Gene 497 (2012) 228–236

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Quality criteria for finding genes with high mRNA–protein expression correlation and coexpression correlation

Gabriel Östlund [a,b,*], Erik L.L. Sonnhammer [a,b,c]

[a] Stockholm Bioinformatics Centre, Science for Life Laboratory, Box 1031, SE-17121 Solna, Sweden
[b] Department of Biochemistry and Biophysics, Stockholm University, Sweden
[c] Swedish eScience Research Center, Sweden

ARTICLE INFO

Article history:
Accepted 19 January 2012
Available online 4 February 2012

Keywords:
mRNA expression
mRNA coexpression
Protein expression
Protein coexpression
mRNA–protein expression concordance
Microarray

ABSTRACT

mRNA expression is widely used as a proxy for protein expression. However, their true relation is not known and two genes with the same mRNA levels might have different abundances of respective proteins. A related question is whether the coexpression of mRNA for gene pairs is reflected by the corresponding protein pairs. We examined the mRNA–protein correlation for both expression and coexpression. This analysis yielded insights into the relationship between mRNA and protein abundance, and allowed us to identify subsets of greater mRNA–protein coherence.

The correlation between mRNA and protein was low for both expression and coexpression, 0.12 and 0.06 respectively. However, applying the best-performing quality measure, high-quality subsets reached a Spearman correlation of 0.31 for expression, 0.34 for coexpression and 0.49 for coexpression when restricted to functionally coupled genes. Our methodology can thus identify subsets for which the mRNA levels are expected to be the strongest correlated with protein levels.

1. Introduction

The function of a protein is modulated by its expression level, or abundance. According to the transcriptional control paradigm, protein expression is governed by factors that control transcription of a gene into an mRNA. Consequently, mRNA measurements are often used as proxies for protein abundance. Thanks to the microarray and sequencing technologies, mRNA abundance is relatively easy and cheap to measure quantitatively. Protein expression data is more resource-consuming to obtain, and also has problems to reach full proteome coverage. Many projects have used mRNA coexpression data to predict functional coupling between the proteins they encode (Alexeyenko and Sonnhammer, 2009; Daub and Sonnhammer, 2008;Huttenhower et al., 2009; van Noort et al., 2003). However, the quantitative relationship between a mRNA and its encoded protein is far from entirely known (de Sousa Abreu et al., 2009). It is in fact highly complex and can be affected by factors such as systematic measurement errors and varying rates of translation, protein degradation, and mRNA degradation.

Much work has previously been done on comparing mRNA and protein abundances in various organisms (de Sousa Abreu et al., 2009; Greenbaum et al., 2003; Le Roch et al., 2004; Lu et al., 2007; Nie et al., 2007), reporting correlation coefficients of up to 0.7. For early studies, the correlation was probably an overestimate because the proteomics techniques could only measure high-abundance proteins (Ghaemmaghami et al., 2003). However, even with the best available techniques, the discrepancy between mRNA and protein abundances remains profound.

Attempts to model translation and protein decay rates based on sequence signals and features have been made (Brockmann et al., 2007; Nie et al., 2006; Schwanhäusser et al., 2011; Tuller et al., 2007; Vogel et al., 2010; Wu et al., 2008). While mRNA abundances provide a high contribution to explaining protein levels, equally much can be explained by features relating to translation and mRNA/protein turnover, and together they can explain up to 80% of protein abundance variation (Schwanhäusser et al., 2011). Still, even for a well-defined system with low measurement errors and a sophisticated model, a large part of the protein abundance factors remains unexplained when externally validated. In light of these results one might question the sensibility of using mRNA measurements as a proxy for protein abundance.

Even if mRNA often is an inaccurate proxy for protein abundance, it might still be possible to identify sets of genes where the coherence of mRNA and protein abundances is greater.

We have investigated distributions of mRNA–protein correlation in several ways. They witness a great heterogeneity in the data, i.e. some genes or gene groups have much higher correlation than others.

There are several potential sources of noise. Technical sources add some uncertainty and could potentially have systematic errors for all/some genes. Biological sources include temporal abundance variation

Abbreviation: HPA, Human Protein Atlas.

* Corresponding author at: Stockholm Bioinformatics Centre, Science for Life Laboratory, Box 1031, SE-17121 Solna, Sweden. Tel.: +46 8 55378566; fax: +46 8 5537 8214.
E-mail address: Gabriel.Ostlund@sbc.su.se (G. Östlund).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2012.01.029

as well as variation between individual cells, composition of cell populations, cell types, conditions and individuals. Biological noise also includes the extent to which protein abundance is regulated by mRNA abundance — the uncertainty of protein abundance would be large if it is only partly dependent on mRNA abundance.

Intrinsic properties of an expression profile such as signal-to-noise ratio and variability can affect the correlation. This suggests that refining the data by filtering, higher mRNA—protein concordance can be obtained. We propose methods to measure the inherent quality of an expression profile, and have evaluated their performance on global datasets. High quality here means profiles with low noise levels and/or with high capacity to detect mRNA—protein correlation if it exists.

The study of mRNA and protein expression concordance has so far generally been done on the level of expression, i.e. by calculating the correlation of expression levels across a set of conditions between the mRNA and its corresponding protein. An alternative approach to study the expression concordance is to calculate the second order correlation, or correlation of coexpression. This signifies how well mRNA—mRNA coexpression corresponds to protein—protein coexpression for the corresponding mRNA and protein pairs (Fig. 1). Coexpression correlation has previously been studied in the contexts of finding functional relationships within a proteome (Zhou et al., 2005), to study functional conservation between species (Dutilh et al., 2006), and to examine mRNA—protein concordance in cancer cell lines (Shankavaram et al., 2007).

As mentioned above, the bottleneck in mRNA—protein expression comparisons is usually the lack of proteomics data and the generally low coverage of the proteome. To meet the need of measuring all human proteins, the Human Protein Atlas (HPA) is being constructed as a global database of protein abundances in over 100 normal and cancer tissues (Uhlen et al., 2005). It is based on antibodies designed

