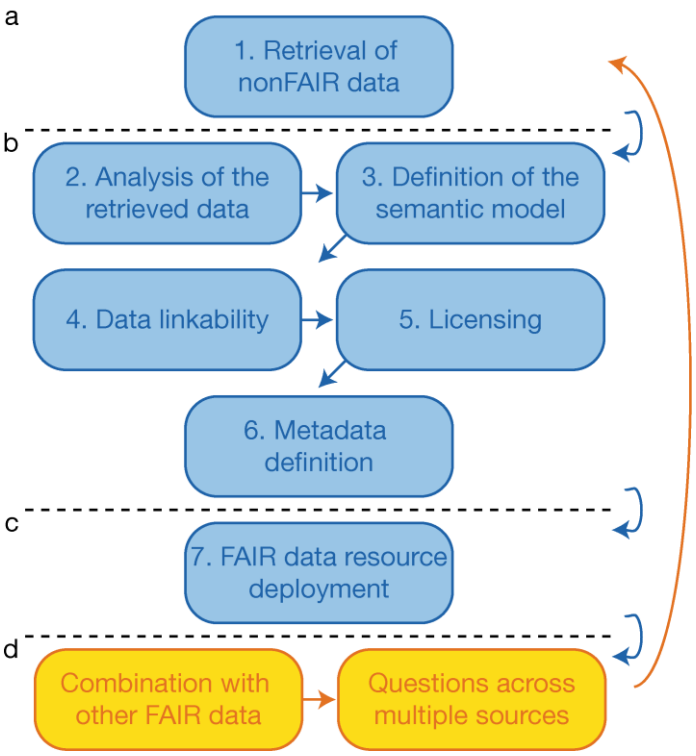

Figure 1.8 Process of FAIRification. a) The data layer. b) The data modelling and FAIRification layer. c) The publication layer. d) The FAIR data use-case layer. Figure acquired and adjusted from www.go-fair.org/fair-principles/fairification-process.

5 Objectives

A large variety of resources exist, each one storing their data using different data models or database management systems. All of them though, are focused in one omics-type or one organism. That can be either due to the complexity of the corresponding omics data that requires the utilization of different data models or even systems or simply because of their main research focus. Even in case of storage of quantitative multi-omics expression data originating from multiple organisms, there are even fewer resources that perform multi-omics data integration, a procedure that can provide the scientific and clinic community with valuable tools and hypotheses. Visualization of such data is also not a trivial procedure, as different omics-types use different plotting mechanisms to deliver the desired visual output to the user. Even more challenging is the visualization of different organisms, especially in the case of quantitative omics data originating from different tissues and organisms, where any kind of visual comparison is needed. The afore-mentioned challenges can be resolved with a generic, well-designed data model that can be supported by a powerful database management system. Proper user interface frameworks along with good software design and technology methods can help in providing generalized visualization tools that would take away the need for separate views and webpages, causing significant code duplication.

ProteomicsDB, built and hosted on an SAP HANA system, has every capability of storing and handling quantitative multi-omics and multi-organism expression data. This thesis is devoted to the design and implementation of a generic data model, capable of storing the aforementioned datasets. Moreover, exploiting the powerful calculation engine of SAP HANA, real-time data analysis and integration tools are implemented and served via ProteomicsDB's analysis tools. Finally, certain modifications on the platforms' user interface enabled the generic visualization of several organisms, allowing the easy future expansion to more organisms.

The first paper describes the initial status of ProteomicsDB before the beginning of this thesis. It later introduces the expansions of the initial schema to include new data models that overcome the shortcomings that were described in the introduction, regarding the storage of other omics types. This paper also presents a model-solution for the problem of mapping gene or protein expression data that use different initial resources for their identifiers. The same model-solution is expanded to store any kind of relation between identifiers, opening the way to storing protein-protein interaction data as well as functional pathway structures. The paper concludes with the inclusion of a final data model, which enables ProteomicsDB the storage of cell viability studies. New visualization and analysis tools were created, allowing the user to query and interact with multi-omics expression data, protein-protein interaction networks and cell viability information on multiple datasets, drugs and cell lines. These extensions, together with the data and metadata that are stored in ProteomicsDB, open the way towards the storage of more data types and the development of new applications regarding multi-omics data integration.

The second paper expands the data wealth of ProteomicsDB with more proteomics and transcriptomic studies. MComBat normalization enabled the mRNA-guided missing value imputation method. The same method can be reversed to impute transcript expression values based on matching proteomics and transcriptomics data. The biochemical assay data that were already stored are now enriched with more protein melting properties and extended with protein turnover data. The cell viability data model stores now one more study raising the number of screened drugs to 20000. The stored omics expression and cell viability data led to modelling drug

sensitivity using elastic net regression models. The fitted models can be used on any sample stored in the database to predict if a drug is effective or not on them. The data model extension to support temporary user-uploaded data opens the way for applying real-time analysis on the platform as well as comparison to ProteomicsDB data. The user uploaded data are slowly integrated into all analysis tools, for example, drug sensitivity prediction on user-uploaded samples or user expression data exploration in the interactive expression heatmap.

All these extensions transform ProteomicsDB into a unique and powerful resource for life science research, and at the same time, set the ground for the development of future analysis and visualization tools.

6 Abbreviations

AC	Alternative current
AFL	Application function library
API	Application programming interface
ATP	Adenosine triphosphate
AUC	Area under the curve
cDNA	Complementary deoxyribonucleic acid
CID	Collision-induced dissociation
CLL	Chronic lymphocytic leukaemia
DB	Database
DBMS	Database management system
DC	Direct current
DDA	Data dependent acquisition
DIA	Data independent acquisition
DNA	Deoxyribonucleic acid
DU	Delivery unit
ELISA	Enzyme-Linked Immunosorbent Assays
ESI	Electrospray ionization
ETD	Electron-transfer dissociation
FAIR	Findability, Accessibility, Interoperability and Reusability
FDR	False discovery rate
FPKM	Fragments Per Kilobase of transcript, per Million mapped reads
FT	Fourier transformation
FTP	File transfer protocol
GB	Gigabyte
GEO	Gene Expression Omnibus
GO	Gene Ontology
GRPC	General-purpose remote procedure call
HCD	Higher energy collision-induced dissociation
HDI	HANA Deployment Infrastructure
HTML	Hypertext markup language
HTTP	Hypertext transfer protocol
IMAC	Immobilized metal affinity chromatography
IPA	Ingenuity pathway analysis
JDBC	Java Database Connectivity
JSON	JavaScript Object Notation
KB	Knowledgebase
LC	Liquid chromatography
MCIA	Multiple co-inertia analysis
MS	Mass spectrometer
MTA	Multi-target application
NF	Normal form
NGS	Next generation sequencing
ODBC	Open Database Connectivity
OWL	Web ontology language
PCA	Principal component analysis
PSI	Proteomics Standards Initiative
PSM	Peptide-Spectrum Match
PTM	Post-translational modification
QC	Quality control
RDBMS	Relational database management system
RDF	Resource Description Framework
RE	Relative Effect
RF	Radio frequency
RMA	Robust Multiarray Average
RNA	Ribonucleic acid
RPKM	Reads Per Kilobase of transcript, per Million mapped reads
RSEM	RNAseq by expectation maximization
SQL	Simple query language
SRM	Selected reaction monitoring
TB	Terabyte
TOF	Time of flight
TPM	Transcripts per Million
URI	Universal resource identifier
URL	Universal resource locator
UTR	Untranslated region
XIC	Extracted ion chromatogram
XML	Extensible Markup Language
XS	Extended services
XSA	Advanced extended services
XSC	Classic extended services

7 References

1. Crick, F. (1970) Central dogma of molecular biology. *Nature*, **227**, 561-563.
2. International Human Genome Sequencing, C. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931-945.
3. Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M.R., Armean, I.M., Bennett, R., Bhai, J., Billis, K., Boddu, S. *et al.* (2019) Ensembl 2019. *Nucleic Acids Res*, **47**, D745-D751.
4. Smith, L.M., Kelleher, N.L. and Consortium for Top Down, P. (2013) Proteoform: a single term describing protein complexity. *Nat Methods*, **10**, 186-187.
5. Aebersold, R., Agar, J.N., Amster, I.J., Baker, M.S., Bertozzi, C.R., Boja, E.S., Costello, C.E., Cravatt, B.F., Fenselau, C., Garcia, B.A. *et al.* (2018) How many human proteoforms are there? *Nat Chem Biol*, **14**, 206-214.
6. Zecha, J., Meng, C., Zolg, D.P., Samaras, P., Wilhelm, M. and Kuster, B. (2018) Peptide Level Turnover Measurements Enable the Study of Proteoform Dynamics. *Mol Cell Proteomics*, **17**, 974-992.
7. Goldberg, A.L. (2003) Protein degradation and protection against misfolded or damaged proteins. *Nature*, **426**, 895-899.
8. Kavallaris, M. and Marshall, G.M. (2005) Proteomics and disease: opportunities and challenges. *Med J Aust*, **182**, 575-579.
9. Harper, J.W. and Bennett, E.J. (2016) Proteome complexity and the forces that drive proteome imbalance. *Nature*, **537**, 328-338.
10. Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H. *et al.* (2014) Mass-spectrometry-based draft of the human proteome. *Nature*, **509**, 582-587.
11. Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S. *et al.* (2014) A draft map of the human proteome. *Nature*, **509**, 575-581.
12. Jonas, S. and Izaurralde, E. (2015) Towards a molecular understanding of microRNA-mediated gene silencing. *Nat Rev Genet*, **16**, 421-433.
13. Filipowicz, W., Jaskiewicz, L., Kolb, F.A. and Pillai, R.S. (2005) Post-transcriptional gene silencing by siRNAs and miRNAs. *Curr Opin Struct Biol*, **15**, 331-341.
14. DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A. and Trent, J.M. (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet*, **14**, 457-460.
15. Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467-470.
16. Pease, A.C., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P. and Fodor, S.P. (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci U S A*, **91**, 5022-5026.
17. Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T. and Solas, D. (1991) Light-directed, spatially addressable parallel chemical synthesis. *Science*, **251**, 767-773.
18. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5**, 621-628.

19. Miller, M.B. and Tang, Y.W. (2009) Basic concepts of microarrays and potential applications in clinical microbiology. *Clin Microbiol Rev*, **22**, 611-633.
20. Chirgwin, J.M., Przybyla, A.E., MacDonald, R.J. and Rutter, W.J. (1979) Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. *Biochemistry*, **18**, 5294-5299.
21. Aviv, H. and Leder, P. (1972) Purification of biologically active globin messenger RNA by chromatography on oligothymidylic acid-cellulose. *Proc Natl Acad Sci U S A*, **69**, 1408-1412.
22. Xiang, C.C., Kozhich, O.A., Chen, M., Inman, J.M., Phan, Q.N., Chen, Y. and Brownstein, M.J. (2002) Amine-modified random primers to label probes for DNA microarrays. *Nat Biotechnol*, **20**, 738-742.
23. Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L. and Liu, C. (2011) Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One*, **6**, e17238.
24. Gharaibeh, R.Z., Fodor, A.A. and Gibas, C.J. (2008) Background correction using dinucleotide affinities improves the performance of GCRMA. *BMC Bioinformatics*, **9**, 452.
25. Goodwin, S., McPherson, J.D. and McCombie, W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, **17**, 333-351.
26. Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nat Rev Genet*, **11**, 31-46.
27. Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.
28. Chang, J.C., Wooten, E.C., Tsimelzon, A., Hilsenbeck, S.G., Gutierrez, M.C., Elledge, R., Mohsin, S., Osborne, C.K., Chamness, G.C., Allred, D.C. *et al.* (2003) Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet*, **362**, 362-369.
29. Anderson, N.L. and Anderson, N.G. (1998) Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis*, **19**, 1853-1861.
30. Blackstock, W.P. and Weir, M.P. (1999) Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol*, **17**, 121-127.
31. Pei, G., Chen, L. and Zhang, W. (2017) WGCNA Application to Proteomic and Metabolomic Data Analysis. *Methods Enzymol*, **585**, 135-158.
32. Voshol, H., Ehrat, M., Traenkle, J., Bertrand, E. and van Oostrum, J. (2009) Antibody-based proteomics: analysis of signaling networks using reverse protein arrays. *FEBS J*, **276**, 6871-6879.
33. Kurien, B.T. and Scofield, R.H. (2009) Introduction to protein blotting. *Methods Mol Biol*, **536**, 9-22.
34. Renart, J., Reiser, J. and Stark, G.R. (1979) Transfer of proteins from gels to diazobenzyloxymethyl-paper and detection with antisera: a method for studying antibody specificity and antigen structure. *Proc Natl Acad Sci U S A*, **76**, 3116-3120.
35. Engvall, E. and Perlmann, P. (1971) Enzyme-linked immunosorbent assay (ELISA). Quantitative assay of immunoglobulin G. *Immunochemistry*, **8**, 871-874.
36. Van Weemen, B.K. and Schuurs, A.H. (1971) Immunoassay using antigen-enzyme conjugates. *FEBS Lett*, **15**, 232-236.

37. Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198-207.
38. Tyers, M. and Mann, M. (2003) From genomics to proteomics. *Nature*, **422**, 193-197.
39. Larsen, M.R., Thingholm, T.E., Jensen, O.N., Roepstorff, P. and Jorgensen, T.J. (2005) Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. *Mol Cell Proteomics*, **4**, 873-886.
40. Pinkse, M.W., Uitto, P.M., Hilhorst, M.J., Ooms, B. and Heck, A.J. (2004) Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-NanoLC-ESI-MS/MS and titanium oxide precolumns. *Anal Chem*, **76**, 3935-3943.
41. Hahne, H., Pachl, F., Ruprecht, B., Maier, S.K., Klaeger, S., Helm, D., Medard, G., Wilm, M., Lemeer, S. and Kuster, B. (2013) DMSO enhances electrospray response, boosting sensitivity of proteomic experiments. *Nat Methods*, **10**, 989-991.
42. Steen, H. and Mann, M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol*, **5**, 699-711.
43. Biemann, K. (1988) Contributions of mass spectrometry to peptide and protein structure. *Biomed Environ Mass Spectrom*, **16**, 99-111.
44. Olsen, J.V., Macek, B., Lange, O., Makarov, A., Horning, S. and Mann, M. (2007) Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods*, **4**, 709-712.
45. Wells, J.M. and McLuckey, S.A. (2005) Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol*, **402**, 148-185.
46. Ma, B. (2015) Novor: real-time peptide de novo sequencing software. *J Am Soc Mass Spectrom*, **26**, 1885-1894.
47. Zolg, D.P., Wilhelm, M., Schmidt, T., Medard, G., Zerweck, J., Knaute, T., Wenschuh, H., Reimer, U., Schnatbaum, K. and Kuster, B. (2018) ProteomeTools: Systematic Characterization of 21 Post-translational Protein Modifications by Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS) Using Synthetic Peptides. *Mol Cell Proteomics*, **17**, 1850-1863.
48. Zolg, D.P., Wilhelm, M., Schnatbaum, K., Zerweck, J., Knaute, T., Delanghe, B., Bailey, D.J., Gessulat, S., Ehrlich, H.C., Weininger, M. *et al.* (2017) Building ProteomeTools based on a complete synthetic human proteome. *Nat Methods*, **14**, 259-262.
49. Gessulat, S., Schmidt, T., Zolg, D.P., Samaras, P., Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A. *et al.* (2019) Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat Methods*, **16**, 509-518.
50. Lu, P., Vogel, C., Wang, R., Yao, X. and Marcotte, E.M. (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol*, **25**, 117-124.
51. Gilchrist, A., Au, C.E., Hiding, J., Bell, A.W., Fernandez-Rodriguez, J., Lesimple, S., Nagaya, H., Roy, L., Gosline, S.J., Hallett, M. *et al.* (2006) Quantitative proteomics analysis of the secretory pathway. *Cell*, **127**, 1265-1281.
52. Higgs, R.E., Knierman, M.D., Gelfanova, V., Butler, J.P. and Hale, J.E. (2005) Comprehensive label-free method for the relative quantification of proteins from biological samples. *J Proteome Res*, **4**, 1442-1450.
53. Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature*, **473**, 337-342.

54. Silva, J.C., Gorenstein, M.V., Li, G.Z., Vissers, J.P. and Geromanos, S.J. (2006) Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics*, **5**, 144-156.
55. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. and Kuster, B. (2007) Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem*, **389**, 1017-1031.
56. Krause, A.W., Carley, W.W. and Webb, W.W. (1984) Fluorescent erythrosin B is preferable to trypan blue as a vital exclusion dye for mammalian cells in monolayer culture. *J Histochem Cytochem*, **32**, 1084-1090.
57. Yip, D.K. and Auersperg, N. (1972) The dye-exclusion test for cell viability: persistence of differential staining following fixation. *In Vitro*, **7**, 323-329.
58. Mosmann, T. (1983) Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays. *J Immunol Methods*, **65**, 55-63.
59. Berg, K., Zhai, L., Chen, M., Kharazmi, A. and Owen, T.C. (1994) The use of a water-soluble formazan complex to quantitate the cell number and mitochondrial function of Leishmania major promastigotes. *Parasitol Res*, **80**, 235-239.
60. Tominaga, H., Ishiyama, M., Ohseto, F., Sasamoto, K., Hamamoto, T., Suzuki, K. and Watanabe, M. (1999) A water-soluble tetrazolium salt useful for colorimetric cell viability assay. *Analytical Communications*, **36**, 47-50.
61. Scudiero, D.A., Shoemaker, R.H., Paull, K.D., Monks, A., Tierney, S., Nofziger, T.H., Currens, M.J., Seniff, D. and Boyd, M.R. (1988) Evaluation of a soluble tetrazolium/formazan assay for cell growth and drug sensitivity in culture using human and other tumor cell lines. *Cancer research*, **48**, 4827-4833.
62. Rampersad, S.N. (2012) Multiple applications of Alamar Blue as an indicator of metabolic function and cellular health in cell viability bioassays. *Sensors (Basel)*, **12**, 12347-12360.
63. O'brien, J., Wilson, I., Orton, T. and Pognan, F. (2000) Investigation of the Alamar Blue (resazurin) fluorescent dye for the assessment of mammalian cell cytotoxicity. *European journal of biochemistry*, **267**, 5421-5426.
64. Duellman, S.J., Zhou, W., Meisenheimer, P., Vidugiris, G., Cali, J.J., Gautam, P., Wennerberg, K. and Vidugiriene, J. (2015) Bioluminescent, Nonlytic, Real-Time Cell Viability Assay and Use in Inhibitor Screening. *Assay Drug Dev Technol*, **13**, 456-465.
65. Niles, A.L., Moravec, R.A. and Riss, T.L. (2009) In vitro viability and cytotoxicity testing and same-well multi-parametric combinations for high throughput screening. *Curr Chem Genomics*, **3**, 33-41.
66. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W. and Stegle, O. (2018) Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*, **14**, e8124.
67. Meng, C., Helm, D., Frejno, M. and Kuster, B. (2016) moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets. *J Proteome Res*, **15**, 755-765.
68. Meng, C., Kuster, B., Culhane, A.C. and Gholami, A.M. (2014) A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*, **15**, 162.
69. Frejno, M., Zenezini Chiozzi, R., Wilhelm, M., Koch, H., Zheng, R., Klaeger, S., Ruprecht, B., Meng, C., Kramer, K., Jarzab, A. *et al.* (2017) Pharmacoproteomic characterisation of human colon and rectal cancer. *Mol Syst Biol*, **13**, 951.
70. Sayers, E.W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K.D. and Karsch-Mizrachi, I. (2019) GenBank. *Nucleic Acids Res*, **47**, D94-D99.

71. Stoesser, G., Baker, W., van den Broek, A., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., Lombard, V., Lopez, R. *et al.* (2003) The EMBL Nucleotide Sequence Database: major new developments. *Nucleic Acids Res*, **31**, 17-22.
72. Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D.J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M. *et al.* (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res*, **47**, D442-D450.
73. The UniProt, C. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res*, **45**, D158-D169.
74. Sigrist, C.J., de Castro, E., Cerutti, L., Cuche, B.A., Hulo, N., Bridge, A., Bougueleret, L. and Xenarios, I. (2013) New and continuing developments at PROSITE. *Nucleic Acids Res*, **41**, D344-347.
75. Thurmond, J., Goodman, J.L., Strelets, V.B., Attrill, H., Gramates, L.S., Marygold, S.J., Matthews, B.B., Millburn, G., Antonazzo, G., Trovisco, V. *et al.* (2019) FlyBase 2.0: the next generation. *Nucleic Acids Res*, **47**, D759-D765.
76. Bult, C.J., Blake, J.A., Smith, C.L., Kadin, J.A., Richardson, J.E. and Mouse Genome Database, G. (2019) Mouse Genome Database (MGD) 2019. *Nucleic Acids Res*, **47**, D801-D806.
77. Berman, H., Henrick, K. and Nakamura, H. (2003) Announcing the worldwide Protein Data Bank. *Nat Struct Biol*, **10**, 980.
78. Andreeva, A., Kulesha, E., Gough, J. and Murzin, A.G. (2020) The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res*, **48**, D376-D382.
79. Dawson, N.L., Lewis, T.E., Das, S., Lees, J.G., Lee, D., Ashford, P., Orengo, C.A. and Sillitoe, I. (2017) CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res*, **45**, D289-D295.
80. Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D. and Petrov, A.I. (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res*, **46**, D335-D342.
81. Kozomara, A., Birgaoanu, M. and Griffiths-Jones, S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res*, **47**, D155-D162.
82. Athar, A., Fullgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., Snow, C., Fonseca, N.A., Petryszak, R., Papatheodorou, I. *et al.* (2019) ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res*, **47**, D711-D715.
83. Griss, J., Perez-Riverol, Y., Lewis, S., Tabb, D.L., Dianes, J.A., Del-Toro, N., Rurik, M., Walzer, M.W., Kohlbacher, O., Hermjakob, H. *et al.* (2016) Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat Methods*, **13**, 651-656.
84. Perez-Riverol, Y., Uszkoreit, J., Sanchez, A., Ternent, T., Del Toro, N., Hermjakob, H., Vizcaino, J.A. and Wang, R. (2015) ms-data-core-api: an open-source, metadata-oriented library for computational proteomics. *Bioinformatics*, **31**, 2903-2905.
85. Uszkoreit, J., Maerkens, A., Perez-Riverol, Y., Meyer, H.E., Marcus, K., Stephan, C., Kohlbacher, O. and Eisenacher, M. (2015) PIA: An Intuitive Protein Inference Engine with a Web-Based User Interface. *J Proteome Res*, **14**, 2988-2997.

86. Bastian, F., Parmentier, G., Roux, J., Moretti, S., Laudet, V. and Robinson-Rechavi, M. (2008), *International Workshop on Data Integration in the Life Sciences*. Springer, pp. 124-131.
87. Euzenat, J. and Shvaiko, P. (2007) *Ontology matching*. Springer.
88. Karsch-Mizrachi, I., Takagi, T., Cochrane, G. and International Nucleotide Sequence Database, C. (2018) The international nucleotide sequence database collaboration. *Nucleic Acids Res*, **46**, D48-D51.
89. Giraldo-Calderon, G.I., Emrich, S.J., MacCallum, R.M., Maslen, G., Dialynas, E., Topalis, P., Ho, N., Gesing, S., VectorBase, C., Madey, G. *et al.* (2015) VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res*, **43**, D707-713.
90. Howe, K.L., Bolt, B.J., Shafie, M., Kersey, P. and Berriman, M. (2017) WormBase ParaSite - a comprehensive resource for helminth genomics. *Mol Biochem Parasitol*, **215**, 2-10.
91. Mason, C.E., Afshinnekoo, E., Tighe, S., Wu, S. and Levy, S. (2017) International Standards for Genomes, Transcriptomes, and Metagenomes. *J Biomol Tech*, **28**, 8-18.
92. Andrews, S., FASTQC. A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, *Accessed* 15 April 2020.
93. Lo, C.C. and Chain, P.S. (2014) Rapid evaluation and quality control of next generation sequencing data with FaQCs. *BMC Bioinformatics*, **15**, 366.
94. Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114-2120.
95. FASTX-Toolkit., http://hannonlab.cshl.edu/fastx_toolkit/, *Accessed* 15 April 2020.
96. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15-21.
97. Bray, N.L., Pimentel, H., Melsted, P. and Pachter, L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*, **34**, 525-527.
98. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. and Kingsford, C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*, **14**, 417-419.
99. Costa-Silva, J., Domingues, D. and Lopes, F.M. (2017) RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS One*, **12**, e0190152.
100. Van Den Berge, K., Hembach, K.M., Soneson, C., Tiberi, S., Clement, L., Love, M.I., Patro, R. and Robinson, M.D. (2019) RNA sequencing data: hitchhiker's guide to expression analysis.
101. Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol*, **11**, R106.
102. Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, **15**, 550.
103. Taylor, C.F., Hermjakob, H., Julian, R.K., Jr., Garavelli, J.S., Aebersold, R. and Apweiler, R. (2006) The work of the Human Proteome Organisation's Proteomics Standards Initiative (HUPO PSI). *OMICS*, **10**, 145-151.
104. Codd, E. (1971) A Relational Database Model for Large Shared Data Banks. *Communications of the ACM*, **13**, 6.
105. Klyne, G. (2004) Resource description framework (RDF): Concepts and abstract syntax. *http://www. w3. org/TR/2004/REC-rdf-concepts-20040210/*.

106. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, **3**, 160018.
107. Garcia, L., Bolleman, J., Gehant, S., Redaschi, N., Martin, M. and UniProt, C. (2019) FAIR adoption, assessment and challenges at UniProt. *Sci Data*, **6**, 175.

Chapter 2
General Methods

Data modelling in ProteomicsDB

Contents

1 Methods and implementation

1.1 Data model extension

1.1.1 Multi-omics data model

The new Multi-omics data model is able to organize existing and future quantitative omics expression data in five main tables: *OmicsProject*, *OmicsExperiment*, *OmicsSample*, *MeasurementPlatform* and *OmicsExpression* (Figure 2.1). Starting from the last one to the first, each table uses a foreign key to the next table, reducing records duplication and keeping

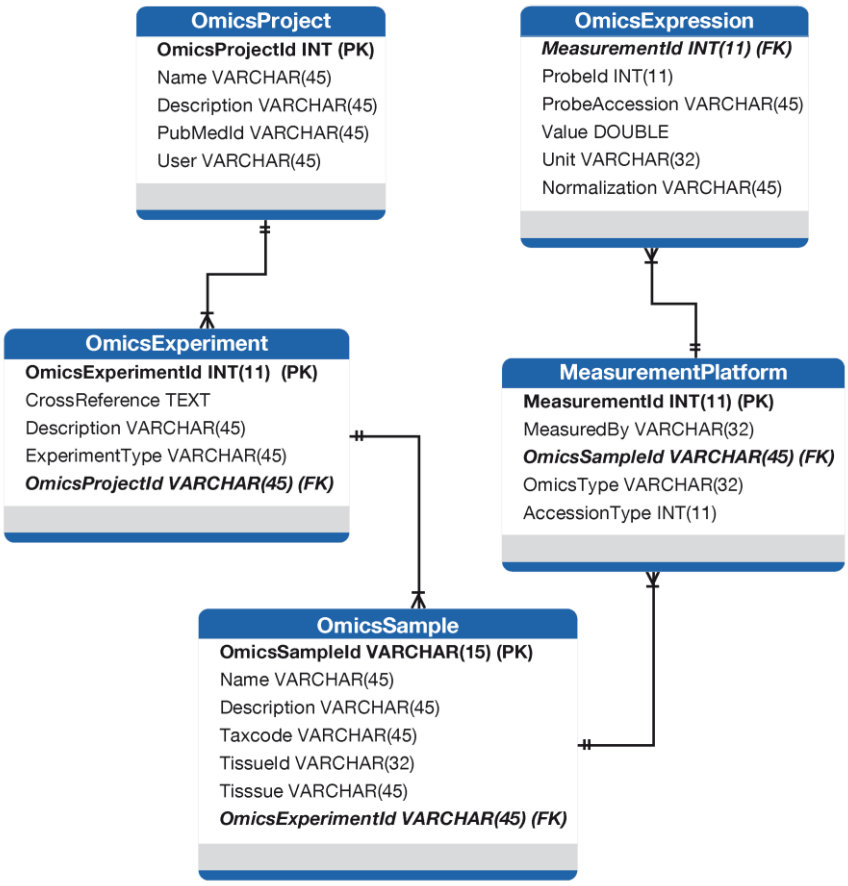

Figure 2.1 Multi-omics Data Model. A generic design, capable of storing any kind of quantitative omics expression data, organized per sample and experiment. The five boxes describe the structure of every table of the data model. Each attribute of the table is defined by the name (e.g. OmicsProjectId), the data type (e.g. INT for Integer) and in case of a string data type, the length of it in a parenthesis (e.g. VARCHAR(45)). The tables are connected by using homonymous identifiers as Primary Keys (PK) and Foreign Keys (FK).

information connected and organized.

The database model shown in Figure 2.1, is designed in a way that a project could contain one or more experiments, while an experiment can also be composed of multiple samples. The model supports the measurement of a sample on different platforms. The *OmicsExpression* table stores the expression values of the quantified entities (in this context, often genes) in the specified sample measurements. The critical part in the *OmicsExpression* table is that it can contain any kind of omics' measurements as long as they can be expressed by a numeric value (e.g. gene

expression, methylation level, etc.). The existence of the "Unit" field also allows the description of the numeric value of each record (e.g. TPM values of RNA-Seq data, iBAQ values of proteomic experiments, etc.).
The datasets that were downloaded already contain an internal probe id, which is stored in the field "ProbeId", along with the provided probe's accession number in the field "ProbeAccession". The first one is necessary to map back against the original dataset, while the accession number is essential to map a probe to a UniProt accession number and associate it to other omics data. The *MeasurementPlatform* table describes each measurement. It stores information about the used platform, the resource type of the probe's accession number (e.g. Unigene, Ensemble Gene Id, etc.), the type of omics data being studied in this measurement and the sample that was used for this measurement. Unigene and Ensemble are standard databases in the field and similar to UniProt but for nucleotide instead of protein sequences. The *OmicsSample* table describes each sample by its name, species (taxonomy code) and the tissue or cell line of origin. Finally, all samples map to an experiment and a project.

1.1.2 Cell Viability data model

In order to be able to take advantage of the plethora of drug sensitivity information available in the public domain, a data model with elements of the Resource Description Framework (RDF) (1) was generated, which will be used to store drug sensitivity datasets in ProteomicsDB (Figure 2.2). Dose-response experiments or *DoseResponses* can be grouped into *DoseResponseDataSets*, which

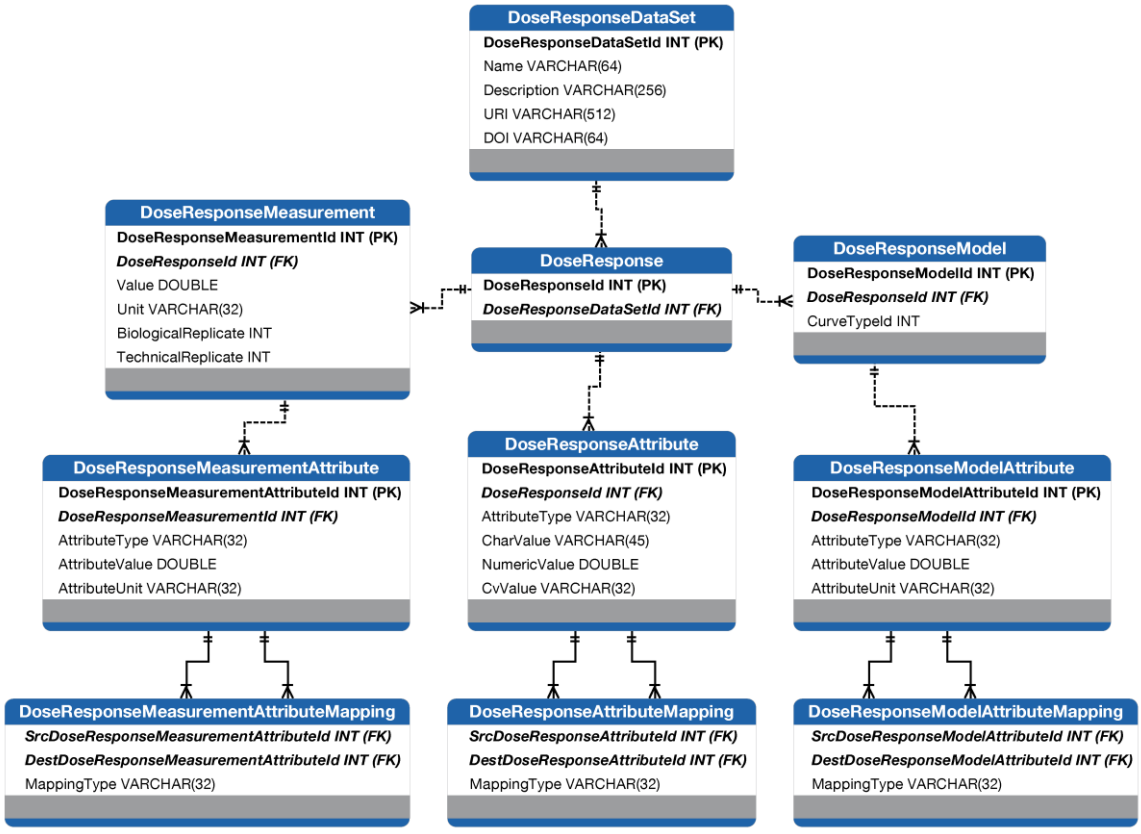

Figure 2.2 The Cell Viability Data Model. It is organized in 3 main branches. The Measurement branch (left) where raw and processed measurements are stored, the fitted model branch (right) where the model attributes and fitted parameter are stored and the main dose response branch (middle) that stores the cell lines and compounds that were used for a specific cell viability experiment

are annotated with meta-data such as a description, URI, DOI, etc. Each *DoseResponse* itself may contain raw data and/or one or multiple dose-response models with associated parameters like IC50, hill slope, etc. Therefore, the data model needed to be flexible enough to support storing raw viability data with their corresponding annotations alongside dose-response models with their corresponding parameters.

