Current Biology 24, 1467–1475, July 7, 2014 ©2014 Elsevier Ltd All rights reserved http://dx.doi.org/10.1016/j.cub.2014.05.044

Article

Deep Proteomics of the *Xenopus laevis* Egg using an mRNA-Derived Reference Database

Martin Wühr,[1,2,4] Robert M. Freeman, Jr.,[1,4] Marc Presler,[1] Marko E. Horb,[3] Leonid Peshkin,[1] Steven P. Gygi,[2,*] and Marc W. Kirschner[1,*]

[1]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA

[2]Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA

[3]Bell Center for Regenerative Biology and Tissue Engineering and National Xenopus Resource, Marine Biological Laboratory, Woods Hole, MA 02543, USA

Summary

Background: Mass spectrometry-based proteomics enables the global identification and quantification of proteins and their posttranslational modifications in complex biological samples. However, proteomic analysis requires a complete and accurate reference set of proteins and is therefore largely restricted to model organisms with sequenced genomes.

Results: Here, we demonstrate the feasibility of deep genome-free proteomics by using a reference proteome derived from heterogeneous mRNA data. We identify more than 11,000 proteins with 99% confidence from the unfertilized *Xenopus laevis* egg and estimate protein abundance with approximately 2-fold precision. Our reference database outperforms the provisional gene models based on genomic DNA sequencing and references generated by other methods. Surprisingly, we find that many proteins in the egg lack mRNA support and that many of these proteins are found in blood or liver, suggesting that they are taken up from the blood plasma, together with yolk, during oocyte growth and maturation, potentially contributing to early embryogenesis.

Conclusion: To facilitate proteomics in nonmodel organisms, we make our platform available as an online resource that converts heterogeneous mRNA data into a protein reference set. Thus, we demonstrate the feasibility and power of genome-free proteomics while shedding new light on embryogenesis in vertebrates.

Introduction

Recent advancements in mass spectrometry (MS)-based proteomics now enable global identification and quantification for up to ~10,000 proteins in a single experiment, along with associated posttranslational modifications [1–3]. The capability to identify proteins and measure their expression levels in an unbiased manner on a proteome-wide scale can revolutionize many areas of biology. However, many of the most interesting biological problems are best studied in nonstandard organisms: limb regeneration in axolotl [4], red blood cell development in ice fish [5], or craniofacial developmental disorders in Darwin's finches [6]. To understand how different processes evolved, it will be important to compare proteomic composition and dynamics in species from diverse clades.

Unfortunately, proteomics is currently very difficult in organisms without well-annotated genomes. In current approaches, proteins are digested with proteases, and the peptides are ionized, fragmented, and detected via MS/MS fragmentation spectra. In principle, these spectra contain sufficient information to deduce a peptide's amino acid sequence. However, this approach is only feasible for subsets of spectra with exceptional quality. The number of interpretable spectra is significantly increased by matching MS/MS spectra with theoretical spectra generated from all proteins encoded in the studied species. This set should be both complete and accurate to achieve maximum sensitivity and specificity. The paucity of high-quality reference databases is the main reason that MS-based proteomics is currently limited largely to species with well-annotated gene models.

Despite the rapid decrease in sequencing costs, obtaining genome-based protein reference sets for new organisms is time intensive and expensive. Creating accurate gene models for a new species relies on faithfully assembling a genome from short-read sequencing data and training gene predictors. Both processes are often met with bioinformatics and species-specific challenges. For example, the size and polyploidy of some species' genome (e.g., lungfish, axolotl, or *Amoebae* [7–9]) make sequencing challenging for the foreseeable future. In contrast, deep coverage RNA sequencing (RNA-seq) is cost effective, and protein-coding transcripts can be reconstructed using established tools and published protocols for any species [10]. Some attempts have been made to generate a protein reference database by six-frame translations of mRNA [11, 12]. Unfortunately, the majority of the obtained protein sequences are biologically irrelevant, unnecessarily increasing the search space for spectral matching and therefore decreasing sensitivity while increasing the need for computational time and resources.

One underexploited model for proteomic experiments is the African clawed frog *Xenopus laevis* [13–16]. Large amounts of material required for deep proteomic experiments (>100 μg of protein) can be obtained easily from *X. laevis* samples but would be very hard or impossible to obtain from other model organisms (e.g., staged embryonic time series or undiluted, metaphase-arrested cytoplasm called egg extract). However, *X. laevis* has rarely been used for MS due to the lack of a released genome, likely due to the difficulty associated with sequencing quasitetraploid genomes [17].

Here, we demonstrate for the *X. laevis* egg that genome-free proteomics is feasible at remarkable depth and that we can extract biological insight from this proteomics data. For our genome-free protein reference set, we combine multiple sources of mRNA information and use knowledge of sequence similarity to proteins from related species for reading frame detection, frameshift correction, and annotation. In proteomic experiments, our database outperforms alternative approaches and even the latest rounds of preliminary gene models based on the unreleased genome. With more than 11,000 proteins identified with 99% confidence, this is by far the deepest proteomic study on *X. laevis* and one of the

[4]Co-first author

*Correspondence: steven_gygi@hms.harvard.edu (S.P.G.), marc@hms.harvard.edu (M.W.K.)

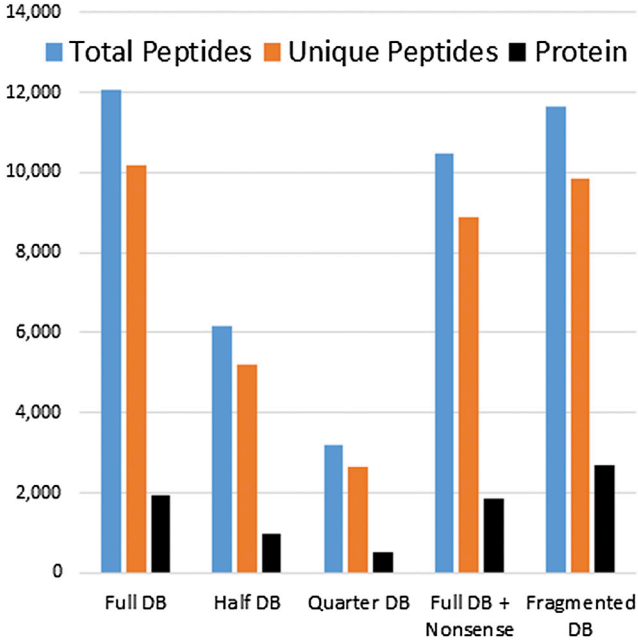

Figure 1. MS Data Can Be Used to Evaluate Relative Reference Database Quality

Spectra from a tryptic digest of yeast lysate were searched against the standard yeast protein database (Full DB). Shown are the number of total peptide spectral matches (blue), unique peptides (orange), or proteins (black) that were confidently identified. To simulate poor reference databases, we removed half (Half DB) or three-quarters of proteins (Quarter DB) from the reference database. The number of identified PSMs and unique peptides scale approximately with the number of proteins in the database. To test how the addition of nonsense sequences would affect the number of identified peptides, we added randomized human proteins to the full yeast database (Full DB + Nonsense). The numbers of peptides and proteins are negatively affected. To simulate a reference database in which proteins are fragmented, we divided at a random position every protein in the reference into two proteins. Whereas the number of identified peptides slightly decreases, the number of identified proteins substantially increases.

deepest analyses performed on any organism. By enumerating the ~11,000 proteins in the *Xenopus* egg and measuring the concentration of each to approximately 2-fold precision, we have produced a valuable resource for the *Xenopus* community. Lastly, we offer the means for researchers to upload and convert mRNA data into a protein reference database for their own proteomic experiments on any organism.

