Can we predict protein from mRNA levels?

ARISING FROM Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–587 (2014).

Nikolaus Fortelny[1,2,3], Christopher M. Overall[1,3,4], Paul Pavlidis[2,5], Gabriela V. Cohen Freue[6,*]

[1] Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, British Columbia, Canada
[2] Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia, Canada.
[3] Centre for Blood Research, University of British Columbia, Vancouver, British Columbia, Canada
[4] Department of Oral Biological and Medical Sciences, University of British Columbia, Vancouver, British Columbia, Canada
[5] Department of Psychiatry, University of British Columbia, Vancouver, British Columbia, Canada.
[6] Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada.

* To whom correspondence should be addressed:
Prof Gabriela V. Cohen Freue, Department of Statistics, University of British Columbia, 3146 Earth Sciences Building, 2207 Main Mall, Vancouver BC, V6T 1Z4, Canada.
Email: gcohen@stat.ubc.ca
Phone: +1-604-822-3710

Prediction of protein levels from mRNA levels has long been fraught with unreliability and lack of precision. However, Wilhelm *et al.*[1] claimed that using estimated gene-specific translation rates together with mRNA levels accurately predicts protein levels in any given tissue, reporting correlations of ~0.9 between predictions and measurements across genes. Here we show that these correlations greatly overestimate the accuracy of per-gene predictions. Using simple and standard statistical evaluation methods, we demonstrate that the gene-specific translation rates estimated by Wilhelm *et al.* are, in general, not useful to predict protein levels from mRNA levels, with a median correlation of 0.21.

Wilhelm *et al.* reported impressive correlations of ~0.9 between predictions and measurements of protein levels (0.91 for Salivary gland and a median of 0.87 from the 12 tissues). From these results the authors concluded that protein abundance in any given tissue can be predicted with good accuracy from the gene's mRNA levels. This is a striking claim because numerous known biological mechanisms exist that decouple protein levels from mRNA levels and need to be considered to predict protein levels [2,3]; yet Wilhelm *et al*.'s results suggest that these mechanisms are negligible. In addition, gene-specific correlations between protein and mRNA levels are far below 0.9 in their data for most genes. This apparent contradiction is resolved by noting that Wilhelm *et al*.'s performance measures were based on the study of the correlations between predicted and measured protein levels *across* genes, whereas their predictions were obtained *within* genes.

The key claim underpinning Wilhelm *et al*.'s[1] interpretations is that the ratio of protein to mRNA levels remains remarkably conserved across tissues for any given gene (at steady state). Indeed, if this ratio (r_g, the translation rate) was a constant, protein levels for a gene *g* in any tissue *t* ($prot_{g,t}$) would be accurately predictable from mRNA ($mRNA_{g,t}$) by using the relation $prot_{g,t} = r_g * mRNA_{g,t}$ suggested by Wilhelm et al. However, since the gene-specific translation rates r_g are unknown, Wilhelm *et al.* estimated them with the median of the per-tissue ratios. This approach is distinct from measuring the translation rates as independent variables to predict protein levels[4,2]. Thus, at the gene level, the only predictor in Wilhelm *et al*. method is mRNA.

Having estimated a gene-specific translation rate and using it to predict protein levels from mRNA levels for each gene, the natural question is how well the given relation

works for each gene. However, this crucial question was not addressed as the authors only evaluated their method by looking at the correlation between the predicted and the measured protein values *across-genes* for each tissue (*e.g.*, Figure 5a, lower right in Wilhelm *et al*.). Thus, they neither quantitatively examined their claim of the constant ratio of protein to mRNA levels nor the accuracy of their predictions on an individual per-gene basis (i.e., *within-genes*).

We demonstrate the problem with their analysis with two control experiments (Figure 1a). In the first control, for every gene *g*, we predict protein levels in all tissues as the median of protein levels of *g* across all 12 tissues *without* using any mRNA data ("mRNA-free"; equivalent to setting mRNA to the constant 1 in all samples; thus $prot_{pred} = r_g = median(prot_{obs})$). In the second control, for every gene *g*, we predict protein levels using Wilhelm *et al*.'s method, but replacing the mRNA values of gene *g* with those of a randomly-selected nonmatching gene. Following Wilhelm *et al.*'s method, we use the (random) mRNA values to estimate the translation rate and to predict protein levels ("Random genes"). The correlations *across* genes for these two controls are 0.84 ("mRNA-free") and 0.83 ("Random genes"), compared to 0.87 (Wilhelm *et al*.'s results) when the true matching mRNA levels of each gene are used throughout the prediction (median across tissues). Thus, we show that it is in fact possible to achieve a high correlation *across* genes without using any mRNA levels and translation rate; or by using the wrong mRNA data to estimate the translation rate and predict protein levels.

