Published online 26 June 2018

Nucleic Acids Research, 2018, Vol. 46, No. 18 ***e109***
doi: 10.1093/nar/gky533

Ribosome elongating footprints denoised by wavelet transform comprehensively characterize dynamic cellular translation events

Zhiyu Xu†, Long Hu†, Binbin Shi, SiSi Geng, Longchen Xu, Dong Wang* and Zhi J. Lu*

MOE Key Laboratory of Bioinformatics, Center for Synthetic and Systems Biology, School of Life Sciences, Tsinghua University, Beijing 100084, China

Received January 24, 2018; Revised May 26, 2018; Editorial Decision May 28, 2018; Accepted May 31, 2018

ABSTRACT

Translation is dynamically regulated during cell development and stress response. In order to detect actively translated open reading frames (ORFs) and dynamic cellular translation events, we have developed a computational method, RiboWave, to process ribosome profiling data. RiboWave utilizes wavelet transform to denoise the original signal by extracting 3-nt periodicity of ribosomes and precisely locate their footprint denoted as Periodic Footprint P-site (PF P-site). Such high-resolution footprint is found to capture the full track of actively elongating ribosomes, from which translational landscape can be explicitly characterized. We compare RiboWave with several published methods, like RiboTaper, ORFscore and RibORF, and found that RiboWave outperforms them in both accuracy and usage when defining actively translated ORFs. Moreover, we show that PF P-site derived by RiboWave shows superior performance in characterizing the dynamics and complexity of cellular translatome by accurately estimating the abundance of protein levels, assessing differential translation and identifying dynamic translation frameshift.

INTRODUCTION

Translation is an essential and energy intensive step of biological process (BP) in cells (1,2). It is dynamically regulated in cell development and stress response (3). For instance, variation in translation initiation sites have delineated a dynamic range of translation regulation in response to different environmental stimuli (4–12). Another alternative translation event that contributes to dynamic translational landscape is ribosomal frameshift, an essential and universal translation process across species (13–18). Additionally, translation can also be regulated via immediate and selective changes in protein translation efficiency (TE) in which cells have developed to encounter different stimuli (19–21). To uncover the dynamic translation landscape of cell, ribosome profiling (Ribo-seq) has been developed to sequence RNA fragments protected by ribosomes and thus monitor translation events with unprecedented resolution (22,23).

Translation regulation usually occurs at the translation initiation phase where cells use different translation initiation sites under stress condition (9,24). Besides special drugs (i.e., harringtonine, lactimidomycin and puromycin) that are used to experientially denoise the input signal and selectively enrich initiating ribosomes (3,24,25), computational methods have been proposed to analyze Ribo-seq data and search for alternative translation processes (3,12,26,27). In addition, statistical tools have been developed to calculate the dynamics of translational efficiency where Ribo-seq signals are normalized by background (i.e., RNA-seq signals) (19–22,28–30). However, it is still hard to identify translation initiation site and calculate TE accurately based on Ribo-seq data alone due to the presence of intrinsic noises that are mainly introduced from experimental procedures and non-specific binding on RNAs (3,31–34). Given the fact that the presence of Ribo-seq reads is not equivalent to the indication of active translation (33), traditional identification of alternative translation process would be ineffective, placing a demand for Ribo-seq denoising.

An intrinsic feature of active translation that can be used for discriminating genuine translational signal against noises is trinucleotide (3 nt) periodicity (32,33). This periodicity originates from the process of codon-anticodon recognition during ribosome translocation (35). Several published tools have utilized this signature to detect actively translated open reading frames (ORFs) based on either uneven distribution among frames (3,26,36–38), uniform distribution across codons (39) or frequency derivation with Fourier transform (33). However, these methods cannot explicitly locate the full track of actively elongation of ribo-

*To whom correspondence should be addressed. Tel: +86 10 62789217; Fax: +86 10 62789217; Email: zhilu@tsinghua.edu.cn
Correspondence may also be addressed to Dong Wang. Email: dwang@biomed.tsinghua.edu.cn
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

somes, which makes them lack of power on detecting local behavior of translation, such as translation initiation, pausing and frameshift.

In this work, we propose that wavelet transform can be used to denoise Ribo-seq data and locate the footprint of active elongating ribosomes by extracting 3-nt periodicity. Wavelet transform is widely used in signal denoise in various fields (40–44). One of its recently biological applications is to eliminate non-experimentally transitions in PAR-CLIP data (45). Different from Fourier transform whose waves usually last for the entire duration of the signal, wavelet transform utilizes multiple small waves that oscillate at certain region along the input signal (46). Thus, wavelet transform gives not only frequency components (i.e., 3-nt periodicity of translating ribosomes) but also the exact positions of these frequency components. It is powerful for studying signal discontinuity and change point (45–49), such as translation initiation and ribosomal frameshift.

Therefore, we have developed a computational method, RiboWave, utilizing wavelet transform to denoise the Ribo-seq raw data and derive a set of Periodic Footprint P-sites (PF P-sites) of actively elongating ribosomes. We compare RiboWave with several published Ribo-seq analysis tools like RiboTaper, ORFscore and RibORF in defining active translated regions (i.e., ORFs) and show that RiboWave outperforms them in both accuracy and usage. More importantly, RiboWave can assay the dynamics of many cellular translation events, where PF P-sites are used to accurately estimate protein abundance, calculate TE and identify ribosomal frameshift.

MATERIALS AND METHODS

Pre-processing of Ribo-seq and RNA-seq

The sequencing data were processed similarly as previous described (33). Ribo-seq reads were firstly stripped from adaptor sequence and then reads that aligning to rRNA sequencing were removed using STAR (50). We aligned the remaining Ribo-seq reads and RNA-seq reads to the prebuild genome index using STAR. We allowed up to three mismatches and up to eight different positions multimapping for Ribo-seq and further eliminated alignments flagged as secondary alignments, ensuring one genomic position per aligned read (33). For RNA-seq, default parameters were used in alignment. The human genome index was obtained from hg19 and built using annotation file from GENCODE (version19) (51). Mouse genome index was built on mm10 and annotation file from GENCODE (version M8) (51). Arabidopsis genome index was built on tair10 (52). For zebrafish dataset, mapping bam file was obtained directly from the download package of RiboTaper (33).

P-sites construction and ORF scanning

P-sites' positions were inferred by investigating the offset of 5′ end of Ribo-seq reads to the first nucleotide of start codon of CCDS transcripts as described in previous study (33). For each read length, an individual aggregate profile was generated in order to determine P-sites positions precisely (Supplementary Figure S1). After that, we applied this offsets-to-read-length rule to all reads of the same length and obtained P-sites positions for Ribo-seq datasets. For RNA-seq, we used an arbitrary 25th position for all reads as the 'P-site' position of RNA-seq. Subsequently, we created P-sites tracks for all transcripts using the inferred P-sites positions for mapped reads.

ORFs were scanned from the FASTA sequence of RNA using custom python scripts. For each transcript, ORFs were scanned by searching all possible AUG start codons and paired with nearest in-frame stop codon (UAA, UAG, UGA). Small ORF (sORF) is defined if its length is shorter than 300 nt.

Denoising procedure

We utilized the wavelet transform to perform frequency-position spectrum analysis. For any function

$$f(t) \in L^2(\mathbb{R}) = \left\{ f(t) \,|\, \int_{-\infty}^{+\infty} |f(t)|^2 dt < \infty \right\},$$

Continuous Wavelet Transform (CWT) of it is

$$W_f(a,b) = \int_{-\infty}^{+\infty} f(t)\, \overline{\psi_{a,b}(t)}\, dt,$$

where $W_f(a,b)$ is the wavelet coefficients and $\psi(t)$ is the mother wavelet that depends on two indices, namely a (scale) and b (position). The wavelet function is defined by

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\, \psi\left(\frac{t-b}{a}\right) \quad (1)$$

By changing factors a and b continuously, we can get all the coefficients at different scales and positions. Thus, we can get a time-scale view of the original signal, and know exactly when a specific frequency occurs. In principal, CWT will decompose the original signal into numerous set of wavelets $\psi(t)$ by calculating coefficients for every a and b. In RiboWave, we use the discrete form of wavelet transform (DWT) to do the analysis to avoid the low efficiency and redundancy brought by CWT:

$$\psi_{j,k}(t) = 2^{j/2}\, \psi\left(2^j t - k\right)$$

$\psi_{j,k}(t)$ is called a series of daughter wavelets which, together with scaling function $\varphi(t)$, is sufficient for decomposing the original signal and obtain corresponding DWT coefficients; besides, these daughter wavelets can be made orthogonal to each other to eliminate the redundancy of CWT (53). Then, the detail $f(t)$ can be represented in the form of

$$f(t) = \sum_{j=j_0}^{\infty} \sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(t) + \sum_{k\in\mathbb{Z}} c_{j_0,k}\varphi_{j_0,k}(t),$$

where DWT coefficients is defined as:

$$d_{j,k} = < f, \psi_{j,k} > = \int_{-\infty}^{+\infty} f(t)\, \psi_{j,k}(t)\, dt,$$

$$c_{j_0,k} = < f, \varphi_{j_0,k} > = \int_{-\infty}^{+\infty} f(t)\, \varphi_{j_0,k}(t)\, dt$$

In our study, we used an analogous of DWT, Discrete Wavelet Packet Transform (DWPT), to perform wavelet transform. The advantage of DWPT is that it decomposes not only low-frequency part of a signal, but also high-frequency part as well (54). It can split the frequency band of 0 Hz ~ 1 Hz into $2^6 = 64$ frequency bands. In RiboWave, we used wavelet function Symlets 4(sym4) and its scaling

function for analysis (55). Particularly, we only considered the frequencies within 0.2Hz ∼ 0.5Hz (covering reads periodicity of 2, 3, 4, 5 nt) and chose 0.328125Hz ∼ 0.34375Hz band to present the 3-nt periodicity. For each frequency band, the coefficients of wavelet function are displayed as a vector, in which each value represents how the wavelet function fits the input signal at each position. The higher the coefficient value is, the closer local signal's frequency to the corresponding wavelet frequency. To track the footprint of elongation, signals with 3-nt frequency's coefficient lower than any other frequency's are treated as noise and thus eliminated. Thus, raw P-sites track is denoised and converted into PF P-sites.

RiboWave framework

RiboWave includes three major procedures: denoising procedure, translating ORF identification and translation initiation site prediction. For a Ribo-seq data, RiboWave first utilizes denoising procedure to convert the P-sites track into Periodic Footprint P-sites (PF P-site) track. The number of Ribo-seq reads before and after denoisng is listed in Supplementary Table 1. The second step, we test the in-frame PF P-sites enrichment over each candidate ORF by *chi-square* test. The definition of in-frame PF P-sites is based on the studied ORF. It refers to all the positions/loci that are located inside the ORF and are of the same reading frame as the ORF itself. Compared to the flanking regions of ORF (i.e., UTRs), we require in-frame PF P-sites must show higher enrichment inside the ORF. At a significant level of 0.05, ORF is predicted to be translated. In the *chi-square* test, the expectation values are the length of ORF and the length of ORF flanking regions; the observed values are the number of in-frame PF P-sites positions inside the ORF and the number of positions of PF P-sites in the flanking regions of ORF. *Chi-square* test is performed by R basic function chisq.test(). If any of the number is lower than five, we utilize the *Fisher-exact* test instead, i.e., R basic function fisher.test().

