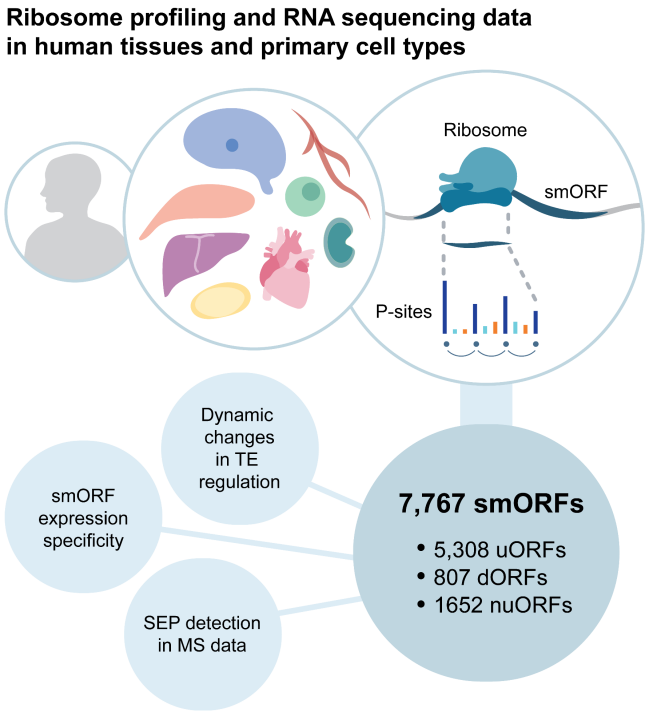
eTOC blurb = “In Brief” (max 50 words, target to non-specialists, in third person)

Chothani et al evaluate ribosome-mRNA interactions across six cell types and five tissues and reveal 7,767 small regions outside of the protein-coding genome that are actively translated in humans. These include specifically expressed and highly conserved small open reading frames. Integration of proteomics reveals more than 600 small proteins.

Highlights (3–4 full sentences as bullet points. Each no more than 85 characters in length, including spaces)

- Ribo-seq in 11 human primary cells and tissues reveals 7,767 high-confidence smORFs
- smORFs exhibit cell-type/tissue specificity and 603 SEPs were detected in MS-data
- Dynamic changes in TE of uORF and mainORF pairs are mostly homodirectional
- An interactive browser for this study can be found at: smorfs.ddnetbio.com

Graphical abstract



A high-resolution map of human RNA translation

Authors

Sonia P. Chothani[1,13], Eleonora Adami[1,2,13], Anissa A. Widjaja[1], Sarah R. Langley[3], Sivakumar Viswanathan[1], Chee Jian Pua[4], Nevin Tham Zhihao[3], Nathan Harmston[5,6], Giuseppe D'Agostino[3], Nicola Whiffin[7], Wang Mao[1], John F. Ouyang[1], Wei Wen Lim[1,4], Shiqi Lim[4], Cheryl Q.E. Lee[1], Alexandra Grubman[8,9,10], Joseph Chen[8,9,10], JP Kovalik[1], Karl Tryggvason[1], Jose M. Polo[8,9,10], Lena Ho[1], Stuart A. Cook[1,4,11,14], Owen J.L. Rackham[1,12,14,*], Sebastian Schafer[1,4,14,15,*]

Affiliations

[1] Program in Cardiovascular and Metabolic Disorders, Duke-National University of Singapore, Singapore 169857, Singapore
[2] Cardiovascular and Metabolic Sciences, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), 13125 Berlin, Germany
[3] Lee Kong Chian School of Medicine, Nanyang Technological University, Clinical Sciences Building, Singapore, 308232, Singapore
[4] National Heart Research Institute Singapore (NHRIS), National Heart Centre Singapore, Singapore 169609, Singapore
[5] Program in Cancer and Stem Cell Biology, Duke-NUS Medical School, Singapore 169857, Singapore
[6] Science Division, Yale-NUS College, Singapore, 138527, Singapore
[7] Wellcome Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK
[8] Department of Anatomy and Developmental Biology, Monash University, Wellington Road, Clayton, VIC 3800, Australia
[9] Development and Stem Cells Program, Monash Biomedicine Discovery Institute, Wellington Road, Clayton, VIC 3800, Australia
[10] Australian Regenerative Medicine Institute, Monash University, Wellington Road, Clayton, VIC 3800, Australia
[11] London Institute of Medical Sciences, London, UK, W12 ONN
[12] School of Biological Sciences, University of Southampton, UK
[13] These authors contributed equally
[14] Corresponding and senior authors (equal contribution)
[15] Lead Contact
[*] Correspondence: owen.rackham@duke-nus.edu.sg (O.J.L.R.), sebastian@duke-nus.edu.sg (S.S.)

Summary

Translated small open reading frames (smORFs) can have important regulatory roles and encode microproteins, yet their genome-wide identification has been challenging. We determined the ribosome locations across six primary human cell types and five tissues and detected 7,767 smORFs with translational profiles matching those of known proteins. The human genome was found to contain highly cell-type- and tissue-specific smORFs and a subset encodes highly conserved amino acid sequences. Changes in translational efficiency of upstream-encoded smORFs (uORFs) and the corresponding main ORFs predominantly occur in the same direction. Integration with 456 mass spectrometry datasets confirms the presence of 603 small peptides at the protein level in humans and provides insights into the subcellular localisation of these small proteins. This study provides a comprehensive atlas of high-confidence translated smORFs derived from primary human cells and tissues in order to provide a more complete understanding of the translated human genome.

Introduction

Human molecular studies typically focus on proteins encoded in annotated open reading frames (ORFs). Ribosome profiling (Ribo-seq) (Ingolia et al., 2009) data has shown that ribosome density in ORFs affects protein levels in cellular (Chothani et al., 2019a) and genetic models of disease (Schafer et al., 2015), as well as in patients (van Heesch et al., 2019). The definition of ORFs is typically based on sequence analysis that considers long stretches between in-frame start and stop codons as a coding region. This approach has historically limited our studies to proteins longer than 100 amino acids. However, recent studies revealed that shorter small open reading frames (smORFs) can be translated and encode small peptides (SEPs) (Couso and Patraquim, 2017; D'Lima et al., 2017; Ho et al., 2017; Lee et al., 2021; Makarewich and Olson, 2017; Matsumoto et al., 2017; Pueyo et al., 2016; Ruiz-Orera et al., 2014). SEPs are implicated in a variety of cellular processes, such as DNA repair or myogenesis (Bi et al., 2017; Quinn et al., 2017; Slavoff et al., 2014; Zhang et al., 2017b), and constitute therapeutic targets (Ho et al., 2017; Lee et al., 2021). Translation of smORFs in the 5' UTR (uORFs) may also serve as a regulatory process for canonical ORFs (Couso and Patraquim, 2017), and their perturbation can cause disease (Whiffin et al., 2020).

Ribo-seq can greatly aid in the identification of translated smORFs (Hao et al., 2017; Hsu et al., 2016; Olexiouk et al., 2018; Wan and Qian, 2014; Xie et al., 2016) and studies have described translation beyond the annotated protein-coding genes as pervasive (Dunn et al., 2013; Ingolia et al., 2011, 2014). However, Ribo-seq protocols are not standardised, and data can vary greatly in depth and quality (Hsu et al., 2016). Computational smORF detection tools (Bartholomäus et al., 2021; Calviello et al., 2020; Ji et al., 2015; Tjeldnes et al., 2021) also differ in their estimates of smORF abundance. Existing repositories range from 100,000s (Hao et al., 2017) to more than a million Ribo-seq smORFs (Olexiouk et al., 2018). Most tools and databases focus on the 3-nucleotide pattern, so called periodicity, of ribosomes translating codons to predict smORFs. To improve their predictions, these tools either model noise in the available data (PRICE, (Erhard et al., 2018)), incorporate multiple data sources (RiboTISH, (Zhang et al., 2017a)) or limit smORFs to canonical start codons (Ribotaper, (Calviello et al., 2016)). Together with the shallow depth of datasets, this has made it difficult to estimate the overall prevalence of translated smORFs in humans and characterise them on a global level. Currently, only 770 smORFs have been incorporated in Ensembl, although there are now efforts to systematically incorporate Ribo-seq smORFs into public gene databases (Mudge et al.). A further major limitation of the existing smORF annotations is that most human Ribo-seq studies have been performed on cell lines. These immortalized cells are not an accurate representation of human physiology (Gillet et al., 2013; Liu et al., 2019). Out of 102 studies profiling human samples currently listed in RPFdb (Xie et al., 2016), only 10 use primary biological material. Our understanding of smORF translation in primary cells or tissues in humans is still very limited.

To address this, we have generated an ultra-high depth RNA- and Ribo-seq dataset across six human primary cell types and five human tissues (**Fig. 1A)** that is quality-matched and can be analysed in its entirety**.** We then developed a tailored smORF pipeline to identify smORFs. Combining this with the integration of sequence analysis, amino-acid conservation and mass spectrometry has allowed us to identify and characterise thousands of human smORFs. We here provide the first comprehensive atlas of smORFs derived from multiple primary human cells and tissues.

Results

An ultra-deep collection of primary human Ribo-seq

At the ribosomal peptidyl site (P-site), the ribosome matches the transfer-RNA (tRNA) to the codon located in the coding frame. Only high-quality Ribo-seq data reveals P-sites at

single-nucleotide resolution, resulting in a prominent 3-nt periodicity signature within the coding sequence (CDS) (**Fig. 1B**). smORFs are also short and thus contain few P-sites compared to long annotated ORFs (Ensembl-ORFs). Thus only very deep and high-quality Ribo-seq data is suitable for accurately distinguishing active translation from background noise (Hsu et al., 2016).

To address this, we generated an ultra-deep, global Ribo-seq compendium of public and newly generated data that allows analysis of translation across multiple primary human samples. We first screened published Ribo-seq datasets and filtered each read length within each dataset for 3nt-periodicity, retaining samples with a periodicity of >60% in the CDS of Ensembl-ORFs. Only a small subset of data curated on RPFdb (Xie et al., 2016) was derived from primary human material (10 out of 102 datasets) and quality is highly variable. Only 3 out of 10 datasets passed our filter criteria (**Table S2**). We next processed our recently published atrial fibroblast (Chothani et al., 2019a) and heart tissue (van Heesch et al., 2019) datasets, which resulted in a total of 10.9B publicly available Ribo-seq reads with suitable 3nt-periodicity. To extend our dataset, we performed matched RNA-seq and Ribo-seq profiling of human embryonic stem cells (ESCs), primary human atrial fibroblasts (AFs), primary human coronary artery endothelial cells (HCAECs), primary human hepatocytes, primary human vascular smooth muscle cells (VSMCs), human umbilical vein endothelial cells (HUVECs), brain (thalamus) tissue as well as human visceral and subcutaneous fat tissue (**Fig. 1A**). We processed all samples uniformly and filtered as above, resulting in a total of 16.7B high-quality Ribo-seq reads which could be taken forward for analysis (**Fig. S1, Table S3, Table S4**).

After removal of abundant RNA species (mtRNA, tRNA, rRNA) and multi-mapping reads, we obtained a global snapshot of RNA translation across multiple primary human cell types and snap-frozen tissues consisting of a total of 1.3B P-sites (**Fig. 1C**). Ribo-seq reads were predominantly ~29nt in length and located within the CDS of Ensembl-ORFs (**Fig. S2**). On average, the 3nt-periodicity was 85% across all datasets (**Fig. 1D, Fig. S3A**). As expected, P-sites, which are inferred from Ribo-seq reads (also called ribosome pretected fragments (RPFs)), were enriched at known start codons, predominantly occupied the first frame in the CDS and were absent after the stop codon (**Fig. 1D, Fig. S3B, Fig. S4A**). Known smORFs such as those encoding *MYMX*, *MOCCI* and *SEHBP* (Bi et al., 2017; Koh et al., 2021; Lee et al., 2021; Zhang et al., 2017b) were well represented (**Fig. 1E, Fig. S4B-E**). Across our combined dataset, we found that more than 79% of codons within Ensembl-ORFs were covered. Taken together, these data show that our global P-site compendium covers a large fraction of the annotated human translatome with high resolution and substantial depth.

Translational signatures inferred from Ribo-seq data guide systematic *de novo* ORF discovery

We detected ~7% of inferred P-sites (91M) outside of Ensembl-ORFs. To identify potential novel translated smORFs we first ran state-of-the-art Ribo-seq tools Ribotaper (Calviello et al., 2016), RiboTISH (Zhang et al., 2017a) and PRICE (Erhard et al., 2018). We then concatenated all human smORFs in sORFs.org (Olexiouk et al., 2018) to create a comprehensive set of 2,621,576 putative smORFs in humans (see Methods, **Fig. S5**). In part, this large number of putative smORFs is a reflection of permissive filtering employed by existing methods that have been tailored to analyse relatively shallow Ribo-seq data. Given the high depth of our compiled human Ribo-seq compendium, we have been able to refine this set of putative smORFs by applying stringent filters and assessing metrics that were not previously possible.

To robustly identify *bona fide* translated regions from this set of putative smORFs we sought to find candidates that mirror the ribosome occupancy in Ensembl-ORFs by considering three criteria: Firstly, we determined the fraction of P-sites in frame 1 (PIF), a metric that has

been utilised commonly for Ribo-seq-based smORF detection. Secondly, we calculated the fraction of codons occupied by P-sites predominantly in frame 1 (Uniformity). The presence of 3nt-periodicity across the entire length of the ORF has not been widely implemented by previous methods, likely due to a lack of high-depth data. Thirdly, we established a score that quantifies the efficient release of translating ribosomes at the stop codon of the smORF (Drop-off) (**Fig. 2A, Fig. S5, S6**). To better understand how these three scores reflect active translation in our dataset, we first determined the distribution of PIF, Uniformity and Drop-off scores across all expressed Ensembl-ORFs. The 95th percentile and mean for each of these scores was determined and subsequently used as a threshold for deciding high-quality evidence of translation. For a putative smORF to be considered actively translated with high confidence, it had to pass these thresholds in all three scores. If a smORF had multiple alternative starts in the coding frame, a single start site was determined based on the uniformity score (see Methods). This resulted in the identification of a total of 7,767 high-confidence smORFs, filling a gap in the current annotations for ORFs shorter than 100aa. (**Fig. 2B, Fig. S7A, Table S5**). These smORFs were located at diverse locations within the transcriptome, but a large proportion (68%, N=5308) of them are within the 5' untranslated regions (UTRs) of known protein-coding genes, making upstream ORFs (uORFs) the most prevalent smORF type. Beyond uORFs, we found a further 1,652 smORFs in transcripts previously annotated as 'non-coding' and will be referred to as novel unannotated ORFs (nuORFs). Less common were the 773 downstream ORFs (dORFs) located in the 3' UTR of known protein-coding genes (**Fig. 2C, Fig. S7B**).

Globally these 7,767 novel smORFs appear indistinguishable from existing protein-coding genes. For instance, plotting of P-sites around the start and stop codons of all high-confidence smORFs showed strong evidence of 3nt-periodicity only within the coding region but not before or after (**Fig. 2D**). We found that 49.9% of all smORFs had a cognate start-site (AUG), with near-cognate start codons CUG (14.9%) and GUG (8.9%) being the next most common (**Fig. 2E, Fig. S7C**). This frequency distribution confirms previous observations on translation initiation sites (Gao et al., 2015; Ingolia et al., 2011; Lee et al., 2012). Further mirroring Ensembl-ORFs, we also observed the Kozak motif sequence (Kozak, 1986) around the smORF start sites (**Fig. 2F**). Since the smORF detection pipeline is agnostic to both the start codon type and its sequence context, identification of sequence motifs known to improve both start codon recognition and translation initiation in eukaryotes provides independent confirmation of the smORF set.