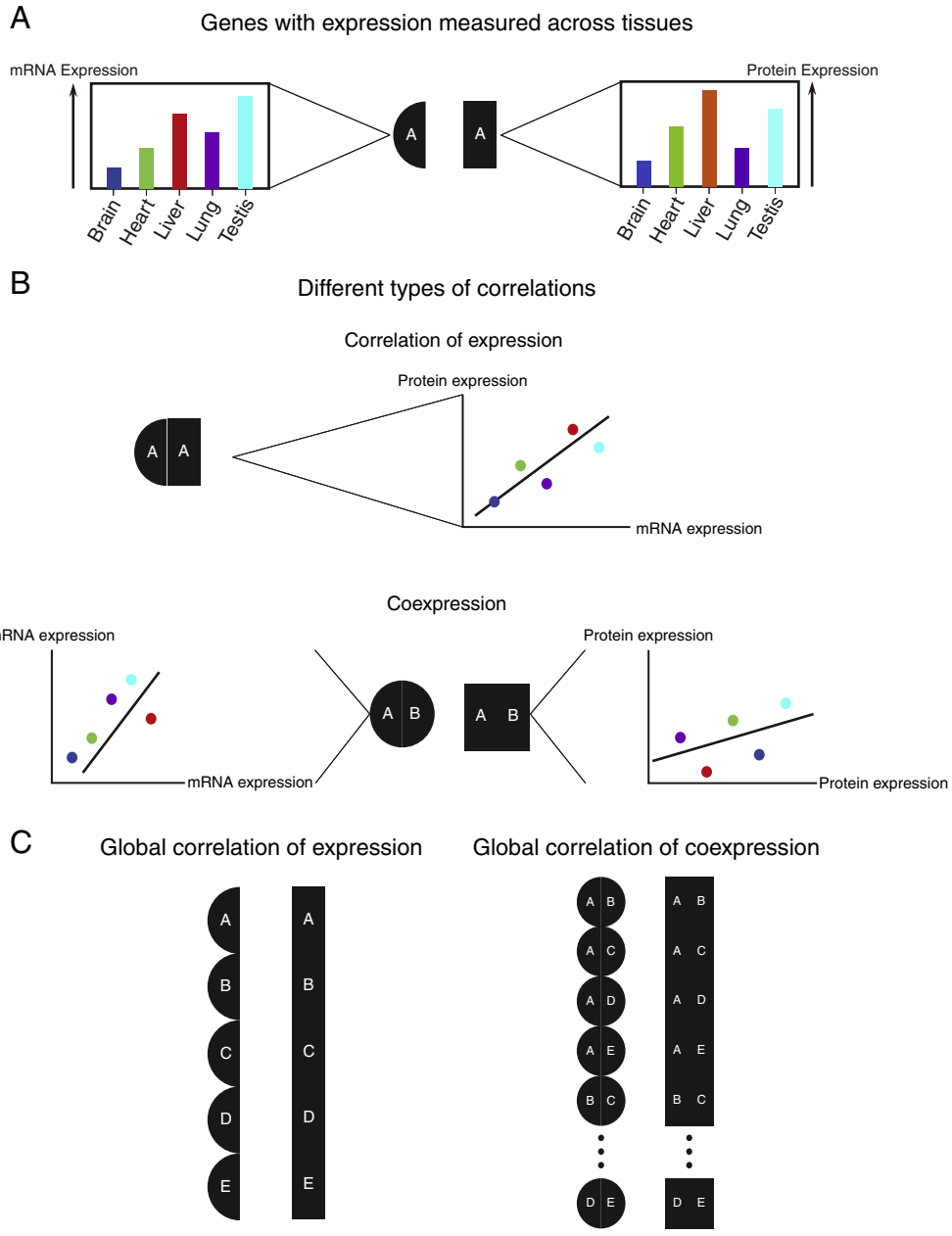

Fig. 1. Principles of mRNA and protein expression comparison. mRNA transcripts are drawn as semicircles and proteins as rectangles. A. If an mRNA and a protein are quantified in the same set of tissues, a correlation coefficient between the mRNA and protein expression profiles can be calculated as shown in B. This is referred to as correlation of expression. Alternatively, correlations between two mRNAs or two proteins can be calculated. This is referred to as coexpression, and sometimes as second order correlation. C. Global mRNA–protein correlation can either be calculated from expression profiles (left), or from correlation profiles (right).

for each human protein that stain tissue samples which are subsequently evaluated by pathologists to quantify the abundance.

To match the HPA tissue-specific protein expression with tissue-specific mRNA expression we used the microarray dataset from the Genomics Institute of the Novartis Research Foundation (Su et al., 2002). It is based on a custom designed array for measuring transcript expression of all human genes in 79 tissues. These datasets represent some of the problems that may be encountered when using publicly available data, e.g. that samples are of related but different nature and that the units of measurement are not directly translatable. In this case many of the tissues are not exactly the same for both datasets, introducing the need to draw a map of corresponding tissues.

To our knowledge there has been no large scale study of mRNA−protein expression concordance across human tissues, and also not on the concordance of mRNA and protein coexpression across human tissues. We here present a study that includes both of these novel aspects. The study was first done as a global comparison, using all mRNA/protein expression or mRNA−mRNA/protein−protein coexpression pairs simultaneously, followed by a genewise analysis, looking at single mRNA/protein expression pairs or single mRNA−mRNA/protein−protein pairs in the case of coexpression.

We show that using quality measures it is then possible to separate genes with high quality from those where any signal is likely to be drowned by noise. When limiting the comparison to the highest quality parts of the mRNA and protein data their correlation can become relatively high, despite their great differences in origin and generation. The main result of our study is the establishment of a methodology to make high-confidence mappings between mRNA and protein levels. The ability to do this is valuable to the vast amount of research that uses the wealth of mRNA expression data to infer protein-level properties since it enables reliable use of mRNA measurements as a proxy for protein levels.

2. Materials and methods

2.1. Protein expression dataset

The protein dataset (HPA) was obtained through a collaboration with the Swedish Human Proteome Resource (HPR) Center. It consists of protein expression measurements for 102 combinations of tissues and cell types (Table S1) derived from antibody-based tissue profiling (Berglund et al., 2008; Uhlen and Ponten, 2005). Specifically, there are 6104 protein epitope signature tags, antigens used in antibody generation designed to be unique for a target protein, corresponding to 4908 genes. Each tag, or probe, is measured in three biological samples of each tissue. The expression values, given as colors, were recoded to 1–4. If the value for all three samples was black, it was coded as value missing; otherwise it was recoded according to the non-black part. If two or more probes corresponded to the same gene, the recoded expression values for the probes were averaged for that gene. Of the 4908 genes in the HPA set, 3646 were also present in the mRNA expression dataset and used for analysis. When used for calculating correlation of expression, values below 1.5, indicating absence, were set to NA. This was done in order to avoid a high proportion of absence values from biasing the correlation. A concordance of absence is not as meaningful as a concordance of abundance. Samples from cancer tissues were omitted because of the higher heterogeneity of tumors as compared with normal tissues (Uhlen et al., 2005).

