Research Article

Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome

Manuel Weiss[1,2,3], *Sabine Schrimpf*[1], *Michael O. Hengartner*[1], *Martin J. Lercher*[4] *and Christian von Mering*[1,2]

[1] Institute of Molecular Life Sciences, University of Zurich, Switzerland
[2] Swiss Institute of Bioinformatics, University of Zurich, Switzerland
[3] PhD program in Molecular Life Sciences, University of Zurich and ETH Zurich, Switzerland
[4] Department of Computer Science, Heinrich-Heine-University Düsseldorf, Germany

Received: June 12, 2009
Revised: December 2, 2009
Accepted: December 14, 2009

Genome-wide, absolute quantification of expressed proteins is not yet within reach for most eukaryotes. However, large numbers of MS-based protein identifications have been deposited in databases, together with information on the observation frequencies of each peptide spectrum (“spectral counts”). We have conducted a meta-analysis using several million peptide observations from five model eukaryotes, establishing a consistent, semi-quantitative analysis pipeline. By inferring and comparing protein abundances across orthologs, we observe: (i) the accuracy of spectral counting predictions increases with sampling depth and can rival that of direct biochemical measurements, (ii) the quantitative makeup of the consistently observed core proteome in eukaryotes is remarkably stable, with abundance correlations exceeding $R_S = 0.7$ at an evolutionary distance greater than 1000 million years, and (iii) some groups of proteins are more constrained than others. We argue that our observations reveal stabilizing selection: central parts of the eukaryotic proteome appear to be expressed at well-balanced, near-optimal abundance levels. This is consistent with our further observations that essential proteins show lower abundance variations than non-essential proteins, and that gene families that tend to undergo gene duplications are less well constrained than families that keep a single-copy status.

Keywords:
Bioinformatics / Evolution / Expression noise / Protein abundance / Shotgun proteomics / Spectral counting

1 Introduction

Most proteins in an organism are tightly regulated in their activity – through spatial and temporal control of their production, compartmentalization in the cell, assembly into protein complexes, post-translational modifications, and/or regulated degradation. Given all these levels of regulation, the molecular abundance of a protein in the cell is only one of several factors controlling its activity, and perhaps not the most important: many gene loci can tolerate copy number polymorphisms that alter their expression dosage [1, 2], there is considerable variation of protein abundances at the cell-to-cell level [3–5], and small changes in gene expression levels during evolution are often argued to be neutral, or nearly neutral [6–9].

Nevertheless, the expression levels of proteins in cells are roughly kept in line with functional requirements [10], and they can be reproducibly measured for a given set of conditions. The measured abundances can differ widely from protein to protein, typically spanning several orders of

Correspondence: Professor Christian von Mering, Institute of Molecular Life Sciences, University of Zurich, Switzerland
E-mail: mering@imls.uzh.ch
Fax: +41-44-635-68-64

Abbreviations: DM, distance to running median; **SILAC,** stable isotope labeling by amino acids in cell culture magnitude (Milo *et al.*, http://bionumbers.org) [11]. The systematic biochemical measurement of absolute protein abundances in a genome-wide fashion is technically demanding, and has only been attempted in yeast [4, 11], an organism of intermediate complexity having the additional advantage of near-complete clone libraries that can provide a consistently quantifiable "tag" at each protein. For restricted sets of proteins or for relatively small bacterial genomes, quantitative approaches are available [12–15], for example in MS *via* the addition of known amounts of mass-shifted control peptides ("spiking") or by peak-detection and integration in ion chromatograms [16]. However, proteome-wide absolute quantification of all expressed proteins in eukaryotes remains difficult and has not yet been attempted for large and complex metazoans, including humans. This leaves many questions unanswered: is there a uniform optimal stoichiometry of the players in the core cellular processes across species? If so, what is this "optimal" abundance for any given eukaryotic protein, and how stringently is it maintained by selection? Are proteins that form functional partnerships typically of the same abundance? How does intrinsic cell-to-cell variation ("noise") in protein expression translate to variability of proteome composition at evolutionary timescales?

While there are currently not enough quantitative data on proteome-wide expression in complex eukaryotes to address these questions, more *qualitative* proteomics approaches do exist, for example systematic MS-based proteome surveys of important model organisms [17–19]. These projects are typically undertaken in order to validate or correct genome annotations, to check the expression status of known or predicted genes in various tissues, and to enable the identification of non-redundant, technically suitable peptides for subsequent, more targeted routine measurements. To cover a large part of the proteome, extensive biochemical fractionation is usually necessary, making these projects large in scope and resource-intensive. Such "shotgun proteomics" experiments generate large lists of tryptic peptide identifications, and, as a side product, information on how often the mass spectrometers have seen any given peptide ("spectral counts"). While in most cases these data were never originally intended for use in quantification, the spectral counts do hold quantitative information, and a number of algorithms have been devised for extracting semi-quantitative abundance estimates from spectral counts [20–25].

Here, we describe an integrated analysis of published shotgun proteomics data, aiming to compare spectral counts in five eukaryotes: yeast, thale cress, human, fruit fly, and nematode (*Saccharomyces cerevisiae, Arabidopsis thaliana, Homo sapiens, Drosophila melanogaster, Caenorhabditis elegans*). We identified gene loci that are common to all five organisms; this set of orthologs roughly defines the "core proteome" likely to be encoded by most contemporary eukaryotes. Because basic cellular processes are largely conserved across eukaryotes, we assume that the relative abundances of proteins in these processes should be conserved as well, roughly maintained by selection. If this is indeed the case, it represents a unique opportunity to externally validate any protein quantification technique: the best technique would be the one that results in the highest correlation of observed abundances across the five organisms. For spectral counting algorithms, comparisons among such distant organisms (*e.g.*: yeast *versus* human) provide another important advantage: the proteins differ sufficiently in sequence, meaning that physical or technical biases per peptide are mostly averaged out.

