Super-resolution ribosome profiling reveals unannotated translation events in *Arabidopsis*

Polly Yingshan Hsu[a], Lorenzo Calviello[b,c], Hsin-Yen Larry Wu[d,1], Fay-Wei Li[a,e,f,1], Carl J. Rothfels[e,f], Uwe Ohler[b,c], and Philip N. Benfey[a,g,2]

[a]Department of Biology, Duke University, Durham, NC 27708; [b]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, 13125 Berlin, Germany; [c]Department of Biology, Humboldt Universität zu Berlin, 10099 Berlin, Germany; [d]Bioinformatics Research Center and Department of Statistics, North Carolina State University, Raleigh, NC 27695; [e]University Herbarium, University of California, Berkeley, CA 94720; [f]Department of Integrative Biology, University of California, Berkeley, CA 94720; and [g]Howard Hughes Medical Institute, Duke University, Durham, NC 27708

Contributed by Philip N. Benfey, September 13, 2016 (sent for review June 30, 2016; reviewed by Pam J. Green and Albrecht G. von Arnim)

Deep sequencing of ribosome footprints (ribosome profiling) maps and quantifies mRNA translation. Because ribosomes decode mRNA every 3 nt, the periodic property of ribosome footprints could be used to identify novel translated ORFs. However, due to the limited resolution of existing methods, the 3-nt periodicity is observed mostly in a global analysis, but not in individual transcripts. Here, we report a protocol applied to *Arabidopsis* that maps over 90% of the footprints to the main reading frame and thus offers super-resolution profiles for individual transcripts to precisely define translated regions. The resulting data not only support many annotated and predicted noncanonical translation events but also uncover small ORFs in annotated noncoding RNAs and pseudogenes. A substantial number of these unannotated ORFs are evolutionarily conserved, and some produce stable proteins. Thus, our study provides a valuable resource for plant genomics and an efficient optimization strategy for ribosome profiling in other organisms.

translation | ribosome footprint | Ribo-seq | ncRNA | sORF

Identifying translated open reading frames (ORFs) is important to understanding the activity of organisms under specific conditions. Until recently, genome-wide mapping of translation has relied primarily on polysome profiling (1). This involves isolation and separation of polysome-associated mRNA through differential centrifugation and fractionation. Actively translated transcripts in the polysome fraction can be identified and quantified using microarrays or RNA sequencing (RNA-seq). However, quantification of these transcripts may not accurately estimate translation levels as the number of ribosomes bound to RNA can vary greatly. In addition, although polysome profiling reveals the identity of translated transcripts, it does not report the translated region of the transcript. These limitations have been overcome by ribosome profiling (2).

Ribosome profiling combines ribosome footprints with deep sequencing (2, 3). After isolating polysomes, the sample is treated with ribonuclease to digest unprotected parts of the RNA. The resulting ribosome-protected RNA fragments (or ribosome footprints) are used to generate a sequencing library (Ribo-seq) (Fig. 1*A*). Sequencing the ribosome footprints reveals the positions and number of ribosomes on a given transcript. When combined with RNA-seq generated from the same starting material, one can accurately determine the average number of ribosomes per mRNA and thus estimate the relative translation levels of a transcript (2). Furthermore, localization of the exact positions of ribosome footprints in the transcriptome provides the opportunity to experimentally identify translated ORFs genome-wide under a specific environment (4–8).

The challenge, however, is to identify real translation events. For example, ribosomes can stall on a specific region of the transcript without translation occurring (9, 10). Also, contaminant RNAs that are highly structured or embedded in ribonucleoprotein complexes [e.g., rRNA, tRNA, and small nucleolar RNA (snoRNA)] are also present in Ribo-seq reads as they resist RNase digestion (7, 11, 12). Therefore, additional features are required to distinguish translation from mere ribosome occupancy and contaminants. Several metrics associated with translation have been exploited (11), for example, the following: (*i*) ribosomes release after encountering a stop codon (9), (*ii*) local enrichment of footprints within the predicted ORF (4, 13), (*iii*) ribosome footprint length distribution (7), and (*iv*) 3-nt periodicity displayed by translating ribosomes (2, 6, 10, 14, 15). Among these features, some work well in distinguishing groups of coding vs. noncoding RNAs, but are insufficient to identify individual transcripts as coding or to define translated regions on a transcript (11, 16). In contrast, 3-nt periodicity is a unique property that allows one to directly define translated regions. It is not observed in RNA-seq data (2, 17). Furthermore, computational pipelines have been developed to identify translated ORFs by interrogating 3-nt periodicity specifically, including "ORF score" (a summary statistic that tests if one particular reading frame is enriched in Ribo-seq by comparing to a uniform distribution) (6) and "RiboTaper" (a spectrum analysis that determines whether footprints on a transcript display 3-nt periodicity) (15). There are additional pipelines that include 3-nt periodicity as part of the analysis (18, 19). The 3-nt periodicity of Ribo-seq has been leveraged to identify novel small ORFs in zebrafish embryos and mouse/human cells (6, 15, 18, 19).

The remaining difficulty is to obtain high-precision ribosome footprints of individual transcripts in the organism of interest.

Significance

Translation is the process by which ribosomes decode information in RNA to produce proteins. The resulting proteins constitute cellular structures and regulate diverse functions in all organisms. Translation also affects mRNA stability. As the final step of the central dogma, translation can alter protein production more rapidly than transcription in a changing environment. However, a robust experimental method to define the landscape of the translatome has not been established in many organisms. We developed an advanced experimental approach and used it to discover proteins missed in the annotation of the *Arabidopsis* genome. This study confirmed computationally predicted noncanonical translation events and uncovered unannotated small proteins that likely have important functions in plants.

Author contributions: P.Y.H., H.L.W., and P.N.B. designed research; P.Y.H. performed research; L.C., H.L.W., and U.O. contributed new reagents/analytic tools; P.Y.H., L.C., H.L.W., F.W.L., C.J.R., U.O., and P.N.B. analyzed data; and P.Y.H. and P.N.B. wrote the paper.

Reviewers: P.J.G., Delaware Biotechnology Institute; and A.G.v.A., University of Tennessee.

The authors declare no conflict of interest.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession nos. GSE81295 and GSE81332). All of the alignments and a script to calculate pairwise sequence identities have been deposited in the Dryad Digital Repository (dx.doi.org/10.5061/dryad.m8jr7).

[1]H.L.W. and F.W.L. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: philip.benfey@duke.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1614788113/-/DCSupplemental.

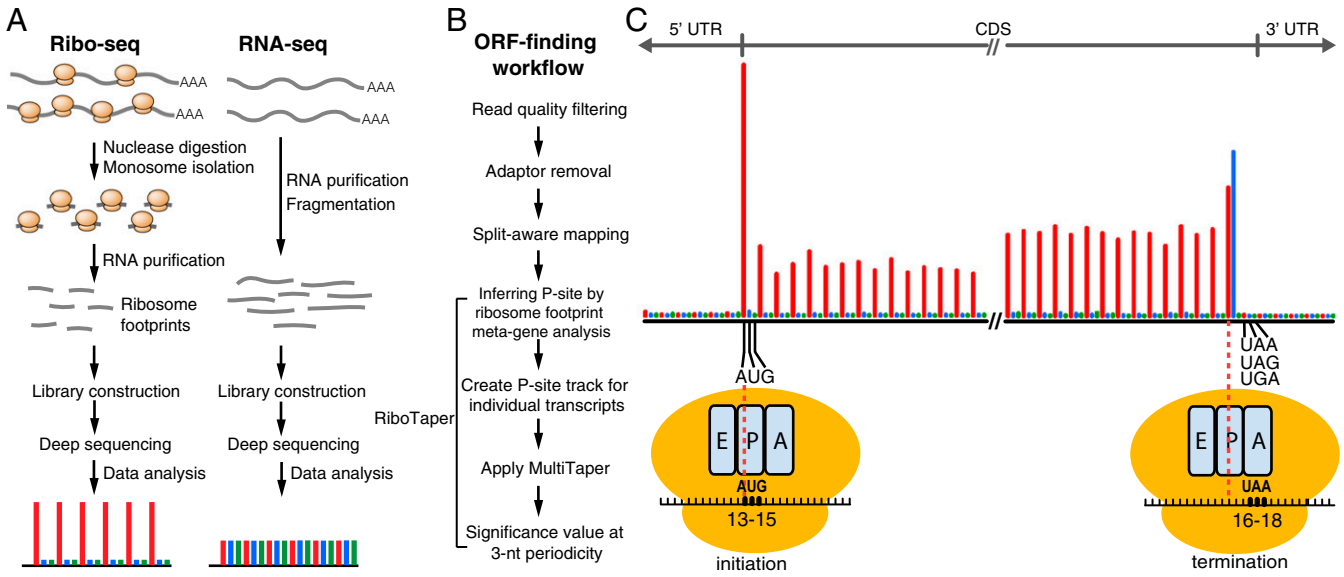

Fig. 1. Identifying translated ORFs using ribosome-profiling data. (*A*) Experimental workflow of ribosome profiling and the expected read distribution among the reading frames. (*B*) Data analysis workflow for ORF finding using RiboTaper. (*C*) Our 28-nt ribosome footprints in the *Arabidopsis* root mapped to the annotated protein-coding genes in TAIR10. Results of other footprint length are shown in Fig. S4*A*. The inferred footprint positions related to the initiating and terminating ribosomes are shown. The A site (the entry point for the aminoacyl-tRNA), P site (where peptide-bond formation occurs), and E site (the exit site of the uncharged tRNA) within ribosomes are shown. A region of 63 nt near the start and stop codon is shown. The position of a ribosome footprint is indicated by its 13th nucleotide within each footprint. Three reading frames are shown in red (the main frame according to the annotated start codon), blue, and green. Most of the footprints are mapped within the CDS and show enrichment for the main reading frame. Footprints at the translation initiation and termination revealed that the ribosomal P site is located between the 13th and 15th nucleotides, whereas the A site is located between 16th and 18th nucleotides.