On the one hand (left column of Figure 2.2), each *DoseResponse* consists of a set of *DoseResponseMeasurements* (i.e. raw viability data) with an associated Unit (stored in the controlled vocabulary of ProteomicsDB), which can be measured in biological and/or technical replicates. Each *DoseResponseMeasurement* itself has associated meta-data like the dose resulting in this *DoseResponseMeasurement*, the Controlled Vocabulary Identifier (CVid) of the particular drug and/or a normalized response, among others. These *DoseResponseMeasurementAttributes* are stored as double with an associated type and unit, stored as varchar in the controlled vocabulary of ProteomicsDB. Mappings between these different attributes are stored using the RDF, which supports storing an arbitrary number of *DoseResponseMeasurementAttributes* together with each *DoseResponseMeasurement*. This allows storing multiple drugs as attributes of a single data point in the case of co-inhibition experiments. On the other hand (right column of Figure 2.2), each *DoseResponse* can have one or more *DoseResponseModels* associated with it. Each of these *DoseResponseModels* (supplied by the authors or fitted using the developed pipelines) has several *DoseResponseModelAttributes* associated with it, which may be different model parameters, lower and upper limit of confidence intervals of these or their standard errors, as well as the CVids of model names with their associated formulas. Again, these *DoseResponseModelAttributes* are stored as a double, with an associated type and unit, stored as varchar in the controlled vocabulary of ProteomicsDB. The RDF is again used to store mappings between these *DoseResponseModelAttributes*. In addition to this data, each *DoseResponse* is stored together with *DoseResponseAttributes* (middle column in Figure 2.2) in order to be able to associate *DoseResponses* with cell lines or assay types. Using *DoseResponseAttributeMappings* (RDF), the same *DoseResponse* can then be associated with multiple cell lines in the case of, e.g. co-culture experiments.

1.1.3 Multi organism implementation

The new data models, along with some existing tables contained by design a field to store the taxonomy code (taxcode) of the species. All organism-specific tables had to be extended by one column to store the corresponding taxcode. That alone did not allow the species separation and visualization in the web interface yet. All appropriate procedures and calculation views were adjusted to either report the result grouped by the taxcode, or require taxcode as an input parameter and filter the required data or table joins by that taxcode. Finally, the backend end endpoints and the frontend calls include now the taxcode as a parameter wherever needed. This required extensive testing before and after the applied changes so that the results would remain the same.

1.2 Resource Identifier Mapping data model

Gene and protein sequences are deposited in well-established repositories such as Ensembl, Unigene and UniProt. These sequences are subject to curation and can, therefore, change between releases. The current model of ProteomicsDB uses static tables to map between different resources' identifiers like Ensembl, Unigene and UniProt. For each new identifier added, an

additional table is necessary to enable their mapping to other identifiers. This also requires the modification of existing procedures every time sequences should be updated. The goal here was to design a new generic data model, which overcomes the shortcomings of the current one and extends its usability to, e.g. incorporate additional semantic information. The same data model will later be extended to support any kind of relations between identifiers. In this chapter, a step by step implementations and extension is described for a better understanding of the applications and the benefits that this model has to offer.

1.2.1 Identifier mapping

The Resource Id Mapping data model (blue part of Figure 2.3) is based on the simple triplet principle <*Source*><*Property*><*Destination*> similar to the RDF framework (1). Instead of modelling the relation between different identifiers with multiple tables, all relations are stored in a triple store structure. This provides a generic and easy-to-use interface to map different entities to each other and enables the structured storage of pathway information (e.g. metabolites involved in certain processes) and the relation between different entities (e.g. drugs mapping to their designated targets). This model will not only provide a unified interface to resource mapping (between identifiers from different omics fields), but it will also serve as an interface to store relations between drug response and biomarkers. Here, *Source* and *Destination* are resource identifiers, whereas the *Property* field describes the relation between different IDs, e.g. “maps _to” or “originates_ from”. To avoid repetitions of large strings and to be able to use the different resource types as foreign keys in other tables, our data model consists of the following tables:

- Resource Type, e.g. UniProt, Ensembl
- Accession, e.g. P00533, ENSG00000143545
- AccessionProperty
- RelationType, e.g. maps_to
- AccessionIdRelation

The table ResourceType stores all available resource types. The table *Accession* stores all related accessions (identifiers) of a resource. Any number of properties associated with an accession number (e.g. version of database or alternative names) is stored in the AccessionProperty table. The table RelationType is used to describe relations between different accessions. Finally, the AccessionIdMapping table stores the corresponding internal IDs in the triplet model <*Source*><*Property*><*Destination*>, but this time as <*SourceId*><*RelationId*><*DestinationId*>. Another advantage is that if a relation is transitive, it is easy to formulate mappings that are not yet stored:

1. <ENSG00000146648> <*maps_to*> <Hs.488293> : **stored**
2. <Hs.488293> <*maps_to*> <P00533> : **stored**
3. <ENSG00000146648> <*maps_to*> <P00533> : **created**

1.2.2 Extension for all relations

As described above, the initial implementation of this model allowed the storage of any kind of relations. As a first step, protein-protein interactions, downloaded and reprocessed from STRING (2), were imported. This included several new RelationType entries and many more AccessionIdMapping entries. At this point, the table AccessionIdMapping was renamed to AccessionIdRelation. Protein-protein interactions, as well as id relations, can be visualized as a graph, where nodes being the identifiers or Accessions and edges being the relations, each entry

in the AccessionIdRelation table is considered an edge. In graphs, edges have properties, for example, weight. In our case interactions have different degrees of certainty as well as scores. For that reason, a new table called EdgeProperty was introduced to store different properties of the stored relations. The second dataset that was downloaded and imported into the Identifier Relation Model is pathway data from KEGG (3). Even with the initial STRING dataset, it was important to create a hierarchy structure to allow search on different levels of this hierarchy. This led to the creation of the RelationIdRelation table, where relations are connected to each other with parent-child relations. The hierarchy is also used to separate the relational datasets, by splitting them into different root-branches or sub-trees (red part of Figure 2.3).

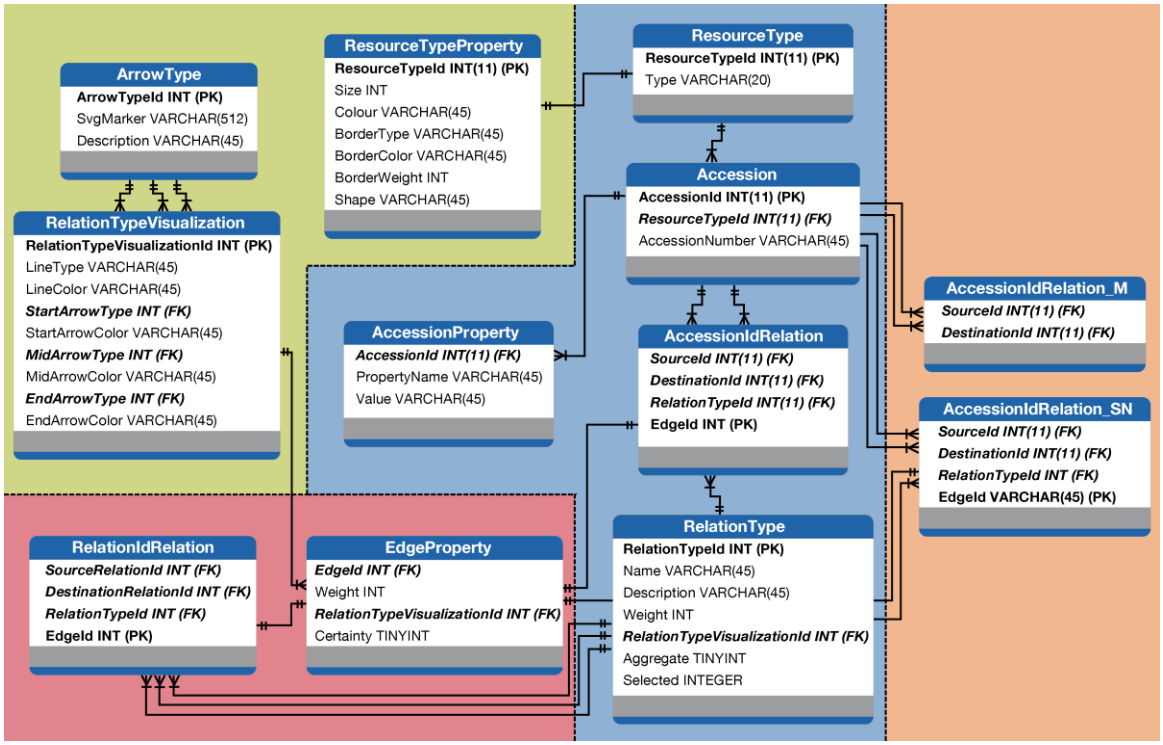

Figure 2.3 The Resource Identifier Relation Data Model. The blue area depicts the initial implementation of the Resource Identifier Mapping data model. It was expanded to store any kind of relations by also allowing the organization of relation in a hierarchy (red area). The model was further expanded with the metadata and visualization part (green area), to enable the visualization of the stored relations between identifiers as a graph. Finally, to minimize querying and retrieval times, a new indexing schema was applied on top of the existing graph, the so-called SuperNodes (orange area).

1.2.3 Metadata

Both the developed hierarchy and the graph that is stored in the above data model needed specific visualizations. Protein interactions are usually visualized with specific symbols and colours. For example, activation is a green arrow marker pointing to the direction the activation is happening while inhibition is a red vertical bar marker close to the inhibited protein. Here we store different types of relations, so we respect the already existing visualizations while adding more for the not specified ones. For all general type of relations like co-expression or relations regarding grouping to a pathway, a blue diamond marker in the middle of the edge was introduced as these are non-directional properties. Assigning in advance marker types, colours and positions on an edge to relations allowed the setup of a generic visualization model storing visualization types. Each

visualization type gets a unique id that gets assigned to the corresponding edges as an edge property. It is crucial to be able to store visualizations that way so that even relations with predefined visualizations can have additional options. As for edges, the same for resource types and accessions had to be designed. Here each resource type gets an entry in the ResourceTypeProperty table storing the shape, colour, etc. of the visualized node. Each accession that belongs to the corresponding resource type inherits this visualization. The new tables that were introduced come to complement the ResourceIdRelation data model (green part of Figure 2.3).

1.2.4 SuperNodes

The implemented data model accompanied by SQL views and procedures was already capable of storing any kind of relations between identifiers. With the storage and retrieval challenges sorted out, query time was raised as an issue.

All entries in the ResourceIdRelation data model can be visualized as a huge graph. In this graph, some areas are sparsely connected, while others are dense. Some areas can also be grouped in one node, such as groups of nodes that correspond to different identifiers from different resources but describe one entity. An example of such nodes is nodes corresponding to the same gene (Figure 2.4ab). The data model is able already to group these nodes by gene, but doing so on the fly is time-consuming and introduces a significant overhead on requests done from the frontend. The absence of a graph engine capable of creating these subgraphs, lead to the creation of several extra tables that would be used as indices to the new subgraphs, so-called SuperNodes (Figure 2.4cd). SuperNodes representing genes will be described here, while the same procedure is applied for every other entity is needed to be saved as a SuperNode.

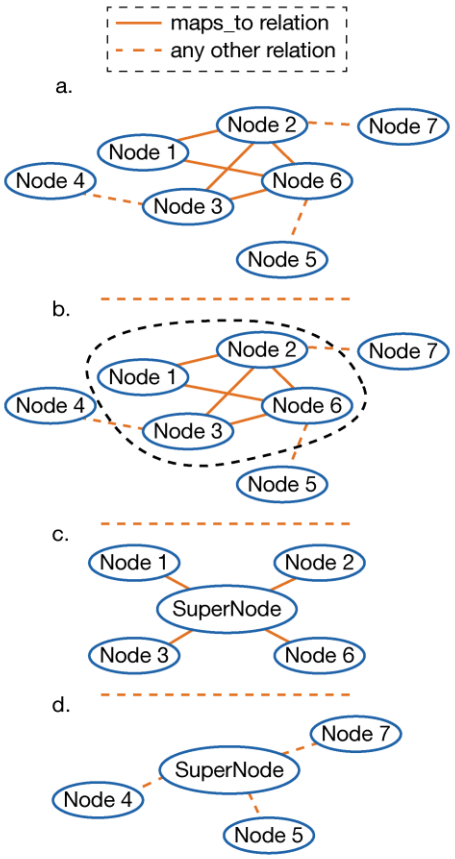

Figure 2.4 SuperNodes creation procedure. Starting from an initial graph (a), nodes that are all connected with the same type of edge are detected (b - circled). The circled nodes are all connected to a new SuperNode (c) that replaces the previous nodes in the initial graph (d).

First, SuperNodes are introduced as a new resource type, so that the default schema is respected. For every gene name that data is stored in ProteomicsDB, a new accession entry is created associated with the resource type SuperNode. Then, in order to keep the existing model clean of new indices that would pollute the semantics of the existing tables, a duplicate table of the AccessionIdRelation was added. The AccessionIdRelation_M table stores all mappings of a SuperNode accession id to the accession ids that correspond to that gene (orange part of Figure 2.3). This allows aggregating any relation info up to the gene level. However, this kind of grouping introduces many edges/relations between two SuperNodes, either replicated edges or edges with different weights. In that case, we select the edge with the highest score/weight. The representative edge is stored in a final table named AccessionIdRelation_SN, as it stores relations between SuperNodes. The edge Id is kept the same as the original edge though so that all edge properties are inherited (orange part of Figure 2.3).

This way of storing data, in the form of triples, is efficient for databases as well as for defining interfaces between procedures and function calls but is extremely complicated for humans. To make data interpretation easier for users but also exploitable in a meaningful way via the ProteomicsDB API, new generic views were implemented as part of the repository. One view projects the data of a node or SuperNode of the aforementioned graph, collecting all the relevant records from the ResourceType, Accession and AccessionProperty tables. A second view projects all edges of the graph with their associated metadata and properties. This view though includes

for every edge not only the source and destination AccessionIds but their full description from the Accession and ResourceType tables. AccessionProperty entries were not relevant at this point as they do not define or describe edges or relations. In case of need of such information, a simple left join on based on the AccessionId is enough. The purpose of this view is to provide relation to or from all relevant AccessionIds, either by querying the AccessionNumber or the AccessionId. Another use-case of this view is the retrieval of all nodes that are linked to each other using a specific set of RelationTypes. Both views perform no filtering on the provided data, as there filtering criteria differ per use-case.
The advantages of this implementation, as well as the results of it, are described in the first manuscript of the Appendix.

1.3 Custom User Data data model

The Custom User Data data model was implemented in order to allow temporary storage of custom user expression data. Data are organized in three tables: *UUID*, *UserDataset* and *UserExpressionData* (Figure 2.5). The table *UUID* stores unique user/session identifiers and the last time a session was accessed. Each user can upload multiple datasets. Expression data of a dataset are stored in the *UserExpressionData* table and then associated with an entry in the *UserDataset* including the dataset name and omics type it includes as well as the creation date. Each uploaded dataset should contain the expression value of a gene in a tissue of origin, accompanied by the quantification and calculation method that was used and the taxonomy code the sample originates from. All tables are interconnected with foreign keys. A deletion in the *UUID* table will cascade to the rest of the tables deleting the relevant data entries. Data here are stored only temporarily while no personal user information will be stored. SAP HANA offers the functionality of temporary tables, but this is not useful here, as in temporary tables, data is stored only within the time-period of an active session. To allow storing for predefined periods, scheduled jobs were designed to remove UUID entries that their access date is no older than the number of days defined in the corresponding procedure. To enable fast retrieval of a single dataset, an xsodata object was created serving a view on the underlying data, enforcing at the same time filtering by DatasetId and UUID.

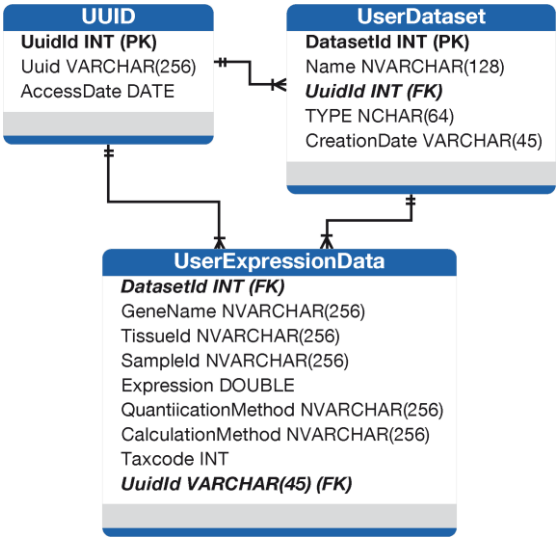

Figure 2.5 The Custom User Upload data model.

1.4 Elastic Net Models data model and prediction procedures

Training of the elastic net models is described in the second manuscript in the Appendix. In short, model training was implemented in R, while hdbprocedures were created for the filtering and preparation of the data. Since the model-fitting procedure needs to run only once for each drug in each drug sensitivity dataset, there is no need for a user-facing endpoint focused on model training. Instead, model training is initiated only upon the addition of new drug sensitivity datasets to ProteomicsDB. Once all models are fit, they are stored in HANA. The data model storing the trained models (Figure 2.6) consists of the following tables: *ModelParameters*, *Coefficients*,

Intercepts and *Effects*. *ModelParameters* stores, as described by the table's name, each separate model and its training parameters, such as the drug and dataset that it was trained on, the omics type of the expression data that was used during training, the number of bootstraps and cross-validation folds as well as the final number of selected predictors/gene names of a model. More parameters are stored here like the *alpha* and *lambda* parameters for the model fitting. After model fitting, the bootstrap coefficients and intercepts are stored in corresponding tables. The *Effects* table stores the corresponding mean effect sizes and selection frequencies of each predictor/gene name of the model.

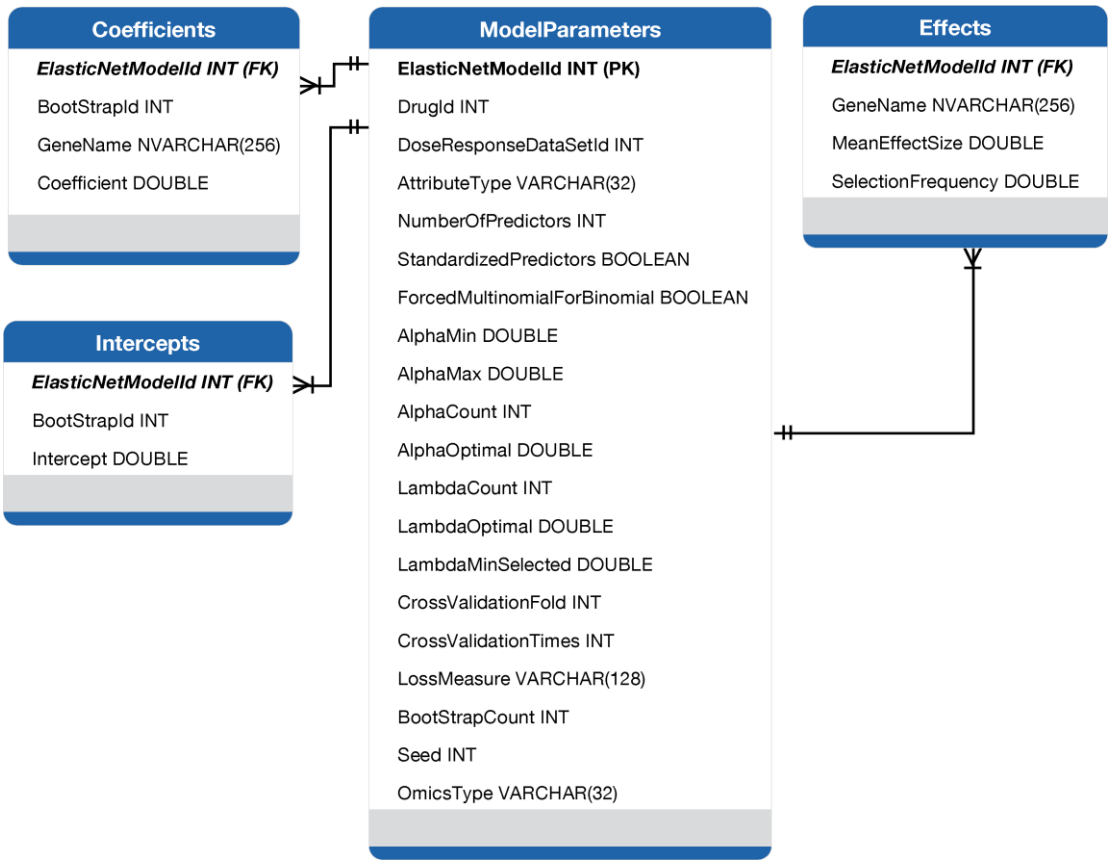

Figure 2.6 The Elastic Net Models data model.

Based on these bootstrap coefficients and intercepts, a graphical calculation view is now used to predict drug sensitivity. This allows the assessment of the variability of the predictions, which was incorporated into the newly developed visualization capabilities of ProteomicsDB (see paper #2 in the Appendix). One advantage here is that elastic net models are a type of generalized linear model, which allows for fast prediction on new data using simple linear algebra operations. In other words, the prediction can be broken down into a series of joins and simple mathematical operations, which makes the prediction of drug sensitivity on new data extremely fast. In contrast to model training, drug sensitivity prediction happens on user request. Users have the opportunity to predict the drug sensitivity of samples (e.g. cell lines) already present in ProteomicsDB which the model has not yet seen, or alternatively upload their own data and predict the drug sensitivity of their model systems.

1.5 Data processing and integration procedures

1.5.1 Intra- and inter-omics data normalization procedure

SAP HANA offers adapters for the direct connection of the database to an R server. That allows the implementation of data processing and transformation procedures that do not exist in the database and its extension packages, so-called RLANG procedures.

A procedure that allows the further omics data comparison is the *omicsMComBatNormalization*. Given two datasets, the first being the reference dataset and the second the one to be normalized, the procedure executes the following steps:

1. Filters the datasets for common genes or proteins,
2. Filters out genes or proteins that show zero variance across all samples of a dataset,
3. Applies the MComBat adjustment to the second dataset and
4. Returns the normalized dataset, as a table, to HANA for further processing in SQL

A detailed review of the literature and a description of the logic around the MComBat and batch effect correction is part of the second manuscript of the Appendix. To enable user-uploaded data comparison to ProteomicsDB stored data, this method is available as a normalization option in the frontend. Users can choose this method to make meaningful comparisons.

1.5.2 The missing value imputation procedure

Taking advantage of the afore-mentioned omics data normalization method, an mRNA-guided missing imputation method described in (4) was implemented. Following their implementation and example, at first, all full- and deep-proteome datasets, stored in ProteomicsDB, were normalized against each other with the aforementioned method to provide a reference dataset. As a second step, all transcriptomic datasets in ProteomicsDB were also normalized against each other. Then, the normalized transcriptomics data were MComBat adjusted based on the proteomics reference dataset. At last, a linear model was fit on the two normalized omics types, leading to a linear equation that will later be used for estimation of missing omics expression values. This feature is implemented and added to the backend endpoint and the frontend functionality of the already existing analysis tool, the interactive expression heatmap.

1.5.3 Drug enrichment analysis procedure

The following method is a part of the Pia Bothe's internship and Master Thesis with the title, "Development of a human tissue ontology based on multi-omics expression patterns to investigate drug responses".

This functionality includes the following steps, each one wrapped in its own procedure:

1. Collection of all full and deep proteomes in ProteomicsDB. There are several filters applied here that need to be provided as input parameters, for example, the minimum number of proteins in a dataset, if only kinases should be considered, etc.
2. Collection of all cell sensitivity data in ProteomicsDB. Here, only cell lines for which we have proteomics expression data are considered. A few filters are also applied here, for example, the minimum number of cell lines screened per drug, etc.
3. Transformation of each drug in step 2 into an equal length vector representation, where each dimension is a cell line. Each dimension is a binary attribute, where 0 represents no effect of the drug to the cell line, while a 1 represents effect of the drug on the cell line. A cell line is considered sensitive to a drug when it fulfils the input based criteria, for example, EC50 less than an input value, or Relative inhibition effect higher than an input

value. This one-hot matrix (5) allows further comparisons between drugs based on their Jaccard similarity (6), as each drug represents a set of strongly affected cell lines.

4. Drug enrichment analysis, using an R procedure, with user-defined of foreground and background selection of cell lines based on their similarity.
5. Multiple test correction, using an R procedure and the Benjamini-Hochberg procedure that provides a q-value per drug.
6. Final projection of a list of drugs, ordered by their q-value in ascending order, provided a cell line or tissue name on which the drug enrichment is performed.

1.5.4 Elastic net regression procedure

Another procedure written in R language that utilizes the package ('glmnet', version 2.0-18) is the *fitElasticNet* procedure. The procedure requires as input parameters:

- An expression table, consisting of 3 columns: GeneName, CellLine and Expression value,
- A drug sensitivity table, consisting of 5 columns: Drug, Cell viability dataset, Cell line, Model attribute type and model attribute value,
- The omics type of the expression data and
- The number of bootstraps, *X*, for the model fitting.

The procedure applies, as a first step, missing value imputation on the expression matrix. Then it fits *X* bootstrap elastic net models, extracts the coefficients and intercepts of each bootstrap and reshapes the output to the corresponding tables (Figure 2.6).

1.6 Frontend adjustments

All created visualization and analysis tools that were developed in the scope of this thesis, are described in the two manuscripts. Here, the logic behind the interaction graph as well as the frontend limitations are described.

1.6.1 Interaction network

The Resource Identifier Relation data model enabled the storing of any kind of relation between different identifiers from different resources. As described in the introduction, protein-protein interaction, along with pathway information, are stored in the same data model. Using the SuperNodes technique and the relevant views, information stored in the data model can be visualized as a graph where nodes are SuperNodes (e.g. Genes or Pathways) and edges are relations between them (e.g. 'activation', 'inhibition', 'belongs to'). The user interface was designed in a protein-centric way. The 'Protein Details' webpage was expanded with a new tab, the 'Interaction network'. As the webpage is focused on a protein of interest, same with the 'Interaction network' tab, any kind of selection begins from that protein. The view is divided into two vertical panels.

The left panel contains three tabs: the Relation tab, the Node information tab and the Options tab. On initialization of the webpage, only the Relation tab is visible. It contains all settings regarding the data retrieval query. More specifically, it contains every relation group and relationships so that the user can preselect the allowed relations in the resulting graph. Upon query result retrieval and rendering of the graph, the Options tab appears. Here the user can show or hide different groups of nodes, such as nodes that represent Genes or Pathways. This menu also offers the option to download the visible graph in three different formats: SVG and PNG for vector graphics and images and in SIF format that is directly importable in Cytoscape (7). The Nodes information tab appears upon selection of one or more nodes. It includes small accordion

menus one for each node and includes information about the UniProt Accession or the String Identifier that the specific node uses in the relations of the graph. For each node there are small buttons that redirect the user to the protein information of that node as well as any other tab of the 'Protein Details' webpage, allowing even the initialization of the graph using a different protein of interest. Finally, in the case that multiple nodes are selected, two more buttons are enabled redirecting to the Interactive Expression Heatmap or the Combination Treatment analysis tools.
The right panel is the graph visualizer, where the queried nodes and edges are displayed. The graph is implemented using the D3js library (version 4) (8) and is a force-directed graph. Nodes have specific weights and charges so that they do not overlap and edges have specific force ranges so that the distance between nodes stays constant, and nodes do not keep increasing their in-between distance forever. When the final positions of the nodes are set by the D3 simulation environment, the simulation is stopped to allow reorganization of the graph. Nodes are drag-able objects. On click and drag of a node, the node changes its position in the graph causing the rest of the nodes to changes their position accordingly as their charges stay the same and the force between them does not allow overlaps. On the release of a node, its current position becomes a permanent position, meaning that this node is locked in space and any alteration in the forces in the graph is not affecting this node. Every node can be locked into position separately. Clicking and dragging in the empty area of the graph allows reposition of the whole graph. Scrolling up and down enabled zooming in and out the appointed area, respectively. After rendering a graph, four in-frame buttons are enabled. The first button is a plus sign ('+') that triggers the expansion of the graph on the selected node with the next five neighbours in every relation group that is selected. The 'bucket' button removes a node and the associated edges from the graph. The 'open-lock' button unlocks again the selected node allowing the calculation of its position by the simulation environment of D3. Deletion and unlocking of multiple nodes are allowed if the nodes are selected by holding down the Shift button on the keyboard and clicking on each node separately. Expanding a node is a single node functionality. Selection of a node is not only locking that node into position but also highlights the selected node by changing its colour to orange and changes the left panel to the Node Information tab.

1.6.2 Frontend framework limitations

As of right now, SAPUI5 is the framework to create user interfaces for ProteomicsDB, but the development of SAPUI5 is stalled, very convenient features of modern frameworks are missing, and the development process is very slow. Moreover, the SAPUI5 theme that ProteomicsDB is developed on, named 'sap_goldreflection', is deprecated and consequent updates of SAPUI5 do not apply for this theme. Upgrading to a newer SAPUI5 theme requires a full re-implementation of the frontend nowadays. Even in that case, the structure of SAPUI5 is limiting the development of new tools as they should comply with specific standards that are not always easy to follow, increasing that way significantly the development time. Therefore, an evaluation of other new frameworks was needed, to future-proof ProteomicsDB and make it more accessible for developers in the future. After comparing React, AngularJS and Vue.js, Vue.js was selected.
Direct integration of SAPUI5 and Vue.js is not possible. However, SAPUI5 allows deployment of pure HTML elements in their XML-like format. On the other side, in Vue.js, when a build is triggered for production usage, it produces two JavaScript files containing the main application and the declaration of the functions and further scripts. It further produces a CSS file regarding the styling of the resulting webpage and an HTML file, called index.html, which loads the JavaScript

code from the two previous files. Placing the content of the index.html file to the HTML element of SAPUI5 allows deploying Vue.js application in the scope of SAPUI5. Further exploration is needed as CSS styles might conflict, and manual adjustments are needed in that case.

2 Abbreviations

API	Application programming interface
CSS	Cascading Style Sheets
HTML	Hypertext markup language
PNG	Portable Network Graphics
RDF	Resource Description Framework
RNA	Ribonucleic acid
SIF	Simple interaction file
SQL	Simple query language
SVG	Scalable Vector Graphics
TPM	Transcripts per Million
URI	Universal resource identifier
XML	Extensible Markup Language

3 References

1. Klyne, G. (2004) Resource description framework (RDF): Concepts and abstract syntax. *http://www. w3. org/TR/2004/REC-rdf-concepts-20040210/*.
2. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P. *et al.* (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*, **47**, D607-D613.
3. Kanehisa, M. and Sato, Y. (2020) KEGG Mapper for inferring cellular functions from protein sequences. *Protein Sci*, **29**, 28-35.
4. Frejno, M., Zenezini Chiozzi, R., Wilhelm, M., Koch, H., Zheng, R., Klaeger, S., Ruprecht, B., Meng, C., Kramer, K., Jarzab, A. *et al.* (2017) Pharmacoproteomic characterisation of human colon and rectal cancer. *Mol Syst Biol*, **13**, 951.
5. Harris, D. and Harris, S. (2010) *Digital design and computer architecture*. Morgan Kaufmann.
6. Jaccard, P. (1912) The distribution of the flora in the alpine zone. 1. *New phytologist*, **11**, 37-50.
7. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, **13**, 2498-2504.
8. Teller, S. (2013) *Data Visualization with d3. js*. Packt Publishing Ltd.

Chapter 3
Publication 1

ProteomicsDB

Citation

The following article titled “ProteomicsDB” has been published in Nucleic Acids Research, Database Issue, on November 02, 2017.

Full citation:
Tobias Schmidt*, Patroklos Samaras*, Martin Frejno, Siegfried Gessulat, Maximilian Barnert, Harald Kienegger, Helmut Krcmar, Judith Schlegl, Hans-Christian Ehrlich, Stephan Aiche, Bernhard Kuster, Mathias Wilhelm, ProteomicsDB, ***Nucleic Acids Research***, Volume 46, Issue D1, 4 January 2018, Pages D1271–D1281, https://doi.org/10.1093/nar/gkx1029

Summary

Mass spectrometry has become the lead technology for proteome research. The produced proteomics data is of high volume and complexity, rendering hypothesis generation a difficult task, especially when one needs to compare results across several experiments or studies. Different repositories have been developed in the last years, being focused though in certain aspects of the data. More often, existing platforms do not bind together peptide and protein identifications with quantification information. ProteomicsDB is a unique resource that stores quantitative mass spectrometry-based proteomics data of human origin. Initially introduced in 2014, ProteomicsDB is capable of organizing and storing identification and quantification data along with metadata regarding the experimental design, such as sample treatment and preparation protocols as well as data acquisition parameters. Proteomics data are organized in samples, experiments and projects, where a project represents a publically available. The protein-centric interface of the platform allows the in-depth investigation of a protein of interest (POI). The existing data model and real-time normalization methods enable cross-dataset comparisons of protein abundance, which can be visualized in the human bodymap, providing at the same time information about the relative abundance of a protein across human tissues. Dose- and temperature-dependent assay data are also available for most of the proteins. Protein-protein interactions and pathway information are offered in an interactive and extendable graph. Validation of identifications is also an option in the mirrored spectrum viewer, where experimentally acquired spectra can be compared to reference spectra originating from synthetic peptides. Finally, proteotypic peptides can be explored for the creation of targeted assays. Another benefit of this platform is the online analysis toolbox that it offers. Multiple proteins’ expression values can be compared in the interactive expression heatmap. Different kinase inhibitors can be compared with each other based on their selectivity against a protein of interest. Combining the dose-dependent assay data in a drug-target interaction network enables the online exploration and design of combination treatments. Different drugs can be explored regarding their effectiveness against multiple cell lines, in the new Cell Viability analysis tool. Finally, the data model of ProteomicsDB was extended to support any kind of quantitative omics expression data. The new data model was populated with MicroArray and RNAseq data from NCBI GEO, ArrayExpress and the Human Protein Atlas. The new omics expression data are available for visualization in the human bodymap and the interactive expression heatmap. A newly developed resource identifier mapping system allows the direct conversion between resource types, allowing consequentially the comparison and analysis of omics data from different resources using

different gene or protein identifiers. The generalized data models, the inter-connected analysis tools and the data diversity render ProteomicsDB a unique resource for the scientific and medical community.

Author contributions

The author of this dissertation had a leading role in the implementation of the work presented. The author was the lead developer regarding the design and implementation of the omics data model and the Resource Identifier Relation data model as well as the frontend adjustments for the support and visualization of the new omics data. The author together with M. Frejno, designed and implemented together the Cell viability data model, while M. Frejno was solely responsible for the retrieval of cell viability raw data and their initial reprocessing. M. Frejno developed the initial pipeline for the acquisition of the cell viability data and the curve fitting, while the author together with T. Schmidt extended the pipeline for the automatic mapping to controlled vocabularies and data transformation and formatting for direct import into the database. The author was involved together with T. Schmidt in the design of the cell viability user interface and backend endpoints. The user interface for the cell viability data as well as the protein-protein interaction graph was implemented by T. Schmidt while the author implemented the database SQL views and procedures as well as the backend endpoints for the two tools. The execution and implementation of the tools presented were performed under the continuous supervision of M. Wilhelm and the doctoral supervisor Prof. Bernhard Kuster. The manuscript was drafted by T. Schmidt and finalized by T. Schmidt, M. Frejno, M. Wilhelm and the author.

