Differential analysis of translation efficiency and usage of open reading frames using DOTSeq

Chun Shen Lim[1,2,3,*], Gabrielle S.W. Chieng[1,2]

[1] Department of Biochemistry, School of Biomedical Sciences, University of Otago, Dunedin, New Zealand.
[2] Genetics Otago, University of Otago, Dunedin, New Zealand.
[3] Maurice Wilkins Centre for Molecular Biodiscovery, Centres of Research Excellence, New Zealand.

* Corresponding author. Email: chunshen.lim@otago.ac.nz

ABSTRACT
Protein synthesis is a key cellular process in which mRNAs are translated into proteins by ribosomes. This process is tightly regulated, enabling cells to control protein output in response to specific cellular states. Ribosome profiling captures translatomic landscapes across conditions, but existing computational tools for differential translation analysis operate at the gene level, overlooking translational control at the level of multiple open reading frames (ORFs). Here, we present DOTSeq, a Differential ORF Translation statistical framework that enables systematic discovery of translational control events within genes. DOTSeq offers differential analyses of ORF usage and translation efficiency across biological conditions. These modules allow global detection of *cis*-regulatory events, such as upstream ORF (uORF)-mediated translational control. Benchmarking on simulated datasets demonstrates DOTSeq's sensitivity to subtle regulatory signals, particularly in modest effect sizes with realistic batch effects, where all tools show limited sensitivity. DOTSeq provides a flexible and powerful approach for dissecting the complexity of translational control.

Keywords: *cis*-regulatory control, beta-binomial, differential ORF usage, differential translation efficiency, generalised linear mixed model, Ribo-seq

INTRODUCTION
Protein synthesis is a key step in the gene expression pathway, where mRNAs are translated into functional proteins. The levels of proteins are tightly regulated during translation, a dynamic process that allows cells to adapt to environmental cues and signaling events. With the advent of ribosome profiling (Ribo-seq), we now have the ability to study translation at nucleotide resolution. Ribo-seq is a deep-sequencing technique that captures ribosome-protected mRNA fragments, typically ~30 nucleotides long, following nuclease digestion. These fragments, or footprints, provide a snapshot of ribosome positions along a stretch of nucleotides, thus providing us an estimate of enabling global quantification of translation and identification of reading frames translated.

An open reading frame (ORF) is a continuous stretch of nucleotide sequence that has a length divisible by three and is bounded by start and stop codons (*1*). An ORF does not necessarily contain a full coding sequence (CDS) which is a sequence that is eventually translated into a protein. While main ORFs (mORFs) encode canonical protein-coding sequences, small ORFs (300 nt or shorter) have recently gained attention due to widespread ribosome occupancy revealed by Ribo-seq (*2–4*). These small ORF include upstream and downstream ORF (uORF and dORF, respectively) (*5*). Some small ORFs encode functional peptides that act as

membrane-associated regulators with important cellular roles (*4*, *6*). Others function as *cis*-regulatory elements, regulating mORF translation through mechanisms such as ribosome stalling, attenuation, or interaction with regulatory proteins. A well-characterised example is ATF4, a key effector of the integrated stress response, where mORF translation is regulated by uORFs (*7–10*). These regulatory mechanisms highlight the complexity of translational control at the level of multiple ORFs. Such complexity, in turn, emphasises the need for genome-wide computational approaches to detect translational regulatory events across biological conditions.

While substantial progress has been made in identifying novel ORFs, such as through our tool RIBOSS (*11*), it is equally important to determine if these ORFs are involved in translational control. However, existing tools for differential translation analysis, including deltaTE, RiboDiff, Riborex, and Xtail, operate at the gene level, and cannot resolve differential usage between ORFs within the same gene (*12–15*). To address this limitation, we developed DOTSeq, an R package that performs differential ORF translation analysis that integrates Ribo-seq and matched RNA-seq data.

DOTSeq allows users to ask two key biological questions: (i) does the translation efficiency of a specific ORF change across conditions? (ii) more importantly, does the relative translation of one ORF compared to others within the same gene change across conditions? DOTSeq addresses these two aspects of translational control by modelling both Differential Translation Efficiency (DTE) and Differential ORF Usage (DOU) at the level of multiple ORFs, enabling the detection of regulatory events in response to distinct cellular states. The DOTSeq framework provides a powerful approach to dissect the complexity of translational control at the global scale.

RESULTS

Overview of DOTSeq modules

Ribo-seq data contains rich information about translation, including complex *cis*-regulatory events such as alternative initiation. These events often contribute to changes in relative usage of ORFs within the same gene, which are not captured by standard gene-level differential translation analyses. To address this, we developed DOTSeq, a computational framework that enables global detection of *cis*-translational control events. DOTSeq utilises ORF-level annotations and statistical modelling strategies that capture both absolute changes in translation and condition-specific shifts in the proportion of Ribo-seq to RNA-seq read counts for individual ORFs, relative to other ORFs within the same gene.

To enable ORF-level analysis, DOTSeq requires preprocessing steps that include: (i) alignment of Ribo-seq and RNA-seq reads using splice-aware short-read aligners (e.g., STAR or HISAT2) (*16*, *17*); and (ii) generation of custom GTF and BED files by "flattening" transcript isoforms into discrete, non-overlapping ORF bins using the RIBOSS engine (*11*) (Fig 1). This step is similar to DEXSeq's step in defining exon counting bins by flattening GTF files (*18*). Flattened annotations based on the GENCODE human reference gene annotation (*19*) are available as part of the package, and custom annotations for other species can be generated using the included orf_to_gtf.py script. Ribo-seq and RNA-seq counts are aggregated per ORF based on the flattened annotation using featureCounts (*20*).

DOTSeq comprises two statistical modules to model differential usage and translation of ORFs (Table 1). The Differential ORF Usage (DOU) module models changes in the relative

usage of ORFs within a gene, allowing users to identify *cis*-regulatory events. It uses a beta-binomial generalised linear mixed model (GLMM) implemented via the glmmTMB package (*21*). It models the proportion of Ribo-seq and RNA-seq counts for a given ORF relative to other ORFs within the same gene, and tests whether this proportion changes between conditions using an interaction term (condition:strategy). This term captures whether the effect of sequencing strategy (Ribo-seq vs RNA-seq) differs between biological conditions, enabling detection of condition-specific shifts in ORF usage driven by *cis*-translation control. Dispersion is modelled separately for Ribo-seq and RNA-seq to account for distinct sources of variability. Post hoc analysis includes pairwise contrasts across conditions, computation of estimated marginal means (EMMs) using the emmeans package (*22*), and adaptive shrinkage of effect sizes via the ashr package (*23*).

