System Wide Analyses have Underestimated Protein Abundances and the Importance of Transcription in Mammals

Large scale surveys in mammalian tissue culture cells suggest that the protein expressed at the median abundance is present at 8,000 - 16,000 molecules per cell and that differences in mRNA expression between genes explain only 10-40% of the differences in protein levels. We find, however, that these surveys have significantly underestimated protein abundances and the relative importance of transcription. Using individual measurements for 61 housekeeping proteins to rescale whole proteome data from Schwanhausser et al., we find that the median protein detected is expressed at 170,000 molecules per cell and that our corrected protein abundance estimates show a higher correlation with mRNA abundances than do the uncorrected protein data. In addition, we estimated the impact of further errors in mRNA and protein abundances using direct experimental measurements of these errors. The resulting analysis suggests that mRNA levels explain at least 56% of the differences in protein abundance for the 4,212 genes detected by Schwanhausser et al., though because one major source of error could not be estimated the true percent contribution could be higher. We also employed a second, independent strategy to determine the contribution of mRNA levels to protein expression. We show that the variance in translation rates directly measured by ribosome profiling is only 12% of that inferred by Schwanhausser et al. and that the measured and inferred translation rates correlate only poorly (R^2=0.13). Based on this, our second strategy suggests that mRNA levels explain ~81% of the variance in protein levels. We also determine the percent contributions of transcription, RNA degradation, translation and protein degradation to the variance in protein abundances using both of our strategies. While the magnitudes of the two estimates vary, they both suggest that transcription plays a more important role than the earlier studies implied and translation a much smaller role. Finally, the above estimates only apply to those genes whose mRNA and protein expression was detected. Based on a detailed analysis by Hebenstreit et al., we

estimate that approximately 40% of genes in a given cell within a population express no mRNA. Since there can be no translation in the absence of mRNA, we argue that differences in translation rates can play no role in determining the expression levels for the ~40% of genes that are non-expressed.

1   Jingyi Jessica Li, Department of Statistics, University of California, Berkeley CA USA[1]

2   Peter J. Bickel, Department of Statistics, University of California, Berkeley CA USA

3   Mark D. Biggin, Genomics Division, Lawrence Berkeley National Laboratory, Berkeley CA USA

4

5   Corresponding author:

6   Mark D. Biggin

7   Genomics Division

8   Lawrence Berkeley National Laboratory

9   1 Cyclotron Road MS 84-171

10 Berkeley CA 94720.

11

12 email mdbiggin@lbl.gov

13 phone 510 486 7606

14

15

16 [1] current address Department of Statistics, University of California, Los Angeles CA USA

17

INTRODUCTION

The protein products of genes are expressed at very different levels from each other in a mammalian cell. Thousands of genes are not detectably expressed. Of those that are, their proteins are present at levels that differ by five orders of magnitude. Cytoplasmic actin, for example, is expressed at 1.5×10^8 molecules per cell (Kislauskis et al. 1997), whereas some transcription factors are expressed at only 4×10^3 molecules per cell (Biggin 2011). There are four major steps that determine differences in protein expression: the rates at which genes are transcribed, mRNAs are degraded, proteins are translated, and proteins are degraded (Fig. 1). The combined effect of transcription and mRNA degradation together determines mRNA abundances (Fig. 1). The joint effect of protein translation and protein degradation controls the relative differences between mRNA and protein concentrations (Fig. 1).

Transcription has long been regarded as a dominant step and is controlled by sequence specific transcription factors that differentially interact with *cis*-regulatory DNA regions. The rates of the other three steps, however, vary significantly between genes as well (Boisvert et al. 2012; Cambridge et al. 2011; Cheadle et al. 2005; de Sousa Abreu et al. 2009; Eden et al. 2011; Guo et al. 2010; Han et al. 2013; Hentze & Kuhn 1996; Hsieh et al. 2012; Ingolia et al. 2011; Kristensen et al. 2013; Loriaux & Hoffmann 2013; Rabani et al. 2011; Schwanhausser et al. 2011; Sharova et al. 2009; Yang et al. 2003). MicroRNAs, for example, differentially interact with mRNAs to alter rates of RNA degradation and protein translation (Ambros 2011; Baek et al. 2008; Elmen et al. 2008; Gennarino et al. 2012; Guo et al. 2010; Hobert 2008; Krutzfeldt et al. 2005; Pillai et al. 2007; Rajewsky 2011; Selbach et al. 2008; Subtelny et al. 2013; Xiao et al. 2007).

To quantify the relative importance of each of the four steps, label free mass spectrometry methods have been developed that measure the absolute number of protein molecules expressed per cell for thousands of genes (Bantscheff et al. 2012; Beck et al. 2011; Maier et al. 2009; Schwanhausser et al. 2011; Vogel et al. 2010; Vogel & Marcotte 2012). By comparing these data to mRNA abundance data, the relative importance of transcription and mRNA degradation versus protein translation and protein degradation can be determined (Maier et al. 2009; Schwanhausser et al. 2011; Vogel & Marcotte 2012) (Fig. 1). By measuring mRNA degradation and protein degradation rates as well, the rates of transcription and translation can be additionally inferred indirectly. Using this approach to study mouse NIH3T3 fibroblasts, Schwanhausser et al. concluded that mRNA levels explain ~40% of the variability in protein levels; that the cellular abundance of proteins is predominantly controlled at the level of

50   translation; that transcription is the second largest determinant; and that the degradation of

51   mRNAs and proteins play a significant but lesser role (Schwanhausser et al. 2011).

52   The above work has provided critically important datasets and an initial framework for analysis.

53   We noticed, however, that Schwanhausser et al.'s protein abundance estimates are mostly

54   lower than established values for individual proteins in the literature and that statistical methods

55   to quantitate the impact of experimental error had not been employed. We therefore set out to

56   explore if we could refine the analysis of these datasets and to compare our results to those of

57   Schwanhausser et al. and other system wide studies.

58

59   **RESULTS AND DISCUSSION**

60   ***A non-linear underestimation of protein abundances***

61   Our starting point was a set of published abundances of 53 mammalian housekeeping proteins,

62   most of which are based on SILAC mass spectrometry or western blot data (Biggin 2011; Brosi

63   et al. 1993; Gregory et al. 2002; Hanamura et al. 1998; Kimura et al. 1999; Kislauskis et al.

64   1997; Princiotta et al. 2003; Wollfe 1998; Wong et al. 2011; Zeiler et al. 2012). On average

65   these established estimates are 16 fold higher than those from Schwanhausser et al.'s original

66   label free mass spectrometry data (Dataset S1). Once we brought this discrepancy to the

67   authors' attention, they upwardly revised their abundance estimates for all 5,028 detected

68   proteins and provided western blot or Selected Reaction Monitoring (SRM) mass spectrometry

69   measurements for eight polypeptides in NIH3T3 cells (see Corrigendum (Schwanhausser et al.

70   2011)). However, Schwanhausser et al.'s second whole proteome abundance estimates are

71   still lower than individual measurements for proteins expressed below 10^6 molecules per cell,

72   with the lowest abundance proteins showing the largest discrepancy (Fig. 2a; Dataset S1).

73   Western blot and SILAC mass spectrometry measurements show the same discrepancy versus

74   the label free whole proteome data (Dataset S1). For example, for proteins expressed below 1

75   million molecules per cell, the 26 SILAC measurements are a median of 2.95 fold higher than

76   Schwanhausser et al.'s second estimates, and the 19 western blot measurements are 3.10 fold

77   higher. This suggests that the discrepancy is not due to error in the individual measurements

78   as a similar bias in two independent methods is unlikely.

79   Of the 61 individual measurements of protein abundance available to us, 15 were made in

80   NIH3T3 cells and 42 were made in HeLa cells. The discrepancy between Schwanhausser et

81   al.'s second whole proteome abundances and these individual measurements is not due to

82   differences in expression levels between HeLa and NIH3T3 cells for the following reasons.

83   One, it is unlikely that such a difference would only occur for lower abundance proteins. Two,

84   five of the individual measurements for lower abundance proteins (Orc2, Orc4, HDAC3, NFkB1,

85   and NFkB2) were made in NIH3T3 cells and are on average 3.7 fold higher than the second

86   whole proteome estimates in this same cell line (Dataset S1). Three, later in the paper we show

87   that collectively the 61 individual proteins measured have on average the same relationship in

88   expression values versus all other cellular proteins in both NIH3T3 and HeLa cells. In addition,

89   as further evidence we note that Schwanhausser et al.'s second estimates for RNA polymerase

90   II and general transcription factors such as TFIIB and TFIIE are only 1.6 fold higher than those

91   in yeast (Borggrefe et al. 2001) and are 7.1 times less than those in HeLa cells (Kimura et al.

92   1999). Yeast cells have 1/40th the volume, 1/200th the amount of DNA and ¼ the number of

93   genes of NIH3T3 and HeLa cells (Milo et al. 2010). Two fold reductions in the concentrations of

94   a single general transcription factor have, in some cases, phenotypic consequence (Aoyagi &

95   Wassarman 2001; Deutschbauer et al. 2005; Eissenberg et al. 2002; Kim et al. 2010). Thus, it

96   is unlikely that a rapidly dividing mammalian cell could function with much larger reductions in

97   the amounts of all of these essential regulators to levels close to those found in yeast.

98   ***Correcting the non-linear bias***

99   Schwanhausser et al. calibrated protein abundances by spiking known amounts of protein

100   standards into a crude protein extract from NIH3T3 cells and then measuring the abundances

101   of several thousand proteins in the mixture by label free mass spectrometry. The 20 "spiked in"

102   protein standards detected in this experiment, however, were present at the equivalent > 8.0 x

103   10^{5} molecules per cell, a level that represents only the most highly expressed 11% of the

104   proteins detected (Fig. 3a) (M. Selbach, personal communication (Schwanhausser et al.

105   2011)). To convert mass spectrometry signals to protein abundances, Schwanhausser et al.

106   assumed that a linear relationship defined using the 20 "spiked in" standards holds true for

107   proteins at all abundances (Fig. 3a). The discrepancy between the resulting estimates and

108   individual protein measurements, however, suggests that this assumption is not valid. We

109   therefore employed the 61 individual protein measurements from the literature as they span a

110   much wider abundance range. In a plot of these data vs Schwanhausser et al.'s second whole

111   proteome estimates, we found that a two-part linear regression gave a statistically better fit

112   over a single regression (Fig. 3b and c) (p-value=0.002, Materials and Methods). We then used

113   this two-part regression to derive new abundance estimates for all 5,028 proteins in

114   Schwanhausser et al.'s dataset (Dataset S1). As Figure 2b shows, the correction removes the

115   non-linear bias.

116 In our rescaled data, the median abundance protein is present at 170,000 molecules per cell

117 (Fig. 2b), considerably higher than Schwanhausser et al.'s original estimate of 16,000

118 molecules per cell and significantly above their second estimate of 50,000 molecules per cell.

119 For low abundance proteins the effect is larger. In our corrected data, the median sequence

120 specific transcription factor is present at 71,000 molecules per cell versus Schwanhausser et

121 al.'s estimates of first 3,500 then 9,300 molecules per cell (Dataset S1). Our correction reduces

122 the range of detected abundances by ~50 fold (unlogged) compared to Schwanhausser et al.'s

123 second estimates (Dataset S1) and the variance in protein levels from 0.97 (log10) to 0.36

124 (log10).

125 ***Corrected protein abundances show an increased correlation with mRNA abundances***

126 As an independent check on the accuracy of our corrected abundances, we compared them to

127 Schwanhausser et al.'s RNA-Seq mRNA expression data. Our corrected protein abundances

128 correlate more highly with mRNA abundances than do Schwanhausser et al.'s second whole

129 proteome estimates (compare Fig. 4a and b). The increase in correlation coefficient is highly

130 significant (p-value<10^{-29}) (Materials and Methods), arguing that our non-linear correction to the

131 whole proteome abundances has increased the accuracy of these estimates. The most

132 dramatic change is that the scatter about the line of best fit is reduced and shows a stronger

133 linear relationship. The 50% prediction band shows that prior to correction the half of proteins

134 whose abundances are best predicted by mRNA levels are expressed over an 11 fold range

135 (unlogged), but after correction they are expressed over a narrower, 4 fold range (Fig. 4a and

136 b). The correction reduces the width of the 95% prediction band even further, by 18 fold.

