TECHNICAL BRIEF

Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines

Mingcong Wang, Christina J. Herrmann*, Milan Simonovic, Damian Szklarczyk and Christian von Mering*

Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Switzerland

Protein quantification at proteome-wide scale is an important aim, enabling insights into fundamental cellular biology and serving to constrain experiments and theoretical models. While proteome-wide quantification is not yet fully routine, many datasets approaching proteome-wide coverage are becoming available through biophysical and MS techniques. Data of this type can be accessed via a variety of sources, including publication supplements and online data repositories. However, access to the data is still fragmentary, and comparisons across experiments and organisms are not straightforward. Here, we describe recent updates to our database resource "PaxDb" (Protein Abundances Across Organisms). PaxDb focuses on protein abundance information at proteome-wide scope, irrespective of the underlying measurement technique. Quantification data is reprocessed, unified, and quality-scored, and then integrated to build a meta-resource. PaxDb also allows evolutionary comparisons through precomputed gene orthology relations. Recently, we have expanded the scope of the database to include cell-line samples, and more systematically scan the literature for suitable datasets. We report that a significant fraction of published experiments cannot readily be accessed and/or parsed for quantitative information, requiring additional steps and efforts. The current update brings PaxDb to 414 datasets in 53 organisms, with (semi-) quantitative abundance information covering more than 300 000 proteins.

Received: September 17, 2014
Revised: December 20, 2014
Accepted: January 30, 2015

Keywords:
Absolute protein abundance / Bioinformatics / Evolution / Spectral counting

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

Data processing and data reuse in proteomics remain challenging, more so than in other fields such as transcriptomics or genomics [1, 2]. On the one hand, this is due to the sheer complexity of the proteome—where cellular proteins are expressed in a large diversity of isoforms and modifications, over a huge dynamic range, and in a variety of cellular localizations and biochemical contexts [3, 4]. On the other hand, the technical and conceptual advances in proteomics currently happen so fast that it remains a challenge to unify and critically appraise all of the data as it arrives [5–7]. Nevertheless, to achieve a deep quantitative coverage of the complete proteome is an essential milestone in the characterization of any model organism or tissue of interest, providing an important baseline for subsequent studies.

A growing number of online resources are dedicated to the processing and dissemination of proteomics data; they are operating at various degrees of postprocessing and data integration. Of these, the largest repositories of primary, raw data are those that are organized in the ProteomeXchange consortium [8]: PRIDE [9], PeptideAtlas [10], MassIVE [11], and PASSEL [10]. Building on these raw data collections as

Correspondence: Dr. Christian von Mering, Institute of Molecular Life Sciences and Swiss, Institute of Bioinformatics, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
E-mail: mering@imls.uzh.ch

Abbreviations: **FDR**, false discovery rate; **PaxDb**, Protein Abundances Across Organisms

*These authors contributed equally.
Colour Online: See the article online to view Figs. 1–3 in colour.

well as on additional curation, submission, and/or reprocessing, a number of additional resources exist. These typically offer higher levels of integration and standardization, but are sometimes also more specialized in terms of scope and coverage. They include GPMDB [12], MOPED [13], ProteomicsDB [14], MaxQB [15], and Human Proteome Map [16]. In addition, other databases whose primary focus is perhaps not exclusively on proteomics may also contain information on protein abundances, notably UniProt [17] and NextProt [18].

Here, we describe the latest updates to (and recent changes in) our protein abundance meta resource Protein Abundances Across Organisms ("PaxDb") ([19], see Fig. 1). Similar to some of the above resources, PaxDb provides access to quantitative proteomics information, but in addition it has some unique priorities and features:

