scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Transcriptome and translatome comparison of tissues from *Arabidopsis thaliana*

Isabel Cristina Vélez-Bermúdez[1,4] ✉, Wen-Dar Lin[2], Shu-Jen Chou[3], Ai-Ping Chen[3] & Wolfgang Schmidt[1]

Translation is one of the multiple complementary steps that orchestrates gene activity. In contrast to the straightforwardness of transcriptional surveys, genome-wide profiles of the translational landscape of plant cells remain technically challenging and are thus less well explored. Protein-coding genes are expressed at a variable degree of efficiency, resulting in pronounced discordance among the regulatory levels that govern gene activity. Ribo-Seq is an extremely useful tool for estimating translation efficiency, but the data sets available for plants are limited. Here, we compare inventories of expressed and translated RNA populations, generated by mRNA sequencing (RNA-Seq) and ribosome footprinting (Ribo-Seq) from shoots and roots of *Arabidopsis thaliana* seedlings. Our data set provides information on the translational fitness of protein-coding mRNAs that may aid in obtaining a comprehensive picture of the regulatory levels governing genes activity across the genome.

Background & Summary

Gene expression is intricately regulated by transcription and translation as well as by the degradation of the products of these processes, aligning gene activity to the prevailing conditions. Measuring transcription at essentially saturating resolutions has become attainable and affordable through the advancements in high-throughput technologies such as microarrays and mRNA sequencing (RNA-Seq). Such resolutions are yet a challenge to discovery-based proteomics approaches relying on mass spectrometry, a limitation that is due to the difficulty to detect instable and low-abundance proteins. Besides incomplete coverage of the proteome, a drawback inherent to proteomic surveys is the fact that the steady-state concentration of proteins is not only determined by the translation rate but heavily affected by their stability[1], which may lead to erroneous assumptions of the translation rate of a given transcript. The supposition that proteomic profiles are significantly biased by protein degradation is corroborated by the fact that in both yeast and Arabidopsis changes in transcript abundance correlated well with alterations in protein levels when upregulated genes were considered, while for downregulated genes no such correlation was observed[1–3].

Proteins are major players in the orchestration of biological activity, ultimately governing physiological and developmental processes. Translation is energetically costly and, therefore, sophisticatedly regulated to allow for rapid and plastic decision-making. Prioritization of responses to environmental or internal cues defines phenotypic readouts, making such decisions particularly important for plants, which cannot escape unfavourable conditions. Techniques aimed at interrogating the translatome (i.e., the assemblage of mRNAs associated with ribosomes) such as polysome profiling or ribosome footprinting (Ribo-Seq) in combination with highly parallel sequencing were designed to investigate particular regulatory processes that govern translation, indirectly addressing the enigma of the notoriously low concordance of mRNA and protein expression. The gap between transcript and protein abundance is particularly wide in plants, possibly owing to the necessity of highly plastic responses to environmental signals[4]. While in mammalian systems about 40% of gene expression variation has been attributed to transcriptional control, in plants estimates are rather close to 10%[5], suggesting a large contribution of other factors such as stochastic noise, degradation, and translational regulation. Thus, investigating

[1]Institute of Plant and Microbial Biology, Academia Sinica, Taipei, 11529, Taiwan. [2]Institute of Plant and Microbial Biology, Bioinformatics Core Lab, Academia Sinica, Taipei, 11529, Taiwan. [3]Institute of Plant and Microbial Biology, Genomic Technology Core, Academia Sinica, Taipei, 11529, Taiwan. [4]Present address: Department of Agricultural, Food, Environmental and Animal Sciences, University of Udine, Di4A, Udine, 33100, Italy. ✉e-mail: isabel.velez@uniud.it