Expression of smORFs in human cell types and tissues

Having identified this novel set of smORFs, we next quantified their transcription (RNA-seq TPM), translation (Ribo-seq TPM) and translational efficiency (TE, see (Chothani et al., 2019b)) levels for Ensembl-ORFs, uORFs, nuORFs and dORFs across all the assayed cell and tissue types. Different datasets with the same cell type or tissue were merged. On average, 4037 uORFs, 878 nuORFs and 432 dORFs are translated in each cell type/tissue (**Fig. 3A**). Comparing the Ribo-seq TPM distributions between smORF subtypes shows that the level of translation varies significantly. We find that uORFs tend to be as highly translated as Ensembl-ORFs, but despite also being located on the same transcript class, dORFs are translated at a significantly lower rate. Similarly, nuORFs, which are encoded on transcripts that have previously been annotated as long non-coding RNAs (lncRNAs), are more lowly translated than both uORFs or Ensembl-ORFs (**Fig. 3B** in fibroblasts**, Fig. S8** and **Table S6** across all other cell types). Overall, the translation efficiency (TE) of smORFs were comparable to the known ORFs with dORFs having lower TE as compared to other types (**Fig. 3C, Fig. S9**).

To extend our understanding of the expression profile of these smORFs beyond the cell types from which Ribo-seq was generated we utilised the FANTOM catalogue which contains gene expression information across 436 cell and tissue types (Abugessaisa et al.,

2017; FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014). To identify specifically expressed smORFs we integrated gene expression information from these samples with the 162 samples from our translation atlas data and grouped these into 48 clusters (see methods). For each cluster, we then calculated the Jensen-Shannon divergence (JSD) for each gene, with a smaller JSD signifying higher specificity to that cluster. This analysis reveals cell-type-specific Ensembl-ORFs such as *PLIN1*, a gene encoding Perilipin which coats lipid storage droplets in adipocytes, as being specific to adipocytes in line with previous literature (Greenberg et al., 1991). Similarly, this analysis also highlights several nuORFs, and genes with uORFs/dORFs, that are highly specific for particular cell types (**Data S2, Table S7**). For instance, CATG00000072615 encodes a translated nuORF, which is highly specific (JSD = 0.47) and highly expressed in endothelial cells (**Fig. 3D**). Revisiting the Ribo-seq data confirms this: we only detect evidence for the translation of this nuORF in both the human umbilical vein endothelial cells and human coronary artery/aortic endothelial cells (**Fig. 3E**) and not in any other cell types. Overall, we identify 240 nuORFs that are highly specific in their gene expression pattern and, as a result, may serve important roles for cell and tissue function and which make them more likely to harbour disease-causing mutations (Magger et al., 2012) or be dysregulated in disease (Lee and Young, 2013).

Upstream open reading frames in translational regulation

Unique transcriptomes establish distinct functional roles of primary human cells (Hon et al., 2017) and tissues (GTEx Consortium et al., 2017). Differential translation efficiency of RNA can also contribute to proteome diversity and disease regulation (Chothani et al., 2019a; Schafer et al., 2015). uORFs may act as translational regulators and have been shown to reduce the protein expression of the main ORF (Calvo et al., 2009). In this case, upregulation of the uORFs should result in the downregulation of the main ORFs and vice versa. To test whether this mechanism contributes to cell or tissue identity, we first identified differentially translated main ORFs in fibroblasts compared to all other cell types using the deltaTE method (Chothani et al., 2019b). Translationally regulated genes in fibroblasts were enriched for the presence of uORFs ($P = 8.62e^{-89}$) and also enriched for differentially translated uORFs ($P = 2.88e^{-67}$) (**Fig. 4A**). However, only a minor fraction of genes showed opposing regulation of the uORF and main ORF, whereas in 91.55% of cases, both the uORF and the main ORF were regulated in the same direction (**Fig. 4B, Table S8**). This trend was also confirmed for Brain tissue, Heart tissue, Kidney tissue and Coronary artery endothelial cells (**Fig. S10**). We have shown previously that TGFB1 leads to translational regulation in fibroblasts (Chothani et al., 2019a). We integrated the smORF annotation and re-analyzed the Ribo-seq time-series experiment of cardiac fibroblasts stimulated with TGFB1. The change in translation efficiency of both uORFs/mainORFs and associated significance p-value was quantified using deltaTE (Chothani et al., 2019b). Genes that are translationally regulated by TGFB1 in fibroblasts were enriched for the presence of uORFs (P = 8.99e-52) and mirroring the previous analysis only a minor fraction of genes showed opposing regulation of the uORF and main ORF. In 92.31% of cases, both the uORF and the main ORF were regulated in the same direction where the uORF and mORF were changing in-tandem across the dynamic transition (**Fig. 4C-G, Fig. S11A**). Splicing junctions in the UTR were found to be covered by RPFs, but a lack of 3nt-periodicity suggests these rarely translate uORfs or dORFs (**Fig. S11B, C**).

Evolutionary conservation of smORF-encoded peptide sequences

Transcript UTRs and lncRNAs are, to a certain degree, conserved across species and serve critical regulatory roles (Siepel et al., 2005). Within these previously thought non-coding genomic regions, we have identified a set of actively translated smORFs. If these give rise to biologically relevant smORF-encoded proteins (SEPs), it is expected that the encoded amino acid sequence would be more conserved than a matched background. To test this hypothesis, we obtained multiple sequence alignments across 100 vertebrates (Kent et al.,

2002) for smORFs that pass our filtering criteria (see **Fig. 2A**, **B**) and also for a background set of smORFs which had low PIF, Uniformity and Drop-off scores (see **Fig. 2A**, **B**). The backgrounds are GC and expression matched and we treat uORFs, dORFs and nuORFs independently in order to ensure that bias from these aspects is minimised. As a measure of AA conservation, we determined the percentage of identical amino acids (AA) across the length of a given smORF for each of the 99 species with respect to humans. The AA sequences of all three classes of smORFs: uORFs, dORFs and nuORFs were significantly ($p < 2.2e^{-16}$) more conserved than their matched background in Primates, Rodents and Carnivores. Overall, uORFs and dORFs were more conserved compared to nuORFs (See **Table S9**), and we observed a sharp decrease in conservation in more distant clades like Fish or Birds (**Fig. 5A, Fig. S12**).

We found 7,596 of the 7,767 newly identified smORFs had more than 60% AA-conservation in primates, with 2,807 of these with more than 60% AA-conservation in rodents. For example, the uORF, located in the 5'UTR of the RAS Guanyl Releasing Protein 3 gene (*RASGRP3*), is highly conserved at the amino acid level in rodents. This is confirmed by the analysis of Ribo-seq P-sites generated from rat heart and liver tissue (Schafer et al., 2015), in which we found clear evidence of translation for the orthologue in the rat. Furthermore, the Ribo-seq scores (PIF, Uniformity and drop-off) were comparable between rats and humans (**Fig. 5B**). We also found 925 smORFs with a positive decibans score according to PhyloCSF (Lin et al., 2011) and 313 smORFs with a significant p-value (<0.1) according to RNAcode (Washietl et al., 2011). These results show that some smORFs are conserved, remain translated across species and encode peptide sequences that are under evolutionary pressure.

Revisiting mass-spectrometry datasets reveals proteins encoded by smORFs

The conservation analyses presented above suggest that the amino acid sequence encoded by some smORFs is important (**Fig. 5A**). One potential explanation for this is that a subset of smORFs encodes stable and biologically relevant small peptides. To identify SEPs on a large scale, we first obtained 26 different mass-spectrometry datasets that match the tissues and cell types profiled in this study. These include the proteome of 16 heart regions (Doll et al., 2017), embryonic stem cells (Shekari et al., 2017) and data generated from the cytoplasm, nucleus and extracellular space of human fibroblasts and endothelial cells (Slany et al., 2016). The latter datasets reveal the subcellular localisation of peptides (**Fig. 6A**).

This data was processed to identify SEPs with matching peptide-spectrum using a two-step approach (see Methods). First, using a target decoy approach, the spectra with peptide-spectrum matches (FDR<1%) to the human Uniprot database were removed. Second, the remaining peaks were tested for peptide-spectrum matches to SEP sequences (FDR<1%). We detected a total of 614 SEPs encoded by 281 uORFs, 47 dORFs and 286 nuORFs with at least one peptide-spectrum match in at least one sample. These included the verification of previously highlighted endothelial cell-specific nuORF in CATG00000072615 (**Fig. 3F**) and conserved uORF in the RASGRP3 gene (**Fig. 5B**). Of the 603 SEPs detected, 111 SEPs had at least 10 different hits, 62 SEPs were found in at least 5 different cell types or regions, and 56 were also found to have multiple unique sequence hits in a given sample. Despite being more lowly expressed and less common than uORFs, nuORF-encoded peptides were more frequently detected using proteomics (**Fig. 6B**). Moreover, we found that smORFs which had a positive deciban score according to PhyloCSF were enriched (p<6.18x10-5, hypergeometric test) among those for which MS evidence could be retrieved.

We detected SEPs predominantly in the cytoplasm (n=138) and interestingly, we also identified 131 SEPs in the nuclear compartment, which included a disproportionately large number of uORF-encoded SEPs (**Fig. 6C**, **Table S10**). The host genes of these SEPs

encode transcription factors and other proteins known to be important for nuclear functions (see **Table S10**), such as Histone Deacetylase 5 (HDAC5) and activating transcription factor 2 (ATF2). Lastly, we also found 27 smORFs in the extracellular region, suggesting secretion of these SEPs after translation (**Table S10**).

Discussion

It is clear that there are many short coding regions of the genome that remain to be uncovered and which may hold important links to our understanding of human health and disease. Many previous efforts to annotate smORFs in humans have focused on immortalised cell lines such as HEK293, which, due to genomic instability, may contain artificial transcripts and spurious translation events that are not relevant for human physiology. In addition, smORF detection has been limited due to lower-depth and low 3nt-periodicity in the available datasets. Compounding this, the way in which smORFs have been identified varies greatly between studies making comparisons between them impossible. Here we combined Ribo-seq datasets from multiple primary human cell types and tissues to create a global P-site repository of high-quality and depth. We then developed a tailored bioinformatics pipeline that can identify distinct characteristics of the translational footprint of coding regions in order to define high-confidence smORFs, with Ensembl-ORFs serving as internal controls. We find that only a small fraction of smORFs currently listed in public databases show signatures of active translation similar to known protein coding regions. However, we do find that translation of smORFs is common and expand the annotated coding genome with 7,767 high-confidence ORFs (**Fig. 2**). This dataset covers ~80% of the known human coding genome and thus also 80% of uORFs and dORFs located on the same transcripts. nuORFs are often encoded on highly specific lncRNAs, which may result in them being slightly less represented in our smORF atlas. Despite this, with our integrative pipeline, we were able to find 1,652 nuORFs located on lncRNAs and which have translational signatures equivalent to Ensembl-ORFs. Our results show that community efforts to integrate Ribo-seq smORFs into genome annotations (Mudge et al.) are necessary but should employ strict quality control measures and uniform processing. To facilitate the future discovery and characterisation of smORFs specific to other cell types, disease states or developmental stages, we provide our analytical approach and all datasets used in this manuscript in raw and processed form for integration with future projects and an interactive online browser at smorfs.ddnetbio.com.

Of the 7,767 smORFs found in this study, we find that 5,347 are translated in all assayed conditions, whilst some are highly specific to certain cell types or tissues. uORFs are the most common subtype and are also the most highly expressed. They have previously been regarded as negative regulators of main ORF translation, however, we find that uORFs and their associated main ORF tend to be regulated in the same direction (**Fig. 4B, E**) in the context of cell- and tissue-specific translation and during fibroblast activation. This suggests either a prominent role for uORFs as positive regulators of translation or that uORFs and their associated main ORF are subject to the same external regulatory mechanisms. If some uORFs encode peptide subunits important for the function of the main ORF-encoded protein, their expression patterns should mirror each other.

Nonsense-mediated mRNA decay (NMD) is a eukaryotic surveillance pathway that triggers the decay of transcripts with premature stop codons (Hentze and Kulozik, 1999). After splicing, an exon junction complex (EJC) is bound to the transcript, which triggers NMD unless translating ribosomes remove the EJC. Premature stop codons initiate early ribosome release, leaving downstream EJCs intact, activating the NMD pathway (**Fig. S11B**). To date, it is unclear why splicing of untranslated regions does not trigger NMD. It has been suggested that NMD cannot be triggered close to the 3' end of the transcript (Hilleren and Parker, 1999; Muhlrad and Parker, 1999). We find that spliced regions are always occupied by ribosome footprints (**Fig S11C**). It thus may be possible that ribosomes prevent NMD in

the UTR. However, we find that most splicing junctions are covered by ribosomes that do not appear to actively translate RNA and move along the transcript without clear 3nt-periodicity. This suggests scanning ribosomes, at least in some instances, might be sufficient to prevent NMD. Ribo-seq bulk data does not reveal whether all splice junctions in all transcripts are covered by ribosomes and translation-unrelated mechanisms that prevent NMD may exist.

Apart from regulatory functions, smORFs can also encode functional peptides. This has been shown for individual SEPs, which act in processes such as DNA repair (Slavoff et al., 2014), myogenesis (Bi et al., 2017; Quinn et al., 2017; Zhang et al., 2017b), inflammation (Lee et al., 2021), cancer (Pang et al., 2020) and metabolism (Chugunova et al., 2019; Friesen et al., 2020; Makarewich et al., 2018; Stein et al., 2018). A stringent conservation analysis revealed that the amino acid sequence encoded in smORFs is significantly more conserved compared to a matched background, with 97.7% of human smORFs having more than 60% of amino acids identical in primates. Furthermore, using Ribo-seq from rat tissues, we found evidence that these conserved smORFs are translated, making rodents a suitable model to explore smORF function *in vivo*. This suggests that many SEPs may be biologically relevant, warranting further follow-up exploration. Functional screens may offer a high-throughput route to provide additional insights into smORF biology when a suitable phenotypic readout is available. However, given the diverse and often highly specific roles of known SEPs, in-depth single gene functional studies are required to better understand the relevance of individual smORFs and SEPs.

It is very challenging to detect short peptides using traditional proteomics approaches, which typically detect less than 100 SEPs that can be confirmed at the protein level (Ma et al., 2016; Slavoff et al., 2013). A more recent approach utilising HLA-I proteomics data was able to verify 320 SEPs at the protein level (Martinez et al., 2020). The integration of our improved smORF annotation and 26 human mass spectrometry datasets confirmed 603 SEPs at the protein level. Further advances in proteomics are needed to better understand how many smORFs encode stable SEPs. Subcellular resolution proteomics revealed that many SEPs are transported into the nucleus. Often these peptides are encoded in the 5' UTR of well-known transcription factor transcripts, and therefore could potentially function as cofactors. A recent study has identified a SEP that regulates more than 15% of the active transcriptome (Koh et al., 2021). Our results suggest that transcriptional regulation by SEPs may be a more common physiological role than previously anticipated.

This comprehensive map of human smORFs generated using primary human cell types and human tissues gives insights into an overlooked part of the genome, revealing potentially new players in health and disease and provides a resource for the scientific community to accelerate discoveries. Overall, we have identified over 7000 new ORFs, many of which are conserved and over 600 of which we have validated at the protein level. Having compiled this resource, it will now be possible for the community to expand on our analysis and utilise the identified smORFs as a starting point for further functional, mutational or evolutionary studies. For example, understanding the rate at which they mutated across the human population will help elucidate their role in disease or understanding the rate at which they acquired or lost in evolution will give a clearer role of these short open reading frames within the wider context of evolutionary dynamics. Our results suggest that smORFs and SEPs are common, have diverse functional roles and provide new opportunities for understanding mammalian physiology and disease.

