bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Extensive post-transcriptional regulation across human tissues

Nikolai Slavov, Alexander Franks, Edoardo Airoldi

Department of Statistics and FAS Center for Systems Biology,
Harvard University, Cambridge, MA 02138, USA
Correspondence: nslavov@alum.mit.edu

Abstract

Transcriptional and post-transcriptional regulation shape tissue-type-specific proteomes, but their relative contributions remain contested. Common estimates of the factors determining protein levels do not distinguish between (i) the factors determining the variability between the abundances of different proteins and (ii) the factors determining the physiological variability of the same protein across different tissue types. We estimated the factors determining these two orthogonal sources of variability and found that the variability between the levels of different proteins is dominated by transcriptional regulation while in contrast the physiological variability across tissue-types is dominated by post-transcriptional regulation. These results reconcile existing estimates in the literature, caution against estimating protein fold-changes from mRNA fold-changes between different cell-types, and highlight the contribution of post-transcriptional regulation in shaping tissue-type-specific proteomes.

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

The relative ease of measuring mRNA levels has facilitated numerous investigations of how cells regulate their gene expression across different pathological and physiological conditions (Sørlie *et al*, 2001; Slavov and Dawson, 2009; Spellman *et al*, 1998; Slavov *et al*, 2011, 2012; Djebali *et al*, 2012). However, often the relevant biological processes depend on protein levels, and mRNA levels are merely proxies for protein levels (Alberts *et al*, 2014). If a gene is regulated mostly transcriptionally, its mRNA level is a good proxy for its protein level. Conversely, post-transcriptional regulation (PTR) can set protein levels independently from mRNA levels, as in the cases of classical regulators of development (Kuersten and Goodwin, 2003), cell division (Hengst and Reed, 1996; Polymenis and Schmidt, 1997) and metabolism (Daran-Lapujade *et al*, 2007; Slavov *et al*, 2014a). Thus understanding the relative contributions of transcriptional and post-transcriptional regulation is essential for understanding their trade-offs and the principles of biological regulation, as well as for assessing the feasibility of using mRNA levels as proxies for protein levels.

Some studies of these relative contributions have concluded that protein levels depend mostly on the mRNA levels (Li *et al*, 2014; Jovanovic *et al*, 2015; Csárdi *et al*, 2015) while other studies have concluded the opposite, i.e., that protein levels depend mostly on post-transcriptional regulation (Gygi *et al*, 1999; Smits *et al*, 2014; Schwanhäusser *et al*, 2011). These differing views arise because of differences in the systems, the methods, and the quantified protein variance. In particular, correlations between absolute levels of mRNA and protein mix/conflate many sources of variation, including variability between the levels of different proteins, variability within the same protein across different conditions and cell-types, and the variability due to measurement error and technological bias.

However, these different sources of variability have very different biological interpretations and implications. A major source of variability in protein and mRNA data arises from differences between the levels of mRNAs and proteins corresponding to different genes. That is, the mean levels (averaged across tissue-types) of different proteins and mRNAs vary widely. We refer to this source of variability as *mean-level variability*. This mean-level variability reflects the fact that some proteins, such as ribosomal proteins, are highly abundant across all profiled conditions while other proteins are orders of magnitude less abundant across all profiled conditions. Another

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

principal source of variability in protein levels, intuitively orthogonal to the mean-level variability, is the variability within a protein across different physiological conditions or cell-types. This variability reflects normal physiological regulation, which we refer to as *physiological variability*, and is usually much smaller in magnitude. However, physiological variability is frequently the most relevant source of variability for understanding different phenotypes across cells-types and physiological conditions.

Here, we separately quantify the contributions of transcriptional and post-transcriptional regulation to the mean-level variability and to the physiological variability across human tissues. Our results suggest that the physiological variability across human tissues is dominated by post-transcriptional regulation, while the mean-level variability is dominated by transcriptional regulation. These results reconcile previous results in the literature (Gygi *et al*, 1999; Schwanhäusser *et al*, 2011; Li *et al*, 2014; Wilhelm *et al*, 2014; Jovanovic *et al*, 2015; Csárdi *et al*, 2015; Smits *et al*, 2014) and highlight the dominance of post-transcriptional regulation in determining the variability in the levels of a protein across cell-types and physiological conditions. We then suggest a simple and general approach for deconvolving the contributions of transcriptional and post-transcriptional regulation to measured protein levels.

Results

The correlation between absolute mRNA and protein levels conflates distinct sources of variability

We start by outlining the statistical concepts underpinning the common correlational analysis and depiction (Gygi *et al*, 1999; Schwanhäusser *et al*, 2011; Wilhelm *et al*, 2014; Csárdi *et al*, 2015) of estimated absolute protein and mRNA levels as displayed in Figure 1a. The magnitude of the correlation between the absolute mRNA and protein levels of different genes and across different tissue-types is used used to estimate the level at which the protein levels are regulated (Wilhelm *et al*, 2014). If the physiological variability of a gene is dominated by transcriptional regulation, its protein-to-mRNA ratio in different tissue-types will be a gene-specific constant. Based on this idea, Wilhelm *et al* (2014) estimated these protein-to-mRNA ratios. They suggested that the

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

median ratio for each gene can be used to scale its tissue-specific mRNA levels and that this "scaled mRNA" predicts accurately tissue-specific protein levels.

Indeed, scaled mRNA levels explain large fraction of the total variance ($R_T^2 = 0.77$, across 6104 measured proteins, Figure 1a) as previously observed (Schwanhäusser *et al*, 2011; Wilhelm *et al*, 2014). However, R_T^2 quantifies the fraction of the total protein variance explained by mRNA levels between genes and across tissue-types; thus, it conflates the mean-level variability with the physiological variability. This conflation is shown graphically in Figure 1b for a subset of 100 genes measured across 12 tissues. The physiological variability is captured by the variability *within* the regression fits and the mean-level variability is captured by the variability *between* the regression fits.

The aggregation of distinct sources of variation, where different subgroups of the data show different trends, may lead to the effect known as Simpson's or amalgamation paradox in the statistical literature, which can lead to counter-intuitive results and incorrect conclusions (Blyth, 1972). To illustrate the Simpson's paradox in this context, we chose a subset of genes for which the scaled mRNA and measured protein levels are negatively correlated across tissues, and the mean-level variability spans the full dynamic range of the data. For this subset of genes, the overall (conflated/amalgamated) correlation is large and positive, despite the fact that all within-gene trends are negative. This counter-intuitive result is possible because the conflated correlation is dominated by the variability with larger dynamical range, in this case the mean-level variability. This conceptual example taken from the Wilhelm *et al* (2014) data demonstrates that, R_T^2 is not necessarily informative about the physiological variability, i.e., the protein variance explained by scaled mRNA *within* a gene (R_P^2). Thus the conflated correlation is not generally informative about the level — transcriptional or post-transcriptional — at which physiological-variability is regulated.

