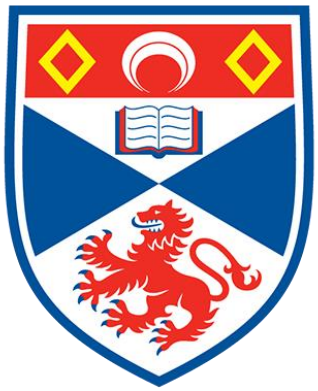
Does differential mRNA expression predict better protein-mRNA correlation? A meta-analysis

Lillian Bates
lgb9@st-andrews.ac.uk



University of
St Andrews

Supervised by Dr. V. Anne Smith

School of Biology, University of St Andrews

August 2023

Word Count (excluding title page, in-text citations, figure legends, table legends, acknowledgements, and references): 2998/3000

Completed as part of the Laidlaw Leadership and Research Scholarship Programme

Introduction

The central dogma of molecular biology describes the flow of genetic information from the DNA genome to messenger RNA (mRNA) to protein (Crick, 1958). mRNA is the bridge between the genetic information encoded in DNA and the characteristics expressed in protein. The genetic information in DNA is copied into mRNA through a process called transcription. The information in mRNA is translated into a sequence of amino acids, creating proteins. At each step throughout the central dogma, several layers of control and regulation allow cells to adjust and respond to changes in their internal and external environments by expressing different genes and producing different proteins (Erdmann et al., 2018).

A gene is expressed when the information in the DNA genome is transferred into mRNA, which can then be translated into a protein. The protein can then perform its job within the biological system. Cells can increase (up-regulate) or decrease (down-regulate) a gene's expression level to address the needs of the cell under different conditions called differential expression (DE). A differentially expressed gene (DEG) is expressed at different levels under different conditions. Genes that are not regulated differently under those different conditions are non-differentially expressed genes (NDEGs).

There is a common assumption in molecular biology that measurements of mRNA levels reflect protein expression levels and that changes in mRNA expression result in parallel changes in protein expression (high protein-mRNA correlation) (Koussounadis et al., 2015). mRNA expression is often used to infer protein expression levels because mRNA is generally easier to measure than protein. Logically, this assumption makes sense; otherwise, it begs the question of why regulation of mRNA expression is needed in the cell. However, a poor correlation

between mRNA and protein has been found (Maier et al., 2009; Vogel & Marcotte, 2012), challenging the use of mRNA expression as an indirect measure of protein expression (Koussounadis et al., 2015). Research into protein expression has many important applications in medicine, drug discovery, and disease research, so it is important to understand when or if this assumption can reliably be used. Two past studies using an ovarian cancer xenograft model (Koussounadis et al., 2015) and in *Pseudomonas aeruginosa* (Erdmann et al., 2018) have found that genes that are differentially expressed on the mRNA level have better mRNA-protein correlations.

This project investigated the relationship between differential expression and mRNA-protein correlation using a meta-analysis of existing studies in various systems. We hypothesised that DEGs would have tighter mRNA-protein correlations. A meta-analysis is a statistical study that synthesises independent data from previously published work to derive more general insights into a question for which conflicting results may exist in the literature. Here, we analysed seven independent previously published datasets with mRNA and protein expression data. First, each gene in each dataset was classified as a DEG or NDEG using the mRNA data. Then, the difference between the mRNA-protein correlations for DEGs and NDEGs was assessed for each study. Finally, the results of the individual analyses were pooled through meta-analysis to obtain a combined result. Overall, the meta-analysis of seven studies suggested tighter mRNA-protein correlations for DEGs than for NDEGs.

Procedures and Results

The data analysis was conducted using R programming language (R Core Team, 2023) in RStudio (Posit Team, 2023). This computational project was divided into six main stages: 1) identification of relevant datasets for analysis, 2) importing data into R, 3) identification of differentially expressed genes, 4) mRNA-protein correlation, 5) significance testing, and 6) meta-analysis (Figure 1). The analysis was completed on each dataset individually, except in stage 6.