Ribosome footprints can show a strong 3-nt periodicity in a global analysis, but signals in individual transcripts are often too noisy to assess periodicity (2, 13). When footprints are out of frame, noise increases and resolution decreases. Thus, to a first approximation, the resolution of Ribo-seq data can be quantified by the fraction of reads in the major reading frame. Studies in several organisms, including *Chlamydomonas*, yeast, zebrafish, and rat, have achieved remarkable resolution with over 80% of the reads mapped to one reading frame (6, 10, 20, 21). In contrast, some organisms such as *Escherichia coli*, *Drosophila*, and plants have very limited resolution to date (22–28). Here, we report optimization of a ribosome profiling protocol and its use in *Arabidopsis*. The resulting data provide super-resolution for ribosome footprints, which enables efficient identification of translated ORFs based on the 3-nt periodicity. Our data not only support many annotated and predicted noncanonical translation events but also uncover evolutionarily conserved novel small ORFs that likely encode functionally important proteins.

Results

Buffer Optimization Greatly Improves Footprint Precision. The resolution of Ribo-seq data can be judged by the 3-nt periodicity that emerges from the analysis. A survey of the literature revealed that published *Arabidopsis* ribosome-profiling methods do not generate optimal 3-nt periodicity (25–27). These protocols use extraction buffers with relatively high ionic strength and buffering capacity, originally designed for polysome isolation (Table S1). Unlike polysome isolation, which emphasizes mRNA integrity, precise ribosome footprints require complete digestion of the unprotected mRNA. We reasoned that the high ionic strength and buffering capacity in the polysome buffer might inhibit the RNase used in ribosome footprinting. To test this hypothesis, we extracted polysomes from *Arabidopsis* using four buffers with varying ionic strength and buffering capacity and examined the resulting polysome profiles to evaluate endogenous RNase activity (Fig. S1 *A* and *B*). We observed similar polysome profiles among samples extracted from the first three buffers (buffers A, B, and C) and a slight increase of monosome-to-polysome ratio when ionic strength decreased in buffers B and C. On the other hand, a clear increase of monosome to polysome ratio was found with buffer D, indicating that the endogenous RNase was most active in this buffer. After adding RNase to polysome extracts to obtain ribosome footprints, we constructed and sequenced eight libraries made from root and shoot samples prepared with the four different buffers. We found that the size distribution of ribosome footprints from buffer A was clearly different and slightly longer than those prepared from the other three buffers (Fig. S1*C*). By quantifying reading frame preference in the most abundant footprints (28 nt long), we observed increased reading frame enrichment as ionic strength/buffering capacity decreased in the four buffers (Fig. S1*D*). This is consistent with previous reports that ionic strength affects ribosome footprint size and enrichment of footprints in the primary reading frame (3, 5). Thus, buffer composition strongly affects footprint precision. However, the same tissues prepared with the four buffers yielded highly correlated footprint counts on individual coding sequences (CDSs) ($r = 0.98–1$; Fig. S1*E*), suggesting that the changes in buffer composition did not affect measurement of ribosome occupancy on mRNAs. Because buffer D yielded the best 3-nt periodicity, we used this buffer for our subsequent experiments.

Optimized Ribosome Profiling Compares Favorably to Published Datasets. We performed ribosome profiling on three biological replicates of root and shoot tissues from *Arabidopsis* seedlings. A strong 3-nt periodicity (Fig. 1*C*) and an excellent correlation across replicates ($r = 0.99–1$, Fig. S2) suggested our protocol was robust. Our method also used fewer starting materials, simpler procedures, and had a shorter preparation time compared with published methods in *Arabidopsis* (Table S1).

To obtain high coverage, we pooled the three replicates of the same tissue for analysis. Compared with previously published *Arabidopsis* ribosome profiling data [see *SI Materials and Methods* for details of individual datasets; Juntawong et al. (26); Liu et al. (25); Merchante et al. (27)], our protocol yields the narrowest

footprint size distribution (Fig. 2*A*), yet still covers expected genomic features of the transcriptome with 96–98% of the footprints mapped to CDSs, and very few footprints mapped to UTRs, introns, or intergenic regions (Fig. S3). As it has been observed that not all footprint sizes display similar 3-nt periodicity (6, 13, 15), we examined the periodicity of footprints with different lengths (Fig. S4 *A–E* for individual datasets; Fig. S5 for summary). Among footprints with a length between 20 and 35 nt, we observed that the 28-nt footprints have the highest in-frame percentage compared with other footprint lengths in our datasets, as well as in the datasets of Liu et al. and Merchante et al. (Fig. S5). In comparisons of 28-nt footprints, our data contained a superior enrichment of footprints in one reading frame, with 96% and 92% of reads in the main reading frame in root and shoot, respectively (Fig. 2*B*).

A ribosome footprint meta-gene analysis, which combines all footprints that map to annotated protein-coding genes (Fig. S4 *A–E*), allows us to infer the corresponding P site (where peptide-bond formation occurs in the ribosomes) within the footprints (2, 13, 15). We assigned the location of footprints according to the first nucleotide within the footprint. By examining footprints near the start codon ("A" in AUG is defined as 0) for 28-nt footprints, it is apparent that footprints cover up to the 12th nucleotide upstream of the AUG (Fig. S4*A*). This indicates that the codon being translated at the P site (in this case, AUG) is located between the 13th and 15th nucleotide within a 28-nt footprint (Fig. 1*C*). Consistent with the start codon position, at translation termination where the A site encounters a stop codon, we observed the last in-frame footprints cover the 15th nucleotide upstream of the stop codon (Fig. S4*A*). This indicates that the A site is located between the 16th and 18th nucleotide within a 28-nt footprint, which is 3 nt downstream of the P-site position inferred above (Fig. 1*C*). Despite some differences in different datasets, the start codon located between the 13th and 15th nucleotide for 28-nt footprints is also observed in the datasets of Liu et al. and Merchante et al. (Fig. S4 *D* and *E*). Furthermore, in our data, we observed that footprints are preferentially digested at the 5′ end when the footprint size is below 28 nt. For example, compared with the 28-nt footprints, which have strong signals up to the upstream 12th nucleotide, the 27-nt footprints have clear signals up to the upstream 11th nucleotide, and 26-nt footprints have signals up to the 10th nucleotide and so on (Fig. S4 *A* and *B*: metaplots; Fig. S4*F*: schematic summary). Because many of the footprints in our data display a robust 3-nt periodicity (Figs. S4 *A* and *B* and S5), we can infer the P-site position for each of these footprint lengths, which is essential for downstream workflow for ORF identification using RiboTaper (15). Overall, our protocol significantly improves the 3-nt periodicity compared with previously published *Arabidopsis* datasets.

