The Perseus computational platform for comprehensive analysis of (prote)omics data

Stefka Tyanova[1], Tikira Temu[1], Pavel Sinitcyn[1], Arthur Carlson[1], Marco Y. Hein[2], Tamar Geiger[3], Matthias Mann[4] and Jürgen Cox[1*]

[1]Computational Systems Biochemistry, Max-Planck Institute of Biochemistry, Martinsried, Germany.

[2]Cellular and Molecular Pharmacology, University of California San Francisco, San Francisco, CA, USA

[3]Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel.

[4]Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Martinsried, Germany.

*Correspondence: cox@biochem.mpg.de

Running title: Perseus platform for proteomics data

Abstract

A main bottleneck in proteomics is the downstream biological analysis of highly multivariate quantitative protein abundance data. The Perseus software supports researchers in interpreting protein quantification, interaction and posttranslational modification data. It contains a comprehensive portfolio of statistical tools for high-dimensional omics data analysis covering normalization, pattern recognition, time series analysis, cross-omics comparisons and multiple hypothesis testing. A machine learning module supports classification and validation of patient groups for diagnosis and prognosis, also detecting predictive protein signatures. Central to Perseus is a user-friendly, interactive workflow environment providing complete documentation of computational methods used in a publication. All activities in Perseus are realized as plugins and users can extend the software by programming their own, which can be shared through a plugin store. Perseus combines a powerful arsenal of algorithms with intuitive usability by biomedical domain experts, making it suitable for interdisciplinary analysis of complex large datasets.

A decade ago proteomics projects were still labor-intensive and cumbersome, and high quality results required semi-manual analysis of spectra for identification and quantification. Today, mass spectrometry (MS)-based shotgun proteomics is reaching a level of maturity that makes it a powerful and broadly applicable technology for researchers in biology and biomedical sciences[1, 2]. Consistent automatic processing of spectra and the identification of peptides, proteins and posttranslational modifications (PTMs) with the help of search engines[3-7] and reliable workflows have become standard computational tasks for which satisfactory solutions exist for single studies as well as community-wide data re-analysis[8-10]. Sophisticated computational proteomics platforms offer complete solutions including the quantification of proteins and PTMs over many samples in a large variety of labeling or label-free formats[11]. Public repositories for the storage and dissemination of MS-based proteomics data exist in practical forms[12, 13]. Yeast systems biology can make use of complete proteome quantification[14] in many different conditions or stimuli with modest measurement effort[15]. Starting with a cohort of human samples protein expression matrices with sample-wise ratios or relative abundances can readily be obtained for more than 10,000 proteins[16-19].

These advances have shifted the bottleneck to the biological interpretation of quantitative abundance and PTM data and to translating the high-dimensional molecular data into relevant findings within the domain of a particular biological or medical investigator. Many potentially important findings are not currently extracted from the data simply because the computational methods and algorithms that would highlight them are not in the hand of the researcher with the necessary domain knowledge to appreciate the meaning of the findings. There are often barriers between informatics and biological researchers, which need to be bridged in order to translate omics technologies to valuable biological or medical discoveries.

Here, we address this problem by creating a computational platform that fulfils two potentially conflicting objectives: (1) All methods should be statistically sound, powerful and comprehensive. (2) It should still be intuitive and easy to use for the domain expert in a biomedical discipline who is not a computational expert. To reach these goals we developed the Perseus platform in close collaboration with biologists, with whom we analyzed projects involving multiple, diverse and distinct data types and experimental approaches. Experienced Perseus users can perform essentially all the computational tasks alone, even with little or no formal bioinformatic training. They can still involve programmers and bioinformatics specialists to extend the functionality of Perseus with plug-ins that add to the Perseus workflow as custom activities. Here we describe the functionalities available in version 1.5.4.0 of Perseus.

Comprehensive workflow-based data analysis platform

Downstream analysis of proteomic data is a multi-faceted and demanding field that integrates many aspects of bioinformatics, statistics and machine learning. It is common practice to hire bioinformaticians with a view to help the biological researchers with various analytical problems. Often these efforts result in multiple small scripts that are tedious to maintain and scale and that require the help of the developer to be re-used or stitched together. This approach is bound to turn downstream data analysis into a major bottleneck for scientific projects and discoveries. Furthermore the results may be of questionable validity when there is no clear documentation and transparency about the methods and scripts employed. We thus set out to develop the Perseus platform as a holistic software that allows continuous expansion of scalable analytical tools, their smooth integration and re-usability while providing the user with explicit documentation of the analysis steps and parameters. Greater detail on the implementation and download of Perseus is provided in **Box 1**.

Perseus offers a wide range of algorithmic activities that cover topics ranging from data normalization through exploratory multivariate data analysis to integration with other omics levels (**Fig. 1**). The following sub-sections describe the various computational and statistical tools in Perseus. Several complete analysis workflows are available on our DokuWiki pages (http://coxdocs.org/doku.php?id=perseus:user:use_cases) that contain step-by-step descriptions of three standard proteomics project types and together with the YouTube videos (https://www.youtube.com/channel/UCKYzYTm1cnmc0CFAMhxDO8w) represent a valuable resource for first time users. Many activities produce interactive graphical output for the visualization of data analysis results, which scale easily to very large sets of input data and therefore allow for thorough inspection by the user even for large-scale experiments with complex experimental designs and many measured variables. Any plot can be exported in a number of graphical formats and edited in standard vector graphics editors upon release of all clipping masks.

The central data type in Perseus is the 'augmented data matrix', which typically represents expression or abundance values of genes or proteins (rows) and biological samples or technical replicates (columns). It is supplemented by additional data containers for annotation of the rows, columns and cells of the matrix (see **Box 2**). These annotation containers are automatically filled in Perseus with gene or protein information derived from the publicly available ontologies, pathways and annotation databases. Sample annotation are used in many activities to define the study design, such as to designate which samples are replicates, or which belong to different treatments or time points in a time series analysis.

The main navigation tool is the workflow panel, which is composed of matrices and activities, and controls the information-flow in a Perseus session (**Supplementary Fig. 1**). The interactive workflow allows the user to keep track of all steps in the analysis and to navigate through data matrices and visualization components. It facilitates revisiting intermediate steps in a complex computational workflow, branching off with alternative parameter settings or a different combination of activities, and comparing results of alternative branches to each other. The matrix objects move through the workflow and are transformed and modified by activities. The workflow itself is a bipartite graph in which every matrix is connected via an activity to the next matrix. A matrix can have interactive local visualizations attached (e.g. plots, histograms and heat maps). Activities can be of a simple single-input structure or they can receive inputs from several matrices for the purpose of data integration when merging data from two or more different omics levels (see **Box 3**).

A session contains a workflow together with all intermediate results and parameter settings for all activities. Session files can be saved and reloaded and also be shared with other researchers who can load them into their Perseus instance for collaborative data analysis. Furthermore the workflow and the session serve as a complete account of the computational methods used in a project representing an accurate and reproducible description of the data analysis for documentation or publication.

Plugins

Perseus is not a static and monolithic software tool, but is instead based on a plug-in architecture that can be extended by the users (**Supplementary Fig. 2**). Perseus and its plug-ins are written in the C# programming language and adhere to a standardized application programming interface (API) that consists of a set of interfaces defining the minimum functionality that a plugin must implement. The five main interfaces in Perseus: data upload, export, processing, analysis and multi-matrix handling form the foundation of the extensible plug-in architecture (**Supplementary Fig. 2**). Plugins implementing these interfaces are visually distributed along the ribbon control menu of Perseus. Program classes that enable data matrix generation, access and export constitute the core code of the software, which is available for download from our GitHub repository (github.com/JurgenCox/perseus-plugins). Using this source code as examples and the plug-in architecture of Perseus, developers can easily expand the current functionalities by programming novel independent modules. The compiled DLL then has to be placed into the main folder of the Perseus installation which will completely integrate them, making them ready for use. A tutorial video on how to program plugins is available at (www.youtube.com/watch?v=MhS4UM1CMwU).