Limitations of the study

Whilst this resource represents the first comprehensive atlas of human smORFs derived from high-quality and high-depth primary cells/tissues, there are a number of aspects that can be further developed. Firstly, despite our efforts to collect a broad range of samples as

possible, the number and diversity of samples is limited by constraints in access to human primary cells and tissues. Further investigation of additional cell types, disease-specific translation would require profiling of other systems. For instance, the current version of the study cannot exclude that there may be other conditions where uORF translation represses main ORF TE levels. Future versions of this resource could also include translation initiation mapping, using drugs such as harringtonine, lactimidomycin to more accurately identify the most used alternative start-site of a given smORF. Although the combined data included in this study has high-depth, lowly translated smORFs, which may or may not be functionally relevant, could have been missed. As our pipeline focuses on identifying ORFs that have high periodicity, uniform coverage and clear dropoff of ribosomes, it may miss ORFs that have overlapping translation with another ORF in a different frame thereby leading to lower periodicity. Functional characterization of individual smORFs is beyond the scope of this study, which is intended to serve as a global resource of translated small ORFs. As a result, beyond providing tissue specificity, conservation and (where possible) identification in mass spectrometry more detailed studies will be required to investigate the role of individual smORFs, or indeed the act of translation that generates them, as functional parts of human physiology. This resource provides a high-confidence set of smORFs that have translation signatures identical to known ORFs, which can be further investigated to answer such questions.

Acknowledgements

We thank Dr. Dennis Kappei for providing expert feedback on proteomics analysis. This work was supported by the Open Fund - Young Individual Research Grant (OFYIRG18nov-0014), Duke-NUS-GCR/2018/0017A Grant and the Academic Clinical Programme Charles Toh Cardiovascular Fellowship Award to S. Schafer and a Singapore National Research Foundation (NRF) Competitive Research Programme (CRP) grant (NRF-CRP20-2017-0002) and National Medical Research Council (NMRC) Young Investigators Research Grant (YIRG) (NMRC/OFYIRG/0022/2016) awarded to O. Rackham. L.H. is supported by NRF-NRFF2017-05 and HHMI IRSP 55008732.

Author Contributions

O.J.L.R, S.A.C, and S.S conceived and arranged funding for the study. O.J.L.R and S.S designed and managed the implementation of the study. S.P.C, E.A, O.J.L.R and S.S wrote the manuscript with input from coauthors. S.P.C performed data processing, ribo-seq analysis and pipeline development and implementation. E.A performed and organised the molecular biology experiments. S.P.C and E.A created the visualisation of results. S.L, A.A.W, S.V, W.M, W.W.L, assisted in molecular biology experiments and in vitro cell culture. S.R.L and N.T.Z assisted with mass spec analysis. G.D, J.F.O and N.H performed aspects of the data processing and N.H provided guidance for the evolutionary analysis. N.W. assisted in the manuscript preparation and genetic aspects of the ribo-seq analysis. C.J.P assisted with sequencing. J.M.P, C.Q.E.L, A.G, J.C, J.P.K, K.T provided samples collection and expertise in cell culture. L.H provided in-depth feedback during the manuscript preparation and provided expertise in smORF detection. Having provided an equal contribution towards this work, the order of the first and senior authors is arbitrary.

Declaration of Interests

S.S. and S.A.C. are co-founders and shareholders of Enleofen Bio PTE LTD. O.J.L.R. and J.M.P. are SAB members and shareholders of Mogrify Ltd. All other authors declare no competing interests.

Main Figure titles and legends

Figure 1: A high-depth and high-resolution dataset of mRNA translation in primary human cell types and

tissues. **A.** Schematic of the data (Ribo-seq and RNA-seq) used in this study. In-house: newly generated datasets. Public: published datasets (Brain (Gonzalez et al., 2014), Atrial fibroblasts (Chothani et al., 2019a), Heart (van Heesch et al., 2019), Skeletal muscle (Wein et al., 2014) and Kidney (Loayza-Puch et al., 2016)). **B.** Schematic illustrating the concepts of RPF (ribosome protected fragment) and relative inferred P-site position, used to determine the frame being read. Abbreviations: A-site (aminoacyl), P-site (peptidyl) and E-site (exit); AUG, start codon, UGA, stop codon. **C.** Bargraph indicating the total amount of sequenced raw reads and the retained reads (%) following each of the indicated pre-processing steps. **D.** Trinucleotide (3-nt) periodicity of the individual datasets, calculated using read lengths of 28-30nt. Data shown as positional heatmap and box and whiskers plot (line at median, the whiskers extend to the most extreme data point which is no more than 1.5*IQR (or interquartile range) from the box) **E.** *MYMX* transcript expression (RNA-seq: light grey) and translation (Ribo-seq: dark grey) are displayed together with inferred ribosomal P-site positions dominating the coding-frame (Frame 1, blue). Abbreviations: 5'UTR (5' untranslated region), CDS (Coding sequence), 3'UTR (3' untranslated region).

Figure 2: Comprehensive and systematic discovery of actively translated ORFs using the human translation dataset identifies smORFs with robust translation signatures similar to Ensembl-ORFs. A. Graphic illustrating the filters adopted to select smORFs with a robust mRNA translation signature: *P-sites In Frame* (PIF), *Uniformity*, *Drop-off* score. For each score, density plots based on the underlying data processed with different smORF-calling tools are shown to illustrate the rationale for their choice. The dotted line represents the threshold determined using Mean -2*SD (95 percentile) of fitted normal distributions for PIF and Uniformity and the Mean for Drop-off score. **B** Length distribution of Ensembl-ORFs (light blue) and filtered smORFs (dark blue). **C.** Bar graph showing smORF classification based on their mapped location. Abbreviations: uORFs (upstream ORFs; located on the 5'UTR of known protein-coding ORFs). dORFs (downstream ORFs; located on the 3'UTR of known protein-coding ORFs). NuORFs (novel unannotated ORFs; located on previously annotated non-coding transcripts). Overlapping CDS indicates cases in which uORFs/dORFs overlapped with the protein-coding ORF on the same transcript. **D.** P-site distribution around the Start and Stop codons of the final combined set of smORFs. **E.** Relative usage of codon start sites among smORFs. **F.** Translation initiation context for Ensembl-ORFs, filtered smORFs and background smORFs. Background smORFs: Randomly selected smORFs from low scoring smORFs (PIF < 40%, Uniformity < 40% and Drop-off < 90%) to match the set-size of filtered smORFs. Kozak consensus is given for reference at the bottom, where R stands for Adenine (A) or Guanine (G), *** denotes the start codon and -3 and +4 positions denote strong Kozak context.

Figure 3: RNA expression and translation of small open reading frames (smORFs) in human cell types and tissues. A. Stacked bar chart showing the number and type of translated smORFs (Ribo-seq TPM > 1) in each dataset. Of the 13 datasets, 10 total cell types and tissues were assessed for expression analysis. AEC and HCAEC merged together as HCAEC; Published and newly generated Brain data merged together; Skeletal muscle tissue removed failing read-depth QC for expression analysis; **B.** Ribo-seq TPM (transcript per million) distribution for each smORF category and known coding ORFs in fibroblasts. uORFs display higher mean translation levels (0.897) than dORFs (0.425; p-value $1.1e^{-97}$) and nuORFs (0.414; p-value $4.6e^{-55}$). Statistics: Student's t-test. **C.** TE (Translation efficiency) distribution for each smORF category and known coding ORFs in fibroblasts. **D.** Scatterplot of Jensen-Shannon divergence and Endothelial cell gene expression. Blue: Genes with uORFs and/or dORFs, Yellow: Genes with nuORFs. Genes in the top-right quadrant (e.g. CATG00000072615) represent genes that are both highly expressed and specifically expressed in endothelial cells. Gene symbols are displayed for genes (top 20) with the highest expression levels and JSD < 0.5. **E.** An example of endothelial cell-specific nuORF, CATG00000072615 encoded in a lncRNA showing Endothelial-specific Ribo-seq read coverage within the cell-types/tissues in this study.

Figure 4: Upstream Open Reading frames in translational regulation. A. Volcano plot showing translational efficiency (TE) for all Ensembl-ORFs in fibroblasts vs background. Log2(ΔTE) denotes the log fold change of TE for the annotated ORF. Log10(adjusted p-value) denotes the significance of the change in TE. Ensembl-ORFs are marked in blue if they also host an upstream ORF (uORF). A chi-square p-value of $8.62e^{-89}$ signified the preferential presence of uORFs hosted in differential-TE genes. **B.** A scatter plot showing TE fold-changes for differential-TE main ORF and uORF pairs in fibroblasts vs background. 91.55% of uORF-main ORF pairs change in the same direction and 8.45% in the opposite direction. uORF-mORF pairs with absolute log2 ΔTE greater than 5 were shown on the axis boundary. **C.** Ribo-seq and RNA-seq of TGFB1 stimulated atrial fibroblasts across a time-series (Baseline, 45mins, 2hrs, 6hrs, 24hrs) **D.** Volcano plot showing translational efficiency (TE) for all Ensembl-ORFs in TGFB1 stimulated fibroblasts (24 hours) vs baseline. Log2(ΔTE) denotes the log fold change of TE for the annotated ORF. Log10(adjusted p-value) denotes the significance of the change in TE.Ensembl-ORFs are marked in blue if they also host an upstream ORF (uORF). A chi-square p-value of $8.99e^{-52}$ signified the preferential presence of uORFs hosted in differential-TE genes. **E.** A scatter plot showing TE fold-changes for differential-TE main ORF and uORF pairs in fibroblasts vs background. 92.31% of significantly changing uORF-mORF pairs change in the same direction. **F-G.** Scatter plot for two exemplar uORF-mORF pairs showing in-tandem change in TE during fibroblast activation. Lines represent median TE values. Red: uORF, Blue: mORF. Header represents Ensembl gene ID hosting both uORF and mORF. Log2 fold change of TE and adjusted p-values are printed for the uORF and mORF respectively.

Figure 5: Evolutionary conservation of small open reading frames across 100 vertebrates. A. Difference in

the average percentage of Amino-Acid conservation for smORFs across 100 vertebrate species with respect to a matched background. Dark blue: upstream ORF; Light blue: downstream ORF; Yellow: novel unannotated ORF. Lighter colour: Background smORFs not detected as being actively translated. **B.** P-site periodicity plot for a 31AA uORF located in the 5'UTR of *RASGRP3* (RAS Guanyl Releasing Protein 3), which was detected to have active translation signature in the human translation dataset (84.55% PIF, 84.38% Uniformity and 99.74% Drop-off) and found to be conserved in the rat (Ribo-seq data from (Schafer et al., 2015);72.49% PIF, 78.12% Uniformity and 98.85% Drop-off) as well as other species (**C**).

Figure 6: **Large-scale re-analysis of Mass-spectrometry data reveals 603 smORF encoded peptides**. **A.** Schematic of the strategy employed for the re-analysis of published mass-spectrometry datasets to identify smORFs. **B.** Bubble chart illustrating smORF encoded proteins (281 uORFs, 47 dORFs and 286 nuORFs) having at least one MS-hit across different samples (Tissues, cell types). Each circle represents an MS-hit for a given smORF in a given sample type. The colour scale indicates the number of total hits and the circle size represents the number of unique peptide sequences found to match the smORF. **C.** MS-evidence of smORF encoded proteins in different subcellular localisations. Abbreviations: ESC (Embryonic stem cells), HUVEC (Human umbilical vein endothelial cells), NHDF (N Human dermal fibroblasts), SmoothP1acells (Smooth muscle cells), AF (Adipose fibroblasts), EC (Endothelial cells), and tissues: AV (Aortic valve), Ao (Aorta), LA (Left atrium), LV (Left ventricle), MV (Mitral valve), PA (Pulmonary artery), PV (Pulmonary valve), PVe (Pulmonary vein), RA (Right atrium), RV (Right ventricle), SepA (atrial septum), SepV (ventricular septum), TV (Tricuspid valve); A3689, A17, CP, KFcells, SV, LH, MK,vcavain as described in Doll et al. (Doll et al., 2017).

KEY RESOURCES TABLE

Reagent or Resource	Source	Identifier
Chemicals		
20/100 marker	IDT	51-05-15-02
Cycloheximide	Sigma Aldrich	C1988
MicroSpin S400HR Columns	GE Healthcare	27-5140-01
Novex TBE-Urea Gels, 15%, 10%	Invitrogen	EC68855BOX, EC68755BOX
Novex TBE Gels 8%	Invitrogen	EC62155BOX
Novex TBE-Urea Sample Buffer (2X)	Invitrogen	LC6876
Novex® Hi-Density TBE sample buffer (5X)	Invitrogen	LC6678
O'Range Ruler 10bp	ThermoScientific	SM1313
Phusion Hi-Fi PCR Master Mix	NEB	M0531S
Sybr Gold	Invitrogen	S11494
TRIzol	Invitrogen	15596018
Critical commercial assays		
TruSeq Ribo Profile (Mammalian) Library Prep Kit	Illumina	RPHMR12126
HT DNA HiSens Reagent Kit	Perkin Elmer	CLS760672
HT DNA 1K/12K/Hi Sens	Perkin Elmer	760517

LabChip		
HT RNA Reagent kit and RNA ladder	Perkin Elmer	760634, CLS960010
HT DNA 5K/RNA/CZE LabChip	Perkin Elmer	760435
Ribo Zero Magnetic Gold	Epicentre Illumina	MRZG12324
RNA clean and concentrator 5	Zymo Research	R1015
TruSeq Stranded mRNA Library preparation	Illumina	RS-122-2101/2
NextSeq 500 High Output kit v2 (150 cycles)	Illumina	20024907
NextSeq 500 High Output kit v2 (75 cycles)	Illumina	20024906
Qubit DsDNA BR Assay kit, 500 assays	Life Technologies	Q32851
Qubit RNA BR Assay Kit,500 assays	Life Technologies	Q10211
KAPA Library Quantification Kit Illumina GA with revised primers-SYBR Fast Universal	Roche	KK4824
Sample purification beads	Beckman coulter	A63881
Published data and annotations		
Ensembl ORFs	Ensembl hg38	Homo_sapiens.GRCh38.86.chr.gtf
FANTOM5 transcript models and CAGE counts	(Hon et al., 2017)	https://fantom.gsc.riken.jp/cat/?fd=source_data
Kidney Ribo-seq/RNA-seq	(Loayza-Puch et al., 2016)	SRP044937: SRR1528686-9, SRR2064424-5
Brain	(Gonzalez et al., 2014)	SRP031501: SRR1562539-SRR1562541 SRR1562544-SRR1562546
Skeletal muscle	(Wein et al., 2014)	SRP040550: SRR1204656, SRR1204658

Fibroblasts	(Chothani et al., 2019a)	GSE123018
Heart tissue	(van Heesch et al., 2019)	EGAS00001003263
Multiple sequence alignment	UCSC https://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz100way/maf/	chr#.maf.gz where #=1:22, X, Y, M
Mass-spec data	See methods for details	See methods for details
Deposited Data		
Raw Ribo-seq data	This paper	GSE182371
Raw RNA-seq data	This paper	GSE182372
Website	This paper	smorfs.ddnetbio.com
Software and Algorithms		
Trimmomatic	(Bolger et al., 2014)	http://www.usadellab.org/cms/index.php?page=trimmomatic
Bowtie	(Langmead et al., 2009)	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
STAR	(Dobin et al., 2012)	https://github.com/alexdobin/STAR/
Feature counts	(Liao et al., 2014)	http://subread.sourceforge.net/
deltaTE	(Chothani et al., 2019b)	https://github.com/SGDDNB/translational_regulation
Ribotaper	(Calviello et al., 2016)	https://ohlerlab.mdc-berlin.de/software/RiboTaper_126/
RiboTISH	(Zhang et al., 2017a)	https://github.com/zhpn1024/ribotish
PRICE	(Erhard et al., 2018)	https://github.com/erhard-lab/price
Sorfs.org	(Olexiouk et al., 2018)	http://sorfs.org/database
MS-GF+	(Kim and Pevzner, 2014)	http://proteomics.ucsd.edu/software-tools/ms-gf/

Star Methods

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact: Sebastian Schäfer (sebastian@duke-nus.edu.sg)

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Raw data can be downloaded from GEO superseries GSE182377 (Ribo-seq: GSE182371, RNA-seq: GSE182372). An interactive browser for all of the identified smORFS can be found at: smorfs.ddnetbio.com
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell culture and tissue collection

Cell types and tissues were selected to constitute a rich and diverse experimental resource, such that most genes would be expressed in at least one cell type, with the assumption that this would lead to most smORFs being expressed in at least one dataset.