(a) Its primary focus is on consistency and comparability, both between datasets as well as between organisms. This is achieved by remapping all abundance information onto the same reference space of protein sequences and genome annotations, and by providing precomputed orthology relationships that allow comparisons between organisms, at the protein family level, across the entire tree of life.
(b) PaxDb is "locus-centric": information on alternative protein isoforms or PTMs is collapsed, down to the level of the single, protein-coding gene locus. This is a conscious decision, aiming to facilitate data interpretation and user interaction, and it should be useful in all scenarios where "proteoform" resolution [3] is not required.
(c) PaxDb introduces a unique quality estimate, which applies at the level of entire datasets, as opposed to individual peptides or proteins. This metric aims to describe how well the observed spread of abundance values in a given dataset covers and delineates known functional groupings of proteins (e.g. protein complexes). The metric is called the "interaction consistency score" [19], and allows comparisons between datasets irrespective of the data source or measurement technique.
(d) When populating PaxDb, datasets are chosen and filtered manually, so as to reflect largely unperturbed, "wild-type" cells, tissues, and organisms.
(e) PaxDb is purely a meta-resource–it does not currently accept user submissions. All its data are imported from primary proteomics databases or from publication supplements; the original search parameters, false discovery rates, and other technical settings are left unchanged.
(f) For each organism or tissue that has already been addressed by multiple available experiments/datasets, PaxDb conducts a weighted averaging to produce an integrated "best-estimate" dataset guided by the above quality estimates [19].
(g) Lastly, PaxDb presents its information in an intuitive and simple web interface, which is enriched with accessory information regarding the annotation, structure, and interaction partners of the various proteins.

2 Data updates

The update process of PaxDb is partly manual, partly automatic, and it occurs on a time-scale of roughly once or twice a year. The growth of datasets and the number of organisms so far is tabulated in Fig. 2A. Care is taken not to exclude nonstandard datasets such as those based on biophysical or single-cell measurements; however, datasets are generally included only if they represent a mostly unperturbed, "normal" and physiological state of cells. Tissues and cell-lines are annotated with controlled vocabularies; in the case of tissues we use the Uberon ontology [20], which natively allows cross-species comparisons of homologous tissues/organs.

For the current update to version 4.0, we started with a manual search for publications describing possible datasets of interest. This included keyword searches and forward citation analysis of landmark papers, but we also systematically scanned all publications in three pertinent journals (MCP, J Prot Res, Proteomics), as well as all publication output of six major labs operating in high-throughput proteomics. The initial results were filtered down to 37 candidate publications, based on the following criteria: (i) studies should be published after August 2012 and not yet be contained in PaxDb, (ii) coverage should be at least 20% of the predicted proteome, or at least 20 000 peptide-spectrum matches in case of MS data, (iii) abundance values must cover at least three orders of magnitude, (iv) there should be no biased subfractionation (e.g. restricted to organelles, compartments, or specific modifications), (v) datasets must be parsable for absolute protein quantification data; this excludes purely relative quantifications, and (vi) datasets must address mainly unperturbed samples in normal, physiological state; this excludes mutated, stressed, or diseases samples.

The same set of criteria were then applied to filter recent datasets stored in three large proteomics data repositories: PRIDE [9], GPMDB [12], and PeptideAtlas [10]. In the case of PRIDE, datasets were accessed via the PRIDE BioMart if "complete" submissions were available, and via the pepXML format in case of selected "partial" submissions. In the case of GPMDB, a recent new feature of the website that allows the aggregated access to peptide information for each taxon/organism, was used. From PeptideAtlas, data were imported via the so-called "builds." GPMDB and PeptideAtlas are convenient data sources, but since the peptides cannot easily be traced back to the original experiments/publications, these data collections have to be taken as they are. Thus they may include some nonphysiological, subfractionated, or mutated samples—although in the case of PeptideAtlas, builds that hinted at this already in their annotation were blocked entirely.

Our data import procedure encountered many proteomics experiments multiple times (see Fig. 2B). Overall, redundancy was avoided by importing a given experiment via the most convenient access route (the two recent, large mapping efforts of the human proteome, for example, were imported via PRIDE).

