Quantification of translation uncovers the functions of the alternative transcriptome

Lorenzo Calviello, Antje Hirsekorn and Uwe Ohler

Translation has a fundamental function in defining the fate of the transcribed genome. RNA-sequencing (RNA-seq) data enable the quantification of complex transcript mixtures, often detecting several transcript isoforms of unknown functions for one gene. Here, we describe ORFquant, a method to annotate and quantify translation at the level of single open reading frames (ORFs), using information from Ribo-seq data. By developing an approach for transcript filtering, we quantify translation transcriptome-wide, revealing translated ORFs on multiple isoforms per gene. For most genes, one ORF represents the dominant translation product, but we also detect genes with translated ORFs on multiple transcript isoforms, including targets of RNA surveillance mechanisms. Measuring translation across human cell lines reveals the extent of gene-specific differences in protein production, supported by steady-state protein abundance estimates. Computational analysis of Ribo-seq data with ORFquant (https://github.com/lcalviello/ORFquant) provides insights into the heterogeneous functions of complex transcriptomes.

Studying gene expression allows us to understand the functions of different molecules and regulatory sequence elements, whether they act at the level of transcription, the transcribed RNA, or the encoded protein. To ensure correct protein synthesis, transcriptional and post-transcriptional regulatory programs determine the identity and amount of mature RNA templates. The translation process ensures the correct identity and amount of synthesized proteins.

The ribosome is the main actor of the translation process, a complex ribonucleoparticle that is not only able to synthesize proteins, but also acts as a quality-control platform for both the nascent peptide and the translated mRNA. Several RNA surveillance mechanisms are known to occur cotranslationally, and their importance for different processes such as differentiation or disease has been investigated.

Ribosome profiling (Ribo-seq) has made it possible to pinpoint the positions of actively translating ribosomes transcriptome-wide, using ribosome footprinting coupled to RNA-seq. In the past decade, Ribo-seq has been extensively used to investigate the molecular mechanisms acting on the ribosome and to identify the entire ensemble of translated regions (the translatome) in multiple organisms and conditions. The resulting rich data sets have triggered a plethora of dedicated analysis methods, which exploit distinct features of Ribo-seq profiles to confidently identify translated ORFs.

In this context, many reports have focused on whether small translated regions are hidden in long noncoding RNAs, with less attention given so far to account for the presence of multiple transcript isoforms per gene.

Transcript diversity can result either from alternative splicing or from alternative transcription start or poly-adenylation site usage, and it is now commonly profiled by RNA-seq experiments, which measure steady-state abundance of (m)RNAs. Large-scale efforts have uncovered the wide spectrum of alternative transcript isoforms, with many being lowly expressed or presenting incomplete ORFs. The contribution of this transcript heterogeneity to an expanded translatome is therefore an intensely debated topic, with much of transcript and protein abundance apparently explained by a single dominant transcript per gene.

The mere presence of multiple transcripts does not imply the presence of a distinct, functional protein translated from each transcript isoform; transcripts might be retained in the nucleus, selectively degraded, or undergo translational repression. From a technical point of view, RNA-seq experiments quantify a complex scenario in which, depending on the protocol used, alternative transcripts may also reflect different steps of RNA processing and not the stable, steady-state cytoplasmic pool of mRNAs available to the ribosome. From a different direction, shotgun proteomics approaches are only recently providing the sensitivity to detect tens of thousands of proteins from a single sample and rarely reach the depth required to investigate alternative protein isoforms.

To close this gap, we developed a strategy to identify and quantify translation on the subset of transcripts that are expressed in the cell. A recent study presented a proof of principle for validating the presence of multiple transcript isoforms in Ribo-seq data, underlining the potential of isoform-aware analysis approaches to fully define the translatome. Following up on this premise, we describe ORFquant, a Ribo-seq analysis approach that detects and quantifies ORF translation across multiple transcript isoforms and zooms in on the potential roles of alternative transcripts.

Results

The ORFquant approach to annotate and quantify translation. Our approach is based on the premise that, despite their short length, Ribo-seq reads are sufficient to support a given set of alternative transcripts (Fig. 1a,b). Single-nucleotide positions corresponding to the peptidyl site for each ribosome (P-site positions), and junction reads are first extracted from the Ribo-seq alignment (Methods), then mapped to flattened gene models from a given annotation (Fig. 1b). In this way, transcript features (for example,
exonic bins or splice junctions) are designated as unique or shared across multiple annotated transcript isoforms.

We first retain a subset of annotated transcripts, which is sufficient to explain all of the observed P sites or junction reads and reduce the occurrence of exons and junctions with no signal, using an Occam’s razor strategy (Methods). In brief, a transcript is filtered out if its features supported Ribo-seq signal can be explained by another transcript with better support (that is, containing more
features with Ribo-seq support or fewer unsupported features). As Ribo-seq reads are largely found in 5′ untranslated regions (UTRs) and coding regions only, this approach might not distinguish between transcripts differing in their 3′ UTR.

This simple yet effective selection strategy leads to a significantly reduced number of transcripts: the observed Ribo-seq signal can be explained by one to three transcript structures for most genes, without showing a strong bias for expression level (Fig. 1c and Extended Data Fig. 1a). This selection dramatically improves the assignments of both exons and junctions to transcripts (Fig. 1d): when considering covered exons or junctions (defined as having at least one Ribo-seq read mapped to them), ~64% of exons mapped to one or two transcripts, compared to ~29% when no selection is performed. Considering only annotated protein-coding transcripts does not substantially improve the mapping of covered features, and it ignores the presence of covered exons and junctions unique to noncoding transcripts. Next, we detect translated ORFs of de novo in each of the selected transcripts using frame preference and the multi taper\textsuperscript{16–18} test to select in-frame signal showing 3-nucleotide (nt) periodicity (Methods), a hallmark of active translation elongation. Detected ORFs are filtered using the same strategy used for transcript filtering.

After calculating coverage on unique and shared ORF features (exonic bins and splice junctions within ORF boundaries), a scaling factor between 0 and 1 is determined using the coverage on unique ORF features or the amount of overlap between ORFs when no unique feature can be detected (Methods). This scaling factor represents the fraction of Ribo-seq signal that can be assigned to that ORF. The scaled number of P sites is then normalized by the ORF length to arrive at transcripts per million (TPM)\textsuperscript{19}–like values, named ORFs per million (ORFs_pM). Moreover, we calculate the relative contribution of each ORF to the overall translation output of each gene (ORF_pct_P_sites, or percentage of gene translation). Additional filtering step discards poorly translated ORFs. ORFs are then annotated according to their position relative to their host transcript, to other detected ORFs in the same gene, and to annotated coding sequence (CDS) regions.