Rights and permissions

Chapter 4
Publication 2

ProteomicsDB: a multi-omics and multi-organism resource for life science research

Citation

The following article titled “ProteomicsDB: a multi-omics and multi-organism resource for life science research” has been published in Nucleic Acids Research, Database Issue, on October 30, 2019.

Full citation:
Patroklos Samaras, Tobias Schmidt, Martin Frejno, Siegfried Gessulat, Maria Reinecke, Anna Jarzab, Jana Zecha, Julia Mergner, Piero Giansanti, Hans-Christian Ehrlich, Stephan Aiche, Johannes Rank, Harald Kienegger, Helmut Krcmar, Bernhard Kuster, Mathias Wilhelm, ProteomicsDB: a multi-omics and multi-organism resource for life science research, ***Nucleic Acids Research***, Volume 48, Issue D1, 08 January 2020, Pages D1153–D1163, gkz974, https://doi.org/10.1093/nar/gkz974

Summary

ProteomicsDB was initially developed as a human-centric mass-spectrometry-based proteomics database. In 2017, the platform was extended to support different omics-types and cell viability data, along with a versatile resource identifier relation data model. The platform is extended not with new data, data models and analysis tools. The data wealth of ProteomicsDB is enriched with one more cell viability study increasing the total number of screen drugs in ProteomicsDB to ~20,000. Moreover, the drug-target interaction data are extended to 1,500 kinase inhibitors and tool compounds. The protein property information is enriched with 13,000 melting points of proteins obtained by thermal proteome profiling. The biochemical assays section contains now also protein turnover data, including synthesis and degradation curves for more than 6,000 proteins. The data extension is completed with the import of more reference spectra, 5 million originating from synthetic peptides and about 40 million from predictions using Prosit. The total of the stored data along with the previous extensions of the platform opened the way for real-time multi-omics data integration tools. Taking advantage of the identifier mapping model and the matching transcriptomic and proteomic datasets, the interactive heatmap is extended with an mRNA-guided missing value imputation method. Elastic net regression models are applied to the cell viability and protein expression datasets to model drug sensitivity. The fitted models are available as a new analysis tool, where users can predict the drug sensitivity of a preferred cell line or tissue that is stored in the database. In order to allow users to bring ProteomicsDB closer to their laboratories and own datasets, a new feature is implemented for the temporary upload of custom-user expression data and their online analysis and side-by-side comparison with data stored in ProteomicsDB. At the time being, user data can be used in the interactive expression heatmap and the drug sensitivity prediction tools, but it is planned to be extended to the whole analysis toolbox of ProteomicsDB. Finally, the platform transforms from human-centric to multi-organism, extending all its functionalities to every other organism, with Arabidopsis thaliana being the first expansion.

Author contributions

The author of this dissertation had a leading role in all aspects of the work presented. The author was the lead developer in all tools and models that were implemented in this work. M. Frejno implemented the R procedures for the elastic net model training. The author together with M. Frejno, designed and implemented the data model for storing the model parameters as well as the SQL views and procedures for the drug sensitivity predictions. The design and implementation of the user interface and backend endpoints for the drug sensitivity prediction tool were done by the author. The author designed and implemented the custom user data upload frontend and backend as well as its integration to existing analysis tools. The author together with M. Wilhelm designed the user interface components that were needed for the protein turnover data as well as the expansion to new organisms. J. Mergner prepared the arabidopsis bodymap, while the author integrated the graphics into the user interface. T. Schmidt imported the predicted spectra for human and arabidopsis peptide sequences in 3 charge states and 3 collision energy settings. The execution and implementation of the tools presented were performed under the continuous supervision of M. Wilhelm and the doctoral supervisor Prof. Bernhard Kuster. The manuscript was drafted by the author and discussed with and finalized by T. Schmidt, M. Wilhelm and B. Kuster.

Rights and permissions

Chapter 5
General discussion and outlook

Contents

1 *Status quo* of ProteomicsDB

The big vision behind ProteomicsDB is to provide a one-stop-shop for solutions in daily issues in life science research. Scientists will be able to upload their expression profiles for online analysis, find relevant data from other omics types, select the appropriate cell line for their experiments and build hypotheses based on their online findings. ProteomicsDB is designed as a collection of data, analysis tools and access points that enable public access to data originating from different omics fields, tissues and organisms all processed with the same pipelines using the same quality control metrics. Different aspects of the platform could be embedded in wetlab workflows helping with the design of experiments, allowing the validation of in-house results based on public studies, assist in biomarker discovery or even be integrated into a full workflow that would process raw data directly retrieved from a mass spectrometer. ProteomicsDB is purposed for helping scientists and clinicians with simple questions like “Which drug is selective and effective against a protein or cell line?”, or “Which cell line should be used for an experiment if the expression of a specific set of proteins is required?”.

The vision is not completed yet, as there are several challenges that need to be faced. ProteomicsDB should be prepared in a generic way as every laboratory uses a different experimental workflow. It needs to be updated continuously from remote sources as a wrong or old identifier for a protein or drug might lead to the generation of false hypothesis triggering a chain of unnecessary experiments. Currently, constant updating needs manual work is time-consuming and prone to errors. Automated solutions have to be designed to overcome this obstacle and bring ProteomicsDB closer to its goal.

The current implementation solves already many issues. It brings together different omics types, cell viability data, and new organisms and expands the data wealth of ProteomicsDB with more studies around proteomics, transcriptomics and drug-target data. However, it lacks some flexibility due to the focus during design-time on specific aspects. Therefore, each data model and extension during this thesis comes with its benefits and drawbacks.

1.1 Support of post-translational modifications

From the beginning, the data model of ProteomicsDB was able to store post-translational modification (PTM) data. PTMs are imported directly from the MaxQuant (1) output, and ProteomicsDB is using UniMod (2) PSI names and identifiers to make them query-able to the scientific community. However, the current implementation allows only the visualization of the position of each detected PTM on a protein and peptide sequence. Another place where PTMs are visible is in the peptide identification list of the spectrum viewer, where the user can filter for allowed fixed or variable modifications. As PTMs have been highly associated with disease and in particular cancer (3-5), ProteomicsDB needs to provide a more direct way to search for them. As a first step, a unique and global naming scheme has to be defined, as every software is using its own format. The naming scheme should contain the type of PTM, the position in the peptide sequence and the amino acid or residue the on which it is bound, for example, “Phospho@S4”, which means there is phosphorylation on the fourth amino acid that is a serine. In case of existence of more than one PTMs in a peptide sequence, they could be presented in a modification string separated by a comma and in ascending order by position. Having this scheme in place allows the query of any peptide identification that includes PTMs from the database. The second step would involve the extension of all current peptide search functionalities in the website to allow existence or filtering on desired modifications. That way, users could browse a specific

protein, select all identifications that contain specific modifications and focus their analysis based on the results.

1.2 Extending the stored drug-target data

The drug-target space of ProteomicsDB is extended so far only by in-house data, using the Kinobeads (6) assays. Integrating data from other resources would extend the target-space of the stored compounds further from kinases, which is currently the case. Nevertheless, the real problem extending any kind of data that are connected to drugs and compounds it the manual mapping of these compounds to public identifiers. Most of the public studies and datasets provide internal identifiers and names of the used compounds, often accompanied by simplified molecular-input line-entry system (SMILES) (7) codes, with a partial mapping to existing drug databases, not referring to the version of this database. Compound databases get continuously updated, and many times, entries are merged or deleted. During implementing ProteomicsDB, ChEMBL (8) was chosen as the database of reference for compounds. It is challenging to connect the correct identifier to the imported drug if no structural information is provided. Matching by name is hard but also prone to errors as there was many times the case of a compound name matching to two or more compound identifiers in the ChEMBL database. Also, the use of SMILES codes is not the most efficient as there is a one-to-many relation between SMLEs and structures. It is also possible a SMILES code with no stereochemistry information to match to more than one compound. A unique representation of the structure of a compound is encoded into InChI (9) codes. However, it is rarely the case that the published studies will also provide InChI codes for their compounds. A SMILES to InChI converting service would solve this problem.

ChEMBL provides a free API nowadays with many functionalities that could be exploited from resources like ProteomicsDB. One of their services is the online and real-time transformation of SMILES to InChI codes. Searching in ChEMBL for compounds that match to an InChI code is still not easy, in any case. UniChem (10) was built as a movement to connect compound identifiers from different online chemical compound databases. It is updated regularly and provides an API for the retrieval of the desired identifier, given an InChI code. Combining the two APIs could help not only the import of new compound-driven studies in the database but also the cross-link connectivity to other databases as well as the real-time retrieval of information regarding the clinical phase of a compound or if it is approved and in which countries.

With the identifier mapping issue solved, more studies could be imported in ProteomicsDB. However, at the same time, certain limitations of the platform's UI should be solved. Checking the biochemical assay tab of a protein and filtering for the Kinobeads assays will result in many entries one for each compound per study for which there exists a binding curve. In a different tool, studying the selectivity of any compound against a selected target-protein, the user will be presented with a long list of compounds, the same as in the previous tool. There is a need to filter these compounds and sort them based on some selectivity score. The implementation of the concentration- and target-dependent selectivity score CATDS (6) would solve this issue, limiting the results only to the selective inhibitors, or giving the opportunity to the user to set the threshold manually and order the results by the selectivity score.

Another issue for this type of data is that Kinobeads assays, same as with proteomics, are blind to particular proteins, which might cause them not to appear as targets of a compound. Clinicians and researchers should be aware that if there is no curve information about a specific protein it does not mean that it is not inhibited by the selected compound.

1.3 Cell viability data model and visualization

ProteomicsDB supports the storage and visualization of dose-response datasets. The data model is generalized to allow storage of different model attributes and mappings between them as it consists of several triple-stores. The user interface (UI) provides a variety of filters to reduce the shown data only to the relevant ones. After filtering on the available data, the UI allows the comparison of several drugs on a cell line or different cell lines treated with a selected drug. Although both the data model and the UI allow the storage and visualization of combination treatments, they are not visualized, however, in an optimal way. More filters could be introduced for the selection of the allowed ratios of the combined drugs. Another filter could be added for the selection of a pool of drugs that users want to explore existing combinations in the stored studies in the allowed ratios. Another shortcoming of the current implementation is that there is no connection of the displayed cell lines to the relations of the tissue ontology of ProteomicsDB. Establishing such a connection would allow the users to filter for cell lines with the same tissue of origin or even "cut" the tissue ontology on a higher level, allowing, for example, all cell lines that originate from tissues form the cardiovascular system.

However, one of the main challenges remains the import of new studies. Drugs of the current imported studies were manually mapped to identifiers of other resources, facing the same problems as described above and extending the time needed to map and import the data in the database. A central online drug identifier mapping solution would fix this issue, bringing along all the other benefits that were described earlier, such as clinical trial information.

1.4 Multi-organism extension

ProteomicsDB is not a human-centric platform anymore. Its functionalities are extended to any other organism that is stored in the database. All tools and visualizations, though, are always available for a selected organism only. There is no direct way of comparing protein or transcript expression values across organisms, for example, between the same tissues of human and mouse. ProteomicsDB lacks currently the information about homologues between species, something that would enable such comparisons not only on expression level but also in protein-protein interaction networks or even comparison of pathways.

The user experience (UX) on the website could also be improved. Currently, the change of the selected organism results in a total reset of the website to the default settings and view. Any produced graphics and analysis is lost upon change to the new organism. Having a mechanism in place for temporarily storing the current results and providing a link for their accessibility would allow at least offline comparisons between species. The next step, of course, could be the incorporation of data from more than one species at any step of the online analysis.

It is already stated that ProteomicsDB was expanded to support any other organism. This is true only for multi-cellular organisms, though. Prokaryotic species or in general unicellular organisms cannot be stored, annotated and visualized properly. One step towards solving this issue is the extension of the tissue ontology of ProteomicsDB with terms and identifiers for organelles or subcellular locations. Finally, the development and integration of more organisms could benefit from the usage of publically available organism bodymaps as vector graphics were tissues-paths are annotated with identifiers existing in tissue ontologies.

1.5 Multi-omics data model

The multi-omics data model was the first extension of ProteomicsDB as a work of this thesis. The design of the model was focused on the already existing but limited support of quantitative transcriptomics data in ProteomicsDB. As a result, this model was designed to replace the existing data model and store transcript expression data in a generic way. At the same time, it was able to capture any other omics-type, where the data are presented as tuples of molecule-identifiers and an expression or amount numeric value. That includes copy number variation data, but there are still many fields that are not supported by this implementation. In the case of genomics, mutation data need a different description than expression data and a complex value including a position in the genome or gene, the type of the mutation (e.g. deletion or insert of a nucleotide) and a character-value representing the new nucleotide. Single nucleotide polymorphism (SNP) data support is also suffering for the same reason. The data model should be extended in order to support further omics-types, including metabolomics, methylomics or lipidomics, fields that can further contribute to multi-omics data integration and lead to interesting results (11-13). A possible solution is the alteration of the expression table to a feature table, not storing anymore expression of a gene, probe or any identifier, but include only fields for the description of this identifier. This could be achieved by using again a table as a triple store that would store attributes or properties of an identifier. An attribute, in this case, would be a position in the genome, the type of a mutation, the value of a mutation or the value of an SNP. The model would still support expression data, as the expression value and the unit of this value could be stored as attributes of this feature.

Another fast-moving field is genomic and proteomic analyses in single-cell measurements. Single-cell studies often explore developmental stages of a cell of a tissue (14,15). The current implementation of ProteomicsDB supports the storage of single-cell expression data, but is lacking a fair description of these studies. ProteomicsDB uses the BRENDA Tissue Ontology (BTO) (16) as for the annotation and mapping of samples to tissues of origin. As BTO does not include description or identifiers for developmental stages of tissues or cells, more ontologies could be integrated into the platform to capture the information around single-cell studies fully.

2 Sustainable infrastructure in a fast-advancing technology

A reason that ProteomicsDB has not reached yet the goal of becoming a one-stop solution store is the advances both in life science and in information technology. New methods, tools and instruments are designed and used for faster or more accurate measurements. To be able to describe these data and provide proper annotation, ProteomicsDB has to be updated and extended continuously.

2.1 Advances in technology in life science

Several new studies are not focused on the proteome or transcriptome of a single organism but a collection of symbiotic bacteria, as in the metaproteomics field (17). Samples in these studies cannot be stored in ProteomicsDB currently as there is no way of annotating a sample with different taxonomy codes (taxcode). Depending on the taxcode of a sample, the appropriate sequence space of that organism is used, which is also not easily extendable in the case of metaproteomics, at least in the current implementation of the platform. There needs to be implemented a new solution for the annotation and integration of complex samples and experiments. That would also affect the processing and import pipelines of ProteomicsDB as the

sequence space of such samples grows significantly resulting in only a few peptides surviving the FDR thresholds and leading to even less inferred proteins.

The proteomic studies that are stored in ProteomicsDB and covered by the underlying pipelines are using the data dependent acquisition (DDA) mode. In principle, storage of studies using data independent acquisition (DIA) mode is also supported, but there are many challenges to be faced regarding the proper FDR control as well as the definition of a standard processing and import pipeline. The proteomics community makes efforts in defining a proper target-decoy approach so that control for false positives is possible. Merely applying the same approach as for DDA is not enough, however. In DDA, the mass spectrometer reports the *m/z* ratio of peptide ions. In DIA peptide ions are measured in 2 dimensions, which are *m/z* ratio and retention time. The decoys that are generated for the database search should follow the same principles in order to give a 50:50 chance to a spectrum to match to them. The problem here is that simply shuffling or reversing the peptide sequence, like in DDA, would cause the actual peptide to elute in different retention time, making the decoy sequence not appropriate for this competition. As ProteomicsDB cares about the quality and the truth of the provided data, it is perhaps not yet the time to support DIA experiments. However, having in place the data model and import pipelines for such experiments will speed up the process when the time comes to support DIA data.

A rather interesting and evolving field in the world of omics technologies is the sequencing technologies. Genomics and transcriptomics have already seen the benefits of nanopore sequencing. Low-cost devices, like the Oxford Nanopore's portable MinION (18,19), are being used for fast full genome sequencing. The DNA or RNA sequence passes through a biological pore, where the electrical conductivity of each nucleotide is measured, therefore identifying each nucleotide sequentially. There are current efforts in bringing the nanopore sequencing technology in proteomics (20,21). However it is more complicated as in DNA and RNA only four different states have to be identified, while in proteomics there are many more states if one considers all amino acids and the modifications they might carry. This technology could shine a bright light into de novo sequencing for proteomics and ProteomicsDB should be ready to integrate such results in the current data model.

2.2 Advances in informatics

Information technology (IT) is also advancing quickly. Electronics and circuits are getting smaller and smaller, allowing the build of more powerful central or graphics processing units (CPU, GPU) using the same space on a chip. More computing power enables the design and implementation of more complex algorithms for the reanalysis of existing data to extract more information or simply to increase the processing speed. Not only hardware evolves, though, but also software and programming languages. In the last years, newer user interface (UI) technologies and frameworks made their appearance, promising an easier and better development experience, shareable code components and a modern way of visualizing data. Due to that, older frameworks get deprecated and discontinued forcing existing applications to update or even change the UI frameworks they use for their development.

ProteomicsDB is currently using an old UI framework that is not used by many other platforms anymore. Parts of the framework are deprecated, and there is a lack of available examples for the development of new tools or applications. Due to the deprecation, updating to a newer version of the current framework needs a major rewriting of the existing code base, which is not meaningful if the new version has the same issues, like limited community usage and support. This

led to the decision of using a newer framework, called VueJS which is supported by a large open-source community. Using such a framework comes with the advantages of reusing existing and published code, finding quick solutions to arising problems, and avoiding common bugs leading to faster development times. The current expansion of ProteomicsDB is based on the integration of the two frameworks, which, although possible, is very hard to maintain. If possible, a full reimplementation of the platform should be followed as a next step, to speed up even more future development. Finally, ProteomicsDB hosts many unique visualization tools that could be rebuilt in a generalized way using the new framework. The generalized plots allow reusing them in other webpages of ProteomicsDB, but they can also be offered to the open-source community and expand the list of available examples and tools.

Another obstacle in the development of the one-stop-shop is the database management system behind the platform. When ProteomicsDB was initially designed, SAP HANA was an innovative solution, providing fast querying of huge tables, by optimizing storage and indexing in the main memory of a system. It was and still is a very good solution to the size of the data and the complexity of the queries that are performed with every call in the frontend. However, during this thesis, the database was expanded with data and procedures that are not suitable to a relational data model, such as the identifier mapping procedures. The current advances in IT brought in the foreground powerful graph DBMS, like Neo4j (22,23), which are able to store this kind of data and provide answers to graph-formed questions, such as the identifier mapping problem. A possible solution in this problem is the decoupling of the single DBMS that is currently used into several DBMSs, each one serving a different cause and providing answers to the corresponding questions.

The steps that were described so far will already provide a stable ground for the future sustainability of the platform. Applying these solutions will result in a reimplementation of a huge part of the existing code. During this procedure, it is worthwhile taking some time to think of other existing technologies that would help with future development, deployment and upgrades. SAP HANA offers the functionality of packing the whole platform in packages, called delivery units (DU). These packages include the database entities, like the schema, the tables, the views and the procedures, the server API and backend calls, and the UI code. The advantage of the DUs is that they are readily deployable to every other SAP HANA system, making migration to newer or better infrastructure easier. They can also include the stored data as files, but it is not recommended as it will increase the size of the DU significantly. The client- and server-side of ProteomicsDB are already parts of a DU including a part of the database schema. While recoding ProteomicsDB, it is worthy of moving the entire database structure in the DU, which will make the whole platform portable. Although the data are not part of the DU, there are ways offered by SAP HANA to import them from an existing database directly.

Re-implementing a full platform is not an easy task or decision. It needs pausing all active tool development or debugging, as double trouble due to fixing the same error twice might arise. It becomes more difficult, especially when the developers behind the platform are researchers and academic progress is measured by the number and impact factor of publications and not with modern-looking visualization tools. Also, there are no grants in academia that would support the hiring of professional developers with the sole purpose of re-designing and implementing an already existing application. The decision is tough though necessary, as it will enable fast future development and extension, even remigration to newer versions or infrastructure when needed.

3 A Resource as a Service

Becoming a F.A.I.R. resource brings ProteomicsDB one step closer to its goal, as it will also provide channels of communication with homemade software used in different wetlab workflows, At the same time it will allow ProteomicsDB to be kept up-to-date with data from other resources.

Many of the online resources described in this thesis offer a web UI but also an API, allowing programmatic access. In both tools, the content that is served is defined by the developers to comply with the needs of the frontend or the estimated usage of their data. The use of open protocols for the definition of an API enables the easy integration of such calls in other services. Usually, though this is not enough as APIs and backend calls are designed in a human-readable way. APIs and online resources that follow the F.A.I.R principles (24) open the way to inter-resource communication. Two F.A.I.R. resources that both depend on each other's data can easily establish communication routes that will speed up routine processes, which otherwise would need manual data handling and even annotation. ProteomicsDB is a nice example, as many of its underlying controlled vocabularies, ontologies and raw data originate from other resources. The protein sequence space, for example, is based on the protein sequences that are stored and evaluated by UniProt (25). A direct communication path would allow the automatic update of the underlying sequence database every time it changes in the original resource. Drug and compound entries could be updated or modified using direct connections to the ChEMBL (8) database. Finally, a direct connection with the raw file repository PRIDE (26), coupled with the data process queue and import pipelines of ProteomicsDB, could automatically extend the data wealth of the platform. Raw files and annotation could be retrieved from PRIDE to the ProteomicsDB servers, processed with the standard pipelines and directly import the results in the database.

A F.A.I.R., modern-looking, easy-to-use and expand ProteomicsDB can play an important role in life science research but also the clinical setting.

3.1 For experiment planning

ProteomicsDB can already support life science in experiment planning. It offers the option to explore the properties of a protein of interest and assist users in designing their experiments under specific conditions, like temperature, dose and duration. By using the existing analysis tools, researchers can explore in the interactive expression heatmap expression patterns of their proteins of interest, or protein targets and off-targets of a specific compound and find the appropriate cell lines for their experiments. The protein-drug interaction graph can provide meaningful insight into designing combination treatments by controlling the concentration of each separate compound and exploring the changes in the target-space of all selected compounds at the same time. Another type of experiments that ProteomicsDB can provide a quick solution is targeted assays. As discussed already, ProteomicsDB stores both protein quantification and identification information. A quick visit to the platform can provide information for proteotypic peptides that can later be used for the design of SRM assays. To ensure the user of the quality of the peptide identifications and therefore the truth behind the proteotypicity of a peptide, researchers can explore the evidence behind the identification of proteins, by checking the spectra of the identified peptides. The experimental spectra can be compared to reference ones originating from synthetic peptides or reference spectra that were predicted by Prosit. Visiting ProteomicsDB and starting from a peptide, protein, tissue or organism of interest, a researcher can start building hypotheses that can be later tested in the laboratory or by analyzing existing data.

However, there are a few things that stall the usage of the platform in a wetlab workflow. Many of these were described already, like the lack of more data or omics-types and the missing or false mappings of molecule names to identifiers. In every experiment, a different protein sequence database is used, and even if the resource of this database is the same, the version always changes. ProteomicsDB needs to face this challenge and provide a way of mapping identifiers across databases and versions so that the results are valid no matter which database was used for the search of the data. Finally, by involving more the scientific community into the design of the platform, organizing user forums and workshops for researchers will allow ProteomicsDB to grow in a more welcome to the wetlab way.

3.2 For online analysis of results

Another field in the vision of usage of ProteomicsDB is the online analysis of in-house produced raw data. If possible, pipelines would allow the direct import of experiments and measurements right after their output from a mass spectrometer. This is not ready yet, however. The current version of the platform allows the upload of expression data, but requires the user to process the raw data offline and then proceed to the upload. Even then, the functionality and analysis that allows user data is currently limited. However, it is planned to be expanded in every part of the resource. ProteomicsDB, as a platform, is not just what is available to the user, including the backend functionalities and procedures. It comes with a variety of scripts and pipelines that take care of metadata annotation, project import, quantification and identification searches, FDR calculation procedures and many other scripts that help the semi-automatic data and metadata retrieval. Even if the user would like to perform offline analysis combining data from ProteomicsDB with in-house data, the current API does not cover the whole database, so not all data is accessible in a programmatic way.

By solving the aforementioned issues, ProteomicsDB could be a valuable addition in every workflow and every stage of performing an experiment. For example, in proteomics, experiment planning and hypothesis generation could be initiated while exploring the data in ProteomicsDB, like finding a cell line the expresses specific targets of a compound. Afterwards, sample preparation would follow and then the measurement of the sample in a mass spectrometer. Binding ProteomicsDB to the output of the MS would start the online identification and quantification pipelines. The researcher would then validate the expression of the protein of interest in their samples and continue the online analysis or comparison to the existing data. After exploring the available studies, the next step could be the design of a combination treatment experiment in ProteomicsDB based on available compounds in the laboratory and the expressed proteins in the processed sample. This can trigger a second circle of experiments, preparing the new samples, performing the combination treatment experiments and measuring again in the mass spectrometer, uploading and searching in ProteomicsDB leading this time the researcher to perform an online differential expression analysis. At this step, the researcher can check if the targets of interest are significantly inhibited to the expected degree and how it affected the general expression patterns in the sample. The workflow could stop here if the initial hypothesis was already proved or trigger the generation of a new hypothesis and initiate a new loop.

ProteomicsDB is envisioned to play a central role in experiment design and online analysis. It is possible to achieve this, by assisting the life science research via loops of experiment planning and performing as well as data integration and analysis that would lead to a new phase of experiment planning.

3.3 For the clinical setting

The final big goal of ProteomicsDB is to be integrated into the clinical setting. There are many challenges that need to be faced in order to begin this integration, though. One obstacle is concerning the drugs and compounds that ProteomicsDB stores data for, where there is no information yet about clinical trials, or if they are commercially available or FDA approved. This information is crucial for clinicians as they would use only approved drugs for treating patients. Nevertheless, even with this information available, the usage of ProteomicsDB would be limited to consulting based on online data. Patient data privacy does not allow, at least not easily, the upload of patient data into online resources for their analysis, as data might be exposed during this process. Therefore, the integration into an experimental workflow as described above is not suitable to the clinical setting as it would require the direct communication of the produced data with an online service and probably the online storage of the data. Fortunately, the solutions that were described already as parts of the future development and sustainability of ProteomicsDB can be proved worthy.

Having ProteomicsDB as a portable version would allow the deployment of the platform locally into the clinical setup. The benefit of having a local platform is that it will be in a totally isolated setting, with no physical access to external individuals and a controlled, if not at all, communication between the platform and online resources. The drawback here is that ProteomicsDB is designed to be hosted and function in an SAP HANA system. The amount of data that the platform includes currently requires a powerful infrastructure to store and analyze the data, which is quite expensive, and what extends this cost is the license for an SAP HANA system. Even if the cost of the license and infrastructure is not an issue, it requires highly-trained personnel for the expansion of the local platform with newer datasets or even tools. There is a clear need for a central “Master” repository that is available to browse online. An extension of the idea of a local platform could be, the installation of a minimal local ProteomicsDB that holds only the imported clinical data, while all other external data are retrieved from the “master” repository, something that SAP HANA enables via direct communication channels, called “remote sources”. In that case, the computational power is split between two servers, the master querying and returning only the relevant data, while the “slave” local setup would perform the analysis on the retrieved and local data. The computational power of a “slave” system will not be on the same scale as the “master” system. This way the cost is dropping drastically as SAP HANA is also offered as an “Express” edition that is free for systems with up to 36GB of main memory. The “slave” server can be updated both with data and tools, but it still needs personnel with expertise on HANA systems. A last important extension to make the “slave” server easily deployable and manageable by any IT personnel is the containerization of the minimal ProteomicsDB with the SAP HANA express edition in a docker container. That way, any clinic could exploit the knowledge that is encoded in the data wealth of ProteomicsDB and use it as patient-focused treatment consultant service, moving one step closer to personalized medicine.

A valuable addition to the platform and in particular to the clinical setup it the new treatment suggestion tool that is based on the molecular similarity of samples. Clinicians could import the proteomics profiles of cancer patients and use this tool in two ways. The first use-case is to find similar to the patient samples, commercially available cell lines and ask ProteomicsDB to provide a list of effective and clinically approved drugs based on the cell viability data that are stored. The second use-case is to make full usage of the treatment suggestion tool and retrieve a set of effective drugs that are significantly enriched in samples and cell lines that are molecularly similar

to the patient sample. This tool was a result of the Master thesis of Pia Bothe with the title "*Development of a human tissue ontology based on multi-omics expression patterns to investigate drug responses*", and the wetlab validation of the resulting compound list was performed during the internship of Lisa Falk and Philipp Hilgendorf. ProteomicsDB can provide significant assistance in patient treatment and complement the existing analysis in molecular tumourboards.

The entirety of ProteomicsDB, though, was built with a focus on the scientific community. There were comments and suggestions from the clinical community to simplify the web interface as it is not comprehensive to them, currently. It is true that scientists with specific questions can visit the resource and easily derive the appropriate answers. However, the way that the platform is presented currently is too specific and might drive clinicians and even scientists from other fields away. There is the need to provide ways and views that would summarize the stored information to the parts that are relevant not only to clinicians but to the corresponding users. Researchers from the proteomics field will be interested in checking the tandem spectra that are stored, but this is not something that would interest a clinician. Scientists for the chemical proteomics field or medicinal chemistry might want to directly check the target space of a compound or the compound that inhibit a specific protein given a concentration, something that is not directly accessible currently, although this information can be derived by using other analysis tools. A way to circumvent these issues is to reformat the UI of ProteomicsDB and introduce different modes of presentation, like a data expert and a data consumer mode. The first mode represents the current view of ProteomicsDB. The second mode would lead the user to the relevant analysis tools in a step-by-step approach, starting from a protein, cell-line or drug of interest and providing simplified and aggregated results in downloadable reports. With regard to the first mode, the current UI is protein-centric, something that also makes it difficult for people starting from a compound or cell-line of interest to explore the database. Adjusting the platform to provide different entry points and linking the results to existing tools or pages would already make the platform friendly to more scientific disciplines.

There is a bright future regarding the contributions of ProteomicsDB to the scientific and clinical community. The vision behind the platform is not science fiction, but more like a scientific need. Tools are built to be used and should be built in a comprehensive way to the user.

If they are not used, then why do we even bother building them?

4 Abbreviations

API	Application programming interface
BTO	BREDA Tissue ontology
CATDS	Concentration- and target-dependent selectivity score
DBMS	Database management system
DDA	Data dependent acquisition
DIA	Data independent acquisition
DNA	Deoxyribonucleic acid
DU	Delivery unit
FAIR	Findability, Accessibility, Interoperability and Reusability
FDR	False discovery rate
InChI	International Chemical Identifier
IT	Information technology
MS	Mass spectrometer
PTM	Post-translational modification
RNA	Ribonucleic acid
SMILES	Simplified molecular-input line-entry system
SNP	Single nucleotide polymorphism
SRM	Selected reaction monitoring
UI	User interface
UX	User experience

5 References

1. Tyanova, S., Temu, T. and Cox, J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc*, **11**, 2301-2319.
2. UniMod, http://www.unimod.org/modifications_list.php, *Accessed* 15 April 2020.
3. Cocchiola, R., Rubini, E., Altieri, F., Chichiarelli, S., Paglia, G., Romaniello, D., Carissimi, S., Giorgi, A., Giamogante, F., Macone, A. *et al.* (2019) STAT3 Post-Translational Modifications Drive Cellular Signaling Pathways in Prostate Cancer Cells. *Int J Mol Sci*, **20**.
4. Hsu, J.M., Li, C.W., Lai, Y.J. and Hung, M.C. (2018) Posttranslational Modifications of PD-L1 and Their Applications in Cancer Therapy. *Cancer Res*, **78**, 6349-6353.
5. Jin, H. and Zangar, R.C. (2009) Protein modifications as potential biomarkers in breast cancer. *Biomark Insights*, **4**, 191-200.
6. Klaeger, S., Heinzlmeir, S., Wilhelm, M., Polzer, H., Vick, B., Koenig, P.A., Reinecke, M., Ruprecht, B., Petzoldt, S., Meng, C. *et al.* (2017) The target landscape of clinical kinase drugs. *Science*, **358**.
7. OpenSMILES, http://opensmiles.org/, *Accessed* 15 April 2020.
8. Mendez, D., Gaulton, A., Bento, A.P., Chambers, J., De Veij, M., Felix, E., Magarinos, M.P., Mosquera, J.F., Mutowo, P., Nowotka, M. *et al.* (2019) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res*, **47**, D930-D940.
9. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D. and Pletnev, I. (2013) InChI - the worldwide chemical structure identifier standard. *J Cheminform*, **5**, 7.
10. Chambers, J., Davies, M., Gaulton, A., Hersey, A., Velankar, S., Petryszak, R., Hastings, J., Bellis, L., McGlinchey, S. and Overington, J.P. (2013) UniChem: a unified chemical structure cross-referencing and identifier tracking system. *J Cheminform*, **5**, 3.
11. Patt, A., Siddiqui, J., Zhang, B. and Mathe, E. (2019) Integration of Metabolomics and Transcriptomics to Identify Gene-Metabolite Relationships Specific to Phenotype. *Methods Mol Biol*, **1928**, 441-468.
12. Li, J., Ren, S., Piao, H.L., Wang, F., Yin, P., Xu, C., Lu, X., Ye, G., Shao, Y., Yan, M. *et al.* (2016) Integration of lipidomics and transcriptomics unravels aberrant lipid metabolism and defines cholesteryl oleate as potential biomarker of prostate cancer. *Sci Rep*, **6**, 20984.
13. Cavill, R., Jennen, D., Kleinjans, J. and Briede, J.J. (2016) Transcriptomic and metabolomic data integration. *Brief Bioinform*, **17**, 891-901.
14. Dou, M., Clair, G., Tsai, C.F., Xu, K., Chrisler, W.B., Sontag, R.L., Zhao, R., Moore, R.J., Liu, T., Pasa-Tolic, L. *et al.* (2019) High-Throughput Single Cell Proteomics Enabled by Multiplex Isobaric Labeling in a Nanodroplet Sample Preparation Platform. *Anal Chem*, **91**, 13119-13127.
15. Zhu, Y., Piehowski, P.D., Zhao, R., Chen, J., Shen, Y., Moore, R.J., Shukla, A.K., Petyuk, V.A., Campbell-Thompson, M., Mathews, C.E. *et al.* (2018) Nanodroplet processing platform for deep and quantitative proteome profiling of 10-100 mammalian cells. *Nat Commun*, **9**, 882.
16. Placzek, S., Schomburg, I., Chang, A., Jeske, L., Ulbrich, M., Tillack, J. and Schomburg, D. (2017) BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res*, **45**, D380-D388.