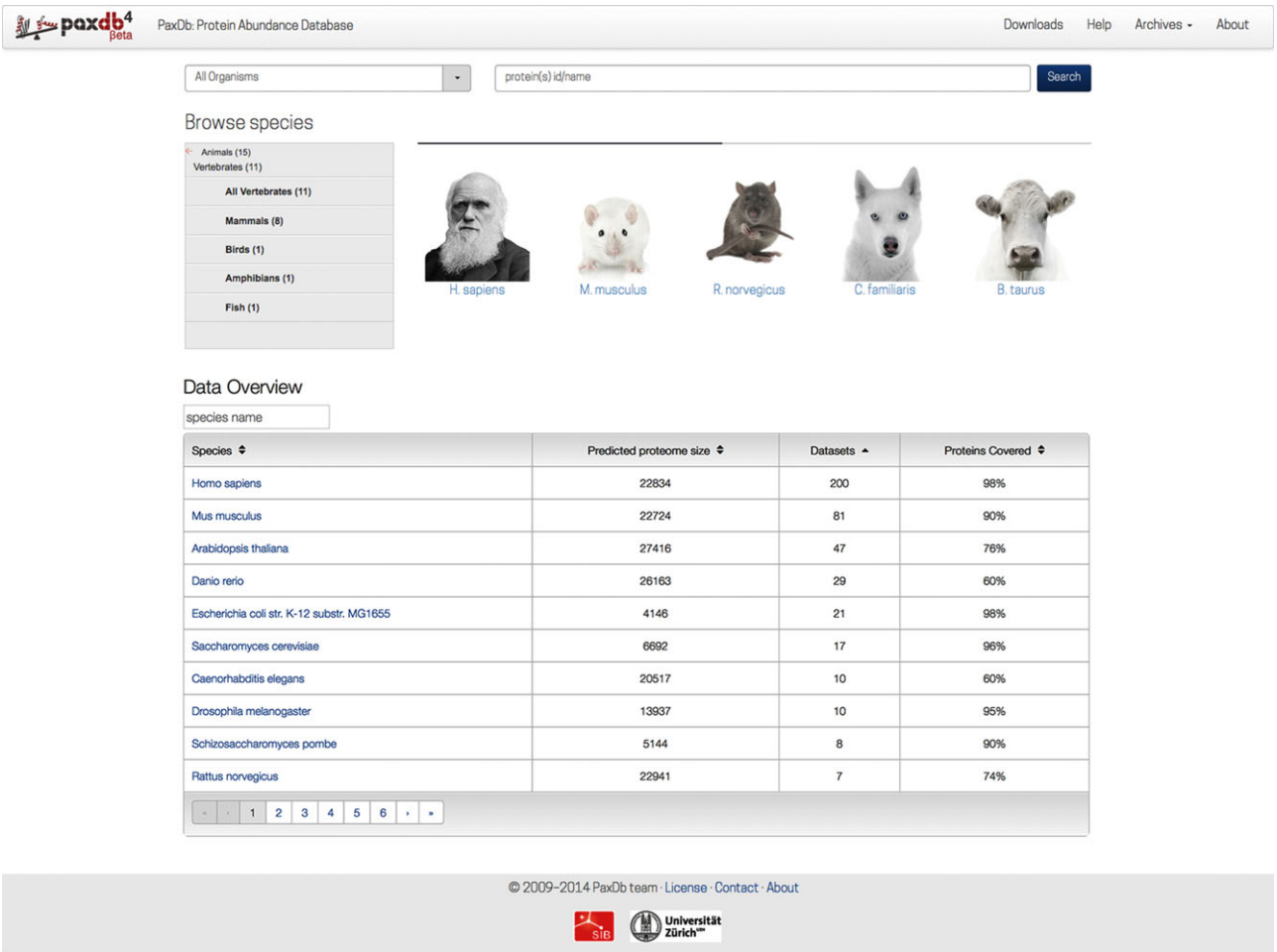

Figure 1. The PaxDb website. Screenshot of the entry page of the PaxDb website, at http://paxdb-org/. Model organisms can be browsed via the navigable taxonomy to the left; proteins of interest are accessible directly as well, via a variety of identifiers and full-text searches (input box at top of the page).