In contrast, the Differential Translation Efficiency (DTE) module quantifies changes in ribosome loading relative to RNA expression at the level of individual ORFs. It is designed to detect DTE in single-ORF/monocistronic genes and in ORFs showing monotonic changes within multi-cistronic genes that the DOU module is not designed to capture. The DTE module uses a similar workflow as deltaTE. This module models raw ORF-level counts directly using a negative binomial generalised linear model (GLM) implemented in DESeq2 (*24*), and tests whether the ratio of Ribo-seq to RNA-seq counts differs across conditions using an interaction term (condition:strategy). Together, these complementary modules offer a genome-wide approach for extracting regulatory information in Ribo-seq experiments across biological conditions, enabling systematic interpretation of complex *cis*-regulatory events that are inaccessible to current gene-level tools.

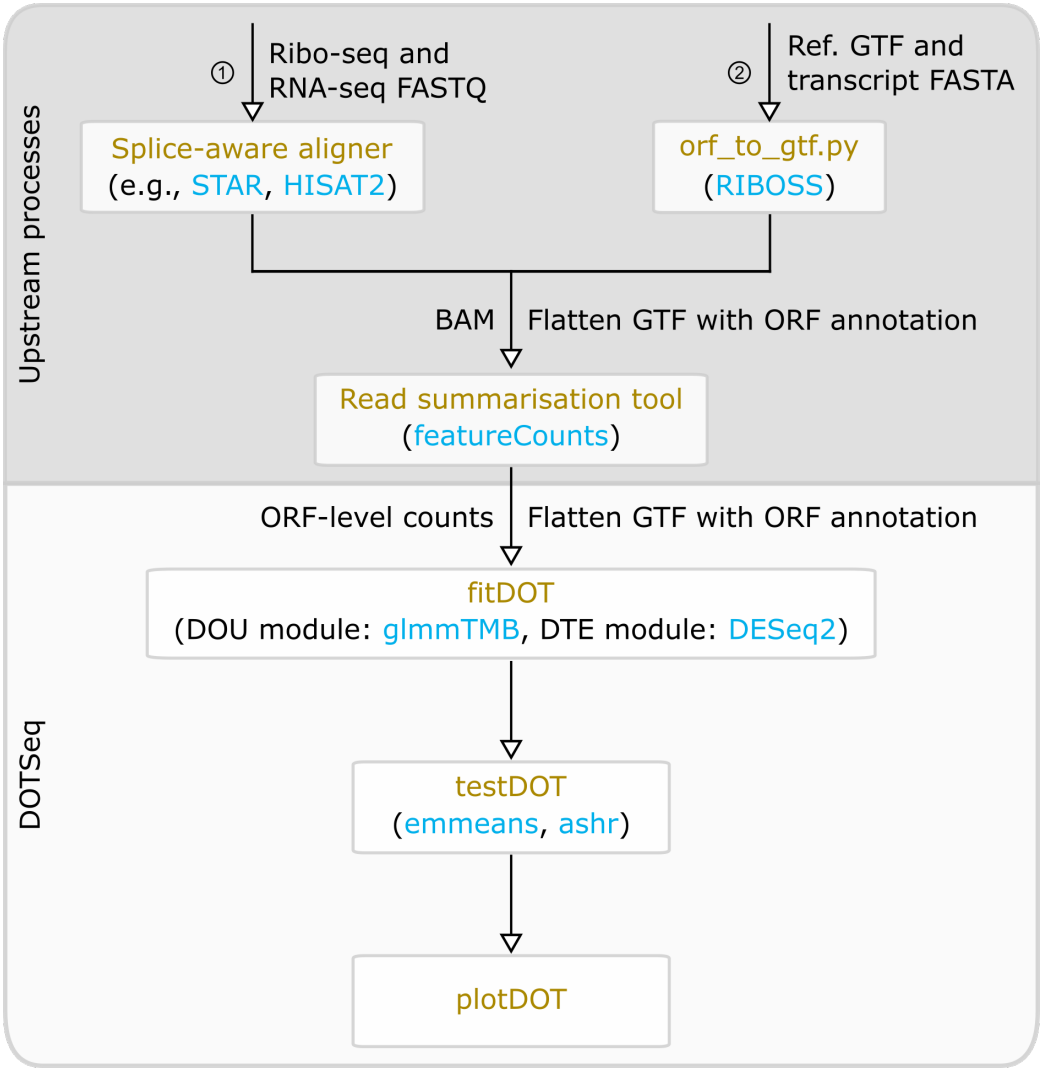

Fig 1. Overview of the DOTSeq workflow. ① Ribo-seq and matched RNA-seq reads are aligned to a reference genome using a splice-aware aligner, such as STAR or HISAT2. ② Reference GTF and transcript FASTA files are parsed using the orf_to_gtf.py wrapper, which leverages the RIBOSS engine to generate flattened GTF and BED files with discrete, non-overlapping ORF annotations. DOTSeq integrates two statistical modules: Differential ORF Usage (DOU) and Differential Translation Efficiency (DTE). The fitDOT function performs model fitting, while testDOT enables post hoc contrasts and empirical Bayesian shrinkage. The plotDOT function generates visualisations including a scatter plot of DOU and DTE effect sizes, a Venn diagram showing overlap between DOU and DTE results, and a heatmap of genes with differential ORF usage.

Table 1. DOTSeq statistical modules for Differential ORF Usage (DOU) and Differential Translation Efficiency (DTE).

Feature	**DOU module**	**DTE module**
Biological question	Does the relative translation of one ORF compared to other ORFs within the same gene change across conditions?	Does the translation efficiency of a specific ORF change across conditions?
Model type	GLMM with beta-binomial distribution via glmmTMB.	Negative binomial GLM via DESeq2.
Response variable	Proportion of counts for an ORF within its gene, modelled as a binomial response $(y_i, n_i - y_i)$, where y_i is the count for ORF_i(Ribo-seq or RNA-seq), and n_i is the total count for all ORFs in the same gene.	Raw ORF-level counts for Ribo-seq and RNA-seq, modelled directly without normalisation to gene-level totals.
Library size normalisation	No normalisation needed as proportional modelling is robust to variations in library size.	Handled internally by DESeq2::estimateSizeFactors.
Dispersion modelling	Library strategy-specific dispersion via dispersion formula, where each ORF has separate dispersion parameters for Ribo-seq and RNA-seq.	Shared dispersion across strategies, where each ORF has a single dispersion parameter. Dispersion is estimated and shrunk based on a mean-dispersion trend.
Interaction term usage	condition:strategy in fixed-effects formula. Captures changes in the proportion of counts for one ORF relative to other ORFs within the same gene, across both Ribo-seq and RNA-seq data.	condition:strategy in design formula. Captures changes in the ratio of Ribo-seq to RNA-seq counts for individual ORFs.
Hypothesis tests	LRT and Wald test.	Wald test (DESeq2 default).
Post hoc contrasts	Via emmeans	Via DESeq2::results.
Empirical Bayesian shrinkage	Adaptive shrinkage for effect sizes (log-odds changes) via ashr	Approximate/adaptive shrinkage effect sizes (log2 fold changes) via apeglm/ashr.