The API allows any user to program activity plugins in their local development environment independent of the central Perseus code repository. We provide a core set of plugins containing more than 100 activities that are bundled with the standard Perseus download and that can also be re-used in newly developed activities (**Supplementary Table 1**). For the majority of them the source code is provided via GitHub. Once users have programmed a new plugin they can make it available through the Perseus plugin store (www.perseus-framework.org/plugins). As an example, the ‘Proteomic ruler’ package combines convenient functionality for the absolute quantification of protein copy numbers per cell from generic label-free shotgun proteomics data[20].

Expression proteomics

Many proteomics projects consist of measuring cells or tissues in two or more conditions, each of them in a certain number of biological replicates, for instance using relative label-free quantification[21] or with a common labeled reference standard[22] for enhancing quantification accuracy. These kinds of proteomics data have similarities to transcriptomics microarray data and their analysis can benefit from the wealth of experience obtained in more than two decades of transcriptome data analysis by a large community. Perseus includes adaptations of some of these algorithms to proteomics workflows.

Before data can be used for the actual data analysis they need to be normalized, filtered and potentially subjected to missing value imputation for which we provide a multitude of options in the standard set of Perseus activities (**Supplementary Fig. 3**). One common task is to determine which proteins are significantly changing between conditions. Perseus adapts a particularly robust method from microarray data analysis that includes a permutation-based false discovery rate and q-value estimation[23]. This enables reliable estimation of the percentage of proteins that are mistakenly indicated as changing.

Another frequent task is to find the main clusters of expression patterns in the data and the sets of proteins responsible for the formation of these patterns. We provide a hybrid k-means-hierarchical clustering algorithm that creates interactive heat-maps and scales to matrices with a very large number of rows and/or columns in a short computing time. As an alternative to clustering, Perseus contains principal component analysis (PCA) based on singular value decomposition[24] – a form that computationally performs very well on high-dimensional data. PCA will detect the main effects in the data and the proteins driving the separation of the proteomic states.

Once an interesting cluster of proteins has been identified, enrichment analysis[25] of biological processes, complexes or pathways is done in a variety of ways, for instance with the Fisher's exact test checking for contingency between cluster membership and the property of interest. The false discovery rate (FDR) is controlled with the Benjamini-Hochberg method[26]. This elucidates what the cluster member proteins have in common and provides clues to the functional role of the cluster. Similar enrichment functionalities have been developed in the context of genomic technologies[27] and Perseus adapts these enrichment analyses specifically tailored to the purpose of proteomics. In particular, the reference space for enrichments is always appropriately chosen to be the subset of measured proteins. Furthermore, proteins which are indistinguishable based on the measured peptides are not double-counted in enrichment tests, by handling the occurrence of multiple alternative identifiers appropriately in enrichment tests.

Posttranslational modifications

Proteomics software typically generates a table for each PTM type of interest, indicating all positions on the identified proteins that are likely to be modified in at least one of the conditions of a study. In addition to scores reflecting the reliability of identification and the confidence in the localization of each site in the protein sequence[28, 29], quantitative information is crucial for understanding the functional role of the modification sites. Relative quantification in the form of site-specific ratios or intensity-based quantification is usually required for the comparison of phosphorylation in different conditions or upon different stimuli. Furthermore, analysis of the proportion of modified to total peptides, i.e. site occupancies, is important for the elucidation of major regulatory phosphorylation events during key cellular processes[30, 31].

Reformatting tools are provided in Perseus that transform the site quantification into a matrix that resembles proteome expression data, which retains information about multiple modification states of peptides. This matrix can then be analyzed with similar methods as introduced in the previous section for expression proteomics, but with some special adaptations. For example, to place phosphorylation events in the context of cellular pathways and signaling events, enrichment analysis of KEGG and Gene Ontology[32] terms can be employed. Importantly, as proteins are often characterized by multiple phosphorylation sites, care should be taken to avoid over-counting of protein-derived annotation in PTM site-based analysis ('protein-relative enrichment').

Integration of external resources is currently a tedious task that requires building access to the databases, parsing the data in the correct format and finally matching identifiers to the in-house data. In Perseus, site-specific annotation, for instance from PhosphoSitePlus[33] or sequence position specific annotation from UniProt are integrated by using an easily operated activity designed for that purpose (**Fig. 2**). This information can be used to generate statistics on which sites in the study are already known from other publications or which are novel, and to import experimentally known kinase-substrate relationships into the matrix. Alternatively, kinase motifs are matched to the sequence window surrounding the phosphorylation site, which when combined with clustering and enrichment analysis often leads to noteworthy conclusions about kinase activity patterns[34]. Reversible phosphorylation is regulated by multiple factors including increased or decreased concentration of kinases and phosphatases and the level of phosphorylation may appear to vary due to changes in the abundance of the modified protein itself. Therefore, Perseus enables straightforward overlaying of modification site and protein abundance to determine the actual quantitative changes in phosphorylation on a certain site and their likely origin.

Interaction proteomics

Affinity enrichment experiments followed by MS for determining interaction partners can nowadays be performed on a large scale involving more than a thousand bait proteins[35, 36]. This works well with intensity based relative label-free quantification[21] but also SILAC or TMT-based quantification can be used. Typically, analysis of such data requires comparison of the quantities of individual proteins in specific samples with those in a control group (**Fig. 3a**). The control may derive from cells not expressing a tagged bait protein[37], or cells in which the bait was knocked down[38]. Alternatively, all samples in which unrelated proteins served as bait can serve as negative control, which we have shown to be the superior in medium[39] and large-scale datasets[35]. Perseus allows the streamlined calculation of large numbers of tests necessary to derive a list of statistically significant outliers specific to each bait, with permutation-based FDR control for each pair of sample and controls. The resulting network of interactions can automatically be formatted to be uploaded to external tools like Cytoscape[40] for visualization (**Fig. 3b**).

For some experimental setups it is necessary to control the FDR globally instead of on the level of individual samples, for instance when interactions are measured under different conditions or over a time course[41]. To this end, Perseus offers a method to combine FDR-based cut-offs for multiple samples (**Fig. 3c**). This is an advantage over methods such as ANOVA because it retains information about the enrichment of each protein in each condition (which is lost in ANOVA), while additionally offering global-level statistics.

Time series analysis

Many biological processes are controlled by characteristic temporal changes in the concentrations of specific biomolecules. For instance, the cell cycle is accompanied by periodic changes in mRNA and protein expression[42-44]. Likewise the circadian cycle[45] involves concerted changes in abundances of proteins, their modifications, mRNAs and metabolites[46]. Perseus contains an FDR-controlled method for detecting expression behavior that is statistically significantly following a given temporal model as for instance expression with a given periodicity (**Fig. 4**). To derive the length of the cycle from the data, a Fourier-based periodicity analysis can be performed that determines the base frequency of periodic expression changes and also allows screening for possible other cycle lengths (e.g. harmonics of the base frequency). The analysis will assign an amplitude of change and a peaking time to each protein. A specialized annotation enrichment analysis designed for periodic expression changes can then determine which biological processes or pathways are switched on at which point along the time axis, detecting clusters of activity in the time dimension. Side-by-side analysis of transcriptome and proteome reveals the time lag between transcription and translation[46].

Cross-omics data analysis

Perseus has activities for comparing proteomics data to other omics dimensions, such as mRNA levels measured by RNA-seq[47]. An importer activity loads next generation sequencing (NGS) short read information as for instance obtained by the Illumina platform into a Perseus session. Reads can be aligned by standard spliced alignment workflows as, for example, provided by the TopHat[48] or STAR[49] suites and read-count based quantification is generated upon upload to Perseus. Multiple reference-genome aligned read files corresponding to data from multiple samples can be used simultaneously and a Perseus matrix will be filled with read count information per gene. The reads can represent RNA-seq or ribosome profiling data[50], which are then converted to quantitative expression profiles, for instance by calculating RPKM values[51]. To investigate the relationship between transcription and translation, this matrix can then be matched to another matrix containing protein expression values, for instance iBAQ values, which are estimates of absolute protein abundances[52, 53]. This enables correlation analysis between the two quantitative omics dimensions (**Supplementary Fig. 4**) and for this purpose we routinely use the vast amounts of freely available NGS data ready for download – e.g. from the ENSEMBL[54] (www.ensembl.org/info/data/ftp), ENA (www.ebi.ac.uk/ena) or SRA (www.ncbi.nlm.nih.gov/sra) databases – most of which are already aligned to the reference genome. Hence, this plug-in enables comprehensive analysis of multiple genomics experiments and comparison with proteomics data in a very short time.

