the plant journal
SEB Society for Experimental Biology

*The Plant Journal* (2017) **89**, 789–804 doi: 10.1111/tpj.13415

RESOURCE

# Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome

**Chia-Yi Cheng[1,†], Vivek Krishnakumar[1,†], Agnes P. Chan[1], Françoise Thibaud-Nissen[2], Seth Schobel[1] and Christopher D. Town[1,*]**

[1]*J. Craig Venter Institute, 9714 Medical Center Drive, Rockville, MD 20850, USA, and*
[2]*National Center for Biotechnology Information, US National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA*

Received 27 July 2016; revised 27 October 2016; accepted 3 November 2016; published online 8 November 2016.
*For correspondence (e-mail cdtown@jcvi.org).
†These authors contributed equally to this work.

## SUMMARY

**The flowering plant *Arabidopsis thaliana* is a dicot model organism for research in many aspects of plant biology. A comprehensive annotation of its genome paves the way for understanding the functions and activities of all types of transcripts, including mRNA, the various classes of non-coding RNA, and small RNA. The TAIR10 annotation update had a profound impact on Arabidopsis research but was released more than 5 years ago. Maintaining the accuracy of the annotation continues to be a prerequisite for future progress. Using an integrative annotation pipeline, we assembled tissue-specific RNA-Seq libraries from 113 datasets and constructed 48 359 transcript models of protein-coding genes in eleven tissues. In addition, we annotated various classes of non-coding RNA including microRNA, long intergenic RNA, small nucleolar RNA, natural antisense transcript, small nuclear RNA, and small RNA using published datasets and in-house analytic results. Altogether, we identified 635 novel protein-coding genes, 508 novel transcribed regions, 5178 non-coding RNAs, and 35 846 small RNA loci that were formerly unannotated. Analysis of the splicing events and RNA-Seq based expression profiles revealed the landscapes of gene structures, untranslated regions, and splicing activities to be more intricate than previously appreciated. Furthermore, we present 692 uniformly expressed housekeeping genes, 43% of whose human orthologs are also housekeeping genes. This updated Arabidopsis genome annotation with a substantially increased resolution of gene models will not only further our understanding of the biological processes of this plant model but also of other species.**

**Keywords: Arabidopsis, annotation, transcriptome.**

## INTRODUCTION

Adopted by the research community over 50 years ago as a model for plant research (Rédei, 1975; Provart *et al.*, 2016), *Arabidopsis thaliana*, a member of the crucifer family, continues to occupy a prominent place in plant biology. It also has an underappreciated influence on medical research and human health. Studies using Arabidopsis have played a leading role in basic biological discoveries (Jones *et al.*, 2008) such as the plant nucleotide-binding, leucine-rich repeat (NB-LRR) proteins and their later identified human orthologs in the innate immune system (Jones and Dangl, 2006), the impact of auxin research on the ubiquitin pathway conserved among eukaryotes (Parry and Estelle, 2006), a light signaling component COP1 whose mammalian orthologs has a role in tumorigenesis (Deng *et al.*, 1991).

The *Arabidopsis thaliana* genome sequence was initially assembled and annotated in 2000 by an international consortium (Arabidopsis Genome Initiative, 2000). The genome sequence derived from the Columbia-0 (Col-0) ecotype was updated twice to remove embedded vector sequences and other sequence errors. The most recent update of the Col-0 genome sequence coincided with the TAIR9 version of the genome sequence. Refined annotations were released by The Institute for Genomic Research

(TIGR, versions 1–4) (Haas *et al*., 2005) and subsequently by The Arabidopsis Information Resource (TAIR, versions 5–10). The TAIR10 genome annotation was informed by *ab initio* gene models, EST sequences from Sanger platforms, and two RNA-Seq datasets available at that time (Lamesch *et al*., 2012). Since TAIR10, around 200 *Arabidopsis thaliana* RNA-Seq studies have been published and deposited in NCBI SRA. In comparison with the EST data that provided the bulk of the TAIR10 annotation, the RNA-Seq data offer single-base resolution and more precise measurement of levels of transcripts and their isoforms (Wang *et al*., 2009). Thus, the publically available RNA-Seq datasets present a compelling opportunity to update the TAIR10 annotation.

Recent literature also reveals a growing amount of information about non-coding RNAs, including long intergenic RNAs, natural antisense transcripts, small RNAs, microRNAs, small nuclear RNAs, small nucleolar RNAs and tRNAs (Wang and Brendel, 2004; Matsui *et al*., 2008; Okamoto *et al*., 2010; Liu *et al*., 2012; Sherstnev *et al*., 2012; Li *et al*., 2013; Csorba *et al*., 2014; Kozomara and Griffiths-Jones, 2014). A complete documentation and uniform annotation of non-coding RNAs will serve as a platform for further understanding their regulatory roles.

Here, we report an update to the annotation of the *Arabidopsis thaliana* Col-0 genome. The new annotation offers gene structure updates applied to the TAIR9 genome sequence and improved the functional descriptions for over 7000 genes. In an effort to exclude false gene models formed by combining isoforms, the analysis is based on tissue-specific assemblies of RNA-Seq data. The new annotation is called Araport11 to maintain version number continuity and to designate the source as Araport, an open-access online resource for Arabidopsis research community (Krishnakumar *et al*., 2015). Araport11 is accessible through Araport (www.araport.org) and GenBank (accessions CP002684–CP002688).

## RESULTS

### The annotation pipeline

*Protein-coding RNA.* The Araport11 re-annotation process included building an augmented gene set extended from TAIR10 annotation followed by incorporating extensive public RNA-Seq data for gene structure and splice isoform updates (Figure 1 and Experimental procedures S1). First we developed a set of reference gene structures by augmenting the TAIR10 dataset with 635 novel gene models from the NCBI RefSeq Gnomon pipeline (Pruitt *et al*., 2012), the MAKER-P pipeline (Campbell *et al*., 2014), and published work (Wang *et al*., 2014; Vidal *et al*., 2013). Next we collected 113 RNA-Seq datasets generated from

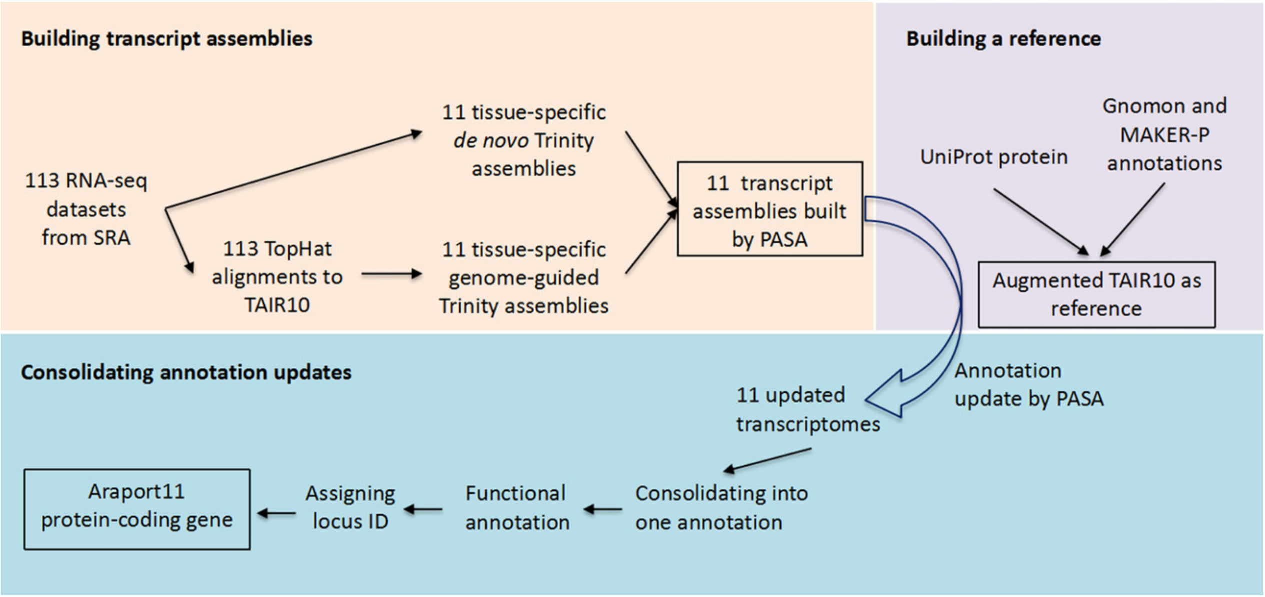

Building transcript assemblies
Building a reference
113 RNA-seq datasets from SRA
11 tissue-specific *de novo* Trinity assemblies
113 TopHat alignments to TAIR10
11 tissue-specific genome-guided Trinity assemblies
11 transcript assemblies built by PASA
UniProt protein
Gnomon and MAKER-P annotations
Augmented TAIR10 as reference
Consolidating annotation updates
Annotation update by PASA
11 updated transcriptomes
Consolidating into one annotation
Functional annotation
Assigning locus ID
Araport11 protein-coding gene

**Figure 1.** Annotation pipeline for Araport11 protein-coding genes.
The RNA-Seq reads obtained from NCBI SRA were grouped into 11 tissue types and assembled by Trinity using a combination of *de novo* and genome-guided assemblies to reconstruct tissue-based transcriptomes. To build Araport11, first an augmented TAIR10 was constructed. TAIR10 annotation was supplemented with novel transcripts from NCBI and MAKER-P assemblies as well as curated protein records from UniProt. Using the augmented TAIR10 as the reference set, PASA annotation updates were run separately on each tissue dataset (to avoid constructing chimeric transcripts across tissues). The resulting 11 transcriptomes were consolidated using a custom Python script to collapse isoforms that share exact exon-intron boundaries but may differ in UTR length. The Araport11 protein-coding genes were re-indexed with appropriate locus and isoform identifiers and submitted to NCBI along with non-coding RNAs and other features as described in the main text.

untreated or mock-treated wild-type Col-0 plants from NCBI SRA. The RNA-Seq datasets were partitioned into 11 groups according to their tissue or organ of origin. These include the aerial part, carpel, dark-grown seedling, leaf, light-grown seedling, pollen, receptacle, root, root apical meristem, stage 12 inflorescence, and shoot apical meristem. We will refer to these sample sources as tissues hereafter following a previous convention (Schmid *et al.*, 2005). In total, 66.1 Gb of RNA-Seq reads were uniquely mapped to the TAIR9 genome sequence and represented 2182-fold coverage of the Araport11 transcriptome (Table S1).

To refine the augmented TAIR10 annotation, eleven tissue-specific transcript assemblies were independently constructed (Experimental procedures S1). The transcript assemblies from each tissue were mapped to the genome, and compatible assemblies were collapsed into individual isoforms. TAIR10 genes that were short and lacked expression support were made obsolete (Data S6). Araport11 contains a final set of 27 655 protein-coding loci with 48 359 transcripts. A comparison of various gene classes present in Araport11 and TAIR10 annotation versions is shown in Table 1. A comparison of gene structure statistics between Araport11 and TAIR10 annotation versions is provided in Figure S1.

To update the functional annotation (gene product names) of novel and existing protein-coding loci, we used a weighted keyword approach (Hoover *et al.*, manuscript in preparation). In brief, each predicted protein sequence was searched against a number of protein and domain databases (Priam, Uniref100, PFAM/TIGRFAM, CAZY, CDD) and scanned with motif finders (TMHMM, InterPro) (Table 2). Keywords were extracted from the definition lines of best matches and scored based on a set of heuristic rules. The maximal scoring definition line became the functional annotation if the score surpassed a threshold. This process generated new functional annotations of 7122 protein-coding loci, including 635 loci not present in TAIR10. Overall, 25 402 of 27 655 protein-coding genes (91.8%) in Araport11 are now annotated with meaningful product names which is a substantial improvement over TAIR10 (79.7%, 21 834 of 27 416). Furthermore, a TAIR 'locus detail' page displays only one of the following attributes: curator summary, computational description, or product name in that order of preference, making product name hard to find, whereas Araport/ThaleMine gene report pages display all three descriptors.