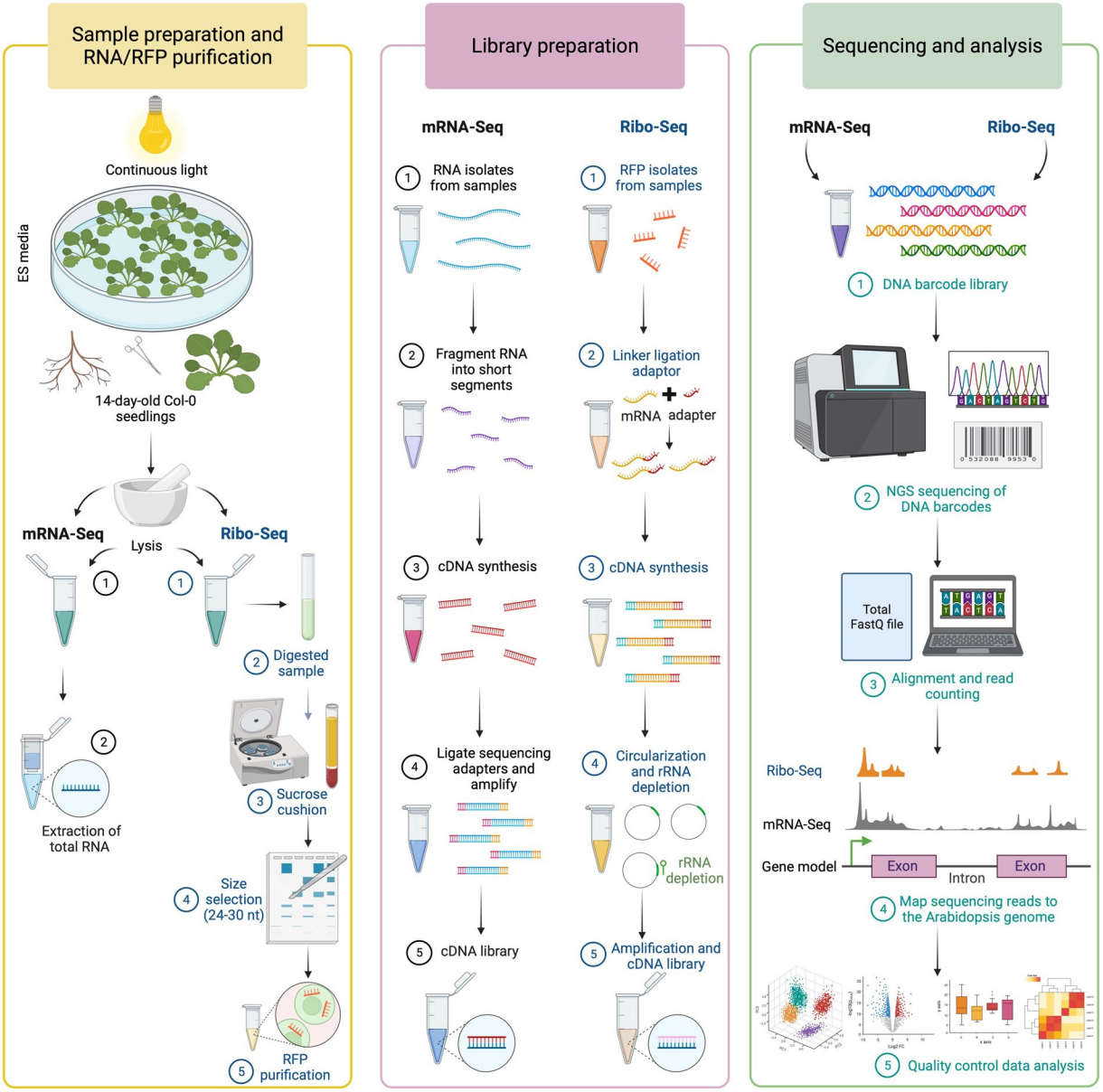

Fig. 1 Overview of the experimental design. The principal steps of the RNA-Seq and Ribo-Seq analyses are depicted in order of their execution. Protocols for Ribo-Seq and RNA-Seq are detailed in the Methods section.

the mechanisms that control protein abundance to understand plant function and development appears to be an obligatory and promising approach.

Interrogations into the *in vivo* translational landscape of plants enable the discovery of the mechanisms underlying translational control, the identification of novel translated short open reading frames (sORFs), previously undiscovered upstream open reading frames (uORFs), non-AUG start codons, the determination of the frame and length of the translated regions, and, consequently, the discovery of novel proteins. Monitoring translation is, however, technically more challenging than measurements of the transcriptome. The invention of the Ribo-Seq methodology, originally developed in the yeast *Saccharomyces cerevisiae*[6], was a big leap forward towards the understanding of translational regulation, but data on what is when (and how efficiently) translated are still scarce, in particular in plants[7]. The incomplete coverage of proteomic approaches renders investigations on translational control of transcripts difficult, even when label-free techniques are employed that allow estimates of absolute protein concentrations. Ribo-Seq provides insights into the density and precise location of translating ribosomes over the entire length of the transcripts, allowing massively parallel sequencing of ribosome-protected footprints (RPFs) after subjecting the remainder of the mRNAs to nucleolytic digestion. Ribo-Seq not only grants access to the inventory of the population of transcripts bounded by ribosomes and thus estimates the 'protein potential' of the cell, but also provides a snapshot of the precise position of the ribosomes at the time when this process (artificially) came to a halt.

The aim of the present study was to provide a comprehensive inventory of Ribo-Seq-generated RPFs form roots and shoots of the reference plant *Arabidopsis thaliana*. The abundance of RPFs is normalized to the steady-state level of mRNA abundance estimated by RNA-Seq, allowing estimates on the translational fitness of transcripts derived from protein-coding genes. The experimental setup is depicted in Fig. 1. Plants were grown on sterile, agar-solidified media for 14 days, dissected into shoots and roots, and immediately frozen in liquid nitrogen. After cell lysis, RNA-Seq samples were generated by extracting total RNA and library construction and subjected to next-generation sequencing. For Ribo-Seq[8], the samples were digested with RNAse I followed by sucrose cushion ultracentrifugation, size selection, ribosome footprint purification, library construction, and sequencing. In a final step, both data sets were analysed and compared to estimate the translation efficiency of the expressed transcripts.

Methods

Plant material, growth conditions, and sample collection. Seeds of the Col-0 accession of *Arabidopsis thaliana* (L.) Heynh were surface sterilized by soaking them in 35% sodium hypochlorite for 5 min, followed by five rinses in sterile water (5 minutes each). The seeds were then placed into sterile, 120 × 120 × 17 mm square Petri dishes containing 100 mL of a growth medium composed of 5 mM KNO_3 (7757-79-1, Merck), 2 mM $MgSO_4 \cdot 7\ H_2O$ (10034-99-8, SIGMA), 2 mM $Ca(NO_3)_2 \cdot 4\ H_2O$ (13477-34-4, Sigma), 2.5 mM KH_2PO_4 (7778-77-0, Merck), 70 µM H_3BO_3 (B0252, Sigma), 14 µM $MnCl_2$ (13446-34-9, Merck), 1 µM $ZnSO_4$ (7446-20-0, Merck), 0.5 µM $CuSO_4$ (7758-98-7, Sigma), 0.01 µM $CoCl_2$ (7646-79-9, Sigma), 0.2 µM Na_2MoO_4 (10102-40-6, Merck), and 40 µM ethylenediaminetetraacetic acid iron(III) sodium salt (NaFe-EDTA; EDFS, Sigma), solidified with 0.4% Gelrite Pure (Supermacy Instrument CO., LTD). Sucrose (1.5% (w/v); Sigma) and 1 g/L MES (M8250, Sigma) were added, and the pH was adjusted to 5.5 with KOH (1310-58-3, Merck). The plates containing the seeds were stratified for 2 days at 4 °C in the dark, transferred to a growth chamber, and grown at 22 °C under continuous illumination (50 mmol m^{-2} s^{-2}; Philips TL lamps) and 70% relative humidity in horizontal position. After 14 days, samples were collected for the experiments. The seedlings were gently removed from the plates, shoots and roots were separated using a razor blade, and the tissues were flash-frozen in liquid nitrogen. Samples were kept at −80 °C until RNA or RPF extraction. A total of thirty-seven shoots and seventy-five independent roots were bulked to form one biological replicate. Shoots and roots were ground and split into aliquots of 100 mg for RNA-Seq and 200 mg for Ribo-Seq analysis.