To compare functional differences between any two ‘omics’ types, we implemented the so called ‘2D annotation enrichment’ activity[55] (**Fig. 5**), which determines annotation terms, whose members show statistically significant outlier behavior in the two dimensions chosen. Genome-wide annotation for this purpose can be membership of proteins in biochemical pathways, gene ontology terms, sub-cellular localization, protein domain content or membership in protein complexes. Processes can be simultaneously up- or down-regulated in both dimensions, or they can lack correlation, such as regulation in one dimension without any corresponding change in the other. We have found 2D enrichment analysis to be a powerful tool to probe regulation for the respective pathways or biological processes, including but not limited to information about the processes that are predominantly transcriptionally, post-transcriptionally or post-translationally regulated.

Machine learning for detecting subtle biological associations and biomarker discovery

Patients can greatly benefit from a more accurate diagnosis and a subsequently more efficient personalized treatment. Perseus combines powerful machine learning and statistical methods for the classification of proteomics samples. For example, we have applied Perseus to study clinical classification of disease subtypes from proteomic data in lymphoma[56] , prostate cancer[57] and breast cancer[58] studies. In Perseus we provide an extensible classification and regression framework that does not rely on a single 'favorite' machine learning technique (**Fig. 6**). Instead at every stage one algorithm can be exchanged for another and rated, making it possible for the non-specialist to determine the machine learning method that is best suited for the particular type of data. In addition to the many algorithms for classification, regression and feature selection that are provided together with the standard Perseus release, including a support vector machine[59] implementation, the machine learning framework is extensible allowing the users to program their own implementations of algorithms. We provide stable APIs for classification and regression models as well as for feature selection algorithms in the context of classification and regression. As an example we adapted the popular LIBSVM[60] implementation of a support vector machine as an open source classification plugin.

The machine learning section of Perseus has a cross validation structure for the purpose of measuring how the prediction performance of classification or regression will generalize to independent data that have not been used for model building, thereby avoiding notorious problems such as over-fitting[61]. The cross validation tools allow robust determination of optimal parameter values in linear or nonlinear models used for prediction. Furthermore, they help in extracting optimal protein sets from the output of a feature selection algorithm that strike a balance between good prediction performance and simplicity. This machine learning based feature selection combined with accurate monitoring of the prediction errors by cross validation offers a complement to t-test-like approaches for determining discriminating protein subsets. It detects multivariate patterns in protein expression profiles, for which the discriminatory power might not be apparent in the expression profiles of single proteins. In this way we can retrieve the members of protein response networks that are invisible to univariate feature selection methods.

Vision and future developments

Perseus integrates a large amount of bioinformatic expertise based on experience in the analysis of diverse types of large-scale proteomics data. It was developed in close collaboration with biological domain experts on the basis of real world and cutting edge life science research. The software offers an intuitive interface that enables researchers without the formal computational skills to analyze their own data, by guiding them through statistical procedures in a rigorous manner thereby equipping them with various tools for extraction of maximum information and biological insights from the data. With a view to easy uptake among diverse users, Perseus also lowers the ‘activation barrier’ by the absence of installation procedures, being completely freely available, the ability to visualize every step with intuitive and interactive plots and an automatic generation of a complete record of each analysis step and the parameters used. We believe the latter feature is crucial for the scientific community as it fosters transparency and reproducibility of the reported results. Moreover, the use of a common platform for analysis allows for unbiased comparison of the results generated in different groups and enhances the collaborations between scientists by simplifying the process of documentation and sharing of protocols. Our guiding principle was to put the expertise of bioinformatics scientists in the hands of all life science researchers, allowing them to focus on their biological questions while benefitting from both powerful statistical tools and cutting edge scalable analytic possibilities without depending on often unavailable specialists.

Through continuous development and maintenance, our goal is to establish Perseus as a comprehensive analysis and visualization tool for systems biology research, similarly as we have done previously with the MaxQuant software for the analysis of mass spectrometric data[11]. As the experimental designs become more and more complex, the functionality of Perseus will be enriched accordingly, building upon its extensible architecture to offer more tools and to support future data types. In particular a comprehensive toolset for the analysis of biological networks[62-64] resulting from co-expression or interaction studies will soon be included. For most of the development of activities in Perseus we started with proteomics data in mind, as well as their comparison to other omics dimensions. However, we have found that many of the techniques implemented in Perseus are applicable to other data types without major modifications and it has already become popular in our group for the analysis of non-proteomics data as well. In the future, metabolomics data with relative quantification profiles for a global set of metabolites over several samples, which is similar to label free quantification proteomics data, will be accommodated by Perseus with only slight adaptations such as customization of the annotation of molecular species.

One major challenge and opportunity that will drive the future development of Perseus is to bridge the currently existing gap between large-scale proteomics data generation and modeling of signaling pathways and biochemical reactions. Modeling studies are still generally performed only on low-throughput data, such as western blots or FACS data. Our goal will be to provide a more automated way to extract quantitative information from large-scale data that can directly be used as input for available modeling platforms[65-67]. Providing automatically meaningful and reliable connections to signaling pathways will also require more extensive knowledge of the behavior of PTM sites in biochemical and signaling pathways than what is currently available in public resources[68, 69].

Perseus has already been ‘battle tested’ in cutting edge proteomics research. We anticipate that it will allow researchers from many areas of life science, including fundamental biology, drug discovery and medical sciences, to increasingly participate directly in sophisticated data analysis. Our hope is that this novel platform will contribute to better communication between disciplines and more effective application of computational tools.

Box 1. Software implementation, download and maintenance

Perseus is implemented in the C# programming language from the .NET Framework 4.5 and runs natively on Windows operating systems. Perseus can be downloaded for free from www.perseus-framework.org under acceptance of our freeware license agreement and user account registration. No installation is required and the software can immediately be used upon download and decompression of the zipped folder. Detailed descriptions of the functions and their parameters are available in the online documentation of Perseus, which is linked to the download page and can also be directly accessed from within the software. Other sources of user support include the active Perseus google group (groups.google.com/forum/#!forum/perseus-list) with more than 1,400 users (May. 2016) and the YouTube videos demonstrating the use of the software (https://www.youtube.com/channel/UCKYzYTm1cnmc0CFAMhxDO8w). Several complete analysis workflows are available on our DokuWiki pages (http://coxdocs.org/doku.php?id=perseus:user:use_cases) that contain step-by-step descriptions of three standard proteomics project types and together with the YouTube videos represent a great resource for first time users. Substantial changes are usually reflected in major releases that happen once a year, however, we recommend updating the annotation files at shorter time intervals. Reproducible bugs in the latest available Perseus version can be reported via the YouTrack bug tracking system (http://maxquant.myjetbrains.com/youtrack/).

Perseus has been co-developed with MaxQuant[11] , which has become a comprehensive and widely accepted environment for the analysis of MS-based proteomics data and which contains further proteomics specific data visualization tools[70]. As a result, integration between Perseus and MaxQuant is excellent, but these environments are independent and can be used together with any upstream data analysis tool. Most of the data structures and algorithms are programmed from scratch and only few external libraries are used. An advantage of this design choice is that it gave us full control over all implementation details and helps improving performance, which can be many times better than the performance achieved in other statistical programming environments[71]. Just like MaxQuant, Perseus will be continuously maintained and developed based of secure long-term funding by the Max Planck Society for the Advancement of Science.

Box 2 Augmented data matrix

The central data format of Perseus is the data matrix, in which biological samples are represented as columns and proteins or other molecular species as rows. Perseus distinguishes several different types of columns. Upon reading new input data, the type of each column needs to be specified. In case the data comes from the MaxQuant environment[11], the suitable type of most columns of the output tables is automatically detected via the column name. The main data are stored in the 'Main columns', which typically contain the protein expression values that are to be subjected to downstream normalization, transformation, etc. and Perseus automatically selects them for statistical tests and data visualization. Other numerical values that serve as annotations such as sequence length, number of identified peptides or posterior error probabilities are stored in 'Numerical columns'. This type of data can also be explored by standard summary statistics and visualization tools, but no statistical tests, e.g. for differential expression, are applied to them. Non-numerical information can be stored as 'Text' or 'Categorical' columns. 'Text' is suitable for storing protein, RNA and gene names and identifiers and these columns are available as data labels in plots. In data integration, this kind of information is interpreted as identifiers to match rows of different matrices to each other or to an external data source. Categorical columns contain data of an enumerable type about each protein, which often signifies membership in biological processes or ontologies. This column type is used in enrichment analysis. The column type 'Multi-numerical' can contain multiple numerical values per entry. Most activities make a pre-selection of columns based on the designated type for a specific context, so it is most convenient that the column types are assigned correctly from the beginning. However, the data type can be changed retrospectively if necessary.