Applying ORFquant, we quantified translation for ~20,800 ORFs in ~12,300 genes profiled in a Ribo-seq data set from the human HEK293 cell line\textsuperscript{4}. Most genes (7,732; Fig. 1e) showed only one translated ORF, with another >5,000 genes showing translation of multiple ORFs. Upon closer inspection (Fig. 1f), we observed that for the majority of genes (~80%), the most translated (that is, major) ORF could explain >80% of the total gene translation, with only 444 genes for which the major ORF explained <50% of the translational output. We did not observe a clear dependency between number of detected ORFs (or percentage of translation of major ORF) and overall Ribo-seq coverage, with the exception of the few dozen genes for which the major ORF accounted for little of the total gene output (Extended Data Fig. 1b,c).

In principle, the final set of ORFs can be provided to any algorithm for transcript quantification. To demonstrate the effectiveness of our simple approach, we compared our estimates with the ones calculated by RSEM\textsuperscript{19}, a well-known statistical approach devoted to transcript quantification. We observed good correlation between the two methods in their estimates of the relative contribution of each ORF to the total output (Extended Data Fig. 1d). Additionally, we observed how RSEM quantification estimates showed high uncertainty (Methods) for ORFs where few unique features are present, which are cases in which ORFquant assigns low translation estimates to the major ORF (Extended Data Fig. 1c). Despite major differences in their quantification strategy (Discussion), both ORFquant and RSEM showed similar performances in determining the contribution of each ORF to the total gene translation output.

To illustrate the consistency of our translation estimates, we annotated the ORF structures with respect to the major (most translated) ORFs in each gene; this allowed us to detect genomic regions (for example, different alternative splice sites) where the Ribo-seq signal should reflect different quantitative estimates of translation coming from different ORF(s). Aggregate profiles of Ribo-seq coverage closely reflected the expected pattern calculated by ORFquant (Fig. 1g and Extended Data Fig. 2). Taken together, the translation of a major ORF accounts for >80% of total gene translation for most of the genes, but distinct translated ORFs are detected from multiple translated transcripts for hundreds of genes.

Quantification of translation as a window into the functional relevance of alternative open reading frames. As translation is a cytoplasmic process, we expected the ensemble of transcript structures selected by ORFquant to represent bona fide cytoplasmic transcripts. To test this hypothesis, we performed a differential exon usage analysis\textsuperscript{20} using RNA-seq data from nuclear and cytoplasmic extracts in HEK293 cells\textsuperscript{21}. Most exons unique to discarded structures showed marked nuclear localization (log_2-fold change (FC) > 0), whereas exons of selected transcripts showed prominent cytoplasmic enrichment (Fig. 2a). Translated transcripts showed more marked cytoplasmic localization. An example of the selection strategy discarding precursor-mRNA structures in favor of cytoplasmic transcripts is shown in Fig. 2b.

When examining the GENCODE annotation\textsuperscript{22} of the transcripts hosting de novo–identified ORFs, we noticed ~2,000 ORFs in noncoding transcript isoforms of protein-coding genes, most of which lacked annotated ORFs (Fig. 2c). Compared with ORFs in annotated protein-coding transcript isoforms, these ORFs showed much lower translation, accounting for a median of 6.8% of gene translation, compared with 87% for ORFs that fully matched annotated CDSs. More than 3,500 N-terminal truncation events were also detected, showing high levels of translation. Upstream ORFs (uORFs) and other small ORFs showed low signal, albeit high when normalized by their lengths. In annotated noncoding genes, we detected 181 ORFs from annotated pseudogenes and 620 ORFs from other noncoding RNA genes, with overall lower translation levels than protein-coding RNAs (Fig. 2c).

Analysis of a deep polysome profiling data set (Trip-Seq\textsuperscript{23}) from the same cell line showed that the quantitative estimates of translation agreed with distinct polysome profiles (Fig. 2d,e and Extended Data Fig. 3); exons uniquely mapping to transcripts harboring lowly translated ORFs accumulated in low polysomes and were depleted in heavier polysomal fractions. Conversely, highly translated transcripts showed sustained levels also in heavy polysomes. Despite the fundamental differences between polysome profiling and Ribo-seq in representing the translated transcriptome, the two techniques therefore agreed in detecting quantitative differences in the translation of multiple transcripts per gene.

The presence of numerous lowly translated ORFs detected in noncoding transcript isoforms (Fig. 2c) suggested inefficient translation or low steady-state abundance of the translated transcript. We wondered whether transcripts subject to RNA surveillance mechanisms (such as nonsense-mediated decay, NMD) might cause such a low but detectable Ribo-seq signal. The presence of a premature termination codon (PTC) is an important feature of many NMD targets\textsuperscript{24}, which is assumed to be recognized as such when the stop codon is located sufficiently upstream of the last splice junction, that is, when a downstream exon junction complex (EJC) is not displaced during translation elongation (Fig. 3a). To investigate the putative action of NMD on PTC-containing transcripts, we divided transcripts based on the presence of a splice site downstream of a detected ORF. A recent study mapped NMD-mediated cleavage events on the transcriptome in HEK293 cells\textsuperscript{25}, by knocking down XRN1, the exonuclease in charge of degrading the cleaved transcripts. When aligning the cleavage sites at the stop codons of (putative)
PTC- and non-PTC-containing transcripts (from the same genes), we observed a clear difference (Fig. 3b): transcripts without a PTC, that is, where all EJCSPs are presumably displaced, showed background-like signal, whereas transcripts harboring a putative PTC showed a marked degradation profile around their stop codon. The degradation signal was less pronounced when SMG6 or URF1 were also knocked down, underlining the effect of known key factors of the NMD pathway on our candidate NMD targets. A clear example of such pattern is observed on a translated ORF in the SNHG17 gene (Fig. 3c).

To further explore the dependency of NMD on the location of PTCs as well as the transcript type, we determined the number of endonucleolytic cuts at the stop codon as a function of PTC distance to the last exon–exon junction. We observed an increase in degradation for NMD candidate ORFs for all of the surveyed ORFs (including uORFs; Fig. 3d). As previously reported, ORFs in snRNA host genes (such as SNHG17, Fig. 3c) showed the highest degradation profile, whereas other categories showed a lower amount of degradation. In summary, ORFquant is able to identify mature mRNAs, quantify the translation output of different transcript isoforms from the same gene, and infer transcript-specific cytoplasmic fates.

A subset of genes translates different major ORFs in different cell lines. To investigate the patterns of alternative ORF usage across different conditions, we ran ORFquant on Ribo-seq data sets from six different human cell lines (Fig. 4a, Supplementary Table 1 and Supplementary Data 1), with newly generated data for K562 and HepG2 cell lines complementing previously published libraries from HEK293, HeLa, U2OS, and Jurkat cells. For each data set, we observed the same trend described in Fig. 1d, with most genes showing translation of one major ORF and hundreds of genes showing sustained translation of multiple ORFs, with a weak dependency on the overall Ribo-seq signal (Extended Data Fig. 4). Across all cell lines, we detected ORF translation for ~17,000 genes (excluding pseudogenes), with ~89% of them annotated as protein-coding genes.

