# Supporting Information

## Hanada et al. 10.1073/pnas.1213958110

### SI Text

**Identification of Coding Small ORFs in Transcripts Verified by Full-Length cDNAs.** A total of 96,358 *Arabidopsis thaliana* full-length cDNAs were collected from the RIKEN *Arabidopsis* full-length (RAFL) cDNA resource (www.brc.riken.jp/lab/epd/catalog/cdnaclone.html) and NCBI GenBank (www.ncbi.nlm.nih.gov/genbank) (1, 2). Full-length cDNAs were cleaned by removing potential vector and poly-A tail contamination using the TIGR SeqClean tool (http://compbio.dfci.harvard.edu/tgi/software/). Cleaned full-length cDNAs were mapped to the TAIR8 *A. thaliana* genome using GMAP software (3). Of 96,358 full-length cDNAs, 93,919 were mapped with more than 95% similarity and 90% coverage. In particular, only 1,346 full-length cDNAs mapped to regions that do not have any coding genes in the TAIR8 version. The 1,346 full-length cDNAs were assumed to be mRNA-like transcripts. Among the transcripts, we searched for small ORFs (sORFs; 30–100 codons) with high-coding potential using the hexamer composition bias between CDS and NCDS.

Coding potential [$P(CDS|F)$] can be defined using the Bayes' theorem as follows:

$$P(CDS|F) = \frac{P(F|CDS)P(CDS)}{P(F|CDS)P(CDS) + P(F|NCDS)P(NCDS)}.$$

Here, $P(F|CDS)$ and $P(F|NCDS)$ are the probabilities that F is derived from CDS and NCDS in a genome, respectively. CDS and NCDS are defined to be *A. thaliana* coding and noncoding (intron) sequences, respectively. Prior probability of CDS $P(CDS)$ and NCDS $P(NCDS)$ are both set to be 0.5. $P(F|CDS)$ and $P(F|NCDS)$ are calculated from the Markov chain models of CDS and NCDS designed in previous reports. To determine whether a target sequence is coding or not, $P(CDS|F)$ was calculated on consecutive windows of 30 bp with a step size of 3 bp. The coding index (CI) for a given sequence is the summed posterior probabilities of all windows within a sequence. The CI is calculated by generating random CDS and NCDS training sequences. The CI is influenced by sequence length; therefore, 10,000 random sequences for each sequence length are generated, and the CI is calculated for all random sequences. When the CI of a sequence is higher than the bottom 1% CI of random CDSs and the top 1% CI in random NCDSs, the target sequence is defined to have a significant coding potential.

**Designed Arrays.** The number of targeted mRNA-like transcripts was 34,546 loci, including 26,254 annotated coding genes, 6,946 coding sORFs identified in previous reports, and 1,346 transcripts verified by full-length cDNAs. For the 34,546 mRNA-like transcriptional loci, the e-array web application (https://earray.chem.agilent.com) was used to identify 10 probe candidates of 60-mer sequences in each transcriptional unit. Probe candidates with more than 50 concatenated nucleotides in the other transcript sequences or probe candidates with simple repeats like AAAAAAA, TTTTTTT, GGGGGGG, and CCCCCCC were disregarded. When any probe candidates could not be identified in a transcript, we examined 60-mer sequences that did not have more than 50 concatenated nucleotides in the other transcript sequences or simple nucleotide repeats within 1 kb from the 3′ end in the transcript. Among these 60-mer sequences, a 60-mer sequence was chosen that was closer to 40% GC content as a probe sequence, because probe candidates identified by e-array web application have ~40% GC content. The specific 60-mer sequences in each transcript were spotted as probe sequences onto a 44K custom array platform produced by Agilent Technologies. However, 12 annotated coding genes and 69 coding sORFs did not have any candidate probe sequence as probes. Taken together, this array can examine expressions of 34,465 mRNA-like loci. Expression atlas for each mRNA-like locus can be examined at HANADB-AT (http://evolver.psc.riken.jp/seiken).

**RNA Samples and Hybridization to Arrays.** Sixteen organ samples from *A. thaliana* accession Col-0 were assayed. For 11 organs, three lines of plants were grown in soil under long-day conditions of 16 h of light at 22 °C. Samples were collected from each of juvenile rosette leaves, adult rosette leaves, cauline leaves, stems, young flower buds (sepals were closed and petals are not visible), mature flower buds (petals are visible), flowers, young siliques filled with white seeds, mature siliques filled with green seeds, old yellowing siliques (2 mo old), and senescence rosettes. Root tissues were collected from 2-wk-old plants growing on 1% agar plates containing Murashige and Skoog (MS). Calli were collected from cultured callus induction medium containing B5 medium supplemented with 20 g/L glucose, 0.5 mg/L 2,4-D, and 0.1 mg/L kinetin. Dry seeds were defined as 4-wk-old after-ripened seeds. Imbibed seeds are collected from seeds imbibed on moistened paper for 24 and 48 h under continuous light at 22 °C. Thus, there were 16 organ samples (dry seeds, 24-h-imbibed seeds, 48-h-imbibed seeds, callus, juvenile rosette, adult rosette, senescence leaves, cauline leaves, stems, root, young buds, mature flower buds, flowers, young siliques, mature siliques and old siliques) processed in triplicate.