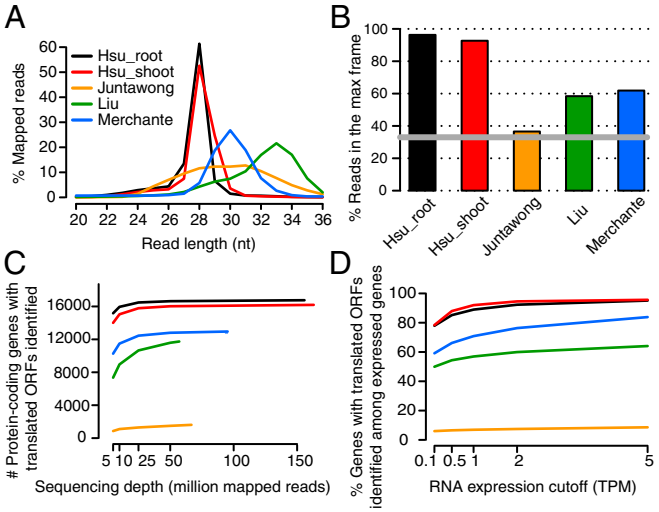

Fig. 2. Comparison between the current study and published *Arabidopsis* ribosome-profiling datasets. (*A*) Length distribution of ribosome footprints in the current study (Hsu_root and Hsu_shoot), compared with three other published datasets in *Arabidopsis* (25–27). See *SI Materials and Methods* for details of the growth conditions for each dataset. Size of footprints isolated in each dataset is compared in Table S1. (*B*) Percentage of Ribo-seq reads in the max reading frame. Data were extracted from the meta-gene analysis using 28-nt footprints in which most of the datasets display the best 3-nt periodicity. The gray line marks 33%, which is the percentage of reads expected if there is no enrichment in any frame. (*C*) Number of protein-coding genes with translated ORFs identified by RiboTaper with different sequencing depths. (*D*) Percentage of protein-coding genes with translated ORFs identified among the expressed protein-coding genes defined by different RNA expression cutoffs. A subset of each dataset (25 million reads) was compared across the studies.

Enhancement of 3-nt Periodicity Improves Identification of Translated ORFs. To identify translated ORFs by taking advantage of the enhanced 3-nt periodicity in our data and to investigate how the periodicity affects ORF identification, we adapted a recently developed pipeline, RiboTaper (15), to *Arabidopsis*. RiboTaper uses the multitaper method (29) to determine the significance of 3-nt periodicity in the P-site signals along an ORF. This approach proved to be effective in detecting active translation from Ribo-seq data, yielding a high-confidence set of translated ORFs in the transcriptome (15). By analyzing the meta-gene plots, we inferred the P-site position of each footprint size in different datasets (Fig. S4 *A–E*) and then created P-site tracks for individual transcripts (Fig. 1*B*). Defining the P-site position for individual footprint lengths based on the meta-gene analysis rather than assigning one presumed position for all footprint lengths improved ORF identification (Fig. S6).

Across different datasets, we observed that deeper sequencing depths result in the identification of more translated ORFs, but once above 50 million mapped reads, the number of ORFs increased only slightly (Fig. 2*C*). In addition, compared with the same sequencing depth, datasets with a better periodicity yield a higher number of identified ORFs (Fig. 2 *B* and *C*). For instance, there are over 16,000 ORFs detected in either our root or shoot data, which is substantially higher than in any other dataset (Fig. 2*C*). It is possible that our datasets have more identified ORFs simply because there are more genes expressed under our experimental conditions. To test this hypothesis, we examined the fraction of ORFs found among the expressed protein-coding genes, defined by varying RNA expression cutoffs across all datasets under the same sequencing depth (Fig. 2*D*). We observed that, under all RNA expression thresholds, our datasets have the highest percentage of ORFs identified among the expressed genes, thus ruling out the possibility that our samples have more identified ORFs due to more expressed genes.

Interestingly, we noticed that the higher the RNA expression levels, the higher the fraction of ORFs found among the expressed genes, suggesting that a transcript with higher expression levels is more likely to have ORFs identified by this method (Fig. 2*D*). However, with lower expression cutoffs, more genes are considered expressed and a higher number of ORFs are found among them, despite the lower fraction (Fig. S7). For example, using transcripts per million (TPM) > 0.1 as RNA expression cutoff, 21,848 protein-coding genes are considered expressed in either the root or shoot, and among them, 18,148 genes have ORFs identified. This results in a fraction of 83% (18,148/21,848), which is considerably lower than the fraction using TPM > 5 (12,714/13,219 = 96%). Nevertheless, the large number of ORFs identified among the expressed genes in our datasets demonstrates that our approach is robust across different RNA expression levels.

RiboTaper determines de novo the start codon of an ORF by examining the in-frame precision of the P-site positions between

candidate AUGs (15). Therefore, an ORF could be identified with a shorter length, that is, truncated at the 5′ end. Although it is possible that a transcript uses a downstream AUG start site rather than the annotated one, the truncation could result from insufficient sequencing coverage or poor periodicity of a given transcript. We therefore examined the ORF length reported by RiboTaper compared with the annotated ORFs across different datasets (Fig. S8). As the sequencing depth increases, we found that datasets with a better periodicity identify ORFs with a higher coverage of the annotated ORF length as seen in our dataset as well as in that of Merchante et al. However, in the datasets with less optimal periodicity, although the number of identified ORFs increases (Fig. 2*C*), the average coverage of annotated ORF length decreases (Fig. S8). Overall, datasets with better periodicity yield higher coverage of the annotated ORF length. Whether the truncated forms of ORFs represent translation events initiated from a downstream AUG remains unclear.

Taken together, our datasets with enhanced 3-nt periodicity correlate with a larger number of ORFs identified, a higher sensitivity to identify ORFs among the expressed transcripts, and an improved ORF length coverage compared with other datasets.

Super-resolution Profiles Can Be Used to Annotate Individual Transcripts. By interrogating the genes annotated in The *Arabidopsis* Information Resource (TAIR10) (30), we found that over 18,000 genes have translated ORFs identified in our data, including a large number of annotated protein-coding genes (18,153 genes), as well as a small set of noncoding RNAs (ncRNAs) (27 genes), pseudogenes (37 genes), and transposable elements (57 genes) (Table 1: summary of ORFs identified; Dataset S1 *A* and *B*: all ORFs identified by RiboTaper in root and shoot). Among the protein-coding genes, in addition to ORFs identified within the annotated CDSs, 187 upstream ORFs (uORFs) were identified within 5′-UTRs and 10 downstream ORFs (dORFs) were found in the 3′-UTRs (Table 1). In contrast to the annotated protein-coding sequences (CDS ORFs) that have a wide range of ORF length, most of the unannotated ORFs (except from transposable elements) have a relatively small length (Fig. S9 *A* and *B*): with uORFs being the smallest. Most of the ORFs identified have a high fraction of P sites mapped to the main reading frame (Fig. S9 *C*–*F*). Thus, by taking advantage of the enhanced 3-nt periodicity, we can use ribosome profiling to identify translated ORFs efficiently.

The strong 3-nt periodicity in our data not only allows efficient identification of ORFs by a statistical method but also provides super-resolution translational profiles of individual transcripts across a wide range of expression levels and ORF lengths (Figs. 3–5). Unlike a well-characterized ncRNA, *HIDDEN TREASURE 1* (*HID1*) (31), for which the P sites do not show a clear 3-nt periodicity along the transcript (Fig. 3*B*), transcripts with translated ORFs have most of the P sites mapped to the main reading frame within the predicted CDSs. This is not only apparent for long and highly expressed transcripts such as *TUBULIN 4* (*TUB4*) (Fig. 3*A*) but also for short and lowly expressed genes, such as *GOLVEN 6* (*GLV6*; also known as *ROOT MERISTEM GROWTH FACTOR 8* or *CLE-LIKE2*) (Fig. 3*C*) (32–34).

To evaluate how sensitive our approach is, we examined our ORF-finding results for known secreted peptide genes and their homologs, which usually have short ORFs and relatively low expression levels. Of the 34 expressed peptide genes with a TPM value greater than 1, we identified translated ORFs in 31 (Dataset S1*C*: summary; Dataset S1*D*: a list of known peptide genes with an ORF identified in the root and shoot). We also confirmed translation of two small peptide genes (*AT4G28460* and *AT4G34600*) that were recently identified through a comparative genomics study combining 32 plant genomes (35) (Dataset S1 *C* and *D*). These results indicate that our improved ribosome profiling combined with the RiboTaper pipeline is able to find small ORFs even in genes with low expression levels.

Ribosome Profiling Supports Noncanonical Translation Events. Previously, several uORFs that encode conserved peptide sequences (CPuORFs) were found to regulate their downstream main ORFs (36, 37). There are 89 CPuORFs predicted in *Arabidopsis*, but only a small number of them have been validated and characterized (36–40). Among the predicted CPuORF genes expressed in our data, there are 39 CPuORFs identified by RiboTaper (Dataset S1*E*: summary, Dataset S1*F*: a list of CPuORFs identified in the root and shoot). For genes possessing multiple CPuORFs such as *SUPPRESSOR OF ACAULIS 51* (*SAC51*) (41), RiboTaper successfully identified all three of the predicted CPuORFs (*CPuORF38*, *39*, and *40*). In addition to CPuORFs, we identified an additional 148 unannotated translated uORFs. Similar to *SAC51*, which has multiple uORFs in the 5′-UTR, we found an extra uORF upstream of *CPuORF51* in the *AT3G53670* gene (Fig. 4*A*).