However, it is exactly the physiological-variability across tissue-types that provides the biological identity of each tissue type. This physiological variability has a dynamic range of about $2-10$ fold and is thus dwarfed by the 10^4 fold dynamic range of abundances across different proteins. To further demonstrate the implications of this vast difference in the dynamic ranges, we generate data from a simple model using the observed between-tissue variability in the protein-to-mRNA ratio. This protein/mRNA ratio has been referred to as a gene's "translational efficiency" because it reflects, in large part, its translational rate. Since this ratio also reflects other layers of regulation,

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

such as protein degradation (Jovanovic *et al*, 2015), we will refer to it as a PTR ratio.

Physiological mRNA variability is a poor predictor of the physiological protein variability

Figure 1 illustrates the statistical problems with using the fraction of the total protein variance explained by scaled mRNA levels (R_T^2) as an indication about the extent to which mRNA changes contribute to protein changes across tissues (i.e., R_P^2). To investigate the significance of this conflation further, we next evaluated the differences between scaling mRNA with the median PTR ratio (as in Figure 1a) and scaling mRNA with the PTR ratio of a specific tissue. That is, instead of using the median PTR ratio, we can use the PTR ratio estimated from one tissue to scale the mRNA level from another tissue. For instance, we correlate the protein levels measured in uterine to the uterine mRNA levels scaled by the prostate PTR ratio, Figure 2a. This correlation is lower compared to the correlation when mRNA is scaled by the median PTR ratio shown in Figure 1a. This reduction underscores that the PTR ratio for a gene varies enough between tissue types to affect even the conflated variability R_T^2. Extending this analysis to more pairs of tissues (Figure 2b, c) indicates very similar results; In all cases, the correlation is around 0.5, substantially smaller than the 0.9 correlation observed when mRNA is scaled by the median PTR ratio, Figure 1a.

Despite the very similar correlations between measured protein and scaled mRNA levels for all 3 comparisons in Figure 2a-c, the corresponding correlations between the protein and mRNA fold-changes differ substantially, Figure 2d-f. Unlike the buffering of mRNA variability observed across species and individuals (Khan *et al*, 2013; Battle *et al*, 2014), the protein levels generally vary much more across tissues than the mRNA levels (Supporting Figure 1); thus the protein and the mRNA fold changes in Figure 2d-f are plotted on different scales. The fold change comparisons in Figure 2d-f demonstrate that in fact, the fraction of variance explained in protein fold-changes by mRNA fold-changes is usually small and depends strongly on the compared pair of tissues. For instance, the mRNA fold-changes between the uterus and prostate have essentially no predictive power for protein fold-changes across these tissues. At the same time, other tissues show a moderate fold-change correlation (e.g., prostate vs. kidney and uterus vs. kidney). The fraction of the variance in protein fold-changes that can be explained by mRNA fold-changes varies significantly across

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the three examples. However, R_T^2 remains high and constants in all three examples (Figure 2a-c), because it is dominated by the mean-level variability. This result underscores the general problem of variance conflation in the analysis of measured mRNA and protein levels.

Next, we sought to evaluate whether physiological variability of mRNAs can serve as a proxy for the physiological variability of proteins. For this analysis, we extend our results on fold-change correlations from Figure 2 to all pairwise combinations of tissue-types. The range of correlations in Figure 3a indicates that for some pairs of tissues, physiological variability in mRNA explains a significant fraction of the physiological protein variance but for other tissue-type pairs it does not. This result indicates that for most genes, the observed variability across at least some tissues is either due to measurement noise or to post-transcriptional regulation. Before focusing on distinguishing between these two possibilities, we investigated whether the physiological variability of some genes is regulated primarily transcriptionally. If so, the protein fold-changes of such genes may be predicted reliably from mRNA fold-changes. To this end, we quantify the error in predicting protein fold-changes from mRNA fold-changes. The cumulative distribution of errors (Figure 3b) indicates that the protein fold-changes for less that 1000 genes can be estimated from mRNA fold-changes with less than 100% error. For over 30% of proteins, estimating protein levels using a single gene-specific PTR ratio results in over 1000% error; see methods.

Coordinated post-transcriptional regulation of functional gene sets

The lack of correlation between protein and mRNA fold-changes can reflect large measurement noise rather than post-transcriptional regulation (Li *et al*, 2014; Franks *et al*, 2015). The noise contribution to the variability of the PTR ratios of a gene is independent from the function of the gene. Conversely, genes with similar functions are likely to be regulated similarly and thus have similar tissue-type-specific PTR ratios. Thus, we explored whether the across-tissues variability of the PTR ratios of functionally related genes reflects such tissue-type-specific and biological-function-specific post-transcriptional regulation.

For this analysis, we define the "relative PTR ratio" (rPTR) of a gene in a given tissue to be the PTR ratio in that tissue divided by the median PTR ratio of the gene across the other 11 tissues. We evaluated the significance of rPTR variability for a gene-set in each tissue-type by comparing the corresponding gene-set rPTR distribution to the rPTR distribution for those same genes pooled

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

across the other tissues (Figure 4); we use the KS-test to quantify the statistical significance of differences in the rPTR distributions; see Methods. Our results indicate that the genes from many GO terms (Consortium *et al*, 2004) have much higher rPTR in some tissues than in others. For example the ribosomal proteins of the small subunit (40S) have high rPTR in kidney but low rPTR in stomach (Figure 4a-b).

Some of these trends can account for fundamental physiological differences between tissue types. For example, the kidney is by far the most metabolically active (energy consuming) tissue among the 12 profiled tissues (Hall, 2010) and it has very high rPTR for many gene sets involved in energy production (Figure 4a). In this case, post-transcriptional regulation very likely plays a functional role in meeting the high energy demands of kidneys. Moreover, the fact that we observe a highly significant (posterior error probability $< 10^{-10}$) mode of rPTR (such as increased TF for mitochondrial genes and decreased rPTR for focal adhesion in kidney) indicates that at least some of the variability in post-transcriptional regulation across tissue-types reflects regulatory activity rather than measurement noise.