DOU and DTE modules show complementary strengths in simulation benchmarks
To evaluate performance under controlled conditions, we designed simulation scenarios modelling DOU using the DOTSeq simDOT function, which builds on the polyester simulation engine (*25*). These scenarios incorporated experimentally derived Ribo-seq and matched RNA-seq count matrices from HeLa cells synchronised at two distinct cell cycle stages: mitotic cycling (cells undergoing DNA replication and mitosis) and interphase (cells replicating DNA without mitosis) (*26*).

DOU is modeled as a change in the proportion of Ribo-seq and RNA-seq counts for a given ORF relative to the total counts aggregated from all the ORFs in the same gene. For example, if an ORF's usage remains stable across conditions, its proportion of Ribo-seq and RNA-seq counts relative to other ORFs will be consistent, and no condition-specific change in its relative contribution to the gene's overall expression and translation will be detected. However, if an ORF becomes more or less translated in one condition, its Ribo-seq proportion will shift relative to its RNA-seq proportion and to other ORFs in the gene, indicating a change in ORF usage within the gene context.

To assess sensitivity across a range of biological signal strengths, we introduced varying magnitudes of log-fold changes in translation efficiency to multi-cistronic genes. These include simulating the absolute magnitude differences between uORFs and mORFs by 1, 2, and 3 log-fold changes, representing low, moderate and large effect sizes, respectively (Fig 2). We also incorporated batch effects to provide realistic experimental variability.

We analysed each simulation using the DOU and DTE modules, and benchmarked them against existing tools for differential translation analysis, including deltaTE (*15*), RiboDiff (*13*), Riborex (*14*), and Xtail (*12*). Performance was assessed using the iCOBRA package, including False Discovery Rate (FDR) and True Positive Rate (TPR) across multiple thresholds (*27*).

The DOU module outperformed the DTE module and other tools at low and moderate effect sizes (Fig 2). This demonstrates the power of the modelling of condition-specific changes in the proportional counts for an ORF relative to other ORFs within the same gene via a beta-binomial framework. While all tools exhibited limited sensitivity at the lowest effect size where the signal-to-noise ratio is low, performance improved with increasing effect size. The DTE module and other tools showed greater sensitivity to effect size and outperformed the DOU module at the largest magnitude. The DTE module is functionally equivalent to deltaTE and, as expected, exhibited identical performance in our benchmarks. Together, these results highlight the combined utility of the two DOTSeq modules in capturing a broad spectrum of translational control events.

To further assess statistical calibration, we examined the distribution of FDR-adjusted p-values across additional 128 simulated scenarios modelling differential ORF usage. These scenarios varied in effect size and batch effect magnitude to reflect diverse biological and technical conditions. The DOU module produced a right-skewed p-value distribution, indicating well-calibrated statistical significance (Fig 3). In contrast, the DTE module exhibited a U-shaped distribution, suggesting p-value inflation and reduced calibration when applied to DOU-type regulatory events. These results highlight the importance of using DOU-specific models for detecting translational control events involving DOU.