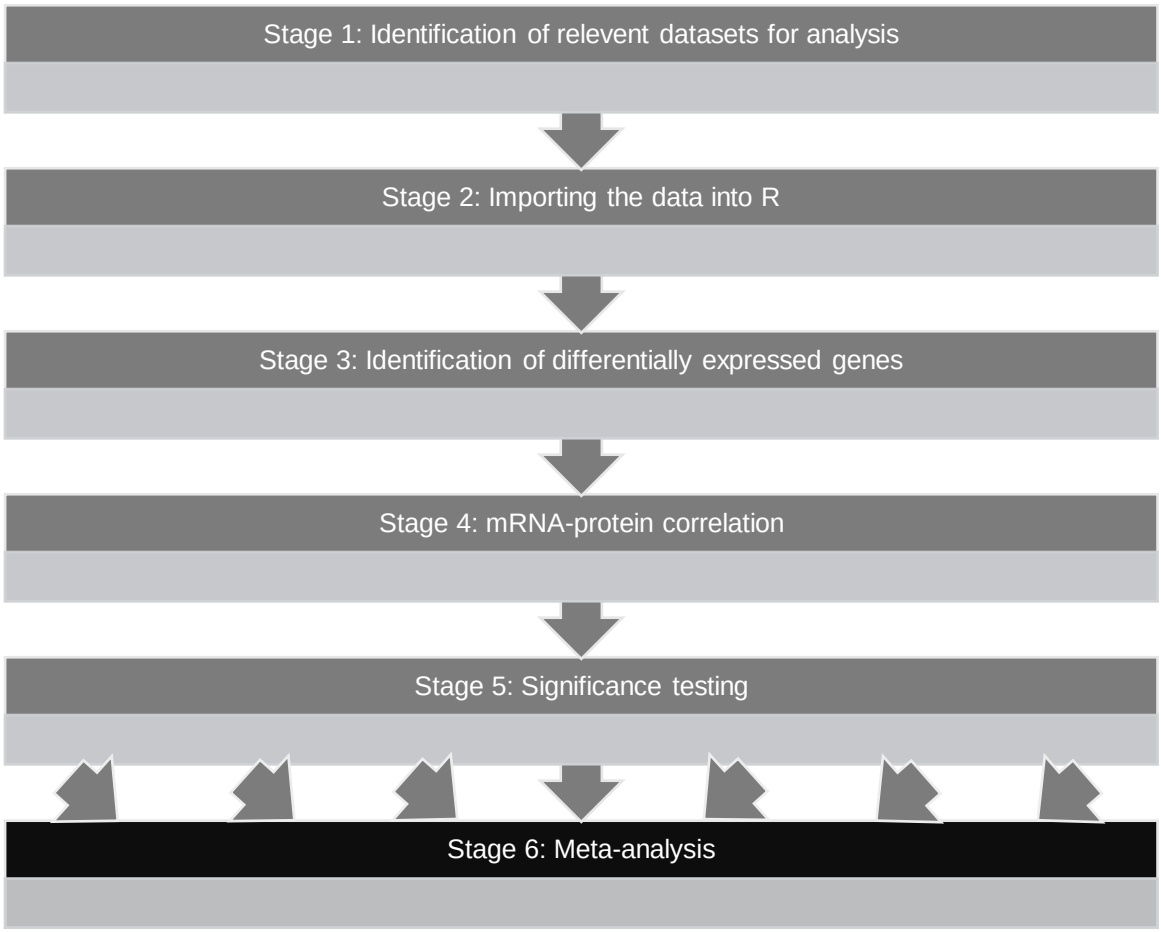


Figure 1. Sequence of project stages. Each box represents a stage of analysis in the project and shows the project method structure. Each stage included several analytical steps. Stages 1-5 were done individually in each dataset. The results of stage 5 for each dataset were pooled into stage 6 to obtain the final results.

Stage 1: Identification of Relevant Datasets for Analysis

Eligible datasets were found by searching the scientific literature in the Web of Science database (Table 1). Here, “dataset” describes the mRNA and protein data extracted from a published study. Eligible datasets needed to measure both mRNA and protein expression of single genes under at least two experimental conditions. For each condition, at least two biological replicates were required (Ritchie et al., 2015; Phipson et al., 2016). The study also needed to measure mRNA and protein expression for ≥ 10 genes, so the gene sample size was large enough to compare the mRNA-protein correlations between the DEGs and NDEGs. Finally, the mRNA and protein measurements needed to be from the same biological replicates. If all these requirements were met, and the data were available in an accessible format, the dataset was eligible to be included.

Four datasets were found during previous proof-of-concept work (Cheng et al., 2016; Darmanis et al., 2016; Genshaft et al., 2016; Lee et al., 2011). To find additional datasets, three database searches were made (Table 2), identifying three datasets for analysis from > 1100 potentially relevant studies (Bai et al., 2021; Caglar et al., 2017; Reimegård et al., 2021).

Table 1. Datasets used in the meta-analysis. *This is the number of genes included in our analysis, which was treated as the sample size for each dataset. This number does not necessarily correspond with the total number of genes measured in the original study (see Stages 2 and 3).

