Published in final edited form as:
Dev Cell. 2015 November 9; 35(3): 383–394. doi:10.1016/j.devcel.2015.10.010.

On the relationship of protein and mRNA dynamics in vertebrate embryonic development

Leonid Peshkin[*,1], **Martin Wühr**[*,1,2], **Esther Pearl**[3], **Wilhelm Haas**[2], **Robert M. Freeman Jr.**[1], **John C. Gerhart**[4], **Allon M. Klein**[1], **Marko Horb**[3], **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]National Xenopus Resource, Marine Biological Laboratory, Woods Hole, MA 02543, USA

[4]Department of Molecular and Cell Biology, University of California, Berkeley, CA 96704, USA

Summary

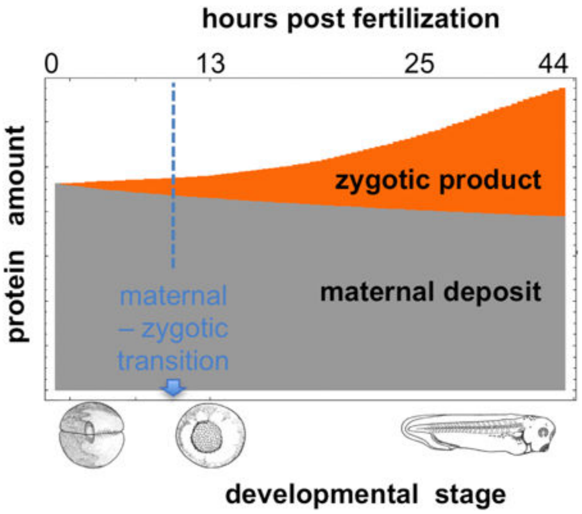
A biochemical explanation of development from the fertilized egg to the adult requires an understanding of the proteins and RNAs expressed over time during embryogenesis. We present a comprehensive characterization of protein and mRNA dynamics across early development in *Xenopus*. Surprisingly, we find that most protein levels change little and duplicated genes are expressed similarly. While the correlation between protein and mRNA levels is poor, a mass action kinetics model parameterized using protein synthesis and degradation rates regresses protein dynamics to RNA dynamics, corrected for initial protein concentration. This study provides detailed data for absolute levels of ∼10K proteins and ∼28K transcripts via a convenient web portal, a rich resource for developmental biologists. It underscores the lasting impact of maternal dowry, finds surprisingly few cases where degradation alone drives a change in protein level, and highlights the importance of transcription in shaping the dynamics of the embryonic proteome.

Graphical abstract

Correspondence: marc@hms.harvard.edu, steven_gygi@hms.harvard.edu.

Author Contributions: Conceptualization: L.P., M.W., M.W.K.; Methodology: M.W., M.W.K.; Software: L.P. B.F.; Formal Analysis: L.P., M.W, J.G., A.M.K.; Resources: E.P., M.H., W.H., S.G.; Data Curation: L.P., E.P., B.F., J.G.; Writing: L.P., M.W., M.W.K.; Supervision: M.H., M.W.K.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



Introduction

Embryonic development had been traditionally described in anatomical terms, tracing organs and structures to reveal lineages and explain morphogenesis. Recently such descriptions have been greatly augmented by RNA expression studies, revealing many molecular events where there were few anatomical markers (Struhl, 1981). When such data are coupled with genetic or pseudo-genetic manipulations, plausible pictures emerge of the regulatory circuits underlying developmental changes. Most recently there have been efforts to incorporate these data into mathematical models of developmental processes (Peter et al., 2012). Their limitation hinge on the difficulty of relating RNA levels directly to the phenotype. Protein is closer to the phenotype than RNA, but protein analysis methods are far less sensitive than those for RNA. Protein abundance may also not be the whole story: posttranslational modifications may provide crucial regulatory input. There are many examples where RNA level is misleading as a measure of protein function, e.g. cyclin proteins in the cell cycle or p53 in tumors. Whether many other misleading examples occur in the embryo is not known. Information on the RNA/protein relationship is generally unavailable at the genome/proteome scale.

Fortunately, methods now allow low mRNA levels to be detected and quantitated accurately by RNA-seq, and specific RNAs to be localized by single-molecule FISH. Although protein methods are more complex, difficult, and expensive, and less sensitive, the relative abundance of proteins in the bulk embryo can also be measured using multiplexed approaches. Major unappreciated pitfalls in the first applications of multiplexed mass spectrometry (MS) have now been circumvented by new analysis methods (McAlister et al., 2014; Wühr et al., 2012). Nevertheless serious limitations in applying these techniques to embryos remain. A single sample requires about 50 μg of protein, which would represent ~1000 mouse embryos. Accurately determining the kinetics of RNA accumulation requires synchronized embryonic samples. Several non-traditional systems are naturally synchronized, but MS methods require a well-curated reference set of protein sequence information, which is often unavailable. Finally, highly abundant proteins like serum or yolk must be removed without depleting other proteins. The *Xenopus* system addresses all these issues: single embryos have about 25 μg of non-yolk protein and *in vitro* fertilization yields very accurate synchrony (Gurdon and Wickens, 1983; Wühr et al., 2014). A good reference genome has recently been generated for *Xenopus* (Bowes et al., 2010), and we now have highly reproducible protocols for efficient removal of yolk while sparing other cellular components. For many years, *Xenopus* was the model of choice for early development in vertebrate species with many experimental results and conceptual findings, generalizable to all vertebrate embryos. Previous attempts at proteomic characterization of Xenopus embryonic development suffered from inferior accuracy of the MS2 methods and covered fewer proteins than we report, at fewer time points (Sun et al., 2014). An initial effort to compare RNA and protein levels found disagreement but provided no satisfactory explanation (Smits et al., 2014).

Experimental embryology has provided extraordinary insight, but little understanding on the biochemical level. Physiological features of embryos were emphasized in the pre-molecular biology era (Brachet, 1950), but have not been explored with modern methods. In this first publication we offer a survey of the economy of the egg and embryo, something not achievable until the elaboration of genome-wide methods. This broad perspective can already be used to inform our understanding of the many biochemical changes underlying embryonic development. In many species including the frog the earliest stages of development proceed without new transcription; suggesting that the control of protein behavior might proceed through unmasking of RNA for translation, or degradation or posttranslational modification of existing proteins. After the mid-blastula transition (8000 cell stage in the frog) transcription is turned on (Newport and Kirschner, 1982), and it has been suggested that the original maternal proteins might rapidly turn over at this point (Howe et al., 1995). Possible hypotheses about the protein economy range from all proteins synthesized on demand, at the right time and location, to stockpiling of all proteins in the egg followed by rearrangement and/or degradation of proteins that are in the wrong place. Using our quantitative time-resolved inventory of RNA and protein, we have developed a picture of the overall strategies used by the egg and embryo. We also provide a deep dataset of individual stories of proteins and RNA that can now be woven, by us and others, into narratives that can help elucidate of development.

Results

Genome-wide measurements of RNA and protein levels across key developmental stages

We profiled developmental stages (Nieuwkoop and Faber, 1994) spanning early development from unfertilized egg (NF 0) through blastula (NF 5 -- 9), gastrula (NF 10 -- 12.5), neurula (NF 13 -- 21) and tailbud. Stage NF 23 is characterized by presence of blood islands and first appearance of olfactory placodes. The last time point (NF 33) is taken when heartbeat has started and the tadpole is ready to hatch. Our processing pipeline for quantitatively measuring levels of RNA and protein is sketched in Figure 1A. Proteins were digested into peptides and change of abundance was measured by isobaric labeling followed by MultiNotch MS3 analysis (McAlister et al., 2014); absolute protein abundance was estimated via MS1 ion-current (Schwanhäusser et al., 2011; Wühr et al., 2014). mRNA levels were measured across eighteen time points starting from the unfertilized egg to stage 33, while protein abundance levels were measured at six key stages (NF 2, 5, 9, 12, 23, 33). RNA level was further measured in two distinct ways: polyadenylated RNA enrichment and ribosomal RNA depletion. mRNA was extracted using standard protocols with bacterial sequence spike-ins for quality control and normalization. Our primary dataset is comprised of 27877 mRNA profiles and 6509 protein profiles, which overlap 6435 gene products (Fig. 1B). The overlap is reduced to 5960 if we use only peptides that uniquely match to a single predicted protein. In addition, we reanalyzed our published (Wühr et al., 2014) egg protein data against the present reference set, resulting in concentration (nM) data for 9728 proteins (Table S1). This is fewer than in the original publication because here we only used unique peptides. Based on overall abundance distribution (Fig. S1A) we estimate that proteins missing from our data are typically present at <10nM.

mRNA measurements are consistent with those previously published—We compared the mRNA time series reported here with a microarray study across 14 stages previously validated and published by us (Yanai et al., 2011). A total of 7806 transcripts were matched between the microarray and the RNA-Seq datasets. The median Pearson correlation coefficient among these transcripts is 0.89 and the median cosine distance is 0.026 (a measure of similarity where zero is coincident and 1 is the most discordant), which suggests confidently reproduced expression profiles. The left panel of Fig. 1C presents a histogram of cosine distances between previously published mRNA abundance time courses and those measured in this study. The right panel provides examples of genes at different levels of agreement: chordin (CHRD: 0.06, near median), tenascin (TNN: 0.008, in lowest 5%) and secernin (SCRN2: 0.3, highest 5%). As we previously showed, some of the discordance in biological repeats is explained by heterochronic developmental timing, i.e. genes which preserve the general expression pattern but show a shift in the onset of expression among different clutches of the same species (Yanai et al., 2011).

