



Gene expression and protein abundance: Just how associated are these molecular traits?

Anahita Samih^{a,b,1}, Maurício Alexander de Moura Ferreira^{a,b,1}, Zoran Nikoloski^{a,b,*}

^a Bioinformatics, Institute of Biochemistry and Biology, University of Potsdam, Potsdam 14476, Germany

^b Systems Biology and Mathematical Modelling, Max Planck Institute of Molecular Plant Physiology, Potsdam 14476, Germany

ARTICLE INFO

Keywords:

Correlation

Regression

Transcriptomics

Proteomics

ABSTRACT

The ability to accurately predict the abundance of proteins from the expression of the corresponding genes has enormous potential for the advancement of biotechnological applications using metabolic engineering and synthetic biology approaches. Addressing this problem has been challenging because of the lag in methodological advances in quantifying protein abundances. Here, we reviewed and classified studies that investigated the relationship between gene expression and protein abundance in different experimental settings and cellular contexts. We focused on comparing and contrasting the findings based on different correlation-based measures, widely used with nominal or transformed transcriptomics and proteomics data. We also included studies that investigated and attempted to explain the observed correlations between gene expression and protein abundance by incorporating data on additional factors, such as translation rate, protein degradation, and post-transcriptional modifications, using various statistical and mechanistic modelling frameworks. Finally, we provided an overview of how the latest advances using data from single-cell analyses have contributed to addressing this pressing question. Our review offers a perspective about how the findings about the relationship between gene expression and protein abundance can propel biotechnological advances, particularly focusing on opportunities resulting from the availability of protein-constrained metabolic models and the complementary machine and deep learning models.

1. Introduction

Many advances in biotechnology are driven by the ability to manipulate gene expression. This allows for engineering cellular processes, such as signalling and metabolic pathways, orchestrated by proteins of different molecular functions and whose abundance varies across cellular contexts and environments. Understanding how gene expression affects the distribution of protein abundances is not only relevant in biotechnology but also helps disentangling the genotype-phenotype map (Ryan et al., 2013).

The interplay of gene expression and protein abundance involves the intricate integration of transcription, post-transcriptional regulation, translation, and post-translational regulation. In the central dogma of molecular biology, which stipulates the flow of information in the cell, the first step is transcription of DNA to RNA, generally of a gene to mRNA. For all domains of life, this process is carried out by the many varieties of the RNA polymerase enzyme. It involves several layers of

regulation, one of which is the activity of transcription factors (TFs) (Alberts et al., 2002). TFs are proteins that control the transcription of genes by inducing or suppressing the expression of the gene by binding to specific DNA sequences named response elements, which can be either promoter or enhancer regions. They function by recruiting additional co-activators or co-repressors, stabilizing or blocking the transcriptional machinery, or can interact with histones and promote the relaxation or coiling of the DNA strand. TFs can also be the endpoints of cell signalling pathways, being activated or repressed in response to cellular and environmental cues. Post-transcriptional regulation involves the processes of maturing, processing, and turnover of RNA molecules. These processes include capping, splicing, polyadenylation, editing, transport, and degradation, and are regulated by RNA-binding proteins and non-coding RNAs (Bentley, 2014; Hentze et al., 2018). As a result, gene expression data capture the diverse transcriptional outcomes of biological systems (Brown and Botstein, 1999) (see Fig. 1).

Translation is the next step, whereby an mRNA molecule is translated

* Corresponding author.

E-mail address: zoran.nikoloski@uni-potsdam.de (Z. Nikoloski).

¹ These authors have contributed equally.

into a peptide or protein. It begins when proteins known as initiation factors bind to the mRNA and recruit the small ribosomal subunit to bind in a specific site. In prokaryotes, this site is the Shine-Dalgarno sequence. Eukaryotes do not have a specific sequence like the prokaryotic Shine-Dalgarno, but the small ribosomal subunit binds to the 5' cap and scans the mRNA molecule for the AUG start codon, using a consensus sequence, named Kozak sequence, that helps guiding the initial scan. The large ribosomal subunit then binds to the small ribosomal subunit and the elongation step begins. At the end of the mRNA sequence, the translation machinery is disassembled, and the emerging peptide or protein is released, which then follows on to post-translational modifications or regulation. This multi-step process, through which genomic information is translated into the proteome, is extensively regulated at the levels of initiation, elongation, localization, and ribosome composition (Hershey et al., 2019; Teixeira and Lehmann, 2019).

Evidence has indicated that the synthesis rate of a protein depends on the concentration and translational efficiency of its mRNA (Gebauer and

Hentze, 2004; Hershey et al., 2019) (see Fig. 1). Like transcription, regulation of translation is crucial for maintaining cellular homeostasis. It finetunes protein levels at different scales, enabling cells to adapt to varying physiological conditions and environmental cues. The resulting proteins exhibit distinct functions; they act as enzymes to drive biochemical reactions, provide structural support, and enable cells to perceive and respond to signals. These molecular functions are modulated by post-translational modifications as well as interactions with other proteins and small molecules (Mann and Jensen, 2003; Ryan et al., 2013). As a result, variations in protein abundance, structure, and function—often due to mutations—directly affect these processes, leading to differences in physical and physiological characteristics (Kidd et al., 2001).

The extent to which each step from transcription to translation contributes to the abundance of proteins is a long-standing open question. Addressing this question has been challenging due to the lag in methodological advances to quantify protein abundances, which

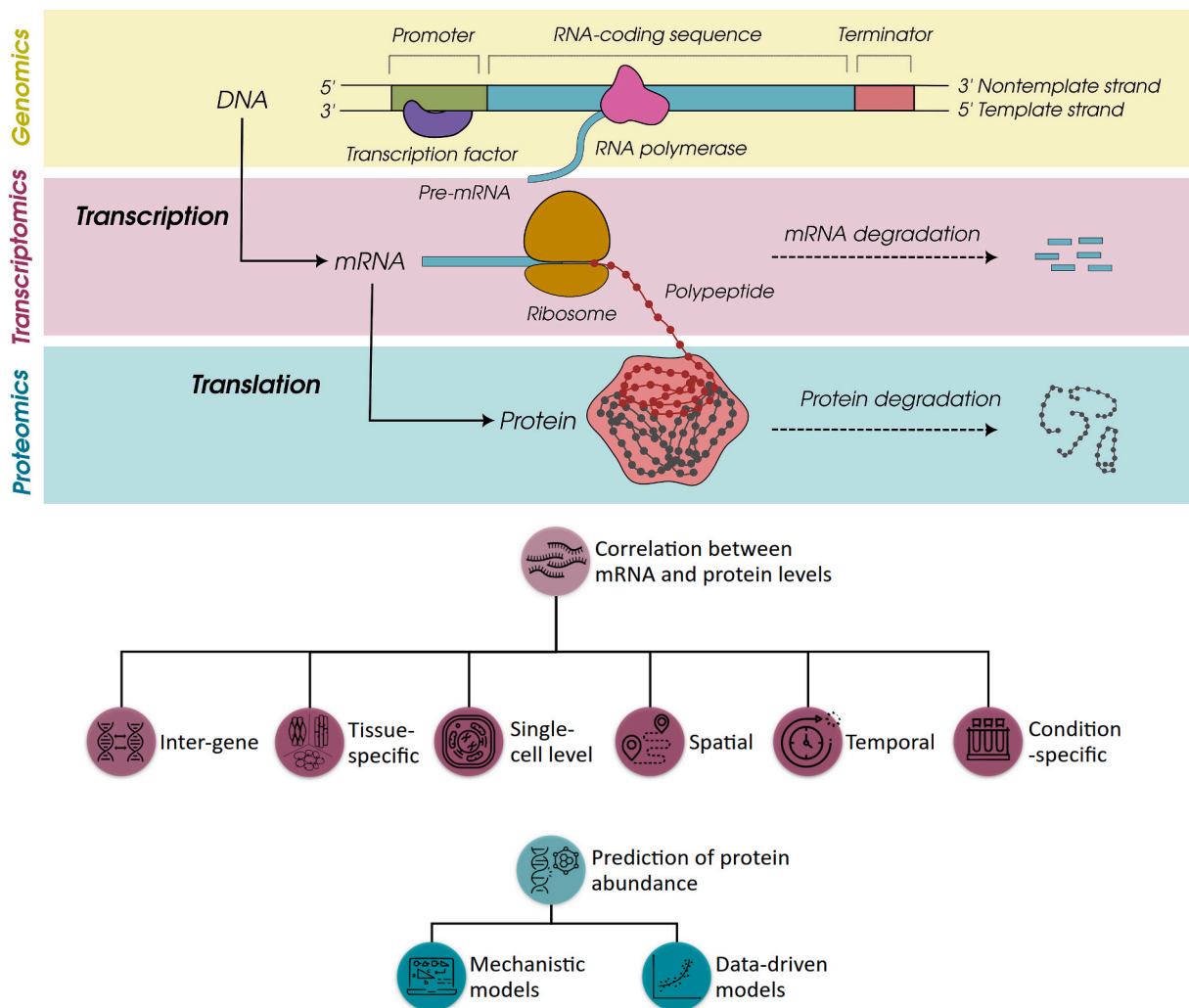


Fig. 1. Classification criteria for studies dissecting the association between protein abundance and gene expression. (a) A schematic representation of transcription and translation. The genomics layer (yellow) illustrates transcription initiation, where transcription factors and RNA polymerase interact with the DNA promoter, leading to pre-mRNA synthesis. The transcriptomics layer (purple) represents the processing of mRNAs, their translation by ribosomes into a polypeptide, and subsequent mRNA degradation. The proteomics layer (blue) depicts protein folding, the assembly of multiple polypeptides into functional protein complexes, and the potential degradation of proteins as part of regulatory processes. (b) A hierarchical classification of transcriptional regulation studies at the correlation between mRNA and protein levels, and prediction of protein abundance. This framework distinguishes between studies that focus on transcriptional mechanisms as the primary determinant of protein abundance and those that consider additional regulatory factors influencing gene expression. The first category is further divided into inter-gene, tissue-specific, single-cell level, condition-specific, spatial, and temporal experimental settings. On the other hand, studies predicting protein abundance utilize data-driven statistical models or mechanistic mathematical models to capture the complexity of gene expression regulation. This classification highlights the complexity of transcriptional control and the integration of different modelling approaches to understand gene expression dynamics.

contrasts the developments in gene expression profiling across different scales. Here, we provide an update on the seminal review that addressed this question (Vogel and Marcotte, 2012). To this end, comprehensively review the key studies from 2012 onward that investigate the relationship between mRNA levels and protein abundance. Our main contribution is a classification and a critical review of the advances in approaches that investigate the extent to which protein abundance can be accurately predicted using transcriptomics data. To this end, we first briefly review the existing techniques for transcriptomics and proteomic profiling with samples from different experimental set-ups. Our review covers experimental studies conducted on cell populations, comparatively addressing the correlation between mRNA and protein abundance across various organisms. We also review studies that examine differences in this relationship in both intra- and inter-gene contexts. Whilst the intra-gene context investigates the correlation between concentrations of proteins and their respective mRNAs across genes, the inter-gene context focuses on the correlation between concentrations from the same gene(s) across different individuals, conditions, or time points. We also include studies that investigate the mRNA-protein correlation by incorporating data on additional factors, such as translation rate, protein degradation, and post-transcriptional modifications, using different modelling frameworks. Finally, we briefly review the latest advances using data from single-cell analyses and offer a perspective about how these findings can propel biotechnological advances.

2. Transcriptome and proteome profiling

Profiling the transcriptome and the proteome offer invaluable insights about the functions of a biological system. Transcriptomics approaches are cost-effective, even for large-scale studies, while proteomics provide more detailed information about protein function, albeit at a higher expense. Further, while transcriptomics provides a full coverage of genes expressed in a cellular context, the coverage of proteins by proteomics techniques remains limited in comparison. Existing, detailed reviews cover the chronology of proteomics (Liu et al., 2016) and transcriptomics (Hrdlickova et al., 2017) approaches, along with their advantages and disadvantages. For completeness, here we provide a brief overview of techniques and approaches currently in use. In addition, we assembled a list of studies providing paired transcriptomics and proteomics data across diverse biological systems (Supplementary Table 1).

2.1. Transcriptome profiling

Compared to proteomics techniques, transcriptomics methods are generally less expensive, more standardized, and straightforward to perform. Microarray-based transcriptomics was the first approach to allow for measuring the expression of a large number of genes simultaneously (Reinartz et al., 2002; Schena et al., 1995; Velculescu et al., 1995). However, it is limited to known transcripts, since it relies on hybridisation. The introduction of RNA-seq technologies revolutionized the field by enabling the measurement of the expression of thousands of genes, including unknown genes and transcripts, which promoted the discovery of novel transcripts and splice variants. Following this breakthrough, many new developments on RNA-seq technology took place, such as single-cell RNA-seq (scRNA-seq) (Islam et al., 2011; Tang et al., 2009), spatial transcriptomics (Ståhl et al., 2016), and long-read RNA Sequencing (Deamer et al., 2016). These techniques allowed for gene expression profiling at single-cell resolution, integrated spatial information with gene expression data, and enabled real-time, direct RNA sequencing without the need for cDNA conversion, respectively. An extensive review highlighted the evolution of single-cell transcriptomics from early methods, like single-cell qPCR (Eberwine et al., 1992; Lambolez et al., 1992), to high-throughput technologies such as STRT-seq (Tang et al., 2009), SMART-seq (Picelli et al., 2014), and CEL-seq (Hashimshony et al., 2012), that enabled comprehensive RNA

profiling at single-cell resolution. Recent advancements include nanodroplet and picowell systems, *in situ* barcoding for analysing thousands of cells, and multi-modal techniques integrating RNA, DNA, and protein data. Spatial transcriptomics, including methods like smFISH (Chen et al., 2018), MERFISH (Xia et al., 2019), and *in situ* sequencing (Ke et al., 2013), now allow mapping of gene expression with spatial context, reaching single-cell resolution in tissues.

The high-throughput sequencing technologies used for RNA-seq outputs sequences known as reads. These reads are mapped to an annotated reference genome, and the number of mappings to specific genomic regions, such as coding sequences, are counted. The counting of gene mappings generates a matrix known as the gene count matrix, from which statistical analysis and differential gene expression analysis is performed. To ensure that gene expression comparisons are fair and not biased, normalization is performed on these gene counts. Traditional RNA-seq normalization methods, such as reads per kilobase million (RPKM), fragments per kilobase million (FPKM) and transcripts per kilobase million (TPM) rely on metrics such as sequencing depth and gene length. Other methods also take into account the RNA composition, such as DESeq2 median-of-ratios (Love et al., 2014) and EdgeR trimmed mean of M-values (TMM) (Robinson and Oshlack, 2010; Wagner et al., 2012). Each normalization method has its pros and cons, rendering it suited for specific purposes. These limitations can be addressed by external RNA spike-ins, which involve adding synthetic RNA molecules at known concentrations before sequencing to correct for variations in total RNA content due to experimental conditions or biological changes (Athanasiadou et al., 2019; Laosuntisuk et al., 2024).

2.2. Proteome profiling

The existing techniques for protein quantification can be divided into two groups, based on whether they provide *relative* or *absolute* quantification, each requiring different measurement approaches. Absolute quantification determines the concentration or content of a specific protein in a sample, essentially measuring the number of protein molecules per unit of volume or dry weight. In contrast, relative quantification compares the abundance of a protein across samples without providing an exact concentration, often focusing on changes in protein levels under different scenarios (e.g., conditions, treatments, and/or time points) (Vogel and Marcotte, 2012). Early methods, like enzyme-linked immunosorbent assay (ELISA) (Engvall and Perlmann, 1972) and western blot, determine the relative intensity of protein bands between different samples (Burnette, 1981) and are generally less expensive; however, they require specific antibodies and instruments, increasing costs for high-throughput profiling.