*Non-coding RNA.* To Araport11 we incorporated existing non-coding RNA datasets from miRBase (Kozomara and Griffiths-Jones, 2014) and PlantRNA database (Cognat *et al.*, 2013) as well as characterized genes collected from publications (e.g. *COOLAIR*) (Csorba *et al.*, 2014). For the long-intergenic RNAs, we used published results of Liu *et al.* (2012) to annotate 2704 loci that fall into the intergenic regions in Araport11. Natural antisense transcripts (NATs) are transcribed from the DNA strand opposite to the sense RNAs and overlap with their sense counterparts. As the accurate detection of NATs relies upon strand-specific sequencing technology combined with statistical-computational approaches, we manually created transcript structures and annotated the 1115 NAT loci identified by combined strand-specific sequencing data and statistical analyses (Li *et al.*, 2013). Overall, Araport11 represents a total of 5178 non-coding RNA loci (Table 1).

*Pseudogenes.* Arabidopsis annotation (TAIR and Araport) has consistently defined pseudogenes as coding sequences that would produce an inactive product, usually due to either stop codon truncation or missense mutation. Due to sequence variation between accessions, pseudogenes are accession-specific. Our analysis shows that almost all of the pseudogenes (>90%) are diverged from their closest paralogs (<97% identity), thus transcript assemblies can be unambiguously mapped to these loci as evidence of transcription. Among the 924 pseudogenic loci annotated in TAIR10, a group of 259 showed evidence of transcription using our mapping pipeline. We manually curated 425 pseudogenic transcript models on those 259 loci using Web Apollo (Lee *et al.*, 2013).

Over 30% of the transcribed pseudogenic loci encode two or more transcript variants, which have clearly defined spliced structures (Figure S2). The level of splicing activity is comparable to that of protein-coding genes (38%), which is remarkably complex for a genomic feature that was traditionally considered non-functional. Furthermore, we updated the functional annotation for 882 pseudogenes using blastx to find the best matched counterparts in the annotated Arabidopsis proteome. We consolidated this dataset with 25 pseudogenic tRNAs curated by the PlantRNA database (Cognat *et al.*, 2013), and obtained a final annotation of 952 pseudogenes in Araport11.

*Upstream open reading frames.* Upstream open reading frames (uORFs) are common genomic features in the 5′ untranslated regions of eukaryotic mRNAs. In Arabidopsis, identification of uORFs has previously relied on the conserved peptides encoded by the uORFs (CPuORFs). These cases are thought to represent less than 1% of all uORFs in plants and animals. In total, 64 CPuORFs associated with 58 loci (Hayden and Jorgensen, 2007) were annotated in TAIR10, while thousands more uORFs have been inferred by computational prediction and ribosome profiling data (Takahashi *et al.*, 2012; Juntawong *et al.*, 2014). Araport11 includes an additional 26 uORF-generating loci from the literature, each with characterized biological function (Saul *et al.*, 2009; Rosado *et al.*, 2012; Takahashi *et al.*, 2012; Laing *et al.*, 2015) (Table 1). Noting that one main ORF

**Table 1** Summary of Araport11

| Type | TAIR10 | Araport11 | Change |
|---|---|---|---|
| (A) Protein-coding genes | | | |
| Total number of loci | 27 416 | 27 655[a] | +239 |
| Number of transcript isoforms | 35 386 | 48 359 | +12 973 |
| Number of loci with two or more splice variants | 5804 | 10 696 | +4892 |
| Number of loci with changes in CDS | – | – | +1158 |
| Number of loci with changes in UTR(s) | – | – | +21 298 |
| Upstream open reading frame (uORF) | 58 | 84 | +26 |
| (B) Non-coding genes | | | |
| Long intergenic non-coding RNA (lincRNA) | 36 | 2444 | +2408 |
| Natural antisense transcripts (NAT) | 223 | 1115 | +892 |
| microRNA (miRNA) | 177 | 325 | +148 |
| Small nucleolar RNA (snoRNA) | 71 | 287 | +216 |
| tRNA | 689 | 689 | 0 |
| Small nuclear RNA (snRNA) | 13 | 82 | +69 |
| rRNA | 15 | 15 | 0 |
| Other RNA | 394 | 221 | −173 |
| Total number of loci | 1359 | 5178 | +3819 |
| ShortStack-identified small RNAs (20-24 nt) | – | 35 846 | +35 846 |
| (C) Other gene types | | | |
| Pseudogenes | 924 | 952 | +28 |
| Transposable Element genes | 3903 | 3901 | −2 |
| (D) Newly instantiated loci | | | |
| Protein-coding genes (novel) | – | 635 | +635 |
| Protein-coding genes (split-inserted) | – | 97[b] | +97 |
| Novel transcribed regions | – | 508 | +508 |
| Total | – | 5080 | +5080 |
| (E) Obsolete loci | | | |
| Protein-coding genes (obsolete) | – | (−) 453[c] | −453 |
| Protein-coding genes (merge-obsolete) | – | (−) 40[d] | −40 |
| Total | – | (−) 493 | −493 |
| (F) Summary | | | |
| Total number of loci | 33 602 | 38 194 | +4592 |

[a] This includes 635 novel genes, 452 obsolete genes, 96 gene split events and 37 gene merge events.

[b] This includes 96 gene split events producing 97 novel split-inserted genes.

[c] This includes 452 obsolete protein-coding genes and 1 obsolete pseudogene.

[d] This includes 37 gene merge events producing 40 merge-obsolete genes.

**Table 2** Databases used in functional annotation

| Database | Description | URL |
|---|---|---|
| Priam | ENZYME-SPECIFIC PROFILES for metabolic pathway prediction | http://priam.prabi.fr/ |
| UniRef100 | UniProt Reference Clusters | http://www.uniprot.org/uniref/ |
| PFAM | Database of Protein Families | http://pfam.xfam.org/ |
| TIGRFAM | Curated Hidden Markov Models (HMMs) | http://www.jcvi.org/tigrfams |
| CAZY | Carbohydrate-Active enZYmes | http://www.cazy.org/ |
| CDD | Conserved Domain Database | https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml |
| TMHMM | Transmembrane helix prediction using Hidden Markov Models | http://www.cbs.dtu.dk/services/TMHMM/ |
| InterPro | Domain based classification of proteins | https://www.ebi.ac.uk/interpro/ |

may have more than one uORF, we adopted a nomenclature (e.g. AT1G67480.uORF1) that associated the uORF with its cognate protein-coding locus and assigned each uORF an isoform number. An additional 6680 loci with predicted uORFs bound by ribosomes at a significant level that have yet to be annotated are publicly available via a JBrowse track at Araport (Data S5) (Bailey-Serres, J., Bazin, J., and Girke, T., personal communication).

### Evaluation of annotation

A quantitative evaluation of the accuracy of the exon-intron structure is a critical step toward maintaining a gold standard annotation. Previously, TAIR used a five-star ranking system to describe the quality of a gene's structural annotation, with five-star the best. In the three- to five-star classes, all splice sites are supported but with decreasing quality of evidence from five to three. Two-star genes have some splice sites supported while one- and zero- star genes have no support for splice sites. Homologous and heterologous protein support is also considered in the ranking score. (https://www.arabidopsis.org/download_files/Genes/TAIR_gene_confidence_ranking/DOCUMENTATION_TAIR_Gene_Confidence.pdf). To evaluate the accuracy of the Araport11 annotation, we used the Annotation Edit Distance (AED) (Eilbeck *et al.*, 2009) to replace the five-star confidence classification used by TAIR. AED is generated by an actively maintained software MAKER-P and has been shown to closely correspond to the confidence classification used by TAIR (Campbell *et al.*, 2014). AED measures the consistency of gene models with the available nucleotide and protein sequence alignments. Each transcript was assigned a computed AED score, between 0 and 1, with 0 denoting complete agreement with the evidence and 1 indicating complete absence of supporting evidence. A comparison of the AED score distribution showed that overall Araport11 gene models are in closer concordance with the underlying evidence than TAIR10 models (Figure 2a).

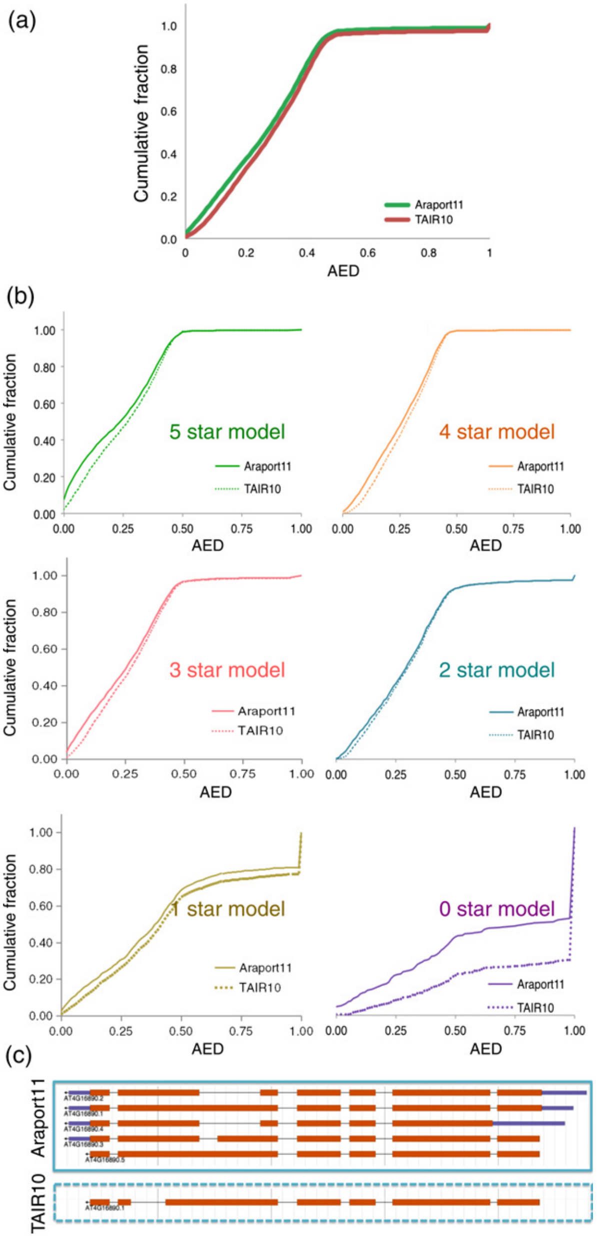


**Figure 2.** Accuracy of annotation.

(a, b) Annotation Edit Distance (AED) measures the consistency of transcript models with the underlying nucleotide/protein alignments. Lower AED scores suggest that the structures are in better agreement with the evidence. The cumulative fraction of AED scores provides a quantitative means to evaluate the annotation of all genes (a) which were broken down into 0–5 star ranks (b) according to TAIR10 annotation.

(c) An example of gene structure improvement made to AT4G16890.1, which was previously annotated with an erroneous 5th intron (reverse strand, transcription right to left). In addition to the gene structure correction, we also added novel isoforms assembled via the Araport11 pipeline.