RNA-Seq. Stranded RNA-Seq analysis was conducted with four biological replicates. Total RNA was extracted from 100 mg of roots or shoots using the RNeasy Plant Mini Kit (Qiagen) following the supplier's instructions. In brief, the plant material was disrupted using a TissueLyser II in 450 µL of RLT buffer containing β-mercaptoethanol per sample, vortexed vigorously and kept on ice. The lysates were transferred to a spin column (Lilac) and centrifuge at 4 °C for 2 minutes at 13,000 rpm. An aliquot (430 µL) of the supernatant of the flow-through was placed into a fresh 1.5 mL Eppendorf tube, 215 µL volume of ethanol (100%) was added, mixed, transferred to a RNeasy Mini spin column (pink) and centrifuged at 4 °C for 15 seconds at 13,000 rpm. The flow-through of the samples was discarded, 700 µL of RW1 buffer was added to each column, and centrifuge at 4 °C for 15 seconds at 13,000 rpm. The flow-through of the samples was discarded, the columns were washed twice with RPE buffer, and centrifuge at 4 °C for 2 minutes at 13,000 rpm each time. The columns were then centrifuged at 4 °C for 2 minutes at 13,000 rpm to remove the remanent ethanol contained in the RPE buffer. The RNeasy spin columns were placed into a fresh 1.5 mL Eppendorf tube, 50 µL of RNase-free water was added to the columns, incubated for 10 minutes on ice, and centrifugated at 4 °C for 2 minutes at 13,000 rpm.

The RNA integrity of the samples was assessed using a Bioanalyzer. Libraries for RNA-seq were prepared using the Illumina TruSeq Stranded mRNA Sample Preparation Kit (RS-122-2103) following the manufacturer's protocol. Four µg of total RNA were used for library construction. PolyA RNA was captured by oligo(dT) beads and fragmented after elution. First-strand cDNA was synthesized by reverse transcriptase (SuperScript III, 18080-093, Invitrogen) using dNTPs and random primers. Second-strand cDNA was generated using a dUTP mix. This double-stranded cDNA then underwent A-tailing at the 3′ end, followed by ligation of barcoded TruSeq adapters. The products were purified and enriched through ten cycles of PCR to create the final double-stranded cDNA library. Final libraries were analyzed using an Agilent High Sensitivity DNA analysis chip (5067-4626, Agilent) to estimate the quantity, checked for size distribution, and were quantified by qPCR using the KAPA Library Quantification Kit (KK4824, KAPA). The prepared library was pooled for single-end sequencing on an Illumina HiSeq 2500 system at YourGene Bioscience Co., New Taipei City, Taiwan, producing 100-bp single-end reads.

Ribo-Seq. Ribosome footprints and library construction for four replicates were performed as described in a detailed step-by-step protocol[8]. Briefly, 200 mg of Arabidopsis roots or shoots were resuspended in 600 µL of lysis buffer (polysome buffer; 20 mM Tris-HCl pH 7.4, 150 mM NaCl, 5 mM MgCl2, 1 mM DTT supplemented with 100 µg/ml cycloheximide, 1% Triton X-100, and Turbo DNase I 25 U/mL), incubated for 15 min on ice during which the samples were vortexed every five minutes. Then, the samples were centrifuged at 16,000 × g for 15 min at 4 °C, and 500 µL of lysate were digested with 12.5 µL RNase I (100 U/µL) for 45 min at room temperature with gentle mixing on a nutator. Subsequently, 16.66 µL of SUPERase*In RNase Inhibitor was added to stop the nuclease digestion. The ribosome pellet was recovered by centrifugation at 70,000 rpm at 4 °C for 4 hours in a TLA 100.4 rotor, and the RPFs were purified using TRIzol® reagent according the manufacturer's instructions. In brief, each ribosomal pellet was resuspended in 1 mL of TRIzol®. After adding 0.2 mL of chloroform, samples were shaken vigorously by hand for 15 sec, incubated for 3 min at room temperature and centrifuged at 12,000 × g for 15 min at 4 °C. Then, the aqueous phase of the samples was very carefully transferred into a fresh 1.5 mL Eppendorf tube,

0.5 mL of 100% isopropanol was added. Samples were vortexed for 5 sec, incubated at −20 °C for 30 min, and centrifuged at 12,000 × g for 10 min at 4 °C. The supernatants were removed and 1 mL of 75% ethanol was added to each sample. Following centrifugation at 7,500 × g for 5 min at 4 °C, the RNA pellets were air dried for 10 min at room temperature, and the individual pellets were resuspended in 20 µL of 10 mM Tris-HCl (pH 8.0). Ten µL of 2X denaturing sample buffer (15 mg bromophenol blue in 1.0 mL of 0.5 M EDTA and add 200 µL to 9.8 mL formamide) was added, and the samples were denatured for 90 sec at 80 °C prior to the electrophoresis. RPFs were then separated on a 15% polyacrylamide TBE-urea gel at 200 V for 65 min and stained with 1X SYBR Gold in 1X TBE running buffer. Gel slices corresponding to 30 nt - 34 nt were then excised in 400 µL RNA gel extraction buffer. Samples were frozen for 1 h at −80 °C and incubated overnight on a nutator at room temperature. Then, the samples were centrifuged at 16,100 × g for 15 min at 4 °C, and 400 µL of each supernatant was transferred into a 1.5 mL non-stick RNase-free micro tube. After addition of 1.5 µg of GlycoBlue the samples were vortexed for 5 sec, 500 µL of isopropanol was added and the samples were again vortexed for 5 sec and kept overnight at −80 °C. The samples were centrifuged at 16,100 × g for 35 min at 4 °C, the supernatants were removed, and the pellets were dried for 10 min at room temperature. Then, the pellets were resuspended in 12 µL of 10 mM Tris-HCl (pH 8.0). The RPF samples were subjected to a quality check using Qubit and Bioanalyzer before proceeding to construct the libraries. Library construction[8] was carried out by dephosphorylation, linker ligation, reverse transcription, circularization and rRNA depletion, PCR amplification, and barcoding of the samples. Four biotinylated oligonucleotides, Oligo 1: 5′/5BioTEG/CATAAACGATGCCGACCAGGGATCAGCGG-3′, Oligo 2: 5′/5BiotinTEG/CTC TGATGATTCATGATAACTCGACGGATCGCATGG-3′, Oligo 3: 5′/5BiotinTEG/CATTAGCATGGGATAACATCAT-3′, and Oligo 4: 5′-/5BiotinTEG/TGCCAAGGATGTTTTCATT AATCAAGAACG-3′ were used to remove contaminating rRNA fragments from the libraries. The libraries were sequenced for 50 bases using an Illumina Hiseq 2500 SR50 system.

Normalization of expression and translational levels. Expression (RNA-Seq) and translational levels (Ribo-Seq) were normalized as RPKM (Reads Per Kilobase of transcript per Million reads mapped) using read counts and feature lengths[9].