In the early 2000s, mass spectrometry-based approaches for absolute and relative quantification like stable isotope labelling by amino acids in cell culture (SILAC) (Ong et al., 2002), tandem mass tagging (TMT) (Thompson et al., 2003), and isobaric tag for relative and absolute quantification (iTRAQ) (Ross et al., 2004), offered high precision and multiplexing; however, they require expensive investments in equipment and reagents. Absolute quantification (AQUA) (Gerber et al., 2003) provides precise protein quantification using synthetic isotope-labelled peptides as standards; however, this method also has a limited dynamic range and sensitivity for low-abundance proteins. Selected reaction monitoring (SRM) (Picotti and Aebersold, 2012) is a cost-effective option mostly for absolute quantification in smaller-scale studies, although assay development is time-consuming. The introduction of sequential acquisition of all theoretical mass spectra (SWATH-MS) (Gillet et al., 2012) enhanced throughput without requiring labels, but instrument costs remain high. Other techniques, like reverse phase protein arrays (RPPA) (Petricoin III et al., 2005) and proximity ligation assay (PLA) that are primarily used for relative quantification, provide high sensitivity and throughput but still require specialized reagents (Fredriksson et al., 2002).

While emerging fields like single-cell proteomics offer innovative

insights, they remain costly due to complex equipment required. While automated analysis pipelines and machine learning methods have reduced labour demands, protein quantification, especially for large-scale studies, remains expensive and labour-intensive compared to simpler techniques. Advancements in two major approaches for single-cell proteomics—next-generation sequencing (NGS)-based methods and mass spectrometry (MS)-based methods—are comprehensively reviewed (Bennett et al., 2023). NGS-based methods use oligonucleotide-antibody conjugates for high-throughput protein detection. Prominent techniques include CITE-seq, which simultaneously profiles RNA and proteins (Stoeckius et al., 2017); proximity extension assays (PEAs) for sensitive protein quantification (Lundberg et al., 2010); and chromatin accessibility methods like ASAP-seq, which integrate chromatin and protein data for multimodal analysis (Mimitou et al., 2021). While these methods provide valuable multimodal insights, they depend on high-quality reagents and may suffer from background noise. MS-based methods offer direct protein quantification and broader proteome coverage. These include label-free analyses leveraging data-dependent or data-independent acquisition (DDA/DIA) workflows (Brunner et al., 2022) and multiplexed analyses using isobaric labelling (e.g., TMT) or multiplexed DIA to improve throughput and mitigate issues, like ratio compression (Petelski et al., 2021). Innovations such as nanoPOTS together with machine learning further enhance MS sensitivity and depth, enabling detection of low-abundance proteins and post-translational modifications (Zhu et al., 2018). While MS excels in uncovering novel proteins and providing in-depth proteomic coverage, NGS-based methods are distinguished by their scalability and multimodal capabilities. The integration of these complementary approaches can enable comprehensive multi-omics analyses (Bennett et al., 2023).

3. Factors affecting the association between mRNA levels and protein abundance

While the nucleotide sequence of a gene defines its mRNA product and the mRNA sequence determines the amino acid sequence of the resulting polypeptide, the relationship between mRNA level and the corresponding protein abundance is not as straightforward (McManus et al., 2014). Using mRNA levels to predict protein abundance is challenging since both mRNA expression and protein abundance are results of dynamic, coupled processes that respond to internal and external cues.

In eukaryote cells, mRNAs are produced much more slowly than proteins. For instance, a mammalian cell generates about two copies of an mRNA per hour, yet dozens of proteins per mRNA can be synthesized in the same time (Schwanhäusser et al., 2011). In addition, in yeast, protein levels are often three orders of magnitude higher than mRNA, with lower mRNA expression linked to regulatory and stress-response genes, and higher expression to essential functions, like translation (Marguerat et al., 2012). This is due to the many steps of mRNA processing, regulation, and transport to the cytosol before initiating translation. In contrast, transcription and translation are more coordinated in prokaryotes, with translation being initiated as soon as the Shine-Dalgarno sequence and start codon are transcribed by the RNA polymerase (Irastortza-Olaziregi and Amster-Choder, 2021). As a result, there is a higher correlation between mRNA levels and protein abundance in prokaryotes than in eukaryotes.

In eukaryotes, mRNAs are less stable and show higher turnover than proteins. For instance, in mammals, mRNAs have an average half-life of 2.6–9 hours, compared to 46 hours for proteins (Marguerat et al., 2012; Schwanhäusser et al., 2011; Sharova et al., 2009; Wang et al., 2019). In addition, evidence from mammals indicate that proteins exhibit a broader dynamic range of stability than mRNAs (Lundberg et al., 2010; Schwanhäusser et al., 2011; Vogel and Marcotte, 2012; Wang et al., 2019). Possible explanations for these observations directly relate to the processes that affect protein abundance (e.g., translational efficiency or half-lives) (Lundberg et al., 2010).

In prokaryotes, mRNA turnover is much faster than in eukaryotes. For example, in *Escherichia coli*, the half-life of a mRNA molecule ranges from less than one minute to around two minutes (Li and Xie, 2011). By contrast, for proteins, the average half-life is around 20 hours (Moran et al., 2012). Despite the brevity of mRNA half-lives, translation rates are still bigger than mRNA decay, since full, mature mRNAs are necessary to produce functional proteins. Given the coupling between transcription and translation, ribosomes can physically block the access of nucleases to the mRNA molecule. They can also prevent the formation of secondary structures on the mRNA molecule that could halt translation. Therefore, mRNA decay is directly linked to transcription and translation efficiency. This is attested by expressing the RNA polymerase of the phage T7 in *E. coli*, which has a higher processivity than the bacterial RNA polymerase. When transcribing *E. coli* genes, the higher rate of transcription results in a destabilization of the coupled transcription-translation machinery, exposing cleave sites and resulting in the mRNA being cleaved and degraded (Iost and Dreyfus, 1995).

Spatial factors may also significantly affect the association between mRNA and protein abundance. Many mRNAs localize to distinct sub-cellular sites within a cell (Martin and Ephrussi, 2009), and protein synthesis may vary depending on mRNA location, as ribosome and tRNA availability can differ spatially (Besse and Ephrussi, 2008; Chai et al., 2014). Consequently, the protein-to-mRNA ratio may fluctuate based on the intracellular localization of these molecules (Ori et al., 2015). Integrating transcriptomics and proteomics at single-cell resolution may provide the means to address the association between mRNA expression and protein abundance (Petrosius and Schoof, 2023), which we explore in Section 6.1.

Since the translation of a protein from mRNA is a dynamic process, any differences in the efficiencies of the involved steps can affect the change of association based on different time snapshots of protein abundances and gene expression. Factors such as cell cycle variability (Buettner et al., 2015), transcriptional bursting (Blake et al., 2003), and the time lag between transcription and translation also contribute to the divergence between mRNA and protein levels (Liu et al., 2016; Vogel and Marcotte, 2012). These challenges underscore the limitations of using mRNA as a proxy for protein levels (Darmanis et al., 2016). Moreover, sequencing based methods inherently lack the ability to detect protein post-translational modifications, which are often essential for protein function.

Lastly, evidence suggests that protein abundances are more conserved than mRNA levels across species, indicating stronger evolutionary constraints on protein abundance to meet essential functional and physiological demands (Khan et al., 2013; Schrimpf et al., 2009; Wang et al., 2019). While mRNA levels can vary over time, post-transcriptional, translational, and protein degradation mechanisms maintain protein levels that are likely optimized to ensure organism's fitness (Laurent et al., 2010). Furthermore, although many genetic variants (e.g. expression quantitative trait loci) affect mRNA levels, their impact on protein abundance is limited, suggesting a buffering effect that stabilizes protein levels (Battle et al., 2015). For instance, essential genes and those involved in molecular complexes exhibit more conserved transcriptional and translational regulation, with translational profiles displaying stronger correlations across isolates compared to transcriptional profiles (Teyssonniere et al., 2024).

4. Classification of studies

The relationship between gene expression and protein levels can vary depending on the type of data used for analysis. In probing the relationship between gene expression and protein levels, one set of approaches make use only of transcript levels and protein abundance, whilst others predict protein abundance using a combined analysis of mRNA expression and other factors related to protein biosynthesis, such as regulatory elements and codon usage bias. Amongst these, a plethora of modelling methods have been employed, that can be grouped into

mechanistic or data-driven methods (see Fig. 1).

4.1. Analysing the correlation between mRNA and protein levels

Numerous studies have examined the correlation between mRNA expression and protein levels in diverse biological systems, from mammalian cells (Cenik et al., 2015; Edfors et al., 2016; Gautier et al., 2016; Jovanovic et al., 2015; Li et al., 2014; Schwanhäusser et al., 2011), other metazoans (Becker et al., 2018; Schrimpf et al., 2009), yeast (Csárdi et al., 2015; Lawless et al., 2016; Wang et al., 2015), and bacteria (Frumkin et al., 2018; Nie et al., 2006) to model plant species (Mergner et al., 2020; Ponnala et al., 2014). To illustrate the extent of correlation between mRNA and protein expression, the Spearman rank coefficient (ρ), Pearson correlation coefficient (r), and the coefficient of determination (R^2) are frequently used. However, the Spearman rank coefficient (ρ), that quantifies non-linear associations, is often considered more appropriate, as protein and mRNA abundances may not covary linearly. Nevertheless, all three metrics are commonly and, often, interchangeably used in publications on mRNA level–protein abundance correlations (see Table 1).

In the following, the existing findings about correlations between gene expression and protein levels are divided based on context in which they have been examined.

Inter-gene correlations. A 2002 study revealed that mRNA levels in *E. coli* do not consistently correlate with protein abundance (Bernstein et al., 2002). This finding was further supported by a 2010 study that measured mRNA and protein levels for over 1,000 *E. coli* genes, reporting no significant correlation within individual cells ($r = 0.01 \pm 0.03$). However, when averaging signals across measurements, mRNA levels could explain 54–77% of the variance in average protein levels (Taniguchi et al., 2010). A notably low correlation between mRNA and protein levels was found in a 2010 study, despite targeting mRNA and proteins from the same genes; here, the average coefficient of determination (R^2) across single cells was only 0.04, with a maximum of 0.35 for specific pairs (Darmanis et al., 2016).

Pioneering inter-gene studies in yeast and human which had suggested a low correlation between mRNA and protein levels in eukaryotes as well (Anderson and Seilhamer, 1997; Beyer et al., 2004; Brockmann et al., 2007; Chen et al., 2002; Ghaemmaghami et al., 2003). A comprehensive study from 2011 in mammalian cells used the quantification of the expression and abundance of over 5,000 genes and proteins using RNA-seq, pSILAC, and absolute protein quantification techniques, respectively. In contrast to the pioneering results, this study reported a value of 0.41 for R^2 between mRNA and protein levels (Schwanhäusser et al., 2011).

A study published in 2014 on mammalian tissue cell culture reported a stronger correlation between mRNA levels and protein abundances after recalibrating data using direct measurements of 61 housekeeping proteins, suggesting that mRNA expression accounts for a greater variance in protein levels than previously thought (Li et al., 2014). This analysis found that mRNA levels explain 56–84% of the variance in protein abundance, with the Pearson correlation coefficient between mRNA levels and protein abundances increasing slightly from 0.626 to 0.642 after applying this correction model. Subsequent analyses further reduced errors by addressing biases in mass spectrometry and mRNA sequencing data, employing multiple independent measurements, and applying advanced statistical methods, such as: Bayesian modelling and error-corrected scaling. By filtering unreliable data, linking analyses to genetic variations, and incorporating direct measurements, like ribosome foot-printing, these studies minimized confounding errors and identified transcription as the dominant factor, explaining up to 73% of the variance in protein expression (Li and Biggin, 2015).

Another investigation across a diverse set of human samples found that the Spearman correlation between mRNA expression and protein levels is not always strong, with a median of 0.22 (Cenik et al., 2015). In addition, a study from 2015 measured absolute protein copy numbers

Table 1

Chronological compilation of studies on protein abundance. The table summarizes the contributions of different regulatory factors to explaining protein abundance across different organisms and studies. The studies are sorted by year of publication.

Factors	Organism	Study
mRNA level ($\rho = 0.46$) CDS length ($\rho = -0.53$) 5' UTR secondary structure stability ($\rho = 0.20$) uORFs ($\rho = -0.18$ to -0.21) 3' UTR length ($\rho = -0.19$) ribosome density ($\rho = 0.19$) Codon bias index ($\rho = 0.08$)	Human	Vogel et al. (2010)
Translation efficiency ($R^2 = 0.07$ – 0.16) mRNA stability ($R^2 = 0.84$)	Mammal	Guo et al. (2010)
mRNA level ($r = 0.6$)	Yeast	Fournier et al. (2010) Lundberg et al. (2010)
mRNA level ($\rho = 0.58$ – 0.63) mRNA level ($R^2 \approx 0.4$) Transcription rate ($R^2 = 0.3$ – 0.4) mRNA degradation ($R^2 \approx 0.06$) Translation rate constant ($R^2 \approx 0.55$)	Human	Schwanhäusser et al. (2011)] Marguerat et al. (2012)
Protein degradation $R^2 \leq 0.05$	Mammal	Kristensen et al. (2013)
mRNA level ($R^2 = 0.36$ – 0.55)	Yeast	Guimaraes et al. (2014)
mRNA level ($R^2 = 0.49$) Protein synthesis rate ($R^2 = 0.42$) Protein degradation rate ($R^2 = 0.13$) mRNA level ($R^2 = 0.53$) Translation elongation efficiency ($R^2 = 0.12$) Translation initiation factors ($R^2 = 0.01$)	Yeast	Li et al. (2014) Zhang et al. (2014) Wilhelm et al. (2014)
mRNA level ($r = 0.642$) mRNA level ($\rho = 0.23$ – 0.47)	Mammal Human	Wang et al. (2015)
mRNA level ($\rho = 0.41$ – 0.55)	Human	Li and Biggin (2015) Cenik et al. (2015) Csárdi et al. (2015)
Ribosomal density ($r = 0.62$) mRNA level ($R^2 = 0.51$) mRNA level ($R^2 = 0.73$) mRNA level ($\rho = 0.22$) mRNA level ($R^2 \leq 0.85$) mRNA levels ($R^2 = 0.59$ – 0.68) Translation ($R^2 = 0.18$ – 0.26) Protein degradation ($R^2 = 0.08$ – 0.22)	Yeast Mammal Human Yeast Mouse	Jovanovic et al. (2015)
mRNA level ($r \approx 0.6$) mRNA level ($\rho = 0.41$ – 0.676) mRNA level ($\rho = 0.58$) Codon bias ($R^2 = 0.53$) RNA-binding protein enrichment ($R^2 = 0.42$) Transcript secondary structure ($R^2 = 0.33$) poly-A tail length ($R^2 = -0.16$)	Human Human Yeast	Edfors et al. (2016) Gautier et al. (2016) Lawless et al. (2016)
Protein degradation (Wilcoxon rank test, $p < 0.05$) Protein turnover (Wilcoxon rank test, $p < 0.05$) mRNA levels ($R^2 \approx 0.94$) TRmIND ($R^2 \approx 0.05$) Protein degradation ($R^2 \approx 0.01$)	Yeast	Li et al. (2017a)
mRNA level ($\rho = 0.21$) mRNA level ($R^2 = 0.46$ – 0.88)	Human Yeast	Fortelny et al. (2017) Lahtvee et al. (2017)
mRNA level ($\rho = 0.54$) mRNA level ($R^2 = 0.34$ – 0.54) mRNA level ($\rho = 0.46$ – 0.62) Ribosome profiling ($\rho = 0.67$ – 0.71)	<i>Drosophila melanogaster</i> Rat Yeast	Becker et al. (2018) Moritz et al. (2019) Blevins et al. (2019)
mRNA level ($r \geq 0.3$) mRNA level ($\rho = 0.42$ – 0.57) Ribosome profiling ($\rho = 0.60$ – 0.69)	Rat Mammal	Shen et al. (2020) Wang et al. (2020)
mRNA level ($r = 0.28$ – 0.70)	<i>Arabidopsis thaliana</i>	Mergner et al. (2020)

Note: R^2 = coefficient of determination; r = Pearson correlation; ρ = Spearman rank correlation; CDS = mRNA coding sequence; uORFs = upstream open reading frames; TRmIND = mRNA abundance-independent translation rate

and probed their Pearson correlation to mRNA levels across human tissues and cell lines. This study found a Pearson correlation of around 0.6 for the tissues analysed, ranging from a minimum of 0.39 in a kidney-derived cell line to a maximum of 0.79 in a breast-derived cell line (Edfors et al., 2016).