137 For our corrected data, the median number of proteins translated per mRNA is 9,800 compared

138 to Schwanhausser et al.'s original estimate of 900 and their second estimate of 2,800. In yeast,

139 the ratio of protein molecules translated per mRNA is 4,200 - 5,600 (Ghaemmaghami et al.

140 2003; Lu et al. 2007). Given that mammalian cells have a higher protein copy number than

141 yeast (Milo et al. 2010), it is not unreasonable that the ratio in mammalian cells would be the

142 higher.

143

144 ***Estimating the impact of molecule specific measurement error***

145 In addition to the above general error in scaling protein abundances, there are additional

146 sources of experimental error that uniquely affect data for each protein and mRNA differently.

147 As a result of these molecule specific measurement errors, the coefficient of determination

148 between measured mRNA and measured protein levels—i.e. R^2 shown in Fig. 4b—is lower

149 than the actual value between true protein and true mRNA levels. With an accurate estimate of

150 the errors, it is possible to calculate the increased correlation expected between true protein

151 and true mRNA abundances. Because the variance in the residuals in Fig. 4b (i.e. the

152 displacement along the y axis of data points about the line of best fit) is composed of both

153 experimental error and the genuine differences in the rates of translation and protein

154 degradation between genes, once the experimental error has been estimated, it is also

155 possible to infer the combined true effects of translation and protein degradation.

156 There are two classes of molecule specific experimental error: stochastic and systematic.

157 Stochastic error, or imprecision, is the variation between replica experiments and is estimated

158 from this variation. Systematic error, or inaccuracy, is the reproducible under or over estimation

159 of each data point, and is estimated by comparing the results obtained with the assay being

160 used to those from gold standard measurements obtained with the most accurate method

161 available.

162 Schwanhausser et al. limited their estimation of experimental error to stochastic errors.

163 Because our correction of the whole proteome abundances reduces the total variance in

164 measured protein expression levels, we first reestimated the proportion of the variance in the

165 residuals in Fig. 4b that is due to stochastic measurement error using replica datasets

166 (Materials and Methods). We find that 7% of this variance results from stochastic protein error

167 and 0.8% from stochastic mRNA error.

168 Schwanhausser et al., however, also noted a significant variance between their whole genome

169 RNA-Seq data and NanoString measurements for 79 genes (R^2=0.79 in Fig. S8A in

170 (Schwanhausser et al. 2011)), though they did not take this into account subsequently. RNA-

171 Seq is well known to suffer reproducible several fold biases in the number of DNA sequence

172 reads obtained for different GC content genomic regions (Cheung et al. 2011; Dohm et al.

173 2008). In contrast, NanoString gives an accurate measure of nucleic acid abundance as

174 correlation coefficients of R^2=0.99 are obtained when NanoString data are compared to known

175 concentrations of nucleic acid standards (Geiss et al. 2008). Thus, it is reasonable to consider

176 NanoString as a gold standard that can be used to assess the systematic error in the RNA-seq

177 data by assuming that the variance between the two methods is due mostly to systematic error

178 in RNA-seq. Using Analysis of Variance (ANOVA), the variance in Schwanhausser et al.'s

179 NanoString/RNA-Seq comparison can be shown to be equivalent to 23.3% of the variation in

180 the residuals in Fig. 4b, 29 fold larger than the stochastic component of mRNA error (see

181 Materials and Methods for a discussion of the assumptions used in this analysis).

182   It is also important to assess the systematic error in the whole proteome abundances as label

183   free mass spectrometry includes such biases (Bantscheff et al. 2012; Kuntumalla et al. 2009;

184   Lu et al. 2007; Peng et al. 2012). In principle the "spiked in" protein standards in

185   Schwanhausser et al.'s calibration experiment (i.e. the data in Fig. 3a) should provide gold

186   standard data. In practice, however, the variance in mass spectrometry estimates for protein

187   standards present at supposedly the same amounts is too high (i.e. the scatter along the x axis

188   in Fig. 3a). This variance would contribute 61% to the variance in the residuals in Fig. 4b, yet

189   the variance of the residuals between the corrected whole proteome estimates and the 61

190   individual protein measurements (i.e. the scatter along the x axis about the solid red line in Fig.

191   3b) would contribute only 44%. Since the western blot and SILAC methods used to make the

192   61 individual protein measurements introduce some experimental error, it seems likely that the

193   commercial protein standards used by Schwanhauser at al. were not as accurately prepared at

194   the correct protein concentrations as one would expect. Since no other suitable gold standard

195   is available, we are thus unable to estimate the systematic protein error, though it is likely to be

196   less 44% of variance in the residuals in Fig. 4b.

197   Taking the stochastic protein error as a minimum estimate of protein error and the variance

198   from the NanoString/RNA-Seq comparison as an estimate of all RNA errors, it can be shown

199   that true mRNA levels explain at least 56% of true protein levels, and by extension protein

200   degradation and translation combined explain no more than 44% (see Materials and Methods).

201   ***Estimating the relative importance of transcription, mRNA degradation, translation and***

202   ***protein degradation***

203   In addition to determining protein and mRNA abundances, Schwanhausser et al. also directly

204   measured mRNA and protein degradation rates and calculated the percentage that each

205   contributed to the variance in protein abundances. Using this information, it is possible to

206   determine the relative importance of transcription, RNA degradation, translation and protein

207   degradation for different scenarios (Table 1, see Materials and Methods). For the 4,212 genes

208   whose protein and mRNA expression was detected, our analysis suggests that transcription

209   explains ~38% of the variance in true protein levels, RNA degradation explains ~18%,

210   translation ~30%, and protein degradation ~14% (Table 1). Clearly these estimates are

211   tentative and depend on the particular assumptions we have made. We believe, though, that

212   they will prove more accurate than Schwanhausser et al.'s suggestion that translation is the

213   predominant determinant of protein expression and that mRNA levels explain around 40% of

214   the variability in protein levels (Schwanhausser et al. 2011) (Table 1).

215

Direct measurements of translation rates support our analysis

Direct measurements of system wide translation rates using ribosome profiling (Guo et al.
2010; Ingolia et al. 2011; Subtelny et al. 2013) provide independent evidence that translation
rates vary less than Schwanhausser et al. suggest. The distributions of the rates of translation
rates measured in mouse embryonic stem cells, mouse neutrophils, mouse NIH3T3 cells and
human HeLa cells are all significantly narrower than Schwanhausser et al. inferred for mouse
NIH3T3 cells (Fig. 5a; Table S1). For NIH3T3 cells the translation rates measured by ribosome
profiling for 95% of the genes detected vary only 5.8 fold, but the rates inferred for 95% of
genes by Schwanhausser et al. vary 115 fold (Fig. 5a). Because each of these datasets contain
differing numbers of genes (Table S1), to provide a more direct comparison we took the
intersection of genes detected by Schwanhausser et al. and by ribosome profiling in NIH3T3
cells (Fig. 5b). The variance in measured translation rates for the genes in the intersection is
only 12% of the variance in rates inferred by Schwanhausser et al. for these same genes (Fig.
5b; Table S1).

Having direct measurements of the variance in translation rates opens up a second strategy to
estimate the relative importance of each step in gene expression (Materials and Methods). In
our first strategy—the measured protein error strategy—protein degradation rates and errors in
protein and mRNA abundances were determined from direct experimental data; and the
variance in true protein levels explained by translation was inferred as that part of the variance
in the residuals in Fig. 4b that is not explained by the three experimentally measured terms. In
our second strategy—the measured translation strategy—translation rates, protein degradation
rates and mRNA errors are determined from direct experimental data; and the variance in
measured protein levels explained by protein error is inferred as that part of the variance in the
residuals in Fig. 4a that is not explained by the sum of variances of the three experimentally
measured components (Materials and Methods). This measured translation strategy is thus
independent of our rescaling of Schwanhausser et al.'s second protein abundance estimates
and of our estimate of stochastic protein measurement error.

According to our second strategy, for NIH3T3 cells the variance in true protein levels is 63% of
the variance in Schwanhausser et al.'s measured protein abundances; mRNA levels contribute
81% to the variance in true protein expression; transcription 71%; RNA degradation 10%;
translation 11%; and protein degradation 8% (Table 1). Despite the significant differences in
the underlying data and assumption used, these results agree broadly with those of our first

248   strategy (Table 1). Both strategies suggest that the variance in Schwanhausser et al.'s second

249   protein abundance estimates is too high. Both suggest that translation contributes less to

250   protein levels and that transcription contributes more that Schwanhausser et al. claimed. In

251   effect, the measured rates of translation provide independent support for our rescaling of

252   Schwanhausser et al.'s protein abundances and our estimates of stochastic protein error, and

253   visa versa.

254   Our second strategy, though, does estimate that mRNA levels and transcription explain a

255   higher percent of protein expression than the first (Table 1), but this is not entirely unexpected.

256   In our first strategy, we were not able to take account of systematic, molecule specific errors in

257   protein abundances because appropriate control measurements were not available. Thus, this

258   first strategy could well have underestimated error. In contrast, our second strategy estimates

259   all types of protein abundance errors in a single term and thus has the potential to be the more

260   accurate if the error in the ribosome profiling and protein degradation data is not too large.

261   To further explore the relationship between our two strategies, we compared the correlation

262   between translation rates inferred by Schwanhausser et al. and those measured by ribosome

263   profiling in NIH3T3 cells (Fig. 6). The coefficient of determination is small (R^2= 0.13), indicating

264   that the ribosome profiling data explain only 13% of the variance in Schwanhausser et al.'s

265   inferred rates. Considered in isolation this result does not establish if the poor correlation is due

266   to errors in either or both datasets. However, our measured protein error strategy shows that

267   the variance in true translation rates contributes no more than 19% to the variance in

268   Schwanhausser et al.'s inferred translation rates, with the remaining 81% of the variance being

269   due to experimental error (Table 1; 0.19 = (0.34x0.30)/(0.97x0.55)). The close agreement of

270   this estimate with the actual correlation between measured and inferred translation rates

271   (R^2<=0.19 vs R^2= 0.13) suggests that the poor correlation is almost entirely due to error in

272   Schwanhausser et al.'s inferred rates. In addition, this result provides further evidence that our

273   two strategies broadly agree, with the measured protein error strategy potentially

274   underestimating the degree of error in Schwanhausser et al.'s data.

275   Ribosome profiling has also shown that translation rates change only several fold upon cellular

276   differentiation and, with the exception of the translation machinery, the change affects all

277   expressed genes to a similar degree (Ingolia et al. 2011). Other system wide studies, including

278   a separate analysis by Schwanhausser et al., also suggest that the differential regulation of

279   translation may be limited to modest changes at a subset of genes (Baek et al. 2008; Hsieh et

280   al. 2012; Kristensen et al. 2013; Schwanhausser et al. 2011; Selbach et al. 2008). This work

281 seems consistent with our analysis and suggests that translation may be used chiefly to fine

282 tune protein expression levels.

283 ***Estimating the number of non-transcribed genes***

284 Both Schwanhausser et al.'s and all of our analyses presented above consider only those

285 genes whose protein and mRNA expression was detected. There are many thousands of other

286 genes, however, which express no mRNA and as a result cannot be translated. To estimate the

287 proportion of such genes in a typical cell, we made use of a detailed analysis by Hebenstreit et

288 al, who showed that there is a trimodal distribution of mRNA expression when the data is

289 derived as an average for a population of cells of a single cell type (Hebenstreit et al. 2012;

290 Hebenstreit et al. 2011) (Fig. S1). The first mode contains Highly Expressed (HE) genes,

291 present at one or more molecules per cell; the second mode is comprised of Low Expressed

292 (LE) genes, which are not expressed in most cells but—as shown by single molecule

293 fluorescent *in situ* hybridization—are present at one to several molecules per cell in a small

294 percent of cells; and the third mode contains genes that are not detectably expressed (NE

295 genes) and thus, given the assays sensitivity, are present at less than one mRNA molecule per

296 100 cells. LE genes tend to be closer to HE genes on the chromosome than are NE genes, and

297 it has been suggested that this proximity may allow escape from repressive chromatin

298 structures in a few cells, explaining the stochastic bursts of rare transcription observed

299 (Hebenstreit et al. 2012; Hebenstreit et al. 2011).