We examined the cumulative fraction of AED in each category of the TAIR10 five-star ranking system (Figure 2b). Among the low-confidence genes, one obvious improvement was made for the zero star models, which, at the time of the TAIR10 annotation, did not have supporting evidence. Many updates were made to the TAIR10 transcripts that had been classified as two stars or one star, signifying that the previously available evidence did not completely cover the junctions or coding regions. One example, AT4G16890.1, is a two-star model previously annotated with an erroneous fifth intron. In Araport11, we corrected the gene structure (Figure 2c) and added novel isoforms detected in multiple tissues. Importantly, the improvements made in Araport11 were not limited to low-confidence models but rather across all annotation classes. This is because the highly ranked models had more expression evidence available as raw material for refinement. As another example, AT2G36480.1 and AT2G36485.1, two-five-star transcripts, were merged into a single transcript since the reads supporting the linking junction were present in multiple tissues.

To assess the completeness of the Araport11 annotation in comparison to TAIR10, we used the Benchmarking Universal Single-Copy Orthologs (BUSCO) toolkit (Simão *et al*., 2015) to examine the concordance of the two proteomes with reference to a curated set of plantae lineage-specific single-copy orthologs. The analysis revealed 945 complete single-copy BUSCOs in Araport11 compared with 942 in TAIR10, with marginal increases in the other categories (Data S8).

## Small RNAs

Small RNAs cover nearly 10% of the Arabidopsis genome yet are under-represented in TAIR10. Most efforts have focused on miRNAs that constitute a minor portion of the small RNA repertoire, resulting in a gap between the knowledge of small RNA expression and the annotation of small RNAs in Arabidopsis (Coruh *et al*., 2014). We analyzed small RNA-Seq datasets generated from root (Hsieh *et al*., 2009; Breakfield *et al*., 2012), leaf (Yu *et al*., 2013), aerial part (Fahlgren *et al*., 2010), flower (Lister *et al*., 2008; Cuperus *et al*., 2010; Law *et al*., 2013), embryo (Lu *et al*., 2012), and silique (Hardcastle *et al*., 2012). We identified *de novo* small RNA clusters using ShortStack (Axtell, 2013) that has been used to annotate small RNA genes in plant and animal species (Jex *et al*., 2014; Coruh *et al*., 2015; Lunardon *et al*., 2016). This method yielded a total of 35 846 clusters which predominantly produced 24-nt small RNAs (Figure 3a). Consistent with previous findings (Zhang *et al*., 2007; Law *et al*., 2013), most of these small RNA clusters are Pol-IV-dependent (Figure 3b) and completely recapitulate the 7631 200-bp static regions that generated small RNAs in stage-12 inflorescence (Law *et al*., 2013). Furthermore, the ShortStack-identified clusters overlapped with 95% of the loci that encode Pol-IV-dependent RNAs (P4RNAs), the precursors of small RNAs (Figure 3d). We hypothesize that the diverse tissue types used in our pipeline contribute to the clusters exclusively detected in

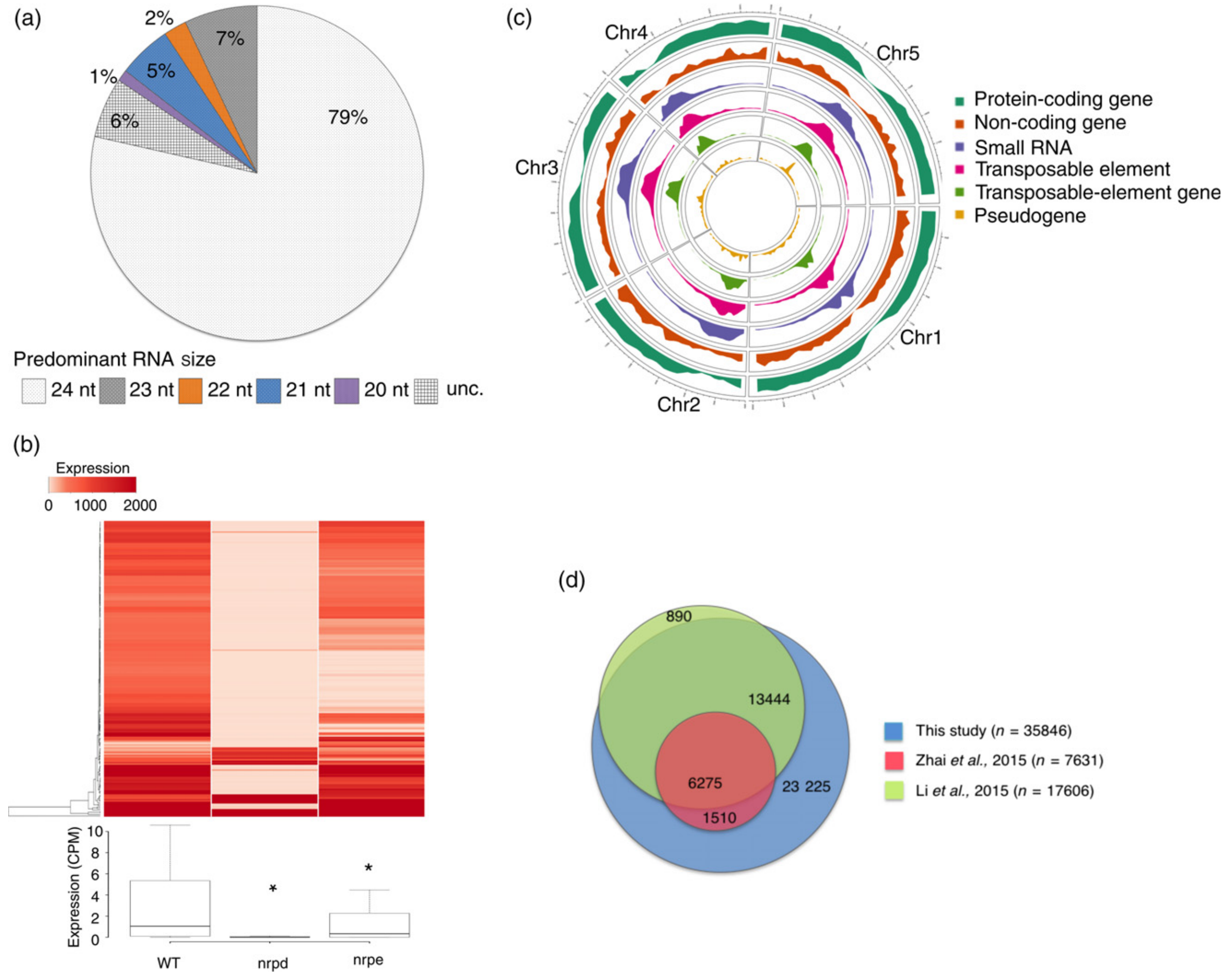


**Figure 3.** Properties of small RNAs.

(a) Size distributions within 35 846 small RNA loci. *unc.*, uncharacterized means that a majority RNA size is lacking for those loci.

(b) Heatmap and boxplot showing the small RNA levels (CPM, counts per million) in wild-type, *nrpd* (*pol-iv*) and *nrpe* (*pol-v*) flowers (* indicates significant reduction; *P*-value $<2.2 \times 10^{-16}$, Wilcoxon rank sum test).

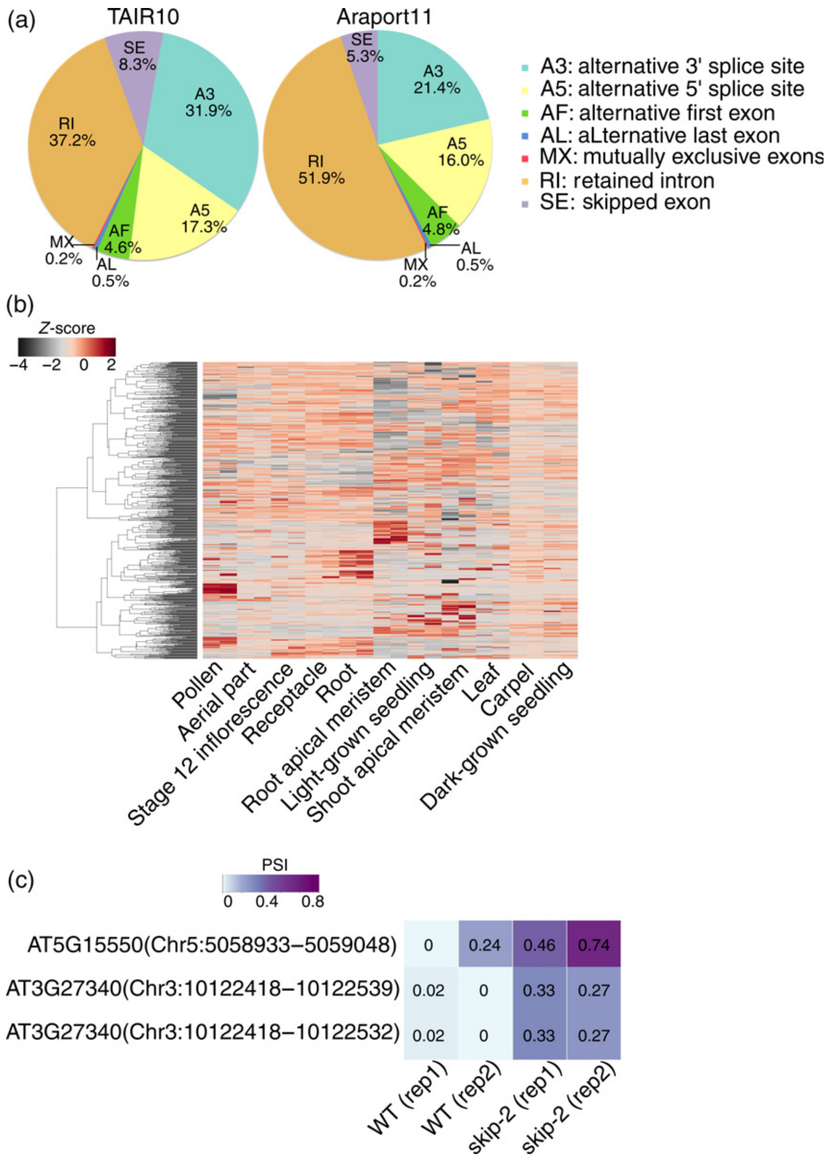
(c) Circular representation of genome-wide distribution of genomic features as indicated by the legend.

(d) A Venn diagram showing the overlap of small RNA generating loci (this study) and P4RNA loci (Li *et al.*, 2015; Zhai *et al.*, 2015).

this work. In line with this hypothesis, around ~43% of the small RNA clusters were detected only in non-flower tissues, while only inflorescence tissues were used in the two P4RNAs studies (Li *et al.*, 2015; Zhai *et al.*, 2015). In summary, we analyzed more than 124 million mapped small RNA-Seq reads and constructed a comprehensive set of small RNA generating loci. The small RNA annotations, including the genomic location and underlying metadata, can be interactively explored via JBrowse at Araport and are also available in Data S1.