17. Rechenberger, J., Samaras, P., Jarzab, A., Behr, J., Frejno, M., Djukovic, A., Sanz, J., Gonzalez-Barbera, E.M., Salavert, M., Lopez-Hontangas, J.L. *et al.* (2019) Challenges in Clinical Metaproteomics Highlighted by the Analysis of Acute Leukemia Patients with Gut Colonization by Multidrug-Resistant Enterobacteriaceae. *Proteomes*, **7**.
18. Lu, H., Giordano, F. and Ning, Z. (2016) Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics Bioinformatics*, **14**, 265-279.
19. Loman, N.J. and Watson, M. (2015) Successful test launch for nanopore sequencing. *Nat Methods*, **12**, 303-304.
20. Ouldali, H., Sarthak, K., Ensslen, T., Piguet, F., Manivet, P., Pelta, J., Behrends, J.C., Aksimentiev, A. and Oukhaled, A. (2020) Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore. *Nat Biotechnol*, **38**, 176-181.
21. Piguet, F., Ouldali, H., Pastoriza-Gallego, M., Manivet, P., Pelta, J. and Oukhaled, A. (2018) Identification of single amino acid differences in uniformly charged homopolymeric peptides with aerolysin nanopore. *Nat Commun*, **9**, 966.
22. Himmelstein, D.S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S.L., Hadley, D., Green, A., Khankhanian, P. and Baranzini, S.E. (2017) Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*, **6**, e26726.
23. Turei, D., Korcsmaros, T. and Saez-Rodriguez, J. (2016) OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat Methods*, **13**, 966-967.
24. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B. and Bourne, P.E. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, **3**.
25. UniProt, C. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*, **47**, D506-D515.
26. Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D.J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M. *et al.* (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res*, **47**, D442-D450.

Publication record

* Authors contributed equally to this work

Publications presented in this thesis:

- Schmidt, T.*, **Samaras, P.***, Frejno, M., Gessulat, S., Barnert, M., Kienegger, H., Krcmar, H., Schlegl, J., Ehrlich, HC, Aiche, S., Kuster, B., Wilhelm, M. (2017). ProteomicsDB. *Nucleic acids research*, *46*(D1), D1271-D1281.
- **Samaras P.**, Schmidt T., Frejno M., Gessulat S., Reinecke M., Jarzab A., Zecha J., Mergner J., Giansanti P., Ehrlich HC., Aiche S., Rank J., Kienegger H., Krcmar H., Kuster B., Wilhelm M. (2019). ProteomicsDB: a multi-omics and multi-organism resource for life science research. *Nucleic acids research, 48*(D1), D1153-1163.

Additional publications:

- Jarzab, A., Kurzawa, N., Hopf, T., Moerch, M., Zecha, J., Leijten, N., Musiol, E., Maschberger, M., Stoehr, G. Daly, C., **Samaras, P.**, Mergner, J., Spanier, B., Angelov, A., Werner, T., Bantscheff, M., Wilhelm, M., Klingenspor, M., Lemeer, S., Liebl, W., Hahne, H., Savitski, M., Bian, Y., Becher, I., Kuster, B. (2020). Meltome Atlas – thermal proteome stability across the tree of life. *Nature methods*, accepted
- Mergner, J., Frejno, M., List, M., Papacek, M., Chen, X., Chaudhary, A., **Samaras, P.**, Richter, S., Shikata, H., Messerer, M., Lang, D., Altmann, S., Cyprys, P., Zolg, D., Mathieson, T., Bantscheff, M., Hazarika, R., Schmidt, T., Dawid, C., Dunkel, A., Hofmann, T., Sprunck, S., Falter-Braun, P., Johannes, F., Mayer, K., Jürgens, G., Wilhelm, M., Baumbach, J., Grill, E., Schneitz, K., Schwechheimer, C., Kuster, B. (2020). Mass-spectrometry-based draft of the Arabidopsis proteome. *Nature*, 579, p.409–414.
- Gessulat, S.*, Schmidt, T.*, Zolg, D.P., **Samaras, P.**, Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A., Reimer, U., Ehrlich, HC., Aiche, S., Kuster, B., Wilhelm, M. (2019). Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature methods*, *16*(6), p.509-518.
- Rechenberger, J.*, **Samaras, P.***, Jarzab, A., Behr, J., Frejno, M., Djukovic, A., Sanz, J., González-Barberá, E., Salavert, M., López-Hontangas, J., Xavier, K., Debrauwer, L., Rolain, JM., Sanz, M., Garcia-Garcera, M., Wilhelm, M., Ubeda, C., Kuster, B. (2019). Challenges in clinical metaproteomics highlighted by the analysis of acute leukemia patients with gut colonization by multidrug-resistant enterobacteriaceae. *Proteomes*, *7*(1), p.2.
- Zecha, J., Meng, C., Zolg, D.P., **Samaras, P.**, Wilhelm, M. and Kuster, B. (2018). Peptide level turnover measurements enable the study of proteoform dynamics. *Molecular & Cellular Proteomics*, *17*(5), p.974-992.

Earlier publications:

- Kavakiotis, I., **Samaras, P.**, Triantafyllidis, A., Vlahavas, I. (2017). FIFS: A data mining method for informative marker selection in high dimensional population genomic data. *Computers in biology and medicine, 90,* p.146-154
- **Samaras, P.**, Fachantidis, A., Tsoumakas, G., Vlahavas, I. (2015). A prediction model of passenger demand using AVL and APC data from a bus fleet. *Proceedings of the 19th Panhellenic Conference on Informatics*, p.129-134
- Kavakiotis, I., Triantafyllidis, A., **Samaras, P.,** Voulgaridis, A., Karaiskou, N., Konstantinidis, E., Vlahavas, I. (2014). Pattern discovery for microsatellite genome analysis. *Computers in biology and medicine*, p.71-78

Acknowledgements

I cannot believe that it is over. A journey that started four years ago. Four years in Germany, in Munich, in Freising. Four awesome years, working with the most amazing group of people, who embraced me from the first moment. We met, became coworkers and even friends. We shared moments of anxiety, but mostly moments of joy. What could I say about all the great parties, the bubbly in the kitchen for every grant and paper that was accepted, or the long game-nights during our winter retreats? Thank you for all these moments! Thank you for helping me through my PhD.

First of all, I would like to thank Bernhard for believing in me and offering me this opportunity to join the Terrific TUM Team, even though I barely knew what a protein is when I first joined the group. Thank you for being there for me whenever I needed your assistance, even when I sent my manuscript to you just a few hours before submission!

Nothing would be possible without Mathias. I hope I met your expectations when you introduced me to the world of ProteomicsDB. Thank you for your constant help, the spontaneous meetings, the great ideas and the crazy topics during our conversations. I will never forget these weekly meetings that were always a mix of science and weird facts.

I could not forget, of course, Tobi. When we started working on the same project, I thought that I was lucky to have a great colleague next to me. I did not foresee, though, that I would end up making a new friend that we would share so much. Thank you for the endless days and nights of coding and debugging together while listening to the same playlist over and over again. It was indeed an “awesome mix”. And then came Martin, a new (at least to me) postdoc in the lab that would join us for the next three years. Another great friend and colleague, who spent every last bit of patience to teach and guide me through the “omics” world. Thank you for meeting you and your lovely family.

This list would become incredibly long if I named everyone from this amazing team. So thank you all, you made every office day special. Many thanks to all the people that trusted me and made me part of their projects. A big Danke schön to Gabi and Silvia, our precious secretaries, who were always there to help me.

I would like to thank Stephan and Christian from SAP for their great collaboration and their precious help. Moreover, I cannot forget Amelie, Pia, Lisa and Philipp who were a great help to me with their Bachelor, Master and internship projects.

I would also like to thank Prof. Dr. Julien Gagneur and PD. Dr. Martin Eisenacher as well as committee chairman Prof. Dr. Dmitrij Frishman for forming my examination committee.

Looking a few years back, I would like to thank the person who introduced me to the world of bioinformatics, my friend Yannis Kavakiotis. Special thanks go to Prof. Dr. Ioannis Vlahavas and Prof. Dr. Alexandros Triantafyllidis.

Furthermore, I would like to thank all the people that were next to me all these years. My friends in Greece, Dimitris, Lina, Vassilis, Christos, Giannis and Vasso. My parents Vangelis and Ntina for always being understanding and supporting, as well as my brothers Thanasis and Tasos. Most importantly, I would like to thank the love of my life, Ermina, for beginning this journey with me and sharing all the happy and challenging moments of it. I couldn’t have done this without you.

Thank you.

Appendix

Published online 2 November 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D1271–D1281***
doi: 10.1093/nar/gkx1029

ProteomicsDB

Tobias Schmidt[1,†], Patroklos Samaras[1,†], Martin Frejno[1], Siegfried Gessulat[1,2], Maximilian Barnert[3,4], Harald Kienegger[3,4], Helmut Krcmar[3,4], Judith Schlegl[5], Hans-Christian Ehrlich[2], Stephan Aiche[2], Bernhard Kuster[1,6,*] and Mathias Wilhelm[1,*]

[1]Chair of Proteomics and Bioanalytics, Technical University of Munich (TUM), Freising, 85354 Bavaria, Germany, [2]Innovation Center Network, SAP SE, Potsdam 14469, Germany, [3]Chair for Information Systems, Technical University of Munich (TUM), Garching 85748, Germany, [4]SAP University Competence Center, Technical University of Munich (TUM), Garching 85748, Germany, [5]PI HANA Platform Core, SAP SE, Walldorf 69190, Germany and [6]Bavarian Biomolecular Mass Spectrometry Center (BayBioMS), Technical University of Munich (TUM), Freising, 85354 Bavaria, Germany

Received September 15, 2017; Revised October 13, 2017; Editorial Decision October 16, 2017; Accepted October 22, 2017

ABSTRACT

ProteomicsDB (https://www.ProteomicsDB.org) is a protein-centric in-memory database for the exploration of large collections of quantitative mass spectrometry-based proteomics data. ProteomicsDB was first released in 2014 to enable the interactive exploration of the first draft of the human proteome. To date, it contains quantitative data from 78 projects totalling over 19k LC–MS/MS experiments. A standardized analysis pipeline enables comparisons between multiple datasets to facilitate the exploration of protein expression across hundreds of tissues, body fluids and cell lines. We recently extended the data model to enable the storage and integrated visualization of other quantitative omics data. This includes transcriptomics data from e.g. NCBI GEO, protein–protein interaction information from STRING, functional annotations from KEGG, drug-sensitivity/selectivity data from several public sources and reference mass spectra from the ProteomeTools project. The extended functionality transforms ProteomicsDB into a multi-purpose resource connecting quantification and meta-data for each protein. The rich user interface helps researchers to navigate all data sources in either a protein-centric or multi-protein-centric manner. Several options are available to download data manually, while our application programming interface enables accessing quantitative data systematically.

INTRODUCTION

Mass spectrometry has developed into the flagship technology for proteome research much akin to what next generation sequencing has become for genomics and transcriptomics (1,2). Since proteins execute and control most biological processes in all domains of life, they are one of the most frequently targeted class of molecules in the context of drug development. Today, scientists and clinicians anticipate that proteins will also become a major source of biomarkers (3) useful to diagnose disease, to stratify patients for treatment and to monitor response to therapy to name a few.

At the same time, the volume and complexity of proteomics data generated by modern mass spectrometers is challenging our ability to turn data into tractable hypotheses, within and, particularly, across larger projects. In order to provide access to previously performed experiments, many different repositories have been developed (4,5). However, their focus is often limited to a particular aspect of the data and frequently, protein identification is decoupled from protein quantification. PRIDE (6) is currently the community-standard for publishing raw data but also peptide and protein identification results (including post-translational modifications). However—until recently—it lacked an intuitive interface for comparing results across different datasets. PeptideAtlas (7), GPMDB (8) and MASSIVE mostly focus on hosting identification results by reprocessing data using their own pipelines. The protein abundance database (PAXDB) (9) stores quantification data from publicly available data, but lacks the underlying peptide identification results. MaxQB (10) does provide both protein identification and quantification data, but is far less comprehensive than any of the other repositories and also does not allow cross-dataset comparison. While most of

*To whom correspondence should be addressed. Mathias Wilhelm. Tel: +49 8161 71 4202; Fax: +49 8161 71 5931; Email: mathias.wilhelm@tum.de
Correspondence may also be addressed to Bernhard Kuster. Email: kuster@tum.de
†These authors contributed equally to this work as first authors.

these databases can store meta-data such as sample preparation and data acquisition protocols, specific treatments and the different conditions used in the experimental setup are not stored in a programmatically accessible format. In addition, none of the aforementioned databases allow storage of other data types. This in turn makes it difficult to systematically explore and mine data across proteomic or multi-omics experiments.

ProteomicsDB is filling this gap by not only enabling cross-dataset comparisons of protein abundance, but also by providing the means to store and analyse proteomics data in contexts other than expression analyses. The protein-centric web interface provides researchers real-time and use-case-specific access to data for single or multiple proteins using interactive visualizations at different levels of detail. The data model of ProteomicsDB is able to store identification and quantification data from almost all conceivable proteomics experiments including meta-data such as sample preparation protocols, data acquisition parameters and sample treatment conditions. More recently, its capabilities have been expanded to also host results from other quantitative omics technologies ranging from drug-protein interaction studies and cell-viability experiments to data from public protein interaction databases and transcriptomes. In this article, we introduce the different analysis options available in ProteomicsDB and highlight the developments accumulated over the past three years.

RESULTS

ProteomicsDB utilizes the in-memory database management system SAP HANA (11) and was developed to enable the real-time interactive exploration of large collections of quantitative mass spectrometry-based proteomics data (12). A major focus during the initial development of ProteomicsDB was to enable the storage of identification and quantification data on both peptide and protein level, irrespective of the experimental setup and analysis method used. Based on 408 experiments resulting from 78 experiments we identify 15721 of 19629 proteins covering 80% of the human proteome. A comparison to the Human Proteome Project (13,14) can be found in the Supplementary Table S1. To this end, ProteomicsDB is able to store the output of any algorithm used for the automatic interpretation of mass spectra (database search). Combined with the ability to map each observed peptide spectrum match (PSM) in any LC–MS/MS raw file transparently to the corresponding sample annotated with information on acquisition and sample preparation parameters, this ensures flexibility during data analysis. The storage of treatment conditions and the overall experimental design facilitate the analysis of more complex relations within and across different datasets, such as dose- and temperature-dependent assays. Efficient access to the data in combination with modern web-based visualization technologies facilitates real-time interactive exploration of heterogeneous data in an intuitive and simple way. All figures and tables available in ProteomicsDB can be downloaded, while an application programming interface allows users to directly interact with the database in order to download raw data for off-line processing or storage (Figure 1).

Because of the in-memory capabilities of SAP HANA, most of the data shown on the website are not precomputed, avoiding the need for monthly or yearly builds and enabling rapid adjustments. The different storage layers and versatile processing capabilities available in HANA enabled the integration of graph and standard relational database features. This facilitated the incorporation of many different data sources and led to the development of a variety of new features. While all protein-related results stored in ProteomicsDB are mapped to UniProt (15) identifiers, a versatile resource identifier mapping system enables a seamless conversion between different resources, which facilitates easy integration of additional data sources not mapped to UniProt (e.g. transcriptomics and interaction data).

In the following sections, we will start by briefly highlighting the data model used by ProteomicsDB and its developments over the past years. Subsequently, we will introduce the main features available on ProteomicsDB, which are organized in protein-centric visualizations for single and multiple proteins.

ProteomicsDB data model

The data model of ProteomicsDB is grouped into 7 major modules (Figure 1): (i) the meta-data, which contains annotations and ontologies; (ii) the repository, which contains the mapping of raw data to samples, experiments and projects, as well as associated meta-data and experimental designs; (iii) peptide and protein identification and quantification data, which stores spectra, the associated database search engine results, as well as peptide and protein abundance information; (iv) reference identification, which contains reference spectra from measurements of synthetic peptide standards; (v) the quantitative omics model; (vi) experiment specific models, such as dose-response models and (vii) cell-viability data. See Supplementary Text 1 for more details about the data models and its internal mechanisms.

Protein-centric web interface

ProteomicsDB is designed to enable researchers to quickly interrogate identification and quantification information of single and multiple proteins. For this purpose, ProteomicsDB offers two major ways to browse all available data. On the one hand, there is the presentation of information available for a single protein of interest. This can be accessed by either searching for a protein or peptide of interest in the 'Human Proteins' or 'Peptides' tab, respectively, or by browsing the human proteome in a 'Chromosome' centric view (Figure 2A). On the other hand, there are visualizations of specific aspects of the data for multiple proteins. This functionality is referred to as 'Analytics' within ProteomicsDB and can be found in the main menu at the top of the website. Currently, four analytical views are implemented and offer the cross-experiment analysis of protein expression, single and multiple drug selection and the exploration of cell viability data. It should be noted that ProteomicsDB is optimized for Firefox and Chrome.

This section focuses on describing the visualizations for single ('Human proteins') and multiple proteins ('Analyt-

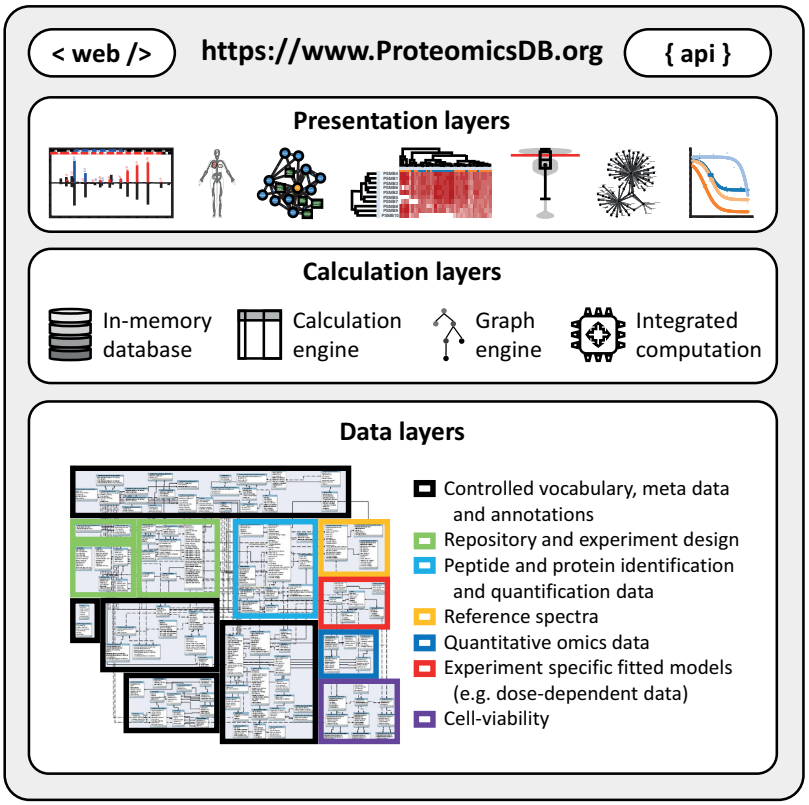

Figure 1. ProteomicsDB consists of three major layers. The bottom layer is the data layer providing information to the calculation layer. It consists of seven major modules enabling the storage and retrieval of meta data, annotations and quantitative information associated with proteins and biological systems. Due to in-memory storage of the data layer, calculations using the calculation engine (structured query language), graph engine and other integrated programming languages (e.g. R and Python) are highly efficient. The results of these calculations can be explored in the presentation layer offering a variety of different interactive visualizations via the web interface or systematic access via the ProteomicsDB application programming interface (API).

ics') in more detail. The features to analyse proteotypicity, reference peptides and FDR estimation for single proteins are fully described in Supplementary Text 2–4 (Supplementary Figures S1–4). Each page also provides a brief description of the functionalities by opening the 'Help'-tab to the right of each page and tab. The feedback-icon, located on the left on each page, can be used to provide direct feedback, comments or report bugs to us. For the purpose of this paper, we will focus on one protein highlighting all available functions and visualizations throughout the manuscript. Discoidin Receptor 1 (DDR1) is a member of a family of receptor tyrosine kinases (RTKs) that is activated in response to collagen and is part of the arsenal of cell surface receptors that mediate tumor cell-environment interactions.

Human proteins

The search field can be used to browse proteins by gene name, accession number or protein description. The resulting table shows all available proteins partially matching the search string. All tables in ProteomicsDB can be filtered and sorted by clicking on a specific column header. Most tables also offer hiding or showing additional columns, which are not shown by default but are always included in downloaded csv files.

Protein summary. Upon selecting a protein of interest, the user sees a brief summary (Figure 2B) about the information available for the protein, including, but not limited to, the number of peptides which were detected (shared and unique on either gene or protein level), the sequence coverage and some basic annotations such as GO terms, chromosomal location, external links and evidence status. The evidence status is either red, yellow or green indicating missing, questionable and strong evidence for its identification, respectively. In addition, the domain structure of the protein is dynamically generated and shown in the middle of the page. Aligned to this, all observed peptides and post-translational modifications (PTMs) are visualized by black bars. This enables users to quickly investigate which part of the protein is (likely) 'MS-accessible' (i.e. produces peptides measureable by mass spectrometry) and which domains were previously observed to harbour post translational modifications (often an indicator of activity modulation). The sequence coverage view can be expanded to investigate which peptides were observed in detail. In addition, the 'Sequence coverage' tab can be opened to view the entire sequence of the protein. Stretches coloured in red indicate that this part is covered by peptides in ProteomicsDB. The theoretical sequence coverage can be explored using the 'Protease map' tab. One or several proteases can be chosen along with different peptide filter criteria, in order to predict which com-

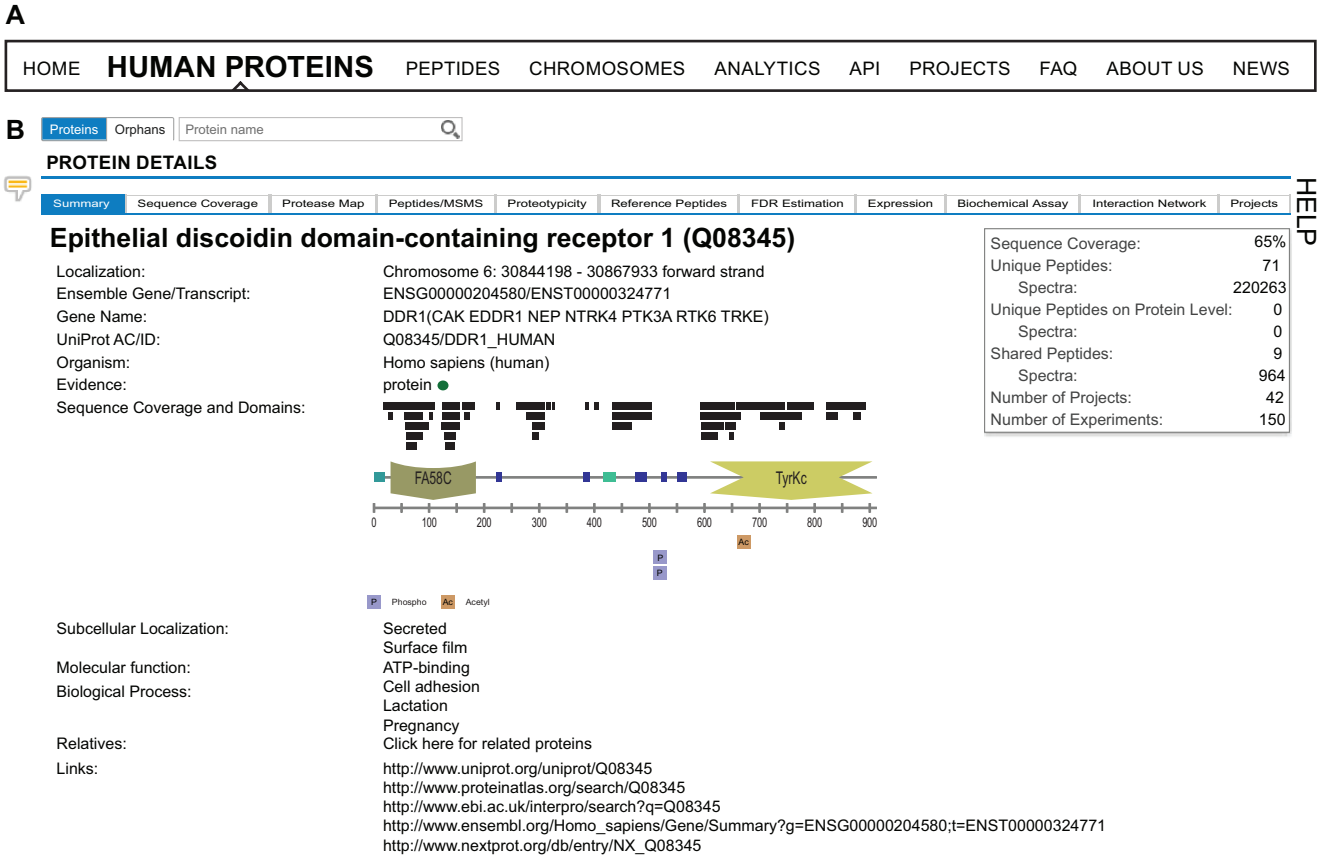

Figure 2. (**A**) ProteomicsDB can be used to interrogate identification and quantification information on either single or multiple proteins. Information about single proteins can be accessed via the 'Human Proteins', 'Peptides', and 'Chromosomes' tabs. Information about multiple proteins can be explored via the 'Analytics' tab. (**B**) On the 'Human Proteins' tab, a brief summary is shown about the information available for a given protein. The corresponding domain structure is dynamically generated and alongside it, all observed peptides and post-translational modifications (PTMs) are displayed.

bination of proteases will lead to the highest (theoretical) cumulative sequence coverage. This feature can guide users in designing experiments that require high sequence coverage such as PTM or variant identification.

Peptides/MSMS. The 'Peptides/MSMS' tab can be used to check individual peptides and their corresponding spectra. The initial view lists all observed peptides including meta-data such as mass, length, uniqueness and the number of observations, as well as different measures of confidence, such as the search engine score. Each spectrum used for protein inference can be visualized in ProteomicsDB using the built-in spectrum browser. In order to view experimental spectra, an overlay containing all available PSMs for the selected peptide (Supplementary Figure S1 top table) can be opened by clicking a peptide of interest. Fragment ions in experimental spectra are annotated on request by an expert system (16). Annotation rules, such as calculated fragment ions and sequence-dependent neutral losses, are stored in the database and can be modified at any time. Annotation options for the spectrum, general visualization options and a fragmentation table can be opened to the left and right of the spectrum (Supplementary Figure S1).

An integrated feature of the spectrum viewer is the mirror representation of a reference spectrum (bottom spectrum) if available. These spectra originate from e.g. synthesized peptides, which were measured independently and can be used to validate the identification of peptides and in turn also proteins. This is especially useful when only a few peptides were identified for a specific protein, since such spurious identifications could originate from false matches during the database search. In case a reference spectrum for the selected peptide is available, the highest scoring PSM matching to the precursor charge and modification status of the selected PSM is chosen and displayed. Already today, ProteomicsDB stores more than 3 million reference spectra acquired as part of the ProteomeTools project (17) and covers more than 250k peptides measured in up to 11 different acquisition methods. For most peptides, multiple reference spectra are available and by default, the one acquired using similar acquisition parameters is shown. However, since parameters such as collision energy are not easily transferable between instruments (18), the user can choose to compare the experimental spectrum against any spectrum acquired under different conditions by selecting a different spectrum to the left of the spectrum viewer in the 'Reference spectrum' tab.

Expression. An essential feature of ProteomicsDB is the storage and visualization of quantitative data from a wide range of biological sources. While the initial development focused on the presentation of proteomics data, the generic

implementation of ProteomicsDB also enables the storage and visualization of other omics data types, such as RNA-Seq data. The 'Expression' tab (Figure 3) can be used to explore the expression pattern of single proteins across the human body. The user can choose the primary data source and can visually explore the expression using a heatmap-like visualization of the human body. This view also superimposes abundance values of cell lines onto their respective tissue of origin and thus allows the integrated analysis of expression values originating from tissues or body fluids and cell lines.

The expression view consists of two major components comprising data selection (Figure 3A) and visualization (Figure 3B–D). To enable meaningful cross-experiment comparison of expression values, only data from similar sources can be selected. For proteomics, MS1 and MS2 quantification techniques (19) cannot be compared directly, thus the filters only support the selection of either type. Likewise, the comparison of protein abundance measures originating from full proteome data (unbiased expression analysis) or affinity type experiments (biased abundance analysis) is not possible.

The data visualization is composed of three interactive and interconnected elements: (i) a heatmap-like body map (Figure 3B), (ii) a cell type aggregated bar chart (Figure 3C) and (iii) a sample specific bar chart (Figure 3D). The expression of DDR1 is restricted to epithelial cells, particularly in the kidney, lung, gastrointestinal tract and brain. Upon selection of a specific tissue in the heatmap, the middle barchart highlights all cell lines and tissues, which are connected to this tissue (e.g. tissue of origin). Likewise, the selection of a bar in the middle bar chart will highlight the corresponding tissue in the bodymap. This will also trigger the display of an additional bar chart, depicting the expression of the selected protein in a sample-specific manner. This view directly enables users to investigate the sample preparation and data acquisition parameters for each measurement by clicking on any bar in the bar chart on the right hand side.

Biochemical assay. Besides visualizing global expression patterns of proteins, ProteomicsDB is also able to make use of the stored experimental design to show changes in protein abundance upon specific treatments and sample handling steps. The 'Biochemical assay' tab (Supplementary Figure S3) provides dedicated views for such data and currently offers the exploration of Kinobeads (20,21) data, a competition binding assay used to decipher kinase:small molecule interactions, and two formats of cellular thermal shift data (22). Here, we will focus on the description of the Kinobeads data. Beyond this specific example, any relative protein abundance measured as a function of e.g. dose, temperature or time can be explored in the same way.

This view lists all available data for the selected protein, including direct and indirect targets as well as background proteins. In order to filter for binders, different filters are available and can be activated or deactivated. The slider can be used to filter for specific EC_{50} ranges or two goodness of fit values: R^2 (R-square) and BIC (Bayesian information criterion). Dose response curves are fitted using a 4-parameter-log-logistic regression (23). Depending on the protein and the selected filters, the table will show multiple potential small molecules, which exhibit a dose-dependent effect. The experimental data is plotted using black circles, whereas the blue line shows the calculated dose response curve. The orange error bar spans ± one standard error of the EC_{50}.

Interaction network. With the addition of protein–protein interaction (PPI) data from STRING and functional annotation data from KEGG, ProteomicsDB now offers interactive exploration of human PPI networks, enriched with functional information. This information can be accessed via the 'Interaction Network' tab (Supplementary Figure S4). We downloaded subscore-resolved protein network data, detailed interaction types as well as directionality information for Homo sapiens from STRING (24) and combined this data with pathway mappings obtained from KEGG (25) through their REST API. This information was mapped to canonical isoforms using UniProt. Therefore, selection of any protein isoform displays the PPI network and functional annotations of the corresponding canonical isoform. This protein-centric analysis allows the exploration of the PPI network with respect to a protein of interest (POI).

For each resource describing relations between proteins and/or functional categories incorporated into ProteomicsDB, it is possible to select only a subset of the included types of relations for visualization in the 'Relations' menu. This reduces the complexity of the PPI network and focuses the attention on relations of interest. Once all desired relation types for the different resources are selected from the menu on the left (not shown; available in the 'Relations' sub-tab), the interaction network for the selected protein can be displayed by clicking the 'Start analysis' button. This generates a force-directed graph in the network window to the right. Proteins and functional annotations are represented by circular and square nodes, respectively, while edges between nodes represent the relations between them. This initial graph only contains the POI together with the top five interacting proteins as determined by STRING subscore, all relations between them, as well as all functional annotations of the POI. Relations between two nodes without directionality information are merged into a single edge to further reduce redundancy in the graph. If desired, resources previously selected for visualization can be hidden by navigating to the 'Options' menu and toggling the corresponding radio buttons. At any point in time, the graph in the network window can be downloaded as a figure (.svg or .png) or a table (.sif) suitable for import into e.g. Cytoscape (26). An in-depth description to control the visualization can be found in Supplementary Text 5.

Analytics

So far, all analyses focused on the exploration of data relating to a single protein. The 'Analytics' section is designed to enable the analysis of data relating to multiple proteins. Currently, it offers four visualizations covering multi-protein expression pattern analysis, drug selection for single and combination treatments and the exploration of cell viability data.