Results

Objective Evaluation of Protein Reference Databases with Peptide Fragmentation Spectra

To construct the best possible reference database for proteomics, we sought a method to evaluate and compare different reference versions objectively. We reasoned that for a given set of peptide fragment spectra, the number of confidently identified peptides is an objective measure of the quality of that reference. To test this assumption, we collected spectra from a trypsin-digested *S. cerevisiae* lysate and searched them against a standard collection of all yeast proteins. We chose yeast, the first sequenced eukaryote [18], because its gene models are exceptionally well annotated. We filtered the spectra, which were matched to peptides (peptide spectrum matches [PSMs]), to 0.5% false discovery rate (FDR) by using the target decoy strategy [1, 19, 20]. Protein grouping was performed with maximum parsimony, with an additional filtering step to 1% FDR at the protein level [1, 21–23]. We then modified the yeast reference set to simulate the effects of searching spectra against low-quality references. First, we randomly removed half or three-quarters of the yeast proteins in the reference database. The number of PSMs, unique peptides, and proteins approximately scales with the number of proteins in the reference database (Figure 1). To test whether irrelevant data would affect the number of identified peptides, we next added shuffled human protein sequences to the yeast reference. As expected, the number of identified peptides and proteins is reduced due to the higher chance of false-positive matches. To simulate a protein reference database with highly fragmented proteins, we bisected each protein from the reference at a random position. With this reference, the number of identified peptides slightly decreased, likely due to the removal of tryptic peptides at the fragmentation site. However, the number of identified apparent proteins increased substantially (Figure 1) because some fragmented proteins were identified once per fragment. To further verify peptide identification as a benchmark for the protein reference set quality, we searched MS spectra obtained from a *X. laevis* sample against the gene models from various species. As expected, the number of identified peptides decreases with evolutionary distance, likely reflecting the lower number of exactly matched peptides in the databases (Figure S1 available online). Thus, we conclude that proteomic data can be used to evaluate the relative quality of a reference protein data set. More specifically, the number of identified peptides, but not the number of identified proteins, can be used as an objective benchmark to compare different reference sets.

Deriving an mRNA-Based Protein Reference Database

For proteomic experiments with *X. laevis*, we needed to obtain a comprehensive, artifact-free reference protein database without using a genome. To guide our approach, we evaluated the success of each processing step by the number of identified peptides when searching our reference against MS/MS data from tryptic peptides of a *X. laevis* egg lysate. With this information, we can evaluate alternative approaches while constructing the database and choose the best possible option to improve our reference incrementally.

An overview of the process we used to generate our reference database, herein called *p*roteomic reference from *h*eterogeneous *R*NA *o*mitting the *g*enome (PHROG), is shown in Figure 2. We combined information from publically available mRNA data and our own RNA-seq data, which we collected to study mRNA dynamics during early development. First, we combined, cleaned, and repeat masked mRNA data from four sources: two RNA-seq de novo assemblies, transcripts from GenBank, and assembled contigs from the *Xenopus* Gene Indices. We then clustered and assembled the preprocessed transcripts by using parameters to maximize assembly, minimize spurious transcript fusions, and collapse homeoalleles that are present in the quasitetraploid *X. laevis*. We compared the assembled transcripts in all six reading frames by using BLASTX against proteins from six vertebrates in order to reveal the most likely translation frame, allowing us to bypass the introduction of large numbers of irrelevant protein sequences when using a six-frame translation. We also used BLASTX alignments to detect and correct for frameshifts that occurred due to sequencing errors. We then translated all transcripts in the BLASTX-hinted frame without regard to start