Third, RiboWave is designed to identify the most confident translation initiation sites among all initiation sites that are predicted to be translated under different situations (Supplementary Figure S7). First, we collect all ORFs that are predicted to be translated and cluster these ORFs according to its stop codon. Next, in case there are multiple start codons for a stop codon, we select the one with less upstream PF P-sites; if still there are multiple start sites remained, we pick the one with the highest P-sites intensity (P-sites of −1, 0, +1 nt of the candidate start codon's first nucleotide) as the final prediction (Supplementary Figure S7).

Performance comparison in identifying translated ORFs

First, the translational signal and the non-translational noise should be fully distinguishable. As previous study did (33), we selected Ribo-seq as translational signal and paired RNA-seq as non-translational signal (GSE51424) (56). For both translational and non-translational signal, we generated a test set using original P-sites track as positive and shuffled P-sites track as negative (33). For both (positive and negative) test sets, we converted the raw P-sites track into PF P-sites track and compared raw P-sites track and PF P-sites track by testing if they were enriched inside the 1000 sampled CCDS transcripts' annotated ORF regions. Similarly, we also tested the performance of RiboWave, RiboTaper, RibORF and ORFscore on the same test set. Default parameters were used for RiboTaper, ORFscore and RibORF as mentioned in published protocol (33,36,39).

Second, as the off-frame overlapping ORFs largely overlap with the translated ORFs, it is important to specify which frame is under translation. To do so, we generated a test set using 1000 sampled CCDS transcripts' annotated ORFs as positives and their off-frame overlapping ORFs as negatives. Since an annotated ORF might have multiple off-frame overlapping ORFs, we balanced the ratio of positive to negative to one by choosing the most overlapped off-frame ORF as negative control. Subsequently, we compared the positives and negatives by calculating the enrichment *chi-square* test *P*-values for raw P-sites and PF P-sites. Similarly, we also tested the performance of RiboWave, RiboTaper, RibORF and ORFscore on the same test set.

For the above two test sets, we also evaluated the influence of different drug treatments using another Ribo-seq dataset from mouse bone marrow-derived dendritic cells (GSE74139) treated with four different kinds of drug: harringtonine, lactimidomycin, cycloheximide and no drug (57).

Third, for a translated ORF, it is important to decide the translation initiation site out of multiple candidate start codons. To do so, we used Ribo-seq as input and predicted the translation initiation site by RiboWave and RiboTaper. We later compared the predicted initiation sites with the experimental validated ones, which were from paired QTI-seq (SRA160745) or paired ORF-RATER method (GSE74139) (24,57). Only ORF with start and stop sites both matched were reported as validated. Only AUG start codon were considered.

Fourth, to validate the prediction results of RiboWave, RiboTaper, RibORF and ORFscore, we used HCT116 cell line Ribo-seq (GSE58207) as input and made prediction for translated CCDS genes, which was later validated by paired mass spectrometry (MS) datasets (PXD000304) (58). Same comparison was performed on sORF as well.

Mass spectrometry data processing

Proteomics MS data (Pride database identifier: PXD000304) was searched against a custom protein database of all human ORFs using MaxQuant following the setting as described in previous study (59). Briefly, heavy labeled arginine (13C6) and lysine (13C6) were additionally selected as fixed modifications, and methionine oxidation to methionine-sulfoxide, pyroglutamate formation of N-terminal glutamine and acetylation of the N-terminus were selected as variable modifications. Precursor ions mass tolerance was set to 10 ppm, while fragment ions mass tolerance was set to 0.5 Da. The maximum charge of peptide was 4+. The minimum peptide length was set at 7, which is the default of MaxQuant and only one miss cleavage was allowed.

Protein abundance estimation

Protein abundance estimation was done on proteins that have enough reads density (>30 PF P-sites) at all time points within the annotated ORF. A total of 1340 proteins were estimated for its RPKM of P-sites and PF P-sites on annotated ORF. In case of multiple annotated ORFs for one single protein, only the longest annotated ORF was selected for further calculation. Relative protein abundance (iTRAQ) throughout the time course (0, 1.5, 3, 6, 9,12 h) was obtained from previous study coupled with paired Ribo-seq data on the cell line MM1.S after the treatment of bortezomib (GSE48785) (60). During comparison, proteins were equally split into three groups (low, medium and high) according to reads abundance. Pearson correlation was considered here. Relative abundance was calculated by Z-score normalizing both RPKM of raw P-sites and PF P-sites and iTRAQ intensity across time series so that iTRAQ and raw P-sites/PF P-sites intensities were at the same scale. Relative deviation against iTRAQ was calculated by measuring the absolute difference between the relative abundance of raw P-sites/PF P-sites and iTRAQ.

Differential translation detection

We used two published datasets: human PC3 cells in response to mTOR signaling perturbation (GSE35469) and dark-grown *Arabidopsis* seedling in response to light stimulus (GSE43703) (20,21). Replicates were combined together for further studies. To identify dysregulated translation genes with high confidence, we required protein coding genes should contain sufficient amount of reads (RNA-seq: RPKM > 0.5, more than 20 mapped reads; Ribo-seq: more than 20 PF P-sites) (28,34,57,61,62). In PC3 data, 10,957 protein coding genes in PC3 data and 13,240 genes in *Arabidopsis* data were subjected to TE calculation. During the calculation of TE, the abundance of translation was represented by the RPKM of either raw P-sites or PF P-sites while the RPKM of raw reads from RNA-seq was used to represent the abundance of transcription. TE was calculated by the ratio of translation abundance and mRNA abundance. Fold change of TE under two conditions were transformed into Z-scores after fitting the data into a normal distribution (28,63,64). P-values were inferred from the distributions. At last, Z-score > 2 were denoted as translation upregulation genes and Z-score < −2 were denoted as downregulation (28,29,65).

Gene ontology analysis

The enriched functional groups were revealed with use of the elim method from the TopGO package to estimate the enrichment of BP terms for a certain gene set (66). The *Fisher exact* test was used to evaluate the representation differences between differentially translated genes detected by raw P-sites and by PF P-sites. Meanwhile, differential translation genes were also subjected to KEGG enrichment analysis using the Database for annotation, visualization and integrated discovery (67).

Gene set enrichment analysis

We used GSEA (v3.0) (68) to assess enrichment of sets of dysregulated translation genes in corresponding gene ontology (GO) geneset. GSEA requires two separate input files: a gene set of interest and differentially translated genes. Here we took different numbers of top ranked genes estimated by either raw P-sites or PF P-sites, i.e., uptranslated and downtranslated, and sorted by its corresponding TE fold change to see the enrichment of the studied gene set over these differentially translated genes. GeneRatio indicates the overlap ratio between the gene set and the input gene list.

Frameshift detection procedure

Frameshift refers to the translation events when the ribosomes exhibit a non 3-nt movement and thus cause a shift in the reading frame (Supplementary Figure S16A, down panel). A change point was defined as the position where the PF P-site's reading frame is different from the last PF P-site's frame. Relative to the change point, we defined the upstream (*up*) region as the region from annotated start codon of the ORF to change point; downstream (*down*) region as the region from change point to the last PF P-site position along the transcript.

To ensure the reading frame before and after the frameshift is different, we came up with a score *Frame change* (Supplementary Figure S16B):

$$\text{inframe\_Perc}_{j,\ j=up,\ down} = \frac{\text{number of PF Psites in annotated frame}}{\text{number of PF Psites in 3 frames}}$$

$$\text{Frame change} = 1 - \frac{2 \times \text{inframe\_Perc}_{up} \times \text{inframe\_Perc}_{down}}{\text{inframe\_Perc}_{up} + \text{inframe\_Perc}_{down}}$$

Frame change can quantify the difference in reading frame from upstream and downstream. To be specific, $inframe\_Perc_{j,\ j=up,\ down}$ calculates the percentage of PF P-sites within the annotated reading frame in either upstream (*up*) or downstream (*down*) region. Difference between $inframe\_Perc_{up}$ and $inframe\_Perc_{down}$ can be incorporated into *Frame change*, with its value approaching 1 indicating the potential of frameshift.

Next, to ensure the consistency of translational signal before and after the change point, we came up with a score *Frame dominancy* (Supplementary Figure S16C):

$$\text{Perc}_{Fi,\ i=0,1,2} = \frac{\text{number of PF Psites in frame}_i}{\text{number of PF Psites in 3 frames}}$$

$$\text{FD}_{j,\ j=up,\ down} = 2 \times \max(\text{Perc}_{F0},\ \text{Perc}_{F1}, \text{Perc}_{F2}) - 1$$

$$\text{Frame dominancy} = \frac{2 \times \text{FD}_{up} \times FD_{down}}{FD_{up} + FD_{down}}$$

Similar to $inframe\_Perc_{j,\ j=up,\ down}$, $Perc_{Fi,\ i=0,1,2}$ calculates the percentage of PF P-sites inside each reading

frame ($i = 0, 1, 2$) in either upstream (*up*) or downstream (*down*) region. $FD_{j,\ j=up,\ down}$ incorporates all three percentage $Perc_{F0}$, $Perc_{F1}$, $Perc_{F2}$ from either the upstream (*up*) or downstream (*down*) region respectively. The value of $FD_{j,\ j=up,\ down}$ indicates the enrichment of PF P-sites within one single frame. The higher $FD_{j,\ j=up,\ down}$, the more consistent PF P-sites are within one reading frame. At last, *Frame dominancy* is able to quantify the PF P-sites consistency in both upstream (*up*) or downstream (*down*) region, with its value approaching 1 indicating the potential of frameshift.

At last, we incorporated both *Frame dominancy* and *Frame change* and came up a new score CRF. The higher CRF, the most likely it is having a frameshift event.

$$\text{CRF} = \text{Frame change} \times \text{Frame dominancy}$$

As one single ORF may contain multiple change points, the highest CRF score was selected as the CRF score of that ORF and corresponding change point was defined as the change point of that ORF. We calculated the CRF score for all annotated ORFs. ORF with CRF score > 0.6 and downstream signal including more than five out-frame PF P-sites was selected as a candidate of frameshift event. In the next, we tried to remove candidates that might be caused by other reasons: downstream off-frame ORFs translation (Supplementary Figure S16E) and chimera P-sites track (Supplementary Figure S16F). To do so, we removed the cases when downstream PF P-sites located fully inside the unannotated ORF or overlapped in the same frame with any other annotated ORFs as well as ORFs predicted to be translated. Further, we required the reads coverage of the downstream region should be similar with the upstream region (>0.75-fold and <1.5-fold). Candidates fulfilled all above criteria were designated as frameshift events.

The frameshift events were detected from HCT116(GSE58207), HeLa(GSE79664), glioblastoma(GBM)(GSE51424), clear cell renal cell carcinoma(ccRCC)(GSE59820) and HEK293T(GSE65778) and validated by paired MS datasets (PXD000304) (58).

