Post-transcriptional regulation across human tissues

Alexander Franks[1], Edoardo Airoldi[2] & Nikolai Slavov[3]

[1]Department of Statistics, University of Washington, Seattle, WA 98195, USA
[2]Department of Statistics, Harvard University, Cambridge, MA 02138, USA
[3]Departments of Bioengineering and Biology, Northeastern University, Boston, MA 02115, USA
Correspondence: nslavov@alum.mit.edu

Abstract

Transcriptional and post-transcriptional regulation shape tissue-type-specific proteomes, but their relative contributions remain contested. Estimates of the factors determining protein levels in human tissues do not distinguish between (i) the factors determining the variability between the abundances of different proteins, i.e., mean-level-variability and, (ii) the factors determining the physiological variability of the same protein across different tissue types, i.e., across-tissue variability. We sought to estimate the contribution of transcript levels to these two orthogonal sources of variability, and found that mRNA levels can account for most of the mean-level-variability but not for across-tissue variability. The precise quantification of the latter estimate is limited by substantial measurement noise. However, protein-to-mRNA ratios exhibit substantial across-tissue variability that is functionally concerted and reproducible across different datasets, suggesting extensive post-transcriptional regulation. These results caution against estimating protein fold-changes from mRNA fold-changes between different cell-types, and highlight the contribution of post-transcriptional regulation to shaping tissue-type-specific proteomes.

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*, 2012a). 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 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*, 2014). 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.

Previous studies have considered single cell-types and conditions in studying variation in absolute mRNA and protein levels genome-wide, often employing unicellular model organisms or mammalian cell cultures (Gygi *et al*, 1999; Smits *et al*, 2014; Schwanhäusser *et al*, 2011; Li *et al*, 2014; Csárdi *et al*, 2015; Jovanovic *et al*, 2015). However, analyzing per-gene variation in relative mRNA and protein expression across different tissue types in a multicellular organism presents a potentially different and critical problem which cannot be properly addressed by examining only genome-scale correlations between mRNA and protein levels. Wilhelm *et al* (2014) and Kim *et al* (2014) have measured protein levels across human tissues, thus providing valuable datasets for analyzing the regulatory layers shaping tissue-type-specific proteomes. The absolute levels of proteins and mRNAs in these datasets correlate well, highlighting that highly abundant proteins have highly abundant mRNAs. Such correlations between the 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 tissues while other proteins, such as cell cycle and signaling regulators, are orders of magnitude less abundant across all profiled conditions (Wilhelm *et al*, 2014). Another principal source of variability in protein levels, intuitively orthogonal to the mean-level variability, is the variability within a protein across different cell-types or physiological conditions and we refer to it as *across-tissue variability*. The across-tissue variability is usually much smaller in magnitude, but may be the most relevant source of variability for understanding different phenotypes across cells-types and physiological conditions.

Here, we sought to separately quantify the contributions of transcriptional and post-transcriptional regulation to the mean-level variability and to the across-tissue variability across human tissues. Our results show that the much of the mean-level protein variability can be explained well by mRNA levels while across-tissue protein variability is poorly explained by mRNA levels; much of the unexplained variance is due to measurement noise but some of it is reproducible across datasets and thus likely reflects post-transcriptional regulation. These results add to 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 suggest that the post-transcriptional regulation is a significant contributor to shaping tissue-type specific proteomes in human.

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 correlation between the absolute mRNA and protein levels of different genes and across different tissue-types has been

used to estimate the level at which the protein levels are regulated (Wilhelm *et al*, 2014).

One measure reflecting the post-transcriptional regulation of a gene is its protein to mRNA ratio, which is sometimes referred to as a gene's "translational efficiency" because it reflects, at least in part, its translational rate. Since this ratio also reflects other layers of regulation, such as protein degradation (Jovanovic *et al*, 2015), and noise we will refer to it descriptively as *protein-to-mRNA* (PTR) ratio. If the across-tissue variability of a gene is dominated by transcriptional regulation, its PTR in different tissue-types will be a gene-specific constant. Based on this idea, Wilhelm *et al* (2014) estimated these protein-to-mRNA ratios and suggested that the 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, mRNA levels scaled by the corresponding median PTR explain large fraction of the total protein variance ($R^2_T = 0.77$, across 6104 measured proteins, Figure 1a) as previously observed (Schwanhäusser *et al*, 2011; Wilhelm *et al*, 2014). However, R^2_T 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 across-tissue variability. This conflation is shown schematically in Figure 1b for a subset of 100 genes measured across 12 tissues. The across-tissue variability is captured by the variability *within* the regression fits and the mean-level variability is captured by the variability *between* the regression fits.

Such aggregation of distinct sources of variability, where different subgroups of the data show different trends, may lead to counter-intuitive results and incorrect conclusions, and is known as the Simpson's or amalgamation paradox (Blyth, 1972). To illustrate the Simpson's paradox in this context, we depicted a subset of genes for which the measured mRNA and protein levels are unrelated across-tissues– the mean-level variability still 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 close to zero. 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^2_T is not necessarily informative about the across-tissue variability, i.e., the protein variance explained by scaled mRNA *within* a gene (R^2_P). Thus the conflated correlation is not generally informative about the level — transcriptional or post-transcriptional — at which

across-tissue variability is regulated. This point is further illustrated in **Supplementary Fig. 1** with data for all quantified genes: The correlations between scaled mRNA and measured protein levels are not informative for the correlations between the corresponding relative changes in protein and mRNA levels.

While across-tissue variability is smaller than mean-level variability, it is exactly the across-tissue variability that contributes to the biological identity of each tissue type. This across-tissue variability has a dynamic range of about $2 - 10$ fold and is thus dwarfed by the $10^3 - 10^4$ fold dynamic range of abundances across different proteins.

Estimates of transcriptional and post-transcriptional regulation across-tissues depend strongly on data reliability

Next, we sought to estimate the fractions of across-tissue protein variability due to transcriptional regulation and to post-transcriptional regulation. This estimate depends crucially on noise in the mRNA and protein data, from sample collection to measurement error. 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 can be 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 across-tissue variability, and we start by estimating the reliability of the relative quantification across human tissues, Figure 2a-d. Reliability is 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.