300 To account for variation in the expression of individual genes between cells, which all LE genes

301 at a minimum must suffer, we assume that the general distribution of mRNA expression levels

302 does not vary from cell to cell even when the expression of individual genes does. The mRNA

303 expression of each LE gene was divided into a component representing expression of one

304 mRNA molecule in some cells and a second component representing the remaining cells that

305 express no mRNA (Materials and Methods). This yields 8,763 NE and LE gene equivalents that

306 are not expressed and 12,546 LE and HE gene equivalents that are expressed. For the 8,763

307 non-expressed gene equivalents, the complete absence of their mRNAs from the cell means

308 that they are not being translated in these cells. Therefore, there can be no variation in the

309 rates at which they are translated. Instead, we assume that the absence of transcription is

310 overwhelmingly the reason why these genes express no protein.

311

312 ***Implication for other system wide studies***

313 Two other system wide estimates of protein abundance in mammalian cells are, like

314 Schwanhausser et al.'s, lower than ours. These two reports suggest that the median

315 abundance protein detected is present at 8,000 (Vogel et al. 2010) or 9,700 (Beck et al. 2011)

316 molecules per cell vs our estimate of 170,000 molecules per cell. Since these lower estimates

317 provide less than 1/10th of the number of histones needed to cover the diploid genome with

318 nucleosomes and are lower than published estimates for a wide array of other housekeeping

319 proteins, it is unlikely that they are accurate.

320 Another study by Wisniewski et al. provided protein abundance estimates for HeLa cells that

321 are generally higher than ours and spread over a broader range (Wisniewski et al. 2012) (Fig.

322 7a). These estimates are 240% higher on average than the set of individual protein

323 measurements from the literature (Dataset S3, Fig. 7b). Since over 80% of these individual

324 measurements were made for proteins in HeLa cells, Wisniewski et al.'s estimates must be

325 incorrectly scaled. Using our two part linear regression strategy, we therefore corrected

326 Wisniewski et al.'s whole proteome data (Materials and Methods, Fig. S2; Dataset S3), bringing

327 the average variation between the whole proteome estimates and individual protein

328 measurements to within 6% of each other (Fig. 7b; Dataset S3). Interestingly, the correction

329 dramatically increases the similarity between the distributions of protein abundances in HeLa

330 and NIH3T3 cells for all orthologous proteins (Fig. 7a). This establishes the important point,

331 mentioned at the beginning of the Results, that in aggregate the 60+ housekeeping proteins

332 show a similar relationship to the expression values of all other cellular proteins in both cell

333 lines, and thus the discrepancies with the uncorrected whole proteome data are not due to

334 differences in expression levels in HeLa versus NIH3T3 cells. The correction also increases the

335 correlation between HeLa cell protein and HeLa mRNA abundances to a statistically significant

336 extent (p-value, 6×10^{-20}) and reduces the 50% and 95% confidence bounds for this

337 relationship by 1.7 fold and 4.6 fold respectively. Wisniewski et al. scaled their protein

338 abundances using the total cellular protein content and the sum of the mass spectrometry

339 signals for all detected polypeptides. They assumed that mass spectrometry signals are

340 proportional to protein abundance. In contrast, our scaling strategy makes no such assumption

341 and instead uses many individual measurements of housekeeping proteins to estimate a

342 multipart (spline) function. The increased correlations obtained with individual protein

343 measurements and with mRNA abundances for two cell lines suggests that our scalings are the

344 more accurate.

345   Other estimates for the contribution of mRNA levels in determining protein expression in

346   mammals are lower than ours, suggesting that mRNA levels contribute 10%-40% (Maier et al.

347   2009; Vogel & Marcotte 2012). In comparison, we estimate that mRNA abundance explains

348   56% - 81% for a set of 4,212 detected proteins. We also have suggested that for the 40% of

349   genes in a given cell that express no mRNA, translation rates likely play no role in determining

350   protein expression levels. The other groups' neither took systematic experimental errors into

351   account or made use of direct measures of translation rates and generally do not discuss non-

352   transcribed genes. For this reason, we suspect their analyses underestimate the contribution of

353   transcription.

354   **CONCLUSIONS**

355   Quantitative whole proteome analyses can offer profound insights into the control of gene

356   expression and provide baseline parameters for much of systems biology. As these important

357   new technologies continue to be refined, it is critical that the data be correctly scaled, that

358   experimental errors be measured and accounted for as much as possible, that all genes be

359   considered, and that direct measurements of each step in gene expression be made. Additional

360   measurements and controls will be needed to derive a more assured system wide

361   understanding of protein and mRNA abundances and the relative importance of each of the

362   four steps in gene expression.

363

364

365   **MATERIALS AND METHODS**

366   ***Correcting protein abundance***

367   For NIH3T3 cells, all credible individual protein abundance measurements available to us for

368   housekeeping proteins (a total of 61 proteins, Dataset S1) were $\log_{10}$ transformed along with

369   the corresponding estimates from Schwanhausser et al.'s second whole proteome dataset.

370   Model selection of different regressive models by leave-one-out cross-validation was used to fit

371   the training data (Bickel & Doksum 2001). This showed that a plausible two-part linear

372   regression with a change point at 10^6 molecules per cell (line<1×10^6...slope=0.56,

373   intercept=2.64; line>1×10^6...slope=1.06, intercept=-0.41) fit the data far better than by chance

374   (likelihood ratio test bootstrap p-value=0.002 (Bickel & Doksum 2001); Fig. 3b and c). The

375   resulting two-part linear model was used to correct all 5,028 protein abundance estimates (Fig.

376   2b, Dataset S1).

377   The null hypothesis that the correlation coefficient of the uncorrected Schwanhausser et al.

378   protein abundance estimates vs mRNA estimates ($R_1=0.626$) is equal to that of our corrected

379   protein estimates vs mRNA estimates ($R_2=0.642$) was tested. The method for comparing

380   dependent correlation coefficients (Olkin & Finn 1990) was employed because both

381   correlations involve the same mRNA-seq data and it is reasonable to assume that the

382   uncorrected and corrected protein abundance estimates and the mRNA estimates have a

383   multivariate Gaussian distribution. The resulting two-sided p-value $< 10^{-29}$ shows that R_2 is

384   significantly larger than R_1.

385   To correct protein abundance estimates for HeLa cells (Wisniewski et al. 2012), the same

386   strategy used for NIH3T3 cells was used. A two-part linear regression with a change point at

387   $10^{6.8}$ molecules per cell fit the data far better than by accident (likelihood ratio test bootstrap p-

388   value=0.001) (Fig. S2). The resulting two-part linear model was used to correct all HeLa cell

389   protein abundance estimates (Fig. 7; Dataset S3). The correlation of HeLa cell protein

390   abundance estimates with mRNA abundances was determined using the mean values of

391   replica HeLa cell RNA-Seq datasets from the ENCODE consortium (Consortium 2011) (GEO

392   Accession ID "GSM765402"). The hypothesis that our corrected protein abundances correlate

393   more highly with these HeLa mRNA abundances than the uncorrected estimates was tested as

394   above, resulting in a two sided p-value of 6×10^{-20}.

395   ***The contribution of mRNA to protein levels: measured protein error strategy***

396   The variance term in a linear model between measured protein abundance (MP) (response)

397   and measured mRNA levels (MR) (predictor) is decomposed in a standard way (ANOVA

398   (Bickel & Doksum 2001)) into three components (Fig. 8). These components of the variance in

399   the residuals represent mRNA measurement error (e_R), protein measurement error (e_P), and

400   the variance in a linear model between true protein abundance (TP) and true mRNA levels (TR)

401   that results from the centered genuine differences in the rates of protein degradation and

402   translation (PDT). The measured protein abundances considered in this case are our rescaled

403   estimates.

404   Statistically, we can write three linear models from Figure 8.

405   $$TR = b_R MR + c_R + e_R \qquad (1)$$

406   $$TP = bTR + c + PDT \qquad (2)$$

407   $$MP = TP + c_P + e_P \qquad (3)$$

408   where TR, MR, TP, MP are abundance values on a $\log_{10}$ scale; the three sources of variation (

409   e_R, e_P and PDT) are assumed to be independent random variables with mean 0; the amount of

410   protein degradation and translation (PDT) is taken to be independent of true mRNA levels (TR)

411   on the basis of partial evidence: the variance in the residuals in Figure 4b is similar for different

412   mRNA abundances; the reversal of the causal relationship between TR and MR in model (1)

413   assumes that TR and MR have an approximately joint Gaussian distribution; the slope of TP in

414   model (3) is assumed to be 1 because the ratios between the 61 protein published abundance

415   measurements and our corrected estimates are close to 1 (Fig. 2b); and finally we note that

416   implicit in the analysis of variance is the assumption that the various datasets employed can

417   thought of as originating from a relatively homogeneous superpopulation. Combining (1)-(3),

418   we write the linear model between measured protein abundance and measured mRNA levels

419   as

$$MP = bb_R MR + bc_R + c + c_P + be_R + PDT + e_P \quad (4)$$

420   Based on model (4)

421   i.   We first estimated $\mathrm{var}(be_R + PDT + e_P)$ as σ_{all}^2 and bb_R as $\hat{b}_{all}$ from fitting the above

422   model with the 8,424 corrected mass spec and RNA-Seq data points pooled from the two

423   replicates (Dataset S1). By independence, we have

$$\mathrm{var}(be_R + PDT + e_P) = b^2 \,\mathrm{var}(e_R) + \mathrm{var}(PDT) + \mathrm{var}(e_P)$$

425   ii.   We next estimated $\mathrm{var}(e_R)$ as $\hat{\sigma}_R^2$ and b_R as $\hat{b}_R$ from fitting model (1) with the 77

426   NanoString ("TR") vs RNA-Seq ("MR") data points, after removing two outliers (Dataset S2).

427   iii. We could not estimate $\mathrm{var}(e_P)$ from directly fitting model (3), as TP data is not available. As

428   a surrogate, we estimated $\mathrm{var}(e_P)$ as $\hat{\sigma}_P^2$ from the following linear model that quantifies the

429   stochastic error in mass spec replicate data:

$$MP_{ij} = avgMP_i + (e_P)_{ij} \quad , \mathrm{j=1,2} \quad (5),$$

431   where MP_{ij} is the corrected mass spec data for the ith protein in the jth replicate in

432   Schwanhausser et al., and $avgMP_i$ is the average of our corrected protein data for the ith

433   protein, i = 1, …, 4,212 (Dataset S1). Please note that $\hat{\sigma}_P^2$ is potentially an underestimate of

434   the protein error as we only consider the stochastic error, not the systematic error.

435   iv. From the estimates σ_{all}^2, $\hat{b}_{all}$, $\hat{\sigma}_R^2$, $\hat{b}_R$ and $\hat{\sigma}_P^2$ above, we estimate var(PDT) as

436
$$\hat{\sigma}^2_{PDT} = \hat{\sigma}^2_{all} - \left(\frac{\hat{b}_{all}}{\hat{b}_R}\right)^2 \hat{\sigma}^2_R - \hat{\sigma}^2_P$$

437   Hence, we have successfully decomposed the variance estimate $\hat{\sigma}^2_{all}$, i.e. the estimated

438   variance of residuals between measured protein levels and measured mRNA levels, into 3

439   components:

440   • $\hat{\sigma}^2_R$—RNA error (23.3% of σ^2_{all})

441   • $\hat{\sigma}^2_P$—protein error (7% of σ^2_{all})

442   • $\hat{\sigma}^2_{PDT}$—protein degradation & translation (69.6% of σ^2_{all})

443   From the diagram and the above calculation, we also derived the percentage of variability in

444   the unobserved true protein levels explained by the unobserved true mRNA levels.

445
$$\frac{\hat{\sigma}^2_{MP} - \hat{\sigma}^2_P - \hat{\sigma}^2_{PDT}}{\hat{\sigma}^2_{MP} - \hat{\sigma}^2_P} = 55.9\%$$

446   where $\hat{\sigma}^2_{MP}$ is the variance of the corrected measured protein levels.

447   We separately estimated the stochastic mRNA error from the replicate RNA-Seq

448   measurements of the 4,212 genes (Dataset S1). The stochastic mRNA error contributes 0.8%

449   of σ^2_{all}.