Human primary atrial fibroblasts were prepared from atrial biopsies of patients undergoing CABG (coronary artery bypass grafting) in keeping with local guidelines (Singhealth Centralized Institutional Review Board 2013/103/C and 2018/2543) and cultured as described previously (Chothani et al., 2019a). Fibroblasts were either untreated or treated with cytokines (5 ng/ml) or antibodies (2 µg/ml) as indicated.

Primary human hepatocytes (5200, ScienCell) were maintained in hepatocyte medium (5201, ScienCell) supplemented with 2% fetal bovine serum, 1% Penicillin-streptomycin at 37°C and 5% CO_2. All experiments with primary cells were carried out at low cell passage.

Vascular smooth muscle cells (VSMCs) Experimental protocols involving human subjects were approved by the SingHealth Centralized Institutional Review Board (CIRB) (CIRB ref: 2013/103/C) in accordance with the ICH Guidelines for Good Clinical Practice and all participants gave written informed consent. Patients undergoing coronary bypass grafting (ages between 21 to 81) at the National Heart Centre Singapore were recruited to the study, and patients with prior valvular heart disease or previous atrial interventions were excluded. Human VSMCs were cultured as previously described (Lim et al., 2020). Aortic biopsies and left internal mammary artery trimmings were used to outgrow primary VSMCs. The tunica media was isolated under a dissecting microscope and minced into 1-2 mm^2 pieces and explanted onto 60 mm cell culture dishes coated with collagen I (C3867, Sigma-Aldrich) and maintained in complete M231 medium (M-231-500) with smooth muscle growth supplement (S-007-25) and 1% antibiotic-antimycotic (15240062) from Life Technologies, in a humidified atmosphere at 37°C and 95% air/5% CO_2. VSMC were negatively selected by magnetic separation with LD columns (130-042-901, Miltenyi Biotec) to exclude CD90+ fibroblasts (130-096-253, Miltenyi Biotec) and CD144+ endothelial cells (130-097-857, Miltenyi Biotec) at passages 1-2.

Pluripotent human embryonic stem cells (hESCs) were seeded and propagated in vitro on LN-521 to maintain pluripotency. Specifically, the hESC line H1 or HS1001 (sourced respectively from WiCell Research Institute and Karolinska Institute; NUS-IRB 12-451) were cultured in monolayer on plates pre-coated with purified human LN isoform overnight at 4°C at 10 µg/ml according to manufacturer's instructions (BioLamina) and maintained in

NutriStem hESC XF (Biological Industries, Israel) medium. Upon confluence, the cells were sub-cultured by trypsinization using TrypLESelect (GIBCO Invitrogen) for 8 min at 37°C, 5% CO_2.

Human umbilical vein endothelial cells (HUVECs): HUVEC cells were purchased from Lonza (C2519A, pooled donor) and were cultured with EGM-2 BulletKit medium (CC3162, Lonza) in a humidified incubator with 5% CO_2. HUVECs (P3 to P4) were harvested at confluence of 80–90% and subcultured into 100 mm petri-dish (Falcon) at the density of 5,000 cells per cm^2.

Human coronary artery endothelial cells (HCAEC): Primary HCAEC (CC2585, Lonza) were cultured and passaged in EGM-2 MV BulletKit medium (CC3202, Lonza). For experiments, HCAEC (P3) were seeded into a 100mm dish at a density of 5000 cells/cm^2.

Human aortic endothelial cells (HAEC): Primary HAECs (PromoCell® C12271) were cultured in endothelial cell growth media MV2 (PromoCell® C22022) with changes every two days. To subculture, HAECS were washed once with PBS, trypsinized (0.25% trypsin; Life Technologies 25200072), and, following neutralization with EGMV2 media, centrifuged at 500xg for 5 min and seeded at a density of 5000 cells/cm^2.

Brain tissue: Brain tissue (thalamus) was obtained from the Victorian Brain Bank and ethics approval was received for patient tissue banking and consent (University of Melbourne HREC Approval No.: 1545740) and for molecular analyses (Monash University MUHREC 2016–0554).

Adipose Tissue: The adipose tissue (visceral and subcutaneous) used in this study was collected from a volunteer (female, 32yo, Chinese, BMI 42, history of polycystic ovarian syndrome) during weight loss surgery at Singapore General Hospital, under approval by the local IRB (Singhealth CIRB 2015-12-14). During surgery, approximately 75 g of visceral and 15g of subcutaneous adipose tissue was removed. The specimens were washed with normal saline, divided into aliquots and stored at – 80°C.

METHOD DETAILS

Library preparation and sequencing

Generation of Ribo-seq libraries: Ribosome profiling was performed as previously described (Chothani et al., 2019a; Schafer et al., 2015).

For brain and adipose tissue, 80-100mg of tissue were placed in chilled tubes containing zirconia beads (11079110zx, Biospec) and 1ml cold lysis buffer supplemented with 0.1 mg/mL cycloheximide (formulation as in TruSeq® Ribo Profile Mammalian Kit, RPHMR12126, Illumina) and lysed using the Magnalyser machine (Roche) in pulses of 20s at 6000g so that the sample would remain cold. Samples were then centrifuged at 20,000g for 10min at 4°C to pellet debris.

Fibroblasts, hepatocytes, hESCs and endothelial cells were grown to 90% confluence in a 10cm culture dish, while VSMCs were pooled from 5 wells of a 6-well culture plate at baseline conditions in basal M231 medium (M-231-500) for 24h, and pelleted before snap-freezing in liquid nitrogen and stored at -80°C prior to lysis for Ribo-Seq. It was ensured that primary cells at low passage (≤ passage 4) were used for these experiments. Cell lysis occurred in the presence of 0.1 mg/mL cycloheximide in 1ml cold lysis buffer (formulation as in TruSeq® Ribo Profile Mammalian Kit, RPHMR12126, Illumina). After immediate repeated pipetting and multiple passes through a syringe with a 21G needle, sample lysates were cleared as described above. 400-800ul of supernatant recovered from homogenized and cleared lysates were then footprinted with Truseq Nuclease (Illumina). Ribosomes were purified using Illustra Sephacryl S400 columns (GE Healthcare), and the protected RNA fragments were extracted with a standard phenol:chloroform:isoamylalcohol technique. Following ribosomal RNA removal (Mammalian RiboZero Magnetic Gold, Illumina), sequencing libraries were prepared out of the footprinted RNA according to the

TruSeq Ribo Profile (Mammalian) Reference Guide, with the additional modification of 8% PAGE purification following the PCR amplification of the final library.

Generation of RNA-seq libraries: Total RNA was isolated using TRIzol Reagent (Invitrogen; 15596018) from cell pellets or 5-10 mg of the same tissue processed for ribosome profiling. Total RNA was DNase-treated and purified using the RNA Clean & Concentrator-5 kit (Zymo Research; R1013). RNA was quantified using a Qubit RNA BR Assay kit (Life Technologies) and its quality was assessed on the basis of their RNA integrity number using the LabChip GX HT RNA Reagent Kit (Perkin Elmer). Per sample, ~1µg was further processed for library preparation with the Truseq Stranded mRNA kit (Illumina) according to the manufacturer's protocols.

Sequencing: The final RNA-seq and ribosome profiling libraries were quantified using KAPA library quantification kits (Roche); the quality and average fragment size of the final libraries were determined using a LabChip GX HT DNA HiSens Reagent Kit (Perkin Elmer). Libraries with unique indexes were pooled and sequenced on a Hiseq / NextSeq 500 Illumina sequencer using 75-bp paired-end [RNA-seq: NextSeq 500 High Output kit v2 (150 cycles)] or 50-bp single-end [Ribo-seq: NextSeq 500 High Output kit v2 (75 cycles)] sequencing chemistry.

Transcript model annotation file construction

Ensembl hg38 transcript models (Yates et al., 2020) and FANTOM5 hg38 robust transcript models (hg38 liftover version provided by the authors of Hon et al. 2017(Hon et al., 2017)) were downloaded as GTF files. Several compatibility issues were resolved before merging the two databases of transcript models: First, for transcripts that had different host gene IDs between hg19 and hg38, their host genes were updated to the latest version (hg38). Second, transcripts with a strand information mismatch between both databases were discarded. Third, transcripts with incorrect chromosome annotation with respect to the host gene were also discarded. After resolving these compatibility issues, the Ensembl transcripts and FANTOM5 transcripts were merged into one GTF annotation file.

Unique and non-overlapping transcripts across both sources were added to a merged file without any changes. Common transcripts were combined by extending the Ensembl transcripts based on FANTOM5 annotation so as to allow maximum search space for smORF calling (see **Fig. S5B-E**). The gene biotype information was retained from the Ensembl annotation file for genes present in Ensembl, whereas for novel genes from the FANTOM5 annotation catalogue the biotype was derived from the FANTOM5 catalogue. The resultant annotation file was formatted to be used as a standard GTF file and was tested with several software for smooth functioning (STAR (Dobin et al., 2012), featureCounts (Liao et al., 2014) and Ribotaper (Calviello et al., 2016)).

Data pre-processing of Ribo-seq and RNA-seq samples

Ribo-seq and RNA-seq data were pre-processed as described previously (Chothani et al., 2019a). See **Table S1** for computational tools used in this study. Raw sequencing data were demultiplexed with bcl2fastq V2.19.0.316 to obtain fastq format files. The fastq file was processed to remove adaptors and low-quality bases using Trimmomatic V0.36 (Bolger et al., 2014). Demultiplexing and trimming of adaptors were carried out for both Ribo-seq and RNA-seq reads. Reads that were shorter than 20 nucleotides for Ribo-seq and 35 nucleotides for RNA-seq were discarded. RPFs represent the actively translated mRNA and the ribosomal RNA (rRNA), mitochondrial RNA (mtRNA) and transfer RNA (tRNA) sequences are considered contaminant sequences. Trimmed Ribo-seq reads were aligned using Bowtie2 (Langmead et al., 2009) to sequences present on RNACentral (release 5.0) database (The RNAcentral Consortium, 2017), a database of known rRNA, mtRNA and

tRNA sequences. The reads aligned to these contaminant sequences were discarded and the remaining unaligned reads were retained for further processing. The removal of contaminant sequences is not carried out for full-length RNA-seq datasets as these contaminant sequences are not prevalent in RNA-seq sequenced reads. After these pre-processing steps, both RNA-seq and Ribo-seq reads were aligned using STAR (Dobin et al., 2012) to the human genome (hg38) using the combined transcript models from Ensembl and FANTOM5. RNA-seq reads were aligned using the default settings of the tool for paired-end reads. Ribo-seq reads were aligned as previously described (Chothani et al., 2019b). Refer **Fig. S5F, Supp. Table 1** for processing pipeline and detailed commands.

Quality check filtering using 3nt-periodicity in known ORFs.

Ribo-seq datasets were further screened for their 3nt-periodicity signal across known ORFs to select high-quality data for smORF detection. RiboTISH (Zhang et al., 2017a) was used to quantify the transcriptome-wide periodicity near the start-codon and stop-codon in the Ribo-seq dataset. Samples with less than 60% 3nt-periodicity across all known ORFs for read lengths between 28nt to 30nt were discarded from any further analysis. High-quality samples with lower sequencing depth were re-sequenced for increased sequencing depth to allow detection of 3nt-periodicity signals even for lowly expressed genes. A heatmap around the start-codon and stop-codon was generated using pheatmap 1.0.8 R package to view the global 3nt-periodicity across different datasets.

Processing and QC of published Ribo-seq dataset

RPFdb v2.0 (Xie et al., 2016), a database of published Ribo-seq dataset, was used to mine dataset derived from human primary cells and tissues. Out of the 102 human Ribo-seq studies, we found 10 that were generated using primary human cells or tissues (See **Table S2**). Cell types or tissues that were redundant to in-house generated data or unavailable published data were not used for further analysis and the remaining four datasets were downloaded and re-processed. Trimming of adaptors was carried out using the recommended tool and adaptor sequence as per the methods described in original publications. The trimmed Ribo-seq and RNA-seq data were processed to obtain alignment files using the same processing pipeline as for the newly generated data in this study. Each dataset was evaluated for various quality check metrics including read depth, sequencing data quality, reads remaining after trimming adaptors or removing rRNA sequences, alignment and importantly, its average 3nt-periodicity. The 3nt-periodicity for each dataset was quantified using Ribo-TISH (see **Data S1**). High-quality datasets with high (60%) 3nt-periodicity were selected for further analysis in this study to obtain high accuracy by reducing the signal-to-noise ratio for smORF detection. Ribo-seq/RNA-seq data from two of our recently published human primary cells/tissue studies (Chothani et al., 2019a; van Heesch et al., 2019) passed QC thresholds and were prepared and processed the same as the newly generated data in this study.

P-site file construction

The ribosome protected mRNA fragment (RPF) does not directly denote the exact position of the codon being translated, i.e, the position of the peptidyl site of the ribosome (Refer **Fig. 1B**). Quantifying P-site positions is required for visualization and quantification of 3-nt periodicity for smORFs. Sample-wise processing of Ribo-seq was carried out to obtain P-site positions for each sequencing read. Each alignment file (.bam) was processed to retain only uniquely mapped reads. They were further processed such that any reads longer than 30nt and shorter than 27nt were discarded. The offset for the P-site position in each read length was calculated based on the known ORFs using RiboTISH (Zhang et al., 2017a). To filter stringently, we tested each read length from 27 to 30 base pairs, the expected length of a ribosome protected fragment (RPF), for 3nt periodicity across the Ensembl-ORFs and only

read lengths with more than 60% 3-nt periodicity were retained within that sample.

Each Ribo-seq aligned read was then processed to determine the position of the P-site by using the determined offset for the given read length and sample. For reads on the positive strand, the P-site position was determined as the read start + offset and for reads on the negative strand, the P-site position was determined as read length - offset + 1. Sequencing reads may have bases that are missing from the reference or bases compared to the reference. This information is stored as a string of characters called, Concise Idiosyncratic Gapped Alignment Report (CIGAR). A CIGAR string records whether the base pair is matched with reference alignment or skipped as a gap. In order to correctly offset the reads to find the P-site position and base-pair aligned to the reference genome, the CIGAR information was also incorporated. These P-site positions were quantified generating a P-site file for each Ribo-seq sample. The calculated P-site positions and sequences were stored as a .bam file with reads of length 1 including base pair sequenced at the P-site position (Refer **Fig. S5G)**.

Merging alignment files to obtain maximum depth for smORF detection

Individual cell-type and tissue RNA-seq alignment files, Ribo-seq alignment files and P-site files were merged together for maximum depth at the nucleotide-resolution for smORF detection. Samples were merged using samtools V0.1.18 to obtain three files: 1. Merged RNA-seq alignment file (.bam), 2. Merged Ribo-seq alignment file (.bam) and 3. Merged P-site file (.bam). These files were used to generate .bedgraph files using genomeCoverageBed for visualization at the nucleotide resolution (**Fig. 1E, 4E**).

Detection of small Open Reading Frames

The merged alignment files were used as input to smORF calling tools for maximised depth in genome-wide P-site coverage thereby increasing the possibility of comprehensive smORF detection. RNA-seq files were downsampled to 10% for maintaining similar depth to Ribo-seq data and quicker runtimes. The smORFs identified in this study were based on *de novo* identification of smORFs from our dataset and further incorporated with smORFs detected and listed previously on sorfDB (Olexiouk et al., 2018), a database of smORFs. For de-novo detection of smORFs, we used three independent tools: Ribotaper, a spectral analysis method that leverages the 3-nt periodicity exhibited by actively translating ribosomes (Calviello et al., 2016); PRICE, a computational method that optimizes noise in the Ribo-seq data to identify smORFs and also is able to detect smORFs overlapping with known-protein coding ORFs (Erhard et al., 2018) and RiboTISH, a toolkit identifying smORFs also with non-canonical start-sites (Zhang et al., 2017a). The tools were used in the default setting on the merged human translation dataset using the Ensembl+FANTOM5 transcript model annotation prepared in this manuscript. The smORF detection was carried out chromosome-wise for parallel processing.