Dataset	Cell/Tissue	Treatment/Condition	mRNA measurement technique	Protein measurement technique	Number of Genes Analysed*
Reimegård et al., 2021	*Homo sapiens* (embryonic stem cell line HS181)	Neural Induction by neural induction medium (NIM)	scRNA-Seq	PEA Assay	78
Lee et al., 2011	*Saccharomyces cerevisiae* (BY4741 Strain)	0.7 M NaCl	Microarray	Mass Spectrometry (MS)	1321
Genshaft et al., 2016	*Homo sapiens* (Human Breast Adenocarcinoma cell line)	1 μM phorbol-12-myristate-13-acetate (PMA)	qRT-PCR	PEA Assay	27
Darmanis et al., 2016	*Homo sapiens* (U3035MG cell line from tumour tissue of patient with grade IV glioblastoma)	10 ng/μl bone morphogenetic protein 4 (BMP4)	qRT-PCR	PEA Assay	24
Cheng et al., 2016	*Homo sapiens* (HeLa cell line)	2.5 mM dithiothreitol (DTT)	Microarray	MS	1237
Caglar et al., 2017	*Escherichia coli* (REL606)	Different growth mediums: Carbon (Glucose, Glycerol, Gluconate, Lactate), Magnesium (Low Mg, Base Mg, High Mg), Sodium (Base Na, High Na)	RNA-Seq	MS	3841
Bai et al., 2021	Maturing *Arabidopsis thaliana* seeds	Seed germination	Microarray	MS	1453

Table 2. Web of Science database search results.

Search Term	# of Results	Date Searched	Datasets found from Search
Single gene expression mRNA protein abundance	722	30 May 2023	Reimegård *et al.,* 2021 Caglar *et al.,* 2017 Bai *et al.,* 2021
Single cell mRNA protein abundance; 2005 to present	412	13 June 2023	Reimegård *et al.,* 2021 Caglar *et al.,* 2017
Multiplexed measurement of mRNA and protein single cell	11	9 June 2023	No datasets found

Stage 2: Importing the Data into R

Each dataset was imported into the R program as data frame coding objects (essentially, tables within R). Once the data were imported as data frames, genes with any missing data were removed. The data frames were restructured to make further analysis more straightforward.

Stage 3: Identification of Differentially Expressed Genes

Each gene in a dataset was classified as either differentially expressed or non-differentially expressed by comparing the change in mRNA expression between a control and experimental condition, resulting in lists of DEGs and NDEGs for each dataset. There were many software options for this analysis and no standard statistical model (Stupnikov et al., 2021). The appropriateness of the software depends on the mRNA measurement technique. After evaluating several options, it was decided that for microarray and RT-PCR, R package limma would be used (Ritchie et al., 2015; Phipson et al., 2016), whereas NOISeq (Tarazona et al., 2015; Tarazona et al., 2011) would be used for RNA-Seq data (Stupnikov et al., 2021).

Data Filtering and Transformation

All the datasets provided data that were already normalised. However, we completed some additional treatments for our analysis. Firstly, a method was developed to filter some genes that did not have enough data for reliable analysis. Many of the genes had expression values of zero in mRNA and protein, which could represent true zeros, but often means the amount of the mRNA or protein present was below the instrument's detection level. Gene measurements composed of

≥ 95% zero data points for both mRNA and protein measurements were removed. However, genes with mRNA measurements ≥ 95% zero but protein measurements < 95% zero (or vice versa) were not removed because these could represent interesting pieces of data. Finally, the logarithm base 2 of each data point was taken, called $\log_2$ transformation. This common transformation for gene expression data makes the data more comparable (Steinhoff & Vingron, 2006).

Choosing Contrast

The next step was to choose the control and experimental trials to compare, called a contrast. For example, Genshaft et al. (2016) measured mRNA and protein abundance at 0 hours (control) before phorbol-12-myistate-13-acetate (PMA) treatment and measured expression again at 24 hours and 48 hours post-PMA treatment, leaving two possible comparisons: control versus 24 hours post-PMA or control versus 48 hours post-PMA. All datasets had multiple possible contrasts, but only one was used to classify DEGs and NDEGs. To choose the contrast for each dataset, a procedure was developed to compare mRNA expression data from the possible contrasts quantitatively.

The DGE software can classify DEGs using multiple contrasts. However, we decided to use only one contrast per dataset to minimise incorporating more variation into the mRNA-protein correlation analysis and meta-analysis later. This decision did introduce a limitation, as not all DEGs would be found, which unfortunately meant that much data in the datasets were not analysed. We were interested in the effect of differential expression on mRNA-protein correlation, less in finding all DEGs from the given treatment, so this trade-off was made to improve the standardisation of analysis between datasets.