Finally, a new development with release 4.0 of PaxDb is the inclusion of protein abundance information from cell-line samples. Cell-lines are unique in that they cannot be considered to be fully physiological and unperturbed, so their inclusion in PaxDb is somewhat of an exception to our normal import rules. However, cell-lines represent a significant fraction of the available data and their proteome expression status is of great interest for everyday lab work. Hence, datasets for 35 different cell-lines (including human induced pluripotent cells, as well as human and mouse embryonic stem cells), were included in this update; their biological origin is annotated using the "Cellosaurus" controlled vocabulary.

A

release	1.0	2.1	3.0	4.0
organisms	10	12	30	53
publications	10	31	55	90
samples	28	99	187	414
spectral counts	16 mio	41 mio	135 mio	2030 mio
proteins covered	70'535	85'452	188'183	> 300'000

B

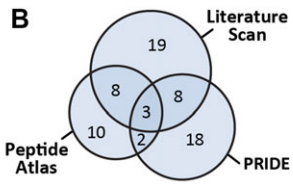

Figure 2. Data updates for version 4.0. (A) Growth of datasets and organisms covered by PaxDb. Note that PaxDb focuses mainly on normal, unperturbed, physiological cells, and tissues—some prominent, published datasets are thus not included. (B) Sources for the new data in PaxDB version 4.0 (i.e. not already contained in version 3.0; the counts in the Venn diagram represent "publications" as opposed to "samples" or "replicates"). Only published data, according to the references stated at the respective sources, is included. Panel (B) focuses on data availability—datasets found at multiple sources are imported only once, from the most convenient/applicable source.

Cell-lines are available for browsing and searching, but typically are not selected to contribute to the "best-estimate" integrated datasets in PaxDb.

3 Rescaling and quality scoring

As introduced and described previously [19], datasets in PaxDb are rescaled to a common abundance metric ("parts per million"), and also ranked via a universally applicable, albeit somewhat indirect quality score. For the rescaling, the datasets are first parsed or processed such that the data reflect proportional abundances of whole protein molecules (i.e. proportionality to counts of complete, individual protein molecules, not to molecular weights, protein volumes, or digested peptides). In the case of spectral counting data, this is done via an in-house pipeline that takes into account protein sizes and estimated relative detectabilities of peptides [19, 21]. For other datasets, the procedures depend on the type of data and the type of quantitative information that is provided (datasets that cannot be converted to proportional abundances of entire protein molecules are discarded). Then, the proportional abundances are rescaled linearly to add up to one million; this means the abundance of each protein of interest is finally expressed in "parts per million," relative to all other proteins in a sample. While this metric cannot be directly converted to "molecules per cell," it has the advantage of being comparable/meaningful across cells of different volumes, or across tissues of different cellular and extracellular compositions.

For the quality scoring, we identified a test that can be applied to any organism and to any abundance dataset, albeit at the cost of providing an indirect quality estimate only [19]. This test relies on the assumption that proteins which are interacting physically in a protein complex, or functionally in a pathway or metabolic process, should have a tendency to be expressed at similar abundance levels. This is merely a tendency, of course, and numerous exceptions to this assumption exist. However, globally, the abundance ratios of functionally interacting proteins are clearly closer to one than those of randomly chosen protein pairs [19], and this signal can be used to provide a relative ranking between datasets, given a constant and externally provided network of functional links between proteins. To compute this score, we first import protein–protein interaction information from the STRING database [22], separately for each organism in PaxDb. For a given protein abundance dataset, we then compute the absolute log abundance ratios of all pairs of proteins annotated to be functionally linked. The median of these absolute log abundance ratios represents an indirect quality metric: the closer it is to zero, the better (i.e. the more there is consistency between abundance values and functional annotations such as protein complexes or pathways). We then compute a background expectation for this metric, by permuting the abundance values in a given dataset randomly, and recomputing the median log abundance ratios. The permutation is repeated several times, yielding a distribution of medians. The actually observed median is then expressed as a Z-score distance to the random distribution of medians—this distance is termed the "interaction consistency score."