We find that our initial assumption indeed holds: proteins in the conserved core are significantly correlated in their abundance and the residual variance is not simply random but instead informative of evolutionary and functional constraints, as described below.

2 Materials and methods

2.1 Data sources

We imported MS/MS protein identifications from the PeptideAtlas database [26], namely from the builds dated March and April 2009 for data from yeast and human, respectively. For *C. elegans*, *D. melanogaster* and *A. thaliana*, we employed MS/MS data directly from dedicated, genome-wide projects published previously [17–19] (Supporting Information Fig. 1; in the case of *C. elegans*, our data includes a number of additional, more recent samples, extending the spectral counts by roughly 50% over the published counts). The data that we assembled cover a variety of tissues, cell-lines, environmental conditions and/or developmental stages, and diverse protocols for biochemical fractionation had been applied. For all five organisms, the Trans-Proteomic Pipeline had originally been used for protein identification (http://tools.proteomecenter.org/TPP.php). In this pipeline, and in PeptideAtlas, a uniform cutoff for the reliability of peptide identifications is used (*i.e.* PeptideProphet score$\geq$0.9).

2.2 Protein abundance quantification

We used a simple spectral counting algorithm employed previously [19], to estimate the protein abundance from the frequencies of observed peptide spectra. For any theoretical tryptic peptide within a protein, we first estimated its likelihood of being successfully identified from MS/MS data, based on its length – a dependency that can be learned from the data and that is roughly the same for all five organisms studied here (Supporting Information Fig. 2). The actual spectral counts were then weighted by this length-based detectability factor (see below). We did not account for more elaborate physical properties of the peptides [23, 24], for three reasons: (i) our analysis contains data from several laboratories, and not all of these have used the exact same setup for sample processing and MS acquisition, (ii) in our hands, length is the most important determinant of peptide detectability, and (iii) sequence-based algorithms could potentially be subtly over-trained, which could result in spuriously high-abundance correlations for orthologs that have high-sequence conservation. Using our simplified approach, the latter problem did not occur: abundance correlations were not higher for proteins of high-sequence conservation (*i.e.* abundance-corrected variance and sequence conservation were not correlated; data not shown). Similarly, we chose not to exclude peptide observations from certain specific experimental setups (such as ICAT, which enriches for Cys-containing peptides). We did test the removal of all Cys-containing peptides, but this resulted in lower correlations against external references, meaning that Cys-containing peptides currently deliver more signal than noise.

Still, there remains the possibility that a given pair of proteins appear to have a similar abundance simply because their constituent tryptic peptides have the same intrinsic "MS-detectability" (*i.e.* beyond the peptide-length effect for which we already correct). To test for this, we artificially suppressed equivalent peptides, by only using alternating sections of the aligned proteins (Supporting Information Fig. 3). As a control, we downsampled the data by the same amount, but this time always considered equivalent peptide positions only. This test does reveal a small, residual effect of shared peptide detectability: for example in the case of the human/yeast comparison, using alternating peptides results in a correlation of $R_S = 0.494$, but using equivalent peptides gives a slightly higher correlation of $R_S = 0.518$. However, this residual effect is quite small, and it has no impact on our further conclusions. Additional support for the validity of our abundance estimates lies in their correlation with independent, known surrogates for protein abundance (Supporting Information Fig. 4): our data show a correlation of $R_S = 0.65$ with the "codon adaptation index" in yeast, and an inverse correlation ($R_S = -0.27$) with protein length.

We describe all individual protein abundances in "parts per million", relative to the molecule counts of all other proteins in the detected proteome (or, alternatively, with reference to the core proteome only). We did not consider splice-variants separately, *i.e.* our measurements are "locus-based": all splice-variants encoded by a locus are aggregated and contribute jointly to a single abundance value for this locus. The actual abundance values were computed as follows: we counted how often any of its amino acids had been identified in a protein, divided by the total number of amino acids in the protein sequence (the latter being length-corrected as described above). An additional length restriction limiting peptides to within ≥ 7 and ≤ 40 amino acids (modified from [23]) was applied, and final counts were normalized and expressed as parts *per* million

$$a = \frac{\sum_i \mathrm{number}(p_i) \cdot \mathrm{length}(p_i)}{\sum_j \mathrm{length}(q_j) \cdot f(q_j)}$$

where a is the protein abundance, p the identified peptides, q the tryptic peptides (*in silico* digest) and $f(q)$ the peptide length correction factor.

It should be stressed that the abundance estimates we thus derive are still subject to considerable error. For example, upon randomly splitting the yeast peptide data in half, the corresponding abundance estimates only correlate to each other with $R_S = 0.92$. From this, we infer a conservative average error of around twofold for the individual estimates (Supporting Information Fig. 5). This error is even larger when splitting the data not randomly, but along different experimental samples (Supporting Information Fig. 5), showing that there are noticeable systematic differences among the various procedures and platforms used.

2.3 Orthologs

We constructed a dedicated set of orthologous groups covering the five organisms we studied here. For this, we imported protein sequences from version 8.0 of the STRING database [27], which contains a complete, pre-computed set of all-against-all BLAST homology relations for these sequences. From the homology relations, we computed orthologous groups essentially as originally proposed by Tatusov *et al.* [28], in an implementation written for the eggNOG database [29]. Briefly, the groups are constructed by joining "triangles" of reciprocal-best-match relations, each involving three species. Triangles are joined when they share one edge, whereby the highest-scoring triangles are joined first. Prior to triangle formation, we search for proteins that are more closely related to each other within an organism than to any of the proteins in the other four organisms ("inparalogs"). These inparalogs are grouped and represented by their highest-scoring member in the subsequent triangle searches. After triangle joining, all pair-wise alignments are tested to verify that the proteins in a group can all be aligned to each other, in a way that defines at least one common sequence segment.