Computing the DEG Lists

Each gene in each dataset was classified by the appropriate software (limma or NOISeq) as a DEG or a NDEG by comparing the control and experimental mRNA data from the chosen contrast (Figure 2A). DEGs were those with a statistically significant difference between the control and experimental mRNA expression, meaning the change in mRNA expression was most likely due to the treatment, not random chance. mRNA expression of NDEGs did not significantly differ between control and experimental. DEGs can be either up-regulated (increase in expression) or down-regulated (decrease in expression) when compared to the control (Figure 2B). Although the current project was not specifically interested in the regulation direction of DEGs, the number of up- or down-regulated genes was still noted.

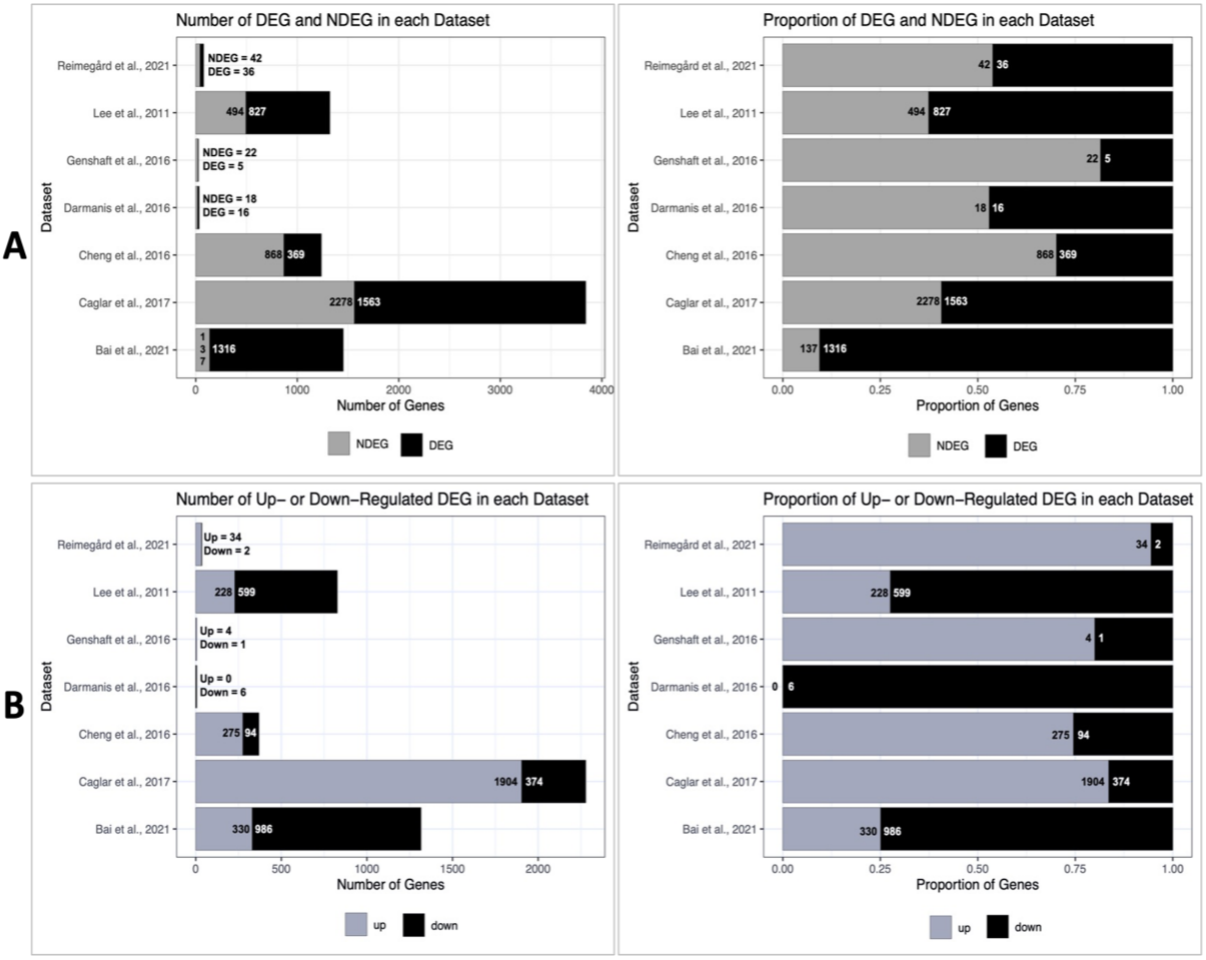


Figure 2. Results of DGE analysis for each dataset. For all the figures, black or white text on or adjacent to the bars states the number of genes the bar represents. Figure 2A (top) shows the breakdown of DEG (black bars) and NDEG (grey bars) for each dataset. 2A (left) compares the breakdown of DEGs/NDEGs scaled by the number of genes. 2A (right) compares the ratio of DEG (black) to NDEG (grey). Figure 2B (bottom) compares the breakdown of up- (grey) and down-regulated (black) DEG for each dataset. 2B (left) compares the breakdown of up-/down-regulated genes scaled by the number of genes. 2B (right) compares the ratio of up-regulated DEG (grey) to down-regulated DEG (black). Plots were made in R using ggplot2 (Wickham, 2016).