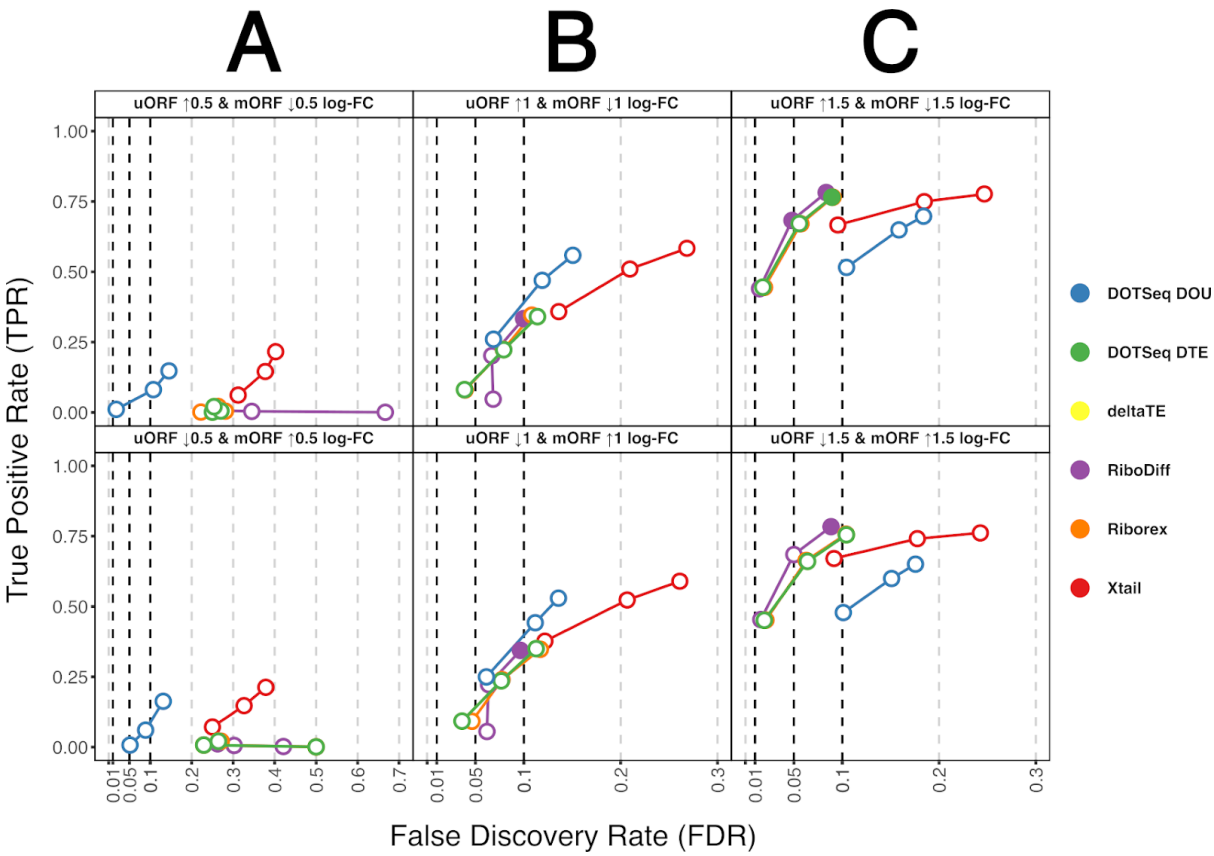

Fig 2. DOTSeq detects simulated *cis*-translational control events at varying effect sizes. The performance of the DOTSeq DOU and DTE modules was benchmarked against existing tools for differential translation analysis across six simulation scenarios, using False Discovery Rate-True Positive Rate (FDR-TPR) plots. Each panel represents a simulation scenario with a distinct combination of *cis*-regulatory event and effect size. These include simulating the absolute magnitude differences between uORFs and mORFs by 1, 2, and 3 log-fold changes (**A**, **B**, and **C**, respectively). Points represent the performance for each tool at the nominal FDR of 1%, 5%, and 10%, and are filled if the observed FDR is below the nominal FDR threshold. **(A)** In low-effect-size scenarios, the DOU module consistently achieves higher TPR than other tools, indicating that it is the best-performing tool although all tools show limited sensitivity. **(B)** In moderate-effect-size scenarios, the curves for the DOU module are closest to the top-left corner of the plot, indicating a higher TPR at a given observed FDR compared to the other tools. These results demonstrate the strength of modelling of Ribo-seq-to-RNA-seq proportions across ORFs within genes, which captures coordinated changes in ORF usage that gene-level methods may miss, especially in noisy datasets. **(C)** In large-effect-size scenarios, the DTE module and other tools outperform the DOU module, indicating their ability to detect strong translational signals. RiboDiff successfully controls the FDR at the 5% and/or 10% nominal level in moderate- and large-effect-size scenarios, whereas the DTE module, deltaTE, and Riborex achieve FDR control at the 10% level in large-effect-size scenarios. Each simulation scenario included Ribo-seq with matched RNA-seq data for two conditions, each in two batches with two biological replicates (eight samples total). A batch effect was introduced on 10% of ORFs to provide realistic experimental variability. For these genes, a random log-fold change was assigned, with a magnitude centred around 0.9. These log-fold changes were applied to the non-reference batches, where the effect is a fold-change relative to the control group.

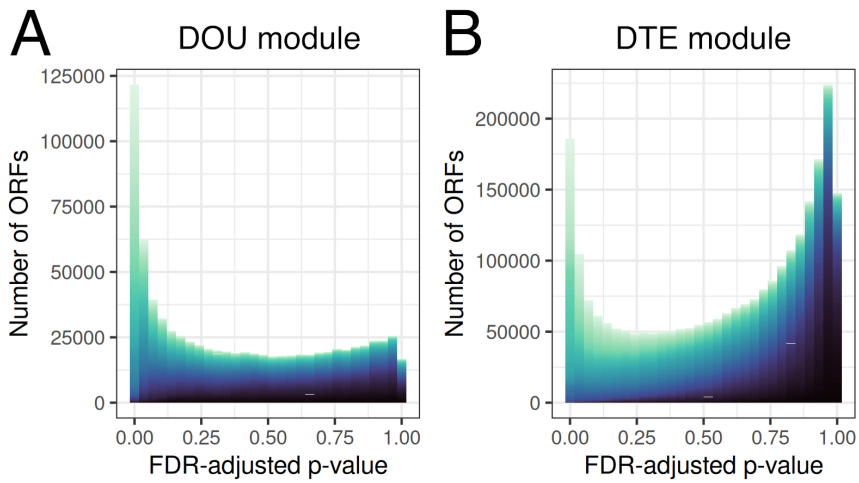

Fig 3. DOTSeq DOU module is well-calibrated for differential ORF usage analysis. Distributions of aggregated FDR-adjusted p-values across 128 simulated scenarios for the DOTSeq DOU **(A)** and DTE **(B)** modules. These scenarios model DOU with varying effect sizes between ORFs within genes and different batch effect magnitudes. Bar height indicates the total number of ORFs across all scenarios, whereas color gradient indicates the composition of scenarios contributing to the corresponding p-value bin. **(A)** The DOU module shows a right-skewed distribution, consistent with well-calibrated statistical significance. **(B)** In contrast, the DTE module exhibits a U-shaped distribution, indicating p-value inflation.

DOTSeq DOU and DTE modules reveal complementary insights into differential ORF translation

To assess whether the DOU and DTE modules capture overlapping or distinct aspects of translational control, we compared effect sizes between the two modules using the HeLa cell cycle data above. Most changes in translation efficiency fell within a modest range (±1 log2-fold change), where 242 of differentially translated ORFs were missed by the DTE module but detected by DOU (Fig 3A and B). However, there was a strong correlation between DOU log-odds changes and DTE log-fold changes, indicating that both modules detect condition-dependent changes in translation through different modelling approaches (Fig 3A, Spearman $\rho = 0.65$, p-value $< 10^{-16}$). Notably, the overlap between the ORFs meeting statistical thresholds was partial: 228 ORFs were unique to DTE, 242 unique to DOU, and 326 shared (Fig 3B, overlap coefficient = 0.56). These findings demonstrate that the two modules provide complementary insights rather than redundant results.

For this cell cycle data, the DOU analysis revealed changes in ORF usage within genes that would be missed by a DTE approach alone. In particular, we observed clear differences in uORF versus mORF usage between mitotic cycling versus interphase (Fig 3C). Such differences highlight the impact of *cis*-regulation, as uORF translation often represses mORFs. In this case study, the DOU module detected a significant shift towards increased uORF and mORF usage in 31 and 5 genes in mitotic cycling, respectively.

These significant changes include a shift in ORF usage from the mORF to the uORF for the initiation factor EIF4G1 and ribosomal protein genes RPL8, RPL14, and RPS24 (Fig 3C). This pattern is consistent with previous studies, showing that global translation is repressed during mitosis, at least partly to a greater extent than transcription (*28–31*). While this repression has been attributed to eIF4G1 phosphorylation (*29*), our findings suggest an

additional layer of translational control mediated by uORFs. In contrast to mitotic cycling, EIF4G1 mORF usage is significantly elevated during interphase (Fig 3C). This observation aligns with functional studies showing that EIF4G1 promotes G1/S transition, as EIF4G1 knockdown decreases CCND1 (Cyclin D1) and phosphorylated Retinoblastoma protein levels, leading to cell-cycle delay, reduced cell viability, and impaired proliferation (*32*, *33*). Together, these findings suggest that the EIF4G1 is regulated through multiple layers, with preferential mORF usage during interphase, supporting its role in cell-cycle progression.

Similar to EIF4G1, we observed an increased uORF usage of the ATF4 gene in mitotic cycling compared to interphase. This is consistent with previous studies, where ATF4 expression is transcriptionally downregulated in M phase when compared to the other phases of the cell cycle, G0/G1, S and G2 phase, indicating cell cycle-dependent regulation (*31*). Moreover, ATF4-dependent transcription has been shown to increase CDKN1A (p21) and decrease CCND1 (Cyclin D1) levels through PERK-mediated stress response, thereby promoting a G1 cell-cycle arrest (*34*). Thus, reduced mORF translation during mitosis may reflect a repressive role of the ATF4 uORF, limiting ATF4 expression to allow normal G1 entry.