Several functions in Perseus rely on additional supplementing data matrices that contain meta-information about the main data matrix (**Supplementary Fig. 5**). Missing values are a common problem of large-scale data in general as some statistical methods cannot handle missing information and therefore require 'imputation' prior to the analysis[72, 73]. Perseus offers several imputation techniques including a method drawing random values from a distribution meant to simulate expression below the detection limit (**Supplementary Fig. 3**). Upon imputation a Boolean background matrix is created (**Supplementary Fig. 5a**), which keeps track of which value was measured and which was imputed. This allows visualization and filtering of imputed values during downstream analysis. Similarly, the user can generate a 'Quality matrix', which will be stored in the background as well. The 'Quality matrix' contains one corresponding value to each entry in the main data matrix and can be used to filter the main matrix (**Supplementary Fig. 5b**). For example a 'Quality matrix' can be generated from the number of peptides used in the quantification of each protein in each sample. This can be useful to mask all cases where a given protein was quantified with less than two peptides in a given sample. The phosphorylation site table is another example, in which such filtering is desirable, as sites with occupancy errors larger than a fixed threshold can be filtered out using a 'Quality matrix' containing the site-specific errors.

Data that characterizes the samples (i.e. information regarding the experimental design) is added to Perseus via row annotations. The groupings used in analysis methods such as t-test statistics and machine learning approaches are set as categorical row annotations (or numerical ones in case of continuous data, such as the time point for time series data) and are automatically recognized by the software in all suitable procedures.

Box 3. Data integration

One of the most laborious and error-prone steps in data analysis is matching and integration of different data types. Through its Multi-processing interface, Perseus offers an easy way to combine matrices and to import information from external databases. Two matrices can be matched based on any identifier that is provided as a column in each of them and the information to be transferred from one matrix to the other can be selected as well. Cases in which multiple entries from one matrix map to a single entry in the other are handled by the software in user-selectable ways, for instance for summarizing multiple numeric values from multiple rows in one matrix to a single entry in the other matrix. Furthermore, different omics data sets can easily be mapped through the pre-built genome lists that can be loaded with a single click.

Interpretation of genome-scale data often incorporates functional information such as pathways, cellular function and localization as well as structural information. In Perseus the user can upload a preprocessed set of annotations from UniProt[74] and use these in filtering and enrichment analysis of the data. Furthermore, PTM-specific annotations such as those obtainable from PhosphoSitePlus[33] and common kinase motifs can be automatically assigned by the software. Integration of user-defined curated annotations is supported in Perseus if certain simple file format requirements are met. The software can read customized annotations from tab-delimited text files, in which the first column contains the identifiers, which will be used for matching the annotations to the main matrix, and the header row contains the names of all annotations to be added. All further columns contain the customized annotations.

ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 686547 (J.C.) and from the FP7 grant agreement GA ERC-2012-SyG_318987–ToPAG (J.C.).

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

References

1. Altelaar, A.F., Munoz, J. & Heck, A.J. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature reviews. Genetics* **14**, 35-48 (2013).
2. Cox, J. & Mann, M. Quantitative, high-resolution proteomics for data-driven systems biology. *Annual review of biochemistry* **80**, 273-299 (2011).
3. Eng, J.K., McCormack, A.L. & Yates, J.R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **5**, 976-989 (1994).
4. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551-3567 (1999).
5. Geer, L.Y. et al. Open mass spectrometry search algorithm. *Journal of proteome research* **3**, 958-964 (2004).
6. Craig, R. & Beavis, R.C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics (Oxford, England)* **20**, 1466-1467 (2004).
7. Bern, M., Cai, Y. & Goldberg, D. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Analytical chemistry* **79**, 1393-1400 (2007).
8. Craig, R., Cortens, J.P. & Beavis, R.C. Open source system for analyzing, validating, and storing protein identification data. *Journal of proteome research* **3**, 1234-1242 (2004).
9. Nesvizhskii, A.I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature methods* **4**, 787-797 (2007).
10. Deutsch, E.W. et al. Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics. Clinical applications* (2015).
11. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* **26**, 1367-1372 (2008).
12. Vizcaino, J.A. et al. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic acids research* **41**, D1063-1069 (2013).
13. Vizcaino, J.A. et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature biotechnology* **32**, 223-226 (2014).
14. de Godoy, L.M. et al. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251-1254 (2008).
15. Hebert, A.S. et al. The one hour yeast proteome. *Molecular & cellular proteomics : MCP* **13**, 339-347 (2014).
16. Nagaraj, N. et al. Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology* **7**, 548 (2011).
17. Beck, M. et al. The quantitative proteome of a human cell line. *Molecular systems biology* **7**, 549 (2011).
18. Munoz, J. et al. The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. *Molecular systems biology* **7**, 550 (2011).

19. Mann, M., Kulak, N.A., Nagaraj, N. & Cox, J. The coming age of complete, accurate, and ubiquitous proteomes. *Molecular cell* **49**, 583-590 (2013).

20. Wisniewski, J.R., Hein, M.Y., Cox, J. & Mann, M. A 'proteomic ruler' for protein copy number and concentration estimation without spike-in standards. *Molecular & cellular proteomics : MCP* (2014).

21. Cox, J. et al. Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Molecular & cellular proteomics : MCP* **13**, 2513-2526 (2014).

22. Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J.R. & Mann, M. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nature methods* **7**, 383-385 (2010).

23. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116-5121 (2001).

24. Alter, O., Brown, P.O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 10101-10106 (2000).

25. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550 (2005).

26. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, 289-300 (1995).

27. Huber, W. et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods* **12**, 115-121 (2015).

28. Beausoleil, S.A., Villen, J., Gerber, S.A., Rush, J. & Gygi, S.P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature biotechnology* **24**, 1285-1292 (2006).

29. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research* **10**, 1794-1805 (2011).

30. Olsen, J.V. et al. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Science signaling* **3**, ra3 (2010).

31. Sharma, K. et al. Ultradeep human phosphoproteome reveals a distinct regulatory nature of tyr and ser/thr-based signaling. *Cell reports* **8**, 1583-1594 (2014).

32. Gene Ontology, C. Gene Ontology Consortium: going forward. *Nucleic acids research* **43**, D1049-1056 (2015).

33. Hornbeck, P.V. et al. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic acids research* **43**, D512-520 (2015).

34. Tyanova, S., Cox, J., Olsen, J., Mann, M. & Frishman, D. Phosphorylation variation during the cell cycle scales with structural propensities of proteins. *PLoS computational biology* **9**, e1002842 (2013).

35. Hein, M.Y. et al. A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell* **163**, 712-723 (2015).

36. Huttlin, E.L. et al. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425-440 (2015).

37. Hubner, N.C. et al. Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *The Journal of cell biology* **189**, 739-754 (2010).

38. Selbach, M. & Mann, M. Protein interaction screening by quantitative immunoprecipitation combined with knockdown (QUICK). *Nature methods* **3**, 981-983 (2006).

39. Keilhauer, E.C., Hein, M.Y. & Mann, M. Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). *Molecular & cellular proteomics : MCP* **14**, 120-135 (2015).

40. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003).

41. Raschle, M. et al. DNA repair. Proteomics reveals dynamic assembly of repair complexes during bypass of DNA cross-links. *Science* **348**, 1253671 (2015).

42. Spellman, P.T. et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* **9**, 3273-3297 (1998).

43. Gauthier, N.P. et al. Cyclebase.org--a comprehensive multi-organism online database of cell-cycle experiments. *Nucleic acids research* **36**, D854-859 (2008).

44. Eser, P. et al. Periodic mRNA synthesis and degradation co-operate during cell cycle gene expression. *Molecular systems biology* **10**, 717 (2014).