By manually inspecting the uORFs, we found that the new uORF identified in *AT5G17460* is actually an ORF from an unannotated gene overlapping with the 5′-UTR of *AT5G17460* (Fig. 4*B*). This unannotated gene is also supported by the EST data (Fig. S10) (42) and is evolutionarily conserved (see below). Therefore, ribosome-profiling data can fine-tune and improve genome annotation.

Although RiboTaper only searches for "AUG" as the start codon, we wanted to see whether our data can validate predicted ORFs that use a non-AUG start, such as a "CUG" codon (43). Among the predicted genes, *AT3G10985* is highly expressed in our root samples. By visualizing its ribosome profile, we found that, in addition to the annotated CDS, many P sites map to the 5′-UTR in frame with a predicted ORF that starts with a CUG codon (Fig. 3*D*). In addition, we confirmed a uORF initiated with a non-AUG codon in *GDP-L-GALACTOSE PHOSPHORYLASE* (*GGP*, *AT4G26850*) in the shoot (Fig. 4*C*). This uORF initiates at 14 aa upstream of previously reported "ACG" start (44) in our data. These examples demonstrate our super-resolution ribosome-profiling data can provide direct experimental support for non-canonical translation events.

Table 1. TAIR10 genes with translated ORFs identified by RiboTaper

	No. of genes with translated ORFs identified					
	Protein-coding genes			Other genes*		
Sample	uORF†	Annotated ORF	dORF	ncRNA	Pseudogene	Transposable elements
Root	136	16,657	2	23	27	31
Shoot	87	16,107	8	14	14	40
Total	187	18,153	10	27	37	57

There are 27,416 protein-coding genes, 394 ncRNAs, 924 pseudogenes, and 3,903 transposable element genes annotated in TAIR10.

*Excluding rRNA, tRNA, and snoRNA genes.

†CPuORFs are annotated as protein-coding genes in TAIR10 and were manually grouped into uORFs here.

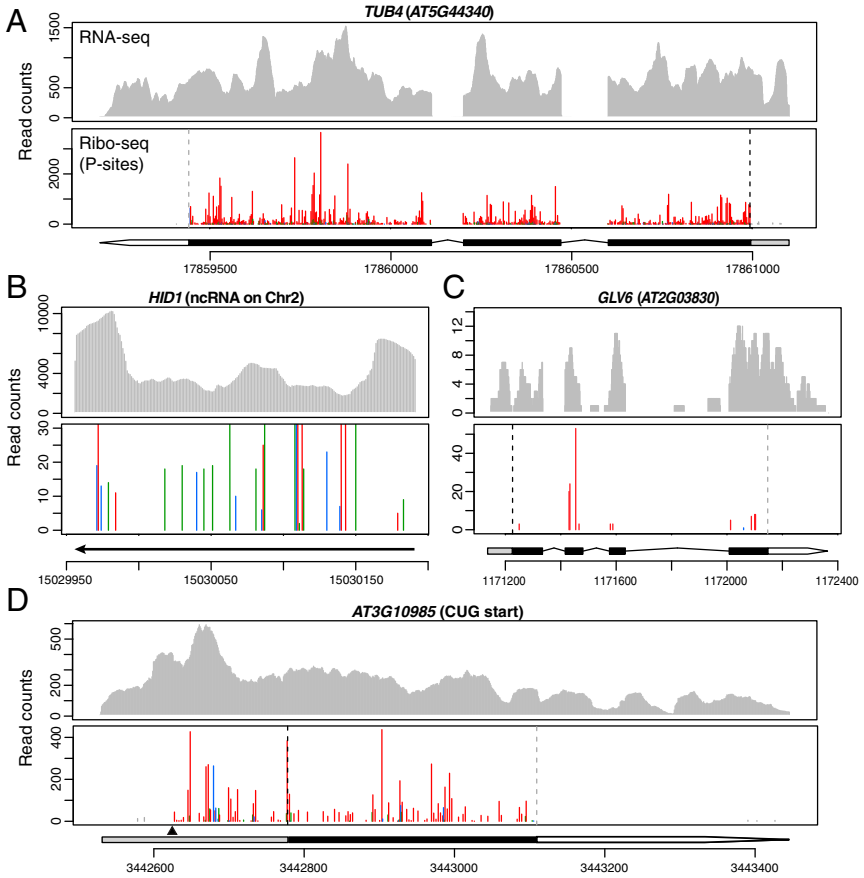

Fig. 3. Distinct profiles of annotated protein-coding genes and a well-characterized ncRNA. RNA-seq and P sites in ribosome footprints in root are shown for the following genes: (*A*) *TUB4*, a highly expressed gene; (*B*) *HID1*, a well-characterized ncRNA whose function is solely contributed by the RNA (31) and whose footprints do not display a clear 3-nt periodicity; the *y* axis is truncated to visualize low-abundance reads; (*C*) *GLV6*, a gene that encodes a secreted peptide with low expression levels and a short ORF; (*D*) *AT3G10985*, which uses an upstream CUG start codon (indicated by a black triangle). Annotated gene model and chromosome coordinates are indicated under each Ribo-seq profile. Within the gene model: gray box, 5′-UTR; black box, CDS; white arrow, 3′-UTR. Ribo-seq reads are shown by plotting their first nucleotide of the P site. Three reading frames are shown in red (the expected frame according to the predicted start codon), blue, and green. Footprints that are outside of the predicted coding sequences are shown in gray. The predicted start codon position is indicated by a black dashed line on each Ribo-seq profile panel; the predicted stop codon position is indicated by a gray dashed line.

Translated ORFs Identified in Annotated ncRNAs and Pseudogenes. We identified ORFs translated in transcripts previously thought to be noncoding, including ncRNAs and pseudogenes. In total, we identified 27 translated small ORFs derived from annotated ncRNAs, which we call *small ORF1* (*sORF1*) to *sORF31* (Table 1 and Dataset S1*G*). These sORFs range from 54 to 312 nt (Fig. S9 *A* and *B*). The P sites clearly show a 3-nt periodicity within the identified ORF range (Fig. 5 *A*–*C*). Interestingly, two sORFs identified in *AT1G79075* (*sORF17*) and *AT3G12965* (*sORF23*) encode a peptide sequence identical to five ribosomal L41 proteins in *Arabidopsis*. Thus, we identified two additional loci of ribosomal L41 genes.

To determine whether these sORFs produce stable proteins in planta, we epitope-tagged their coding regions and transformed them into *Arabidopsis*. To ensure that these transgenes behave similarly to the endogenous genes, we built the constructs using genomic sequences including their native promoter/5′-UTR/introns/3′-UTRs, so that the only difference between the transgenes and endogenous genes is the HA tag right before the stop codon in the transgenes (Fig. 5*D*). Of the four sORFs we tested (*sORF2*, *sORF12*, *sORF23*, and *sORF3*), we detected proteins from three of them in root extracts by Western blot (Fig. 5*D*). We also found that 37 annotated pseudogenes are actually expressed and translated (Table 1 and Dataset S1*H*). Mining publicly available proteomics data, four ORFs that we identified in either annotated ncRNAs or pseudogenes also have unique peptides detected by mass spectrometry (Dataset S1*I*) (45). The Western blots and the mass spectrometry data not only support the translation of these unannotated ORFs but also demonstrate that some of the ORFs produce stable proteins in plants.

Many unannotated ORFs Identified Are Evolutionarily Conserved. If the unannotated ORFs we identified encode functionally important proteins, we expect their homologs to be conserved in other plant genomes. We surveyed 15 other plant genomes, including 6 from Brassicaceae and 9 from other major lineages: eudicots (asterids and rosids), monocots, *Amborella* (the earliest diverging flowering plant), and *Selaginella* (a lycophyte). We used tBLASTn to search against whole-genome assemblies, and because pseudogenes could be a truncated form of other genes, we excluded ORFs that have more than 50 hits from the downstream analysis. After obtaining BLAST hits, we aligned all homolog sequences together to confirm they have similar protein sequences and similar start/stop positions (Fig. 6 and stringency described in *SI Materials and Methods*; all alignments are available in the Dryad Digital Repository).