450

451   ***The contributions of transcription, translation and protein and mRNA degradation:***

452   ***measured error strategy***

453   To determine the relative contributions of measured RNA degradation (RD) and measured

454   protein degradation (PD) to the variance in true protein expression (TP), we estimated their

455   variances, var(RD) and var(PD). We took Schwanhausser et al.'s calculated percentages for

456   the contribution of RD and PD to explain the variance of their uncorrected mass whole

457   proteome abundances (Schwanhausser et al. 2011) (6.4% for RD and 4.9% PD, Matthias

458   Selbach personal communication). Since the variance of the 8,424 uncorrected mass spec

459   data points from the two replicates is 0.97, we thus calculated var(RD) and var(PD) as 0.062

460   and 0.048 respectively. The relative contributions of var(RD) and var(PD) to var(TP) (estimated

461   as $\hat{\sigma}^2_{MP} - \hat{\sigma}^2_P$) was calculated (Table 1). We also determined the contribution of transcription

462   (var(TXN)) to var(TP) as (var(TR)-var(true RD))/var(TP), where var(TR) was estimated as

463    $\hat{\sigma}^2_{MP} - \hat{\sigma}^2_P - \hat{\sigma}^2_{PDT}$, and the contribution of translation as (var(TP)-var(TR)-var(true PD))/var(TP)

464   (Table 1).

The contributions of each step of gene expression to protein levels: measured
translation strategy

We calculated the relative contributions of each of the four steps in gene expression by an independent, second approach that does not rely either on our rescaling of Schwanhausser et al.'s protein abundance estimates or on our estimate of stochastic protein errors. Instead, our second approach infers true protein abundance based on Subtelny et al.'s direct measurements of translation rates in NIH3T3 cells by ribosome profiling (Subtelny et al. 2013) and on our estimate of RNA measurement error. The measured protein abundances considered are thus Schwanhausser et al.'s second estimates, not our rescaling of these estimates. A central assumption is that since the variance in Subtelny et al.'s measured translation rates is 12% of the variance in the rates of translation inferred by Schwanhausser et al., then the contribution of translation to the variance in true protein levels is 12% of the value provided by Schwanhausser et al.

The variance term in a linear model between measured protein abundance (MP) and measured mRNA levels (MR) was decomposed as before (Fig. 8) except that the variance in the linear model between true protein abundance (TP) and true mRNA levels (TR) that results from the variance in the rates of protein degradation (PD) and protein translation (PT) were considered separately as cPD and dPT respectively. Similar to our measured error strategy, we can write three linear models using the same assumptions.

484         $TR = b_R MR + c_R + e_R$ (1)

485         $TP = bTR + cPD + dPT + f$ (2)

486         $MP = TP + c_P + e_P$ (3)

Thus, we can write the linear model between measured protein abundance (MP) and measured mRNA levels (MR) for the measured translation strategy as

489         $MP = bb_R MR + bc_R + f + c_P + be_R + cPD + dPT + e_p$ (4)

490   Based on this revised model (4)

491   i.   We first estimated $\mathrm{var}(be_R + cPD + cPT + e_P)$ as $\hat{\sigma}^2_{all}$ and bb_R as $\hat{b}_{all}$ from fitting the above

492   model with the 8,424 mass spec and RNA-Seq data points pooled from the two replicates

493   using Schwanhausser's second estimates (Dataset S1). By independence, we thus have

494    $$\mathrm{var}(be_R + cPD + cPT + e_P) = b^2\,\mathrm{var}(e_R) + \mathrm{var}(cPD) + \mathrm{var}(dPT) + \mathrm{var}(e_P)$$

495   ii. The values of $\mathrm{var}(e_R)$ and b_R are the same as those derived previously by our measured

496   error strategy. Thus, we can estimate $\hat{b} = \hat{b}_{all}/\hat{b}_R$

497   iii. We used the estimate of $\mathrm{var}(cPD)$ from Schwanhausser et al., i.e., 0.97 x 5% = 0.0475.

498   iv. From Schwanhausser et al.'s results, we have $\mathrm{var}(dPT) = d^2\,\mathrm{var}(PT)$ estimated as 0.97 x

499   55% = 0.54. From Schwanhausser et al.'s estimates for each of 3,633 genes (Dataset S1,

500   second tab, column AG) $\mathrm{var}(PT)$ has an estimate of 0.29. Hence, the estimate of d^2 is

501   1.86. From Subtelny et al., we have a separate, directly measured estimate of $\mathrm{var}(PT)$ as

502   0.03533, which we obtained by slightly increasing the variance of their data for the 3,126

503   genes in the intersected dataset (Fig. 5B; Table S1) by the ratio of the variances for

504   Schwanhusser et al.'s inferred rates for the 3,633 genes and the 3,126 genes (Table S1).

505   Using this value to replace that of Schwanhausser et al., we obtained a new estimate of

506         $\mathrm{var}(dPT) = d^2\,\mathrm{var}(PT)$ as 1.86 x 0.03533 = 0.06593132.

507   v.   Now we can estimate $\mathrm{var}(e_P)$ as $\hat{\sigma}^2_P = \hat{\sigma}^2_{all} - \hat{b}\hat{\sigma}^2_R - \hat{\sigma}^2_{cPD} - \hat{\sigma}^2_{dPT}$ were $\hat{\sigma}^2_{cPD}$ is an estimate of

508   var(cPD) and $\hat{\sigma}^2_{dPT}$ an estimate of var(dPT).

509   vi. Given Schwanhausser et al.'s second 8,424 uncorrected mass spec data, we can also

510   estimate var(TP) as $\hat{\sigma}^2_{TP} = \hat{\sigma}^2_{MP} - \hat{\sigma}^2_P$, where $\hat{\sigma}^2_{MP}$ is an estimate of var(MP).

511   Given the estimates $\hat{\sigma}^2_{cPD}$ and $\hat{\sigma}^2_{dPT}$ and Schwanhausser et al.'s estimate of the contribution of

512   the variance in RNA degradation (defined as $\hat{\sigma}^2_{gRD}$), we can decompose $\hat{\sigma}^2_{TP}$ as:

513   • variance explained by PD: $\hat{\sigma}^2_{cPD}/\hat{\sigma}^2_{TP}$

514   • variance explained by PT: $\hat{\sigma}^2_{dPT}/\hat{\sigma}^2_{TP}$

515   • variance explained by TR: $1 - \dfrac{\hat{\sigma}^2_{cPD}}{\hat{\sigma}^2_{TP}} - \dfrac{\hat{\sigma}^2_{dPT}}{\hat{\sigma}^2_{TP}}$

516 • variance explained by RD: $\hat{\sigma}^2_{gRD} / \hat{\sigma}^2_{TP}$

517 • variance explained by TXN: $1 - \frac{\hat{\sigma}^2_{cPD}}{\hat{\sigma}^2_{TP}} - \frac{\hat{\sigma}^2_{dPT}}{\hat{\sigma}^2_{TP}} - \frac{\hat{\sigma}^2_{gRD}}{\hat{\sigma}^2_{TP}}$

518 ***The number of genes not transcribed in a typical cell within a population.***

519 To estimate the number of genes not transcribed in a typical cell within a population, we

520 employed a deep RNA-Seq dataset that detected polyA+ mRNA for 15,325 protein coding

521 genes in mouse Th2 cells (Hebenstreit et al. 2011). To place these abundance estimates on

522 the same scale as those of Schwanhausser et al.'s data the 3,841 mRNAs expressed above 1

523 RPKM (reads per kilobase of exon per million mapped reads) in common between the two

524 datasets were identified. The Th2 cell data were then scaled to have the same median and

525 variance for these common genes in numbers of mRNA molecules per cell (Fig. S3). Following

526 Hebenstreit et al., we divided the expressed genes into 11,301 Highly Expressed (HE) genes,

527 present at one or more mRNA molecule per cell, and 4,024 Low Expressed (LE) genes,

528 expressed below one molecule per cell. The remaining 5,984 genes whose expression was not

529 detected were designated Not Expressed (NE) genes. We then divided each LE gene into two:

530 a fraction of a gene expressed at 1 molecule per cell with a weight *w* and a fraction of a gene

531 that is not expressed in any cells with a weight 1-*w*. The 4,024 LE genes were thus

532 decomposed into 1,245 gene equivalents expressed at 1 molecules per cell and 2,779 gene

533 equivalents that are not expressed. Combining these with the 11,301 HE genes and 5,984 NE

534 genes, we obtained 12,546 HE and LE expressed gene equivalents and 8,763 NE and LE non-

535 expressed gene equivalents.

536

545

545   **REFERENCES**

546   Ambros V. 2011. MicroRNAs and developmental timing. *Curr Opin Genet Dev* 21:511-517.

547   Aoyagi N, and Wassarman DA. 2001. Developmental and transcriptional consequences of

548   mutations in Drosophila TAF(II)60. *Mol Cell Biol* 21:6808-6819.

549   Baek D, Villen J, Shin C, Camargo FD, Gygi SP, and Bartel DP. 2008. The impact of

550   microRNAs on protein output. *Nature* 455:64-71.

551   Bantscheff M, Lemeer S, Savitski MM, and Kuster B. 2012. Quantitative mass spectrometry in

552   proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem*

553   404:939-965.

554   Beck M, Schmidt A, Malmstroem J, Claassen M, Ori A, Szymborska A, Herzog F, Rinner O,

555   Ellenberg J, and Aebersold R. 2011. The quantitative proteome of a human cell line.

556   *Mol Syst Biol* 7:549.

557   Bickel PJ, and Doksum KA. 2001. *Mathematical Statistics: Basic Ideas and Selected Topics*.

558   Upper Saddle River: Prentice Hall.

559   Biggin MD. 2011. Animal transcription networks as highly connected, quantitative continua. *Dev*

560   *Cell* 21:611-626.

561   Boisvert FM, Ahmad Y, Gierlinski M, Charriere F, Lamont D, Scott M, Barton G, and Lamond

562   AI. 2012. A quantitative spatial proteomics analysis of proteome turnover in human

563   cells. *Mol Cell Proteomics* 11:M111 011429.

564   Borggrefe T, Davis R, Bareket-Samish A, and Kornberg RD. 2001. Quantitation of the RNA

565   polymerase II transcription machinery in yeast. *J Biol Chem* 276:47150-47153.

566   Brosi R, Hauri HP, and Kramer A. 1993. Separation of splicing factor SF3 into two components

567   and purification of SF3a activity. *J Biol Chem* 268:17640-17646.

568   Cambridge SB, Gnad F, Nguyen C, Bermejo JL, Kruger M, and Mann M. 2011. Systems-wide

569   proteomic analysis in mammalian cells reveals conserved, functional protein turnover. *J*

570   *Proteome Res* 10:5275-5284.

571   Cheadle C, Fan J, Cho-Chung YS, Werner T, Ray J, Do L, Gorospe M, and Becker KG. 2005.

572   Control of gene expression during T cell activation: alternate regulation of mRNA

573   transcription and mRNA stability. *BMC Genomics* 6:75.

574   Cheung MS, Down TA, Latorre I, and Ahringer J. 2011. Systematic bias in high-throughput

575   sequencing data and its correction by BEADS. *Nucleic Acids Res* 39:e103.

576   Consortium TEP. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS*

577   *Biology* 9:e1001046.