Indel simulation

To systematically evaluate the ability of CRF score in detecting frameshift event, we performed a simulation to randomly introduce indels (insertion/deletion) in the annotated ORF's DNA sequence so as to create an artificial frameshift. Later, we quantified this frameshift by calculating the CRF score before and after the indels. Here we took 25,554 annotated ORFs that were predicted to be translated in HCT116 sample. The indel positions were treated as the change point during the calculation of CRF score.

RESULTS

RiboWave identifies the periodic footprint of ribosomes by wavelet transform

Given the fact that actively elongating ribosomes usually have a consistent, 3-nt periodic codon-by-codon pattern, extraction of the 3-nt periodic footprint would be a critical point during the denoising procedure. We use wavelet transform for purpose because it derives both frequency and position information of the ribosomes from Ribo-seq data, i.e., it shows not only what frequencies are present, but also which ribosomes are present with such frequency (see 'Materials and Methods' section and Figure 1A). Based on wavelet transform, we have developed a powerful and rigorous Ribo-seq analysis tool, RiboWave, that denoises Ribo-seq data (peptidyl-site track, P-site track) by extracting 3-nt periodicity of ribosomes. The denoised data track, denoted as Periodic Footprint P-sites (PF P-sites), is capable of capturing the full track of ribosomes' 3-nt periodic footprint, i.e., start, movement and stop of the actively elongating ribosomes (Figure 1A).

The detail pipeline of PF P-sites generation is illustrated in Figure 1B, in which we first converted Ribo-seq's raw reads into positions of peptidyl-sites (P-sites) as described in previous study (33) (Figure 1B and Supplementary Figure S1). Then, we used wavelet transform to generate a coefficient matrix of frequency at each nucleotide, i.e., frequency-position spectrum matrix. In the frequency-position spectrum matrix, values are the coefficients of different wavelet functions (i.e., different frequencies, rows of matrix) fitting the input signal (i.e. P-sites intensities) at given positions (i.e. nucleotide positions, columns of the matrix). Subsequently, raw P-sites were denoised into PF P-sites by removing noises with frequencies other than 3 nt: for each nucleotide, raw P-sites' signal is neglected if it shows dominant frequency other than 0.33Hz. Therefore, the denoised PF P-sites represent a high-resolution 3-nt periodic footprint of elongating ribosomes (Figure 1B).

We illustrated the signals of raw P-sites and PF P-sites on two example transcripts, a mRNA and a long noncoding RNA (lncRNA) (Figure 1C and Supplementary Figure S2). The P-sites signal is split into three reading frames, frames 0, 1 and 2, by counting the position of nucleotide from the start of a transcript. The example mRNA, which is annotated by GENCODE (51) as consensus coding sequence (CCDS) transcript, shows clear enrichment of PF P-sites inside the annotated ORF on frame 1 (*chi-square* test P-value < 0.05), while the raw P-sites are noisy and scatter over all three frames. On the contrary, we see lots of raw P-sites falling inside a putative ORF on the example lncRNA, while the signal track of PF P-sites is clean. This illustration suggests that RiboWave can remove those ambiguous raw P-sites and produce a track of high-resolution periodic footprint, PF P-sites, for actively elongating ribosomes.

RiboWave accurately identifies actively translated ORFs

One of the direct application of Ribo-seq is to find actively translated regions. It includes the following steps: distinguish translational signal from noise, specify correct reading frame and pinpoint translation initiation and termination sites. Here, we demonstrate that PF P-site shows superior performance in defining translating regions when non-translational noises are removed by RiboWave.

First, we aim at evaluating the performance of PF P-sites in distinguishing translational signal from non-translational noise. To do so, we randomly sampled 1000 CCDS transcripts' Ribo-seq P-site tracks as translational signals and their shuffled P-site tracks as noise and com-

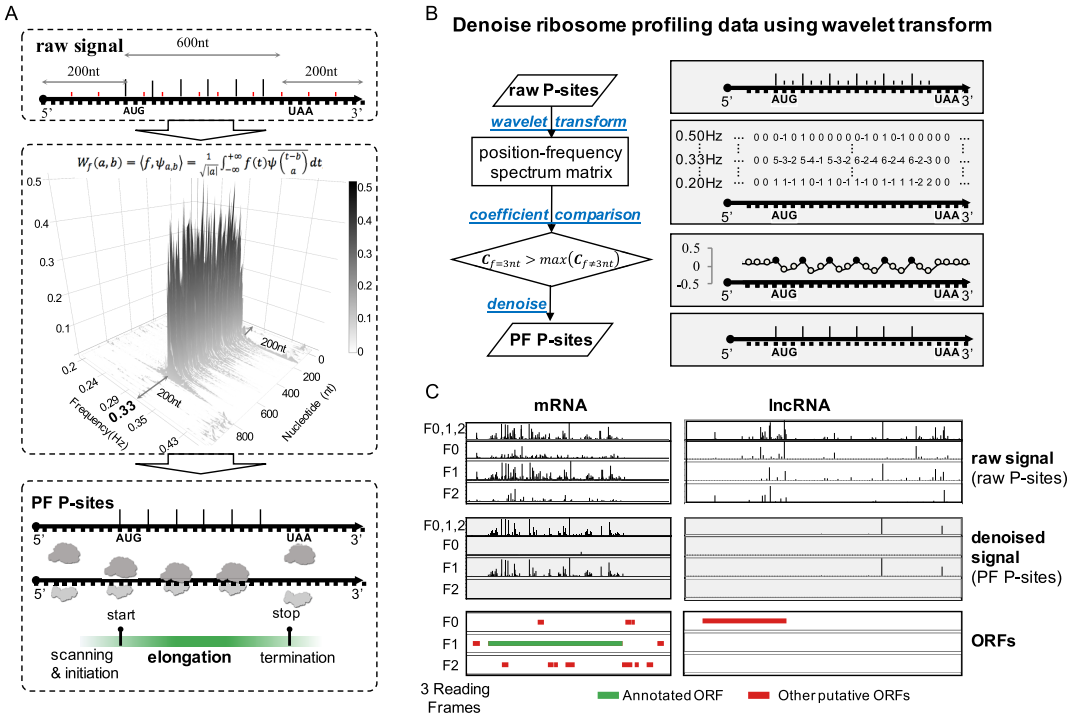

Figure 1. RiboWave detects high-resolution periodic footprint, PF P-sites, from ribosome profiling data using wavelet transform. (**A**) Wavelet transform decomposes the raw Ribo-seq signal into position-frequency matrix, where *x*-axis indicates frequency, *y*-axis indicates nucleotide position and *z*-axis indicates the coefficient of each nucleotide for each frequency. Based on that matrix, we are able to extract 3-nt periodicity of actively elongating ribosomes and precisely locate their footprint denoted as Periodic Footprint P-site (PF P-site). The denoised signal, PF P-site track, provides valuable information about the start, movement and stop of the elongating ribosomes. (**B**). Detailed procedure of denoising. First, raw P-sites' track is directly inferred from reads of Ribo-seq. Second, a position-frequency spectrum matrix is generated by wavelet transform. Values in the matrix are the coefficients of different wavelet functions (i.e., different frequencies) for each nucleotide. Thirdly, we compare the coefficients and smooth the signals other than 3-nt periodicity (0.33 Hz). At last, we derive the Periodic Footprint P-sites (PF P-sites). $C_{f=3\,\mathrm{nt}}$ indicates the coefficient of 0.33Hz (3-nt periodicity); $C_{f\neq 3\,\mathrm{nt}}$ indicates the coefficient of other frequencies. (**C**) Raw and PF P-sites on two example transcripts, mRNA (ENSG00000059804) and lncRNA (ENSG00000258644). The original signals of P-sites (F 0,1,2) are distributed at three frames, F0, F1 and F2. Annotated ORF is colored in green. Putative ORFs derived by directly sequence scanning are colored in red.

pared two types of signals' enrichment over annotated ORFs by *chi-square* test *P*-values (Figure 2A). Both raw P-site track (before denoising) and PF P-site track (after denoising) showed significant enrichment over annotated ORFs for translational signal and no enrichment for shuffled noise, indicating both raw P-sites and PF P-sites can distinguish translational signal from shuffled noise (Figure 2A). The same comparison was done by taking the same transcripts' signal from paired RNA-seq dataset as non-translational signal and their shuffled P-sites as the noise (Figure 2B) (33). No enrichment was detected for PF P-sites on non-translational RNA-seq data, suggesting its unique specificity in identifying translational signal (Figure 2B down panel). On the contrary, significant enrichment over ORFs on RNA-seq suggests the use of raw P-sites might lead to false identification, which is consistent with previous study (Figure 2B up panel) (33). We summarized the above comparisons by AUC (area under the receiver operating characteristic curve) scores using the original tracks of Ribo-seq or RNA-seq as positive and their shuffled tracks as negative. For the Ribo-seq dataset, both raw P-site track and PF P-site track achieved high AUC scores (0.96 and 0.94, respectively) (Figure 2C and Supplementary Figure S3). For the RNA-seq dataset, only the PF P-site track showed an expected AUC score of approaching 0.5 (0.499); raw P-site track showed a AUC score as high as 0.73 (Figure 2C and Supplementary Figure S3). These results show that PF P-site is able to specifically distinguish the footprint of actively elongating ribosomes from non-translational noises.

Next, we want to track the trace of actively elongating ribosomes by identifying the correct reading frame out of three candidate frames. Performances were compared between positive set: CCDS transcripts' annotated ORFs and negative set: the most overlapped off-frame ORFs from the same transcript. By the ratio of overlapping, we divided the test sets to five levels: 0 ∼ 100%, 0 ∼ 25%, 25 ∼ 50%, 50 ∼ 75% and 75 ∼ 100%. With the increase of overlapping ratios, PF P-site track was shown to have much better performance than raw P-sites. (PF P-sites: 0.94, 0.94, 0.93, 0.90 and 0.89; raw P-sites: 0.84, 0.84, 0.76, 0.67 and 0.68) (Figure 2D and Supplementary Figure S3). Collectively, the results show that PF P-site is able to recognize the correct reading frame of translational footprint.