To estimate the *within study* reliability of mRNA levels, we split each dataset into two subsets, each of which contain measurements for all tissues. The levels of each mRNA were estimated from each subset and the estimates correlated, averaging across tissues (Figure 2a). These correlations

provide estimates for the reliability of each mRNA and their median provides a global estimate for the reliability of relative RNA measurement, not taking into account noise due to sample collection and handling.

To estimate the *within study* reliability of protein levels, we computed separate estimates of the relative protein levels within a dataset. For each protein, Estimate 1 was derived from 50 % of the quantified peptides and Estimate 2 from the other 50 %. Since much of the analytical noise related to protein digestion, chromatographic mobility and peptide ionization is peptide-specific, such non-overlapping sets of of peptides provide mostly, albeit not completely, independent estimates for the relative protein levels. The correlations between the estimates for each protein (averaging across 12 tissues) are displayed as a distribution in Figure 2b.

In addition to the *within study* measurement error, protein and mRNA estimates can be affected by study-dependable variables such as sample collection and data processing. To account for these factors, we estimated *across study* reliability by comparing estimates for relative protein and mRNA levels derived from independent studies, Figure 2c-d. For each gene, we estimate the reliability for each protein by computing hte the empirical correlation between mRNA abundance reported by the ENCODE (Djebali *et al*, 2012b) and by (Fagerberg *et al*, 2014). The distribution of correlations Figure 2c is shifted towards lower values compared to the within-study correlations, indicating that much of the noise in mRNA estimates is study-dependent.

To estimate the *across study* reliability of protein levels, we compared the protein levels estimated from data published by Wilhelm *et al* (2014) and Kim *et al* (2014). To quantify protein abundances, Wilhelm *et al* (2014) used iBAQ scores and Kim *et al* (2014) used spectral counts. To ensure uniform processing of the two datasets, we downloaded the raw data and analyzed them with maxquant using identical settings, and estimate protein abundances in each dataset using iBAQ; see Methods. The corresponding estimates for each protein were correlated to estimate their reproducibility. The distribution of correlations Figure 2d is shifted towards significantly lower values compared to the within-study correlations Figure 2d, indicating that, as with mRNA, the vast majority of the noise is study-dependent.

The across tissue correlations and the reliability of the measurements can be used to estimate the across tissue variability in protein levels that can be explained by mRNA levels (i.e., transcriptional regulation) as shown in Figure 2e and proven in the methods. As the reliability of

the protein and the mRNA estimates decrease, the sensitivity of the estimated transcriptional contribution increases. Although the average across-tissue mRNA protein correlation was only 0.29 ($R^2 = 0.08$), because of massive noise attenuation the data are consistent with approximately 50% of the variance being explained by transcriptional regulation and approximately 50% coming from post-transcriptional regulation, although the low reliability of the data and large sampling variability precludes making this estimate precise. Thus, we next considered analyses that can provide estimates for the scope of post-transcriptional regulation even when the reliability of the data is low.

Coordinated post-transcriptional regulation of functional gene sets

The low reliability of estimates across datasets limits the reliability of estimates of transcriptional and post-transcriptional regulation for individual proteins, Figure 2. Thus, we focused on estimating the post-transcriptional regulation for sets of functionally related genes as defined by the gene ontology (Consortium *et al*, 2004). By considering such gene sets, we may be able to average out some of the measurement noise and see regulatory trends shared by functionally related genes. Indeed, some of the noise contributing to the across-tissue variability of a gene is likely independent from the function of the gene; see Methods. 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.

Since this analysis aims to quantify across-tissue variability, we define the "relative protein to mRNA 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 across the other tissues (Figure 3); we use the KS-test to quantify the statistical significance of differences in the rPTR distributions; see Methods. The results indicate that the genes from many GO terms 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 3a-b).

While the strong functional enrichment of rPTR suggests functionally concerted post-transcriptional

regulation, it can also reflect systematic dataset-specific measurement artifacts. To investigate this possibility, we obtained two estimates for rPTR from independent datasets: Estimate 1 is based on data from Wilhelm *et al* (2014) and Fagerberg *et al* (2014), and Estimate 2 is based on data from Kim *et al* (2014) and Djebali *et al* (2012b). These two estimates are highly reproducible for most tissues, as shown by the correlation between the median rPTR for GO terms in Figure 3d; **Supplementary Fig. 2** shows the reproducibility for all tissues. The correlations between the two rPTR estimates remain strong when computed with all GO terms (not only those showing significant enrichment) as shown in Table S1, as well as when computed between the rPTRs for all genes Table S2.

Consensus protein levels

Given the low reliability of protein estimates across studies Figure 2, we sought to increase it by deriving consensus estimates. Indeed, combining data from both studies can allow to average out some of the noise, and thus provide more reliable consensus estimates; see Methods. As expected for protein estimates with increased reliability, the consensus protein levels correlate better to mRNA levels than the corresponding protein levels estimated from a either dataset alone, Figure 4.

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 1 and **Supplementary Fig. 1**). This underscores the role of transcription for setting the full dynamic range of protein levels. In stark contrast, differences in the proteomes of distinct human tissues are poorly explained by transcriptional regulation, Figure 1. This is due to measurement noise (Figure 2) but also to post-transcriptional regulation. Indeed, large and reproducible rPTR ratios suggest that the mechanisms shaping tissue-specific proteomes involve post-transcriptional regulation, Figure 3. 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.

As with all analysis of empirical data, the results depend on the quality of the data and the estimates of their reliability. This dependence on data quality is particularly strong given that some conclusions rest on the failure of across-tissue mRNA variability to predict across-tissue protein variability. Such inference based on unaccounted for variability is substantially weaker than measuring directly and accounting for all sources of variability. The low across study reliability suggest that the signal is strongly contaminated by noise, especially systematic biases in sample collection and handling, and thus the data cannot accurately quantify the contributions of different regulatory mechanisms, Figure 2. Another limitation of the data is that isoforms of mRNAs and proteins are merged together, i.e., using razor proteins. This latter limitation is common to all approaches quantifying proteins and mRNAs from peptides/short-sequence reads. It stems from the limitation of existing approaches to their to infer isoform and quantify them separately.

The strong enrichment of rPTR ratios within gene sets (Figure 3) demonstrates a functionally concerted regulation at the post-transcriptional level. Some of the rPTR trends can account for fundamental physiological differences between tissue types. For example, the kidney 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 3a). In this case, post-transcriptional regulation very likely plays a functional role in meeting the high energy demands of kidneys.