A study in 2016 also observed a modest Spearman correlation, ranging from 0.41 to 0.676, between mRNA levels and absolute abundance of 6130 proteins measured during the differentiation of human erythroid progenitors (Gautier et al., 2016).

In addition to studies on mammalian systems, several studies explored the correlation between mRNA and protein in yeast. For instance, a 2015 study found that the Pearson correlation between the transcriptome and proteome in *Saccharomyces cerevisiae* was at 0.51 (Wang et al., 2015). Further, reanalysis of data from 24 yeast studies, using methods that considered effects of noise, demonstrated that mRNA levels explain more than 85% of the variation in steady-state protein levels across many different genes (Csárdi et al., 2015). Further, a study from 2016 quantified nearly 2,000 proteins in *S. cerevisiae* using selected reaction monitoring (SRM) mass spectrometry with stable isotope labelled (SIL) QconCAT standards (Lawless et al., 2016). By integrating RNA-seq data, this study found that 70% of protein abundance in yeast could be explained by mRNA levels, further highlighting the significant role of transcriptional control in protein expression. Lastly, in plant studies, *Arabidopsis thaliana* has been used to investigate the correlation between mRNA and protein levels. Across different tissues, Pearson correlation coefficients were found to range from 0.28 to 0.7 (Mergner et al., 2020).

An alternative approach to inter-gene associations examines how changes in mRNA levels of a single gene drive alterations in the abundances of the corresponding protein. This is performed by correlating protein abundance with mRNA levels from the corresponding gene(s) across various states measured in different individuals, cell types, time points, or environments. Such comparisons can provide insights into average properties and global trends across genes. Given that specific proteins may exhibit abundances that differ significantly from the average, exploration of these variations can offer valuable biological insights, potentially pointing to strong transcriptional or post-transcriptional regulation of proteins and cellular processes (Vogel and Marcotte, 2012).

Tissue-specific analyses. Comparing mRNA levels and protein abundances across tissues and organs is challenging due to the diverse cell types present at different developmental stages in most tissues. To address this issue, a 2010 study conducted a quantitative comparison of the human transcriptome and proteome in three well-characterized human cell lines from different functional origins. This study found Spearman correlation values ranging from 0.58 to 0.63 between transcript level and protein abundance differences across these cell lines, suggesting that changes in transcript levels were, in general, reflected by corresponding changes in protein levels. However, protein levels were found to show a broader dynamic range than mRNA levels (Lundberg et al., 2010).

A 2014 study on human colon and rectal tumours found that while mRNA and protein abundance were positively correlated within individual colorectal cancer samples (average value of 0.47), the average Spearman correlation across different tumours was much lower (value of 0.23), suggesting that mRNA levels alone are poor predictors of protein abundance (Zhang et al., 2014).

These observations led to investigating transformations of the data on mRNA levels and protein abundances. For instance, the protein-to-mRNA (PTR) ratio for a given gene has gained interest and led to numerous studies that investigated its importance in improving the accuracy of protein abundance predictions based on mRNA levels (Edfors et al., 2016; Erasan et al., 2019; Fortelny et al., 2017; Franks et al., 2017; Mergner et al., 2020; Wilhelm et al., 2014). Following this approach, a 2014 study examined the Spearman correlation between mRNA levels and protein abundances across 12 human tissues, finding a

moderate overall correlation ranging from 0.41 to 0.55. This study also found that, despite variations in expression levels, the PTR ratio for each gene was consistent across tissues. Using the median PTR ratio, they demonstrated that protein abundance could be reliably inferred from mRNA levels, with a strong correlation (around 0.9) between predicted and measured protein levels (Wilhelm et al., 2014). A 2015 study performed a similar analysis but expanded the number of samples from 12 to 30 human samples, identifying proteins corresponding to 17,294 genes (84% of human protein-coding genes) (Kim et al., 2014).

However, the previous finding was challenged in two studies from 2017 (Fortelny et al., 2017; Franks et al., 2017). The first study used control experiments to show that the high correlations between predicted and observed protein levels were largely due to between-gene variations rather than accurate predictions for individual genes. Given that a much lower median Spearman correlation within genes ($\rho = 0.21$) was found, this study indicated that gene-specific translation rates do not reliably predict protein levels at the individual gene level (Fortelny et al., 2017). It further suggested that the previous findings may serve as an illustration for the Simpson's paradox, where a trend observed in aggregated data does not hold within individual subsets. Here, the high correlations across genes resulted from large variations in protein levels between genes, while within each gene, mRNA levels did not consistently predict protein abundance (Fortelny et al., 2017).

The second critical study highlighted that both data sets from the 2014 studies (Kim et al., 2014; Wilhelm et al., 2014) suffered from substantial measurement noise, undermining the reliability of protein-mRNA correlations. They also argued that scaling mRNA levels by a median protein-to-mRNA ratio relies on the assumption of consistent protein translation rates across tissues. This method overlooks tissue-specific variability in these ratios, potentially conflating mean-level variability with cross-tissue variability, which may fail to capture the dynamic regulation unique to different tissues (Franks et al., 2017).

A 2016 study examined if transcript levels of a given gene can serve as proxies for the corresponding protein levels both across tissues and genes. For the tissue-level analysis, introducing a gene-specific mRNA-to-protein conversion factor improved the mRNA-protein correlation to a median Pearson correlation of 0.93 (Edfors et al., 2016). This finding was further critically discussed, confirming that the protein-to-mRNA ratio was largely conserved across tissues for a given gene, but that it can vary substantially between genes (Silva and Vogel, 2016).

Single-cell level. Traditional bulk analyses provide only population-level averages, limiting the capacity to capture cell-to-cell variability and obscuring critical insights into altered cell population dynamics and shifts in cell-type-specific transcriptomes and proteomes (Singh, 2021). By aggregating data from all cells within a population, these approaches cannot reveal the heterogeneity inherent in cellular responses, a limitation particularly significant in studies of cancer stem cells (Bonnet and Dick, 1997). Increasing evidence shows that distinct cellular states can respond differently to identical extrinsic signals, with variability often observed even among cells of the same type (Aissa et al., 2021; van Galen et al., 2019). Advances in single-cell technologies, mentioned above, have enabled precise quantification of transcriptomes and proteomes with single-molecule sensitivity (Li and Xie, 2011), facilitating the investigation of the correlation between mRNA levels and protein abundance. Methods for simultaneous measurements of both mRNA and protein include PEA/STA (Alex S Genshaft, 2016), PLAYR (Andreas P Frei, 2016), CITE-seq (Stoeckius et al., 2017), REAP-seq (Vanessa M Peterson, 2017) and RAID (Jan et al., 2019), which were reviewed in detail by Flynn et al. (2023) and Xuefei Wang et al. (2024).

The earliest attempts at measuring both mRNA and protein abundances simultaneously in a single cell relied on fluorescence assays. In the 2010 study of Taniguchi et al. (2010), they constructed a yellow fluorescent protein fusion library for *Escherichia coli* and analysed fluorescence images at the molecule level. For 127 genes with high expression they could detect, they assessed the correlation between their mRNA and protein concentrations. The observed an almost zero Pearson

correlation for these genes ($r = 0.1$). Another study was the work of Ståhlberg et al. (2012). In this study, they performed a combination of measurements from the same single cell in the human fibro-sarcoma cell line HT1080. They performed reverse transcription followed by quantitative PCR to measure mRNA levels, and proximity ligation assays for protein quantity in GFP tagged proteins. When comparing both measurements, they observed a Spearman correlation of 0.31 ($p = 0.01$).

Later attempts combined high-throughput multi-omics approaches to measure mRNA and protein levels in single cells. Specht et al. (2021) analysed both mRNA and protein levels in macrophages to probe for cell heterogeneity and obtained Pearson correlations between the transcriptomics and proteomics measurements ranging from -0.2 to 0.4. Similarly, the study of Gayoso et al. (2021) employed CITE-seq and measured mRNA and surface proteins from mouse lymph node cells. The Pearson correlations between mRNA and surface protein levels ranged between 0.0 and 0.8. In the study of Fulcher et al. (2024) they performed parallel measurements of both transcriptome and proteome from single cells using an approach they named nanodroplet splitting for linked-multimodal investigations of trace samples (nanoSPLITS). This study analysed samples from murine cell lines NAL1A and C1C10, and from primary cells isolated from human pancreatic islets. The Pearson correlation between mRNA and protein levels across cells and conditions ranged from 0.35 to 0.45. Despite these advancements, the correlations between transcript and protein levels are similar to those observed in bulk omics studies.

Investigations based on temporal data. If a population of cells is exposed to a stimulus, such as stress, protein abundances fluctuate over a certain period, moving the system out of steady state until an equilibrium is (re)established (Vogel and Marcotte, 2012). Here, we review results on the association of mRNA levels and protein abundances from studies that focused on state transitions and various perturbations affecting the steady state.

The temporal scales under consideration can greatly influence the relationship between time resolved protein levels and their coding mRNAs (Liu et al., 2016). A 2010 study revealed a time-shifted correlation between mRNA and protein expression in yeast cells treated with rapamycin, with early mRNA changes (1-2 hours) aligning best with later protein responses (4-6 hours), reaching a peak Pearson correlation coefficient of 0.60 (Fournier et al., 2010).

In 2012, a comprehensive transcriptomics and proteomics study on fission yeast found that mRNA and protein levels are generally well-correlated, but this correlation varies with cell state (Marguerat et al., 2012). In proliferating cells, the coefficient of determination (R^2) was 0.55, indicating a moderate to strong relationship. In quiescent cells, R^2 was weaker, with a value of 0.36, suggesting that protein levels are less directly tied to mRNA levels, probably due to their longer half-lives (Marguerat et al., 2012).

In 2015, a study explored how protein expression in mammalian cells is regulated through various stages of the protein life cycle, most importantly transcription (Jovanovic et al., 2015). This study found that in pre-stimulation, mRNA levels significantly influence overall protein expression, more than protein translation and degradation combined. For instance, mRNA levels explained 59-68% of the variance in protein levels (Jovanovic et al., 2015). A 2018 study investigated the relationship between mRNA levels and protein abundances during embryogenesis of *Drosophila melanogaster*, using data for 3,761 genes at 14 stages of embryonic development. This study found moderate correlations between mRNA and protein levels ($p = 0.54$) (Becker et al., 2018).

Investigations based on spatially resolved data. A 2019 study explored the correlation between mRNA levels and protein abundances in different regions of the rat brain (Moritz et al., 2019). The study found that mRNA and protein levels exhibited poor to moderate correlation (R^2 values between 0.34 and 0.54), largely attributed to neuronal polarity. Specifically, the transport of proteins across long distances within neurons often results in proteins being present in regions where their corresponding mRNAs are not. This study identified neuronal polarity as

a third factor contributing to poor transcript protein correlation, alongside translational regulation and protein turnover (Moritz et al., 2019).

Investigations based on data from different conditions. A 2012 study investigated changes in mRNA levels, translation, and protein abundances in fission yeast, *Schizosaccharomyces pombe*, during oxidative stress, heat shock, and DNA damage over multiple time points (15–120 minutes). This temporal analysis in different conditions highlighted dynamic regulatory changes and delayed effects, such as the lag between mRNA transcription and protein synthesis. A strong Pearson correlation was observed between mRNA and protein levels for up-regulated genes ($r = 0.74$, $p < 1e^{-13}$), while down-regulated genes showed no significant Pearson correlation ($r = 0.07$, $p \approx 0.27$). Oxidative stress primarily relies on transcriptional regulation, whereas heat shock and DNA damage involve distinct translational responses, underscoring the tailored nature of stress adaptation (Lackner et al., 2012).

Another study in 2017 quantified the absolute levels of mRNA and proteins in *Saccharomyces cerevisiae* under ten different environmental conditions. The authors found that while the overall correlation between mRNA and protein abundances across all conditions was moderate ($R^2 = 0.46$), a much stronger correlation was observed for differentially expressed proteins, with a median R^2 of 0.88. This finding suggests that while general protein levels are influenced by multiple regulatory factors beyond mRNA abundance, genes that undergo significant expression changes tend to exhibit a more direct transcriptional control over protein synthesis (Lahtvee et al., 2017).

A 2020 study provided a detailed analysis of how hypoxic stress affects cardiac cells in rats at both the transcriptional and protein synthesis levels. Pearson correlation coefficients were used to examine the relationship between gene expression and active translation, identifying genes with positive, neutral, or negative correlations between these two processes. Notably, positive Pearson correlations ($r \geq 0.3$) were observed for 14% (i.e., 603) of the genes, indicating a closer alignment between their transcription and translation levels. This analysis highlighted the intricate regulatory mechanisms that control protein production in response to hypoxia in cardiac cells (Shen et al., 2020).

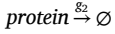
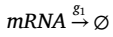
4.2. Prediction of protein abundance from RNA-seq and other data

System-wide studies measuring transcripts and proteins across the genome have emphasized the critical role of multiple factors beyond transcript levels on protein abundance. Overall, protein abundance is influenced by the rates of transcription, mRNA turnover, translation rates, codon usage bias, and protein degradation (Liu et al., 2016; McManus et al., 2014). The studies reviewed above consistently reported a moderate to low correlation between mRNA expression and protein abundance. These findings suggest that variability in protein abundance may be influenced, in part, by factors other than just mRNA levels. For instance, a 2020 study in *A. thaliana* have shown that the relative contribution of mRNA levels to predict protein abundance is only 19%, while codon usage bias has a relative contribution of 21%, protein interactions, mutation rates, regulatory elements and other factors amount to 12%, and 48% remain attributed to unknown factors (Mergner et al., 2020). Therefore, it is essential to explore additional physiological and regulatory mechanisms along the central dogma to better predict protein abundance.

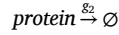
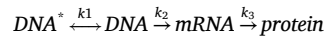
To unravel the contributions of various regulatory processes, two primary approaches have proven effective. The first entails the development of mechanistic mathematical models that incorporate all potential factors influencing protein abundance to provide a comprehensive predictive framework. The second approach involves statistical methods, such as regression analysis, which link variations in protein abundance to specific mRNA and protein sequence features associated with distinct regulatory mechanisms (Vogel et al., 2010).