Cell has evolved exquisite mechanism to regulate translation initiation (3). However, it has been pointed out that standard Ribo-seq is not suitable for detecting translation initiation sites due to the existence of non-translational noise (24,25,69). As demonstrated in Figure 2E, a metagene pattern was plotted around the annotated start codons

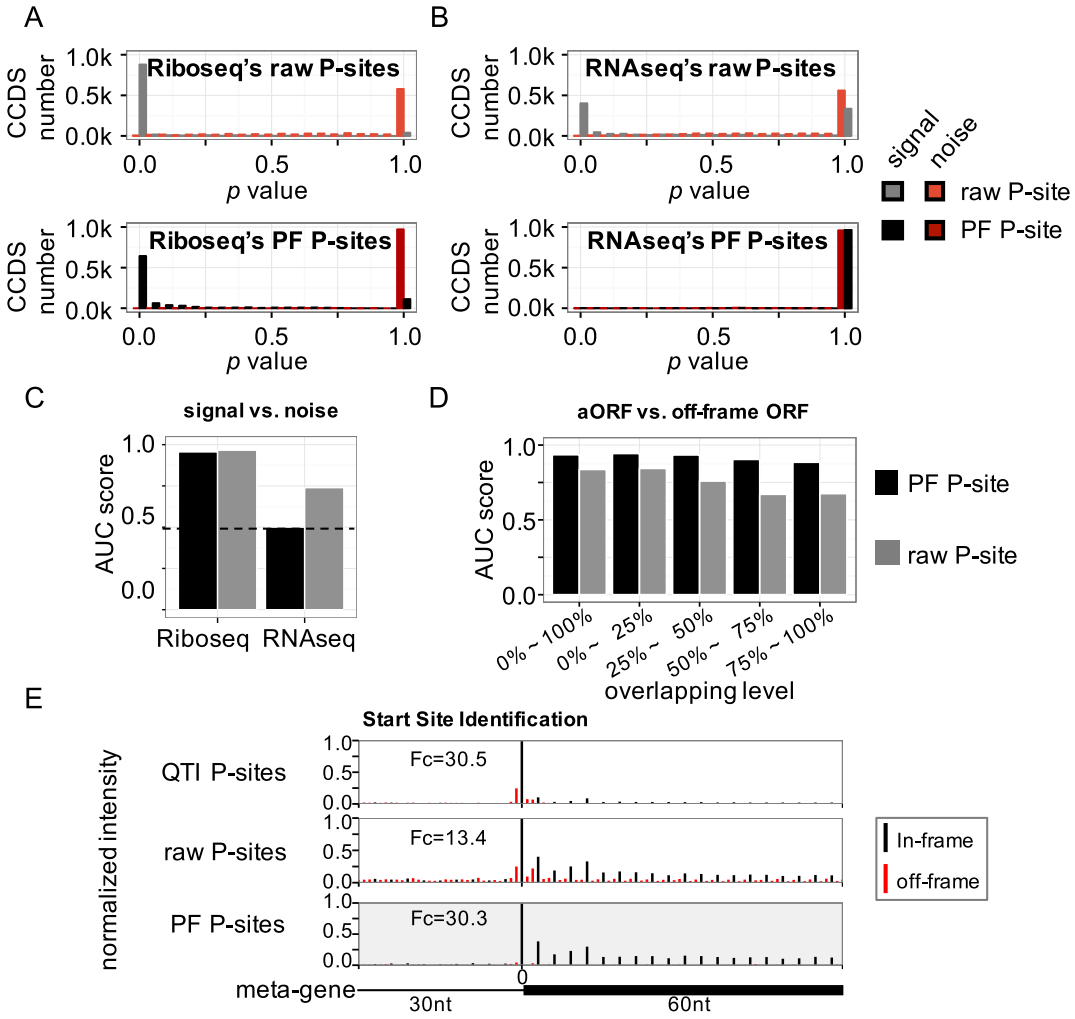

Figure 2. PF P-sites derived from RiboWave identify translational signal, locate reading frame and pinpoint translation initiation site. Distribution of raw and PF P-sites within the annotated ORFs. *Chi-square* enrichment test *P*-values are calculated for signal (original signal) versus noise (shuffled signal) based on 1000 CCDS transcripts. The *P*-value is calculated from (**A**) Ribo-seq data and (**B**) RNA-seq data (used as control). (**C**) The AUC scores are also calculated when distinguishing signals from noises. (**D**) AUC scores are calculated for the annotated ORFs (aORFs) versus overlapped off-frame ORFs at different overlapping levels. (**E**) The meta-gene patterns around the annotated start sites for QTI-seq P-sites, Ribo-seq raw P-sites and Ribo-seq PF P-sites. Fold-change (Fc) of each type of signal is the intensity of start codon (−1,0,+1 nt) over upstream region (−30 to −2 nt).

(−30 to +60 nt) of CCDS transcripts for three types of signals: P-site track from quantitative translation initiation sequencing (QTI-seq) (24), raw P-site track from Ribo-seq data and PF P-site track denoised from the same Ribo-seq data. QTI-seq was used as a positive control because it utilizes specific drug treatments, lactimidomycin and puromycin, to enhance the translation initiation signal with its P-sites mostly enriched around the initiation sites (3,24). Though a peak at the annotated start codon was observed, raw P-site track was coupled with unneglectable signal scattered everywhere, indicating its limited power in detecting translation initiation sites. In contrast, PF P-site track showed barely signal before initiation, but clean and consistent codon-by-codon in-frame signal after initiation. The advantage of PF P-sites was also revealed by calculating the intensity fold-change ratios (Fc) between start codon region (−1, 0 and +1 nt of the first nucleotide of start codon) and upstream region (−30 to −2 nt of the first nucleotide of start codon). PF P-site track achieved comparable fold-change ratio (Fc = 30.3) with QTI-seq (Fc = 30.5) and much higher than raw P-site track (Fc = 13.4) (Figure 2E). These results suggest that PF P-site is of great sensitivity in identifying translation initiation site using standard Ribo-seq as input, without additional drug treatment procedures. In addition, the denoised PF P-site of Ribo-seq data might be a nice supplement to QTI-seq in two situations, when the starting peak does not lead to a complete translation (34,70), and when the starting peak is caused by experimental containments (31,71).

Comparison of RiboWave and other Ribo-seq analysis tools on ORF identification

Above results highlight the novelty and superiority of PF P-site in identifying actively translated regions, prompting us to further compare RiboWave with other published Ribo-seq analysis tools. The ORF detection pipeline of RiboWave

begins with the derivation of PF P-sites, followed by in-frame enrichment test to see if PF P-sites are enriched in any candidate ORFs (Figure 3A). RiboWave start codon prediction is set for the purpose to identify the most confident translation initiation site based on Ribo-seq data alone (Supplementary Figure S7). The real translation initiation site should remark the emergence of actively elongating ribosomes. As a result, we reason it should have least upstream active translating footprints, PF P-sites, even in the occasion when upstream ORF (uORF) exists (Supplementary Figure S7). Following these steps, we are able to determine whether the transcript is translated and if does, which ORF is under translation and where the translation initializes. We compared the performance of RiboWave with three other well-known Ribo-seq analysis tools, RiboTaper, ORFscore and RibORF, in characterizing ORF translation (33,36,39).

First, we compared these tools in distinguishing translation signal from non-translational noise (Figure 3B). Like we did in Figure 2C, we compared the AUC scores of four tools on Ribo-seq dataset: RiboWave showed the highest AUC score as 0.94; RiboTaper, RibORF and ORFscore showed lower AUC scores of 0.93, 0.83 and 0.80. When applied on non-periodic RNA-seq test set, RiboWave showed the closest AUC score to 0.5: 0.498; RiboTaper, RibORF and ORFscore showed greater deviations from 0.5 (0.66, 0.32 and 0.43, respectively), indicating their weakened power in discriminating non-translational noises. Similarly, we also conducted the comparison by treating different drugs (no drug, cycloheximide, lactimidomycin and harringtonine) (Figure 3C). Despite the usage of lactimidomycin and harringtonine in specifying translation initiation sites, RiboWave still showed robust and top performance. Similar result was also observed on different P-sites coverage and RNA expression level (Supplementary Figure S4). Both RiboWave and RiboTaper detect the 3-nt periodicity, while RibORF and ORFscore utilize the uneven distribution of three frames. Thus, when separating translational signal from noise, only RiboTaper was comparable to RiboWave.

Next, we evaluated the power of eliminating off-frame overlapping ORFs for four tools (Figure 3D and Supplementary Figure S5). Like we did in Figure 2D, we compared the AUC scores of four tools on different overlapping levels: $0 \sim 100\%$, $0 \sim 25\%$, $25 \sim 50\%$, $50 \sim 75\%$ and $75 \sim 100\%$. For every overlapping level, RiboWave (0.94, 0.94, 0.93, 0.90 and 0.89, respectively), RibORF (0.95, 0.96, 0.93, 0.92 and 0.88, respectively) and ORFscore (0.95, 0.95, 0.93, 0.93 and 0.88, respectively) showed comparable AUC scores and much higher AUC scores than RiboTaper (0.81, 0.82, 0.69, 0.60 and 0.56, respectively). Such result was consistent when considering different drug treatment or different P-sites coverage, RNA expression level (Supplementary Figures S5 and 6). RiboWave, RibORF and ORFscore explicitly utilize reading frame information, while RiboTaper merely detects the existence of 3-nt periodicity for a given ORF using Fourier transform, lack of location information, e.g. which specific reading frame. Thus, the former three methods performed better than RiboTaper on separating in-frame ORFs from overlapped off-frame ORFs.

At last, we compared RiboWave with RiboTaper on detecting translation initiation sites based on Ribo-seq data (RibORF and ORFscore do not provide utility to identify translation initiation sites). We used QTI-seq as validation set and paired Ribo-seq as input set. For 2829 initiation sites detected by QTI-seq, RiboWave matched 2093(74%) of them and RiboTaper matched 1966 (69%) (Figure 3E and Supplementary Figure S8). Since the detection of QTI peak does not guarantee the completion of translation, we next compared the accuracy of determining the correct initiation sites out of all ORFs that were predicted to be translated: RiboWave showed a rate as high as 91% compared to RiboTaper (85%) (Supplementary Figure S9). We repeated the comparison on another validation set ORF-RATER (57), which incorporated Ribo-seq with different treatments (harringtonine, lactimidomycin, cycloheximide treated or no-drug treated) to predict translation initiation sites (57). Using paired untreated Ribo-seq data as input set, RiboWave matched 12,153 (81%) out of 15,055 initiation sites as detected by ORF-RATER, while RiboTaper matched only 8599 (57%) of them (Figure 3E and Supplementary Figure S8). Besides, in terms of accuracy, RiboWave showed much higher rate of 93% than RiboTaper (77%). These combined results suggest RiboWave prediction highly agrees with experiments defined translation initiation sites and thus might be more suitable in identifying translation initiation sites from Ribo-seq data, especially when QTI-seq or other drug treatment data not available. Figure 3F provides an example of translation initiation site where transcript's QTI-seq P-site track, raw P-site track and PF P-site track are plotted. Translation initiation site candidates are highlighted in the top panel where the first AUG start codon is designated as the real translation initiation site by QTI-seq. Comparison of RiboWave and RiboTaper shows that RiboWave detects the correct initiation site from which PF P-sites start to emerge while RiboTaper tends to pick the downstream start site instead (Figure 3F).

At last, MS data were used to validate the translation of predicted ORFs by four tools (Figure 3G and Supplementary Figure S10). Out of 2351 CCDS genes validated by MS, ORFscore showed highest recovery rate as 99.9% (2348) followed by RiboWave 96.9% (2278), RiboTaper 96.5% (2268) and RibORF 95.1% (2236), respectively. In terms of positive predicted value (PPV), RiboWave showed highest PPV as 0.197 (11,562) while ORFscore showed lowest PPV as 0.184 (12,778). The result suggests that the predicted ORFs using different tools on Ribo-seq data have similar overlap with MS data. In terms of predicting translation on sORFs, RiboWave also shows comparable result and highest PPV as validated by MS (Supplementary Figure S11). Furthermore, RiboWave have also successfully verified the translation of sORF in a lincRNA, *toddler* (Supplementary Figure S12) (33,72).