The rPTR patterns and the across tissue correlations in **Supplementary Fig. 1** 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 strongly on the context. For example transcriptional regulation is contributing significantly to the dynamical responses of dendritic cells (Jovanovic *et al*, 2015) and to the differences between kidney and prostate gland (**Supplementary Fig. 1**b) but less to the differences between kidney and thyroid gland (**Supplementary Fig. 1**a). All data, across all profiled tissues, suggest that post-transcriptional regulation contributes substantially to the across-tissue variability of protein levels. The degree of this contribution 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 (Xue *et al*, 2015; Slavov *et al*, 2015; Preiss, 2016), 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; 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, tissue-type specific proteomes may be shaped by substantial post-transcriptional regulation; in contrast, cell stimulation that preserves the cell-type, may elicit a strong transcriptional remodeling but weaker post-transcriptional remodeling.

Acknowledgments

We thank M. Jovanovic, E. Wallace, J. Schmiedel, and D. A. Drummond for discussions and constructive comments. This work was partially funded by a SPARC grant from the Broad Institute to N.S. and E.A., the Washington Research Foundation Fund for Innovation in Data-Intensive Discovery and the Moore/Sloan Data Science Environments Project at the University of Washington, and NIGMS of the NIH under Award Number DP2GM123497.

Methods

Data and scaled mRNA levels

We used data from Wilhelm *et al* (2014); Kim *et al* (2014); Fagerberg *et al* (2014); Djebali *et al* (2012b) 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. For these genes, about 8% of the mRNA measurements and about 40% of the protein measurements are missing.

First, denote m_{it} the log mRNA levels for gene i in tissue t. Similarly, let p_{it} 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 absolute differences between data from the tissue and the first tissue (arbitrarily chosen as a baseline). That is, for all $t > 1$, we define

$$p^n_{it} = (p^u_{it} - \hat{\mu}_t)$$

with

$$\hat{\mu}_t = \underset{\mu}{\text{argmin}} \sum_i |p^u_{i1} - (p^u_{it} - \mu)|$$

Where p^n_{it} and p^u_{it} represent the normalized and non-normalized protein measurements respectively. For each t, the value of μ_t which minimizes the absolute difference is

$$\hat{\mu}_t = \underset{u}{\text{median}}(p_{i1} - p^u_{it})$$

We use the same normalization for mRNA. This normalization, which corresponds to a location shift of the log abundances for each tissue, corrects for any multiplicative differences in the raw (unlogged) mRNA or protein. We normalize these measurements by aligning the medians rather than the means, as the median is more robust to outliers.

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

$$\hat{T_i} = \underset{t}{\text{median}}(p_{it} - m_{it})$$

In such a situation the log "scaled mRNA" (or mean protein level) can be defined as

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

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

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

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

Across-Tissue Correlations

For each gene, i, we can compute the correlation between mRNA and protein across tissues. Unlike the between gene correlations which are consistently large after scaling for each tissue (Figure 1a), across-tissue correlations are highly variable between genes. Although this could be in part because true mRNA/protein correlations vary significantly between genes, a huge amount of the heterogeneity can be explained sampling variability. There are only between 10 and 12 tissues in common across datasets and for many genes the abundances are missing, which means that the empirical estimates of across tissue correlation for each gene are very noisy. To find a representative estimate of the across-tissue correlation we can take the median over all genes.

As an alternative, if the correlation was roughly constant between genes, we could pool information to yield a representative estimate of this across-tissue correlation. For a gene i, we compute the fisher transformed within gene correlation as $z_i = \text{arctanh}(\hat{r}_i)$ which is approximately normal:

$$z_i \sim N\left(\frac{1}{2}log(\frac{1+\rho}{1-\rho}), \frac{1}{\sqrt{N_i - 3}}\right)$$

where N_i are the number of observed mRNA-protein pairs for gene i (at most 11) and ρ corresponds to the population correlation. We can then easily find the maximum likelihood estimate of the Fisher transformed population correlation by weighting each observation by its variance:

$$\begin{aligned} \omega_i &= \frac{1}{n_i - 3} \\ W_i &= \frac{\omega_i}{\sum_j \omega_j} \\ \hat{z}_{pop} &= \sum W_i z_i \end{aligned}$$

We can then transform this estimate back to the correlation scale

$$\hat{\rho} = \frac{e^{2\hat{z}_{pop}} - 1}{e^{2\hat{z}_{pop}} + 1}$$

Depending on the datasets used, using this method we estimated the population across-tissue mRNA/protein correlation to be between 0.21 for data by Wilhelm *et al* (2014) and 0.29 for data by Kim *et al* (2014). This correlation cannot be used as direct evidence for the relationship between mRNA and protein levels since both mRNA and protein datasets are unreliable due to measurement noise. This measurement noise attenuates the true correlation. Below we address this by directly estimating data reliability and correcting for noise.

Noise Correction

measured mRNA and protein across tissues. 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 (Csárdi *et al*, 2015).

Note that it is simple to show that the empirical variance is the sum of the variance of the signal and the variance of the noise:

- e_i - Expectation at the i^{th} data point; $\tilde{e}_i = e_i - \langle e \rangle$
- ζ_i - Noise at the i^{th} data point; $\langle \zeta \rangle = 0$
- x_i - Observation at the i^{th} data point; $\tilde{x}_i = x_i - \langle x \rangle$, $x_i = e_i + \zeta_i$;

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

Spearman's correction is based on estimates of the "reliability" of the measurements, which is defined as the fraction of total measured variance due to signal rather than to noise:

$$\text{Reliability} = \frac{\sigma^2_{signal}}{\sigma^2_{total}} \quad (1)$$

$$= 1 - \frac{\sigma^2_{err}}{\sigma^2_{err} + \sigma^2_{signal}} \quad (2)$$

If X and Y are noisy measurements of two quantities, we can comptue the noise corrected correlation between them as

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

(4)

In practice, reliabilities are not known but we can often estimate them. In this application, for both mRNA and protein we need measurements in which all steps, from sample collection to level estimation, are repeated independently. In order to estimate the mRNA reliabilities we use independent measurements from Fagerberg *et al* (2014) and Djebali *et al* (2012b). For estimating protein reliabilities we use measurements from Wilhelm *et al* (2014) and Kim *et al* (2014). Across-tissue reliabilities are computed per gene whereas within-tissue reliabilities are computed per tissue across genes. If two independent measurements have the same reliability, it can be estimated by computing the correlation between the two measurements (Spearman, 1904; Zimmerman and Williams, 1997; Csárdi *et al*, 2015). We estimated the approximate across-tissue protein reliability to be 0.21 and the across-tissue mRNA reliability to be 0.77. Given the estimated across-tissue mRNA/protein correlation of 0.29 (calculated using data from Kim *et al* (2014) and Fagerberg *et al* (2014)) we estimated the noise-corrected fraction of across-tissue protein variance explained by mRNA to be approximately 50% 2. Note that if both mRNA or both protein datasets share biases, then the estimated reliabilities will be too small, thus deflating the inferred fraction of protein variance explained by mRNA. Moreover, because the reliabilities are low, sampling variability is large, missing data is prevalent, and mRNA/protein correlation likely vary by gene there is significant uncertainty about this estimate.