Quantifying post-transcriptional regulation across human tissues

The results in Figure 4 demonstrate the some of the physiological variability of protein levels is due to post-transcriptional regulation, not noise. To further quantify the fractions of physiological protein variability due to transcriptional regulation, post-transcriptional regulation, and noise, we need to take noise into account. Both RNA-seq (Marioni *et al*, 2008; Consortium *et al*, 2014) and mass-spectrometry (Schwanhäusser *et al*, 2011; Peng *et al*, 2012) have relatively large and systematic error in estimating *absolute* levels of mRNAs and proteins, i.e., the ratios between different proteins/mRNAs. These errors originate from DNA sequencing GC-biases, and variations in protein digestion and peptide ionization. However, relative quantification of the same gene across tissue-types by both methods is much more accurate since systematic biases are minimized when taking ratios between the intensities/counts of the same peptide/DNA-sequence measured in different tissue types (Ong *et al*, 2002; Blagoev *et al*, 2004; Consortium *et al*, 2014; Jovanovic *et al*, 2015). It is this relative quantification that is used in estimating physiological variability, and thus noise levels are much smaller compared to the noise of absolute quantification.

To quantify the transcriptional and post-transcriptional contributions to physiological protein

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

variability, we start by estimating the reliability of the measurements, Figure 5a, b. Reliability is simply defined as the fraction of the observed/empirical variance due to signal. Thus reliability is proportional to the signal strength and decreases with the noise levels. For both protein and mRNA, we use independent estimates of their fold changes between salivary and adrenal glands. In the case of mRNA, these independent estimates correspond to replica RNA-seq measurements, Figure 5a. For proteins, the independent estimates where derived by non-overlapping sets of of peptides; that is, the fold change of each protein with multiple quantified peptides in the salivary and adrenal glands was estimated from from two non-overlapping sets of of peptides, Figure 5b.

Taking into account the reliability of the measurements, we depict the upper-bound for the fractions of physiological protein variability that can be explained by mRNA levels (i.e., transcriptional regulation) in Figure 5c. To account for any uncertainty in the reliability estimates, we depict the fraction of explained variance for a wide range of reliability estimates. At the reliability estimated for this dataset (Figure 5a, b), at most 30 % of the physiological variability – the variability of protein levels across-tissue types that is left after accounting for the measurement noise – can be explained by the mRNA levels. The remaining 70 % is most likely due to post-transcriptional regulation. This result underscores the different modes of regulation for the mean-level variability (mostly transcriptional) and for the physiological variability (mostly post-transcriptional). Since the exact estimate depends on the noise/reliability levels, we show estimates for higher and lower levels of reliability. Even if measurement error is larger and most measured variance in both mRNA and protein levels is due to noise, not signal, i.e., reliability $< 50\ \%$, transcriptional regulation still can explain at most about 50 % of the physiological protein variability. Thus even in this extreme case, post-transcriptional regulation is likely a major determinant of physiological protein variability, and thus tissue-type specific proteomes.

Discussion

Highly abundant proteins have highly abundant mRNAs. This dependence is consistently observed (Jovanovic *et al*, 2015; Csárdi *et al*, 2015; Gygi *et al*, 1999; Smits *et al*, 2014; Schwanhäusser *et al*, 2011) and dominates the explained variance in the estimates of absolute protein levels (Figure 2a, Figure 2a-c, Figure 5c). This underscores the role of transcription for setting the full dynamic range

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of protein levels. In stark contrast, differences in the proteomes of distinct human tissues are poorly explained by transcriptional regulation (Figure 2d-f, Figure 5a-b). Rather, the mechanisms shaping the tissue-specific proteomes involve post-transcriptional regulation. This result underscores the role of translational regulation and of protein degradation for mediating physiological functions within the range of protein levels consistent with life.

The estimates of absolute protein levels are affected by technological biases and measurement error (Peng *et al*, 2012; Franks *et al*, 2015) which can contribute to overestimating post-transcriptional regulation. These biases can difficult to estimate and influential (Csárdi *et al*, 2015), potentially leading to underestimates of the variance in protein levels explained by transcription. However, such systematic biases do not affect the relative changes of protein levels and the estimates of physiological variability. Indeed, the strong enrichment of rPTR within gene sets (Figure 4) demonstrates a concerted regulation at the post-transcriptional level. It is thus unlikely that bias and measurement error alone explain the weak correlations between tissue-specific differences in mRNA and protein levels (Figure 2).

As with all analysis of empirical data, all results depend on the quality of the data and the estimates of their reliability. If the reliability of the data are significantly below 50 %, the data would be consistent with mRNA levels accounting for most of the physiological variability, as an upper limit estimate for the transcriptional contribution. In that case, however, the signal is dominated by noise, and thus the data cannot accurately quantify the contributions of different regulatory mechanisms. The strong functional enrichment for rPTR (Figure 4) and the error estimates by Wilhelm *et al* (2014) suggest that the physiological variance not explained by mRNA is likely due to post-transcriptional regulation, not to signal dwarfed by noise.

The correlations between the fold changes of mRNAs and proteins in Figure 3 indicate that the relative contributions of transcriptional and post-transcriptional regulation can vary substantially depending on the tissues compared. Thus, the level of gene regulation depends at least to come extent on context. For example transcriptional regulation is contributing significantly to the dynamical responses of dendritic cells (Jovanovic *et al*, 2015) and to the differences between spleen and kidney (Figure 3a) but much less to the differences between spleen and thyroid gland (Figure 3a). All data, across all profiled tissues, suggest that post-transcriptional regulation contributed very substantially to the physiological variability of protein levels. The degree of this large contri-

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bution depends on the context.

Indeed, if we only increase the levels for a set of mRNAs without any other changes, the corresponding protein levels must increase proportionally as demonstrated by gene inductions (McIsaac *et al*, 2011). However, the differences across cell-types are not confined only to different mRNA levels. Rather, these differences include different RNA-binding proteins, alternative untranslated regions (UTRs) with known regulatory roles in protein synthesis, specialized ribosomes, and different protein degradation rates (Mauro and Edelman, 2002; Gebauer and Hentze, 2004; Rojas-Duran and Gilbert, 2012; Castello *et al*, 2012; Arribere and Gilbert, 2013; Slavov *et al*, 2014b; Katz *et al*, 2014). The more substantial these differences, the bigger the potential for post-transcriptional regulation. Thus cell-type differentiation and commitment may result in much more post-transcriptional regulation than observed during perturbations preserving the cellular identity. Consistent with this possibility, mRNA fold-changes can account for less than 50 % of the measured physiological variability; the remaining variability is likely due to substantial tissue-specific post-transcriptional regulation; in contrast, stimulating dendritic cells elicits a strong transcriptional response but not change in the cell-type and thus less cell-type-specific post-transcriptional regulation (Jovanovic *et al*, 2015).

Acknowledgments

We thank M. Jovanovic, E. Wallace, and J. Schmiedel for discussions and constructive comments. This work was funded by a SPARC grant from the Broad Institute, a grant from the National Institutes of Health (R01-GM096193), and an Alfred Sloan Research Fellowship to E.M.A.