Mapping of RNA-seq and ribo-seq reads. RNA-Seq reads were mapped to the TAIR10 genome using BLAT[10] with default parameters. RPKM values were computed using the RackJ toolkit (http://rackj.sourceforge.net/) based on reads mapped with at least 95% identity.

Ribo-Seq reads were first processed by Cutadapt[11] for adaptor removal, the command executed was as follows: cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -m 15 -O 4 -j 16 -o SampleID.fastq.gz. The parameters were set to specify the adapter sequence with -a, set a minimum length requirement for the trimmed reads with -m 15, ensuring a minimum overlap of four nucleotides between read and adapter sequence with -O 4, utilizing 16 processing cores with -j 16 and an output of the trimmed reads to the specified file with -o. To gain alignments of Ribo-Seq cleaned reads as precise as possible, the following prioritized approach was adopted: (i) reads were mapped to the Araport11[12] transcriptome using Bowtie2[13] with parameters "-k 30–gbar 151–sensitive" and only perfect matches were accepted. (ii) For splicing junctions supported by RNAseq data, reads were mapped to junction-spanning fragments using Bowtie2 with the same parameters and only perfect matches were accepted. (iii) The two steps described above were repeated by accepting alignments with an identity of at least 95% and exact matches at the two ends. (iv) Reads were mapped to the Araport11 genome using BLAT[14] with default parameters and accepted alignments with an identity of at least 95%, no alignment blocks shorter than 8 bps, and exact matches at the two ends. (v) The two steps described above were repeated with BLAT parameters "-minMatch = 1 -minScore = 14 -tileSize = 9" for a higher sensitivity. Note that steps (iii)~(v) requested matches at both sequence ends. Step (vi) was introduced to repeat the steps described above without requesting exact matches at the two ends. In step (ii), junctions were inferred based on RNA-Seq data, confirmed by a minimum of ten reads each with at least four different mapping start positions. Thereby, alignments of Ribo-Seq reads spanning junctions unknown to the Araport11 database were recovered.

Multi-reads with two or more best alignments were subjected to further filtering. Covering regions were identified and the corresponding multi-read counts were computed. As a result, five regions covered by more than 95% of multi-reads were identified, and reads with alignments in these regions were removed from the computation, as well as reads mapping to non-coding genes. The alignments produced by this procedure were used for RPKM computation[9] (See 'Code availability').

Principal component analysis (PCA) analysis. The PCA plot[9] was generated using the R software package[15]. For PCA, the input matrix was filled with TMM (Trimmed Mean of M values)[16]-normalized RPKM values with rows for samples and columns for genes, where genes with constant values were removed. The R function prcomp was applied for PCA with the input matrix and parameter 'scale = TRUE'. Excel was used for the visualization of the principal components.

Pearson correlation calculations. Pearson correlation analysis was based on log-RPKM and performed for every four biological replicates from roots and shoots (RNA-Seq and Ribo-Seq) using script RPKMcorrelation.pl from the RackJ toolkit (See 'Code availability').

Hierarchical clustering. Heatmaps and hierarchical clustering of detected genes derived from Ribo-Seq and RNA-Seq were generated using the Next-Generation Clustered Heat Map Viewer[17] (HG-CHM). For building the heatmaps, we input matrixes of RPKM values into the NG-CHM web interface separately for the Ribo-Seq and RNA-Seq datasets. In both cases, the following pre-processing steps were performed within the NG-CHM Transform Matrix page: (i) log2 transformation to RPKM values, (ii) replacement of all invalid values with N/A,

Sample	Number of raw reads	Number of reads uniquely mapped to the genome	Number of reads multi-mapped to genes
RNA-Seq Roots R1	44,798,444	42,317,504	758,944
RNA-Seq Roots R2	45,622,719	42,963,830	730,198
RNA-Seq Roots R3	48,045,644	45,467,118	677,274
RNA-Seq Roots R4	40,316,088	38,222,726	570,717
RNA-Seq Shoots R1	44,650,350	42,462,624	570,111
RNA-Seq Shoots R2	42,707,942	40,479,344	559,643
RNA-Seq Shoots R3	39,429,288	37,452,617	495,109
RNA-Seq Shoots R4	43,249,322	41,054,352	507,391
Ribo-Seq Roots R1	58,082,938	17,134,771	36,708,935
Ribo-Seq Roots R2	67,623,514	21,207,755	43,451,374
Ribo-Seq Roots R3	53,329,668	12,330,618	37,789,284
Ribo-Seq Roots R4	56,012,462	17,921,517	34,627,507
Ribo-Seq Shoots R1	59,738,467	8,041,415	48,086,813
Ribo-Seq Shoots R2	53,278,422	7,090,468	43,400,598
Ribo-Seq Shoots R3	61,826,359	9,782,578	46,095,325
Ribo-Seq Shoots R4	64,248,867	5,686,319	55,810,228

Table 1. Read numbers of the RNA-Seq and Ribo-Seq data sets[20]. reads that mapped uniquely to the genome.

(iii) removal of all rows with at least one N/A value, and (iv) keeping 4,500 rows (genes) with the highest standard deviations. In the NG-CHM Cluster Matrix page, the ordering options for rows and columns were both set to 'Hierarchical Clustering' with default options of Euclidean distance metric and Ward agglomeration to obtain the hierarchical clustering results of rows (genes) and columns (samples).

Calculation of length distribution and nucleotide periodicity of RPFs. RPF length distribution and 3-nt periodicity were determined using a number of Perl one-liner commands and Excel (See 'Code availability'). Three-nt periodicity was estimated by counting the base on the 13th bp positions of the Ribo-Seq reads[9].

Comparison of the number of genes detected by mass spectrometry, Ribo-seq, and RNA-seq. The comparison among the total detected genes (RPKM > 0) in roots and shoots derived from RNA-Seq, Ribo-Seq, and tier1-canonical proteins[18] data sets was made using a Venn diagram online tool (https://bioinformatics.psb.ugent.be/webtools/Venn/)[19].

Data Records

Raw reads FASTQ files, trimmed reads FASTQ files, and mapped read BAM files (mapped to Araport11) derived from the RNA-Seq and Ribo-Seq surveys of roots and shoots of *Arabidopsis thaliana* were deposited at the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) (BioProject accession PRJNA990964)[20]. Detailed information on the analysis of RNA-Seq and Ribo-Seq data has been deposited in Figshare[9].