Creating an Aggregated Dataset

We use the two independent protein datasets to create a single aggregated dataset which is of arguably higher reliability than either dataset individually. To create this dataset, we take a weighted average of the two protein abundance datasets, by tissue. We compute the weights based on measurement reliabilities for each tissue in each of the two datasets.

Assume we have two random variables, $\widetilde{X}_1$ and $\widetilde{X}_2$, corresponding to measurements on the same quantity (e.g. two independent protein measurements) with $\widetilde{X}_i = X + \epsilon_i$ where $X \sim N(0, \sigma_X^2)$ is the signal which is independent of $\epsilon_i \sim N(0, \sigma_{\epsilon_i}^2)$, the measurement error for sample i. We have a third random variable corresponding to a different quantity (e.g. an mRNA measurement), $\widetilde{Y}$ that is typically positively correlated with $\widetilde{X}_1$ and $\widetilde{X}_2$ with the same covariance σ_{XY}^2. To create the aggregated dataset we first compute the reliability of $\widetilde{X}_i$ $Rel(\widetilde{X}_i) = \frac{\sigma_X^2}{\sigma^2_{\widetilde{X}_i}} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_{\epsilon_i}^2}$ for both datasets.

Note that

$$\begin{aligned} \text{Cor}(\widetilde{X}_1, \widetilde{X}_2) &= \frac{\sigma_X^2}{\sigma_{\widetilde{X}_1}\sigma_{\widetilde{X}_2}} \\ \text{Cor}(\widetilde{X}_i, Y) &= \frac{\sigma_{XY}^2}{\sigma_{\widetilde{X}_i}\sigma_Y} \end{aligned}$$

Thus,

$$\begin{aligned} \text{Cor}(\widetilde{X}_1, \widetilde{X}_2)\frac{\text{Cor}(\widetilde{X}_1, \widetilde{Y})}{\text{Cor}(\widetilde{X}_2, \widetilde{Y})} &= \frac{\sigma_X^2}{\sigma^2_{\widetilde{X}_1}} \\ &= \frac{\sigma_X^2}{\sigma_X^2 + \sigma_{\epsilon_1}^2} \\ &= Rel(\widetilde{X}_1) \end{aligned}$$

Similarly, $\text{Cor}(\widetilde{X}_1, \widetilde{X}_2)\frac{\text{Cor}(\widetilde{X}_2, \widetilde{Y})}{\text{Cor}(\widetilde{X}_1, \widetilde{Y})} = \text{Rel}(\widetilde{X}_2)$. We use these facts and compute the empirical correlations between datasets to independently estimate the across gene reliabilites for each tissue from each dataset. We then Fisher weight the protein abundances based on their reliabilities. That is,

for each tissue t, the aggregated dataset, X_{At} is

$$\begin{aligned} X_{At} &= w\widetilde{X}_{1t} + (1-w)\widetilde{X}_{2t} \\ w &= \frac{Rel(\widetilde{X}_{1t})}{Rel(\widetilde{X}_{1t}) + Rel(\widetilde{X}_{2t})} \end{aligned}$$

When the reliability of $\widetilde{X}_{1t}$ and $\widetilde{X}_{2t}$ are close, each dataset is weighted equally. When one reliability dominates the other, that dataset contributes more to the aggregated dataset. We found that the full aggregated dataset has a higher median per gene correlation with mRNA than either of the protein datasets individually (0.34) which suggests that it may be of higher reliability. When computing the per tissue reliabilities, we found that the reliabilities of the lung and pancreas datasets from Wilhelm *et al* (2014) had lower reliability relative to the data from Kim *et al* (2014). This offers a partial explanation for why the independent estimates of the rPTR ratios for these tissues did not have a significant positive correlation (Figure 3d).

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:

$$\hat{r}_{it} = p_{it} - \underset{t' \neq t}{\text{median}}(p_{it'} - m_{it'})$$

To exclude the possibility that $\hat{r}_{it} = 0$ exactly, we require that $t' \neq t$. When the estimated rPTR is larger than zero, the measured protein level in tissue t is larger than the estimated mean protein level. Likewise, when this quantity is smaller than zero, 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 error 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

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

where i_1 to i_{n_k} index the genes in set k and t 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 (t) relative to all other tissues.

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

$$\rho_{kt} = \text{KS}(\mathcal{G}_{kt}, \bigcup_{t' \neq t} \mathcal{G}_{kt'})$$

To correct for testing multiple hypotheses, we computed the false discovery rate (FDR) for all gene sets in tissue t (Storey, 2003). In Figure 3a-c, we present only the functional groups with FDR less than 1% and report their associated p-values. Note that the test statistics for each gene set are positively correlated since the genesets are not disjoint, but Benjamini *et al* (2001) prove that the Benjamini-Hochberg procedure applied to positively correlated test statistics still controls FDR. Thus, the significance of of certain functional groups suggests that not all of the variability in rPTR is due to measurement noise. We also calculated rPTR using two pairs of measurements: one set of rPTR estimates was calculated using protein data from Wilhelm *et al* (2014) and mRNA from Fagerberg *et al* (2014) and the other was calculated using data from Kim *et al* (2014) and Djebali *et al* (2012b). rPTR of the significant sets was largely reproducible across estimates from independent datasets (Figure 3d and **Supplementary Fig. 2**) and across genes (Table S2).