578   de Sousa Abreu R, Penalva LO, Marcotte EM, and Vogel C. 2009. Global signatures of protein
579   and mRNA expression levels. *Mol Biosyst* 5:1512-1526.
580   Deutschbauer AM, Jaramillo DF, Proctor M, Kumm J, Hillenmeyer ME, Davis RW, Nislow C,
581   and Giaever G. 2005. Mechanisms of haploinsufficiency revealed by genome-wide
582   profiling in yeast. *Genetics* 169:1915-1925.
583   Dohm JC, Lottaz C, Borodina T, and Himmelbauer H. 2008. Substantial biases in ultra-short
584   read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 36:e105.
585   Eden E, Geva-Zatorsky N, Issaeva I, Cohen A, Dekel E, Danon T, Cohen L, Mayo A, and Alon
586   U. 2011. Proteome half-life dynamics in living human cells. *Science* 331:764-768.
587   Eissenberg JC, Ma J, Gerber MA, Christensen A, Kennison JA, and Shilatifard A. 2002. dELL
588   is an essential RNA polymerase II elongation factor with a general role in development.
589   *Proc Natl Acad Sci U S A* 99:9894-9899.
590   Elmen J, Lindow M, Schutz S, Lawrence M, Petri A, Obad S, Lindholm M, Hedtjarn M, Hansen
591   HF, Berger U, Gullans S, Kearney P, Sarnow P, Straarup EM, and Kauppinen S. 2008.
592   LNA-mediated microRNA silencing in non-human primates. *Nature* 452:896-899.
593   Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, Fell HP, Ferree S,
594   George RD, Grogan T, James JJ, Maysuria M, Mitton JD, Oliveri P, Osborn JL, Peng T,
595   Ratcliffe AL, Webster PJ, Davidson EH, Hood L, and Dimitrov K. 2008. Direct
596   multiplexed measurement of gene expression with color-coded probe pairs. *Nat
597   Biotechnol* 26:317-325.
598   Gennarino VA, D'Angelo G, Dharmalingam G, Fernandez S, Russolillo G, Sanges R, Mutarelli
599   M, Belcastro V, Ballabio A, Verde P, Sardiello M, and Banfi S. 2012. Identification of
600   microRNA-regulated gene networks by expression analysis of target genes. *Genome
601   Res* 22:1163-1172.
602   Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, and
603   Weissman JS. 2003. Global analysis of protein expression in yeast. *Nature* 425:737-
604   741.
605   Gregory SG, Sekhon M, Schein J, Zhao S, Osoegawa K, Scott CE, Evans RS, Burridge PW,
606   Cox TV, Fox CA, Hutton RD, Mullenger IR, Phillips KJ, Smith J, Stalker J, Threadgold
607   GJ, Birney E, Wylie K, Chinwalla A, Wallis J, Hillier L, Carter J, Gaige T, Jaeger S,
608   Kremitzki C, Layman D, Maas J, McGrane R, Mead K, Walker R, Jones S, Smith M,
609   Asano J, Bosdet I, Chan S, Chittaranjan S, Chiu R, Fjell C, Fuhrmann D, Girn N, Gray
610   C, Guin R, Hsiao L, Krzywinski M, Kutsche R, Lee SS, Mathewson C, McLeavy C,
611   Messervier S, Ness S, Pandoh P, Prabhu AL, Saeedi P, Smailus D, Spence L, Stott J,
612   Taylor S, Terpstra W, Tsai M, Vardy J, Wye N, Yang G, Shatsman S, Ayodeji B, Geer

613   K, Tsegaye G, Shvartsbeyn A, Gebregeorgis E, Krol M, Russell D, Overton L, Malek JA,

614   Holmes M, Heaney M, Shetty J, Feldblyum T, Nierman WC, Catanese JJ, Hubbard T,

615   Waterston RH, Rogers J, de Jong PJ, Fraser CM, Marra M, McPherson JD, and Bentley

616   DR. 2002. A physical map of the mouse genome. *Nature* 418:743-750.

617 Guo H, Ingolia NT, Weissman JS, and Bartel DP. 2010. Mammalian microRNAs predominantly

618   act to decrease target mRNA levels. *Nature* 466:835-840.

619 Han K, Jaimovich A, Dey G, Ruggero D, Meyuhas O, Sonenberg N, and Meyer T. 2013.

620   Parallel measurement of dynamic changes in translation rates in single cells. *Nat*

621   *Methods*.

622 Hanamura A, Caceres JF, Mayeda A, Franza BR, Jr., and Krainer AR. 1998. Regulated tissue-

623   specific expression of antagonistic pre-mRNA splicing factors. *RNA* 4:430-444.

624 Hebenstreit D, Deonarine A, Babu MM, and Teichmann SA. 2012. Duel of the fates: the role of

625   transcriptional circuits and noise in CD4+ cells. *Curr Opin Cell Biol* 24:350-358.

626 Hebenstreit D, Fang M, Gu M, Charoensawan V, van Oudenaarden A, and Teichmann SA.

627   2011. RNA sequencing reveals two major classes of gene expression levels in

628   metazoan cells. *Mol Syst Biol* 7:497.

629 Hentze MW, and Kuhn LC. 1996. Molecular control of vertebrate iron metabolism: mRNA-

630   based regulatory circuits operated by iron, nitric oxide, and oxidative stress. *Proc Natl*

631   *Acad Sci U S A* 93:8175-8182.

632 Hobert O. 2008. Gene regulation by transcription factors and microRNAs. *Science* 319:1785-

633   1786.

634 Hsieh AC, Liu Y, Edlind MP, Ingolia NT, Janes MR, Sher A, Shi EY, Stumpf CR, Christensen C,

635   Bonham MJ, Wang S, Ren P, Martin M, Jessen K, Feldman ME, Weissman JS, Shokat

636   KM, Rommel C, and Ruggero D. 2012. The translational landscape of mTOR signalling

637   steers cancer initiation and metastasis. *Nature* 485:55-61.

638 Ingolia NT, Lareau LF, and Weissman JS. 2011. Ribosome profiling of mouse embryonic stem

639   cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147:789-802.

640 Kim DU, Hayles J, Kim D, Wood V, Park HO, Won M, Yoo HS, Duhig T, Nam M, Palmer G,

641   Han S, Jeffery L, Baek ST, Lee H, Shim YS, Lee M, Kim L, Heo KS, Noh EJ, Lee AR,

642   Jang YJ, Chung KS, Choi SJ, Park JY, Park Y, Kim HM, Park SK, Park HJ, Kang EJ,

643   Kim HB, Kang HS, Park HM, Kim K, Song K, Song KB, Nurse P, and Hoe KL. 2010.

644   Analysis of a genome-wide set of gene deletions in the fission yeast

645   Schizosaccharomyces pombe. *Nat Biotechnol* 28:617-623.

646   Kimura H, Tao Y, Roeder RG, and Cook PR. 1999. Quantitation of RNA polymerase II and its
647   transcription factors in an HeLa cell: little soluble holoenzyme but significant amounts of
648   polymerases attached to the nuclear substructure. *Mol Cell Biol* 19:5383-5392.
649   Kislauskis EH, Zhu X, and Singer RH. 1997. beta-Actin messenger RNA localization and
650   protein synthesis augment cell motility. *J Cell Biol* 136:1263-1270.
651   Kristensen AR, Gsponer J, and Foster LJ. 2013. Protein synthesis rate is the predominant
652   regulator of protein expression during differentiation. *Mol Syst Biol* 9:689.
653   Krutzfeldt J, Rajewsky N, Braich R, Rajeev KG, Tuschl T, Manoharan M, and Stoffel M. 2005.
654   Silencing of microRNAs in vivo with 'antagomirs'. *Nature* 438:685-689.
655   Kuntumalla S, Braisted JC, Huang ST, Parmar PP, Clark DJ, Alami H, Zhang Q, Donohue-
656   Rolfe A, Tzipori S, Fleischmann RD, Peterson SN, and Pieper R. 2009. Comparison of
657   two label-free global quantitation methods, APEX and 2D gel electrophoresis, applied to
658   the Shigella dysenteriae proteome. *Proteome Sci* 7:22.
659   Loriaux PM, and Hoffmann A. 2013. A protein turnover signaling motif controls the stimulus-
660   sensitivity of stress response pathways. *PLoS Comput Biol* 9:e1002932.
661   Lu P, Vogel C, Wang R, Yao X, and Marcotte EM. 2007. Absolute protein expression profiling
662   estimates the relative contributions of transcriptional and translational regulation. *Nat
663   Biotechnol* 25:117-124.
664   Maier T, Guell M, and Serrano L. 2009. Correlation of mRNA and protein in complex biological
665   samples. *FEBS Lett* 583:3966-3973.
666   Milo R, Jorgensen P, Moran U, Weber G, and Springer M. 2010. BioNumbers--the database of
667   key numbers in molecular and cell biology. *Nucleic Acids Res* 38:D750-753.
668   Olkin I, and Finn JD. 1990. Testing correlated correlations. *Psychological Bulletin* 108:330-333.
669   Peng M, Taouatas N, Cappadona S, van Breukelen B, Mohammed S, Scholten A, and Heck
670   AJ. 2012. Protease bias in absolute protein quantitation. *Nat Methods* 9:524-525.
671   Pillai RS, Bhattacharyya SN, and Filipowicz W. 2007. Repression of protein synthesis by
672   miRNAs: how many mechanisms? *Trends Cell Biol* 17:118-126.
673   Princiotta MF, Finzi D, Qian SB, Gibbs J, Schuchmann S, Buttgereit F, Bennink JR, and
674   Yewdell JW. 2003. Quantitating protein synthesis, degradation, and endogenous
675   antigen processing. *Immunity* 18:343-354.
676   Rabani M, Levin JZ, Fan L, Adiconis X, Raychowdhury R, Garber M, Gnirke A, Nusbaum C,
677   Hacohen N, Friedman N, Amit I, and Regev A. 2011. Metabolic labeling of RNA
678   uncovers principles of RNA production and degradation dynamics in mammalian cells.
679   *Nat Biotechnol* 29:436-442.
680   Rajewsky N. 2011. MicroRNAs and the Operon paper. *J Mol Biol* 409:70-75.

681   Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, and Selbach M.
682   2011. Global quantification of mammalian gene expression control. *Nature* 473:337-
683   342.

684   Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, and Rajewsky N. 2008.
685   Widespread changes in protein synthesis induced by microRNAs. *Nature* 455:58-63.

686   Sharova LV, Sharov AA, Nedorezov T, Piao Y, Shaik N, and Ko MS. 2009. Database for mRNA
687   half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and
688   differentiating mouse embryonic stem cells. *DNA Res* 16:45-58.

689   Subtelny AO, Eichhorn SW, Chen GR, Sive H, and Bartel DP. 2013. Poly(A)-tail lengths and a
690   developmental switch in translational control. *Nature* in press.

691   Vogel C, Abreu Rde S, Ko D, Le SY, Shapiro BA, Burns SC, Sandhu D, Boutz DR, Marcotte
692   EM, and Penalva LO. 2010. Sequence signatures and mRNA concentration can explain
693   two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol* 6:400.

694   Vogel C, and Marcotte EM. 2012. Insights into the regulation of protein abundance from
695   proteomic and transcriptomic analyses. *Nat Rev Genet* 13:227-232.

696   Wisniewski JR, Ostasiewicz P, Dus K, Zielinska DF, Gnad F, and Mann M. 2012. Extensive
697   quantitative remodeling of the proteome between normal colon tissue and
698   adenocarcinoma. *Mol Syst Biol* 8:611.

699   Wollfe A. 1998. *Chromatin: Structure and Function*. San Diego: Academic Press.

700   Wong PG, Winter SL, Zaika E, Cao TV, Oguz U, Koomen JM, Hamlin JL, and Alexandrow MG.
701   2011. Cdc45 limits replicon usage from a low density of preRCs in mammalian cells.
702   *PLoS One* 6:e17533.

703   Xiao C, Calado DP, Galler G, Thai TH, Patterson HC, Wang J, Rajewsky N, Bender TP, and
704   Rajewsky K. 2007. MiR-150 controls B cell differentiation by targeting the transcription
705   factor c-Myb. *Cell* 131:146-159.

706   Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, and Darnell
707   JE, Jr. 2003. Decay rates of human mRNAs: correlation with functional characteristics
708   and sequence attributes. *Genome Res* 13:1863-1872.

709   Zeiler M, Straube WL, Lundberg E, Uhlen M, and Mann M. 2012. A Protein Epitope Signature
710   Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed
711   determination of protein copy numbers in cell lines. *Mol Cell Proteomics* 11:O111
712   009613.

713

714

Table 1 (on next page)

The contribution of different steps in gene expression to the variance in protein abundances between genes.

[*] In this column, the value given for Schwanhausser et al.'s 2nd data is the variance in their measured protein abundances; the remaining values are our estimate for the variance in true protein levels for different scenarios.

[a] Estimates from Schwanhausser et al. based on the 4,212 genes for which NIH3T3 cell protein and mRNA abundance data are available.

[b] Our estimates for same the 4,212 genes studied by Schwanhausser et al. after correcting the overall scaling of the NIH3T3 cell protein abundance data and taking molecule specific stochastic and systematic experimental error into account.

[c] Our estimates for same the 4,212 genes studied by Schwanhausser et al. derived using measured translation rates from Subtelny et al.