Technical Validation

Experimental design. *A. thaliana* (Col-0) seedlings were grown under controlled conditions for 14 days. In order to perform the RNA-Seq and Ribo-Seq experiments, we separately collected shoot and root samples. The plant material was collected without wounding or damaging the plant tissue; undamaged tissue is crucial to obtaining high quality data in particular for Ribo-Seq surveys. For RNA-Seq experiments, we bulked twelve shoots and twenty-five roots; Ribo-Seq required twenty-five shoots and fifty independent roots per replicate. All samples were arranged in four biological replicates. An overview of the experimental design and explanation of RNA-Seq and Ribo-Seq steps is depicted in Fig. 1.

Quality control of the RNA-seq and ribo-seq data. RNA-Seq and Ribo-Seq libraries were generated in four biological replicates and sequenced using the Illumina platform (Table 1). RNA-Seq libraries yielded 44,798,444 (Roots R1), 45,622,719 (Roots R2), 48,045,644 (Roots R3), and 40,316,088 (Roots R4) reads from mRNA extracted from roots, and 44,650,350 (Shoots R1), 42,707,942 (Shoots R2), 39,429,288 (Shoots R3), and 43,249,322 (Shoots R4) reads from mRNA extracted from shoots[20]. For Ribo-Seq, the number of yielded reads were 58,082,938 (Roots R1), 67,623,514 (Roots R2), 53,329,668 (Roots R3), 56,012,462 (Roots R4), 59,738,467 (Shoots R1), 53,278,422 (Shoots R2), 61,826,359 (Shoots R3), and 64,248,867 (Shoots R4)[20]. For RNA-Seq samples, reads that mapped uniquely to the genome were 42,317,504 (Roots R1), 42,963,830 (Roots R2), 45,467,118 (Roots R3), 38,222,726 (Roots R4), 42,462,624 (Shoots R1), 40,479,344 (Shoots R2), 37,452,617 (Shoots R3), and 41,054,352 (Shoots R4), with a length of the trimmed reads of 101 nt[20]. For Ribo-Seq, only reads that directly mapped to more than 95% to the Araport11 genome annotation were considered; reads with alignment blocks shorter than 8 bp were removed. For protein-coding genes, this procedure contributed to less than 1% of the unique Ribo-Seq reads. For Ribo-Seq, the reads mapped uniquely to genome were 17,134,771 (Roots R1), 21,207,755 (Roots R2), 12,330,618 (Roots R3), 17,921,517 (Roots R4), 8,041,415 (Shoots R1), 7,090,468 (Shoots

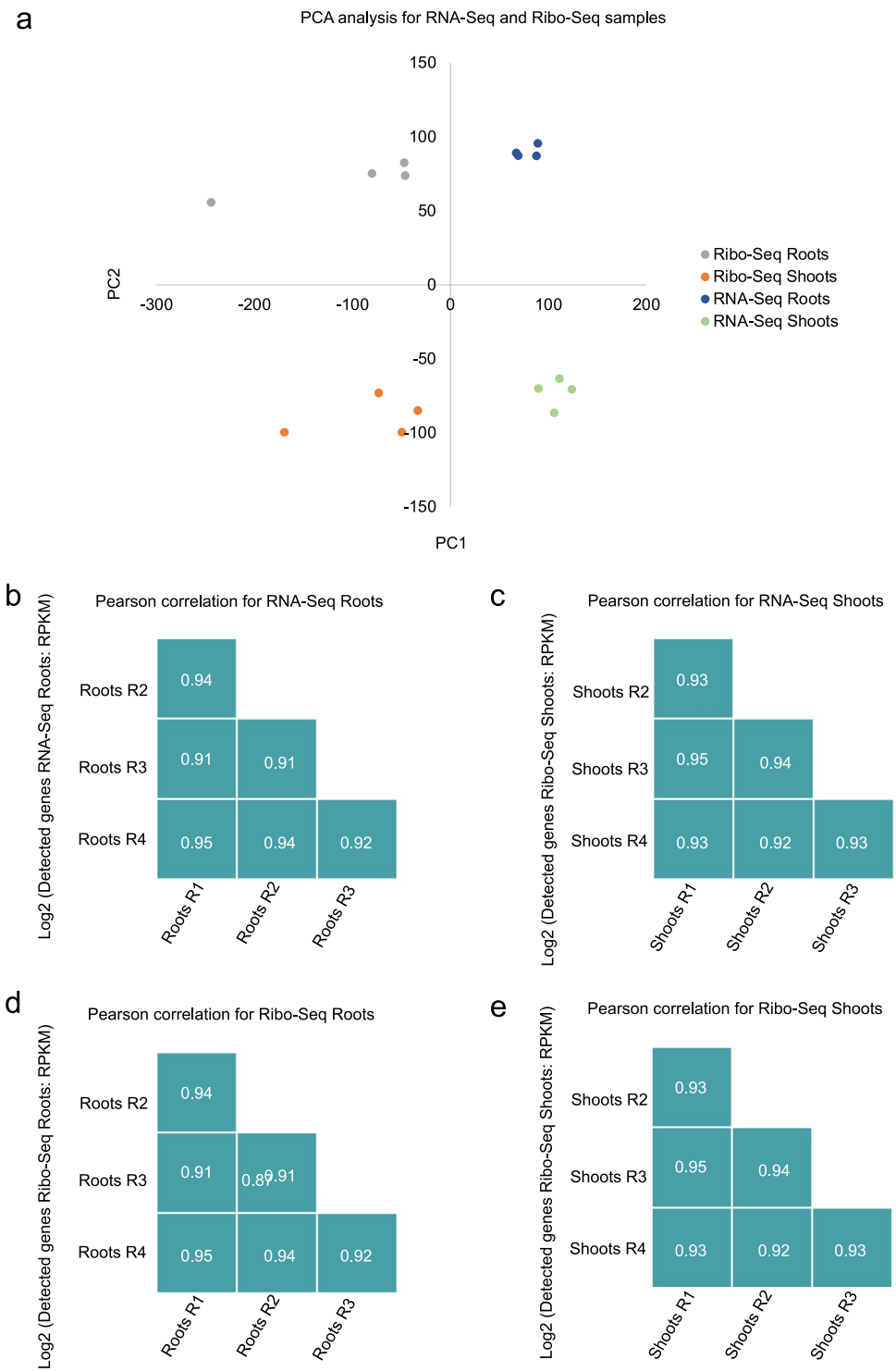

Fig. 2 Technical validation. (**a**) Principal-component analysis (PCA) of the RNA-Seq and Ribo-Seq data derived from roots and shoots. The analysis demonstrates high repeatability of both surveys. The samples were categorised by the individual biological replicates of each tissue for each technique. (**b,c**) Pearson correlation coefficient values of the RNA-Seq (**b**) and Ribo-Seq samples (**c**).