2.2. mRNA expression dataset

The microarray dataset, the Human U133A + GNF1H (MAS5-condensed) dataset, was downloaded from http://symatlas.gnf.org (Su et al., 2002). The dataset consists of two parts; the HG-U133A Affymetrix chip and the custom made GNF1H chip. Together, the two contains 44,775 probe sets corresponding to 33,698 gene models. For each probe set there are two replicates for 79 human tissues (Table S2), organs and cell lines. Probe identifiers were mapped to gene identifiers followed by a pruning of probes mapping to more than one gene. After mapping and pruning, 22,572 probes corresponding to 13,733 genes remained. Where several probes mapped to the same gene, expression values for the probes were averaged for that gene. Of the 13,733 genes, 3646 were also present in the protein expression dataset and used for analysis.

2.3. Quality measures

In order to stratify the data into subsets of different quality, a number of measures were calculated from each gene's expression profile. For the mRNA data the following measures were used:

- Mean expression: the arithmetic average of the gene's expression in all tissues.
- Replicate correlation: Spearman correlation between the two biological replicates across all tissues. This quality measure indicates the levels of biological and measurement noise.
- Entropy:

$\sum p_i \log p_i$, $p_i = e_i / \sum e_i$ where e_i is the gene's expression level in tissue i. This corresponds to the negative Shannon entropy with tissues representing symbols and frequency approximated by expression levels. This quality measure will separate uniform expression profiles from those with a wide span of expression values.

- Span: the maximum expression minus the minimum. This quality measure indicates the maximum absolute span of expression values in the expression profile.
- Normalized span: the maximum expression minus the minimum divided by the mean expression. This quality measure is the same as the span, but takes the mean level of expression into account.
- Median expression: the median of the gene's expression in all tissues.
- Sd: standard deviation of the gene's expression in all tissues. This quality measure indicates the variability of values in the expression profile.
- Normalized sd: standard deviation divided by the mean expression. This quality measure is the same as sd, but takes the mean level of expression into account.
- Mean + entropy: the combined sum of mean expression and entropy ranks. This quality measure separates low uniform expression profiles from those with high non-uniform expression profiles.

In all cases where multiple probes existed for a given gene, the arithmetic average of these values was used. Quality measures were generated separately for all tissues, for mapped tissues and for non-mapped tissues.

For the protein data, the antibody validation score provided in the HPA database (Berglund et al., 2008) was used, which is based on immunohistochemistry, immunoflourescence, protein array and western blot data. It is discretized at four levels: very low, low, moderate, and high. The number of proteins in each level was 487, 1524, 1173, and 462.

2.4. Mapping of tissues

Tissues of the protein dataset were mapped to tissues of the mRNA dataset. Where tissues in one dataset corresponded to several tissues in the other dataset an average was taken to represent the pool of tissues, e.g. the mRNA dataset contains expression for lymph node while the protein dataset has separate measurements for follicle cells and non-follicle cells of the lymph node. Similarly where several tissues of one dataset corresponded to several tissues in the other dataset averages were taken for both pools of tissues, e.g. the mRNA dataset has measurements for cerebellum and cerebellum peduncles while the protein dataset has measurements for cells in granual layer,

molecular layer and purkinje cells of the cerebellum. The mapping resulted in 21 common tissues/tissue pools, see Table S3.

2.5. Quality measure defined subsets

Quality measure defined subsets were generated in order to filter out genes for which noise and noise sensitivity would obscure mRNA—protein correlations. Genes were partitioned into four equally sized bins according to quartiles for each continuous mRNA quality measure, and also according to protein antibody validation. 16 mRNA—protein subsets were thus generated by intersecting each mRNA subset with each protein subset. For more information on individual subset sizes, see Table S5.

2.6. Global correlation of expression

Global correlation of expression was calculated using the entire data of mapped tissues. This was done correlating mRNA expression for all genes in all tissues with corresponding protein expression for the same genes in the same tissues. In addition to doing this for the entire data, the analysis was also performed for quality measure defined subsets. More specifically, the correlation for the subset was calculated using only the mRNA and protein expression for the genes of the subset. Resampling based P-values were determined for each subset by randomly sampling the same number of genes as found in the subset. This resampling was repeated 10,000 times for each subset. A z-score for the correlation of the subset was computed based on the distribution of correlations in the random samples, and a resampling p-value for the subset's correlation significance was computed from the z-score.

2.7. Genewise correlation of expression

Correlation between protein- and mRNA expression for a gene was calculated by correlating the protein expression with the mRNA expression of that gene in the 22 tissues/tissue pools mapped between the datasets. This was calculated for all genes common to both datasets.

2.8. Global correlation of coexpression

The global correlation of coexpression was calculated, similarly to Shankavaram et al. (2007), by first calculating the coexpression for all possible pairs of each quality subset, separately for mRNA and protein. Following that, the correlation between the mRNA—mRNA and protein—protein coexpressions was calculated for all pairs where both mRNAs and proteins existed in both datasets. For both the mRNA data as well as the protein data, the coexpression correlation was calculated using both mapped tissues, non-mapped tissues as well as all tissues. The obtained correlation of coexpression is an indication of to what extent genes with coexpressed mRNA also have coexpressed proteins. Resampling based P-values were determined in the same manner as for global correlation of expression.

2.9. Pairwise correlation of coexpression

The pairwise correlation of coexpression was obtained by first calculating all pairwise coexpression correlations for one gene and all other members of the quality subset the gene belonged to. This was done separately for mRNA and protein data. Finally the correlation of coexpression was obtained by calculating the correlation of mRNA coexpression correlations with protein coexpression for that gene. This was done for all genes. The obtained coexpression correlations for a gene is an indication of to what extent the same genes with which it has a coexpression of mRNA, it also has a coexpression with on the protein level.

2.10. Global and pairwise correlation of coexpression restricted to functional pairs

Global as well as pairwise correlation of coexpression was done using only functionally coupled gene pairs. These were obtained by taking all gene pairs connected by a link in FunCoup (Alexeyenko and Sonnhammer, 2009) with a confidence above 0.25, after support from mRNA and protein coexpression had been removed.

2.11. Generation of percentile scores

The correlation of true (corresponding) mRNA—protein profile pairings was compared to false pairings, i.e. where an mRNA or protein is compared with a non-corresponding protein or mRNA, respectively. The placement of the true correlation in the distribution of all possible false pairings was used to obtain the percentile score, i.e. how a large proportion of the false pairs were closer to zero than the true correlation. This was done for all genes and was used both for genewise correlations of expression and similarly for pairwise correlations of coexpression. For a globally fair comparison, the correlations in each subset were compared to false pairs generated from the entire dataset. As an example, even if the mRNA for a gene has a high correlation of 0.9 to its protein, even higher correlations might exist to other proteins. There may also be higher correlations between the protein of that gene and the mRNA of other genes. Thus, even if the correlation between mRNA and protein is high, it may not be significant.

2.12. Calculation of correlations

Calculations of correlation were done with the R (Ihaka and Gentleman, 1996) functions cor()/cor.test() (with/without estimated p-value) for pairwise complete observations. The Spearman correlation was used throughout.