Expression heatmap. The comparison of protein expression profiles across different tissues, fluids and cell lines can

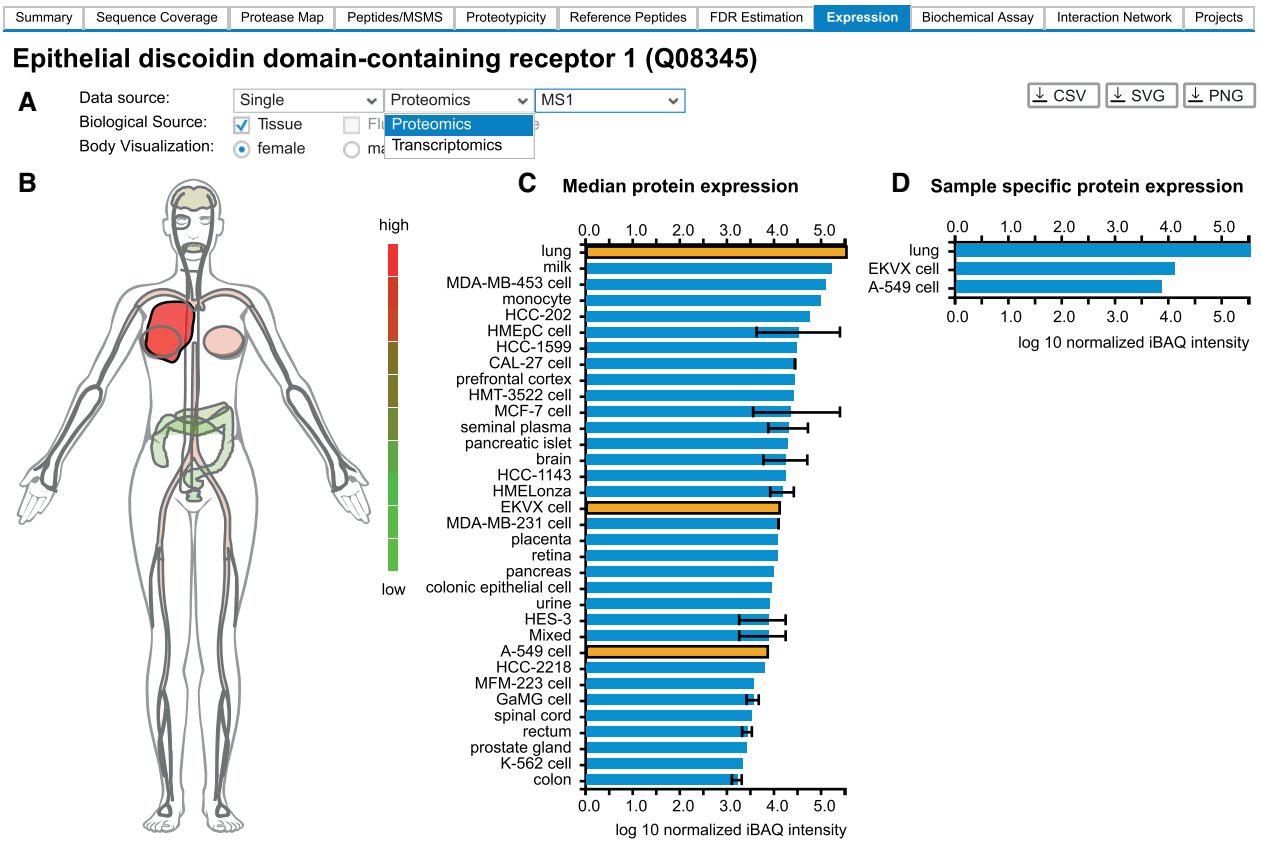

Figure 3. (**A**) ProteomicsDB can visualize expression data from different omics technologies. (**B**) A heatmap-like bodymap superimposing abundance values of tissues, fluids and cell lines (biological sources) onto their respective tissues of origin. (**C**) A bar chart resolving the expression data of b) on the level of their biological source. If multiple measurements for the same biological source are available, the error bar indicates the lowest and highest abundance observed for the selected protein. The bar chart and the bodymap are linked to each other, enabling the selection of either a tissue of origin in the bodymap (highlighted in dark red) or a biological source in the barchart (highlighted in orange). Here, the lung (high expression of DDR1), was selected in the bodymap, which automatically highlights all corresponding tissues and cell lines in the bar chart (EKVX cell and A-549 cell originated from lung tissue). (**D**) A bar chart visualizing sample-specific abundance values of the sources selected in middle bar chart (highlighted in orange). On click on one of the bars, the corresponding sample preparation protocol can be examined.

give rise to new hypotheses and puts protein expression into context. While the expression tab of a single protein allows the analysis of expression patterns over multiple biological sources, it does not enable the analysis of multiple proteins simultaneously. This analysis is possible with the help of the 'Expression heatmap' tab (Figure 4), which shows proteins and biological sources as rows and columns, respectively. For this application, any list of gene names or UniProt identifiers can be supplied to ProteomicsDB. Similar to the 'Expression' tab, a user can choose between multiple available data sources and quantification methods.

Figure 4 shows the resulting heatmap when searching for beta subunits of the proteasome 'PSMB*' in tissues and cell lines using protein expression values estimated by the iBAQ approach. The heatmap is fully interactive and provides multiple options to adjust and explore the data. Additional features of the heatmap are explained in Supplementary Text 6.

Inhibitor potency/selectivity analysis. One topic of great scientific interest is finding the most selective and potent drug against a specific target of interest. For this purpose, ProteomicsDB enables the interactive exploration of dose-dependent competition-binding data in a multi-protein-centric view (Figure 5). Starting with the selection of a protein of interest (here DDR1) the user can filter models based on dose-dependent data available for this protein using several criteria: the EC_{50} range, the R^2 and BIC (similar as in the 'Biochemical assay' tab) (Figure 5A). The pEC50 ($-\log_{10}$ EC_{50} in nM) distribution of all targets meeting the filter criteria for each drug showing a dose-dependent effect on the selected target are plotted in separate violin charts (Figure 5B). The red marker indicates the EC_{50} of the selected protein for each drug. The selectivity of each compound can be evaluated by the numbers above and below the red marker, which depict the number of targets with higher or lower potency compared to the selected protein, respectively.

Users can inspect the pEC50 distribution of all targets for a given drug in an ordered bar chart by selecting the radio button underneath the corresponding violin plot. Targets depicted in (i) green, (ii) blue and (ii) gray are (a) more potent, (b) have similar potency or are (c) at least 10× less potent than the selected target (red), respectively (Figure 5C). This bar chart enables the investigation of all other targets of the selected drug, which could—depending on

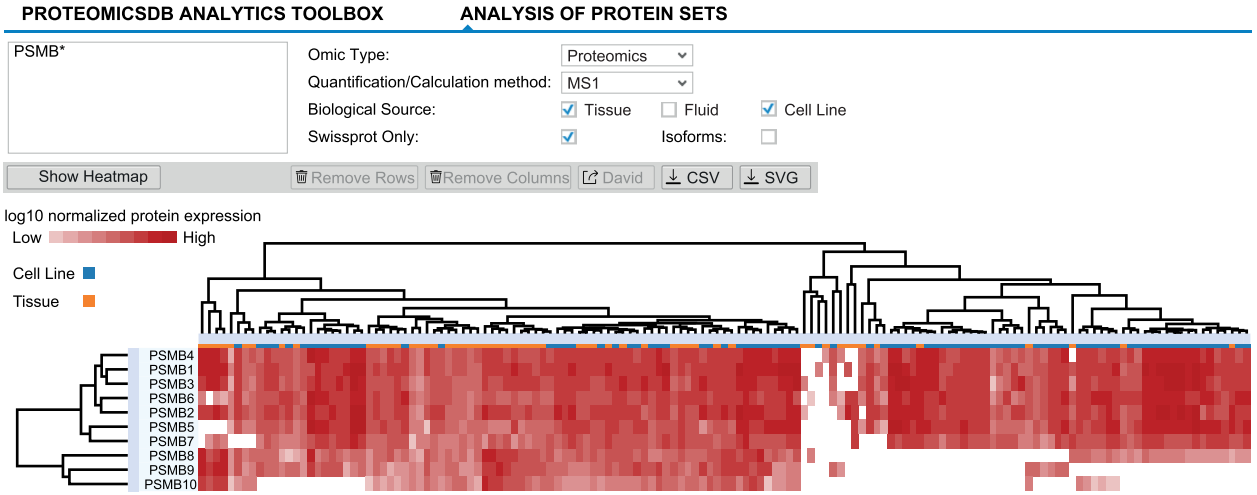

Figure 4. Expression heatmaps of multiple proteins across different tissues, fluids and cell lines can be displayed via the 'Expression heatmap' functionality of the 'Analytics' tab. Proteins and biological sources are shown as rows and columns, respectively. The dendrograms show the result of hierarchically clustering proteins and biological sources, respectively. Branches can be selected and either removed or used to perform GO-enrichment analyses (proteins). Here, all beta-units of the proteasome are displayed, suggesting differential expression of the canonical (expression of PSMB5, 6 and 7) and induced (expression of PSMB8, 9 and 10) proteasome across tissues and cell lines.

its use—increase the risk of unwanted side effects. Individual dose–response plots can be investigated by selecting a specific drug:protein interaction in the bar chart. On click—similar to the 'Biochemical assay' tab—a scatter plot depicting the individual measurements (black dots) and the fitted dose-response model (blue curve) with its estimated EC_{50} and standard error is shown to the right of the bar chart (Figure 5D).

Dose-dependent protein-drug interaction analysis. The potency analysis provides an interface to select an inhibitor for a single protein of interest. However, in some applications, targeting multiple proteins can lead to a more effective treatment (e.g. to suppress resistance formation). The 'Dose-dependent protein–drug interaction analysis' tab (Figure 6) provides an interface to explore the predicted dose-dependent effects of multiple drugs on multiple proteins. This enables the selection of the most promising drug-combination to inhibit a set of proteins, while maintaining the lowest number of off-targets to decrease the chances of unwanted side-effects. Two views are available, which show the predicted target profile of the selected drugs at a certain dose as (i) a protein–drug interaction graph and (ii) a table showing the predicted inhibition effects. Both views are based on the dose-dependent models stored in ProteomicsDB.

The "Proteins" search field accepts sets of protein names. On this basis, all drugs showing at least one inhibitory effect on one of the proteins are taken into consideration. Alternatively, the "Drugs" search field can be used to manually add/select a set of drugs. In case both fields are used, the union of all drugs, either inhibiting at least one of the target proteins or selected manually, is used.

The graph-view shows the protein-drug interaction landscape of the selected drugs. Proteins (circles) are connected to drugs (squares) if a binding/inhibition curve is available for this combination. Each drug selected for the analysis is displayed on the left hand side of the view. The checkbox can be used to disable (hide) a drug from both views. In addition, the dose of each drug can be adjusted by moving the slider or by manually entering a desired drug-concentration. The predicted inhibition of a particular protein in both the graph and the table view are updated in real-time based on the given concentration of a drug. Predicted inhibitory effects are highlighted in the graph by grey edges of varying thickness (proportional to EC_{50}) and blue proteins, the shading of which indicates the level of inhibition. Predicted inhibitory effects are only shown in case they surpass a user-defined cutoff (left vertical slider). In addition to the manual drug concentration, selecting an edge between a protein:drug pair sets the concentration of the drug to the EC_{50} of that interaction.

Cell viability data exploration. With the inclusion of dose-resolved viability data from several large-scale drug sensitivity studies (27–30), ProteomicsDB is now providing tools for fast exploration of dose-response curves quantifying sensitivity and resistance of hundreds of cell lines across hundreds of inhibitors (Figure 7). For each dose-response dataset, ProteomicsDB offers inhibitor- and cell line-centric analysis tools, which allow the identification of sensitive/resistant cell lines for a given inhibitor, while also enabling the identification of potent/impotent inhibitors for a given cell line, respectively. We downloaded dose-resolved viability data from various sources and converted them to relative viabilities, in order to bring the different datasets onto the same scale. Subsequently, the classical symmetric four-parameter log-logistic model was fitted to each inhibitor/cell line combination in each dataset, followed by parameter extraction and calculation of several summary statistics.

After selecting a dataset of interest, analyses can be either cell line- or inhibitor-centric. For this purpose, either one cell line can be chosen, comparing all available inhibitors

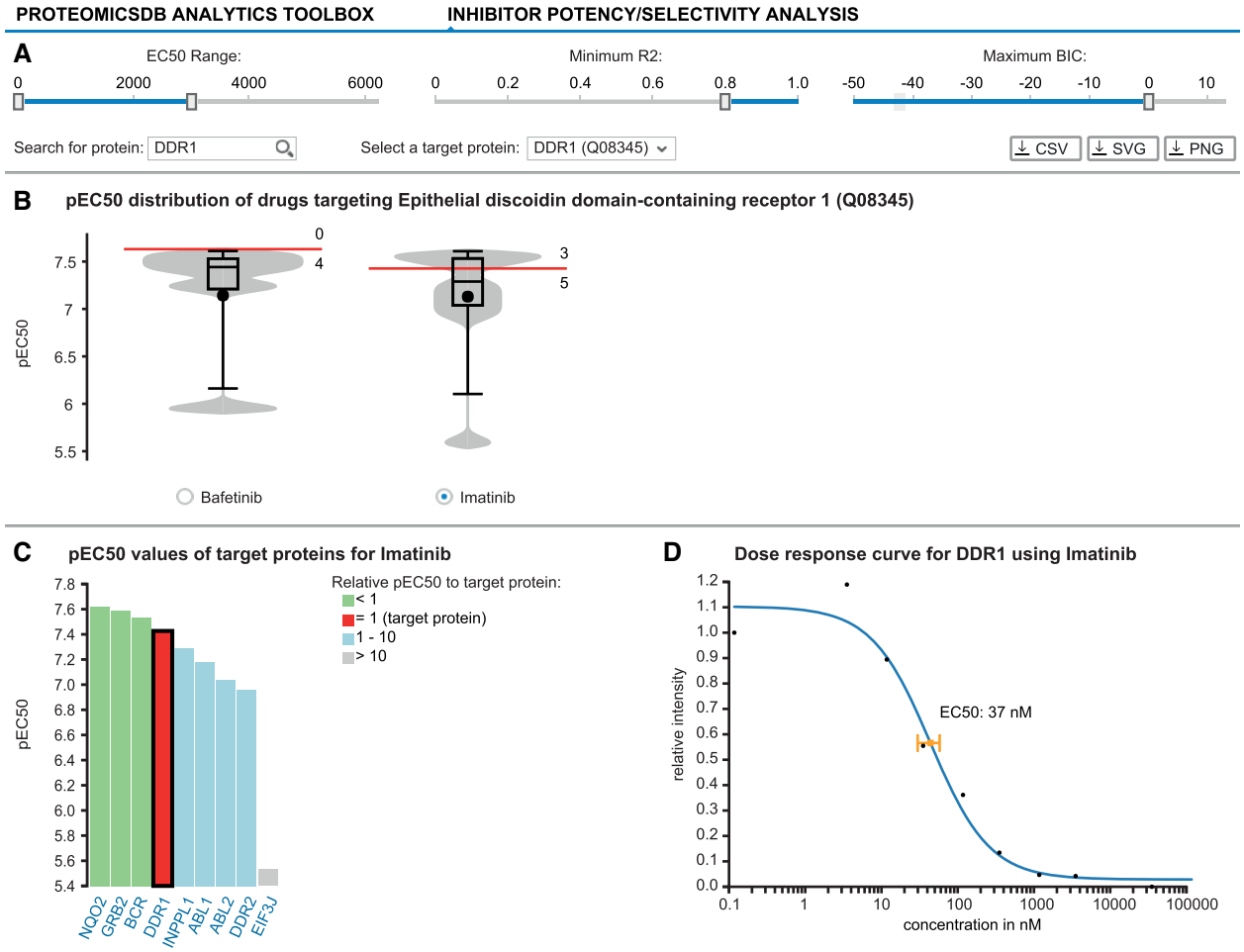

Figure 5. ProteomicsDB enables the exploration of drug selectivity data from various sources. (**A**) Starting with the selection of a target protein, the user can filter fitted selectivity curves using several criteria: the EC_{50} range, the R^2 and BIC. (**B**) Violin plots depicting the pEC_{50} ($-\log_{10}$ EC_{50}) distributions for all compounds targeting the selected protein given the filter criteria from (A). The red marker indicates the EC_{50} of the selected protein for each drug. Numbers above and below the red marker indicate the number of other target proteins with higher or lower potency, respectively. At the time of writing, Bafetinib shows the most potent and selective inhibition of DDR1 with the given filters. (**C**) Bar chart displaying the distribution of pEC_{50} values for Imatinib depicting all of its protein:drug interactions available in ProteomicsDB. (**D**) The underlying raw data and the fitted model can be investigated on click on one of the bars (black border). The scatter plot highlights the EC_{50} for the selected protein:drug pair.

on this single cell line, or, vice versa, an inhibitor can be selected in order to compare the viabilities of all tested cell lines (Figure 7A). Selection of one cell line and one inhibitor is also possible, enabling direct investigation of a specific cell line/inhibitor pair. Using a parallel coordinates plot, it is possible to filter for multiple model parameters and different summary statistics simultaneously (Figure 7B). The exact distribution for each of these variables can be investigated across all cell lines/inhibitors, while selection of one or more cell lines/inhibitors allows the inspection of the underlying dose-resolved viability data (Figure 7C). Visualization of dose-resolved data for multiple cell lines/inhibitors—while essential for judging the reliability of the experimental data—is a feature largely missing from the web portals associated with the original publications (Figure 7D).

FUTURE DIRECTIONS

The large collection of experimental and reference spectra stored in ProteomicsDB opens the door for the development of new functionalities. For example, since the ProteomeTools project covers the entire human proteome with reference spectra, a systematic orthogonal evaluation of protein FDR using synthetic spectra becomes possible. This will provide further validation of protein identification events in ProteomicsDB. Similarly, spectra in ProteomicsDB could be downloaded or compared directly to user data in order to validate the identification of proteins with no prior observations. Furthermore, the combination of reference and experimental spectra and their chromatographic properties will enable the development of tools to guide the development of directed and targeted experiments by custom data-driven spectral library generation. Such tools could make use of the cell line- and tissue-specific protein background identified before and could provide experiment-driven estimates of interfering peptides.

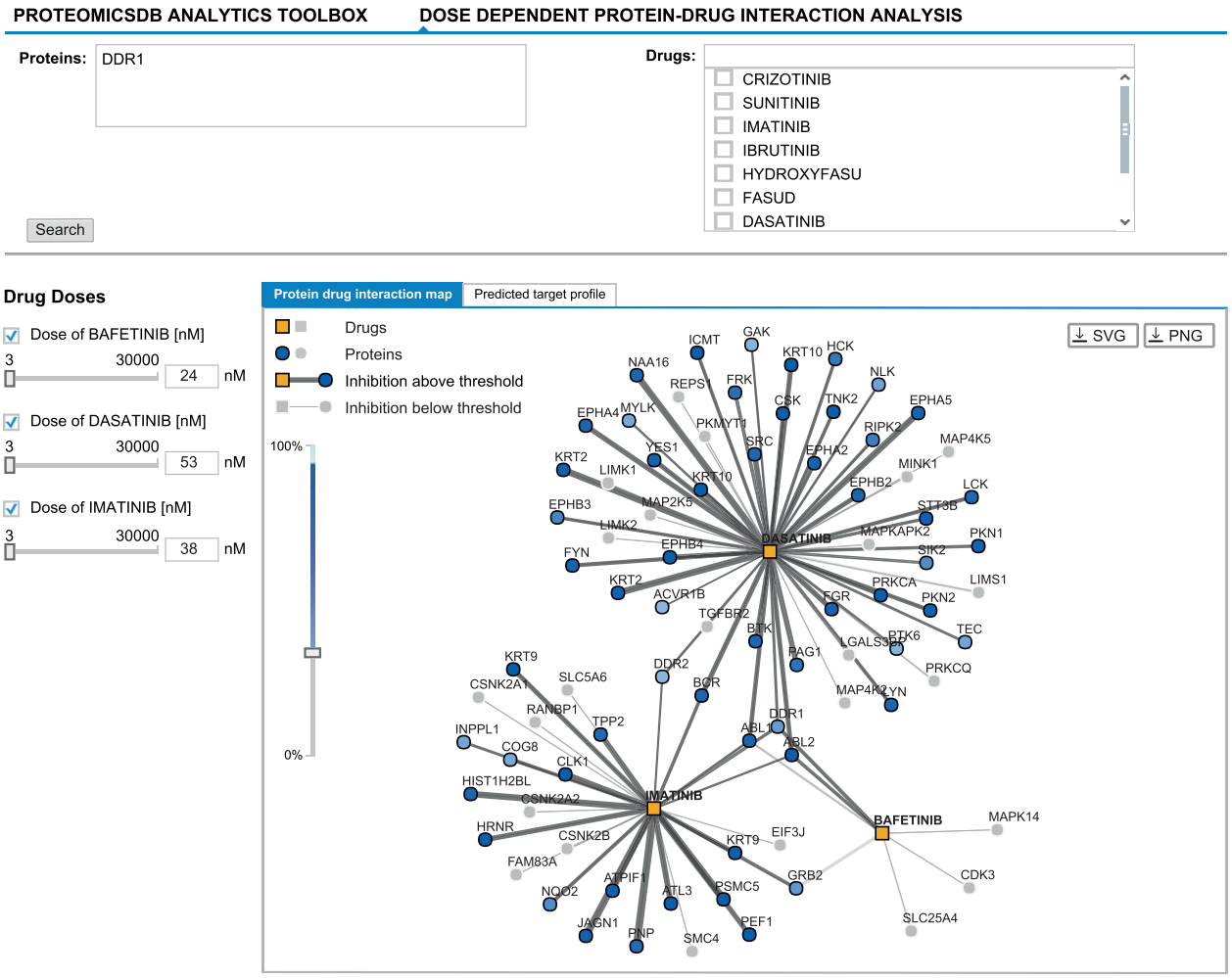

Figure 6. The 'Dose-dependent protein-drug interaction analysis' enables exploring protein:drug interaction data in a multi-drug fashion. It allows the selection of promising drug combinations suitable to inhibit a given target protein (here DDR1). The graph-view shows the protein-drug interaction landscape of selected drugs. Drugs (squares) and proteins (circles) are connected if binding/inhibition curves ('Biochemical Assay' data) are available. Predicted inhibitory effects are highlighted in the graph by dark grey edges of varying thickness (proportional to the EC_{50}) and proteins coloured in different shades of blue (indicates the level of inhibition). Predicted inhibitory effects are only shown in case they surpass a user-defined cutoff (left vertical slider). The concentration of a drug can be adjusted by either clicking an edge (sets the concentration of the drug to the EC_{50} of that interaction), by manually adjusting the concentration using the sliders on the left or by entering the desired concentration into the textbox (left; next to sliders). Again, Bafetinib shows the most selective inhibition of DDR1 at an EC_{50} of 24 nM in comparison to the other two available inhibitors Imatinib (38 nM) and Dasatinib (53 nM).

Due to ProteomicsDB's in-memory architecture, performing database-wide protein inference on all proteins at the same time is possible. This was essential in the development of the picked-FDR approach (31). While a significant proportion of the data stored in ProteomicsDB is already programmatically accessible, an obvious next step is the extension of this service to enable systematic access to all data. By extending the accessibility of data, ProteomicsDB might become an important infrastructure for computational scientists to develop and test new algorithms and for biologists to generate and test new hypotheses.

We will further broaden the scope of ProteomicsDB over the next years. One of the upcoming extensions is to provide protein abundance estimates for other organisms, such as *Mus musculus* and *Arabidopsis thaliana*. While the data model already supports the import of data from other model organisms, the user interface will need to be adjusted. With the ability to store expression patterns from other organisms, cross-species comparisons becomes possible enabling a plethora of questions to be asked and answered.

With the integration of other data sources into ProteomicsDB, the comparative visualization of multiple orthogonal dataset is within reach. We have already started to cross-link visualization tools within ProteomicsDB in the 'Interaction network' tab. However, the integrated data model of ProteomicsDB will also enable the interactive visualization of multiple data sources at once in order to provide a more comprehensive view on omics data. For example, the annotation rows in the 'Expression heatmap' can be extended to show drug selectivity data for proteins or memberships in pathways and signalling cascades, while the annotation columns can show cell viability data for biological sources. Similarly, the 'Expression heatmap' could make use of the imported protein–protein interac-

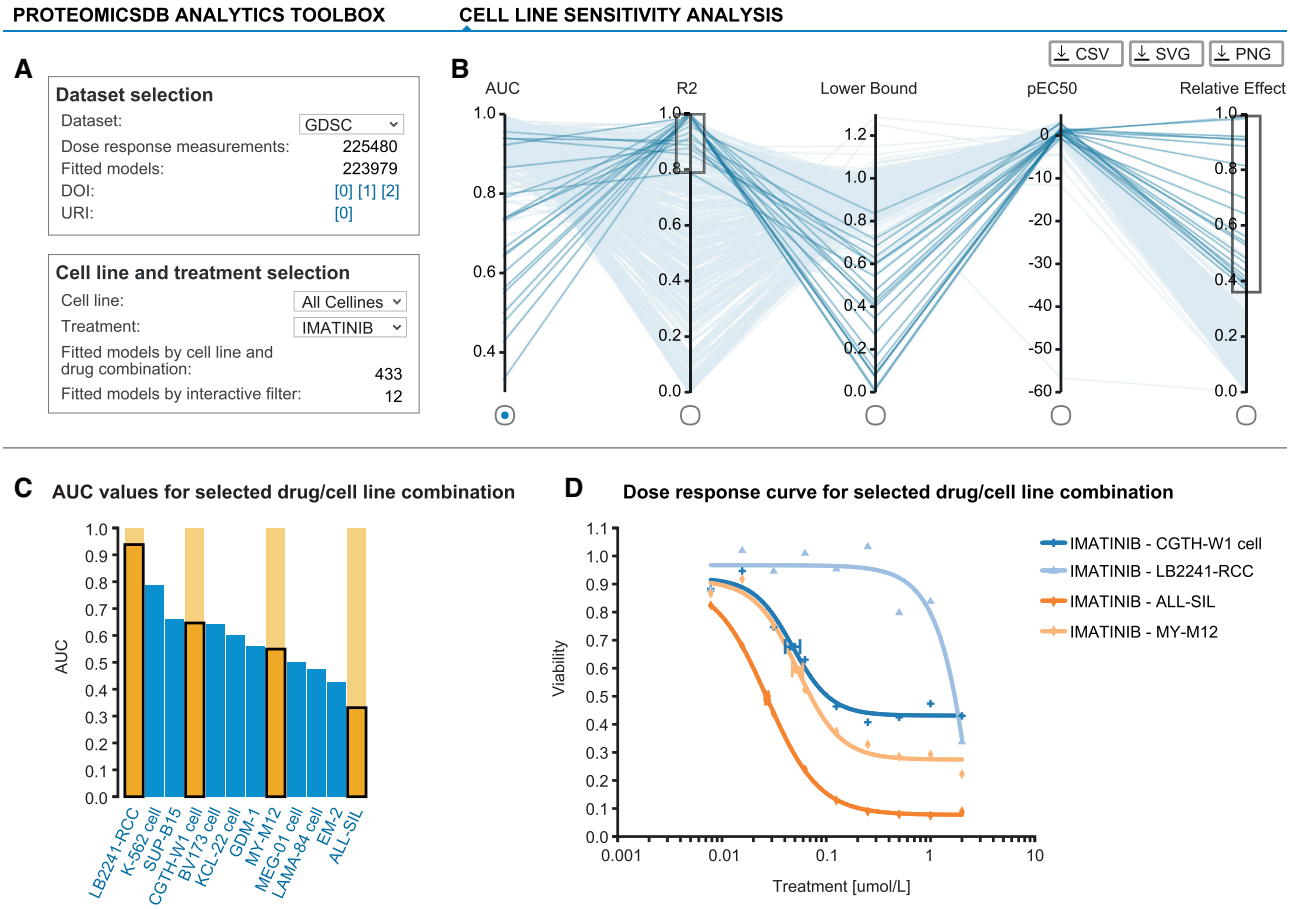

Figure 7. ProteomicsDB incorporates several publicly available large-scale drug sensitivity screens. (**A**) Each drug sensitivity dataset in ProteomicsDB can be explored in a cell-line- or inhibitor-centric way and general statistics are shown for a given selection. (**B**) Users can interactively filter dose-response models based on multiple parameters such as AUC, R^2, lower bound, pEC_{50} and relative effect (percent decrease in viability over the tested concentration range). (**C**) The distribution of a given parameter is visualized in a bar chart on selection of an axis in (B). (**D**) The underlying raw and fitted data can be investigated on click on one or many of the bars (highlighted in orange). The scatter plot highlights the EC_{50} for the selected cell line:drug pairs. The cell lines CGTH-W1, LB2241-RCC, ALL-SIL and MY-M12 show a clear dose-dependent effect on their viability upon Imatinib treatment. However, their EC_{50} values vary, highlighting that these cell lines show differential sensitivity/resistance to Imatinib.

tion and pathway data and allow users to add proteins to the heatmap based on the network neighborhood of the selected protein. Integrated models of drug-selectivity, cell-viability and protein/mRNA expression could be trained to predict treatment outcomes and estimate missing values in either dataset. Given the computational power of the underlying hardware, it is even conceivable to provide the infrastructure and interfaces to users to upload their own data for direct comparison and for direct model training on their phenotypic measurements (32).

AVAILABILITY

ProteomicsDB is available under https://www.ProteomicsDB.org.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

The authors wish to thank all previous members of the ProteomicsDB project team, Hannes Hahne (TUM), Adelya Fatykhova, Anja Gerstmair, David Weese, Emanuel Ziegler, Franz Faerber, Helmut Cossmann, Ingrid Hurbain, Jan Huenges, Joos-Hendrik Boese, Lars Butzmann, Lars Rueckert, Marcus Lieberenz, Mohammed AbuJarour, Wilhelm Becker (SAP), Marcus Bantscheff, Mikhail Savitski, Toby Mathieson (GSK), and the Kuster laboratory for fruitful discussions and technical assistance.

FUNDING

German Federal Ministry of Education and Research (BMBF) [031L0008A] (in part). Funding for open access charge: German Federal Ministry of Education and Research (BMBF) [031L0008A].

Conflict of interest statement. M.W. and B.K. are founders and shareholders of OmicScouts, which operates in the field of proteomics. They have no operational role in the com-

pany and their involvement had no impact on the current study. S.G., J.S., H.-C.E. and S.A. are employees of SAP SE.

REFERENCES

1. Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.
2. Han,X., Aslanian,A. and Yates,J.R. 3rd (2008) Mass spectrometry for proteomics. *Curr. Opin. Chem. Biol.*, **12**, 483–490.
3. Hawkridge,A.M. and Muddiman,D.C. (2009) Mass spectrometry-based biomarker discovery: toward a global proteome index of individuality. *Annu. Rev. Anal. Chem. (Palo Alto Calif.)*, **2**, 265–277.
4. Riffle,M. and Eng,J.K. (2009) Proteomics data repositories. *Proteomics*, **9**, 4653–4663.
5. Perez-Riverol,Y., Alpi,E., Wang,R., Hermjakob,H. and Vizcaino,J.A. (2015) Making proteomics data accessible and reusable: current state of proteomics databases and repositories. *Proteomics*, **15**, 930–949.
6. Vizcaino,J.A., Csordas,A., del-Toro,N., Dianes,J.A., Griss,J., Lavidas,I., Mayer,G., Perez-Riverol,Y., Reisinger,F., Ternent,T. *et al.* (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.*, **44**, D447–D456.
7. Desiere,F., Deutsch,E.W., King,N.L., Nesvizhskii,A.I., Mallick,P., Eng,J., Chen,S., Eddes,J., Loevenich,S.N. and Aebersold,R. (2006) The PeptideAtlas project. *Nucleic Acids Res.*, **34**, D655–D658.
8. Craig,R., Cortens,J.P. and Beavis,R.C. (2004) Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.*, **3**, 1234–1242.
9. Wang,M., Weiss,M., Simonovic,M., Haertinger,G., Schrimpf,S.P., Hengartner,M.O. and von Mering,C. (2012) PaxDb, a database of protein abundance averages across all three domains of life. *Mol. Cell Proteomics*, **11**, 492–500.
10. Schaab,C., Geiger,T., Stoehr,G., Cox,J. and Mann,M. (2012) Analysis of high accuracy, quantitative proteomics data in the MaxQB database. *Mol. Cell Proteomics*, **11**, doi:10.1074/mcp.M111.014068.
11. Färber,F., Cha,S.K., Primsch,J., Bornhövd,C., Sigg,S. and Lehner,W. (2012) SAP HANA database. *ACM SIGMOD Record.*, **40**, 45–51.
12. Wilhelm,M., Schlegl,J., Hahne,H., Gholami,A.M., Lieberenz,M., Savitski,M.M., Ziegler,E., Butzmann,L., Gessulat,S., Marx,H. *et al.* (2014) Mass-spectrometry-based draft of the human proteome. *Nature*, **509**, 582–587.
13. Legrain,P., Aebersold,R., Archakov,A., Bairoch,A., Bala,K., Beretta,L., Bergeron,J., Borchers,C.H., Corthals,G.L., Costello,C.E. *et al.* (2011) The human proteome project: current state and future direction. *Mol. Cell Proteomics*, **10**, doi:10.1074/mcp.M111.009993.
14. Gaudet,P., Michel,P.A., Zahn-Zabal,M., Britan,A., Cusin,I., Domagalski,M., Duek,P.D., Gateau,A., Gleizes,A., Hinard,V. *et al.* (2017) The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res.*, **45**, D177–D182.
15. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
16. Neuhauser,N., Michalski,A., Cox,J. and Mann,M. (2012) Expert system for computer-assisted annotation of MS/MS spectra. *Mol. Cell Proteomics*, **11**, 1500–1509.
17. Zolg,D.P., Wilhelm,M., Schnatbaum,K., Zerweck,J., Knaute,T., Delanghe,B., Bailey,D.J., Gessulat,S., Ehrlich,H.C., Weininger,M. *et al.* (2017) Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods*, **14**, 259–262.
18. Zolg,D.P., Wilhelm,M., Yu,P., Knaute,T., Zerweck,J., Wenschuh,H., Reimer,U., Schnatbaum,K. and Kuster,B. (2017) PROCAL: A set of 40 peptide standards for retention time indexing, column performance monitoring and collision energy calibration. *Proteomics*, doi:10.1002/pmic.201700263.
19. Bantscheff,M., Schirle,M., Sweetman,G., Rick,J. and Kuster,B. (2007) Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.*, **389**, 1017–1031.
20. Lemeer,S., Zorgiebel,C., Ruprecht,B., Kohl,K. and Kuster,B. (2013) Comparing immobilized kinase inhibitors and covalent ATP probes for proteomic profiling of kinase expression and drug selectivity. *J. Proteome Res.*, **12**, 1723–1731.
21. Bantscheff,M., Eberhard,D., Abraham,Y., Bastuck,S., Boesche,M., Hobson,S., Mathieson,T., Perrin,J., Raida,M., Rau,C. *et al.* (2007) Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nat. Biotechnol.*, **25**, 1035–1044.
22. Savitski,M.M., Reinhard,F.B., Franken,H., Werner,T., Savitski,M.F., Eberhard,D., Martinez Molina,D., Jafari,R., Dovega,R.B., Klaeger,S. *et al.* (2014) Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science*, **346**, 1255784.
23. Ritz,C. and Streibig,J.C. (2005) Bioassay Analysis using R. *J. Stat. Softw.*, **12**, doi:10.18637/jss.v012.i05.
24. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M. *et al.* (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.
25. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
26. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
27. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehar,J., Kryukov,G.V., Sonkin,D. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.
28. Iorio,F., Knijnenburg,T.A., Vis,D.J., Bignell,G.R., Menden,M.P., Schubert,M., Aben,N., Goncalves,E., Barthorpe,S., Lightfoot,H. *et al.* (2016) A landscape of pharmacogenomic interactions in cancer. *Cell*, **166**, 740–754.
29. Medico,E., Russo,M., Picco,G., Cancelliere,C., Valtorta,E., Corti,G., Buscarino,M., Isella,C., Lamba,S., Martinoglio,B. *et al.* (2015) The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets. *Nat. Commun.*, **6**, 7002.
30. Rees,M.G., Seashore-Ludlow,B., Cheah,J.H., Adams,D.J., Price,E.V., Gill,S., Javaid,S., Coletti,M.E., Jones,V.L., Bodycombe,N.E. *et al.* (2016) Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat. Chem. Biol.*, **12**, 109–116.
31. Savitski,M.M., Wilhelm,M., Hahne,H., Kuster,B. and Bantscheff,M. (2015) A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. *Mol. Cell Proteomics*, **14**, 2394–2404.
32. Gujral,T.S., Peshkin,L. and Kirschner,M.W. (2014) Exploiting polypharmacology for drug target deconvolution. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 5048–5053.