We also observed a shift toward mORF usage in the H2AC6 gene, an isotype of the histone 2A, during mitotic cycling. Previous studies have shown that H2A deubiquitination is required for proper progression through the M phase (*35*). Thus, the shift in mORF usage supports the notion that H2AC6 plays a crucial role in mitotic cycling. Taken together, these results demonstrate that the DOU module is capable of revealing *cis*-translational control events.

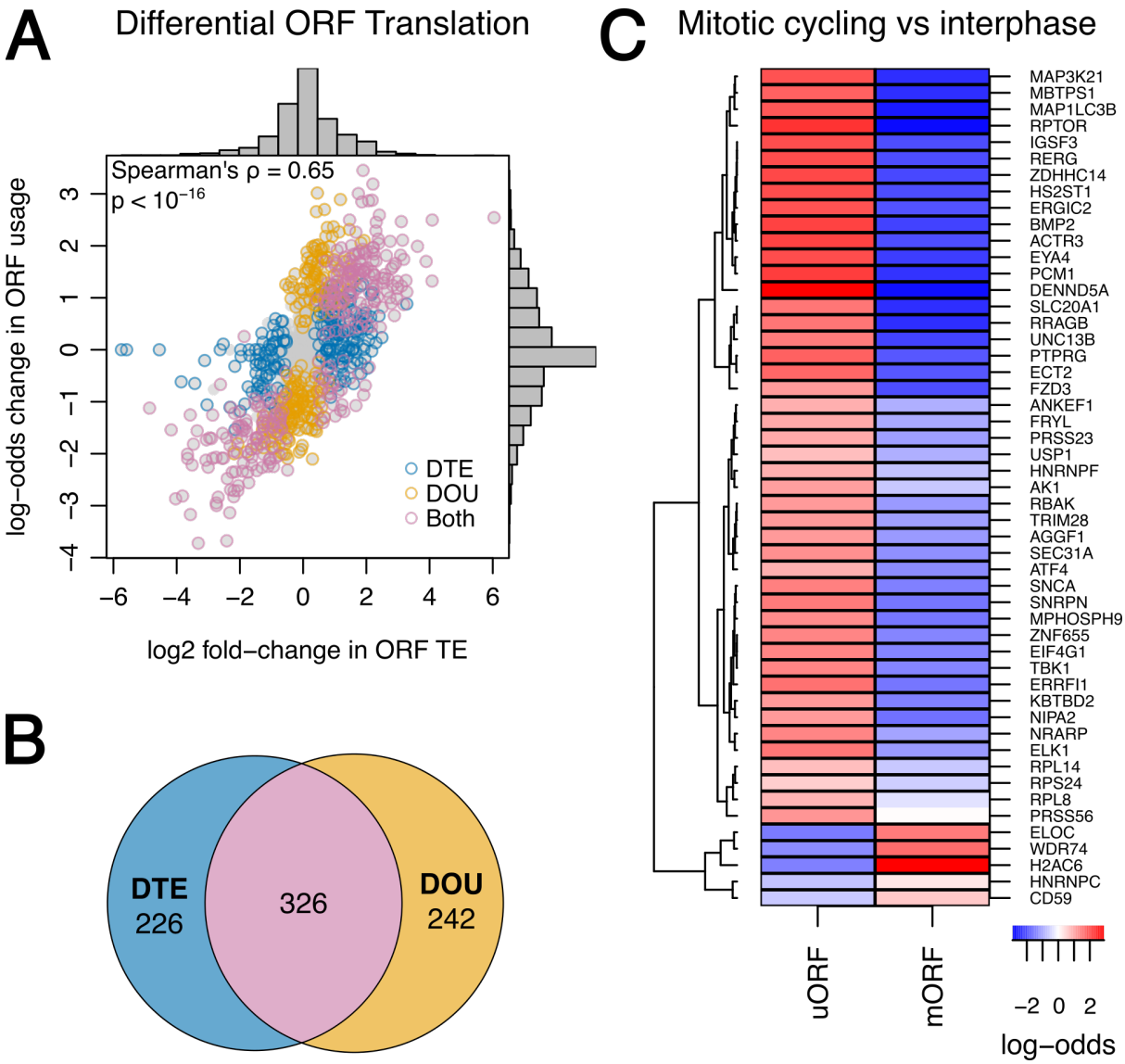

Fig 3. DOTSeq reveals distinct patterns of differential ORF translation between mitotic cycling and interphase. Analysis of Ribo-seq and RNA-seq data comparing HeLa cells synchronised at interphase and mitotic cycling. **(A)** Scatter plot showing a significant positive correlation between DOU log-odds change and DTE log2 fold-change (Spearman $\rho = 0.65$, p-value $< 10^{-16}$). Effect sizes for both DOU and DTE have been shrunk using empirical Bayesian estimators to improve stability and interpretability. **(B)** Venn diagram illustrating overlap between ORFs detected by DOU and DTE. Of the total, 228 ORFs are unique to DTE, 242 to DOU, and 326 are shared, yielding an overlap coefficient of 0.59. **(C)** Heatmap of genes with differential ORF usage (DOU), showing log-odds values ranging approximately from –2 to +2. Rows represent genes, highlighting differences in uORF and mORF usage between mitotic cycling and interphase conditions.

DISCUSSION

The ability to resolve changes in ORF usage across biological conditions opens up new avenues to exploring translational control mechanisms that are obscured in gene-level analyses. DOTSeq addresses this need by enabling differential analyses of both ORF usage and translation efficiency. This enhanced sensitivity and resolution enables discovery of important regulatory events (e.g., uORF-mediated repression) that are often overlooked in gene-level analyses.

To evaluate DOTSeq's ability in capturing these regulatory events, we have conducted simulations to benchmark its performance. These simulations demonstrate the complementary strengths of the two DOTSeq modules: the DOU module is effective at detecting subtle, condition-specific shifts in ORF usage in noisy datasets, whereas the DTE module is more sensitive to large-magnitude changes in translation efficiency. While low sensitivity at modest effect sizes remains a common challenge, the DOU module consistently outperforms existing tools in this regime, even with high technical noise.

To demonstrate a real-world application of DOTSeq, we have applied the framework to identify changes in ORF usage across cell cycle phases. This analysis revealed key *cis*-regulatory events contributing to global translation repression during mitosis, including uORF-mediated repression of EIF4G1, a master regulator of protein synthesis, and several ribosomal protein genes. These findings highlight DOTSeq's ability to uncover an additional regulatory layer of protein synthesis in Ribo-seq experiments.