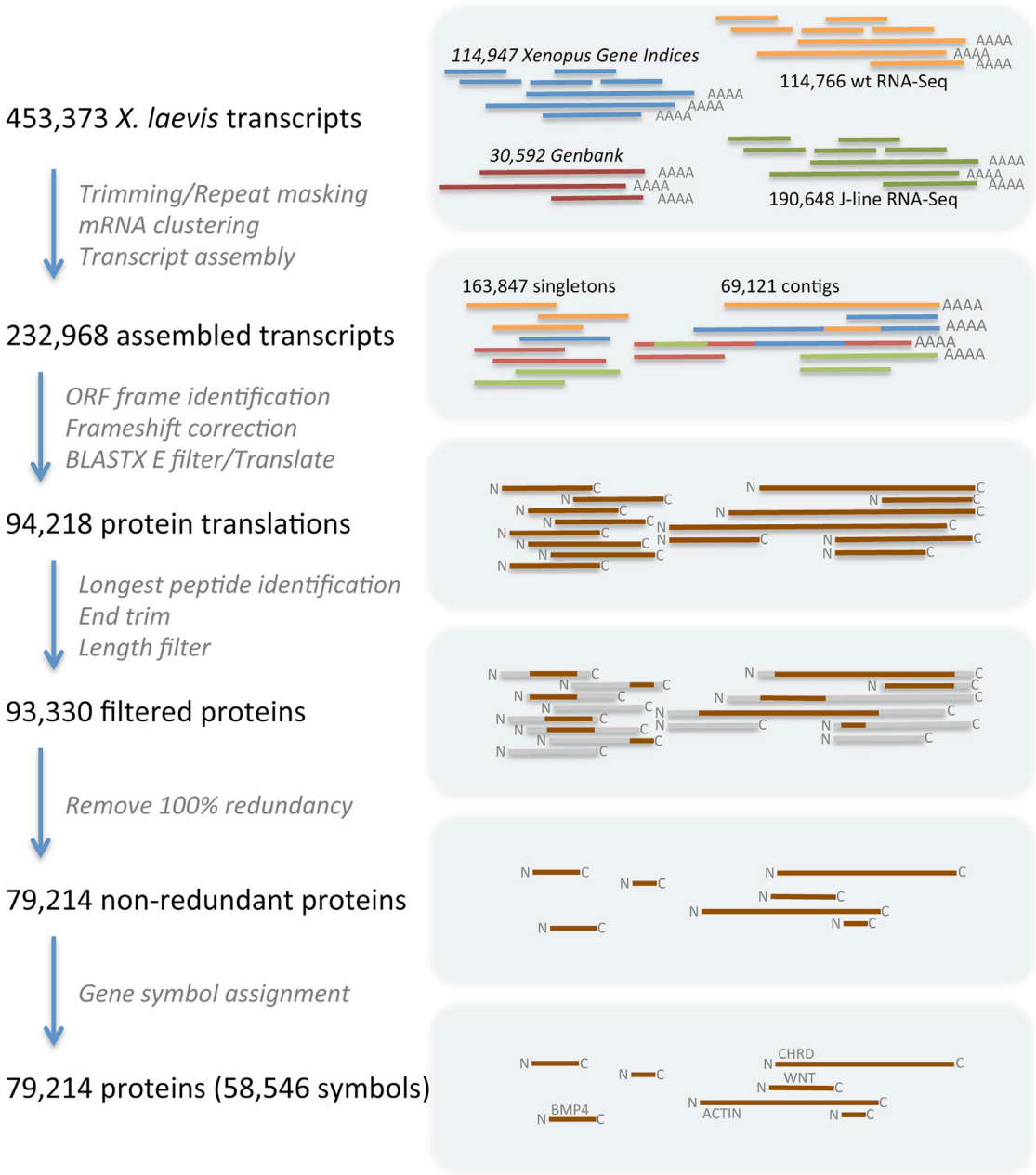

Figure 2. Overview of the Steps for Constructing the High-Quality Protein Reference Set PHROG

Transcripts from four different sources were combined, trimmed and cleaned using SeqClean, masked using RepeatMasker, and clustered and assembled using TGICL/CAP3. The assembled transcripts were aligned against a collection of model vertebrate proteins using BLASTX. The results were used for identifying the correct translation frame, for frameshift correction (if appropriate), and for removing sequences without significant similarity to known proteins. Once translated using BioPerl, the longest peptide for each protein is identified, and the ends are trimmed to match tryptic peptides. The collection is processed to remove 100% redundant proteins using CD-HIT, and gene symbols are assigned to the remaining members using the reciprocal or single best BLAST hit against human proteins. The numbers indicate the numbers of transcripts or proteins in each group.

and stop codons, ignoring translations from transcripts with a BLASTX E value > 1×10^{-5} as we hardly ever detected peptides from proteins with the indicated threshold (Figures S2A and S2B). This hinted translation and filtering method significantly outperformed alternative translation attempts (Table S1A). To determine whether hidden or never-before-seen proteins might be present in these removed transcripts, we used bona fide protein-coding transcripts as an initial training set and used the hidden Markov model (HMM)-based translation method. Adding these trained translations back did not increase our peptide discovery numbers (Table S1A). Next, the translated sequences' N termini and C termini were trimmed to remove sequence ends that were either upstream or downstream of start and stop codons or that would produce fractions of tryptic peptides and therefore could not be detected via standard MS searches. The removal of invalid sequences from the database increased the number of identified peptides (Table S1B). After eliminating redundancies, our final

Table 1. Comparison of Different Reference Databases

	X. tropicalis Gene Models	*X. laevis* Xenbase	*X. laevis* Gene Models	PHROG	PHROG + *X. laevis* Gene Models	PHROG Six-Frame	PHROG RNA-Seq Only
Proteins in database	43,455	34,178	44,159	79,214	123,373	610,557	71,716
Amino acid in database	22,546,772	14,676,179	15,683,803	25,605,893	41,289,696	76,509,919	24,281,510
PSMs	9,300	16,142	17,354	18,867	19,030	17,564	17,156
Unique peptides identified	7,847	13,381	14,531	15,894	16,043	14,791	14,510
Proteins identified	1,850	2,505	2,969	3,130	3,176	3,098	2,923

Comparison of different reference databases. We evaluated the performance of the different reference databases by testing against a tryptic-digested *X. laevis* egg lysate.

database PHROG contained 79,214 proteins (Figure 2). Finally, to facilitate interpretation of identified protein sequences, we assigned protein names and gene symbols by using a modified reciprocal best-BLAST-hit approach based on a target reference of curated human proteins. A summary of the composition of our database and its performance in a proteomic experiment compared to alternative reference sets is shown in Table 1. Judging by the number of identified peptides, via MS, our database outperforms the protein reference from Xenbase, the gene models from *X. tropicalis*, a six-frame translated database, and even the gene models from the unreleased genome assembly version 7.0 (kindly provided by Dan Rokhsar). One alternative to PHROG is using a better-annotated reference set from a related species (e.g., *X. tropicalis*). However, when using MS, a single amino acid mismatch makes it impossible to identify a peptide. By using the *X. laevis* published proteins from Xenbase, we identify ~70% more peptides compared to the *X. tropicalis* reference (Table 1). The preliminary gene models provide a significant improvement for peptide identification over previously known proteins. Surprisingly, even with the latest assembly of the genome, our mRNA-based approach identifies ~10% more peptides. When we combine PHROG with the preliminary gene models as protein reference, we only identify an additional ~1% of peptides compared to using PHROG alone. The PHROG six-frame translated reference database is much larger than all other databases and identifies ~10% fewer peptides compared to PHROG, likely because of additional false-positive hits with irrelevant database entries, which hurts sensitivity (Table S2).

One major advantage of our approach is that we combine mRNA information from various sources, thereby maximizing coverage. Besides our own RNA-seq data, we used publicly available mRNA sources for *X. laevis*, including expressed sequence tags, which are available for many nonstandard model organisms in large quantities [24]. To demonstrate that the mRNA-based proteomics approach is also feasible without public mRNA data, we created a reference relying only on our own RNA-seq data. This database identifies 90% of peptides that the PHROG identifies and approximately the same number of peptides as the *X. laevis* preliminary gene models (Table 1).

Deep Genome-Free Proteomics Demonstrated on the *X. laevis* Egg

To demonstrate the power of the genome-free proteomics approach, we determined the proteomic content of the metaphase-arrested *X. laevis* egg. To obtain the deepest possible coverage, we digested the proteins with both LysC and trypsin or with LysC alone, fractionated each sample with a medium pH reverse-phase column, and analyzed the fractions with liquid chromatography followed by MS (LC-MS). The acquired spectra were searched against our PHROG reference set, the preliminary gene models, and Xenbase protein database for comparison. The results are summarized in Figure 3. By using Xenbase's GenBank proteins known at the time of this writing, we identified 97,999 unique peptides. With the *X. laevis* 7.0 gene models, we identified 26% more peptides. With our PHROG reference, we identified 143,476 unique peptides, an increase of 46% over Xenbase. When we matched these peptides to the minimal number of proteins and filtered to 1% FDR on the protein level, we identified 6,455 proteins from Xenbase, 9,720 proteins with the genome, and 11,103 proteins from PHROG (Figure 3B). Unexpectedly, the relative increase of proteins when comparing PHROG to Xenbase is larger than the relative increase in unique peptides. We believe that this is mostly due to an overrepresentation of the highest-abundant proteins in Xenbase (i.e., the proteins for which most MS/MS spectra will be collected) (Figure S3). In contrast, PHROG seems to allow us to identify many lower-abundant proteins, which would be missed with the Xenbase reference set. Furthermore, PHROG might identify multiple splice forms or proteins with slightly different sequences (e.g., alloaleles), which may be missing in Xenbase. Importantly, the numbers obtained with the very stringent filtering criteria used here indicate that this study is among the deepest proteomic analyses ever performed on any species.