Supplementary Material

ProteomicsDB

Tobias Schmidt[1,§], Patroklos Samaras[1,§], Martin Frejno[1], Siegfried Gessulat[1,2], Maximilian Barnert[3,4], Harald Kienegger[3,4], Helmut Krcmar[3,4], Judith Schlegl[5], Hans-Christian Ehrlich[2], Stephan Aiche[2], Bernhard Kuster[1,6,*], Mathias Wilhelm[1,*]

[1] Chair of Proteomics and Bioanalytics, Technical University of Munich (TUM), Freising, 85354, Bavaria, Germany

[2] Innovation Center Network, SAP SE, Potsdam, 14469, Germany

[3] Chair for Information Systems, Technical University of Munich (TUM), Garching, 85748, Germany

[4] SAP University Competence Center, Technical University of Munich (TUM), Garching, 85748, Germany

[5] PI HANA Platform Core, SAP SE, Walldorf, 69190, Germany

[6] Bavarian Biomolecular Mass Spectrometry Center (BayBioMS), Technical University of Munich (TUM), Freising, 85354, Bavaria, Germany

[§] Authors contributed equally (shared first)

* To whom correspondence should be addressed. Tel: +49 8161 71 4202; Fax: +49 8161 71 5931; Email: mathias.wilhelm@tum.de, kuster@tum.de

Supplementary text

1 Data model

ProteomicsDB uses multiple ontologies and controlled vocabularies (CV) for internal representation, annotation and meta-data. For this purpose, multiple publicly available CVs such as PSI-MS (1) (terms for proteomics and mass spectrometry), BTO (2) (BRENDA tissue ontology), UO (3) (unit ontology), GO (4) (gene ontology) and sep (5) (terms for chromatographic/separation methods) were imported to enable cross-platform data exchange. Terms not defined in imported ontologies are created manually. Terms, their definitions and relations are stored in the triple-store format, enabling the representation of complex relations. In addition to the CVs, similar storage mechanisms were implemented to enable storing data from other sources, such as STRING (6), KEGG (7) and CHEMBL (8). These ontologies are primarily used in the repository of ProteomicsDB. The main purpose of the repository is to provide a mapping of raw files to samples. Each sample can be annotated to track sample preparation and data acquisition methods. Some of these methods include stable isotope labelling and sample multiplexing technologies such as SILAC (9) and TMT (10), which allow the combined analysis of multiple samples in the same experiment (files). Each of these files can therefore contain quantitative information from different biological systems (e.g. cell lines or treatments). ProteomicsDB supports all major isotope labelling and multiplexing technologies used in the proteomics community and – by providing such a mapping – is able to visualize and analyse identification and quantification results with respect to their biological origin and experimental setting (e.g. full proteome or affinity purification). Samples are then grouped into experiments, which are themselves aggregated to form projects, in order to enable a hierarchical organization of the data.

In addition to the mapping of samples to the biological origin, some experiments require the storage of treatment conditions for each sample. To model the design of an experiment, ProteomicsDB defines treatments as sets of predefined experimental factors, which represent the sequential steps performed during sample preparation. An experimental factor is either a numeric value such as ‘time’ or ‘dose’ associated with a unit, or a (controlled) free-text, such as ‘drugs’ and ‘baits’. For example, a drug treatment is defined by two experimental factors, namely the drug and the concentration at which it was used. Multiple samples assigned to an experiment can be annotated with different drug doses of the same drug, enabling the representation of dose-dependent experiments. Some treatment conditions are predefined but the underlying model allows the generation of any treatment condition necessary for storing the full experimental design. This data model enables the analysis of more complex relationships between samples within an experiment and their corresponding protein and peptide abundance.

As mentioned above, the data model representing peptide and protein identifications is designed to enable the storage of any database search results, including annotated MS/MS spectra. For this purpose, protease-specific in-silico digests of UniProt are stored within ProteomicsDB, enabling the mapping of identified spectra and their associated peptide sequences to proteins. Due to the use of an in-memory database, the sequence file (UniProt) can be efficiently modified and adjusted to newer versions of the human genome. This can be triggered (internally) by uploading an updated UniProt

sequence file, which results in a new in-silico digest. De-validated protein sequences and their corresponding unique peptide identifications are not removed but are instead flagged as outdated and moved to archive tables. ProteomicsDB differentiates between three types of uniqueness: Shared peptides are peptides, which are observed in multiple proteins mapping to at least two gene loci. Unique peptides are classified into isoform- and gene-unique, describing their ability to differentiate different isoforms (peptide only maps to one protein accession) and genes (peptide only maps to one gene locus). New proteins and peptides are added to the in silico-digest and afterwards the peptide uniqueness for every peptide is re-computed. This will also trigger the re-calculation of protein abundances. By incorporating a timestamp for each protein, outdated results are still accessible in the database but are not shown on the web page for reasons of consistency.

In addition to the identified experimental spectra, ProteomicsDB allows the storage of reference spectra acquired from e.g. synthetic peptide standards. These are stored separately from the experimental spectra, but contain similar information on PSM and spectrum level. While experimental spectra are annotated on-the-fly by using a peptide fragmentation model stored in ProteomicsDB, the annotation of fragment ion signals in reference spectra can be stored directly to enable prior manual annotation in order to avoid wrong assignments.

Extracted quantitative data (e.g. peptide precursor ion intensity area or fragment ion reporter ion intensities) are stored at the peptide level to enable protease-specific aggregation of peptide intensities into protein intensities, which can be efficiently re-computed every time the in-silico digest is modified because of modifications of the underlying protein sequence space. For protein abundance estimation, two of the most popular approaches were implemented in ProteomicsDB: the iBAQ (11) well as the top3 intensity (12) methods. Protein abundance is calculated separately for different isotope label types (e.g. light and heavy SILAC), as these represent different samples. In the case of label-free quantification, SILAC and dimethyl labelling experiments (13), iBAQ and top3 intensities are derived from the precursor intensity area measurements, while for isobaric stable isotope labelling (e.g. TMT or iTRAQ), iBAQ and top3 values are calculated based on the respective reporter intensities.

Besides abundance estimates of proteins, ProteomicsDB was extended to enable the storage of other omics data as well. Similar to the proteomics repository, the omics data model organizes samples into experiments and projects. In order to reflect the variety of other omics technologies, this model stores the abundance measure and the measured entity (e.g. transcript) alongside the technology platform and unit provided by the author. Multiple measurements can be attached to a single sample, which facilitates storing e.g. transcript abundances in conjunction with e. g. DNA methylation levels.

The ID conversion functionality – a part of the metadata model – was initially designed to enable the inter-resource conversion of IDs and thus supplements the omics data model. However, due to its generic implementation, it now also serves as an interface to store relations between for example proteins and other proteins, drugs and proteins, as well as a protein’s membership in pathways or regulatory networks. All imported entities are automatically clustered into so-called super-nodes to

enable efficient and easy navigation of this complex graph of different biological entities (e.g. genes, transcripts, proteins, metabolites) and relations between them (e.g. ‘interacts with’ or ‘activates’).

Combining the experimental design and the integrated abundance estimation of proteins enables the calculation of e.g. dose- or temperature-response models. The module storing these relations is flexible in terms of the underlying model (e.g. 4-parameter-log-logistic (14) or linear fits) and enables the visualization, query and analysis of such data. For this purpose, the mathematical formula and it’s parameters are stored in a triple-store-like model, connecting the model-parameters to the observed protein abundance and sample description. The advantages of this data model are showcased by the visualization of competition-binding assay data and CETSA (15) experiments.

In order to be able to take advantage of the plethora of drug sensitivity information available in the public domain, another triple-store-like model is used to store public drug sensitivity datasets in ProteomicsDB. In addition to the protein-centric data described above, ProteomicsDB currently stores phenotypic data from four drug sensitivity screens (16-19). This allows the association of proteomics data on cell lines included in these screens with their sensitivity/resistance towards thousands of drugs, enabling the discovery of pharmacoproteomic markers of drug response. These drug sensitivity datasets are collections of dose-response experiments measuring the viability (the response) of cancer cell lines as a function of the concentration of a specific drug (the dose) or drug combination. They are annotated with meta-data such as URI and DOI in order to link them to the original publication. ProteomicsDB stores high-level information such as dose-response models and their parameters alongside dose-resolved viability data after normalisation, in order to enable the user to estimate the variability of the underlying data. The data model is flexible enough to store experiments with multiple drugs (drug combinations) and can easily be expanded to support e.g. co-culture experiments (cell line combinations) in the future. In cases where raw data are available, this data model allows the comparison of drug sensitivity of the same cell line treated with the same drug across different drug sensitivity datasets, which increases confidence in the data and reduces the number of spurious associations with drug sensitivity in pharmacoproteomic studies further down the line.

2 Proteotypicity

Targeted proteomics is an emerging field and offers reproducible and consistent identification and quantification across many samples. The “Proteotypicity” tab is designed to aid researchers during the selection of appropriate peptides for so-called targeted measurements. For this purpose, peptides are sorted by their experimental proteotypicity. The experimental proteotypicity of a peptide is the ratio of the number of experiments in which this peptide was identified to the number of experiments in which the protein was identified. High values indicate good accessibility during sample preparation (e.g. solubility) and acquisition (e.g. chromatographic retention, MS detection), thus rendering these peptides suitable for targeted measurements. However, the use of multiplexing technologies (e.g. dimethyl or TMT) could either improve or impair the ‘MS-accessibility’ of peptides and thus, proteotypicity is also dependent on the multiplexing strategy used. To differentiate between peptides carrying different labels/modifications, the user can choose for which class of peptides the proteotypicity should be calculated. This use-case strongly benefits from using an in-memory database. Pre-calculating all possible combinations of options would result in a large storage overhead. ProteomicsDB instead calculates the experimental proteotypicity in real-time on request using the search results of more than 19k LC-MS/MS runs. This capability allows the incorporation of further options, such as user defined PSM- or peptide FDR-cutoffs, peptide charge states or biological sources without losing performance.

3 Reference peptides

The “Reference peptides” tab lists all available reference peptides. Similar to the “Peptides/MSMS” tab, each of the available PSMs can be investigated by selecting a peptide of interest. This tab also offers the option to compare two reference spectra against each other since the selected reference spectrum is plotted on the top (similar to the experimental one), and the lower spectrum is again a reference spectrum that can be chosen from a list. This feature can be used to investigate, for example, collision-energy-dependent fragmentation and the change of abundance of fragment ions. In turn, users can take advantage of this information and manually optimize the collision energy for targeted experiments by selecting an energy, which leads to highly abundant heavier fragment ions.

4 FDR estimation

Estimating the proportion of false matches in an experiment is important to assess and maintain the quality of peptide and protein identifications. The most widely used strategy to estimate the false discovery rate (FDR) (20) in proteomics is the target-decoy approach. However, the classical target-decoy approach overestimates protein FDR in large collections of data due to an accumulation effect of false positive peptide identifications. ProteomicsDB was essential in the development of the so-called “picked protein FDR” approach (21) which alleviates this issue. The picked protein FDR approach treats target and decoy sequences of the same protein as a pair rather than as individual

entities and chooses either the target or the decoy sequence depending on which receives the highest score. This method avoids the accumulation of decoy identifications in classical target-decoy-based models for large datasets.

The “FDR estimation” tab (Supplementary Figure S2) shows the results of applying this method. Two visualizations are available to judge the gene- (left column) or isoform-level (right-column) identification of a protein. The upper plot in both columns depicts the database-wide distribution of protein scores. Highlighted are the protein of interest (blue dot) and its respective decoy (orange dot). If the target was observed with a higher protein-score, the q-value (22) indicates the confidence of this identification. Otherwise, the corresponding decoy was “identified” with a higher score, rendering the identification of the target invalid. The bottom plots show the distribution of either the gene-unique or isoform-unique target and decoy PSMs as a function of their identification score. Here, DDR1 was confidently identified on gene-level, but since no observed isoform-specific peptides exist for the canonical version of DDR1 (all observed peptides are shared with at least one other isoform of DDR1), the explicit identification of the canonical isoform Q08345 of DDR1 cannot be ascertained.

Briefly, each dataset/experiment in ProteomicsDB is analyzed separately with Mascot and Maxquant/Andromeda, applying the standard 1% PSM FDR per individual raw file. Afterwards, the data in ProteomicsDB is aggregated and peptides shorter than 7 amino acids are filtered out, before peptides are grouped by length. For each peptide length bin, a 5% local FDR is applied, rejecting all PSMs below the respective search engine score threshold, which is much more stringent than a global FDR criterion. The resulting number of proteins are in line with the originally reported number of proteins at 1% protein FDR (23).

In order to account for the accumulation of false positives on a database-wise level, the ‘picked’ protein FDR approach described by Savitski et al (21) is used on the aggregated data in ProteomicsDB. Briefly, for each identified target and decoy protein hit, the Q-score (- log10 q-value) of the most confidently identified gene-specific peptide is used as the protein score. For each protein, the scores of the target is compared to the score of its corresponding decoy. If the score of the decoy was higher than that of the target, the protein was counted as a decoy hit and the target was disregarded. If the score of the target was higher than that of the decoy, the protein was counted as a target hit and the decoy was disregarded. Finally, the protein FDR was calculated using the remaining target and decoy hits in the same way as for the classical target-decoy approach.

5 Features of the interaction network visualization

Scrolling up/down or double-clicking/shift-double-clicking inside the network window zooms in/out of the network. Dragging inside the network window allows repositioning of the entire network. Hovering over a node highlights all nodes of the network connected to it and the corresponding relations, visualized as edges. Hovering over an edge highlights the two nodes it connects. Clicking on an edge displays a layover with meta-information considering the selected relation. Since multiple relations without directionality information are merged into a single edge, this table can contain multiple entries. One/more nodes can be selected by clicking/shift-clicking on them, which fixes their positions with respect to the force-directed layout and displays meta-information for protein nodes in the 'Node information' section to the left of the interaction graph. This section is automatically populated with links to different parts of ProteomicsDB, which enables quick navigation to protein-centric (e.g. biochemical assay data) or multi-protein-centric (e.g. heatmap) analyses. Selected nodes can be 1) moved by dragging one of them to a new position, 2) unselected by clicking anywhere in the network window but on the graph, 3) deleted by clicking the trashcan in the control panel (top-left corner of the network window) and 4) unfixed (exposed to force-directed layout) by clicking the lock in the control panel. After selecting a single protein node, the network can be expanded by clicking the '+' in the control panel. This will add the next five interaction partners/functional annotations per resource (ordered by their subscores from high to low) to the network, enabling its dynamic exploration.

6 Features of the expression heatmap visualization

If space permits, the leafs of the dendrogram are expanded to show the respective cell and protein annotation. Specific branches of the trees can be selected and removed from the visualization by selecting one and then using the "Remove Rows" or "Remove Columns" button at the top. In addition, one or multiple branches in the protein-dendrogram can be selected in order to perform a GO-enrichment analysis by using the "DAVID" (24,25) link at the top. The biological origin of the samples is color-coded in blue, orange and green for tissues, body fluids and cell lines. The size of the heatmap is defined by the size of the browser. In order to investigate specific patterns in the heatmap, navigation is possible via click-and-drag (panning) and the mouse-wheel (zooming).

Supplementary Figures

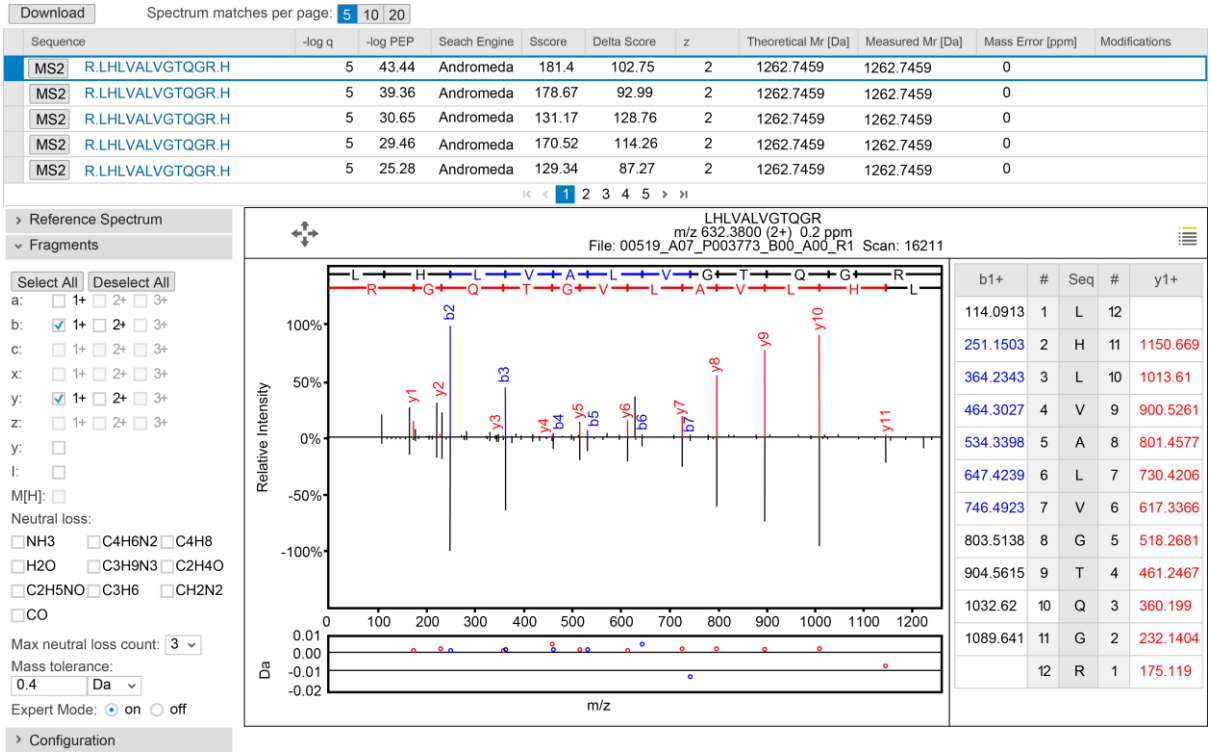

Figure S1 The spectrum viewer visualizes and dynamically annotates available experimental peptide spectrum matches (top table) stored in ProteomicsDB according to user-specified settings (tab to the left). A mirror representation of a reference spectrum (bottom spectrum) is shown automatically if available and can be used to validate the identification of a peptide. The sequence and fragmentation table (superimposed on the right) depicts the presence of y- (red) and b-ions (blue). Here, the peptide LHLVALVGTQGR of DDR1 shows a very strong correlation to its reference spectrum acquired as part of the ProteomeTools project, confirming its valid identification.

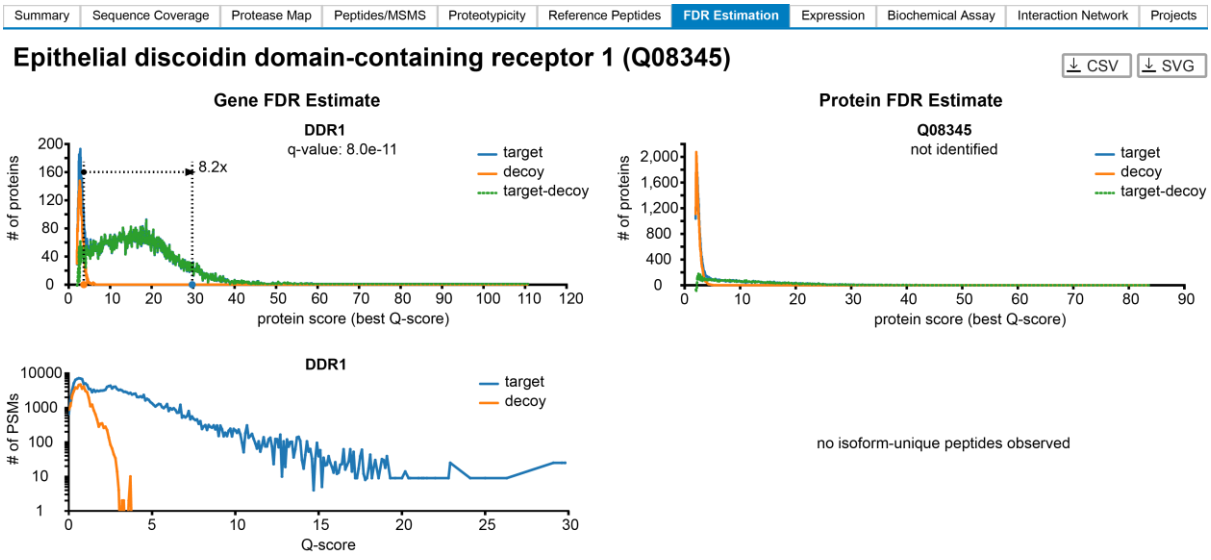

Figure S2 The FDR estimation for a given protein (here: DDR1/Q08345) using the ‘picked protein FDR’ approach can be inspected on gene- (DDR1; left column) or isoform-level (Q08345; right-column). The upper plots depict the database-wide distribution of protein scores on either of these levels. The q-value indicates the confidence of a given identification, if the target was observed with a higher protein-score. The bottom plots show the distribution of either the gene-unique or isoform-unique target and decoy PSMs as a function of their identification score. Here, DDR1 was confidently identified on the gene-level, but the explicit identification of the canonical isoform Q08345 cannot be ascertained.

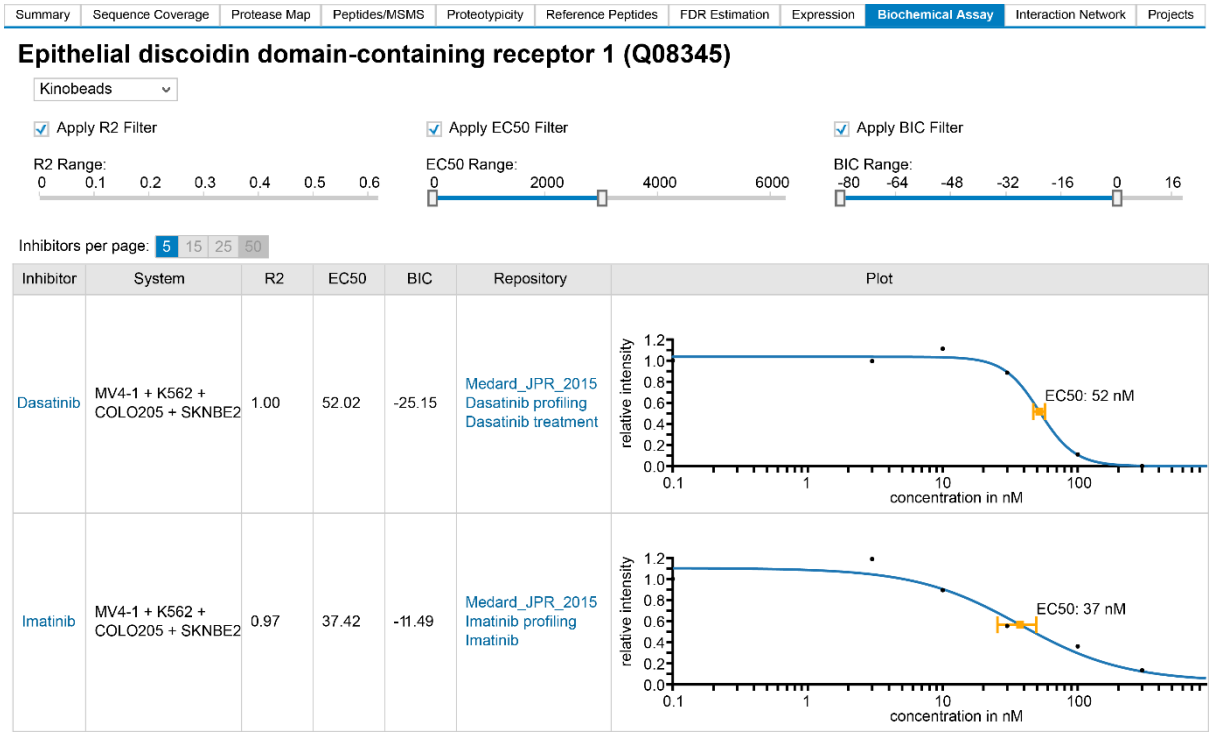

Figure S3 Chemoproteomics technologies such as Kinobeads enable the elucidation of the target space of drugs in a systematic fashion. ProteomicsDB visualizes raw and fitted data of such experiments for a given protein in the “Biochemical Assay” tab. Users can filter fitted curves for specific EC_{50} ranges and two measures quantifying goodness of fit: R^2 and BIC. Here, Dasatinib and Imatinib show a dose-dependent effect on the relative intensity (residual binding to Kinobeads) of DDR1, suggesting a high affinity of these compounds to DDR1 (EC_{50} of 52 nM and 37 nM, respectively).

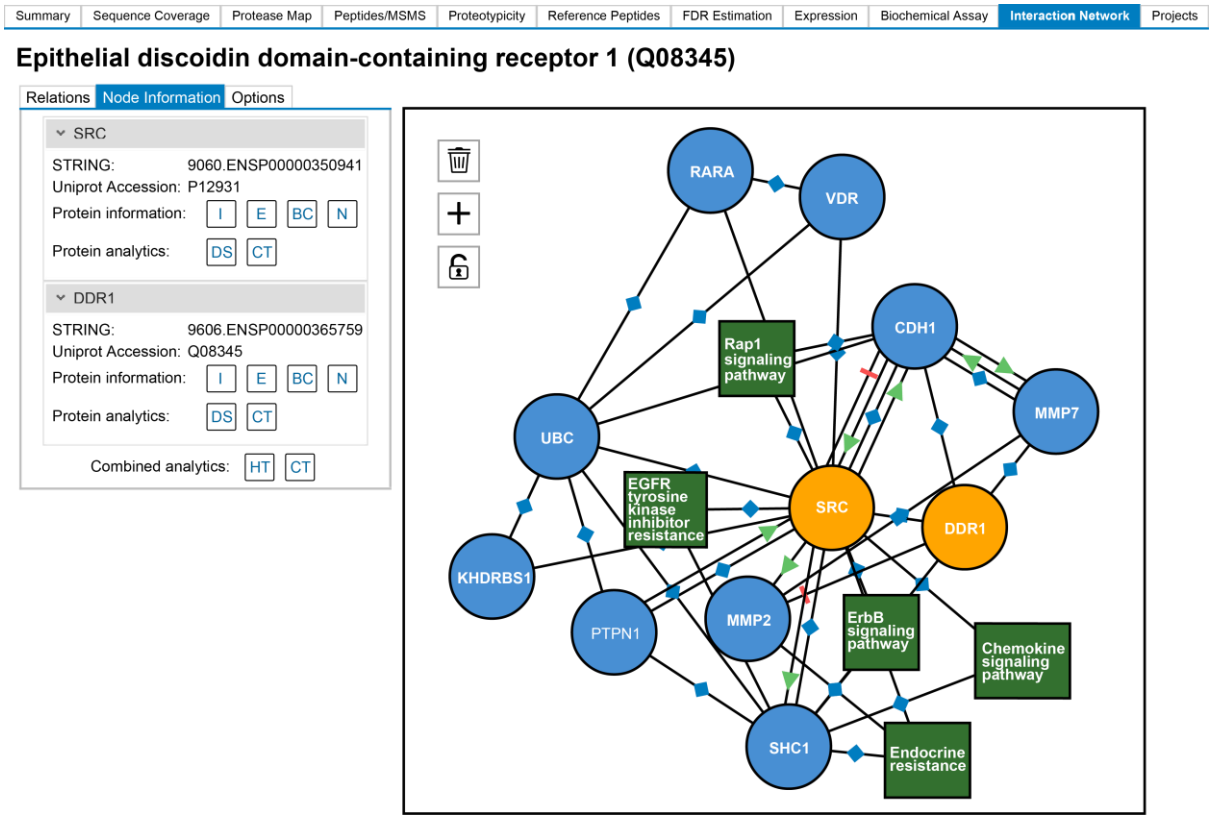

Figure S4 The interaction graph allows to quickly navigate protein-protein interaction networks and pathway annotations. Proteins and pathways are shown as blue spheres and green rectangles, respectively. Shapes on edges inform about the type of interaction: blue diamonds symbolize known interactions without directionality information as well as functional annotations, red bars indicate inhibitory effects (e.g. SRC inhibits CDH1) and green arrows represent activating effects between two nodes (e.g. SRC activates MMP2). Selected subgraphs and/or proteins (marked in orange) can be directly used for multi-protein centric analyses via “Combined analytics” links (HT: Heatmap; CT: Combination treatment) in the “Node Information” panel on the left, which also enables quick navigation to protein-centric analyses (I: Summary page; E: Expression; BC: Biochemical assay; N: Interaction network; DS: Drug selectivity; CT: Combination treatment).

References

1. Mayer, G., Montecchi-Palazzi, L., Ovelleiro, D., Jones, A.R., Binz, P.A., Deutsch, E.W., Chambers, M., Kallhardt, M., Levander, F., Shofstahl, J. *et al.* (2013) The HUPO proteomics standards initiative- mass spectrometry controlled vocabulary. *Database (Oxford)*, **2013**, bat009.
2. Gremse, M., Chang, A., Schomburg, I., Grote, A., Scheer, M., Ebeling, C. and Schomburg, D. (2011) The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res*, **39**, D507-513.
3. Gkoutos, G.V., Schofield, P.N. and Hoehndorf, R. (2012) The Units Ontology: a tool for integrating units of measurement in science. *Database (Oxford)*, **2012**, bas033.
4. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25-29.
5. Gibson, F., Hoogland, C., Martinez-Bartolome, S., Medina-Aunon, J.A., Albar, J.P., Babnigg, G., Wipat, A., Hermjakob, H., Almeida, J.S., Stanislaus, R. *et al.* (2010) The gel electrophoresis markup language (GelML) from the Proteomics Standards Initiative. *Proteomics*, **10**, 3073-3081.
6. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M. *et al.* (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, **37**, D412-416.
7. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*, **45**, D353-D361.
8. Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*, **40**, D1100-1107.
9. Ong, S.-E., Kratchmarova, I. and Mann, M. (2003) Properties of13C-Substituted Arginine in Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC). *Journal of Proteome Research*, **2**, 173-181.
10. Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T. and Hamon, C. (2003) Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Analytical Chemistry*, **75**, 1895-1904.
11. Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature*, **473**, 337-342.
12. Silva, J.C., Gorenstein, M.V., Li, G.Z., Vissers, J.P. and Geromanos, S.J. (2006) Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics*, **5**, 144-156.
13. Boersema, P.J., Raijmakers, R., Lemeer, S., Mohammed, S. and Heck, A.J. (2009) Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat Protoc*, **4**, 484-494.
14. Ritz, C. and Streibig, J.C. (2005) Bioassay Analysis using R. *Journal of Statistical Software*, **12**.
15. Martinez Molina, D., Jafari, R., Ignatushchenko, M., Seki, T., Larsson, E.A., Dan, C., Sreekumar, L., Cao, Y. and Nordlund, P. (2013) Monitoring drug target engagement in cells and tissues using the cellular thermal shift assay. *Science*, **341**, 84-87.
16. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, J., Kryukov, G.V., Sonkin, D. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603-607.
17. Iorio, F., Knijnenburg, T.A., Vis, D.J., Bignell, G.R., Menden, M.P., Schubert, M., Aben, N., Goncalves, E., Barthorpe, S., Lightfoot, H. *et al.* (2016) A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, **166**, 740-754.

18. Medico, E., Russo, M., Picco, G., Cancelliere, C., Valtorta, E., Corti, G., Buscarino, M., Isella, C., Lamba, S., Martinoglio, B. *et al.* (2015) The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets. *Nat Commun*, **6**, 7002.
19. Rees, M.G., Seashore-Ludlow, B., Cheah, J.H., Adams, D.J., Price, E.V., Gill, S., Javaid, S., Coletti, M.E., Jones, V.L., Bodycombe, N.E. *et al.* (2016) Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol*, **12**, 109-116.
20. Beausoleil, S.A., Villen, J., Gerber, S.A., Rush, J. and Gygi, S.P. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol*, **24**, 1285-1292.
21. Savitski, M.M., Wilhelm, M., Hahne, H., Kuster, B. and Bantscheff, M. (2015) A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. *Mol Cell Proteomics*, **14**, 2394-2404.
22. Kall, L., Storey, J.D., MacCoss, M.J. and Noble, W.S. (2008) Posterior error probabilities and false discovery rates: two sides of the same coin. *J Proteome Res*, **7**, 40-44.
23. Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H. *et al.* (2014) Mass-spectrometry-based draft of the human proteome. *Nature*, **509**, 582-587.
24. Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, **37**, 1-13.
25. Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, **4**, 44-57.

Published online 30 October 2019 *Nucleic Acids Research, 2020, Vol. 48, Database issue* ***D1153–D1163***
doi: 10.1093/nar/gkz974

ProteomicsDB: a multi-omics and multi-organism resource for life science research

Patroklos Samaras[1], Tobias Schmidt[1], Martin Frejno[1], Siegfried Gessulat[1,2], Maria Reinecke[1,3,4], Anna Jarzab[1], Jana Zecha[1], Julia Mergner[1], Piero Giansanti[1], Hans-Christian Ehrlich[2], Stephan Aiche[2], Johannes Rank[5,6], Harald Kienegger[5,6], Helmut Krcmar[5,6], Bernhard Kuster[1,7,*] and Mathias Wilhelm[1,*]

[1]Chair of Proteomics and Bioanalytics, Technical University of Munich (TUM), Freising, Bavaria, Germany, [2]Innovation Center Network, SAP SE, Potsdam, Germany, [3]German Cancer Consortium (DKTK), Partner Site Munich, Munich, Germany, [4]German Cancer Research Center (DKFZ), Heidelberg, Germany, [5]Chair for Information Systems, Technical University of Munich (TUM), Garching, Germany, [6]SAP University Competence Center, Technical University of Munich (TUM), Garching, Germany and [7]Bavarian Biomolecular Mass Spectrometry Center (BayBioMS), Technical University of Munich (TUM), Freising, Bavaria, Germany

Received September 14, 2019; Revised October 11, 2019; Editorial Decision October 11, 2019; Accepted October 15, 2019

ABSTRACT

ProteomicsDB (https://www.ProteomicsDB.org) started as a protein-centric in-memory database for the exploration of large collections of quantitative mass spectrometry-based proteomics data. The data types and contents grew over time to include RNA-Seq expression data, drug-target interactions and cell line viability data. In this manuscript, we summarize new developments since the previous update that was published in Nucleic Acids Research in 2017. Over the past two years, we have enriched the data content by additional datasets and extended the platform to support protein turnover data. Another important new addition is that ProteomicsDB now supports the storage and visualization of data collected from other organisms, exemplified by *Arabidopsis thaliana*. Due to the generic design of ProteomicsDB, all analytical features available for the original human resource seamlessly transfer to other organisms. Furthermore, we introduce a new service in ProteomicsDB which allows users to upload their own expression datasets and analyze them alongside with data stored in ProteomicsDB. Initially, users will be able to make use of this feature in the interactive heat map functionality as well as the drug sensitivity prediction, but ultimately will be able to use all analytical features of ProteomicsDB in this way.