Collectively, the comparison results show that RiboWave is accurate and robust, and can be used to systemically identify actively translated ORFs. In addition to this, a novel and unique feature of RiboWave is that it explicitly derives a track of active translation signals, PF P-sites, along each transcript. In the following sections, we illustrate the usages of RiboWave in understanding the dynamics of trans-

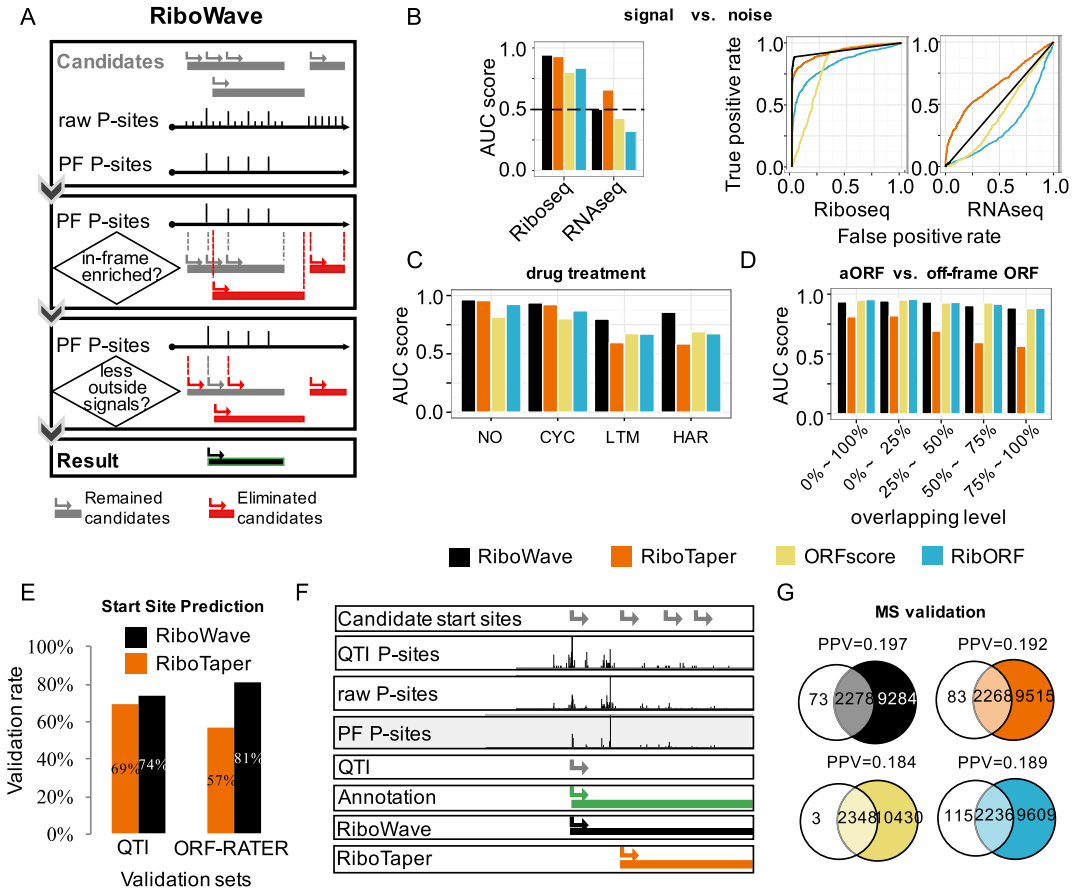

Figure 3. Comparison between RiboWave and three other Ribo-seq analysis tools. (**A**) Workflow of RiboWave for identifying translating ORF based on PF P-sites. RiboWave first denoises raw P-sites into PF P-sites. Second, it tests if PF P-sites are in-framely enriched in a certain ORF. Third, RiboWave determines the translation initiation site by minimizing upstream PF P-sites. At last, one ORF having the highest intensity within the initiation site is identified as final prediction. (**B**) AUC scores and ROC curves of distinguishing signal (original signal) from noise (shuffled signal) for four tools, RiboWave, RiboTaper, RibORF and ORFscore, based on two datasets, Ribo-seq and RNA-seq (used as control). (**C**) Performance of four tools in distinguishing translation signal (Ribo-seq) from noise (shuffled signal) on special drug treatment (no drug, cycloheximide, lactimidomycin and harringtonine). (**D**) AUC scores of distinguishing annotated ORFs from overlapped off-frame ORFs at different overlapping levels. (**E**) Validation rate of the translation initiation sites predicted by RiboWave and RiboTaper, using two validation sets, QTI-seq and ORF-RATER. (**F**) An example of identifying translation initiation site. (**G**) MS validation for the ORFs predicted by different tools on gene level.

lational landscape, where PF P-site plays an indispensable role.

RiboWave accurately estimates dynamics of protein abundance

Translation regulation plays a pivotal role in the control of protein synthesis (9). Intensity measured from Ribo-seq is known to have better correlation with protein abundance than RNA-seq does (22,73–76). To assess the performance and utility of RiboWave on protein abundance estimation based on denoised Ribo-seq data (i.e., PF P-sites), we first compared the intensities of raw P-sites and PF P-sites via the correlation to absolute protein abundance. In this case, RPKM of the annotated ORF was used to estimate the reads intensity of either raw P-sites and PF P-sites for each protein.

We first investigated the correlation between PF P-sites intensity and protein levels in multiple time points. We curated Ribo-seq data at six individual time points (0, 1.5, 3, 6, 9, 12 h) with its protein abundance quantified by the approach of isobaric tag for relative and absolute quantitation (iTRAQ) (76). Comparison was done across the time course and showed that the intensities of PF P-sites and iTRAQ measured protein abundance shared consistent dynamics, while raw P-sites intensities showed limited correlation with iTRAQ data (Figure 4A up panel, 4B and 4C). We have achieved the same result when averaging the intensity of raw P-sites, PF P-sites and iTRAQ at each time point (Figure 4A bottom panel). Both correlation (Figure 4B) and deviation (Figure 4C) against iTRAQ intensity were stable within proteins of different levels of reads intensities (low, medium and high), suggesting PF P-site's robustness in estimation protein abundance. The example of protein SEC31A is shown in Figure 4D, illustrating the advantage of PF P-sites over raw P-sites. To sum up, the above results suggest that PF P-site serves as a better estimator of protein abundance than raw P-site.

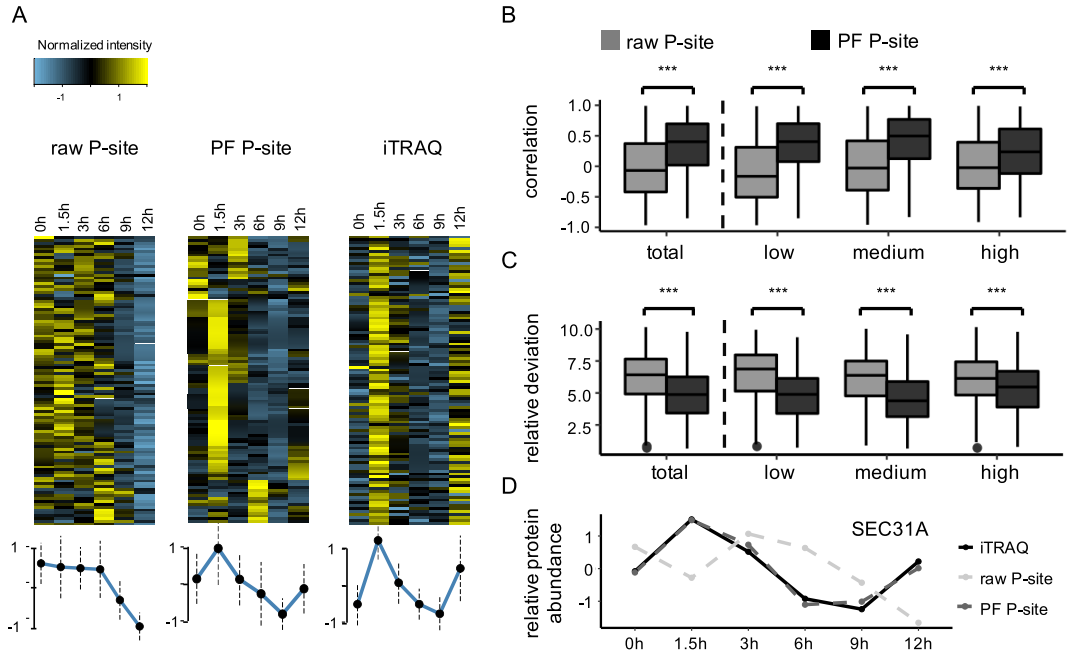

Figure 4. PF P-sites derived from RiboWave accurately evaluate the dynamics of protein abundance. (**A**) Heat maps show the expression (RPKM) of raw P-sites (left), PF P-sites(middle) and iTRAQ intensity (right) for 100 randomly sampled proteins across time series. Average relative intensity of sampled proteins is shown in the bottom. (**B**) Correlation between iTRAQ intensities and raw P-sites/PF P-sites RPKM across time course. Proteins are equally split into three individual groups (low, medium and high) according to reads abundance. Pearson correlation coefficient (PCC) is considered here. (**C**) Relative deviation of raw P-sites/PF P-sites RPKM against iTRAQ intensities on the level of different reads abundance. (**D**) An example of protein SEC31A with its relative raw P-sites, PF P-sites and iTRAQ intensities plotted throughout the time course.

RiboWave improves the detection of differentially translated genes

Ribo-seq is usually accompanied by RNA-seq to estimate TE and identifies genes that are subjected to translational dysregulation. TE is defined by the ratio of Ribo-seq read intensity and RNA-seq read intensity (22,28,29,63,65,76,77). However, the convention way of calculating TE might bring some false discoveries because the signal of raw P-sites contains not only active translational signal but also unspecific binding noises (31,32).

To demonstrate the performance of PF P-sites in assessing differential TE, we used two ribosome profiling datasets under different biological conditions. One is from human PC3 cells in response to mTOR signaling perturbation (19,20); the other is from dark-grown *Arabidopsis* seedling in response to light stimulus (21). Replicates were combined and subjected to the same pre-processing procedure, followed by TE calculation. During the calculation of TE, both datasets used RPKM of either raw P-sites or PF P-sites from Ribo-seq data to estimate the activity of translation before normalization by RNA-seq. We selected differentially translated genes (Z-score > 2 or Z-score < -2) and picked top 200 translationally up/downregulated genes for further studies (28,29,65,76). In general, PF P-sites and raw P-sites detected similar amount of differential translated genes (Supplementary Figure S13). Next, we used GO and KEGG pathway enrichment analyses to identify relevant biological functions and processes.