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

Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I (2001) Controlling the false discovery rate in behavior genetics research. *Behavioural brain research* **125**: 279–284

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

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* (2012a) Landscape of transcription in human cells. *Nature* **489**: 101–108

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, *et al* (2012b) 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

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

Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, *et al* (2014) A draft map of the human proteome. *Nature* **509**: 575–581

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

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

Preiss T (2016) All Ribosomes Are Created Equal. Really? *Trends in biochemical sciences* **41**: 121–123

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 (2014) 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

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 (2015) Differential stoichiometry among core ribosomal proteins. *Cell Reports* **13**: 865 – 873

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

Spearman C (1904) The proof and measurement of association between two things. *Am J Psychol* **15**: 72–101

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

Xue S, Tian S, Fujii K, Kladwang W, Das R, Barna M (2015) RNA regulons in Hox 5'-UTRs confer ribosome specificity to gene regulation. *Nature* **517**: 33–38

Zimmerman D, Williams R (1997) Properties of the spearman correction for attenuation for normal and realistic non-normal distributions. *Applied Psychological Measurement* **21**: 253270

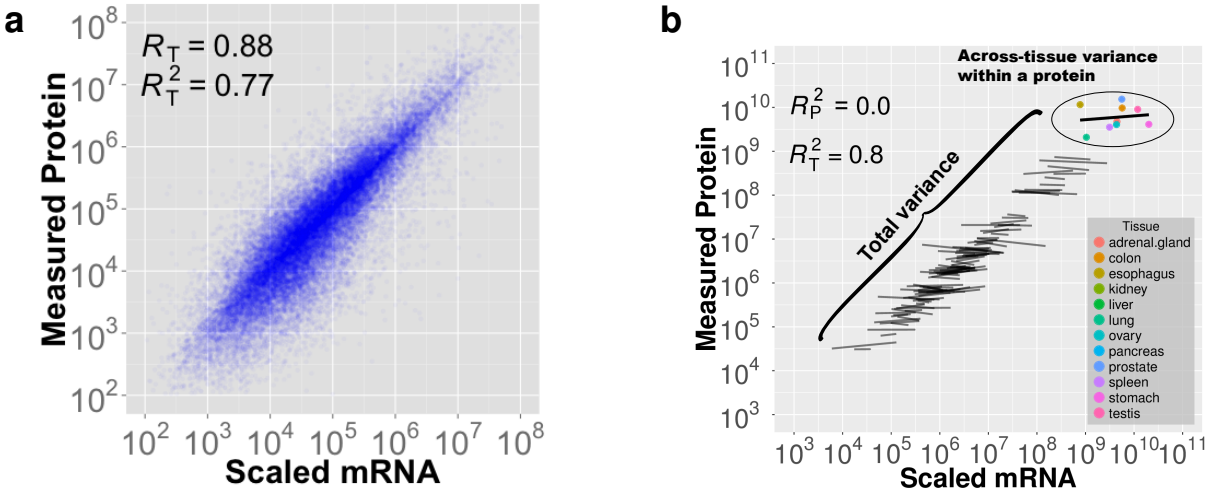

Figure 1. The fraction of total protein variance explained by scaled mRNA levels is not informative about the across-tissue variance explained by scaled mRNA levels. (**a**) mRNA levels scaled by the median protein-to-mRNA (PTR) ratio correlate strongly with measured protein levels ($R_T^2 = 0.77$ over 6104 measured mRNAs and 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 and across-tissue variability. 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, mRNA levels scaled by the median PTR ratio are *not* correlated to the corresponding measured protein levels ($R_P \approx 0$). See Supporting Information and **Supplementary Fig. 1**.