Mechanistic mathematical models. The association between

mRNA levels and protein abundance can be expressed in terms of kinetic constants of gene expression. Constitutive gene expression illustrates the simplest model for this association:



where k_1 denotes the transcription rate (from DNA to mRNA), k_2 , the translation rate (from mRNA to protein), g_1 stands for the mRNA degradation rate, and g_2 for protein degradation rate (primarily due to cell division for stable proteins like YFP fusions). This model is defined as a two-stage model (Fig. 2a). This model can also consider the activation stage of a gene, defined as:



where DNA^* is the gene in its inactive form, and DNA is the gene in its active form. This variation is defined as a three-stage model (Fig. 2b). Foundational stochastic frameworks for modelling gene expression based on these principles were pioneered by studies such as [McAdams and Arkin \(1997\)](#) and [Paulsson et al. \(2000\)](#). The former addressed the question about the factors that influence the dynamics of protein concentration changes over time after its corresponding gene is activated. This study introduced the concept of switching delay, referring to the time interval between promoter activation and expression response, which was evaluated using Monte Carlo simulations. The latter study

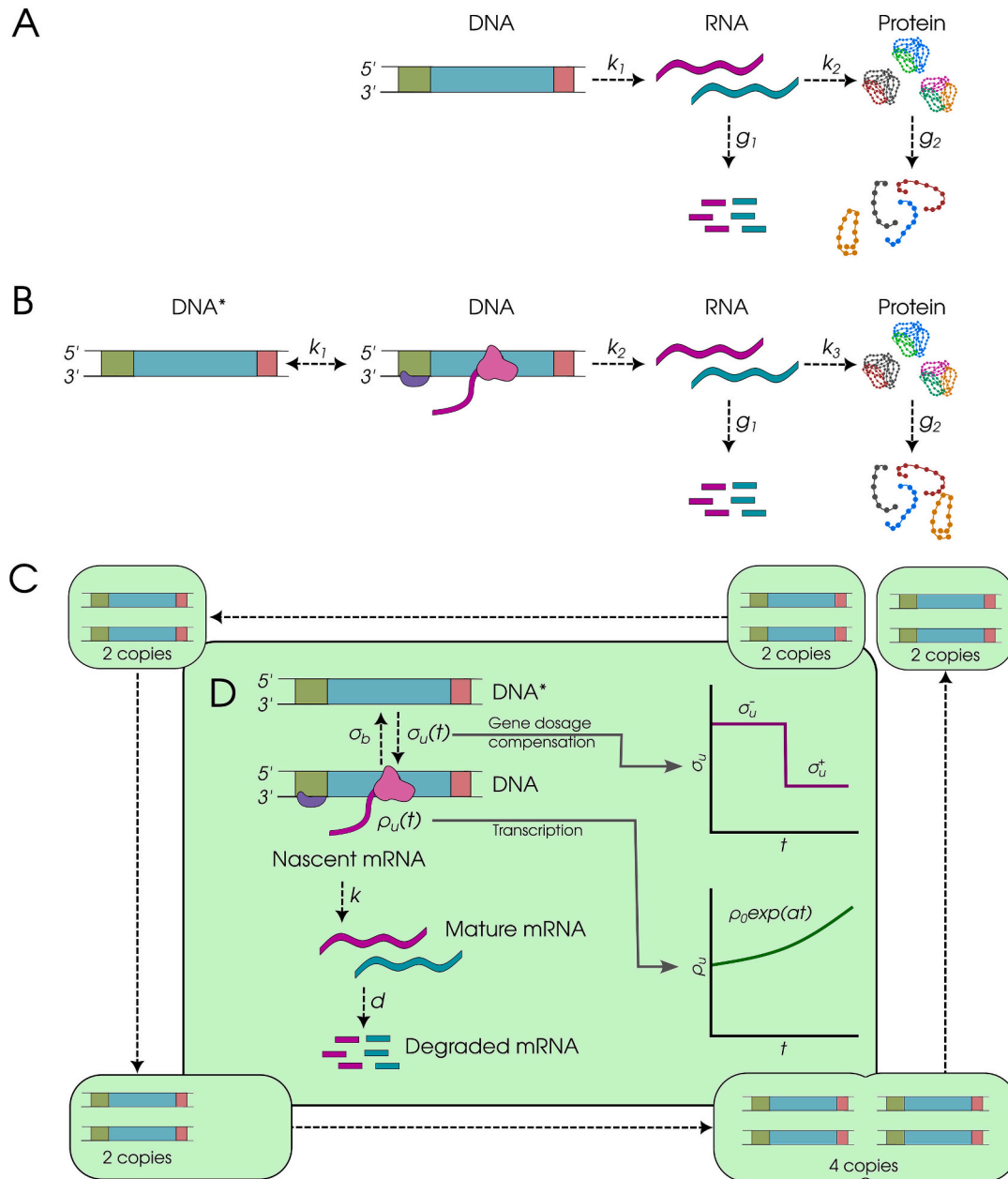


Fig. 2. Models of stochastic gene expression. (a) Two-stage model. (b) Three-stage model. (c) Expanded model to account for growth, gene replication and gene dosage. Adapted from [Cao and Grima, 2020](#).

deals with how sensitivity in regulatory mechanisms is affected by signal noise, originating from random fluctuations in regulatory responses due to low copy numbers of these signalling molecules. It found that despite these molecules being present in low number, they increase rather than decrease regulatory sensitivity, defined as stochastic focusing.

A breakthrough study in 2006 derived the gamma distribution to describe the steady-state protein concentration in a population of cells, linking it explicitly to the stochastic dynamics of gene expression (Friedman et al., 2006). The gene expression process is represented by the aforementioned two-stage model. The distribution of protein copy numbers, x , is then described by a gamma distribution detailed in Cai et al. (2006):

$$p(x) = \frac{x^{a-1} e^{-x/b}}{\Gamma(a)b^a}$$

where:

$$a = \frac{m_p^2}{\sigma_p^2}$$

$$b = \frac{\sigma_p^2}{m_p^2}$$

with a associated with the “burst frequency” $\left(\frac{k_1}{g_2}\right)$, b represents the “burst size” $\left(\frac{k_2}{g_1}\right)$, and m_p^2 and σ_p^2 denote the mean and variance of the protein distribution, respectively. It is important to highlight that the gamma distribution is a continuum approximation, since the protein copy numbers are integers. This model establishes a quantitative connection between kinetic parameters and measurable metrics, such as mean protein abundance (x) and variance (σ^2), and provides a representation of stochastic chemical dynamics considering molecule number (discrete) and time (continuous). It elucidates how transcriptional noise propagates to protein levels, linking mRNA and protein dynamics to reveal gene expression variability and underlying regulatory mechanisms. The model showed remarkable agreement with stochastic simulations, accurately capturing protein distribution mean and variance with relative errors as low as 1% under biologically relevant conditions (Friedman et al., 2006). This theoretical framework was empirically validated in a study that developed a robust technique for real-time analysis of low copy number protein expression with single-molecule sensitivity (Cai et al., 2006). A follow-up study expanded on these findings and provided a simplification of the master equation of gene expression (Shahrezaei and Swain, 2008). Assuming protein synthesis occurs in bursts, mRNA fluctuations can then be implicitly represented in the chemical master equation. Further, the lifetime of proteins is also assumed to be much longer than the lifetime of mRNA molecules (minutes for mRNA compared to the length of the cell cycle for proteins). This is expressed through a parameter, gamma, denoting the ratio between the probability per unit time of degradation of an mRNA and the probability per unit time of degradation of a protein molecule; gamma takes values larger than 1 for about 80% of genes in yeast. For large values of gamma, mRNA levels can be considered to be at a steady-state for most of the lifetime of a protein., provided gammas much larger than 1:

$$\frac{dP_n}{d\tau} = a \left[\left(1 - \frac{b}{1+b}\right) \sum_{r=0}^n \left(\frac{b}{1+b}\right)^r P_{n-r} - P_n \right] + (n+1)P_{n+1} - nP_n.$$

Therefore, by simplifying the dynamics of a stochastic system, they derive an approach to analyse general fluctuation in gene expression data, avoiding large numbers of simulations necessary and aiding in deconvolving factors that can affect protein distributions. The solution is a discrete, negative binomial distribution that approximates to the gamma distribution found by Cai et al. (2006).

Later in 2010, there was another extension to this model, which extended gamma distribution modelling to a genome-wide scale and identified two distinct noise regimes: (1) Intrinsic Noise: At low protein levels ($m_p < 10$), noise scales inversely with mean protein abundance, resembling Poisson behaviour, and (2) Extrinsic Noise: At higher protein levels ($m_p > 10$), noise levels off, indicating the dominance of extrinsic factors such as global resource availability. The noise in protein expression (h_p^2) is defined as:

$$h_p^2 = \frac{\sigma_p^2}{m_p^2}$$

where σ_p^2 and m_p^2 are the variance and mean of protein abundance, respectively. When incorporating extrinsic factors, the total noise can then be decomposed as:

$$h_p^2 = \frac{\langle b \rangle + \langle b \rangle h_b^2}{m_p} + h_a^2 + h_a^2 h_b^2 + h_b^2$$

where h_a^2 and h_b^2 represent normalized variances for burst frequency (a) and burst size (b), respectively. For mRNA, noise (h_p^2) scales inversely with mean mRNA abundance, reflecting its distinct stochastic dynamics. This detailed noise analysis provides a clearer understanding of how variability in gene expression arises, particularly in bacteria, where low mRNA and protein copy numbers make noise a significant factor affecting the relation between these quantities (Taniguchi et al., 2010).

In 2011, a mathematical model was developed to predict protein abundance from transcriptomics data by integrating experimental measurements of mRNA and protein levels along with their respective half-lives. The model assumes steady state, where synthesis and degradation rates are balanced, leading to simplified equations for mRNA and protein dynamics:

$$\text{mRNA dynamics : } [mRNA] = \frac{v_{sr}}{k_{dr}}$$

$$\text{Protein dynamics : } [protein] = \frac{k_{sp}[mRNA]}{k_{dp}}$$

where v_{sr} is the mRNA synthesis rate, k_{dr} is the mRNA degradation rate constant, k_{sp} is the translation rate constant, k_{dp} is the protein degradation rate constant, $[protein]$ represents the concentration of protein. By substituting the steady-state expression for $[mRNA]$ into the protein equation, the steady-state protein concentration is then given by:

$$[protein] = \frac{v_{sr}k_{sp}}{k_{dr}k_{dp}}$$

This model enables the prediction of mRNA and protein levels under steady-state conditions based on synthesis and degradation rates, providing insights into gene expression dynamics in mammalian cells (Schwanhäusser et al., 2011).

Another study published in 2011 investigated the dynamic relationship between mRNA and protein levels in *S. cerevisiae* during osmotic stress (Lee et al., 2011). The temporal dynamics of protein abundance were modelled using a mass-action kinetic equation, accounting for mRNA-driven synthesis, protein degradation, and dilution due to cell growth and division:

$$\frac{d[P_r]}{dt} = k_t[mRNA_r] - (k_d + \mu)[P_r]$$

where $[P_r]$ is the protein abundance for a given gene r , $[mRNA_r]$ is the mRNA abundance for the same gene, k_t denotes the translation rate, which varies dynamically over time, k_d is the protein degradation rate, assumed to be constant, and μ is the dilution rate due to cell division. This model was used to show that temporary cell-division arrest maintains protein levels, preventing the significant decrease in proteins that would occur with reduced transcripts. For example, without division

arrest, proteins encoded by reduced mRNA would decline by approximately 1.5-fold. Transient mRNA bursts, peaking 30 minutes after stress, were shown to accelerate protein acclimation; their absence delayed protein adjustment by an average of 53 minutes. Testing alternative scenarios, such as constant translation rates or unchanged mRNA abundance, confirmed that transcript reduction contributes minimally compared to translational redistribution and division arrest. Additionally, post-transcriptional regulation was observed to buffer variability, with approximately 24% of proteins exhibiting reduced noise relative to mRNA, ensuring stable production under stress conditions. These findings highlight the dynamic interplay of transcriptional and post-transcriptional processes during cellular adaptation.

An analysis from 2012 addressed how delays in protein synthesis reduce the correlation between mRNA and protein levels in *E. coli* (Gedeon and Bokes, 2012). Using an extended two-stage stochastic model, the study showed that translational and maturation delays align theoretical predictions with experimental observations of low correlation. The mRNA dynamics was described by:

$$M(t) = M(t-1) + k_1 - g_1 M(t)$$

where $M(t)$ denotes mRNA abundance at time t , k_1 stands for the transcription rate, and g_1 denotes mRNA degradation rate. The protein dynamics with delay was modelled by

$$N(t) = N(t-1) + k_2 M(t-d) - g_2 N(t)$$

where $N(t)$ stands for protein abundance at time t , k_2 for translation rate, d for translational delay, and g_2 denotes the protein degradation rate. The key result derived from this mechanistic model is the expression for the correlation coefficient r between mRNA and protein abundances in the delayed model, i.e.

$$r = e^{-g_1 d} r_{const}$$

where r denotes correlation coefficient between mRNA and protein levels in the presence of delay, r_{const} the correlation coefficient in the absence of delay, g_1 stands for mRNA degradation rate (estimated as $\frac{1}{5} \text{ min}^{-1}$) and d for translational delay (estimated as 7.5 minutes). If r_{const} denotes the correlation coefficient in the absence of delay and r the correlation in its presence, then using the mRNA degradation rate $g_1 \approx \frac{1}{5} \text{ min}^{-1}$ and a translational delay $d \approx 7.5 \text{ min}$, we find that when $r_{const} = 0.13$, the observed correlation drops to $r = 0.029$, and when $r_{const} = 0.16$, it decreases to $r = 0.035$. Thus, the correlation between mRNA and protein levels is reduced to 2.9–3.5% due to translational delay. This resolves the apparent paradox of low correlation between mRNA levels and protein abundance in *E. coli* by demonstrating that translational delay is the primary factor reducing correlation, without requiring additional sources of noise. Additionally, the mRNA degradation rate strongly influences correlation by controlling mRNA turnover, while other factors such as transcription rate, translation rate, and protein degradation affect protein abundance but do not significantly alter correlation.

A study published in 2015 integrated experimental data with a computational framework to quantify the contributions of mRNA levels, translation rates, and degradation rates to protein abundance in bone marrow-derived dendritic cells of mouse (*Mus musculus*), both at steady state and in response to stimulation. The authors modelled changes in mRNA and protein levels over time using ordinary differential equations (ODEs), which were fitted to experimental data. By comparing predicted protein levels to observed data using Spearman-corrected coefficients of determination, the study revealed that, before stimulation, mRNA levels

accounted for 59–68% of protein variance, while translation and protein degradation rates contributed 18–26% and 8–22%, respectively. Upon stimulation, mRNA levels explained 87–92% of protein fold changes, demonstrating that transcriptional changes primarily drive dynamic responses, particularly in immune-related proteins. In contrast, protein modules involved in basic cellular functions were more dependent on translation and degradation for regulation (Jovanovic et al., 2015).

Another study from 2015 combined biological insights with mathematical modelling to describe the relationship between mRNA levels, translation, and protein levels in yeast. The authors initially assumed a simplified mathematical model:

$$\frac{dP_i}{dt} = \tau_i M_i - \delta_i P_i$$

where P_i denotes the steady-state protein level of gene i , M_i is the mRNA level of gene i , τ_i stands for the translation rate, and δ_i is the protein degradation rate. At steady state, the equation simplifies to:

$$P_i = \frac{\tau_i}{\delta_i} M_i$$

This suggests that protein levels are proportional to mRNA levels if translation and degradation rates are constant. However, translation rates often vary with mRNA abundance, leading to a nonlinear relationship. To address this issue, the model was modified to include variable translation rates dependent on mRNA levels:

$$P_i = \frac{\tau_i(M_i)}{\delta_i} M_i$$

where $\tau_i(M_i)$ is the translation rate as a function of M_i . This study further incorporated corrections for noise and advanced statistical methods, such as Spearman's correction and structured covariance modelling (SCM), to account for missing data and measurement errors. Regression techniques, like ranged major-axis regression, are also used to reduce biases. Through simulations, the study validated these approaches and demonstrated that mRNA levels explain over 85% of the variation in protein levels, revising earlier models that attributed more variation to post-transcriptional effects (Csárdi et al., 2015).