3. Results

Our goal was to establish the level of expression and coexpression correlations between mRNA and protein on a global transcriptomic and proteomic scale (Fig. 1), and to investigate principles that may cause such correlations to be reduced. For mRNA expression we used the Novartis database, measured on 79 tissue types, and for protein expression we used the HPA database, measured on 102 tissue types. These data were generated by independent groups using tissue samples from different individuals, hence the comparison is considerably more challenging than in previous studies using single cell lines or tissues. Because many of the tissue types differ between the datasets, we defined 22 tissue classes to which most of the data could be mapped (Table S3). Before analyzing the mRNA—protein correlation of expression and coexpression, we first validated both datasets functionally to assert that they capture biologically relevant information (Supplementary materials).

3.1. Global correlation of expression

Using the entirety of the tissue-mapped data, the global Spearman correlation of expression between mRNA and protein was calculated across the corresponding tissues. The resulting correlation was merely 0.12 ($p<2.2e-16$) which is lower than observed in previous studies but not surprising given the greater challenges of this study. This correlation is probably affected by only having the four discrete expression values in the protein data. For example, it might be positively affected by the predominance of zero expression (24.7% of the protein data). While Spearman correlation is less sensitive to outliers than Pearson correlation, such a bias can persist for discrete data like the protein expression used here. In fact, looking only at non-zero protein expression the correlation dropped to 0.08 ($p<2.2e-16$).

Given the low global correlation, taking measures to reduce sources of noise may lead to better results.

3.2. Designing quality measures for mRNA data

Since measurement errors and sample differences are likely to introduce noise, identifying a subset of genes where there is a higher capacity to detect correlation if it exists should improve the results. By using measures of quality, genes for which noise is likely to dampen correlation could be filtered out. For the protein expression data, the validation score that is provided for each HPA antibody was used as a quality measure. A low HPA validation score may for instance be the result of high cross reactivity. For the mRNA data, no quality or validation was provided, but rationales exist to create such measures.

Some expression patterns might be problematic for estimating correlation and can indicate poor quality. For example, a gene with a very narrow expression interval (i.e. the same expression level in all tissues) may obtain a low correlation because the effect of the noise would be larger than the signal. When differences in expression levels are relatively small between tissues, noise would be more likely to cause erroneous ranking than if the differences in expression levels are relatively large. This can be gauged e.g. by the entropy of the gene's expression values, the span between maximum and minimum values, or by the standard deviation of the expression profile. Also, genes with an overall low expression level suffer more from noise effects, and therefore two genes with high levels of expression are more likely to result in a more accurate correlation. Because of this, mean or median expression level can be used as a quality measurement.

The mRNA data consist of a set of one or more probes for each gene and contain two biological replicates for each tissue. Ideally, the measurements for the same gene in both replicates would be the same. This can be measured by the correlation between the tissue expressions in the two biological replicates and used as a quality measure for each gene.

Different quality measures could capture different properties, and a combination may be superior to any single measure. To investigate the scope for this, we calculated the correlation between all measure pairs across all genes (Table 1). This showed that some measures are strongly correlated, for instance mean expression was strongly correlated with both minmax and sd. To counteract this dependency we normalized them by dividing with the mean expression, which instead made them highly correlated with entropy. However, other measure pairs were hardly correlated, for instance entropy and mean expression, indicating that a combination of these two could lead to further quality refinement.

3.3. Quality measures can improve mRNA−protein expression correlation

Does a subset of genes with lower measurement noise have a higher mRNA−protein expression correlation? To examine this, the global correlation was calculated for subsets of genes based on the different quality measures (Table 2). The subsets are of equal size for mRNA data but not for protein data. As a general trend, the subset combining the highest mRNA and protein quality had the highest correlation, but in some cases where the difference between neighboring subsets was small this did not hold.

As expected, a higher mean mRNA expression resulted in higher mRNA−protein correlation. The best subset had a Spearman correlation of 0.23 ($p<2e-16$). Entropy gave even higher correlation, 0.28 ($p=9.90e-38$). Replicate correlation resulted in a maximal correlation of 0.19 ($p<2e-16$), but surprisingly it was found in the second quartile (25−50%) and with moderate antibody quality. This may be due to the fact that the replicated correlation in the highest quartiles is almost 1.0, suggesting some sort of artifact such as outliers. Median surprisingly had results quite different from mean which could be due to differences in ranking profiles with e.g. bimodal expression pattern.

As mentioned above, the quality measures mean expression and entropy were largely independent, and were therefore combined into a new measure called mean + entropy. This resulted in higher correlations than using mean and information content separately. As seen in Fig. 2, genes with a high mean + entropy mRNA quality and high protein expression quality tend to have a higher correlation, up to 0.31 ($p<2e-16$). The correlations of the subsets were compared to distributions from randomly sampled subsets of equal size. As seen in Table S7 the correlations for the highest and lowest quality subsets are significantly higher and lower, respectively, than the correlation for the entire data.

While 0.31 is still a fairly low correlation, it is substantially higher than for the whole data set, showing that if coherent mRNA–protein expression is present it can be extracted by restricting analysis to higher quality subsets.

3.4. Genewise correlation of expression

Examining the correlation of entire datasets used as a pool of measurements gives the overall correlation for all genes in a given data(sub)set. It is of interest to stratify the dataset by single genes instead, to investigate the distribution of mRNA–protein correlation across all genes. Correlation of expression was calculated for all genes using the expressions from the mapped tissues. The genes were partitioned into the same subsets as above based on mean + entropy and HPA antibody validation quality measures, and within each subset the correlation distribution was generated. Fig. 3A shows the correlation distribution for the subset with the highest quality in both mRNA and protein expression. Although the average is only 0.32, about the same as the global correlation for this subset, it is clear that many genes have a considerably higher correlation, some even approaching 1.0.

Since the genewise correlations could suffer from some of the problems inherent to the global correlation, it might be prudent to not look at the absolute correlation value, but instead how unexpected it is given the dataset. mRNAs with an equally high or higher correlation to randomly chosen proteins than to the corresponding protein are obviously not biologically significant. Artificially high correlations could be caused by biases inherent to the data such as outlier data points or small

Table 1
Correlations between mRNA quality measures, across all genes in the mRNA expression dataset. A correlation close to zero indicates independence whereas one close to 1 or −1 indicates a high interdependence.