Estimation of the Concentration of Individual Proteins in the *X. laevis* Egg

Beyond providing a comprehensive list of identified proteins, we also wanted to estimate each protein's concentration. The difficult-to-predict ionization efficiency of peptides prevents us from directly measuring absolute protein abundance via MS. However, we can estimate each protein's concentration by summing up the ion current in the MS1 spectrum for all peptides of a protein and normalizing by the number of theoretical tryptic and LysC peptides [25]. We collected published concentrations for 50 proteins in *Xenopus* egg extract from the literature (Table S3) and plotted the concentration against the normalized ion current (Figure 4A). The detected proteins with published concentrations range over four orders of magnitude from 30 μM for Nucleoplasmin [26] to 3 nM for the MAPKKK Mos [27]; from our panel, we only failed to detect the 20 pM Axin [28]. The Pearson correlation for published protein concentration and normalized ion current in log-log space is 0.92 (Figure 4A). We confirmed that we did not overfit our data by performing a 10-fold cross-validation, obtaining essentially the same result (data not shown). Using this correlation, we regressed the protein concentration for all detected proteins (Figure 4B and Table S4). With this approach, the estimated protein concentration differs on average by 1.9-fold compared to the published protein concentrations. The histogram for all estimated protein concentrations shows a median of ~30 nM (Figure 4B).

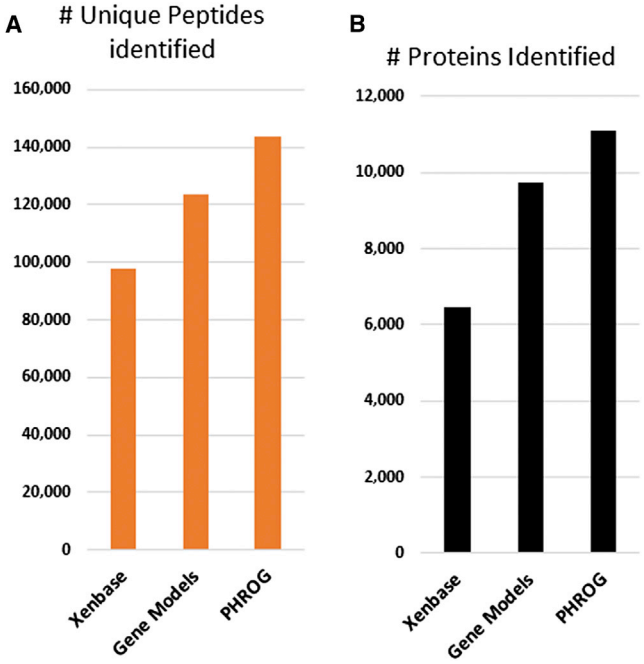

Figure 3. Comparison of Protein Reference Databases for the Fractionated *X. laevis* Egg Sample

(A) Number of unique peptides identified with 0.5% FDR on the peptide level. PHROG significantly outperforms the publically available proteins from Xenbase and even the preliminary gene models from the 7.0 genome assembly as reference database.

(B) Comparison of the number of proteins identified in the egg, with additional filtering to 1% FDR at the protein level and maximal parsimony.

As an additional resource, we provide the protein concentrations summed by their assigned human gene symbols (Table S5). Several distinct *Xenopus* proteins were mapped to the same human gene symbol. This is because similar but distinct proteins in *X. laevis* matched the same human gene during gene symbol assignment. The search results from the preliminary genome indicate that we identified nearly 10,000 distinct *X. laevis* genes (gene models do not contain splice variants).

For further validation, we asked whether subunits of stable protein complexes tend to have similar predicted concentrations. For ten stable complexes [29–31], we plotted the concentration of the subunits for each complex identified via the assigned gene symbols. Remarkably, the complexes' subunits cluster around similar concentrations, as shown in Figure 4C. At first glance, the anaphase promoting complex (APC) subunits are scattered relatively widely. However, some of the APC subunits are known to be dimeric, whereas some are monomeric [30]. Our precision is not good enough to separate these populations, but the dimeric subunits tend to have higher concentrations than the monomeric subunits (Figure 4C). Interestingly, when we perform a similar analysis with components of metabolic pathways, the component's concentrations often vary by many orders of magnitude (Figure 4D).

Relationship of mRNA Abundance and Protein Abundance

Given our previous work in *Xenopus* transcriptomics [32], we sought to understand the relationship between mRNA and protein abundance. Using standard methods to estimate the abundance of the RNA-seq transcripts, we calculate the Pearson correlation of mRNA and protein abundance to be 0.32, whereas the Spearman correlation is 0.30 (in log-log space; Figure S4); these values are low compared to previous studies in tissue culture cells [2, 25, 33]. Unlike tissue culture cells, the *X. laevis* egg, which originates from the oocyte, emerges with a potentially different proteome and transcriptome after maturation. Although the correlation of protein and mRNA abundance is weak, we are more likely to observe the corresponding protein the more abundant the mRNA is (Figure 5A). We asked whether there were systematically overrepresented classes of genes that could only be seen via RNA-seq [34]. After mapping 4,675 gene symbols to our RNA-seq data, we found that membrane proteins (2,013 gene symbols), proteins involved in cell differentiation (894 gene symbols), transcription factors (316 gene symbols), and extracellular matrix proteins (189 gene symbols) are significantly overrepresented in the mRNA-only set (E values $< 1 \times 10^{-10}$). Membrane proteins are known to be harder to detect via MS than soluble proteins, but we currently cannot distinguish whether membrane proteins are overrepresented as RNA because of MS sensitivity issues or because they are not expressed in the egg and are stockpiled for later translation. The same is true for the typically low-abundant transcription factors. For proteins used in differentiation and for extracellular matrix proteins, it seems more likely that the mRNA is present in the egg and will be expressed only during later stages of development.

With the current state of technology, RNA-seq is more sensitive than protein detection via MS. Therefore, we were surprised to find 368 proteins for which we could not find any mRNA support. After running gene set enrichment analyses with these proteins, we found that they were significantly enriched for blood plasma and liver proteins (Figure 5B and Table S6). During oocyte maturation, the yolk protein vitellogenin is synthesized in the liver and transported via the blood plasma to the oocyte, where it is endocytosed [35, 36]. We conclude that many proteins, besides vitellogenin, are also likely to be taken up via endocytosis from the blood plasma during oogenesis. Metabolic labeling experiments in the 1960s noted a small uptake of serum proteins in whole ovary but did not identify any of them [37]. It will be important to evaluate the intracellular role of these proteins during embryonic development.