A 2017 study investigated the relationship between mRNA abundance and protein expression in *S. cerevisiae*, introducing a refined model for quantifying translational control. The steady-state equation for protein expression is given by:

$$prot_i = RNA_i TR_i PnD_i$$

where RNA_i is the mRNA abundance (molecules per cell), TR_i is the translation rate (number of protein molecules per mRNA molecule), PnD_i is the fraction of protein not degraded per cell cycle. Taking the logarithm of both sides:

$$\log_{10}(prot_i) = \log_{10}(RNA_i) + \log_{10}(TR_i) + \log_{10}(PnD_i)$$

The translation rate (TR) is separated into two components:

$$TR_i = TRmD_i \cdot TRmIND_i$$

where $TRmD_i$ (mRNA-dependent translation rate) affects how protein levels scale with mRNA levels and $TRmIND_i$ (mRNA-independent translation rate) captures variation in translation efficiency that is unrelated to mRNA levels. Assuming that the amplification exponent ($b_{prot-RNA}$) describes how protein levels scale with mRNA levels, the above-mentioned equation can be rewritten as:

$$\log_{10}(prot_i) = \log_{10}(a) + b_{prot-RNA} \log_{10}(RNA_i) + \log_{10}(TRmIND_i) + \log_{10}(PnD_i)$$

where a is a constant and $b_{\text{prot-RNA}}$ is the amplification exponent. The relationship between translation rate and mRNA abundance is then given by:

$$b_{\text{prot-RNA}} = 1 + b_{\text{TR-RNA}}$$

where $b_{\text{TR-RNA}}$ is the slope of the log-transformed translation rate in terms of the log-transformed mRNA level. The variance in total translation rate (TR) is split into:

$$\text{var}(\log_{10}(\text{TR})) = \text{var}(\log_{10}(\text{TRmD})) + \text{var}(\log_{10}(\text{TRmIND}))$$

TRmD explains $\sim 20\%$ of translation variance and TRmIND explains $\sim 80\%$ of translation variance. To quantify the contributions of mRNA, TRmIND , and protein degradation, the study applies ordinary least squares (OLS) regression:

$$\log_{10}(\text{prot}) = a + \beta \log_{10}(\text{RNA}) + \gamma \log_{10}(\text{TRmIND}) + \eta \log_{10}(\text{PnD}) + \epsilon$$

where β represents the contribution of mRNA abundance, γ represents the contribution of TRmIND , η represents the contribution of protein degradation, ϵ is the unexplained variance (likely due to measurement error). The study estimates the amplification exponent as $b_{\text{prot-RNA}} = 1.20$ (95% CI: [1.14, 1.26]) and the correlation between mRNA and protein levels as $R^2_{\text{prot-RNA}} \approx 0.94$. The contributions to protein expression variability includes mRNA levels contributing $\sim 94\%$, $\text{TRmIND} \sim 5\%$, protein degradation $\sim 1\%$ of the variance. This mathematical model refined the understanding of translational control by distinguishing mRNA-dependent and independent contributions. It demonstrates that mRNA levels predominantly determine protein expression, while translation plays a secondary fine-tuning role rather than a dominant one. The study challenged past overestimations of the role of translation and strengthened the argument that transcription is the primary driver of protein abundance (Li et al., 2017a).

Later, in 2018, a model based on ODEs was used to describe and classify the dynamic relationship between mRNA and protein levels during *D. melanogaster* embryogenesis. The core equation is given by:

$$\frac{dP(t)}{dt} = \alpha \bullet \text{mRNA}(t) - \lambda \cdot P(t)$$

where $P(t)$ stands for the protein concentration at time t , $\text{mRNA}(t)$ denotes mRNA concentration at time t , α is the translation rate constant (controls protein synthesis rate from mRNA), and λ denotes protein degradation rate constant. This equation models protein concentration as a balance between synthesis, proportional to mRNA levels, and degradation, proportional to protein concentration. To capture different regulatory scenarios, four modelling scenarios were considered. The production model assumes continuous synthesis and degradation, where protein levels closely follow mRNA dynamics. The delayed-production model introduces a fixed delay (Δt) between mRNA production and the onset of protein synthesis, suitable for transitions like the maternal-to-zygotic transition (MZT). The degradation model assumes negligible synthesis ($\alpha = 0$) and focuses solely on protein degradation, explaining maternally deposited or pre-existing proteins. The stationary model assumes neither synthesis ($\alpha = 0$) nor degradation ($\lambda = 0$), resulting in constant protein levels. For each mRNA-protein pair, parameters (α , λ , and Δt) were estimated by fitting the model to experimental time-course data. Based on these fits, mRNA-protein pairs were classified into one of the four regulatory scenarios or marked as “rejected” if none of the models provided a good fit. The majority of pairs (84%) were explained by one of these models, while 16% were identified as involving complex post-transcriptional regulation (Becker et al., 2018).

Although the two-stage and three-stage models discussed so far have been useful, even expanded to include factors such as mRNA processing and maturation, cell division, DNA replication, gene dosage, and growth, they can only be explored through stochastic simulations. The work of Cao and Grima (2020) proposes an analytically tractable stochastic model of mRNA dynamics in eukaryotic cells. In this model, they

include the aforementioned factors as well as growth-dependent transcription (Fig. 2c). To account for gene dosage, two gene copies can independently switch between ON and OFF and replication results in four copies at a certain cell age t . The change in rate in which each gene is switched from OFF to ON is also considered. The maturation of mRNA is represented by production of an unspliced mRNA, denoted by N , which then becomes a mature, spliced mRNA denoted by M under rate k . This spliced mRNA then decays at rate d . Another factor included is growth-dependent transcription, where the rate of transcription ρ is proportional to cell volume V . Lastly, to account for cell division, it is assumed that both types of mRNA can be independently segregated between the new cells, and it follows binomial partitioning, where the probability of a mRNA molecules to end up in which cell follows a binomial distribution. Using experimental data from eukaryotic cells, namely yeast, mouse and human, this study found that variations in mRNA degradation rates and promoter ON-OFF switching to be the most significant factors in explaining variability in gene expression across cell populations.

Although now analytically tractable, two- and three-stage models, even when extended to include cell-cycle effects, mRNA maturation, or multi-step promoter activation, cannot uniquely resolve intrinsic bursting from extrinsic heterogeneity, because any fixed-parameter Telegraph or negative-binomial distribution can be re-expressed as a compound of a “simpler” model with parameter noise. In this scenario, Ham et al. (2021) have demonstrated that population-level mRNA distributions alone inflate apparent burstiness, confounding mechanistic inference. To overcome this issue, they introduce the Noise Decomposition Principle and pathway-reporter method, showing that measuring covariances between successive species in the expression pathway (nascent versus mature RNA, or mRNA versus protein) directly isolates extrinsic noise on transcriptional parameters, without requiring identically regulated dual reporters. By framing extrinsic noise as the normalised covariance of conditional expectations, this approach recovers the true burst frequency and size parameters of the underlying two- or three-state models, even in the presence of complex cell-cycle dynamics. Integrating such pathway-covariance measurements into mechanistic ODE or stochastic frameworks thus provides a principled route to disentangle intrinsic transcriptional kinetics from ensemble variability, thereby refining burst-parameter estimates and enhancing the predictive power of mechanistic gene-expression models.

Building on the delay-driven bursting frameworks discussed above, Jiang et al. (2021) introduce a different mechanistic approximation that preserves non-Markovian memory effects without expanding state-space by training a neural network to learn time-dependent effective propensities for degradation events. In their NN-CME approach, the delay chemical master equation, whose two-time probability kernels render direct analysis intractable, is projected onto a time-inhomogeneous Markovian master equation by parameterizing the removal propensity as a neural-network function, $\text{NN}\theta(n, t)$. Crucially, the network is trained on a modest number of stochastic simulations (or experimental snapshots), such that the learned propensities encode the full delay distribution as dynamic corrections to first-order decay. Once trained, the NN-CME yields analytic or finite-state-projection solutions that match delay-CME distributions across regimes of constitutive expression, bursting, and promoter switching, at a fraction of the computational cost. Moreover, by including kinetic rate constants in the training objective, this method simultaneously infers burst frequency, burst size, and delay parameters directly from data, integrating parameter estimation with model approximation. This hybrid mechanistic-machine-learning framework thereby generalizes the two-state bursting models reviewed above, offering a principled route to derive low-dimensional, yet delay-aware, kinetic descriptions of gene expression.

Further advancing the quantification of transcriptional dynamics, Grima and Esmenjaud (2024) addressed a critical limitation in inferring parameters (synthesis rate α , switching-on rate β , switching-off rate θ)

from single-cell mRNA count distributions using the two-stage model. Their work demonstrated that unaccounted extrinsic noise such as cell-to-cell variability in kinetic parameters introduces systematic biases in estimated values, distorting interpretations of transcriptional bursting. For instance, extrinsic noise in synthesis rates overestimates α , β and θ while underestimating burst size, whereas noise in switching-on rates has the opposite effect. Notably, near a critical noise threshold, inferred parameters diverge to infinity, erroneously suggesting extreme bursting. The authors also established "bias signatures" tied to noise sources and derived a correction method leveraging gene-gene covariances in scRNA-seq data to rescale burst frequencies and sizes. This framework, validated across mammalian transcriptomes, challenges prior estimates (e.g., Larsson et al., 2019) and underscores that traditional snapshot fitting conflates intrinsic kinetics with extrinsic heterogeneity, a fundamental constraint echoed in earlier multi-scale models (Cao and Grima, 2020). By integrating moment-based inference with noise decomposition principles, their approach bridges mechanistic accuracy and population-level variability.

Statistical (regression) models. The limited correlation between mRNA levels and protein abundances is primarily due to the regulation of translation. As a result, directly monitoring translation can offer valuable insights into gene expression by revealing post-transcriptional regulation and protein synthesis mechanisms (Teyssonniere et al., 2024). Ribosome profiling addresses the gap between transcriptomics and proteomics by sequencing ribosome-protected mRNA fragments, enabling the identification of actively translated regions. This technique quantifies translational activity for each mRNA based on read counts, also referred to as translation efficiency, which is defined as the ratio of ribosome density (measured through ribosome profiling) to mRNA levels (measured through mRNAseq) (Ingolia et al., 2009; Johannes et al., 1999). Combining ribosome profiling with techniques such as translating ribosome affinity purification (TRAP) for cell-specific analysis—particularly in complex tissues like the brain—can yield more precise and quantitative measurements of translation, maintaining tissue specificity (Guo et al., 2010; Ingolia, 2014).

In a seminal 2010 study, ribosome profiling and microRNA (miRNA) measurements were used to assess the impact of translation efficiency and mRNA levels on protein production in mammals. Specifically, translation efficiency was calculated by comparing changes in ribosome occupancy (ribosome protected fragments, or RPFs) to changes in mRNA levels. For each gene, the change in ribosome occupancy was normalized by the change in mRNA levels, enabling the isolation of the effect of translation itself (i.e., the number of ribosomes bound to the mRNA) from changes in mRNA levels. The study concluded that miRNAs primarily affect protein production through mRNA destabilization rather than direct inhibition of translation. This destabilization reduces mRNA levels, which accounts for approximately 84% of the observed decrease in protein output, while changes in translation efficiency, such as ribosome occupancy, contribute to a smaller extent. Thus, for most miRNA-targeted genes, the reduction in mRNA levels is the dominant factor influencing protein production, with translational repression playing a much smaller role. In this context, mRNA stability plays a critical role in determining protein production levels (Guo et al., 2010).

A 2014 study examined the impact of codon bias on translation efficiency and regulatory divergence between *S. cerevisiae* and *S. paradoxus*. Using the tRNA adaptation index (tAI), which measures how well codon usage aligns with tRNA availability, the study found that genes with high tAI values used codons efficiently recognized by abundant tRNAs, enhancing translation efficiency. Codon bias was strongly correlated with mRNA abundance, ribosome occupancy, and translation efficiency within both species. However, its inter-species differences were weakly correlated with translation efficiency divergence, suggesting that other factors, such as transcript leader features, may play a bigger role (McManus et al., 2014).

Another 2014 study explored how genetic variation impacts the relationship between protein and mRNA levels in yeast, revealing a

strong correlation between ribosome footprint abundance (protein synthesis proxy) and mRNA levels (Spearman's correlation $\rho = 0.71$, rising to 0.77 for certain genes). Translation efficiency varied but typically reinforced or subtly adjusted mRNA differences without causing major discrepancies. Translational differences modestly amplified protein-level variations but were not the main driver of mRNA-protein mismatches. Both *cis*- and *trans*-regulatory variations similarly influenced mRNA and translation, highlighting their close connection (Albert et al., 2014).

A 2014 study using ribosome profiling on two yeast species (*S. cerevisiae* and *S. paradoxus*) and their hybrid showed that mRNA levels alone cannot fully predict protein abundance due to the crucial role of translational regulation. While *cis*-regulatory divergence in mRNA levels was found in 61% of orthologs, translational divergence affected 35%, with similar magnitudes of change (median log₂ *cis*-ratio: 0.288 for mRNA, 0.325 for translation). Translational regulation often buffered mRNA changes, stabilizing protein production, as seen in 561 genes with opposing effects compared to 256 with reinforcing effects. This buffering caused mRNA levels to overestimate protein divergence by 15%, highlighting the combined role of transcriptional and translational regulation in maintaining protein levels (Artieri and Fraser, 2014).

A 2015 study examined mRNA levels and ribosomal density levels in a study focusing on a hybrid of *S. cerevisiae* and *S. bayanus*. Protein abundance correlated more strongly with ribosomal density ($r = 0.62$) than with mRNA levels ($r = 0.51$), while the Pearson correlation between mRNA and ribosomal density was 0.77. These findings highlight that protein abundance is more closely linked to translation efficiency than to mRNA levels. Therefore, translation serves as a key regulatory step, buffering against transcriptional fluctuations and maintaining stability in protein expression despite variations in mRNA levels. This result underscores the critical need to study gene regulation across multiple levels to fully understand the mechanisms underlying phenotypic stability (Wang et al., 2015).

A 2018 study examined the impact of codon usage bias on translation efficiency and its effect on the relationship between protein abundance and mRNA levels. Synonymous codons, which encode the same amino acid, are used with varying frequencies, with optimal codons matching abundant tRNAs to enable efficient translation. In contrast, rare codons corresponding to scarce tRNAs can slow translation, reducing protein output even when mRNA levels are high. Translation efficiency, defined as the ratio of protein abundance to mRNA abundance for a given gene, reflects how effectively ribosomes synthesize proteins from available mRNA transcripts and depends on the alignment between codons and available tRNAs. The study found that highly expressed genes tend to use optimal codons, ensuring high protein production and a strong correlation between mRNA levels and protein abundance, whereas genes with non-optimal codons exhibit a weaker correlation due to slower translation or ribosome stalling (Frumkin et al., 2018).

Ribosome profiling (Ribo-seq) has been used to investigate the role of post-transcriptional buffering in gene expression in *S. cerevisiae* under severe oxidative stress. In control conditions, the Spearman correlation between gene expression, measured by RNA-seq, and protein abundance was 0.46, while the correlation between Ribo-seq (ribosome-protected RNA fragments) and protein abundance was significantly higher, at 0.71. Under oxidative stress, the RNA-seq and proteomics correlation improved to 0.62, whereas the Ribo-seq and proteomics correlation slightly decreased to 0.67. These findings highlighted the superior predictive power of Ribo-seq in estimating protein abundance compared to RNA-seq. They also observed that changes in mRNA levels are counterbalanced by opposite changes in ribosome density. This mechanism prevents significant alterations in protein levels and underscores the prevalence and importance of post-transcriptional regulation, which may be more widespread than previously understood (Blevins et al., 2019).

In a subsequent study conducted in 2020, Spearman correlation

coefficient (ρ) was used to assess the relationships among the mRNA levels, ribosome profiling data, quantifying ribosome-protected mRNA fragments to estimate protein synthesis rates, and protein abundances across multiple mammalian species and organs. The analysis revealed significantly stronger correlations between ribosome profiling and proteomics data than between transcriptomics and proteomics data in all three organs, namely, brain, liver, and testis. This highlights the crucial role of translational regulation as a more accurate predictor of protein abundance than mRNA levels alone (Wang et al., 2020).

Similarly, a 2021 study highlighted a substantial disconnect between transcription (mRNA levels) and translation (measured as the number of ribosome-protected fragments (RPF)) across various tissues and developmental stages in mice, showing that mRNA changes do not always correlate with protein production. This study highlighted the complexity of translational control and the critical role of post-transcriptional mechanisms in tissue-specific functions and developmental transitions. The study specifically focused on alternative splicing as an important translational regulation mechanism, demonstrating how it generates mRNA isoforms with varying translational efficiencies, leading to tissue- and stage-specific expression patterns and impacting the mRNA-protein relationship (Wang et al., 2021).