Methods

Data and scaled mRNA levels

We used data from Wilhelm *et al* (2014); Fagerberg *et al* (2014) containing estimates for the mRNA levels (based on RNA-seq) and for the protein levels (based on mass-spectrometry) of $N = 6104$ genes measured in each of twelve different human tissues: adrenal gland, esophagus, kidney, ovary, pancreas, prostate, salivary gland, spleen, stomach, testis, thyroid gland, and uterus.

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For these genes, about 8% of the mRNA measurements and about 40% of the protein measurements are missing.

First, denote m_{ij} the log mRNA levels for gene i in condition j. Similarly, let p_{ij} denote the corresponding log protein levels. First, we normalize the columns of the data, for both protein and mRNA, to different amounts of total protein per sample. Any multiplicative factors on the raw scale correspond to additive constants on the log scale. Consequently, we normalize data from each tissue-type by minimizing the sum of squared differences between data from that tissue and the first tissue (chosen to serve as a baseline). Specifically, for all proteins and conditions $j > 1$, we normalize each measurement by setting

$$p^n_{ij} \leftarrow p^u_{ij} - \frac{1}{N}\sum_i (p^u_{i1} - p^u_{ij})$$

Where p^n_{ij} and p^u_{ij} represent the normalized and non-normalized protein measurements respectively. We conduct the same normalization for mRNA. This normalization corrects for any multiplicative differences in the raw mRNA or protein.

After normalization, we define $r_{ij} = p_{ij} - m_{ij}$ as the log PTR ratio of gene i in condition j. If the post-transcriptional regulation the ith gene were not tissue-specific, then the ith PTR ratio would be independent of tissue-type and can be estimated as

$$\hat{T}_i = \underset{j}{\text{median}}(p_{ij} - m_{ij})$$

Then the log "scaled mRNA" (or mean protein level) can be defined as

$$\overline{p}_{ij} = m_{ij} + T_i$$

On the raw scale this amounts to scaling each mRNA by its median PTR ratio and represents and estimate of the mean protein level. The residual difference between the log mean protein level and the measured log protein level

$$r_{ij} = p_{ij} - \overline{p}_{ij}$$

consists of both tissue-specific post-transcriptional regulation and measurement noise.

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Noise correction

Measurement noise attenuates estimates of correlations between mRNA and protein level (Franks *et al*, 2015). A simple way to quantify this attenuation of correlation due to measurement error is via Spearman's correction. Spearman's correction is based on the fact that the variance of the measured data can be decomposed into the sum of variance of the noise and the signal. If the noise and the signal are independent, this decomposition and the Spearman's correction are exact. Below is a simple proof that the observed empirical variance is the sum of the variance of the signal and the variance of the noise:

- e_i - Expectation at the ith data point; $\tilde{e}_i = e_i - \langle e \rangle$
- ζ_i - Noise at the ith data point; $\langle \zeta \rangle = 0$
- x_i - Observation at the ith data point; $\tilde{x}_i = x_i - \langle x \rangle$

$$\sigma_x^2 = \frac{1}{n}\sum_i \tilde{x}_i^2 = \frac{1}{n}\sum_i (\tilde{e}_i + \zeta_i)^2 =$$
$$= \underbrace{\frac{1}{n}\sum_i \tilde{e}_i^2}_{\sigma_e^2} + \underbrace{\frac{1}{n}\sum_i \zeta_i^2}_{\sigma_\zeta^2} + \underbrace{\frac{2}{n}\sum_i \tilde{e}_i\zeta_i}_{\approx 0}$$

To use this additivity and make Spearman's correction, we need to estimate the "reliability" of the measurements, which is defined as the fraction of total measured variance due to signal rather than to noise:

$$\text{Reliability} = \frac{\sigma_{signal}^2}{\sigma_{total}^2} \quad (1)$$
$$= 1 - \frac{\sigma_{err}^2}{\sigma_{err}^2 + \sigma_{signal}^2} \quad (2)$$

The noise corrected correlation is then simply

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\frac{Cor(X,Y)}{\sqrt{Rel(X)Rel(Y)}} \tag{3}$$

$$\tag{4}$$

We estimated the reliabilities of the mRNA and the protein measurements from independent estimates for the mRNA and the protein levels, Figure 5a, b. Given these estimates of mRNA and protein reliabilities, we computed the de-noised fraction of physiological protein variability explained by transcript levels using Equation 3. Figure 5c depicts the R^2 regions as a function of measurement reliablities.

Functional gene set analysis

To identify tissue-specific PTR for functional sets of genes, we analyzed the distributions of PTR ratios within functional gene-sets using the same methodology as Slavov and Botstein (2011). We restrict our attention to functional groups in the GO ontology (Consortium *et al*, 2004) for which at least 10 genes were quantified by Wilhelm *et al* (2014). Let k index one of these approximately 1600 functional gene sets. First, for every gene in every tissue we estimate the relative PTR (rPTR) or equivalently, the difference between log mean protein level and measured protein level:

$$\hat{r}_{ij} = p_{ij} - \underset{j' \neq j}{\text{median}}(p_{ij'} - m_{ij'})$$

To exclude the possibility that $\hat{r}_{ij} = 0$ exactly, we require that $j' \neq j$. When the estimated rPTR is larger than zero, the measured protein level in tissue j is larger than the estimated mean protein level. Likewise, when this quantity is smaller than onezero, the measured protein is smaller than expected. Measured deviations from the mean protein level are due to both measurement noise and tissue specific PTR. To eliminate the possibility that all of the variability in the rPTR ratios is due to measurement we conduct a full gene set analysis.

For each of the gene sets we compute a vector of these estimated log ratios so that a gene set is comprised of

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\mathcal{G}_{kj} = \{\hat{r}_{i_1 j}, ..., \hat{r}_{i_{n_k} j}\}$$

where i_1 to i_{n_k} index the genes in set k and j indexes the tissue type.

Let $KS(\mathcal{G}_1, \mathcal{G}_2)$ be the function that returns the p-value of the Kolmogorov-Smirnov test on the distribution in sets $\mathcal{G}_1$ and $\mathcal{G}_2$. The KS-test is a test for a difference in distribution between two samples. Using this test, we identify gene sets that show systematic differences in PTR ratio in a particular tissue (j) relative to all other tissues.

Specifically, the p-value associated with gene set k in condition j is

$$\rho_{kj} = \text{KS}(\mathcal{G}_{kj}, \bigcup_{j' \neq j} \mathcal{G}_{kj'})$$

To correct for multiple hypotheses testing, we computed the false discovery rate (FDR) for all gene sets in tissue j (Storey, 2003). In Figure 4a-c, we present only the functional groups with FDR less than 2% and report their associated p-values. The significance of many of these groups, controlling for false discoveries suggests that not all of the variability in rPTR is due to measurement noise.