4 False discovery rates

Since PaxDb does not reprocess raw MS data, and since it does not reexecute peptide-spectrum matching searches, the search parameter settings and false discovery rates (FDR) of the original submitters are always retained. However, there is controversy as to what extent false discovery rates represent a problem, especially as they propagate through larger integrated data collections [23]. To reestimate FDRs in an independent way, Ezkurdia et al. proposed to focus on a set of proteins that should not be expected to be observed in the vast majority of human tissues, namely human olfactory receptors [23]. Because there are several hundred of these receptors encoded in the human genome, they do represent a broad and universal test set of likely "false-positive" protein identifications (except, of course, for samples originating in nasal tissues and perhaps some other, inherently "leaky" tissues [24]). We have implemented this test on all human abundance datasets in PaxDb 4.0, and indeed observe variable levels of inferred FDR across datasets (Supporting Information Table 1). This led us to block a small number of datasets from further inclusion in PaxDb, and the remainder of the data usually exhibit estimated FDRs of 5% or better, many even at 1% or better. Despite reasonably low FDRs throughout, false discovery identifications generally remain a pressing problem, since they will disproportionally affect the abundance estimates of lowly expressed proteins.

5 Stoichiometries and abundances on the tree of life

One of the unique features of PaxDb is its seamless comparability of data across organisms. This allows insights in protein abundance evolution, such as abundance conservation in the eukaryotic core proteome [21, 25, 26], cost-diversity tradeoffs during evolution [27], or the fate of paralogs during evolutionary network rewiring [28]. Orthology relationships in PaxDb are precomputed, through the *eggNOG* mechanism [29], and can be browsed at various levels of phylogenetic depth (e.g. "mammals," "animals," or "eukaryotes"). To illustrate the usefulness of these comparisons, Fig. 3 shows abundance comparisons at various levels of organismal relatedness. At one end of the spectrum, closely related organisms such as mouse and human show a relatively high level of abundance conservation, with more than 3400 orthologous proteins observed in both, at an overall abundance correlation of 0.7 (Fig. 3A). At the other end of the spectrum, distant comparisons across the root of the tree of life (e.g. Bacteria versus Eukaryotes) reveal a universal core of

