Can we predict protein from mRNA levels?

ARISING FROM M. Wilhelm *et al. Nature* **509,** 582–587 (2014); doi:10.1038/nature13319

Prediction of protein levels from mRNA levels has long been fraught with unreliability and a 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 approximately 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. There is a Reply to this Comment by Wilhelm, M. *et al. Nature* **547,** http://dx.doi.org/10.1038/nature23294 (2017).

Wilhelm *et al.* reported impressive correlations of approximately 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]; however, the results of Wilhelm *et al.*[1] 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 the performance measures of Wilhelm *et al.*[1] 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 the interpretation of Wilhelm *et al.*[1] 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) were a constant, protein levels for a gene g in any tissue t ($\text{prot}_{g,t}$) would be accurately predictable from mRNA ($\text{mRNA}_{g,t}$) by using the relation $\text{prot}_{g,t} = r_g \times \text{mRNA}_{g,t}$ suggested by Wilhelm *et al.* However, as 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[2,4]. Thus, at the gene level, the only predictor in the 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 evaluated their method only by looking at the correlation between the predicted and the measured protein values across genes for each tissue (for example, figure 5a, lower right in Wilhelm *et al.*[1]). Thus, they quantitatively examined neither their claim for a constant ratio of protein to mRNA levels nor the accuracy of their predictions on an individual per-gene basis (that is, within genes).

We demonstrate the problem with their analysis with two control experiments (Fig. 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 (called mRNA-free in Fig. 1a; equivalent to setting mRNA to the constant 1 in all samples; thus $\text{prot}_{\text{pred}} = r_g = \text{median}(\text{prot}_{\text{obs}})$). In the second control, for every gene g, we predict protein levels using the method of Wilhelm *et al.*[1], but replacing the mRNA values of gene g with those of a randomly selected, nonmatching gene. Following the method in ref. 1, we use the (random) mRNA values to estimate the translation rate and to predict protein levels (called Random genes in Fig. 1a). The correlations across genes