For light irradiation, WT seeds were plated on MS plates. For the white-light time course experiment, the plates were placed at 4 °C in darkness for at least 2 d before receiving 6-h red light irradiation to synchronize germination. Plates were then placed in the dark for 3 d and before being exposed to white fluorescent light (100 $\mu mol \cdot m^{-2} \cdot s^{-1}$) for 1, 6, and 24 h. For the monochromatic light experiment, WT seedlings were grown in continuous red light, blue light, far-red light, or darkness for 6 d after red light irradiation. The light intensity used was 0.17 $\mu mol \cdot m^{-2} \cdot s^{-1}$ for red light, 1.16 $\mu mol \cdot m^{-2} \cdot s^{-1}$ for blue light, and 0.53 $\mu mol \cdot m^{-2} \cdot s^{-1}$ for far-red light. There were eight light irradiation samples (0 h white, 1 h white, 6 h white, 24 h white, dark, blue, far-red, and red lights) performed in triplicate.

All abiotic stress samples were collected from plants grown for 2 wk on MS plates under 2 and 6 h of abiotic stress conditions under continuous light at 22 °C. For drought stress treatment, plants were transferred and dehydrated onto dried filter paper in covered plastic dishes. For heat and cold stress treatments, covered plastic dishes containing plants were transferred directly from 22 °C to 37 °C and 2 °C, respectively. For salinity stress treatment, plants were transferred onto filter paper moistened on MS plates including 200 mM NaCl.

Approximately 500 μg total RNA was extracted from each sample in triplicate using TRIzol (Invitrogen). cRNA amplification and fluorescence labeling were performed using the Agilent Low RNA Input Linear Amplification Kit (Agilent Technology), according to the manufacturer's protocol. A primer containing poly dT and a T7 polymerase promoter was annealed to the poly(A)+ RNA. First- and second-strand cDNA was reverse-transcribed from 500 ng total RNA using MMLV-RT enzyme. Cyanine 3–labeled cRNA was synthesized using T7 RNA polymerase. Labeled cRNA was purified by an RNeasy Mini Kit (Qiagen). Hybridization was performed using the in situ Hybridization Kit Plus (Agilent Technologies) and GeneChip Eukaryotic Hybridization Control Kit (Affymetrix) on our array (Agilent) and ATH1 GeneChips (Affymetrix), respectively. For the custom and ATH1 arrays, the hybridized and washed material on each glass slide was scanned by an Agilent G2505 B DNA microarray scanner (Agilent Technologies) and Fluidics Station 400 (Affymetrix), respectively. Data analysis was performed by Feature Extraction and Image Analysis software (Agilent Technologies) and Microarray Suite (Affymetrix).

***Arabidopsis* Gene Expression Atlas.** The array intensities were processed using the Bioconductor (www.bioconductor.org) package in the R software environment (www.r-project.org). Hybridization intensities in our arrays were normalized among different arrays by quantile normalization. Curve fitting of lower and higher Gaussian distribution by expectation-maximization was performed on the log10-transformed expression intensities of annotated genes using the R library "Mclust." The top 5% values in lower distribution of expression intensities were inferred using the R function "dnorm" in each of 33 conditions,

The ATH1 array is the gold standard microarray platform. In this array, hybridization intensities were adjusted to reduce background signals using the MAS5 function. The ATH1 platform includes an expression atlas of 20,033 loci among our targeted 34,465 transcriptional loci. To examine whether our arrays produced a comparable atlas or not, the expression intensities in 20,033 genes inferred by our array were compared with those of ATH1 array for the same sample (2-wk-old seedlings). The expression intensities between our array and ATH1 array were compared.

**RT-PCR and Quantitative Real-Time PCR Analysis.** In RT-PCR, RNA was mixed from all samples used in this study. cDNA was synthesized using a QuantiTect Reverse Transcription Kit (QIAGEN) according to the manufacturer's instructions. PCR was performed with Ex Taq polymerase (TAKARA BIO) using the following program: 2 min at 94 °C, 40 cycles of 30 s at 94 °C, 30 s at 55 °C, 30 s at 72 °C, and 5 min at 72 °C. The primers were designed to amplify the entire ORF. RT-PCR products were analyzed by the electrophoresis system MultiNA (SHIMADZU) with a DNA-1000 kit according to the manufacturer's protocol.

In quantitative real-time PCR, RNAs were extracted from 2-wk-old WT and overexpression mutants of five sORFs (sORF2146, sORF2989, sORF5697, sORF2686, and sORF3416). Real-time PCR was performed according to the protocol of the StepOnePlus (Life Technology) using SYBR Green Realtime PCR Master Mix (TOYOBO). All values were normalized to the expression of the *SDG16* gene (AT4G27910) as the internal control. The primer sets used in this study are shown as follows (sORF2146_Fw: TTCGGAGAGTGTTCAGTGCAG, sORF2146_Re: TGGTTACTCGATAGATCTTCCCC, sORF2989_Fw: CTTTGAAAGATTTGTGGTTCCG, sORF2989_Re: TCGGTCCCCATATCAATCCT, sORF5697_Fw: GGTCGGAGTAGCGTAACAAGAA, sORF5697_Re: ATCGTAAGTTGGGACGACGA, sORF2686_Fw: TCCGGTCTTTTGGGTACTGA, sORF2686_Re: AAACGAGTCAAAAGCATTATTCTG, sORF3416_Fw: ATGGGTCACTATTGTTTTGTCTTG, sORF3416_Re: AGGACCCCATGTGAAGATTT, SDG16_Fw: GAGAGAAGCACGTTATCGCA, SDG16_Re: GAATGATTGATGAGCCGAGC).