For translated ORFs identified in the annotated ncRNAs or the unannotated gene mentioned above (Fig. 4*B*), we found 15 of the 19 single-exon ORFs have at least one homolog outside of *Arabidopsis thaliana* (Fig. 7). These ORFs can be further classified into six groups: (I) homologs only found in *Arabidopsis lyrata*; (II) homologs in multiple species within Brassicaceae; (III) homologs in other eudicots

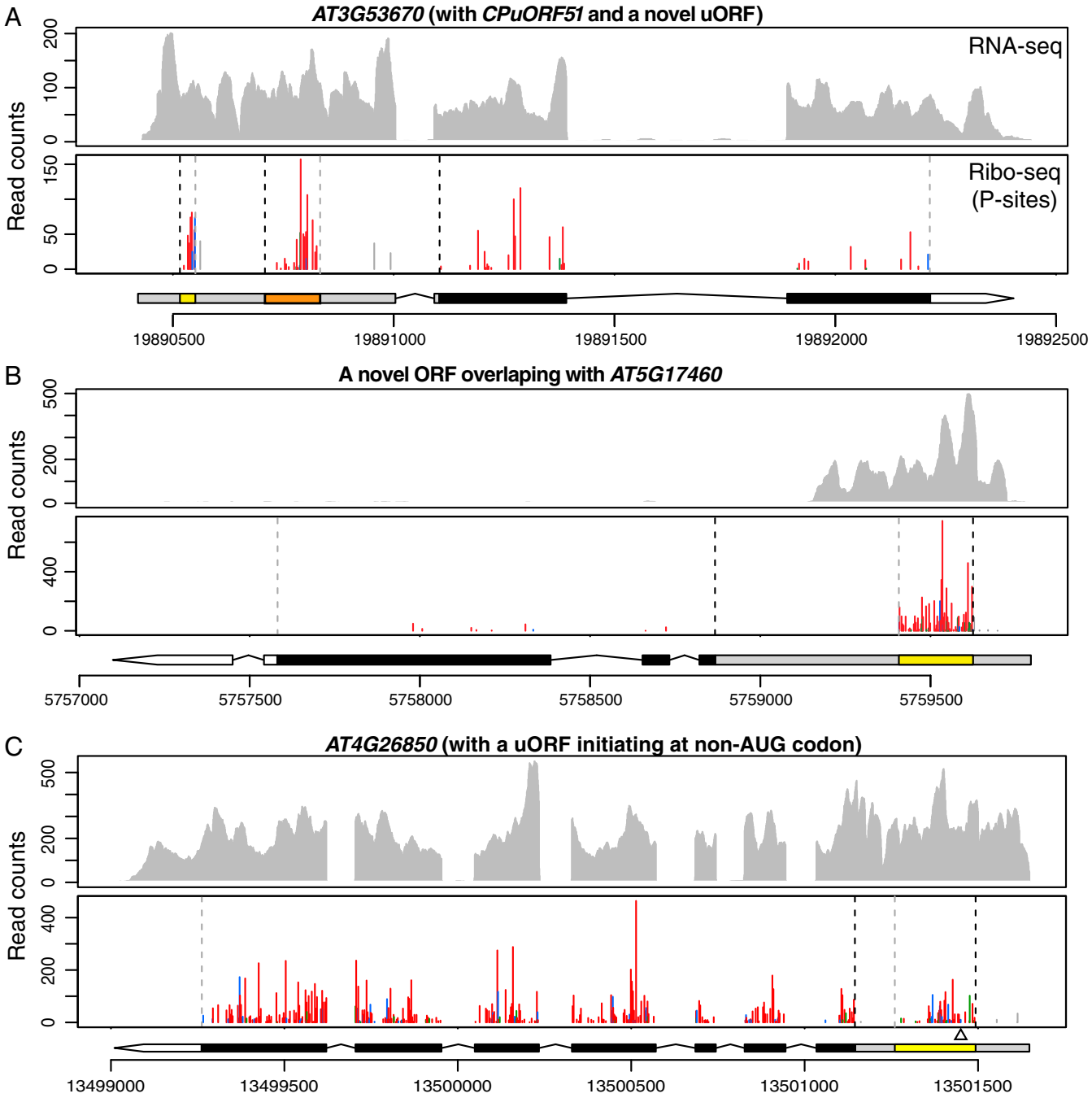

Fig. 4. uORFs and an unannotated ORF revealed by ribosome profiling. RNA-seq and P sites in ribosome footprints in root (*A* and *B*) or shoot (C) for the following genes: (*A*) *CPuORF51* (orange box) and an unannotated uORF (yellow box) within the 5′-UTR of *AT3G53670*. (*B*) An unannotated ORF identified as a uORF within the 5′-UTR of *AT5G17460* appears to be an ORF for an unannotated gene. The RNA-seq reads only cover a portion of the 5′-UTR of *AT5G17460*, suggesting the ORF identified (yellow box) represents the CDS of an unannotated gene, rather than a uORF of *AT5G17460*. (*C*) A uORF initiating at a non-AUG codon within the 5′-UTR of *AT4G26850* in the shoot. The uORF is marked as a yellow box in the 5′-UTR; the previously reported start codon (ACG; ref. 44) is indicated by an empty triangle underneath. Gene model and data presentation are the same as described in the legend of Fig. 3.

besides Brassicaceae; (IV) homologs in eudicots and *Amborella trichocarpa*, but not in monocots; (V) homologs in almost all flowering plants examined; and (VI) homologs in almost all plants examined including *Selaginella*. For translated ORFs identified in pseudogenes, some also have homologs in multiple Brassicaceae species, and some have homologs in almost all plants examined (Fig. S11). The total number of homologs found for each unannotated ORF is summarized in Dataset S1 *J* and *K*. Taken together, many of the unannotated ORFs are present and conserved in diverse plant lineages, as distant as *Amborella* and *Selaginella*, which diverged from *A. thaliana* over 200 and 445 million years ago, respectively (46). These findings indicate that these unannotated ORFs likely produce functionally important proteins.

Discussion

Ribosome profiling is a powerful technique providing precise information about translation in vivo. The resolution of Ribo-seq determines the amount of information that can be extracted,