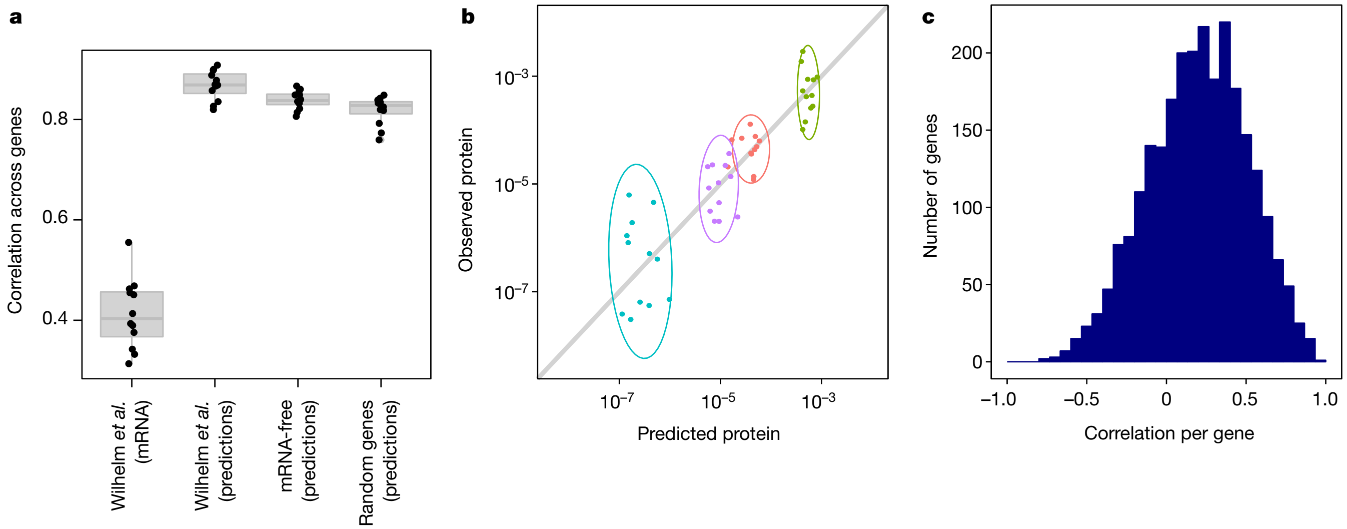


Figure 1 | Across-genes correlations versus within-genes correlations between observed and predicted protein levels. a, Boxplots of the correlations across genes, one correlation per tissue (overlapped points), between observed mRNA and protein levels (Wilhelm *et al.* (mRNA)), and between predicted and observed protein levels as measured in the original publication (Wilhelm *et al.* (predictions)) and in our control experiments (mRNA-free (predictions)) and Random genes (predictions)). **b**, Predicted and observed protein values of four example genes (indicated by colour, $n = 12$ 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 genes is low. The grey line corresponds to the 45° line. **c**, Histogram of Spearman correlations of predicted-to-observed protein levels across samples (tissues), resulting in one value per gene.

are 0.84 for the mRNA-free control and 0.83 for the Random genes control, compared to 0.87 in the results of Wilhelm *et al.*[1] 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 for this result is that these three high (across-gene) correlations, and in particular those obtained by the method of Wilhelm *et al.*[1] (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 (Extended Data Figs 1, 2). 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}) (Fig. 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 Fig. 1b), an effect similar to Simpson's paradox[5,6]. Thus, the correlations studied by Wilhelm *et al.*[1] are uninformative about the performance of their method and the validity of their constant ratios claim.

An appropriate approach to evaluate the per-gene method of Wilhelm *et al.*[1] 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, Fig. 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. Furthermore, 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 the protein predictions reported by Wilhelm *et al.*, we built the accompanying web application for 5,895 genes (https://dakep.shinyapps.io/central-dogma/).

In a recent review, Liu, Beyer, and Aebersold[3] emphasized 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 evaluated only across genes[1,8,9] must be reconsidered using evaluations within genes instead. While it is conceivable that additional data on other factors that influence protein levels (for example, degradation rates) will permit more accurate predictions, the current data do not support high accuracy for most genes when using mRNA alone.

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.
email: gcohen@stat.ubc.ca

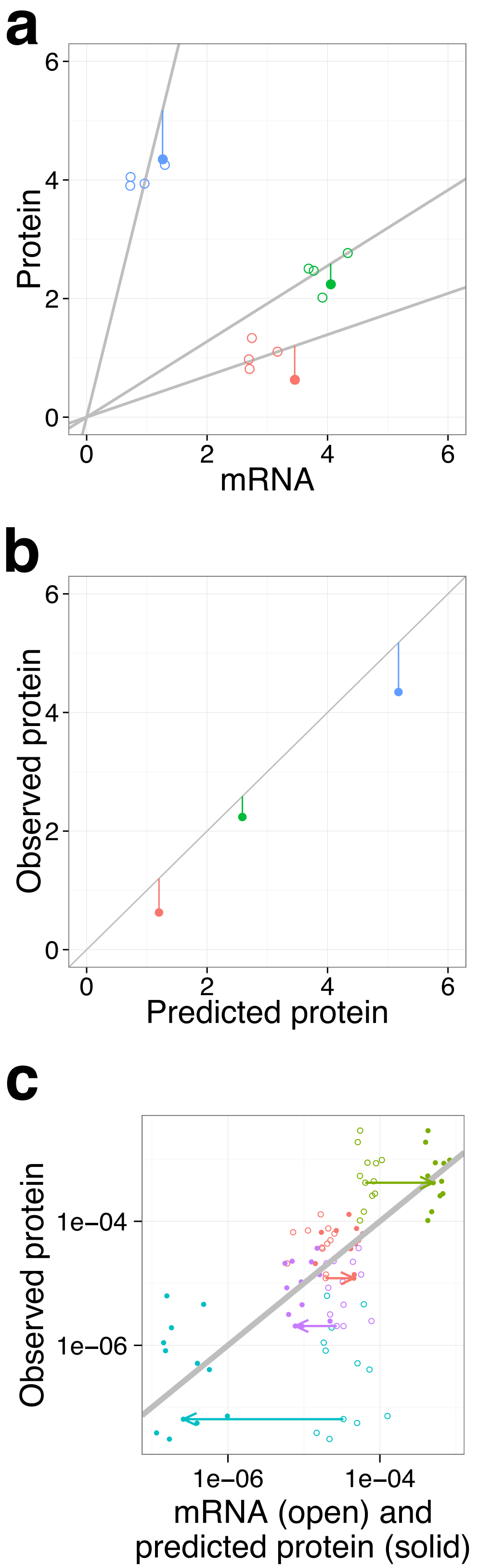
Received 2 November 2016; accepted 1 June 2017.