These discoveries rely on DOTSeq's ability to statistically model overdispersed and sparse count data from Ribo-seq and RNA-seq. DOTSeq leverages state-of-the-art statistical modelling to handle the complexity and noise inherent in Ribo-seq and RNA-seq data. However, fitting these beta-binomial GLMMs on sparse Ribo-seq data can lead to convergence issues, including a non-positive definite Hessian. To mitigate these issues, we have implemented multiple dispersion modelling approaches within our framework. (i) The default approach models dispersion as a function of sequencing strategy (Ribo-seq with RNA-seq as reference), allowing for strategy-dependent overdispersion. (ii) As a simpler alternative, dispersion is assumed to be constant across all predictor levels. This approach reduces model complexity and is less prone to convergence issues. (iii) To balance model stability and flexibility, both models are fitted and compared using likelihood ratio tests. The simpler model is selected when no significant difference is detected. (iv) For maximum flexibility, users may specify a custom formula for dispersion modelling. Users may also enable model diagnostics to assess model fit, including tests for overdispersion, zero inflation, and residual properties.

While our framework is flexible, fitting fully parameterised models comes with a computational cost. For example, it took 14 min 10 s (± 1 min 44 s) to fit the beta-binomial GLMMs for 35,863 ORFs in the HeLa cell cycle dataset, running on a single thread of a MacBook Pro M4 with 16 GB RAM. In contrast, the DTE module runs at a speed comparable to existing tools such as deltaTE (32 ± 0.05 s). Despite these challenges, DOTSeq offers a flexible and powerful framework for dissecting the complex landscape of translational control. DOTSeq is broadly applicable to studies of translational regulation across conditions within biological contexts such as development, disease, and stress responses.

METHODS

Data

The Ribo-seq and matched RNA-seq data for HeLa cells synchronised at interphase, mitotic arrest, and mitotic cycling were retrieved from PRJNA957808 (*26*). There are two biological replicates for each condition.

Data preprocessing

Adapter trimming was performed using Cutadapt v5.1 (*36*). Reads were mapped to the UCSC Genome Browser's hg38 with GENCODE v47 annotation using STAR v2.7.11b (*16*, *19*, *37*).

For ORF-level quantification, flattened GTF and BED files containing ORF annotations were generated by processing the GENCODE annotation and transcript FASTA files using the orf_to_gtf.py script from DOTSeq, with the RIBOSS v0.45 engine (*11*). Reads mapped to individual ORFs were counted using featureCounts v2.1.1 (*20*), guided by this flatten GTF file. The resulting flatten annotation files and count matrices were used as input for DOTSeq.

Statistical modelling

The DOTSeq framework builds on a beta-binomial GLMM implemented in glmmTMB v1.1.12 (*21*). The proportion of counts for an ORF within its gene is modelled by using a logit link function:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

The full fixed-effects formula is:

$$\text{logit}(p_i) = \beta_0 + \beta_{\text{cond}} \cdot x_{\text{cond},i} + \beta_{\text{strat}} \cdot x_{\text{strat},i} + \beta_{\text{int}} \cdot x_{\text{cond},i} \cdot x_{\text{strat},i}$$

Here, p_i is the probability that a read maps to a particular ORF (e.g., mORF or uORF) for observation i. The covariates $x_{cond,i}$ and $x_{strat,i}$ are dummy variables representing the condition (e.g., mitotic cycling vs interphase) and strategy (RNA-seq vs Ribo-seq) for observation i, respectively. β_0 is the intercept, β_{cond} is the condition coefficient, β_{strat} is the strategy coefficient, and β_{int} is the interaction coefficient.

This modelling approach fits a unified mean structure across Ribo-seq and RNA-seq levels using a single joint model. For the RNA-seq, the value of $x_{strat,i} = 0$, and the model simplifies to:

$$\text{logit}(p_{\text{rna},i}) = \beta_0 + \beta_{\text{cond}} \cdot x_{\text{cond},i}$$

For the Ribo-seq level, the value of $x_{strat,i} = 1$, and the model becomes:

$$\text{logit}(p_{\text{ribo},i}) = \beta_0 + \beta_{\text{cond}} \cdot x_{\text{cond},i} + \beta_{\text{strat}} + \beta_{\text{int}} \cdot x_{\text{cond},i}$$

This model can be simplified as:

$$\text{logit}(p_{\text{ribo},i}) = (\beta_0 + \beta_{\text{strat}}) + (\beta_{\text{cond}} + \beta_{\text{int}}) \cdot x_{\text{cond},i}$$

to isolate the effective intercept and slope for the Ribo-seq level. The condition effect at the Ribo-seq level is modified by β_{int}. Therefore, when $\beta_{int} = 0$, the condition effect is the same for both Ribo-seq and RNA-seq levels. When $\beta_{int} \neq 0$, the condition affects Ribo-seq differently than RNA-seq, indicating a translation-specific effect.

Since DOTSeq models the proportion of reads mapped to an ORF within a gene, a contrast between Ribo-seq and RNA-seq levels is performed using emmeans v1.11.2-8 with Wald test (*22*) to isolate translation-specific effects:

$$\text{logit}(p_{\text{ribo},i}) - \text{logit}(p_{\text{rna},i}) = \beta_{\text{strat}} + \beta_{\text{int}} \cdot x_{\text{cond},i}$$

This contrast quantifies how much more (or less) likely a read is to map to an ORF at the Ribo-seq level compared to the RNA-seq level, condition-dependently. It is not equivalent to translation efficiency, which is typically defined as a ratio of Ribo-seq to RNA-seq read counts. Instead, this contrast captures DOU, reflecting translation-specific regulation of individual ORFs within genes (*cis*-translational control events).

By default, a null model without the interaction term is included, and a likelihood ratio test is performed to assess the significance of translation-specific effects (i.e., whether the interaction term improves model fit). As part of the workflow, Wald tests for condition-specific contrasts are performed when there are more than two conditions. To stabilise effect size estimates, adaptive empirical Bayesian shrinkage is applied to the log-odds contrasts using ashr v2.2-63 (*23*).

To account for overdispersion in proportional modelling, DOTSeq models variability using the dispersion formula:

$$\log\left(\frac{\phi_i}{1-\phi_i}\right) = \gamma_{0,i} + \gamma_{1,i} \cdot x_{\text{strat},i}$$

where ϕ_i is the overdispersion parameter for ORF_i, γ_0 is the dispersion intercept, and γ_1 captures the additional dispersion at the Ribo-seq level. This allows the model to fit a shared mean structure across Ribo-seq and RNA-seq levels and estimate separate dispersion parameters for these two levels. The fitted dispersions for ORF_i are computed as:

$$\phi_{\text{rna},i} = \frac{1}{1+\exp(\gamma_{0,i})}, \quad \phi_{\text{ribo},i} = \frac{1}{1+\exp(\gamma_{0,i}+\gamma_{1,i})}$$

where $\gamma_{0,i}$ and $\gamma_{1,i}$ are the intercept and strategy effect for ORF_i, respectively. This approach accommodates varying technical noise and biological heterogeneity in Ribo-seq and RNA-seq levels to improve the accuracy of statistical inference.