R2), 9,782,578 (Shoots R3), and 5,686,319 (Shoots R4)[20]. It thus appears that the Ribo-Seq libraries exhibited sufficient reads mapping to the genome, with an average length of filtered reads between 28–34 nt. However, due to the shorter length of the reads, the number of uniquely mapped reads was about 2-fold lower for Ribo-Seq-derived reads than those derived from RNA-Seq (Table 1).

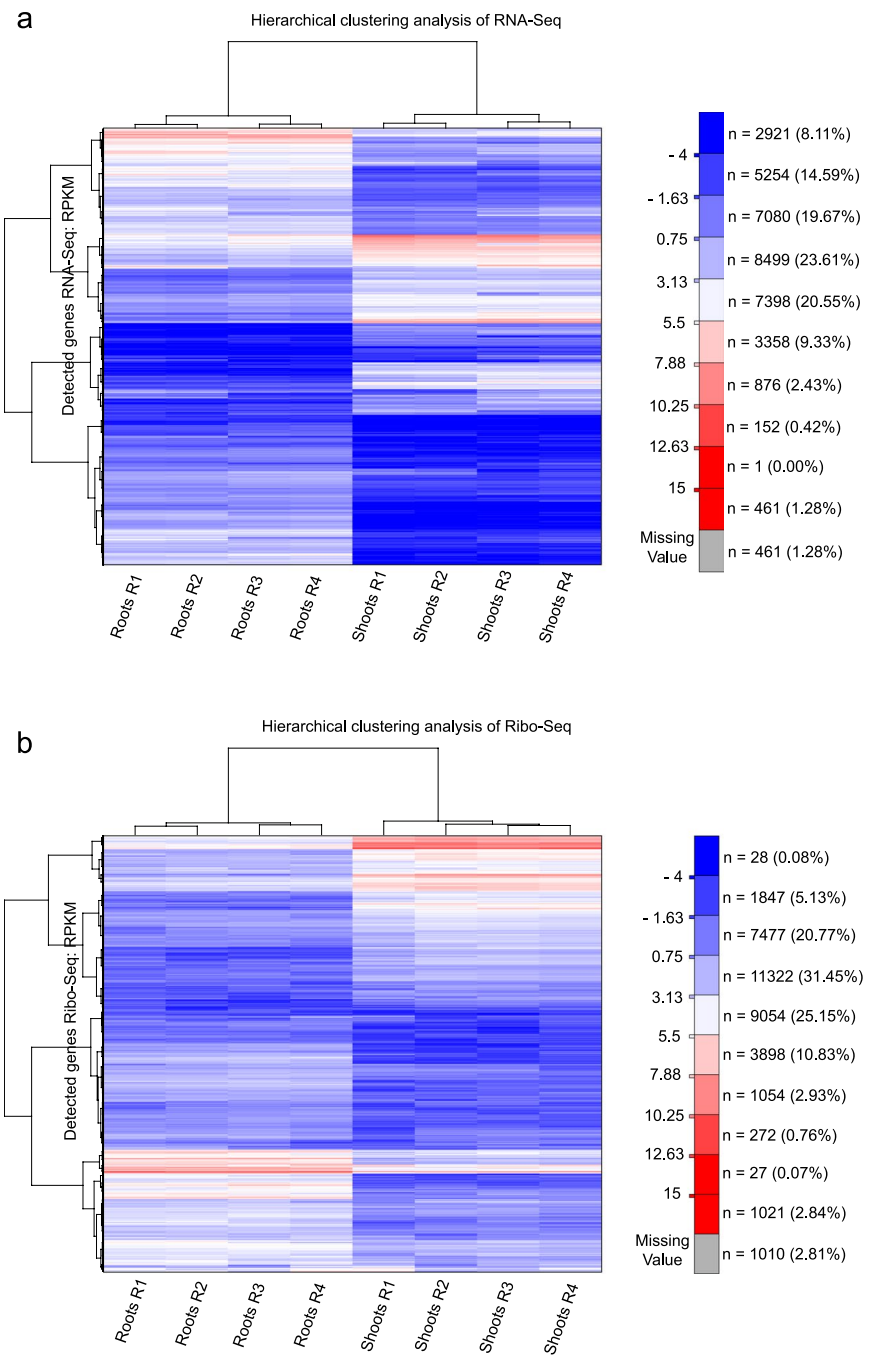

Fig. 3 Quality validation of the data. (**a**) Hierarchical clustering of RNA-Seq samples from roots and shoots. (**b**) Hierarchical clustering of Ribo-Seq samples from roots and shoots. Note that the values in the heatmaps were all log2 transformed. The colour map for a and b is indicated as follows: threshold 1: -4; blue; threshold 2: 6; white, and threshold 3: 15; red.

Reproducibility of the RNA-seq and ribo-seq data. To evaluate the quality of the data, we performed a Principal Component Analysis (PCA) for the data derived from the RNA-Seq (total mRNA) and Ribo-Seq (RPFs) profiling experiments, using individual biological replicates for each tissue under study[9] (Fig. 2a). The PCA plot shows low bias associated with the biological replicates for both RNA-Seq and Ribo-Seq, supporting the robustness of the data. The results reveal that the variability of the Ribo-Seq replicates was somewhat higher when compared to the high reproducibility of the RNA-Seq data, an observation that is possibly associated with the more complex protocol of the former methodology.