Discussion

We present here the deepest proteomic study ever performed on *X. laevis* and one of the deepest performed on any organism. We identified ~11,000 proteins and estimated each protein's concentration, ranging more than four orders of magnitude, with an approximate average error of 2-fold. It might be possible to further improve protein concentration predictions by combining normalized ion current with peptide detectability prediction algorithms [38–40]. Our results will be a highly valuable resource for the *Xenopus* egg extract community for data mining, planning new experiments, and complementing previous knowledge. For the development community, it begins to define the dowry of the egg and widens the opportunity for study of translational control, fertilization, and the maternal-zygotic transition. The large amount of material obtainable from *Xenopus* eggs and embryos, coupled with this new resource, should encourage the use of proteomics in development.

We started working on *X. laevis* proteomics in 2011 without access to a genome. We wanted to take advantage of

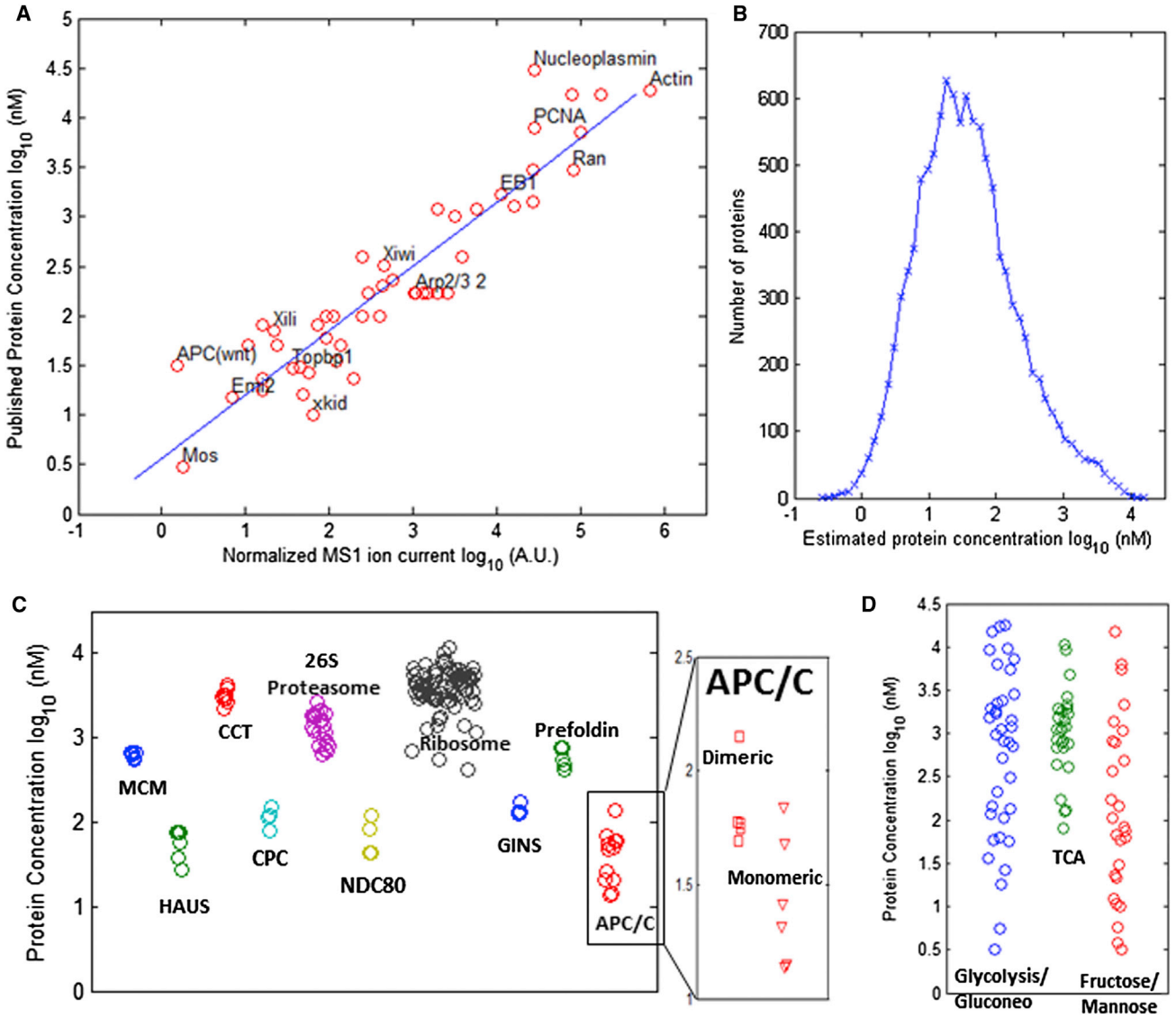

Figure 4. Estimation of Protein Abundance in the *Xenopus* Egg

(A) Previously published protein concentrations for 49 proteins versus measured ion current in MS1 spectrum normalized to protein length. The Pearson correlation is 0.92. On average, the predicted protein concentration is approximately 2-fold different from the reported protein concentration.

(B) Histogram of concentration for all identified proteins regressed from normalized MS1 ion current. Median concentration of measured proteins is approximately 30 nM.

(C) Estimated concentration for subunits of stable complexes is similar. For the APC/C, we additionally distinguished between subunits that were reported to be dimeric (square) or monomeric (triangle) within the complex. Although our accuracy is not good enough to separate the two populations, the estimated concentrations for dimeric subunits tend to be higher than those for monomeric subunits.

(D) Concentrations for enzymes of a metabolic pathway can vary widely. For each metabolic pathway, the predicted concentrations of its members are plotted (based on the Kyoto Encyclopedia of Genes and Genomes).

proteomics in this unique system and had to develop the genome-free methods presented in this study out of necessity. Although this was intended as a preliminary effort, we were surprised by how well the approach worked, especially because we can compare it now to the early gene models. Ultimately, a high-quality genome with well-annotated gene models will likely provide the highest-quality reference set possible for RNA and protein analysis. However, reference sets based on mRNA are much cheaper and faster to obtain than gene models from genomic data. Based on this study, we now believe that mRNA-derived proteomic data could assist in building gene models that are more accurate by using identified peptide sequences to confirm exons. Recent studies suggest that even for model organisms with well-annotated genomes (e.g., rat or mouse), utilizing gene models based on RNA-seq evidence increases the information that can be gained from proteomic experiments [41, 42]. Furthermore, the relative quality of gene models, generated with different parameters, could be evaluated and potentially improved by utilizing the number of identified peptides from a proteomic experiment as a benchmark.