## Diversity of alternative splicing

Nearly 40% of Araport11 protein-coding loci encode two or more splicing isoforms. This number is lower than the corresponding fraction in human (~81%), worm (~74%), and fly (~47%) (Gerstein *et al.*, 2014) and is expected to increase when additional treatments, genotypes and tissues are included in the analysis. While the majority of the protein-coding loci is transcribed into two to three different variants, there are 388 loci capable of encoding between seven and 27 isoforms. These 388 genes are overrepresented in the organic cyclic compound metabolic process (GO:1901360), nucleic acid metabolic process (GO:0090304), and cellular aromatic compound metabolic process (GO:0006725), which may imply a correlation between versatile transcript structures and these metabolic processes. In line with previous findings (Palusa *et al.*, 2007), the serine/arginine-rich proteins, a family of splicing regulators conserved in eukaryotes, are also overrepresented in these 388 loci with high numbers of isoforms.



| | WT (rep1) | WT (rep2) | skip-2 (rep1) | skip-2 (rep2) |
|---|---|---|---|---|
| AT5G15550(Chr5:5058933–5059048) | 0 | 0.24 | 0.46 | 0.74 |
| AT3G27340(Chr3:10122418–10122539) | 0.02 | 0 | 0.33 | 0.27 |
| AT3G27340(Chr3:10122418–10122532) | 0.02 | 0 | 0.33 | 0.27 |

**Figure 4.** Features of splicing events.
(a) Pie charts showing proportions of different classes of alternative splicing events in the Araport11 and TAIR10 annotations.
(b) Unsupervised clustering of splicing events across eleven tissues. The scale bar indicates *z*-scores of $\Psi$.
(c) Comparison of $\Psi$ for three experimentally verified retained introns between wild-type and splicing-defective *skip-2* plants. [Colour figure can be viewed at wileyonlinelibrary.com]

To examine the general nature of alternative splicing (AS), we further characterized the AS events using the SUPPA software (Alamancos *et al.*, 2015) and obtained a total of 19 915 splicing events in Araport11 gene models, over 65% of which were not annotated in TAIR10. The AS events consist of 10 375 retained introns (RI), 4230 alternative 3′ splice sites, 3172 alternative 5′ splice sites, 1051 skipped exons, 944 alternative first exons, 113 alternative last exons, and 30 mutually exclusive exons (Figure 4a). Unless ribosome occupancy data are available (see for example, exitrons, below), the possibility that some of the retained introns are found in incompletely processed nuclear transcripts cannot be excluded (Yu *et al.*, 2016).

To examine the variation of splicing events throughout development, we computed a 'percent splicing index' (PSI or $\Psi$) across the eleven tissues using SUPPA. PSI is a value that denotes the efficiency of a given splicing event and calculated as the fraction of the supporting isoform(s) to the total isoforms. The distribution of PSI values across tissues revealed the diverse nature of splicing activity. Unsupervised hierarchical clustering showed that many splicing events are specific to particular tissue type(s) and thus may be regulated in a tissue-specific manner (Figure 4b).

Various experimental approaches and computational analyses have conclusively shown that intron retention is the most prevalent AS event in Arabidopsis and rice, ranging from 40 to 65% (Iida *et al.*, 2004; Ner-Gaon *et al.*, 2004). This is different from human AS events, in which intron retention represents the least common (3.5%) type of event. We explored the dependency of retained introns on spliceosome activity by calculating the PSI values of RI events in wild type and a *skip-2* mutant defective in a splicing factor *SKIP* conserved in plants, yeast, and human. For the retained introns whose abundance increased in *skip-2* as verified by RT-PCR, Wang *et al.* found that their PSI values also increased (Wang *et al.*, 2012). This demonstrated the concordance of PSI values with experimental evidence of transcript abundance (Figure 4c). We extended the analysis to a genome-wide level and found that compared with wild type, the PSI values for 6641 RI events in *skip-2* were significantly higher (Wilcoxon rank sum test, *P*-value $<2.2 \times 10^{-16}$), suggesting a positive role of spliceosomes in the recognition and removal of these introns.

Nonsense-mediated decay (NMD) is a conserved eukaryotic quality-control mechanism that eliminates both normal and aberrant transcripts with premature termination codons (Shaul, 2015). It was assumed that intron retention would introduce premature termination codon and thus turnover by NMD-pathway. However, many intron-retained transcripts were not sensitive to NMD as their transcript abundance did not increase in a *upf1 upf3* mutant defective in NMD (Kalyna *et al.*, 2012). We globally examined the relations of NMD and intron retention by comparing the PSI values for wild-type and NMD-defective *upf1 upf3* seedlings using published RNA-Seq datasets (Drechsel *et al.*, 2013). Our analysis showed that the overall PSI values for 5499 RI events were reduced in *upf1 upf3* (Wilcoxon rank sum test, *P*-value = 0.0005112) compared to wild type, suggesting that the intron-retained transcripts were not necessarily targeted by NMD in Arabidopsis.

### The exonic introns are occupied by ribosomes

About 17% (1504) of the retained introns in Araport11 are exonic introns, or exitrons, which are 'introns with both splice sites inside an annotated coding exon' (Marquez *et al.*, 2012). This subset of non-constitutive introns, originally designated as 'cryptic exons' (Berget, 1995), tends to be shorter and flanked by weaker splice sites in vertebrates and plants (Stamm *et al.*, 2000; Zavolan *et al.*, 2003; Marquez *et al.*, 2015). The impact of exitrons on the coding sequence of the transcript is dependent on their lengths (Figure 5a–d). Nearly 60% of the exitrons have the lengths divisible by three ($\mathrm{EI}_{x3}$); therefore, retaining $\mathrm{EI}_{x3}$ does not alter the reading frame and will only increase the length of the coding sequence. Even for the exitrons with lengths not divisible by three (non-$\mathrm{EI}_{x3}$), retention does not necessarily introduce a stop codon downstream from the splice junctions (Figure 5c, d). Taken together, our analyses show that over 80% of the exitron-retained transcripts have longer coding sequences than their exitron-spliced counterparts. To further explore their potential for being translated, we compared the ribosome occupancies of exitrons, introns and coding exons within the same locus. The ribosome occupancy was calculated as the ratio of the relative abundance of each feature (intron, exitron, or coding exon) in the ribosome footprint library (Liu *et al.*, 2013) to those of the control RNA library. We found that in dark-grown seedlings, the ribosome occupancies of exitrons were significantly higher than those of introns (Figure 5e, f), confirming their contribution to the final protein product.

### Discovery of novel transcribed regions

The increasing sensitivity of transcript detection, both tiling arrays and now RNA-Seq, has revealed the presence of transcriptional events in regions of the genome not annotated with functional genes. It is a matter of debate as to whether this intergenic transcription has functional significance or is merely transcriptional noise arising from random initiation events (Ponting and Belgard, 2010). In addition to observing RNA-Seq support for almost all (96%) of the novel protein-coding genes predicted by the Gnomon (Souvorov *et al.*, 2010) and MAKER-P (Campbell *et al.*, 2014) pipelines, we identified an additional 508 intergenic regions of Araport11 to which RNA-Seq data could be mapped, designated these novel transcribed regions (NTRs) and assigned AGI locus identifiers to them. The majority (82.7%, 430/508) of the NTRs are expressed at a level of TPM (transcript per million) >1 in one or more tissues. Over 60% of the NTRs encode multi-exonic transcripts, some of which are transcribed into two or more isoforms with splice variants (Figure 6a). Furthermore, over a quarter of the NTRs contained a transcription start site reported in a large-scale analysis using whole root samples (Morton *et al.*, 2014), providing additional support for their structures. Most transcripts encoded from NTRs were expressed at low levels and/or only in a limited number of tissues (Figure 6b), which may explain why they escaped prior annotation.

Some novel transcripts have sequence similarity to known plant proteins. However, less than a quarter of these transcripts cover over 70% of the subject proteins, most of which are annotated in other species as uncharacterized. It is plausible that these NTRs that partially cover the subject proteins could encode small peptides, or represent pseudogenes or non-coding RNAs. To further explore these possibilities, we examined the ribosome occupancies of the NTRs using ribosome footprinting datasets generated from light- and dark-grown seedlings (Liu *et al.*, 2013; Juntawong *et al.*, 2014). In 55% (281/508) of the NTRs we observed aligned reads associated with ribosomes, implying a likelihood of translation. We also compared the NTRs with an atlas of about 90 000 conserved non-coding sequences (CNS) generated from a nine-way genome alignment of Brassicaceae species with *Arabidopsis thaliana* (Haudry *et al.*, 2013) and found 123 NTRs located in those conserved non-coding regions. Collectively, the attributes of these NTRs strongly suggest that they are not simply transcriptional noise. However, we have not specified their protein-coding or non-coding classification as diagnostic evidence is lacking. Whether they function as

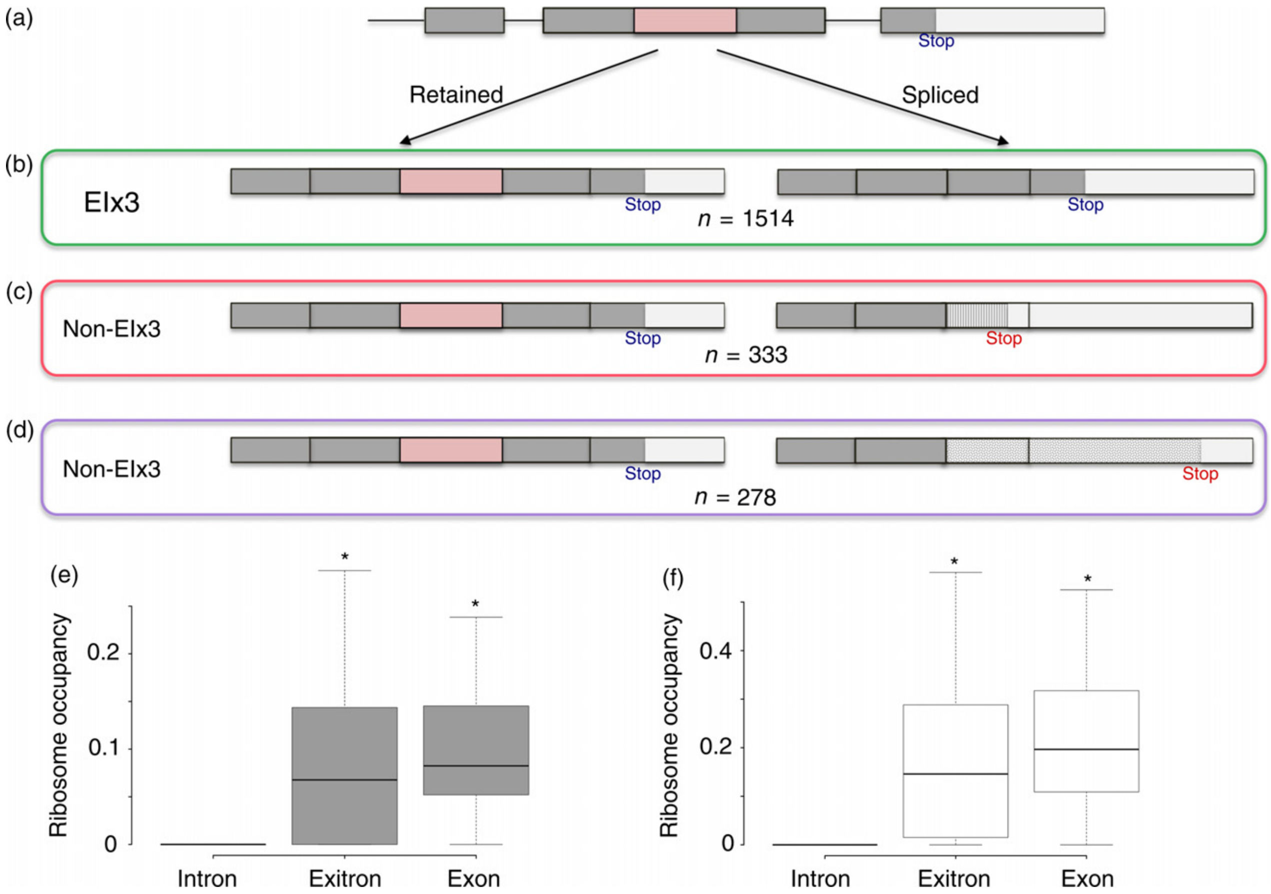


**Figure 5.** Exitrons.

(a) Both splice sites of an exitron (pink) are localized within the coding region (shaded rectangles). The impact of exitron retention or splicing on the CDS is dependent on the length of the exitron (b–d).