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. Gene expression. Statistics requantitates the central dogma. Science **347,** 1066–1067 (2015).
3. Liu, Y., Beyer, A. & Aebersold, R. On the dependency of cellular protein levels on mRNA abundance. *Cell* **165,** 535–550 (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. & Lucker, J. Simpson's paradox: a cautionary tale in advanced analytics. *Significance* https://www.statslife.org.uk/the-statistics-dictionary/2012-simpson-s-paradox-a-cautionary-tale-in-advanced-analytics (2012).
7. Hocking, R. R. *Methods and Applications of Linear Models: Regression and the Analysis of Variance* (Wiley, 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).

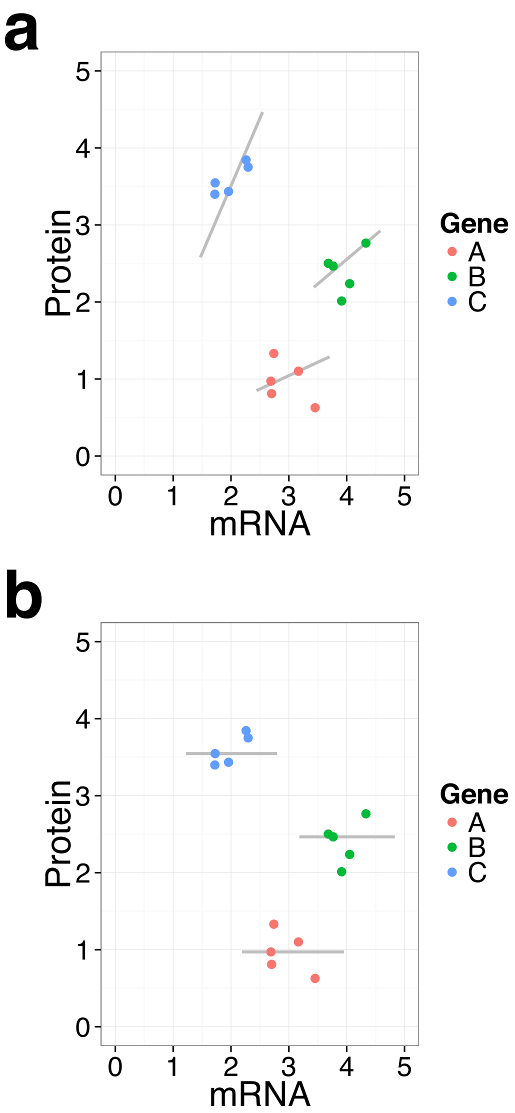
Author Contributions N.F. and G.V.C.F. designed the research project. C.M.O. and P.P. contributed to the design of the project and provided funding. N.F. designed the manuscript and analysed the data. P.P. and G.V.C.F. wrote the paper. G.V.C.F. supervised the research project. All authors contributed to extensive discussions and revisions of all drafts of the paper.

Competing Financial Interests Declared none.

doi:10.1038/nature23293



Extended Data Figure 1 | Relation between gene-specific protein predictions and observed protein levels. a, mRNA and protein in simulated data for three genes (colours) 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 (colours) in one tissue from **a**. The error in the prediction is indicated by the distance from the point to the 45° 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 for four example genes. Data points from one tissue and their modification by the prediction of Wilhelm *et al.*[1] are indicated by an error.



Extended Data Figure 2 | mRNA contribution to protein prediction. mRNA and protein in simulated data for three genes (A, B, and C, colours) in five tissues. **a**, Three gene-specific models (grey lines) to predict protein levels from mRNA levels as in Wilhelm *et al.* **b**, Three gene-specific models (grey lines) to predict protein levels without using mRNA.