In PC3 data, top 200 downtranslated genes were selected when mTOR signaling was inhibited. Most enriched biological processes were related to translation process and biosynthesis, which is consistent with the result of mTOR signaling perturbation (20,78) (Figure 5A and Supplementary Figure S14). Within the same biological process, the enrichments from raw P-sites were much lower than that of PF P-sites (Figure 5A and Supplementary Figure S14). Similarly, we also performed the Gene Set Enrichment Analysis (GSEA) on the gene set of translation initiation (GO:0006413) by combining different numbers of top dysregulated translation genes. In agreement with previous result, downtranslated genes identified by PF P-sites were significantly enriched in translation initiation gene set than those from raw P-sites (Figure 5B). We further examined the enrichment at different expression levels and found that the performance of PF P-sites was very robust (Supplementary Figure S15).

Similarly, we performed the same analysis on *Arabidopsis thaliana* where the seedlings were pre-grown under the dark environment (21). Top 200 uptranslated genes were picked as candidates. We found that exposure to light induced enhanced translation of genes related to photosynthesis and chloroplast organization, which is in good agreement with previous reports (21) (Figure 5C and Supplementary Figure S14). Again, PF P-sites achieved better enrichments than raw P-sites (Figure 5C). A GSEA analysis on the gene set of chloroplast organization (GO:0009658) further demonstrated the advantage of PF P-sites (Figure 5D and Supplementary Figure S15).

Taken together, the function enrichment results above suggest that PF P-sites derived by RiboWave is able to de-

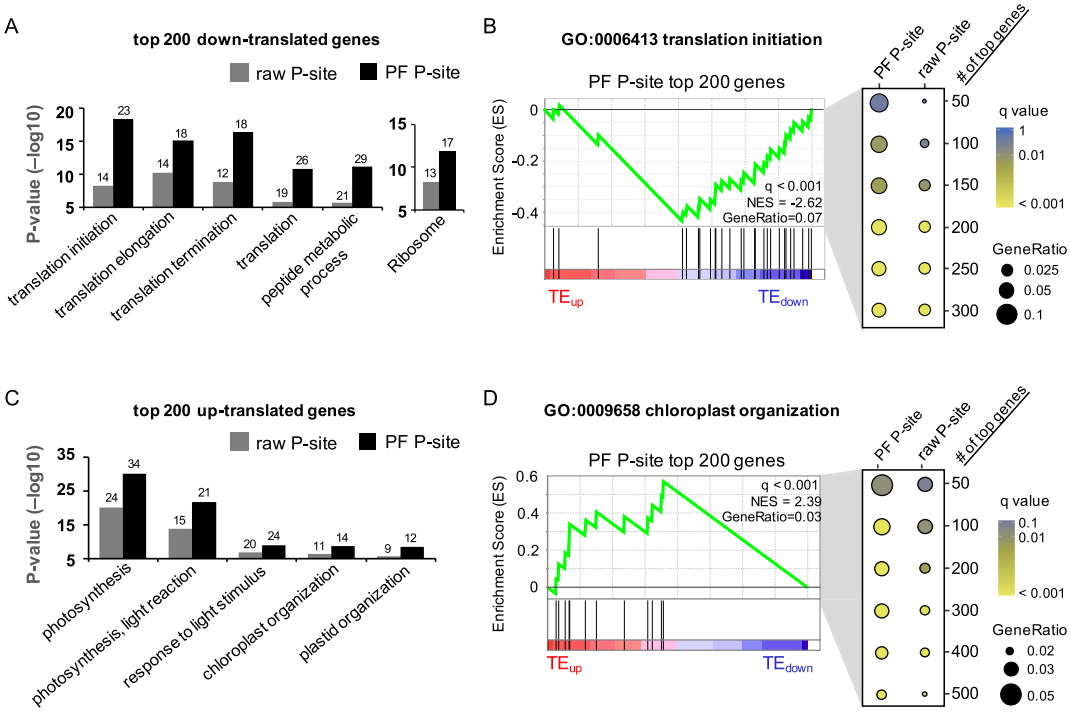

Figure 5. TE calculated from PF P-sites improves the identification of differentially translated genes. (**A**) GO enrichment (BP) of the top 200 downregulated genes identified by the TE based on either raw P-sites or PF P-sites in a PC3 dataset. KEGG pathway enrichment is also indicated in the right. *P*-values of the enrichment are shown on the vertical axes in −log10 scale. The number on top of each bar represents the number of genes, within the downregulated genes, falling in the corresponding BP or pathway. (**B**) GSEA performed on a gene set of translation initiation (GO:0006413). Different numbers of top ranked genes estimated by either raw P-sites or PF P-sites (# of top genes), i.e., up- and downtranslated genes, are merged as the input gene list and pre-ranked by its fold change of TE. *q*-value and ratios of genes overlapped within the input gene list (GeneRatio) are labeled, respectively. (**C**) GO enrichment (BPs) of the top 200 upregulated genes identified in *Arabidopsis* dataset. (**D**) GSEA analysis performed on a gene set of chloroplast organization (GO:0009658).

tect differentially translation events with better biological relevance and higher sensitivity than raw P-sites.

RiboWave identifies ribosomal frameshift event explicitly

Frameshift refers to an alternative regulation of protein translation, which is observed as a shift in translation reading frame (79). Previous method used a periodicity transition score (PTS) (26) to identify such event based on Ribo-seq (12,73). But PTS is not able to capture the whole picture of frameshift by pointing out the shifting site and changed frames accurately and explicitly. Now with the help of RiboWave and PF P-sites, it is possible to trace the movement of actively elongating ribosomes and thus identify the frameshift events explicitly. To do so, we have defined a score to quantify the extent of frameshift during translation (Figure 6A and Supplementary Figure S16). The change of reading frame (CRF) score utilizes two properties of frameshift: (i) *Frame dominancy* requires consistent PF P-sites within one reading frame for both upstream and downstream of the change point; and (ii) *Frame change* requires PF P-sites of upstream and downstream come from different frames (Figure 6A and Supplementary Figure S16; 'Materials and Methods' section). By definition, CRF score would be approximating 1 for frameshift event and *vice versa*. To evaluate the performance of CRF score, we performed a simulation by introducing one nucleotide of indels (insertion/deletion) randomly in the annotated CCDS ORFs' DNA sequences to cause a frameshift in the context of reference genome. Significant difference was observed when comparing the CRF scores before and after the indels (Figure 6A). Conversely, using raw P-sites instead of PF P-sites during the calculation of CRF score showed only limited power for identifying frameshift events (Supplementary Figure S16D).

Application of CRF score on detecting novel frameshift events requires further efforts. As the CRF could be caused by other ambiguities. For example, a high CRF score is expected if the downstream off-frame ORF is also translated (Supplementary Figure S16E). Similarly, if the change point's upstream and downstream signals come from two separate transcripts, in other words, if the P-site track is a chimera, a high CRF score is also expected (Supplementary Figure S16F). Thus, we developed a pipeline to *de novo* identify genuine frameshift events. We started by calculating the CRF score of each change point for all annotated ORFs and selected those change points with high CRF score (Figure 6B). Subsequently, we ruled out the cause of downstream off-frame ORFs translation and chimera P-site tracks (see Methods). Candidates showing significantly unbalanced P-sites coverage before and after the change point were also eliminated (see 'Materials and Methods' section).

Based on the pipeline, we identified five genes with frameshift events, PEG10, GPATCH4, APHGAP35, SAFB2 and DBP in HCT116 cell (Figure 6C and Supplementary Table 2). These five genes were scanned in multiple

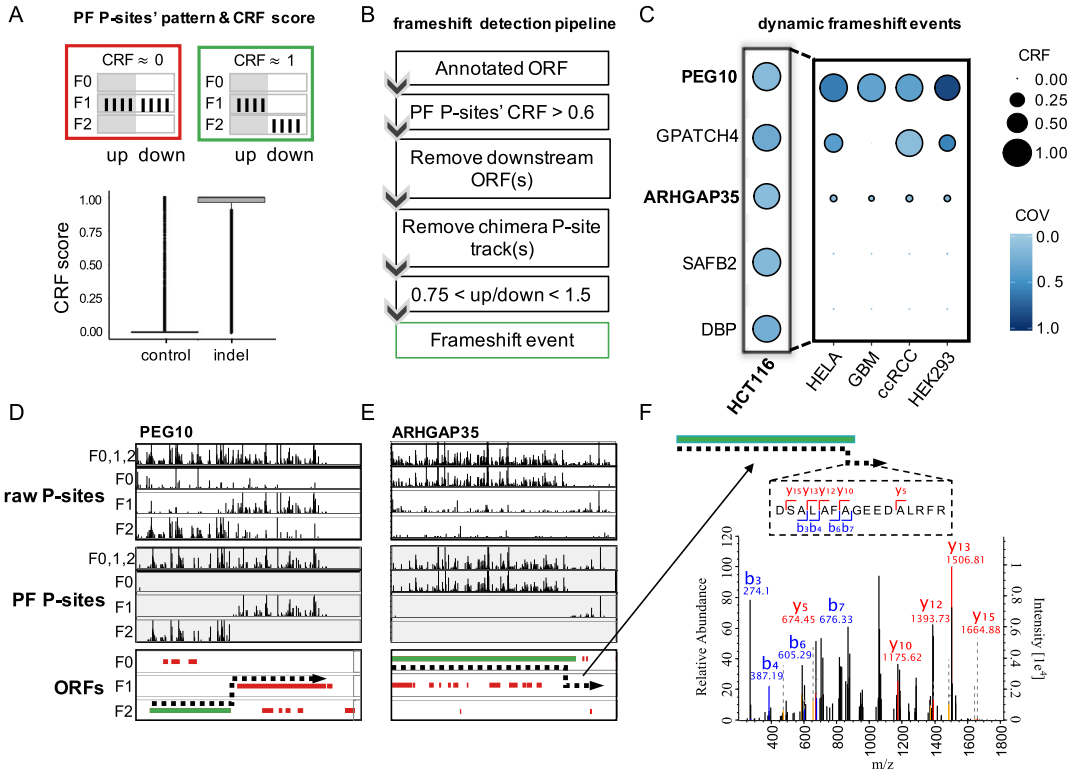

Figure 6. PF P-sites derived from RiboWave explicitly identify translation frameshift events. (**A**) Schematic diagram for CRF score; and its performance on detecting frameshift events simulated by artificial insertions and deletions (indels). (**B**) *De novo* frameshift detecting pipeline based on PF P-sites. We use the CRF score of PF P-sites to define candidate frameshift events. Ambiguous events, such as translated downstream ORFs and chimera tracks, are removed. The coverage of PF P-sites from either upstream and downstream of the change point need to be similar (0.75- to 1.5-fold change) in order to remove ambiguity. (**C**) PF P-site tracks have identified five frameshift events in HCT116. These five genes demonstrate dynamic frameshift potential within different cells. CRF score(CRF) and ORF coverage(COV) are labeled, respectively. GBM: glioblastoma; ccRCC: clear cell renal cell carcinoma. (**D**) A well-known programmed frameshift event in PEG10 detected by PF P-sites. Dashed line indicates the translation track of the ribosomes. (**E**) A cell specific +1 frameshift we found in gene ARHGAP35, which is caused by a genetic deletion. (**F**) MS/MS validation for a peptide fragment, DSALAFAGEEDALRFR, after the frameshift.

tissues/cells to reveal its dynamic frameshift potential under different circumstances (Figure 6C). PEG10 and GPATCH4 displayed consistent high CRF scores in multiple cell lines, while the frameshifts of ARHGAP35, SAFB2 and DBP were predicted to be more cell specific (Figure 6C). PEG10 is reported as a programmed frameshift gene that involves two ORFs, ORF1 and ORF2 (80). Its translation starts in frame 2 until position 1437, where a programmed frameshift event occurs, and hereafter, the subsequent translation proceeds in the frame 1 (80). Compared to raw P-site track where reads could be observed in all three frames, PF P-site track showed a clear signal of frameshift at the right position (Figure 6D). Similarly, the identified frameshift of GPATCH4, which is caused by genetic insertion (annotated by dbSNP (81)), was confirmed by paired MS data (Supplementary Figure S17).