Normalized reads derived from the RNA-Seq and Ribo-Seq surveys were used to define detected genes in the transcriptome and translatome[9]. The reproducibility of both experiments was confirmed by high Pearson correlation between sample pairs[9] (Fig. 2b,c) and hierarchical clustering of the data, which corroborated the robustness of both methods[9] (Fig. 3a,b). To further evaluate the quality of the Ribo-Seq data, we analysed the

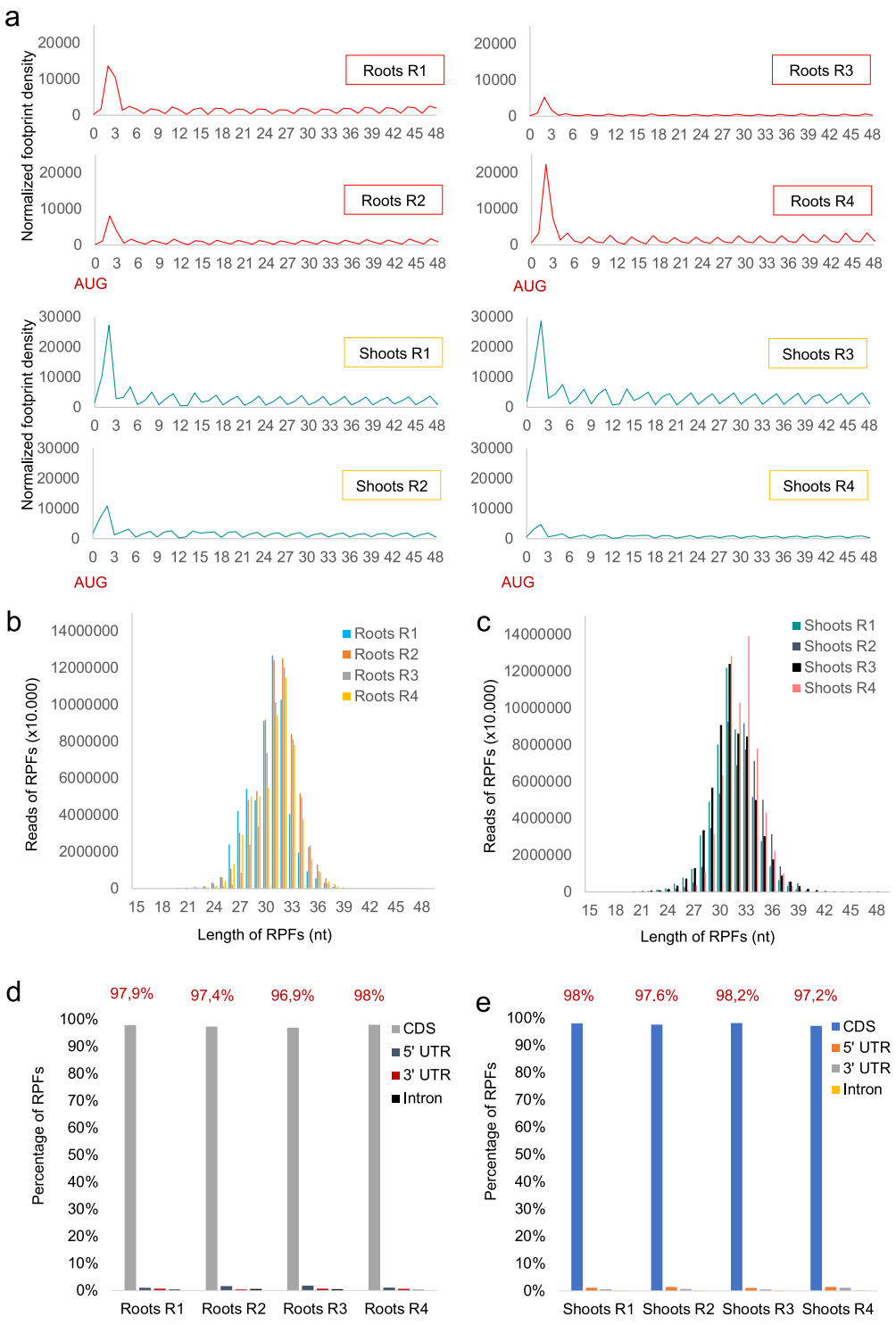

Fig. 4 Reproducibility of the data. (**a**) Three-nucleotide periodicity of the Ribo-Seq data. The position of the ribosomes along the transcripts was computed based on the 13th base pair for individual biological replicates of roots (upper panel) and shoots (lower panel). (**b,c**) Length distribution of RPFs from roots (**b**) and shoots (**c**) showing a peak at 30 nt. (**d,e**) Percentage of RPFs mapped to CDS, 5′ UTR, 3′ UTR, and introns in samples from roots (**d**) and shoots (**e**).

3-nucleotide (nt) periodicity—a phenomenon that allows predicting open reading frames—using individual biological replicates for roots and shoots[9] (Fig. 4a). The triplet periodicity is one of the most important features of Ribo-Seq data derived from protein-coding genes. Such periodicity is usually not observed in non-coding regions. We computed and aggregated read depths of the first 50 bps from all representative coding models, where read depths were computed based on the 13th base pair of every mapped Ribo-Seq read. In each sample,

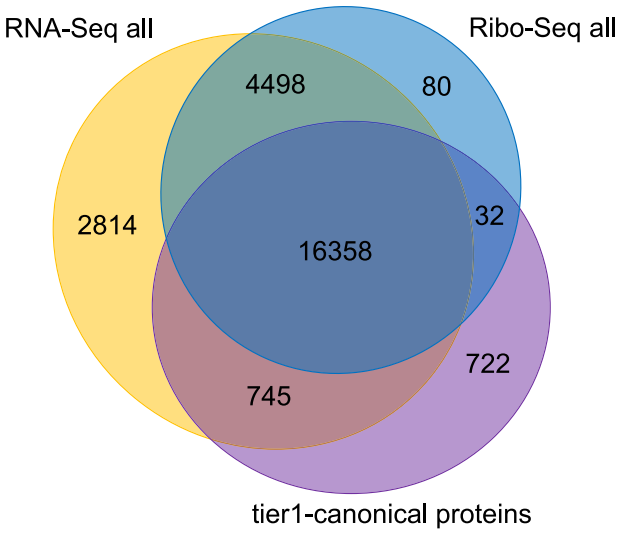

Fig. 5 Comparison of the genes detected by RNA-Seq and Ribo-Seq with expressed proteins. The proteomic data set was derived from a previous study[18].

the plots show a pronounced 3-nt periodicity (Fig. 4a). In addition, the first codon showed stronger signals than the other codons, a further feature that is typically observed in Ribo-Seq reads.