Stage 4: mRNA-Protein Correlation & Stage 5: Significance Tests

The objective of stage 4 was to calculate the mRNA-protein correlation for each gene. Then, in stage 5, the results of stages 3 and 4 were compared to assess the effect of differential expression on the mRNA-protein correlations for each dataset and, specifically, test our hypothesis that DEGs would have better mRNA-protein correlations.

Correlation

The mRNA-protein correlations for each gene were calculated using mRNA and protein data from the trials compared to find the DEGs. The result was a single correlation coefficient value per gene. For this process, we used Spearman's rank correlation (Spearman, 1904), a non-parametric test, meaning it does not assume the data were normally distributed (bell-shaped curve). The correlation coefficient represents how linked two variables are. Spearman correlation coefficients can be between +1 (positive correlation) and -1 (negative correlation). In this case, coefficients closer to +1 suggested a positive mRNA-protein correlation for that gene. Coefficients closer to -1 indicated a negative correlation between mRNA and protein expression for that gene (e.g., higher mRNA expression associated with lower protein expression and vice versa). Coefficients closer to zero suggested no correlation, positive or negative. We hypothesised that DEGs would have more positive mRNA-protein correlation coefficients.

For some genes, all of the replicates for mRNA expression or protein expression were measured as zero. Spearman's rank correlation coefficient cannot be calculated if all the data points for mRNA or protein were identical, such as all zero. Hence, the coefficient for these genes was NA (missing data) (Smirnov, 1948). The correlation coefficients were needed for stage 5, and the software would ignore coefficients of NA, effectively excluding these genes from the final calculation. However, we did not want to exclude these genes because they could represent interesting pieces of data. To avoid excluding these genes, we replaced all data points measured as zero with random small numbers close to zero generated between 1.0×10^{-400} and 1.0×10^{-300} because zeros can often be interpreted as an amount below the instrument's detection limit, not necessarily precisely 0.

Correlation coefficients were calculated 100 times for each zero-containing gene, using newly generated random numbers each time to minimise the impact of the random number replacements. Then, the average coefficient for all 100 repeats was used as the overall correlation coefficient for the gene. Comparing the correlations between the trials without zero replaced (those that were not NA) did not change much. In the future, other methods to address this issue could be developed.

The next step was to compare the mRNA-protein correlation coefficients of DEGs and NDEGs in each dataset, which can be visualised by plotting the median mRNA-protein correlation for DEGs and NDEGs for each dataset (Figure 3). In addition to visualising any differences, their significance was quantified using statistical tests in stage 5.