For each gene and cell line, we defined the major ORF as the most translated ORF from a gene, regardless of its positional features and existing ORF annotation. For ~77% of the quantified genes, the same ORF was consistently identified as the major translated ORF in all of the assayed cell lines (Fig. 4b). For ORFs in noncoding RNAs, we detected a more cell-specific pattern of major ORF usage. However, a few dozen noncoding genes showed translation of the same major ORF. One such example is, again, SNHG17, where the...
Fig. 3 | De novo annotation of NMD candidates. a, Schematic annotation of NMD candidates. b, Aggregate profiles of 5′ fragments around stop codons of NMD candidates (continuous lines) and control ORFs (dotted lines) from the same genes. c, Example of a (not previously annotated) translated ORF in the SNHG77 gene; coverage tracks of 5′-end fragments in control and different knockdown experiments are shown in the bottom track. d, Number of 5′ fragments observed in an XRN1-knockdown experiment around stop codons (y axis) versus the distance between stop codons and the last exon–exon junction (x axis) for ORFs on different transcripts. Smoothing was carried out by a generalized additive model (gam in R, with default parameters). The red vertical line indicates 50 nucleotides upstream of the last exon–exon junction. Source data for b,d are available online.

As expected, genes translated in all cell lines showed overall higher Ribo-seq signal. However, we did not observe a clear dependence between number of distinct major ORFs across cell lines and overall gene translation (Extended Data Fig. 5). Genes exhibiting translation of multiple major ORFs showed an enrichment for Gene Ontology (GO) categories like GTPase regulator (Extended Data Fig. 5), a category also enriched in genes expressing multiple major transcript isoforms across human tissues39. Two or more distinct major ORFs were identified in 18% and 5% of genes, representing candidate major ORF-switching events across cell lines (Fig. 4b). Upon closer examination, we observed that genes translating multiple major ORFs also showed a more complex mixture of translated ORFs. Consequently, translation of the major ORF for those genes accounted for a lower percentage of total gene translation (Fig. 4b).

ORF diversity is created by different mechanisms: differences in alternative splicing of internal coding exons, alternative transcriptional start sites, or alternative usage of last exons (Fig. 4d and Extended Data Figure 6). However, in ~40% of the cases, distinct major ORFs translated across cell lines showed a low degree of overlap (Fig. 4e) despite coming from the same genes, that is, largely unrelated to differences in local alternative splicing events. This low overlap reflected the presence of alternative usage of uORFs or other small ORFs (Fig. 4f and Extended Data Figure 6), which can represent the major translation product of a gene in specific cell lines.

Taken together, these translation estimates indicate that the presence of one dominant ORF agrees across multiple cell lines for the majority of genes. For ~20% of the translated genes, however, highly translated small ORFs and several transcripts expressed at sustained levels create a substantial level of complexity in protein synthesis, with distinct ORFs accounting for the majority of the gene translational output in different cell lines.

Agreement between protein abundance and synthesis estimates depends on proteome coverage and transcriptome complexity. Ribo-seq reflects the density of elongation-competent 80S ribosomes, and thus, active protein synthesis, but an increased signal at a specific location may also represent stalled, inactive ribosomes. We therefore examined whether our translation quantification reflects the abundance of the synthesized protein product. Using a comprehensive custom protein database derived from the set of identified ORFs (Extended Data Figure 7 and Supplementary Data 2), we estimated proteome-wide steady-state protein abundance using published deep MS data9,21 for the same cell lines outlined above (Fig. 4a). We detected between 7,000 and 8,000 proteins per cell line (Extended Data Fig. 8 and Supplementary Data 2), and performed label-free quantification using signal from unique peptides only (Methods). To estimate the ability of both
Fig. 4 | Diversity in gene translation across cell lines. a, Workflow for the analysis of the different data sets. b, Number of genes (top), average number of detected ORFs (middle), and average percentage of gene translation (bottom) for the number of cell lines in which the gene harbored a detected ORF. Colors indicate the gene biotype. Genes translating one or more distinct major ORFs across cell lines are shown in different panels. The maximum width of each violin plot is the same for each panel, and the median value is shown as a black bar. c, Detected ORFs and Ribo-seq signal in the SNHG17 gene in HeLa cells. d, Detected ORFs and Ribo-seq signal in the C5orf63 gene in HepG2 and U2OS cells. e, Distribution of overlap between multiple major ORFs from the same gene. f, Length (in nucleotides) of major ORFs with low and high overlap. g, Enrichment of different categories of major ORFs with high and low degrees of genomic overlap; numbers of ORFs with high and low overlap are shown in parentheses. Source data for b,e–g are available online.

When comparing the fraction of total gene translation to the fraction of total protein abundance for the few dozen genes with MS matches to multiple detected protein isoforms, we observed a correlation of 0.58 (Fig. 5c). Here, few proteins harbored >8 uniquely mapping peptides, thus limiting our ability in reliably estimating their abundance (Fig. 5a). We observed lower correlations when skipping the ORF-specific scaling step during translation quantification, highlighting the importance of accounting for the presence of multiple translated ORFs per gene (Fig. 5c). The same pattern was observed for the other cell lines analyzed (Fig. 5d and Extended Data Fig. 9). A slight increase in correlations was detected when using all Ribo-seq reads (instead of uniquely mapping reads only) to derive translation estimates (Extended Data Fig. 10), likely resulting from a better quantification in repetitive regions.

Taken together, these results show excellent correlations between ORFquant quantification and steady-state protein abundance at the single ORF level, and mostly subject to the technical limitations in
While using RNA-seq to bias Ribo-seq mapping to different isoforms has been proposed as a successful strategy\textsuperscript{34,35}, we show how (Fig. 1d), allowing ORF-specific translation estimates. Exonic and junction reads to their possible transcripts of origin that are either rapidly degraded or retained in the nucleus. Our strategy to identify the subset of translated ORFs across transcript isoforms in the assayed condition or they represent pre-mRNA intermediates not present at translating ribosomes; that is, they are not expressed but discarded RNAs (Fig. 2a) indicates that these structures are not unique, annotated transcript structures are exported to the cytoplasm and eventually translated into functional proteins. Only a fraction of known, annotated transcript structures are quantified using iterative methods (such as the expectation-maximization algorithm) to resolve the mixture resulting from multiple transcripts\textsuperscript{19,32,33}. While using RNA-seq to bias Ribo-seq mapping to different isoforms has been proposed as a successful strategy\textsuperscript{14,15}, we show how Ribo-seq alone has the unique advantage of measuring exclusively cytoplasmic mRNAs, enabling a clean representation of translatable mRNAs. Moreover, resolving the mixture of multiple transcript isoforms with short-read data can be challenging for many genes, especially in the absence of coverage on unique transcript features. A top-performing algorithm designed to solve this problem with high accuracy displays high variability in its estimates for such cases (Extended Data Fig. 1). The rapidly increasing availability of full-length transcript sequence data based on long-read sequencing\textsuperscript{26} holds great promise in solving these complex scenarios.