45. Partch, C.L., Green, C.B. & Takahashi, J.S. Molecular architecture of the mammalian circadian clock. *Trends Cell Biol* **24**, 90-99 (2014).

46. Robles, M.S., Cox, J. & Mann, M. In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism. *PLoS genetics* **10**, e1004047 (2014).

47. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* **10**, 57-63 (2009).

48. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* **14**, R36 (2013).

49. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15-21 (2013).

50. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. & Weissman, J.S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218-223 (2009).

51. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**, 621-628 (2008).

52. Schwanhäusser, B. et al. Global quantification of mammalian gene expression control. *Nature* **473**, 337-342 (2011).

53. Aviner, R., Shenoy, A., Elroy-Stein, O. & Geiger, T. Uncovering Hidden Layers of Cell Cycle Regulation through Integrative Multi-omic Analysis. *PLoS genetics* **11**, e1005554 (2015).

54. Yates, A. et al. Ensembl 2016. *Nucleic acids research* **44**, D710-716 (2016).

55. Cox, J. & Mann, M. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC bioinformatics* **13 Suppl 16**, S12 (2012).

56. Deeb, S.J. et al. Machine Learning Based Classification of Diffuse Large B-cell Lymphoma Patients by their Protein Expression Profiles. *Molecular & cellular proteomics : MCP* (2015).

57. Iglesias-Gato, D. et al. The Proteome of Primary Prostate Cancer. *Eur Urol* (2015).

58. Tyanova, S. et al. Proteomic maps of breast cancer subtypes. *Nat Commun* **7**, 10259 (2016).

59. Vapnik, V.N. The nature of statistical learning theory. (Springer, New York; 1995).

60. Chang, C. & Lin, C. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* **2**, 1-27 (2011).

61. Hastie, T., Tibshirani, R. & Friedman, J.H. The elements of statistical learning: data mining, inference, and prediction. (Springer, New York; 2001).

62. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* **4**, Article17 (2005).

63. Ideker, T. & Krogan, N.J. Differential network biology. *Molecular systems biology* **8**, 565 (2012).

64. Mutation, C. & Pathway Analysis working group of the International Cancer Genome, C. Pathway and network analysis of cancer genomes. *Nature methods* **12**, 615-621 (2015).

65. Hoops, S. et al. COPASI--a COmplex PAthway SImulator. *Bioinformatics (Oxford, England)* **22**, 3067-3074 (2006).

66. Angermann, B.R. et al. Computational modeling of cellular signaling processes embedded into dynamic spatial contexts. *Nature methods* **9**, 283-289 (2012).

67. Cowan, A.E., Moraru, II, Schaff, J.C., Slepchenko, B.M. & Loew, L.M. Spatial modeling of cell signaling networks. *Methods Cell Biol* **110**, 195-221 (2012).

68. Croft, D. et al. The Reactome pathway knowledgebase. *Nucleic acids research* **42**, D472-477 (2014).

69. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic acids research* **44**, D457-462 (2016).

70. Tyanova, S. et al. Visualization of LC-MS/MS proteomics data in MaxQuant. *Proteomics* **15**, 1453-1456 (2015).

71. Ihaka, R. & Gentleman, R. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* **5**, 299-314 (1996).

72. Troyanskaya, O. et al. Missing value estimation methods for DNA microarrays. *Bioinformatics (Oxford, England)* **17**, 520-525 (2001).

73. Liew, A.W., Law, N.F. & Yan, H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Brief Bioinform* **12**, 498-513 (2011).

74. UniProt, C. UniProt: a hub for protein information. *Nucleic acids research* **43**, D204-212 (2015).

75. Hosp, F. et al. A Double-Barrel Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS) System to Quantify 96 Interactomes per Day. *Molecular & cellular proteomics : MCP* **14**, 2030-2041 (2015).

76. Stingele, S. et al. Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. *Molecular systems biology* **8**, 608 (2012).

Key references

[3] This publication describes the earliest approach to correlate tandem mass spectra of peptides to theoretical fragment ion series calculated from in silico digests of known protein sequences with the aim of identifying peptides and proteins.

[11] Perseus has been developed in conjunction with MaxQuant which comprises a complete quantitative workflow for the analysis of shotgun proteomics data including support for a large variety of experimental techniques.

[15] In this paper the authors demonstrate that the yeast proteome can be analyzed within one hour measurement time recovering nearly all expressed cellular proteins.

[21] Here the MaxLFQ algorithm for relative label-free protein quantification is described. It enabled many researchers to conduct large proteomics studies with complex experimental designs without the need for labeling of their samples.

[23] A pioneering method is described for the robust detection of significantly changing biomolecules in large omics datasets. It uses repeated permutations of the data to determine false discovery rates.

[25] GSEA is the forerunner of many methods for analyzing molecular profiling data to determine which sets of genes or proteins are correlated with a phenotypic class distinction.

[26] In this seminal paper a simple yet powerful procedure is shown to control the false discovery rate for multiple testing of many independent hypotheses.

[52] In this publication a large scale quantitative analysis of transcription and translation rates is performed introducing the iBAQ technique for estimating protein abundances from mass spectrometry data.

Figure legends

Figure 1 | The Perseus data analysis platform. The core data structure of Perseus is the data matrix, containing samples in columns and expression values (e.g. protein, mRNA) in its cells. Additional information such as GO terms, KEGG pathways and other database sources can be added for each row entry in the form of annotation columns. Perseus incorporates data cleansing and normalization and multiple methods for exploratory analysis such as histogram charts, intensity curves, scatter plots. Classical expression omics data analysis is supported by robust statistical tools including t-tests, PCA, correlation analysis as well as enrichment analysis. Beyond the standard methods Perseus supports more complex tasks, among which are supervised learning, PTM data analysis and multiple omics data integration.

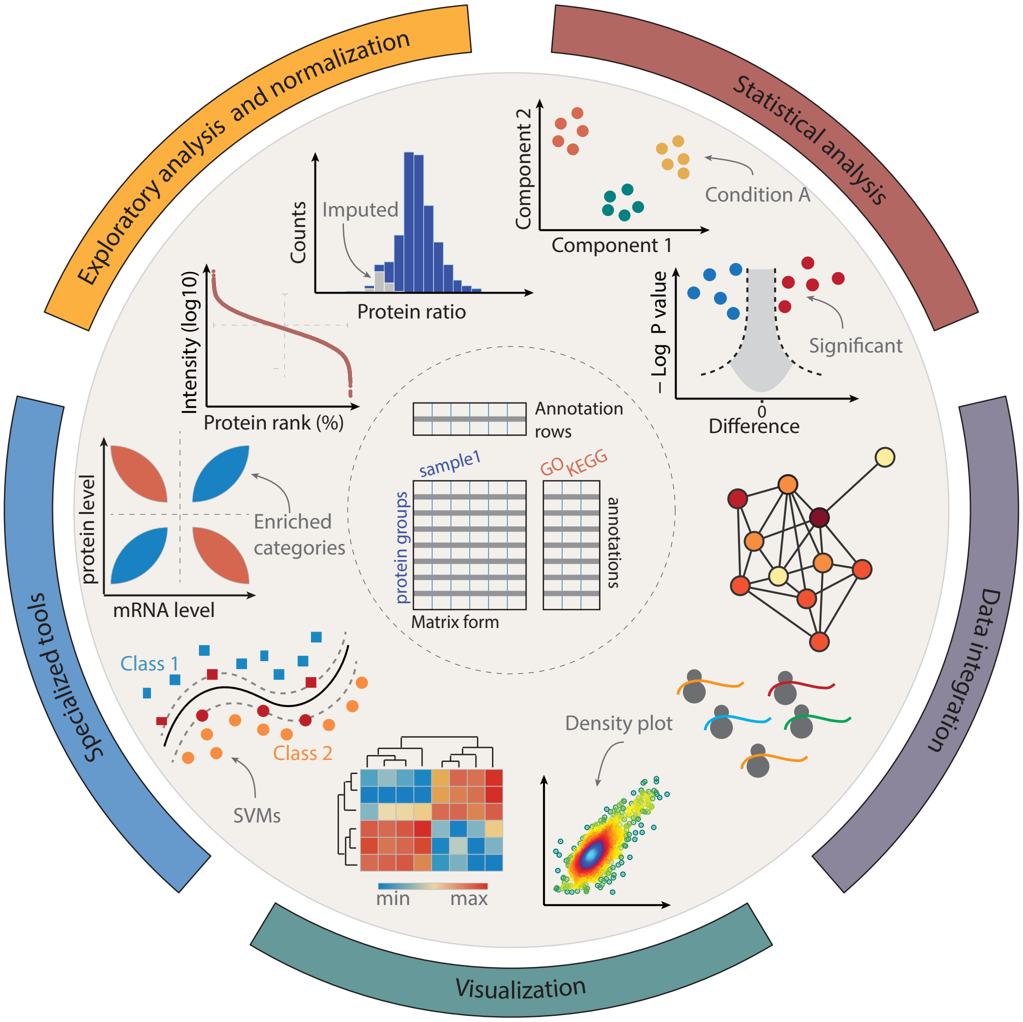
Figure 2 | Posttranslational modifications. (**a**) Annotations from various resources including UniProt and PhosphoSitePlus (PSP) can be mapped onto each phosphorylation site via the protein identifier, the modified amino acid and its position. Multiple site-specific annotations from UniProt including protein secondary structure, information if a site is known to be biologically important and domain information can be easily imported. (**b**) Estimation of the number of novel phosphorylation sites detected in an experiment as compared to already known sites stored in public repositories. (**c**) A set of short sequences surrounding a modification site can be used to generate a sequence logo and scale it by entropy in order to identify possible recognition motifs. (**d**) Comparison of the protein intensity distributions of matched total and phospho-tyrosine proteomes showing that phospho-tyrosines preferentially appear on more abundant proteins[31]).

