Review

Viewing the Proteome: How to Visualize Proteomics Data?

Eystein Oveland[1,2,3], Thilo Muth[4,5], Erdmann Rapp[4,5], Lennart Martens[6,7,*], Frode S. Berven[1,2,8] and Harald Barsnes[1]

[1] Proteomics Unit, Department of Biomedicine, University of Bergen, Norway

[2] KG Jebsen Centre for Multiple Sclerosis Research, Department of Clinical Medicine, University of Bergen, Norway

[3] Department of Clinical Medicine, University of Bergen, Norway

[4] Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany

[5] glyXera GmbH, Magdeburg, Germany

[6] Department of Medical Protein Research, VIB, Ghent, Belgium

[7] Department of Biochemistry, Ghent University, Ghent, Belgium

[8] Norwegian Multiple Sclerosis Competence Centre, Department of Neurology, Haukeland University Hospital, Bergen, Norway

[*] Corresponding author

Corresponding author: Prof. Dr. Lennart Martens, Department of Medical Protein Research, Ghent University, Albert Baertsoenkaai 3, B-9000 Ghent Belgium; e-mail: lennart.martens@UGent.be

Abbreviations:

Keywords: Visualization / Proteome Databases / Graphs / Raw Data

Abstract

Proteomics has become one of the main approaches for analyzing and understanding biological systems. Yet similar to other high-throughput analysis methods, the presentation of the large amounts of obtained data in easily interpretable ways remains challenging. In this review we therefore present an overview of the different ways in which proteomics software supports the visualization and interpretation of proteomics data. The unique challenges and current solutions for visualizing the different aspects of proteomics data, from acquired spectra *via* protein identification and quantification to pathway analysis, are discussed, and examples of the most useful visualization approaches are highlighted. Finally, we offer our ideas about future directions for proteomics data visualization.

Main text

Background

In recent years, mass spectrometry-based proteomics has undergone an immense technological progress and computational software tools have struggled to keep up with advancing high-throughput experimental methods [1]. In the previous decades, proteomics data analyses have mainly been performed with the goal to confirm a predefined hypothesis, whereas today the data itself is often the source from which the assumptions and conclusions originate. The focus has correspondingly shifted from mainly relying on *a priori* knowledge to a more unbiased data investigation, referred to as discovery proteomics, a move that is only partially counteracted by the emergence of targeted proteomics approaches [2, 3]. As a result, data analysis in proteomics has also moved increasingly towards data mining, in turn having direct implications on the way information is presented.

Advances in computer hardware and graphics have supported this evolution, and have led to substantial progress in the field of computational proteomics, both for standalone and for web-based applications [4, 5]. The development of new community standard mass spectrometry (MS) data formats such as mzML and mzIdentML has furthermore made it much easier to exchange proteomics data [6, 7]. Yet despite this constantly improving technology, the increasing size and complexity of proteomics data still poses its challenges – not only to the researchers, but also to the developers of visualization tools [8, 9].

We here therefore present existing useful and innovative approaches to tackle the visualization challenges in the field of mass spectrometry based proteomics, and provide our ideas about necessary future developments in the field. Yet in order to keep the review to a reasonable length, we have to restrict the scope of this text. While very interesting in their own right, we therefore cannot discuss the visualization of data from data independent

acquisition [10], from MALDI imaging mass spectrometry [11, 12], or from structural proteomics [13] in any detail. Also, as our focus is on the end-user, statistical packages such as R (http://www.r-project.org), Matlab (MathWorks, MA,USA) and SPSS (IBM, NY, USA), are not covered here either. Finally, it should be noted that this review does not attempt to provide a complete overview of all available tools for proteomics data processing or analysis, see instead [14-17], but rather focuses on tools that do something visually interesting with proteomics data.

The overall structure of this review is organized to follow the common workflow of proteomics experiments: from raw data, *via* identification and quantification, through pathway and network analysis, to whole proteome databases (**Figure 1**).

Visualizing Raw Data

The wet-lab part of proteomics experiments ends with the acquisition of raw data, i.e., the unprocessed data from the mass spectrometer, and it is vital to ensure that these data are of high quality and that any initial processing does not introduce deleterious effects. The instrument vendors have therefore developed their own tools to access, visualize and convert raw data, e.g., Thermo raw files can be viewed with Xcalibur (Thermo Scientific), and RawMeat (VAST Scientific and Thermo Scientific) can be used to visualize and compare raw files.

However, the proprietary raw files are in a poorly accessible format, often not directly compatible with existing open source software [18]. Thus conversion to the open standard format mzML [6] is recommended in order to allow interaction with, visualization of, and downstream analyses of these data [19]. Most raw files can be converted to open source formats using MSConvertGUI from ProteoWizard [20].

Quality control of samples and LC-MS performance

Quality control of the raw data is essential in order to obtain reliable identification and quantification, and reproducible results [21, 22]. First, the total ion chromatogram (TIC) should be examined to verify satisfactory injection and ionization of the sample. Second, the MS1 spectra should be investigated to detect contamination and confirm high resolution with sharp peaks, and to ensure high performance of both the mass spectrometric and the chromatographic part. Third, the MS2 spectra should be checked to ensure proper fragmentation and resolution.

Viewing the TIC is a quick way to ensure proper sample loading which can be inspected by SeeMS from the ProteoWizard package [23] (**Figure 2A**). Furthermore, while contamination by polymers and other non-peptide molecules can be detected in the TIC, the base peak intensity (BPI) is better for this purpose, as it shows the intensities of the most intense peaks in the MS1 spectra. Using MZmine [24] the typical 44 Da moieties of the frequent MS contaminant polyethylene glycol (PEG) [25] can be seen in a BPI chromatogram; such contaminants can also easily be visualized in 2D plots by viewing the m/z, retention time (RT) and intensity of the MS1 in the same chart (**Figure 2B**). Such 2D plots are also important to detect chromatographic shifts in RT introduced by poor chromatography.

To obtain better data acquisition, statistics about charge distribution, spray current, target fill-times and m/z with charge distributions can be visualized and investigated. Visualization of the (MS1) electrospray ion current trace is exemplified by the software RawMeat comparing three different raw files (**Figure 2C**).

Processing and comparison of raw data

Visual inspection and comparison of LC-MS runs is useful to uncover differences between runs. By selecting an MS1 feature in two different runs in the 2D view and inspecting them in detail in 3D, the differences become evident as illustrated by MSight [26] (**Figure 2D**). MSight also allows the manual measurement of distances in m/z and RT in the 2D plot and displays the MS1 spectrum for the selected RT-window.

Applying algorithms to remove noise in LC-MS data, centroiding of spectrum peaks and peak picking is important for downstream analyses, and visualization of these processes is useful to assess the impact of these steps. MZmine [24] enables visual interaction with the filtering prior to follow-up analyses. To reduce the file size and remove noise in the data, the beginning and end of LC-MS runs can be discarded, and the interactive visualization helps to set the thresholds (**Figure 2E, left**).

To compare peak detection methods one can visualize raw data together with peak picking and identification results. Using MZmine, peak picking is performed in three steps (mass detection, chromatogram building and peak deconvolution) and each step can be visualized graphically. A preview of the mass detection with a given noise level threshold illustrates which peaks will be discarded if the process is executed (**Figure 2E, right**).

Peak picking alters profile MS data in such a way that the volume of the peaks is represented by centroided peak height. The effect can be visualized in 3D using TOPPView from OpenMS [27] (**Figure 2F**). Viewing the peak-picked file can be useful to verify that low intensity peaks have been picked and that no artifacts have been introduced. TOPPView enables interactive visualization and comparison of two LC-MS runs at the MS1 (**Figure 2G**) and MS2 levels.

Quantification and analytical work

During MS1 quantification it is important that the respective MS2 spectra are linked to the correct MS1 feature. The data from proprietary raw files (Thermo, AB Sciex and Bruker) or open source mzXML (the current standard mzML is not yet supported) can be viewed, searched and quantified in MaxQuant [14]. The tool allows viewing LC-MS data in 2D (RT, m/z and intensity) and the MS1 and MS2 spectra can also be viewed. Progenesis LC-MS (Nonlinear Dynamics) is commercial software for label-free quantification, with powerful 2D and 3D visualization of the LC-MS data (proprietary or open source files) similar to MaxQuant. The software updates the visualization of the LC-MS 2D map after alignment of the LC-MS runs to be compared, and the degree of alignment is indicated by color. The volumes of the MS1 peaks can also be compared visually after assigning peptide identities.

In order to visualize and analyze chromatographic features and spectra in detail, the open source OpenChrom [28] tool allows users to implement their own methods, algorithms, filters, detectors or integrators, and supports manual interaction such as peak integration and quantification.

Visualizing Proteomics Identifications

The basis of any mass spectrometry based identification is a peptide to spectrum match (PSM), and most tools for analyzing proteomics data provide some form of spectrum viewer to visualize this match, allowing an expert to assess how well the peptide matches the spectrum, see for example [29-35]. A good spectrum viewer represents an excellent tool to check the quality of individual PSMs. Some viewers also go a step further by showing related details for the peptide to spectrum match, e.g., the spectrum viewer in PeptideShaker (http://peptide-shaker.googlecode.com) also includes a sequence fragmentation plot (**Figure 3A**).

Comparing multiple PSMs

In many cases it is necessary to look at more than a single PSM, for example to compare multiple PSMs mapping to the same peptide. And while it is possible to open one viewer per PSM, more advanced options are available. If two PSMs are to be compared, so-called mirrored spectra can be used, where one spectrum is shown above the x-axis and another spectrum below the x-axis (**Figure 3B**). Mirrored spectra can be for instance be created in TOPPView, compomics-utilities [29] and MS Manager (Advanced Chemistry Development Laboratories), amongst others.