Another example is a cell specific frameshift identified in ARHGAP35, a GTPase-activating protein, that has been designated as a mutational cancer driver in several cancer types (82–85). In HCT116, a cell specific deletion occurs at position 4330 as annotated by COSMIC (86), which gives rise to the change of reading frame. This cell specific frameshift was accurately predicted by PF P-sites (Figure 6E) and validated by paired MS data where a mutated peptide sequence caused by frameshift, DSALAFAGEEDALRFR, was detected (Figure 6F). In summary, the above results suggest that RiboWave is able to explicitly identify frameshift events and predict its dynamics in multiple cell lines.

DISCUSSION

Based on wavelet transform, we have developed RiboWave, a Ribo-seq analysis tool that denoises the original Ribo-seq data into a high-resolution periodic footprint, PF P-sites. This footprint enables us to explore many important and interesting translational regulation events in cells.

Although Ribo-seq contains different source of non-translational noises (3,31–34), few methods are specially designed to denoise the raw signal. We notice that a newly published method, Ribo-TISH (3), also improves the performance of ORF prediction by removing reads with low quality. Ribo-TISH utilizes a set of statistical metrics, like, distribution of RPF (ribosome-protected mRNA fragment) counts across three reading frames, meta-gene profiles of RPF counts to indicate low quality reads. But it does not, as RiboWave does, explicitly reveals the biological processes of active elongation by locating the 3-nt periodicity of RPF. It

is also worth noticing that although we denote these filtered signals from Ribo-seq as noise in the context of translation, the composition and biological meaning of them requires further study. For instance, they may be associated with the processes of noncoding RNAs (31).

Techniques like QTI-seq have been considered as the gold standard when it comes to predicting translation initiation sites. However, in most cases, the treatment of harringtonine and lactimidomycin is not common (87), placing a demand for a computational method that is able to identify translation initiation sites based on standard Ribo-seq data alone. In our study, we have demonstrated high level of consistency with the translation initiation sites defined by QTI-seq. As a result, we think that RiboWave might be a nice option for predicting translation initiation sites and further studying alternative translation initiation especially when there is no QTI-seq data available.

In our study we used MS data as validation in a couple of situations. Although MS is pointed out to be the most widely used technology for systemically quantification of protein abundance, it still faces significant technical challenges, especially in the cases that the amounts of analytes are too low to be detected by canonical MS (88). Therefore, it would be impossible for MS to provide a genome-wide snapshot of all protein levels. Alternatively, our method RiboWave could provide a solution to this problem as we have shown that PF P-site is sensitive and robust in estimating protein abundance. Meanwhile, ribosome profiling is technically easier and economically cheaper than MS-based proteomic profiling.

In this study, we have shown that PF P-site is able to improve the estimation of protein TE. Besides the direct quantification of TE by normalizing Ribo-seq to paired RNA-seq, sophisticated statistical strategies, such as Xtail (19) and Riborex (30), have also been proposed to search for differential translation. In the future, we would incorporate PF P-sites into these statistical models to explore dysregulated translation process.

DATA AVAILABILITY

The software of RiboWave is available at https://lulab.github.io/Ribowave. It can be used to denoise Ribo-seq raw data into PF P-site track, define translated ORFs, estimate reads occupancy as well as TE and identify potential ribosomal frameshift.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank R. Qu, B. Hu for contributing ideas and discussing with us.

Author Contributions: Z.X. and L.H. developed the method and applied it to various cellular translation events. L.X. helped on the wavelet transform models. S.G. performed frameshift validation using MS data. B.S. performed indel identification using RNA-seq data. All authors contributed to the manuscript writing.

FUNDING

National Key Research and Development Plan of China [2016YFA0500803]; National Natural Science Foundation of China [31522030, 31771461]; Fok Ying-Tong Education Foundation; Beijing Advanced Innovation Center for Structural Biology; Bio-Computing Platform of Tsinghua University Branch of China National Center for Protein Sciences (Beijing). Funding for open access charge: National Key Research and Development Plan of China [2016YFA0500803]; National Natural Science Foundation of China [31522030, 31771461]; Fok Ying-Tong Education Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Brar,G.A. (2016) Beyond the triplet code: context cues transform translation. *Cell*, **167**, 1681–1692.
2. Schwanhausser,B., Busse,D., Li,N., Dittmar,G., Schuchhardt,J., Wolf,J., Chen,W. and Selbach,M. (2011) Global quantification of mammalian gene expression control. *Nature*, **473**, 337–342.
3. 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.* (2017) Genome-wide identification and differential analysis of translational initiation. *Nat. Commun.*, **8**, 1749.
4. Munoz,A. and Castellano,M.M. (2012) Regulation of translation initiation under abiotic stress conditions in Plants: Is it a conserved or not so conserved process among eukaryotes? *Comp. Funct. Genome*, **2012**, 406357.
5. Sonenberg,N. and Hinnebusch,A.G. (2009) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, **136**, 731–745.
6. Jackson,R.J., Hellen,C.U.T. and Pestova,T.V. (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Bio.*, **11**, 113–127.
7. Willems,P., Ndah,E., Jonckheere,V., Stael,S., Sticker,A., Martens,L., van Breusegem,F., Gevaert,K. and Van Damme,P. (2017) N-terminal proteomics assisted profiling of the unexplored translation initiation landscape in arabidopsis thaliana. *Mol. Cell. Proteomics*, **16**, 1064–1080.
8. Wan,J. and Qian,S.B. (2014) TISdb: a database for alternative translation initiation in mammalian cells. *Nucleic Acids Res.*, **42**, D845–D850.
9. Spriggs,K.A., Bushell,M. and Willis,A.E. (2010) Translational regulation of gene expression during conditions of cell stress. *Mol. Cell*, **40**, 228–237.
10. Touriol,C., Bornes,S., Bonnal,S., Audigier,S., Prats,H., Prats,A.C. and Vagner,S. (2003) Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. *Biol. Cell*, **95**, 169–178.
11. Claus,P., Doring,F., Gringel,S., Muller-Ostermeyer,F., Fuhlrott,J., Kraft,T. and Grothe,C. (2003) Differential intranuclear localization of fibroblast growth factor-2 isoforms and specific interaction with the survival of motoneuron protein. *J. Biol. Chem.*, **278**, 479–485.
12. Wang,H., Wang,Y. and Xie,Z. (2017) Computational resources for ribosome profiling: from database to Web server and software. *Brief Bioinform.*, doi:10.1093/bib/bbx093.
13. Atkins,J.F., Loughran,G., Bhatt,P.R., Firth,A.E. and Baranov,P.V. (2016) Ribosomal frameshifting and transcriptional slippage: from genetic steganography and cryptography to adventitious use. *Nucleic Acids Res.*, **44**, 7007–7078.
14. Chen,J., Petrov,A., Johansson,M., Tsai,A., O'Leary,S.E. and Puglisi,J.D. (2014) Dynamic pathways of—1 translational frameshifting. *Nature*, **512**, 328–332.
15. Jorgensen,F. and Kurland,C.G. (1990) Processivity errors of gene expression in Escherichia coli. *J. Mol. Biol.*, **215**, 511–521.
16. Cho,C.P., Lin,S.C., Chou,M.Y., Hsu,H.T. and Chang,K.Y. (2013) Regulation of programmed ribosomal frameshifting by co-translational refolding RNA hairpins. *PLoS One*, **8**, e62283.