References

Alberts B, Johnson A, Morgan JLD, Raff M, Roberts K, Walter P (2014) *Molecular Biology of the Cell*. Garland, 6th edition

Arribere JA, Gilbert WV (2013) Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome research* **23**: 977–987

Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, Gilad Y (2014) Impact of regulatory variation from RNA to protein. *Science* : 1260793

Blagoev B, Ong SE, Kratchmarova I, Mann M (2004) Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nature biotechnology* **22**: 1139–1145

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Blyth CR (1972) On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association* **67**: 364–366

Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, *et al* (2012) Insights into RNA biology from an atlas of mammalian mRNA–binding proteins. *Cell* **149**: 1393–1406

Consortium GO, *et al* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic acids research* **32**: D258–D261

Consortium SI, *et al* (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology* **32**: 903–914

Csárdi G, Franks A, Choi DS, Airoldi EM, Drummond DA (2015) Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *PLoS Genetics* **11**: e1005206

Daran-Lapujade P, Rossell S, van Gulik WM, Luttik MA, de Groot MJ, Slijper M, Heck AJ, Daran JM, de Winde JH, Westerhoff HV, *et al* (2007) The fluxes through glycolytic enzymes in *Saccharomyces cerevisiae* are predominantly regulated at posttranscriptional levels. *Proceedings of the National Academy of Sciences* **104**: 15753–15758

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, *et al* (2012) Landscape of transcription in human cells. *Nature* **489**: 101–108

Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K, *et al* (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular Cellular Proteomics* **13**: 397–406

Franks AM, Csárdi G, Drummond DA, Airoldi EM (2015) Estimating a structured covariance matrix from multi-lab measurements in high-throughput biology. *Journal of the American Statistical Association* **110**: 27–44

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gebauer F, Hentze MW (2004) Molecular mechanisms of translational control. *Nature reviews Molecular cell biology* **5**: 827–835

Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. *Molecular and cellular biology* **19**: 1720–1730

Hall JE (2010) *Guyton and Hall Textbook of Medical Physiology: Enhanced E-book*. Elsevier Health Sciences

Hengst L, Reed SI (1996) Translational control of p27Kip1 accumulation during the cell cycle. *Science* **271**: 1861–1864

Jovanovic M, Rooney MS, Mertins P, Przybylski D, Chevrier N, Satija R, Rodriguez EH, Fields AP, Schwartz S, Raychowdhury R, *et al* (2015) Dynamic profiling of the protein life cycle in response to pathogens. *Science* **347**: 1259038

Katz Y, Li F, Lambert NJ, Sokol ES, Tam WL, Cheng AW, Airoldi EM, Lengner CJ, Gupta PB, Yu Z, *et al* (2014) Musashi proteins are post-transcriptional regulators of the epithelial-luminal cell state. *eLife* **3**: e03915

Khan Z, Ford MJ, Cusanovich DA, Mitrano A, Pritchard JK, Gilad Y (2013) Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* **342**: 1100–1104

Kuersten S, Goodwin EB (2003) The power of the 3 UTR: translational control and development. *Nature Reviews Genetics* **4**: 626–637

Li JJ, Bickel PJ, Biggin MD (2014) System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* **2**: e270

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research* **18**: 1509–1517

Mauro VP, Edelman GM (2002) The ribosome filter hypothesis. *Proceedings of the National Academy of Sciences* **99**: 12031–12036

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

McIsaac RS, Silverman SJ, McClean MN, Gibney PA, Macinskas J, Hickman MJ, Petti AA, Botstein D (2011) Fast-acting and nearly gratuitous induction of gene expression and protein depletion in *Saccharomyces cerevisiae*. *Molecular biology of the cell* **22**: 4447–4459

Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular cellular proteomics* **1**: 376–386

Peng M, Taouatas N, Cappadona S, van Breukelen B, Mohammed S, Scholten A, Heck AJ (2012) Protease bias in absolute protein quantitation. *Nature methods* **9**: 524–525

Polymenis M, Schmidt E (1997) Coupling of cell division to cell growth by translational control of the G1 cyclin CLN3 in yeast. *Genes development* **11**: 2522

Rojas-Duran MF, Gilbert WV (2012) Alternative transcription start site selection leads to large differences in translation activity in yeast. *Rna* **18**: 2299–2305

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M (2011) Global quantification of mammalian gene expression control. *Nature* **473**: 337–342

Slavov N, Airoldi EM, van Oudenaarden A, Botstein D (2012) A conserved cell growth cycle can account for the environmental stress responses of divergent eukaryotes. *Molecular Biology of the Cell* **23**: 1986 – 1997

Slavov N, Botstein D (2011) Coupling among growth rate response, metabolic cycle, and cell division cycle in yeast. *Molecular Biology of the Cell* **22**: 1997–2009

Slavov N, Budnik B, Schwab D, Airoldi E, van Oudenaarden A (2014a) Constant Growth Rate Can Be Supported by Decreasing Energy Flux and Increasing Aerobic Glycolysis. *Cell Reports* **7**: 705 – 714

Slavov N, Dawson KA (2009) Correlation signature of the macroscopic states of the gene regulatory network in cancer. *Proceedings of the National Academy of Sciences* **106**: 4079–4084

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Slavov N, Macinskas J, Caudy A, Botstein D (2011) Metabolic cycling without cell division cycling in respiring yeast. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 19090–19095

Slavov N, Semrau S, Airoldi E, Budnik B, van Oudenaarden A (2014b) Variable stoichiometry among core ribosomal proteins. *Under review* **1**: arXiv:1406.0399

Smits AH, Lindeboom RG, Perino M, van Heeringen SJ, Veenstra GJC, Vermeulen M (2014) Global absolute quantification reveals tight regulation of protein expression in single Xenopus eggs. *Nucleic acids research* **42**: 9880–9891

Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, *et al* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences* **98**: 10869–10874

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B (1998) Comprehensive identification of cell cycle–regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular biology of the cell* **9**: 3273–3297

Storey JD (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of statistics* : 2013–2035