Protein measurements are also reliable—We compiled previously published Western blots for 35 proteins displaying distinct patterns during the course of development and compared these to the quantitative data obtained by mass spectrometry. Overall, our data agree very well with established information on protein dynamics (Fig. S1C, Table S1). Figure 1D shows a histogram of cosine distances between previously published protein abundance changes and changes measured in this study for those proteins (left); examples of three protein dynamics patterns quantified via Western blot (solid) and multiplexed proteomics (- -) with representative cosine distances (right). The corresponding protein distances are color-coded. Three examples are shown: ITLN1 (red) which agrees very well between the two methods and has a cosine value of 0.005; LIN28A (magenta), which is at the median cosine distance (0.03); and XNF7 (khaki), which shows the lowest level of agreement between the two methods (cosine distance 0.15).

Absolute abundance of mRNA and protein—In addition to relative changes, we estimated absolute mRNA concentration by dividing the total messenger RNA abundance in the embryo proportionally to FPKM counts. We estimate protein concentration based on MS1 ion current prorated to the isobarically labeled fractions (Schwanhäusser et al., 2011, Wühr et al., 2014). The Pearson correlation between previously published protein concentration and normalized ion-current is 0.92 (Fig. S1B).

Allo-alleles at the protein and mRNA level show no sign of sub-functionalization

The whole *X. laevis* genome was duplicated about 50MYa -- as a result, many genes have a close paralog referred to as the 'homeolog' or 'allo-allele'. Single gene duplications, as well as whole genome duplications have a special place in evolutionary theory, where it is asserted that they provide a way for new functions to arise through subfunctionalization (Barton, 2007; Force et al., 1999). We have compared protein expression patterns across 164 pairs of homeologs obtained from Xenbase (Bowes et al., 2010), for which the expression comparison is possible thanks to unique peptides in each sequence (Fig. 2A). Figure 2B shows a typical example of a pair of allo-alleles of gene DAPL1. Protein is shown in green and mRNA in blue. There is remarkable concordance across homeologs: the median Pearson correlation is 0.94, the median cosine distance is .006. We selected peptides that are both unique and differ by only a single amino-acid across the homeologs (see Fig. 2A, Table S1). Based on 90 such paired peptides, we again obtain exceptional agreement in expression across homeologs with a median Pearson correlation of 0.92 – see the histogram of cosine distances for 164 protein pairs (Fig. 2C.left) and 630 mRNA pairs (Fig. 2C.right) where a colored arrow shows the position of DAPL1. Gray histograms show the baseline distribution obtained by randomly matching pairs of allo-alleles to one another. In this representative set of allo-alleles there is no evidence for sub-functionalization. This apparent redundancy in conjunction with the dosage difference (Fig. 2D) is consistent with observations in other systems of similar time-scale (Dean et al., 2008).

Abundant proteins are stockpiled rather than produced on demand

Most developmental studies have focused on genes expressed at different times, places and circumstances. What is not clear is whether these are exceptional cases or whether embryos are constantly changing the mix of proteins in the embryo. In *X. laevis* there is little new protein synthesis from fertilization up to neurulation(Lee et al., 1984). Overall protein synthesis does not change appreciably throughout these periods and remains at approximately 100 ± 20 (sd.) ng per hour or about 0.4% per hour of the total non-yolk protein content. Based on these mesurements in 24 hours, at most an additional 9% of protein could be synthesized Proteins that appear stable throughout our experiment are therefore likely to be made early and not degraded, rather than maintaining a constant level through high production rates and high turnover. Nevertheless, bulk measurements bias the interpretation toward the most abundant proteins. MS analysis allows us to see which proteins are stable and which are dynamic. Figure 3A presents nine main temporal trends of relative protein abundance via the medians of clusters (K-means clustering using cosine distance, also see Fig. S2A). The thickness of the median line reflects the number of proteins that fall into the respective cluster. The two largest clusters (together 3215 or ~54%) contain proteins whose abundances are essentially flat. Except for one dynamic red cluster, all trends are either induction or degradation; the more dynamic the trend, the fewer proteins fall into that category.

Many proteins change little in abundance during development from the egg to hatching tadpole. To quantitate this behavior we computed a parameter we call "**dynamicity**", **Δ**. For each pattern, **Δ** is defined as the cosine distance between a flat line and the abundance curve. **Δ** is 0 for flat proteins and increases with more active dynamics. Using the value of protein abundance discussed above we analyzed **Δ** as a function of abundance. Figure 3B shows a histogram of **Δ** for detected proteins. As is evident from this histogram, most proteins do not change much within the surveyed period. The insert for Fig. 3B presents four examples: transportin, the flattest observed (TNPO2: **Δ**=1.0e-04); ribosomal protein RPL11, at the median of the distribution (**Δ**=0.017), which represents less than 1 degree between vectors; an isoform of hemoglobin zeta (HBZ), one of the most dynamic proteins (**Δ**=0.57; ~35 degrees between vectors) and Oncomodulin (OCM2) which shows the same pattern.

Dynamicity decreases with abundance—Figure 3C shows a density plot of protein absolute abundance against **Δ**. This plot illustrates that high abundance proteins are generally flat, while low abundance proteins are mostly dynamic. Black circles show two abundant isoforms of TPI1 (Triosephosphate Isomerase 1) (1 and 5 μM) which are very flat proteins (**Δ**=0.002 and 0.004). Red circles show positions of two very low abundance isoforms of OCM2 (a calmodulin family gene) (6 - 10 nM) which are very dynamic (**Δ** between 0.51 and 0.57) and have identical patterns of accumulation. In particular, of proteins whose abundance is less than 100nM, 75% have a dynamicity over 0.1. The Spearman correlation between abundance and dynamicity is -0.55. We further confirm this trend by subdividing the proteins into ten quantile bins by concentration in $\log_{10}$ scale and plotting the mean dynamicity in each bin against the concentration (Fig. S3B), there is a clear monotonic trend. To ensure that this trend is not an artifact of measuring the abundant protein levels via many constituent peptides, while rare proteins are often measured via only a single peptide we present the same plot using only one randomly chosen peptide for each protein -- the result is very similar (Fig. S3B).

Specific examples—The general pattern, whereby abundant proteins show very little change throughout development into the hatching stage, makes intuitive sense in terms of the general function of these proteins. Metabolic enzymes are one group of abundant and flat proteins, e.g. complete sets of enzymes are present in the egg for glycolysis, TCA cycle, and fatty acid metabolism. These abundant enzymes remain at about the same level throughout early development. There is no indication that the formation of tissues of high metabolic demand, such as muscle and nerve perturbs the pervasive constancy of the levels of enzymes for central metabolism. A few metabolic enzymes with tissue-specific isoforms, such as the brain isoform of aldolase (ALDOC) and the liver isoform of carnitine palmitoyl transferase (CPT1A), are expressed dynamically once the respective cell types are generated (Fig. S2B). Only a small fraction of abundant proteins is gradually degraded throughout gastrulation and neurulation. They represent a group composed largely of liver-specific proteins found in the oocyte with no measurable mRNA counterpart (Wühr et al., 2014). They are likely endocytosed from the bloodstream, along with the yolk protein, vitellogenin, and gradually degraded. Liver proteins such as albumin may have no function in the oocyte; hence, it is not be surprising that they are degraded and not resynthesized in early development. However, other proteins like glycogen phosphorylase have homologs that are found in every cell type. The homologs behave as expected: the abundance of the endogenous protein is lower (muscle PYGM and brain PYGB at 0.3 μM and 1.3 μM) than the putatively endocytosed protein (liver PYGL 16.1 μM). It would be interesting to know how the degradation machinery eliminates specifically the endocytosed protein. Finally, some of the most dynamic proteins are transcription factors, such as NFKBIA (**Δ**=0.29) and two isoforms of Y-box protein YBX1 (**Δ** of 0.28 and .30).