INTRODUCTION

ProteomicsDB (https://www.ProteomicsDB.org) is an in-memory database initially developed for the exploration of large quantities of quantitative human mass spectrometry-based proteomics data including the first draft of the human proteome (1). Among many features, it allows the real-time exploration and retrieval of protein abundance values across different tissues, cell lines, and body fluids via interactive expression heat maps and body maps. Today, ProteomicsDB supports multiple use cases across different disciplines and covering a wide range of data (2). For instance, tandem mass spectra, peptide identifications and peptide proteotypicity values can be used as starting points to develop targeted mass spectrometry assays. Because of the recent incorporation of a large amount of reference spectra from the ProteomeTools project (3,4) as well as spectra predicted by the artificial intelligence Prosit (5), both experimental and reference spectra can be used for assay development and to validate the identification of so far unobserved, or in fact any proteins. The integration of phenotypic data allows the exploration of the dose-dependent effect of drugs of interest (e.g. clinically approved drugs) on multiple cell lines (6–9). The dynamic identifier mapping in ProteomicsDB allows the integration of transcriptomics data from e.g. the Human Protein Atlas project (10) and Bgee (11), and thus facilitates the automated integration of different data sources within ProteomicsDB. This, in turn, allows the development of new tools. A wide range of drug-target interaction data can be visualized in ProteomicsDB as well, which enables the exploration of combination treatments in a dose-dependent protein-drug interaction graph *in-silico*.

*To whom correspondence should be addressed. Tel: +49 8161 71 4202; Fax: +49 8161 71 5931; Email: mathias.wilhelm@tum.de
Correspondence may also be addressed to Bernhard Kuster. Email: kuster@tum.de

ProteomicsDB is becoming an increasingly valuable resource in (proteomic) life science research, evidenced by the increasing number of external resources linking to ProteomicsDB, such as UniProt (12) and GeneCards (13), as well as resources making use of our application programming interface (API) to show e.g. protein expression information, as done by OmniPathDB (14) and Gene Info eXtension (GIX) (15).

In this version, we expanded the data content of ProteomicsDB by including additional publically available as well as in-house generated proteomic and transcriptomic studies. Furthermore, we expanded the drug-target interaction data now covering ∼1500 kinase inhibitors and tool compounds. The cell line viability data were enriched with an additional large dataset (16) now covering >20 000 drugs against 1500 cell lines. We further increased the amount of protein property information that is stored in ProteomicsDB, such as 13 000 melting points of proteins obtained by thermal proteome profiling (17). In addition, we expanded the biochemical assays section to include protein turnover data with synthesis and degradation curves for >6000 proteins. We further increased the number of reference tandem mass spectra in ProteomicsDB to >5 million from synthetic peptides and 40 million from predictions, which, in total, are represented by 3 billion fragment ions.

RESULTS

Overview

ProteomicsDB aims to provide real-time analytical functions to users, including computationally challenging tasks. For this purpose, ProteomicsDB was carefully designed and organized (Figure 1). It consists of a production unit, a computing unit, and a storage unit, all intra-connected via a 16Gbit local network. The production unit hosts the production server as well as the entire development and testing environment. The computing unit is one machine with a fully dockerized environment which currently handles two main tasks. First, an R server that handles R-procedures from ProteomicsDB such as the clustering available in the heat map. Second, a docker container with various services handling requests to our deep learning tool Prosit which is connected to two NVIDIA P100 GPU cards.

Over the past two years, the user interface and data content of ProteomicsDB were updated to accommodate new requirements such as hosting data from other organisms. Figure 2A shows the changes that were made to the front page such that users can select the organism of interest. Parts of the webpage have been renamed to be more generic and cover every organism, such as the 'Human Proteins' tab, which was renamed to 'Proteins'. The front page statistics lists new information about the quantity of the data that is available for the chosen organism, including information about tissue coverage, quantitative multi-omics expression values, biochemical assay measurements as well as cell viability measurements. The main pane of the front page was redesigned to show the main features of the platform. It is now split into two sections. The left section provides direct links to the protein centric visualizations, the analytics toolbox, the new feature to upload custom data and a link to Prosit. The right section includes links that trigger the selection of the corresponding organism. To make organism selection available throughout the web interface, we additionally adjusted the left sidebar to show one icon per available organism. The 'Feedback' button that that was previously located in that position was transferred to the right pane below the 'Help' button. In light of these changes, all internal procedures and endpoints (e.g. API) were adjusted to support the new data types and organisms.

Figure 2B depicts the data expansion in ProteomicsDB since 2017, grouped by categories. By re-analyzing and uploading more publically available proteomics studies, we increased the tissue coverage of ProteomicsDB by ∼70 human tissues and cell lines (+∼30%), to a total of almost 300 tissues and cell lines. The broader coverage of biological systems has direct impact on visualizations like the human body map or expression heat map. The plethora of data in ProteomicsDB allows not only the further online exploration of the proteome and its properties but also enables the development of new tools integrating different omics data sources. Currently, human proteomics and transcriptomics data are available for ∼17 000 genes and ∼60 tissues (Figure 2C, D). This large overlap enabled the implementation of a new missing value imputation approach which makes use of transcriptomics or proteomics data to estimate the presence and abundance of protein or RNA not covered in individual data sets. For ∼13 000 proteins, additional information derived from other biochemical assays such as melting behavior or synthesis or degradation curves are available. By integrating additional publicly available datasets, the overlap at the tissue- and protein level will increase further over the next years and eventually cover all the >1000 (cancer) cell lines for which we already have cell viability data. This, in turn, will aid the development of a better understanding of the molecular factors that govern the life of a particular cell.

New biochemical assay data, covering more protein properties

In addition to importing additional expression profile datasets, we further extended our biochemical assay portal by integrating the results of three additional studies covering target information of small molecule kinase inhibitors, melting (thermal aggregation) behavior of proteins and turnover data. First, in order to extend knowledge on druggable protein kinases (18), we imported ∼500 000 kinase inhibitor dose-response curves (Figure 3) covering 243 kinase inhibitors that are either approved for use or are in clinical trials (18) and ∼1300 tool compounds targeting kinases (unpublished). This data gives users a broader coverage and thus more options to select inhibitors to study a particular protein kinase. Various learnings might arise from such analysis, such as assessing the repurposing potential of clinical kinase inhibitors. Moreover, users can discover an appropriate molecule/inhibitor with respect to potency and selectivity to study the function of a particular kinase (19). Another use case is to identify inhibitors which share the same target(s) but have different off-targets, which can be used to identify and study the core signaling pathway of the shared target(s) or general on-target effects (18). In addition, the biochemical assay data and tools provided

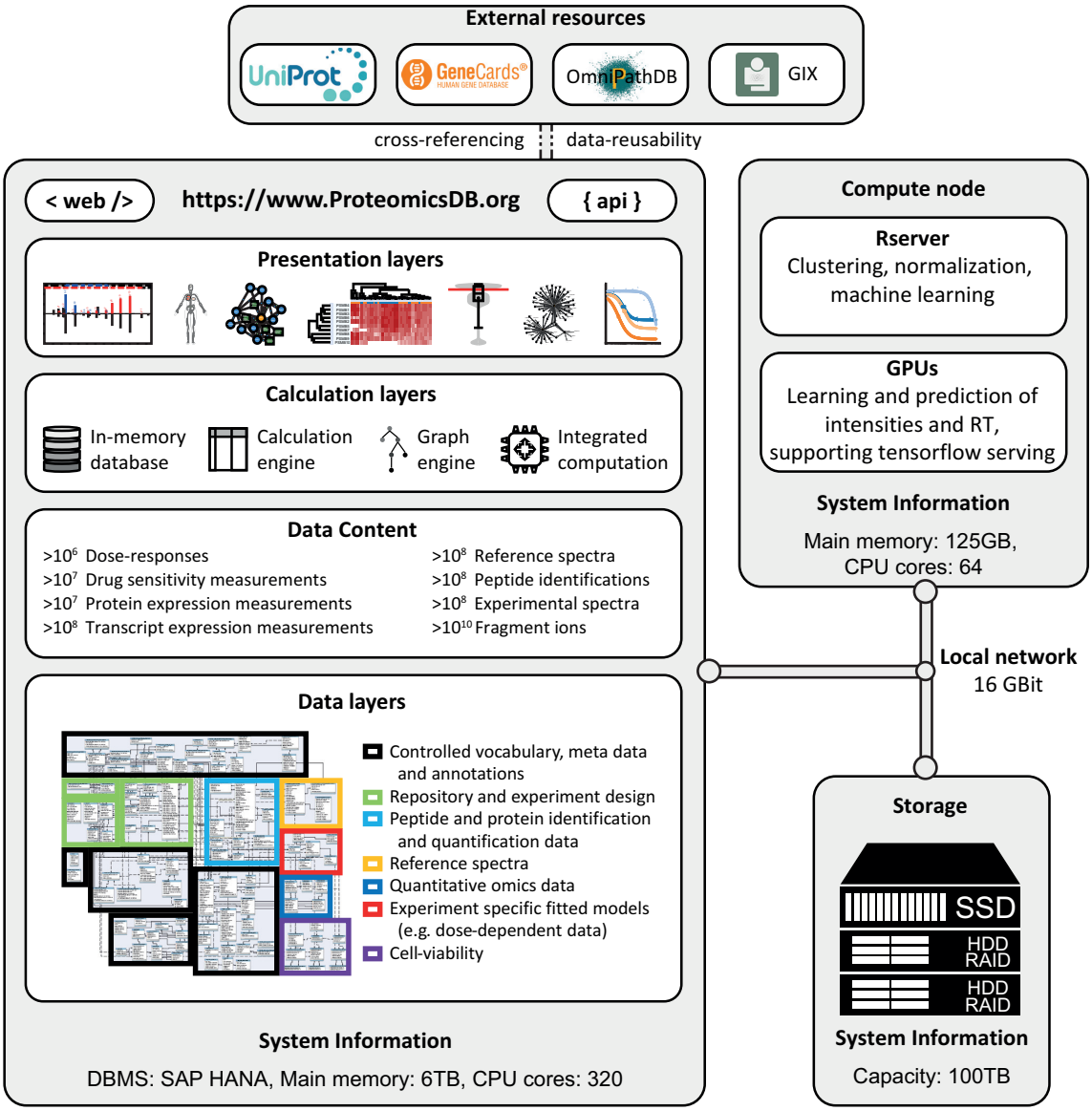

Figure 1. The architecture of ProteomicsDB. The production unit hosts the SAP HANA in-memory database management system which involves three of the presented layers: the data layers, data content and the calculation layers. Parts of the calculation layers are shared between the production unit and the compute node, such as the clustering and correlation procedures for the interactive expression heat map which are calculated by the Rserver. Part of the data content is stored in the network storage unit, so that data are always available throughout the network if needed. The entire infrastructure is intra-connected via a 16 Gbit bandwidth local network that enables rapid communication and data transfer between units.

in ProteomicsDB (e.g. Inhibitor potency/selectivity analysis) can be used to discover new lead compounds for medicinal chemistry programs targeting a specific kinase of interest (20,21). The dose-response curves can be explored in the 'Biochemical assay' tab of the protein details view. This view allows users to filter the data by different properties, so that only compounds that fit the desired criteria will be displayed. For all curves, full experimental designs are stored for the users to browse and explore. For dose-response curves that belong to studies that are not published yet, the curve information is available but the experimental design, although fully imported, will only be shown when these studies are published. Second, the meltome data of ProteomicsDB was enriched with another study that covers the protein melting properties for many organisms (unpublished). Therefore, users can more thoroughly study the effect of temperature on selected proteins. We now cover the melting properties of ∼13 000 human proteins. ProteomicsDB thus provides an extensive resource and data-driven guidance on which temperature range should be used for e.g. a thermal shift assay or which temperature would be suitable for an isothermal dose response assay (ITDR). Third, we introduced a new assay type in the 'Biochemical Assay' tab which covers data from protein turnover measurements (synthesis and degradation). Users can obtain the half-life time of proteins of interest to assess their stability (22). This data can support the analysis of the mode of action of drugs (23) and might provide additional avenues

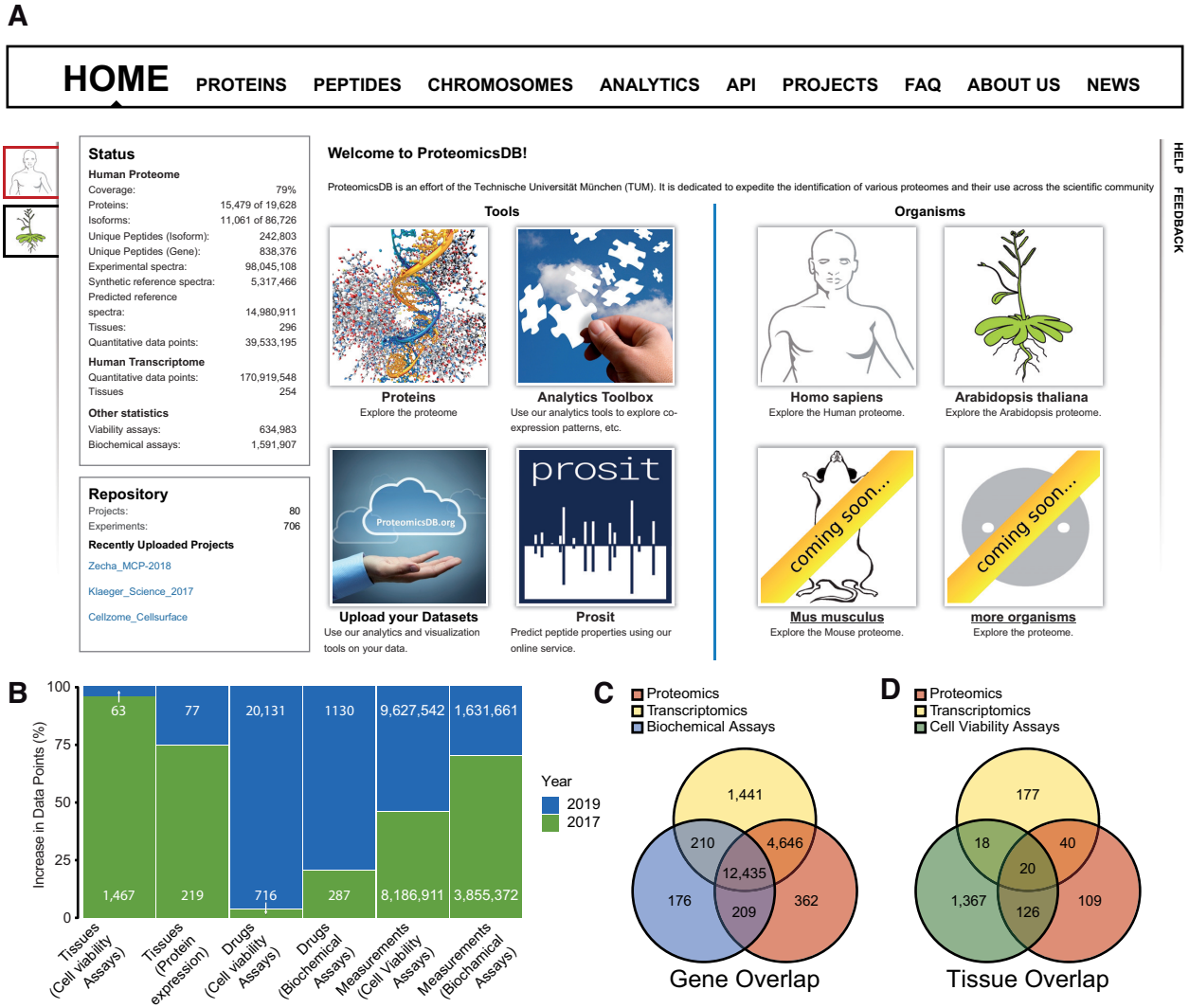

Figure 2. Additions to ProteomicsDB. (**A**) The front page of ProteomicsDB has been adjusted to host new organisms as well as provide information about the quantity of the different data types that are stored in the database. (**B**) Barplot depicting the proportion and absolute number of data points added to ProteomicsDB (in blue) since the previous update manuscript in 2017 (green). (**C**) Venn diagram showing the number and overlap of genes for which proteomics, transcriptomics or biochemical assay data is available in ProteomicsDB. (**D**) Venn diagram showing the number and overlap of tissues (as well as cell lines and body fluids) for which the respective data types are available in ProteomicsDB.

into understanding the effectiveness of drugs in light of the stability of on- or off-target proteins (18). In total, ∼20 000 proteins (including isoforms) are covered by at least one and ∼3000 by all three biochemical assay types, providing potentially valuable insight into additional aspects of a protein's life cycle. As ProteomicsDB visualizes every curve (accessible via the 'Biochemical assay' tab in the 'Protein Details' view), users can assess the quality of each individual curve and underlying data points themselves.

Upload and online analysis of user expression data

Uploading expression profiles. ProteomicsDB's ability to interconnect and cross-reference data from various sources is one of its core features. However, this was so far only possible for data already stored in ProteomicsDB, limiting its usefulness for the interpretation of data acquired in a user laboratory. In order to fill this gap, we implemented a new feature called 'Custom User Data Upload' (Figure 4). Here, users can temporarily upload their expression profiles and optionally normalize them to the data stored in ProteomicsDB. On upload of a dataset, a temporary session is created in the database which can be accessed by a unique session ID. This session will automatically expire after 14 days, which will result in the permanent and not recoverable deletion of all corresponding data unless the user chooses to extend this period. Users can save and use their session ID to load their session to any other computer or browser. Data stored in such sessions are available via ODATA (https://www.odata.org) services within ProteomicsDB and will ultimately allow the integration into any existing analytical pipeline.

The first use case we highlight is the comparison of custom expression data to expression data stored in Pro-

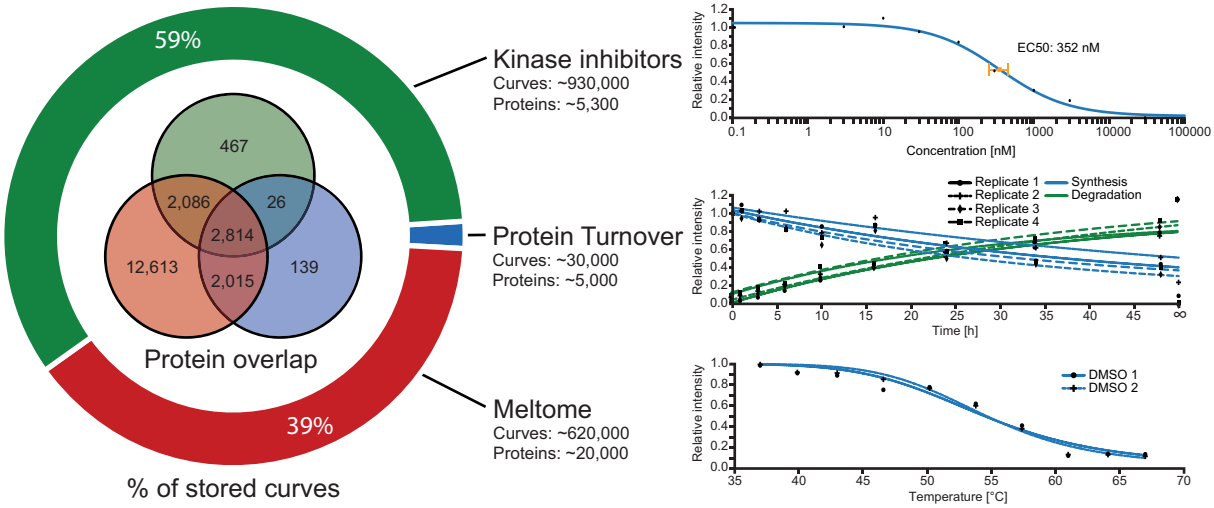

Figure 3. New biochemical assay data. The pie chart on the left shows the distribution of biochemical assay data available for three different applications. The Venn diagram inside the pie chart shows the overlap of proteins for which biochemical assay data of the respective type is available. The diagrams on the right show exemplary fitted curves for each biochemical assay type, accompanied by the number of curves and proteins that each assay covers.

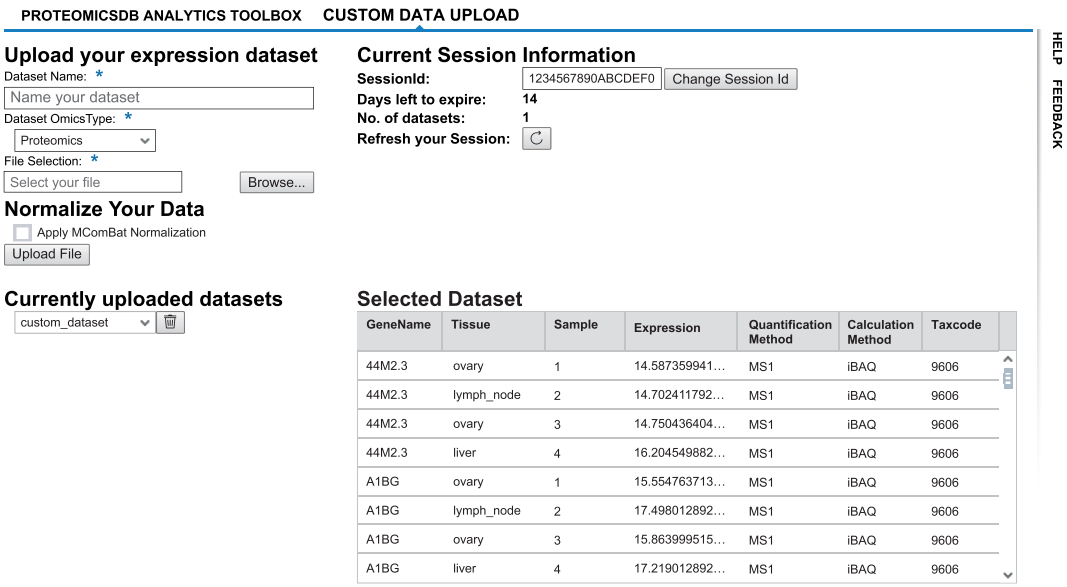

Figure 4. Custom data analysis area of ProteomicsDB. The 'Custom Data Upload' tab enables users to upload their own expression datasets temporarily to ProteomicsDB. The datasets are session-specific so that no other user has access to this uploaded data.

teomicsDB. For this to be successful, we highly recommend making use of the normalization feature available upon upload. The uploaded expression profiles are normalized via MComBat (24) using the total sum normalized proteomics expression values of ProteomicsDB as a reference set. Because MComBat normalization depends on the calculation of a mean and variance for any given protein, only datasets with three or more samples can be normalized using this method. Every uploaded dataset has to adhere to a pre-defined comma-separated format (.csv files) where each row must provide the following information. (i) A gene name—HGNC symbol as the identifier, which will help us associate the uploaded proteins to the ones stored in ProteomicsDB and enable cross-dataset comparisons. (ii) A tissue or cell line name representing the origin of the measured sample, which will be used for visualizations. (iii) A sample name, which is important to separate samples with the same tissue of origin especially for the normalization step, as samples with the same sample and tissue/cell line name will be automatically aggregated as there is no way to separate them. (iv) The expression value of the corresponding protein in the sample in $\log_{10}$ scale, accompanied by the quantification and calculation method that was used, which will help with further comparisons of matching in-ProteomicsDB data. (v) The taxonomy code of each sample, which will allow dataset separation based on the selected organism, a feature which is discussed below. A detailed documentation on how to use this functionality as well as on

the data upload format, can be found by clicking the 'Help' button that accompanies every view in ProteomicsDB (Figure 4).

Use of analytical tools on uploaded datasets. By uploading an expression dataset, back-end procedures take care of the data modelling and transformation, so that they are compatible to existing tools with no major differences to the data available in ProteomicsDB. The first tool making use of this is the interactive expression heat map. The heat map allows interactive visualization of expression patterns of multiple groups of proteins. Upon upload, users can choose a data source and focus their analysis on either data from ProteomicsDB, their own datasets noted as 'User Data' or the integration of both, noted as 'Combined'. Because the heat map automatically aggregates tissues, duplicated tissue names provided in the custom dataset will appear as one column. The automatic mapping enables users to use all functionalities of the heat map, such as direct links to the ProteomicsDB's protein summary views and perform GO enrichment analysis on the selected proteins. The 'Combined' option allows users to compare their data to data stored in ProteomicsDB. They can further allow a comparison of some or all datasets that they have uploaded to the in-database data. Users should expect that uploaded datasets that were not subjected to normalization during uploading, will clustered together. If the normalization step was enabled, then user samples should cluster with tissues or cell lines that have similar expression profiles in ProteomicsDB, ideally from the same origin. Figure 4 shows such an example where a custom dataset was co-clustered with data stored in ProteomicsDB. Some of the uploaded expression profiles of cell lines co-cluster with the respective cell lines stored in ProteomicsDB (here lung and liver samples). There are cases though (here ovary) that cluster with other tissues (here uterus). This feature enables users to find the closest cell lines for which ProteomicsDB contains, e.g. phenotypic information and explore compounds that may be effective in user cell lines.

Extended heat map features—missing values imputation. ProteomicsDB stores a large collection of transcriptomics expression profiles alongside the respective proteomic profiles. Having access to expression data from both sources and to the automatic mapping using the built-in Resource Identifier Relation Model, ProteomicsDB is able to perform data-driven missing value imputation using either data type. Especially proteomics data (depending on the depth of measurement) can show a large number of missing values. Data selected for imputation might come from different projects for both omics types. Even projects of the same omics type might differ in the distribution of their expression values. This phenomenon is commonly referred to as 'batch effect' and results in additional variance by the fact that we aggregate data across multiple 'batches'. Here, the term 'batch' refers to experiments processed in one laboratory over a short time period using the same technological platform (25). We performed intra-omics normalization and batch effect correction using ComBat (26). Next, we apply MComBat (24) to perform inter-omics correction of systematic differences. MComBat, in contrast to ComBat, allows selecting a reference dataset so that all other datasets will be normalized based on the reference. Transcriptomics data are then transferred to the same scale of the proteomics expression data. Previous experiments showed that the correlation across all tissues between mRNA and protein expression data is higher with than without such an adjustment (27). Finally, we implemented the mRNA-guided missing value imputation method, described in (27). For this purpose, we train linear regression models and extrapolate protein abundance from transcriptomics abundance. To validate the performance of the generated models, we created artificial missing values in a random subset of the protein expression data that are stored in ProteomicsDB. We then used our models to extrapolate the protein abundances and compared them to two other common missing value imputation strategies: (a) replacing missing values with the minimum protein abundance of the corresponding sample and (b) random sampling from the corresponding sample's protein abundance distribution, as the created missing values originate from the whole abundance distribution. The mRNA-guided missing value imputation method showed the best correlation to the measured values (Supplementary Figure S1) which is why we implemented it. The entire procedure, from data normalization to training the regression model is performed by the R server (Figure 1). This is possible because the SAP HANA in-memory database management system supports direct connections to the R-server via proper adapters. Missing value imputation is available in the interactive heat map (Figure 5) and can be activated by the respective button. Once activated, and only if matching expression profiles are available, the model trained above and the adjusted transcriptomics expression data are used to fill in missing values in the protein expression matrix. The authors point out that missing value imputation can lead to issues and should therefore be carefully considered and evaluated on a case by case basis. Especially in the case of mRNA-guided missing value imputation, it becomes less accurate if the RNA dataset or protein expression data has a limited number of samples. Moreover, not all missing values can be imputed if RNASeq matching data is missing.

Drug sensitivity prediction for proteomic profiles

ProteomicsDB already covers a lot of phenotypic drug sensitivity information (Figure 2B) and to the best of our knowledge, no other platform exists which shows the full dose response curves across multiple resources including filters to the extent as ProteomicsDB's cell viability viewer does. However, the list of cell lines for which this data is available is necessarily incomplete and likely entirely unavailable or impossible to generate if cells lines were derived from say patient tissue in a particular laboratory. In order to obtain an estimate of the susceptibility of such cell lines to drugs, without performing an experiment, ProteomicsDB provides a tool to model and estimate drug sensitivity, based on expression profiles. Recent proteome profiling of the NCI60 (28) and the CRC65 (27) cancer cell line panels, and an additional panel of 20 breast cancer cell lines (29) showed that protein signatures can predict drug sensitivity or resistance. On this basis, we implemented elastic net regression (30) in ProteomicsDB to model drug sensi-

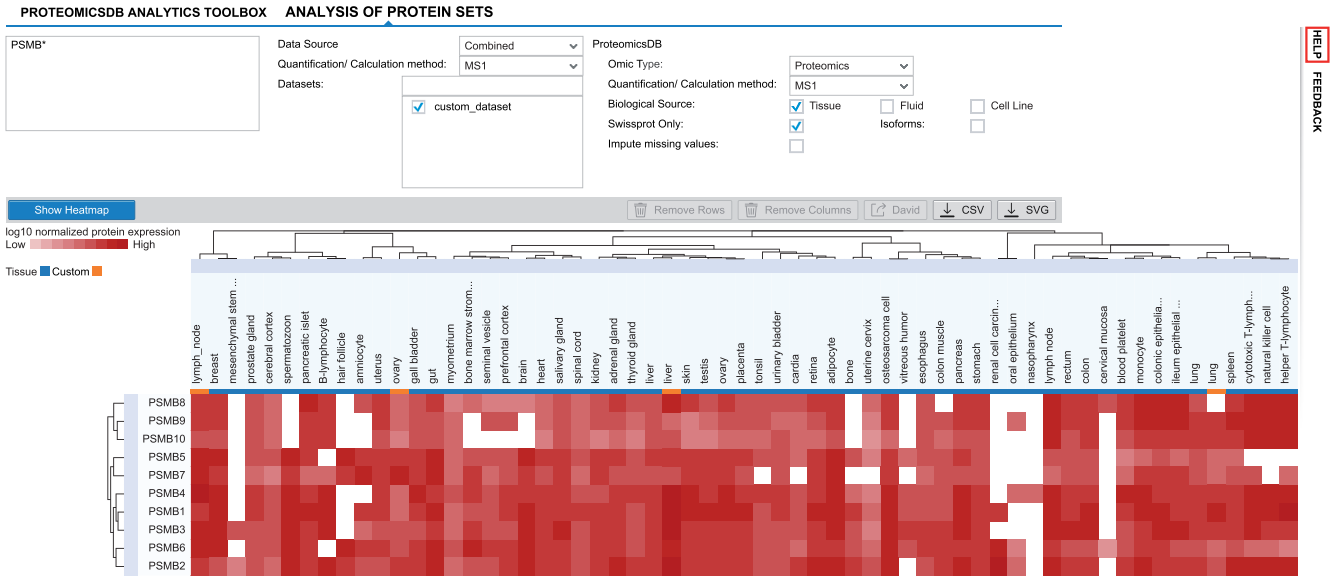

Figure 5. Combined interactive expression heat map. User datasets can be clustered along with data stored in ProteomicsDB for a combined analysis. User datasets (marked in orange) that were normalized using MComBat subsequent to upload, cluster close to samples in ProteomicsDB (in blue) that were generated from the same or similar tissues or cell types.

tivity as a function of quantitative protein expression profiles. This functionality can be used in the 'Drug Sensitivity Prediction' view (Figure 6). Here, users can select from a variety of tissues and cell lines whose proteomic profiles are stored in ProteomicsDB. Next, a drug or compound can be selected to check for its effect on the selected cell line (Figure 6A). Figure 6B shows the result of the prediction as bar plots - one for each predicted feature (area under the curve, pEC50, relative effect). Error bars show the range of the predictions of all bootstraps of the corresponding model. Each drug in ProteomicsDB might be accompanied by multiple models (multiple bars in each bar plot), because the drug may have been used in more than one drug sensitivity screen which was imported into ProteomicsDB (max. 4). It is important to point out that each model includes a certain set of predictor-proteins. If the sample on which a user wants to predict drug sensitivity does not contain some of the required proteins, prediction from some models is not possible. Selecting a bar of any bar plot generates a volcano plot (Figure 6C), which shows information for the interpretation of the trained model. The x-axis shows how strong the expression of a particular protein is associated with drug sensitivity or resistance, analogous to a correlation. The y-axis shows the number of bootstrap models contained the particular protein as a predictor, when training the elastic net model. Proteins that appear in the top left and right areas of the volcano plot (Figure 6C) are frequently selected from the models as predictors, as they have a high positive or negative correlation with drug sensitivity or resistance and can, therefore, represent potential biomarkers. Instead of predicting drug sensitivity on tissues or cell lines from ProteomicsDB, users also have the option to use this functionality on their own datasets, uploaded using the 'Custom User Data Upload' tab. Predictions can be applied to all user datasets, although it is highly recommended to use normalization upon uploading, as the models were trained on data stored in ProteomicsDB and expect values from the same or similar expression distributions.

Real-time analytics and visualization for any organism

ProteomicsDB was initially developed for the exploration of the human proteome. As a result, every database view and endpoint was designed without explicit support for multiple organisms. In order to support the storage, handling and visualization of data from multiple organisms, all layers of ProteomicsDB (Figure 1) required modifications and extensive testing. In the new version presented here, we modified all backend procedures to support querying of data for a specific taxonomy. The API endpoints were modified to require a taxcode in order to respond with the desired data. With this functionality in place, we prepared the database and the data models to support and handle the protein sequence space of any organism. Similarly, the user interface was modified to support the visualization of data from a selected organism. Users can change the selected organisms by using the respective icons on the left hand side of each view, or directly on the front page of ProteomicsDB (Figure 2A). For the protein expression visualization, new interactive body maps for *Arabidopsis thaliana* and *Mus musculus* were generated (Figure 7A, Supplementary Figure S2) and function in the same way as the human body map.