17. Caliskan,N., Peske,F. and Rodnina,M.V. (2015) Changed in translation: mRNA recoding by—1 programmed ribosomal frameshifting. *Trends Biochem. Sci.*, **40**, 265–274.
18. Ketteler,R. (2012) On programmed ribosomal frameshifting: the alternative proteomes. *Front. Genet.*, **3**, 242.
19. Xiao,Z., Zou,Q., Liu,Y. and Yang,X. (2016) Genome-wide assessment of differential translations with ribosome profiling data. *Nat. Commun.*, **7**, 11194.
20. Hsieh,A.C., Liu,Y., Edlind,M.P., Ingolia,N.T., Janes,M.R., Sher,A., Shi,E.Y., Stumpf,C.R., Christensen,C., Bonham,M.J. *et al.* (2012) The translational landscape of mTOR signalling steers cancer initiation and metastasis. *Nature*, **485**, 55–61.
21. Liu,M.J., Wu,S.H., Wu,J.F., Lin,W.D., Wu,Y.C., Tsai,T.Y., Tsai,H.L. and Wu,S.H. (2013) Translational landscape of photomorphogenic Arabidopsis. *Plant Cell*, **25**, 3699–3710.
22. Ingolia,N.T., Ghaemmaghami,S., Newman,J.R. and Weissman,J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.
23. Ingolia,N.T. (2016) Ribosome footprint profiling of translation throughout the genome. *Cell*, **165**, 22–33.
24. 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.
25. 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.
26. Michel,A.M., Choudhury,K.R., Firth,A.E., Ingolia,N.T., Atkins,J.F. and Baranov,P.V. (2012) Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res.*, **22**, 2219–2229.
27. Zupanic,A., Meplan,C., Grellscheid,S.N., Mathers,J.C., Kirkwood,T.B., Hesketh,J.E. and Shanley,D.P. (2014) Detecting translational regulation by change point analysis of ribosome profiling data sets. *RNA*, **20**, 1507–1518.
28. Su,X., Yu,Y., Zhong,Y., Giannopoulou,E.G., Hu,X., Liu,H., Cross,J.R., Ratsch,G., Rice,C.M. and Ivashkiv,L.B. (2015) Interferon-gamma regulates cellular metabolism and mRNA translation to potentiate macrophage activation. *Nat. Immunol.*, **16**, 838–849.
29. Vu,L.P., Pickering,B.F., Cheng,Y., Zaccara,S., Nguyen,D., Minuesa,G., Chou,T., Chow,A., Saletore,Y., MacKay,M. *et al.* (2017) The N(6)-methyladenosine (m(6)A)-forming enzyme METTL3 controls myeloid differentiation of normal hematopoietic and leukemia cells. *Nat. Med.*, **23**, 1369–1376.
30. Li,W., Wang,W., Uren,P.J., Penalva,L.O.F. and Smith,A.D. (2017) Riborex: fast and flexible identification of differential translation from Ribo-seq data. *Bioinformatics*, **33**, 1735–1737.
31. Ji,Z., Song,R., Huang,H., Regev,A. and Struhl,K. (2016) Transcriptome-scale RNase-footprinting of RNA-protein complexes. *Nat. Biotechnol.*, **34**, 410–413.
32. Andreev,D.E., O'Connor,P.B.F., Loughran,G., Dmitriev,S.E., Baranov,P.V. and Shatsky,I.N. (2017) Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. *Nucleic Acids Res.*, **45**, 513–526.
33. 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**, 165–173.
34. Guttman,M., Russell,P., Ingolia,N.T., Weissman,J.S. and Lander,E.S. (2013) Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell*, **154**, 240–251.
35. Wurmbach,P. and Nierhaus,K.H. (1979) Codon-anticodon interaction at the ribosomal-P (Peptidyl-Transfer Rna) site. *Proc. Natl. Acad. Sci. U.S.A.*, **76**, 2143–2147.
36. Bazzini,A.A., Johnstone,T.G., Christiano,R., Mackowiak,S.D., Obermayer,B., Fleming,E.S., Vejnar,C.E., Lee,M.T., Rajewsky,N., Walther,T.C. *et al.* (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.*, **33**, 981–993.
37. Duncan,C.D. and Mata,J. (2014) The translational landscape of fission-yeast meiosis and sporulation. *Nat. Struct. Mol. Biol.*, **21**, 641–647.
38. Raj,A., Wang,S.H., Shim,H., Harpak,A., Li,Y.I., Engelmann,B., Stephens,M., Gilad,Y. and Pritchard,J.K. (2016) Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife*, **5**, e13328.
39. 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. *Elife*, **4**, e08890.
40. Guo,X., Li,Y.L., Suo,T. and Liang,J. (2017) De-noising of digital image correlation based on stationary wavelet transform. *Opt. Laser Eng.*, **90**, 161–172.
41. Aggarwal,R., Singh,J.K., Gupta,V.K., Rathore,S., Tiwari,M. and Khare,A. (2011) Noise reduction of speech signal using wavelet transform with modified universal threshold. *Int. J. Comput. Appl.*, **20**, 14–19.
42. Ramakrishnan,A.G. and Saha,S. (1997) ECG coding by wavelet-based linear prediction. *IEEE Trans. Biomed. Eng.*, **44**, 1253–1261.
43. Malmurugan,N., Shanmugam,A., Jayaraman,S. and Chander,V.D. (2005) A new and novel image compression algorithm using wavelet footprints. *Acad. Open Internet*, **14**.
44. Akansu,A.N., Serdijn,W.A. and Selesnick,I.W. (2010) Emerging applications of wavelets: a review. *Phys. Commun.*, **3**, 1–18.
45. Sievers,C., Schlumpf,T., Sawarkar,R., Comoglio,F. and Paro,R. (2012) Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. *Nucleic Acids Res.*, **40**, e160.
46. Polikar,R. (1996) The wavelet tutorial, https://cseweb.ucsd.edu/~baden/Doc/wavelets/polikar_wavelets.pdf.
47. Venkatakrishnan,P. and Sangeetha,S. (2014) Singularity detection in human EEG signal using wavelet leaders. *Biomed. Signal. Process*, **13**, 282–294.
48. Kargol,A. (2013) Wavelet-based protocols for ion channel electrophysiology. *BMC Biophys.*, **6**, 3.
49. Ashida,H., Asai,K. and Hamada,M. (2012) Shape-based alignment of genomic landscapes in multi-scale resolution. *Nucleic Acids Res.*, **40**, 6435–6448.
50. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
51. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
52. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.
53. Valens,C. (1999) A really friendly guide to wavelets, https://www.cs.unm.edu/~williams/cs530/arfgtw.pdf.
54. Coifman,R.R. and Wickerhauser,M.V. (1992) Entropy-based algorithms for best basis selection. *IEEE Trans. Inform. Theory*, **38**, 713–718.
55. Michel,M., Yves,M., Georges,O. and Jean-Michel,P. (2009) Wavelet toolbox 4 user's guide. The MathWorks, Inc.
56. 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.
57. Fields,A.P., Rodriguez,E.H., Jovanovic,M., Stern-Ginossar,N., Haas,B.J., Mertins,P., Raychowdhury,R., Hacohen,N., Carr,S.A., Ingolia,N.T. *et al.* (2015) A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Mol. Cell*, **60**, 816–827.
58. Crappe,J., Ndah,E., Koch,A., Steyaert,S., Gawron,D., De Keulenaer,S., De Meester,E., De Meyer,T., Van Criekinge,W., Van Damme,P. *et al.* (2015) PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res.*, **43**, e29.
59. Koch,A., Gawron,D., Steyaert,S., Ndah,E., Crappé,J., De Keulenaer,S., De Meester,E., Ma,M., Shen,B. and Gevaert,K. (2014) A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables

the discovery of alternative translation start sites. *Proteomics*, **14**, 2688–2698.
60. Wiita,A.P., Ziv,E., Wiita,P.J., Urisman,A., Julien,O., Burlingame,A.L., Weissman,J.S. and Wells,J.A. (2013) Global cellular response to chemotherapy-induced apoptosis. *Elife*, **2**, e01236.
61. Juntawong,P., Girke,T., Bazin,J. and Bailey-Serres,J. (2014) Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E203–E212.
62. Merchante,C., Brumos,J., Yun,J., Hu,Q., Spencer,K.R., Enriquez,P., Binder,B.M., Heber,S., Stepanova,A.N. and Alonso,J.M. (2015) Gene-specific translation regulation mediated by the hormone-signaling molecule EIN2. *Cell*, **163**, 684–697.
63. Wang,X., Zhao,B.S., Roundtree,I.A., Lu,Z., Han,D., Ma,H., Weng,X., Chen,K., Shi,H. and He,C. (2015) N(6)-methyladenosine modulates messenger RNA translation efficiency. *Cell*, **161**, 1388–1399.
64. Quackenbush,J. (2002) Microarray data normalization and transformation. *Nat. Genet.*, **32**, 496–501.
65. Rubio,C.A., Weisburd,B., Holderfield,M., Arias,C., Fang,E., DeRisi,J.L. and Fanidi,A. (2014) Transcriptome-wide characterization of the eIF4A signature highlights plasticity in translation regulation. *Genome Biol.*, **15**, 476.
66. Alexa,A., Rahnenfuhrer,J. and Lengauer,T. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.
67. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
68. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
69. Morris,D.R. and Geballe,A.P. (2000) Upstream open reading frames as regulators of mRNA translation. *Mol. Cell. Biol.*, **20**, 8635–8642.
70. Pauli,A., Valen,E. and Schier,A.F. (2015) Identifying (non-)coding RNAs and small peptides: challenges and opportunities. *Bioessays*, **37**, 103–112.
71. Ingolia,N.T., Brar,G.A., Stern-Ginossar,N., Harris,M.S., Talhouarne,G.J., 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.
72. Pauli,A., Norris,M.L., Valen,E., Chew,G.L., Gagnon,J.A., Zimmerman,S., Mitchell,A., Ma,J., Dubrulle,J., Reyon,D. *et al.* (2014) Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science*, **343**, 746–755.
73. Calviello,L. and Ohler,U. (2017) Beyond read-counts: ribo-seq data analysis to understand the functions of the transcriptome. *Trends Genet.*, **33**, 728–744.
74. Cenik,C., Cenik,E.S., Byeon,G.W., Grubert,F., Candille,S.I., Spacek,D., Alsallakh,B., Tilgner,H., Araya,C.L., Tang,H. *et al.* (2015) Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. *Genome Res.*, **25**, 1610–1621.
75. Zur,H., Aviner,R. and Tuller,T. (2016) Complementary post transcriptional regulatory information is detected by PUNCH-P and ribosome profiling. *Sci. Rep.*, **6**, 21635.
76. Wiita,A.P., Ziv,E., Wiita,P.J., Urisman,A., Julien,O., Burlingame,A.L., Weissman,J.S. and Wells,J.A. (2013) Global cellular response to chemotherapy-induced apoptosis. *Elife*, **2**, e01236.
77. Liu,T.Y., Huang,H.H., Wheeler,D., Xu,Y., Wells,J.A., Song,Y.S. and Wiita,A.P. (2017) Time-resolved proteomics extends ribosome Profiling-Based measurements of protein synthesis dynamics. *Cell Syst.*, **4**, 636–644.
78. Reiter,A.K., Anthony,T.G., Anthony,J.C., Jefferson,L.S. and Kimball,S.R. (2004) The mTOR signaling pathway mediates control of ribosomal protein mRNA translation in rat liver. *Int. J. Biochem. Cell B*, **36**, 2169–2179.
79. Craigen,W.J., Cook,R.G., Tate,W.P. and Caskey,C.T. (1985) Bacterial peptide chain release factors: conserved primary structure and possible frameshift regulation of release factor 2. *Proc. Natl. Acad. Sci. U.S.A.*, **82**, 3616–3620.
80. Clark,M.B., Jänicke,M., Gottesbühren,U., Kleffmann,T., Legge,M., Poole,E.S. and Tate,W.P. (2007) Mammalian gene PEG10 expresses two reading frames by high efficiency–1 frameshifting in embryonic-associated tissues. *J. Biol. Chem.*, **282**, 37359–37369.
81. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
82. Gonzalez-Perez,A., Perez-Llamas,C., Deu-Pons,J., Tamborero,D., Schroeder,M.P., Jene-Sanz,A., Santos,A. and Lopez-Bigas,N. (2013) IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods*, **10**, 1081–1082.
83. Liu,Y., Zhang,J., Li,L., Yin,G., Zhang,J., Zheng,S., Cheung,H., Wu,N., Lu,N., Mao,X. *et al.* (2016) Genomic heterogeneity of multiple synchronous lung cancer. *Nat. Commun.*, **7**, 13200.
84. Kondo,T. (2017) Molecular mechanisms involved in gliomagenesis. *Brain Tumor Pathol.*, **34**, 1–7.
85. Rajendran,B.K. and Deng,C.X. (2017) Characterization of potential driver mutations involved in human breast cancer by computational approaches. *Oncotarget*, **8**, 50252–50272.
86. Forbes,S.A., Beare,D., Boutselakis,H., Bamford,S., Bindal,N., Tate,J., Cole,C.G., Ward,S., Dawson,E. and Ponting,L. (2016) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.*, **45**, D777–D783.
87. Xiao,Z., Huang,R., Xing,X., Chen,Y., Deng,H. and Yang,X. (2018) De novo annotation and characterization of the translatome with ribosome profiling data. *Nucleic Acids Res.*, **46**, e61.
88. Du,R., Zhu,L., Gan,J., Wang,Y., Qiao,L. and Liu,B. (2016) Ultrasensitive detection of low-abundance protein biomarkers by mass spectrometry signal amplification assay. *Anal. Chem*, **88**, 6767–6772.