2.4 Expression variance

The 1581 orthologous groups present in all five organisms were used to define the core proteome, and for 1172 of these we were able to infer all five abundance estimates (one for each organism). In the case of organisms having more than one protein in an orthologous group, the abundances of these inparalogs in the organism were added up. The observed normalized variance *per* group ("CV") was inversely related to protein abundance (Fig. 3; $p < 10^{-15}$). To account for this effect and to obtain an "intrinsic" measure of variance that could differentiate the variance of proteins independent of their abundance, we ranked CV values by abundance and computed a running median (Fig. 3). Then, for each protein family, the difference between the observed variance and the median variance at that particular abundance range (the "DM" value: "distance to running median") was introduced as a measure for the "intrinsic variance" (similar to the procedure in [4]). We express DM values in normalized form (*i.e.* in percent of the median value), with negative values indicating less variance than expected, and positive values indicating more variance than expected.

2.5 Functional annotations

Since yeast is arguably among the best-annotated organisms (in terms of molecular biology details), we categorize orthologous groups of proteins based on the annotation of their yeast protein member(s). Gene ontology annotations of yeast proteins were imported from *Saccharomyces* Genome Database [30], and an updated catalog of protein complexes was imported from CYC2008 [31].

2.6 Statistical tests

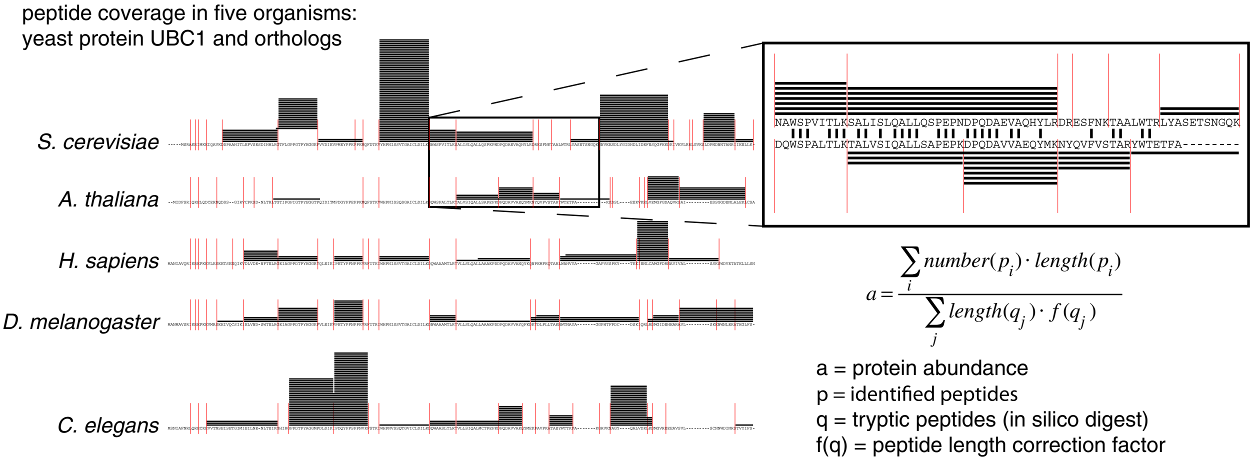
Unless otherwise noted, two-sided Kolmogorov–Smirnov tests were used to gauge the significance of observed differences in distributions. To correct for multiple testing, *p*-values were adjusted according to Benjamini and Hochberg [32]. All abundance correlations are rank-based (Spearman correlations), and *p*-values for these correlations were computed by the "AS 89" algorithm as implemented in the R software package.

3 Results

In order to characterize the eukaryotic core proteome and its compositional stability, we first defined a set of protein families universally encoded in the five organisms we studied (*S. cerevisiae*, *A. thaliana*, *H. sapiens*, *D. melanogaster*, and *C. elegans*). We found 1581 such protein families, usually represented by exactly one protein-coding locus in a given organism (584 families contained a single locus in all five genomes; the average representation of all 1581 groups is 1.41 loci *per* organism). This set of proteins likely represents a nearly universal, ancient core of eukaryotic proteins – comprised mostly of proteins with information processing functions, metabolic functions, and cellular maintenance functions.

Independent of the orthology detection, we also estimated the abundance of all detectable proteins in these five organisms, based on about six million peptide spectra available from public databases (Fig. 1 and Supporting Information Fig 1; see also Section 2). This resulted in protein abundance values extending over more than four orders of magnitude, from less then 1 ppm (parts per million) to more than 10000 ppm. Of the 1581 protein families having representatives in the five organisms, 1172 received an abundance estimate in all five organisms – this is the set of protein families we used for all evolutionary analyses described below (Supporting Information Table 1).