RPF length distribution is another feature that we measured to determinate the quality of the libraries. We found that the vast majority of RPFs from each biological replicate showed read lengths ranging from about 26 nt to 34 nt with a peak at 30 nt (Fig. 4b,c), a distribution that has been observed in other studies using Arabidopsis tissue[21]. When RPFs were mapped to the Arabidopsis genome[9] (Fig. 4d,e), 97.9%, 97,4%, 96,9%, and 98% of the reads from the root samples R1-R4 localised to annotated CDS, the remaining RPFs were located in the 5′ UTR, 3′ UTR, or in intronic regions. For shoot samples, 98%, 97,6%, 98,2%, and 97.2% of the reads were located in the CDS, while the remainder of RPFs were aligned to 5′ UTR, 3′ UTR, or intronic regions.

Robustness of the ribo-seq data. A multi-omics comparation using the total list of detected genes (roots and shoots) in the RNA-Seq and Ribo-Seq data sets against a previously published inventory of the Arabidopsis proteome[17] that comprised proteins with at least two uniquely mapping non-nested peptides showed that a large number of genes (16,358) were found in all three data sets (Fig. 5). A subset of 7,312 genes detected by RNA-Seq was not covered by the proteomic data set. For Ribo-Seq, a subset of 4,578 genes was not comprised in the expressed proteome. A subset of 4,498 genes was comprised in the RNA-Seq and Ribo-Seq data but not in the proteome, suggesting that these genes are translated to lowly abundant or unstable proteins.

Code availability

The detailed procedure and code used for mapping Ribo-seq reads and Pearson correlations analysis are available at https://github.com/wdlingit/cop/wiki/Riboseq-multi-stage-mapping-and-filtering. The code used to calculate the length distribution and nucleotide periodicity of RPFs is available at https://github.com/wdlingit/cop/wiki/Ribo-seq-3bp-periodicity-preference.

Received: 20 May 2024; Accepted: 12 March 2025;
Published online: 25 March 2025

References

1. Lackner, D. H., Schmidt, M. W., Wu, S., Wolf, D. A. & Bähler, J. Regulation of transcriptome, translation, and proteome in response to environmental stress in fission yeast. *Genome Biology* **13**, R25, https://doi.org/10.1186/gb-2012-134-r25 (2012).
2. Lee, M. V. *et al.* A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Molecular Systems Biology* **7**, 514, https://doi.org/10.1038/msb.2011.48 (2011).
3. Lan, P., Li, W. & Schmidt, W. Complementary proteome and transcriptome profiling in phosphate-deficient Arabidopsis roots reveals multiple levels of gene regulation. *Molecular & Cellular Proteomics* **11**, 1156–1166, https://doi.org/10.1074/mcp.M112.020461 (2012).
4. Vélez-Bermúdez, I. C. & Schmidt, W. The conundrum of discordant protein and mRNA expression. Are plants special? *Frontiers in Plant Science* **5** https://doi.org/10.3389/fpls.2014.00619 (2014).
5. Pan, I. C. *et al.* Post-transcriptional coordination of the Arabidopsis iron deficiency response is partially dependent on the E3 ligases RING DOMAIN LIGASE1 (RGLG1) and RING DOMAIN LIGASE2 (RGLG2). *Molecular & Cellular Proteomics* **14**, 2733–2752, https://doi.org/10.1074/mcp.M115.048520 (2015).
6. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223, https://doi.org/10.1126/science.1168978 (2009).
7. Wu, H. L., Jen, J. & Hsu, P. Y. What, where, and how: Regulation of translation and the translational landscape in plants. *Plant Cell* **36**(5), 1540–1564, https://doi.org/10.1093/plcell/koad197 (2023).
8. Vélez-Bermúdez, I. C., Chou, S.-J., Chen, A.-P., Lin, W.-D. & Schmidt, W. Protocol to measure ribosome density along mRNA transcripts of *Arabidopsis thaliana* tissues using Ribo-seq. *Star Protocols* **4**(3), 102520 (2023).

9. Vélez-Bermúdez, I. C., Lin, W.-D., Chou, S.-J., Chen, A.-P. & Schmidt, W. Transcriptome and translatome comparison of tissues from *Arabidopsis thaliana*. *figshare* https://doi.org/10.6084/m9.figshare.23936259.v1 (2023).
10. Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome Research* **12**, 656–664, https://doi.org/10.1101/gr.229202 (2002).
11. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal.* **17**, 10–12, https://doi.org/10.14806/ej.17.1.200 (2011).
12. Cheng, C.-Y. *et al.* Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant Journal* **89**, 789–804, https://doi.org/10.1111/tpj.13415 (2017).
13. Langmead, B. & Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359, https://doi.org/10.1038/nmeth.1923 (2012).
14. Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome research* **12**(4), 656–664, https://doi.org/10.1101/gr.229202 (2002).
15. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (2021).
16. Robinson, M. D. & Oshlack A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**, R25, https://doi.org/10.1186/gb-2010-11-3-r25 (2010).
17. Ryan, M. C. *et al.* Interactive Clustered Heat Map Builder: An easy web-based tool for creating sophisticated clustered heat maps. *F1000Research* **8**, ISCB Comm J–1750, https://doi.org/10.12688/f1000research.20590.2 (2019).
18. van Wijk, K. J. *et al.* The Arabidopsis PeptideAtlas: Harnessing worldwide proteomics data to create a comprehensive community proteomics resource. *Plant Cell* **33**, 3421–3453, https://doi.org/10.1093/plcell/koab211 (2021).
19. VIB / UGent. Bioinformatics & Evolutionary Genomics. https://bioinformatics.psb.ugent.be/webtools/Venn/
20. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP447248 (2023).
21. Mahboubi, A., Delhomme, N., Häggström, S. & Hanson, J. Small-scale sequencing enables quality assessment of Ribo-Seq data: an example from Arabidopsis cell culture. *Plant Methods* **17**, 92, https://doi.org/10.1186/s13007-021-00791-w (2021).

Acknowledgements

This work was supported by grants from Academia Sinica (AS-TP-109-L01) and the National Science and Technology Council (111-2313-B-001-01) to W.S. Figure 1 was created with BioRender.com.

Author contributions

W.S. and I.C.V.-B. designed and conceived the experiments. I.C.V.-B., S.-J.C. and A.-P.C. conducted the experiments. W.-D.L. performed the computational analysis. W.S., I.C.V.-B. and W.-D.L., S.-J.C. and A.-P.C. analysed the results. I.C.V.-B. and W.S. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to I.C.V.-B.

Reprints and permissions information is available at www.nature.com/reprints.

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