**Sequence Analyses of Coding sORFs.** The following genomes were used to assess conservation of coding sORFs: *Physcomitrella patens* (http://genome.jgi-psf.org/Phypa1_1); *Selaginella moellendorffii* (http://genome.jgi-psf.org/Selmo1); *Zea mays* (www.maizesequence.org); *Sorghum bicolor* (http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html); *Brachypodium distachyon* (www.jgi.doe.gov); *Oryza sativa* (http://rapdb.dna.affrc.go.jp/download/irgsp1.html); *Mimulus guttatus* (http://genome.jgi-psf.org/mimulus); *Vitis vinifera* (www.genoscope.cns.fr/externe/GenomeBrowser/Vitis); *Ricinus communis* (http://castorbean.jcvi.org/downloads.php); *Manihot esculenta* (ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v5.0/Mesculenta); *Populus trichocarpa* (ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v5.0/Ptrichocarpa); *Cucumis sativus* (ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v5.0/Csativus); *Glycine max* (ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v4.0/Gmax); *Medicago truncatula* (www.medicago.org/genome/downloads/Mt3); *Carica papaya* (http://asgpb.mhpcc.hawaii.edu/papaya/); and *Arabidopsis lyrata* (http://genome.jgi-psf.org/Araly1). Sequence pairs were regarded as homologous sequences if they had less than one e-value by BLAST search (tblastn) (4). When there was a stop codon in the translated genomic sequence match, the 5′ sequence not truncated by the stop codon was used as the homologous sequence. These conserved pairs were aligned by CLUSTALW (5), and the synonymous and nonsynonymous substitution rates ($K_a$ and Ks) were calculated using PAML. To determine the null distribution of $K_a$/Ks values in the same procedure, we generated 1,000 random sequences (0.1 MB) to be the similar nucleotide composition of all coding sORFs. Sequence pairs were regarded as similar sequences if they had less than 10 e-values by BLAST search (tblastn). When there was a stop codon in the translated genomic sequence match, the 5′ sequence not truncated by the stop codon was used as the homologous sequence. These conserved pairs were aligned by CLUSTALW, and the synonymous and nonsynonymous substitution rates ($K_a$ and Ks) were calculated using PAML. The median $K_a$/Ks ratio (0.32) was defined to be the $K_a$/Ks ratio that represents neutrality in our procedure. To determine that the $K_a$/Ks ratio was significantly <0.32, a likelihood ratio–based procedure was applied to the sequence pairs. For each pair, two maximum likelihood values were calculated with the $K_a$/Ks ratio fixed at 0.32 and with the $K_a$/Ks ratio as a free parameter. The ratio of the maximum likelihood values was then compared with the $\chi^2$ distribution. To correct for multiple testing, the false discovery rate (FDR) was estimated by using Q-VALUE software. The null hypothesis was rejected if FDR values were <0.05.

To define gene families among coding sORFs, all-against-all similarity searches were conducted using a BLAST search (blastp) and an e-value cutoff of 10. Based on the blast scores, gene clusters representing gene families were generated with the "mcxdeblast" and "mcl" programs of the Markov clustering package (6). Consequently, median and minimum sequence similarity within clusters is 84% and 40%, respectively.

**Construction of Overexpression Mutants.** To make transgenic plants that overexpress each of the coding sORFs, we designed forward primers from the region around the start codon and reverse primers from the region around the stop codon. Each sORF amplified by PCR was introduced into pMDC32, which includes a double 35S promoter (7). The recombinant binary vector was then introduced into the *Agrobacterium tumefaciens* (strain GV3101). *Agrobacterium* was infected into *Arabidopsis* using the floral dip method (8). Transformed seedlings were selected on a medium containing half-strength MS medium containing 20 mg/L hygromycin B and 100 mg/L cefotaxime, and 17 randomly chosen seedlings were transferred to soil for each coding sORFs. Visible phenotypes, such as plant color, flowering time, and fertility were monitored in all constructed overexpression lines. When more than three overexpression lines showed the same phenotypes, the transformed sORF was considered responsible for the morphologies.

1. Seki M, et al. (2002) Functional annotation of a full-length Arabidopsis cDNA collection. *Science* 296(5565):141–145.
2. Yamada K, et al. (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* 302(5646):842–846.
3. Wu TD, Watanabe CK (2005) GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21(9):1859–1875.
4. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.
5. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22(22):4673–4680.
6. Dongen SV (2000) Graph clustering by flow simulation. PhD thesis (Univ of Utrecht, Utrecht, The Netherlands).
7. Curtis MD, Grossniklaus U (2003) A gateway cloning vector set for high-throughput functional analysis of genes in planta. *Plant Physiol* 133(2):462–469.
8. Clough SJ, Bent AF (1998) Floral dip: A simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant J* 16(6):735–743.