In the case of yeast, four independent experimental data sets on protein abundance are available [4, 11, 23, 33], two of which are not based on MS and can therefore be used for validation. We find that those two latter sets correlate reasonably well with our estimates based on spectral counting ($R_S = 0.65$ and $R_S = 0.58$, respectively; (Supporting Information Fig. 6)). However, why are these correlations not higher? They could in principle be limited by errors on both sides – errors within the experimental measurements and/or errors within our spectral counting technique. To



$$a = \frac{\sum_i number(p_i) \cdot length(p_i)}{\sum_j length(q_j) \cdot f(q_j)}$$

a = protein abundance
p = identified peptides
q = tryptic peptides (in silico digest)
f(q) = peptide length correction factor

Figure 1. Tryptic peptide observations on aligned orthologs in five species. The yeast protein UBC1 (a Ubiquitin-conjugating enzyme), and four of its orthologs in multicellular eukaryotes are shown aligned; the predicted tryptic cleavage sites are marked in red. Peptide observations, from a collection of published shotgun proteomics experiments, are pooled and shown as horizontal lines above the protein sequences; each line represents one peptide observation. Inset: magnified section showing details for the yeast/plant alignment; conserved residues are indicated and the plant peptides are shown below the sequence for clarity. Lower right: abundance computation for individual proteins. The observed peptides are normalized by the theoretically expected tryptic peptides in a given protein; the latter are corrected for "observability" based on their length.

investigate this, we used transcript abundances as an independent "arbiter": we compared the data sets against absolute transcript abundances in yeast (and also against transcript abundances in *C. elegans,* the latter being an arbiter at a much larger evolutionary distance). Despite the limited overall correlation between transcript and protein abundances [34, 35], transcripts can serve as "arbiters" because they share little technical biases with proteins: a gene that encodes a problematic protein (*e.g.* a transmembrane protein) may still encode a transcript that can be reliably measured, and *vice versa* (*e.g.* a transcript with problematic secondary structure may still encode a protein that is well-suited for MS). Remarkably, in all tests our spectral-counting-derived protein abundances showed the best correlations against the transcript levels (Supporting Information Fig. 7). Together with the fact that the spectral counting data have a much higher coverage than the biochemical experimental data, this suggests that they are indeed among the best quantitative protein abundance data available in any eukaryote to date (rivaled, in yeast only, by recent stable isotope labeling by amino acids in cell culture (SILAC) data [33], which correlates slightly lower but has better coverage). Globally, we covered about 56% of all predicted proteins encoded in the five genomes (ranging from only 48% in plants to about 60% in fly). Note that, within yeast, the best correlation is seen between the two MS-based data sets (our spectral counting *versus* the SILAC data; $R_S = 0.70$; Supporting Information Fig. 6). This is notable because distinct measurement strategies have been used for the two sets (area under the peak, and spectral counting, respectively). Still, this correlation is far from perfect, and is a reminder of the relatively high level of noise in any genome-wide protein abundance quantification (see Section 2 for noise quantification).

By mapping our abundance values onto the orthology information, we then obtained five independent snapshots of essentially the same core proteome, separated not only by hundreds of million years of independent evolution, but also by differences in sampling material and procedures (*e.g.* various growth conditions, differences in handling, tissue selection, and sample preparation). Because of these evolutionary and environmental/procedural differences, the five snapshots should represent a good view of the actual compositional variance of the core proteome. It should be noted that the core proteome represents only a rather small part of the full proteomes (covering 20.7% of all proteins in yeast, but only 7.6% of the proteins in human). In addition, we are effectively "averaging" over various conditions and stages, meaning that we cannot study the varied expression of individual proteins in response to external stimuli. Nevertheless, the averaged core proteome enables comparisons between the five organisms on an equal footing. In addition, it consists mainly of "house-keeping" genes whose tissue-to-tissue expression variability is limited, and it

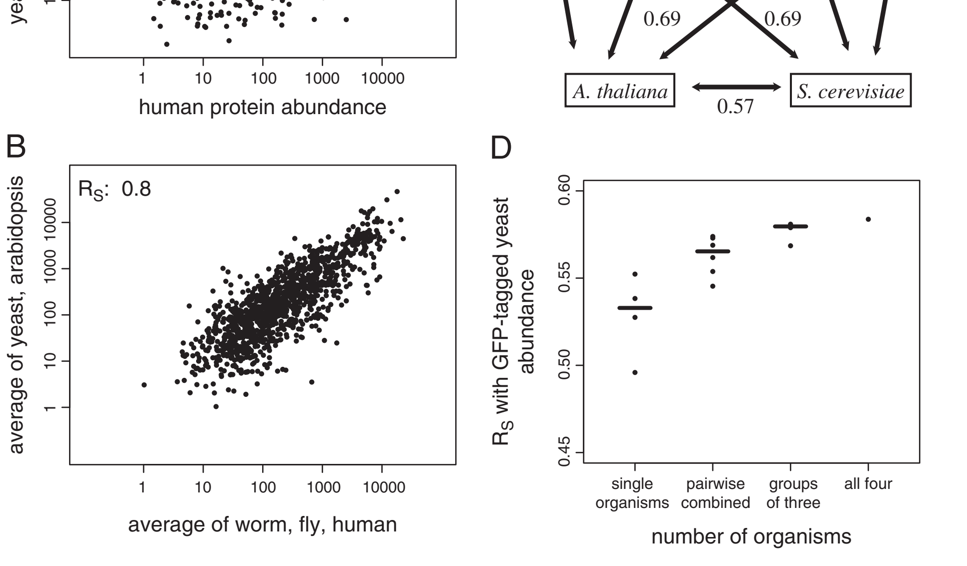


Figure 2. Protein abundance correlations. (A) Comparison of the inferred abundances of orthologous proteins in two organisms (yeast *versus* human). (B) Increased abundance correlation upon data aggregation: comparison of the average abundances of yeast and plant, *versus* the average abundance of the three animals. (C) Spearman rank correlations for all pairwise comparisons. (D) Data aggregation across organisms also improves the correlation against non-MS derived, independently measured protein abundances in yeast (in this case, the yeast MS data were left out). All abundances in this figure are indicated relative to the core proteome only (*i.e.* non-conserved proteins are not contributing to the total when computing the abundances; the indicated values denote relative molecule counts in parts *per* million).

encompasses many essential functions such as information processing, cellular structure, metabolism, and signal transduction.