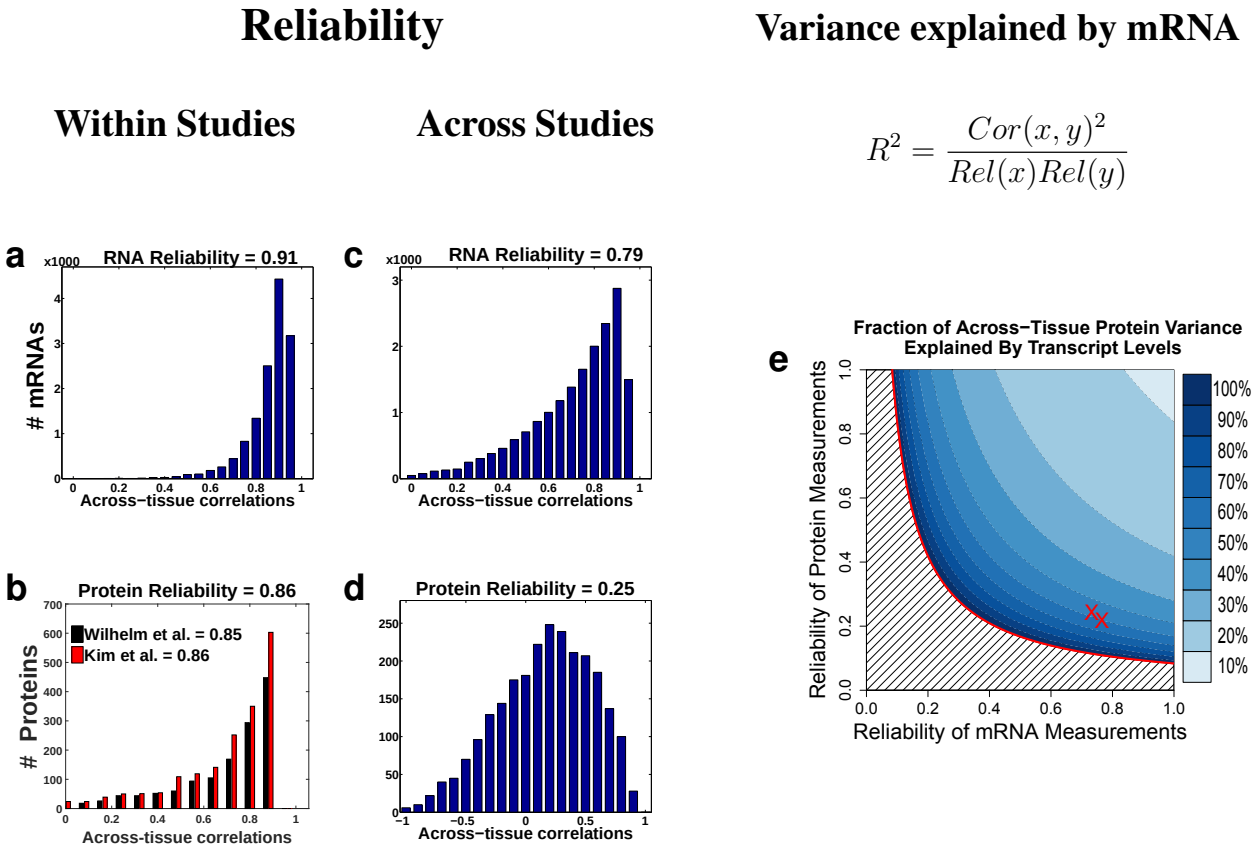

Figure 2. Data reliability crucially influences estimates of transcriptional and post-transcriptional regulation across-tissue. (**a**) The within-study reliability – defined as the fraction of the measured variance due to the signal – of relative mRNA levels is estimated as the correlation between the mRNA levels measured in the twelve different tissues. Replicate estimates from (Fagerberg *et al*, 2014) for the levels of each transcript were correlated (averaging across the 12 tissue-types) and the results for all analyzed transcripts displayed as a distribution. (**b**) The within-study reliability of relative protein levels is estimated as the correlation between the protein levels measured in 12 different tissues by Wilhelm *et al* (2014). Separate estimates for each protein were derived from non-overlapping sets of peptides and were correlated (averaging across the 12 tissue-types) and the results for all analyzed proteins displayed as a distribution; see Methods. (**c**) The across-study reliability of mRNA was estimated by correlating estimates as in (a) but these estimates came from different studies (Djebali *et al*, 2012b) and (Fagerberg *et al*, 2014). (**d**) The across-study reliability of proteins was estimated by correlating estimates as in (b) but these estimates came from different studies (Wilhelm *et al*, 2014) and (Kim *et al*, 2014). (**e**) The fraction of across-tissue protein variance that can be explained by mRNA levels is plotted as a function of the reliability of the estimates of mRNA and protein levels, given an mRNA/protein correlation of 0.29. The red Xs correspond to two estimates of reliability of the mRNA and protein measurements computed from both independent mRNA and protein datasets.

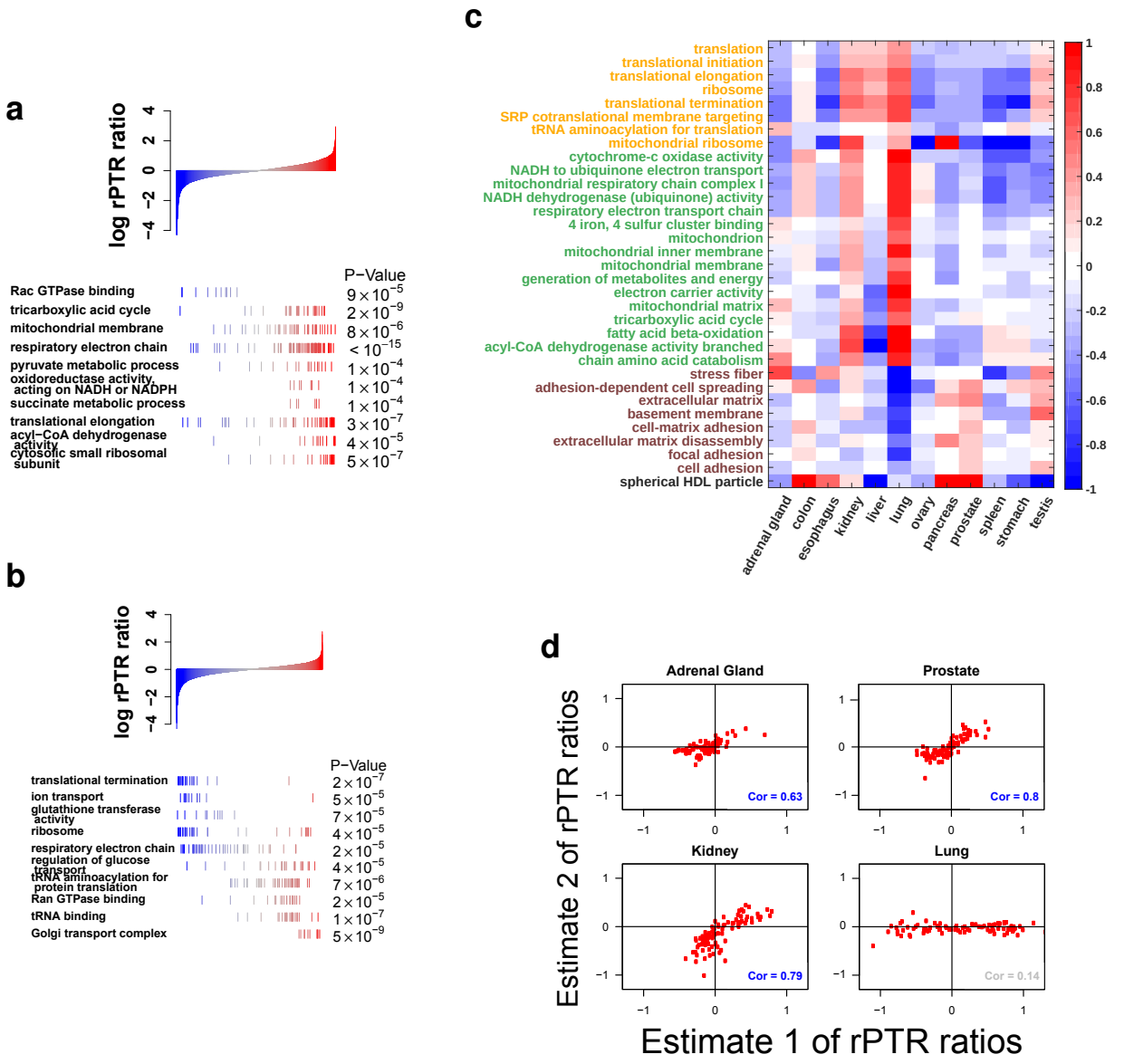

Figure 3. Concerted and reproducible variability in the relative protein-to-RNA (rPTR) ratio of functional gene-sets across tissue-types (**a**) mRNAs coding for the small ribosomal subunit, NADH dehydrogenase and respiratory proteins have much higher protein-to-mRNA ratios in kidney as compared to the 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. All data are displayed on a $\log_{10}$ scale, and functionally related gene-sets are marked with the same color. (**d**) The reproducibility of rPTR estimates across estimates from different studies is estimated as the correlation between the median rPTRs for GO terms showing significant enrichment as shown in panels (a-c). See Supporting Information and **Supplementary Fig. 2**.