	Replicate correlation	Mean expression	Median expression	Span	Sd	Entropy	Normalized span	Normalized sd
Replicate correlation	1.00	x	x	x	x	x	x	x
Mean expression	0.59	1.00	X	X	X	X	X	X
Median expression	0.51	0.90	1.00	X	X	X	X	X
Span	0.63	0.84	0.62	1.00	X	X	X	X
Sd	0.66	0.86	0.64	0.99	1.00	X	X	X
Entropy	0.38	0.10	−0.19	0.56	0.55	1.00	X	X
Normalized span	0.31	0.09	−0.20	0.57	0.53	0.96	1.00	X
Normalized sd	0.36	0.10	−0.20	0.56	0.55	0.99	0.98	1.00

Table 2
Global mRNA–protein correlations and their significance for quality measure defined subsets. The mRNA subsets (columns) were defined as quartiles using seven different quality measures, while the protein subsets (rows) were defined as antibody validation categories throughout. The highest correlation for each mRNA quality measure is shown with bold numbers.

Correlations					P-values				
Replicate correlation	0–25	26–50	51–75	76–100	**Replicate correlation**	0–25	26–50	51–75	76–100
V. low	−0.03	−0.07	−0.05	0.01	V. low	0.14	1.18E-003	6.45E-002	7.09E-001
Low	0.02	−0.03	0.05	0.12	Low	0.13	1.12E-002	1.02E-005	1.17E-017
Moderate	0.04	**0.19**	0.12	0.08	Moderate	0.02	2.43E-034	2.15E-015	1.86E-010
High	−0.01	0.17	0.16	0.14	High	0.72	6.45E-009	5.39E-011	2.80E-013
Mean expression	0–25	26–50	51–75	76–100	**Mean expression**	0–25	26–50	51–75	76–100
V. low	−0.06	0.05	0	0.02	V. low	3.60E-003	1.80E-002	9.48E-001	6.06E-001
Low	0.02	0.03	0.07	0.12	Low	9.31E-002	6.85E-003	1.14E-007	2.29E-019
Moderate	0.08	0.1	0.1	0.12	Moderate	1.21E-007	3.34E-010	5.86E-013	1.99E-019
High	0.1	0.17	**0.23**	0.21	High	4.03E-003	2.24E-012	8.02E-022	1.59E-027
Median expression	0–25	26–50	51–75	76–100	**Median expression**	0–25	26–50	51–75	76–100
V. low	−0.07	0	0.04	0.03	V. low	9.64E-004	9.29E-001	1.05E-001	3.10E-001
Low	0.03	0.04	0.05	0.12	Low	6.79E-003	4.06E-004	1.56E-005	1.57E-019
Moderate	0.06	0.07	0.03	0.02	Moderate	1.32E-004	2.51E-005	2.39E-002	1.44E-001
High	**0.13**	0.07	0.11	0.08	High	1.66E-005	5.07E-003	7.56E-006	1.11E-004
Entropy	0–25	26–50	51–75	76–100	**Entropy**	0–25	26–50	51–75	76–100
V. low	−0.07	0	0.02	−0.09	V. low	1.31E-003	9.86E-001	3.79E-001	2.20E-003
Low	0.02	0.05	0.02	0.04	Low	1.46E-001	1.37E-004	1.14E-001	7.98E-003
Moderate	0.12	0.07	0.1	0.11	Moderate	6.55E-016	3.83E-007	6.12E-012	1.55E-013
High	−0.02	0.08	0.24	**0.28**	High	5.46E-001	1.55E-003	1.78E-023	9.90E-038
Normalized span	0–25	26–50	51–75	76–100	**Normalized span**	0–25	26–50	51–75	76–100
V. low	−0.06	−0.03	0.06	−0.12	V. low	1.60E-003	1.71E-001	1.26E-002	8.24E-005
Low	0	0.02	0.07	0.02	Low	7.81E-001	6.03E-002	7.95E-008	7.30E-002
Moderate	0.11	0.07	0.11	0.12	Moderate	2.87E-013	1.12E-006	4.50E-015	7.44E-015
High	−0.06	0.14	0.21	**0.27**	High	1.98E-002	2.29E-009	2.92E-016	4.13E-035
Normalized sd	0–25	26–50	51–75	76–100	**Normalized sd**	0–25	26–50	51–75	76–100
V. low	−0.06	−0.02	0.04	−0.11	V. low	1.73E-003	3.91E-001	1.49E-001	4.35E-004
Low	0.02	0.04	0.03	0.03	Low	1.44E-001	8.63E-004	7.66E-003	3.28E-002
Moderate	0.12	0.05	0.12	0.12	Moderate	1.81E-016	2.93E-004	1.91E-016	1.76E-014
High	−0.01	0.08	**0.27**	0.26	High	6.13E-001	1.60E-003	3.88E-029	1.34E-032
Mean + entropy	0–25	26–50	51–75	76–100	**Mean + entropy**	0–25	26–50	51–75	76–100
V. low	−0.13	0	−0.05	0.08	V. low	1.57E-011	9.39E-001	5.93E-002	3.78E-002
Low	0	−0.02	0.08	0.12	Low	8.73E-001	9.41E-002	3.30E-010	4.20E-019
Moderate	0.07	0.13	0.18	0.16	Moderate	8.97E-006	1.95E-016	2.01E-037	1.70E-031
High	−0.01	0.19	0.25	**0.31**	V. low	8.52E-001	1.63E-010	1.22E-026	3.49E-057

number of data points. In order to examine this problem the correlation between each true mRNA–protein pairs was compared to the correlations of all false mRNA–protein pairs of the same mRNA and protein.

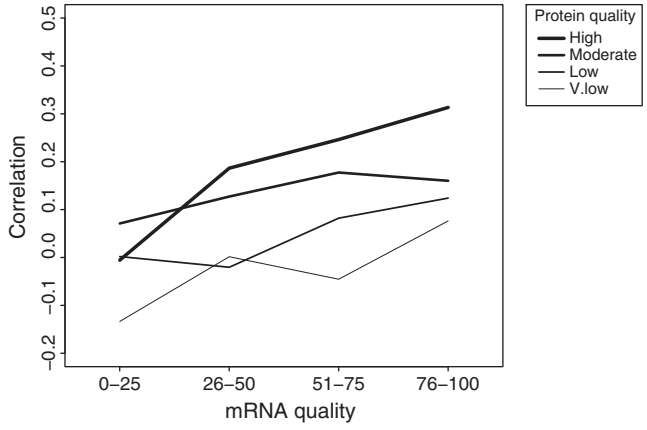

Fig. 2. Global correlation of mRNA–protein expression for 16 subsets of different quality. The Spearman correlations were obtained from all mRNA–protein expression level pairs of the same gene and tissue for each subset. The mRNA subsets on the x-axis are defined by the combined mRNA quality measure mean + entropy (combined rank of mean expression and entropy), which gave the largest improvement in correlation. For protein quality, HPA antibody validation score was used as quality measure, shown by increasing line thickness.