If more than two PSMs are to be compared simultaneously, one option is to use PSM bubble plots [36], showing each fragment ion as a bubble where the size represents the intensity of the peak, the x-axis the m/z value and the y-axis the mass error (**Figure 3C**). With this approach it is possible to visualize hundreds of PSMs at the same time, e.g., to analyze fragmentation variability [37].

De novo sequencing

A problem closely related to visualizing PSMs is the inspection of *de novo* results, i.e., the mappings between a (partial) peptide sequence and a spectrum. The most popular commercial software for this purpose is PEAKS [35]. Recently an open source alternative became available called DeNovoGUI [38], which contains an easily interpretable way of displaying the *de novo* annotations on the spectrum (**Figure 3D**).

Linking spectra, peptides and proteins

The next step involves the visualization of the connections between spectra and peptides, and between peptides and proteins. This is complicated by the fact that a peptide can map to more than one protein, known as the protein inference problem [39]. The most

valuable approach is an interactive visualization combining tables and spectrum viewers, where the user can interact with the data by selecting the proteins, peptides and spectra to inspect. An example of this approach is found in PRIDE Inspector [30], which allows the user to select a protein to see all matched peptides, then select a peptide to see the spectrum itself, all inside the same display. PRIDE Inspector is particularly interesting as it allows this type of interactive visualization for both local data (in various standard formats) and for all the public data in PRIDE [40]. Loading local data is particularly interesting as a means of validating the file content before submitting the data to public repositories. Similar approaches are used by other tools, e.g., Proteome Discoverer (Thermo Scientific), Scaffold (http://www.proteomesoftware.com), PeptideShaker (http://peptide-shaker.googlecode.com) and the web-based MS-Viewer [41].

Post-translational modifications

Post-translational modifications (PTMs) and their site assignment scores can often assist the interpretation of biological activity [42]. Thus, when analyzing protein and peptide identifications the detection and visualization of PTMs are important, and the modification sites are usually visualized on the sequence or spectrum. For in depth analysis and visualization of PTM sites the Scaffold extension Scaffold PTM (http://www.proteomesoftware.com) can be used.

Targeted identification

For targeted LC-MS data the mass spectrometer vendors have developed commercial software to view the data, e.g., MultiQuant (AB SCIEX) and MassHunter (Agilent). However, Skyline [43] is the most frequently used free software to analyze and visualize selected reaction monitoring (SRM) data. The peaks can be investigated with respect to shape,

retention time of elution and intensity, and the results can be compared to previously identified spectra (**Figure 3E**).

Repurposing raw data

A final example in this section is an intriguing combination of raw data and peptide identifications. In large data sets, looking up the data for a particular identified peptide sequence can be time-consuming. Furthermore, if the expert decides upon visual inspection of these data that a second search should be carried out with additional options for post-translational modifications (PTM) that were not included in the original search, this search would have to be carried out on the full data complement. Systems such as Slice (http://slice.ionomix.com) and DICE (http://research.ionomix.com/cptac) developed by Askenazi *et al.*, therefore make it possible to quickly delve into the raw data supporting a particular peptide identification and, if required, to extract these data for targeted re-analysis.

Visualizing Quantitative Proteomics

Visualization of quantitative proteomics data concerns relating quantity and accuracy of the protein measurements, represented by the detected peptide MS1 or MS2 spectra [44], to the analyzed conditions in experimentally meaningful ways. Most software tools for protein quantification provide workflows containing graphical result representations, e.g., Progenesis, Proteome Discoverer (Thermo Scientific), MaxQuant [45] and IsobariQ [46] for discovery analyses; and Skyline [43], MultiQuant (AB SCIEX) and MassHunter (Agilent Technologies) for targeted analyses. However, post-processing steps using other software are often required to obtain conclusive quantitative results. We will here focus on the most common visualization techniques employed by such downstream software and show examples of how these can be used.

Protein ratio distributions

Quantitative proteomics data have to be normalized in order to be comparable. This normalization process can be visualized through scatter plots where the protein quantity values (here ratios) for all proteins are shown before processing, after log transformation, and after normalization (**Figure 4A**). Proteins with extreme values, i.e., outliers, stand out in the scatter plot as indicated in the figure. If the distribution is uneven or appears as several subpopulations, there is most likely an issue with these data, resulting from either data recording, or data processing. A limitation of scatter plots is that a high number of quantified proteins will result in many data points, which can be difficult to separate visually.

Histograms show how many values are present in predefined bin-intervals, and a continuous density plot presents a profile distribution of the ratios independent of pre-defined bins (**Figure 4A**). The histogram or density plot can visualize thousands of values, and can illustrate whether they are normally distributed. Normality can also be analyzed using Q-Q plots, comparing the actual probability quantiles (y-axis) to predicted quantiles (x-axis); if the two distributions are similar, the points will approximate the line y = x. Q-Q-plots can be generated in InfernoRDN (supersedes DAnTE [47]) and Excel. Histograms can be created using GProX [48], Perseus provided with MaxQuant [45] and the commercial software GraphPad Prism (GraphPad Software Inc.). Any statistical package such as R or SPSS will also be able to generate these visualizations.

Boxplots can display the spread of data across replicates (including possible outliers) for a condition while simultaneously providing a comparison to other conditions [49]. The box ranges from the first quartile (Q1, 25%) to the third quartile (Q3, 75%) of the distribution and represents the interquartile range; the line across the box indicates the median. The whiskers are lines extending from Q1 and Q3 to the most extreme data points defined based

on statistical formulas, e.g. Tukey, Altman or Spear, the latter extends to the minimum and maximum values [49]. Boxplots can be generated using BoxPlotR [50] (http://boxplot.tyerslab.com), a free web-tool that uses R in the background to plot data from delimited text files (**Figure 4B**). The spread in the data can be visualized further by including the individual data points as a so-called bee swarm which also highlights outliers. Boxplots can also be created using GProX and GraphPad, amongst others.

Another way to illustrate quantitative proteomics data is to plot the fold change (using a $\log_2$ transformation) versus the p-value ($-\log_{10}$ transformed) for all the quantified proteins. This generates a so-called volcano plot, which highlights the proteins with high fold changes and low p-values, and is well-suited for illustrating changes in large datasets (**Figure 4C**). It also shows that proteins with high fold changes do not necessarily have low p-values and are not necessarily the most trusted candidates. On the other hand, small changes with high p-values may be statistically better, but are not necessarily biologically interesting. For the generation of volcano plots tools such as GProX and Perseus can be used.

Protein/sample sub groups

Proteins with similar regulation between conditions are often biologically interesting. In Perseus, the expression level of proteins in different conditions can be presented as parallel line charts (conditions on the x-axis, expression level on the y-axis), and proteins with similar profiles will be visually linked. In order to extract proteins with similar profiles, a curve shape can be selected, and matching proteins listed (**Figure 4D**). Other software tools generating expression profiles are Progenesis and IsobariQ [46].

The individual expression levels of thousands of proteins across multiple conditions can be visualized by color intensity in heatmaps. A red-green color scheme (for down- and up-regulation) is often used, despite its obvious limits for colorblind people. Unsupervised

hierarchical clustering of the data enables grouping of conditions and/or proteins using dendrograms. In Perseus, it is straightforward to create heatmaps and investigate protein clusters using interactive dendrograms (**Figure 4E**). Progensis LC-MS and GProX are also capable of creating such heatmpas.

By using unsupervised clustering, i.e., by not relying on previous knowledge, it is possible to find subgroups not otherwise considered, thus allowing the visualization to guide the data exploration. Principal component analysis (PCA) is a way to investigate underlying differences between replicates and conditions in quantitative proteomics results [51]. It is generally most useful to look at the first two principal components in a two dimensional scatter plot (**Figure 4F**), but 3D plots with three components are also used.

PCA biplots illustrate whether replicates are reproducible, and if they differ in comparison to other groups. In Progenesis, features appearing close to a replicate in the PCA biplot have high abundance in that replicate, and features that cluster together have similar abundance profiles (**Figure 4F**). PCA plots can also be used to detect outliers. Other software generating PCA plots are Perseus and the Excel add-in Multibase (www.numericaldynamics.com).

Protein sets and intersections

It is often useful to use Venn diagrams to illustrate proteins that are uniquely present in one condition and not in others, or proteins only found in two or more conditions, e.g., to compare protein quantification lists based on accession numbers. There are 2^n possible intersections for n conditions, meaning that visually Venn diagrams are limited to comparing four conditions (2^4=16 intersections) as discussed in [52]. Venny (http://bioinfogp.cnb.csic.es/tools/venny) is a simple online interactive tool for comparing up

to four lists, and the common elements can easily be extracted for follow-up analysis (**Figure 4G**). Other software capable of generating Venn diagrams are Scaffold and Perseus.

Comparison of protein lists is also possible using Euler diagrams, a type of Venn diagram typically drawn with overlapping ellipses with their area proportional to the number of elements. Euler diagrams comparing three conditions can be created using eulerAPE [53] (http://www.eulerdiagrams.org).

When creating Venn or Euler diagrams it is crucial to make sure that unique peptides are the basis for quantification, and that the accession number does not represent a protein group, as this will completely ruin the basis for the comparison. It is also worth mentioning that an unequal distribution and degree of PTMs between compared conditions could obscure quantitative comparisons of data sets [42]. Dedicated software, such as the Scaffold extension Scaffold PTM (http://www.proteomesoftware.com), therefore ought to be used to ensure correction interpretation of the data.