For ribotaper, the merged P-site alignment file, the merged RNA-seq alignment file and an annotation index based on Ensembl+FANTOM5 transcript annotation (prepared using Ribotaper supplementary script: Create_annotations_files.bash) was used. An offset of 0 and read length 1 was used as the P-site alignment file included pre-computed P-site positions inferred for each length based on the offset quantified. PRICE uses multiple read lengths to model the noise hence, the merged Ribo-seq alignment file consisting of all read lengths and an annotation index (prepared using PRICE, gedi) was used. For RiboTISH, the merged Ribo-seq alignment file, the Ensembl+FANTOM5 GTF file (prepared in this manuscript), human genome fasta file and parameter file generated by RiboTISH quality function were used. Lastly, smORFs found in humans as in sorfs.orf (Olexiouk et al., 2018), a database of detected smORFs based on published Ribo-seq dataset (intergenic smORFs were removed) were also added to the list of de-novo smORFs identified in this study. There

was no minimum or maximum length cutoff that was used for identifying high-confidence smORFs in this study.

Nomenclature for smORFs

In order to understand the overall extent of translation on the transcriptome, we combined smORFs identified by all three tools and the database. A common nomenclature of smORFs was defined to allow combining and cross-comparison of smORFs from different sources with different naming systems. The four smORF result files obtained using the three tools and the database were each processed to a standard GTF format to allow merging of the smORFs detected by different methods. The highest level in the hierarchy of the nomenclature is ORF. An ORF is defined as having a stop-site not shared by any known protein-coding gene. It is labelled using an ORF id, which is a combination of the host gene id and the position of the stop-codon, HostGeneID_StopPos. Each ORF is defined to have several isoform-ORFs or iORFs based on different start codon positions in the coding frame identified for a given stop codon position. This can be due to different splicing structures or due to alternative canonical or non-canonical start-codons present in the same coding frame. The iORF_id was determined from the unique identifier assigned for each smORF identified by the tools or database that detected the isoform. Similar to Ensembl-ORFs, smORFs may span over multiple exons. This information was further stored in the form of multiple orfCDSs, which included positions of smORF covered across different exonic regions. Grouping of similar and duplicate smORFs was carried out by assigning each detected smORF, an ORF_id, iORF_id and orfCDS locations and storing them according to the hierarchical structure of ORF->iORF->orfCDS in a standard GTF file. This is similar to the previously used structure for Ensembl-ORFs: gene->transcript->CDS (see **Fig. S6A**).

Translation signature scores

Three scores were developed to decipher actively translating smORFs from likely false-positive smORFs which are random occurrences of ORFs. All scores were calculated for each smORF identified by the three tools and the database using their defined start and stop site and the orfCDS information was used to skip intronic regions while calculating scores. The high-quality inferred P-sites determined in this manuscript, merged across all datasets, were used as input for the scores. First, *P-sites in frame (PIF)* is defined as the percentage of inferred P-sites in the coding-frame of the iORF across its length. PIF was quantified for each smORF by calculating the sum of inferred P-sites across the iORF's length and dividing by the total number of inferred P-sites in the iORF. Second, *Uniformity* is defined as the uniform coverage of the 3nt-periodicity signal across the complete iORF. In order to quantify uniformity, each codon (3 nucleotides) was tested for the percent inferred P-sites in the coding-frame and was flagged as PASS if this was found to be larger than 33.33% (random), otherwise as FAIL. Finally, the uniformity was calculated by quantifying the percentage of codons that were flagged as PASS across the length of the smORF. Third, *Drop-off* score is defined as the quantification of ribosome disengagement at the stop-codon. For each smORF, a 15 nucleotide (exon only) bin before and after the stop-codon was determined after accounting for splice junctions. Inferred P-sites were quantified in the coding frame. The drop-off score was quantified as the percentage of the number of inferred P-sites in the coding frame before the stop codon to the total inferred P-sites in the coding frame before and after the stop codon. A drop-off of 100%, which means 100% of inferred P-sites in the coding-frame are before the stop-codon and no inferred P-sites after the stop-codon, signifies a sharp disengagement of ribosomes at the stop-codon. (**Fig. 2A**).

The three scores were calculated for Ensembl-ORFs and smORFs. The Ensembl-ORFs were used as the true-set to establish the properties of the scores. Two normal distributions were fit to each PIF and Uniformity scores across Ensembl-ORFs using normalmixEM function from the mixtools package, version 1.2.0. A 95 percentile value (mean - 2*standard

deviations of the normal distribution for Ensembl-ORFs) was calculated and used as a threshold for high-quality PIF and Uniformity scores (See **Fig. S6C-E**). For the Drop-off score, the mean was quantified and used as a threshold for a high-quality score. Each smORF at the iORF level was tested for PIF, Uniformity and Drop-off and discarded if either of the three scores were smaller than the threshold values. smORFs which were completely inside an annotated ORF or had in-frame overlap with an annotated ORF were also discarded.

Isoform prioritization

Only one copy of iORFs was retained in instances of duplicates, i.e., with exactly the same start-codon, stop-codon, orfCDS positions and thus exactly the same nucleotide sequence. All iORFs that failed any of the three translational signature scores were discarded. Each ORF could have multiple iORFs that pass the three translational signature scores, which are alternative start-codons in-frame for a given stop-codon. To determine the most actively used start-site in-frame for the given stop-codon, we used the sequence length and Uniformity scores. The iORFs were sorted based on their sequence length. The uniformity score of the iORF with the smallest length was stored and compared iteratively with the next iORF until the score dropped. The iORF selected before the Uniformity score dropped was prioritized as the iORF for the given ORF, i.e., the longest iORF before the uniform coverage of P-sites is dropped.

Categorization of smORFs

ORFs were categorized into nuORFs (smORFs present on previously annotated non-coding RNA), uORFs (smORFs present upstream of an Ensembl-ORF) or dORFs (smORFs present downstream of an Ensembl-ORF) based on their location with respect to the current transcript annotations as described in **Fig. S7**. Bedtools intersect was used to determine possible overlaps of given smORFs with any known annotations (Ensembl hg38). ORFs that overlapped with five_prime_utr or three_prime_utr were assigned as uORFs and dORFs respectively. Furthermore, as the FANTOM5 catalogue extensions did not include UTR annotation, we found ORFs that did not overlap with five_prime_utr or three_prime_utr but were hosted in gene biotypes “coding_mRNA" or “protein_coding”. These were tested for their relative position w.r.t to the start codon and stop codon of the host gene’s annotated ORF. smORFs upstream of the start codon were assigned as uORFs and smORFs downstream of the stop codon were assigned as dORFs. ORFs present on the 5'UTR but partially overlapping with the CDS and/or start-codon were considered as overlapping uORFs. Similarly, ORFs which were located on the 3'UTR but also overlapped with the CDS and/or stop-codon were categorized as overlapping dORFs. ORFs present on genes that did not have the gene biotype coding_mRNA or protein_coding and had no overlap with any known current annotation such as five_prime_utr, three_prime_utr or CDS were assigned as nuORFs. smORFs overlapping multiple gene annotations spanning multiple genes were assigned first as uORF or dORF or nuORF and multiple host gene IDs were documented in **Table S5**, as alternate Gene IDs.

Sequence-based annotations

Nucleotide sequence for each smORF was obtained using Bedtools getfasta on the human genome fasta file using the spliced orfCDS locations for each smORF. The nucleotide sequence of smORFs was used to obtain the amino-acid sequence using the nucleotide-to-codon translation table. A fasta file for the smORF amino-acid sequences was generated. SignalP (Armenteros et al., 2019) and NLStradamus (Nguyen Ba et al., 2009) were run in default settings for the smORF fasta file to detect presence of signal peptide, nuclear localization signal for each smORF. Nucleotide sequences at a 15bp window around the start of the smORFs were obtained using bedtools getfasta on human genome fasta file

downloaded from Ensembl: Homo_sapiens.GRCh38.dna.primary_assembly.fa. The sequence was plotted using makePWM in R package seqLogo (Bembom O et al,. 2021 R package version 1.58.0) to observe kozak motifs.

Annotated ORF and smORF expression

The generation of smORF nomenclature and GTF allowed read counting on individual smORFs which was carried out using Feature Counts (Liao et al., 2014). The counting was carried out using iORFs and orfCDS as features and meta-features. Read counts were calculated for each iORF for each Ribo-seq and RNA-seq (full-length and Clipped) sample. Read counts for Ensembl-ORFs and novel smORFs were combined to quantify Transcripts per million (TPM) values in each sample. Distributions of Ribo-seq TPM for each cell type were plotted using the density function in R. Expressed smORFs and annotated smORFs in each cell-type/tissue within the dataset were determined using a Ribo-seq TPM of larger than 1. Distribution of TE was plotted using TE values ($TPM_{Ribo\text{-}seq}/TPM_{RNA\text{-}seq}$)(Schafer et al., 2015) for all ORFs with RNA-seq TPM larger than 1.

Cell-type specificity analysis

For cell-type/tissue specificity analysis, we considered fibroblasts, artery endothelial cells (HCAEC and AEC), human umbilical vein endothelial cells (HUVEC), vascular smooth muscle cells (VSMC), Embryonic stem cells, Hepatocytes, Kidney tissue, Fat tissue, Heart tissue, Brain tissue (Inhouse and published). Skeletal muscle tissue data were not used due to low total read depth as it would limit the evidence of expression of a given gene. The number of cell-types/tissues for each gene where it is expressed (full-length RNA-seq TPM>1) was determined. The genes were binned into groups from specific to ubiquitous, ranging from expression in 1 to 10 cell-types/tissues. The genes were segregated based on whether they hosted an Ensembl-ORF and whether they also hosted uORFs/dORFs, or if they hosted nuORFs in previously annotated long non-coding RNAs. Similarly, the number of cell-types/tissues for each smORF and annotated ORF where they were translated (Ribo-seq TPM>1) was determined. The smORFs/Ensembl-ORFs were binned into whether they were found specifically or ubiquitously in 1 to 10 cell-types/tissues.

To obtain cell-type-specific genes, we incorporated the FANTOM5 gene expression atlas (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014) which used the cap analysis of gene expression (CAGE) technique to profile the transcriptome of more than 500 samples across >170 cell types. Specifically, we used the updated FANTOM5 data which has been re-processed for the hg38 genome (Abugessaisa et al., 2017). Low-quality samples, namely samples with zero median expression and biological replicates that do not cluster together, were removed, resulting in a final dataset of 436 samples. The gene expression is then normalised using the variance stabilizing transformation (VST) function in the DESeq2 R package. Hierarchical clustering is then performed on the VST-normalised expression, resulting in 48 different cell types. For each gene, its expression value based on full-length RNA-seq counts was normalized across the 436 samples ($TPM_{gene1,\ sample\ 1}/Sum_{436\_samples}(TPM_{gene1})$) such that the values ranged between 0 and 1. Furthermore, the 162 samples in this paper were assigned to each of the 48 clusters by matching the source cell-type/tissue with the cell types in each cluster. Similarly, for each gene, its expression value based on full-length RNA-seq counts was normalized across the 162 samples ($TPM_{gene1,\ sample\ 1}/Sum_{162\_samples}(TPM_{gene1})$) such that the values ranged between 0 and 1. After normalization, both the FANTOM5 gene expression and RNA-seq gene expression generated in this paper were combined to finally obtain a matrix of 598 samples over 85,162 genes. For each of the 48 clusters, an ideal vector (Length=598) was created with gene expression as 1 for samples within the cluster and 0 for samples outside of the cluster, albeit an ideal scenario of specific gene expression in the given cluster. For a given gene and cluster, the expression patterns across the 598 samples were compared to its ideal case

using Jenssen-Shannon divergence (JSD, quantified by R package https://github.com/tillbe/jsd). The smaller the divergence from the ideal vector, the more the specificity of the gene's expression in the given cluster. Gviz (Hahne and Ivanek, 2016) was used to visualize translation of a Endothelial cell-specific nuORF in the lncRNA (CATG00000072615) across the 10 different cell-type/tissues in the Ribo-seq data generated and/or combined in this paper (See **Fig. 3E**).

Differential translation analysis

A 29bp length clipped RNA-seq dataset was prepared for differential translation analysis as described previously (Chothani et al., 2019b). Trimmed RNA-seq reads were clipped using FASTX Toolkit *V0.0.14* to 29nt to allow a fair comparison with Ribo-seq reads in translation efficiency (TE) analysis to minimize technical differences in data analysis. Read counting in the CDS region was carried out for each gene using Feature Counts (Liao et al., 2014). Ribo-seq read counts, RNA-seq read counts for the Ensembl-ORFs and smORFs identified in this manuscript were used. Differentially-TE genes (DTEGs) for each cell type or tissue were identified individually. For instance, for a given cell-type, fibroblasts, a DTEG was defined as an Ensembl-ORF or a smORF that had significantly changing TE in fibroblasts as compared to the background. The background used was all samples from all other cell types except fibroblasts. The background samples were down-sized by using only two samples per cell-type which were selected as the two with the highest library size. Cell-type and tissues were analysed separately. The change in translation efficiency and associated significance p-value was quantified using deltaTE (Chothani et al., 2019b). Ribo-seq and RNA-seq data for TGFB1 stimulated fibroblast time-series experiment was processed as described previously (Chothani et al., 2019a) for differential-TE analysis. Read counting for known ORFs and smORFs was carried out using Feature Counts (Liao et al., 2014). Differentially-TE genes (DTEGs) for each time-point (45mins, 2hrs, 6hrs, 24hrs) with respect to baseline were identified along with quantification of translation efficiency and significance p-value using deltaTE (Chothani et al., 2019b). Volcano plots were generated using the EnhancedVolcano function in R (kevinblighe).

Evolutionary conservation analysis

Multiple sequence alignment files (maf) across 100 vertebrates were downloaded from UCSC (Rosenbloom et al., 2015). Alignments to 100 vertebrate species for the smORF locations on the human genome (hg38) were obtained using the Bio.AlignIO.MafIO module in BioPython. The nucleotide sequence was ungapped (ungap(sequence)) and translated (translate(sequence)) to amino-acid using the Bio.seq module in BioPython and stored as a multi-specie fasta file for each smORF. The percentage of amino-acid conservation with respect to the hg38 sequence was quantified for sequences in other species with the same length as the smORF identified in this paper. A background set of smORFs for each subtype, namely, uORF, dORF and nuORF were selected from the list of smORFs with low scores in PIF, Uniformity and Drop-off. AA-sequence fasta and percentage AA conservation were quantified for the background locations using the same method as for smORFs. The multiple sequence alignment for a given smORF was visualized using the msaR package (2021) in R using the multi-species amino-acid sequence fasta file. The evolutionary tree for 100 vertebrates was downloaded in the newick format from UCSC (Kent et al., 2002), (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz100way/) and plotted using the ape package in R (Paradis and Schliep, 2019). Rat Ribo-seq data was downloaded (Schafer et al., 2015) and mapped to Ensembl gene annotations (rn6, Release 86). This was used to infer P-sites in the same way as described previously for human Ribo-seq data. Read lengths 28-30 were used and offset was determined using RiboTISH. The generated P-site alignment file was used to quantify P-sites in Frame, Uniformity and Drop-off scores for the uORF in RASGRP3 (**Fig. 5B**) as defined earlier in this study for human data. PhyloCSF and RNA code were run using default settings using 100 vertebrate phylogeny for all the 7,767

smORF sequences.

Mass-spectrometry validation

The NHDF and HUVEC processed mgf files from Slany et al. (Slany et al., 2016) were downloaded from PRIDE (PXD003406 to PXD003417). The ES processed mgf files from (Shekari et al., 2017) were downloaded from PRIDE (PXD006271). Heart proteome RAW data files from Doll et al. (Doll et al., 2017) were downloaded from PRIDE (PXD006675). For the heart proteome RAW data files, the top 100 peaks by intensity from MS level 2 were centroided and exported to .mgf files with Protowizards msconvert. The database search and peptide-spectrum matching was then performed with a two-step approach. First, the peaks were matched to the human Uniprot database (UniProt Reference 2017_4, 21,007 protein entries; http://www.uniprot.org/) using MS-GF+ (Kim and Pevzner, 2014). Carboxyamidomethylation of cysteine was chosen as a fixed modification, and oxidation of methionine and proline were chosen as variable modifications. The parent mass tolerance was set at 10 ppm and an isotope error range was set to -1,2. The number of tolerable (tryptic) termini was set to 2 and the minimum and maximum peptide lengths were set to 5 and 50, respectively. Using a target-decoy approach, the spectra with peptide-spectrum matches of an FDR < 1% were removed and the remaining spectra were used in the second step.