In a recent study, ribosome profiling and RNA sequencing were performed to investigate the transcriptome and translation dynamics in eight genetically diverse *S. cerevisiae* natural isolates from various ecological environments. The study found a moderate to strong Spearman correlation, ranging from 0.52 to 0.80, between RNA-seq and Ribo-seq data across samples, indicating a clear relationship between transcription and translation levels. However, these findings suggested that additional regulatory mechanisms, such as post-transcriptional buffering, modulate this correlation. Rather than focusing on a direct measurement of correlation between mRNA levels and protein abundance, the study highlights the broader patterns of buffering that help maintain protein levels despite transcriptional variability (Teyssonniere et al., 2024).

While translation has received the most attention as a complementary factor to transcription in protein prediction, several other key factors also play a significant role. These include protein degradation, dilution during differentiation, and various mRNA features—such as 5' UTR secondary structures, flanking nucleotides, open reading frame (ORF) length, and codon usage bias (CUB). A 2010 study found a moderate correlation between mRNA levels and protein abundance, quantified by the Spearman's rank correlation $\rho = 0.46$, concluding that post-transcriptional regulation, including mRNA sequence length, amino acid properties, upstream open reading frames (uORFs), and secondary structures in the 5' untranslated region (UTR), play a significant role in protein abundance variation in human cells. Consequently, a non-parametric regression technique, called multivariate adaptive regression splines (MARS), was used to integrate various sequence features to predict steady-state protein concentrations in human cells. This analysis revealed that while mRNA levels alone explained ~ 30% of protein variation, additional sequence-based factors accounted for another 30 – 40%, with mRNA sequence length (~ 20%), translation-related features (~ 9%), and protein degradation signals (~ 8%) as the strongest contributors. Collectively, these factors explained ~ 67% of total protein abundance variation, while the remaining ~ 33% remained unexplained, likely due to unaccounted regulatory mechanisms, cell cycle effects, and measurement noise (Vogel et al., 2010).

A 2013 study utilized partial least squares (PLS) regression to model the contributions of transcription, protein synthesis, and protein degradation to protein expression changes during differentiation. The weak correlation often observed between mRNA and protein levels was clarified by incorporating synthesis and degradation rates into the model. The study revealed that synthesis rates accounted for 41% and degradation rates for 13% of the variance in protein expression changes not explained by transcription. The PLS decomposition of predictors and response is described as:

$$X = TP^T + E$$

$$y = Tq + f$$

where y denotes the change in protein expression, X is a matrix of transcriptional changes, synthesis rates, and degradation rates, T stands for latent variables combining predictors to best explain protein expression changes, P^T is a loading matrix, linking predictors to latent variables, E is the residual matrix for X (unexplained variation), q denotes the regression vector for y , showing influence of T , and f stands for the residual of y , indicating unexplained variation. This approach significantly improved the prediction of protein abundance dynamics (Kristensen et al., 2013).

A similar analysis of over 800 genes in *E. coli* challenges the notion that translation initiation is the primary bottleneck by employing PLS regression on more than 100 mRNA sequence features. The study presents a predictive model that explains 66% of protein abundance variation, identifying mRNA transcript level as the dominant factor, accounting for 53% of the variation. Translation elongation efficiency contributes 12%, with codon usage and amino acid composition playing a crucial role, favouring codons translated by abundant tRNAs. In contrast, translation initiation factors account for only 1% (Guimaraes et al., 2014).

In 2016, a linear regression model was developed to describe the relationship between transcript abundance and protein abundance in yeast. The model incorporates seven features related to translation and post-transcriptional regulation to better predict protein levels based on mRNA data. These features include transcript abundance, the translation adaptation index (tAI) (which measures codon bias), RNA-binding protein interaction, protein degradation, transcript secondary structure (PARS score), poly-A tail length, and protein turnover. The final model, combining these features, explained about 70% of the variance in protein abundance across the quantified proteins. Three of the seven features along with transcript abundance contributed to improving the prediction accuracy. The predictive strength of the model was evaluated using a Spearman rank correlation of 0.83 between observed and predicted protein abundances (Lawless et al., 2016).

A 2019 study developed a multivariate linear regression model to predict the protein-to-mRNA ratio (PTR) from sequence features. This model integrates various mRNA and protein sequence elements known to influence translation initiation, elongation, termination, and protein stability. It includes features such as codon usage, mRNA sequence motifs (in the 5' UTR, coding sequence, and 3' UTR), protein sequence features, GC content, mRNA half-life, and protein degradation signals. The multivariate linear tissue-specific PTR model used in the study is represented by the following equation:

$$y_{ij} = \beta_{0j} + x_i^T \beta_j + \epsilon_{ij}$$

where y_{ij} is the tissue-specific protein-to-mRNA ratio ($\log_{10}$) for gene i in tissue j , x_i^T is the row vector of sequence feature predictors for gene i , β_j is the vector of tissue-specific coefficients, β_{0j} is the intercept for tissue j , and ϵ_{ij} is the error term. The model was evaluated using data from 29 human tissues and 11,575 genes, highlighting substantial variation in PTR ratios across different tissues (Wang et al., 2019). Initially, their LR model accounted for an average of 22% of the variance based solely on sequence information. By integrating additional experimentally validated interactions and modifications—including mRNA methylation, miRNA and RBP binding sites, and post-translational modifications—the model achieved a median precision improvement of 3.2-fold in predicting PTR ratios. Their results further validated previously identified mRNA regulatory elements as key contributors to PTR ratio variability (Eraslan et al., 2019).

5. Future perspectives

Quantifying protein abundance is crucial for understanding the ultimate effects of gene expression, as proteins are the functional products of genes and directly influence cellular behaviour and biotechnological applications. While mRNA levels provide a snapshot of transcription, protein levels ultimately determine biological function. Precise control of protein abundance is essential for optimizing and controlling the function of genetic circuits and metabolic pathways, as performed by methods from synthetic biology and metabolic engineering. As a result, by understanding factors that contribute to the accurate prediction of protein abundance, synthetic biologists can enhance bio-production efficiency, reduce metabolic burden, and improve the reliability of engineered biological systems for applications in biotechnology, medicine, and environmental sustainability (Fig. 3).

The past two decades have seen significant advancements in high-throughput transcriptomics and proteomics technologies, enabling large-scale quantification of gene expression and protein abundance across diverse biological systems. Despite these advancements, the fundamental question of how closely gene expression (mRNA levels) predicts protein abundance remains only partially resolved. A broad spectrum of studies has established that while transcription is a key determinant of protein levels, post-transcriptional, translational, and degradation mechanisms introduce variability, rendering the accurate prediction of protein abundance as a challenging problem (Liu et al., 2016).

For instance, existing evidence has demonstrated that correlations between steady-state mRNA and protein levels are moderate, often ranging from 0.4 to 0.6 in bulk-level studies; these values can go up to ~ 0.85 by accounting for noise measurement errors (Csárdi et al., 2015). The correlation weakens significantly at the single-cell level, where stochastic fluctuations, transcriptional noise, and cell cycle effects introduce further complexity.

Gene-specific variability in mRNA-protein correlations Given that there is no universally high correlation between mRNA and protein levels across all genes, categorizing genes based on function, expression dynamics, and regulatory mechanisms can provide more nuanced insights into factors that shape these correlations. Existing evidence shows that the correlation between mRNA and protein levels can vary depending on gene function, sequence conservation, and untranslated region (UTR) lengths. For instance, studies on human CD8⁺ T cells have revealed that while overall mRNA expression is a poor predictor of protein output, function-specific correlations can also be identified (Nicolet and Wolkers, 2022). These findings suggest that certain gene classes may exhibit more predictable mRNA-protein relationships due to shared regulatory mechanisms or structural features. Here, techniques based on biclustering of correlations and functional similarity could be useful to make systematic investigations of the existing data sets. Furthermore, differentially expressed mRNAs often show a stronger

correlation with their corresponding protein products than non-differentially expressed mRNAs (Koussounadis et al., 2015). This suggests that genes undergoing active expression changes may exhibit tighter coupling between mRNA and protein abundances, reinforcing the need for predictive models focusing on specific classes of genes to improve the accuracy of predicted protein levels.

Machine learning approaches and large language models. In multiple organisms, steady-state protein levels are primarily determined by mRNA levels. However, factors such as spatial and temporal variations, local resource availability for protein biosynthesis, and protein stability and degradation rates add complexity to the relationship between levels of protein and the corresponding transcripts (Liu et al., 2016). Given that in many cases, transcript levels alone are insufficient to predict protein abundance, incorporating mRNA sequence features into machine learning models (e.g., non-linear regressions) represents a promising approach to improving protein abundance predictions (Zrimec et al., 2021). For instance, several regulatory and coding regions have already been integrated into machine and deep learning models for protein abundance prediction. These include codon usage (Ferreira et al., 2021; Trösemeier et al., 2019), transcription factor binding sites (de Boer et al., 2020), Shine-Dalgarno sequences in prokaryotes (Bonde et al., 2016; Salis et al., 2009), and Kozak sequences in eukaryotes (Li et al., 2017b) and DNA sequence (Ding et al., 2018; Fu et al., 2020; Kotopka and Smolke, 2020). However, these opportunities are not fully explored, particularly with respect to the consideration of natural variability in protein abundance and structural variants in genomes of representatives (e.g., strains and accessions) of a species with advanced machine learning and deep learning approaches.

The advent of large language models (LLM) trained on sequence data have provided new means to analyse, predict, and design protein properties. Amongst those with the task of analysing the relationship between RNA and protein levels as well as predicting the latter, Stefanini et al. (2023) adapted the Perceiver IO (Andrew Jaegle et al., 2021) model to predict both mRNA and protein levels from sequence data alone. Testing on glioblastoma and lung cancer tissues, the R^2 score for the predictions ranged from 0.026 to 0.161. The study of Carlos Outeiral (2024) went in another direction, leveraging codon usage and nucleotide sequences instead of amino acid sequence to train the LLM. They developed a LLM named CaLM and compared its performance to LLMs trained on amino acid sequences, such as ESM, ProtTrans and BERT. They report that CaLM outperformed all other LLMs on predicting protein abundance, achieving a Pearson correlation of over 0.5 for *S. cerevisiae*.

Ribosome competition and translational resource allocation A crucial factor affecting protein abundance predictability is the competition for cellular translational resources. Historically, many studies have examined the relationship between the mRNA levels of a single gene and abundance of the corresponding protein. However, gene expression and protein synthesis occur in a competitive environment,

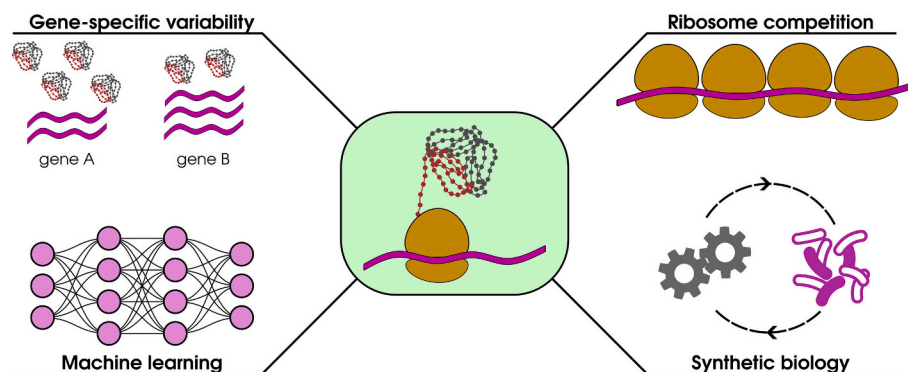


Fig. 3. Future perspectives on the study of mRNA and protein abundance correlations. The presented perspectives could help shedding light on the complex relationship between mRNA and protein levels.

where limited translational resources such as ribosomes, tRNAs, and translation factors must be shared among multiple genes. For instance, Mather et al. (2013) made use of ideas from queuing theory to describe how different mRNAs compete for limited translational resources. Namely, when a particular transcript is significantly upregulated, the total number of transcripts may increase, leading to competition among mRNAs for translation by ribosomes. This competition introduces translational crosstalk, where the expression of one protein is influenced not only by its own mRNA abundance but also by the expression levels of other genes in the network. To fully capture the complexity of gene expression, future research making use of machine and deep learning approaches must consider protein abundance as a function of all expressed genes rather than analysing gene-protein relationships *per gene* basis. This requires integrating translational resource availability into predictive models and accounting for ribosome competition and binding kinetics.

Synthetic biology and gene expression regulation. Understanding the factors that regulate protein abundance is especially important for biotechnological and synthetic biology applications. For instance, it has been shown that the integration of protein abundance data into protein-constrained metabolic models can improve the accuracy of predicted physiological traits under different environmental and genetic perturbations (de Moura Ferreira et al., 2023). For instance, knowledge of how the expression of transcription factors affects the expression of downstream target genes, and how their expression determines the abundance of the corresponding enzymes can be used in the design of precise metabolic engineering strategies. This approach would require the integration of gene regulatory network models with models of metabolism, which has already been attempted based on the large-scale gene expression compendia (Chandrasekaran and Price, 2010). Note that this research direction can benefit from having paired transcriptomics and proteomics data to make better use of the integrated models in the design of metabolic engineering strategies that do not only include enzyme-coding genes, but also transcription factors.

6. Conclusion

Regulation of gene expression is one of the key factors to the prediction of physiological and metabolic traits. We argue that this can be achieved by making use of accurate machine or deep learning models for protein abundance based on gene expression and gene sequence features in the context of protein-constrained metabolic models. The systems-level interplay between mRNA competition, ribosome availability, and protein stability represents an essential but under-explored aspect of gene expression. As a result, we advocate that future work should shift from single-gene analyses to holistic models that account for the entire gene regulatory networks, considering both transcriptional regulation and resource-mediated translational constraints. As innovative technologies and computational models evolve, we can move closer to a quantitative, predictive understanding of how gene expression dictates protein levels in diverse biological systems and experimental settings.

Funding sources

We acknowledge the ALFAFUELS project funded by the European Union under Grant Agreement Number 191122224 (to Z.N.) and REG-UMET project funded by the NovoNordisk Foundation under Project Number 23OC0085412 (to Z.N.).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.biotechadv.2025.108720>.