Tissue-specific proteins are typically produced on demand

The elaboration of complex tissues is expected to be accompanied by changes in the levels of proteins that pre-existed in the egg and by the synthesis of new proteins. We would expect tissue specific proteins to be synthesized as the embryo reaches the stages where there is frank expression of a suite of proteins characteristic of that tissue type. Indeed, such examples of the tissue specific proteins are present: HAL (Histidine ammonia-lyase) has three isoforms, one of which is predominantly (88% of total) present in stage 33 and is known to be predominantly expressed in fetal liver. Neurogenesis genes are exemplified by such genes as FABP7 (Fatty acid binding protein 7) and OCM2 (Oncomodulin). HBZ (Zeta-globin) is a polypeptide first synthesized in the yolk sac of the early embryo. In order to analyze how tissue-specific gene expression is distributed in embryogenesis, we introduced a tissue specificity index **T** which ranges between 0 (nonspecific) and 1 (highly specific e.g. rhodopsins). This index is based on the Gini index, which has been widely used in economics for assessing income distribution in a population and has also been used in biology for kinase specificity (Gujral et al., 2014). Tissue specific expression data is not available for *Xenopus*. Instead we have used the data available for 96 tissues and cell types in mouse, grouping together similar tissues e.g. different neuronal tissues. Figure 4A shows the histogram of tissue specificity over all proteins we find in *Xenopus* embryos with khaki and magenta areas showing the lowest and the highest 25% quantile. The proteins in this lowest quartile and the highest quartile are chosen to represent nonspecific and tissue-specific genes respectively, without regard to which particular tissue. We further clustered all temporal patterns of protein change using the cosine distance measure to see how tissue specificity depends on temporal pattern. Figure 4B shows the two most populated clusters: a flat cluster of 1260 proteins and a temporal increasing cluster of 140. Each cluster is labeled with a fraction N/S representing how many (N)onspecific and tissue (S)pecific proteins are found in each. There is a clear bias (Fisher's test P-value 1e-4) towards nonspecific proteins in the flat cluster and towards tissue specific proteins in the dynamic cluster.

We again find some tissue specific proteins in the egg that could be best explained as having been endocytosed with vitellogenin (Wühr et al., 2014). As examples of other tissue-specific proteins, we also see highly abundant epithelial keratins KRT8 (7 μM) and KRT19 (5 μM), long before the appearance of differentiated epithelial cells. Importantly, we do not find any other tissue specific intermediate filament proteins that are abundant in differentiated tissues: neurofilament protein (L, M, N), desmin, peripherin, and internexin; nor do we find other widely accepted neuronal markers, such as TAU or MAP2.

Evidence for dynamic post translational modification in early development

We made no special effort to examine posttranslational modification, yet we found about a thousand spectra for modified peptides (Table S1). Two special cases we briefly consider are phosphorylation and acetylation. Eight proteins (e.g. aldolase and nucleoplasmin) show both types of modifications. A specific search for phospho-peptides quantifies 731 spectra corresponding to about 225 proteins. Figure S4 shows the result of K-means clustering into nine clusters. These clusters are not mutually exclusive, since for many genes some peptides are de-phosphorylated while others are phosphorylated. E.g. nucleolin has peptides in both 1^{st} and 7^{th} clusters, while nucleoplasmin (NPM2) in 7^{st} and 9^{nd} clusters. One dramatic pattern is rapid de-phosphorylation between fertilization and pre-MBT – cluster 7. Two key groups stand out among genes with dynamic phosphorylation patterns. First, 23 genes involved in splicing machinery ACIN1, CWC27, CD2BP2, CLNS1A, GEMIN5, DHX16, KHSRP, NSRP1, PABPN1, PAPOLA, PRPF3, RBM25, SF1, SF3A1, SRSF4, SRSF11, SRRM1, TFIP11, THRAP3, TCERG1, SLU7, ZCCHC8 and ZRANB. Second, 15 genes located in nucleoli and involved in ribosomal biogenesis namely: ANP32A, DKC1, ESF1, KRI1, NOLC1, NOP58, NPM1, NPM2, NUCKS1, NSUN2, RRP12, LYAR, TOP2A, UTP18 and nucleolin (NCL). We also observed acetylated peptides. There is a dramatic change in acetylation of Lys27 in histone H3 around the MBT, when transcription starts, consistent with studies of histone acetylation in the regulation of transcription (Stasevich et al., 2014). Acetylation is known to regulate metabolism in glycolysis, fatty acid synthesis, urea cycle, and TCA cycle (Zhao et al., 2010). We find four enzymes in glycolysis that each show a major acetylation increase at NF23, while their protein abundances show no significant change. These proteins are present at micromolar concentrations (ALDOA, ALDOC, LDHb and PGK1 at over 5 μM). Phosphoglucomutase is known to be positively regulated by acetylation in the C-terminus and we see a C-terminal acetylation, suggesting activation. A more detailed study will require enrichment for peptides harboring such modifications.

The correspondence between mRNA and protein in the developing embryo

Generally, mRNA abundance is a poor predictor of protein abundance in the embryo—

We find that when analyzed stage-by-stage, mRNA concentrations typically only modestly correlate with respective protein concentrations. This observation is consistent with previous publications in bacteria, yeast and human cell culture (Smits et al., 2014; Vogel and Marcotte, 2012). In these studies agreement is quantified as rank correlation between abundance of mRNA and protein in a given sample. When calculated this way stage-by-stage separately for six stages for which we measured both mRNA and protein (Fig. 5A), the median Spearman correlation for each stage is modest, with values similar to these previously reported for somatic cells (0.42). Poly-(A) enriched mRNA shows worse agreement with protein than ribo-depleted (Fig. 5A) in the early stages, which are known to have a lot of poly-A elongation and shortening, while the agreement at later stages is somewhat better. These results correspond to our intuition of what might be expected, since translational efficiency is affected not only by the total level of mRNA message, but also by polyadenylation status.

The dynamics of protein and RNA are very poorly correlated—An alternative test of agreement between mRNA and protein is to look at the correlation of changes across time points. In agreement with previously published results, we found mRNA-protein correlation to be poor in a majority of cases: the mean Pearson correlation coefficient between mRNA and respective protein time series is close to zero (0.2) (See Fig 4B). To exemplify extremes in the correlation histogram we provide individual examples of mRNA and protein (Fig 4C). One explanation for the general lack of correlation is given by examining the ratio of concentration between mRNA and respective protein. The rate of protein synthesis is limited by the amount of mRNA available; when the level of RNA is very low relative to the level of protein, fluctuations in mRNA lead to small changes in translation rates that have little impact on the protein level. Confirming this general trend, when we divided all genes into ten bins according to the mRNA/protein ratio (Fig. S3A) we observed that genes that show higher mRNA/protein ratios have better agreement between mRNA and protein dynamics.

The correspondence of RNA to protein for dynamic proteins—Coarse co-clustering mRNA and protein patterns into 3-by-3 matrix reveals the mutual information (Fig. 5D). Generally protein dynamics across development can be coarsely classified into three categories: those that stay flat, those that disappear, and those that accumulate. The more dynamic the protein pattern (left to right in Fig. 5D), the better the agreement between protein and mRNA patterns. However this mutual information does not suggest a simple temporal correlation. The process of protein synthesis takes time, so protein would be expected to be synthesized and accumulated after a delay relative to mRNA synthesis. In a related group of cases mRNA levels spike and fade away, while protein is accumulated (as in Fig. 6A). We hypothesize that a few of the truly anti-correlated patterns (when mRNA is gradually disappearing as protein levels are increasing at and after the MBT) are due to packaging of RNA in granules, such as P-bodies (Hogan et al., 2008). Several RNA-binding proteins are in this group: RBM7, RBM27, LARP1B, LARP7, KIN, NOL12, YTHDF1. These observations suggest the hypothesis that when mRNA granules break up mRNA simultaneously becomes available for translation and degradation, leading to the paradoxical behavior of RNA decline and protein accumulation. Further, selecting 720 genes where protein is dynamic and some maternal mRNA is present we found no cases where pre-MBT and post-MBT translation rates (estimated as the ratio of protein increment over the mRNA level) are sufficiently different to suggest mRNA “masking” i.e. translational control. Overall, though mRNA and protein dynamics typically correlate poorly, one can still gain information about likely protein behavior from mRNA dynamics, and vice versa.

Mass-action kinetics equation links RNA changes to protein changes

Although mRNA and protein dynamics poorly correlate, they clearly contain mutual information. To test how well we can predict protein dynamics for a given mRNA dynamics, we modeled embryonic protein turnover using mass-action kinetics. Under simple assumptions of temporal and spatial invariance of synthesis and degradation, the expected change in protein levels over time is given by $\frac{dp}{dt}=\mathrm{K}_{\mathrm{S}}\mathrm{r}(\mathrm{t})-\mathrm{K}_{\mathrm{D}}\mathrm{p}(\mathrm{t})$, where p(t) is the amount (moles) of protein per embryo, K_S is the translation rate (mole per mole per hour) for protein at time t, r(t) is the amount of mRNA for the transcript encoding that protein, K_D

is the decay rate (hour $^{-1}$) of the protein. For each protein, the parameters K_S and K_D can be fit to the measurements of r(t) and p(t) subject to the initial concentration fixed at p_0 so as to minimize the difference between the observed protein level p_i at time t_i and the predicted protein level $p(t_i, K_S, K_D)$ on average over all observed time points *i* (see Fig. 6A):

$$\min_{\{\mathrm{K}_S, \mathrm{K}_D, p_0\}} \sum\nolimits_i \{p(t_i|\mathrm{K}_S, \mathrm{K}_D, p_0) - \hat{p}_i\}^2$$

To prevent over-fitting we also consider simplified models with no synthesis $(\mathrm{K_S}=0)$: $\frac{dp}{dt}=\mathrm{K_D}\,\mathrm{p(t)}$; no degradation $(\mathrm{K_D}=0)$: $\frac{dp}{dt}=\mathrm{K_S}\,\mathrm{r(t)}$; and a degenerate model $\frac{dp}{dt}=0$, selecting the best model according to a Bayesian information criterion. The optimization search results in so-called MLE (maximum likelihood estimate) values for parameters and also estimates confidence intervals for these parameters (Supp Meth).