In the second step, the remaining peaks were matched using MS-GF+ to a custom smORF database developed in this study. Similarly, carboxyamidomethylation of cysteine was chosen as fixed modification, and oxidation of methionine and proline were chosen as variable modifications. The parent mass tolerance was set at 10 ppm, the isotope error range was set to -1,2 and the minimum and maximum peptide lengths were set to 5 and 50, respectively. For the custom smORF database, the number of tolerable (tryptic) termini was set to 1. Using a target-decoy approach and a relaxed threshold, the smORFs with at least one peptide that had a spectrum with peptide-spectrum matches of an FDR < 1% were deemed to be identified

Shiny web application

The genomic coordinates and scores for smORFs were shared as part of a web application created using the R Shiny package. A ui.R and server.R file was created to display an R data-table to browse iORF genomic coordinates, length, peptide sequence, PIF, uniformity and drop-off scores for filtered smORFs. The header of the datatable was incorporated with searching and sorting functions. Gene-level and smORF expression information across cell-types and tissues in this study and testing for presence of smORFs in a given gene ID list are provided.

Quantification and Statistical analysis

The software tools used in this study have been listed in Table S1 and described in the respective methods subsections. Details of samples in each cell type or tissue can be found in Table S3 and S4. The boxplot error bars and centres are defined in figures and figure legends. Statistical p-values associated with the significance tests are described in the figure legends.

Supplementary Files

Data S1: Related to Figure 1. RiboTISH quality images for Ribo-seq samples.
Data S2: Related to Figure 3. Jenssen-Shannon divergence and mean expression showing cell-type specific genes and smORFs.
Table S4: Related to Figure 1. Ribo-seq sample information and processing statistics.
Table S5: Related to Figure 2. Annotations and metadata for smORFs identified in this study.

Table S6: Related to Figure 3. Transcripts per million based on RNA- and Ribo- seq for annotated ORFs and smORFs identified in this study.
Table S7: Related to Figure 3. Cell-type specificity of smORFs quantified by Jenssen-Shannon divergence score for each gene in the 48 clusters of cell-types.
Table S8: Related to Figure 4. Differential translation of Upstream ORF and main ORF pairs identified with a significant change in TE in fibroblasts w.r.t a background.
Table S9: Related to Figure 5. Percentage of amino-acids (AA) identical in 99 vertebrates w.r.t human sequence.
Table S10: Related to Figure 5. SmORF-encoded peptides found in Mass-spectrometry data across various cell-types/tissues and subcellular localizations.

References

Abugessaisa, I., Noguchi, S., Hasegawa, A., Harshbarger, J., Kondo, A., Lizio, M., Severin, J., Carninci, P., Kawaji, H., and Kasukawa, T. (2017). FANTOM5 CAGE profiles of human and mouse reprocessed for GRCh38 and GRCm38 genome assemblies. Sci Data *4*, 170107. .

Armenteros, J.J.A., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., and Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. Nature Biotechnology *37*, 420–423. https://doi.org/10.1038/s41587-019-0036-z.

Bartholomäus, A., Kolte, B., Mustafayeva, A., Goebel, I., Fuchs, S., Benndorf, D., Engelmann, S., and Ignatova, Z. (2021). smORFer: a modular algorithm to detect small ORFs in prokaryotes. Nucleic Acids Res. *49*, e89–e89. .

Bi, P., Ramirez-Martinez, A., Li, H., Cannavino, J., McAnally, J.R., Shelton, J.M., Sánchez-Ortiz, E., Bassel-Duby, R., and Olson, E.N. (2017). Control of muscle formation by the fusogenic micropeptide myomixer. Science *356*, 323–327. .

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. .

Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B., and Ohler, U. (2016). Detecting actively translated open reading frames in ribosome profiling data. Nat. Methods *13*, 165–170. .

Calviello, L., Hirsekorn, A., and Ohler, U. (2020). Quantification of translation uncovers the functions of the alternative transcriptome. Nat. Struct. Mol. Biol. *27*, 717–725. .

Calvo, S.E., Pagliarini, D.J., and Mootha, V.K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc. Natl. Acad. Sci. U. S. A. *106*, 7507–7512. .

Chothani, S., Schäfer, S., Adami, E., Viswanathan, S., Widjaja, A.A., Langley, S.R., Tan, J., Wang, M., Quaife, N.M., Jian Pua, C., et al. (2019a). Widespread Translational Control of Fibrosis in the Human Heart by RNA-Binding Proteins. Circulation *140*, 937–951. .

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

Chugunova, A., Loseva, E., Mazin, P., Mitina, A., Navalayeu, T., Bilan, D., Vishnyakova, P., Marey, M., Golovina, A., Serebryakova, M., et al. (2019). LINC00116 codes for a

mitochondrial peptide linking respiration and lipid metabolism. Proceedings of the National Academy of Sciences *116*, 4940–4945. https://doi.org/10.1073/pnas.1809105116.

Couso, J.-P., and Patraquim, P. (2017). Classification and function of small open reading frames. Nat. Rev. Mol. Cell Biol. *18*, 575–589. .

D'Lima, N.G., Ma, J., Winkler, L., Chu, Q., Loh, K.H., Corpuz, E.O., Budnik, B.A., Lykke-Andersen, J., Saghatelian, A., and Slavoff, S.A. (2017). A human microprotein that interacts with the mRNA decapping complex. Nat. Chem. Biol. *13*, 174–180. .

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2012). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. .

Doll, S., Dreßen, M., Geyer, P.E., Itzhak, D.N., Braun, C., Doppler, S.A., Meier, F., Deutsch, M.-A., Lahm, H., Lange, R., et al. (2017). Region and cell-type resolved quantitative proteomic map of the human heart. Nat. Commun. *8*, 1469. .

Dunn, J.G., Foo, C.K., Belletier, N.G., Gavis, E.R., and Weissman, J.S. (2013). Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. Elife *2*, e01179. .

Erhard, F., Halenius, A., Zimmermann, C., L'Hernault, A., Kowalewski, D.J., Weekes, M.P., Stevanovic, S., Zimmer, R., and Dölken, L. (2018). Improved Ribo-seq enables identification of cryptic translation events. Nat. Methods *15*, 363–366. .

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest, A.R.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J.L., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., et al. (2014). A promoter-level mammalian expression atlas. Nature *507*, 462–470. .

Friesen, M., Warren, C.R., Yu, H., Toyohara, T., Ding, Q., Florido, M.H.C., Sayre, C., Pope, B.D., Goff, L.A., Rinn, J.L., et al. (2020). Mitoregulin Controls β-Oxidation in Human and Mouse Adipocytes. Stem Cell Reports *14*, 590–602. .

Gao, X., Wan, J., Liu, B., Ma, M., Shen, B., and Qian, S.-B. (2015). Quantitative profiling of initiating ribosomes in vivo. Nat. Methods *12*, 147–153. .

Gillet, J.-P., Varma, S., and Gottesman, M.M. (2013). The Clinical Relevance of Cancer Cell Lines. JNCI Journal of the National Cancer Institute *105*, 452–458. https://doi.org/10.1093/jnci/djt007.

Gonzalez, C., Sims, J.S., Hornstein, N., Mela, A., Garcia, F., Lei, L., Gass, D.A., Amendolara, B., Bruce, J.N., Canoll, P., et al. (2014). Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. J. Neurosci. *34*, 10924–10936. .

Greenberg, A.S., Egan, J.J., Wek, S.A., Garty, N.B., Blanchette-Mackie, E.J., and Londos, C. (1991). Perilipin, a major hormonally regulated adipocyte-specific phosphoprotein associated with the periphery of lipid storage droplets. Journal of Biological Chemistry *266*, 11341–11346. https://doi.org/10.1016/s0021-9258(18)99168-4.

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, et al. (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213. .

Hahne, F., and Ivanek, R. (2016). Visualizing Genomic Data Using Gviz and Bioconductor. Methods Mol. Biol. *1418*, 335–351. .

Hao, Y., Zhang, L., Niu, Y., Cai, T., Luo, J., He, S., Zhang, B., Zhang, D., Qin, Y., Yang, F., et al. (2017). SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. Brief. Bioinform. https://doi.org/10.1093/bib/bbx005.

van Heesch, S., Witte, F., Schneider-Lunitz, V., Schulz, J.F., Adami, E., Faber, A.B., Kirchner, M., Maatz, H., Blachut, S., Sandmann, C.-L., et al. (2019). The Translational Landscape of the Human Heart. Cell *178*, 242–260.e29. .

Hentze, M.W., and Kulozik, A.E. (1999). A perfect message: RNA surveillance and nonsense-mediated decay. Cell *96*, 307–310. .

Hilleren, P., and Parker, R. (1999). mRNA surveillance in eukaryotes: kinetic proofreading of proper translation termination as assessed by mRNP domain organization? RNA *5*, 711–719. .

Ho, L., van Dijk, M., Chye, S.T.J., Messerschmidt, D.M., Chng, S.C., Ong, S., Yi, L.K., Boussata, S., Goh, G.H.-Y., Afink, G.B., et al. (2017). ELABELA deficiency promotes preeclampsia and cardiovascular malformations in mice. Science *357*, 707–713. .

Hon, C.-C., Ramilowski, J.A., Harshbarger, J., Bertin, N., Rackham, O.J.L., Gough, J., Denisenko, E., Schmeier, S., Poulsen, T.M., Severin, J., et al. (2017). An atlas of human long non-coding RNAs with accurate 5' ends. Nature *543*, 199–204. .

Hsu, P.Y., Calviello, L., Wu, H.-Y.L., Li, F.-W., Rothfels, C.J., Ohler, U., and Benfey, P.N. (2016). Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. Proc. Natl. Acad. Sci. U. S. A. https://doi.org/10.1073/pnas.1614788113.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223. .

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell *147*, 789–802. .

Ingolia, N.T., Brar, G.A., Stern-Ginossar, N., Harris, M.S., Talhouarne, G.J.S., Jackson, S.E., Wills, M.R., and Weissman, J.S. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep. *8*, 1365–1379. .

Ji, Z., Song, R., Regev, A., and Struhl, K. (2015). Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. https://doi.org/10.7554/eLife.08890.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006. .

kevinblighe GitHub - kevinblighe/EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling.

Kim, S., and Pevzner, P.A. (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun. *5*, 5277. .

Koh, M., Ahmad, I., Ko, Y., Zhang, Y., Martinez, T.F., Diedrich, J.K., Chu, Q., Moresco, J.J., Erb, M.A., Saghatelian, A., et al. (2021). A short ORF-encoded transcriptional regulator.

Proceedings of the National Academy of Sciences *118*, e2021943118. https://doi.org/10.1073/pnas.2021943118.

Kozak, M. (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell *44*, 283–292. .

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25. .

Lee, T.I., and Young, R.A. (2013). Transcriptional Regulation and Its Misregulation in Disease. Cell *152*, 1237–1251. https://doi.org/10.1016/j.cell.2013.02.014.

Lee, C.Q.E., Kerouanton, B., Chothani, S., Zhang, S., Chen, Y., Mantri, C.K., Hock, D.H., Lim, R., Nadkarni, R., Huynh, V.T., et al. (2021). Coding and non-coding roles of MOCCI (C15ORF48) coordinate to regulate host inflammation and immunity. Nat. Commun. *12*, 2130. .

Lee, S., Liu, B., Lee, S., Huang, S.-X., Shen, B., and Qian, S.-B. (2012). Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc. Natl. Acad. Sci. U. S. A. *109*, E2424–E2432. .

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

Lim, W.-W., Corden, B., Ng, B., Vanezis, K., D'Agostino, G., Widjaja, A.A., Song, W.-H., Xie, C., Su, L., Kwek, X.-Y., et al. (2020). Interleukin-11 is important for vascular smooth muscle phenotypic switching and aortic inflammation, fibrosis and remodeling in mouse models. Sci. Rep. *10*, 17853. .

Lin, M.F., Jungreis, I., and Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics *27*, i275–i282. .

Liu, Y., Mi, Y., Mueller, T., Kreibich, S., Williams, E.G., Van Drogen, A., Borel, C., Frank, M., Germain, P.-L., Bludau, I., et al. (2019). Multi-omic measurements of heterogeneity in HeLa cells across laboratories. Nat. Biotechnol. *37*, 314–322. .

Loayza-Puch, F., Rooijers, K., Buil, L.C.M., Zijlstra, J., Oude Vrielink, J.F., Lopes, R., Ugalde, A.P., van Breugel, P., Hofland, I., Wesseling, J., et al. (2016). Tumour-specific proline vulnerability uncovered by differential ribosome codon reading. Nature *530*, 490–494. .

Ma, J., Diedrich, J.K., Jungreis, I., Donaldson, C., Vaughan, J., Kellis, M., Yates, J.R., 3rd, and Saghatelian, A. (2016). Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. Anal. Chem. *88*, 3967–3975. .

Magger, O., Waldman, Y.Y., Ruppin, E., and Sharan, R. (2012). Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. PLoS Comput. Biol. *8*, e1002690. .

Makarewich, C.A., and Olson, E.N. (2017). Mining for Micropeptides. Trends Cell Biol. *27*, 685–696. .

Makarewich, C.A., Baskin, K.K., Munir, A.Z., Bezprozvannaya, S., Sharma, G., Khemtong, C., Shah, A.M., McAnally, J.R., Malloy, C.R., Szweda, L.I., et al. (2018). MOXI Is a Mitochondrial Micropeptide That Enhances Fatty Acid β-Oxidation. Cell Reports *23*, 3701–3709. https://doi.org/10.1016/j.celrep.2018.05.058.

Martinez, T.F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M.N., and Saghatelian, A. (2020). Accurate annotation of human protein-coding small open reading frames. Nat. Chem. Biol. *16*, 458–468. .

Matsumoto, A., Pasut, A., Matsumoto, M., Yamashita, R., Fung, J., Monteleone, E., Saghatelian, A., Nakayama, K.I., Clohessy, J.G., and Pandolfi, P.P. (2017). mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. Nature *541*, 228–232. .

Mudge, J.M., Ruiz-Orera, J., Prensner, J.R., Brunet, M.A., Gonzalez, J.M., Magrane, M., Martinez, T., Schulz, J.F., Yang, Y.T., Mar Albà, M., et al. A community-driven roadmap to advance research on translated open reading frames detected by Ribo-seq. https://doi.org/10.1101/2021.06.10.447896.

Muhlrad, D., and Parker, R. (1999). Aberrant mRNAs with extended 3' UTRs are substrates for rapid degradation by mRNA surveillance. RNA *5*, 1299–1307. .

Nguyen Ba, A.N., Pogoutse, A., Provart, N., and Moses, A.M. (2009). NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. BMC Bioinformatics *10*, 202. .

Olexiouk, V., Van Criekinge, W., and Menschaert, G. (2018). An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. *46*, D497–D502. .

Pang, Y., Liu, Z., Han, H., Wang, B., Li, W., Mao, C., and Liu, S. (2020). Peptide SMIM30 promotes HCC development by inducing SRC/YES1 membrane anchoring and MAPK pathway activation. J. Hepatol. *73*, 1155–1169. .

Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics *35*, 526–528. https://doi.org/10.1093/bioinformatics/bty633.

Pueyo, J.I., Magny, E.G., and Couso, J.P. (2016). New Peptides Under the s(ORF)ace of the Genome. Trends Biochem. Sci. *41*, 665–678. .

Quinn, M.E., Goh, Q., Kurosaka, M., Gamage, D.G., Petrany, M.J., Prasad, V., and Millay, D.P. (2017). Myomerger induces fusion of non-fusogenic cells and is required for skeletal muscle development. Nat. Commun. *8*, 15665. .

Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., et al. (2015). The UCSC Genome Browser database: 2015 update. Nucleic Acids Res. *43*, D670–D681. .

Ruiz-Orera, J., Messeguer, X., Subirana, J.A., and Alba, M.M. (2014). Long non-coding RNAs as a source of new peptides. Elife *3*, e03523. .

Schafer, S., Adami, E., Heinig, M., Rodrigues, K.E.C., Kreuchwig, F., Silhavy, J., van Heesch, S., Simaite, D., Rajewsky, N., Cuppen, E., et al. (2015). Translational regulation shapes the molecular landscape of complex disease phenotypes. Nat. Commun. *6*, 7200. .

Shekari, F., Nezari, H., Larijani, M.R., Han, C.-L., Baharvand, H., Chen, Y.-J., and Salekdeh, G.H. (2017). Proteome analysis of human embryonic stem cells organelles. J. Proteomics *162*, 108–118. .

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson,

H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050. .

Slany, A., Bileck, A., Kreutz, D., Mayer, R.L., Muqaku, B., and Gerner, C. (2016). Contribution of Human Fibroblasts and Endothelial Cells to the Hallmarks of Inflammation as Determined by Proteome Profiling. Mol. Cell. Proteomics *15*, 1982–1997. .

Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z., Karger, A.D., Budnik, B.A., Rinn, J.L., and Saghatelian, A. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat. Chem. Biol. *9*, 59–64. .

Slavoff, S.A., Heo, J., Budnik, B.A., Hanakahi, L.A., and Saghatelian, A. (2014). A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. J. Biol. Chem. *289*, 10950–10957. .

Stein, C.S., Jadiya, P., Zhang, X., McLendon, J.M., Abouassaly, G.M., Witmer, N.H., Anderson, E.J., Elrod, J.W., and Boudreau, R.L. (2018). Mitoregulin: A lncRNA-Encoded Microprotein that Supports Mitochondrial Supercomplexes and Respiratory Efficiency. Cell Rep. *23*, 3710–3720.e8. .

The RNAcentral Consortium (2017). RNAcentral: a comprehensive database of non-coding RNA sequences. Nucleic Acids Res. *45*, D128–D134. .

Tjeldnes, H., Labun, K., Torres Cleuren, Y., Chyżyńska, K., Świrski, M., and Valen, E. (2021). ORFik: a comprehensive R toolkit for the analysis of translation. BMC Bioinformatics *22*, 1–16. .

Wan, J., and Qian, S.-B. (2014). TISdb: a database for alternative translation initiation in mammalian cells. Nucleic Acids Res. *42*, D845–D850. .

Washietl, S., Findeiss, S., Müller, S.A., Kalkhof, S., von Bergen, M., Hofacker, I.L., Stadler, P.F., and Goldman, N. (2011). RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA *17*, 578–594. .

Wein, N., Vulin, A., Falzarano, M.S., Szigyarto, C.A.-K., Maiti, B., Findlay, A., Heller, K.N., Uhlén, M., Bakthavachalu, B., Messina, S., et al. (2014). Translation from a DMD exon 5 IRES results in a functional dystrophin isoform that attenuates dystrophinopathy in humans and mice. Nature Medicine *20*, 992–1000. https://doi.org/10.1038/nm.3628.

Whiffin, N., Genome Aggregation Database Production Team, Karczewski, K.J., Zhang, X., Chothani, S., Smith, M.J., Gareth Evans, D., Roberts, A.M., Quaife, N.M., Schafer, S., et al. (2020). Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. Nature Communications *11*. https://doi.org/10.1038/s41467-019-10717-9.

Xie, S.-Q., Nie, P., Wang, Y., Wang, H., Li, H., Yang, Z., Liu, Y., Ren, J., and Xie, Z. (2016). RPFdb: a database for genome wide information of translated mRNA generated from ribosome profiling. Nucleic Acids Res. *44*, D254–D258. .

Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., et al. (2020). Ensembl 2020. Nucleic Acids Res. *48*, D682–D688. .

Zhang, P., He, D., Xu, Y., Hou, J., Pan, B.-F., Wang, Y., Liu, T., Davis, C.M., Ehli, E.A., Tan, L., et al. (2017a). Genome-wide identification and differential analysis of translational initiation. Nat. Commun. *8*, 1749. .

Zhang, Q., Vashisht, A.A., O'Rourke, J., Corbel, S.Y., Moran, R., Romero, A., Miraglia, L., Zhang, J., Durrant, E., Schmedt, C., et al. (2017b). The microprotein Minion controls cell fusion and muscle formation. Nat. Commun. *8*, 15664. .

(2021). Multiple Sequence Alignment for R Shiny [R package msaR version 0.5.0].

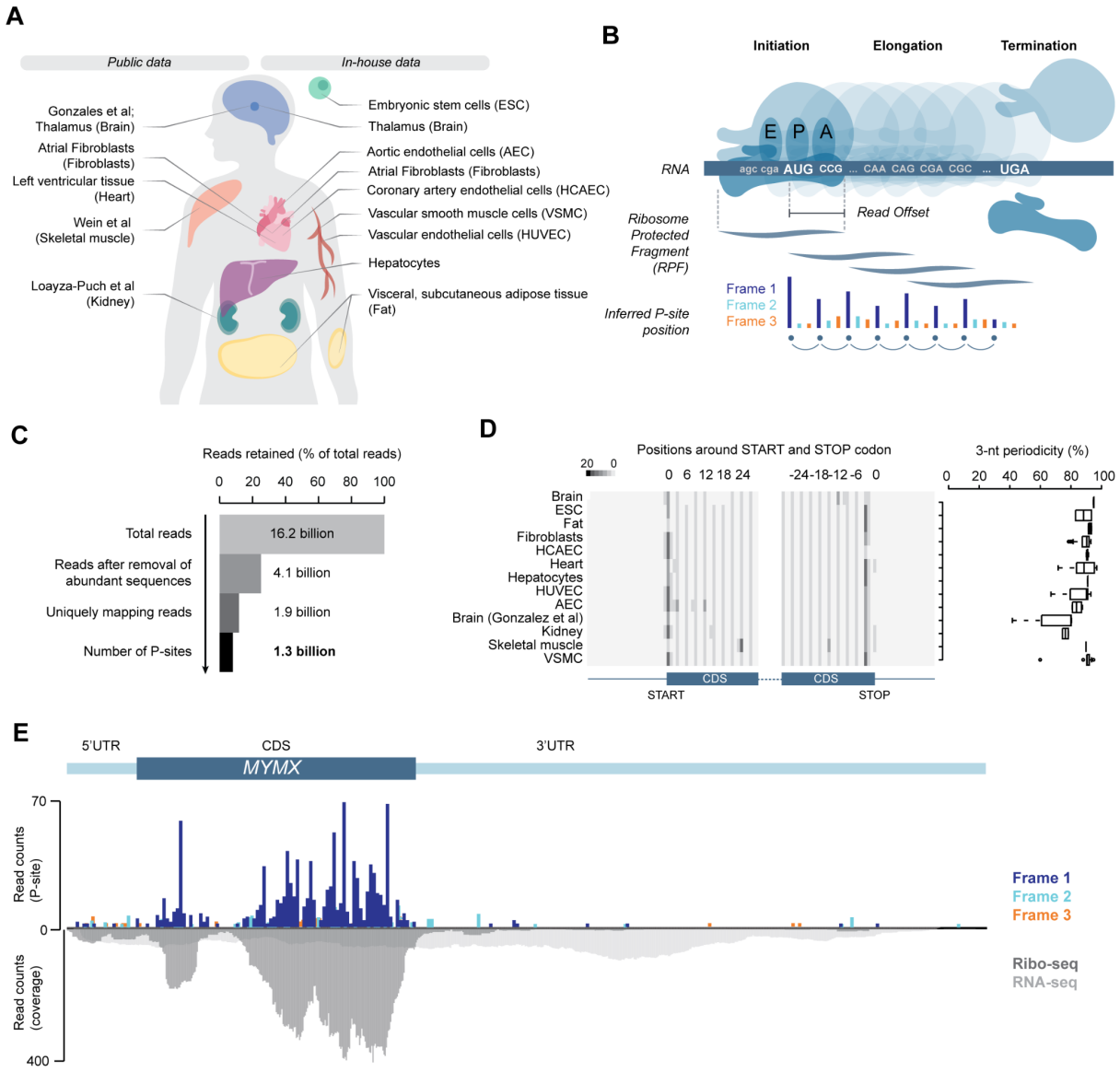


Figure 1: A high-depth and high-resolution dataset of mRNA translation in primary human cell types and tissues. **A.** Schematic of the data (Ribo-seq and RNA-seq) used in this study. In-house: newly generated datasets. Public: published datasets (Brain (Gonzalez et al., 2014), Atrial fibroblasts (Chothani et al., 2019a), Heart (van Heesch et al., 2019), Skeletal muscle (Wein et al., 2014) and Kidney (Loayza-Puch et al., 2016)). **B.** Schematic illustrating the concepts of RPF (ribosome protected fragment) and relative inferred P-site position, used to determine the frame being read. Abbreviations: A-site (aminoacyl), P-site (peptidyl) and E-site (exit); AUG, start codon, UGA, stop codon. **C.** Bargraph indicating the total amount of sequenced raw reads and the retained reads (%) following each of the indicated pre-processing steps. **D.** Trinucleotide (3-nt) periodicity of the individual datasets, calculated using read lengths of 28-30nt. Data shown as positional heatmap and box and whiskers plot (line at median, the whiskers extend to the most extreme data point which is no more than 1.5*IQR (or interquartile range) from the box) **E.** *MYMX* transcript expression (RNA-seq: light grey) and translation (Ribo-seq: dark grey) are displayed together with inferred ribosomal P-site positions dominating the coding-frame (Frame 1, blue). Abbreviations: 5'UTR (5' untranslated region), CDS (Coding sequence), 3'UTR (3' untranslated region).

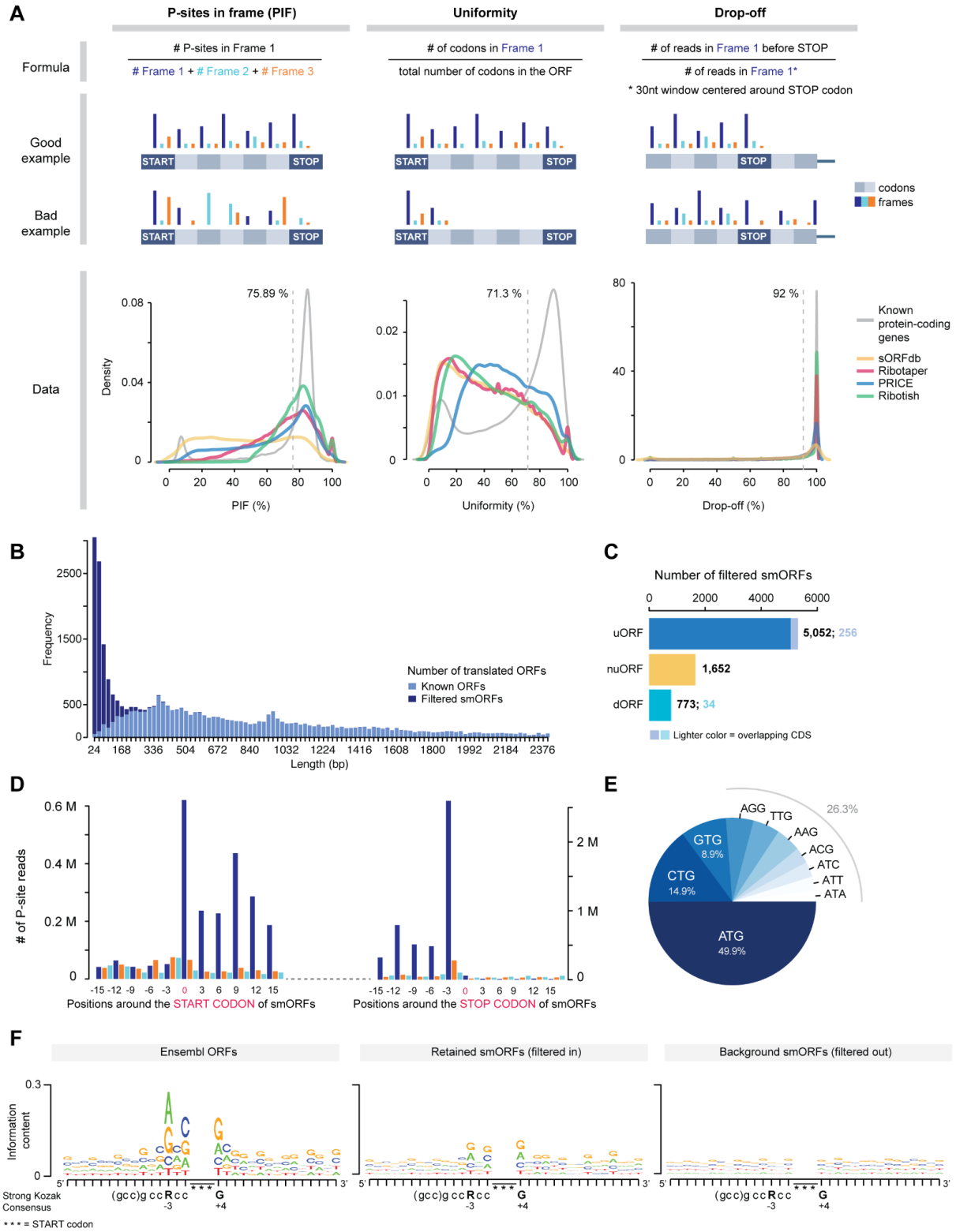


Figure 2: Comprehensive and systematic discovery of actively translated ORFs using the human translation dataset identifies smORFs with robust translation signatures similar to Ensembl-ORFs. A. Graphic illustrating the filters adopted to select smORFs with a robust mRNA translation signature: *P-sites In Frame* (PIF), *Uniformity*, *Drop-off* score. For each score, density plots based on the underlying data processed with different smORF-calling tools are shown to illustrate the rationale for their choice. The dotted line represents the threshold determined using Mean -2*SD (95 percentile) of fitted normal distributions for PIF and Uniformity and the Mean for Drop-off score. **B** Length distribution of Ensembl-ORFs (light blue) and filtered smORFs (dark blue). **C.** Bar graph showing smORF classification based on their mapped location. Abbreviations: uORFs (upstream ORFs; located on the 5'UTR of known protein-coding ORFs). dORFs (downstream ORFs; located on the 3'UTR of known protein-coding ORFs). NuORFs (novel unannotated ORFs; located on previously annotated

non-coding transcripts). Overlapping CDS indicates cases in which uORFs/dORFs overlapped with the protein-coding ORF on the same transcript. **D.** P-site distribution around the Start and Stop codons of the final combined set of smORFs. **E.** Relative usage of codon start sites among smORFs. **F.** Translation initiation context for Ensembl-ORFs, filtered smORFs and background smORFs. Background smORFs: Randomly selected smORFs from low scoring smORFs (PIF < 40%, Uniformity < 40% and Drop-off < 90%) to match the set-size of filtered smORFs. Kozak consensus is given for reference at the bottom, where R stands for Adenine (A) or Guanine (G), *** denotes the start codon and -3 and +4 positions denote strong Kozak context.