To analyze the results, we first computed the rank abundance correlations between the various core proteomes, which we interpret as a lower limit of their quantitative conservation. As reported earlier for the case of *C. elegans versus D. melanogaster* [19], we generally found these correlations to be surprisingly high, ranging from $R_S = 0.57$ for the comparison yeast *versus* plant, to $R_S = 0.85$ for the comparison nematode *versus* fruit fly (Fig. 2). Most likely, the underlying biological correlations are in reality even higher, since spectral counting is presumably limited by the numerical noise that is associated with under-sampling. If limited sampling is indeed an issue, then we should be able to increase the correlations by including more data – and this is what we observe: we can improve the correlation between the non-MS-derived, independently measured protein abundances in yeast and our MS-derived abundances by averaging the MS data from an increasing number of other species (Fig. 2D). The correlations across large evolutionary distances were also improved by aggregation, for example by grouping the three animals and comparing them to a grouping of yeast and arabidopsis ($R_S = 0.80$; this compares to R_S in the range of 0.64–0.70 for the individual comparisons at this distance; note that yeast and arabidopsis are not a meaningful phylogenetic grouping but were merely grouped because both are at large distances from animals). This latter comparison of course does not only increase sampling depth, but also smoothes out individual differences among the organisms.

Given the remarkable evolutionary conservation of protein abundances, we next tested whether groups of functionally linked proteins would be expressed at distinct, typical abundances. For this, we grouped proteins according to their membership in stable yeast protein complexes [31]. Indeed, we observe that members of the same protein complex tend to have similar abundances (Supporting Information Fig. 8). This observation suggests that we can expect to further decrease

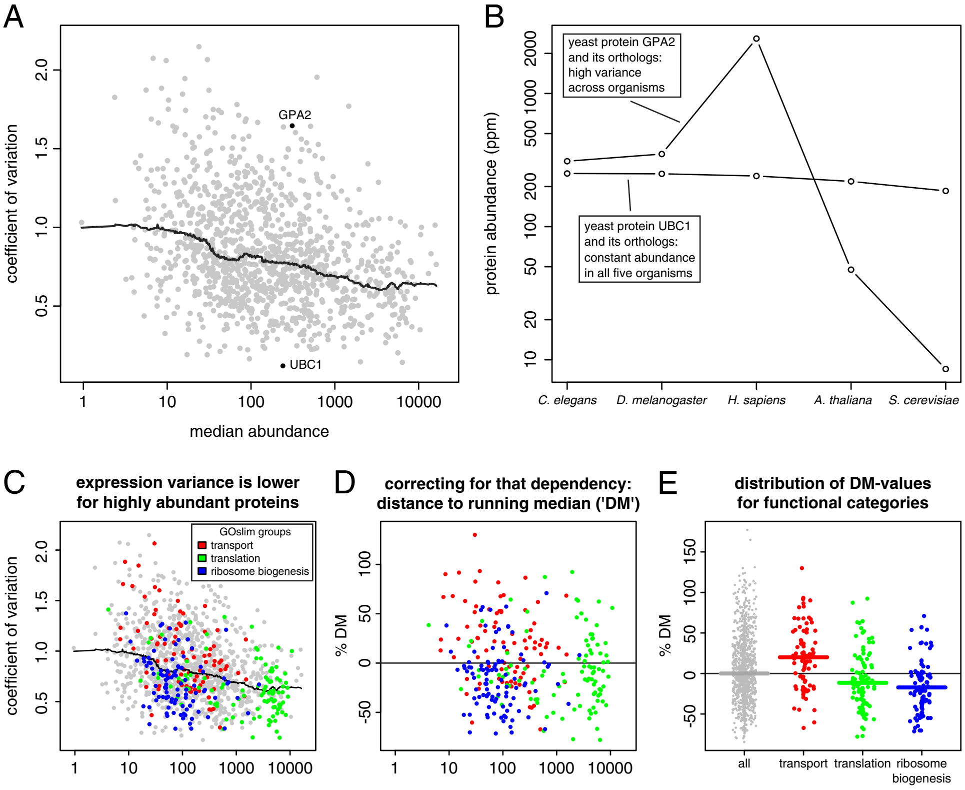


Figure 3. Variance of protein abundance. (A) Coefficient of variation (*i.e.* variance divided by the mean) across five organisms, plotted against median abundance (each dot represents an orthologous group of proteins present in five organisms). A running median is shown, with a window size of 200 and truncated window-sizes at both ends. (B) Two exemplary proteins with similar median abundance, but large differences in their coefficients of variation are shown. GPA2 is a G-protein subunit involved in glucose sensing, whereas UBC1 is an Ubiquitin-conjugating enzyme that helps degrade short-lived or abnormal proteins. GPA2 is presumably strongly regulated; hence the extreme differences in observed abundances. (C) Same plot as in (A), with three high-level functional categories marked in color. (D) Normalization against the abundance dependency: In this plot, variance is now expressed as ''distance to the running median'' (DM), indicated in percent of the median. (E) Distributions of DM values, for all proteins (gray) and for three functional categories shown in color (all three differ significantly from the background distribution; $p = 0.002$, $p = 0.003$, and $p < 1e-05$, respectively).

numerical noise by grouping proteins by their membership in protein complexes – this is what we find, resulting for example in an abundance correlation of $R_S = 0.92$ between entire protein complexes in the worm on the one hand and the fly on the other hand (Supporting Information Fig. 9). Interestingly, we could not further increase the correlation across the animal *versus* fungi/plant grouping, suggesting that our observed correlation of $R_S = 0.80$ is already close to the actual biological correlation at this large distance.