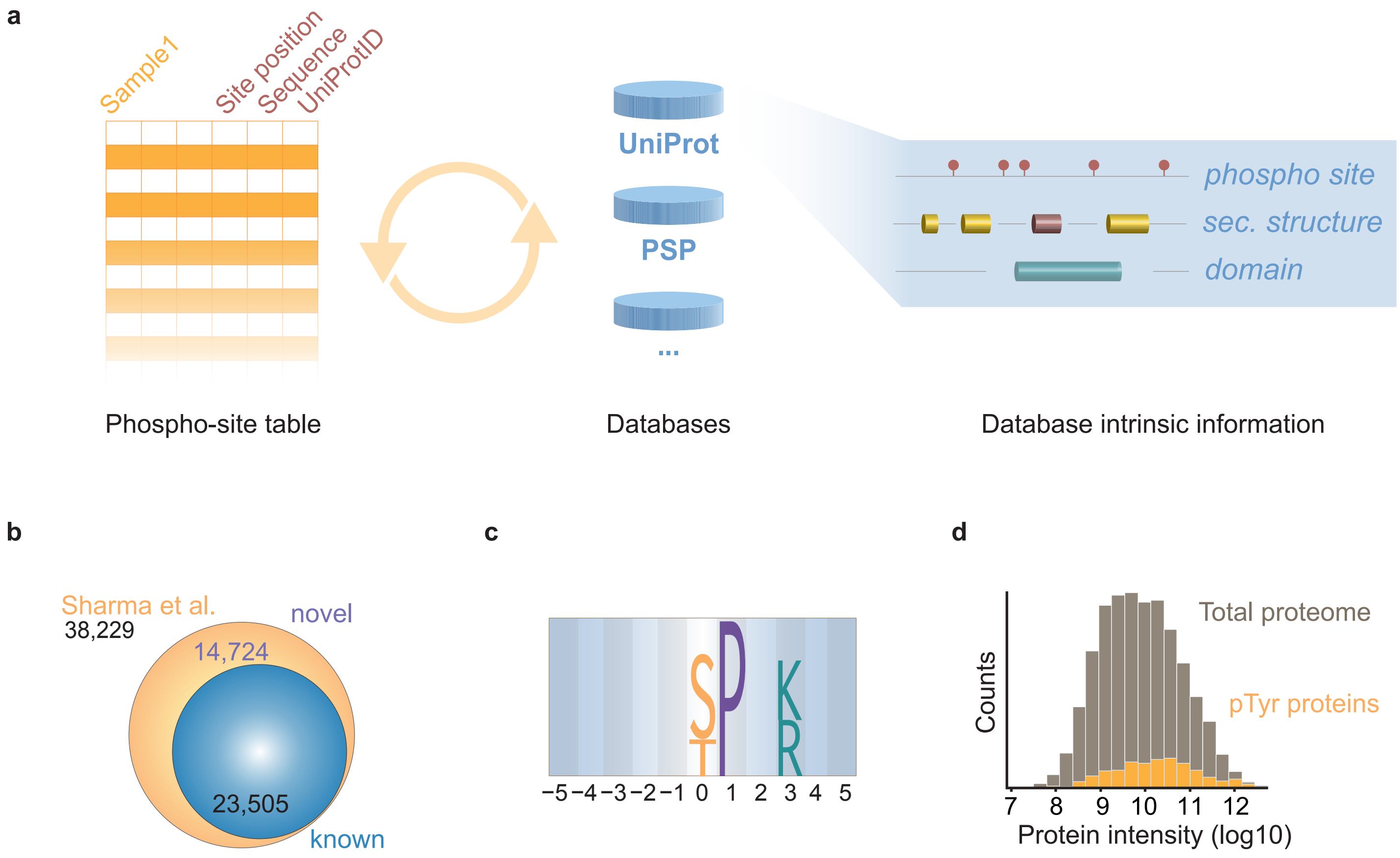
Figure 3 | Interaction proteomics. a. The interaction partners of the baits in a large set of pulldowns are determined in a multi-volcano analysis. Control groups can be defined in multiple ways: (**a**) common control group for all pulldowns (as shown), specific controls for each pulldown or the complement group of each pulldown set (i.e. the union of all other pulldowns).(**b**) Cytoscape visualization of the interaction network generated by Perseus with the affinity enrichment data from ref[75]. (**c**) Here the total set of interactors and interactions between them is determined by a global permutation-based FDR approach. For each condition a two-sample test is performed with all other conditions serving as control. The global set of interactors at a given value for the FDR is obtained by a permutation involving all conditions.

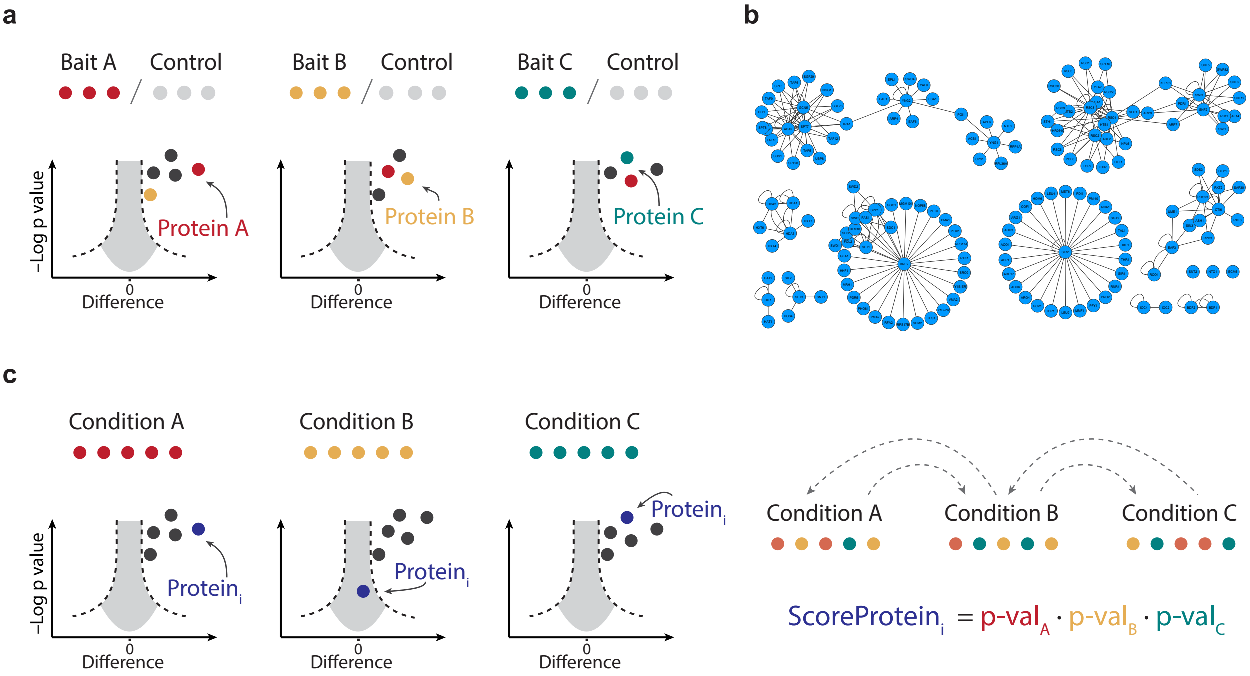
Figure 4 | Time series analysis. The time series set of plug-ins of Perseus contains a periodicity analysis component that allows detection of periodic oscillations in protein expression over time. (**a**) The amplitude (expression level) and phase (up- or down-regulation) are determined by the software by optimizing a cosine function fit to the data. A permutation-based approach, in which the time points are randomly reshuffled multiple times, identifies the statistically significantly oscillating proteins, exemplified by global circadian oscillations of the proteome in mouse liver[46]. (**b**) A total of 180 proteins were found to follow circadian rhythm over two cycles and characteristic phases of up- and down-regulation were clearly characterized as illustrated by the red and blue clusters.