The explanation of this result is that these three high (across-gene) correlations, and in particular those obtained by Wilhem *et al.*'s method (median of 0.87) are driven by the large degree of variation in protein levels *between* genes. Thus, the high correlations reported by Wilhelm *et al*. do not merely reflect the accuracy of the predictions (Figures S1 and S2). The between-gene variation greatly exceeds the within-gene variation (mean of per-gene variances across tissues equals $3.2*10^{-6}$ and mean of per-tissue variances across genes equals $1.4*10^{-5}$) (Figure 1b). This generates a high correlation between predicted and observed protein levels *across* genes (median of 0.87) even when these correlations are low for individual genes (see also ellipses in Figure 1b), an effect similar to the Simpson's paradox[5,6]. Thus, the correlations studied by Wilhelm *et al*. are uninformative about the performance of their method and the validity of their constant ratios claim.

An appropriate approach to evaluate Wilhelm *et al*.'s per-gene method is to measure the correlation between predicted and observed protein levels *within each gene and across tissues*[7]. We note that Wilhelm *et al.* used the median ratio to estimate the translation rate and predict protein levels of all genes, even of those with almost invariant mRNA and protein levels. Thus, all within-gene correlations between their predictions and the measured protein levels must be evaluated. These correlations are low for most genes (median correlation 0.21, Figure 1c), indicating that the gene-specific translation rates estimated by the authors together with mRNA levels form, in general, a poor predictor of protein abundance levels. Further, these results also suggest that the ratios of mRNA and protein are not constant for most genes. To help visualize individual per-gene mRNA and protein data together with protein predictions reported by Wilhelm *et al.*, we built the accompanying web application for 5895 genes (https://dakep.shinyapps.io/central-dogma/).

In an recent review Liu, Beyer, and Aebersold[3] emphasize that the difference between across-gene and within-gene correlations of observed mRNA and protein levels are a potential point of confusion. Our contribution emphasizes that to evaluate gene-specific predictions, one must consider gene-specific accuracy measures. In particular, across-gene and within-gene correlations of *predicted* and observed protein levels have distinct interpretations as well and have often been confused in the literature. Analyses proposing gene-specific predictions but only evaluated across genes[1,8,9] must be reconsidered using evaluations *within* genes instead. While it is conceivable that additional data on other factors influencing protein levels (*e.g.*, degradation rates) will permit more accurate predictions, the current data does not support high accuracy for most genes when using mRNA alone.

Figure Legends

Figure 1

(a) Correlation between observed mRNA and protein levels (mRNA), or predicted and observed protein levels (predictions) across genes, one correlation per tissue, as measured in the original publication. (b) Predicted and observed protein values of four example genes (indicated by color, n=12 tissues per gene). Correlation *across* genes is high because the variation between genes largely exceeds the within-gene variation (illustrated with ellipses). Correlation across tissues within-gene is low. The grey line corresponds to the 45º line. (c) Histogram of Spearman correlation of predicted to observed protein levels across samples (tissues) resulting in one value per gene.

Figure S1

(**a**) mRNA and protein in simulated data for three genes (colors) in five tissues. The data points for one tissue are highlighted and the error from the ratio-based prediction is indicated. (**b**) Predicted and observed protein in simulated data for three genes (colors) in one tissue from (**a**). The error in the prediction is indicated by the distance of the point to the 45 degree line. (**c**) mRNA (open symbols) and predicted protein (solid symbols) on the x-axis and observed protein on the y-axis. The plot shows real data of four example genes. Data points from one tissue and their modification by Wilhelm *et al.*'s prediction are indicated by an error.

Figure S2

mRNA and protein in simulated data for three genes (colors) in five tissues. Three models are shown by grey lines: (**a**) a model predicting protein levels from mRNA levels per gene as in Wilhelm et al. (**b**) predicting protein levels per gene without using mRNA.

References

1. Wilhelm, M. et al. Mass-spectrometry-based draft of the human proteome. Nature 509, 582–587 (2014).
2. Li, J. J. & Biggin, M. D. Statistics requantitates the central dogma. Science 347, 1066–1067 (2015).
3. Liu, Y. & Aebersold, R. The interdependence of transcript and protein abundance: new data–new complexities. Mol. Syst. Biol. 12, 856–856 (2016).
4. Li, J. J., Bickel, P. J. & Biggin, M. D. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ 2, e270 (2014).
5. Friendly, M., Monette, G. & Fox, J. Elliptical Insights: Understanding Statistical Methods through Elliptical Geometry. Stat. Sci. 28, 1–39 (2013).
6. Berman, S., DalleMule, L., Greene, M., and Lucker, J. Simpson's Paradox: a cautionary tale in advanced analytics. Significance. (Wiley-Blackwell, 2012).
7. Hocking, R. R. Methods and Applications of Linear Models: Regression and the Analysis of Variance. (John Wiley and Sons, 2013).
8. Schwanhäusser, B. et al. Global quantification of mammalian gene expression control. Nature 473, 337–342 (2011).
9. Edfors, F. et al. Gene-specific correlation of RNA and protein levels in human cells and tissues. Mol. Syst. Biol. 12, 883 (2016).

Competing Financial Interests. Declared none.

Author Contributions NF and GCF designed the research project. CMO and PP contributed to the design of the project and provided funding. NF designed the manuscript and analyzed the data. PP and GCF wrote the paper. GCF supervised the research project. All authors contributed to extensive discussions and revisions of all drafts of the paper.