Visualizing Pathways and Networks

Proteomics analyses often result in a large amount of protein identifications, which, without a biological context, can be rather overwhelming. An additional refinement of the results is therefore needed to unravel the underlying biological knowledge. This section provides an overview of methods and visualization tools for the post-processing analysis of protein identification data.

Functional annotation

In general, protein identifications can be enriched by adding functional knowledge via protein identifiers commonly referred to as accession numbers. Various web resources and

databases exist for querying comprehensive meta-information that has been assigned to a magnitude of the protein accessions. One of these is the Gene Ontology (GO) database featuring three structured vocabularies (ontologies) describing biological process, cellular component and molecular function [54]. A protein can be assigned to one or more GO terms and frequency-based calculations can be performed. The standalone tool Ontologizer features a table and graph visualization of data linking to over-represented GO terms [55]. In the graph view, each of the terms is further connected to associated child terms (**Figure 5A**). The software also provides methods to search for GO term over-representation and to perform correction for multiple testing.

The web-based application DAVID [56] offers functional annotations based on GO terms, but also provides the visualization of proteins on BioCarta (http://www.biocarta.com) and KEGG pathway maps [57]. As an example of the tools offered on the DAVID website, the functional annotation clustering allows for the grouping of redundant annotation terms. Here, the reported associations between genes/proteins and terms can be inspected *via* a heatmap visualization (**Figure 5B**).

Protein interactions and pathways

An important prerequisite to understand cellular processes is the knowledge about proteins and their interaction partners. This information is available *via* various public protein-protein interaction databases [58, 59]. The STRING [60] database holds data on known and predicted protein-protein interactions, and the website displays uploaded proteins together with interacting entities in a network (**Figure 5C**). The commercial software Ingenuity Pathway Analysis (IPA) (Ingenuity Systems, http://www.ingenuity.com) generates hypothetical protein-protein interactions based on a knowledge database and integrates data

from various omics formats. The inferred proteins can also be visualized in protein-protein interaction networks and canonical pathways.

An additional phylogenetic view enables insights on how well proteins are conserved in the taxonomy. Protein interaction and biological pathway analyses can be performed using the Reactome database [61], containing curated data and access to reaction-based relationships, mainly from *Homo sapiens*. The interactive website allows the detailed investigation of pathways and the analysis of user-supplied experimental data. One option is to submit a list of proteins (or genes), which can be used in an enrichment analysis to find enriched pathways in the supplied data. It is also possible to submit numerical expression data, resulting in color-coded nodes in the Reactome pathway display according to the quantitative values from the provided experimental data (**Figure 5D**). Expression data can also be visualized in tissues, e.g., via a human body visualization.

The KEGG database integrates genomic, biochemical and functional data by focusing on intermediate pathways [57]. Identified proteins can be mapped to their corresponding KEGG ontologies by the KEGG Automated Annotation Server (KAAS) [62] and representative pathways are visualized directly on the website. The interactive pathway explorer iPath 2.0 [63] presents a web-based application to visualize and analyze data in several regulatory and metabolic pathway maps (based upon KEGG information).

Going one step further in terms of customization capabilities, PathVisio [64] represents a useful stand-alone software package with the possibility to draw and edit biological pathways. The user can also load expression data and fully modify the visualization results.

Graphs and networks

The standalone visualization platform Cytoscape allows for the visualization of biomolecular interaction networks and integration of related attribute data [65]. Cytoscape supports a plethora of plugins [66]: among the most popular is BiNGO for calculating overrepresented GO terms and displaying them in a network [67]. The software also includes access to various interaction databases by means of the Proteomics Standard Initiative Common QUery InterfaCe (PSICQUIC) web service [68] – providing the basis for protein-protein interaction visualization.

In general, protein data can be modeled and visualized as graphs to reflect the high degree of connectedness with external meta-information, such as pathways and ontologies. Graphs have the advantage of showing different types of information and their relations at the same time, whereas a representation using multiple tables will increase in complexity with each dimension. However, graphs with a high number of entities can be hard to read. As a consequence, data filtering and querying strategies are needed to present the given information in a useful way. Graphs can be directly visualized in tools such as Cytoscape or Gephi (https://gephi.org), for further data exploration.

Cytoscape is currently the most used graph analysis software for biological data, mainly due to its numerous biology related plugins and its links to relevant ontologies and databases. Gephi on the other hand is a more general tool for displaying and exploring various types of networks and graphs, and is not as strongly linked to a specific use case. For example it provides a set of generic layout algorithms to change the shape of a graph during analysis, and includes various clustering and dynamic filtering options, allowing the user to work interactively with complex datasets. Both Cytoscape and Gephi are plugin-based and provide APIs (application programmer interfaces) for users wanting to create their own plugins.

Another example where graphs can be useful is when exploring protein inference. Each peptide and protein can be represented as nodes in a graph with an edge added between a peptide and a protein if the protein can be inferred by the given peptide. In this way the complexity of protein inference issues can be visualized (**Figure 5E**)**.** For an example of how this can be achieved using Cytoscape, see [69].

Note that in addition to choosing a useful visualization strategy, one should keep in mind the reliability of the underlying data and the often automatically inferred pathway information. A study by Müller *et al.* [70] demonstrated that suggested pathway information ought to be used with caution, as incorrect application of software may result in data misinterpretation. For further details on challenges related to pathway analysis, see [71].

Visualizing Proteome Databases

There are numerous repositories for proteomics data, for example: PRIDE, PeptideAtlas, the Global Proteome Machine Database (GPMDB), neXtProt, ProteomicsDB, and the Human Proteome Map [40, 72-76]. Common between all these repositories is that they contain large amounts of mass spectrometry generated proteomics data to be browsed and displayed in various ways. All of them display the proteins identified in the database, usually in tabular format, the peptide coverage and the MS/MS spectra associated with the identifications. In the following we focus on human data, as well as visualization options for large scale proteome data. PRIDE, GPMDB and neXtProt will not be discussed further as they have limited options for visualizing complex protein data, except through tabular formats.

<u>Mapping proteins to chromosomes</u>

An interesting way of displaying a large gathering of data as found in these databases is by linking the proteins to the chromosome where their corresponding genes are located [77]. Such a view is useful to gain an overview of the density of identified proteins on a given area of a given chromosome, and could be used as a viewing tool in the chromosome-centric human proteome project (C-HPP) [78], where one of the goals is to identify at least one expressed protein variant of all human genes for each chromosome.

A CircAtlas view, as used by the chromosome explorer in PeptideAtlas, is one way of making such a display. Here two circles are divided into the different chromosomes (genomic information) with PeptideAtlas density and observations plots in addition to SwissProt protein locations (**Figure 6A**).

Another way of linking the identified proteins to the chromosomes is by using bar charts, as in ProteomicsDB [72] (**Figure 6B**), where each chromosome is divided into sections, and the number of the identified genes and proteins is displayed for each section. Common for both approaches is that they give a quick overview of the chromosome coverage and how many proteins have been found. The challenge in such views is to obtain more detailed information about individual identified or not yet identified proteins.

Human body protein maps

Another aspect of large proteome databases is how to visualize where in the (human) body the protein/proteoform(s) have been found, in what quantity, and under which conditions. Due to the immense complexity of the data, this presents an enormous challenge, but is of great importance as it would allow the researcher to use proteomics data in a more systematic way, e.g., in systems biology. The complexity of this task is indicated by having hundreds of cell types in the body, and inside each cell multiple subcellular compartments. It is estimated that up to one million proteoforms may exists, and the number of conditions is

also very high, considering gender, age, diseases, and other biological and environmental factors [76, 79]. Considering this great complexity, the limited annotation of the data present in the databases is a major hurdle for allowing broader and more detailed visualizations. In our opinion, proper data annotation is therefore a subject that should be given increased attention in the future.

ProteomicsDB [72] has come up with a way to display some of this complex information through a human body heatmap (**Figure 6C**), represented by the amount of the selected protein observed from the proteomics experiments. The amount of protein in each location is based on iBAQ (intensity based absolute quantification) values calculated from the datasets, and gives the researcher a quick overview of where the protein has been observed in the body and approximately at which levels. They also give the users the option to select data from either males or females, and limit the selection to specific tissue, fluid or cell lines. Together this gives the researcher a quick overview of the complex collection of data.

The Human Proteome Map [73] has a similar display for viewing the proteome information based on the body location and quantity using data from label-free analysis displayed through a heat map. Here it is also possible to see the peptide coverage and the individual peptide abundances for the 30 locations analyzed, which can be useful, for example, when selecting signature peptides for targeted proteomics assays (**Figure 6D**).

Both these databases are based on gathering and storing raw data and results from mass spectrometry based proteomics experiments on human material, and covers observations of more than 90% of the human proteome (17-18 000 proteins). It should, however, be noted that the background data should be taken into consideration when determining what the displayed information really tells the user, what it can be used for, and how much trust one

can have in the displayed data. For example, a recent publication concluded that the experimental data from the two above-mentioned databases should be used with caution [80].

Protein size distributions

CSF-PR (cerebrospinal fluid proteome resource) is another recently launched database, containing more than 3000 protein entries identified from human cerebrospinal fluid [81]. An interesting visualization option in this database is the viewing of the protein size distribution based on SDS-PA gels (**Figure 6E**), which gives a quick impression of the possible presence of proteoforms and truncation products. The estimated amount of protein in each gel fraction is visualized either based on the number of peptides or spectra observed, or the average precursor intensity. This display of the data thus provides easy access to information that can be very useful when selecting sample processing methods or proteotypic peptides for targeted proteomics experiments.