The proteomic data from the *X. laevis* egg illustrate the feasibility of genome-free proteomics, which can be extended to any nonstandard organism. One advantage of our methods

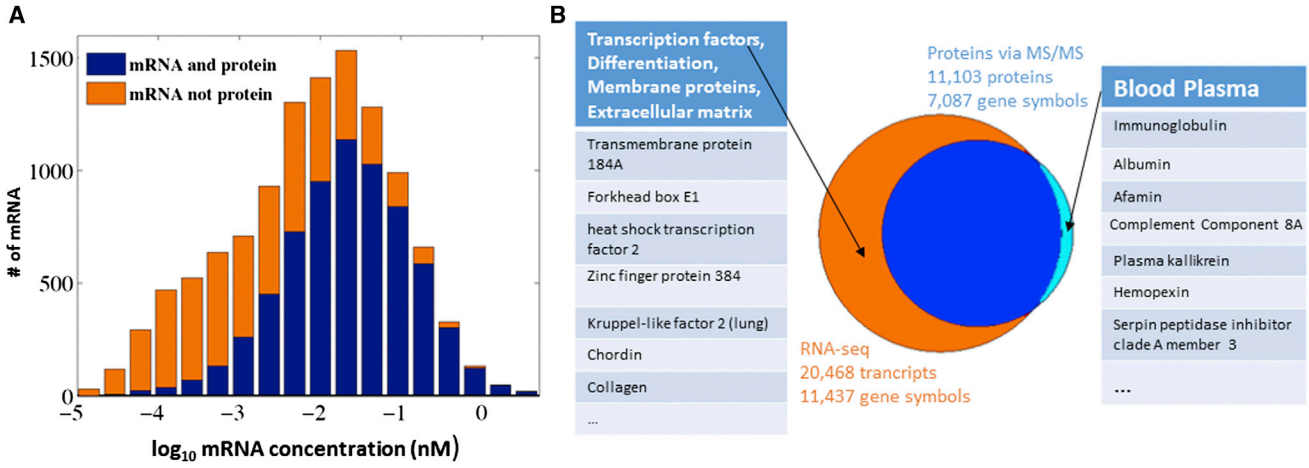

Figure 5. mRNA and Protein Abundance

(A) Histogram of mRNA levels in the egg. mRNA for which the protein was also detected is colored blue. Orange indicates that only mRNA was detected. The median of mRNA concentration is approximately 1,000-fold lower than the median for protein abundance. Although we see only a weak correlation between mRNA and protein abundance (0.32 Pearson correlation), the lower the mRNA concentration, the less likely we are to detect the corresponding protein.

(B) mRNA and protein were matched via assigned gene symbols. MS is able to identify approximately 60% of all gene symbols for which we could detect mRNA. The proteins that we cannot detect via MS are overrepresented by transcription factors, proteins involved in differentiation, and transmembrane proteins. On the contrary, for ~350 gene symbols, we could identify only proteins, but not mRNA. This group is highly enriched for blood plasma and liver proteins and was likely endocytosed during oocyte growth.

is that mRNA data can be combined from heterogeneous sources. For many species, multiple expressed sequence tags and some full-length sequence information are available (e.g., http://compbio.dfci.harvard.edu/tgi/tgipage.html) [24]. However, only relying on our RNA-seq data, genome-free proteomics is possible. Approximately 10% of unique peptide data were lost by only using RNA-seq data from embryonic development; however, based on the findings in this paper, one could likely minimize this loss by adding mRNA data from the adult liver. We have integrated our series of pipeline scripts into an online resource that creates a high-quality protein reference database from heterogeneous mRNA sources, and that resource can be found at http://kirschner.med.harvard.edu/tools/mz_ref_db.html.

Amino acid sequence information alone is not very informative. Rather, one needs to integrate that information with previous knowledge of proteins and their functions, e.g., which proteins bind to form a complex, which proteins are part of a metabolic pathway, or simply what the protein's name from which one can access the literature is. For nonstandard model organisms, it is unlikely that there is much previous knowledge of proteins from that species. However, by relating sequence similarity to human proteins, one can assign proteins to gene symbols and then interpret protein levels for development. One unexpected finding in *Xenopus* is that many proteins, which could be identified by MS, had no observable mRNA in the egg. We found that these were almost certainly proteins produced in the liver and endocytosed from blood. We also found mRNAs without protein, and this suggests that certain transcripts may be stockpiled in the egg for translation at later stages of development.

This study demonstrates the power of genome-free proteomics, and our online tool increases the scope of proteomic experiments. Knowledge of the level of protein expression can offer new insight into molecular regulation and provides a valuable resource for both biochemical and developmental work in *Xenopus*.

Experimental Procedures

Sample Preparation for MS

The research with *X. laevis* was performed under the oversight of the Harvard Medical Area Institutional Animal Care and Use Committee. Female *X. laevis* were induced with 700 U HCG. After 14 hr, eggs were harvested, washed with 1× MMR, and dejellied with Cysteine (2% w/v) (pH 8.0). Sixty eggs were flash frozen with liquid nitrogen. Eggs were lysed with 250 mM sucrose, 1% NP40 substitute (Sigma), 5mM EDTA (pH 7.2), 1 Roche complete mini tablet (EDTA-free), 20 mM HEPES (pH 7.2), 10 μM Combretastatin 4A, and 10 μM Cyochalasin D. For lysis, eggs were vortexed at maximum speed for 10 s, pipetted ten times up and down with a 200 μL pipette tip, incubated on ice for 10 min, and again vortexed for 10 s. Lysates were clarified by centrifugation at 4,500 RCF at 4°C for 4 min in a tabletop centrifuge. The cytoplasmic and lipid layers were mixed by gentle flicking and removed from the pelleted yolk. To the lysate, HEPES (pH 7.2) was added to 100 mM, and SDS was added to 2% (w/v). The sample was reduced with 5 mM DTT for 20 min at 60°C and then alkylated with 15 mM NEM for 20 min at room temperature (RT). Excess NEM was reacted with an additional 5 mM DTT at RT. Proteins were isolated by methanol-chloroform precipitation [43]. The protein pellet was resuspended (~5 mg/mL) in 6 M Guanidine HCl in 50 mM HEPES (pH 8.5) and sonicated for 5 min. The sample was diluted to 2 M Guanidine with 50 mM HEPES (pH 8.5) and digested with LysC (Wako Chemicals) at 20 ng/μL at RT for 14 hr. Next, we diluted Guanidine HCl with 50 mM HEPES (pH 8.5) to 0.5 M and digested further with 10 ng/μL of sequencing grade trypsin (Roche) at 37°C for 8 hr or LysC at an additional 20 ng/μL at RT. Samples were subjected to C18 solid-phase extraction (SPE) (SepPak, Waters) to desalt and isolate peptides. To reduce sample complexity, ~1mg LysC peptides and 0.5 mg LysC/trypsin peptides were resuspended in a 10 mM sodium carbonate buffer (pH 8.0) and then fractionated by medium pH reverse-phase HPLC (Zorbax 300Extend-C18, 4.6 mm × 250 mm column, Agilent) using an Acetonitrile gradient from 6%–31%. With a flow rate of 0.8 mL/min, fractions were collected into a 96-well plate every 38 s and then pooled into 24 fractions by combining alternating wells from each column of the plate. Each fraction was dried and resuspended in 20 μL of 1% phosphoric acid. Peptides from each fraction were desalted and extracted once more with reverse-phase purification [44], resuspended in 10 μL 1% formic acid. Approximately 4 μL per fraction was analyzed by LC-MS.