Next, we studied the abundance *variance* across species, for each individual protein family (Fig. 3; see Section 2). We first observed that the variance scales inversely with protein abundance, as has been observed for cell-to-cell expression variance as well [4]. In our case, it is likely that this observation is at least partly due to technical limitations – instrument error and spectral counting inaccuracies are presumably more prevalent for proteins of low abundance. Therefore, we normalized the variance data for protein abundance and obtained an "intrinsic" variance measure termed DM (distance from a running median of variances), in accordance with [4] (Supporting Information Table 2). These variance values are still very noisy in themselves (given that each variance calculation is based on only five data points), but they provide – for the first time – an objective view on the intrinsic quantitative conservation of the eukaryotic core proteome. In order to be biologically meaningful, these variances should correlate with externally described functional properties of the proteins, which is what we investigated next.

First, we grouped proteins into broad functional categories as described in the Gene Ontology database (GOslim) [36]. We observed that proteins in the biological process category "Transport" show unusually high variance, whereas proteins in the categories "Translation" and in particular also "Ribosome biogenesis" showed unusually low variance (Fig. 3). Similarly, we observed significant differences among cellular localizations and molecular functions, most notably a lower-than-expected variance of proteins in the categories "Transcriptional regulator" and "Translational regulator". We also observed a weak but significant correlation of our variance with cell-to-cell expression noise ($R_S = 0.11$, $p = 0.003$; (Supporting Information Fig. 10)). Next, we turned to essentiality data in yeast [37]; here again, we observed differences in variance: the variance is significantly lower for essential genes as compared to non-essential genes ($p = 0.0001$; Fig. 4).

For functional groupings, the combination of low variance across species and good intra-group agreement of abundance values across genes suggest that global stoichiometries between functional groups might be relatively constant across organisms. As a case in point, we computed for all five organisms the relative stoichiometries of the categories "Translation" *versus* "Ribosome biogenesis". We find that this ratio is about 20:1, and remarkably, it is relatively stable across all five organisms (ranging from 13:1 in human and arabidopsis to 23:1 in fruit fly).

The last aspect we addressed was gene duplicability – many historical instances of gene duplications presumably affected the abundance of the encoded proteins. Thus, protein families whose abundance variations are more tightly controlled might have experienced a stronger purifying selection against fixation of a duplicate copy. To test this, we classified orthologous groups according to the number of paralogs they contain (this roughly reflects the frequency of retained gene duplication events in the five lineages). We observe that the groups with the lowest numbers of paralogs indeed showed the lowest variance ($p = 0.0013$; Fig. 4), suggesting that the need to maintain a given protein abundance level may at least partly be responsible for restrictions on gene duplicability.

4 Discussion

Biological variation of protein abundances has been studied extensively, but so far only at the level of individual cells [4, 5, 38, 39], or across closely related species [40]. In contrast, here we study the extent to which protein expression levels can vary

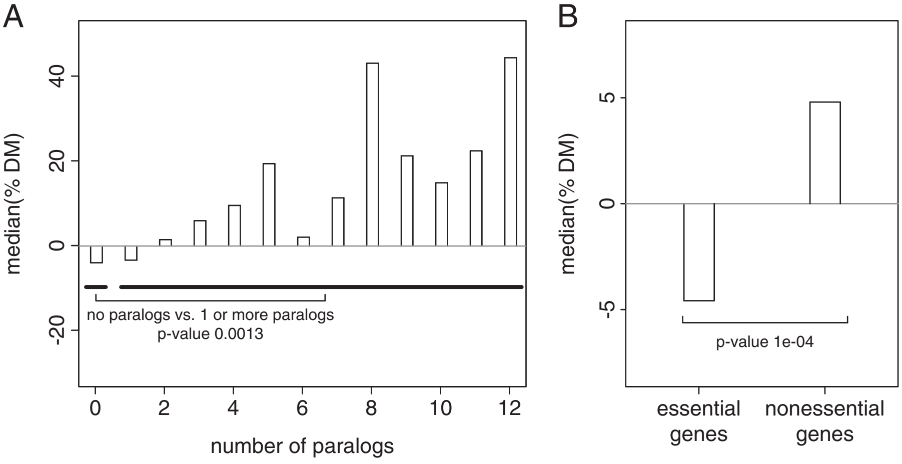


Figure 4. Abundance variance and evolutionary signals. (A) The abundance variance of an orthologous group increases with the number of inparalogs in that group. (B) Essential genes have a lower abundance variance than non-essential genes. Both plots are based on abundance-corrected variance measures (DM), as in Fig. 3.

over large evolutionary distances, while we average over distinct tissues and/or environmental conditions. Thus, we are essentially probing the upper limit of the quantitative compositional flexibility of the "core" proteome in the course of evolution.

Many parts of the proteome are of course heavily regulated, and will change their expression levels over orders of magnitude in response to external stimuli or developmental cues. However, what we study here is a unique and stable subset of the proteome – the "core" as defined by ubiquity across eukaryotic kingdoms. A significant part of this core is likely expressed in all cells and under all conditions. Indeed, we observe that of the 1581 protein families we studied, 75% can be detected by MS in all five organisms (this figure increases to 92% when lowering the requirement to four organisms). Proteins in the core appear generally better conserved and/or better suited for quantification: within yeast, for example, the two MS-based data sets (SILAC and spectral counting) correlate better for core-proteins ($R_S = 0.75$) than for non-core proteins ($R_S = 0.62$), the former correlation being higher than any inter-organism correlation involving yeast.