This was used to obtain percentile scores for all genes, i.e. the percentage of the false pairs with a correlation lower than the true pair. While one would expect the correlation of the true pair to be higher than all false pairs, this may not happen due to either noise or other strongly coexpressed genes, e.g. complex members, that also have high correlations. Since it is difficult to estimate what would be expected by chance, picking a cutoff for significance is not possible. However, conclusions can still be drawn from the distribution of percentile scores.

Fig. 3B shows the percentile distribution for the highest quality subset based on mean + entropy and HPA antibody validation quality measures. The curve has a sharp peak at 90% with a steep decline toward 100%, and surprisingly no true mRNA–protein pairs were ranked highest. The fact that most of the density is below the 95th percentile suggests that there is a high degree of noise, causing false mRNA–protein pairs to often be ranked higher than the true pairs. With this type of analysis it would however be possible to extract genes that have the highest correlation relative to what is expected by chance.

3.5. Global correlation of coexpression

An alternative way to analyze the concordance between mRNA and protein expression is to look at the correlation of coexpression, or second-order correlation of expression (Dutilh et al., 2006; Shankavaram et al., 2007; Zhou et al., 2005). This can be seen as an

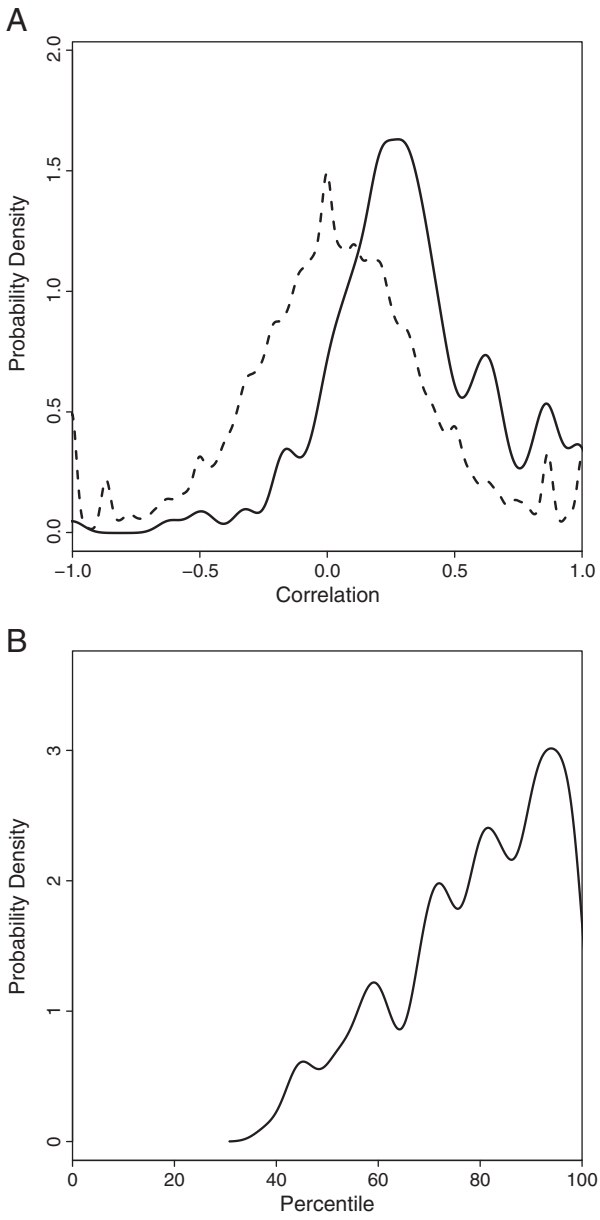

Fig. 3. A. Distribution of genewise mRNA–protein expression correlations for the subset with the highest mRNA and protein quality in Fig. 2. Distribution for true pairs of mRNA and protein corresponding to the same gene is shown with a solid line. Distribution for false pairs of mRNA and protein corresponding to different genes is shown with a dashed line. Other subsets have a similar distribution shape but are shifted to the left. B. How often does the true mRNA–protein pair have a higher correlation than false pairs with the same mRNA or protein? Shown is the distribution of percentile scores for true mRNA–protein pairs, i.e. each true pair's correlation's rank as a percentage of all false pair ranks. If the true pair is ranked highest then the percentile score would be 100. The plot is for the subset with the highest mRNA and protein quality in Fig. 2. Other subsets have a similar distribution shape but are shifted to the left.

orthogonal approach to direct expression correlation, as it first analyzes coexpression of pairs within the mRNA and protein data separately, and then analyzes the pair–pair correlations between the datasets. Coexpression correlation has some advantages. Because the mapping between mRNA and protein is done by pairs, there is no need to restrict the tissues used to the intersecting subset. In fact, the approach would work even with zero overlap between tissues sampled for mRNA and protein.

After verifying that coexpression from each dataset carries functional information (see Supplementary materials), the global correlation of coexpression was calculated for the same subsets as for expression correlation (Fig. 4) and compared to randomly sampling the data (Table S8). As expected, the trend of increasing correlations for higher quality subsets was observed here as well. The correlations were however higher and more significant for most of the higher quality subsets. The highest Spearman correlation was 0.32 ($p<2e\text{-}16$), substantially higher than the correlation of 0.06 obtained using the entirety of the data. We redid the analysis for mapped tissues only which gave lower correlations, maximally 0.22 ($p<2e\text{-}16$) (Table S4). Redoing it for only nonmapped tissues gave similar correlations as with mapped tissues only. This suggests that the correlation is independent of whether the tissues intersect or not, but depends on the total amount of data. On a global scale, mRNA–protein coexpression appears to be at the same level as expression.

3.6. Coexpression limited to functionally coupled genes

Coexpression correlation using all possible pairs suffers from the problem that most genes are not functionally related to each other and are therefore not coexpressed. All functionally unrelated pairs introduce noise to the global correlation. The simple approach of restricting the between-set comparison to pairs with a high coexpression in either mRNA or protein would likely result in an artificially high correlation of coexpression, hence it is not suitable. However, restricting the analysis to only consider "true" pairwise coexpression correlations, e.g. supported by a functional coupling, would be a reasonable way to reduce the noise. We redid the analysis restricting the coexpression correlations to those pairs supported by a link in FunCoup (Alexeyenko and Sonnhammer, 2009) with confidence above 0.25. Due to the sparsity of the network the data was severely restricted, resulting in correlations that were only significant in the highest quality bins (Table S6). The global correlation for the best subset obtained in this manner was 0.49, substantially higher than when considering all pairs, thus confirming that noise from non-related pairs can reduce the correlation.