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

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

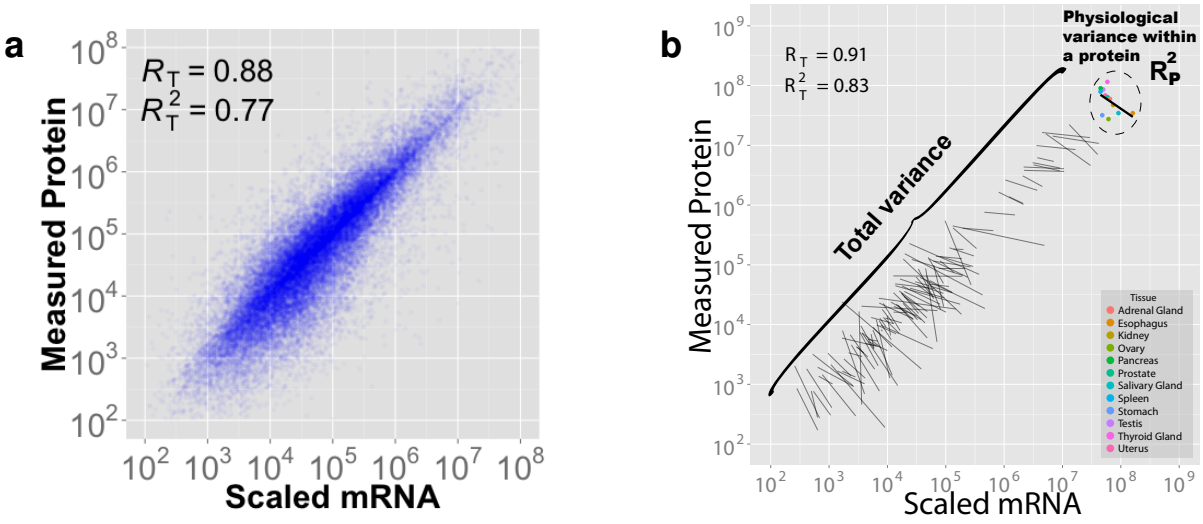

Figure 1. The fraction of total protein variance explained by scaled mRNA levels is not informative about the physiological variance explained by scaled mRNA levels. (**a**) mRNA levels scaled by the median protein-to-mRNA ratio correlate strongly with measured protein levels ($R_B^2 = 0.77$ over 6104 measured proteins in each of 12 different tissues). (**b**) A subset of 100 genes are used to illustrate an example Simpson's paradox: regression lines reflect within gene physiological variation across each tissue. Despite the fact that the overall correlation between scaled mRNA and measured protein levels is large and positive $R_T = 0.89$, for any single gene in this set, scaled mRNA is *negatively* correlated with measured protein levels ($R_P < 0$).

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

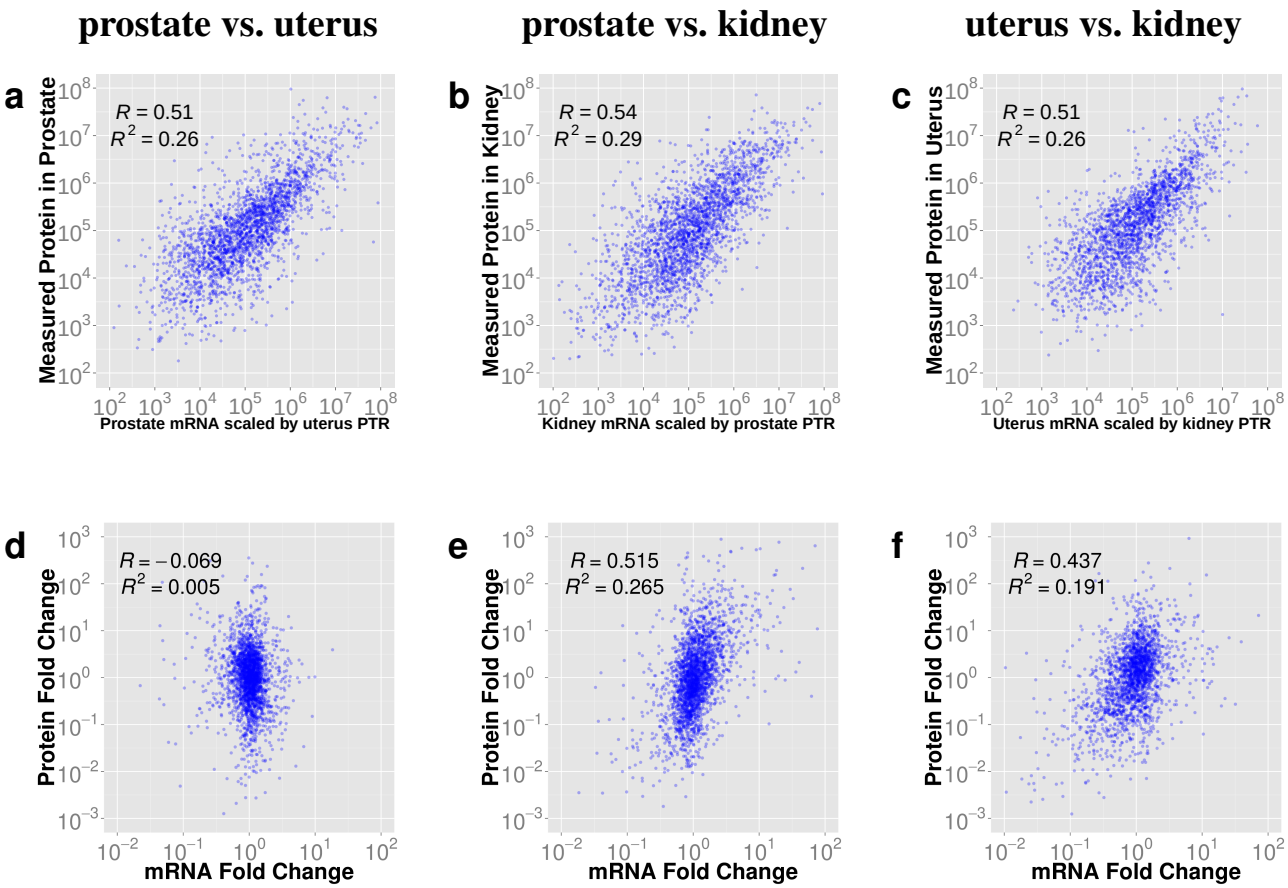

Figure 2. mRNA fold-changes are weakly correlated with protein fold-changes in some tissue-types but uncorrelated in others While scaled mRNA is predictive of the absolute protein levels (a-c, top row), the accuracy of these predictions does not generally reflect the accuracy of protein fold-changes predicted from mRNA fold-changes (d-f, bottom row).