Table 1. The contribution of different steps in gene expression to the variance in protein abundances between genes

	variance in protein levels(log10)[*]	Percent contribution to variance in protein levels				
		mRNA	Transcription	RNA degradation	Translation	Protein degradation
Schwanhausser 2nd data[a]	0.97	40%	34%	6%	55%	5%
Measured protein error strategy[b]	0.34	56%	38%	18%	30%	14%
Measured translation strategy[c]	0.61	81%	71%	10%	11%	8%

Figure 1

The steps regulating protein expression.

The steady state abundances of proteins and mRNAs are each determined by their relative rates of production (i.e. transcription or translation) and their rates of degradation.
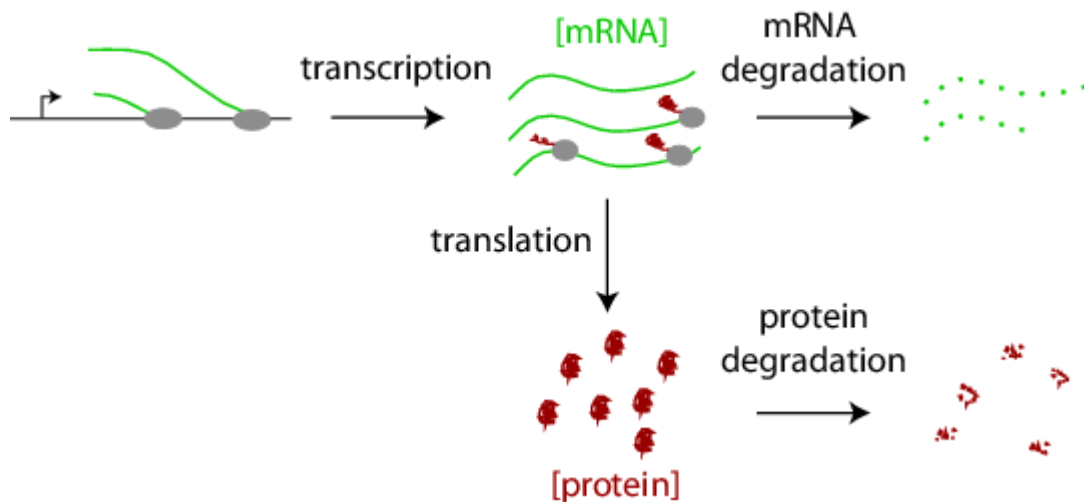
Figure 2

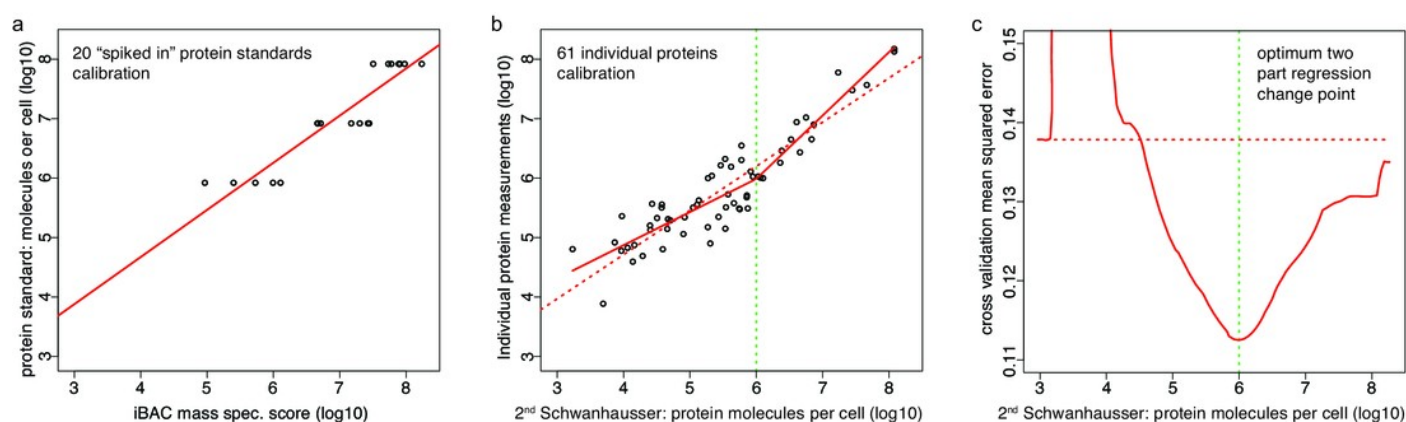A non-linear bias in protein abundance estimates and its correction.

a, The y axis shows the ratios of 61 individually derived protein abundance estimates divided by the abundance estimates from Schwanhausser et al.'s second whole proteome dataset. The x axis shows Schwanhausser et al.'s second whole proteome abundance estimates. The red line indicates the locally weighted line of best fit (Lowess parameter f=1.0), and the vertical dotted grey lines show the locations of the 1st quartile, median and 3rd quartile of the abundance distribution of the 5,028 proteins detected in the whole proteome analysis. **b,** The same as panel a. except that the whole proteome estimates of Schwanhausser et al. have been corrected using a two-part linear model and the abundances from the 61 individual protein measurements, see Fig 4b.

a

median

1st quartile   3rd quartile

individual protein / 2nd Schwanhausser

2nd Schwanhausser estimates: protein molecules per cell (log10)

b

median

1st quartile   3rd quartile

individual protein / corrected estimates

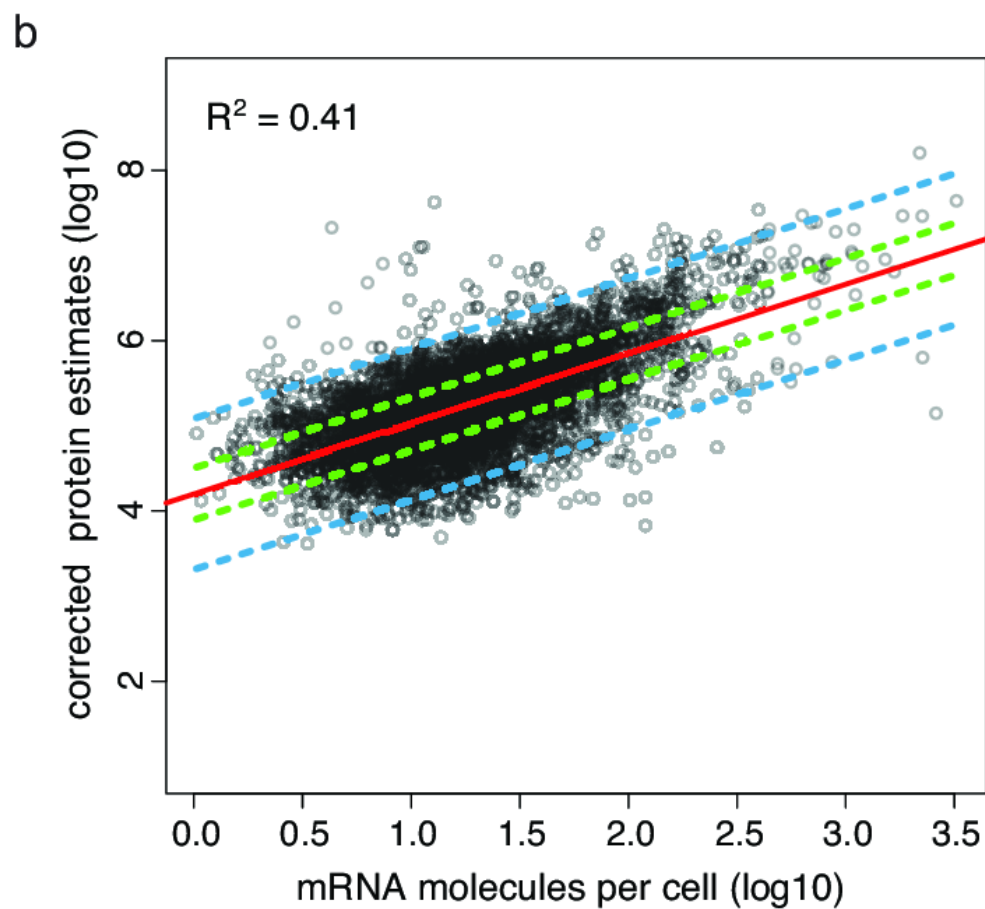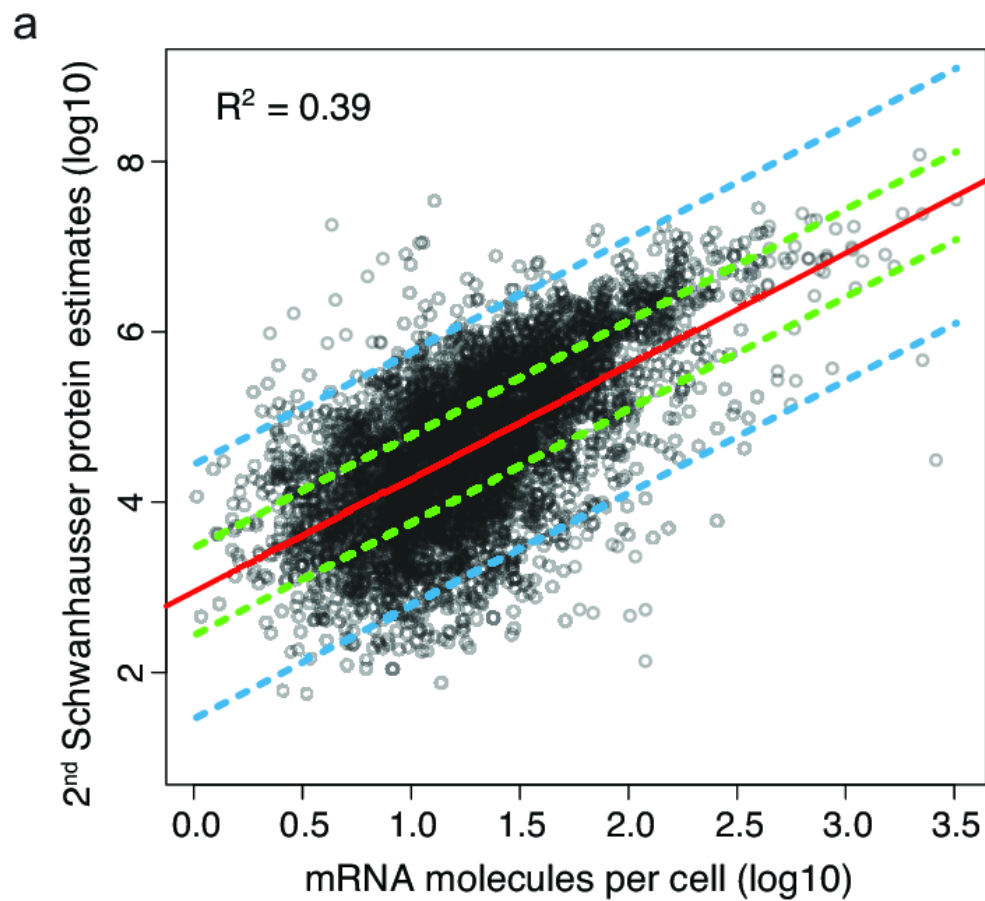corrected estimates: protein molecules per cell (log10)
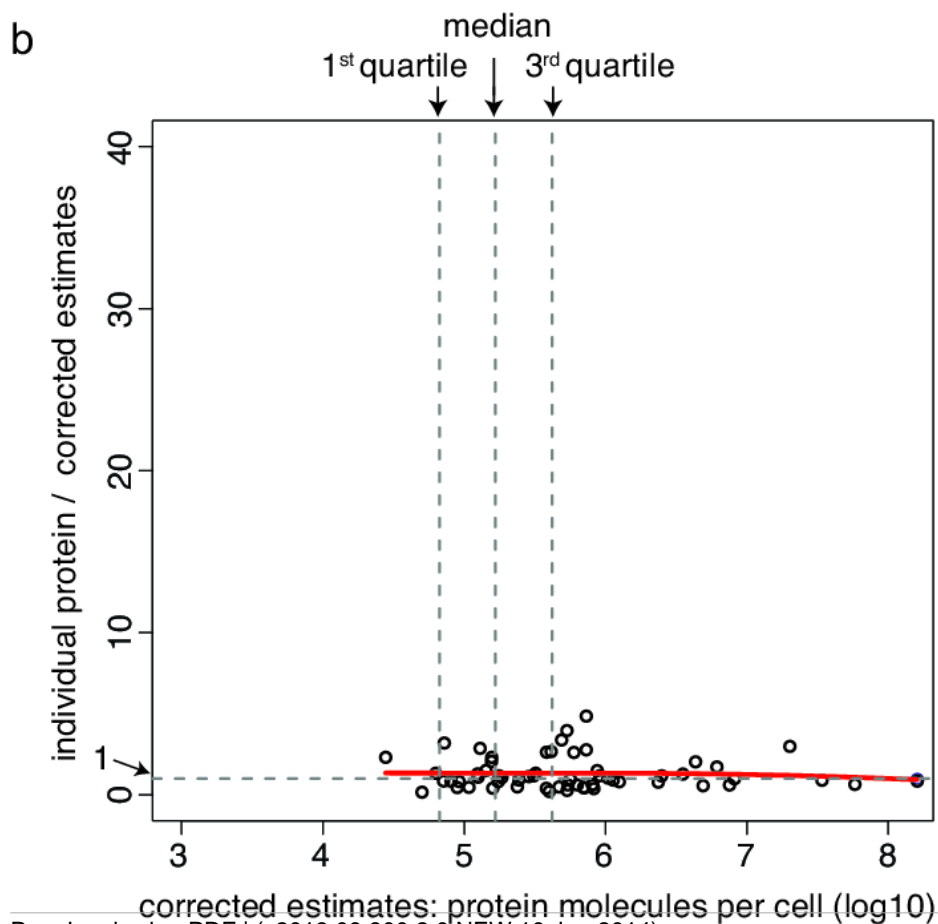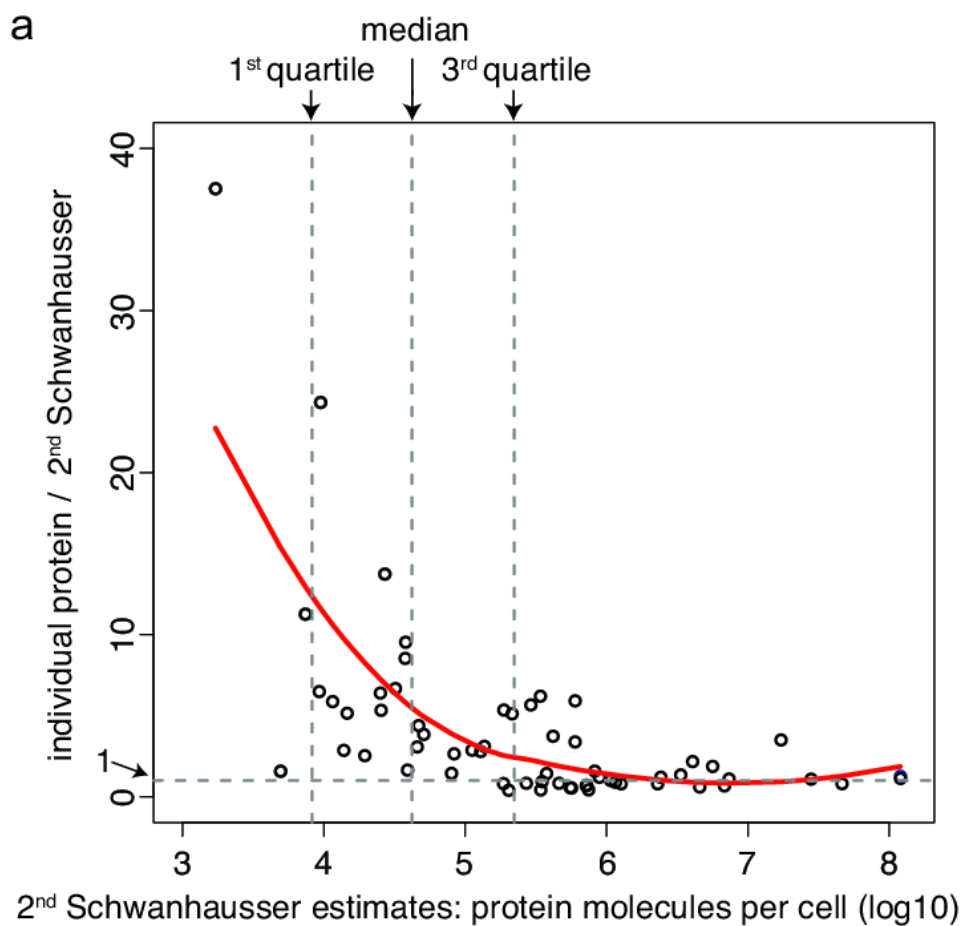
Figure 3