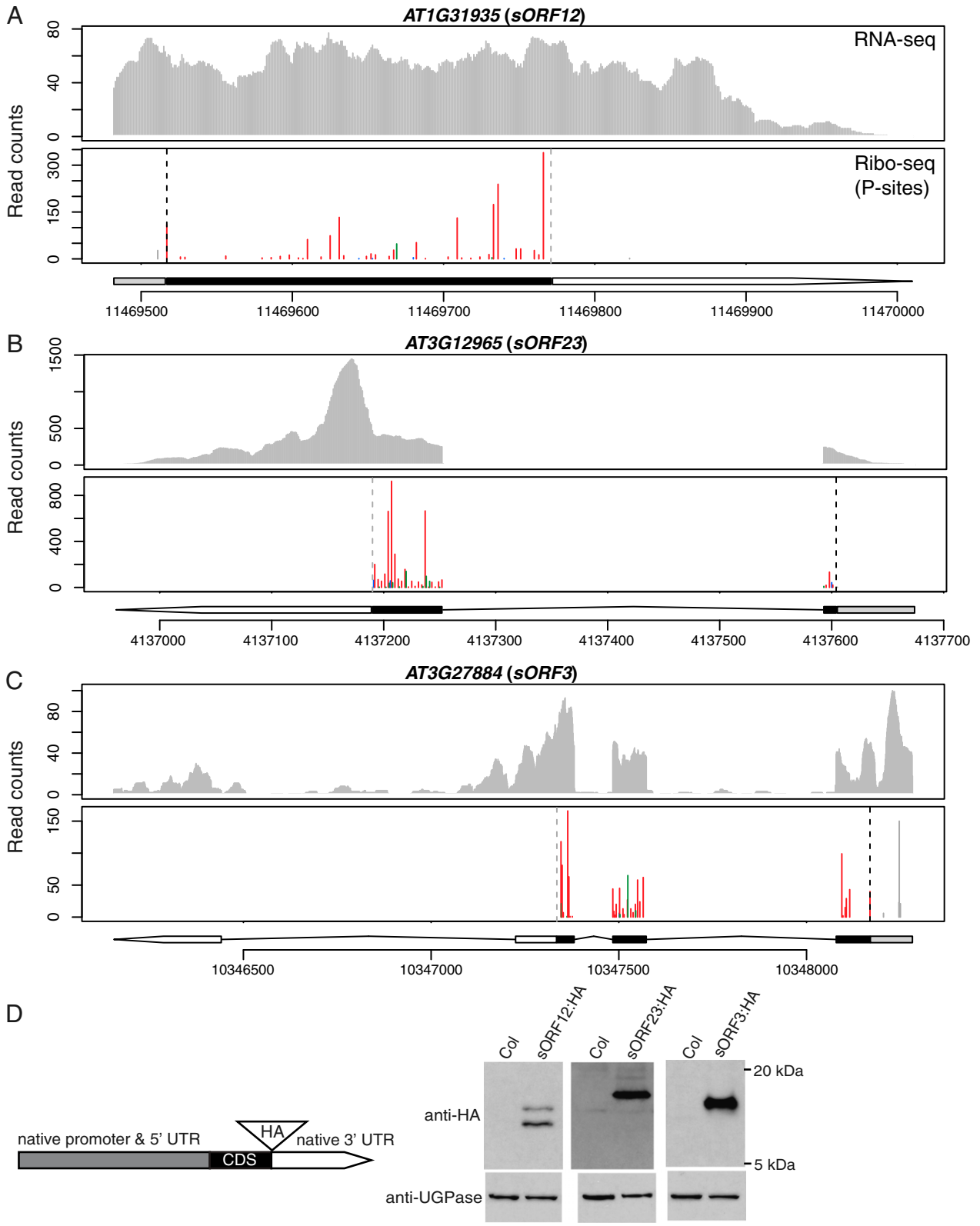

Fig. 5. Translated ORFs identified within annotated ncRNAs. (*A–C*) RNA-seq and P sites in ribosome footprints in root for three ORFs identified within annotated ncRNAs. The predicted CDS and 5′-UTR are depicted as black and gray boxes, respectively. The 3′-UTR is represented by a white arrow. Data presentation is the same as described in the legend of Fig. 3. (*D*) A schematic diagram of HA-tagged constructs and Western blot analysis of proteins produced by the three annotated ncRNAs in *A–C*. Total protein in root of control plants (Col-0) and transgenic plants expressing individual HA-tagged proteins was isolated and analyzed with either anti-HA antibody or anti-UGPase antibody as a loading control.

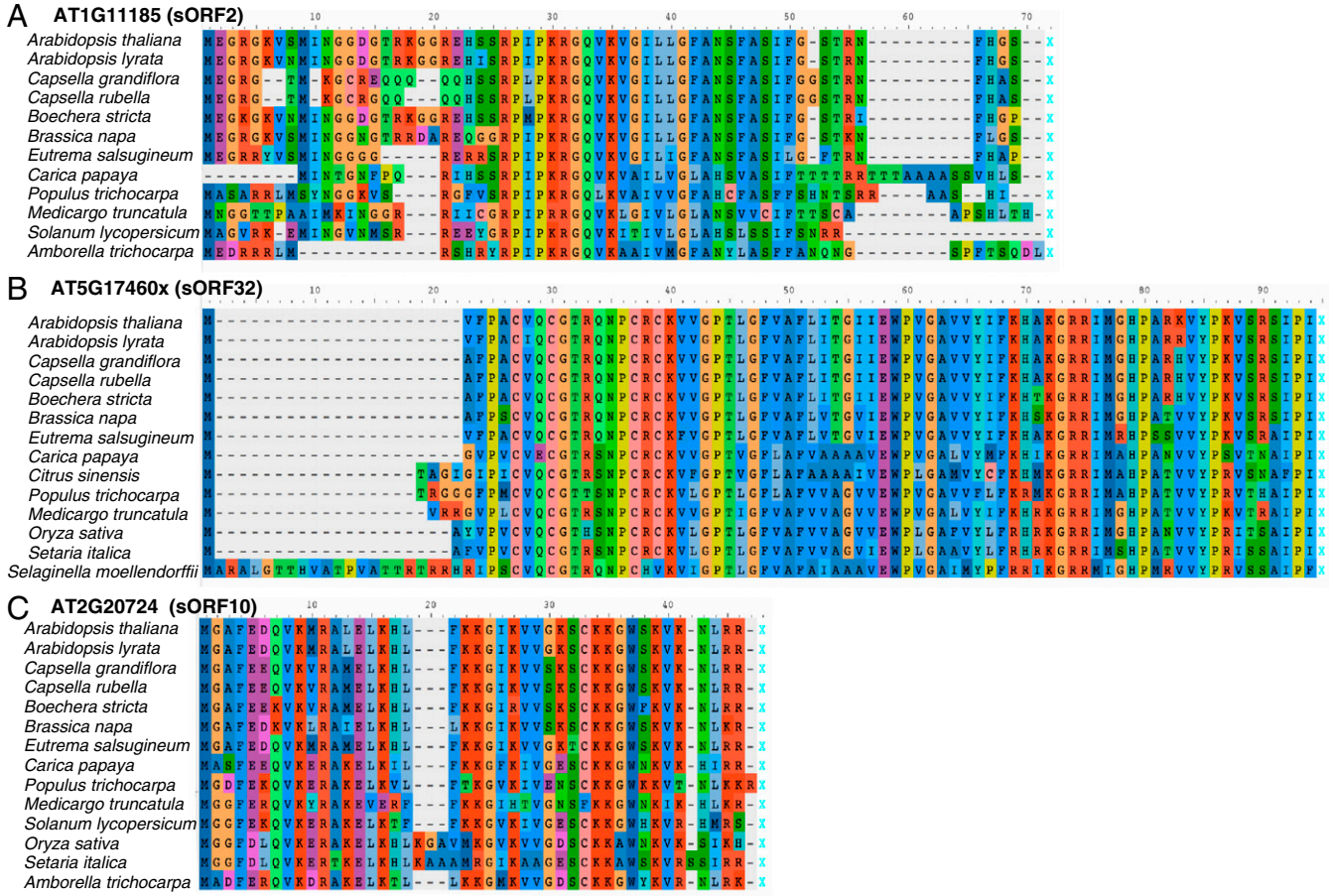

Fig. 6. Representative sequence alignments of unannotated ORFs in *A. thaliana* with corresponding homologs in 15 other plants. (*A*) An ORF identified in an annotated ncRNA. (*B*) An ORF identified in an unannotated gene overlapping with *AT5G17460* (denoted as *AT5G17460x*; also known as *sORF32*). (*C*) An ORF identified in a pseudogene. If there are multiple homologs identified in one genome, the homolog with the highest sequence identity to *A. thaliana* is shown. Amino acids with the same functional groups are shown in similar colors. Note that all these protein sequences have very similar start (the left-most methionine) and stop positions (X).

especially when identifying translated ORFs. Our datasets from *Arabidopsis* root and shoot display a super resolution even in individual transcripts. Compared with three published methods for *Arabidopsis* (Table S1), our protocol requires fewer starting materials and fewer sample processing steps, and yields dramatically better resolution. The key to obtaining precise footprints is complete digestion of unprotected portions of RNA. Judging the levels of digestion based on conversion of polysomes to monosomes within a polysome profile does not appear to be a reliable indication of complete digestion (26). By contrast, judging sharpness of RNA bands around 28 nt in a denaturing gel is a good indicator of complete digestion (27). Consistent with ionic strength being an important determinant of footprint precision in mammalian cell and human cytomegalovirus ribosome profiling (3, 5), we found ionic strength/buffering capacity in the extraction buffer had profound effects on footprint periodicity in *Arabidopsis*. With an optimized low ionic strength/buffering capacity extraction buffer, our protocol yielded a substantial improvement in Ribo-seq resolution compared with other methods in *Arabidopsis*. The resolution of our data are also among the best for all organisms.

Previous bioinformatics studies reported that ∼35% of *Arabidopsis* genes have at least one uORF (47, 48), and therefore over 9,000 uORF-containing genes would be expected in TAIR10. However, how many of these predicted uORFs are actually translated was an open question. Liu et al. (25) found 1,996 genes have at least one Ribo-seq read within predicted uORFs. Using RiboTaper, we identified 187 uORFs translated among 18,745 expressed genes (TPM > 1) in our data. Because RiboTaper examines 3-nt periodicity along the potential uORF, it is possible that some translated uORFs were missed due to their short length, insufficient sequencing coverage potentially due to low expression levels, or because they overlapped with other uORFs. Although the number of uORFs identified might be an underestimate, those identified are of high confidence. For example, our list includes 44% of the predicted CPuORFs, several of which are known to play an important role in regulating downstream main ORFs involved in diverse functions in plants (36).

Perhaps of greatest interest is the identification of small translated ORFs within annotated ncRNAs. Because computational approaches typically exclude ORFs that are less than 100 aa (49, 50), small proteins are likely missed, and their transcripts may be classified as ncRNAs (51, 52). As shown by Western blot and mass spectrometry, at least some of the small ORFs we identified produce stable proteins. Evolutionary conservation further suggests that many of these unannotated ORFs encode functionally important proteins. Even species-specific ORFs might play an important role (52), as translation can have essential regulatory functions in addition to producing stable proteins (53, 54). Recently, several peptides derived from annotated ncRNAs were found to play important roles in signaling and development, such as Toddler in zebrafish embryo development (55) and DWORF in heart muscle contraction (56). How the sORFs we identified function in plants requires further investigation.