For this subset – the "eukaryotic core proteome" – we have now established a reference of typically observed protein abundance ranges, and this can serve as a baseline against which to compare cellular protein abundance states in the future. To us, the observed evolutionary conservation suggests purifying selection: proteins would be roughly maintained at optimal abundance levels – whereby functional requirements would limit the amount of underproduction, and both toxicity and the metabolic cost of protein production would limit over-production. The metabolic cost of protein production can be visible to selection in eukaryotes, at least in species with large effective population sizes [41]. While complex eukaryotes such as humans may be relatively indifferent to the metabolic cost of protein production, at least for lowly expressed individual proteins, the core proteome examined here consists largely of highly expressed proteins, for which total production cost may be appreciable. That expression levels are to some extent maintained by purifying selection has already been proposed based on transcript abundance data, although the extent of selection at the transcript level remains controversial [7, 8, 42–45].

Invoking selection in general is also compatible with our observation that essential genes are better maintained at their typical abundance level than non-essential genes, presumably because of both abundance mismatches having greater fitness consequences for these genes and essential proteins being often highly expressed [46, 47] and thus imposing a larger metabolic burden on the cell. Furthermore, as the overall cost of protein production is proportionally higher for more abundant proteins, a role of production costs in limiting abundance variation is also consistent with our finding of a lower CV for highly abundant proteins. Most striking, however, is the implicit conclusion that core protein stoichiometries actually do have an optimum that applies across eukaryotic kingdoms. This suggests a detailed quantitative conservation of the core cellular processes, independent of vastly different physiologies and cellular organizations. Thus, to modify a quote of Jacques Monod, "what is true for yeast is true for the elephant", not just in principle but also in rather surprising quantitative detail.

Interestingly, we have not observed a correlation between our abundance variance and protein evolutionary rate (data not shown). This appeared puzzling at first, as proteins with more conserved expression levels might intuitively also be functionally more constrained, and thus perhaps also in their abundance variance. However, evolutionary rate at the sequence level appears to be largely dominated by constraints from protein folding and not necessarily by functional constraints [48, 49]. In contrast, our variance would mainly be controlled by constraints at the level of gene regulation, for example by keeping production levels and turnover rates of both transcript and protein relatively constant. From the outset, protein evolutionary rate could have been a potential confounding factor of our analysis, as it is known to be correlated with abundance itself, and to some extent also with essentiality (which, in turn, is correlated with gene duplicability). However, when we performed a step-wise multiple linear regression, testing these and other variables, they were not detected as confounding factors (Supporting Information Fig. 11). For abundance variance, the most important predictor remains abundance itself, as shown in Fig. 3A. After correcting for this, the number of duplicated genes follows, and then essentiality and cell-to-cell noise (Fig. 4; Supporting Information Fig. 11).

Despite the strong overall correlations of protein abundances, we did of course observe differences in the abundance of individual proteins across species. Do these differences reflect adaptations, short-term gene regulation due to distinct stimuli, or are they the consequences of genetic drift? All else being equal, drift in abundance may be expected to be stronger for proteins that can tolerate higher levels of expression noise (cell-to-cell variation at given environmental conditions). That our variance at evolutionary timescales correlates only very weakly with cell-to-cell protein expression noise therefore suggests that many of our observed differences are perhaps not due to drift, but may reflect regulation or adaptations to the substantial physiological differences among the eukaryotes studied here.

Finally, we observed that gene families with more duplicates tend to show higher variance in protein expression levels. This argues against a strict model of sub-functionalization, in which the functions of the ancestral protein would be simply divided up between the duplicate copies; in this case, their total abundance should remain identical and our variance would not be affected – since we add up the contributions of all paralogs in a family. On the other hand, total abundance levels do not grow linearly with the number of paralogs either [19], and hence duplications most likely fix due to a mixture of sub- and neo-functionalization.

In summary, our analysis provides the first quantitative overview of the core eukaryotic proteome, and helps to further establish spectral counting as a semi-quantitative measure (provided that a good sampling depth can be achieved). It will be intriguing to extend the analysis to more organisms and to deeper spectral counts, in order to achieve a more fine-grained view on protein abundance stability, on purifying selection, and perhaps also on ways to objectively separate the ''constitutive'' from the ''regulated'' parts of any given proteome.

This work has been supported by the University of Zurich through its Research Priority Program ''Systems Biology and Functional Genomics'', by the Gerbert Rüf Foundation, by the Ernst Hadorn Foundation and by the Swiss National Science Foundation.

The authors have declared no conflict of interest.

5 References

[1] Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E. *et al.*, Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 2007, *315*, 848–853.

[2] Henrichsen, C. N., Chaignat, E., Reymond, A., Copy number variants, diseases and gene expression. *Hum. Mol. Genet.* 2009, *18*, R1–R8.

[3] Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M. *et al.*, Noise in protein expression scales with natural protein abundance. *Nat. Genet.* 2006, *38*, 636–643.

[4] Newman, J. R., Ghaemmaghami, S., Ihmels, J., Breslow, D. K. *et al.*, Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* 2006, *441*, 840–846.

[5] Cohen, A. A., Geva-Zatorsky, N., Eden, E., Frenkel-Morgenstern, M. *et al.*, Dynamic proteomics of individual cancer cells in response to a drug. *Science* 2008, *322*, 1511–1516.

[6] Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I. *et al.*, A neutral model of transcriptome evolution. *PLoS Biol.* 2004, *2*, E132.

[7] Bedford, T., Hartl, D. L., Optimization of gene expression by natural selection. *Proc. Natl. Acad. Sci. USA* 2009, *106*, 1133–1138.

[8] Yanai, I., Graur, D., Ophir, R., Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *OMICS* 2004, *8*, 15–24.

[9] Fay, J. C., Wittkopp, P. J., Evaluating the role of natural selection in the evolution of gene regulation. *Heredity* 2008, *100*, 191–199.