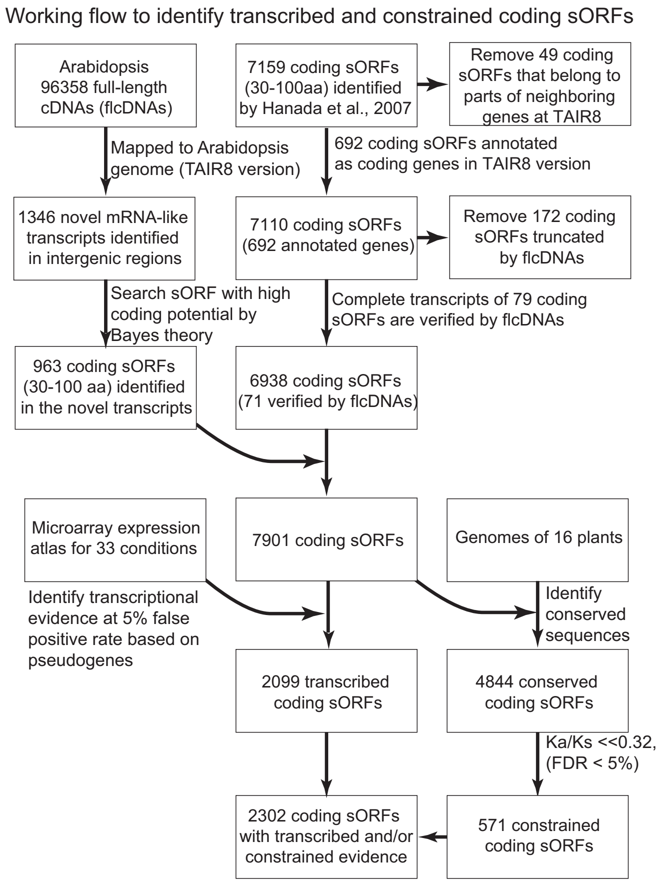

Working flow to identify transcribed and constrained coding sORFs
Arabidopsis 96358 full-length cDNAs (flcDNAs)
7159 coding sORFs (30-100aa) identified by Hanada et al., 2007
Remove 49 coding sORFs that belong to parts of neighboring genes at TAIR8
Mapped to Arabidopsis genome (TAIR8 version)
692 coding sORFs annotated as coding genes in TAIR8 version
1346 novel mRNA-like transcripts identified in intergenic regions
7110 coding sORFs (692 annotated genes)
Remove 172 coding sORFs truncated by flcDNAs
Search sORF with high coding potential by Bayes theory
Complete transcripts of 79 coding sORFs are verified by flcDNAs
963 coding sORFs (30-100 aa) identified in the novel transcripts
6938 coding sORFs (71 verified by flcDNAs)
Microarray expression atlas for 33 conditions
7901 coding sORFs
Genomes of 16 plants
Identify transcriptional evidence at 5% false positive rate based on pseudogenes
Identify conserved sequences
2099 transcribed coding sORFs
4844 conserved coding sORFs
Ka/Ks <<0.32, (FDR < 5%)
2302 coding sORFs with transcribed and/or constrained evidence
571 constrained coding sORFs

**Fig. S1.** Working flow to identify transcribed and constrained coding sORFs. Identified coding sORFs that have a qualifying expression atlas in 33 conditions and sequence conservation in 16 plant species.

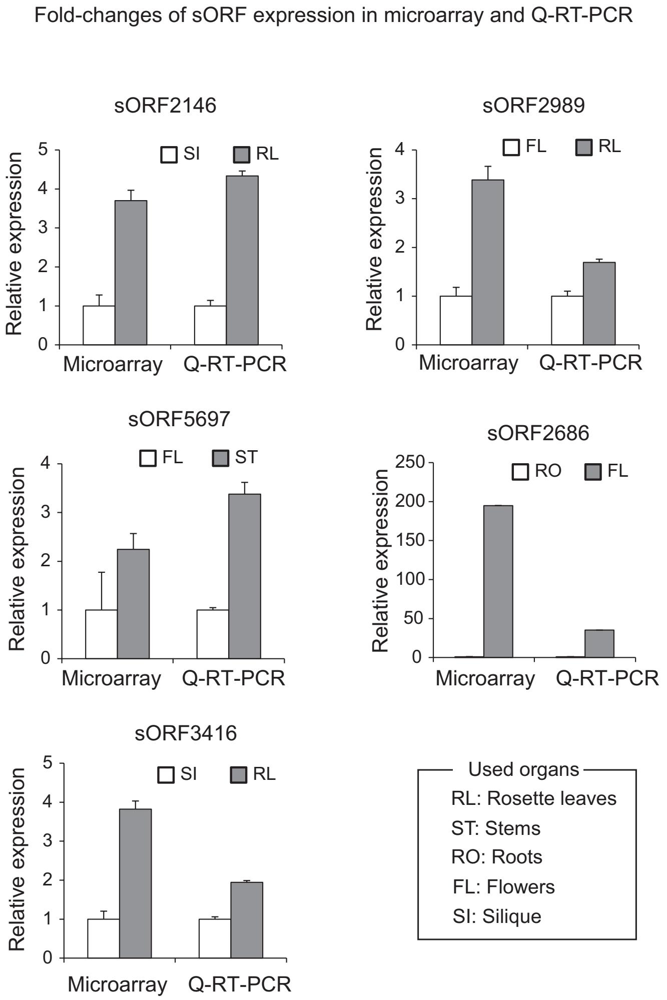


**Fig. S2.** Fold-changes of sORF expression in microarray and quantitative RT-PCR. Fold-changes of expression intensities in manually chosen five sORFs (sORF2146, sORF2989, sORF5697, sORF2686, and sORF3416) are compared in real-time quantitative RT-PCR and our microarray among five organs.

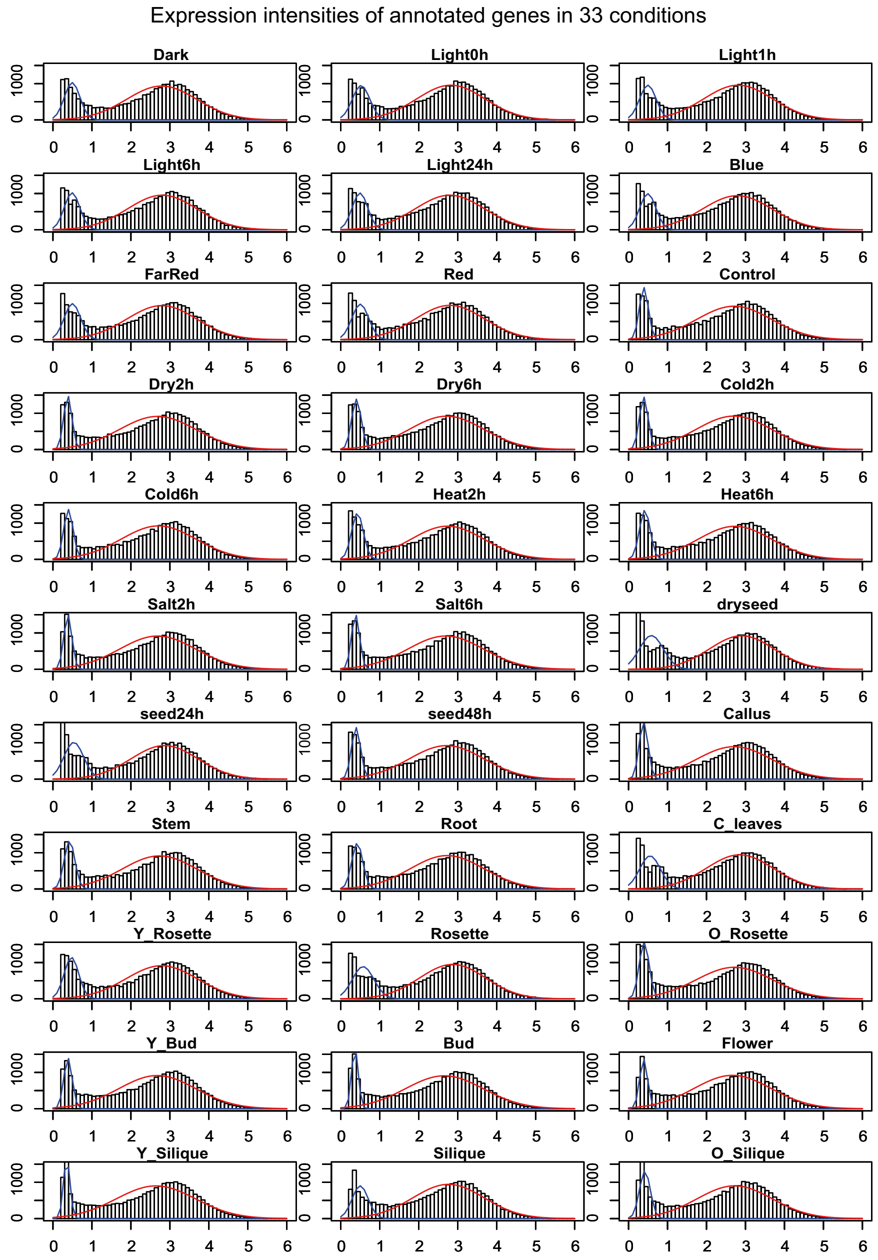


**Fig. S3.** Expression intensities of annotated genes in 33 conditions. *x* axis represents log10 values of annotated genes in 33 conditions (continuous dark light, white 0 h, white 1 h, white 6 h, white 24 h, continuous blue light, continuous far-red light, continuous red light, control, drought 2 h, drought 6 h, cold 2 hoursm cold 6 h, heat 2 h, heat 6 h, salt 2 h, salt 6 h, dry seeds, 24-h-imbibed seeds, 48-h-imbibed seeds, callus, stems, root, cauline leaves, juvenile rosette, adult rosette, senescence leaves, young buds, mature flower buds, flowers, young siliques, mature siliques, and old siliques). In each condition, expression intensities of annotated genes were fitted to lower (blue solid line) and higher (red solid line) Gaussian distribution.

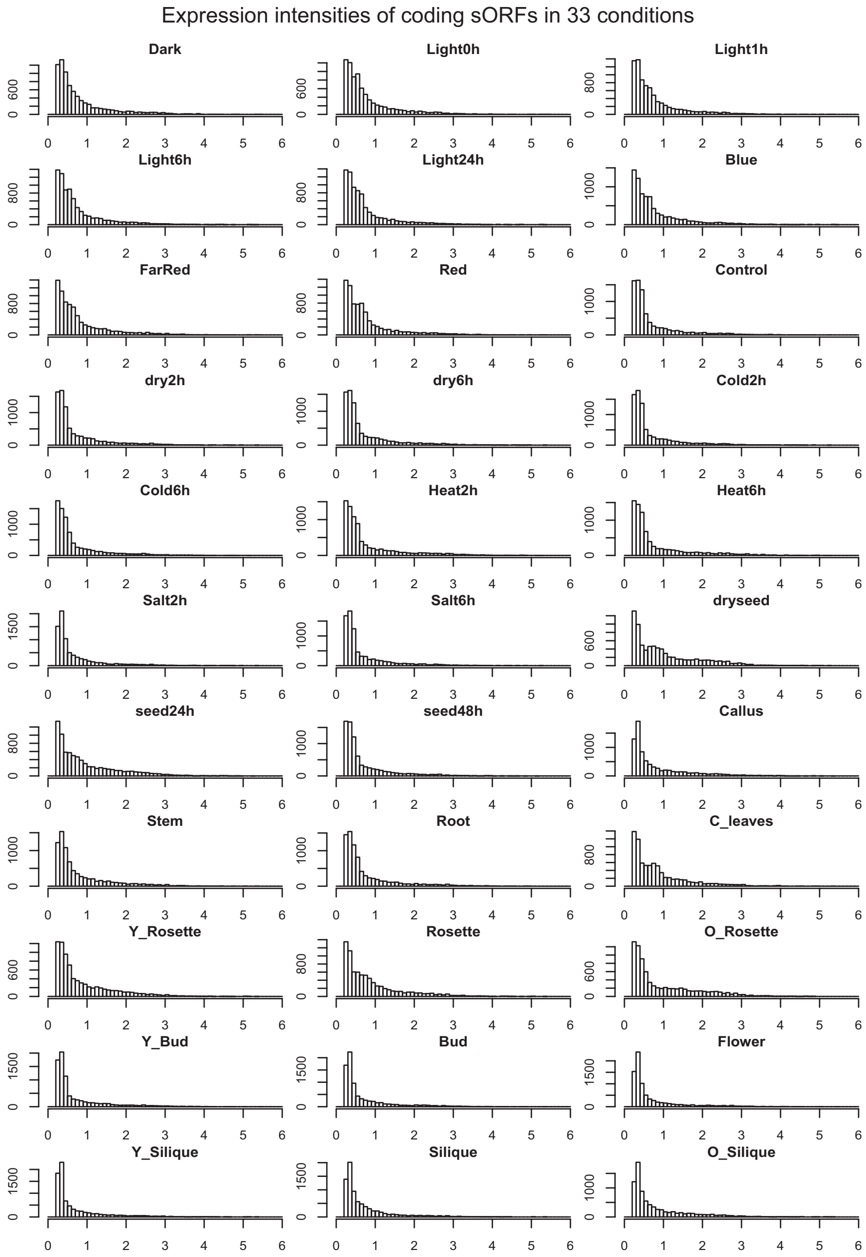


**Fig. S4.** Expression intensities of coding sORFs in 33 conditions. *x* axis represents log10 values of coding sORFs in 33 conditions (continuous dark light, white 0 h, white 1 h, white 6 h, white 24 h, continuous blue light, continuous far-red light, continuous red light, control, drought 2 h, drought 6 h, cold 2 hoursm cold 6 h, heat 2 h, heat 6 h, salt 2 h, salt 6 h, dry seeds, 24-h-imbibed seeds, 48-h-imbibed seeds, callus, stems, root, cauline leaves, juvenile rosette, adult rosette, senescence leaves, young buds, mature flower buds, flowers, young siliques, mature siliques, and old siliques).

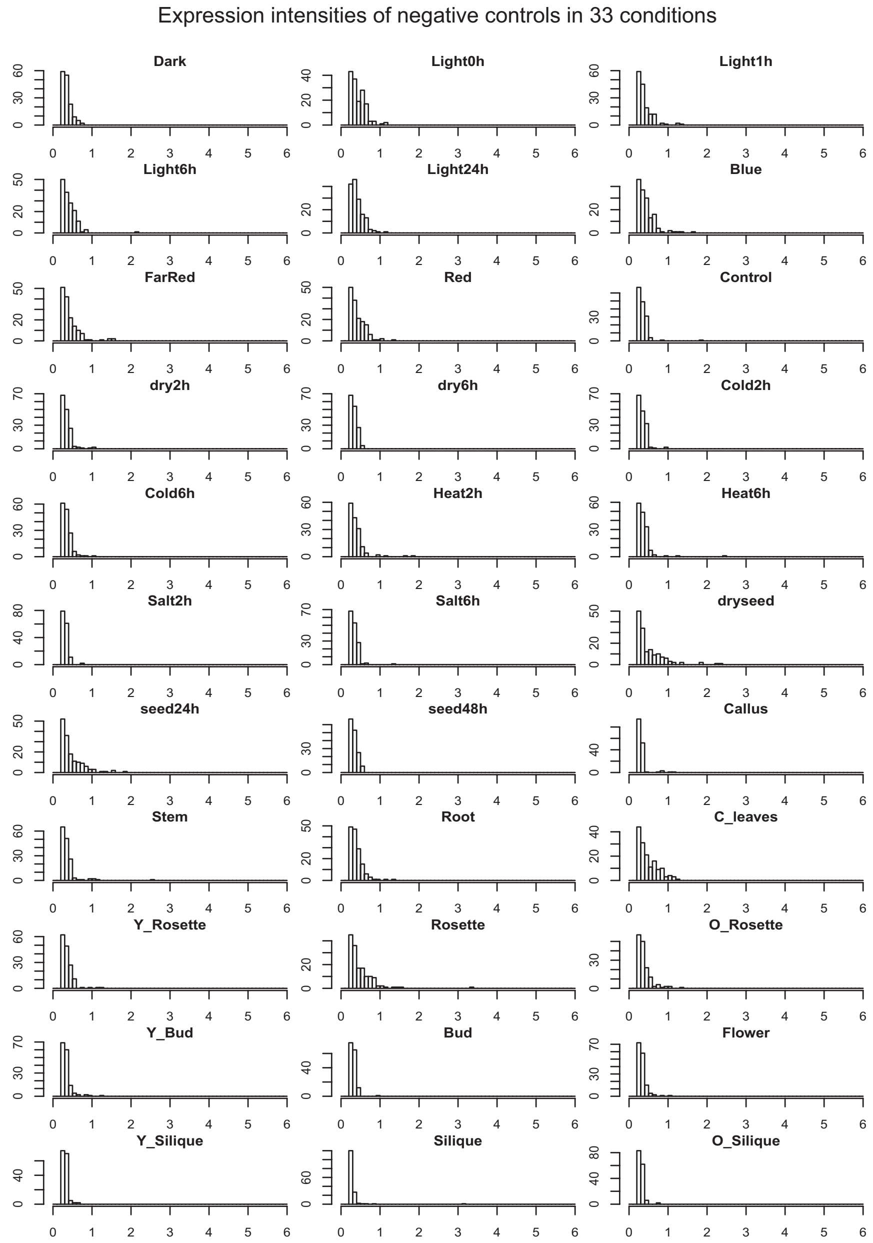
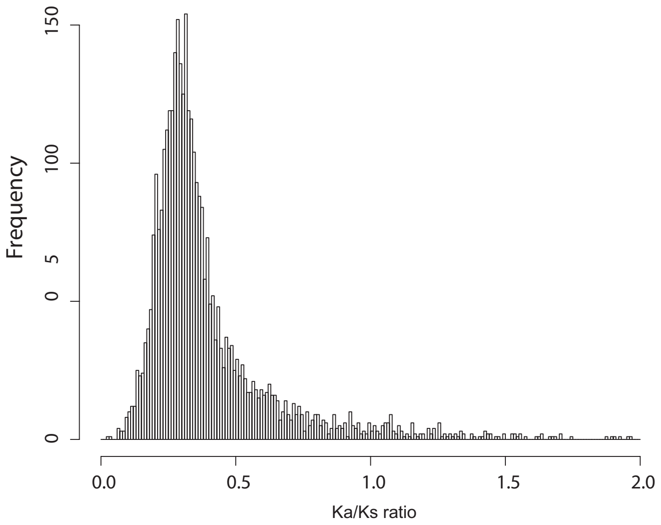
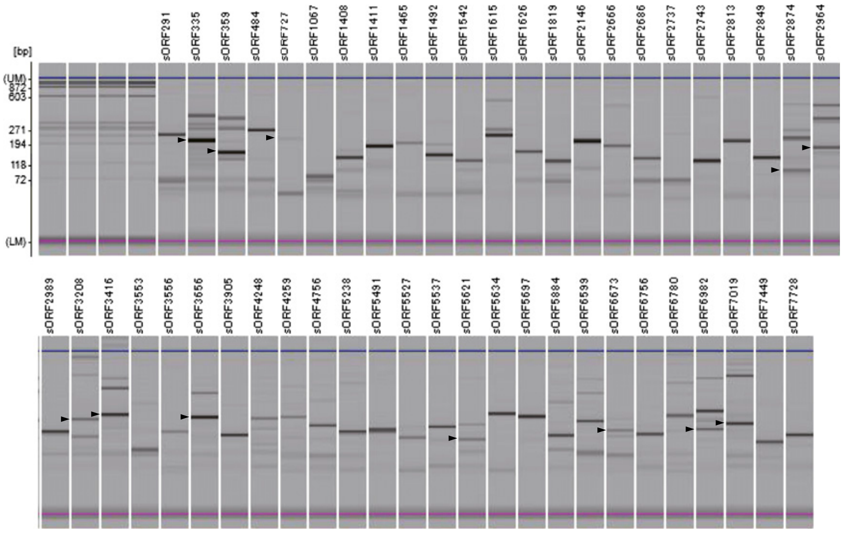


**Fig. S5.** Expression intensities of negative controls in 33 conditions. *x* axis represents log10 values of negative controls in 33 conditions (continuous dark light, white 0 h, white 1 h, white 6 h, white 24 h, continuous blue light, continuous far-red light, continuous red light, control, drought 2 h, drought 6 h, cold 2 hoursm cold 6 h, heat 2 h, heat 6 h, salt 2 h, salt 6 h, dry seeds, 24-h-imbibed seeds, 48-h-imbibed seeds, callus, stems, root, cauline leaves, juvenile rosette, adult rosette, senescence leaves, young buds, mature flower buds, flowers, young siliques, mature siliques, and old siliques).



Ka/Ks ratio of sequences similar to sORFs in random sequences

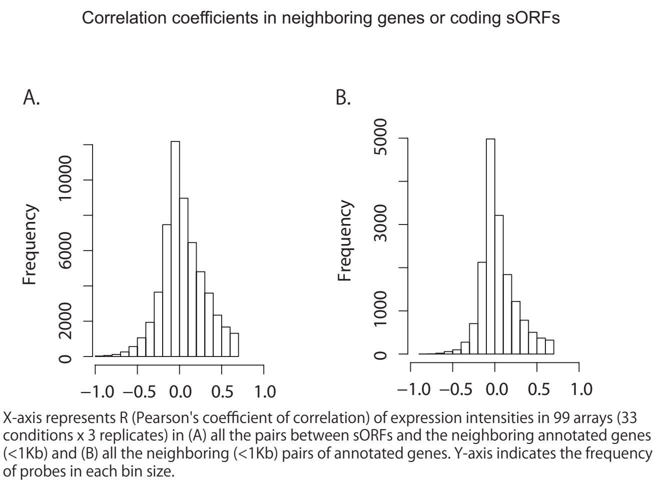
**Fig. S6.** $K_a$/Ks ratio of sequences similar to sORFs in random sequences. A total of 4,265 similar sequences of coding sORFs are identified against these random sequences with similar nucleotide composition to that of the coding sORF. The median $K_a$/Ks ratio (0.32) was defined to be the $K_a$/Ks ratio, which represents neutrality in our procedure.



RT-PCR of 49 coding sORFs whose transgenic plants showed phenotypic effects.

Arrows show the bands of expected size.

**Fig. S7.** RT-PCR of 49 coding sORFs whose transgenic plants showed phenotypic effects. RT-PCR analyses with specific primers are conducted in 49 coding sORFs whose transgenic plants showed phenotypic effects. Arrow shows the bands of expected size.



Correlation coefficients in neighboring genes or coding sORFs

X-axis represents R (Pearson's coefficient of correlation) of expression intensities in 99 arrays (33 conditions x 3 replicates) in (A) all the pairs between sORFs and the neighboring annotated genes (<1Kb) and (B) all the neighboring (<1Kb) pairs of annotated genes. Y-axis indicates the frequency of probes in each bin size.

**Fig. S8.** Correlation coefficients in neighboring genes or coding sORFs. Expression patterns between neighboring annotated genes within <1 Kb distance and coding sORF was evaluated by *R* value (Pearson's coefficient of correlation) of expression intensities in 99 arrays (33 conditions × 3 replicates). The *R* values for all neighboring pairs of annotated genes were not significantly different from those for the pairs of coding sORFs and their neighboring annotated genes (0.07 ± 0.24 in coding sORFs, 0.08 ± 0.28 in neighboring annotated genes, *P* = 0.24, Wilcoxon test).

**Table S1. Comparisons among number of sORFs with and without higher expression compared with low distribution of expression intensities (LE) in annotated genes and those with and without evidence of translation**

| Evidence of translation | sORFs with higher expression compared with LE in annotated genes | sORFs without higher expression compared with LE in annotated genes | Ratio* | *P* value ($\chi^2$ test) |
|---|---|---|---|---|
| sORFs with evidence of translation | 69 | 8 | 8.6 | 0.83 |
| sORFs without evidence of translation | 6,952 | 872 | 8.0 | |

*Ratio of number of sORFs with higher expression to number of those without higher expression.

**Table S2. Comparisons among number of sORFs with and without higher expression compared with negative controls and those with and without evidence of translation**

| Evidence of translation | sORFs with higher expression compared with negative controls | sORFs without higher expression compared with negative controls | Ratio* | *P* value ($\chi^2$ test) |
|---|---|---|---|---|
| sORFs with evidence of translation | 75 | 2 | 37.5 | 0.51 |
| sORFs without evidence of translation | 7,506 | 318 | 23.6 | |

*Ratio of number of sORFs with higher expression to number of those without higher expression.

**Table S3. Comparisons among number of sORFs with and without higher expression compared with pseudogenes and those with and without evidence of translation**

| Evidence of translation | With higher expression in comparison with pseudogenes | Without higher expression in comparison with pseudogenes | Ratio* | *P* value ($\chi^2$ test) |
|---|---|---|---|---|
| sORFs with evidence of translation | 30 | 47 | 0.6 | 0.01 |
| sORFs without evidence of translation | 2,069 | 5,755 | 0.4 | |

*Ratio of number of sORFs with higher expression to number of those without higher expression.

**Table S4. Comparisons among number of sORFs with and without purifying selection and those with and without evidence of translation**

| Evidence of translation | With purifying selection | Without purifying selection | Ratio* | *P* value ($\chi^2$ test) |
|---|---|---|---|---|
| sORFs with evidence of translation | 15 | 62 | 0.24 | $3.3 \times 10^{-5}$ |
| sORFs without evidence of translation | 566 | 7,258 | 0.08 | |

*Ratio of number of sORFs with purifying selection to number of those without purifying selection.

## Other Supporting Information Files

Datasets S1–S3 (XLS)