(b) When the exitron length is a multiple of three, splicing out an exitron results in the deletion of coding sequence.

(c, d) When the exitron length is not a multiple of three, splicing out an exitron changes the C-terminal coding sequence and the new stop codon could be upstream (c) or downstream (d) in comparison to the exitron-retained transcript.

(e, f) Boxplots showing the ribosome occupancies for dark-grown seedling without (d) or with (f) 4 h exposure to light (* indicates significant increase compared to intron; *P*-value $<2.2 \times 10^{-16}$, Wilcoxon rank sum test).

(small) peptide-coding RNAs or non-coding RNAs remains an open question for future studies.

### Expression dynamics across the 11 tissues

Over 99% of all genes, including 27 596 protein-coding genes, 944 pseudogenes, and 4560 non-coding RNAs were detected at the level above 0.1 TPM in at least one of the 113 RNA-Seq datasets, demonstrating that the use of extensive RNA-Seq data improved the overall sensitivity of detection in comparison with microarrays (Schmid *et al.*, 2005). The less than 1% of genes without detectable expression include 118 non-coding RNAs, 44 hypothetical proteins, 16 pseudogenes, and a few published/curated proteins. Inspection of additional RNA-Seq datasets showed that a subset of these genes was detectable under various experimental regimes such as hormone treatment, pathogen infection, and/or nutrient stress, suggesting the expression of these genes is specific to environmental stimuli.

To comprehensively explore the expression dynamics throughout development, we separately examined the expression profiles of protein-coding genes, non-coding RNAs, and pseudogenes. We will refer to expressed genes as those with greater than one TPM for the remaining analyses and describe the expression features of these three categories of transcripts in the following sections.

*Protein-coding genes.* Over 84% of protein-coding genes were expressed in one or more tissue. About 59% to 70% of the total genes were expressed in most tissues except for pollen, in which only 23% of genes were expressed. Furthermore, nearly 20% of the expressed genes (TPM >1) are shared by eleven tissues (Data S2) and this rises to 57% when pollen is excluded. This indicates a significant

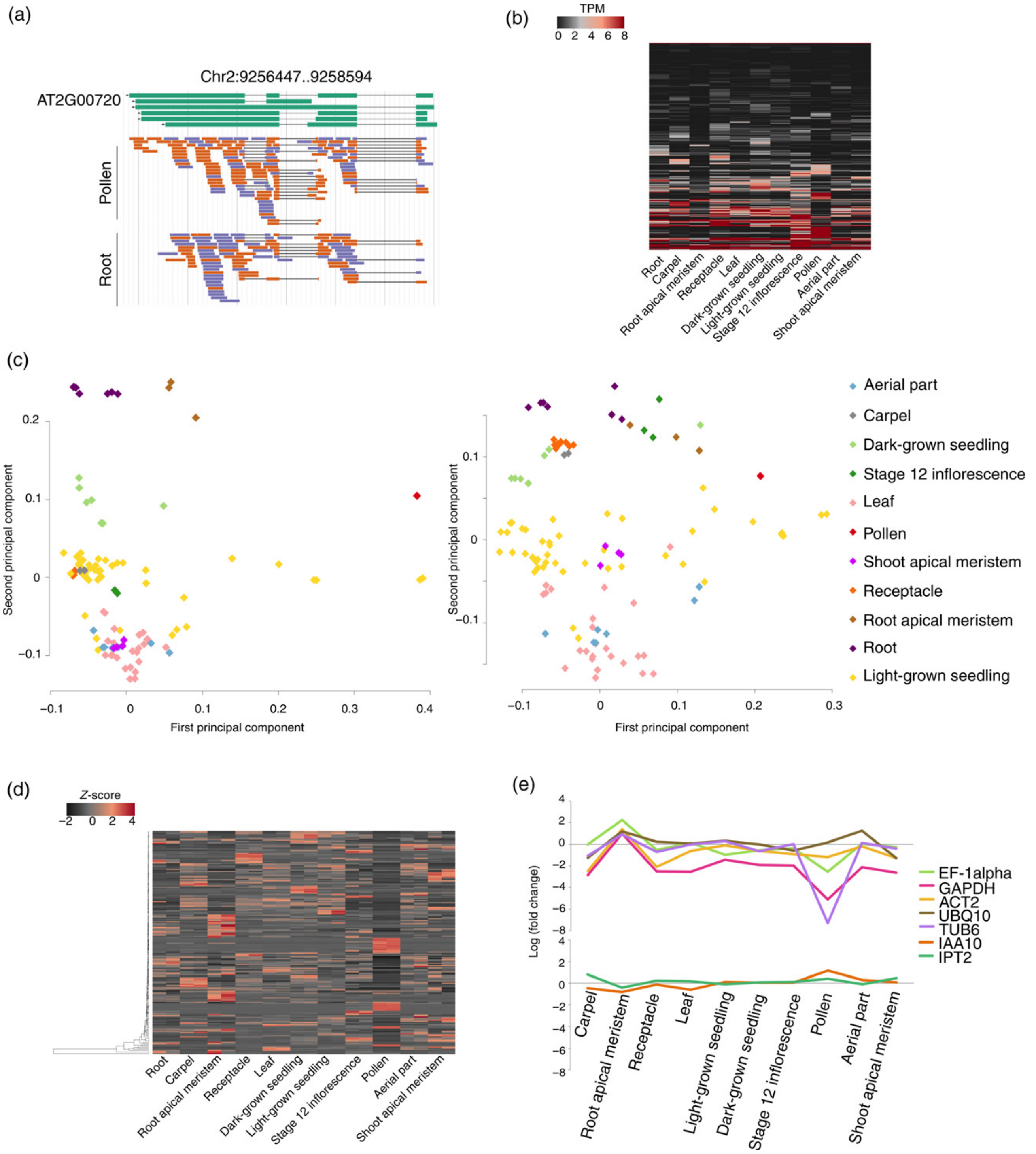


**Figure 6.** Genome-wide expression profile of several classes of transcripts.
(a) An example of novel transcribed region that encodes multiple splice variants.
(b) Heat map showing the expression levels of 508 novel transcribed regions across eleven tissues.
(c) PCA of protein-coding genes (left) and pseudogenes (right) on 113 RNA-Seq datasets revealed clusters sharing similar expression signatures. Note the loose cluster formed by light-grown seedling samples (yellow).
(d) Heat map showing the expression of non-coding RNAs across eleven tissues. The scale bar indicates the *z*-score of TPM.
(e) Expression levels of newly identified (IAA10, IPT2) or conventionally used (EF-1α, GAPDH, ACT2, UBQ10, TUB6) housekeeping genes. The isoforms in this plot are AT1G04100.1 (*IAA10*), AT1G04100.1 (*IPT2*), AT1G07920.1 (*EF-1α*), AT3G04120.1 (*GAPDH*), AT3G18780.2 (*ACT2*), AT4G05320.3 (*UBQ10*), and AT5G12250.1 (*TUB6*). For *ACT2* and *UBQ10* that have more than one isoform, the isoforms with highest average expression were depicted in this plot. Fold change is shown in $\log_2$ scale. [Colour figure can be viewed at wileyonlinelibrary.com]

overlap of genes expressed in morphologically distinct tissues. Still, each tissue maintains a characteristic transcriptional profile as reflected in principal component analysis (PCA). Leaf samples, for example, clustered together with each other and with samples made from the aerial part and shoot apical meristem, regardless of the growth conditions or harvesting age. Notably, light-grown seedling samples, the most widely used materials, appeared to be the most heterogeneous despite the similar plate-grown conditions and harvesting age (Figure 6c).

Next, we explored genes that were expressed in only one tissue. Pollen has the highest percentage of such tissue-specific genes despite the smallest size of transcriptome; the reproductive tissues and root have higher fractions of tissue-specific genes (Table S3). We performed GO enrichment analyses on each set of tissue-specific genes and found enriched GO terms in several tissues. For example, in the receptacle of stage 15 flowers where abscission zone cells responsible for programmed separation of plant organs were dominant, we found overrepresented GO terms related to cell death and response to stimulus in the tissue-specific genes. A complete list of tissue-specific genes and the enriched GO terms is shown in Data S3.

*Pseudogenes.* In Arabidopsis and rice, most pseudogenes are expressed but at a relatively low level compared to protein-coding genes; 2–36% of the pseudogenes in Arabidopsis were detected using EST, massively parallel signature sequencing (MPSS), and microarray evidence (Zou *et al.*, 2009). The RNA-Seq data reveal the expression of 98, 80 or 20% of all pseudogenes at levels of >0.1, >0.5 or >1.0 TPM in one or more tissues, similar to the widespread expression of pseudogenes in human (Kalyana-Sundaram *et al.*, 2012). Pseudogene expression also shows tissue-specific signatures, as anatomically-related samples cluster closer together in principal component analysis (Figure 6c).

*Non-coding RNAs.* While 4560 non-coding RNAs were detected at levels above 0.1 TPM, only 1644 reached a level above 1 TPM. The spatial expression of non-coding RNAs is more restricted as each non-coding RNA was expressed on average in five tissues, compared to eight tissues for protein-coding genes. Furthermore, many non-coding RNAs appeared to have peak expression in specific tissues (Figure 6d). Nearly 23% of the non-coding RNAs were expressed in only one tissue, compared with 6% of the protein-coding genes showed tissue specificity. These tissue-specific non-coding RNA were predominantly expressed in reproductive tissues and root. The preferential expression of non-coding RNAs in reproductive organs was also observed in *Drosophila* (Brown *et al.*, 2014).

Recent studies have provided some insight into the responsiveness of long intergenic RNA and natural antisense RNA to environmental stimuli (Liu *et al.*, 2012; Wang *et al.*, 2014). Therefore, we analyzed the expression profiles of non-coding RNAs annotated in Araport11 in response to abscisic acid (ABA) treated seedlings (GSE65016) and identified 101 non-coding RNAs that are differentially expressed in response to ABA treatment. We further examined the expression profile of these ABA-responsive non-coding RNAs after mild salt treatment (Sani *et al.*, 2013). A subset of these genes (AT4G09695, AT5G00990, and AT4G13505) is induced in the root sample, but not shoot, by salt treatment (Figure S3).

### Identification of housekeeping genes

Uniformly expressed genes are crucial internal references for large expression datasets, and can also shed light on the mechanisms underlying basic cell maintenance. It has recently become apparent that the expression patterns of many transcript isoforms change with the developmental stages and/or environmental conditions (Staiger and Brown, 2013). Therefore, our approach was to measure expression at the transcript level and define a housekeeping gene for which at least one transcript has fulfilled the following selection criteria. We used two RNA-Seq datasets per tissue type because the numbers of RNA-Seq datasets for each tissue ranged from two to 45, thus avoiding any analysis bias brought about by favoring the more highly represented tissues. To accommodate the heterogenous sources of the RNA-Seq datasets, we used RUVseq (Risso *et al.*, 2014; Peixoto *et al.*, 2015) to remove technical effects including batch effect, library preparation, and other nuisance effects. The adjustment of unwanted variation was followed by an ANOVA-like test in edgeR (McCarthy *et al.*, 2012) to identify transcripts whose expression levels were not significantly different among any of the tissues. We further restricted the variations between tissues to be less than three fold. Overall, we identified 705 transcripts encoded from 692 loci that are hereafter referred to housekeeping genes (Data S4). These housekeeping genes were enriched in pentatricopeptide repeat (IPR002885) and tetratricopeptide-like helical domain (IPR011990) motifs. GO terms associated with basal biological processes such as single-organism intracellular transport, establishment of localization in cell, vesicle-mediated transport, organelle organization, and cytoplasmic transport were also overrepresented. Additional housekeeping genes include elements involved in auxin signaling (*IAA10*), auxin transport, and cytokinin biosynthesis (*IPT2*), consistent with the essential roles of these two phytohormones. The loci that do not encode proteins may not have housekeeping functions in the traditional sense.