Discussion

From the examples above it is clear that interactive visualization already plays a crucial role in proteomics data analysis. Due to the strong increase in the amount of data, plus its growing size and complexity, this will become even more important in the coming years. However, there are still obstacles that need to be solved in order to achieve a more streamlined use of available visualization techniques. The first set of challenges is related to the way the proteomics data is stored. There is a need for standard formats so that data can be easily shared between different tools. This point is gradually being addressed by the implementation of the new standard data formats in proteomics, such as mzML, mzIdentML and qcML [7, 19, 22], plus the development of common resources for sharing data in these standard formats such as ProteomeXchange [82].

But even though the new formats are optimized for capturing as much information as possible about proteomics experiments and results, they are not tailored towards visualization and interaction. The demand for easy-to-use interactive visualizations to explore the growing amounts of available proteomics data imposes a very different set of requirements for data storage and accessibility, e.g., the need for faster reading times [83]. One example is the so called in-memory database for quick data access used by ProteomicsDB [72]. Another example is the novel way in which the data is stored in tools such as Slice (http://slice.ionomix.com) in order to ensure rapid access to the information needed for the visualizations.

In addition to challenges in the way data is stored, hardware capabilities also come into play when visualizing very large proteomics data sets. Visualizing an entire proteomics data set in a single display is typically not feasible, unless one has access to very high performance computing infrastructure. One is therefore frequently left with one of two choices: either filter the data in some way or buy more powerful hardware. And while the latter in many cases will solve the problem, it is not always possible or desirable. Additionally, the software has to be capable of utilizing more powerful resources, which is not always the case. More clever ways of interacting with data are therefore needed, allowing the user to focus on those elements of the data that are interesting to their specific research questions, while at the same time not hiding any important data. Indeed, often it is too easy to focus too much on the question (or desired outcome), and thus forget about the bigger context, for instance in the case of pathway analysis tools that allow the user to filter the data to only display interactions related to a specific pathway.

Interactivity is a key element and is expected by most users experienced with modern visual interactive displays. Being able to interact with the visualizations and thus continue exploring the data is therefore essential. Interactivity dramatically increases the usefulness of

most visualizations, e.g., allowing the user to easily zoom in on interesting areas or simply get information about a specific data point in a plot. Linking multiple interactive visualizations based on the same data is therefore the next obvious step, i.e., allowing the user to see and interact with different aspects or elements of a dataset at the same time, for example by simultaneously displaying a PCA plot, a protein profile plot and a table representing a set of protein ratios comparing two or more groups. The user can then locate protein clusters in the PCA plot, select a cluster to inspect the related protein profiles, and finally see the table for details about the selected proteins. This type of interaction, a long-time feature of dedicated data analysis tools such as Spotfire (http://spotfire.tibco.com), is already supported in tools such as Perseus and the hope is that more tools will follow.

Having sophisticated tools is not enough however, as proper visualizations also depend on high quality and well-annotated data. Requesting that data submitted to online proteomics repositories contains the required annotations [84] to make better use of the data should therefore be mandatory before publishing. With tools such as the ProteomeXchange submission tool [82] the annotation of information such as species is mandatory, but further information about the samples, e.g., gender, age, disease state, is not, and ought to be easier to annotate. Having such information available would greatly increase the value of the data and open up for new ways of visualizing and interacting with the data. It should however, also be noted that privacy issues or ethical concerns can stand in the way of thoroughly annotating patient-derived samples.

Visualizing proteomics data also results in new opportunities for crowdsourcing [85, 86], i.e., "hiring" the crowd to participate in proteomics research. The most successful crowdsourcing project related to proteomics is Foldit (http://fold.it), an online puzzle game dedicated to protein folding. By visualizing three-dimensional protein structures and letting players interact with these to find the optimal fold (and beat other players to the best score!),

the project has already resulted in both improved teaching material related to protein structures [87] and led to novel scientific knowledge [88, 89]. This shows the huge (and largely untapped) potential of crowdsourcing, which in most cases starts with being able to visualize and interact with the data.

We have just started to see the huge potential of interactive visualizations in proteomics. Although the technology of instruments and computer hardware is improving rapidly, the visualization methods for proteomics data are still lagging behind their real potential. However, with the development of common data standards to simplify data sharing, more efficient hardware usage, smarter software based on multiple linked and interactive visualizations, plus potentially enrolling the crowd as part of the work, the future of proteomics will hopefully soon become even more interactive and more visually enticing.

Acknowledgements

E.O. and F.S.B. acknowledges the support by the Western Norway Regional Health Authority, the Meltzer Foundation, Kjell Alme's Legacy for Research in Multiple Sclerosis, the Frank Mohn Foundation, the Research Council of Norway and the Kristian Gerhard Jebsen Foundation. T.M. and E.R. acknowledge the support by Max Planck Society. L.M. acknowledges the support of Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks") and the IWT SBO grant 'INSPECTOR' (120025). H.B. is supported by the Research Council of Norway.

The authors have no competing financial or commercial interests.

References

[1] Martin, S. F., Falkenberg, H., Dyrlund, T. F., Khoudoli, G. A., *et al.*, PROTEINCHALLENGE: Crowd sourcing in proteomics analysis and software development. *J Proteomics* 2013, *88*, 41-46.

[2] Pan, S., Aebersold, R., Chen, R., Rush, J., *et al.*, Mass spectrometry based targeted protein quantification: methods and applications. *J Proteome Res* 2009, *8*, 787-797.

[3] Gallien, S., Duriez, E., Domon, B., Selected reaction monitoring applied to proteomics. *J Mass Spectrom* 2011, *46*, 298-312.

[4] Kelchtermans, P., Bittremieux, W., De Grave, K., Degroeve, S., *et al.*, Machine learning applications in proteomics research: how the past can boost the future. *Proteomics* 2014, *14*, 353-366.

[5] Verheggen, K., Barsnes, H., Martens, L., Distributed computing and data storage in proteomics: many hands make light work, and a stronger memory. *Proteomics* 2014, *14*, 367-377.

[6] Deutsch, E., mzML: a single, unifying data format for mass spectrometer output. *Proteomics* 2008, *8*, 2776-2777.

[7] Jones, A. R., Eisenacher, M., Mayer, G., Kohlbacher, O., *et al.*, The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol Cell Proteomics* 2012, *11*, M111 014381.

[8] O'Donoghue, S. I., Gavin, A. C., Gehlenborg, N., Goodsell, D. S., *et al.*, Visualizing biological data-now and in the future. *Nat Methods* 2010, *7*, S2-4.

[9] Gehlenborg, N., O'Donoghue, S. I., Baliga, N. S., Goesmann, A., *et al.*, Visualization of omics data for systems biology. *Nat Methods* 2010, *7*, S56-68.

[10] Dang, X., Scotcher, J., Wu, S., Chu, R. K., *et al.*, The first pilot project of the consortium for top-down proteomics: A status report. *Proteomics* 2014, *14*, 1130-1140.

[11] Alexandrov, T., MALDI imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC bioinformatics* 2012, *13 Suppl 16*, S11.

[12] Gessel, M. M., Norris, J. L., Caprioli, R. M., MALDI imaging mass spectrometry: Spatial molecular analysis to enable a new age of discovery. *Journal of proteomics* 2014, *107C*, 71-82.

[13] Hyung, S. J., Ruotolo, B. T., Integrating mass spectrometry of intact protein complexes into structural proteomics. *Proteomics* 2012, *12*, 1547-1564.

[14] McHugh, L., Arthur, J. W., Computational methods for protein identification from mass spectrometry data. *PLoS Comput Biol* 2008, *4*, e12.

[15] Perez-Riverol, Y., Wang, R., Hermjakob, H., Muller, M., *et al.*, Open source libraries and frameworks for mass spectrometry based proteomics: a developer's perspective. *Biochim Biophys Acta* 2014, *1844*, 63-76.

[16] Hoopmann, M. R., Moritz, R. L., Current algorithmic solutions for peptide-based proteomics data generation and identification. *Curr Opin Biotechnol* 2013, *24*, 31-38.

[17] Vaudel, M., Sickmann, A., Martens, L., Current methods for global proteome identification. *Expert review of proteomics* 2012, *9*, 519-532.

[18] Martens, L., Nesvizhskii, A. I., Hermjakob, H., Adamski, M., *et al.*, Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories. *Proteomics* 2005, *5*, 3501-3505.

[19] Martens, L., Chambers, M., Sturm, M., Kessner, D., *et al.*, mzML--a community standard for mass spectrometry data. *Mol Cell Proteomics* 2011, *10*, R110 000133.

[20] Chambers, M. C., Maclean, B., Burke, R., Amodei, D., *et al.*, A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotech* 2012, *30*, 918-920.

[21] Tabb, D. L., Quality assessment for clinical proteomics. *Clin Biochem* 2013, *46*, 411-420.

[22] Walzer, M., Pernas, L. E., Nasso, S., Bittremieux, W., *et al.*, qcML: An Exchange Format for Quality Control Metrics from Mass Spectrometry Experiments. *Mol Cell Proteomics* 2014, *13*, 1905-1913.

[23] Kessner, D., Chambers, M., Burke, R., Agus, D., Mallick, P., ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* 2008, *24*, 2534-2536.