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

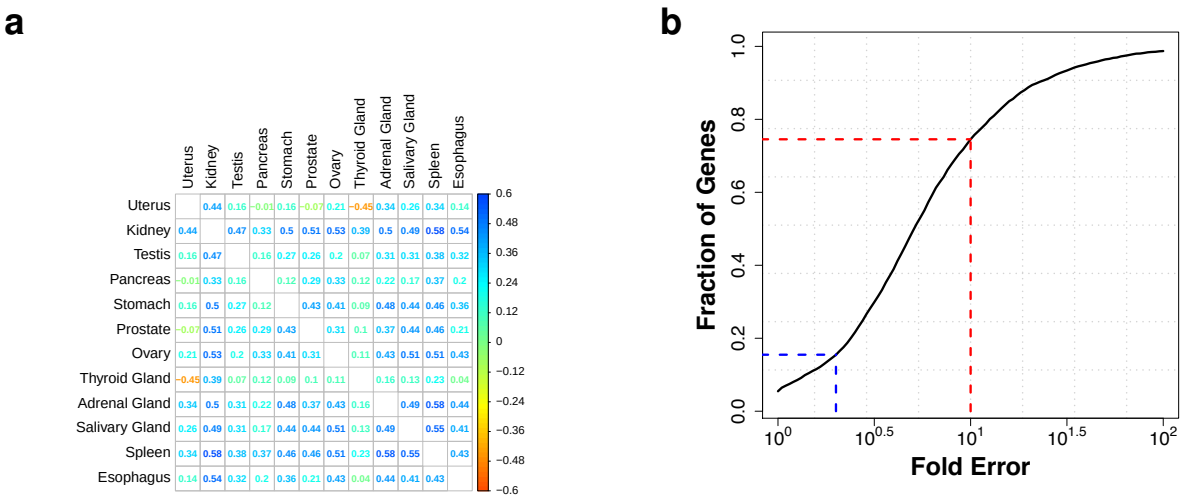

Figure 3. Measured protein fold-changes can be predicted from the corresponding mRNA fold-changes only for very few genes (**a**) As in Figure 2d-f, we computed all pairwise correlations between the mRNA and protein fold-changes for all 12 tissue-types. (**b**) The cumulative distribution of the maximum fold error that results from estimating protein fold-changes by mRNA fold-changes. The error in quantifying protein fold change from mRNA exceeds 10-fold for 1862 proteins (31%, red) and exceeds two-fold for 5106 proteins (85 %, blue); the fold-changes of less than 1000 proteins can be estimated from mRNA fold-changes with less than twofold (100%) error (blue).

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

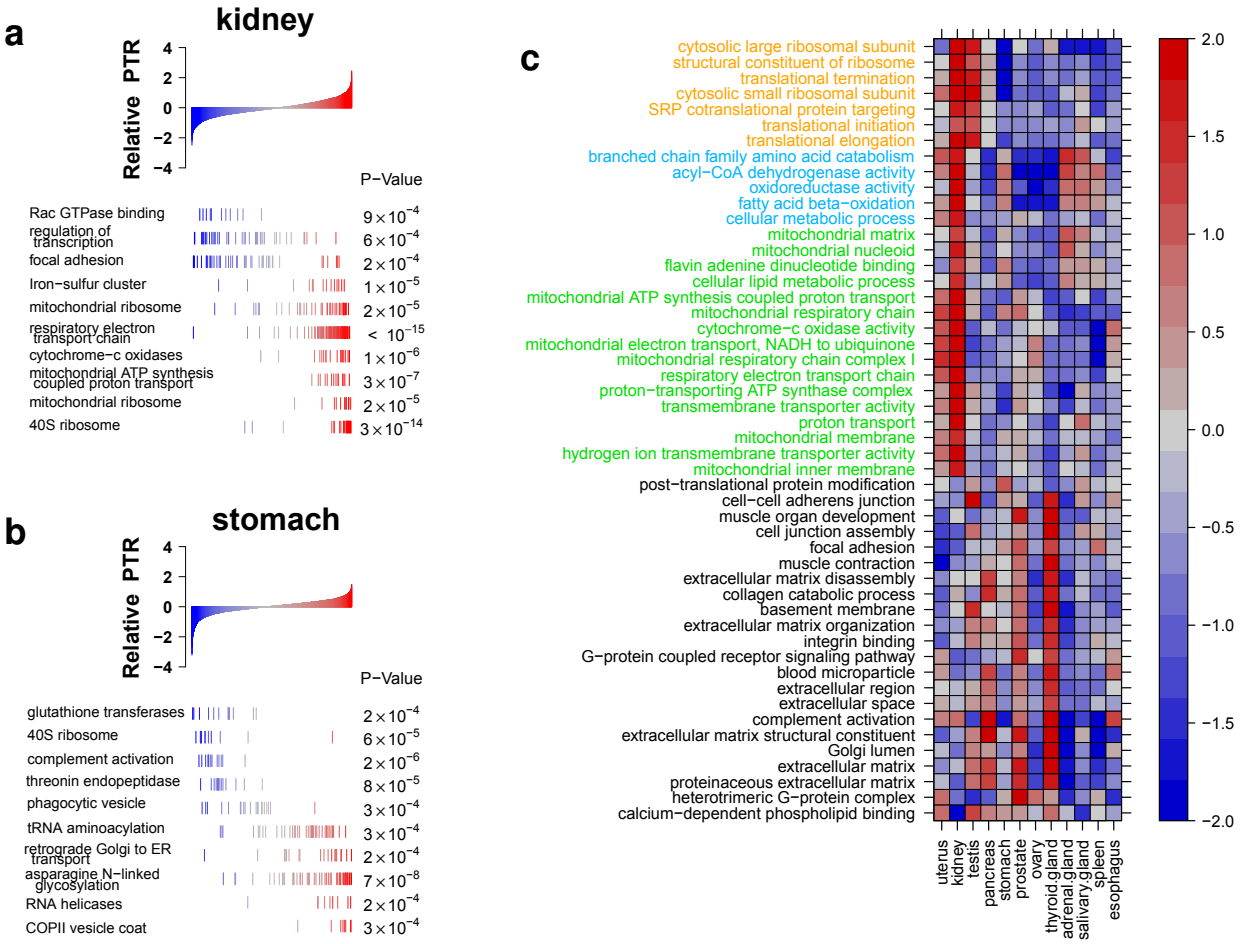

Figure 4. Concerted variability in the relative PTR ratio of functional gene-sets across tissue-types (**a**) mRNAs coding for the small ribosomal subunit, NADH dehydrogenase and respiratory proteins have much higher protein-to-mRNA ratios in kidney as compared to the median across the other 11 tissues (FDR $< 2\%$. In contrast mRNAs coding for focal adhesion have lower protein-to-mRNA ratios (FDR $< 2\%$). (**b**) The stomach also shows very significant rPTR variation, with low rPTR for the small ribosomal subunit and high rPTR for tRNA-aminoacylation (FDR $< 2\%$). (**c**) Summary of rPTR variability, as depicted in panel (a-b), across all tissues and many gene ontology (GO) terms. Metabolic pathways and functional gene-sets that show statistically significant (FDR $< 2\%$) variability in the relative protein-to-mRNA ratios across the 12 tissue types. Functionally related gene-sets are marked with the same color.