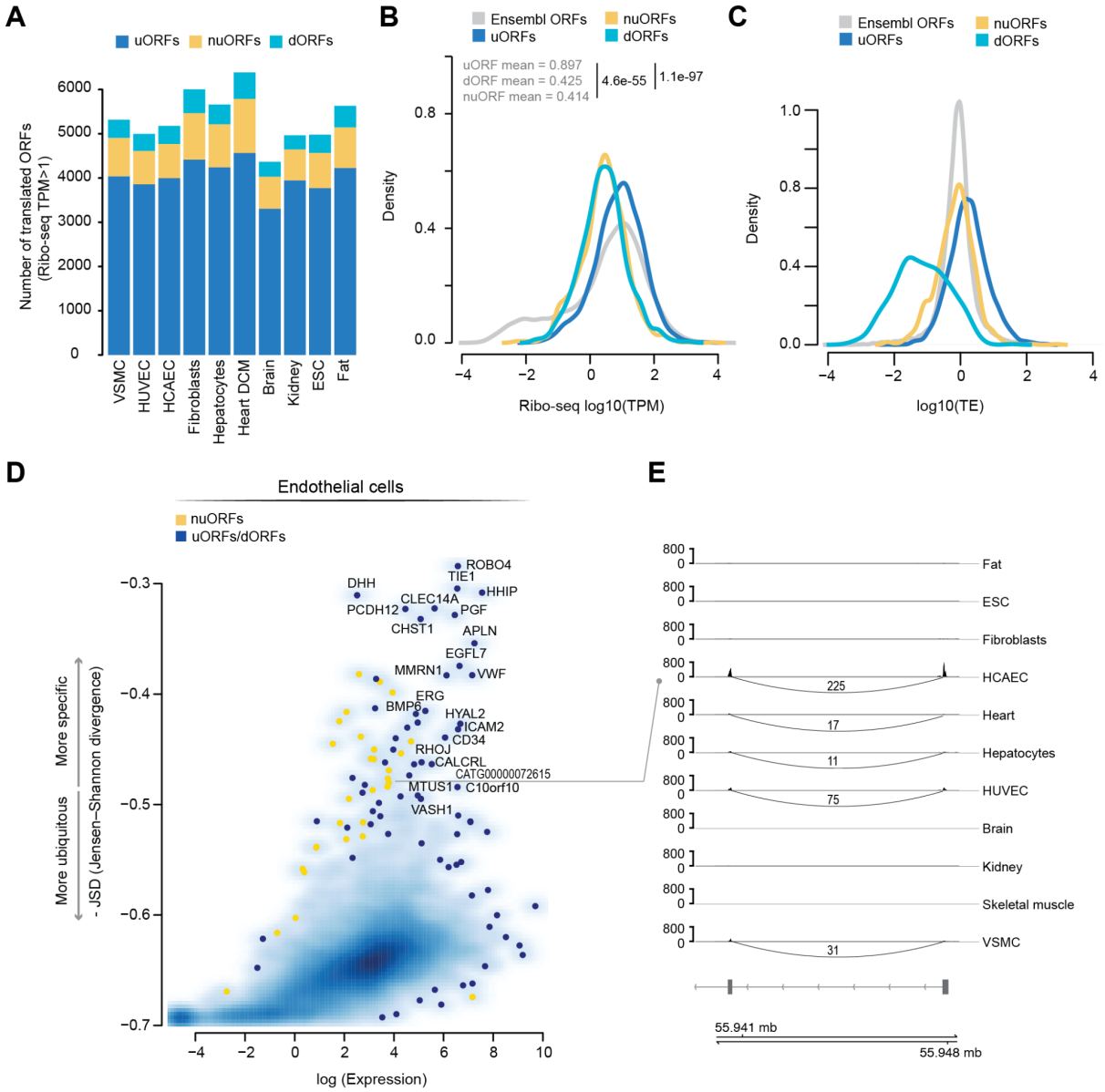


Figure 3: RNA expression and translation of small open reading frames (smORFs) in human cell types and tissues. A. Stacked bar chart showing the number and type of translated smORFs (Ribo-seq TPM > 1) in each dataset. Of the 13 datasets, 10 total cell types and tissues were assessed for expression analysis. AEC and HCAEC merged together as HCAEC; Published and newly generated Brain data merged together; Skeletal muscle tissue removed failing read-depth QC for expression analysis; **B.** Ribo-seq TPM (transcript per million) distribution for each smORF category and known coding ORFs in fibroblasts. uORFs display higher mean translation levels (0.897) than dORFs (0.425; p-value $1.1e^{-97}$) and nuORFs (0.414; p-value $4.6e^{-55}$). Statistics: Student's t-test. **C.** TE (Translation efficiency) distribution for each smORF category and known coding ORFs in fibroblasts. **D.** Scatterplot of Jensen-Shannon divergence and Endothelial cell gene expression. Blue: Genes with uORFs and/or dORFs, Yellow: Genes with nuORFs. Genes in the top-right quadrant (e.g. CATG00000072615) represent genes that are both highly expressed and specifically expressed in endothelial cells. Gene symbols are displayed for genes (top 20) with the highest expression levels and JSD < 0.5. **E.** An example of endothelial cell-specific nuORF, CATG00000072615 encoded in a lncRNA showing Endothelial-specific Ribo-seq read coverage within the cell-types/tissues in this study.

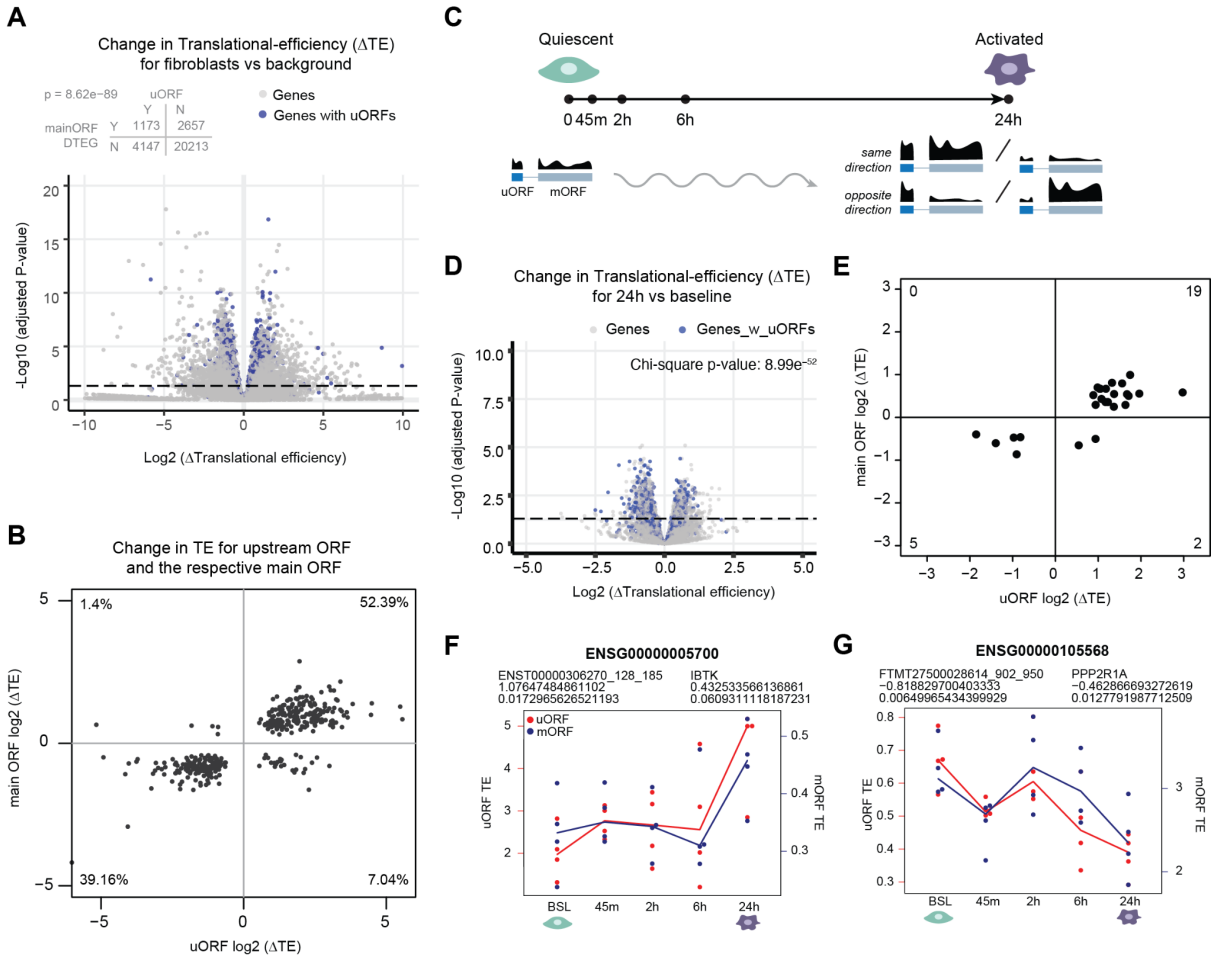


Figure 4: Upstream Open Reading frames in translational regulation. A. Volcano plot showing translational efficiency (TE) for all Ensembl-ORFs in fibroblasts vs background. Log2(ΔTE) denotes the log fold change of TE for the annotated ORF. Log10(adjusted p-value) denotes the significance of the change in TE. Ensembl-ORFs are marked in blue if they also host an upstream ORF (uORF). A chi-square p-value of $8.62e^{-89}$ signified the preferential presence of uORFs hosted in differential-TE genes. **B.** A scatter plot showing TE fold-changes for differential-TE main ORF and uORF pairs in fibroblasts vs background. 91.55% of uORF-main ORF pairs change in the same direction and 8.45% in the opposite direction. uORF-mORF pairs with absolute log2 ΔTE greater than 5 were shown on the axis boundary. **C.** Ribo-seq and RNA-seq of TGFB1 stimulated atrial fibroblasts across a time-series (Baseline, 45mins, 2hrs, 6hrs, 24hrs) **D.** Volcano plot showing translational efficiency (TE) for all Ensembl-ORFs in TGFB1 stimulated fibroblasts (24 hours) vs baseline. Log2(ΔTE) denotes the log fold change of TE for the annotated ORF. Log10(adjusted p-value) denotes the significance of the change in TE.Ensembl-ORFs are marked in blue if they also host an upstream ORF (uORF). A chi-square p-value of $8.99e^{-52}$ signified the preferential presence of uORFs hosted in differential-TE genes. **E.** A scatter plot showing TE fold-changes for differential-TE main ORF and uORF pairs in fibroblasts vs background. 92.31% of significantly changing uORF-mORF pairs change in the same direction. **F-G.** Scatter plot for two exemplar uORF-mORF pairs showing in-tandem change in TE during fibroblast activation. Lines represent median TE values. Red: uORF, Blue: mORF. Header represents Ensembl gene ID hosting both uORF and mORF. Log2 fold change of TE and adjusted p-values are printed for the uORF and mORF respectively.

Figure 5: Evolutionary conservation of small open reading frames across 100 vertebrates. A. Difference in the average percentage of Amino-Acid conservation for smORFs across 100 vertebrate species with respect to a matched background. Dark blue: upstream ORF; Light blue: downstream ORF; Yellow: novel unannotated ORF. Lighter colour: Background smORFs not detected as being actively translated. **B.** P-site periodicity plot for a 31AA uORF located in the 5'UTR of *RASGRP3* (RAS Guanyl Releasing Protein 3), which was detected to have active translation signature in the human translation dataset (84.55% PIF, 84.38% Uniformity and 99.74%

Drop-off) and found to be conserved in the rat (Ribo-seq data from (Schafer et al., 2015);72.49% PIF, 78.12% Uniformity and 98.85% Drop-off) as well as other species (**C**).

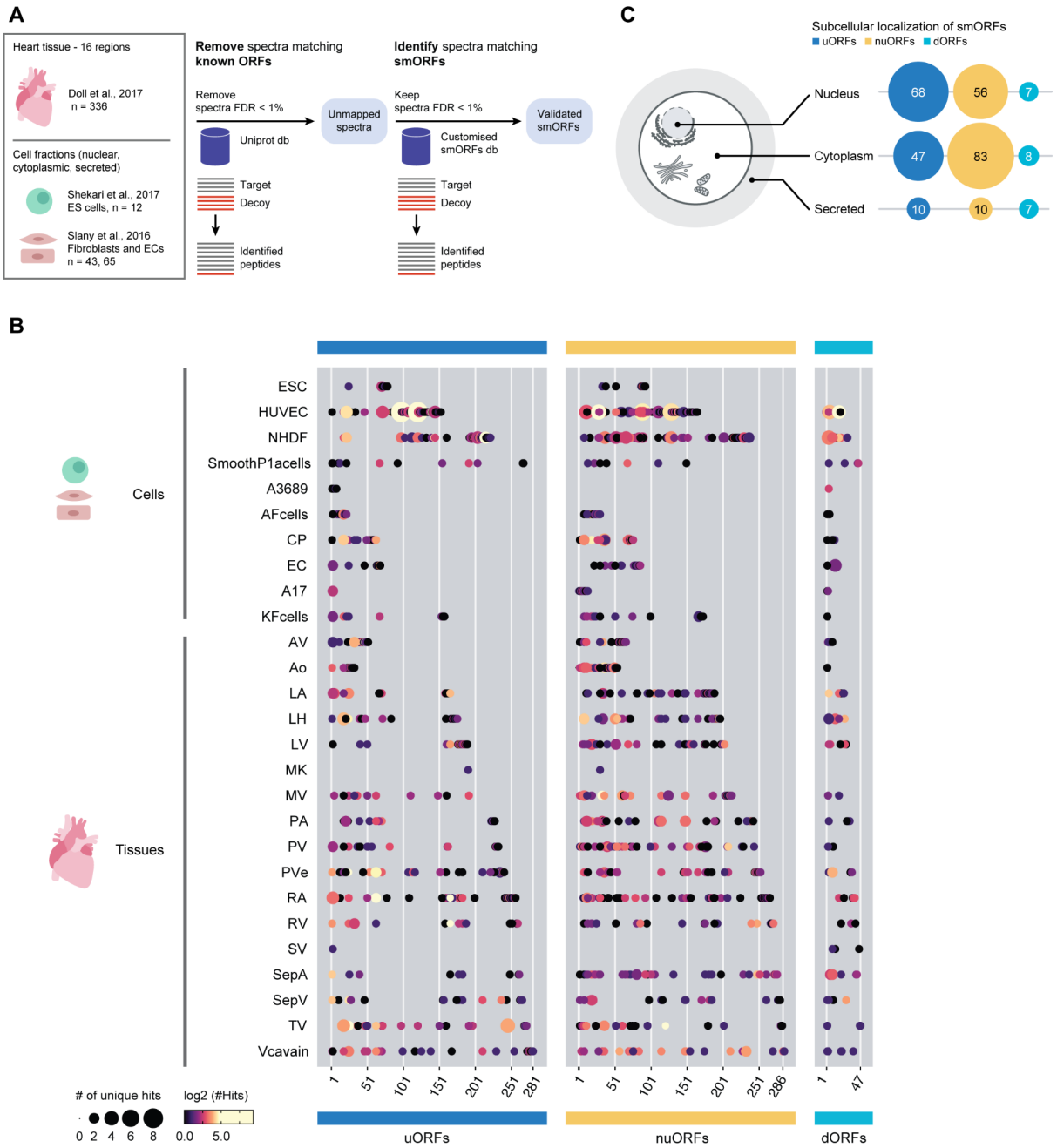


Figure 6: **Large-scale re-analysis of Mass-spectrometry data reveals 603 smORF encoded peptides**. **A.** Schematic of the strategy employed for the re-analysis of published mass-spectrometry datasets to identify smORFs. **B.** Bubble chart illustrating smORF encoded proteins (281 uORFs, 47 dORFs and 286 nuORFs) having at least one MS-hit across different samples (Tissues, cell types). Each circle represents an MS-hit for a given smORF in a given sample type. The colour scale indicates the number of total hits and the circle size represents the number of unique peptide sequences found to match the smORF. **C.** MS-evidence of smORF encoded proteins in different subcellular localisations. Abbreviations: ESC (Embryonic stem cells), HUVEC (Human umbilical vein endothelial cells), NHDF (N Human dermal fibroblasts), SmoothP1acells (Smooth muscle cells), AF (Adipose fibroblasts), EC (Endothelial cells), and tissues: AV (Aortic valve), Ao (Aorta), LA (Left atrium), LV (Left ventricle), MV (Mitral valve), PA (Pulmonary artery), PV (Pulmonary valve), PVe (Pulmonary vein), RA (Right atrium), RV (Right ventricle), SepA (atrial septum), SepV (ventricular septum), TV (Tricuspid valve); A3689, A17, CP, KFcells, SV, LH, MK,Vcavain as described in Doll et al. (Doll et al., 2017).