[24] Pluskal, T., Castillo, S., Villar-Briones, A., Oresic, M., MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* 2010, *11*, 395.
[25] Keller, B. O., Sui, J., Young, A. B., Whittal, R. M., Interferences and contaminants encountered in modern mass spectrometry. *Analytica chimica acta* 2008, *627*, 71-81.
[26] Palagi, P. M., Walther, D., Quadroni, M., Catherinet, S., *et al.*, MSight: an image analysis software for liquid chromatography-mass spectrometry. *Proteomics* 2005, *5*, 2381-2384.
[27] Sturm, M., Bertsch, A., Gropl, C., Hildebrandt, A., *et al.*, OpenMS - an open-source software framework for mass spectrometry. *BMC bioinformatics* 2008, *9*, 163.
[28] Wenig, P., Odermatt, J., OpenChrom: a cross-platform open source software for the mass spectrometric analysis of chromatographic data. *BMC bioinformatics* 2010, *11*, 405.
[29] Barsnes, H., Vaudel, M., Colaert, N., Helsens, K., *et al.*, compomics-utilities: an open-source Java library for computational proteomics. *BMC Bioinformatics* 2011, *12*, 70.
[30] Wang, R., Fabregat, A., Rios, D., Ovelleiro, D., *et al.*, PRIDE Inspector: a tool to visualize and validate MS proteomics data. *Nat Biotechnol* 2012, *30*, 135-137.
[31] Chambers, M. C., Maclean, B., Burke, R., Amodei, D., *et al.*, A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* 2012, *30*, 918-920.
[32] Barsnes, H., Huber, S., Sickmann, A., Eidhammer, I., Martens, L., OMSSA Parser: an open-source library to parse and extract data from OMSSA MS/MS search results. *Proteomics* 2009, *9*, 3772-3774.
[33] Muth, T., Vaudel, M., Barsnes, H., Martens, L., Sickmann, A., XTandem Parser: An open-source library to parse and analyse X!Tandem MS/MS search results. *Proteomics* 2010, *10*, 1522-1524.
[34] Colaert, N., Barsnes, H., Vaudel, M., Helsens, K., *et al.*, Thermo-msf-parser: an open source Java library to parse and visualize Thermo Proteome Discoverer msf files. *J Proteome Res* 2011, *10*, 3840-3843.
[35] Ma, B., Zhang, K., Hendrie, C., Liang, C., *et al.*, PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2003, *17*, 2337-2342.
[36] Barsnes, H., Eidhammer, I., Martens, L., FragmentationAnalyzer: An open-source tool to analyze MS/MS fragmentation data. *Proteomics* 2010, *10*, 1087-1090.
[37] Barsnes, H., Eidhammer, I., Martens, L., A global analysis of peptide fragmentation variability. *Proteomics* 2011, *11*, 1181-1188.
[38] Muth, T., Weilnbock, L., Rapp, E., Huber, C. G., *et al.*, DeNovoGUI: an open source graphical user interface for de novo sequencing of tandem mass spectra. *J Proteome Res* 2014, *13*, 1143-1146.
[39] Nesvizhskii, A. I., Aebersold, R., Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* 2005, *4*, 1419-1440.
[40] Martens, L., Hermjakob, H., Jones, P., Adamski, M., *et al.*, PRIDE: the proteomics identifications database. *Proteomics* 2005, *5*, 3537-3545.
[41] Baker, P. R., Chalkley, R. J., MS-viewer: a web-based spectral viewer for proteomics results. *Mol Cell Proteomics* 2014, *13*, 1392-1396.
[42] Olsen, J. V., Mann, M., Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol Cell Proteomics* 2013, *12*, 3444-3452.
[43] MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., *et al.*, Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 2010, *26*, 966-968.
[44] DeSouza, L. V., Siu, K. W., Mass spectrometry-based quantification. *Clin Biochem* 2013, *46*, 421-431.
[45] Cox, J., Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 2008, *26*, 1367-1372.
[46] Arntzen, M. O., Koehler, C. J., Barsnes, H., Berven, F. S., *et al.*, IsobariQ: software for isobaric quantitative proteomics using IPTL, iTRAQ, and TMT. *J Proteome Res* 2011, *10*, 913-920.
[47] Polpitiya, A. D., Qian, W. J., Jaitly, N., Petyuk, V. A., *et al.*, DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics* 2008, *24*, 1556-1558.

[48] Rigbolt, K. T., Vanselow, J. T., Blagoev, B., GProX, a user-friendly platform for bioinformatics analysis and visualization of quantitative proteomics data. *Molecular & cellular proteomics : MCP* 2011, *10*, O110 007450.
[49] Streit, M., Gehlenborg, N., Bar charts and box plots. *Nature methods* 2014, *11*, 117.
[50] Spitzer, M., Wildenhain, J., Rappsilber, J., Tyers, M., BoxPlotR: a web tool for generation of box plots. *Nature methods* 2014, *11*, 121-122.
[51] Ringner, M., What is principal component analysis? *Nat Biotechnol* 2008, *26*, 303-304.
[52] Lex A, Gehlenborg, N., Points of view: Sets and intersections. *Nat Methods* 2014, *11*.
[53] Micallef, L., Rodgers, P., eulerAPE: Drawing Area-Proportional 3-Venn Diagrams Using Ellipses. *PloS one* 2014, *9*, e101717.
[54] Blake, J. A., Harris, M. A., The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* 2008, *Chapter 7*, Unit 7 2.
[55] Bauer, S., Grossmann, S., Vingron, M., Robinson, P. N., Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* 2008, *24*, 1650-1651.
[56] Huang da, W., Sherman, B. T., Tan, Q., Kir, J., *et al.*, DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res* 2007, *35*, W169-175.
[57] Kanehisa, M., Goto, S., KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000, *28*, 27-30.
[58] Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., *et al.*, IntAct: an open source molecular interaction database. *Nucleic Acids Res* 2004, *32*, D452-455.
[59] Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., *et al.*, DIP: the database of interacting proteins. *Nucleic Acids Res* 2000, *28*, 289-291.
[60] Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., *et al.*, STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 2013, *41*, D808-815.
[61] Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., *et al.*, Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 2005, *33*, D428-432.
[62] Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., Kanehisa, M., KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 2007, *35*, W182-185.
[63] Yamada, T., Letunic, I., Okuda, S., Kanehisa, M., Bork, P., iPath2.0: interactive pathway explorer. *Nucleic Acids Res* 2011, *39*, W412-415.
[64] van Iersel, M. P., Kelder, T., Pico, A. R., Hanspers, K., *et al.*, Presenting and exploring biological pathways with PathVisio. *BMC bioinformatics* 2008, *9*, 399.
[65] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., *et al.*, Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003, *13*, 2498-2504.
[66] Saito, R., Smoot, M. E., Ono, K., Ruscheinski, J., *et al.*, A travel guide to Cytoscape plugins. *Nature methods* 2012, *9*, 1069-1076.
[67] Maere, S., Heymans, K., Kuiper, M., BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 2005, *21*, 3448-3449.
[68] Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F. S., *et al.*, PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nature methods* 2011, *8*, 528-529.
[69] Vaudel, M., Venne, A. S., Berven, F. S., Zahedi, R. P., *et al.*, Shedding light on black boxes in protein identification. *Proteomics* 2014, *14*, 1001-1005.
[70] Muller, T., Schrotter, A., Loosse, C., Helling, S., *et al.*, Sense and nonsense of pathway analysis software in proteomics. *J Proteome Res* 2011, *10*, 5398-5408.
[71] Khatri, P., Sirota, M., Butte, A. J., Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol* 2012, *8*, e1002375.
[72] Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., *et al.*, Mass-spectrometry-based draft of the human proteome. *Nature* 2014, *509*, 582-587.
[73] Kim, M. S., Pinto, S. M., Getnet, D., Nirujogi, R. S., *et al.*, A draft map of the human proteome. *Nature* 2014, *509*, 575-581.
[74] Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., *et al.*, The PeptideAtlas project. *Nucleic Acids Res* 2006, *34*, D655-658.

[75] Craig, R., Cortens, J. P., Beavis, R. C., Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res* 2004, *3*, 1234-1242.
[76] Lane, L., Argoud-Puy, G., Britan, A., Cusin, I., *et al.*, neXtProt: a knowledge platform for human proteins. *Nucleic Acids Res* 2012, *40*, D76-83.
[77] Guo, F., Wang, D., Liu, Z., Lu, L., *et al.*, CAPER: a chromosome-assembled human proteome browsER. *J Proteome Res* 2013, *12*, 179-186.
[78] Paik, Y. K., Jeong, S. K., Omenn, G. S., Uhlen, M., *et al.*, The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat Biotechnol* 2012, *30*, 221-223.
[79] Vickaryous, M. K., Hall, B. K., Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol Rev Camb Philos Soc* 2006, *81*, 425-455.
[80] Ezkurdia, I., Vazquez, J., Valencia, A., Tress, M., Analyzing the First Drafts of the Human Proteome. *J Proteome Res* 2014.
[81] Guldbrandsen, A., Vethe, H., Farag, Y., Oveland, E., *et al.*, In-depth characterization of the cerebrospinal fluid proteome displayed through the CSF Proteome Resource (CSF-PR). *Mol Cell Proteomics* 2014.
[82] Juan A Vizcaíno, Eric W Deutsch, Rui Wang, Attila Csordas, *et al.*, ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* 2014, *32*, 223–226.
[83] Teleman, J., Dowsey, A. W., Gonzalez-Galarza, F. F., Perkins, S., *et al.*, Numerical compression schemes for proteomics mass spectrometry data. *Mol Cell Proteomics* 2014, *13*, 1537-1542.
[84] Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P. A., *et al.*, The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol* 2007, *25*, 887-893.
[85] Barsnes, H., Martens, L., Crowdsourcing in proteomics: public resources lead to better experiments. *Amino Acids* 2013, *44*, 1129-1137.
[86] Good, B. M., Su, A. I., Crowdsourcing for bioinformatics. *Bioinformatics* 2013, *29*, 1925-1933.
[87] Farley, P. C., Using the computer game "FoldIt" to entice students to explore external representations of protein structure in a biochemistry course for nonmajors. *Biochem Mol Biol Educ* 2013, *41*, 56-57.
[88] Eiben, C. B., Siegel, J. B., Bale, J. B., Cooper, S., *et al.*, Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nat Biotechnol* 2012, *30*, 190-192.
[89] Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., *et al.*, Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nat Struct Mol Biol* 2011, *18*, 1175-1177.
[90] Sturm, M., Bertsch, A., Gröpl, C., Hildebrandt, A., *et al.*, OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics* 2008, *9*.