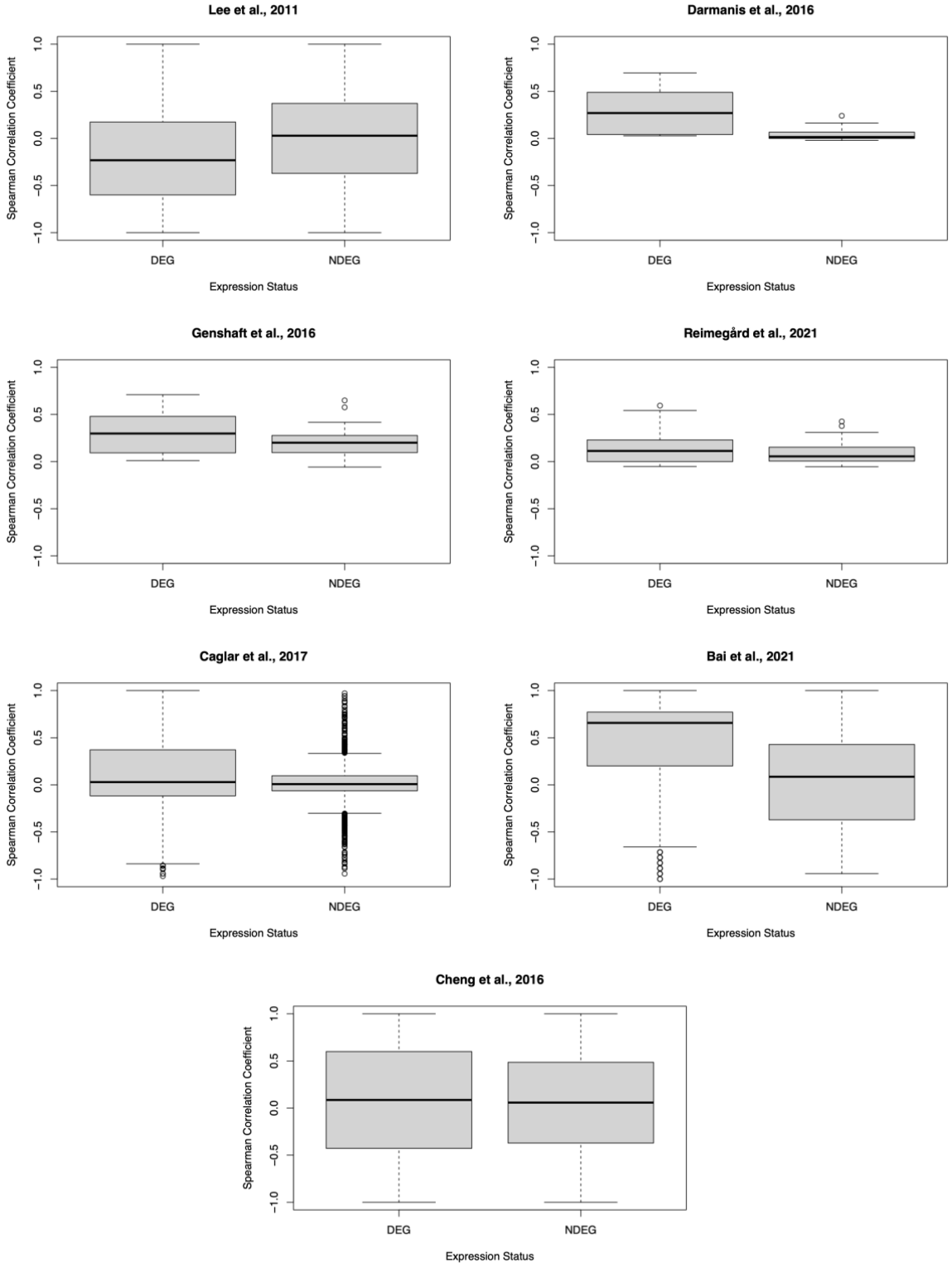


Figure 3. mRNA-protein correlation of DEGs versus NDEGs for each dataset. Boxplots show the median Spearman mRNA and protein correlation coefficient for each dataset for DEG groups versus the NEG groups. These plots show the differences in the correlations between these groups. Values closer to 1.0 are more tightly correlated. Values closer to 0 are less correlated, and values closer to -1 are more negatively correlated.

Significance Tests

The significance tests aimed to evaluate our hypothesis if DEGs had significantly higher mRNA-protein correlation coefficients. Two non-parametric tests were used: the Kolmogorov–Smirnov test (KS Test) (Smirnov, 1948) and a Brunner-Munzel (BM Test) (Brunner & Munzel, 2000). The result of each test was a *P*-value, which quantified how likely it was that the observed correlation could be attributed to random chance. For this project, *P*-values less than the threshold of 0.05 were considered statistically significant. A *P*-value of 0.05 meant a 95% chance of an actual increase in mRNA-protein correlation for DEGs than NDEGs not caused by random chance. As such, a study with a *P*-value < 0.05 suggested that DEGs had significantly higher mRNA-protein correlations than NDEGs in support of our hypothesis. Two *P*-values for each study were calculated, one for each test (Table 3).

For the KS-Test, Caglar et al. (2017) ($P < 2.2x10^{-16}$) and Bai et al. (2021) ($P < 2.2x10^{-16}$) returned significant values. The test did not return exact values because they were below the level that the test can calculate precise *P*-values, suggesting a very significant result. The other studies did not have significant *P*-values, suggesting there was not a significant link between DEGs and a higher mRNA-protein correlation. For the BM-Test, the same studies showed significant and non-significant results (Table 3), except for Darmanis et al. (2016), which was significant for the BM-Test (*P*-value = $6.82x10^{-3}$) but was not significant for the KS-Test (*P*-value = 0.051).

Interestingly, Lee et al. (2011) had a *P*-values close to 1 for both tests (KS-Test *P* = 0.904; BM-Test *P* = 0.999999998), suggesting that NDEGs had higher mRNA-protein correlations, the opposite of the hypothesis of this project. Lee et al. (2011) had some unique properties compared to the other datasets that could have

contributed, such as cell type and the lack of a proper control trial. Whether this result reflects a true biological phenomenon or can be explained by technical considerations would be interesting to examine more in the future.

Table 3. **Results of KS-Tests and BM-Tests** of the relationship between DGE and mRNA-protein correlation. Significant *P*-values were < 0.05. A significant *P*-value suggests DEGs had better mRNA-protein correlations.