References

- Aissa, A.F., Islam, A.B., Ariss, M.M., Go, C.C., Rader, A.E., Conrardy, R.D., Gajda, A.M., Rubio-Perez, C., Valyi-Nagy, K., Pasquinelli, M., et al., 2021. Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nat. Commun.* 12, 1628.
- Albert, F.W., Muzzey, D., Weissman, J.S., Kruglyak, L., 2014. Genetic influences on translation in yeast. *PLoS Genet.* 10, e1004692.
- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., 2002. *Molecular motors*. In: *Molecular Biology of the Cell*, 4th Edition. Garland Science.
- Anderson, L., Seilhamer, J., 1997. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* 18, 533–537.
- Andrew Jaegle, J.C., Borgeaud, Sebastian, Alayrac, Jean-Baptiste, Doersch, Carl, Ionescu, Catalin, Ding, David, Koppula, Skanda, Zoran, Daniel, Brock, Andrew, Shelhamer, Evan, Hénaff, Olivier, Botvinick, Matthew M., Zisserman, Andrew, Vinyals, Oriol, 2021. Perceiver IO: A General Architecture for Structured Inputs & Outputs. *arXiv*. <https://doi.org/10.48550/arXiv.2107.14795>.
- Artieri, C.G., Fraser, H.B., 2014. Evolution at two levels of gene expression in yeast. *Genome Res.* 24, 411–421.
- Athanasiadou, R., Neymotin, B., Brandt, N., Wang, W., Christiaen, L., Gresham, D., Tranchina, D., 2019. A complete statistical model for calibration of RNA-seq counts using external spike-ins and maximum likelihood theory. *PLoS Comput. Biol.* 15, e1006794.
- Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K., Gilad, Y., 2015. Impact of regulatory variation from RNA to protein. *Science* 347, 664–667.
- Becker, K., Bluhm, A., Casas-Vila, N., Dinges, N., Dejung, M., Sayols, S., Kreutz, C., Roignant, J.-Y., Butter, F., Legewie, S., 2018. Quantifying post-transcriptional regulation in the development of *Drosophila melanogaster*. *Nat. Commun.* 9, 4970.
- Bennett, H.M., Stephenson, W., Rose, C.M., Darmanis, S., 2023. Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. *Nat. Methods* 20, 363–374.
- Bentley, D.L., 2014. Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.* 15, 163–175.
- Bernstein, J.A., Khodursky, A.B., Lin, P.-H., Lin-Chao, S., Cohen, S.N., 2002. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl. Acad. Sci.* 99, 9697–9702.
- Besse, F., Ephrussi, A., 2008. Translational control of localized mRNAs: restricting protein synthesis in space and time. *Nat. Rev. Mol. Cell Biol.* 9, 971–980.
- Beyer, A., Hollunder, J., Nasheuer, H.-P., Wilhelm, T., 2004. Post-transcriptional expression regulation in the yeast *Saccharomyces cerevisiae* on a genomic scale. *Mol. Cell. Proteomics* 3, 1083–1092.
- Blake, W.J., Kærn, M., Cantor, C.R., Collins, J.J., 2003. Noise in eukaryotic gene expression. *Nature* 422, 633–637.
- Blevins, W.R., Tavella, T., Moro, S.G., Blasco-Moreno, B., Closa-Mosquera, A., Díez, J., Carey, L.B., Albà, M.M., 2019. Extensive post-transcriptional buffering of gene expression in the response to severe oxidative stress in baker's yeast. *Sci. Rep.* 9, 11005.
- Bonde, M.T., Pedersen, M., Klausen, M.S., Jensen, S.I., Wulff, T., Harrison, S., Nielsen, A. T., Herrgård, M.J., Sommer, M.O., 2016. Predictable tuning of protein expression in bacteria. *Nat. Methods* 13, 233–236.
- Bonnet, D., Dick, J.E., 1997. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat. Med.* 3, 730–737.
- Brockmann, R., Beyer, A., Heinisch, J.J., Wilhelm, T., 2007. Posttranscriptional expression regulation: what determines translation rates? *PLoS Comput. Biol.* 3, e57.
- Brown, P.O., Botstein, D., 1999. Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* 21, 33–37.
- Brunner, A.-D., Thielert, M., Vasilopoulou, C., Ammar, C., Coscia, F., Mund, A., Hoerning, O.B., Bache, N., Apalategui, A., Lubeck, M., et al., 2022. Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. *Mol. Syst. Biol.* 18, e10798.
- Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., Stegle, O., 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* 33, 155–160.
- Burnette, W.N., 1981. "Western blotting": electrophoretic transfer of proteins from sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. *Anal. Biochem.* 112, 195–203.
- Cai, L., Friedman, N., Xie, X.S., 2006. Stochastic protein expression in individual cells at the single molecule level. *Nature* 440, 358–362.
- Cao, Z., Grima, R., 2020. Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proc. Natl. Acad. Sci.* 117, 4682–4692. <https://doi.org/10.1073/pnas.1910888117>.

- Carlos Outeiral, C.M.D., 2024. Codon language embeddings provide strong signals for use in protein engineering. *Nat. Mach. Intelligence*. <https://doi.org/10.1038/s42256-024-00791-0>.
- Cenik, C., Cenik, E.S., Byeon, G.W., Grubert, F., Candille, S.I., Spacek, D., Allsallakh, B., Tilgner, H., Araya, C.L., Tang, H., et al., 2015. Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. *Genome Res.* 25, 1610–1621.
- Chai, L.E., Loh, S.K., Low, S.T., Mohamad, M.S., Deris, S., Zakaria, Z., 2014. A review on the computational approaches for gene regulatory network construction. *Comput. Biol. Med.* 48, 55–65. <https://doi.org/10.1016/j.combiomed.2014.02.011>.
- Chandrasekaran, S., Price, N.D., 2010. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA* 107, 17845–17850. <https://doi.org/10.1073/pnas.1005139107>.
- Chen, G., Gharib, T.G., Huang, C.-C., Taylor, J.M., Misek, D.E., Kardia, S.L., Giordano, T.J., Iannettoni, M.D., Orringer, M.B., Hanash, S.M., et al., 2002. Discordant protein and mRNA expression in lung adenocarcinomas. *Mol. Cell. Proteomics* 1, 304–313.
- Chen, J., McSwiggen, D., Únal, E., 2018. Single molecule fluorescence in situ hybridization (smFISH) analysis in budding yeast vegetative growth and meiosis. *J. Visualized Exp. JoVE* 57774.
- Csárdi, G., Franks, A., Choi, D.S., Airolidi, E.M., Drummond, D.A., 2015. Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *PLoS Genet.* 11, e1005206.
- Darmanis, S., Gallant, C.J., Marinescu, V.D., Niklasson, M., Segerman, A., Flamourakis, G., Fredriksson, S., Assarsson, E., Lundberg, M., Nelander, S., et al., 2016. Simultaneous multiplexed measurement of RNA and proteins in single cells. *Cell Rep.* 14, 380–389.
- de Boer, C.G., Vaishnav, E.D., Sadeh, R., Abeyta, E.L., Friedman, N., Regev, A., 2020. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.* 38, 56–65. <https://doi.org/10.1038/s41587-019-0315-8>.
- de Moura Ferreira, M.A., Wendering, P., Arend, M., Batista da Silveira, W., Nikoloski, Z., 2023. Accurate prediction of *in vivo* protein abundances by coupling constraint-based modelling and machine learning. *Metab. Eng.* 80, 184–192. <https://doi.org/10.1016/j.ymben.2023.09.014>.
- Deamer, D., Akeson, M., Branton, D., 2016. Three decades of nanopore sequencing. *Nat. Biotechnol.* 34, 518–524.
- Ding, W., Cheng, J., Guo, D., Mao, L., Li, J., Lu, L., Zhang, Y., Yang, J., Jiang, H., 2018. Engineering the 5' UTR-Mediated Regulation of Protein Abundance in Yeast Using Nucleotide Sequence Activity Relationships. *ACS Synth. Biol.* 7, 2709–2714. https://doi.org/10.1021/ACSSYNBIO.8B00127/ASSET/IMAGES/LARGE/SB-2018-00127_0004.JPEG.
- Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., Zettl, M., Coleman, P., 1992. Analysis of gene expression in single live neurons. *Proc. Natl. Acad. Sci.* 89, 3010–3014.
- Edfors, F., Danielsson, F., Hallström, B.M., Käll, L., Lundberg, E., Pontén, F., Forsström, B., Uhlen, M., 2016. Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol. Syst. Biol.* 12, 883.
- Engvall, E., Perlmann, P., 1972. Enzyme-linked immunosorbent assay, ELISA: III. Quantitation of specific antibodies by enzyme-labeled anti-immunoglobulin in antigen-coated tubes. *J. Immunol.* 109, 129–135.
- Eraslan, G., Avsec, Ž., Gagneur, J., Theis, F.J., 2019. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* 20, 389–403. <https://doi.org/10.1038/s41576-019-0122-6>.
- Ferreira, M., Ventrone, R., Almeida, E., Silveira, S., Silveira, W., 2021. Protein Abundance Prediction Through Machine Learning Methods. *J. Mol. Biol.* 433, 167267. <https://doi.org/10.1016/j.jmb.2021.167267>.
- Flynn, E., Almonte-Loya, A., Fragiadakis, G.K., 2023. Single-Cell Multiomics. *Annual Rev. Biomed. Data Sci.* 6, 313–337. <https://doi.org/10.1146/annurev-biodatasci-020422-050645>.
- Fortelny, N., Overall, C.M., Pavlidis, P., Freue, G.V.C., 2017. Can we predict protein from mRNA levels? *Nature* 547, E19–E20. <https://doi.org/10.1038/nature22293>.
- Fournier, M.L., Paulson, A., Pavelka, N., Mosley, A.L., Gaudenz, K., Bradford, W.D., Glynn, E., Li, H., Sardi, M.E., Fleharty, B., et al., 2010. Delayed correlation of mRNA and protein expression in rapamycin-treated cells and a role for Ggc1 in cellular sensitivity to rapamycin. *Mol. Cell. Proteomics* 9, 271–284.
- Franks, A., Airolidi, E., Slavov, N., 2017. Post-transcriptional regulation across human tissues. *PLoS Comput. Biol.* 13, e1005535.
- Fredriksson, S., Gullberg, M., Jarvius, J., Olsson, C., Pietras, K., Gústafsdóttir, S.M., Östman, A., Landegren, U., 2002. Protein detection using proximity-dependent DNA ligation assays. *Nat. Biotechnol.* 20, 473–477.
- Frei, Andreas P., Bava, Felice-Alessio, Zunder, Eli R., Hsieh, Elena W.Y., Chen, Shih-Yu, Nolan, Garry P., 2016. Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat. Methods* 13. <https://doi.org/10.1038/nmeth.3742>.
- Friedman, N., Cai, L., Xie, X.S., 2006. Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression. *Phys. Rev. Lett.* 97, 168302.
- Frumkin, I., Lajoie, M.J., Gregg, C.J., Hornung, G., Church, G.M., Pilpel, Y., 2018. Codon usage of highly expressed genes affects proteome-wide translation efficiency 1–10. <https://doi.org/10.1073/pnas.1719375115>.
- Fu, H., Liang, Y., Zhong, X., Pan, Z.L., Huang, L., Zhang, H.L., Xu, Y., Zhou, W., Liu, Z., 2020. Codon optimization with deep learning to enhance protein expression. *Sci. Rep.* 10, 1–9. <https://doi.org/10.1038/s41598-020-74091-z>.
- Fulcher, J.M., Markillie, L.M., Mitchell, H.D., Williams, S.M., Engbrecht, K.M., Degnan, D.J., Bramer, L.M., Moore, R.J., Chrisler, W.B., Cantlon-Bruce, J., Bagnoli, J.W., Qian, W.-J., Seth, A., Pasa-Tolić, L., Zhu, Y., 2024. Parallel measurement of transcriptomes and proteomes from same single cells using nanodroplet splitting. *Nat. Commun.* 15, 10614. <https://doi.org/10.1038/s41467-024-54099-z>.
- Gautier, E.-F., Ducamp, S., Leduc, M., Salnot, V., Guillonnet, F., Dussiot, M., Hale, J., Giarratana, M.-C., Raimbault, A., Douay, L., et al., 2016. Comprehensive proteomic analysis of human erythropoiesis. *Cell Rep.* 16, 1470–1484.
- Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K.L., Streets, A., Yosef, N., 2021. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods* 18, 272–282. <https://doi.org/10.1038/s41592-020-01050-x>.
- Gebauer, F., Hentze, M.W., 2004. Molecular mechanisms of translational control. *Nat. Rev. Mol. Cell Biol.* 5, 827–835.
- Gedeon, T., Bokes, P., 2012. Delayed protein synthesis reduces the correlation between mRNA and protein fluctuations. *Biophys. J.* 103, 377–385.
- Genshaft, Alex S., Li, Shuqiang, Gallant, Caroline J., Darmanis, Spyros, Prakadan, Sanjay M., Ziegler, Carly G.K., Lundberg, Martin, Fredriksson, Simon, Hong, Joyce, Regev, Aviv, Livak, Kenneth J., Landegren, Ulf, 2016. Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction. *Genome Biol.* 17. <https://doi.org/10.1186/s13059-016-1045-6>.
- Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W., Gygi, S.P., 2003. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci.* 100, 6940–6945.
- Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., Weissman, J.S., 2003. Global analysis of protein expression in yeast. *Nature* 425, 737–741.
- Gillet, L.C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., Aebersold, R., 2012. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics* 11.
- Grima, R., Esmeijaud, P.-M., 2024. Quantifying and correcting bias in transcriptional parameter inference from single-cell data. *Biophys. J.* 123, 4–30. <https://doi.org/10.1016/j.bpj.2023.10.021>.
- Guimaraes, J.C., Rocha, M., Arkin, A.P., 2014. Transcript level and sequence determinants of protein abundance and noise in *Escherichia coli*. *Nucleic Acids Res.* 42, 4791–4799.
- Guo, H., Ingolia, N.T., Weissman, J.S., Bartel, D.P., 2010. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 466, 835–840.
- Ham, L., Jackson, M., Stumpf, M.P., 2021. Pathway dynamics can delineate the sources of transcriptional noise in gene expression. *eLife* 10, e69324. doi: <https://doi.org/10.7554/eLife.69324>.
- Hashimshony, T., Wagner, F., Sher, N., Yanai, I., 2012. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* 2, 666–673.
- Hentze, M.W., Castello, A., Schwarzl, T., Preiss, T., 2018. A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.* 19, 327–341.
- Hershey, J.W., Sonenberg, N., Mathews, M.B., 2019. Principles of translational control. *Cold Spring Harb. Perspect. Biol.* 11, a032607.
- Hrdlickova, R., Toloue, M., Tian, B., 2017. RNA-Seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA* 8, e1364.
- Ingolia, N.T., 2014. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* 15, 205–213. <https://doi.org/10.1038/nrg3645>.
- Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., Weissman, J.S., 2009. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* 324, 218–223.
- Iost, I., Dreyfus, M., 1995. The stability of *Escherichia coli* lacZ mRNA depends upon the simultaneity of its synthesis and translation. *EMBO J.* 14, 3252–3261. <https://doi.org/10.1002/j.1460-2075.1995.tb07328.x>.
- Irastortza-Oláziregi, M., Amster-Choder, O., 2021. Coupled Transcription-Translation in Prokaryotes: An Old Couple With New Surprises. *Front. Microbiol.* 11.
- Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., Linnarsson, S., 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* 21, 1160–1167.
- Jan, P., Gerlach, K.W.M., van Buggenum, Jessie A.G., Tanis, Sabine E.J., Hogeweg, Mark, Heuts, Branco M.H., Muraro, Mauro J., Elze, Lisa, Rivello, Francesca, Rakszewska, Agata, van Oudenaarden, Alexander, Huck, Wilhelm T.S., Stunnenberg, Hendrik G., 2019. Combined quantification of intracellular (phospho-) proteins and transcriptomics from fixed single cells. *Sci. Rep.* 9. <https://doi.org/10.1038/s41598-018-37977-7>.
- Jiang, Q., Fu, X., Yan, S., Li, R., Du, W., Cao, Z., Qian, F., Grima, R., 2021. Neural network aided approximation and parameter inference of non-Markovian models of gene expression. *Nat. Commun.* 12, 2618. <https://doi.org/10.1038/s41467-021-22919-1>.
- Johannes, G., Carter, M.S., Eisen, M.B., Brown, P.O., Sarnow, P., 1999. Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray. *Proc. Natl. Acad. Sci.* 96, 13118–13123.
- Jovanovic, M., Rooney, M.S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E.H., Fields, A.P., Schwartz, S., Raychowdhury, R., et al., 2015. Dynamic profiling of the protein life cycle in response to pathogens. *Science* 347, 1259038.
- Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wahlby, C., Nilsson, M., 2013. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* 10, 857–860.
- Khan, Z., Ford, M.J., Cusanovich, D.A., Mitran, A., Pritchard, J.K., Gilad, Y., 2013. Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* 342, 1100–1104.
- Kidd, D., Liu, Y., Cravatt, B.F., 2001. Profiling serine hydrolase activities in complex proteomes. *Biochemistry* 40, 4005–4015.

- Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaekady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al., 2014. A draft map of the human proteome. *Nature* 509, 575–581.
- Kotopka, B.J., Smolke, C.D., 2020. Model-driven generation of artificial yeast promoters. *Nat. Commun.* 11, 2113.
- Koussounadis, A., Langdon, S.P., Um, I.H., Harrison, D.J., Smith, V.A., 2015. Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. *Sci. Rep.* 5, 10775.
- Kristensen, A.R., Gsponer, J., Foster, L.J., 2013. Protein synthesis rate is the predominant regulator of protein expression during differentiation. *Mol. Syst. Biol.* 9, 689.
- Lackner, D.H., Schmidt, M.W., Wu, S., Wolf, D.A., Bähler, J., 2012. Regulation of transcriptome, translation, and proteome in response to environmental stress in fission yeast. *Genome Biol.* 13, 1–14.
- Lahtvee, P.J., Sánchez, B.J., Smialowska, A., Kasvandik, S., Elsemman, I.E., Gatto, F., Nielsen, J., 2017. Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast. *Cell Systems* 4, 495–504.e5. <https://doi.org/10.1016/j.cels.2017.03.003>.
- Lambole, B., Audinat, E., Bochet, P., Crépel, F., Rossier, J., 1992. AMPA receptor subunits expressed by single Purkinje cells. *Neuron* 9, 247–258.
- Laosuntisuk, K., Vennapusa, A., Somayanda, I.M., Leman, A.R., Jagadish, S.K., Doherty, C.J., 2024. A normalization method that controls for total RNA abundance affects the identification of differentially expressed genes, revealing bias toward morning-expressed responses. *Plant J.* 118, 1241–1257.
- Larsson, A.J.M., Johnsson, P., Hagemann-Jensen, M., et al., 2019. Genomic encoding of transcriptional burst kinetics. *Nature* 565, 251–254. <https://doi.org/10.1038/s41586-018-0836-1>.
- Laurent, J.M., Vogel, C., Kwon, T., Craig, S.A., Boutz, D.R., Huse, H.K., Nozue, K., Wallia, H., Whiteley, M., Ronald, P.C., Marcotte, E.M., 2010. Protein abundances are more conserved than mRNA abundances across diverse taxa. *Proteomics* 10, 4209–4212. <https://doi.org/10.1002/ps.1000327>.
- Lawless, C., Holman, S.W., Brownridge, P., Lanthaler, K., Harman, V.M., Watkins, R., Hammond, D.E., Miller, R.L., Sims, P.F.G., Grant, C.M., Eysers, C.E., Beynon, R.J., Hubbard, S.J., 2016. Direct and absolute quantification of over 1800 yeast proteins via selected reaction monitoring. *Mol. Cell. Proteomics* 15, 1309–1322. <https://doi.org/10.1074/mcp.M115.054288>.
- Lee, M.V., Topper, S.E., Hubler, S.L., Hose, J., Wenger, C.D., Coon, J.J., Gasch, A.P., 2011. A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Mol. Syst. Biol.* 7, 1–12. <https://doi.org/10.1038/msb.2011.48>.
- Li, J.J., Biggin, M.D., 2015. Statistics requantitates the central dogma. *Science* 347, 1066–1067.
- Li, G.-W., Xie, X.S., 2011. Central dogma at the single-molecule level in living cells. *Nature* 475, 308–315.
- Li, J.J., Bickel, P.J., Biggin, M.D., 2014. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* 2, e270.
- Li, J.J., Chew, G.-L., Biggin, M.D., 2017a. Quantitating translational control: mRNA abundance-dependent and independent contributions and the mRNA sequences that specify them. *Nucleic Acids Res.* 45, 11821–11836.
- Li, J., Liang, Q., Song, W., Marchisio, M.A., 2017b. Nucleotides upstream of the Kozak sequence strongly influence gene expression in the yeast *S. cerevisiae*. *J. Biol. Eng.* 11, 1–14.
- Liu, Y., Beyer, A., Aebersold, R., 2016. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* 165, 535–550. <https://doi.org/10.1016/j.cell.2016.03.014>.
- Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 1–21. <https://doi.org/10.1186/s13059-014-0550-8>.
- Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Ålgenäs, C., Lundberg, J., Mann, M., Uhlen, M., 2010. Defining the transcriptome and proteome in three functionally different human cell lines. *Mol. Syst. Biol.* 6, 450.
- Mann, M., Jensen, O.N., 2003. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* 21, 255–261.
- Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R., Bähler, J., 2012. Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell* 151, 671–683.
- Martin, K.C., Ephrussi, A., 2009. mRNA localization: gene expression in the spatial dimension. *Cell* 136, 719–730.
- Mather, W.H., Hasty, J., Tsimring, L.S., Williams, R.J., 2013. Translational cross talk in gene networks. *Biophys. J.* 104, 2564–2572.
- McAdams, H.H., Arkin, A., 1997. Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci.* 94, 814–819.
- McManus, C.J., May, G.E., Spealman, P., Shteyman, A., 2014. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res.* 24, 422–430.
- Mergner, J., Frejmo, M., List, M., Papacek, M., Chen, X., Chaudhary, A., Samaras, P., Richter, S., Shikata, H., Messerer, M., Lang, D., Altmann, S., Cyprius, P., Zolg, D.P., Mathieson, T., Bantscheff, M., Hazarika, R.R., Schmidt, T., Dawid, C., Dunkel, A., Hofmann, T., Sprunck, S., Falter-Braun, P., Johannes, F., Mayer, K.F.X., Jürgens, G., Wilhelm, M., Baumbach, J., Grill, E., Schneitz, K., Schwechheimer, C., Kuster, B., 2020. Mass-spectrometry-based draft of the Arabidopsis proteome. *Nature* 579, 409–414. <https://doi.org/10.1038/s41586-020-2094-2>.
- Mimitou, E.P., Lareau, C.A., Chen, K.Y., Zorzet-Fernandes, A.L., Hao, Y., Takeshima, Y., Luo, W., Huang, T.-S., Yeung, B.Z., Papalexi, E., et al., 2021. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.* 39, 1246–1258.
- Moran, M.A., Satinsky, B., Gifford, S.M., Luo, H., Rivers, A., Chan, L.-K., Meng, J., Durham, B.P., Shen, C., Varaljay, V.A., Smith, C.B., Yager, P.L., Hopkinson, B.M., 2012. Sizing up metatranscriptomics. *The ISME J.* 7, 237–243. <https://doi.org/10.1038/ismej.2012.94>.
- Moritz, C.P., Mühlhaus, T., Tenzer, S., Schulenburg, T., Friauf, E., 2019. Poor transcript-protein correlation in the brain: negatively correlating gene products reveal neuronal polarity as a potential cause. *J. Neurochem.* 149, 582–604.
- Nicolet, B.P., Wolters, M.C., 2022. The relationship of mRNA with protein expression in CD8+ T cells associates with gene class and gene characteristics. *PLoS One* 17, e0276294.
- Nie, L., Wu, G., Brockman, F.J., Zhang, W., 2006. Integrated analysis of transcriptomic and proteomic data of *Desulfovibrio vulgaris*: zero-inflated Poisson regression models to predict abundance of undetected proteins. *Bioinformatics* 22, 1641–1647. <https://doi.org/10.1093/BIOINFORMATICS/BTL134>.
- Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., Mann, M., 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* 1, 376–386.
- Ori, A., Toyama, B.H., Harris, M.S., Bock, T., Iskar, M., Bork, P., Ingolia, N.T., Hetzer, M. W., Beck, M., 2015. Integrated transcriptome and proteome analyses reveal organ-specific proteome deterioration in old rats. *Cell Syst.* 1, 224–237.
- Paulson, J., Berg, O.G., Ehrenberg, M., 2000. Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation. *Proc. Natl. Acad. Sci.* 97, 7148–7153.
- Petelski, A.A., Emmott, E., Leduc, A., Huffman, R.G., Specht, H., Perlman, D.H., Slavov, N., 2021. Multiplexed single-cell proteomics using SCoPE2. *Nat. Protoc.* 16, 5398–5425.
- Peterson, Vanessa M., Zhang, Kelvin Xi, Kumar, Namit, Wong, Jerelyn, Li, Lixia, Wilson, Douglas C., Moore, Renee, McClanahan, Terrill K., Sadekova, Svetlana, 2017. Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* 35. <https://doi.org/10.1038/nbt.3973>.
- Petricoin III, E.F., Bichsel, V.E., Calvert, V.S., Espina, V., Winters, M., Young, L., Belluco, C., Trock, B.J., Lippman, M., Fishman, D.A., et al., 2005. Mapping molecular networks using proteomics: a vision for patient-tailored combination therapy. *J. Clin. Oncol.* 23, 3614–3621.
- Petrosius, V., Schoof, E.M., 2023. Recent advances in the field of single-cell proteomics. *Transl. Oncol.* 27, 101556.
- Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S., Sandberg, R., 2014. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* 9, 171–181.
- Picotti, P., Aebersold, R., 2012. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* 9, 555–566.
- Ponnala, L., Wang, Y., Sun, Q., van Wijk, K.J., 2014. Correlation of mRNA and protein abundance in the developing maize leaf. *Plant J.* 78, 424–440.
- Reinartz, J., Bruyns, E., Lin, J.-Z., Burcham, T., Brenner, S., Bowen, B., Kramer, M., Woychik, R., 2002. Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. *Brief. Funct. Genom.* 1, 95–104.
- Robinson, M.D., Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11, 1–9.
- Ross, P.L., Huang, Y.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., et al., 2004. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* 3, 1154–1169.
- Ryan, C.J., Cimermančić, P., Szpiech, Z.A., Sali, A., Hernandez, R.D., Krogan, N.J., 2013. High-resolution network biology: connecting sequence with function. *Nat. Rev. Genet.* 14, 865–879.
- Salis, H.M., Mirsky, E.A., Voigt, C.A., 2009. Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.* 27, 946–950.
- Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470.
- Schrimpf, S.P., Weiss, M., Reiter, L., Ahrens, C.H., Jovanovic, M., Malmström, J., Brunner, E., Mohanty, S., Lercher, M.J., Hunziker, P.E., et al., 2009. Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biol.* 7, e1000048.
- Schwahnhauser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., Selbach, M., 2011. Global quantification of mammalian gene expression control. *Nature* 473, 337–342.
- Shahrezaei, V., Swain, P.S., 2008. Analytical distributions for stochastic gene expression. *Proc. Natl. Acad. Sci.* 105, 17256–17261. <https://doi.org/10.1073/pnas.0803850105>.
- Sharova, L.V., Sharov, A.A., Nedorezov, T., Piao, Y., Shaik, N., Ko, M.S., 2009. Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA Res.* 16, 45–58.
- Shen, Z., Zeng, L., Zhang, Z., 2020. Translatome and transcriptome profiling of hypoxic-induced rat cardiomyocytes. *Molecular Therapy-Nucleic Acids* 22, 1016–1024.
- Silva, G.M., Vogel, C., 2016. Quantifying gene expression: the importance of being subtle. *Mol. Syst. Biol.* 12, 885.
- Singh, A., 2021. Towards resolving proteomes in single cells. *Nat. Methods* 18, 856.
- Specht, H., Emmott, E., Petelski, A.A., Huffman, R.G., Perlman, D.H., Serra, M., Kharchenko, P., Koller, A., Slavov, N., 2021. Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biol.* 22, 50. <https://doi.org/10.1186/s13059-021-02267-5>.
- Stahl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., et al., 2016. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353, 78–82.
- Stahlberg, A., Thomsen, C., Ruff, D., Åman, P., 2012. Quantitative PCR Analysis of DNA, RNAs, and Proteins in the Same Single Cell. *Clin. Chem.* 58, 1682–1691. <https://doi.org/10.1373/clinchem.2012.191445>.

- Stefanini, M., Lovino, M., Cucchiara, R., Ficarra, E., 2023. Predicting gene and protein expression levels from DNA and protein sequences with Perceiver. *Comput. Methods Prog. Biomed.* 234, 107504. <https://doi.org/10.1016/j.cmpb.2023.107504>.
- Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., Smibert, P., 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* 14, 865–868.
- Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al., 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382.
- Taniguchi, Y., Choi, P.J., Li, G.-W., Chen, H., Babu, M., Hearn, J., Emili, A., Xie, X.S., 2010. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *science* 329, 533–538.
- Teixeira, F.K., Lehmann, R., 2019. Translational control during developmental transitions. *Cold Spring Harb. Perspect. Biol.* 11, a032987.
- Teyssonniere, E.M., Shichino, Y., Mito, M., Friedrich, A., Iwasaki, S., Schacherer, J., 2024. Translation variation across genetic backgrounds reveals a post-transcriptional buffering signature in yeast. *Nucleic Acids Res.* 52, 2434–2445.
- Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Hamon, C., 2003. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* 75, 1895–1904.
- Trösemeier, J.H., Rudolf, S., Loessner, H., Hofner, B., Reuter, A., Schulenburg, T., Koch, I., Bekeredjian-Ding, I., Lipowsky, R., Kamp, C., 2019. Optimizing the dynamics of protein expression. *Sci. Rep.* 9. <https://doi.org/10.1038/S41598-019-43857-5>.
- van Galen, P., Hovestadt, V., Wadsworth II, M.H., Hughes, T.K., Griffin, G.K., Battaglia, S., Verga, J.A., Stephansky, J., Pastika, T.J., Story, J.L., et al., 2019. Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity. *Cell* 176, 1265–1281.
- Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzler, K.W., 1995. Serial analysis of gene expression. *Science* 270, 484–487.
- Vogel, C., Marcotte, E.M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* 13, 227–232.
- Vogel, C., de Sousa Abreu, R., Ko, D., Le, S.-Y., Shapiro, B.A., Burns, S.C., Sandhu, D., Boutz, D.R., Marcotte, E.M., Penalva, L.O., 2010. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* 6, 400.
- Wagner, G.P., Kin, K., Lynch, V.J., 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* 131, 281–285.
- Wang, Z., Sun, X., Zhao, Y., Guo, X., Jiang, H., Li, H., Gu, Z., 2015. Evolution of gene regulation during transcription and translation. *Genome Biol. Evol.* 7, 1155–1167.
- Wang, D., Eraslan, B., Wieland, T., Hallström, B., Hopf, T., Zolg, D.P., Zecha, J., Asplund, A., Li, L., Meng, C., et al., 2019. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* 15, e8503.
- Wang, Z.-Y., Leushkin, E., Liechti, A., Ovchinnikova, S., Mößinger, K., Brünig, T., Rummel, C., Grütner, F., Cardoso-Moreira, M., Janich, P., et al., 2020. Transcriptome and translome co-evolution in mammals. *Nature* 588, 642–647.
- Wang, H., Wang, Y., Yang, J., Zhao, Q., Tang, N., Chen, C., Li, H., Cheng, C., Xie, M., Yang, Y., et al., 2021. Tissue-and stage-specific landscape of the mouse translome. *Nucleic Acids Res.* 49, 6165–6180.
- Wilhelm, M., Schlegel, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., et al., 2014. Mass-spectrometry-based draft of the human proteome. *Nature* 509, 582–587.
- Xia, C., Babcock, H.P., Moffitt, J.R., Zhuang, X., 2019. Multiplexed detection of RNA using MERFISH and branched DNA amplification. *Sci. Rep.* 9, 7721.
- Xuefei Wang, W.J., Xinchao, Wu, Hong, Ni, 2024. Progress in single-cell multimodal sequencing and multi-omics data integration. *Biophys. Rev.* 16. <https://doi.org/10.1007/s12551-023-01092-3>.
- Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S., et al., 2014. Proteogenomic characterization of human colon and rectal cancer. *Nature* 513, 382–387.
- Zhu, Y., Piehowski, P.D., Zhao, R., Chen, J., Shen, Y., Moore, R.J., Shukla, A.K., Petyuk, V.A., Campbell-Thompson, M., Mathews, C.E., Smith, R.D., Qian, W.-J., Kelly, R.T., 2018. Nanodroplet processing platform for deep and quantitative proteome profiling of 10–100 mammalian cells. *Nat. Commun.* 9, 882. <https://doi.org/10.1038/s41467-018-03367-w>.
- Zrimec, J., Buric, F., Kokina, M., Garcia, V., Zelezniak, A., 2021. Learning the Regulatory Code of Gene Expression. *Front. Mol. Biosci.* 8, 530. <https://doi.org/10.3389/fmolb.2021.673363>.