Wilhelm *et al.* reply

REPLYING TO N. Fortelny, C. M. Overall, P. Pavlidis & G. V. Cohen Freue *Nature* **547,** http://dx.doi.org/10.1038/nature23293 (2017)

In the accompanying Comment, Fortelny *et al.*[1] present a re-analysis of a particular aspect of our draft human proteome[2], notably our claim that "Having learned the protein/mRNA ratio for every protein and transcript, it now becomes possible to predict protein abundance in any given tissue with good accuracy from the measured mRNA abundance." Their key criticism is that the correlation analysis used at the time "…greatly overestimates the accuracy of per-gene predictions," and hence concluded that "…the current data do not support high accuracy when using mRNA alone." While we agree with parts of the analysis, we do not agree with all of the conclusions.

First, the controversy may have arisen from our use and interpretation of the word 'prediction'. In statistics, prediction is always done within the experimental unit (for example, the liver) and allows statements about (relative) protein abundance variation between biological replicates of (many) liver samples. This is neither what we did nor what we meant to imply, because our data did not contain any replicates of the same tissue and the proteomic and transcriptomic data originated from different samples. Instead, our analysis was designed to estimate (perhaps the better word in this context) the absolute abundance of proteins for tissues for which no proteomic data are available. Given the ease of obtaining transcriptomic profiles of tissues, we still think this is a useful and practical approach.

Second, the accuracy with which absolute protein levels can be estimated by our approach depends on the technical variation in the data. As we analysed in extended data figure 5 of the original publication[2] on the basis of stable isotope-labelled and absolutely quantified peptides, the median fold error within the assembled (heterogeneous) proteomic data is about 3. The average median absolute deviation of the protein/mRNA ratios is about 2.8. Therefore, the technical variation in our data limits the accuracy and precision with which absolute protein abundance can be estimated and may not suffice to determine variations in protein abundance within biological replicates of the same tissue.

Third, it has been shown that the number of proteins with tissue-specific expression is surprisingly low[3] and that many proteins show similar absolute expression values across different tissues and cell lines. This explains why Fortelny *et al.*[1] observe high correlation when using the median protein abundance ('mRNA-free') as a proxy for the expression of a protein across all analysed tissues (figure 1a of ref. 1). Despite this observation, using measured mRNA levels is obviously meaningful because some proteins show vast (absolute) expression differences between tissues and cell lines (highlighted in figure 3a of our original work[2]), a biological fact that would be missed if mRNA levels were not considered.

Fourth, for technical and biological reasons, figure 1b and 1c of Fortelny *et al.*[1] needs careful interpretation. For example, a per-gene correlation of measured versus predicted protein abundance of around zero across different tissues may simply mean that the protein is actually similarly expressed in many tissues and thus the correlation of zero has no particular meaning. Conversely, a correlation of close to 1 for a particular protein may imply a biological function requiring tissue-dependent regulation.

In conclusion, we agree with Fortelny *et al.*[1] that more accurate data are necessary to enable prediction of protein abundance variation within a particular cell type or tissue. We also fully acknowledge that further transcriptional and post-translational factors have to be considered in order to explain protein levels in cells accurately. We therefore welcome the re-analysis of our data as it stimulates discussion on an important scientific topic and further highlights the value of creating and sharing large-scale biological data resources.

Note: The author list of this Reply contains only those individuals most closely involved in the matter discussed in this BCA.

Mathias Wilhelm[1], Hannes Hahne[1]†, Mikhail Savitski[2], Harald Marx[1]†, Simone Lemeer[1]†, Marcus Bantscheff[3] & Bernhard Kuster[1]
[1]Chair of Proteomics and Bioanalytics, Technical University of Munich, Emil-Erlenmeyer Forum 5, 85354 Freising, Germany.
[2]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany.
[3]Cellzome GmbH, Meyerhofstrasse 1, 69117 Heidelberg, Germany.
†Present addresses: OmicScouts GmbH, Lise-Meitner-Str. 30, 85354 Freising, Germany (H.H.); Department of Chemistry, University of Wisconsin, 4426 Biotech, 425 Henry Mall, Madison, Wisconsin 53706, USA (H.M.); Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands (S.L.).

1. Fortelny, N., Overall, C. M., Pavlidis, P. & Cohen Freue, G. V. Can we predict protein from mRNA levels? *Nature* **547,** http://dx/doi.org/10.1038/nature23293 (2017).
2. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–587 (2014).
3. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347,** 1260419 (2015).

doi:10.1038/nature23294