To bring *Arabidopsis thaliana* into ProteomicsDB, we downloaded, processed and imported the protein sequence space from UniProt, following the same mechanism as for human proteins. Upon import, appropriate decoy sequences were created for every protease, to allow false discovery (FDR) estimation by the picked FDR approach already implemented in ProteomicsDB (31). We furthermore imported the Plant Ontology (PO) (32) to be able to make use of ontologies for the different plant tissues. This step was not necessary for *Mus musculus*, since the

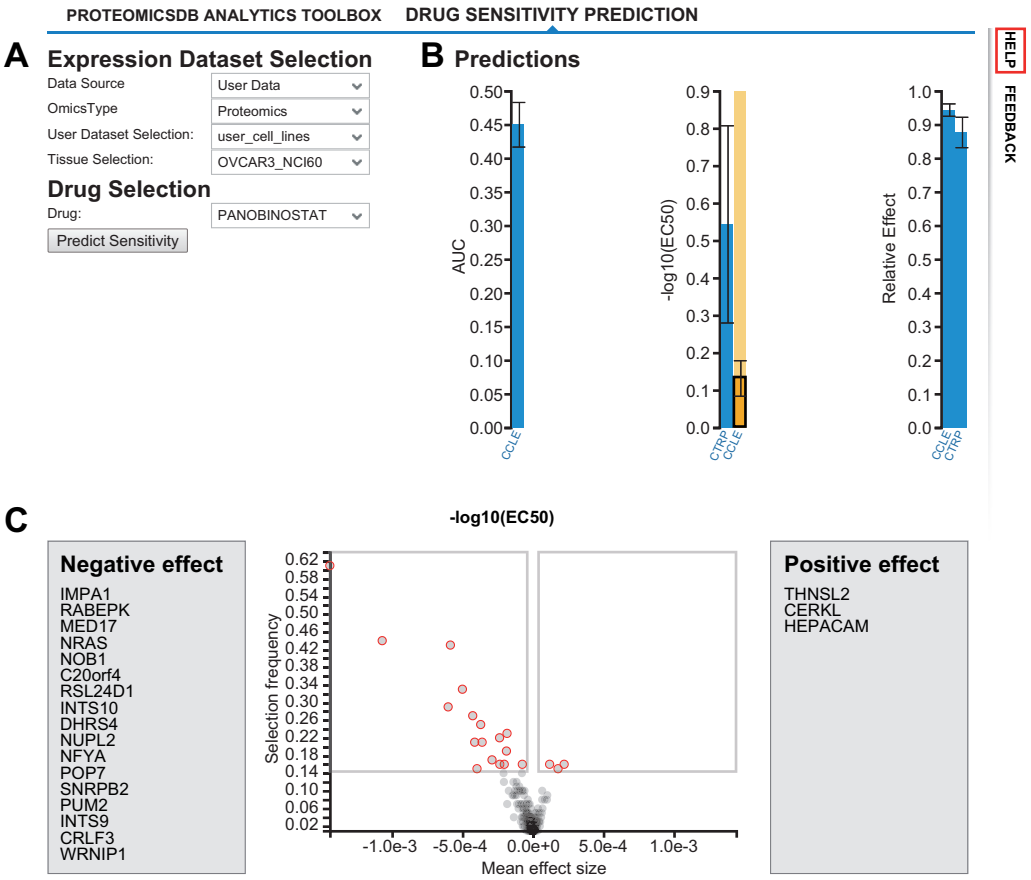

Figure 6. Drug sensitivity prediction. (**A**) Prediction is enabled for both, data stored in ProteomicsDB or user uploaded datasets. (**B**) This view visualizes the predicted sensitivity of a chosen cell line to a chosen drug expressed by area under the curve (AUC, left bar), the negative log of the effective concentration of the drug (EC50, middle bars) and the relative (cell killing) effect (right bars). If more than one bar is shown, more than one training data set was available for the particular drug and either one or several predictions are shown. (**C**). Each dot in the volcano plot, represents a protein that is associated to drug sensitivity or resistance on the basis of the elastic net model generated during training.

Brenda Tissue Ontology (BTO) (33) that was previously imported into ProteomicsDB to support the analysis of human proteins covers any mammalian tissue. To complete the protein information and meta-data panel, we downloaded and imported protein domain information from SMART (34) using their RESTful API and GO annotations using the QuickGo-API of the European Bioinformatics Institute (EBI). Protein-protein interactions and functional pathway information were downloaded from STRING (35) and KEGG (36), respectively. The latter data were processed and transformed for import into our triple-store data model, which allows the automatic mapping of the respective STRING and KEGG identifiers to the corresponding UniProt accessions and our internal protein identifiers. With the meta-data imported, the proteomics and transcriptomics expression profiles for *Arabidopsis thaliana* were imported. The project covers 30 different tissues, including a tissue-derived cell line that was derived from callus tissue. Because of the generic design of ProteomicsDB, any analytical view (e.g. heat map) will work without further modifications for any other organism. However, due to the limited datasets available for phenotypic drug responses (and the respective drug targets), other views do not show any *A. thaliana* or *M. musculus* data yet.

As mentioned before, we have imported >5 million reference spectra acquired from synthetic human peptides in the ProteomeTools project. As a next step, we imported more than 10 million Prosit-predicted peptide spectra, in three different charge states and 3 different collision energies. By chance, these spectra also represent 70 000 peptides from *Arabidopsis thaliana* because their sequences are identical in either organism. In addition, we added predicted spectra for all peptides present in the experimental data set. Thus, akin to the human case, these reference spectra can be used to validate peptide identifications in experimental data using the mirror spectrum viewer integrated in ProteomicsDB. First, these are directly accessible in the 'Peptides/MSMS' tab of the 'Protein Details' view, where users can validate or invalidate i.e. one hit wonders (proteins which are only identified by a single peptide/spectrum), and more generally validate proteins/peptides in case the user wants confirmation that the protein is actually present in the sample of a project and consequently in a cell line or tissue in ProteomicsDB. Since ProteomicsDB contains up to 14 different types of reference spectra (11 fragmentation settings from ProteomeTools and 3 normalized collision energies from Prosit) as indicated in the list of available reference spectra, users can select the optimal match (37). Second, in the 'Reference Pep-

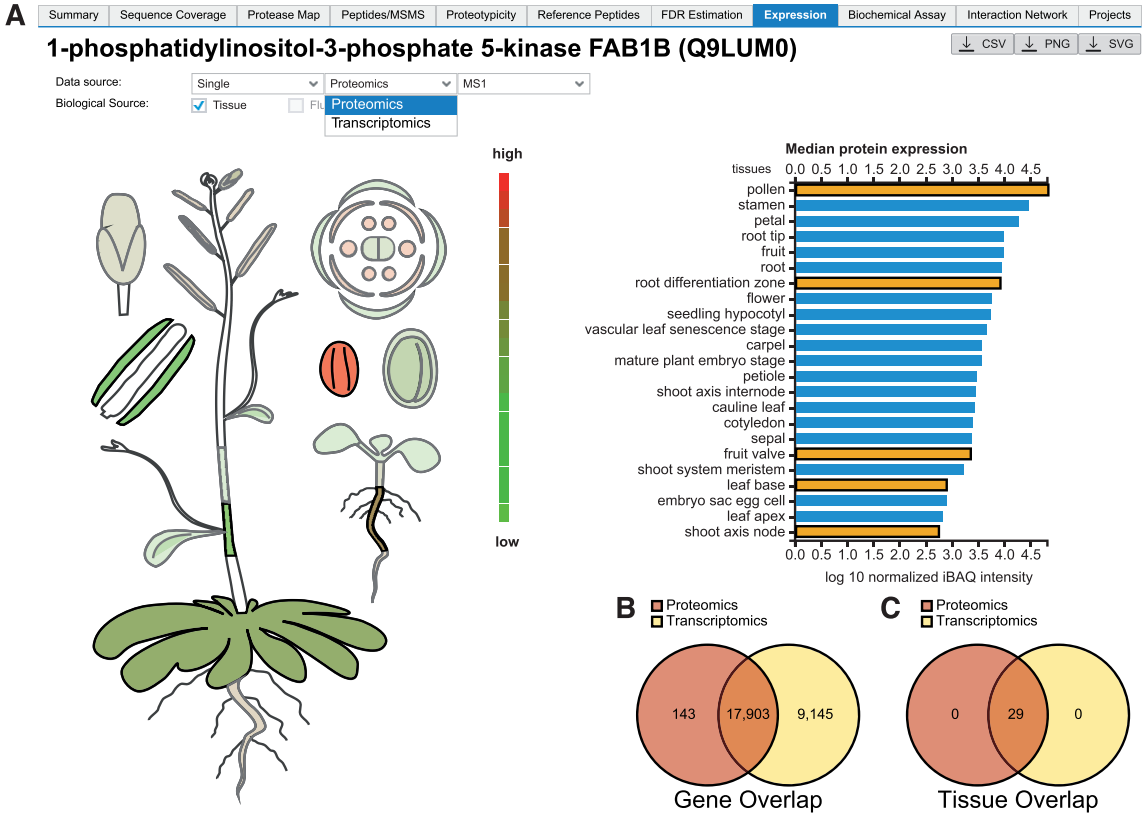

Figure 7. ProteomicsDB as a multi-organism and multi-omics platform. (**A**) Proteome or transcriptome expression data are visualized in the tissues of a chosen organism (left) and numerical expression data (medians in case multiple samples of the same tissue are available) are shown on the right for each tissue the protein was found in. Tissue bars selected by users turn orange and the respective tissue is highlighted on the body map on the left view projects the tissue aggregated omics expression values to the corresponding organism's body map. (**B**) Venn diagram is showing the overlap of gene-level data available for proteomics and transcriptomics for *Arabidopsis thaliana*. (**C**) Venn diagram showing the overlap of tissues for which proteomics and transcriptomics expression values are available in ProteomicsDB.

tides' tab, where users can browse ProteomeTools and Prosit spectra for e.g. designing targeted mass spectrometric assays. The two separate views exist because for some proteins, no experimental spectra of endogenous proteins might be available, while many reference spectra might be available because the ProteomeTools synthesized all meaningful peptides for a hitherto unobserved protein. For proteins where experimental data from endogenous proteins is available, users can take experimental proteotypicity of peptides into account and thus rationalize which peptide to choose for an assay. Additionally, this view can be used to compare spectra created by different fragmentation methods and, more importantly, different collision energies to optimize their targeted assays for collision energies which generate desired fragment ions (e.g. highly intense and high m/z ions). Furthermore, spectra can now be downloaded in the mirrored spectrum viewer as msp-files. Finally, as mentioned above, ProteomicsDB is also ready to support *Mus musculus* data. However, the selection of mouse in ProteomicsDB will only be enabled once the data has been published.

FUTURE DIRECTIONS

The continuous updates introduced over the last years have transformed ProteomicsDB into a multi-omics resource for life science research covering proteomic and transcriptomic expression, pathway, protein-protein and protein-drug interactions, and cell viability data (Supplementary Figure S3). Many aspects of ProteomicsDB are already respecting the FAIR principles (38). For example, e.g. findability (F) is supported by unique identifiers, accessibility (A) via API endpoints including meta-data and reusability (R) by way of multiple online services taking advantage of ProteomicsDB's API endpoints. However, more efforts are currently made to transform ProteomicsDB into a fully FAIR resource, e.g. by extending the API to allow access to all data stored in ProteomicsDB. One particular strength of ProteomicsDB is its versatile mapping service allowing the seamless connection between different data types. This enables subsequent modelling and data mining to further evolve ProteomicsDB from an information database to a knowledge platform. Along these lines, we plan to extend our analytical toolbox such that scientists in life science research can directly benefit from the wealth of data stored in ProteomicsDB. Here, we show the first steps into this direction by extending the toolbox as well as enabling users to upload their own expression data. Combined with ProteomicsDB's flexible infrastructure, this will provide ease of use for data analysis, interpretation and machine learning

capabilities not accessible to every laboratory or scientist. For this purpose, we are also planning to further extend the data content of ProteomicsDB to include, e.g. protein structures integrated with drug–target affinity data (20) or develop tools which allow the prediction of the target spaces of kinase inhibitors (39).

Two more extensions are planned that will allow the further integration and exploitation of reference spectra. The first one is to use synthetic or predicted reference spectra to systematically validate and assess the confidence of experimental data by evaluating their spectral similarity. As shown earlier, the integration of intensity information can lead to drastic improvements in either the number of identified peptides or the ability to differentiate correct from incorrect matches (5). Especially the latter will help to increase the confidence of each peptide identification and thus also increase the quality of identification and quantification results stored in ProteomicsDB. The second extension is the implementation of a smart tool which will allow users to build targeted assays based on data stored in ProteomicsDB as described.

Ultimately, the collected data and generated knowledge should culminate in actionable hypotheses. These may drive the design of laboratory experiments or eventually aid decision making in patient care. One way how ProteomicsDB could be used for the latter is by providing tools that assist molecular tumor boards. We plan to provide pipelines where researchers and clinicians will be able to upload the protein profiles of patient samples in a fully anonymized fashion and have in-depth bioinformatic analysis reports returned, spiked with a wide range of information including, e.g. protein and RNA abundance levels, biomarkers that predict sensitivity or resistance, potential off-label uses based on approved kinase inhibitors as well as general sample characterization, classification or origin identification based on similarities of molecular fingerprints.

DATA AVAILABILITY

ProteomicsDB is available at https://www.ProteomicsDB.org.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to thank all members of the Kuster laboratory for fruitful discussions and technical assistance.

FUNDING

German Science Foundation [SFB924, SFB1309, SFB1321]; German Federal Ministry of Education and Research (BMBF) [031L0008A, 031L0168]; SAP. Funding for open access charge: BMBF [031L0168].

Conflict of interest statement. T.S., S.G. and M.F. are founders and shareholders of msAId, which operates in the field of proteomics. M.W. and B.K. are founders and shareholders of OmicScouts and msAId, which operate in the field of proteomics. They have no operational role in the company. S.G., H.-C.E. and S.A. are employees of SAP SE. Neither company affiliation had any influence on the results presented in this study.

REFERENCES

1. Wilhelm,M., Schlegl,J., Hahne,H., Gholami,A.M., Lieberenz,M., Savitski,M.M., Ziegler,E., Butzmann,L., Gessulat,S., Marx,H. *et al.* (2014) Mass-spectrometry-based draft of the human proteome. *Nature*, **509**, 582–587.
2. Schmidt,T., Samaras,P., Frejno,M., Gessulat,S., Barnert,M., Kienegger,H., Krcmar,H., Schlegl,J., Ehrlich,H.C., Aiche,S. *et al.* (2018) ProteomicsDB. *Nucleic Acids Res.*, **46**, D1271–D1281.
3. Zolg,D.P., Wilhelm,M., Schmidt,T., Medard,G., Zerweck,J., Knaute,T., Wenschuh,H., Reimer,U., Schnatbaum,K. and Kuster,B. (2018) ProteomeTools: Systematic characterization of 21 Post-translational protein modifications by liquid chromatography tandem mass spectrometry (LC-MS/MS) using synthetic peptides. *Mol. Cell. Proteomics*, **17**, 1850–1863.
4. Zolg,D.P., Wilhelm,M., Schnatbaum,K., Zerweck,J., Knaute,T., Delanghe,B., Bailey,D.J., Gessulat,S., Ehrlich,H.C., Weininger,M. *et al.* (2017) Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods*, **14**, 259–262.
5. Gessulat,S., Schmidt,T., Zolg,D.P., Samaras,P., Schnatbaum,K., Zerweck,J., Knaute,T., Rechenberger,J., Delanghe,B., Huhmer,A. *et al.* (2019) Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods*, **16**, 509–518.
6. Iorio,F., Knijnenburg,T.A., Vis,D.J., Bignell,G.R., Menden,M.P., Schubert,M., Aben,N., Goncalves,E., Barthorpe,S., Lightfoot,H. *et al.* (2016) A landscape of pharmacogenomic interactions in cancer. *Cell*, **166**, 740–754.
7. Rees,M.G., Seashore-Ludlow,B., Cheah,J.H., Adams,D.J., Price,E.V., Gill,S., Javaid,S., Coletti,M.E., Jones,V.L., Bodycombe,N.E. *et al.* (2016) Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat. Chem. Biol.*, **12**, 109–116.
8. Medico,E., Russo,M., Picco,G., Cancelliere,C., Valtorta,E., Corti,G., Buscarino,M., Isella,C., Lamba,S., Martinoglio,B. *et al.* (2015) The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets. *Nat. Commun.*, **6**, 7002.
9. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehar,J., Kryukov,G.V., Sonkin,D. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.
10. Uhlen,M., Oksvold,P., Fagerberg,L., Lundberg,E., Jonasson,K., Forsberg,M., Zwahlen,M., Kampf,C., Wester,K., Hober,S. *et al.* (2010) Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.*, **28**, 1248–1250.
11. Komljenovic,A., Roux,J., Wollbrett,J., Robinson-Rechavi,M. and Bastian,F.B. (2018) BgeeDB, an R package for retrieval of curated expression datasets and for gene list expression localization enrichment tests [version 2; peer review: 2 approved, 1 approved with reservations]. *F1000Res*, **5**, 2748.
12. UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
13. Stelzer,G., Rosen,N., Plaschkes,I., Zimmerman,S., Twik,M., Fishilevich,S., Stein,T.I., Nudel,R., Lieder,I., Mazor,Y. *et al.* (2016) The GeneCards Suite: From gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinformatics*, **54**, 1.30.1–1.30.33.
14. Turei,D., Korcsmaros,T. and Saez-Rodriguez,J. (2016) OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods*, **13**, 966–967.
15. Knight,J.D.R., Samavarchi-Tehrani,P., Tyers,M. and Gingras,A.C. (2019) Gene Information eXtension (GIX): effortless retrieval of gene product information on any website. *Nat. Methods*, **16**, 665–666.
16. Monga,M. and Sausville,E.A. (2002) Developmental therapeutics program at the NCI: molecular target and drug discovery process. *Leukemia*, **16**, 520–526.
17. Savitski,M.M., Reinhard,F.B., Franken,H., Werner,T., Savitski,M.F., Eberhard,D., Martinez Molina,D., Jafari,R., Dovega,R.B., Klaeger,S. *et al.* (2014) Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science*, **346**, 1255784.

18. Klaeger,S., Heinzlmeir,S., Wilhelm,M., Polzer,H., Vick,B., Koenig,P.A., Reinecke,M., Ruprecht,B., Petzoldt,S., Meng,C. *et al.* (2017) The target landscape of clinical kinase drugs. *Science*, **358**, eaan4368.
19. Koch,H., Busto,M.E., Kramer,K., Medard,G. and Kuster,B. (2015) Chemical proteomics uncovers EPHA2 as a mechanism of acquired resistance to small molecule EGFR kinase inhibition. *J. Proteome Res.*, **14**, 2617–2625.
20. Heinzlmeir,S., Kudlinzki,D., Sreeramulu,S., Klaeger,S., Gande,S.L., Linhard,V., Wilhelm,M., Qiao,H., Helm,D., Ruprecht,B. *et al.* (2016) Chemical proteomics and structural biology define EPHA2 inhibition by clinical kinase drugs. *ACS Chem. Biol.*, **11**, 3400–3411.
21. Heinzlmeir,S., Lohse,J., Treiber,T., Kudlinzki,D., Linhard,V., Gande,S.L., Sreeramulu,S., Saxena,K., Liu,X., Wilhelm,M. *et al.* (2017) Chemoproteomics-Aided medicinal chemistry for the discovery of EPHA2 inhibitors. *Chem. Med. Chem*, **12**, 999–1011.
22. Zecha,J., Meng,C., Zolg,D.P., Samaras,P., Wilhelm,M. and Kuster,B. (2018) Peptide level turnover measurements enable the study of proteoform dynamics. *Mol. Cell. Proteomics*, **17**, 974–992.
23. Savitski,M.M., Zinn,N., Faelth-Savitski,M., Poeckel,D., Gade,S., Becher,I., Muelbaier,M., Wagner,A.J., Strohmer,K., Werner,T. *et al.* (2018) Multiplexed proteome dynamics profiling reveals mechanisms controlling protein homeostasis. *Cell*, **173**, 260–274.
24. Stein,C.K., Qu,P., Epstein,J., Buros,A., Rosenthal,A., Crowley,J., Morgan,G. and Barlogie,B. (2015) Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. *BMC Bioinformatics*, **16**, 63.
25. Chen,C., Grennan,K., Badner,J., Zhang,D., Gershon,E., Jin,L. and Liu,C. (2011) Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One*, **6**, e17238.
26. Johnson,W.E., Li,C. and Rabinovic,A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.
27. Frejno,M., Zenezini Chiozzi,R., Wilhelm,M., Koch,H., Zheng,R., Klaeger,S., Ruprecht,B., Meng,C., Kramer,K., Jarzab,A. *et al.* (2017) Pharmacoproteomic characterisation of human colon and rectal cancer. *Mol. Syst. Biol.*, **13**, 951.
28. Gholami,A.M., Hahne,H., Wu,Z., Auer,F.J., Meng,C., Wilhelm,M. and Kuster,B. (2013) Global proteome analysis of the NCI-60 cell line panel. *Cell Rep.*, **4**, 609–620.
29. Lawrence,R.T., Perez,E.M., Hernandez,D., Miller,C.P., Haas,K.M., Irie,H.Y., Lee,S.I., Blau,C.A. and Villen,J. (2015) The proteomic landscape of triple-negative breast cancer. *Cell Rep*, **11**, 630–644.
30. Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc.: B (Stat. Methodol.)*, **67**, 301–320.
31. Savitski,M.M., Wilhelm,M., Hahne,H., Kuster,B. and Bantscheff,M. (2015) A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Mol. Cell. Proteomics*, **14**, 2394–2404.
32. Walls,R.L., Cooper,L., Elser,J., Gandolfo,M.A., Mungall,C.J., Smith,B., Stevenson,D.W. and Jaiswal,P. (2019) The plant ontology facilitates comparisons of plant development stages across species. *Front. Plant Sci.*, **10**, 631.
33. Gremse,M., Chang,A., Schomburg,I., Grote,A., Scheer,M., Ebeling,C. and Schomburg,D. (2011) The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.*, **39**, D507–D513.
34. Letunic,I. and Bork,P. (2018) 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.*, **46**, D493–D496.
35. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P. *et al.* (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
36. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
37. Zolg,D.P., Wilhelm,M., Yu,P., Knaute,T., Zerweck,J., Wenschuh,H., Reimer,U., Schnatbaum,K. and Kuster,B. (2017) PROCAL: a set of 40 peptide standards for retention time indexing, column performance monitoring, and collision energy calibration. *Proteomics*, **17**, 1700263.
38. Wilkinson,M.D., Dumontier,M., Aalbersberg,I.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.W., da Silva Santos,L.B., Bourne,P.E. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
39. Li,X., Li,Z., Wu,X., Xiong,Z., Yang,T., Fu,Z., Liu,X., Tan,X., Zhong,F., Wan,X. *et al.* (2019) Deep learning enhancing kinome-wide polypharmacology profiling: model construction and experiment validation. *J. Med. Chem.*, doi:10.1021/acs.jmedchem.9b00855.

Supplementary Material

ProteomicsDB: A multi-omics and multi-organism resource for life science research

Patroklos Samaras[1], Tobias Schmidt[1], Martin Frejno[1], Siegfried Gessulat[1, 2], Maria Reinecke[1, 3, 4], Anna Jarzab[1], Jana Zecha[1], Julia Mergner[1], Piero Giansanti[1], Hans-Christian Ehrlich[2], Stephan Aiche[2], Johannes Rank[5, 6], Harald Kienegger[5, 6], Helmut Krcmar [5, 6], Bernhard Kuster[1, 7, *], Mathias Wilhelm[1, *]

[1] Chair of Proteomics and Bioanalytics, Technical University of Munich (TUM), Freising, 85354, Bavaria, Germany

[2] Innovation Center Network, SAP SE, Potsdam, 14469, Germany

[3] German Cancer Consortium (DKTK), Partner Site Munich, Munich, Germany

[4] German Cancer Research Center (DKFZ), Heidelberg, Germany

[5] Chair for Information Systems, Technical University of Munich (TUM), Garching, 85748, Germany

[6] SAP University Competence Center, Technical University of Munich (TUM), Garching, 85748, Germany

[7] Bavarian Biomolecular Mass Spectrometry Center (BayBioMS), Technical University of Munich (TUM), Freising, 85354, Bavaria, Germany

* To whom correspondence should be addressed. Tel: +49 8161 71 4202; Fax: +49 8161 71 5931; Email: mathias.wilhelm@tum.de, kuster@tum.de

Supplementary Figures

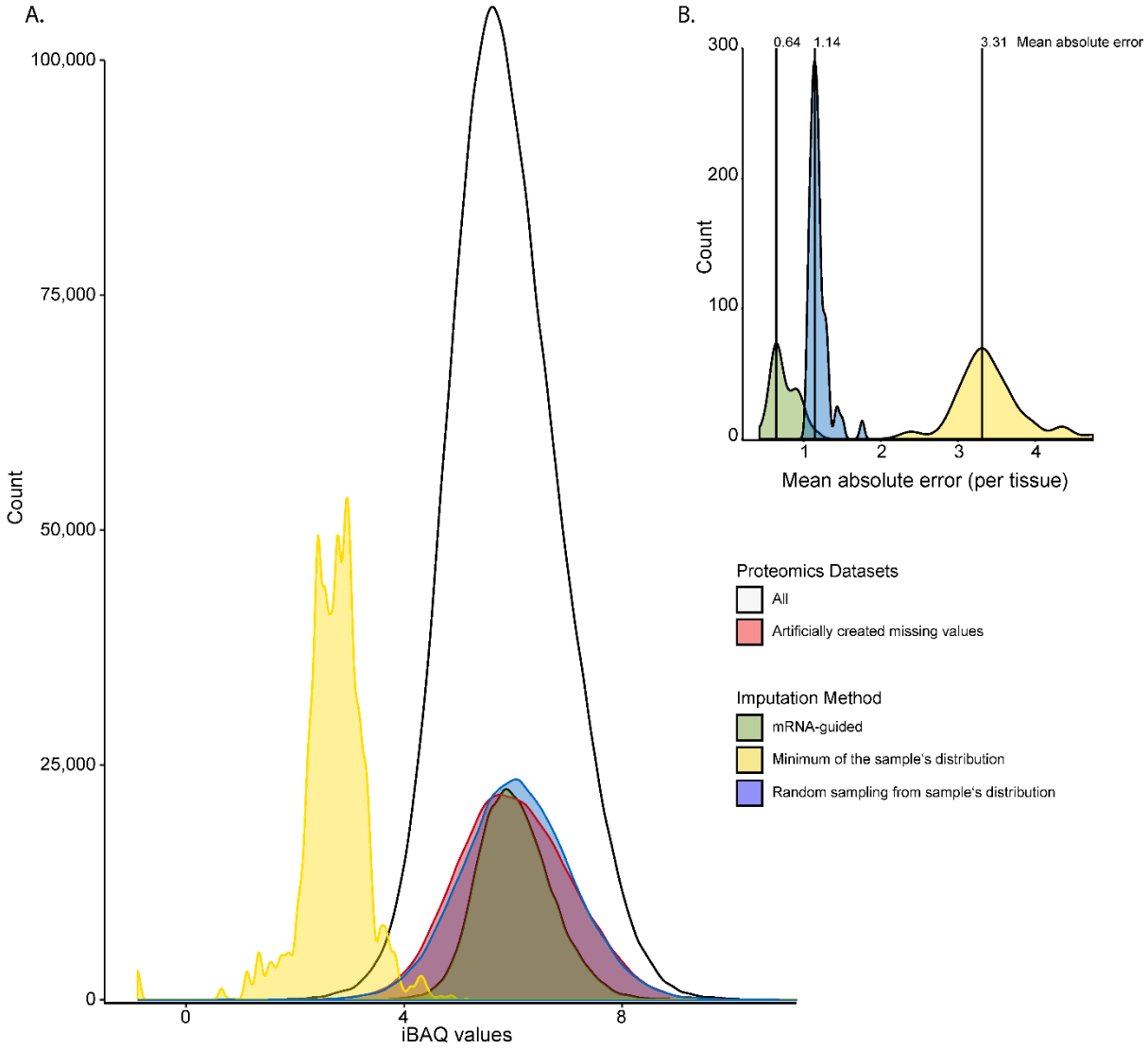

Figure S1. Histograms for comparing different imputation methods. A. The white distribution represents all iBAQ protein expression values in ProteomicsDB. The red distribution represents the missing values we created by random sampling 10% of the white distribution. The green, yellow and blue distributions represent the imputed missing values based on the 3 different methods respectively: mRNA-guided, minimum of a sample's distribution and random sampling from each sample's distribution. B. Histograms of the mean absolute error per tissue for each imputation method.

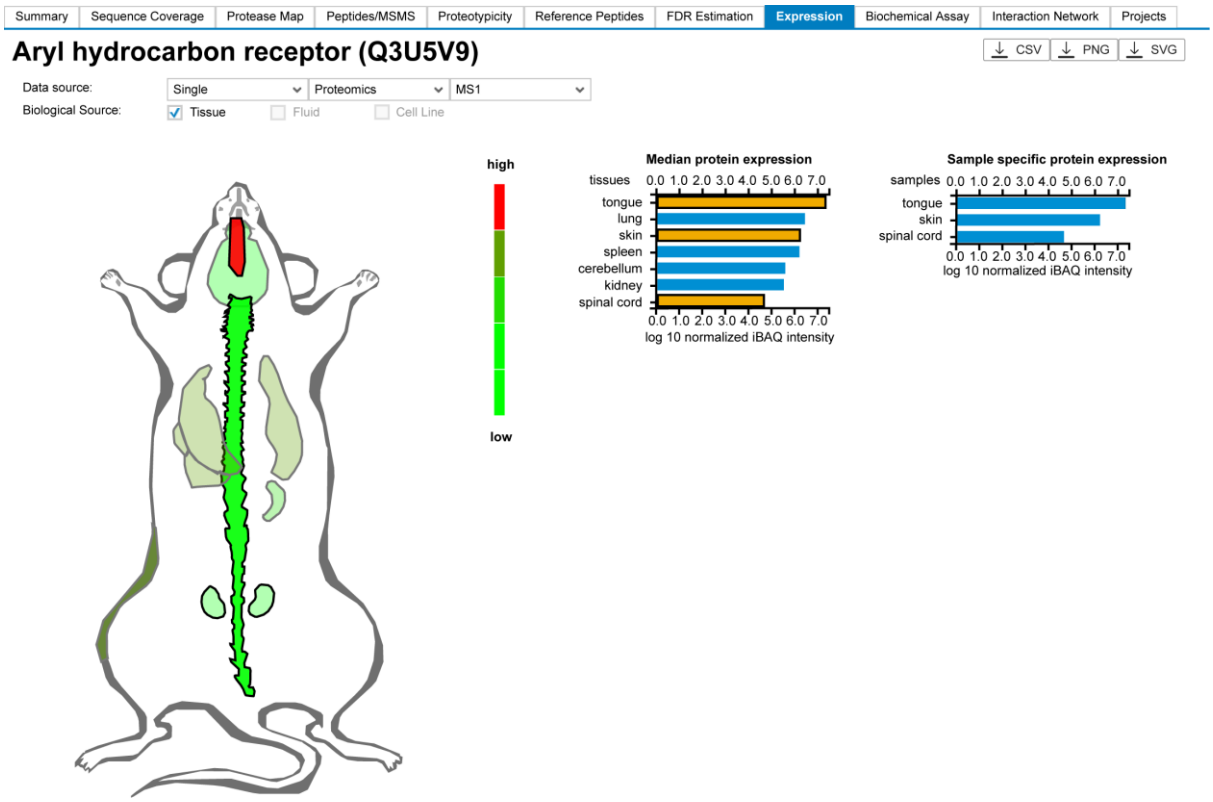

Figure S2. Interactive expression body map for *Mus musculus.* As previously available for *Homo sapiens,* the same interactive body map idea is used for visualizing quantitative expression data for every organism hosted by ProteomicsDB.

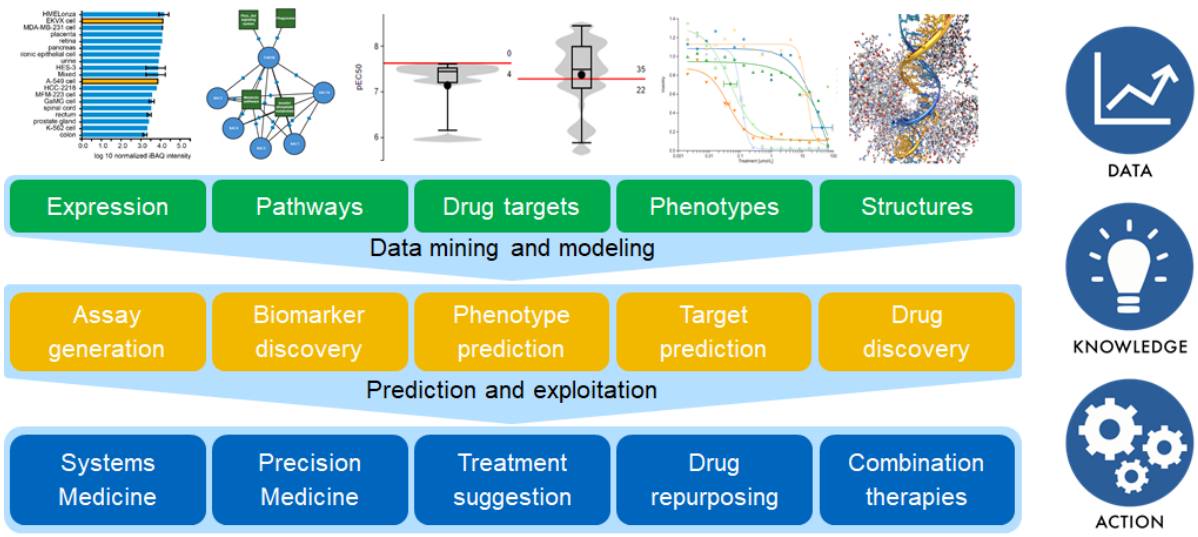

Figure S3. The future of ProteomicsDB: From data to knowledge to action. Future releases of ProteomicsDB will contain protein structure data. Applying data mining and knowledge discovery methods, new information will be generated which will be used to build new tools that will deliver the extracted knowledge to the user. The final goal of ProteomicsDB is to provide tools that will combine all data coming from the previous layers and take part in the decision making in modern research.