3.7. Pairwise correlation of coexpression

Pairwise correlation of coexpression can be calculated in a similar way as genewise correlation of expression. This way we can analyze the distribution of correlations across all pairs. It was not possible to restrict this analysis to functionally coupled pairs due to the sparsity of the network because this resulted in too few data points per

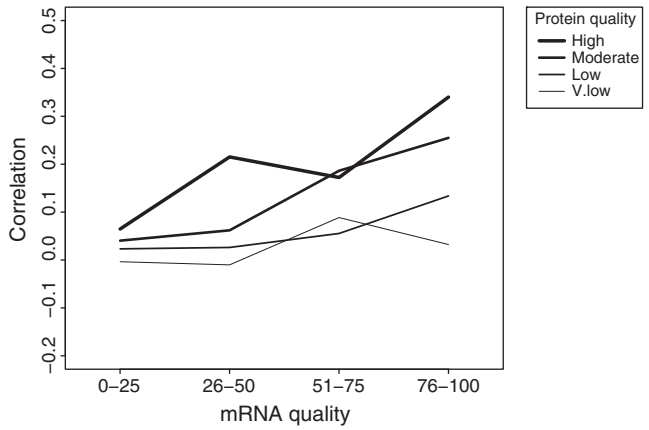

Fig. 4. Global correlation of mRNA and protein coexpressions, i.e. second order expression correlation, for the same 16 quality subsets as in Fig. 2. The correlations were obtained by first calculating coexpression of pairs within the mRNA and protein datasets separately. The mRNA–protein correlation was then obtained as the global correlation of pair–pair correlations for all mRNAs and proteins in the subset.

gene. Fig. 5A shows the correlation distribution for the subset with the highest quality in both mRNA and protein expression. The average was 0.28, slightly lower than the average genewise correlation, probably due to the lack of very high correlations found in the genewise analysis.

To assess how unexpected the observed coexpression correlations are, we calculated percentile scores in the same manner as for expression, i.e. each true pair's correlation was expressed as the percentage of false pairs ranking below it. Compared to the genewise analysis, the pairwise correlation of coexpression was strongly shifted toward higher percentile scores in the higher quality subsets. For the highest quality subset (Fig. 5B), almost all pairs are above the 95th percentile. This suggests that although the absolute average pairwise correlations are somewhat lower than the genewise, they are considerably more robust in the sense that they are much less expected by chance.

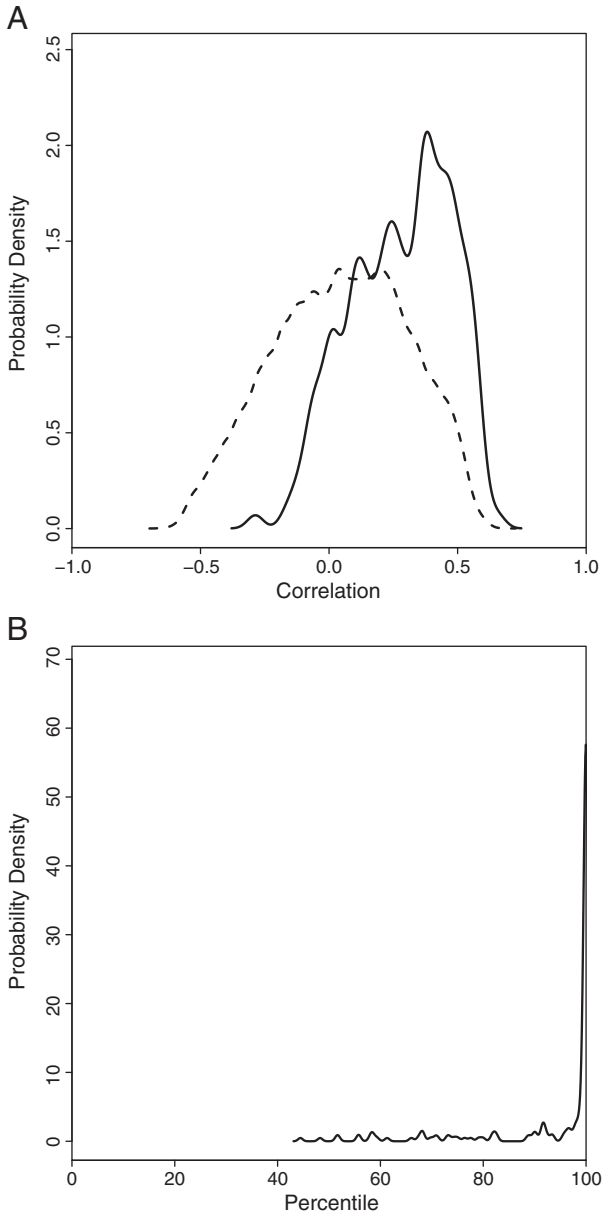

Fig. 5. A. Distribution of pairwise (mRNA–mRNA)–(protein–protein) coexpression correlations for the subset with the highest mRNA and protein quality in Fig. 4. Distribution for true pairs of mRNA and protein corresponding to the same gene is shown with a solid line. Distribution for false pairs of mRNA and protein corresponding to different genes is shown with a dashed line. Other subsets have a similar distribution shape but are shifted to the left. B. How often does the true (mRNA–mRNA)–(protein–protein) pair have a higher correlation than false pairs with the same mRNA–mRNA or protein–protein pair? Shown is the distribution of percentile scores for true pairs, i.e. each true pair's correlation's rank as a percentage of all false pair ranks. If the true pair is ranked highest then the percentile score would be 100. The plot is for the subset with the highest mRNA and protein quality in Fig. 4. Other subsets have a similar distribution shape but are shifted to the left.

4. Discussion

We have explored the distributions of mRNA–protein correlation for expression and coexpression using two large-scale tissue-based datasets. While most cases have a low correlation, the variation is high and some do have a high correlation. To exploit this we devised and benchmarked a number of quality measures that can define subsets of highly correlating genes/proteins and pairs. The most useful quality measures were based on either the magnitude or the variation of expression.

There are both technical and biological reasons why one would expect such quality measures to correctly identify subsets where coherence between mRNA and protein is greater. Such profiles would be less sensitive to noise. Additionally, if the mRNA level is generally high and varies greatly, this could indicate that a lot of energy is spent on regulating mRNA levels, which only makes sense if there is a direct connection between mRNA level and protein abundance.

The primary reason for low quality data for both mRNA and protein are noise levels that interfere with accurate estimation of low expression levels (Shankavaram et al., 2007). This includes measurement errors, different degrees of translational and post-translational regulation of protein abundance, sample collection issues, biological variation and systematic biases. In our study, using the semi-quantitative HPA protein expression measurements provided much higher coverage of genes than other resources, but the fact that it only provides four discrete expression levels can be a drawback for correlation analyses. As long as the four values are more or less equally represented (which is the case in HPA) this is not a problem, but if nearly all points were assigned to one value, then even the Spearman correlation would become unreliable.