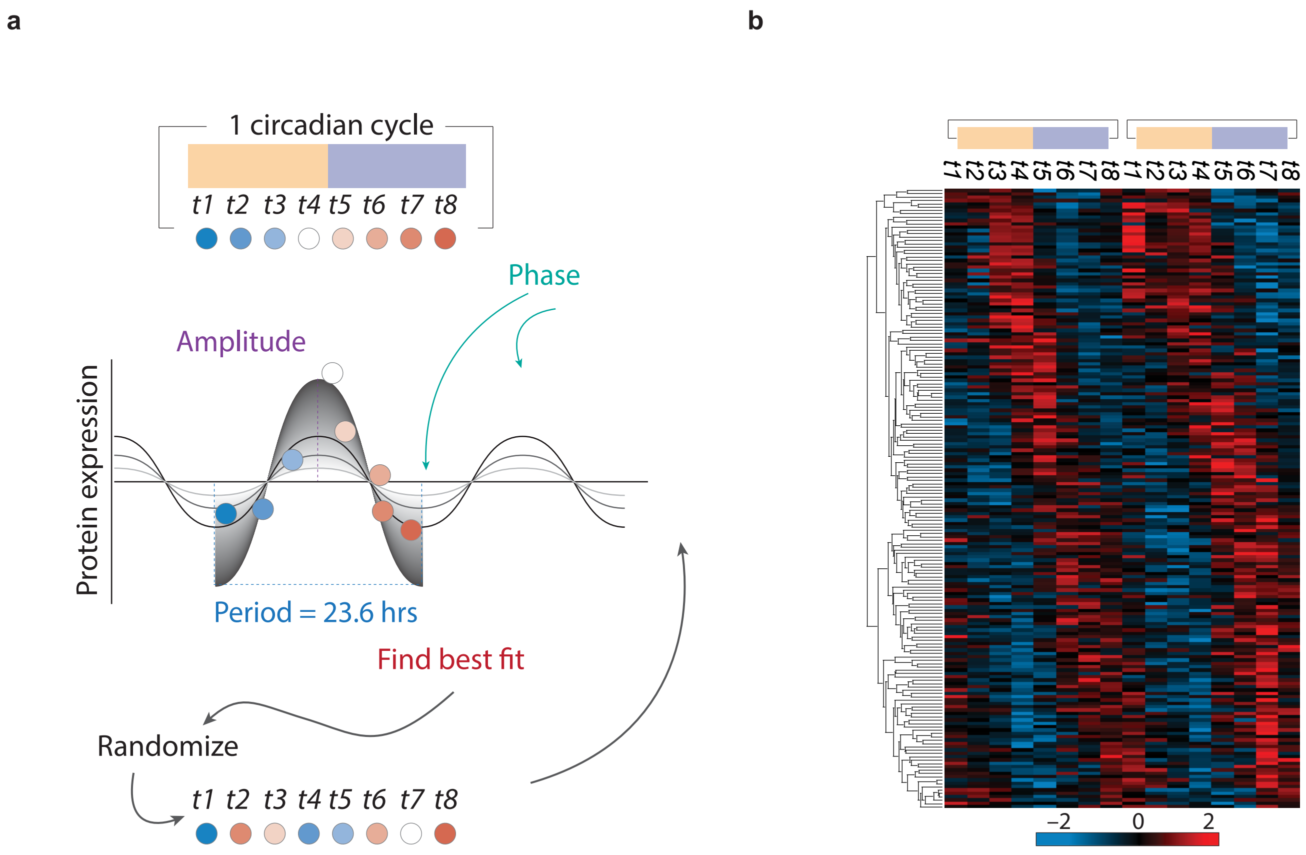
Figure 5 | Cross-omics data comparison by 2D annotation enrichment analysis. (**a**) Proteome and transcriptome expression data are joined into one Perseus matrix. Both 'omics' columns are sorted and transformed into ranks. A bivariate test is performed on each annotation term checking if the protein-mRNA pairs belonging to a certain process show a common trend, for instance if they are up-regulated in both dimensions., (**b**) The processes and locations represented by green dots show common up-regulation at both mRNA and protein levels, whereas the yellow dots indicate simultaneous down-regulation (data from ref[76]). The processes represented by purple dots exhibit up-regulation at the protein level, while the corresponding mRNA levels are collectively down-regulated.

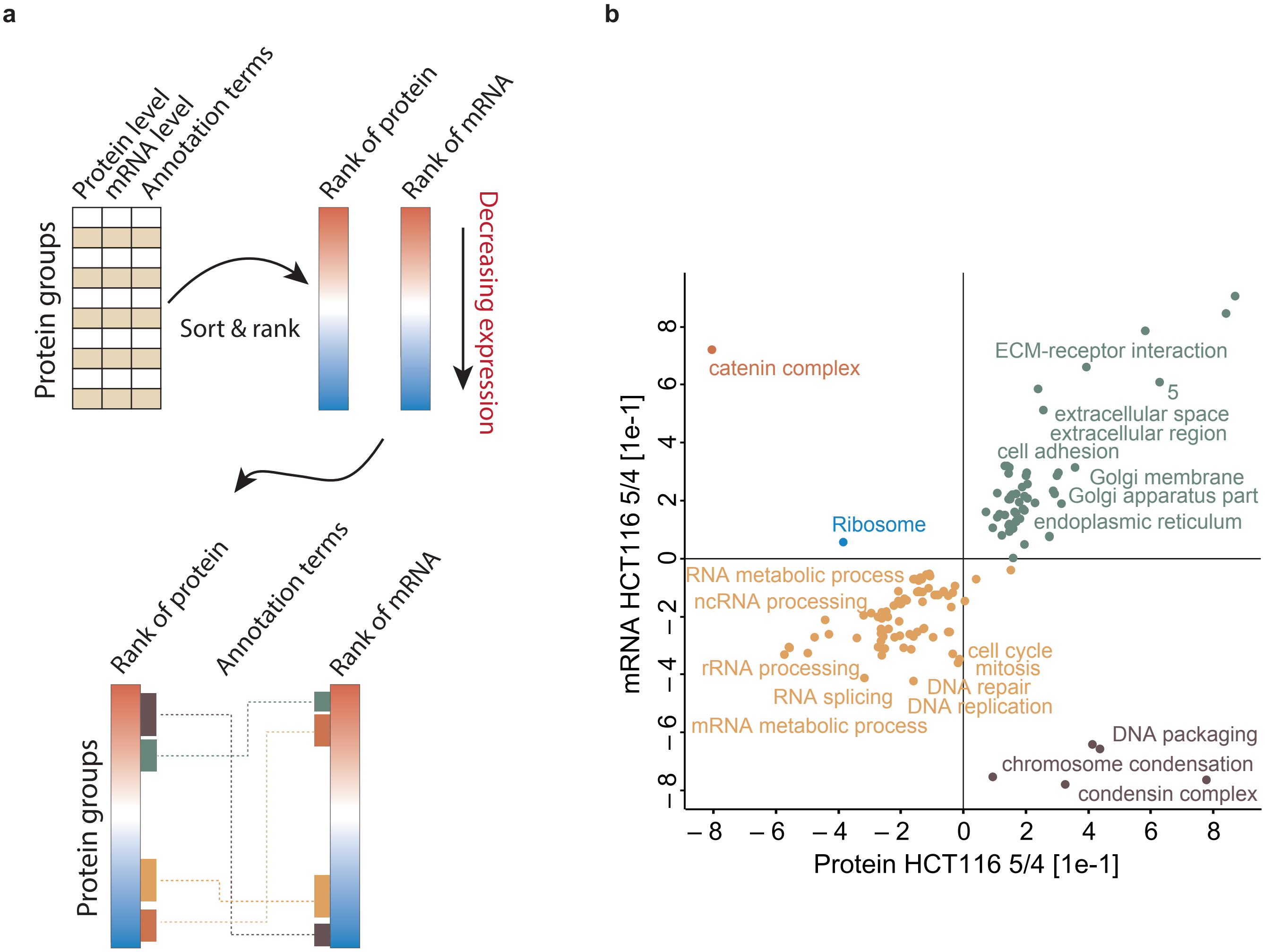
Figure 6 | Machine learning for clinical proteomics and biomarker discovery. The Learning plug-in in Perseus provides implementation of classification and regression analyses and implements various feature selection methods. Estimation of the accuracy of a trained predictor, including the feature selection step is performed in a cross-validation procedure, in which the dataset is first split into training and test subsets and the classifier is trained on the train set and its performance is then estimated on the test set. After training, the classification/regression model then assigns a predicted class to the samples of unknown class. The feature selection procedure outputs the ranks for all proteins with best ranks corresponding to the most discriminative proteins in the data. The learning module is complemented by an algorithm for screening for the optimal parameters of the different classification algorithms to maximize the classifier's performance.