Figure legends

Figure 1: Visualizing Proteomics Data. An overview of the main topics covered in this review.

Figure 2: Visualizing Raw Data. A) TIC from an LC-MS sample run on an Orbitrap instrument and analyzed with SeeMS (ProteoWizard [31]). The TIC is on average 1xE9 which is in the upper limit of a proper peptide sample injection. **B)** Base peak intensity (BPI) and 2D plot of the MS1 from a PEG contaminated sample (MZmine [24]). **C)** Typical (MS1) electrospray ion current trace of three LC-MS runs illustrated by RawMeat (VAST Scientific and Thermo Scientific). **D)** MS1 feature comparison in 2D showing a peak with elution width of 0.7 min compared to a peak with 0.3 min elution. The respective peaks are shown in 3D enabling investigation of co-eluting peaks (MSight [26]). **E)** Filtering of LC-MS raw data by RT window and intensity level thresholds using MZMine [24]. **F)** 3D view of an LC-MS run before (left) and after peak picking (right) (TOPPView [90]). **G)** Illustration of a MS1 native peak (i) and a MS1 peak-picked (ii) by mirror view (left) and an overlay view zoomed in on one peak (right). The MS1 native peak has a bell-shaped curve and a peak width of approximately 0.02 Th (m/z) (TOPPView [90]).

Figure 3: **Visualizing Proteomics Identifications. A)** Sequence fragmentation plot, linking the intensity of the detected fragment ions to the amino acid sequence of the identified peptide. B-ions in blue blow the sequence and y-ions in red above the sequence. (PeptideShaker, http://peptide-shaker.googlecode.com). **B)** Mirrored spectra with asterisks indicating unique peaks (compomics-utilities [29]). **C)** Bubble plot showing two peptide-spectrum matches for the same spectrum (FragmentationAnalyzer [36]). **D)** De novo sequence

annotation in DeNovoGUI [38]. **E)** The peak curves and areas of three SRM transitions (overlaid) for quantifying the respective peptide in three different conditions (Skyline [43]).

Figure 4: Visualizing Quantitative Proteomics. A) Scatter plots illustrating transformation and normalization of a proteomics data set (Microsoft Excel), the same data set alternatively viewed as histogram (GraphPad Prism, GraphPad Software Inc.) and density plot. **B)** Boxplots comparing three different conditions (top), and with the data points presented as a bee swarm (bottom) (BoxPlotR [50]). **C)** Volcano plot showing the distribution of quantified proteins according to p-value and fold change, indicating significance level with a red line and color coded degree of fold change (Microsoft Excel). **D)** Line chart illustrating unsupervised clustering of protein expression profiles for four different conditions. The proteins in red fit the user-selected expression profile and can be investigated further (Perseus, http://www.perseus-framework.org). **E)** Heatmap showing the expression levels of the proteins and unsupervised clustering of the conditions (x-axis) and proteins (y-axis) as dendrograms. E.g., the proteins that group together (asterisk) are upregulated in condition 2 and 3 and not in 1 and 4, contributing to that 2 and 3 are more similar and group together in the dendrogram (Perseus, http://www.perseus-framework.org). **F)** PCA biplot showing peptide features as a cloud and replicates from the same condition as filled color-coded circles, enabling detection of outliers (Progenesis LC-MS, Nonlinear Dynamics). **G)** Venn diagram comparing the accession numbers of proteins identified in four different conditions (A, B, C, D) illustrating the intersections between the proteomes (Venny, http://bioinfogp.cnb.csic.es/tools/venny).

Figure 5: **Visualizing Pathways and Network. A)** Ontology graph view in the Ontologizer. Overrepresented GO terms are shown in green. The example shows a test dataset attributed to the small molecule metabolic process ontology (Ontologizer [55]). **B)** Heatmap view in the DAVID functional classification tool. The green boxes represent reported associations between annotation terms and proteins (DAVID [56]). **C)** STRING protein-protein interaction display (with known and predicted functional partners), showing the WNT7a signaling protein with interaction partners (STRING [60]). **D)** Pathway diagram view from Reactome, showing expression data on the metabolism of nucleotides: Ecto-5-prime-nucleotidase (CD73) catalyzes the conversion of purine 5-prime mononucleotides to nucleosides. A color range is used to distinguish the user-supplied numerical expression values, from yellow (highest values) to dark blue (lowest values) (Reactome [61]). **E)** Protein inference visualization in Cytoscape, showing the inference of proteins (in red) from peptides (in blue).

Figure 6: **Visualizing Proteome Databases. A)** CircAtlas interactive view from the chromosome explorer in PeptideAtlas, where two circles are divided into chromosomes (genomic information) with the ratio PeptideAtlas/UniProt density (blue) and the number of PeptideAtlas observations (green). A chromosome (the circled numbers) can be investigated in detail and a specific area of interest (pink highlight) chosen by the user. Clicking a specific protein or peptide opens the respective data in UniProt or PeptideAtlas (PeptideAtlas [74]). **B)** ProteomicsDB has divided the chromosome into sections, and the number of genes and proteins identified is displayed as bar charts (ProteomicsDB, https://www.proteomicsdb.org). **C)** Heatmap of the protein expression in the human body from ProteomicsDB, here exemplified by serum albumin (ProteomicsDB, https://www.proteomicsdb.org). **D)** The Human Proteome Map visualizing information about the peptide coverage and the individual peptide abundances for the 30 locations

analyzed (Human Proteome Map [73]). **E)** CSF-PR displays information about the size distribution of CSF proteins from an SDS-PA gel. Fraction number (from heavy to light) on the x-axis, peptide count on the y-axis. Blue bars represent molecular weight standards, peptides in green. Top: Apolipoprotein B-100, bottom: Secretogranin-1.(CSF-PR [81])

Figure 1

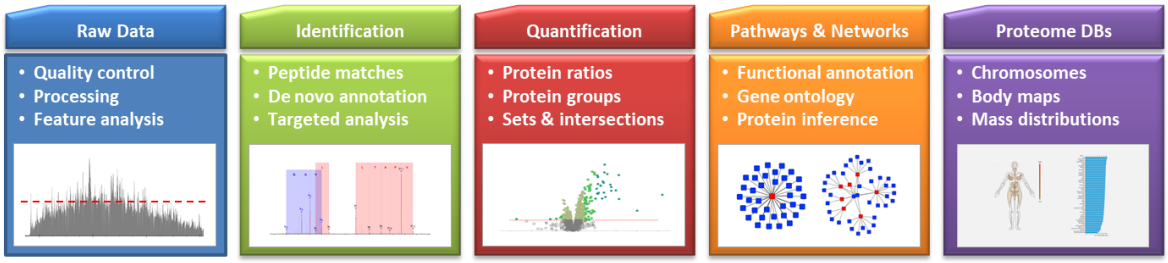

Raw Data
• Quality control
• Processing
• Feature analysis
Identification
• Peptide matches
• De novo annotation
• Targeted analysis
Quantification
• Protein ratios
• Protein groups
• Sets & intersections
Pathways & Networks
• Functional annotation
• Gene ontology
• Protein inference
Proteome DBs
• Chromosomes
• Body maps
• Mass distributions