Estimation of Protein Concentration

Published protein concentrations were collected from the literature (Table S3). To obtain the MS1 ion current, we divided the MS1 precursor peptide

intensities by the corresponding noise value (Thermo raw file). This signal to noise ratio is a proxy for the number of charges in an Orbitrap analyzer [45]. To convert charges into ion current, we divided by the MS1 ion-injection time. For each PSM, we recorded the maximum ion current during a peptide's elution. These ion currents were summed for all PSMs that matched a protein [25] and normalized to the number of theoretically calculated tryptic plus LysC peptides, with at least 7 amino acids and at most 25 amino acids (missed cleavages were not allowed for theoretical peptides). The published protein name was searched on the Human Genome Organisation gene name database to assign gene names (http://www.genenames.org/). If multiple proteins that had been matched with the same gene symbol were found in the MS data set, their MS1 ion currents were summed. On occasion, multiple gene symbols were combined. For complete description of which gene symbols were combined and for further assumptions required for converting published values into cytoplasmic concentrations, see Table S3.

PHROG Final Build

X. laevis transcripts from GenBank, *X. laevis* Gene Indices version 11 [24, 46], and the de novo assemblies from the wild-type and J line RNA-seq data were combined (ensuring unique identifiers), cleaned and trimmed using SeqClean (http://compbio.dfci.harvard.edu/tgi/software/), and masked for common repeat motifs using RepeatMasker (http://www.repeatmasker.org) with its default libraries. The cleaned sequences were clustered with TGICL [46], using default parameters (93% identity) but requiring a 100 bp overlap, and assembled using CAP3 [47] with default parameters (92% identity). The contigs and singletons were searched against a small database of model chordate proteins (*H. sapiens*, *M. musculus*, *G. gallus*, *D. rerio*, *X. tropicalis*, and *X. laevis*) using BLASTX [48], and the full BLASTX reports were parsed for strand, translation frame, expectation (E) value, bit score, and alignment coordinates of both query and subject. Before translation, the parsed data were processed to select transcripts that show possible frame shifts, as determined by translation frames of the high-scoring pairs (HSPs); the sequences of such transcripts were adjusted to compensate for and to retain the translation frame of the best HSP. All transcripts (corrected and not corrected) that showed conserved alignments ($E \leq 1 \times 10^{-5}$) were fully translated, without regard to the best open reading frame (ORF), in the hinted frame; those above this E value were discarded. The translated proteins were subsequently processed as follows: (1) the longest peptide from the full translation was retained; (2) protein ends were trimmed to reflect potential trypsin-digested peptides; and (3) any resulting protein fragments <7 amino acids were discarded. Finally, the remaining proteins were processed by CD-HIT [49], with a threshold of 100%, to collapse the group into a nonredundant data set. The alternative references were generated as follows: we performed the six-frame translation of the PHROG according to Evans et al. [11] by using the transcripts after TGICL/CAP3 clustering and assembly but prior to any filtering and/or trimming. We performed the HMM-based translation of PHROG on the same transcripts by using TransDecoder from the Trinity suite, translating on the positive strand only with a minimum size of 24 amino acids. The best-guess translation was performed using Virtual Ribosome [50], using parameters to translate on any strand and return the longest ORF. All translations were also processed by CD-HIT with a threshold of 100% [49].

Resources

The scripts and short protocol for usage, the protein database generation pipeline, and the PHROG FASTA file are available as online resources at http://kirschner.med.harvard.edu/tools/mz_ref_db.html.

Accession Numbers

The MS proteomics data have been deposited to the ProteomeXchange Consortium [51] via the PRIDE partner repository with the data set identifier PXD000926.

Supplemental Information

Supplemental Information includes Supplemental Experimental Procedures, four figures, and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2014.05.044.

Author Contributions

M.W., R.M.F., M.P., M.E.H., L.P., S.P.G., and M.W.K. designed the experiments. M.W. and M.P. performed the experiments. M.W., R.M.F., and L.P. analyzed the data. M.W., R.M.F., M.P., M.E.H., L.P., S.P.G., and M.W.K wrote the manuscript.

Acknowledgements

We would like to thank R. Harland, D. Rokhsar, and the *X. laevis* genome consortium for making the preliminary gene models of *X. laevis* available. We thank Ramin Rad for his programming help and the RITG team for their HPC assistance. Thanks to Woong Kim, Robert Everley, and Joao Paulo for help with mass spectrometers and to the S.P.G. computer room for bioinformatics support. This work was supported by NIH grants R01GM103785, R01HD073104, P40OD010997, and R01DK077197. We thank MBL for their support of this project.

Received: March 17, 2014
Revised: May 1, 2014
Accepted: May 19, 2014
Published: June 19, 2014

References

1. Huttlin, E.L., Jedrychowski, M.P., Elias, J.E., Goswami, T., Rad, R., Beausoleil, S.A., Villén, J., Haas, W., Sowa, M.E., and Gygi, S.P. (2010). A tissue-specific atlas of mouse protein phosphorylation and expression. Cell *143*, 1174–1189.
2. Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol. Syst. Biol. *7*, 548.
3. Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., and Aebersold, R. (2011). The quantitative proteome of a human cell line. Mol. Syst. Biol. *7*, 549.
4. Kragl, M., Knapp, D., Nacu, E., Khattak, S., Maden, M., Epperlein, H.H., and Tanaka, E.M. (2009). Cells keep a memory of their tissue origin during axolotl limb regeneration. Nature *460*, 60–65.
5. di Prisco, G., Cocca, E., Parker, S., and Detrich, H. (2002). Tracking the evolutionary loss of hemoglobin expression by the white-blooded Antarctic icefishes. Gene *295*, 185–191.
6. Abzhanov, A., Protas, M., Grant, B.R., Grant, P.R., and Tabin, C.J. (2004). Bmp4 and morphological variation of beaks in Darwin's finches. Science *305*, 1462–1465.
7. Thomson, K.S. (1972). An attempt to reconstruct evolutionary changes in the cellular DNA content of lungfish. J. Exp. Zool. *180*, 363–371.
8. Straus, N.A. (1971). Comparative DNA renaturation kinetics in amphibians. Proc. Natl. Acad. Sci. USA *68*, 799–802.
9. McGrath, C.L., and Katz, L.A. (2004). Genome diversity in microbial eukaryotes. Trends Ecol. Evol. *19*, 32–38.
10. Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.
11. Evans, V.C., Barker, G., Heesom, K.J., Fan, J., Bessant, C., and Matthews, D.A. (2012). De novo derivation of proteomes from transcriptomes for transcript and protein identification. Nat. Methods *9*, 1207–1211.
12. Looso, M., Borchardt, T., Krüger, M., and Braun, T. (2010). Advanced identification of proteins in uncharacterized proteomes by pulsed in vivo stable isotope labeling-based mass spectrometry. Mol. Cell. Proteomics *9*, 1157–1166.
13. Newport, J., and Kirschner, M. (1982). A major developmental transition in early Xenopus embryos: II. Control of the onset of transcription. Cell *30*, 687–696.
14. Desai, A., Murray, A., Mitchison, T.J., and Walczak, C.E. (1999). The use of Xenopus egg extracts to study mitotic spindle assembly and function in vitro. Methods Cell Biol. *61*, 385–412.
15. Murray, A.W., and Kirschner, M.W. (1989). Cyclin synthesis drives the early embryonic cell cycle. Nature *339*, 275–280.
16. Wühr, M., Tan, E.S., Parker, S.K., Detrich, H.W., 3rd, and Mitchison, T.J. (2010). A model for cleavage plane determination in early amphibian and fish embryos. Curr. Biol. *20*, 2040–2045.
17. Hughes, M.K., and Hughes, A.L. (1993). Evolution of duplicate genes in a tetraploid animal, Xenopus laevis. Mol. Biol. Evol. *10*, 1360–1369.
18. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. Science *274*, 546–567, 563–567.