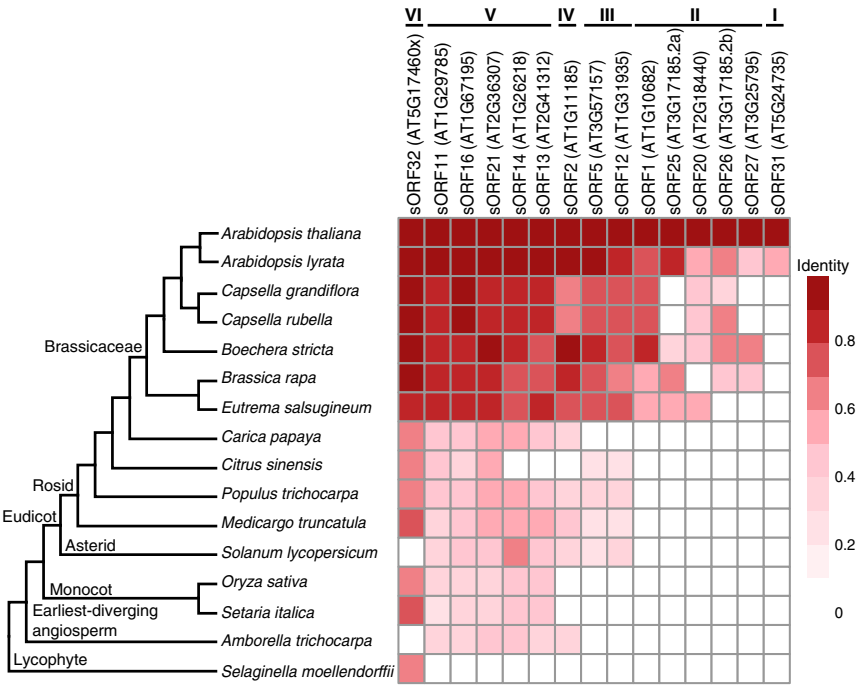

Fig. 7. Homolog sequence identities of translated ORFs found in annotated ncRNAs or in an unannotated gene. A heat map showing amino acid sequence identities between translated ORFs within annotated ncRNAs/unannotated gene (*sORF32*) in *A. thaliana* and their corresponding homologs in 15 other plant species. A phylogenetic tree showing evolutionary divergence is on the *Left*. One homolog with the best sequence identity in each genome is represented here. The ORFs can be further grouped based on their homologs identified in other species (I to VI).

Although RiboTaper only searches for AUG start codons, our super-resolution ribosome-profiling data also provide an invaluable resource to study noncanonical start codons and alternative start sites. The data may also be useful for characterizing translation of different transcript isoforms. With its high sensitivity for identification of translated ORFs and its quantitative nature, ribosome profiling can also serve as a proxy for the proteome or assist proteomics studies (57–59). Finally, as the *Arabidopsis* genome is among the best annotated, we expect ribosome profiling will be an even more powerful approach to uncovering novel ORFs and improving genome annotation when applied to less well-characterized organisms.

Materials and Methods

Detailed information on materials and methods used in this study is provided in *SI Materials and Methods*.

Plant Materials and Growth Conditions. *Arabidopsis* seeds were surface sterilized, imbibed at 4 °C for 2 d, and grown hydroponically with sterile liquid media (2.15 g/L Murashige and Skoog salt, 1% sucrose, 0.5 g/L MES, pH 5.7) and shaken at 85 rpm under 16-h light and 8-h dark at 22 °C.

Ribo-seq and RNA-seq Library Construction. Detailed procedures are provided in *SI Materials and Methods*. Four polysome extraction buffers were tested (Fig. S1), and buffer D was used to extract three biological replicates of 4-d-old root and shoot samples. Polysomes were extracted from 0.1 g of root or shoot pulverized tissue with buffer D [100 mM Tris·HCl (pH 8), 40 mM KCl, 20 mM $MgCl_2$, 2% polyoxyethylene (10) tridecyl ether (v/v), 1% deoxycholic acid (w/v), 1 mM DTT, 100 μg/mL cycloheximide, and 10 unit/mL DNase I]. The nuclease digestion was performed at 23 °C for 1 h. Ribosomes were isolated by size exclusion columns (illustra MicroSpin S-400 HR Columns; GE Healthcare). After RNA isolation and rRNA depletion, footprints from 28 to 30 nt separated by a denaturing gel were recovered. Ribo-seq and RNA-seq libraries were constructed using the ARTseq/TruSeq Ribo Profile Kit (illumina). The libraries were barcoded and pooled for single-end 50-bp sequencing in a HiSeq 2000 or 2500 machine.

Ribo-seq and RNA-seq Data Analysis. Quality filtering and adaptor clipping were performed by FASTX_toolkit (60). The rRNA, tRNA, and snoRNA sequences were removed in Ribo-seq data using bowtie2 (61). Mapping to the *Arabidopsis* genome [TAIR10 (30)] was carried out by STAR (62). Statistical presentations of the data were plotted in R using various R packages. TPM values were determined by RSEM (63). de novo ORF finding was performed by RiboTaper (15).

Western Blotting. C-terminus HA-tagging constructs were built by Gibson assembly (64) and transformed into Col-0 plants. Total proteins were extracted from root of 4-d-old Col-0 and T4 transgenic plants. Protein samples were analyzed by immunoblotting, using anti-HA antibody or anti-UGPase antibody followed by a secondary antibody conjugated to HRP.

BLAST and Sequence Alignment. tBLASTn (65) was performed in 15 plant genomes. Multiple sequence alignments for each ORF and its homologs were constructed.

ACKNOWLEDGMENTS. We thank Nicholas T. Ingolia, Gene-Wei Li, Ariel A. Bazzini, Jose M. Alonso, and Catharina Merchante for helpful discussions on ribosome profiling; Meng Chen for sharing the pJHA212B-RbcS terminator vector; Wen-Ping Hsieh for advice on protein extraction; Emily Jie-Ning Yang for suggestions on Gibson Assembly; Carmen Wilson for technical assistance; and members of the P.N.B. laboratory for critical reading and helpful discussions, especially Jazz Dickinson and Eric Rogers. This work used the Vincent J. Coates Genomics Sequencing Laboratory at University of California, Berkeley, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303. This research was funded by NIH Grant R01-GM043778 (to P.N.B.), the Howard Hughes Medical Institute and the Gordon and Betty Moore Foundation through Grant GBMF3405 (to P.N.B.), and Deutsche Forschungsgemeinschaft-Transregios 175 (to U.O.). Additional support was provided by US Department of Agriculture - National Institute of Food and Agriculture Postdoctoral Fellowship Award 2016-67012-24720 (to P.Y.H.).

1. King HA, Gerber AP (2016) Translatome profiling: Methods for genome-scale analysis of mRNA translation. *Brief Funct Genomics* 15(1):22–31.
2. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324(5924):218–223.
3. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* 7(8):1534–1550.
4. Brar GA, et al. (2012) High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* 335(6068):552–557.