While polysome profiling experiments (Fig. 2d) and label-free quantification of the protein product (Fig. 5b) support the Ribo-seq-based estimates of relative ORF translation levels, we believe that additional efforts can improve ORF-specific quantification of translation. A more accurate approach will have to address the issue of variable Ribo-seq coverage along the ORFs, which reflects the complex dynamics of translation. However, the impact of different features on Ribo-seq coverage, such as experimental biases, codon composition, or RNA structural features\textsuperscript{41}, remain to be understood. Our approach also uses a strict definition of ORFs that requires a canonical start codon and does not account for overlapping frames. It is still an open question how to correctly define the precise boundaries of translated elements that account for non-canonical start codons and signals from overlapping frames, such as from upstream ORFs\textsuperscript{38} or complex gene structures in compact genomes such as found in viruses and organelles.

**Discussion**

Only a fraction of known, annotated transcript structures are expressed in a specific context, and only a fraction of those structures are exported to the cytoplasm and eventually translated into functional proteins. This observation inspired us to devise a simple strategy to identify the subset of translated ORFs across transcript isoforms from Ribo-seq data, by discarding a substantial fraction of transcript structures with no support. The marked nuclear localization of annotated but discarded RNAs (Fig. 2a) indicates that these structures are not present at translating ribosomes; that is, they are not expressed in the assayed condition or they represent pre-mRNA intermediates that are either rapidly degraded or retained in the nucleus. Our strategy therefore resulted in a markedly improved mapping of Ribo-seq exonic and junction reads to their possible transcripts of origin (Fig. 1d), allowing ORF-specific translation estimates.

The quantification of transcript isoform expression is a well-studied problem in RNA-seq, with popular methods applying iterative methods (such as the expectation-maximization algorithm) to resolve the mixture resulting from multiple transcripts\textsuperscript{19,32,33}. While using RNA-seq to bias Ribo-seq mapping to different isoforms has been proposed as a successful strategy\textsuperscript{14,15}, we show how Ribo-seq alone has the unique advantage of measuring exclusively cytoplasmic mRNAs, enabling a clean representation of translatable mRNAs. Moreover, resolving the mixture of multiple transcript isoforms with short-read data can be challenging for many genes, especially in the absence of coverage on unique transcript features. A top-performing algorithm designed to solve this problem with high accuracy displays high variability in its estimates for such cases (Extended Data Fig. 1). The rapidly increasing availability of full-length transcript sequence data based on long-read sequencing\textsuperscript{26} holds great promise in solving these complex scenarios.

While polysome profiling experiments (Fig. 2d) and label-free quantification of the protein product (Fig. 5b) support the Ribo-seq-based estimates of relative ORF translation levels, we believe that additional efforts can improve ORF-specific quantification of translation. A more accurate approach will have to address the issue of variable Ribo-seq coverage along the ORFs, which reflects the complex dynamics of translation. However, the impact of different features on Ribo-seq coverage, such as experimental biases, codon composition, or RNA structural features\textsuperscript{41}, remain to be understood. Our approach also uses a strict definition of ORFs that requires a canonical start codon and does not account for overlapping frames. It is still an open question how to correctly define the precise boundaries of translated elements that account for non-canonical start codons and signals from overlapping frames, such as from upstream ORFs\textsuperscript{38} or complex gene structures in compact genomes such as found in viruses and organelles.
Our strategy enabled us to detect thousands of lowly translated ORFs in transcript isoforms of protein-coding genes that are annotated as noncoding, consistent with current models for mRNA surveillance such as NMD (Fig. 3). Similarly, we observed that many detected ORFs in noncoding RNAs show high degradation profiles at their stop codons, especially pronounced in snoRNA host genes (Fig. 3d), an important aspect to consider when addressing the protein-coding ability of transcripts based only on ribosome occupancy. Moreover, the ability to experimentally identify transcript functions de novo, also in conjunction with de novo transcriptome reconstruction methods, allows comprehensive reannotation of the transcribed genome when pre-existing annotation models are deemed inadequate.

Expanding our analysis across multiple cell lines allowed us to assess the complexity of translation per gene for both coding and noncoding genes (Fig. 4b). We found the majority of genes to be translating the same major ORF (including highly translated ORFs in noncoding RNAs; Fig. 4c) and detected distinct ORFs used for the major translation product in different cell lines in thousands of genes. These genes showed an overall more complex pattern of transcript expression, with sustained translation of many transcripts, thus posing difficulty in defining clear isoform switching events. In this context, the presence of highly translated small ORFs in protein-coding genes (Extended Data Fig. 6), which may play gene regulatory roles rather than expand the proteome, adds further complexity. Unfortunately, the limited amount of data at hand (often lacking replicate information) and the heterogeneity of protocols adopted by different labs pose challenges in identifying the mechanisms promoting diversity (or lack thereof) in protein synthesis for each gene.

By deriving the entire set of translated ORFs from Ribo-seq data, irrespective of protein-coding status, we decided to apply a global false discovery rate (FDR) cutoff for peptide identification (Methods). Different strategies for group-specific or local FDR control might represent a viable alternative to improve on peptide identification and quantification. By providing an approach requiring Ribo-seq data alone, we believe that ORFQuant can easily complement other dedicated pipelines for proteogenomics applications, such as the PROTEOFORMER pipeline, adding novel features such as assessment of NMD sensitivity, transcript-level and genome-level ORF annotation, and ORF-specific quantification estimates.

We observed a substantial agreement between our estimates of translation and steady-state protein abundance. The level of agreement between mRNAs and proteins has been subject to intense debate; our results indicate that for thousands of genes, shotgun proteomics experiments and sequencing of ribosome-occupied RNA fragments do show excellent agreement, albeit with expected dependencies on the reliability with which we can quantify the levels of translation and protein abundance (Fig. 5a). An increasing availability of Ribo-seq and proteomics data in a single controlled environment will improve our understanding of the relationships between translation and protein synthesis and help pinpoint cases in which this correlation robustly deviates from expectations.

While our analyses provide a promising starting point for the investigation of transcript-specific protein production, the current scarcity of matching data specifically limits our ability to validate the translation of alternative protein isoforms per gene. A recent study demonstrated how protein isoforms engage with distinct protein-protein interaction networks, with such interactions being as different as the ones involving proteins from distinct genes. With both proteomics and transcriptomics techniques rapidly advancing at a fast pace, our study demonstrates the unique advantage of ribosome profiling in characterizing and quantifying cytoplasmic gene expression programs, at the interface between RNA and protein.
28. Jang, C., Lahens, N. F., Hogness, J. B. & Sehgal, A. Ribosome profiling reveals an important role for translational control in circadian gene expression. *Genome Res.* **25**, 1836–1847 (2015).

29. Tapial, J. et al. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Res.* **27**, 1759–1768 (2017).

30. Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics* **11**, M111.014050 (2012).

31. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).

32. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

33. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

34. Wang, H., McManus, J. & Kingsford, C. Isoform-level ribosome occupancy estimation guided by transcript abundance with Ribomap. *Bioinformatics* **32**, 1880–1882 (2016).

35. Reixachs-Solé, M., Ruiz-Orra, J., Albà, M. M. & Eyras, E. Ribosome profiling at isoform level reveals evolutionary conserved impacts of differential splicing on the proteome. *Nat. Commun.* **11**, 1768 (2020).

36. Legnini, I., Alles, J., Ayoub, S., Karaikos, N. & Rajewsky, N. FLAM-seq: full-length mRNA sequencing reveals principles of poly(A) tail length control. *Nat. Methods* **16**, 879–886 (2019).

37. Fang, H. et al. Scikit-ribo enables accurate estimation and robust modeling of translation dynamics at codon resolution. *Cell Syst.* **6**, 180–191.e4 (2018).

38. Michel, A. M. et al. Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res.* **22**, 2219–2229 (2012).

39. Kovaka, S. et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).

40. Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **11**, 1114–1125 (2014).

41. Verbruggen, S. et al. PROTEOFORMER 2.0: further developments in the ribosome profiling-assisted proteogenomic hunt for new proteoforms. *Mol. Cell. Proteomics* **18**, S126–S140 (2019).

42. Airoldi, E., Franks, A. & Slavov, N. Post-transcriptional regulation across human tissues. *PLoS Comput. Biol.* **13**, e1005535 (2017).

43. Yang, X. et al. Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* **164**, 805–817 (2016).

44. Collins, B. C et al. Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat. Commun.* **8**, 291 (2017).

45. Martens, L. & Vizcaíno, V. A. A golden age for working with public proteomics data. *Trends Biochem. Sci.* **42**, 333–341 (2017).

46. Garaldé, D. R. et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2020.
Methods

ORFquant – transcript and ORF filtering. Gene models from the GTF annotation are flattened to obtain coordinates about exonic bins or junctions, together with the set of transcripts they map to. Next, P-site positions and junction reads (from all read lengths) are mapped to such features, to obtain positive (with at least one read count) or negative features (with no reads). Internal features are then defined as features contained between the coordinates of the first (most upstream) and last (most downstream) positive features.

The filtering procedure is then applied. Initially, an empty vector of positive features is created, and such a vector is updated at each step, adding (when present) new positive features contained by the analyzed transcript. After creating the empty vector, the set of annotated transcripts is analyzed, applying the following rules for each transcript \( T_x \):

1. \( T_x \) contains a novel positive feature: \( T_x \) is selected, and each previously selected \( T_x \) is re-analyzed: If all the positive features of \( T_x \) are also contained in \( T_x \), \( T_x \) is discarded.
2. \( T_x \) does not contain a novel positive feature: \( T_x \) is initially selected, but it is compared with each previously selected structure \( T_x \). Two possible scenarios are evaluated:
   i. All the positive features of \( T_x \) are also contained in \( T_x \); if \( T_x \) has more positive features than \( T_x \), or fewer negative internal features than \( T_x \), \( T_x \) is discarded.
   ii. All the positive features of \( T_x \) are also contained in \( T_x \); if \( T_x \) has fewer negative internal features than \( T_x \), \( T_x \) is discarded.

This greedy strategy reduces the number of transcripts that is necessary to cover all the positive features (features with reads), trying to minimize the presence of negative features (features with no reads). We select ORFs following the same rules, this time using exonic bins and splice junctions derived from the ORF structures.

ORFquant – ORF finding. As in the RiboTaper\textsuperscript{18} method, only ATG is considered as a potential start codon, and the \( P \) value for the multitaper method applied to the candidate ORF P-site track must be below 0.05. To select ORFs with in-frame P sites and account for local off-frame effects, we require the average signal on each covered codon to be >50% in frame. The same strategy is used to select the start codon for each ORF, requiring >50% average in-frame codon signal between each candidate ATG and the next.

ORFquant – ORF quantification. After the ORF-finding step, ORF filtering and quantification is subsequently performed, using the length-normalized Ribo-seq coverage, \( \text{cov} \), on each ORF feature:

\[
\text{cov} = \frac{\# \text{ reads}}{\text{Length}}
\]

P-site positions are used to calculate coverage on exonic regions, and spliced reads are used for junctions. Length is set to 60 nt for junctions, according to the possible nucleotide space covered by a spliced read of ~30 nt.

A feature \( F \) can be unique to one ORF or shared between multiple ORFs. For each ORF, we calculate the average coverage on unique features, \( \text{AvCovUn} \), using the coverage \( \text{cov}_F \), on each of the unique features \( F \).

\[
\text{AvCovUn} = \frac{\sum F \cdot \text{cov}_F}{\# F}
\]

The same calculation is performed for all features, \( F \), mapping to the ORF:

\[
\text{AvCovAll} = \frac{\sum F \cdot \text{cov}_F}{\# F}
\]

A scaling factor \( \text{CovF} \) (with a minimum value of 0 and a maximum of 1) is calculated for each ORF, using the ratio between \( \text{AvCovUn} \) and \( \text{AvCovAll} \). Such scaling factor represents the fraction of Ribo-seq signal that can be attributed to the ORF.

\[
\text{CovF} = \frac{\text{AvCovUn}}{\text{AvCovAll}}
\]

When no unique feature is present in one ORF (all regions are shared with other ORFs), the signal at each feature is adjusted using the quantification performed on other ORFs, as follows: the coverage \( \text{AdCov}_F \) on each feature \( F \) attributed to that ORF is calculated by subtracting the expected signal coming from other ORFs (\( \text{ORF}_{\text{rest}} \)) overlapping that feature, using their scaling factors. In such cases, the calculation of the adjusted coverage for each feature \( F \) is as follows:

\[
\text{AdCov}_F = \text{cov}_F - \left( \text{cov}_F \times \sum \text{ORF}_{\text{rest}} \cdot \text{CovF}_{\text{ORF}_{\text{rest}}} \right)
\]

After calculating the adjusted coverage for each feature, the average of such coverage values is calculated.

\[
\text{AvAdCov} = \frac{\sum F \cdot \text{AdCov}_F}{\# F}
\]

The final scaling factor is here defined by the ratio of the adjusted coverage (coverage belonging to the ORF) to the total coverage (coverage coming from all ORFs).

\[
\text{C}_{\text{ORF}} = \frac{\text{AvAdCov}}{\text{AvCovAll}}
\]

If no unique region is present in any detected ORF in the gene (all regions are shared among ORFs and no \( \text{CovF} \) value can be initially calculated), the scaling factor is derived assuming uniform Ribo-seq coverage on each ORF. The shared coverage \( \text{ShCovF} \) is now simply calculated, dividing it by the number of ORF\text{rest} mapping to the feature \( F \).

\[
\text{ShCovF} = \frac{\# \text{ reads}}{\text{Length}_F} / \# \text{ ORF}_{\text{rest}},
\]

\[
\text{AvShCov} = \frac{\sum F \cdot \text{ShCov}_F}{\# F}
\]