A non-linear model of protein dynamics with up to three parameters fits most of the data—Most of the protein patterns fit the model well. The goodness of fit is characterized by the cosine distance between the measured and the predicted pattern as well as by adjusted R^2 (Fig. S5A). Consider one sample protein Calpain-8 (CAPN8) presented in Fig. 6A. The beige stripe shows 95% confidence band for protein dynamics which corresponds to the 95% confidence range in synthesis and degradation rates. The no-degradation model (K_D = 0) is selected for this gene, synthesis rate is estimated at a maximum constrained value of K_S = 1200. Compared to the baseline model of predicting protein from mRNA, we improve the Pearson correlation from .469 to 0.999. The cosine distance between actual quantitative protein measurements marked via green discs and predicted protein level shown by the continuous green curve is 0.0028 (c.f. 0.38 for mRNA – protein) -- about 50% of all protein patterns are fit better than that. Adjusted R^2 for this fit is .70 which is worse than about 75% of all fits.

There are several limitations to this approach. Naturally, the accuracy with which we assign half-lives is limited by the observation period of our experiments i.e. ~50 hours. A flat protein is easily explained by setting an initial protein concentration at the right level, then assuming a zero degradation rate and a zero synthesis rate, which simply disregards the mRNA profile. That trivial model was selected for about 24% of the genes all of which were not fit well ($R^2 < .7$). However we get the most information out of dynamic rather than static protein patterns. About 18% of the protein patterns needed a complete three–parameter model. Another 15% ignored protein synthesis and only assigned a degradation rate, while remaining 43% assumed negligible degradation and only used synthesis rate (Fig. 6B).

When we look at the complete collection of RNA and protein measurements, 80% of all well-fit ($R^2 > .7$) models use synthesis to explain the protein pattern, and three quarters of these do not use the degradation rate. Moreover for about 60% of all proteins the half-life is estimated to be longer than the duration of our experiment, suggesting that protein levels during this early period are largely controlled by protein synthesis rather than degradation.

This suggests that there is one broad class of proteins which are deposited in the egg and do not need to be localized, while tissue specific proteins are localized by the means of mRNA localization or spatially defined mRNA expression and subsequent protein synthesis. Finally, there are very few genes for which the protein pattern is dynamic but not regressed to mRNA via our simple model (median **Δ**=0.004 for the degenerate model $\frac{dp}{dt}=0$ genes, c.f. **Δ**=0.072 for genes explained by full $\frac{dp}{dt}=\mathrm{K_S\,r(t)-K_D\,p(t)}$ model) suggesting that additional layers of translational regulation, or disjoint protein-mRNA localization are not very common or not significant in the early embryo.

The distribution of synthesis and degradation rates is physiologically plausible—As a result of fitting the model to the data we obtain a wide range of synthesis and degradation rates spanning four orders of magnitude. Fig. 6C shows histograms of half-life and synthesis rate, where the half-life is given in hours, while synthesis rate in molecules of protein synthesized per molecule of mRNA per hour (see Table S1). The observed distribution is biologically plausible. In particular it resembles the similar distribution obtained for mouse cell culture using metabolic labeling (Schwanhäusser et al., 2011). We observe a median half-life of 43 hours (over non-zero estimates) and median synthesis of 213 molecules of protein per molecule of mRNA per hour (*c.f.* 40h and 140m/m/h for mammalian cell culture, and synthesis rates for sea urchin of 120 m/m/h at 15°C (Ben-Tabou de-Leon and Davidson, 2009)). The long median half-life indicates that most proteins are very stable during the period of our time series. There is a general trend that rapidly synthesized proteins have shorter half-life (Fig. S5B) with a rank correlation of -0.7. The 995 short lived proteins (lower 25%) are strongly enriched (43 genes, multiple hypothesis adjusted P-value of 3e^{-20}) for the cell cycle genes such as CHEK1, GMNN, kinases PLK1, TLK1, CHEK1, AURKB and DNA-binding (52 genes, P-value 5e^{-15}). The long-lived slow turnover proteins (2209 estimated half life over 50 h) include proteins such as metabolic enzymes and tubulins and are strongly enriched (90 genes, adjusted P-value of 1e^{-29}) for mitochondrial proteins (121 genes P-value 5e^{-71}) such as ATP synthases (e.g. ATP5J, J2, C1, A1) and NADH dehydrogenases (e.g. NDUF A3, A6, A9, B9).

mRNA dynamics can be used to predict protein dynamics—Having shown that three parameter models can typically encode protein dynamics, we next asked if we can predict protein dynamics throughout development given the mRNA profile and the initial protein concentration in the egg. As a proof-of-principle we employ a simplified predictor, which uses the median rates of synthesis and degradation for all proteins. We use this model to forecast the protein expression for the genes where we had measured the protein patterns and compared the predicted and measured patterns. The predictive power of this model is best measured by a cosine distance between predicted and measured temporal pattern of protein expression (Fig. S5C). Figure 6D provides a histogram of Pearson correlation for model-based vs measured protein expression for a model assuming median synthesis and median degradation rates while using the actual initial concentration. The median correlation of 0.72 is a striking improvement over simply using the mRNA dynamics as described above (Fig. 5B), which gave a correlation of 0.24. Furthermore, the mRNA dynamics pattern can be used to improve the prediction power. For example the median synthesis rate for proteins whose mRNA is broadly degraded (see bottom mRNA cluster of Fig. 5D) is only 17 m/m/h, while for proteins whose mRNA is in the top cluster (sharply induced) the median is 287 m/m/h. By conditioning the synthesis rate on mRNA pattern category we improved the modeling accuracy to .84 which could likely be further improved by considering more than three categories.

Note that our method immediately allows us to make predictions for over 3000 additional proteins that could not be detected in the developmental series but whose concentration in the egg was measured. For all remaining genes that were not detected in the egg we can assume the low expected 1nM concentration and still apply our method (see Supplemental Data). Yet another application of our approach is to predict the protein dynamics in a part of the embryo if spatially resolved information on mRNA levels is available (Junker et al., 2014). If the fraction of the embryonic volume where a given gene is expressed can be estimated from e.g. in-situ hybridization, the projected protein dynamics can be adjusted by pro-rating both the initial protein level and the mRNA expression level. Such predictions would be valuable for planning morpholino and RNAi experiments (Heasman et al., 2000). The ability to calculate protein levels will be especially important for classes of genes that are expressed at low levels and are hard to detect in MS measurements, such as transcription factors, receptors and secreted signaling molecules.

Embryonic protein economy expressed as gradual replacement of maternal by zygotic protein—As the embryo develops, maternally deposited proteins are degraded and replaced by zygotic products. For each individual protein where synthesis and degradation rates were recovered from modeling, turnover dynamics can be obtained by solving a simple system of equations where at each point total protein concentration is factored into two components, maternal and zygotic:

$$\frac{\mathrm{dp}_N}{\mathrm{dt}}=K_S r(t) - K_D p_N(t);$$
$$\frac{\mathrm{dp}_M}{\mathrm{dt}}=-K_D p_M(t);$$

subject to boundary conditions: $p_N(t_0) = 0$; $p_M(t_0) = p_0$; where p_0 is the concentration of zygotic protein product (which includes protein translated from maternal mRNA stored in the egg); p_M is concentration of maternal protein product. This system can be solved for p_M(t) and p_N(t) for each gene. The turnover can now be integrated over all proteins fitted by our model and extrapolated to the whole embryo, as illustrated in Fig. 7. This turnover analysis suggests that (not counting yolk) most of the protein composition of a complex highly differentiated organism 50 hours after fertilization was originally provided to it via maternal deposit. Rather than synthesizing most of the building material from scratch, degrading or secreting a lot of material, the embryo makes careful use of what is provided maternally. Presumably some of that protein is simply stockpiled until it is useful, raising the question of exactly how the stockpiled protein is maintained in an inactive state and prevented from premature degradation. One well-studied example is yolk, which is stored in granules; most of the maternal yolk supply persists through the period of our experiment (Jorgensen et al., 2009). Other proteins may similarly be compartmentalized, or maintained in an inactive state via post-translational modification. The question of how this is achieved, and how needed protein is eventually released, opens up a number of research directions, including research on positional signaling and shuttling mechanisms.

Discussion

In situ hybridization (Gall and Pardue, 1969) enabled the revolutionary developments of Drosophila genetics to be applied at the molecular level. Together with other techniques, such as RTPCR and microarray analysis, we have a deeper understanding of vertebrate and invertebrate development. Yet, it is still uncertain how closely mRNA changes correlate with the activation of specific developmental processes. To address this, we need to take on the daunting task of measuring both protein expression and posttranslational modification. The optimal system in which to do this is *Xenopus*, where synchrony is easy to achieve and each egg has sufficient protein for deep analysis.