The DTE module in DOTSeq uses the DESeq2 framework to model differential translation efficiency across conditions directly through the condition:strategy interaction term (*15*, *24*). For each ORF, raw Ribo-seq and RNA-seq read counts are normalised and modelled using a negative binomial GLM. This includes a standard DESeq2 workflow: size factor estimation, dispersion fitting, Wald tests for condition-specific contrasts, and shrinkage of effect sizes.

This complements the DOU module by identifying ORFs with overall changes in translation, using transcript abundance as baseline.

Simulation

To simulate *cis*-translational control, 134 datasets were generated using the DOTSeq's simDOT function, which is based on the deltaTE source code (*15*) and uses the polyester v1.29.1 engine (*25*). The HeLa cell cycle count matrices were used as input to preserve realistic baseline expression profiles.

Simulation scenarios included predefined regulatory events such as "uORF_up_mORF_down" and "uORF_down_mORF_up", which assigned log-fold changes ranging from ±0.5 to ±2 to specific ORF types within multi-cistronic genes. To simulate experimental variability, batch effects were introduced as additional log-fold changes within the same range (±0.5 to ±2). Each simulation included two experimental conditions, with 1–2 biological replicates per condition and 2–3 batches. The simulation output comprised combined count matrices for Ribo-seq and RNA-seq, sample-level metadata (including condition, replicate, strategy, and batch), ground truth labels for differentially regulated ORFs, and the corresponding log-fold changes.

Benchmarking and analysis

Benchmarks were performed on a Ubuntu 24.04.2 LTS server equipped with 2 × AMD EPYC 7763 64-Core Processors (128 physical cores, 128 threads), running at up to 3.53 GHz, with 1 TB DDR4 RAM. The performance of DOTSeq was benchmarked using simulation datasets against four existing tools for detecting differential translation: deltaTE (*15*), RiboDiff v0.2.2 (*13*), Riborex v2.4.0 (*14*), and Xtail v.1.20 (*12*). The DOU and DTE modules of DOTSeq were evaluated independently.

Performance was assessed using iCOBRA v1.37.1 (*27*), which was used to generate FDR-TPR curves. The onlyshared argument in the calculate_performance function was set to FALSE as default, allowing all features in the truth table to be used, irrespective of whether they were detected by all methods.

Overlap between significant ORFs identified by the DOU and DTE modules was visualised using eulerr v7.0.2 (*38*). Additional plots were produced using the base R functions (*39*) and ggplot2 v3.5.2 (*40*).

Code availability

DOTSeq is available at https://github.com/compgenom/DOTSeq.git.

Author contributions

C.S.L. conceptualised and led the study. C.S.L. developed the statistical framework and conducted data analysis. C.S.L. and G.S.W.C. wrote the manuscript.

Funding
This work was supported by a Marsden Fund Fast-Start Grant (MFP-UOO-2111) from Government funding, administered by the Royal Society of New Zealand Te Apārangi, and by research funds from a University of Otago Research Grant and the Otago School of Biomedical Sciences Dean's Fund awarded to C.S.L.

Conflicts of interest: The authors declare that they have no conflicts of interest.

REFERENCES

1. P. Sieber, M. Platzer, S. Schuster, The Definition of Open Reading Frame Revisited. *Trends Genet* **34**, 167–170 (2018).

2. J. M. Mudge, J. Ruiz-Orera, J. R. Prensner, M. A. Brunet, F. Calvet, I. Jungreis, J. M. Gonzalez, M. Magrane, T. F. Martinez, J. F. Schulz, Y. T. Yang, M. M. Albà, J. L. Aspden, P. V. Baranov, A. A. Bazzini, E. Bruford, M. J. Martin, L. Calviello, A.-R. Carvunis, J. Chen, J. P. Couso, E. W. Deutsch, P. Flicek, A. Frankish, M. Gerstein, N. Hubner, N. T. Ingolia, M. Kellis, G. Menschaert, R. L. Moritz, U. Ohler, X. Roucou, A. Saghatelian, J. S. Weissman, S. van Heesch, Standardized annotation of translated open reading frames. *Nat Biotechnol* **40**, 994–999 (2022).

3. E. W. Deutsch, L. W. Kok, J. M. Mudge, J. Ruiz-Orera, I. Fierro-Monti, Z. Sun, J. G. Abelin, M. M. Alba, J. L. Aspden, A. A. Bazzini, E. A. Bruford, M. A. Brunet, L. Calviello, S. A. Carr, A.-R. Carvunis, S. Chothani, J. Clauwaert, K. Dean, P. Faridi, A. Frankish, N. Hubner, N. T. Ingolia, M. Magrane, M. J. Martin, T. F. Martinez, G. Menschaert, U. Ohler, S. Orchard, O. Rackham, X. Roucou, S. A. Slavoff, E. Valen, A. Wacholder, J. S. Weissman, W. Wu, Z. Xie, J. Choudhary, M. Bassani-Sternberg, J. A. Vizcaíno, N. Ternette, R. L. Moritz, J. R. Prensner, S. van Heesch, High-quality peptide evidence for annotating non-canonical open reading frames as human proteins, *bioRxiv* (2024). https://doi.org/10.1101/2024.09.09.612016.

4. R. P. Hellens, C. M. Brown, M. A. W. Chisnall, P. M. Waterhouse, R. C. Macknight, The Emerging World of Small ORFs. *Trends Plant Sci* **21**, 317–328 (2016).

5. Q. Wu, M. Wright, M. M. Gogol, W. D. Bradford, N. Zhang, A. A. Bazzini, Translation of small downstream ORFs enhances translation of canonical main open reading frames. *EMBO J* **39**, e104763 (2020).

6. C. Baena-Angulo, A. I. Platero, J. P. Couso, Cis to trans: small ORF functions emerging through evolution. *Trends Genet* **41**, 119–131 (2025).

7. K. M. Vattem, R. C. Wek, Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc Natl Acad Sci U S A* **101**, 11269–11274 (2004).

8. P. D. Lu, H. P. Harding, D. Ron, Translation reinitiation at alternative open reading frames regulates gene expression in an integrated stress response. *J Cell Biol* **167**, 27–33 (2004).

9. G. Neill, G. R. Masson, A stay of execution: ATF4 regulation and potential outcomes for the integrated stress response. *Front Mol Neurosci* **16**, 1112253 (2023).

10. A. M. Smirnova, V. Hronová, M. P. Mohammad, A. Herrmannová, S. Gunišová, D. Petráčková, P. Halada, Š. Coufal, M. Świrski, J. Rendleman, K. Jendruchová, M. Hatzoglou, P. Beznosková, C. Vogel, L. S. Valášek, Stem-loop-induced ribosome queuing in the uORF2/ATF4 overlap fine-tunes stress-induced human ATF4 translational control. *Cell Rep* **43**, 113976 (2024).