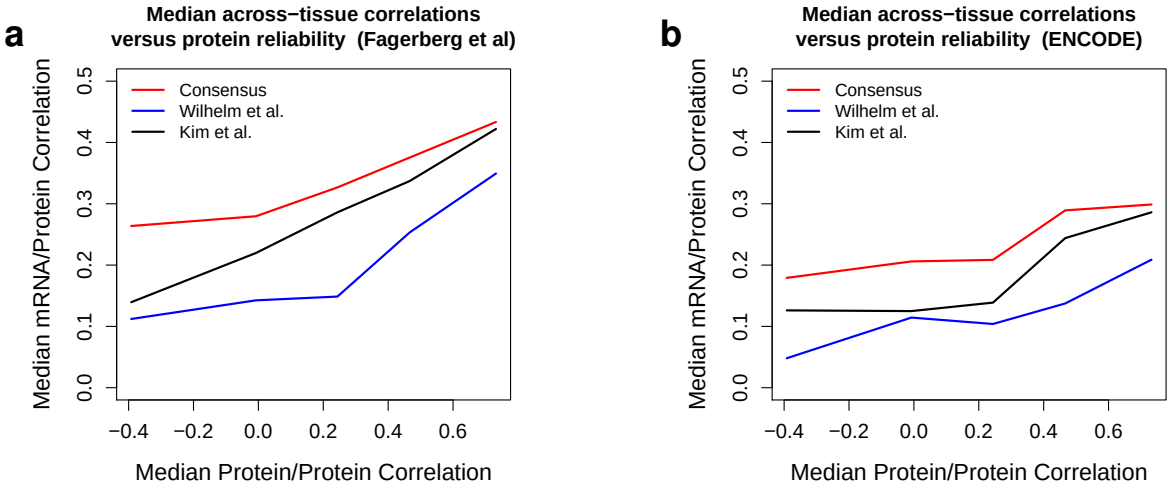

Figure 4. Deriving a consensus protein dataset for improved quantification of human tissue proteomes We compiled a consensus protein dataset my merging data from Wilhelm *et al* (2014) and Kim *et al* (2014) as described in Methods. The relative protein levels estimated from (Wilhelm *et al*, 2014), (Kim *et al*, 2014), and the consensus dataset were correlated to mRNA levels from Fagerberg *et al* (2014) (**a**) or to mRNA levels from (Djebali *et al*, 2012b) (**b**). The correlations are shown as a function of the median correlation between protein estimates from Wilhelm *et al* (2014) and Kim *et al* (2014). The consensus dataset exhibits the highest correlations, suggesting that it has averaged out some of the noise in each dataset and provides a more reliable quantification of of human tissue proteomes.

Tables

	adrenal	colon	esophagus	kidney	liver	lung	ovary	pancreas	prostate	testis
Corr.	0.38	0.33	0.14	0.34	0.16	0.14	0.07	0.18	0.39	0.27
Lower	0.34	0.30	0.09	0.31	0.12	0.10	0.03	0.13	0.36	0.24
Upper	0.41	0.36	0.19	0.38	0.20	0.18	0.10	0.22	0.42	0.31

Table S1. Estimates of relative protein-to-RNA (rPTR) ratio for GO terms reproduce across different datasets Pearson correlations between two estimates of the median rPTR ratios for all GO terms indicate reproducible effects in all tissues. As in Figure 2, rPTR estimates are derived using independently data sources. The lower and upper estimates are the endpoints of the 95% confidence interval.

	adrenal	colon	esophagus	kidney	liver	lung	ovary	pancreas	prostate	testis
Corr.	0.39	0.36	0.18	0.32	0.49	0.18	0.07	0.22	0.38	0.27
Lower	0.36	0.33	0.13	0.28	0.46	0.14	0.03	0.18	0.35	0.24
Upper	0.42	0.39	0.22	0.35	0.52	0.22	0.10	0.27	0.42	0.31

Table S2. Estimates of relative protein-to-RNA (rPTR) ratio for genes reproduce across different datasets Correlations between the two estimates of rPTR ratios for all genes indicate reproducible effects in all tissues. The rPTR ratios were estimated independently from different datasets (as in Figure 2). The lower and upper estimates are the endpoints of the 95% confidence interval.