We have focused here on protein dynamics in the early embryo and on comparing the dynamics of proteins to their mRNAs in the early development of frog eggs from fertilization to just before hatching (stage NF33). We find that two kinds of protein patterns dominate the early embryo: a stable set of maternally inherited proteins, many of them abundant; and a very dynamic set of lower abundance proteins, which most often strongly track with RNA levels. Transcription factors are an example of this latter class. Such proteins are characterized by rapid synthesis changes driven by transcription and rapid protein degradation. We were able to track proteins that show dramatic changes in the post-translational modifications, namely phosphorylation and acetylation. We have made all this data available in an easily accessible browser.

There are of course inherent limitations to our interpretations. Bulk measurements limit us in ascribing changes to specific regions of embryos. Relative protein quantitation is limited to 6509 gene products; the total number of detected proteins is slightly greater, about 7000. By comparison to our own single sample proteomics (11,300 proteins in the egg (Wühr et al., 2014)) we know that much depth is left to be explored. This difference is mostly due to the significant increase in the duty cycle of the Multinotch MS3 method which we employed for accurate multiplexed quantification as compared to the label-free MS2 approach we used previously. In the present study RNA quantitation goes roughly three times deeper than protein. Despite these limitations, the work presented here is by far the deepest known exploration of relative protein changes in embryogenesis.

Our data allowed us to resolve the apparent conflict between protein measurements and mRNA measurements, using a simple model for expression kinetics that assumes that the observed median rates of protein synthesis and degradation apply to all proteins. Given the initial protein level and mRNA kinetics, we can then make an accurate prediction of protein levels throughout development. The excellent agreement between model and experiment indicates that the spread around the mean values for protein synthesis and degradation mostly represents measurement error. Using this model, appropriate localized mRNA measurements could allow tissue- and region-specific protein dynamics to be calculated, aiding in the interpretation of morpholino experiments that could be confounded by the presence of lingering maternal protein. However any specific protein might have an atypical rate of synthesis (per mole of RNA) or degradation. The most drastic modifications of the composition of the egg take place in the least abundant proteins. By stage 33 about 85% of the less abundant proteins are newly synthesized, as compared to under 30% for the most abundant proteins. Much of the change closely tracks RNA expression and appears to be driven by transcription, rather than translational control.

We designed our study of RNA around its intersection with protein data, and we have therefore focused solely on coding sequence. We have not yet attempted to distinguish among splice variants. The rich dataset we have made available can now be used to investigate these issues. Our rather limited study of PTMs could also be greatly expanded by enriching for modified peptides with antibody precipitation or chromatography.

Our protein data represent generally the most abundant genes with coverage down to about the 10 nM range. This level of analysis offers insight into the general strategy of protein regulation during development from egg to hatching. The unfertilized egg is provisioned with many materials that are maintained without much loss up to the feeding tadpole stage. As yolk is not consumed until after gastrulation (Jorgensen et al., 2009; Vastag et al., 2011), the protein complement of the embryonic cells must be similar to that in the earliest cleavage divisions. The non-yolk protein made before the tailbud stage is very small compared to the non-yolk endowment from the egg. To change its protein composition the embryo must thus either transcribe and translate new genes or degrade or modify old proteins, but new transcription is very rare before the MBT. Our data show that many proteins remain virtually unchanged until the beating heart stage (2 days of development, corresponding to about 10 days of mouse development). In many cases an unchanging protein level stands in contrast to large excursions of individual mRNA levels that have no known consequence, raising the question of whether these are gratuitous and simply not selected against (Gerhart and Kirschner, 1997). It is not known whether these RNAs are being translated, whether proteins are being degraded at a rate that would compensate for their synthesis, or whether transcription is highly localized. Unless we hypothesize precise cytoplasmic localization of proteins in the egg or the existence of intercellular shuttling mechanisms, we must assume that many protein deposits end up mis-placed and are slowly degraded and diluted without much effect.

There is an extensive literature speculating that in the embryo, with its rapid nuclear proliferation and small nuclear to cytoplasm ratio, extensive protein regulation is occurring at the level of translational and protein degradation. We found virtually no convincing examples of this, outside the cell cycle. Although we can predict most protein data accurately from mRNA levels, the outliers in our analysis may be the most interesting, providing information on unusual types of translational control. Our work relieves many concerns about the reported discordance of RNA and protein expression seen in many publications, but provokes questions about which proteins are chosen to be maternal and which are chosen to be actively regulated by transcription. Finally, these studies should now focus our attention on protein modification as a probable source of the regulation of the many very stable proteins that are maternally provisioned and maintained throughout the early stages of development.

Experimental Procedures

Xenopus laevis J-line embryos were collected according to NF system (Nieuwkoop and Faber, 1994) at stages 0, 2, 6, 6.5, 7, 8, 8.5, 9, 10, 12, 14, 16, 18, 20, 23, 26, 30 and 33. Embryos were de-jellied in 2% cysteine, pH 7.8, and flash frozen for later preparation.

Total RNA was isolated using TRIzol. Two distinct rounds of RNA sequencing were performed: the first using poly(A) enrichment, the second using ribosomal RNA depletion. For both libraries: barcodes for multiplexing were added during the amplification PCR. Epicenter FailSafe PCR enzyme mix was used in the amplification step. Libraries were run on High Sensitivity DNA chips on the Bioanalyzer 1000. Size selection 350-600 bp was performed using the Pippin Prep automated electrophoresis system from Sage Science with 2% agarose cassettes. Samples were purified post size selection using MinElute columns and run again on High Sensitivity DNA chips on the Bioanalyzer.

Sequencing was performed on Illumina HiSeq-1000 instruments. Paired-end 100 bp reads from mRNA libraries were adapter/quality trimmed and filtered. Ribosomal reads were removed and the remaining high quality paired reads were aligned to the reference set using Bowtie (Langmead et al., 2009) with default parameters. RSEM package (Li and Dewey, 2011) was used to determine abundance estimates for all transcripts, and those transcripts having little read support were filtered out.

MS sample preparation and data-analysis was performed essentially as previously described (Wühr et al., 2015). Embryos were lysed and yolk removed via centrifugation (Wuhr et al. 2014). Proteins were purified via methanol chloroform extraction (Wessel and Flügge, 1984), digested with LysC and labeled with six-plex TMT. LC-MS experiments were performed on an Orbitrap Elite (Thermo Fischer Scientific) using the MultiNotch MS3 method (McAlister et al., 2014). For quantification we only used peptides that matched to only one protein in the reference database. For the quantification of each protein we used a weighted sum of TMT Signal/FT-Noise intensities of its assigned peptides.

For mapping of both mRNA and protein data (respectively the short sequences for RNA-Seq and peptide-Spectra matches for MS) we used as a main reference *X. laevis* genome assembly (DoE JGI REF; v6r1: a total of 43,013 sequences) downloaded from Xenbase (Bowes et al., 2010).

Statistical Analysis and Modeling

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them.

We estimate protein concentration based on MS1 ion current prorated to the isobarically labeled fractions (Wühr et al., 2014). We estimated absolute mRNA concentration by dividing the total messenger RNA abundance in the embryo proportionally to FPKM counts.

MATHEMATICA was used to fit protein synthesis and degradation rates using the respective mRNA and protein concentration data. To prevent over-fitting we used BIC to compare goodness of fit for alternative models. MATHEMATICA notebook which allows for interactive exploration of the model setting for any gene is available upon request.

Data—The mRNA and mass spectrometry proteomics data have been deposited to the GEO repository with the dataset identifiers GSE73905, GSE73870. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (Vizcaíno et al., 2014) via the PRIDE partner repository with the dataset identifier PXD002349. As part of this publication we provide a proteomic and transcriptomic data Web browser: http://kirschner.med.harvard.edu/MADX.html

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH grants **R01GM103785**, **R01HD073104**, P40OD010997, and R01DK077197. MW was supported by the Charles A. King Trust Postdoctoral Fellowship Program, Bank of America, N.A., Co-Trustee. AMK holds a Career Award at the Scientific Interface from Burroughs-Wellcome Fund and an Edward Mallinckrodt Jr. Foundation Grant. We thank the PRIDE team for proteomic data distribution.

References

Barton, NH. Evolution. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory Press; 2007. Gene Duplication and Divergence Allows for Functional Diversification without Loss of Previous Functions.

Bowes JB, Snyder KA, Segerdell E, Jarabek CJ, Azam K, Zorn AM, Vize PD. Xenbase: gene expression and improved integration. Nucleic Acids Res. 2010; 38:D607–D612. [PubMed: 19884130]

Brachet, J. Chemical embryology. Interscience Publishers; 1950.

Dean EJ, Davis JC, Davis RW, Petrov DA. Pervasive and persistent redundancy among duplicated genes in yeast. PLoS Genet. 2008; 4:e1000113. [PubMed: 18604285]

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 1999; 151:1531–1545. [PubMed: 10101175]

Gall JG, Pardue ML. Formation and detection of RNA-DNA hybrid molecules in cytological preparations. Proc Natl Acad Sci U S A. 1969; 63:378–383. [PubMed: 4895535]

Gerhart, J.; Kirschner, M. Cells, embryos, and evolution: toward a cellular and developmental understanding of phenotypic variation and evolutionary adaptability. Malden, Mass: Blackwell Science; 1997.