Figure 2

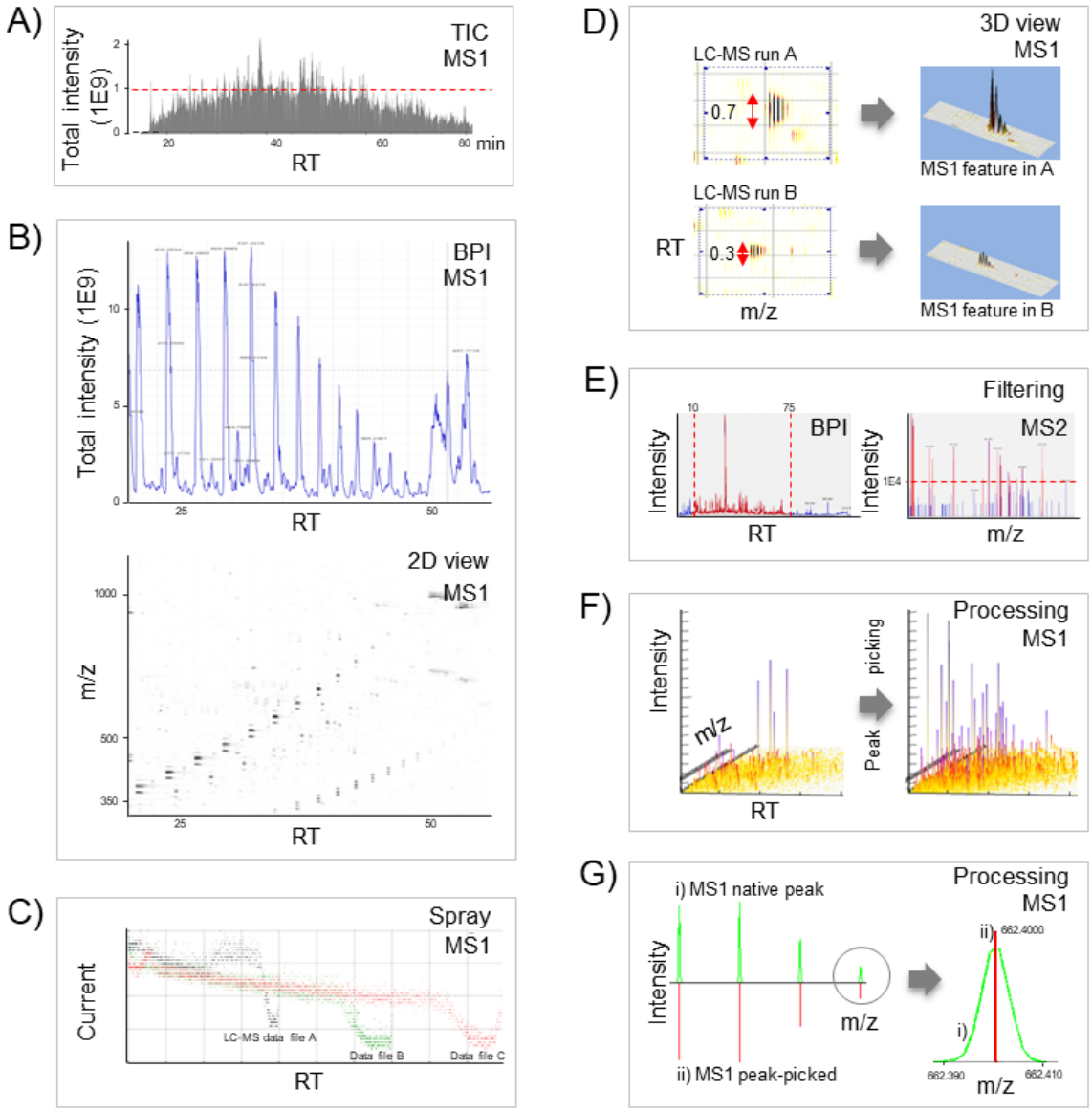

A)
TIC
MS1
Total intensity (1E9)
RT
min
B)
BPI
MS1
Total intensity (1E9)
RT
2D view
MS1
m/z
RT
C)
Spray
MS1
Current
LC-MS data file A
Data file B
Data file C
RT
D)
3D view
MS1
LC-MS run A
0.7
MS1 feature in A
LC-MS run B
RT
0.3
m/z
MS1 feature in B
E)
Filtering
BPI
MS2
Intensity
RT
Intensity
m/z
F)
Processing
MS1
Intensity
m/z
RT
Peak picking
G)
Processing
MS1
i) MS1 native peak
Intensity
m/z
ii) MS1 peak-picked
ii)
662.4000
i)
662.390
662.410
m/z

Figure 3

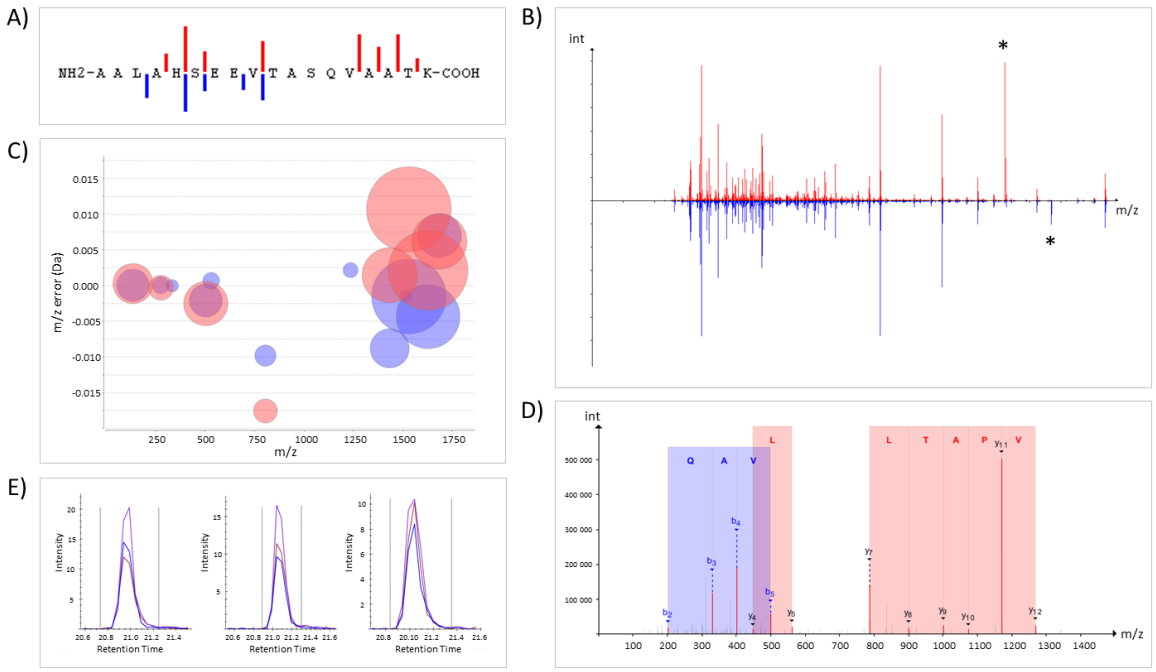

A)
NH2-A A L A H S E E V T A S Q V A A T K-COOH
B)
int
m/z
C)
m/z error (Da)
m/z
D)
int
m/z
E)
Intensity
Retention Time

Figure 4

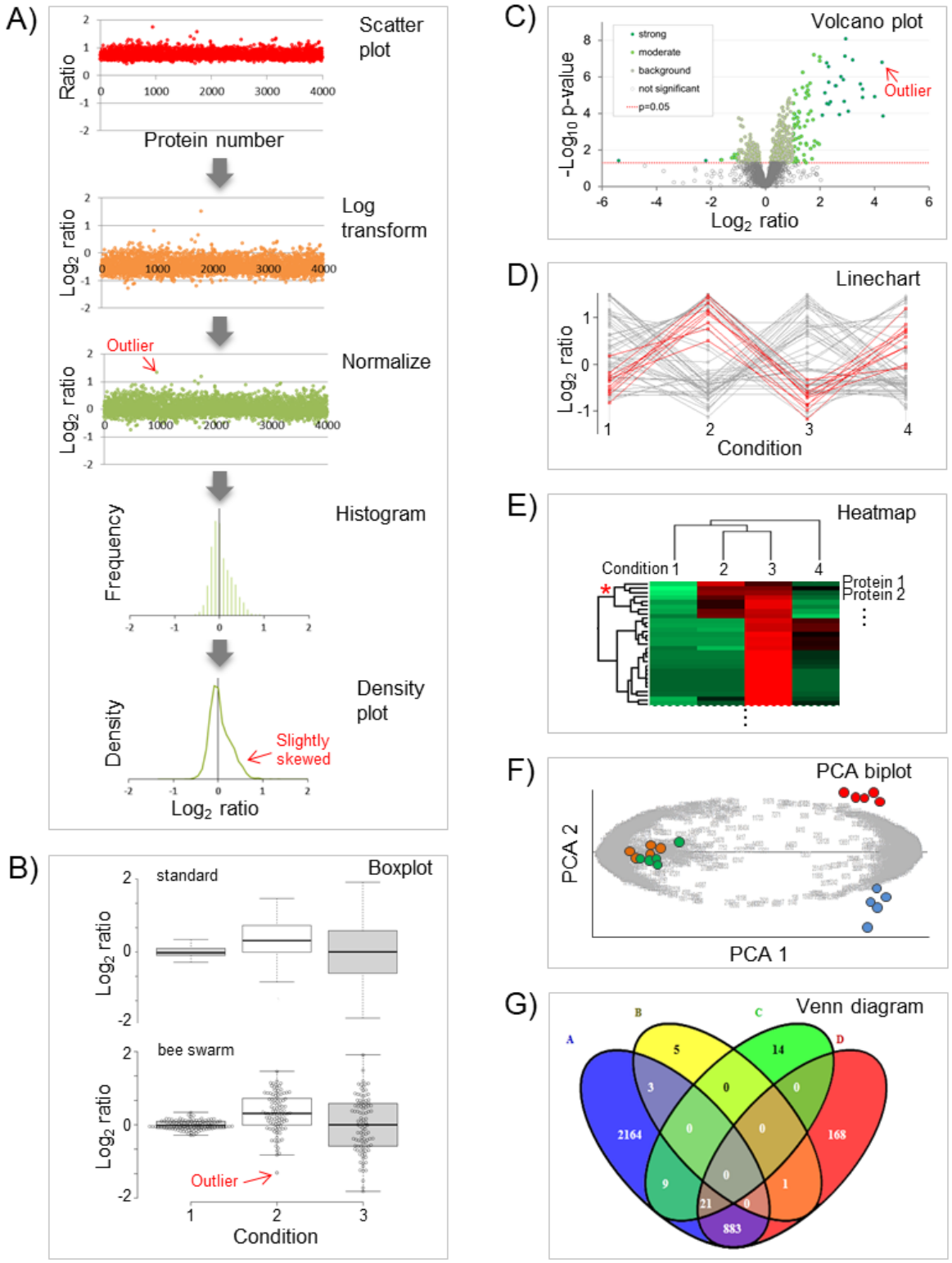

A)
Scatter plot
Ratio
Protein number
Log transform
Log2 ratio
Normalize
Outlier
Histogram
Frequency
Density plot
Density
Slightly skewed
Log2 ratio
B)
Boxplot
standard
bee swarm
Log2 ratio
Outlier
Condition
C)
Volcano plot
strong
moderate
background
not significant
p=0.05
-Log10 p-value
Outlier
Log2 ratio
D)
Linechart
Log2 ratio
Condition
E)
Heatmap
Condition 1 2 3 4
Protein 1
Protein 2
F)
PCA biplot
PCA 2
PCA 1
G)
Venn diagram
A
B
C
D
5
14
3
0
0
2164
0
0
168
9
0
1
21
0
883