Supplementary Figures

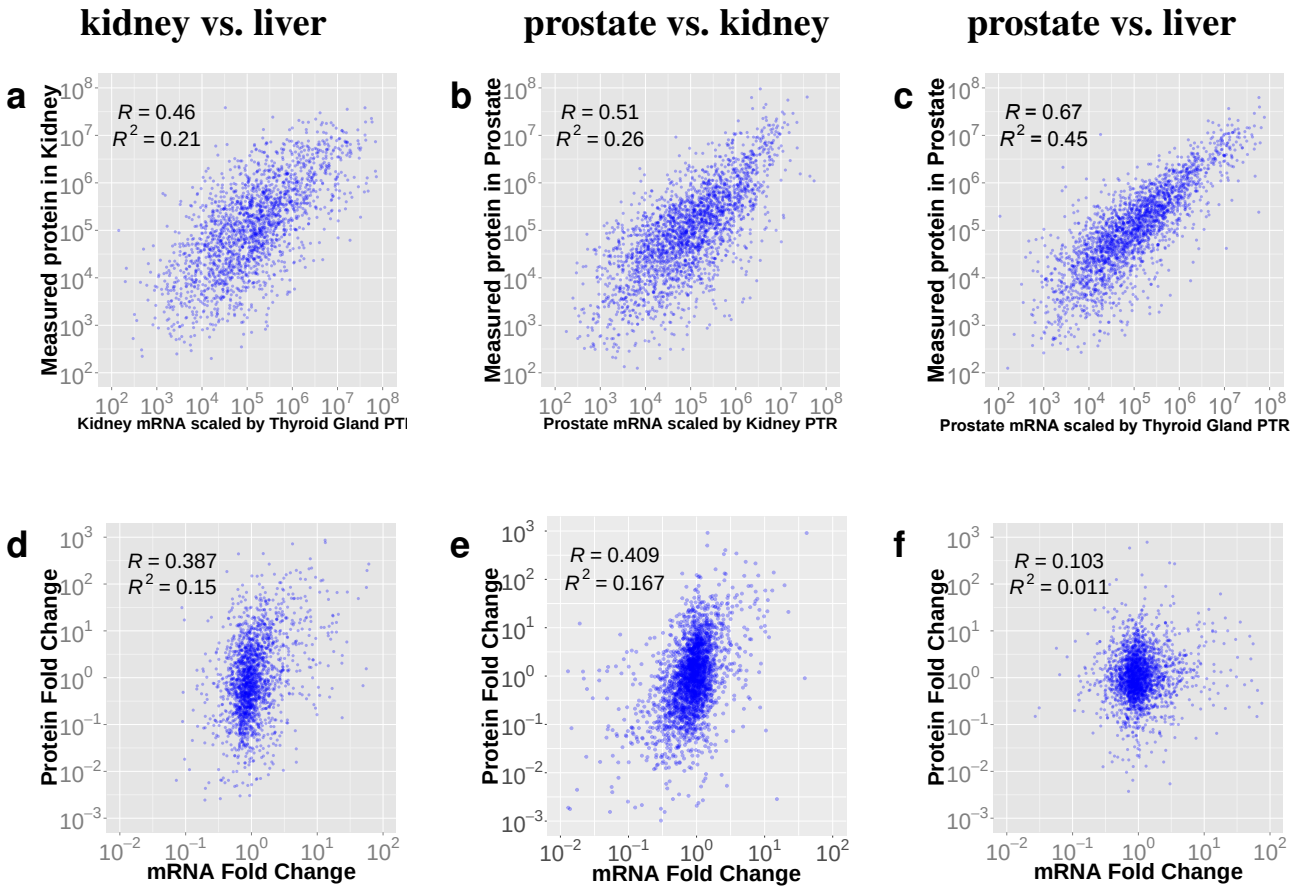

Figure S1. The total protein variance explained by scaled mRNA levels is not indicative of the correlations between mRNA and protein fold-changes across the corresponding tissue pairs. While scaled mRNA is predictive of the absolute protein levels (a-c, top row), the accuracy of these predictions does not generally reflect the accuracy of protein fold-changes across tissues that are predicted from the corresponding mRNA fold-changes (d-f, bottom row).

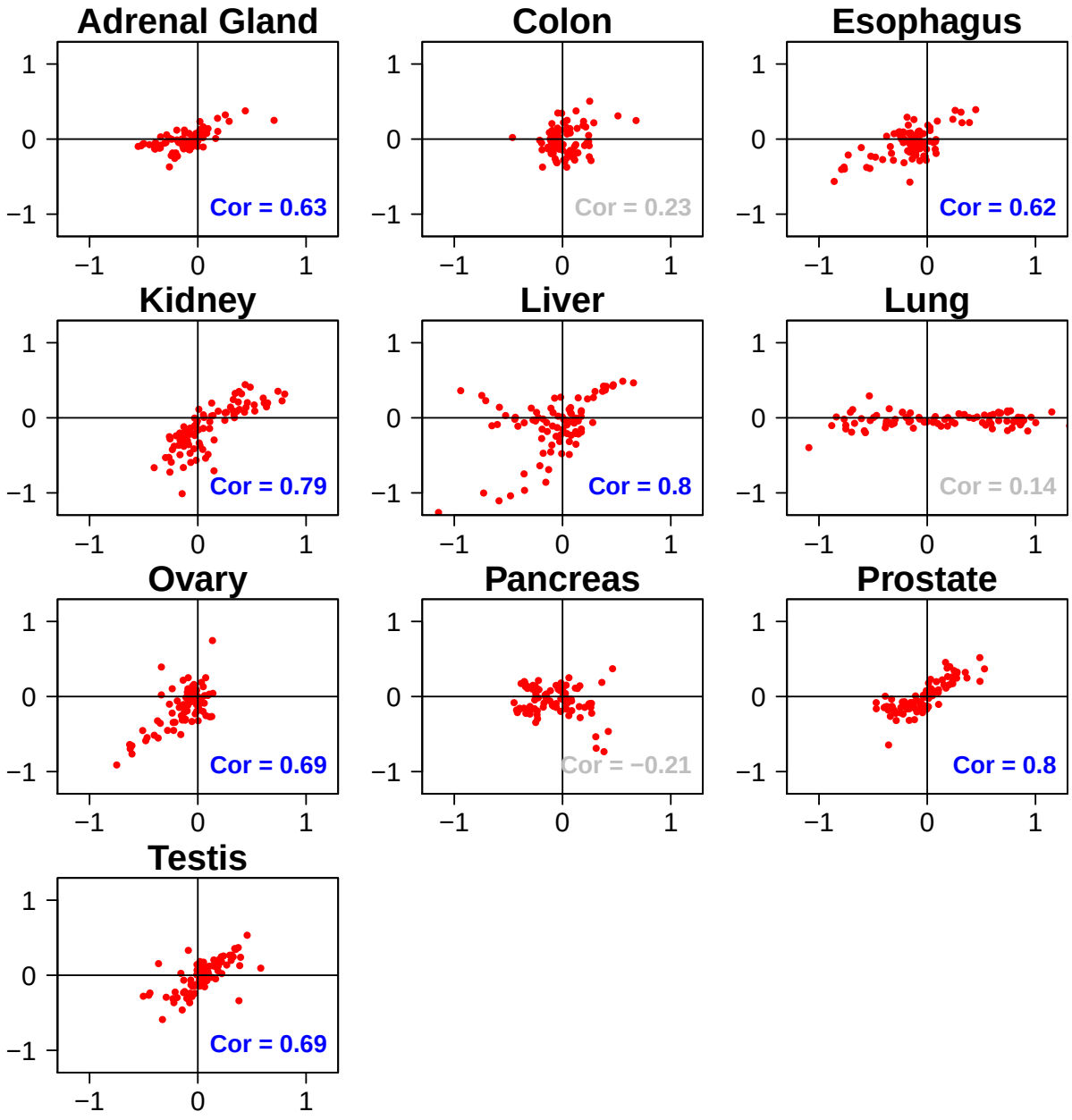

Figure S2. Reproducibility of rPTR ratios estimated from different datasets