Correlation of coexpression was found to be more robust than the correlation of expression in the sense that the coexpression pairs were much more likely to be true than for expression. Still, the global correlation was not higher for coexpression. The reason for this is simply that the global correlation is mostly based on mRNA–mRNA and protein–protein pairs that are not functionally related, and therefore have a low correlation. We managed to alleviate this effect by restricting the analysis to functionally coupled pairs but this approach is limited by the amount of known and predicted protein interactions.

Coexpression carries an advantage in that mRNA and protein can be compared with less regard to overlap of conditions between datasets. Also, it is likely less sensitive to systematic differences between samples and measurements. Even with systematic errors in one dataset, coexpression should be detectable for gene pairs with a strong functional coupling. Unless systematic errors are the same for both protein and mRNA, the errors will have a larger effect on expression concordance.

The technical limitations on the measurements could dampen the ability to detect concordance. The measurements are performed on the level of tissues, averaged over multiple cell types. This may dilute a strong signal stemming from a single distinct cell type in an organ made up of many cell types. In fact, anything larger than single cell measurement enforces an averaging and signal dilution. Also, splice forms are not accounted for in the datasets which could result in over- or underestimation of the true expression levels.

Entropy is by itself not an ideal quality measure. The highest (negative) entropy (0) is measured for genes expressed only in one tissue. Such expression profiles would be sensitive to show false correlation to non-related genes with high expression in that tissue. This exemplifies why it is advantageous to examine percentile scores instead

of the raw correlations. However, by combining the mean and entropy quality measures this can largely be avoided because the mean expression counteracts the entropy's preference of fewer tissues with expression.

While external validation would be desirable to show generalizability of the presented approach, the lack of other large-scale datasets with mRNA or protein expression across multiple tissues renders this impossible. One can however generate subsets using multiple-tissue data and apply the sets to independent single-tissue mRNA and protein data. We applied the subsets defined by the datasets in this study to single-tissue data from Schwanhäusser et al. (2011). The lower quality sets showed significantly lower mRNA–protein correlations than expected (see Supplementary materials), showing that the quality subsets can be used to avoid low coherence genes. This serves as a partial confirmation that the sets are generalizable for global expression correlation. Unfortunately single-tissue data cannot give information about genewise or coexpression correlations.

There are several problems when comparing multiple-tissue with single-tissue mRNA–protein correlations. Even if a gene across many tissues generally has a high mRNA–protein concordance, one would not necessarily expect a high concordance in every single tissue. In any given tissue, the gene might not be expressed, be expressed in only in a subset of cells or have post-transcriptional regulation specific to that tissue. Additionally, properties such as tissue-specific translation efficiency can strongly affect the correlation within one tissue but would have less impact across multiple tissues. Thus while we would expect the high-quality sets to be more reliable in general, it would not necessarily hold true for each single tissue.

Previous work has shown that protein abundance is only partly regulated through mRNA abundance. Therefore one should always be cautious when using mRNA as a proxy for protein abundance. We have here shown that it is possible to identify subsets with greater coherence between mRNA and protein abundance by restricting the analysis to genes with a high quality score. This can help reduce the danger of drawing erroneous conclusions for genes with low mRNA–protein concordance.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.gene.2012.01.029.

References

Alexeyenko, A., Sonnhammer, E.L., 2009. Global networks of functional coupling in eukaryotes from comprehensive data integration. Genome Res. 19 (6), 1107–1116.

Berglund, L., et al., 2008. A genecentric Human Protein Atlas for expression profiles based on antibodies. Mol. Cell. Proteomics 7 (10), 2019–2027.

Brockmann, R., Beyer, A., Heinisch, J.J., Wilhelm, T., 2007. Posttranscriptional expression regulation: what determines translation rates? PLoS Comput. Biol. 3 (3), e57.

Daub, C.O., Sonnhammer, E.L., 2008. Employing conservation of co-expression to improve functional inference. BMC Syst. Biol. 2, 81.

de Sousa Abreu, R., Penalva, L.O., Marcotte, E.M., Vogel, C., 2009. Global signatures of protein and mRNA expression levels. Mol. Biosyst. 5 (12), 1512–1526.

Dutilh, B.E., Huynen, M.A., Snel, B., 2006. A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation. BMC Genomics 7, 10.

Ghaemmaghami, S., et al., 2003. Global analysis of protein expression in yeast. Nature 425 (6959), 737–741.

Greenbaum, D., Colangelo, C., Williams, K., Gerstein, M., 2003. Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol. 4 (9), 117.

Huttenhower, C., et al., 2009. Exploring the human genome with functional maps. Genome Res. 19 (6), 1093–1106.

Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. J. Comput. Graph. Stat. 5 (3), 299.

Le Roch, K.G., et al., 2004. Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. Genome Res. 14 (11), 2308–2318.

Lu, P., Vogel, C., Wang, R., Yao, X., Marcotte, E.M., 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat. Biotechnol. 25 (1), 117–124.

Nie, L., Wu, G., Zhang, W., 2006. Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in *Desulfovibrio vulgaris*: a quantitative analysis. Genetics 174 (4), 2229–2243.

Nie, L., Wu, G., Culley, D.E., Scholten, J.C., Zhang, W., 2007. Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications. Crit. Rev. Biotechnol. 27 (2), 63–75.

Schwanhäusser, B., et al., 2011. Global quantification of mammalian gene expression control. Nature 473 (7347), 337–342.

Shankavaram, U.T., et al., 2007. Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. Mol. Cancer Ther. 6 (3), 820–832.

Su, A.I., et al., 2002. Large-scale analysis of the human and mouse transcriptomes. Proc. Natl. Acad. Sci. U. S. A. 99 (7), 4465–4470.

Tuller, T., Kupiec, M., Ruppin, E., 2007. Determinants of protein abundance and translation efficiency in *S. cerevisiae*. PLoS Comput. Biol. 3 (12), e248.

Uhlen, M., Ponten, F., 2005. Antibody-based proteomics for human tissue profiling. Mol. Cell. Proteomics 4 (4), 384–393.

Uhlen, M., et al., 2005. A human protein atlas for normal and cancer tissues based on antibody proteomics. Mol. Cell. Proteomics 4 (12), 1920–1932.

van Noort, V., Snel, B., Huynen, M.A., 2003. Predicting gene function by conserved co-expression. Trends Genet 19 (5), 238–242.

Vogel, C., et al., 2010. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol. Syst. Biol. 6, 400.

Wu, G., Nie, L., Zhang, W., 2008. Integrative analyses of posttranscriptional regulation in the yeast *Saccharomyces cerevisiae* using transcriptomic and proteomic data. Curr. Microbiol. 57 (1), 18–22.

Zhou, X.J., et al., 2005. Functional annotation and network reconstruction through cross-platform integration of microarray data. Nat. Biotechnol. 23 (2), 238–243.