Calibrating absolute protein abundances.

a, The relationship between iBAC mass spectrometry signal (x axis) and the amounts of the 20 "spiked in" protein standards (y axis) used by Schwanhausser et al. to calibrate their whole proteome abundances (data kindly provided by Matthias Selbach, Dataset S2). The line of best fit is shown (red). **b,** The relationship between individually derived estimates for 61 housekeeping proteins (y axis) and Schwanhausser et al.'s second whole proteome estimates (x axis). The two part line of best fit used to correct the second whole proteome estimates is shown (solid red line) as is the single linear regression (dashed red line). **c,** The fit of different regression models for the data in panel b. The y axis shows the leave-one-out cross validation root mean square error for each model. The x axis shows the protein abundance used to separate the data for two part linear regressions. The red curve shows the optimum change point for a two part linear model is at an abundance of ~10^6 molecules per cell. The dashed red horizontal line shows the root mean square error for the single linear regression.

Figure 4

Protein abundance estimates versus mRNA abundances.

a, The relationship between Schwanhausser et al.'s second protein abundance estimates vs mRNA levels for 4,212 genes in NIH3T3 cells. The linear regression of the data is shown in red, the 50% prediction band by dashed green lines, and the 95% prediction band by dashed blue lines. **b,** The relationship between our corrected estimates of protein abundance vs mRNA levels. The linear regression and prediction bands are labeled as in panel a.

a

$R^2 = 0.39$

2nd Schwanhausser protein estimates (log10)

mRNA molecules per cell (log10)

b

$R^2 = 0.41$

corrected protein estimates (log10)

mRNA molecules per cell (log10)

Figure 5

Measured versus inferred translation rates.

a. The relative density of ribosomes per mRNA for each gene directly measured by ribosome profiling (Guo et al. 2010; Ingolia et al. 2011; Subtelny et al. 2013) (colored lines) compared to the translation rates for each gene inferred by Schwanhausser et al. (Schwanhausser et al. 2011) (black lines). The distribution of values from the ribosome profiling experiments was scaled proportionally to have the same median as that of the Schwanhausser et al. values, and the gene frequencies of the each distribution was normalized to have the same total. The locations of the 2.5 and 97.5 percentiles of the two distributions for NIH3T3 cells are shown as dashed lines. **b.** As panel a. except that the data for all genes in the Schwanhausser et al. dataset are shown in the solid black line and data for the genes in the intersection of the Schwanhausser et al. and Subtelny et al.'s datasets are shown in dashed lines. The variances and numbers of genes for each dataset are given in Table S1.

a

b
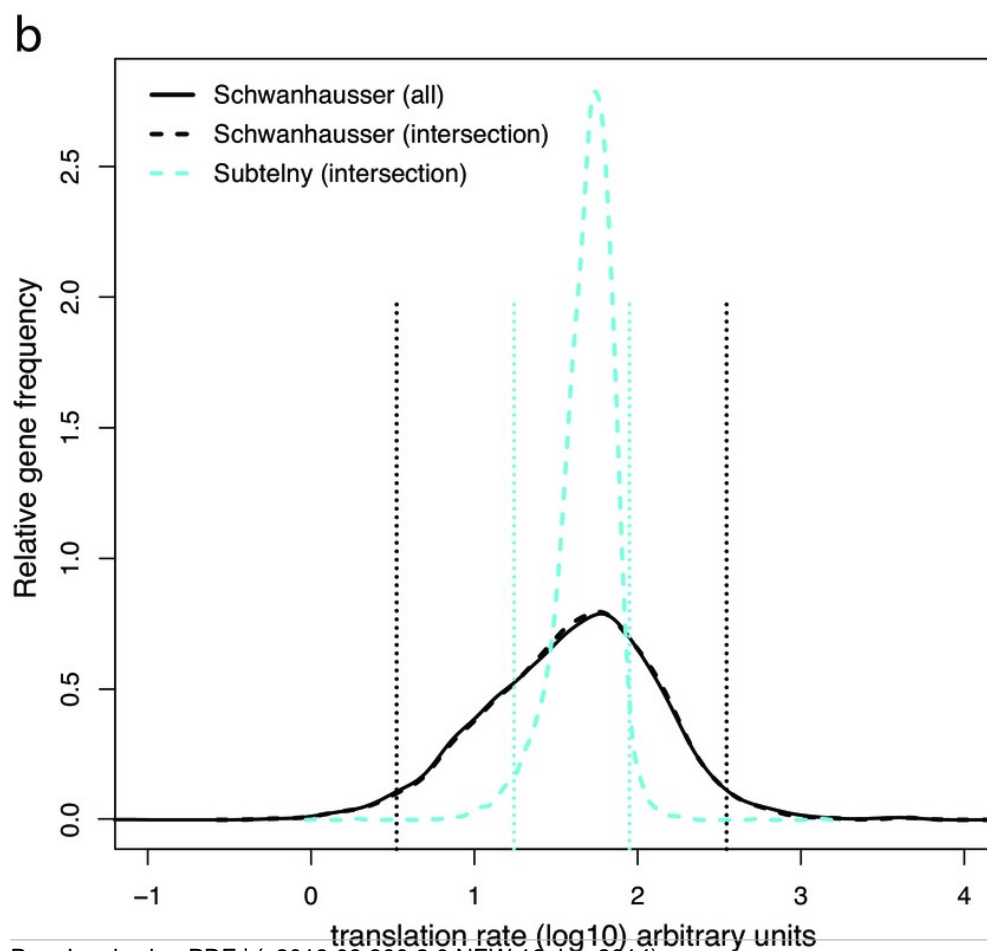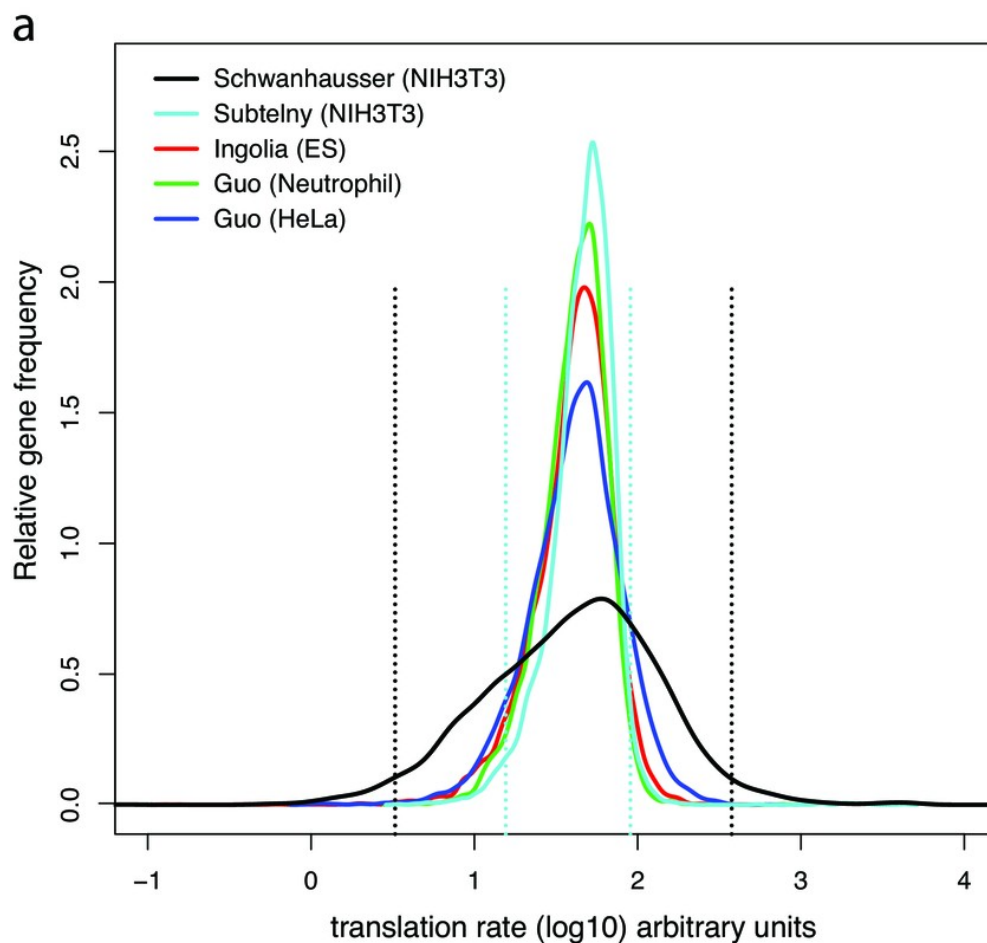
Figure 6