Gujral TS, Peshkin L, Kirschner MW. Exploiting polypharmacology for drug target deconvolution. Proc Natl Acad Sci U S A. 2014; 111:5048–5053. [PubMed: 24707051]

Gurdon JB, Wickens MP. The use of Xenopus oocytes for the expression of cloned genes. Methods Enzymol. 1983; 101:370–386. [PubMed: 6193395]

Heasman J, Kofron M, Wylie C. Beta-catenin signaling activity dissected in the early Xenopus embryo: a novel antisense approach. Dev Biol. 2000; 222:124–134. [PubMed: 10885751]

Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. PLoS Biol. 2008; 6:e255. [PubMed: 18959479]

Howe JA, Howell M, Hunt T, Newport JW. Identification of a developmental timer regulating the stability of embryonic cyclin A and a new somatic A-type cyclin at gastrulation. Genes Dev. 1995; 9:1164–1176. [PubMed: 7758942]

Jorgensen P, Steen JAJ, Steen H, Kirschner MW. The mechanism and pattern of yolk consumption provide insight into embryonic nutrition in Xenopus. Dev Camb Engl. 2009; 136:1539–1548.

Junker JP, Noël ES, Guryev V, Peterson KA, Shah G, Huisken J, McMahon AP, Berezikov E, Bakkers J, van Oudenaarden A. Genome-wide RNA Tomography in the zebrafish embryo. Cell. 2014; 159:662–675. [PubMed: 25417113]

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25. [PubMed: 19261174]

Lee G, Hynes R, Kirschner M. Temporal and spatial regulation of fibronectin in early Xenopus development. Cell. 1984; 36:729–740. [PubMed: 6697394]

Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011; 12:323. [PubMed: 21816040]

McAlister GC, Nusinow DP, Jedrychowski MP, Wühr M, Huttlin EL, Erickson BK, Rad R, Haas W, Gygi SP. MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. Anal Chem. 2014; 86:7150–7158. [PubMed: 24927332]

McGivern JV, Swaney DL, Coon JJ, Sheets MD. Toward defining the phosphoproteome of Xenopus laevis embryos. Dev Dyn Off Publ Am Assoc Anat. 2009; 238:1433–1443.

Newport J, Kirschner M. A major developmental transition in early Xenopus embryos: I. characterization and timing of cellular changes at the midblastula stage. Cell. 1982; 30:675–686. [PubMed: 6183003]

Nieuwkoop, PD.; Faber, J. Normal table of Xenopus laevis (Daudin): a systematical and chronological survey of the development from the fertilized egg till the end of metamorphosis. New York: Garland Pub; 1994.

Peter IS, Faure E, Davidson EH. Predictive computation of genomic logic processing functions in embryonic development. Proc Natl Acad Sci U S A. 2012; 109:16434–16442. [PubMed: 22927416]

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. Global quantification of mammalian gene expression control. Nature. 2011; 473:337–342. [PubMed: 21593866]

Smits AH, Lindeboom RGH, Perino M, van Heeringen SJ, Veenstra GJC, Vermeulen M. Global absolute quantification reveals tight regulation of protein expression in single Xenopus eggs. Nucleic Acids Res. 2014; 42:9880–9891. [PubMed: 25056316]

Stasevich TJ, Hayashi-Takanaka Y, Sato Y, Maehara K, Ohkawa Y, Sakata-Sogawa K, Tokunaga M, Nagase T, Nozaki N, McNally JG, et al. Regulation of RNA polymerase II activation by histone acetylation in single living cells. Nature. 2014; 516:272–275. [PubMed: 25252976]

Struhl G. A gene product required for correct initiation of segmental determination in Drosophila. Nature. 1981; 293:36–41. [PubMed: 7266657]

Sun L, Bertke MM, Champion MM, Zhu G, Huber PW, Dovichi NJ. Quantitative proteomics of Xenopus laevis embryos: expression kinetics of nearly 4000 proteins during early development. Sci Rep. 2014; 4:4365. [PubMed: 24626130]

Ben-Tabou de-Leon S, Davidson EH. Modeling the dynamics of transcriptional gene regulatory networks for animal development. Dev Biol. 2009; 325:317–328. [PubMed: 19028486]

Vastag L, Jorgensen P, Peshkin L, Wei R, Rabinowitz JD, Kirschner MW. Remodeling of the metabolome during early frog development. PloS One. 2011; 6:e16881. [PubMed: 21347444]

Vizcaíno JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Ríos D, Dianes JA, Sun Z, Farrah T, Bandeira N, et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol. 2014; 32:223–226. [PubMed: 24727771]

Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet. 2012; 13:227–232. [PubMed: 22411467]

Wessel D, Flügge UI. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. Anal Biochem. 1984; 138:141–143. [PubMed: 6731838]

Wühr M, Haas W, McAlister GC, Peshkin L, Rad R, Kirschner MW, Gygi SP. Accurate multiplexed proteomics at the MS2 level using the complement reporter ion cluster. Anal Chem. 2012; 84:9214–9221. [PubMed: 23098179]

Wühr M, Freeman RM, Presler M, Horb ME, Peshkin L, Gygi SP, Kirschner MW. Deep proteomics of the Xenopus laevis egg using an mRNA-derived reference database. Curr Biol CB. 2014; 24:1467–1475. [PubMed: 24954049]

Wühr M, Güttler T, Peshkin L, McAlister GC, Sonnett M, Ishihara K, Groen AC, Presler M, Erickson BK, Mitchison TJ, et al. The Nuclear Proteome of a Vertebrate. Curr Biol CB. 2015

Yanai I, Peshkin L, Jorgensen P, Kirschner MW. Mapping gene expression in two Xenopus species: evolutionary constraints and developmental flexibility. Dev Cell. 2011; 20:483–496. [PubMed: 21497761]

Zhao S, Xu W, Jiang W, Yu W, Lin Y, Zhang T, Yao J, Zhou L, Zeng Y, Li H, et al. Regulation of cellular metabolism by protein lysine acetylation. Science. 2010; 327:1000–1004. [PubMed: 20167786]

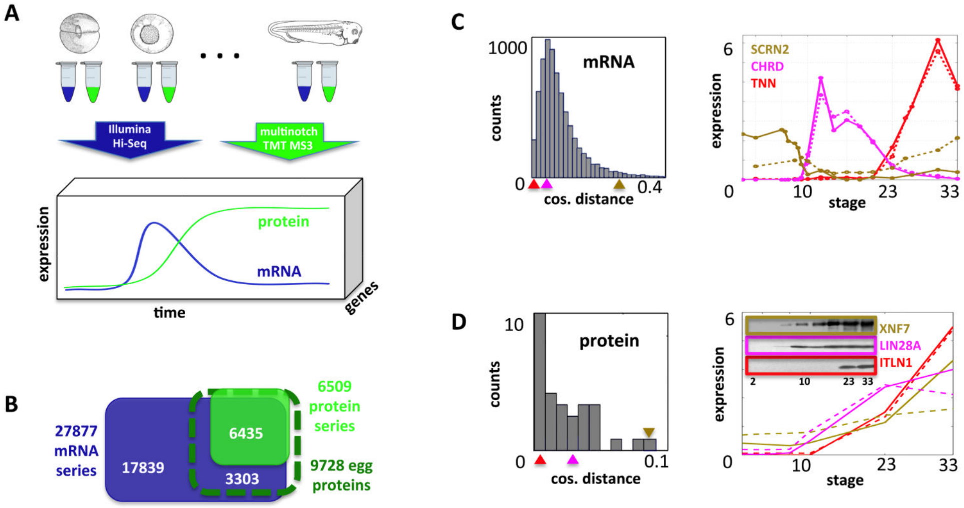
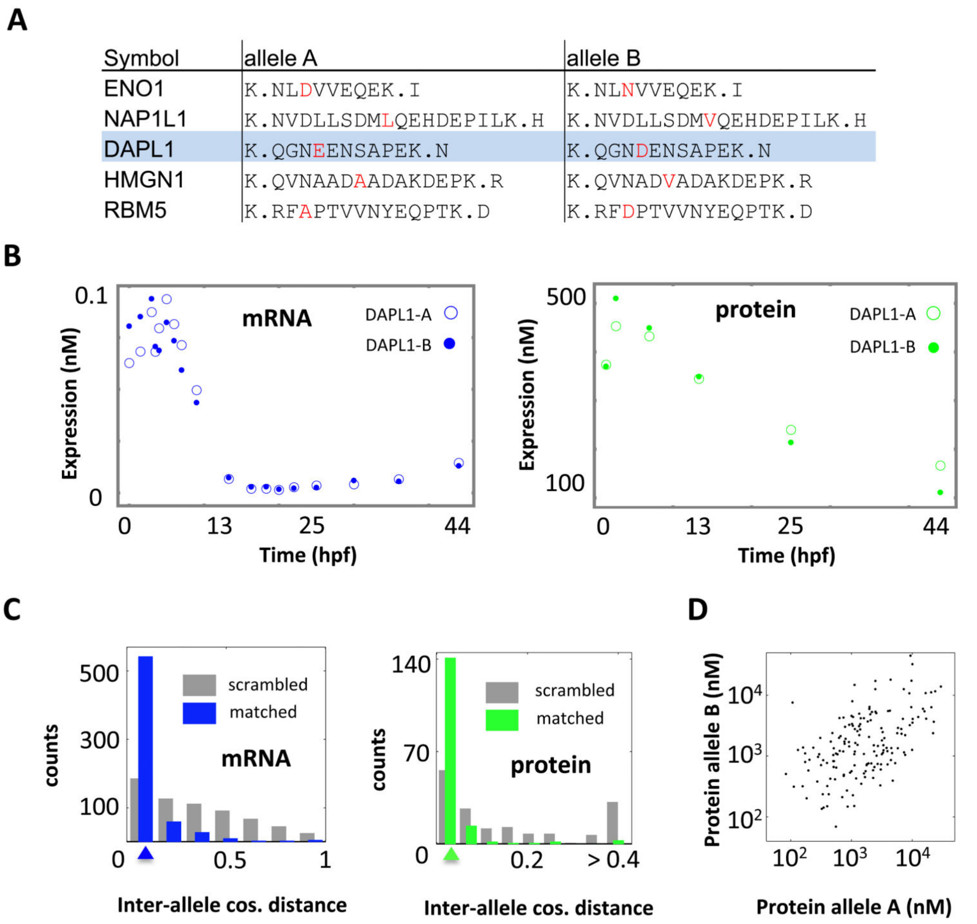