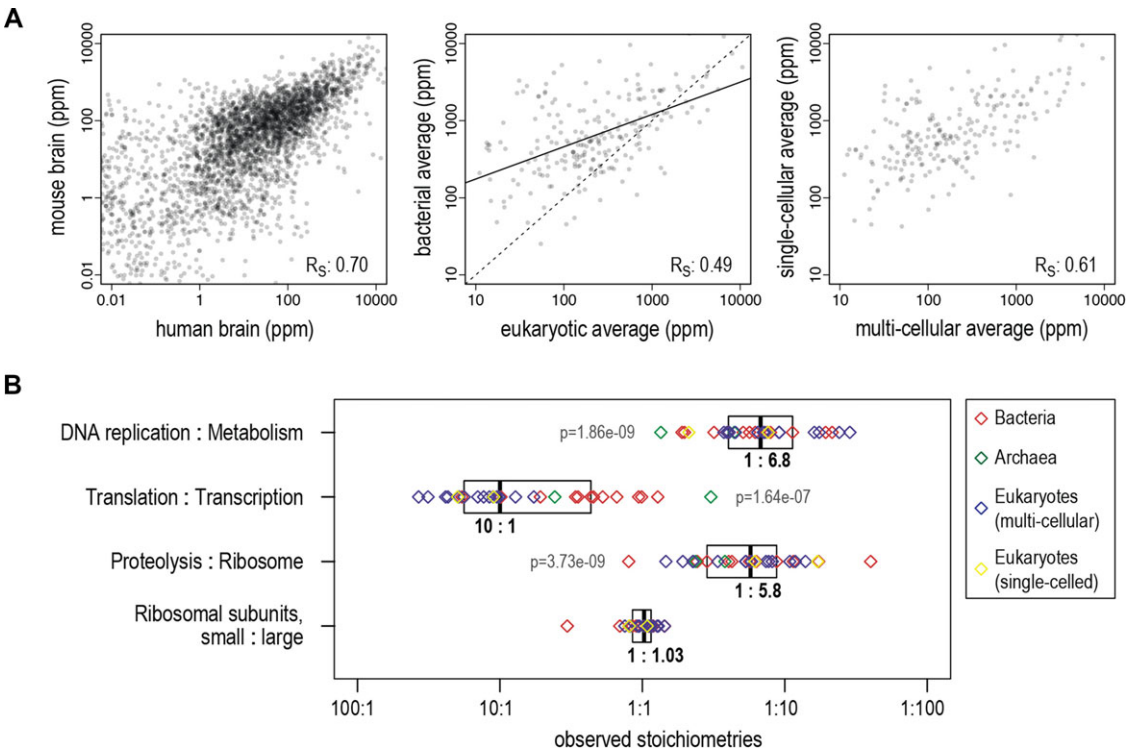

Figure 3. Abundance conservation and stoichiometry (A) Abundance correlations among orthologous proteins, across different evolutionary distances. (B) Inferred stoichiometry ratios between functionally related cellular processes, across multiple datasets, and organisms. In this plot, each data point denotes one organism; the abundance averages were taken from the integrated datasets (where available). Boxplots denote the medians, as well as the 25 and 75% percentiles, respectively. Below each boxplot, the median is also indicated textually. With the exception of the two ribosomal subunits, all stoichiometries are significantly different from 1:1 (*p*-values are indicated). All data in this figure are from version 3.0 of PaxDb.

the proteome—mostly involved in information processing, but still with an abundance correlation approaching 0.5 overall.

When combined with information on protein complexes or pathways, the view across multiple organisms might unravel general pathway-stoichiometries and scaling laws in proteome composition. With individual proteins and pathways, the measurement noise is likely still too large to allow many meaningful conclusions [30], but integration across datasets and organisms may allow to constrain global stoichiometries and min/max levels of regulation. This is explored in Fig. 3B: it shows the relative abundance ratios between functionally connected protein complexes (or processes), such as between the two subunits of the ribosome, between the ribosome and the proteasome, or between translation and transcription. Since multiple organisms, datasets, and proteins contribute data points to this plot, the ratios are statistically meaningful and reveal the expected differences of scale. Strikingly, the final abundance ratio of the two subunits of the ribosome comes down to 1:1.03 (Fig. 3B), which is very close to the theoretical expectation, and illustrates the quantitative power of data aggregation. Another notable observation concerns the stoichiometry between the core machineries of translation and transcription. Overall, this ratio is about 10:1, but it is significantly higher in eukaryotes than in prokaryotes ($p < 2e\text{-}05$; Fig. 3B), perhaps owing to larger cell-sizes and slower growth.

These and similar studies represent some of the use cases that PaxDb was designed for, but many other usage scenarios will undoubtedly surface with each new data release.

The author has declared no conflict of interest.

6 References

[1] Perez-Riverol, Y., Alpi, E., Wang, R., Hermjakob, H., et al., Making proteomics data accessible and reusable: Current state of proteomics databases and repositories. *Proteomics* 2015, *15*, 930–949.

[2] Martens, L., Bioinformatics challenges in mass spectrometry-driven proteomics. *Methods Mol. Biol.* 2011, *753*, 359–371.

[3] Smith, L. M., Kelleher, N. L., Linial, M., Goodlett, D., et al., Proteoform: a single term describing protein complexity. *Nat. Methods* 2013, *10*(3), 186–187.

[4] Zubarev, R. A., The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics* 2013, *13*, 723–726.

[5] Breker, M., Schuldiner, M., The emergence of proteome-wide technologies: systematic analysis of proteins comes of age. *Nat. Rev. Mol. Cell. Biol.* 2014, *15*, 453–464.

[6] Altelaar, A. F., Munoz, J., Heck, A. J., Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* 2013, *14*, 35–48.