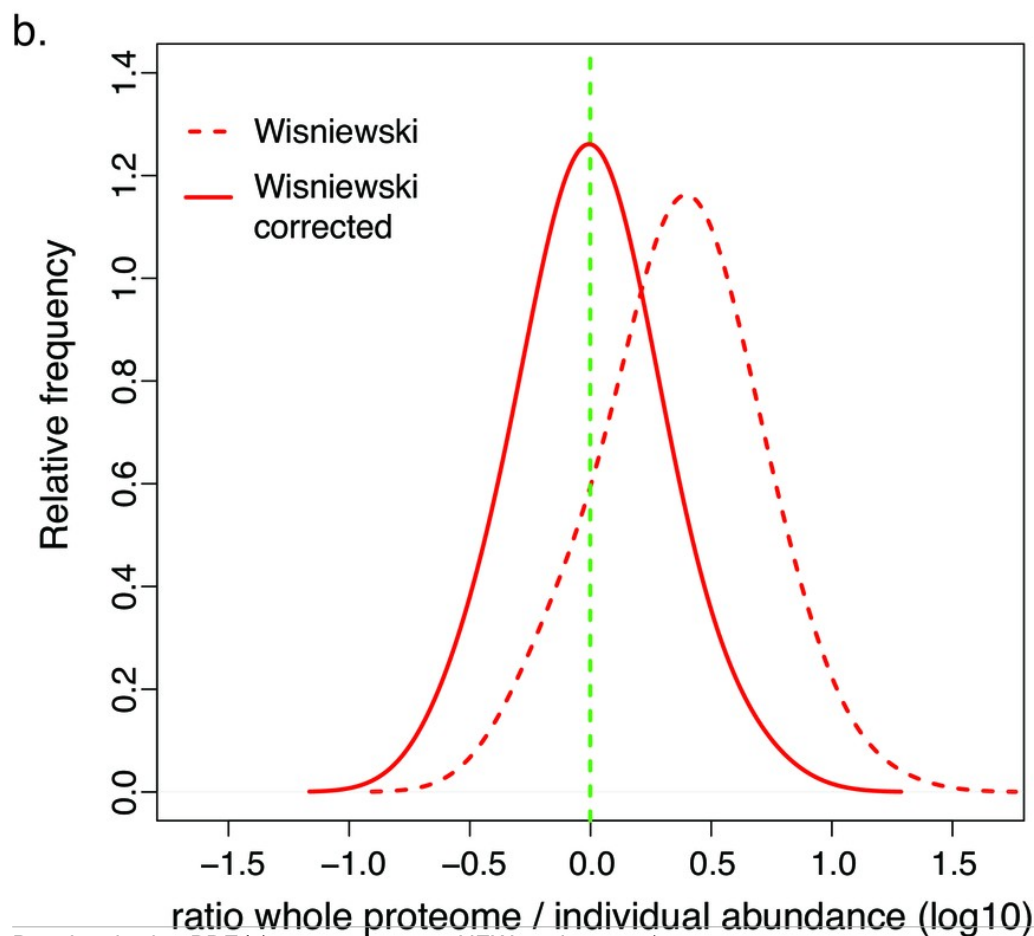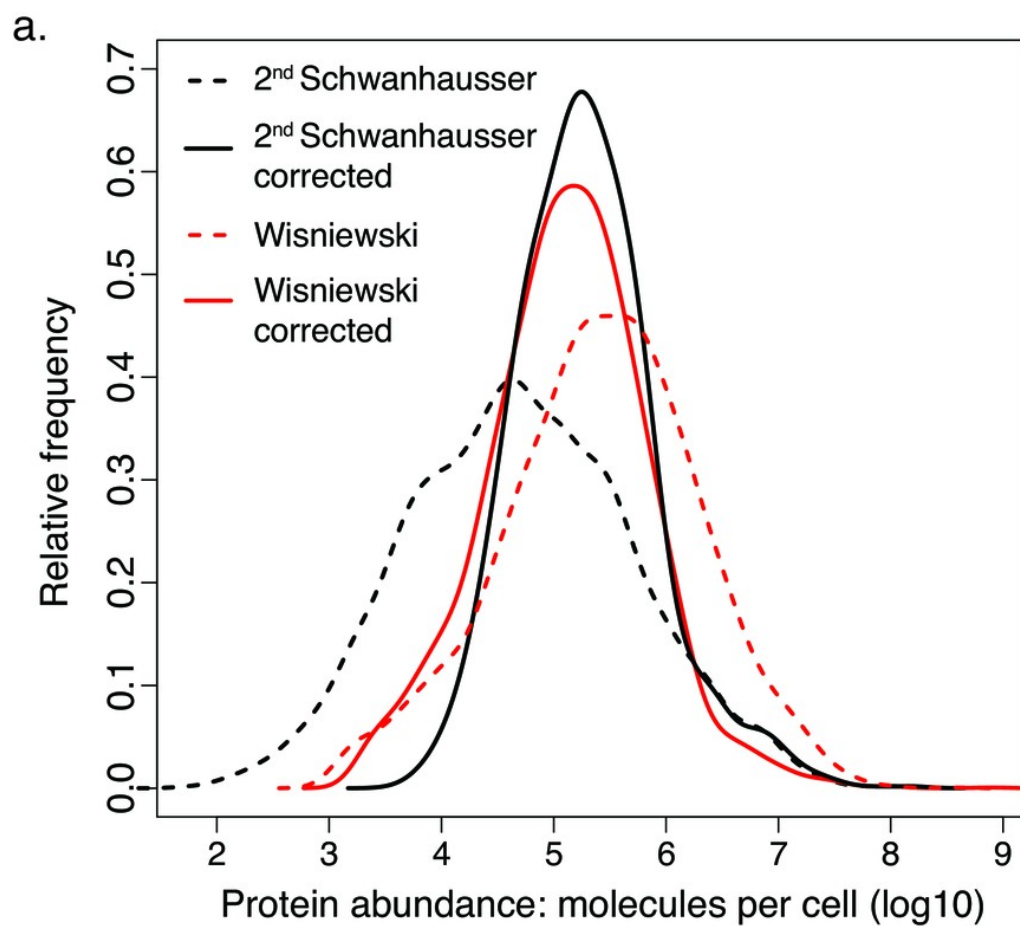Correlation between measured versus inferred translation rates.

The relationship between the measured rates of translation determined by Subtelny et al. using ribosome footprinting vs the inferred rates of translation determined by Schwanhausser et al for the same set of 3,126 genes in NIH3T3 cells, see Table S1 for further details. The units shown are those provided in the original datasets. The linear regression is shown.
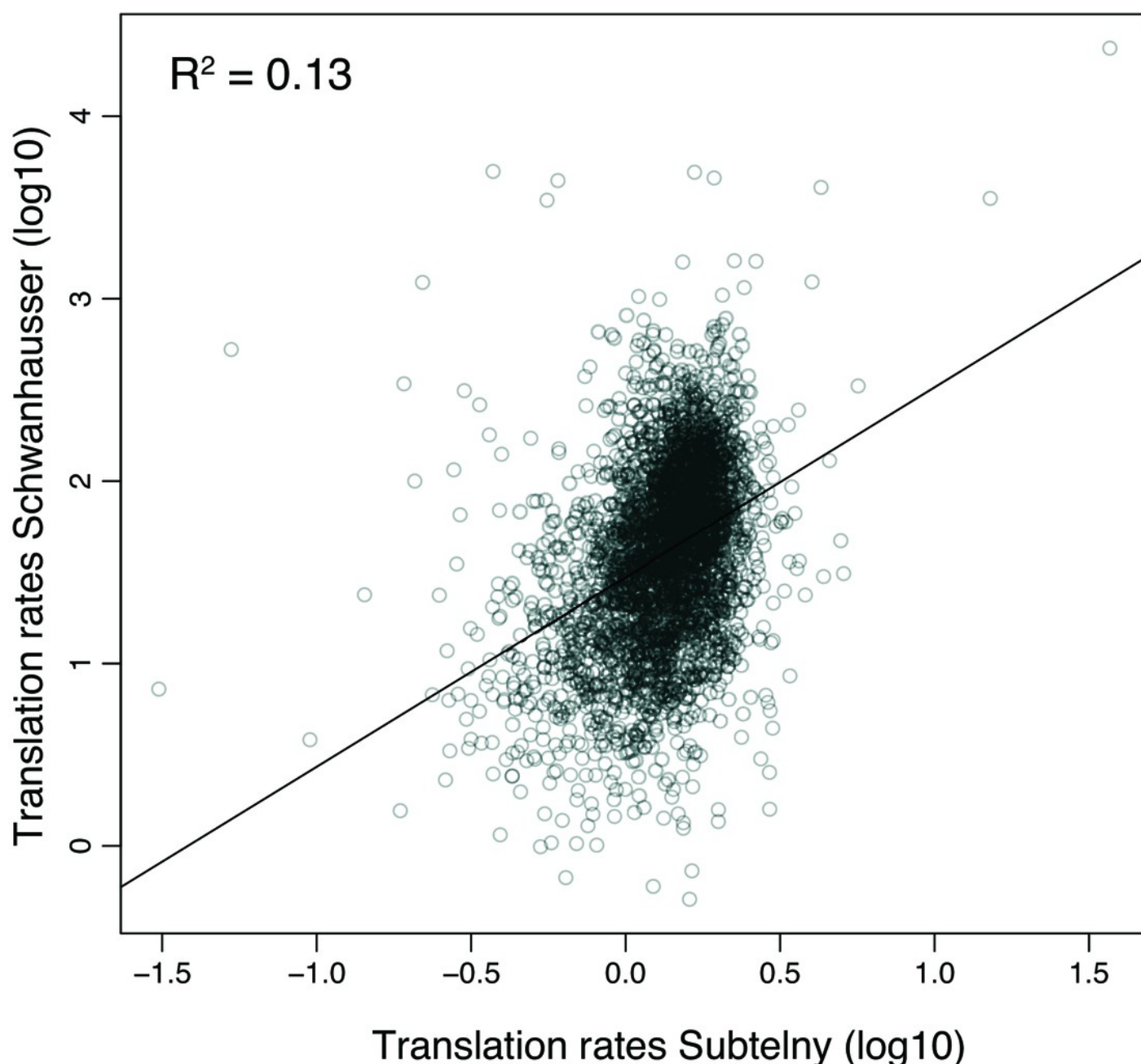
Figure 7

Comparison of corrected and uncorrected whole proteome abundance estimates.

a. The distributions of protein abundance estimates for 4,680 orthologous proteins in NIH3T3 cells (black lines) or HeLa cells (red lines). The values from Schwanhausser et al.'s second estimates and Wisniewski et al.'s estimates are shown as dashed lines. The values for our corrected abundance estimates are shown as solid lines. **b.** The ratios of HeLa cell whole proteome abundance estimates divided by individual measurements from the literature for 66 proteins. Results for the original data from Wisniewski et al. (dashed line) and after these values have been corrected (solid line) are plotted. The green dashed vertical line indicates a ratio of 1.

a.

Relative frequency

- - - 2nd Schwanhausser
— 2nd Schwanhausser corrected
- - - Wisniewski
— Wisniewski corrected

Protein abundance: molecules per cell (log10)

b.

Relative frequency

- - - Wisniewski
— Wisniewski corrected

ratio whole proteome / individual abundance (log10)

Figure 8

The relationship between true and measured protein and mRNA levels.