Figure 1. Early Embryonic Stages in *X. laevis*

(A) mRNA and protein were collected from various stages of development. **(B)** The dataset combines temporal profiles of 27877 mRNA and 6509 proteins and egg concentration data for 9728 proteins. **(C)** A histogram of ~8000 cosine distances between published and new mRNA profiles. Three sample mRNA profiles – Chordin, Tenascin N and Secernin are given as published (solid) and new RNA-Seq data (dashed). **(D)** A histogram of 35 cosine distances between published and new protein abundance changes. Three proteins quantified via Western blot (solid) and multiplexed proteomics (dashed) with representative cosine distances color coded.



Symbol	allele A	allele B
ENO1	K.NLDVVEQEK.I	K.NLNVVEQEK.I
NAP1L1	K.NVDLLSDMLQEHDEPILK.H	K.NVDLLSDMVQEHDEPILK.H
DAPL1	K.QGNEENSAPEK.N	K.QGNDENSAPEK.N
HMGN1	K.QVNAADAADAKDEPK.R	K.QVNADVADAKDEPK.R
RBM5	K.RFAPTVVNYEQPTK.D	K.RFDPTVVNYEQPTK.D

Figure 2. Allo-alleles are Concordant in both Protein and mRNA Expression

(A) Peptides with a single amino-acid difference (red) used to distinguish the allo-alleles. **(B)** mRNA and protein expression in allo-alleles of DAPL1. **(C)** Histogram of cosine distance over temporal expression in 164 allo-allele pairs of proteins (left) and 630 pairs of mRNA (right). Median cosine distances are 0.006 and 0.04 respectively. Median Pearson correlations are 0.94 and 0.85 respectively. Cosine distance between protein and mRNA pair of DAPL1 profiles is 0.004 and 0.03 respectively, exemplifying the median discordance as shown by colored triangle positions. Gray histograms show the baseline distribution obtained by randomly re-matching allo-alleles. **(D)** Scatter plot of cumulative protein concentration for allo-alleles. Overall rank correlation between allo-alleles is 0.50.

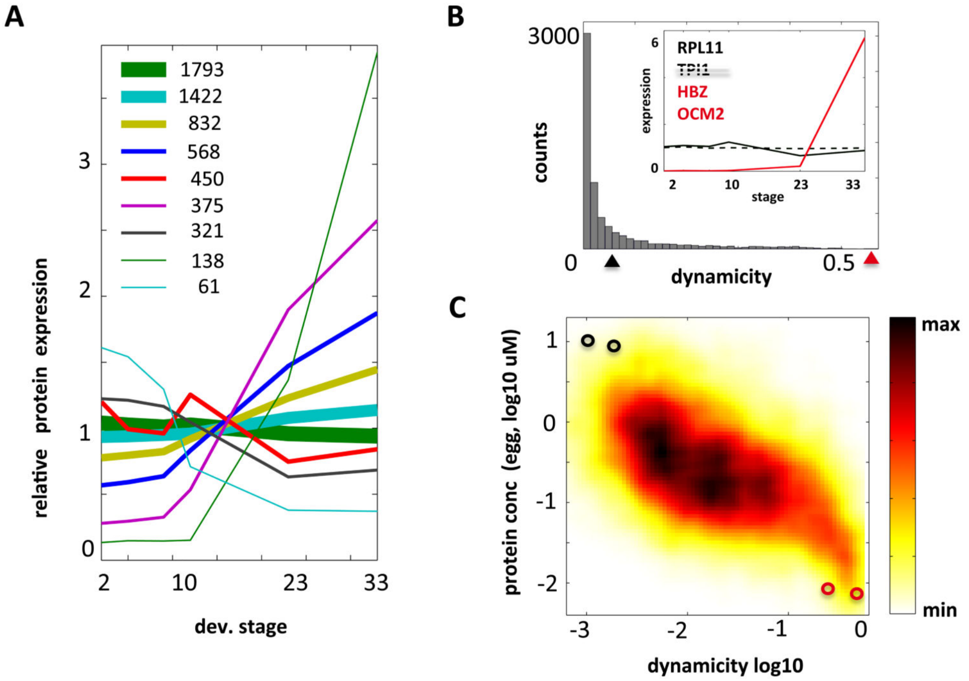


Figure 3. Most Proteins Change Little in Level from Egg through Tailbud Stages

(A) K-means clustering of relative protein abundance into nine clusters using cosine distance, labeled by the number of proteins which fall into each cluster represented by the median curve. Thickness of the median line reflects the number of proteins in the cluster. **(B)** Histogram of protein dynamicity shows that most proteins do not change much within the surveyed period. The insert shows representative examples: flat (gray dashed line) TPI1 the triosephosphate Isomerase 1 is among the flattest possible with **Δ** = 1.0e-04. RPL11 (black) is at the median of the dynamicity distribution (**Δ** = 0.0162; 1 degree difference. OCM2 (a calmodulin) and one of the isoforms of hemoglobin zeta (HBZ), a form of alpha globin produced in the yolk sack of mammals are among the most dynamic (**Δ** = 0.571; 35 degrees difference) proteins. Color code: red for dynamic, black for flat. **(C)** Highly abundant proteins are generally flat, while low abundance proteins are mostly dynamic. Density plot of protein absolute concentration in the egg against dynamism. Black circles show two TPI1 (triosephosphate isomerase 1) which are very highly abundant (1 and 5 μM in the egg) while flat (**Δ** = 0.002 and 0.004) and red circles show positions of OCM2 (a calmodulin) and HBZ (hemoglobin zeta)- very dynamic (**Δ** of 0.47 and 0.52) low abundance genes (2 - 20 pM in the egg).

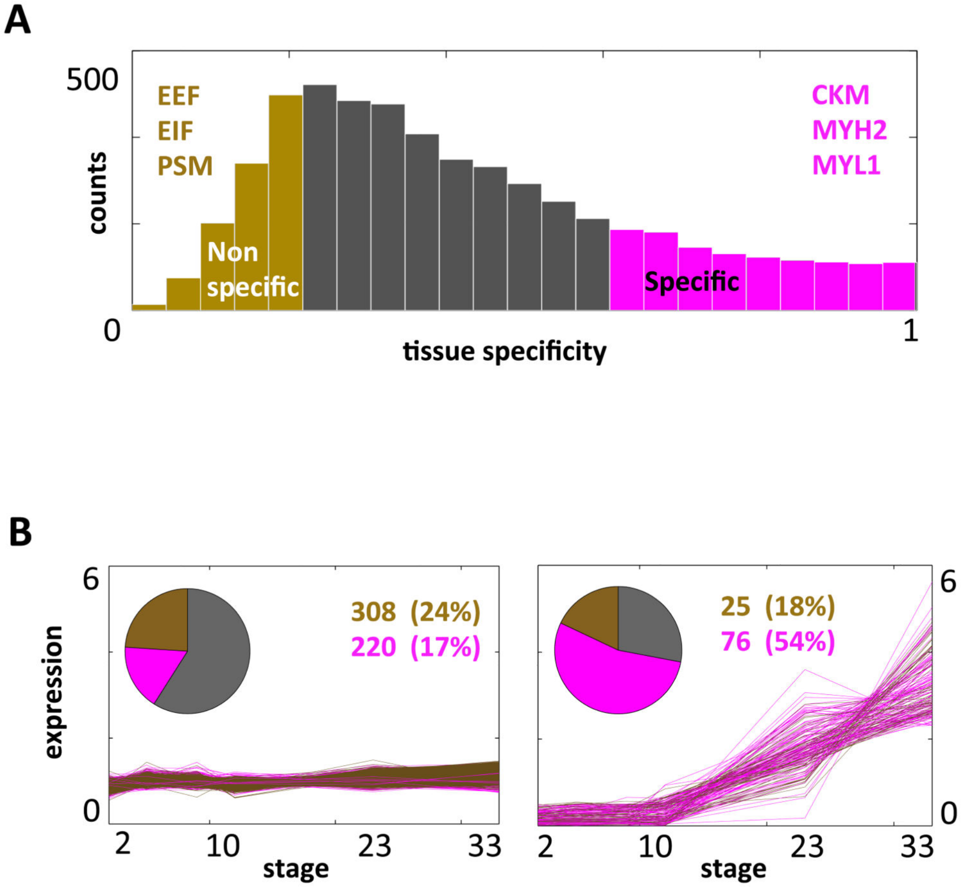


Figure 4. Temporal Expression of Tissue-specific proteins

(A) Histogram of tissue specificity over all measured proteins with the lowest and the highest 25% quantiles color-coded. Sample unspecific genes are elongation factors and proteosome while specific are myosin and creatine kinase. **(B)** Fraction of "non-Specific / Specific" proteins found in two most representative clusters.