Figure 5

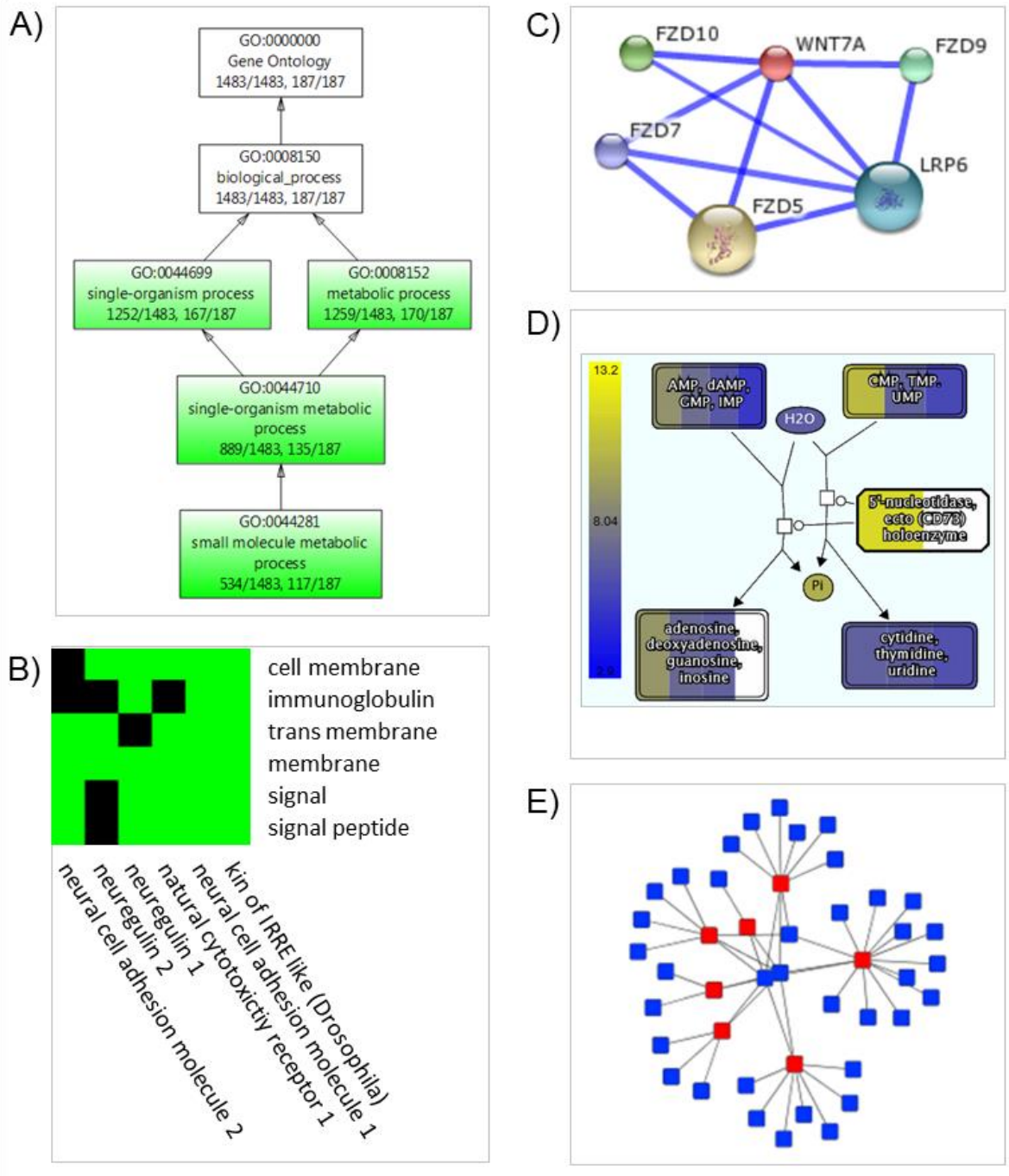
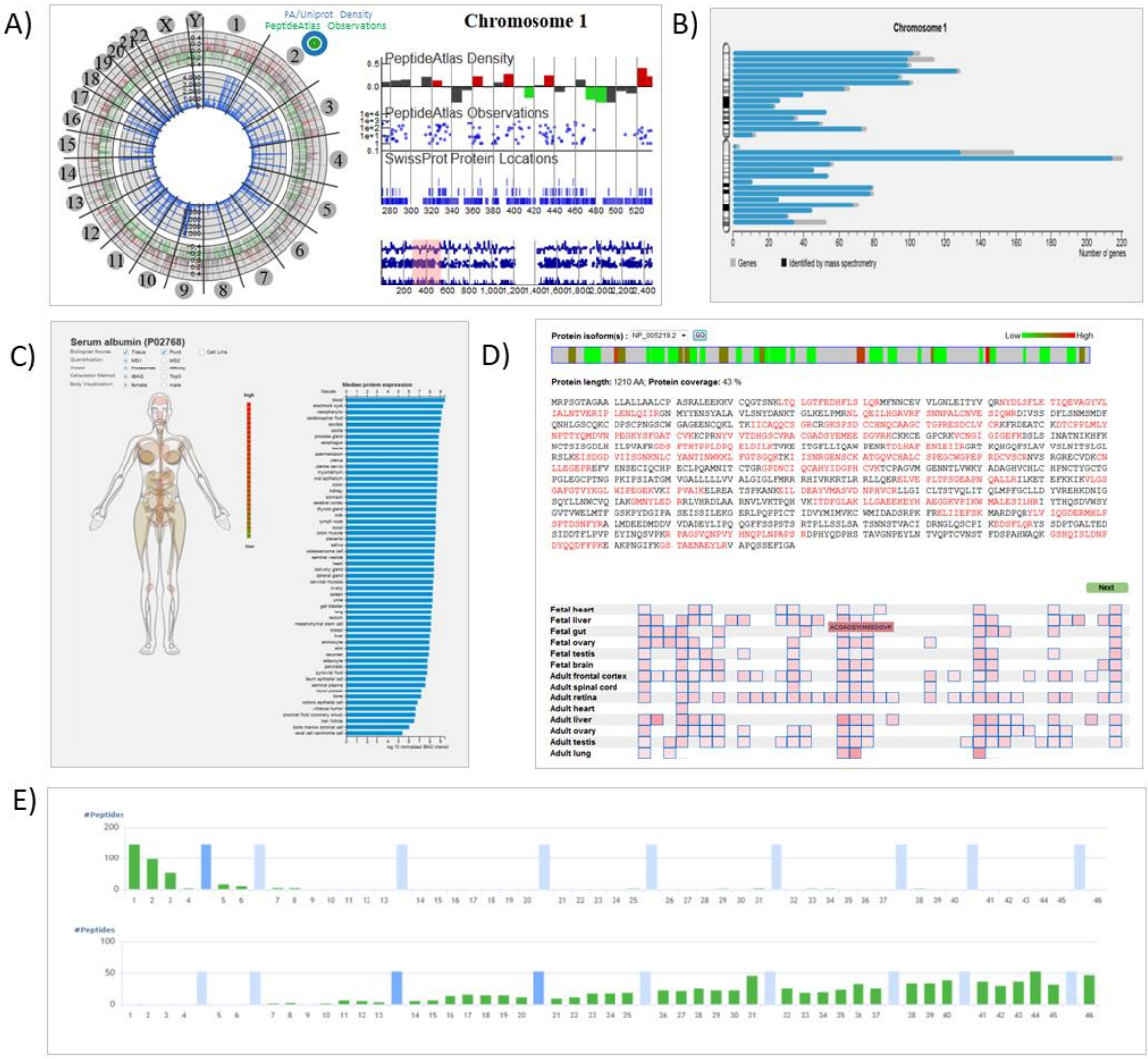

A)
GO:0000000
Gene Ontology
1483/1483, 187/187
GO:0008150
biological_process
1483/1483, 187/187
GO:0044699
single-organism process
1252/1483, 167/187
GO:0008152
metabolic process
1259/1483, 170/187
GO:0044710
single-organism metabolic process
889/1483, 135/187
GO:0044281
small molecule metabolic process
534/1483, 117/187
B)
cell membrane
immunoglobulin
trans membrane
membrane
signal
signal peptide
kin of IRRE like (Drosophila)
neural cell adhesion molecule 1
natural cytotoxicity receptor 1
neuregulin 1
neuregulin 2
neural cell adhesion molecule 2
C)
FZD10
WNT7A
FZD9
FZD7
LRP6
FZD5
D)
13.2
8.04
AMP, dAMP, CMP, IMP
H2O
CMP, TMP, UMP
5'-nucleotidase, ecto (CD73) holoenzyme
Pi
adenosine, deoxyadenosine, guanosine, inosine
cytidine, thymidine, uridine
E)

Figure 6


A)
Chromosome 1
PeptideAtlas Density
PeptideAtlas Observations
SwissProt Protein Locations
B)
Chromosome 1
Genes
Identified by mass spectrometry
Number of genes
C)
Serum albumin (P02768)
D)
Protein length: 1210 AA; Protein coverage: 43 %
Fetal heart
Fetal liver
Fetal gut
Fetal ovary
Fetal testis
Fetal brain
Adult frontal cortex
Adult spinal cord
Adult retina
Adult heart
Adult liver
Adult ovary
Adult testis
Adult lung
E)
#Peptides
#Peptides