The scaling factor is again derived dividing the average shared coverage (attributed to the ORF) to total average coverage.

\[
\text{C}_{\text{ORF}} = \frac{\text{AvCovSh}}{\text{AvCovAll}}
\]

After the calculation of \( \text{C}_{\text{ORF}} \), the adjusted number of \( P \) sites for each ORF (\( \text{P}_{\text{ORF}} \)) is calculated using the raw number of \( P \) sites mapping to the ORF multiplied by the scaling factor, to obtain ORF-specific quantification estimates.

\[
\text{P}_{\text{ORF}} = \text{P}_{\text{sites}} \times \text{C}_{\text{ORF}}
\]

For each ORF of length \( \text{Length}_F \), the scaled numbers of \( P \) sites \( \text{P}_{\text{ORF}} \) is normalized over the entire set of detected ORFs to obtain TPM-like values, named ORFs per million (ORFs_pM), using this formula:

\[
\text{ORFs_pM}_{\text{ORF}} = \frac{\text{P}_{\text{ORF}}}{\text{Length}_F} \times 10^9 \frac{\# \text{ ORF}_{\text{rest}}}{\sum \# \text{ ORF}_{\text{rest}}}.
\]

Moreover, we calculated the contribution of each ORF to the overall translation output of a single gene. Such metric, named \( \text{ORF}_{\text{pct}} \), (or percentage of gene translation), is calculated dividing \( \text{P}_{\text{ORF}} \) by the sum of \( \text{P}_{\text{ORF}} \) of all ORFs (\# ORFs) detected in a gene.

\[
\text{ORF}_{\text{pct}}_{\text{ORF}} = \frac{\text{P}_{\text{ORF}}}{\# \text{ ORF}_{\text{rest}}} \times 10^9 \frac{\# \text{ ORF}_{\text{rest}}}{\sum \# \text{ ORF}_{\text{rest}}}
\]

Normalization by length is here not applied, as this metric wants to quantify the amount of translation per gene coming from each ORF. The \( \text{ORF}_{\text{pct}}_{\text{P}_{\text{sites}}} \) metric indicates length-normalized \( \text{ORF}_{\text{pct}}_{\text{P}_{\text{sites}}} \) values (for example, they can be high for a short, highly translated ORF).

After quantification, ORFs are subjected to a filtering step, and quantification is performed again until all ORFs are being retained.

ORFquant parameters. For all cell lines, ORFquant was run using a cutoff of 2% of total gene translation and using only uniquely mapping reads.

RSEM quantification. RSEM 1.3.1 was run in strand-specific mode on Ribo-seq data using a seed length of 20, using Bowtie2 as an aligner, and enabling the calculation of confidence intervals together with posterior mean estimates. ORFquant-derived ORF positions were used to specify the transcript sequences to use as reference. When possible, an additional 15 nucleotides were added to start and end coordinates to allow for the mapping of Ribo-seq reads. The “TPM, coefficient_of_quartile_variation” column of the RSEM output was used as a proxy to monitor the variability in RSEM quantification estimates.

Cell lines employed. K562 and HepG2 cell lines were obtained from the ATCC catalog. The K562 cell line was authenticated by Eurofins genomics. Both cell lines tested negative for mycoplasma contamination using the LookOut kit, Sigma Aldrich #P0035–1KT.

Ribosome profiling experiments. Ribo-seq was performed as described previously\textsuperscript{10} and adapted for HepG2 and K562 cell lines. A total of 5 × 10\textsuperscript{6} K562 suspension cells and two 80% confluent 10-cm tissue culture dishes of adherent HepG2 cells (DSMZ #ACC-10 and #ACC-180, respectively) were used.
Adherent cells were washed with ice-cold PBS supplemented with 100 µg/ml -cyclheximide (Sigma Aldrich #C8599) and immediately snap frozen by immersing the dishes in liquid nitrogen. The dishes were then transferred to wet ice, and 400 µl of lysis buffer (1x polyosme buffer (20 mM Tris-Cl, pH 7.4, 150 mM NaCl, 5 mM MgCl2, with 1 mM DTT (Sigma Aldrich #48316) and 100 µg/ml -cyclheximide added freshly; keep on ice), 1% (v/v) Triton X-100 (Calbiochem #648466), 25 µl 1× TURBO DNase (Life Technologies #AM2295) was immediately dripped onto the frozen cells. The cells and buffer were then scraped off and thawed by a side of the dish, mixed using a pipet tip. Suspension cells were supplemented with 100 µg/ml -cyclheximide, pelleted for 5 min at 300 × g, and washed with ice-cold PBS + 100 µg/ml -cyclheximide. The washed cell pellet was immediately snap frozen in liquid nitrogen. Four-hundred microliters of ice-cold lysis buffer was added, and the cells were put on wet ice to thaw, mixed using a pipet tip. The cells were left to lyse for 10 min on ice, followed by 10x trituration through a 26-gauge needle. After centrifugation for 10 min at 20,000g at 4°C, the clarified supernatant was transferred to a pre-cooled tube on ice. For nuclease footprinting, 400 µl of lysate was supplemented with 1000 U of RNase I (Life Tech. #AM2295) and incubated in a thermomixer set to 23 °C, shaking at 500 r.p.m. for 45 min. Footprinting was stopped by adding 260 µl of SUPERase-In (Life Technologies #2696).

To recover ribosomes, two MicroSpin S-400 HR columns (GE Healthcare #27-5140-01) per 400 µl of sample were equilibrated with a total of 3 ml of polysome buffer. The columns were drained by spinning for 10 min at 600 × g at 4°C. The cell lysate was supplemented with 200 µl of polysome buffer. The columns were drained by spinning for 4 min at 600 × g, then the cell lysate was supplemented (pH 7.4, 150 mM NaCl, 20 mM Tris-Cl) for 2 min at 60°C. Three volumes of Trizol LS (Life Technologies #10926010) were added to the flow-through, and RNA was extracted using the Direct-zol RNA Mini-Prep kit (Zymo Research #R2052) as per the manufacturer’s instructions. RNA was quantified using the Qubit RNA Broad Range Assay (Life Technologies #Q10211).

Ribosomal RNA was removed from 10 µg of footprinted RNA using the RiboZero Magnetic Gold kit (Illumina #MRG12328) as per the manufacturer’s instructions. Footprintted RNA was precipitated from the supernatant (90%) using 1.5 µl of GlycoBlue (Life Technologies #9515), 9 µl of 3 M sodium acetate, and 300 µl of ethanol by incubation for 1 h at −80°C and pelleted for 30 min at maximum speed at 4°C. The RNA pellet was dissolved in 10 µl of RNase-free water.

To recover the ribosome-protected RNA fragments, the sample was loaded onto two lanes of a 1-mm 17.5% urea − TBE at 250 V for 80 min and stained with 0.5% (w/v) SDS for 2 h, rotating at room temperature. The supernatant was supplemented with 1.5 µl of GlycoBlue and 500 µl of isopropanol and incubated on dry ice for 30 min. The RNA was then pelleted for 30 min at 20,000g at 4°C. The pellet was dissolved in 40 µl of water.

To prepare the RNA sample for use in a smallRNA library preparation kit, the sample was phosphorylated using 5 µl of 10X T4 PNK buffer and 1 µl of T4 PNK (NEB #M0201), 1 µl of SUPERase-In, 2.5 µl of 10 mM ATP, and 0.5 µl of 1% Triton X-100. After incubation for 1 h at 37°C, RNA was precipitated and pelleted by adding 41 µl of water, 1.5 µl of GlycoBlue, 8 µl of 5 M NaCl and 150 µl of isopropanol, as described previously. Libraries were prepared using the NEXTflex Small RNA-seq Kit v3 (BioScientific #5132-06) as per the manufacturer’s instructions and sequenced on an Illumina NextSeq500 machine with 3 libraries pooled at 1.8 pM using one High Output Kit v2 (Illumina #FC-404-2005) with 75 cycles single-end.

Ribo-seq and RNA-seq data processing. Ribo-seq reads were stripped of their adapters using cutadapt. Randomized unique molecular identifier sequences (where present) were removed, and reads were collapsed. Reads aligning to rRNA, snRNA, and tRNA sequences were removed with Bowtie2 (ref. #). Unaligned reads were then mapped with STAR, using the hg38 genome and the GENCODE 25 annotation in GTF format. For RNA-seq and Ribo-seq, a maximum of four and two mismatched bases was allowed, respectively, and multimappping to up to 20 different positions was permitted. Alignments flagged as secondary alignments were filtered out, ensuring one mapping position per aligned read. P-site positions and junction reads were extracted using the Ribo-seQC tool with default parameters. Statistics about the different Ribo-seq libraries are available as Supplementary Data 1. Uvis” was used to visualize data tracks and transcript annotation.

Polysome profiling analysis. DEXSeq was run to detect differential exon usage between each of the polysome fraction and the cytoplasmic abundance. Transcripts were divided based on the translation levels of their translated ORF(s) and intersected with differential exons (FDR < 0.01 in at least one polysome fraction). Only genes with multiple translated transcripts were used.

Nuclear – cytoplasmic comparison. DEXSeq was run to detect differential exon usage between the nuclear and the cytoplasmic fraction. Differential exons (FDR < 0.01) were intersected with transcript structures, and only exons uniquely mapping to one transcript group (for example, discarded transcripts, selected transcripts) were selected.

5’ end of endonucleolytic cuts. Bigwig files for the different libraries were normalized by library size. Coordinates were lifted to hg38 and overlapped with ORFquant-identified stop codon positions, for both NMD candidates and controls (canonical’ stop codons taken from the same genes). A window of 50 nucleotides was used to derive spatial profiles and count the number of reads mapping around stop codons in the different conditions.

Merging ORFquant result across cell lines. ORFs were considered to be distinct if they ended at different stop codons or could not be mapped to the same transcript. Enrichments for ORF categories at different levels of overlap were calculated using normalized residuals from a chi-squared test. GO enrichment was performed using the clusterProfiler r and topGO packages.

Proteomics database search. Raw data were searched using MaxQuant r version 1.6.0.13, using carbamidomethyl as fixed modification, and oxidation of methionine and acetylation at protein N termini as variable modifications. An FDR cutoff of 0.01 was used at the level of both peptide-spectrum match and protein identification. Quantification was performed using only unique peptides. Between runs was enabled. We used a custom database to perform the peptide search. ORFquant-detected ORFs were merged in a unique database, choosing only ORFs explaining a minimum of 10% of gene translation (ORF_pctl_P_sites) in at least one cell line.

Comparison between protein abundance and translation estimates. For each protein group, iBAQ values were summed up for each replicate. ORFs whose iBAQ values were summed for all ORFs mapping to each protein group. ORF_pctl_iBAQ values were obtained by dividing each iBAQ value for the sum iBAQ values for that gene. Protein groups mapping to multiple genes were discarded. The same procedure was applied to ORFs whose iBAQ values, to compare protein and translation estimates for each protein isoform. Only proteins detected by Ribo-seq (or RNA-seq) and proteomics were used. Gene-level TPM values in the Extended Data Fig. 8 were calculated using kallisto with default parameters.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
Ribo-seq data for HepG2 and K562 are available at GEO under the accession code GSE129061. Public Ribo-seq data sets for other cell lines were downloaded through the accession codes GSE79664 (HeLa), GSE57136 (HEK293), GSE74279 (Jurkat) and GSE56924 (U2OS). Nuclear and cytoplasmic RNA-seq data were accessed at the European Nucleotide Archive using the accession code PRREIB4197. TriP-seq data were downloaded from GEO using the accession code GSE69352. Bigwig tracks of 5’ ends were accessed using the accession code GSE57433. Proteomics data were downloaded from the PRIDE repository under accession code PXD002395. The list of P-site positions and junction reads and a list of quantified ORFs in the cell types analyzed are available in Supplementary Data 1. The final protein database is available in Supplementary Data 2, together with the parameters used to perform the MaxQuant search and the set of identified peptides and proteins. Source data are provided with this paper.

Code availability
ORFquant is available at https://github.com/lclavelle/ORFquant.

References
47. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBNet J. 17, 10–12 (2011).
48. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25 (2009).
49. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).
50. Calviello, L., Sydow, D., Harnett, D. & Ohler, U. Ribo-seQC: comprehensive analysis of cytoplasmic and organellar ribosome profiling data. Preprint at bioRxiv https://doi.org/10.1101/601468.
51. Hahne, F. & Ivanek, R. In Methods in Molecular Biology Vol. 1418 (eds Mathé, E. & Davis, S.) 335–351 (Humana Press, 2016).
52. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 16, 284–287 (2012).
53. Alexa, A. & Rahnenfuhrer, J. topGO: enrichment analysis for gene ontology. R Package version 2.26.0. (2016).

Acknowledgements
The authors acknowledge funding from the German Federal Ministry of Education and Research (BMBF grant 031 A538 A RBC) and the German Research Foundation (DFG)
grant TR175). L.C. thanks Stephen Floor (UCSF) for support and feedback during the preparation of this manuscript.

**Author contributions**

Initial study was conceived by L.C. and U.O. L.C. ideated and implemented the ORFquant pipeline, with supervision from U.O. All data analysis and visualization was performed by L.C. Ribosome footprinting libraries in K562 and HepG2 were performed by A.H. The manuscript was written by L.C. and U.O., with additional input by A.H.

**Competing interests**

The authors declare no competing interests.

**Additional information**

Extended data is available for this paper at https://doi.org/10.1038/s41594-020-0450-4.

Supplementary information is available for this paper at https://doi.org/10.1038/s41594-020-0450-4.

Correspondence and requests for materials should be addressed to L.C. or U.O.

Peer review information Anke Sparmann was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | Transcript selection and ORF quantification statistics. In a) the number of selected transcripts per gene (x-axis) against the number of genes and their TPM levels. In b) Number of genes and their TPM values are plotted against the number of detected ORFs. In c) the number of genes (top), their TPM values, the number of unique ORF features and the variability of RSEM quantification estimates (averaged across all the ORFs per gene) are plotted against the contribution (in percentages) of the major ORF. As for b), the maximum width of each violin plot is the same for each panel, and the median value is shown as a black bar. In d) RSEM quantification (on the x axis, IsoPct_from_PME_TPMs) plotted against ORFquant quantification estimates (ORF_pct_P_sites_pN). Both values correspond to length-normalized quantification estimates.
Extended Data Fig. 2 | ORFquant quantifies translation on alternatively spliced isoforms. Aggregate plots of Ribo-seq coverage (normalized 0–1 per each region) and ORF coverage (ORF_pct_P_sites_pN) over different candidate alternative splice sites. No mixture indicates the presence of a single ORF only, while other lines indicate the presence of additional ORFs, divided by their translation values. Explanatory schemes are depicted at the bottom of each plot, with blue representing the major ORF and red the additional ORF(s).
Extended Data Fig. 3 | Polysome profiles of alternative isoform with different translation output. Distributions of exonic log2 fold changes between different polysome profiles and cytoplasmic abundance. Lowly translated ORFs are depleted in heavier polysome fractions, while highly translated ORFs show signal in all fractions.
Extended Data Fig. 4 | Statistics on ORFs detection and quantification in the assayed cell lines. In a) number of genes and average Ribo-seq signal per gene (y-axes) against the number of detected ORFs (x-axis), for the assayed cell lines. In b) number of genes and average Ribo-seq signal per gene (y-axes) against the translation of the major ORF (x-axis), for the assayed cell lines. As in a), the maximum width of each violin plot is the same for each panel, and the median value is shown as a black bar.
Extended Data Fig. 5 | Genes expressing one or multiple major ORFs across different cell lines. In a) average length-normalized translation of the major ORF and average Ribo-seq gene signal per gene (y-axes), plotted against the number of cell lines where the gene harbored a detected ORF. Colors indicate the gene biotype. Values were plotted dividing genes according the number of district major ORF detected across cell lines. The maximum width of each violin plot is the same for each panel, and the median value is shown as a black bar. In b) top enriched GO categories for genes translating one (right) or multiple (left) major ORFs across cell lines.
Extended Data Fig. 6 | Examples of genes exhibiting translation of multiple major ORFs. Translation on multiple major ORFs per gene for the a) EP400NL, b) GAS7, and c) IFRD1 genes, representing examples of alternative splicing, transcription start site or small ORF usage. Displayed tracks represent, in descending order: gene annotation and (for each cell line): P-sites positions, junction reads, ORF coverage (defined as % of gene translation) and quantified ORFs. In b) and c) intron sizes were scaled to a maximum of 300 and 150 nt respectively.
Extended Data Fig. 7 | Ribo-seq assisted protein detection and quantification across cell lines. In a) the overlap between in-silico generated tryptic peptides using the full GENCODE25 database or ORFquant-derived protein sequences in the 6 cell lines analyzed. Up to two missed cleavage events were allowed. In b) iBAQ values (y-axis) are plotted against length-normalized translation quantification estimates, for each cell line. Each plot is divided by number of unique peptides (from proteomics) and unique features with reads (from Ribo-seq).
Extended Data Fig. 8 | Agreement between transcript and protein abundance. iBAQ values (y-axis) plotted against (x-axis) gene-level TPM values from RNA-seq. Plots are divided according to the number of unique peptides detected. Number of proteins, together with Pearson and Spearman correlations, is shown for each plot.
Extended Data Fig. 9 | Translation quantification and protein abundance for multiple gene products. Pearson and Spearman correlation coefficients between ORFquant-derived % of gene translation and % of gene protein abundance. In each plot, translation quantification is show with (right) and without (left) adjusting for the presence of multiple ORFs. Size and color of each data point indicate the number of unique peptides detected.
Extended Data Fig. 10 | Increased agreement between translation and protein abundance using signal from all reads. Correlation between translation quantification and protein abundance using all reads, as shown in Fig. 5 using uniquely mapping reads only. Values are shown for all proteins (left), or as percentages of gene output for protein isoforms (right).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒ The statistical test(s) used AND whether they are one- or two-sided
  Only common tests should be described solely by name; describe more complex techniques in the Methods section.
☒ A description of all covariates tested
☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  Give P values as exact values whenever suitable.
☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒ Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection | https://github.com/calviello/ORFquant
Data analysis | https://github.com/calviello/ORFquant

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the ‘Nature Research guidelines for submitting code & software’ for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Ribo-seq data for HepG2 and K562 is available at GEO under the accession GSE129061. Ribo-seq datasets for other cell lines were previously published, and accessed using the accessions GSE79664 (HeLa), GSE73136 (HEK293), GSE74279 (Jurkat) and GSE69924 (U2OS); more details about the analyzed samples can be found in Supplementary Table 1. Nuclear and cytoplasmic RNA seq was accessed at the European Nucleotide Archive using the accession PRJEB4197. TriP-seq data was downloaded from GEO using the accession GSE69352. Transcriptome-wide tracks of 5’ ends were accessed using the accession GSE57433. Proteomics data was downloaded from the PRIDE repository under accession PXD002395. The list of P-sites positions and junction reads in the cell types analyzed is available in Supplementary Data 1. The list of quantified ORFs in the different cell lines is available in Supplementary Data 1. The final protein database is available in Supplementary Data 2, together with the parameters used to perform the MaxQuant search and the set of identified peptides and proteins.
Field-specific reporting

Please select the one that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size ☐ No sample size calculation was performed.
Data exclusions ☐ No data was excluded from the analysis.
Replication ☐ Multiple datasets from different techniques were used to validate our findings.
Randomization ☐ Not relevant. The study focuses on the development of an analytical pipeline for ORF detection and quantification using Ribo-seq data.
Blinding ☐ Not relevant. The study focuses on the development of an analytical pipeline for ORF detection and quantification using Ribo-seq data.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems | Methods |
|----------------------------------|---------|
| n/a | Involved in the study |
| ☒ Antibodies | ☒ ChIP-seq |
| ☒ Eukaryotic cell lines | ☒ Flow cytometry |
| ☒ Palaeontology | ☒ MRI-based neuroimaging |
| ☒ Animals and other organisms | |
| ☐ Human research participants | |
| ☒ Clinical data | |

Eukaryotic cell lines

Policy information about cell lines.

Cell line source(s) | ATCC
Authentication | K562 authenticated by Eurofins genomics
Mycoplasma contamination | Tested negative using the LookOut kit, Sigma Aldrich #MP0035-1KT
Commonly misidentified lines (See ISLAC register) | Cell lines used in our laboratory are not present in the list of misidentified cell lines.