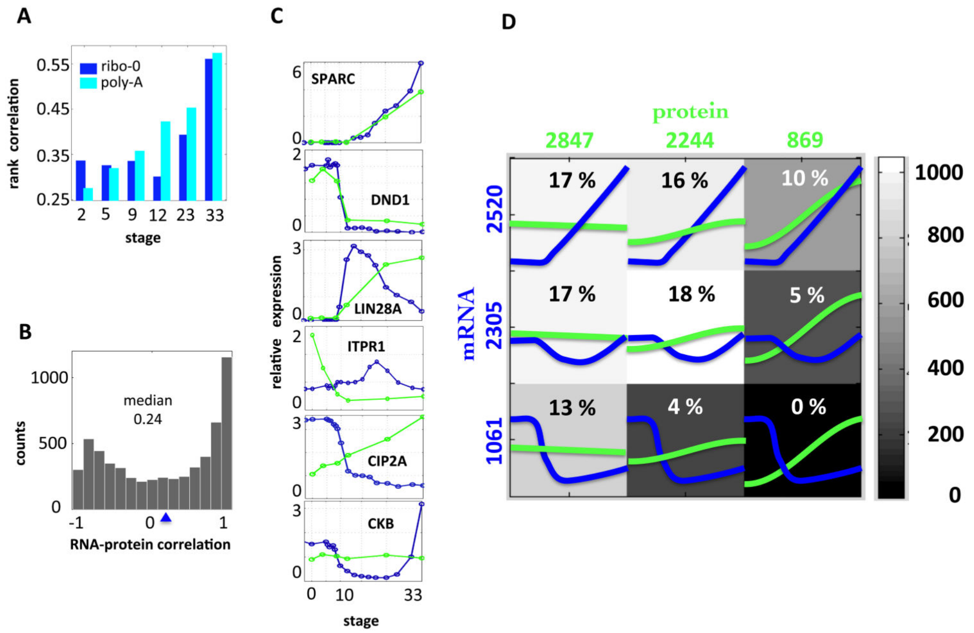


Figure 5. Discordance of Temporal Patterns in mRNA and Protein Expression

(A) Rank correlation (Spearman) within developmental stage between protein and mRNA temporal patterns for ribo-depleted and poly-A enriched methods of mRNA measurement.

(B) Histogram of Pearson correlation between protein and mRNA temporal change patterns.

(C) Exemplary mRNA/protein time series. The ordinates represent relative concentration of protein to mRNA. Each plot shows estimated absolute concentration of mRNA and protein.

(D) Mutual information between the temporal pattern of expression for mRNA and protein presented as co-clustering into three key trends. Grey scale background reflects the number of genes in each cluster. The left column illustrates that a flat protein pattern may correspond to any mRNA pattern, but if the protein is dynamic, it usually follows respective change in the mRNA concentration – see top of the right column for induction. Criss-cross patterns of anti-correlation are rarely observed – bottom of the right column.

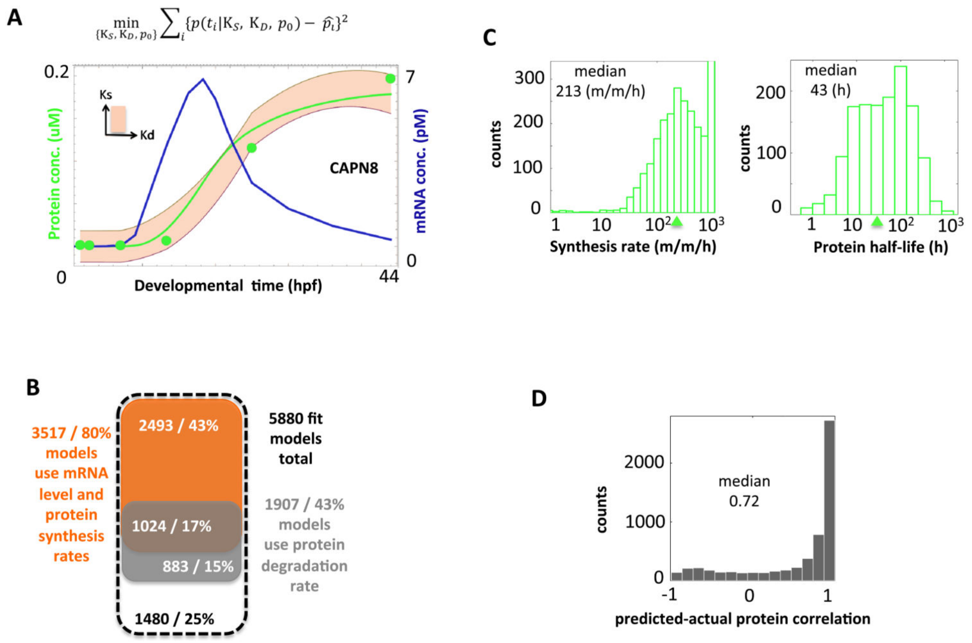


Figure 6. Mass-action Kinetics Equation Results in a Plausible Model of Embryonic Protein Economy

(A) Robust fitting of solution to the equation $\frac{dp}{dt} = \mathrm{K_S r(t)} - \mathrm{K_D p(t)}$ is done by searching a combination of synthesis and degradation rates minimizing the mean square difference in protein level (see equation above the plot). Beige stripe shows 95% confidence band for protein dynamics which corresponds to the 95% confidence range in synthesis and degradation rates. This region includes actual protein measurements marked via green discs. The no-degradation model is selected. **(B)** Venn diagram of models of different complexity. **(C)** Histograms of half-life (left) and synthesis rate (right). Half-life is given in hours, while synthesis rate in moles of protein synthesized per mole of mRNA per hour. **(D)** Histogram of Pearson correlation for model-based vs measured protein expression for a model assuming median synthesis and median degradation rates while using the actual initial concentration.

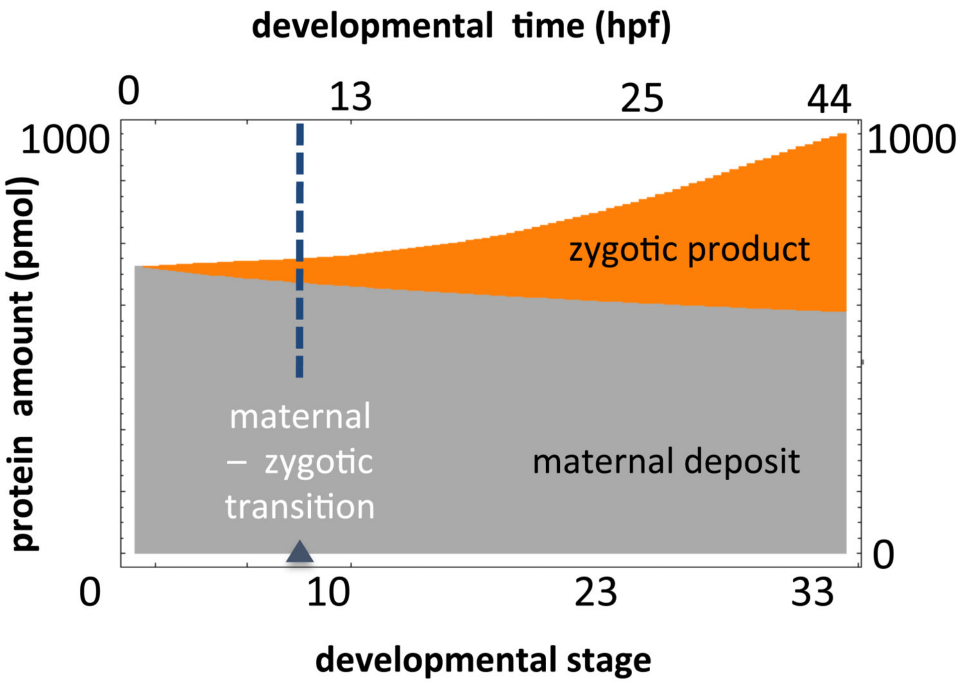


Figure 7. Embryonic protein economy

expressed as gradual replacement of maternal by zygotic protein, integrated over all proteins fitted by our model and extrapolated to the whole embryo.