19. Elias, J.E., and Gygi, S.P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods *4*, 207–214.
20. Wühr, M., Haas, W., McAlister, G.C., Peshkin, L., Rad, R., Kirschner, M.W., and Gygi, S.P. (2012). Accurate multiplexed proteomics at the MS2 level using the complement reporter ion cluster. Anal. Chem. *84*, 9214–9221.
21. Nesvizhskii, A.I., Keller, A., Kolker, E., and Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem. *75*, 4646–4658.
22. Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372.
23. Chvatal, V. (1979). A greedy heuristic for the set-covering problem. Math. Oper. Res. *4*, 233–235.
24. Quackenbush, J., Liang, F., Holt, I., Pertea, G., and Upton, J. (2000). The TIGR gene indices: reconstruction and representation of expressed gene sequences. Nucleic Acids Res. *28*, 141–145.
25. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.
26. Mills, A.D., Laskey, R.A., Black, P., and De Robertis, E.M. (1980). An acidic protein which assembles nucleosomes in vitro is the most abundant protein in Xenopus oocyte nuclei. J. Mol. Biol. *139*, 561–568.
27. Huang, C.Y., and Ferrell, J.E., Jr. (1996). Ultrasensitivity in the mitogen-activated protein kinase cascade. Proc. Natl. Acad. Sci. USA *93*, 10078–10083.
28. Lee, E., Salic, A., Krüger, R., Heinrich, R., and Kirschner, M.W. (2003). The roles of APC and Axin derived from experimental and theoretical analysis of the Wnt pathway. PLoS Biol. *1*, E10.
29. Lawo, S., Bashkurov, M., Mullin, M., Ferreria, M.G., Kittler, R., Habermann, B., Tagliaferro, A., Poser, I., Hutchins, J.R., Hegemann, B., et al. (2009). HAUS, the 8-subunit human Augmin complex, regulates centrosome and spindle integrity. Curr. Biol. *19*, 816–826.
30. Zhang, Z., Yang, J., Kong, E.H., Chao, W.C., Morris, E.P., da Fonseca, P.C., and Barford, D. (2013). Recombinant expression, reconstitution and structure of human anaphase-promoting complex (APC/C). Biochem. J. *449*, 365–371.
31. Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.W. (2010). CORUM: the comprehensive resource of mammalian protein complexes—2009. Nucleic Acids Res. *38* (Database issue), D497–D501.
32. Yanai, I., Peshkin, L., Jorgensen, P., and Kirschner, M.W. (2011). Mapping gene expression in two Xenopus species: evolutionary constraints and developmental flexibility. Dev. Cell *20*, 483–496.
33. Tian, Q., Stepaniants, S.B., Mao, M., Weng, L., Feetham, M.C., Doyle, M.J., Yi, E.C., Dai, H., Thorsson, V., Eng, J., et al. (2004). Integrated genomic and proteomic analyses of gene expression in Mammalian cells. Mol. Cell. Proteomics *3*, 960–969.
34. Wang, J., Duncan, D., Shi, Z., and Zhang, B. (2013). WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res. *41* (Web Server issue), W77–W83.
35. Opresko, L.K., and Karpf, R.A. (1987). Specific proteolysis regulates fusion between endocytic compartments in Xenopus oocytes. Cell *51*, 557–568.
36. Opresko, L., Wiley, H.S., and Wallace, R.A. (1980). Differential postendocytotic compartmentation in Xenopus oocytes is mediated by a specifically bound ligand. Cell *22*, 47–57.
37. Wallace, R.A., and Jared, D.W. (1969). Studies on amphibian yolk. 8. The estrogen-induced hepatic synthesis of a serum lipophosphoprotein and its selective uptake by the ovary and trasformation into yolk platelet proteins in Xenopus laevis. Dev. Biol. *19*, 498–526.
38. Arike, L., Valgepea, K., Peil, L., Nahku, R., Adamberg, K., and Vilu, R. (2012). Comparison and applications of label-free absolute proteome quantification methods on Escherichia coli. J. Proteomics *75*, 5437–5448.
39. Vogel, C., and Marcotte, E.M. (2012). Label-free protein quantitation using weighted spectral counting. Methods Mol. Biol. *893*, 321–341.
40. Fusaro, V.A., Mani, D.R., Mesirov, J.P., and Carr, S.A. (2009). Prediction of high-responding peptides for targeted protein assays by mass spectrometry. Nat. Biotechnol. *27*, 190–198.
41. Low, T.Y., van Heesch, S., van den Toorn, H., Giansanti, P., Cristobal, A., Toonen, P., Schafer, S., Hübner, N., van Breukelen, B., Mohammed, S., et al. (2013). Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. Cell Rep *5*, 1469–1478.
42. Menschaert, G., Van Criekinge, W., Notelaers, T., Koch, A., Crappé, J., Gevaert, K., and Van Damme, P. (2013). Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. Mol. Cell. Proteomics *12*, 1780–1790.
43. Wessel, D., and Flügge, U.I. (1984). A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. Anal. Biochem. *138*, 141–143.
44. Lohse, M.M., Bolger, A.M.A., Nagel, A.A., Fernie, A.R.A., Lunn, J.E.J., Stitt, M.M., and Usadel, B.B. (2012). RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res. *40* (Web Server issue), W622–W627.
45. Makarov, A., and Denisov, E. (2009). Dynamics of ions of intact proteins in the Orbitrap mass analyzer. J. Am. Soc. Mass Spectrom. *20*, 1486–1495.
46. Pertea, G.G., Huang, X.X., Liang, F.F., Antonescu, V.V., Sultana, R.R., Karamycheva, S.S., Lee, Y.Y., White, J.J., Cheung, F.F., Parvizi, B.B., et al. (2003). TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics *19*, 651–652.
47. Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Res. *9*, 868–877.
48. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.
49. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659.
50. Wernersson, R. (2006). Virtual Ribosome—a comprehensive DNA translation tool with support for integration of sequence feature annotation. Nucleic Acids Res. *34* (Web Server issue), W385–W388.
51. Vizcaíno, J.A., Deutsch, E.W., Wang, R., Csordas, A., Reisinger, F., Ríos, D., Dianes, J.A., Sun, Z., Farrah, T., Bandeira, N., et al. (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat. Biotechnol. *32*, 223–226.