Dataset	**KS-Test *P*-value**	**KS-Test Significance**	**BM-Test *P*-value**	**BM-Test Significance**
Reimegård et al., 2021	0.126	Not Significant	0.183	Not Significant
Lee et al., 2011	0.904	Not Significant	0.999999998	Not Significant
Genshaft et al., 2016	0.251	Not Significant	0.296	Not Significant
Darmanis et al., 2016	0.051	Not Significant	6.83×10^{-3}	Significant
Cheng et al., 2016	0.167	Not Significant	0.355	Not Significant
Caglar et al., 2017	$< 2.2 \times 10^{-16}$	Significant	1.27×10^{-7}	Significant
Bai et al., 2021	$< 2.2 \times 10^{-16}$	Significant	7.99×10^{-29}	Significant

Stage 6: Meta-Analysis

The results of these tests varied greatly in significance across the studies, providing conflicting information about the relationship between mRNA-protein correlation and differential expression. A meta-analysis was used to find the consensus of these tests by pooling the individual results using a method called *P*-value combination. This method takes *P*-values from statistical tests asking the same question and combines them to generate a new *P*-value, assessing the tested phenomenon across all the studies.

All of the *P*-values from each test were combined using Stouffer's *P*-value combination method (Stouffer et al.,1949) to find the consensus about the relationship between DEG and mRNA-protein correlation (Heard & Rubin-Delanchy, 2018; Rice, 1990; Dewey, 2022). As mentioned, Caglar et al. (2017) and Bai et al.

(2021) returned significant inexact *P*-values < $2.2x10^{-16}$ for the KS-Tests, so there were no values to input for the KS combination. To avoid their exclusion from the *P*-value combination, $2.2x10^{-16}$ was input for these into the meta-analysis. The final combined *P*-value of the KS-Tests was $1.23x10^{-13}$, much less than 0.05, representing a significant result. This result supports the hypothesis that DEGs have a tighter correlation between mRNA and protein expression. Replacing the inexact *P*-values meant that the combined KS *P*-value was likely not exact. However, if we had the precise *P*-values, it is unlikely the combined *P*-value would have been above the 0.05 threshold. The final combined *P*-value of the BM-Test (which did not require any *P*-values to be replaced) was $1.49x10^{-8}$, which again provides support for the hypothesis.

Conclusions and Future Work

Conclusions

Overall, the sequential analysis of these datasets demonstrates statistically significant higher correlations between mRNA and protein expression for DEGs than NDEGs. Similar results have been found in previous studies of only one system (Erdmann et al., 2018; Koussounadis et al., 2015). It is often assumed that a system's mRNA level can inform the amount of protein in the system based on the central dogma of molecular biology. This assumption is useful in research because mRNA is easier to measure than protein. However, it has been shown that mRNA and protein have a poor correlation on a genome-wide level, suggesting this assumption is invalid (Maier et al., 2009; Vogel & Marcotte, 2012). This calls into question some of the fundamentals of the central dogma of molecular biology. If mRNA levels do not in some way correlate to protein levels in a system, what is the

biological point of regulating mRNA expression from the DNA genome? From this question stems the hypothesis that genes differentially regulated in response to a change in environment (DEG) would have a tighter correlation between mRNA and protein than those not differentially regulated on the mRNA level.

Our findings support this hypothesis and suggest that regulating a gene at the level of mRNA expression has a downstream effect on the protein level, leading to a higher mRNA-protein correlation for DEGs. For NDEGs, other mechanisms may be in place to regulate protein abundance and function, so the amount of mRNA does not necessarily reflect the amount of protein. These additional regulatory mechanisms could also impact DEGs and lead to a lower correlation. However, the findings here support that DEGs generally have tighter correlations between mRNA and protein. Here, the goal was to assess the presence of this relationship, less about the effect size, which would be interesting to examine in the future.

Our application of meta-analysis aimed to provide a more general result, using datasets across many species, from mammals to bacteria, plants, and fungi. The majority of the datasets that we identified were from experiments using human cells. However, the human datasets were not individually significant, except for the BM-test on data from Darmanis et al. (2016). This result suggests that mRNA may not be a particularly good indicator of protein expression in mammalian systems. Further research into this question in mammal models could help address this point.