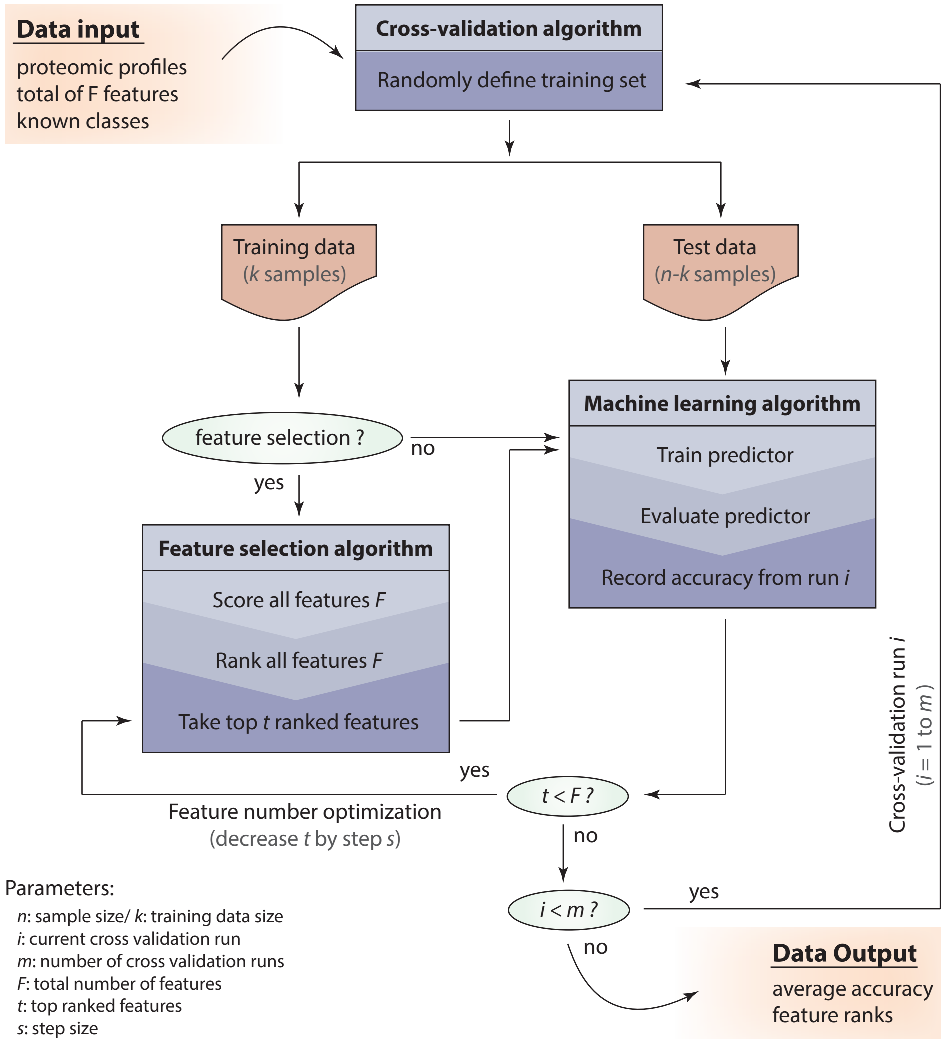

Exploratory analysis and normalization
Statistical analysis
Data integration
Visualization
Specialized tools
Counts
Imputed
Protein ratio
Component 2
Component 1
Condition A
Intensity (log10)
Protein rank (%)
− Log P value
Significant
0
Difference
Annotation
rows
sample1
GO
KEGG
protein groups
annotations
Matrix form
protein level
mRNA level
Enriched
categories
Class 1
Class 2
SVMs
Density plot
min
max


a
Sample1
Site position
Sequence
UniProtID
UniProt
PSP
...
phospho site
sec. structure
domain
Phospho-site table
Databases
Database intrinsic information
b
Sharma et al.
38,229
novel
14,724
23,505
known
c
−5 −4 −3 −2 −1 0 1 2 3 4 5
d
Total proteome
pTyr proteins
Counts
7 8 9 10 11 12
Protein intensity (log10)


a
Bait A
Control
Bait B
Control
Bait C
Control
−Log p value
Protein A
Protein B
Protein C
0
Difference
0
Difference
0
Difference
b
c
Condition A
Condition B
Condition C
−Log p value
Protein$_i$
Protein$_i$
Protein$_i$
0
Difference
0
Difference
0
Difference
Condition A
Condition B
Condition C
ScoreProtein$_i$ = p-val$_A$ · p-val$_B$ · p-val$_C$


a
b
1 circadian cycle
t1 t2 t3 t4 t5 t6 t7 t8
Phase
Amplitude
Protein expression
Period = 23.6 hrs
Find best fit
Randomize
t1 t2 t3 t4 t5 t6 t7 t8
t1 t2 t3 t4 t5 t6 t7 t8 t1 t2 t3 t4 t5 t6 t7 t8
−2
0
2


a
Protein level
mRNA level
Annotation terms
Protein groups
Sort & rank
Rank of protein
Rank of mRNA
Decreasing expression
Rank of protein
Annotation terms
Rank of mRNA
Protein groups
b
catenin complex
ECM-receptor interaction
5
extracellular space
extracellular region
cell adhesion
Golgi membrane
Golgi apparatus part
endoplasmic reticulum
Ribosome
RNA metabolic process
ncRNA processing
cell cycle
mitosis
rRNA processing
RNA splicing
DNA repair
DNA replication
mRNA metabolic process
DNA packaging
chromosome condensation
condensin complex
mRNA HCT116 5/4 [1e-1]
8
6
4
2
0
−2
−4
−6
−8
−8
−6
−4
−2
0
2
4
6
8
Protein HCT116 5/4 [1e-1]


Data input
proteomic profiles
total of F features
known classes
Cross-validation algorithm
Randomly define training set
Training data
(k samples)
Test data
(n-k samples)
feature selection ?
no
yes
Machine learning algorithm
Train predictor
Evaluate predictor
Record accuracy from run i
Feature selection algorithm
Score all features F
Rank all features F
Take top t ranked features
yes
t < F ?
no
Feature number optimization
(decrease t by step s)
Cross-validation run i
(i = 1 to m)
yes
i < m ?
no
Data Output
average accuracy
feature ranks
Parameters:
n: sample size/ k: training data size
i: current cross validation run
m: number of cross validation runs
F: total number of features
t: top ranked features
s: step size