5. Stern-Ginossar N, et al. (2012) Decoding human cytomegalovirus. *Science* 338(6110): 1088–1093.
6. Bazzini AA, et al. (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* 33(9):981–993.
7. Ingolia NT, et al. (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Reports* 8(5):1365–1379.
8. Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM (2014) Long non-coding RNAs as a source of new peptides. *eLife* 3:e03523.
9. Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES (2013) Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* 154(1):240–251.
10. Guydosh NR, Green R (2014) Dom34 rescues ribosomes in 3′ untranslated regions. *Cell* 156(5):950–962.
11. Brar GA, Weissman JS (2015) Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol* 16(11):651–664.
12. Ji Z, Song R, Huang H, Regev A, Struhl K (2016) Transcriptome-scale RNase-footprinting of RNA-protein complexes. *Nat Biotechnol* 34(4):410–413.
13. Chew G-L, et al. (2013) Ribosome profiling reveals resemblance between long noncoding RNAs and 5′ leaders of coding RNAs. *Development* 140(13):2828–2834.
14. Michel AM, et al. (2012) Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res* 22(11):2219–2229.
15. Calviello L, et al. (2016) Detecting actively translated open reading frames in ribosome profiling data. *Nat Methods* 13(2):165–170.
16. Ingolia NT (2016) Ribosome footprint profiling of translation throughout the genome. *Cell* 165(1):22–33.
17. Guo H, Ingolia NT, Weissman JS, Bartel DP (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 466(7308):835–840.
18. Fields AP, et al. (2015) A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Mol Cell* 60(5):816–827.
19. Ji Z, Song R, Regev A, Struhl K (2015) Many lncRNAs, 5′UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* 4:e08890.
20. Chung BY, et al. (2015) The use of duplex-specific nuclease in ribosome profiling and a user-friendly software package for Ribo-seq data analysis. *RNA* 21(10):1731–1745.
21. Schafer S, et al. (2015) Translational regulation shapes the molecular landscape of complex disease phenotypes. *Nat Commun* 6:7200.
22. Oh E, et al. (2011) Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell* 147(6):1295–1308.
23. Dunn JG, Foo CK, Belletier NG, Gavis ER, Weissman JS (2013) Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila melanogaster*. *eLife* 2:e01179.
24. Aspden JL, et al. (2014) Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *eLife* 3:e03528.
25. Liu M-J, et al. (2013) Translational landscape of photomorphogenic *Arabidopsis*. *Plant Cell* 25(10):3699–3710.
26. Juntawong P, Girke T, Bazin J, Bailey-Serres J (2014) Translational dynamics revealed by genome-wide profiling of ribosome footprints in *Arabidopsis*. *Proc Natl Acad Sci USA* 111(1):E203–E212.
27. Merchante C, et al. (2015) Gene-specific translation regulation mediated by the hormone-signaling molecule EIN2. *Cell* 163(3):684–697.
28. Lei L, et al. (2015) Ribosome profiling reveals dynamic translational landscape in maize seedlings under drought stress. *Plant J* 84(6):1206–1218.
29. Thomson DJ (1982) Spectrum estimation and harmonic analysis. *Proc IEEE* 70(9): 1055–1096.
30. Berardini TZ, et al. (2015) The *Arabidopsis* information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis* 53(8):474–485.
31. Wang Y, et al. (2014) *Arabidopsis* noncoding RNA mediates control of photomorphogenesis by red light. *Proc Natl Acad Sci USA* 111(28):10359–10364.
32. Whitford R, et al. (2012) GOLVEN secretory peptides regulate auxin carrier turnover during plant gravitropic responses. *Dev Cell* 22(3):678–685.
33. Matsuzaki Y, Ogawa-Ohnishi M, Mori A, Matsubayashi Y (2010) Secreted peptide signals required for maintenance of root stem cell niche in *Arabidopsis*. *Science* 329(5995):1065–1067.
34. Meng L, Buchanan BB, Feldman LJ, Luan S (2012) CLE-like (CLEL) peptides control the pattern of root growth and lateral root development in *Arabidopsis*. *Proc Natl Acad Sci USA* 109(5):1760–1765.
35. Ghorbani S, et al. (2015) Expanding the repertoire of secretory peptides controlling root development with comparative genome analysis and functional assays. *J Exp Bot* 66(17):5257–5269.
36. Jorgensen RA, Dorantes-Acosta AE (2012) Conserved peptide upstream open reading frames are associated with regulatory genes in angiosperms. *Front Plant Sci* 3:191.
37. Ebina I, et al. (2015) Identification of novel *Arabidopsis thaliana* upstream open reading frames that control expression of the main coding sequences in a peptide sequence-dependent manner. *Nucleic Acids Res* 43(3):1562–1576.
38. Hayden CA, Jorgensen RA (2007) Identification of novel conserved peptide uORF homology groups in *Arabidopsis* and rice reveals ancient eukaryotic origin of select groups and preferential association with transcription factor-encoding genes. *BMC Biol* 5(1):32.
39. Takahashi H, Takahashi A, Naito S, Onouchi H (2012) BAIUCAS: A novel BLAST-based algorithm for the identification of upstream open reading frames with conserved amino acid sequences and its application to the *Arabidopsis thaliana* genome. *Bioinformatics* 28(17):2231–2241.
40. Vaughn JN, Ellingson SR, Mignone F, Arnim Av (2012) Known and novel post-transcriptional regulatory sequences are conserved across plant families. *RNA* 18(3):368–384.
41. Imai A, et al. (2006) The dwarf phenotype of the *Arabidopsis* acl5 mutant is suppressed by a mutation in an upstream ORF of a bHLH gene. *Development* 133(18):3575–3585.
42. Campbell MS, et al. (2014) MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol* 164(2):513–524.
43. Simpson GG, et al. (2010) Noncanonical translation initiation of the *Arabidopsis* flowering time and alternative polyadenylation regulator FCA. *Plant Cell* 22(11): 3764–3777.
44. Laing WA, et al. (2015) An upstream open reading frame is essential for feedback regulation of ascorbate biosynthesis in *Arabidopsis*. *Plant Cell* 27(3):772–786.
45. Castellana NE, et al. (2008) Discovery and revision of *Arabidopsis* genes by proteogenomics. *Proc Natl Acad Sci USA* 105(52):21034–21038.
46. Clarke JT, Warnock RCM, Donoghue PCJ (2011) Establishing a time-scale for plant evolution. *New Phytol* 192(1):266–301.
47. Kim B-H, Cai X, Vaughn JN, von Arnim AG (2007) On the functions of the h subunit of eukaryotic initiation factor 3 in late stages of translation initiation. *Genome Biol* 8(4):R60.
48. von Arnim AG, Jia Q, Vaughn JN (2014) Regulation of plant translation by upstream open reading frames. *Plant Sci* 214:1–12.
49. Basrai MA, Hieter P, Boeke JD (1997) Small open reading frames: Beautiful needles in the haystack. *Genome Res* 7(8):768–771.
50. Claverie JM (1997) Computational methods for the identification of genes in vertebrate genomic sequences. *Hum Mol Genet* 6(10):1735–1744.
51. Hellens RP, Brown CM, Chisnall MAW, Waterhouse PM, Macknight RC (2015) The emerging world of small ORFs. *Trends Plant Sci* 21(4):317–328.
52. Andrews SJ, Rothnagel JA (2014) Emerging evidence for functional peptides encoded by short open reading frames. *Nat Rev Genet* 15(3):193–204.
53. Gaba A, Jacobson A, Sachs MS (2005) Ribosome occupancy of the yeast CPA1 upstream open reading frame termination codon modulates nonsense-mediated mRNA decay. *Mol Cell* 20(3):449–460.
54. Arribere JA, Gilbert WV (2013) Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Res* 23(6):977–987.
55. Pauli A, et al. (2014) Toddler: An embryonic signal that promotes cell movement via Apelin receptors. *Science* 343(6172):1248636.
56. Nelson BR, et al. (2016) A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science* 351(6270):271–275.
57. Ingolia NT, Lareau LF, Weissman JS (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147(4): 789–802.
58. Menschaert G, et al. (2013) Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol Cell Proteomics* 12(7):1780–1790.
59. Crappé J, et al. (2015) PROTEOFORMER: Deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res* 43(5):e29.
60. Pearson WR, Wood T, Zhang Z, Miller W (1997) Comparison of DNA sequences with protein sequences. *Genomics* 46(1):24–36.
61. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–359.
62. Dobin A, et al. (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1): 15–21.
63. Li B, Dewey CN (2011) RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12(1):323.
64. Gibson DG, et al. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* 6(5):343–345.
65. Camacho C, et al. (2009) BLAST+: Architecture and applications. *BMC Bioinformatics* 10(1):421.
66. Mustroph A, Juntawong P, Bailey-Serres J (2009) Isolation of plant polysomal mRNA by differential centrifugation and ribosome immunopurification methods. *Methods Mol Biol* 553:109–126.
67. Wei T, Simko V (2016) corrplot: Visualization of a correlation matrix. Available at https://cran.r-project.org/web/packages/corrplot/index.html. Accessed April 27, 2016.
68. Akalin A, Franke V, Vlahoviček K, Mason CE, Schübeler D (2015) Genomation: A toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics* 31(7): 1127–1129.
69. Lawrence M, et al. (2013) Software for computing and annotating genomic ranges. *PLoS Comput Biol* 9(8):e1003118.
70. Adler D (2005) vioplot: Violin plot. Available at https://cran.r-project.org/web/packages/vioplot/vioplot.pdf. Accessed April 27, 2016.
71. Kolde R (2015) pheatmap: Pretty Heatmaps. Available at https://cran.r-project.org/web/packages/pheatmap/index.html. Accessed April 27, 2016.
72. Gibson DG (2011) Enzymatic assembly of overlapping DNA fragments. *Methods Enzymol* 498:349–361.
73. Yoo SY, et al. (2005) The 35S promoter used in a selectable marker gene of a plant transformation vector affects the expression of the transgene. *Planta* 221(4):523–530.
74. Zhang X, Henriques R, Lin SS, Niu QW, Chua NH (2006) Agrobacterium-mediated transformation of *Arabidopsis thaliana* using the floral dip method. *Nat Protoc* 1(2):641–646.
75. Silverstone AL, et al. (2001) Repressing a repressor: Gibberellin-induced rapid reduction of the RGA protein in *Arabidopsis*. *Plant Cell* 13(7):1555–1566.
76. Goodstein DM, et al. (2012) Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res* 40:D1178–D1186.
77. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.
78. Larsson A (2014) AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 30(22):3276–3278.
79. Krishnakumar V, et al. (2015) Araport: The *Arabidopsis* information portal. *Nucleic Acids Res* 43:D1003–D1009.