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Reliability

$$\mathrm{Rel}(x) = \frac{\mathrm{Var(Signal)}}{\mathrm{Var(Signal)} + \mathrm{Var(Noise)}}$$

Variance explained by mRNA

$$R^2 = \frac{Cor(x,y)^2}{Rel(x)Rel(y)}$$

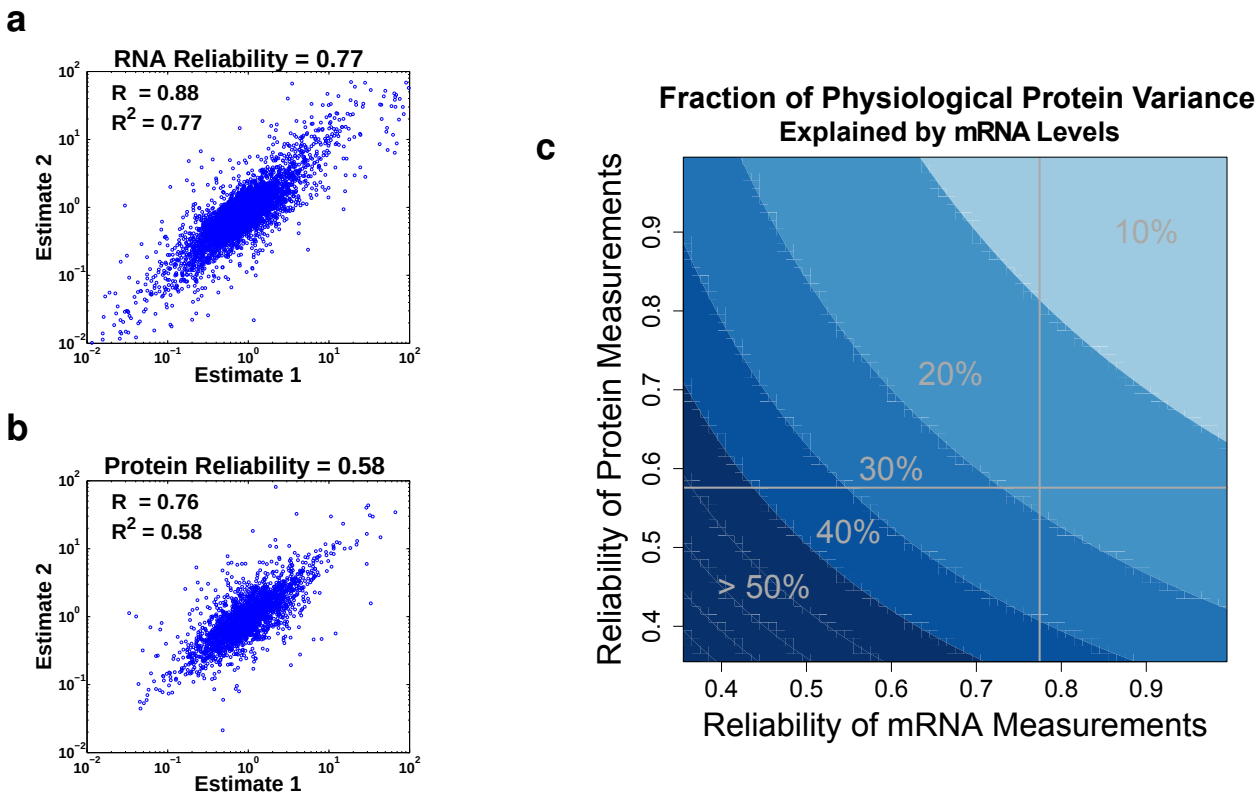

Figure 5. Contributions of transcriptional and post-transcriptional regulation to the observed physiological variability in protein levels (**a**) The reliability – defined as the fraction of the measured variance due to the signal – of relative mRNA levels is estimates as the R^2 between two independent estimates of the fold-changes between the salivary and the adrenal glands. (**b**) The reliability of relative protein levels is estimates as the R^2 between two independent estimates of the fold-changes between the salivary and the adrenal glands. The two estimates are derived from non-overlapping sets of peptides. (**c**) The fraction of physiological protein variance that can be explained by transcriptional regulation (mRNA levels) is plotted for reliability levels of the measurements ranging from 35% to 100 %. The reliability estimates from panels (a) and (b) are marked by the gray lines.

bioRxiv preprint doi: https://doi.org/10.1101/020206; this version posted May 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

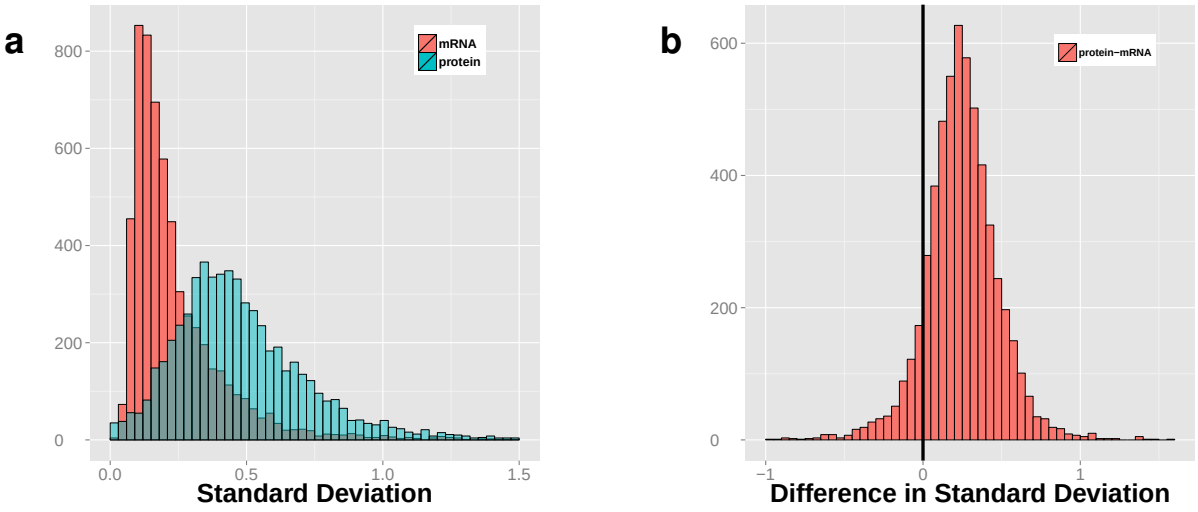

Figure S1. The dynamic range of physiological variability is larger for proteins than for mRNAs (**a**) Distributions of dynamic ranges for mRNAs and proteins quantified by the standard deviations computed on a log scale. Note that the log-scale makes the standard deviation independent of scalar scaling on the linear scale. (**b**) Distribution of differences between the standard deviations of proteins and their corresponding mRNAs. The larger than zero median indicates that the physiological variability of most genes is larger at the protein level than at the mRNA level.