[10] Dekel, E., Alon, U., Optimality and evolutionary tuning of the expression level of a protein. *Nature* 2005, *436*, 588–592.

[11] Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W. *et al.*, Global analysis of protein expression in yeast. *Nature* 2003, *425*, 737–741.

[12] Ong, S. E., Mann, M., Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* 2005, *1*, 252–262.

[13] 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.

[14] Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., Kuster, B., Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* 2007, *389*, 1017–1031.

[15] Malmstrom, J., Beck, M., Schmidt, A., Lange, V. *et al.*, Proteome-wide cellular protein concentrations of the human pathogen *Leptospira interrogans*. *Nature* 2009, *460*, 762–765.

[16] 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.

[17] Brunner, E., Ahrens, C. H., Mohanty, S., Baetschmann, H. *et al.*, A high-quality catalog of the *Drosophila melanogaster* proteome. *Nat. Biotechnol.* 2007, *25*, 576–583.

[18] Baerenfaller, K., Grossmann, J., Grobei, M. A., Hull, R. *et al.*, Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science* 2008, *320*, 938–941.

[19] 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.

[20] Gao, J., Opiteck, G. J., Friedrichs, M. S., Dongre, A. R., Hefta, S. A., Changes in the protein expression of yeast as a function of carbon source. *J. Proteome Res.* 2003, *2*, 643–649.

[21] Liu, H., Sadygov, R. G., Yates,J. R., 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 2004, *76*, 4193–4201.

[22] Ishihama, Y., Oda, Y., Tabata, T., Sato, T. *et al.*, Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* 2005, *4*, 1265–1272.

[23] Lu, P., Vogel, C., Wang, R., Yao, X., Marcotte, E. M., Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* 2007, *25*, 117–124.

[24] Mallick, P., Schirle, M., Chen, S. S., Flory, M. R. *et al.*, Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* 2007, *25*, 125–131.

[25] Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K. *et al.*, Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J. Proteome Res.* 2006, *5*, 2339–2347.

[26] Deutsch, E. W., Lam, H., Aebersold, R., PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.* 2008, *9*, 429–434.

[27] Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S. *et al.*, STRING 8--a global view on proteins and their functional

interactions in 630 organisms. *Nucleic Acids Res.* 2009, *37*, D412–D416.

[28] Tatusov, R. L., Koonin, E. V., Lipman, D. J., A genomic perspective on protein families. *Science* 1997, *278*, 631–637.

[29] Jensen, L. J., Julien, P., Kuhn, M., von Mering, C. *et al.*, eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.* 2008, *36*, D250–D254.

[30] Hong, E. L., Balakrishnan, R., Dong, Q., Christie, K. R. *et al.*, Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res.* 2008, *36*, D577–D581.

[31] Pu, S., Wong, J., Turner, B., Cho, E., Wodak, S. J., Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.* 2009, *37*, 825–831.

[32] Benjamini, Y., Hochberg, Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 1995, *57*, 289–300.

[33] de Godoy, L. M., Olsen, J. V., Cox, J., Nielsen, M. L. *et al.*, Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* 2008, *455*, 1251–1254.

[34] Gygi, S. P., Rochon, Y., Franza, B. R., Aebersold, R., Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* 1999, *19*, 1720–1730.

[35] Fu, X., Fu, N., Guo, S., Yan, Z. *et al.*, Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* 2009, *10*, 161.

[36] Harris, M. A., Clark, J., Ireland, A., Lomax, J. *et al.*, The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 2004, *32*, D258–D261.

[37] Giaever, G., Chu, A. M., Ni, L., Connelly, C. *et al.*, Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 2002, *418*, 387–391.

[38] Raser, J. M., O'Shea, E. K., Noise in gene expression: origins, consequences, and control. *Science* 2005, *309*, 2010–2013.

[39] Becskei, A., Kaufmann, B. B., van Oudenaarden, A., Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nat. Genet.* 2005, *37*, 937–944.

[40] Fu, N., Drinnenberg, I., Kelso, J., Wu, J. R. *et al.*, Comparison of protein and mRNA expression evolution in humans and chimpanzees. *PLoS ONE* 2007, *2*, e216.

[41] Bragg, J. G., Wagner, A., Protein material costs: single atoms can make an evolutionary difference. *Trends Genet.* 2009, *25*, 5–8.

[42] Chan, E. T., Quon, G. T., Chua, G., Babak, T. *et al.*, Conservation of core gene expression in vertebrate tissues. *J. Biol.* 2009, *8*, 33.

[43] Yanai, I., Hunter, C. P., Comparison of diverse developmental transcriptomes reveals that co-expression of gene neighbors is not evolutionarily conserved. *Genome Res.* 2009, *12*, 2214–2220.

[44] Xing, Y., Ouyang, Z., Kapur, K., Scott, M. P., Wong, W. H., Assessing the conservation of mammalian gene expression using high-density exon arrays. *Mol. Biol. Evol.* 2007, *24*, 1283–1285.

[45] Gilad, Y., Oshlack, A., Rifkin, S. A., Natural selection on gene expression. *Trends Genet.* 2006, *22*, 456–461.

[46] Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M. *et al.*, Protein abundance profiling of the *Escherichia coli* cytosol. *BMC Genomics* 2008, *9*, 102.

[47] Schmidt, M. W., Houseman, A., Ivanov, A. R., Wolf, D. A., Comparative proteomic and transcriptomic profiling of the fission yeast *Schizosaccharomyces pombe*. *Mol. Syst. Biol.* 2007, *3*, 79.

[48] Drummond, D. A., Wilke, C. O., Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 2008, *134*, 341–352.

[49] Powers, E. T., Balch, W. E., Costly mistakes: translational infidelity and protein homeostasis. *Cell* 2008, *134*, 204–206.