The results of this project can provide further guidance to researchers in how they use mRNA data to predict protein expression levels. Due to the presence of many post-transcriptional regulatory mechanisms, it is important to measure protein directly rather than relying on mRNA levels. However, if this is not possible, mRNA expression data of DEGs can more reliably be used than for NDEGs, where mRNA

data would not be an appropriate reflection of protein expression. Our findings have implications in many areas that measure mRNA and protein expression, such as disease research, drug discovery and development, and more. Assumptions are unavoidable in research, but it is important to minimise them and understand their impact. This research helps to provide more information about the use case of this common assumption. Future research into the specifics of the observed relationship between differential expression and mRNA-protein correlation would be beneficial in further defining the use of mRNA as a measure of protein expression and provide further insights into gene expression regulation on the mRNA and protein level.

Future Work

Our meta-analysis showed significant results with the seven datasets analysed. We have not found any reports in the literature addressing this question using a meta-analysis. The set of seven datasets used here is relatively small. As we completed our analysis presented in this essay, additional datasets have been identified for use in a future meta-analysis in preparation for publication.

Acknowledgements

I would like to thank my research advisor, Dr. V. Anne Smith, for her invaluable support, guidance, and mentorship on my project. I would also like to thank the Laidlaw Foundation for funding my project through my participation in the Laidlaw Scholars Leadership and Research Programme.

References

Bai, B. et al., 2021. Delayed Protein Changes During Seed Germination. *Frontiers in Plant Science,* Volume 12, pp. 1-11.

Brunner, E. & Munzel, U., 2000. The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation. *Biometrical Journal,* 42(1), pp. 17-25.

Caglar, M. U. et al., 2017. The E. coli molecular phenotype under different growth conditions. *Scientific Reports,* 7(1), p. 45303.

Cheng, Z. et al., 2016. Differential dynamics of the mammalian mRNA and protein expression response to misfolding stress. *Molecular Systems Biology,* 12(1), p. 855.

Darmanis, S. et al., 2016. Simultaneous Multiplexed Measurement of RNA and Proteins in Single Cells. *Cell Reports,* 14(2), pp. 380-389.

Dewey, M., 2022. *metap: meta-analysis of significance values.* s.l.:R package version 1.8.

Erdmann, J. et al., 2018. Environment-driven changes of mRNA and protein levels in Pseudomonas aeruginosa. *Environmental Microbiology,* 20(11), pp. 3952-3963.

Feng, C., Wang, H., Lu, N. & Tu, X. M., 2013. Log transformation: application and interpretation in biomedical research. *Statistics in Medicine,* 32(2), pp. 230-239.

Genshaft, A. S. et al., 2016. Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction. *Genome Biology,* 17(1), p. 188.

Heard, N. A. & Rubin-Delanchy, P., 2018. Choosing between methods of combining p-values. *Biometrika,* 105(1), pp. 239-246.

Hunt, G. P. et al., 2022. GEOexplorer: a webserver for gene expression analysis and visualisation. *Nucleic Acids Research,* 50(W1), pp. W367-W374.

Koussounadis, A. et al., 2015. Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. *Scientific Reports,* 5(1), p. 10775.

Lee, V. M. et al., 2011. A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Molecular Systems Biology,* 7(1), p. 514.

Maier, T., Güell, M. & Serrano, L., 2009. Correlation of mRNA and protein in complex biological samples. *FEBS Letters,* 583(24), pp. 3966-3973.

Phipson, B. et al., 2016. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *The Annals of Applied Statistics,* 10(2), p. 946.

Posit Team, 2023. *RStudio: Integrated Development Environment for R,* Boston: Posit Software, PBC.

R Core Team, 2023. *R: A Language and Environment for Statistical Computing,* Vienna: R Foundation for Statistical Computing.

Reimegård, J. et al., 2021. A combined approach for single-cell mRNA and intracellular protein expression analysis. *Communications Biology,* 4(1), p. 624.

Rice, W. R., 1990. A Consensus Combined P-Value Test and the Family-Wide Significance of Component Tests. *Biometrics,* 46(2), p. 303.

Ritchie, M. E. et al., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research,* 42(7), p. e47.

Spearman, C., 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology,* 15(1), p. 72.

Steinhoff, C. & Vingron, M., 2006. Normalization and quantification of differential expression in gene expression microarrays. *Briefings in Bioinformatics,* 7(2), pp. 166-177.

Stouffer, S. et al., 1949. *The American Soldier: Adjustment during Army Life.* Princeton(NJ): Princeton University Press.

Stupnikov, A. et al., 2021. Robustness of differential gene expression analysis of RNA-seq. *Computational and Structural Biotechnology Journal,* Volume 19, pp. 3470-3481.

Tarazona, S. et al., 2015. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research,* 43(21), p. e140.

Tarazona, S. et al., 2011. Differential expression in RNA-seq: A matter of depth. *Genome Research,* 21(12), pp. 2213-2223.

Vogel, C. & Marcotte, E. M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics,* 12(4), pp. 227-232.

Wickham, H., 2016. *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag.