SaTAnn quantifies translation on the functionally heterogeneous transcriptome

Lorenzo Calviello[1,2,4,*], Antje Hirsekorn[1], Uwe Ohler[1,2,3,*]

[1]Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine, Hannoversche Strasse 28, Berlin 10115, Germany

[2]Department of Biology, Humboldt-Universität zu Berlin, Unter den Linden 6, Berlin 10117, Germany

[3]Department of Computer Science, Humboldt-Universität zu Berlin, Unter den Linden 6, Berlin 10117, Germany

[4]Present address: University of California San Francisco, Department of Cell & Tissue Biology, 513 Parnassus Avenue, 94143, San Francisco, California, United States

[*]Corresponding authors: calviello.l.bio@gmail.com, Uwe.Ohler@mdc-berlin.de

Abstract

Deep sequencing methods have matured to comprehensively detect the full set of transcribed loci, but there is a gap to determine the function of the resulting highly complex transcriptomes. At the center of the gene expression cascade, translation is fundamental in defining the fate of much of the transcribed genome. We have developed a new approach (*SaTAnn*, *Splice-aware Translatome Annotation*) to annotate and quantify translation at the single open reading frame (ORF) level, that uses information from ribosome profiling to determine the translational state of each isoform in a comprehensive annotation. For most genes, one ORF represents the dominant translation product, but our approach also detects translation from ORFs belonging to multiple transcripts per gene, including targets of RNA surveillance mechanisms such as nonsense-mediated decay. Diversity in the translation output across human cell lines reveals the extent of gene-specific differences in protein production, which are supported by steady-state protein abundance estimates. Computational analysis of Ribo-seq data with *SaTAnn* (available at https://github.com/lcalviell/SaTAnn) provides a window into the functions of the heterogeneous transcriptome.

Introduction

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 then 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 quality control platform for both the nascent peptide[1] and the translated mRNA[2]. Several RNA surveillance mechanisms are known to occur co-translationally, and their importance for different processes such as differentiation or disease has been investigated[3].

Ribosome profiling (Ribo-seq) has made it possible to pinpoint the positions of actively translating ribosomes transcriptome-wide, using ribosome footprinting coupled to RNA sequencing[4]. In the last 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 datasets have triggered a plethora of dedicated analysis methods, which exploit distinct features of Ribo-seq profiles to confidently identify translated ORFs[5,6]. Many reports have focused on whether small translated regions are hidden in long non-coding RNAs, with less attention given so far to account for the presence of multiple transcript isoforms per gene.

Transcript diversity can result from either alternative splicing (AS) 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 a wide spectrum of multiple isoforms expressed by thousands of genes, with many isoforms being lowly expressed and/or presenting incomplete open reading frames[7]. The contribution of this transcript heterogeneity to an expanded translatome is therefore an intensely debated topic[8,9], with much of transcript and protein abundance apparently explained by a single dominant transcript per gene[10].

The mere presence of multiple isoforms does not imply the presence of a distinct, functional protein translated from each transcript structure: 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 another perspective, shotgun proteomics approaches are only recently

providing the sensitivity to detect tens of thousands of proteins from a single sample and cannot be expected to deliver the complete set of proteins, especially of lowly abundant ones[11].

To close this gap, we aimed to develop a strategy to identify and quantify translation on the subset of isoforms 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[12], underlining the potential of isoform-aware analysis approaches to fully define the translatome. Following up on this premise, we here describe a splice-aware translatome annotator, or *SaTAnn*, a Ribo-seq analysis approach that detects and quantifies translation of ORFs in multiple transcript isoforms and zooms in on the potential roles of alternative transcripts.

Results

The SaTAnn 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 isoforms (Figure 1a, b). We applied our workflow on a previously published Ribo-seq dataset in HEK293 cells[13]. Single-nucleotide positions corresponding to the peptidyl-site for each ribosome (P-sites positions) and junction reads are first extracted from the Ribo-seq alignment (Methods) and then mapped to flattened gene models from a given annotation (Figure 1b). In this way, transcript features (e.g. exonic bins or splice junctions) are designated as unique or shared across multiple annotated isoforms.

We first select a subset of annotated isoforms, which is sufficient to explain all 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 with Ribo-seq signal are not unique to that transcripts, and those features can be explained by another isoform with better support. Strictly nested structures are also eliminated. As Ribo-seq reads generally map to 5'UTRs and coding regions only, this approach cannot distinguish isoforms differing in their 3'UTR.

This simple yet effective selection strategy leads to a significantly reduced number of transcripts. The observed Ribo-seq signal could be explained by 1 to 3 transcript structures for most genes (Figure 1c). Reassuringly, the number of selected transcripts showed only a mild correlation with overall Ribo-seq coverage, with signal in highly covered genes explained by more transcript structures. This selection

dramatically improves the assignments of both exons and junctions to transcripts (Figure 1d). When considering covered exons or junctions (defined as having at least one Ribo-seq read mapped to them), ~63% of exons mapped to 1 or 2 transcripts, compared to ~31% when no selection was performed. Considering only annotated protein-coding transcripts does not improve the mapping of covered features, while it ignores the presence of covered exons and junctions unique to non-coding transcripts. Next, we detected translated ORFs *de novo* in each of the selected transcripts, using the multitaper[13–15] method to select in-frame signal displaying 3nt periodicity. Detected ORFs are further filtered by the same strategy used for transcript filtering.

After calculating coverage on unique and shared ORF features (exonic bins and splice junctions), a scaling factor between 0 and 1 is determined from the coverage on unique ORF features, and the amount of overlap between ORFs when no unique feature is present on all ORFs (Methods). This scaling factor represents the fraction of Ribo-seq signal which 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)[16]-like values, named P-sites per Nucleotide per Million (*P_sites_pNpM*). Moreover, we calculated the relative contribution of each ORF to the overall translation output of each gene (*ORF_pct_P_sites* value or percentage of gene translation). An 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 CDS regions.

Applying *SaTAnn*, we quantified translation for ~24,000 ORFs in ~15,000 genes in the HEK293 data. Most genes (9,183, Figure 1e, *left*) displayed only one translated ORF, with another >5,800 genes showing translation of multiple ORFs. Upon closer inspection (Figure 1e, *right*), we observed that for the majority of genes (~80%), the translation of the most translated (i.e., major) ORF could explain >80% of the total gene translation, with only 554 genes in which the major ORF explained <50% of the translational output. We did not observe a dependency between number of detected ORFs (or % of translation of major ORF) and overall Ribo-seq coverage (Figure 1e, *right*).

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 (e.g. 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 reflect the expected pattern calculated by *SaTAnn* for each individual ORF (Figure 1f). Additional profiles over different genomic locations are shown in Supplementary Figure 1. 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 isoforms for hundreds of others.

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 *SaTAnn* to represent *bona fide* cytoplasmic transcripts. To test this hypothesis, we performed a differential exon usage analysis with DEXSeq[17], using RNA-seq data from nuclear and cytoplasmic extracts in HEK293 cells[18]. Most exons unique to discarded structures showed marked nuclear localization (log2FC>0), while exons of selected transcripts showed a prominent cytoplasmic enrichment (Figure 2a). Transcripts in which we identified a complete translated ORF displayed a more marked cytoplasmic localization. An example of the selection strategy discarding pre-mRNA structures in favor of cytoplasmic transcripts is shown in Figure 2b.

When examining the GENCODE annotation[19] of the transcripts hosting the de novo identified ORFs, we noticed ~2,000 ORFs in non-coding isoforms of protein-coding genes (Figure 2c). Compared to ORFs in annotated protein-coding isoforms, these ORFs exhibited much lower translation, accounting for a median of 6.8 % of host gene translation, compared to 87% for ORFs that fully matched annotated CDSs. More than 3,500 N-terminal truncation events were also detected, showing high levels of translation. uORFs and other small ORFs exhibited low overall levels of translation, albeit high when normalized by their length. In annotated non-coding genes, we detected >2,300 ORFs from annotated pseudogenes and ~900 ORFs from other non-coding RNA genes. Based on length-normalized translation values (*P_sites_pNpM*), ORFs located in long ncRNAs and non-coding isoforms show considerably lower translation than protein-coding isoforms (Figure 2c).

Analyzing a deep polysome profiling dataset (Trip-Seq[20]) from the same cell line showed that the quantitative estimates of translation agreed with distinct polysome profiles (Figure 2d,e; Supplementary Figure 2): Exons uniquely mapping to transcripts harboring lowly translated ORFs accumulated in low polysomes and were depleted in heavier polysomal fractions. Conversely, highly translated transcripts exhibited 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 non-coding isoforms (Figure 2c) suggests inefficient translation and/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[21], which is assumed to be recognized as such when located sufficiently upstream of the last splice junction, i.e. when a downstream Exon Junction Complex (EJC) is not displaced during translation elongation (Figure 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[22], by knocking down the exonuclease *XRN1* in charge of degrading the cleaved transcripts. When aligning the cleavage sites at the stop codons of (putative) PTC- and non-PTC-containing transcripts (taken from the same genes), we observed a clear difference (Figure 3b): transcripts without PTC, i.e. where all EJCs are presumably displaced, showed background-like signal, while transcripts harboring a putative PTC showed a marked degradation profile around their stop codon[21]. The degradation signal was less pronounced when *SMG6* or *UPF1* were also knocked-down, underlining the effect of known key factors of the NMD pathway at our candidate NMD targets. A clear example of such pattern is visible on a translated ORF in the *GAS5* gene (Figure 3c).

To further explore the dependency of NMD with regards to 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 the surveyed ORFs (including uORFs; Figure 3d). As reported[22], ORFs in snoRNA host genes (such as *GAS5*, Figure 3c) showed the highest degradation profile, while other categories exhibited a lower amount of degradation.

In summary, Ribo-seq data can serve as an excellent means to identify mature mRNAs, quantify different levels of translation output within single genes, and infer transcript-specific cytoplasmic fates.

A subset of genes translates different major ORFs across cell lines

To investigate the patterns of alternative ORF usage across different conditions, we ran *SaTAnn* on Ribo-seq datasets from 6 different human cell lines (Supplementary Table 1, Supplementary Data 1, Figure 4a),

with newly generated data for K562 and HepG2 cell lines complementing previously published libraries from HEK293, HeLa, U2OS and Jurkat cells[23–25]. For each dataset we observed the same trend described in Figure 1d, with most genes showing translation of one major ORF, and hundreds of genes showing sustained translation of multiple ORFs, with a weak dependency of the overall Ribo-seq signal (Supplementary Figures 3, 4). Across all cell lines, we detected ORF translation for ~17,800 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 the assayed cell lines (Figure 4b). For ORFs in non-coding RNAs, we detect a more cell-specific pattern of major ORF usage. However, a few dozen non-coding genes displayed translation of the same major ORF: one such example is *GAS5*, where the translation of an ORF terminating at a PTC (Figure 3c) is consistently detected across the assayed cell lines (Figure 4c).

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 (Supplementary Figure 5). Two or more distinct major ORFs were identified in 18% and 4.7% of genes, representing candidate major ORF switching events across cell lines (Figure 4b). At a closer look, we observed that genes translating multiple major ORFs also displayed 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 (Figure 4b, lower panel).

ORF diversity is created by different mechanisms: differences in alternative splicing of internal coding exons (Supplementary Figure 6), alternative transcriptional start sites (Supplementary Figure 7), or alternative usage of last exons (Figure 4d). Genes exhibiting translation of multiple major ORFs showed an enrichment for GO categories like GTPase regulator (Supplementary Figure 8), a category also enriched in genes expressing multiple transcript isoforms across human tissues[26], thus likely representing the result of transcript heterogeneity across cell lines. However, in ~40% of the cases, distinct major ORFs translated across cell lines showed a low degree of overlap (Figure 4e) despite coming from the same genes, unlikely to result from differences in local alternative splicing events. Such a low overlap reflected the presence of alternative usage of small ORFs (Figure 4f), such as uORFs, which can represent the major gene translation product of a gene (Figure 4g, Supplementary Figure 9).

Taken together, these translation estimates indicate that the presence of one (or multiple) dominant ORFs agree across multiple cell lines for the majority of genes. For around 20% of the translated genes,

however, highly translated small ORFs and/or several isoforms expressed at sustained levels create a substantial level of complexity in protein synthesis, for which multiple ORFs can represent the majority of the gene translational output.

Agreement between protein abundance and synthesis estimates depends on proteome coverage and transcriptome complexity

Ribo-seq reflects the density of active 80S ribosomes and thus ongoing protein production, but the signal at a specific locus may represent both elongating or stalled ribosomes. We therefore examined whether the relative abundance of different translated ORFs would agree between Ribo-seq and proteomics data. We estimated proteome-wide steady-state protein abundance using published deep mass spectrometry data[27,28] for the same cell lines outlined above (Figure 4a). We detected between 7,000 and 8,000 proteins per cell line (Supplementary Figure 10, Supplementary Data 2, Methods), and performed label-free quantification using the signal from unique peptides only. To estimate the ability of both techniques in quantifying protein synthesis/abundance, we divided proteins based on the number of exon or junction features covered by Ribo-seq (with >=1 read mapping, i.e. independent of the exact number of mapping reads), and by the number of detected unique peptides (independent of their intensity). In cases where 0 - 3 unique peptides were detected, the correlation (in log space) between HEK293 *SaTAnn*-derived estimates of translation (*P_sites_pNpM*) and steady-state protein abundance (*iBAQ*) ranged from 0.57 to 0.60 (Figure 5a). However, for proteins having >9 uniquely mapping peptides and >8 covered features (n>1900) the correlation between *SaTAnn* estimates of translation and protein abundance reached the value of ~0.87. The same phenomenon was observed for all the assayed cell lines (Figure 5b). A clear dependency on the number of unique peptides was also observed when correlating *iBAQ* values with transcript abundance estimates from RNA-seq, albeit to a lower extent (Supplementary Figure 11).
When comparing the fraction of total gene translation to the fraction of total gene protein abundance for the few dozen genes with multiple detected protein isoforms, we observed a correlation of 0.4. Here, only few proteins harbored >9 uniquely mapping peptides (Figure 5c), thus limiting our ability in reliably estimating their abundance (Figure 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 (Figure 5d). In only one out of the six cell lines, the correlations did not improve, likely as result of few detected peptides (Jurkat, Supplementary Figure 12). Taken together,

these results show a good agreement of splice-aware ORF-centric quantification estimates with steady-state protein abundance, when enough signal is available to perform such a correlative analysis.

Discussion

Only a fraction of known, annotated gene 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 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 (Figure 2a) indicates that these structures are not present at translating ribosomes, i.e. that they are not expressed in the assayed condition or that they represent pre-mRNA intermediates which 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 (Figure 1d), allowing for ORF-specific translation estimates. However, optimally deconvolving the mixture of multiple transcript isoforms can be challenging for many genes, especially in the absence of coverage or unique transcript features allowing us to estimate ORF usage. The rapidly increasing availability of complete isoform sequence data based on long-read sequencing[29] will soon make this a common starting point, even for model systems without extensively curated annotation.

While polysome profiling experiments (Figure 2d) and label-free quantification of the protein product (Figure 5b) support the Ribo-seq estimates of relative ORF translation levels, we believe that additional efforts can improve ORF-specific quantification of translation. The quantification of 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 isoforms[16,30,31]. An accurate approach will however have to address the issue of variable Ribo-seq coverage along the ORFs, which reflects the complex dynamics of translation. Its exact causes, including the experimental biases, codon composition or RNA structural features[32], remain to be understood. Our approach also uses a simplified 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[33] 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 isoforms of protein coding genes annotated as non-coding, consistent with current models for mRNA surveillance such as NMD (Figure 3). Similarly, we observed that many detected ORFs in non-coding RNAs show high degradation profiles at their stop codons, especially pronounced in snoRNA host genes (Figure 3d). This well-known phenomenon is therefore important to consider when addressing the protein-coding ability of transcripts based on ribosome occupancy. In turn, the ability to identify NMD target candidates can provide an advantageous starting point for further research into defining the features of co-translational mRNA surveillance and its links to protein quality control[34].

Expanding our analysis across multiple cell lines allowed us to assess the complexity of translation per gene for both coding and non-coding genes (Figure 4b). We found the majority of genes to be translating the same major ORF (including highly translated ORFs in non-coding RNAs, Figure 4c), but we 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 protein-coding transcripts, thus suggesting a cautionary note when trying to define translation coupled to clear isoform switching events. In this context, the presence of highly translated small ORFs even in protein-coding genes (Supplementary Figure 9), which may play gene regulatory roles rather than expand the proteome, adds further complexity. Unfortunately, the limited amount of data at hand (often without replicate information) and the heterogeneity of protocols adopted by different labs, poses challenges to precisely quantify the contribution of different mechanisms, such alternative TSS usage, alternative splicing, or translation regulatory phenomena (such as differential uORF translation), in promoting diversity (or lack thereof) in protein synthesis.

Despite these potential limitations, 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[35]; 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 detect and quantify the levels of translation and protein abundance (Figure 5a). An increasing availability of Ribo-seq and proteomics data in a single controlled environment will improve our understanding of this relationship and help to pinpoint interesting cases in which this correlation deviates from expectation. The current scarcity of matching data, however, limits our ability to validate the translation of multiple protein isoforms per gene, as most detected peptides are shared between different isoforms (Supplementary Figure 12). Yet, the currently observed agreement between translation and protein abundance at the protein isoform level provides a

promising starting point for the investigation of isoform-specific protein production. 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 different genes[36]. With both proteomics[37,38] and transcriptomics[39] 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.

References

1. Brandman, O. & Hegde, R. S. Ribosome-associated protein quality control. *Nat Struct Mol Biol* **23**, 7–15 (2016).
2. Hu, W., Sweet, T. J., Chamnongpol, S., Baker, K. E. & Coller, J. Co-translational mRNA decay in Saccharomyces cerevisiae. *Nature* **461**, 225–229 (2009).
3. Shoemaker, C. J. & Green, R. Translation drives mRNA quality control. *Nat. Struct. Mol. Biol.* **19**, 594–601 (2012).
4. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-Wide Analysis in Vivo of Resolution Using Ribosome Profiling. *Science (80-.).* **324**, 218–23 (2009).
5. Wang, H., Wang, Y. & Xie, Z. Computational resources for ribosome profiling: from database to Web server and software. *Brief. Bioinform.* (2017). doi:10.1093/bib/bbx093
6. Calviello, L. & Ohler, U. Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome. *Trends in Genetics* **33**, 728–744 (2017).
7. Braunschweig, U. *et al.* Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res.* **24**, 1774–1786 (2014).
8. Blencowe, B. J. The Relationship between Alternative Splicing and Proteomic Complexity. *Trends Biochem. Sci.* **4**, e07794 (2017).
9. Tress, M. L., Abascal, F. & Valencia, A. Most Alternative Isoforms Are Not Functionally Important. *Trends in Biochemical Sciences* (2017). doi:10.1016/j.tibs.2017.04.002
10. Gonzalez-Porta, M., Frankish, A., Rung, J., Harrow, J. & Brazma, A. Transcriptome analysis of

human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol* (2013). doi:10.1186/gb-2013-14-7-r70

11. Bekker-Jensen, D. B. *et al.* An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Syst.* **4**, 587–599.e4 (2017).
12. Weatheritt, R. J., Sterne-Weiler, T. & Blencowe, B. J. The ribosome-engaged landscape of alternative splicing. *Nat. Struct. Mol. Biol.* 1–9 (2016). doi:10.1038/nsmb.3317
13. Calviello, L. *et al.* Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods* **13**, 1–9 (2015).
14. Thomson, D. J. & J., D. Spectrum Estimation and Harmonic Analysis. *Proc. IEEE, Vol. 70, p. 1055-1096* **70**, 1055–1096 (1982).
15. Rahim, K. J., Burr, W. S. & Thomson, D. J. Appendix A: Multitaper R Package in 'Applications of Multitaper Spectral Analysis to Nonstationary Data,' PhD diss., Queen's University,. 149–183 (2014).
16. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
17. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22**, 2008–2017 (2012).
18. Sultan, M. *et al.* Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics* **15**, 675 (2014).
19. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–74 (2012).
20. Floor, S. N. & Doudna, J. A. Tunable protein synthesis by transcript isoforms in human cells. *Elife* **5**, (2016).
21. Lykke-Andersen, S. & Jensen, T. H. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat. Rev. Mol. Cell Biol.* **16**, 665–677 (2015).
22. Lykke-Andersen, S. *et al.* Human nonsense-mediated RNA decay initiates widely by endonucleolysis and targets snoRNA host genes. *Genes Dev.* **28**, 2498–517 (2014).
23. Park, J. E., Yi, H., Kim, Y., Chang, H. & Kim, V. N. Regulation of Poly(A) Tail and Translation during the Somatic Cell Cycle. *Mol. Cell* **62**, 462–471 (2016).
24. Gawron, D., Ndah, E., Gevaert, K. & Van Damme, P. Positional proteomics reveals differences in N-terminal proteoform stability. *Mol. Syst. Biol.* **12**, 858–858 (2016).
25. Jang, C., Lahens, N. F., Hogenesch, J. B. & Sehgal, A. Ribosome profiling reveals an important role

for translational control in circadian gene expression. *Genome Res* **25**, 1836–1847 (2015).

26. 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.* 1–10 (2017). doi:10.1101/gr.220962.117
27. 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-M111.014050 (2012).
28. 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).
29. Garalde, D. R. *et al.* Highly parallel direct RN A sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).
30. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
31. 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).
32. 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).
33. Michel, A. M. *et al.* Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res.* **22**, 2219–29 (2012).
34. Feng, Q., Jagannathan, S. & Bradley, R. K. The RNA Surveillance Factor UPF1 Represses Myogenesis via Its E3 Ubiquitin Ligase Activity. *Mol. Cell* **67**, 239–251.e6 (2017).
35. Franks, A., Airoldi, E. & Slavov, N. Post-transcriptional regulation across human tissues. *PLoS Comput. Biol.* (2017). doi:10.1371/journal.pcbi.1005535
36. Yang, X. *et al.* Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* **164**, 805–817 (2016).
37. Collins, B. C. *et al.* Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat. Commun.* **8**, (2017).
38. Martens, L. & Vizcaíno, J. A. A Golden Age for Working with Public Proteomics Data. *Trends Biochem. Sci.* (2017). doi:10.1016/j.tibs.2017.01.001
39. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).

40. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

41. 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).

42. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

43. Calviello, L., Sydow, D., Harnett, D. & Ohler, U. Ribo-seQC: comprehensive analysis of cytoplasmic and organellar ribosome profiling data. doi:10.1101/601468

Methods

SaTAnn - Transcript/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-sites positions and junction reads 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 explained by the selected transcripts is created, and it is updated at each selection step. After creating the empty vector, the list of annotated transcripts is analyzed, applying the following rules for each transcript Tx_i:

1) Tx_i contains an unexplained positive feature:

   Tx_i is selected and each previously selected Tx_j is re-analyzed:

   If all the positive features of Tx_j are also contained in Tx_i, Tx_j is discarded.

2) Tx_i does not contain an unexplained positive feature:

   Tx_i is initially selected, but it is compared with each previously selected structure Tx_j. Two possible scenarios are evaluated:

   i) All the positive features of Tx_i are also contained in Tx_j:

      if Tx_j has more positive features than Tx_i, or fewer negative *internal* features than Tx_i, Tx_i is discarded

   ii) All the positive features of Tx_j are also contained in Tx_i:

      If Tx_i has fewer negative *internal* features than Tx_i, Tx_j is discarded.

This greedy strategy reduces the number of transcripts that is necessary to explain 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.

SaTAnn - ORF finding

As in the RiboTaper[13] method, only ATG is considered as potential start codon, and the p-value for the multitaper method applied to the candidate ORF P-sites 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.

SaTAnn - ORF quantification

After the ORF finding step, ORF filtering and quantification is subsequently performed, using the length-normalized Ribo-seq coverage Cov on each ORF feature.

$$Cov = \frac{\#reads}{length}$$

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

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 $AvCovUn$, using the coverage Cov_{Fu} on each of the unique features F_u.

$$AvCovUn = \frac{\sum_1^{\#Fu} Cov_{Fu}}{\#F_u}$$

The same calculation is performed for all features $Fall$ mapping to the ORF

$$AvCovAll = \frac{\sum_1^{\#F_{all}} Cov_{Fall}}{\#F_{all}}$$

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

$$C_{ORF} = \frac{AvCovUn}{AvCovAll}$$

When no unique feature is present in one ORF (all regions are shared with other ORFs), the coverage Cov_{Fadj} on each feature $Fadj$ attributed to that ORF is calculated subtracting the expected signal coming from other ORF_{Fadj} mapping to that feature, using their scaling factors. In such cases, the calculation of the adjusted coverage for each feature $Fadj$ is as follows:

$$Cov_{Fadj} = \frac{\#reads_{Fadj}}{length_{Fadj}} - \frac{\#reads_{Fadj}}{length_{Fadj}} * \sum_{ORF_{Fadj}}^{\#ORFadj} C_{ORF_{Fadj}}$$

$$AvCovAdj = \frac{\sum_{1}^{\#F_{adj}} Cov_{Fadj}}{\#F_{adj}}$$

$$C_{ORF} = \frac{AvCovAdj}{AvCovAll}$$

If no unique region is present in any detected ORF in the gene (all regions are shared among ORFs and no C_{ORF} value can be calculated), the scaling factor is calculated assuming uniform Ribo-seq coverage on each ORF. Coverage Cov_{Fsh} is now simply divided by the number of ORF_{Fsh} mapping to the feature Fsh.

$$Cov_{Fsh} = \frac{\#reads_{Fsh}}{length_{Fsh}} / \#ORF_{Fsh}$$

$$AvCovSh = \frac{\sum_{1}^{\#F_{sh}} Cov_{Fsh}}{\#Fsh}$$

$$C_{ORF} = \frac{AvCovSh}{AvCovAll}$$

After the calculation of C_{ORF}, the adjusted number of P-sites for each ORF (P_{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.

$$P_{ORF} = Psites * C_{ORF}$$

For each ORF of length L_{ORF}, the scaled numbers of P-sites P_{ORF} is normalized over the entire set of detected ORFs $\#RFN$, to obtain TPM-like values, named P-sites per Nucleotide per Million (*P_sites_pNpM*), using this formula:

$$P_sites_pNpM_{ORF} = \frac{P_{ORF}}{L_{ORF}} * \frac{10^6}{\sum_{ORF}^{\#ORF} \frac{P_{ORF}}{L_{ORF}}}$$

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

$$ORF_pct_P_sites_{ORF} = \frac{P_{ORF}}{\sum_{ORF}^{\#ORFg} P_{ORF}}$$

Normalization by length is here not applied, as this metric wants to quantify the amount of translation per gene coming from each ORF. The *ORF_pct_P_sites_pN* metric indicates length-normalized *ORF_pct_P_sites* values (e.g. 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.

SaTAnn parameters used

For all cell lines, *SaTAnn* was run using a cutoff of 2% of total gene translation. The set of identified ORFs in each cell line is available in Supplementary Data 1.

Ribosome profiling:

Ribo-seq was performed as described previously[13] and adapted for HepG2 and K562 cell lines. 5x10^6 K562 suspension cells and two 80% confluent 10 cm TC 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 ug/ml cycloheximide (Sigma Aldrich #C4859) and immediately snap-frozen by immersing the dishes in liquid nitrogen. The dishes were then transferred to wet ice and 400 ul of lysis buffer (1X polysome buffer (20 mM Tris-Cl pH 7.4, 150 mM NaCl, 5 mM MgCl2, with 1 mM DTT (Sigma Aldrich #43816) and 100 ug/ml cycloheximide added freshly; keep on ice), 1% (v/v) Triton X-100 (Calbiochem #648466), 25 U/ml TURBO DNase (Life Tech. #AM2238)) was immediately dripped onto the frozen cells. The cells and buffer were then scraped off and left to thaw on one side of the dish, mixing them using a pipet tip.

Suspension cells were supplemented with 100 ug/ml cycloheximide, pelleted for 5 min at 300 g and washed with ice-cold PBS + 100 ug/ml cycloheximide. The washed cell pellet was immediately snap-frozen in liquid nitrogen. 400 ul of ice-cold lysis buffer was added, and the cells were put on wet ice to thaw, mixing them using a pipet tip.

The cells were left to lyse for 10 min on ice, followed by 10x trituration through a 26-G 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 ul of lysate were supplemented with 1000 U of RNase I (Life Tech. #AM2295) and incubated in a thermomixer set to 23°C, shaking at 500 rpm for 45 min. Footprinting was stopped by adding 260 U of SUPERASE-In (Life Tech. #2696).

To recover ribosomes two MicroSpin S-400 HR columns (GE Healthcare #27-5140-01) per 400 ul of sample were equilibrated with a total of 3 ml of polysome buffer. The columns were drained by spinning for 4 min at 600 g, then the sample was applied and spun for 2 min at 600g. Three volumes of Trizol LS (Life Tech. #10296010) 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 Tech. #Q10211).

Ribosomal RNA was removed from 10 ug of footprinted RNA using the RiboZero Magnetic Gold kit (Illumina #MRZG12324) as per the manufacturer’s instructions. Footprinted RNA was precipitated from the supernatant (90 ul) using 1.5 ul of glycoblue (Life Tech. #9515), 9 ul of 3 M sodium acetate and 300 ul

of ethanol by incubation for 1h at -80°C and pelleted for 30 min at max. speed at 4°C. The RNA pellet was dissolved in 10 ul 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-PAGE gel along with 27 nt and 30 nt RNA markers. The gel was run in 1X TBE at 250 V for 80 min and stained for 3 min in 1X SYBR gold (Life Tech. #S11494) in 1X TBE. Sample bands between 27 nt and 30 nt were excised and crushed by spinning through a punctured tube. RNA was extracted by soaking the gel pieces in 400 ul of RNA extraction buffer (400 mM NaCl, 1 mM EDTA, 0.25% (wt/v) SDS) for 2 h, rotating at room temperature. The supernatant was supplemented with 1.5 ul of glycoblue and 500 ul of isopropanol and incubated on dry ice for 30 min, followed by pelleting of the RNA for 30 min at 20'000 g at 4°C. The pellet was dissolved in 40 ul of water.

To prepare the RNA sample for use in a smallRNA library preparation kit the sample was phosphorylated using 5 ul of 10X T4 PNK buffer and 1 ul of T4 PNK (NEB #M0201), 1 ul of SUPERASE-In, 2.5 ul of 10 mM ATP and 0.5 ul of 1% Triton X-100. After incubation for 1 h at 37°C RNA was precipitated and pelleted by adding 41 ul of water, 1.5 ul of glycoblue, 8 ul of 5M NaCl and 150 ul of isopropanol as described before. Libraries were prepared using the NEXTflex Small RNA-Seq Kit v3 (BiooScientific #5132-06) as per the manufacturer's instructions and sequenced on an Illumina NextSeq500 machine with 13 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[40]. Randomized UMI sequences (where present) were removed, and reads were collapsed. Reads aligning to rRNA, snoRNAs and tRNA sequences were removed with Bowtie2[41]. Unaligned reads were then mapped with STAR[42] using the hg38 genome and the GENCODE 25 annotation in GTF format. For RNA-seq and Ribo-seq, a maximum of four and two mismatches was allowed, and multimapping of to up to 20 different positions was permitted. Alignments flagged as secondary alignments were filtered out, ensuring one mapping position per aligned read. P-sites positions and junction reads were extracted using Ribo-seQC[43] with default parameters. Statistics about the different Ribo-seq libraries are available as Supplementary Data 1. *SaTAnn* input files are available in Supplementary Data 1. *Gviz* was used to visualize data tracks and transcript annotation.

Polysome profiling:

DEXSeq[17] 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). Only genes with multiple translated transcripts were used.

Nuclear-cytoplasmic comparison:

DEXSeq[17] 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 (e.g. discarded transcripts, selected transcripts etc…) were selected.

5'end of endonucleolytic cuts:

Bigwig files for the different libraries were obtained from the GEO accession GSE57433 and normalized by library size. Coordinates were lifted to hg38 and overlapped with *SaTAnn*-identified stop codon positions, for both NMD candidates and controls ("canonical" stop codons taken from the same genes). Stop codon regions of NMD candidates overlapping known CDS regions were removed.

Merging *SaTAnn* 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. Enrichment for ORF categories at different level of overlap were calculated using normalized residuals from a chi-squared test. GO enrichment was performed using the *topGO* package.

Proteomics database search:

Raw data was downloaded using the PRIDE accession PXD002395. RAW data was searched using MaxQuant[28] version 1.6.0.13, using Carbamidomethyl as fixed modification, and oxidation of Methionine and acetylation protein N-termini as variable modifications. Quantification was performed using only unique peptides. Matching between runs was enabled. We used a custom database to perform the

peptide search: *SaTAnn*-detected ORFs were merged in a unique database, choosing only ORFs explaining a minimum of 10% of gene translation in at least one cell line. The final protein database, consisting of roughly 44,500 entries, is available as Supplementary File 2, together with the XML file used to perform the MaxQuant search and the set of identified proteins.

Comparison between protein abundance and translation estimates:

iBAQ values for summed up for each replicate. *P_sites_pNpM* values were summed for all ORFs mapping to each protein group. *ORF_pct_iBAQ* values were obtained by dividing each iBAQ value for the total of iBAQ values mapping to each gene, discarding protein groups mapping to multiple genes. The same procedure was applied to *P_sites_pNpM* values to compare protein and translation estimates. Only proteins detected by Ribo-seq (or RNA-seq) and proteomics were used. Gene-level TPM values in Supplementary Figure 11 were calculated using kallisto[30].

Data availability

SaTAnn is available at https://github.com/lcalviell/SaTAnn. Ribo-seq for HepG2 and K562 can be accessed at GEO under the accession number XXX.

Author contributions

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

Acknowledgements

L.C. thanks Stephen Floor (UCSF) for support during the preparation of this manuscript.

Supplementary files

Supplementary Table 1: Summary of Ribo-seq datasets analyzed in this study.

Supplementary Data 1: Archive containing all P-sites positions and junction reads, together with the set of *SaTAnn* identified ORFs, for each cell line.

Supplementary Data 2: Archive containing the set of identified proteins, including their Ribo-seq statistics, the XML file used for the MaxQuant run and the custom protein database.

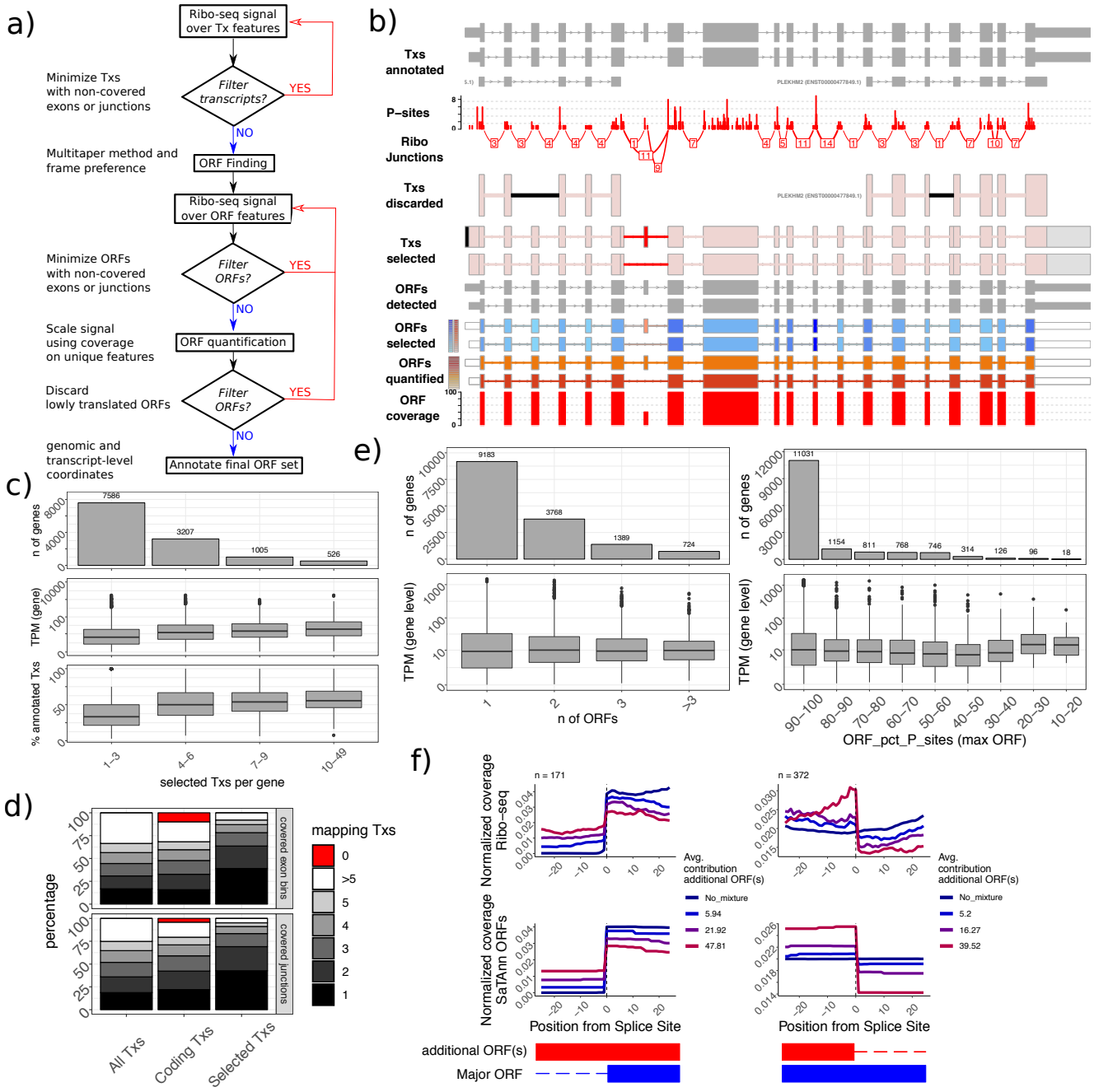

Figure 1: The *SaTAnn* strategy to quantify translation on selected transcripts. a) The *SaTAnn* workflow; b) the *PHLEKM2* gene as an example: displayed tracks represent, from top to bottom: 1) complete annotation, 2) P-sites positions, 3) junction reads, 4) discarded transcripts, 5) selected transcripts, 6) detected ORFs, 7) selected ORFs, 8) quantified ORFs and 9) ORF coverage (defined here using the fraction of gene translation). Colors for discarded and selected transcripts indicate unique features with no signal (black); shared features with no signal (grey); unique features with signal (red); and shared features with signal (pink). Colors for discarded and selected ORFs indicate signal in shared features (blue heatmap) and signal in unique features (red heatmap). For the quantified ORFs, the heatmap indicates ORF coverage values (0-100). Thick bars indicate CDS regions, as defined by the annotation or by *SaTAnn* (de-novo). c) Number of selected transcripts per gene (x-axis) against number of genes (top), TPM levels (middle), and % of annotated transcripts (bottom). d) Percentage of covered junctions (top), or covered exons (bottom) mapping to a different number of structures using all transcripts, protein-coding transcripts only or selected transcripts only. e) The number of detected ORFs (x-axis) is shown against number of genes and their TPM values (left). On the right, the number of genes (y-axis) (and again their TPM values) are plotted against the contribution (in percentages) of their major ORF. e) Aggregate plot of Ribo-seq coverage (normalized 0-1 per each region) and ORF coverage (*ORF_pct_P_sites_pN*, Methods) over candidate alternative splice sites regions (left) as defined by SaTAnn. No mixture indicates one ORF only, while other tracks indicate the presence of additional ORFs, divided by their summed translation values. Explanatory scheme at the bottom, with blue representing the major ORF and red the additional ORF(s). On the right, same plot for a different alternative splice sites conformation.

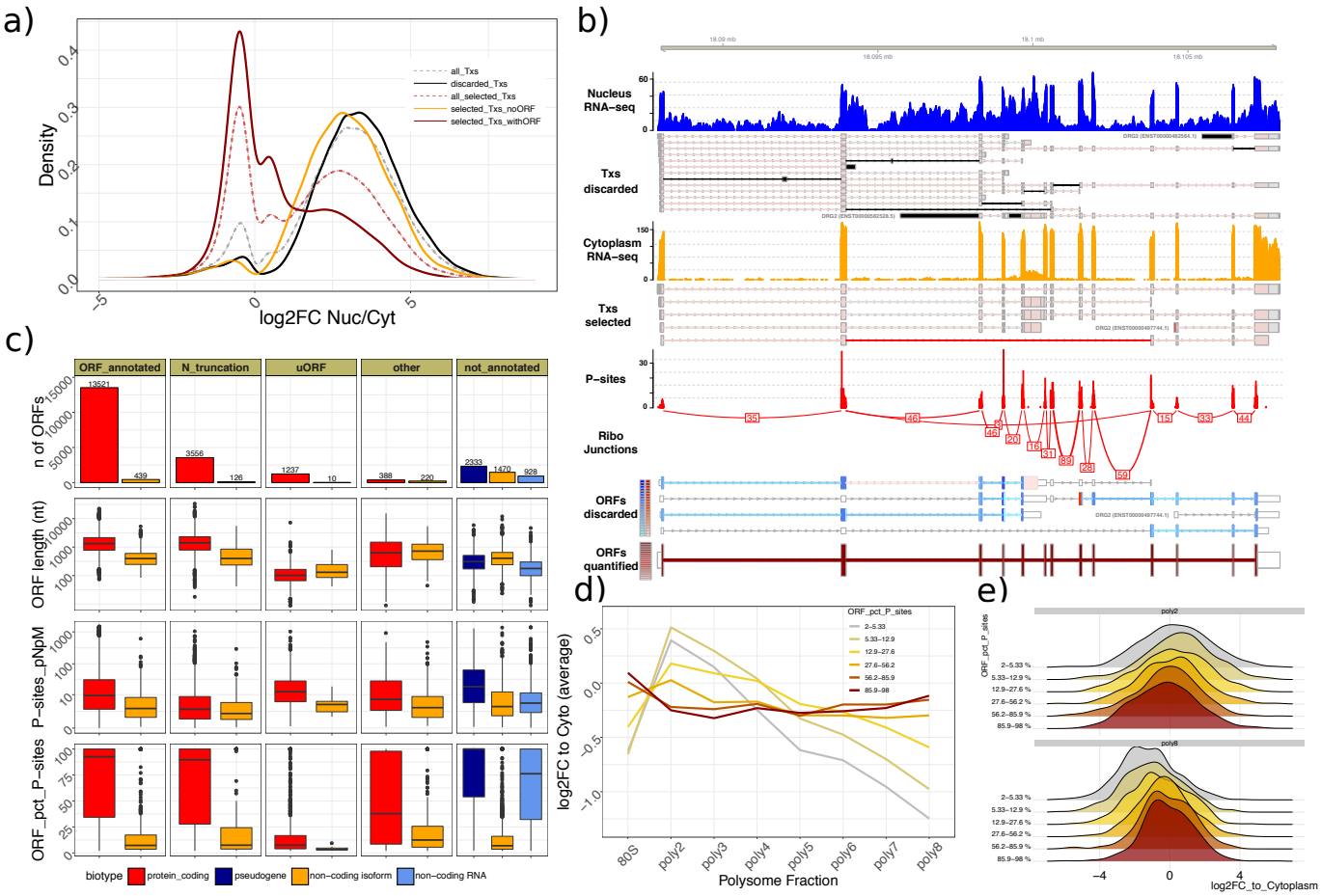

Figure 2: Quantification of translation on cytoplasmic mRNAs. a) Density of exonic fold changes for nuclear and cytoplasmic RNA-seq for different transcript classes. Negative values indicate more cytoplasmic abundance, while positive values indicate enrichment in the nucleus. b) The *DGR2* locus as example: tracks represent, in descending order: 1) Nuclear RNA-seq coverage, 2) discarded transcripts, 3) cytoplasmic RNA-seq coverage, 4) selected transcripts, 5) P-sites positions, 6) junction reads, 7) discarded ORFs, 8) quantified ORFs. Color representation as in Figure 1b. c) Overview of the *SaTAnn*-derived translatome. Number of ORFs, ORF length in nucleotides, length-normalized quantification and % of gene translation are shown, stratified by ORF category and annotated biotype. ORF_annotated represents ORFs whose structure perfectly matches the annotated CDS; other represents additional ORFs (e.g. nested ORFs, overlapping ORFs, dORFs), while not_annotated represents ORFs in transcripts with no CDS annotation. d) Average exonic fold changes with respect to cytoplasmic abundance (y-axis) for different polysome fractions (x-axis) for ORFs exhibiting different levels of translation within the same genes. e) Density plot of aforementioned exonic fold changes for two polysome fractions and for different ORF classes.

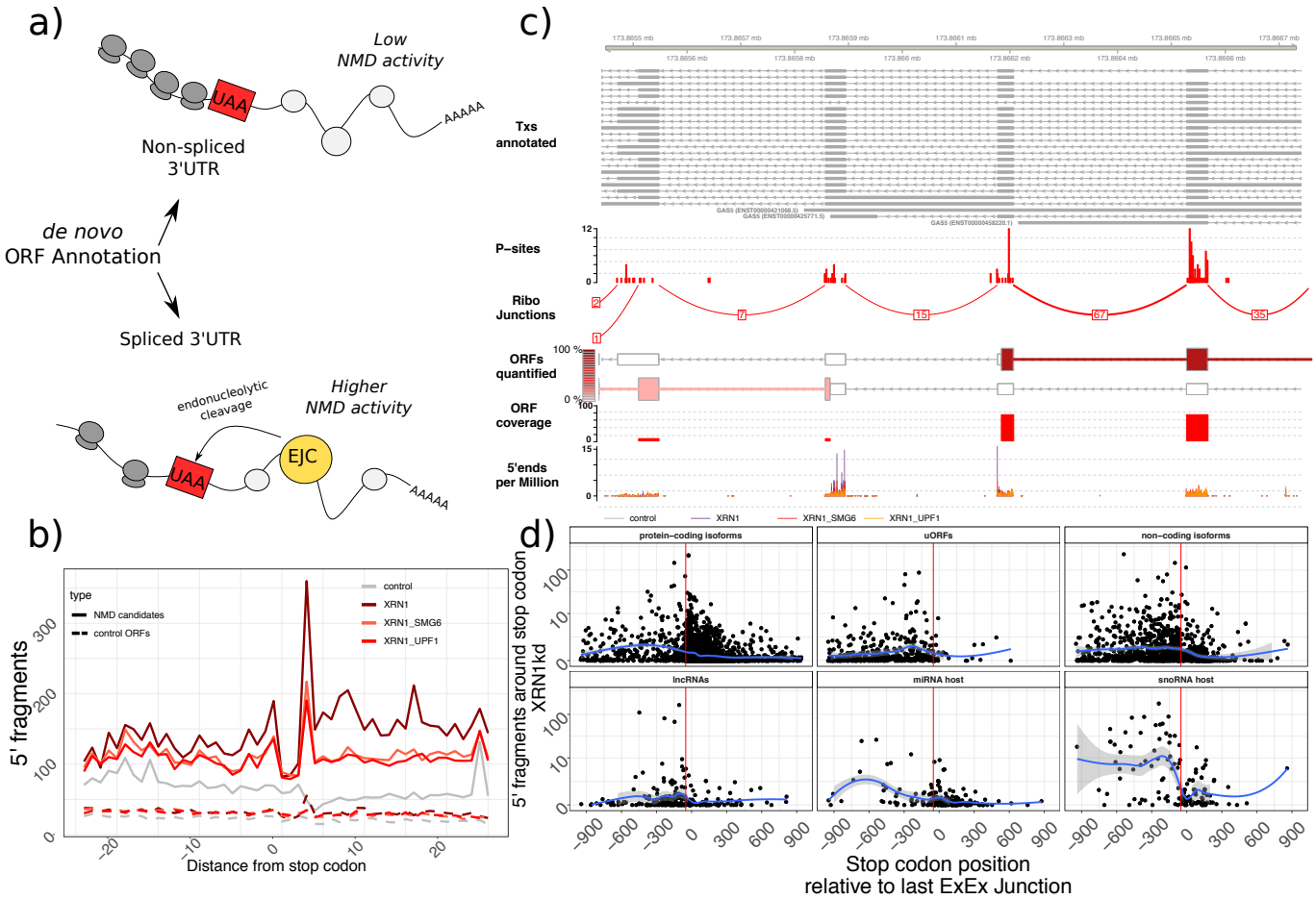

Figure 3: de-novo annotation of NMD candidates. a) Schematic annotation of NMD candidates. b) Aggregate profiles of 5' fragments around stop codons of NMD candidates and controls from the same genes. c) Example of a translated ORF in the *GAS5* gene. d) Number of 5' fragments observed in an *XRN1* knockdown experiment around stop codons (y-axis), versus the distance between stop codons and the last exon-exon junction (x-axis), for different transcripts/ORF classes. 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.