[7] Mann, M., Kulak, N. A., Nagaraj, N., Cox, J., The coming age of complete, accurate, and ubiquitous proteomes. *Mol. Cell* 2013, *49*, 583–590.

[8] Vizcaino, J. A., Deutsch, E. W., Wang, R., Csordas, A., et al., ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* 2014, *32*, 223–226.

[9] Vizcaino, J. A., Cote, R. G., Csordas, A., Dianes, J. A., et al., The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* 2013, *41*, D1063–D1069.

[10] Kusebauch, U., Deutsch, E. W., Campbell, D. S., Sun, Z., et al., Using PeptideAtlas, SRMAtlas, and PASSEL: comprehensive resources for discovery and targeted proteomics. *Curr. Protoc. Bioinformatics* 2014, *46*, 13251–132528.

[11] Mass Spectrometry Interactive Virtual Environment. Available from: http://massive.ucsd.edu.

[12] 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.

[13] Montague, E., Stanberry, L., Higdon, R., Janko, I., et al., MOPED 2.5—an integrated multi-omics resource: multi-omics profiling expression database now includes transcriptomics data. *OMICS* 2014, *18*, 335–343.

[14] Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., et al., Mass-spectrometry-based draft of the human proteome. *Nature* 2014, *509*, 582–587.

[15] Schaab, C., Geiger, T., Stoehr, G., Cox, J., et al., Analysis of high accuracy, quantitative proteomics data in the MaxQB database. *Mol. Cell Proteomics* 2012, *11*, M111 014068.

[16] 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.

[17] UniProt, C., Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 2014, *42*, D191–D198.

[18] Gaudet, P., Argoud-Puy, G., Cusin, I., Duek, P., et al., neXtProt: organizing protein knowledge in the context of human proteome projects. *J. Proteome Res.* 2013, *12*, 293–298.

[19] Wang, M., Weiss, M., Simonovic, M., Haertinger, G., et al., PaxDb, a database of protein abundance averages across all three domains of life. *Mol. Cell Proteomics* 2012, *11*, 492–500.

[20] Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E., et al., Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* 2012, *13*, R5.

[21] Weiss, M., Schrimpf, S., Hengartner, M. O., Lercher, M. J., et al., Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome. *Proteomics* 2010, *10*, 1297–1306.

[22] Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., et al., STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 2014, *43*, D447–D452.

[23] Ezkurdia, I., Vazquez, J., Valencia, A., Tress, M., Analyzing the first drafts of the human proteome. *J. Proteome Res.* 2014, *13*, 3854–3855.

[24] Feldmesser, E., Olender, T., Khen, M., Yanai, I., et al., Widespread ectopic expression of olfactory receptor genes. *BMC Genomics* 2006, *7*, 121.

[25] Laurent, J. M., Vogel, C., Kwon, T., Craig, S. A., et al., Protein abundances are more conserved than mRNA abundances across diverse taxa. *Proteomics* 2010, *10*, 4209–4212.

[26] Schrimpf, S. P., Weiss, M., Reiter, L., Ahrens, C. H., et al., Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biol.* 2009, *7*, e48.

[27] Krick, T., Verstraete, N., Alonso, L. G., Shub, D. A., et al., Amino acid metabolism conflicts with protein diversity. *Mol. Biol. Evol.* 2014, *31*, 2905–2912.

[28] Gagnon-Arsenault, I., Blanchet, F. C. M., Rochette, S., Diss, G., et al., Transcriptional divergence plays a role in the rewiring of protein interaction networks after gene duplication. *J. Proteomics* 2013, *81*, 112–125.

[29] Powell, S., Forslund, K., Szklarzcyk, D., Trachana, K., et al., eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.* 2013, *42*, D231–D239.

[30] Matalon, O., Horovitz, A., Levy, E. D., Different subunits belonging to the same protein complex often exhibit discordant expression levels and evolutionary properties. *Curr. Opin. Struct. Biol.* 2014, *26C*, 113–120.