The housekeeping genes we identified include 12 of the previously reported top 100 most stably expressed genes in Arabidopsis (Czechowski *et al.*, 2005). The marginal overlap may be due to the differences in the source data (RNA-Seq vs. microarray) and the identification criteria. Furthermore, we found that the traditional reference genes, *ACT2* (AT3G18780), *TUB6* (AT5G12250), *EF-1α* (AT1G07920), *UBQ10* (AT4G05320), and *GAPDH* (AT3G04120) fluctuate across different tissues (Figure 6e) and are not present in our list. It is noteworthy that 297 of these 692 housekeeping genes have human orthologs also characterized as housekeeping genes with uniform expression across 16 human tissues using RNA-Seq data (Eisenberg and Levanon, 2013), indicating a considerable conservation of core processes between plant and mammalian systems.

## DISCUSSION

Continuous refinement and routine updates of annotation are prerequisites for correctly interpreting the functional elements of the genome. Even for well-annotated genomes such as Arabidopsis, fly, human, mouse, rice, and yeast, the advantages of RNA-Seq have offered an unprecedented resolution of their transcriptomes (Nagalakshmi *et al.*, 2008; Wilhelm *et al.*, 2008; Lu *et al.*, 2010; Graveley *et al.*, 2011). The new Araport11 annotation consists of 37 686 genes (27 655 protein-coding, 5178 non-coding, 952 pseudogenic, and 3901 transposable element-related loci) and 508 novel transcribed regions which altogether totaled to 38 194 loci (Table 1). The annotated loci span 67.7 Mb (56.6% of the genome sequence), an increase from 61.2 Mb (51.2%) in TAIR10.

The updated annotation revealed a sophisticated landscape of the transcriptional structures of genes, in terms of their splicing patterns and UTRs. The number of protein-coding genes expressing multiple splice isoforms increased from 21% in TAIR10 to 39% in Araport11. This figure is lower than a previous estimate of 54%, reported as 61% of intron-containing genes that are alternatively spliced (Marquez *et al.*, 2012). Therefore, we analyzed the same normalized cDNA libraries made from mixed flowers and seedlings (Marquez *et al.*, 2012) using the Araport11 pipeline and found ~16% of genes have more the one splice isoform. This indicated that the criteria used in the two studies, instead of the starting material and normalization procedures, plays a major role in the discrepancy. Nevertheless, we believe the annotation of splice isoforms is not saturated for two reasons. First, we only analyzed datasets generated from wild-type plants under normal growth conditions. Thus splicing events that occur only under stressed conditions and/or mutants may not be present and annotated. Second, PASA requires consensus splice sites (GT/GC donor with an AG acceptor, or the AT–AC U12-type dinucleotide pairs) and eight perfect matches on each side of a splice site to minimize incorrect transcript structures. Enforcing these filters is likely to have missed some splice variants with non-consensus splice sites.

Of the exitrons identified in this work, 209 overlapped with those previously reported (Marquez *et al.*, 2012). We investigated the 793 exitrons exclusively detected by Marquez *et al.* and found that in those loci, the exitron-retained transcripts, but not the exitron-spliced counterparts, were annotated in Araport11. This indicated that the exitron-retained transcripts in those loci were the major isoforms. Depending on the exon-oriented or intron-oriented perspective one might choose, the retained introns can be perceived as intronic exons, or cryptic exons (Berget, 1995), and splicing out of exitrons is a type of exon skipping behavior. The exonic nature of these exitrons is in line with our results showing that they are associated with ribosome at a level comparable to that of constitutive exons (Figure 5e, f).

Approximately one-third of the splicing activities occur in the UTRs and do not affect protein sequence. For example, phytochrome-interacting factor 3 (*PIF3*) encodes six transcript variants with identical coding sequence; three of which were specifically present in dark-grown seedling samples. The complex UTR landscape that we captured in Araport11 is in agreement with the reported heterogeneity of transcription start sites and polyadenylation sites (Sherstnev *et al.*, 2012; Morton *et al.*, 2014). In addition, we noted that there are approximately 3000 cases where genes overlap at their boundaries (Data S7), a phenomenon also observed in other eukaryotic species (Wang *et al.*, 2009). One challenge we encountered was that the short read lengths (≤100 bp) and non-strandedness of the 113 RNA-Seq datasets make a completely automatic pipeline inadequate to provide unambiguous gene boundaries for overlapping genes. We found that manual curation with consideration of additional datasets a necessary and worthwhile effort to avoid invasive UTR annotations. Likewise, increasing effort has also been placed on manual curation over the past decade to maintain the most trusted metazoan genomes (Schurch *et al.*, 2014).

We identified 508 novel transcribed regions based exclusively on RNA-Seq read mapping, most of which were not reported by comparative sequence analysis. This reinforces the importance of using transcriptome profiling to complement comparative analyses for genome annotation (Haudry *et al.*, 2013), similar to the phenomenon observed in fly (Graveley *et al.*, 2011). Interestingly, many of the NTRs that overlap with conserved non-coding sequences are also occupied by ribosomes in our analysis. Given that ribosome occupancy alone is insufficient to classify transcripts as coding or non-coding (Guttman *et al.*, 2013), the next challenge is to uncover whether these NTRs encode small peptides as found in Arabidopsis and other species (Cohen, 2014; Lauressergues *et al.*, 2015; Quinn and Chang, 2015).

Although the transcription and function of pseudogenes remain limitedly understood, recent results have started to challenge the old label of pseudogene as 'junk' DNA (Pink *et al.*, 2011; Milligan and Lipovich, 2014). In Arabidopsis, several pseudogenes and transposons are transcriptionally activated in response to stress conditions (Zeller *et al.*, 2009). Here, we showed that many of the pseudogenic loci are indeed transcribed, and the percentage of pseudogenes with more than one isoform (~30%) is comparable with that of protein-coding genes (~38%). Furthermore, we also observed uniform expression of a subset of pseudogenes and transposable element genes. Whether transcribed pseudogenes and transposable element genes can serve as a source of non-coding RNA, as shown in other species (Milligan and Lipovich, 2014), remain to be elucidated. Future functional studies will undoubtedly enhance our understanding of their roles in growth and development as well as stress responses.

Despite the substantial increase in structural diversity and expression profiling presented in Araport11, the annotation of *Arabidopsis thaliana* should be an ongoing project. Additional samples from developmentally specific stages (i.e. embryonic and senescent tissues) and environmental stimuli, and advancement in long read sequencing technology will most certainly result in the discovery of novel features as proven in the latest annotation release for maize, sorghum, human and mouse (Sharon *et al.*, 2013; Abdel-Ghany *et al.*, 2016; Bussotti *et al.*, 2016; Wang *et al.*, 2016). Furthermore, the small RNA annotation demonstrates that the integration of published datasets and analytic tools could provide a comprehensive and uniform annotation for the research community. Finally, advances in the bioinformatics analysis of the ribosome profiling data will provide valuable new insights into the translational activities and classification of both novel and existing transcripts.

## EXPERIMENTAL PROCEDURES

### Datasets

We assessed over 400 RNA-Seq datasets available as of September 2014 from the NCBI SRA. Among these, 113 Illumina-based datasets generated from wild-type Col-0 plants with untreated or mock-treated conditions were used in the annotation pipeline. A complete list of SRA accessions with detailed sample descriptions is available at Araport (https://www.araport.org/rna-seq-read-datasets-used-araport11). We describe the processing steps in detail in the Experimental procedures S1.

### Tissue-specific transcriptomes

To instantiate tissue-specific transcript isoforms and avoid possible chimeras between tissues, each tissue bin was assembled independently using Trinity (release 20140413) (Grabherr *et al.*, 2011) using both *de novo* and genome-guided modes. For the latter, only uniquely mapped reads were used by Trinity to guide the *de novo* assembly of overlapping read clusters at each locus. Both procedures used the default options except for two values: a minimum contig length of 183 bp, as 95% of Arabidopsis transcripts are greater than or equal to this length, and a maximum intron length of 2000 bp which is the 99th percentile for intron sizes in TAIR10 gene models. The *de novo* and genome-guided results generated in parallel were then collapsed, mapped back to the reference genome and merged into compatible transcript clusters by PASA (Haas *et al.*, 2003) using its alignment assembly module. These transcript clusters were sub-assembled into compatible transcript structures, which adhered to the following criteria: maximum intron length of 2000 bp, ≥90% of transcript aligned at ≥95% identity, both sides of the splice boundary are supported by at least 8 bp and the splice sites are canonical. We describe refining approaches in the Experimental procedures S1 'Structural annotation refinement'.

### Novel transcribed regions

Overlapping novel transcripts localized within Araport11 intergenic regions were clustered into novel transcribed regions. We used featureCounts (v1.5.0-p1) (Liao *et al.*, 2014) to calculate the count of reads aligned to the novel transcribed regions in the 113 RNA-Seq datasets. We only retained the regions that either contained more than one read count per million (CPM) mapped reads in any given tissue or encoded intron-containing transcript(s). For the latter, we did not apply an expression threshold because we expected these regions to be lowly expressed since they escaped prior annotation, thus imposing a expression threshold may be too stringent. Overall, we proposed 508 novel transcribed regions that fulfilled the above criteria.

### Annotation edit distance

We used MAKER (2.31.8) (Holt and Yandell, 2011) under default settings to compute AED for each transcript of protein-coding genes, non-coding RNAs, and pseudogenes. We used the default setting with the following supporting evidence: (i) PASA-assembled transcripts generated in this study, (ii) EST, full-length cDNAs and RNA-Seq data including non-control and additional ecotype samples provided by MAKER-P (Campbell *et al.*, 2014), and (iii) UniRef90 protein sequences for ten plant species (see 'Predicting proteins based on transcript models' in the Experimental procedures) aligned to TAIR9 genome sequence using the protein2genome option of MAKER.

### Small RNAs

We used Cutadapt (Martin *et al.* 2011) to trim the reads from 9 studies generated from multiple tissues (Table S2). We used ShortStack (Axtell, 2013) to identify *de novo* small RNA clusters (miRNA search disabled,—nohp). The reads from replicate samples were coupled in a single ShortStack run, and only the *DCL*-derived loci called out by ShortStack were retained for the remaining analyses. We used the default settings in ShortStack (v3.3) except that the clusters of small RNA must have one alignment per million mapped reads (—mincov 1 rpm). Next, we used BEDtools (Quinlan and Hall, 2010) to merge overlapping clusters from different studies and adopted the consensus regions as the small RNA reference loci (Data S1). ShortStack computes the dominant 20- to 24-nt small RNA size for each locus where mixtures of different sized RNA are possibly generated (Coruh *et al.*, 2015). During the merging step, we used the predominant RNA size within each consensus cluster, and designated the locus as 'uncharacterized' (Figure 3a) if a dominant size was not available. We computed the small RNA abundance using ShortStack count mode under default settings for wild-type, *nrpd (pol-iv)*, *nrpe (pol-v)* libraries (Figure 3b) made from stage 1–12 flowers (Law *et al.*, 2013). The circular visualization (Figure 3c) was generated using the circlize package (Gu *et al.*, 2014). We used BEDtools (Quinlan and Hall, 2010) to compare the genomic regions of the small RNA clusters with the P4RNA loci (Li *et al.*, 2015; Zhai *et al.*, 2015).

### Expression profiles

Gene expression values were computed on 113 RNA-Seq datasets at the gene and transcript levels using the default setting of Salmon (v0.6.0) quasi-mapping-based mode with the option (—useVBOpt) for the variational Bayesian EM algorithm (Patro *et al.*, 2015). The expression data are available at Araport and also via ThaleMine. Additional information about expression analyses is available in the Experimental procedures S1.

### Alternative splicing events

We calculated alternative splicing events on Araport11, TAIR10 and human (Ensembl release 83) annotation using the generateEvents option of SUPPA (v1.2b) (Alamancos *et al.*, 2015). To calculate the PSI value for each splicing event, we ran the psiPerEvent function of SUPPA using the Salmon-generated expression file as input. To compare the splicing activities across 11 tissues, we selected two datasets per tissue (detailed in Experimental procedures S1 'Identifying housekeeping genes') and calculated the variance of each PSI in those 22 datasets. For unsupervised hierarchical clustering, we generated the dendrogram using the default setting of hclust and plotted the heatmap (Figure 4b) using heatmap.2 in gplots. We removed events with missing PSI values and compared the PSI distributions of retained intron events between wild-type and mutant samples using Wilcoxon rank sum test (Figure 4c).

## ACCESSION NUMBERS

Araport11 has been published in NCBI GenBank under the accession numbers CP002684 (Chr1), CP002685 (Chr2), CP002686 (Chr3), CP002687 (Chr4), and CP002688 (Chr5). The corresponding NCBI BioProject and BioSample numbers encapsulating the above GenBank accessions are PRJNA10719 and SAMN03081427 respectively. The data can also be accessed via Araport and a list of links to each dataset is available in Data S5.

## ACKNOWLEDGEMENTS

We would like to thank Michael Axtell, Eduardo Eyras, and Carson Holt for helpful discussions on software utilization (ShortStack, SUPPA, and MAKER-P), Kim Pruitt and Ho-Ming Chen for constructive comments, and Jason Miller and Sergio Contrino for critical reading of the manuscript. This work was funded by the National Science Foundation (DBI-1262414).

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

**Figure S1.** A summary of gene structure statistics of Araport11 and comparison with TAIR10 showing improvements in both lengths and number of transcripts, CDSs, exons and introns.

**Figure S2.** Pseudogenic transcript variants encoded from the gene locus AT3G25495.

**Figure S3.** Heat map showing the expression levels of non-coding RNAs in shoot and root samples after salt priming treatment.

**Table S1.** The mapping results of the 113 RNA-seq datasets used in Araport11 annotation.

**Table S2.** A list of small RNA-seq data sets used in Araport11 annotation.

**Table S3.** Summary of protein-coding genes in eleven tissues.

**Data S1.** A GFF3 file containing 35 846 small RNA generating loci identified in this study.

**Data S2.** A list of 4577 genes expressed (TPM >1) in 11 tissues used in this study.

**Data S3.** Tissue-specific genes and enriched GO terms.

**Data S4.** Housekeeping loci identified in this study.

**Data S5.** Links to datasets available at Araport.

**Data S6.** A complete list of obsolete AGI identifiers.

**Data S7.** Overlapping Genes in Araport11 Annotation.

**Data S8.** BUSCO analysis and comparison of TAIR10 and Araport11 annotation.

**Experimental procedures S1.** The following additional analyses are detailed in this document: Processing RNA-seq libraries, Structural annotation refinement, Functional annotation, Assigning locus and isoform identifiers, Annotation of Novel Transcribed Regions, Identifying housekeeping genes, and Enrichment analysis.

## REFERENCES

**Abdel-Ghany, S.E., Hamilton, M., Jacobi, J.L., Ngam, P., Devitt, N., Schilkey, F., Ben-Hur, A. and Reddy, A.S.N.** (2016) A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.* **7**, 11706.

**Alamancos, G.P., Pagès, A., Trincado, J.L., Bellora, N. and Eyras, E.** (2015) Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA*, **21**, 1521–1531.

**Arabidopsis Genome Initiative** (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796–815.

**Axtell, M.J.** (2013) ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA*, **19**, 740–751.

**Berget, S.M.** (1995) Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270**, 2411–2414.

**Breakfield, N.W., Corcoran, D.L., Petricka, J.J., Shen, J., Sae-Seaw, J., Rubio-Somoza, I., Weigel, D., Ohler, U. and Benfey, P.N.** (2012) High-resolution experimental and computational profiling of tissue-specific known and novel miRNAs in Arabidopsis. *Genome Res.* **22**, 163–176.

**Brown, J.B., Boley, N., Eisman, R. *et al.*** (2014) Diversity and dynamics of the Drosophila transcriptome. *Nature*, **512**, 393–399.

**Bussotti, G., Leonardi, T., Clark, M.B. *et al.*** (2016) Improved definition of the mouse transcriptome via targeted RNA sequencing. *Genome Res.* **26**, 705–716.

**Campbell, M.S., Law, M., Holt, C. *et al.*** (2014) MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164**, 513–524.

**Cognat, V., Pawlak, G., Duchêne, A.-M. *et al.*** (2013) PlantRNA, a database for tRNAs of photosynthetic eukaryotes. *Nucleic Acids Res.* **41**, D273–D279.

**Cohen, S.M.** (2014) Everything old is new again: (linc)RNAs make proteins! *EMBO J.* **33**, 937–938.

**Coruh, C., Shahid, S. and Axtell, M.J.** (2014) Seeing the forest for the trees: annotating small RNA producing genes in plants. *Curr. Opin. Plant Biol.* **18**, 87–95.

**Coruh, C., Cho, S.H., Shahid, S., Liu, Q., Wierzbicki, A. and Axtell, M.J.** (2015) Comprehensive annotation of *Physcomitrella patens* small RNA loci reveals that the heterochromatic short interfering RNA pathway is largely conserved in land plants. *Plant Cell*, **27**, 2148–2162.

**Csorba, T., Questa, J.I., Sun, Q. and Dean, C.** (2014) Antisense COOLAIR mediates the coordinated switching of chromatin states at FLC during vernalization. *Proc. Natl Acad. Sci. USA*, **111**, 16160–16165.

**Cuperus, J.T., Carbonell, A., Fahlgren, N. *et al.*** (2010) Unique functionality of 22-nt miRNAs in triggering RDR6-dependent siRNA biogenesis from target transcripts in Arabidopsis. *Nat. Struct. Mol. Biol.* **17**, 997–1003.

**Czechowski, T., Stitt, M., Altmann, T., Udvardi, M.K. and Scheible, W.-R.** (2005) Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. *Plant Physiol.* **139**, 5–17.

**Deng, X.W., Caspar, T. and Quail, P.H.** (1991) cop1: a regulatory locus involved in light-controlled development and gene expression in Arabidopsis. *Genes Dev.* **5**, 1172–1182.

**Drechsel, G., Kahles, A., Kesarwani, A.K., Stauffer, E., Behr, J., Drewe, P., Rätsch, G. and Wachter, A.** (2013) Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant of the Arabidopsis steady state transcriptome. *Plant Cell*, **25**, 3726–3742.

**Eilbeck, K., Moore, B., Holt, C. and Yandell, M.** (2009) Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics*, **10**, 67.

**Eisenberg, E. and Levanon, E.Y.** (2013) Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574.

**Fahlgren, N., Jogdeo, S., Kasschau, K.D. *et al.*** (2010) MicroRNA gene evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*. *Plant Cell*, **22**, 1074–1089.

**Gerstein, M.B., Rozowsky, J., Yan, K.-K. *et al.*** (2014) Comparative analysis of the transcriptome across distant species. *Nature*, **512**, 445–448.

**Grabherr, M.G., Haas, B.J., Yassour, M. *et al.*** (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652.

**Graveley, B.R., Brooks, A.N., Carlson, J.W. *et al.*** (2011) The developmental transcriptome of *Drosophila melanogaster*. *Nature*, **471**, 473–479.

**Gu, Z., Gu, L., Eils, R., Schlesner, M. and Brors, B.** (2014) Circlize implements and enhances circular visualization in R. *Bioinformatics*, **30**, 2811–2812.

**Guttman, M., Mitchell, G., Pamela, R., Ingolia, N.T., Weissman, J.S. and Lander, E.S.** (2013) Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell*, **154**, 240–251.

**Haas, B.J., Delcher, A.L., Mount, S.M. *et al.*** (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666.

**Haas, B.J., Wortman, J.R., Ronning, C.M. *et al.*** (2005) Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. *BMC Biol.* **3**, 7.

**Hardcastle, T.J., Kelly, K.A. and Baulcombe, D.C.** (2012) Identifying small interfering RNA loci from high-throughput sequencing data. *Bioinformatics*, **28**, 457–463.

**Haudry, A., Platts, A.E., Vello, E. *et al.*** (2013) An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat. Genet.* **45**, 891–898.

**Hayden, C.A. and Jorgensen, R.A.** (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**, 32.

**Holt, C. and Yandell, M.** (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, **12**, 491.

**Hsieh, L.-C., Lin, S.-I., Shih, A.C.-C., Chen, J.-W., Lin, W.-Y., Tseng, C.-Y., Li, W.-H. and Chiou, T.-J.** (2009) Uncovering small RNA-mediated responses to phosphate deficiency in Arabidopsis by deep sequencing. *Plant Physiol.* **151**, 2120–2132.

**Iida, K., Seki, M., Sakurai, T., Satou, M., Akiyama, K., Toyoda, T., Konagaya, A. and Shinozaki, K.** (2004) Genome-wide analysis of alternative pre-mRNA splicing in *Arabidopsis thaliana* based on full-length cDNA sequences. *Nucleic Acids Res.* **32**, 5096–5103.

**Jex, A.R., Nejsum, P., Schwarz, E.M. *et al.*** (2014) Genome and transcriptome of the porcine whipworm Trichuris suis. *Nat. Genet.* **46**, 701–706.

**Jones, J.D.G. and Dangl, J.L.** (2006) The plant immune system. *Nature*, **444**, 323–329.

**Jones, A.M., Chory, J., Dangl, J.L., Estelle, M., Jacobsen, S.E., Meyerowitz, E.M., Nordborg, M. and Weigel, D.** (2008) The impact of Arabidopsis on human health: diversifying our portfolio. *Cell*, **133**, 939–943.

**Juntawong, P., Girke, T., Bazin, J. and Bailey-Serres, J.** (2014) Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis. *Proc. Natl Acad. Sci. USA*, **111**, E203–E212.

**Kalyana-Sundaram, S., Kumar-Sinha, C., Shankar, S. *et al.*** (2012) Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell*, **149**, 1622–1634.

**Kalyna, M., Simpson, C.G., Syed, N.H. *et al.*** (2012) Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic Acids Res.* **40**, 2454–2469.

**Kozomara, A. and Griffiths-Jones, S.** (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–D73.

**Krishnakumar, V., Hanlon, M.R., Contrino, S. *et al.*** (2015) Araport: the Arabidopsis information portal. *Nucleic Acids Res.* **43**, D1003–D1009.

**Laing, W.A., Martínez-Sánchez, M., Wright, M.A. *et al.*** (2015) An upstream open reading frame is essential for feedback regulation of ascorbate biosynthesis in Arabidopsis. *Plant Cell*, **27**, 772–786.

**Lamesch, P., Berardini, T.Z., Li, D. *et al.*** (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210.

**Lauressergues, D., Couzigou, J.-M., Clemente, H.S., Martinez, Y., Dunand, C., Bécard, G. and Combier, J.-P.** (2015) Primary transcripts of microRNAs encode regulatory peptides. *Nature*, **520**, 90–93.

**Law, J.A., Du, J., Hale, C.J., Feng, S., Krajewski, K., Palanca, A.M.S., Strahl, B.D., Patel, D.J. and Jacobsen, S.E.** (2013) Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. *Nature*, **498**, 385–389.

**Lee, E., Eduardo, L., Helt, G.A. *et al.*** (2013) Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.* **14**, R93.

**Li, S., Liberman, L.M., Mukherjee, N., Benfey, P.N. and Ohler, U.** (2013) Integrated detection of natural antisense transcripts using strand-specific RNA sequencing data. *Genome Res.* **23**, 1730–1739.

**Li, S., Vandivier, L.E., Tu, B., Gao, L., Won, S.Y., Li, S., Zheng, B., Gregory, B.D. and Chen, X.** (2015) Detection of Pol IV/RDR2-dependent transcripts at the genomic scale in Arabidopsis reveals features and regulation of siRNA biogenesis. *Genome Res.* **25**, 235–245.

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

**Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H. and Ecker, J.R.** (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523–536.

**Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., Arenas-Huertero, C. and Chua, N.H.** (2012) Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *Plant Cell*, **24**, 4333–4345.

**Liu, M.-J., Wu, S.-H., Wu, J.-F., Lin, W.-D., Wu, Y.-C., Tsai, T.-Y., Tsai, H.-L. and Wu, S.-H.** (2013) Translational landscape of photomorphogenic Arabidopsis. *Plant Cell*, **25**, 3699–3710.

**Lu, T., Lu, G., Fan, D. *et al.*** (2010) Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res.* **20**, 1238–1249.

**Lu, J., Zhang, C., Baulcombe, D.C. and Chen, Z.J.** (2012) Maternal siRNAs as regulators of parental genome imbalance and gene expression in endosperm of Arabidopsis seeds. *Proc. Natl Acad. Sci. USA*, **109**, 5529–5534.

**Lunardon, A., Forestan, C., Farinati, S., Axtell, M. and Varotto, S.** (2016) Genome-wide characterization of maize small RNA loci and their regulation in the required to maintain repression6-1 (rmr6-1) mutant and long-term abiotic stresses. *Plant Physiol.* **170**, 1535–1548.

**Marquez, Y., Brown, J.W.S., Simpson, C., Barta, A. and Kalyna, M.** (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res.* **22**, 1184–1195.

**Marquez, Y., Höpfler, M., Ayatollahi, Z., Barta, A. and Kalyna, M.** (2015) Unmasking alternative splicing inside protein-coding exons defines exitrons and their role in proteome plasticity. *Genome Res.* **25**, 995–1007.

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

**Matsui, A., Ishida, J., Morosawa, T. *et al.*** (2008) Arabidopsis transcriptome analysis under drought, cold, high-salinity and ABA treatment conditions using a tiling array. *Plant Cell Physiol.* **49**, 1135–1149.

**McCarthy, D.J., Chen, Y. and Smyth, G.K.** (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297.

**Milligan, M.J. and Lipovich, L.** (2014) Pseudogene-derived lncRNAs: emerging regulators of gene expression. *Front. Genet.* **5**. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4316772/ [Accessed October 21, 2016].

**Morton, T., Petricka, J., Corcoran, D.L., Li, S., Winter, C.M., Carda, A., Benfey, P.N., Ohler, U. and Megraw, M.** (2014) Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures. *Plant Cell*, **26**, 2746–2760.

**Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M.** (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.

**Ner-Gaon, H., Halachmi, R., Savaldi-Goldstein, S., Rubin, E., Ophir, R. and Fluhr, R.** (2004) Intron retention is a major phenomenon in alternative splicing in Arabidopsis. *Plant J.* **39**, 877–885.

**Okamoto, M., Tatematsu, K., Matsui, A. *et al.*** (2010) Genome-wide analysis of endogenous abscisic acid-mediated transcription in dry and imbibed seeds of Arabidopsis using tiling arrays. *Plant J.* **62**, 39–51.

**Palusa, S.G., Ali, G.S. and Reddy, A.S.N.** (2007) Alternative splicing of pre-mRNAs of Arabidopsis serine/arginine-rich proteins: regulation by hormones and stresses. *Plant J.* **49**, 1091–1107.

**Parry, G. and Estelle, M.** (2006) Auxin receptors: a new role for F-box proteins. *Curr. Opin. Cell Biol.* **18**, 152–156.

**Patro, R., Duggal, G. and Kingsford, C.** (2015) Accurate, fast, and model-aware transcript expression quantification with Salmon. *BioRxiv*, 021592. Available at: http://biorxiv.org/content/early/2015/10/03/021592 [Accessed January 12, 2016].

**Peixoto, L., Risso, D., Poplawski, S.G., Wimmer, M.E., Speed, T.P., Wood, M.A. and Abel, T.** (2015) How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets. *Nucleic Acids Res.* **43**, 7664–7674.

**Pink, R.C., Wicks, K., Caley, D.P., Punch, E.K., Jacobs, L. and Carter, D.R.F.** (2011) Pseudogenes: Pseudo-functional or key regulators in health and disease? *RNA*, **17**, 792. Available at: [Accessed October 21, 2016].

**Ponting, C.P. and Belgard, T.G.** (2010) Transcribed dark matter: meaning or myth? *Hum. Mol. Genet.* **19**, R162–R168.

**Provart, N.J., Alonso, J., Assmann, S.M. *et al.*** (2016) 50 years of Arabidopsis research: highlights and future directions. *New Phytol.* **209**, 921–944.

**Pruitt, K.D., Tatusova, T., Brown, G.R. and Maglott, D.R.** (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.

**Quinlan, A.R. and Hall, I.M.** (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

**Quinn, J.J. and Chang, H.Y.** (2015) Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* **17**, 47–62.

**Rédei, G.P.** (1975) Arabidopsis as a genetic tool. *Annu. Rev. Genet.* **9**, 111–127.

**Risso, D., Davide, R., John, N., Speed, T.P. and Sandrine, D.** (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902.

**Rosado, A., Li, R., van de Ven, W., Hsu, E. and Raikhel, N.V.** (2012) Arabidopsis ribosomal proteins control developmental programs through translational regulation of auxin response factors. *Proc. Natl Acad. Sci. USA*, **109**, 19537–19544.

**Sani, E., Herzyk, P., Perrella, G., Colot, V. and Amtmann, A.** (2013) Hyperosmotic priming of Arabidopsis seedlings establishes a long-term somatic memory accompanied by specific changes of the epigenome. *Genome Biol.* **14**, R59.

**Saul, H., Elharrar, E., Gaash, R. *et al.*** (2009) The upstream open reading frame of the Arabidopsis AtMHX gene has a strong impact on transcript accumulation through the nonsense-mediated mRNA decay pathway. *Plant J.* **60**, 1031–1042.

**Schmid, M., Markus, S., Davison, T.S. *et al.*** (2005) A gene expression map of Arabidopsis thaliana development. *Nat. Genet.* **37**, 501–506.

**Schurch, N.J., Cole, C., Sherstnev, A. *et al.*** (2014) Improved annotation of 3′ untranslated regions and complex loci by combination of strand-specific direct RNA sequencing, RNA-Seq and ESTs. *PLoS ONE*, **9**, e94270.

**Sharon, D., Tilgner, H., Grubert, F. and Snyder, M.** (2013) A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* **31**, 1009–1014.

**Shaul, O.** (2015) Unique aspects of plant nonsense-mediated mRNA decay. *Trends Plant Sci.* **20**, 767–779.

**Sherstnev, A., Duc, C., Cole, C., Zacharaki, V., Hornyik, C., Ozsolak, F., Milos, P.M., Barton, G.J. and Simpson, G.G.** (2012) Direct sequencing of Arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation. *Nat. Struct. Mol. Biol.* **19**, 845–852.

**Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M.** (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

**Souvorov, A., Kapustin, Y., Kiryutin, B., Chetvernin, V., Tatusova, T. and Lipman, D.** (2010) *Gnomon–NCBI eukaryotic gene prediction tool*, NCBI. Available at: http://www.ncbi.nlm.nih.gov/core/assets/genome/files/Gnomon-description.pdf.

**Staiger, D. and Brown, J.W.S.** (2013) Alternative splicing at the intersection of biological timing, development, and stress responses. *Plant Cell*, **25**, 3640–3656.

**Stamm, S., Zhu, J., Nakai, K., Stoilov, P., Stoss, O. and Zhang, M.Q.** (2000) An alternative-exon database and its statistical analysis. *DNA Cell Biol.* **19**, 739–756.

**Takahashi, H., Takahashi, A., Naito, S. and 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**, 2231–2241.

**Vidal, E.A., Moyano, T.C., Krouk, G., Katari, M.S., Tanurdzic, M., McCombie, W.R., Coruzzi, G.M. and Gutiérrez, R.A.** (2013) Integrated RNA-seq and sRNA-seq analysis identifies novel nitrate-responsive genes in *Arabidopsis thaliana roots*. *BMC Genomics*, **14**, 701.

**Wang, B.-B. and Brendel, V.** (2004) The ASRG database: identification and survey of Arabidopsis thaliana genes involved in pre-mRNA splicing. *Genome Biol.* **5**, R102.

**Wang, Z., Zhong, W., Mark, G. and Michael, S.** (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63.

**Wang, X., Wu, F., Xie, Q. *et al.*** (2012) SKIP is a component of the spliceosome linking alternative splicing and the circadian clock in Arabidopsis. *Plant Cell*, **24**, 3278–3295.

**Wang, H., Chung, P.J., Liu, J., Jang, I.-C., Kean, M.J., Xu, J. and Chua, N.-H.** (2014) Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in Arabidopsis. *Genome Res.* **24**, 444–453.

**Wang, B., Tseng, E., Regulski, M. *et al.*** (2016) Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* **7**, 11708.

**Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J. and Bähler, J.** (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.

**Yu, A., Lepère, G., Jay, F. *et al.*** (2013) Dynamics and biological relevance of DNA demethylation in Arabidopsis antibacterial defense. *Proc. Natl Acad. Sci. USA*, **110**, 2389–2394.

**Yu, H., Tian, C., Yu, Y. and Jiao, Y.** (2016) Transcriptome survey of the contribution of alternative splicing to proteome diversity in *Arabidopsis thaliana*. *Mol. Plant*, **9**, 749–752.

**Zavolan, M., Kondo, S., Schonbach, C., Adachi, J., Hume, D.A., Hayashizaki, Y. and Gaasterland, T., RIKEN GER Group and GSL Members** (2003) Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res.* **13**, 1290–1300.

**Zeller, G., Henz, S.R., Widmer, C.K., Sachsenberg, T., Rätsch, G., Weigel, D. and Laubinger, S.** (2009) Stress-induced changes in the *Arabidopsis thaliana* transcriptome analyzed using whole-genome tiling arrays. *Plant J.* **58**, 1068–1082. Available at: [Accessed October 21, 2016].

**Zhai, J., Bischof, S., Wang, H. *et al.*** (2015) A one precursor one siRNA model for Pol IV-dependent siRNA biogenesis. *Cell*, **163**, 445–455.

**Zhang, X., Henderson, I.R., Lu, C., Green, P.J. and Jacobsen, S.E.** (2007) Role of RNA polymerase IV in plant small RNA metabolism. *Proc. Natl Acad. Sci. USA*, **104**, 4536–4541.

**Zou, C., Lehti-Shiu, M.D., Thibaud-Nissen, F., Prakash, T., Buell, C.R. and Shiu, S.-H.** (2009) Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. *Plant Physiol.* **151**, 3–15.