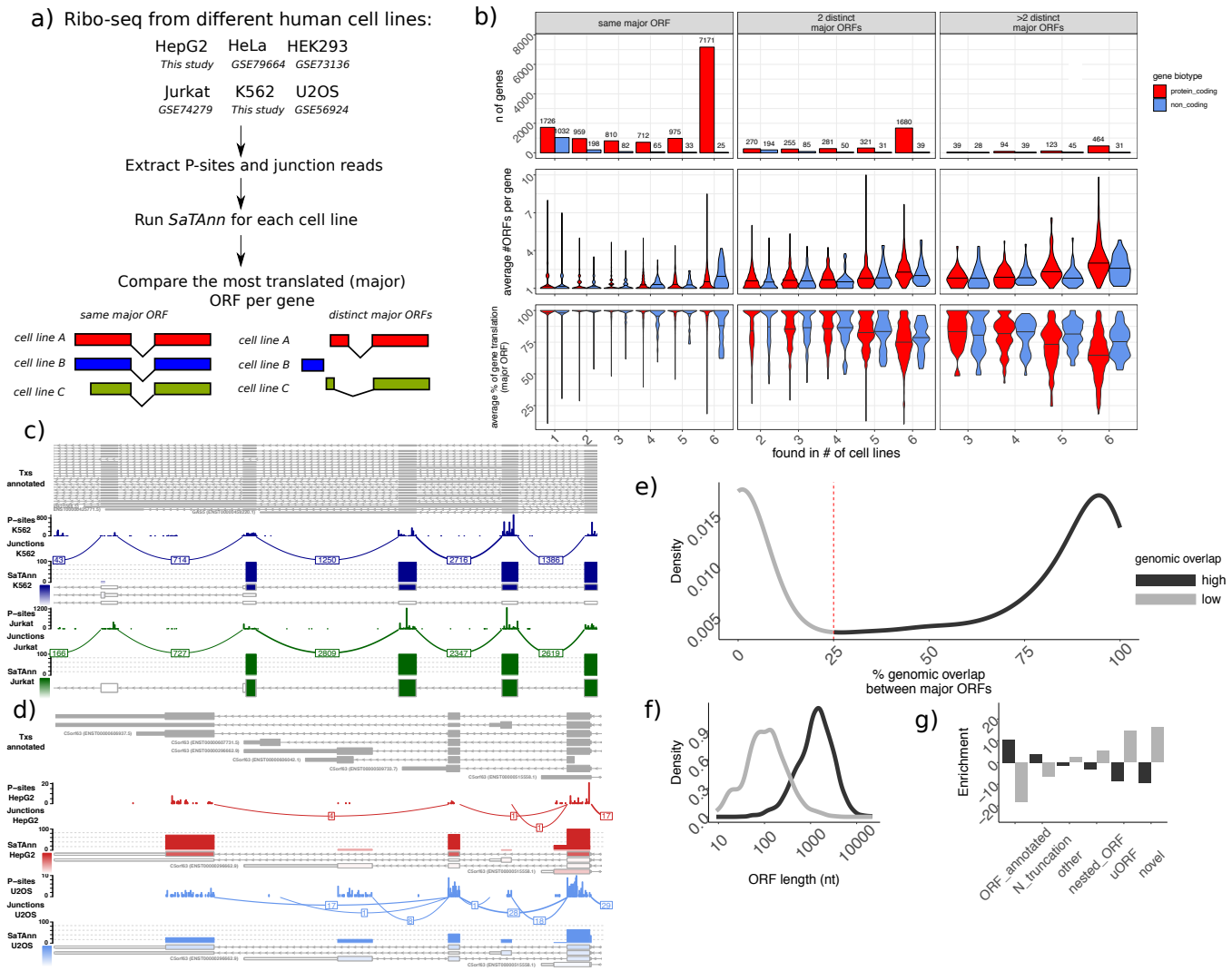

Figure 4: Diversity in gene translation across cell lines. a) Workflow for the analysis of the different datasets. b) Number of genes (top), average number of detected ORFs (middle) and average % of gene translation (bottom), for each number of cell lines where the gene harbored a detected ORF. Colors indicate the gene biotype. Genes translating one or more distinct major ORFs across cell lines are shown separately. c) Detected ORFs and Ribo-seq signal in the *GAS5* gene in K562 and Jurkat 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 for major ORFs with high and low degree of overlap.

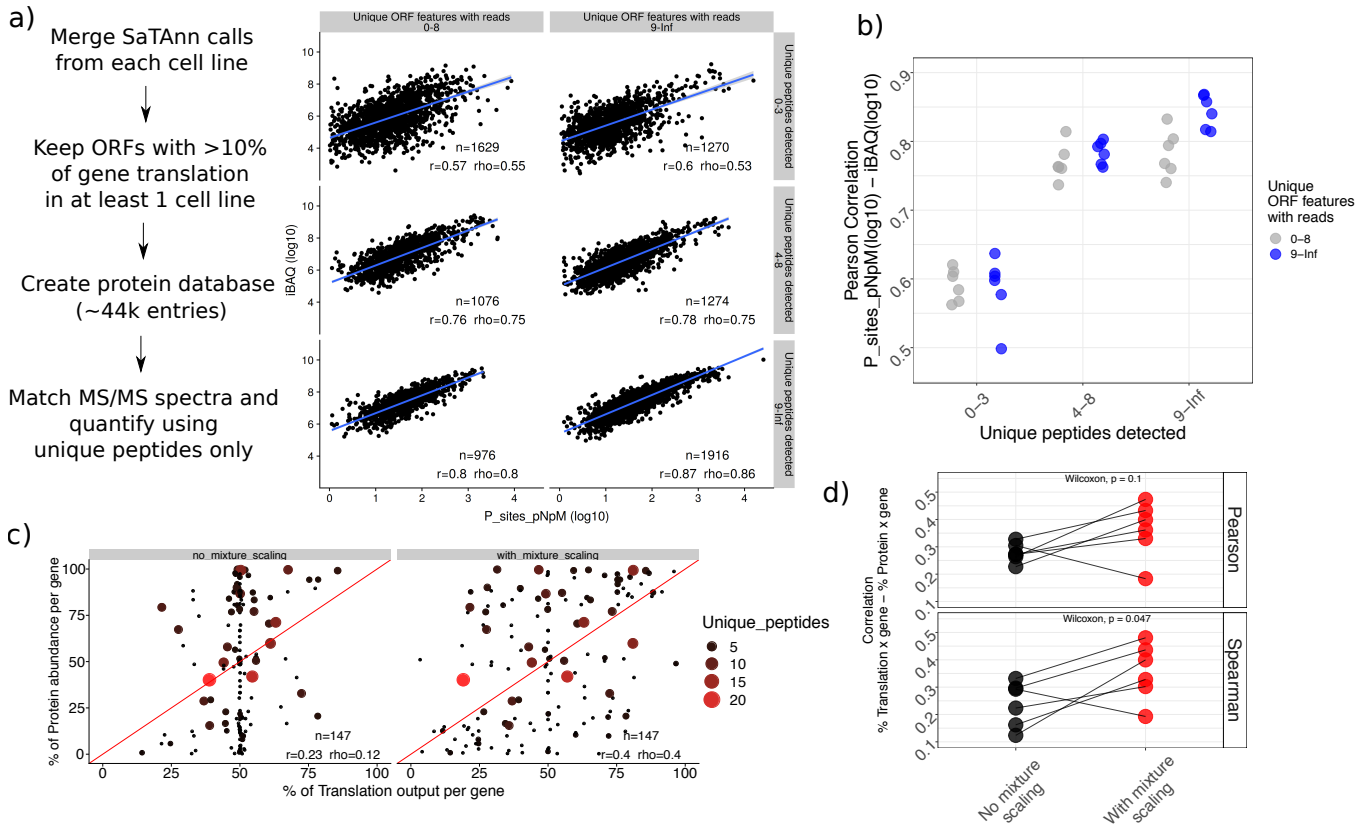

Figure 5: Agreement of protein synthesis with steady-state protein abundance estimates. a) Workflow of the proteomics analysis (left). On the right, *iBAQ* values (y-axis) versus length-normalized translation quantification estimates; proteins are split in multiple groups based on the number of detected unique peptides (proteomics) or unique covered features (Ribo-seq). b) Correlation values (as in a) shown for all the assayed cell lines. c) % of gene protein abundance (y-axis) plotted against % of gene translation (x-axis), with translation quantification performed 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. d) Correlations from c) shown for the 6 cell lines. P-values derived from Wilcoxon rank-sum test (one-sided).