11. C. S. Lim, A. K. Gibbon, A. T. Tran Nguyen, G. S. W. Chieng, C. M. Brown, RIBOSS detects novel translational events by combining long- and short-read transcriptome and translatome profiling. *Brief Bioinform* **26** (2025).

12. Z. Xiao, Q. Zou, Y. Liu, X. Yang, Genome-wide assessment of differential translations with ribosome profiling data. *Nat Commun* **7**, 11194 (2016).

13. Y. Zhong, T. Karaletsos, P. Drewe, V. T. Sreedharan, D. Kuo, K. Singh, H.-G. Wendel, G. Rätsch, RiboDiff: detecting changes of mRNA translation efficiency from ribosome footprints. *Bioinformatics* **33**, 139–141 (2017).

14. W. Li, W. Wang, P. J. Uren, L. O. F. Penalva, A. D. Smith, Riborex: fast and flexible identification of differential translation from Ribo-seq data. *Bioinformatics* **33**, 1735–1737 (2017).

15. S. Chothani, E. Adami, J. F. Ouyang, S. Viswanathan, N. Hubner, S. A. Cook, S. Schafer, O. J. L. Rackham, deltaTE: Detection of Translationally Regulated Genes by Integrative Analysis of Ribo-seq and RNA-seq Data. *Curr Protoc Mol Biol* **129**, e108 (2019).

16. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

17. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).

18. S. Anders, A. Reyes, W. Huber, Detecting differential usage of exons from RNA-seq data. *Genome Res* **22**, 2008–2017 (2012).

19. J. M. Mudge, S. Carbonell-Sala, M. Diekhans, J. G. Martinez, T. Hunt, I. Jungreis, J. E. Loveland, C. Arnan, I. Barnes, R. Bennett, A. Berry, A. Bignell, D. Cerdán-Vélez, K. Cochran, L. T. Cortés, C. Davidson, S. Donaldson, C. Dursun, R. Fatima, M. Hardy, P. Hebbar, Z. Hollis, B. T. James, Y. Jiang, R. Johnson, G. Kaur, M. Kay, R. J. Mangan, M. Maquedano, L. M. Gómez, N. Mathlouthi, R. Merritt, P. Ni, E. Palumbo, T. Perteghella, F. Pozo, S. Raj, C. Sisu, E. Steed, D. Sumathipala, M.-M. Suner, B. Uszczynska-Ratajczak, E. Wass, Y. T. Yang, D. Zhang, R. D. Finn, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, A. Kundaje, B. Paten, M. L. Tress, E. Birney, F. J. Martin, A. Frankish, GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Res* **53**, D966–D975 (2025).

20. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

21. M. McGillycuddy, D. I. Warton, G. Popovic, B. M. Bolker, Parsimoniously fitting large multivariate random effects in glmmTMB. *J. Stat. Softw.* **112** (2025).

22. Estimated Marginal Means, aka Least-Squares Means. https://rvlenth.github.io/emmeans/.

23. M. Stephens, False discovery rates: a new deal. *Biostatistics* **18**, 275–294 (2017).

24. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

25. A. C. Frazee, A. E. Jaffe, B. Langmead, J. T. Leek, Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics* **31**, 2778–2784 (2015).

26. J. Ly, K. Xiang, K.-C. Su, G. B. Sissoko, D. P. Bartel, I. M. Cheeseman, Nuclear release of eIF1 restricts start-codon selection during mitosis. *Nature* **635**, 490–498 (2024).

27. C. Soneson, M. D. Robinson, iCOBRA: open, reproducible, standardized and live method benchmarking. *Nat Methods* **13**, 283 (2016).

28. M. E. Tanenbaum, N. Stern-Ginossar, J. S. Weissman, R. D. Vale, Regulation of mRNA translation during mitosis. *Elife* **4** (2015).

29. M. I. Dobrikov, M. Shveygert, M. C. Brown, M. Gromeier, Mitotic phosphorylation of eukaryotic initiation factor 4G1 (eIF4G1) at Ser1232 by Cdk1:cyclin B inhibits eIF4A helicase complex binding with RNA. *Mol Cell Biol* **34**, 439–451 (2014).

30. A. Contreras, C. Perea-Resa, Transcriptional repression across mitosis: mechanisms and functions. *Biochem Soc Trans* **52**, 455–464 (2024).

31. Q. Hao, X. Zong, Q. Sun, Y.-C. Lin, Y. J. Song, S. Hashemikhabir, R. Y. C. Hsu, M. Kamran, R. Chaudhary, V. Tripathi, D. K. Singh, A. Chakraborty, X. L. Li, Y. J. Kim, A. V. Orjalo, M. Polycarpou-Schwarz, B. S. Moriarity, L. M. Jenkins, H. E. Johansson, Y. J. Zhu, S. Diederichs, A. Bagchi, T. H. Kim, S. C. Janga, A. Lal, S. G. Prasanth, K. V. Prasanth, The S-phase-induced lncRNA SUNO1 promotes cell proliferation by controlling YAP1/Hippo signaling pathway. doi: 10.7554/eLife.55102 (2020).

32. P. K. Jaiswal, S. Koul, P. S. T. Shanmugam, H. K. Koul, Eukaryotic Translation Initiation Factor 4 Gamma 1 (eIF4G1) is upregulated during Prostate cancer progression and modulates cell growth and metastasis. *Scientific Reports* **8**, 1–12 (2018).

33. Y. Lu, S. Yu, G. Wang, Z. Ma, X. Fu, Y. Cao, Q. Li, Z. Xu, Elevation of EIF4G1 promotes non-small cell lung cancer progression by activating mTOR signalling. *Journal of Cellular and Molecular Medicine* **25**, 2994–3005 (2021).

34. M. S. Köberlin, Y. Fan, C. Liu, M. Chung, A. F. M. Pinto, P. K. Jackson, A. Saghatelian, T. Meyer, A fast-acting lipid checkpoint in G1 prevents mitotic defects. *Nat Commun* **15**, 2441 (2024).

35. H.-Y. Joo, L. Zhai, C. Yang, S. Nie, H. Erdjument-Bromage, P. Tempst, C. Chang, H. Wang, Regulation of cell cycle progression and gene expression by H2A deubiquitination. *Nature* **449**, 1068–1072 (2007).

36. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10 (2011).

37. G. Perez, G. P. Barber, A. Benet-Pages, J. Casper, H. Clawson, M. Diekhans, C. Fischer, J. N. Gonzalez, A. S. Hinrichs, C. M. Lee, L. R. Nassar, B. J. Raney, M. L. Speir, M. J. van Baren, C. J. Vaske, D. Haussler, W. J. Kent, M. Haeussler, The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res* **53**, D1243–D1249 (2025).

38. J. Larsson, Area-Proportional Euler and Venn Diagrams with Ellipses [R package eulerr version 7.0.2]. (2024).

39. The R Project for Statistical Computing. https://www.R-project.org/.

40. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer Science & Business Media, 2009).