Genes Genet. Syst. (2016) **91**, p. 111–125

Identification of reference genes for quantitative expression analysis using large-scale RNA-seq data of *Arabidopsis thaliana* and model crop plants

Toru Kudo[1], Yohei Sasaki[1], Shin Terashima[1], Noriko Matsuda-Imai[1], Tomoyuki Takano[1], Misa Saito[1], Maasa Kanno[1], Soichi Ozaki[1], Keita Suwabe[2], Go Suzuki[3], Masao Watanabe[4], Makoto Matsuoka[5], Seiji Takayama[6] and Kentaro Yano[1]*

[1]*School of Agriculture, Meiji University, 1-1-1 Higashi-mita, Tama, Kawasaki, Kanagawa 214-8571, Japan*
[2]*Graduate School of Bioresources, Mie University, 1577 Kurimamachiya-cho, Tsu, Mie 514-8507, Japan*
[3]*Division of Natural Science, Osaka Kyoiku University, 4-698-1 Asahigaoka, Kashiwara, Osaka 582-8582, Japan*
[4]*Graduate School of Life Sciences, Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai, Miyagi 980-8577, Japan*
[5]*Bioscience and Biotechnology Center, Nagoya University, Furo, Chikusa, Nagoya, Aichi 464-8601, Japan*
[6]*Graduate School of Biological Sciences, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan*

(Received 16 September 2015, accepted 10 December 2015; J-STAGE Advance published date: 1 April 2016)

In quantitative gene expression analysis, normalization using a reference gene as an internal control is frequently performed for appropriate interpretation of the results. Efforts have been devoted to exploring superior novel reference genes using microarray transcriptomic data and to evaluating commonly used reference genes by targeting analysis. However, because the number of specifically detectable genes is totally dependent on probe design in the microarray analysis, exploration using microarray data may miss some of the best choices for the reference genes. Recently emerging RNA sequencing (RNA-seq) provides an ideal resource for comprehensive exploration of reference genes since this method is capable of detecting all expressed genes, in principle including even unknown genes. We report the results of a comprehensive exploration of reference genes using public RNA-seq data from plants such as *Arabidopsis thaliana* (Arabidopsis), *Glycine max* (soybean), *Solanum lycopersicum* (tomato) and *Oryza sativa* (rice). To select reference genes suitable for the broadest experimental conditions possible, candidates were surveyed by the following four steps: (1) evaluation of the basal expression level of each gene in each experiment; (2) evaluation of the expression stability of each gene in each experiment; (3) evaluation of the expression stability of each gene across the experiments; and (4) selection of top-ranked genes, after ranking according to the number of experiments in which the gene was expressed stably. Employing this procedure, 13, 10, 12 and 21 top candidates for reference genes were proposed in Arabidopsis, soybean, tomato and rice, respectively. Microarray expression data confirmed that the expression of the proposed reference genes under broad experimental conditions was more stable than that of commonly used reference genes. These novel reference genes will be useful for analyzing gene expression profiles across experiments carried out under various experimental conditions.

Key words: Gene expression analysis, reference gene, RNA-seq

INTRODUCTION

Gene expression levels are spatiotemporally regulated to maintain and control biological processes, such as developmental events and responses to environmental stimuli. Therefore, monitoring gene expression levels has played an important role in research addressing the molecular mechanisms underlying biological functions. Among the methods for gene expression analysis, reverse-

Edited by Tetsu Kinoshita
* Corresponding author. E-mail: kyano@meiji.ac.jp
DOI: http://doi.org/10.1266/ggs.15-00065

transcription quantitative real-time polymerase chain reaction (RT-qPCR) has been considered the gold standard method to quantify transcript levels because of its wide dynamic range and its high specificity, sensitivity and throughput (Bustin et al., 2005; Hoebeeck et al., 2007; Schmittgen et al., 2008). Whereas comprehensive methods using microarray and next-generation sequencing technologies have been developed for high-throughput analysis of gene expression profiles, RT-qPCR still has advantages in quickness, sensitivity, accuracy and cost in targeting analysis.

While RT-qPCR is accepted as an accurate method, special attention must be paid to errors caused by machine, enzyme and pipet variation (Bustin, 2002; Vandesompele et al., 2002). To eliminate such errors and prevent misinterpretation of results, normalization using a reference gene as an internal control is frequently performed. A reference gene will ideally have been validated to be expressed at a constant level over a range of experimental conditions (Vandesompele et al., 2002). Traditionally, since so-called 'housekeeping' genes that play an essential role in the cell, such as in the cytoskeleton and glycolysis, were expected to be expressed stably, they have been used as the reference genes. On the other hand, contrary to the desirability of stable expression for reference genes, it has also been demonstrated that expression of the classic housekeeping genes is actually not always constant (Thellin et al., 1999). This suggests that reference genes should be validated under each experimental condition. However, in principle, evaluating the expression stability of a candidate gene is fairly difficult, unless we have a reliable way to quantify gene expression level without normalization by a reference gene; thus, a circularity problem arises (Vandesompele et al., 2002; Andersen et al., 2004). Similarly, selecting a reference gene simply on the basis that it shows the most stable expression in the samples under analysis is inadequate because the errors to be normalized are totally unknown. This is the reason why a reference gene is needed. Thus, as a practical solution, it is desirable that stable expression of a reference gene should be validated by independent evidence, such as public transcriptomic data.

Several computational tools have been published to help researchers select suitable reference genes. For instance, geNorm evaluates a given set of genes by comparing variation in the expression ratios in all combinations of two genes in a given set of genes and samples (Vandesompele et al., 2002). This method is based on the principle that "the expression ratio of two ideal internal control genes is constant in all samples", enabling us to avoid the circularity problem (Vandesompele et al., 2002). However, depending on the preselected gene set, geNorm may overestimate the expression stability of genes that are actually not expressed stably (Andersen et al., 2004). Normfinder, another tool addressing the circularity problem, employs a mathematical model to describe expression stability by considering variation both within a sample group and among sample groups (Andersen et al., 2004). In the use of this tool, there is a requirement that the preselected genes are not expressed differently (Andersen et al., 2004). Therefore, before analysis with Normfinder, Andersen et al. (2004) prepared a gene set by genome-wide evaluation of gene expression stability with microarray transcriptomic data using the coefficient of variation (CV) value as an index.

To survey superior reference genes in *Arabidopsis thaliana* (Arabidopsis), Czechowski and colleagues utilized comprehensive microarray data of seven series, including 323 experimental conditions. Expression stability of all genes on the Affymetrix GeneChip Arabidopsis ATH1 Genome Array (NCBI accession number GPL198) was evaluated with these microarray data using the CV (Czechowski et al., 2005). The top-ranked reference gene candidates in this survey outperformed classic housekeeping genes in expression stability (Czechowski et al., 2005), indicating that comprehensive exploration is essential to gain the best choice for reference genes. Whereas the Arabidopsis ATH1 microarray, the predominant commercial microarray used for expression analysis in Arabidopsis, contains approximately 22,000 gene-specific probe sets (Redman et al., 2004), a total of 33,602 genes are now annotated on the Arabidopsis genome (TAIR10; Lamesch et al., 2012). Hence, it is possible that further superior reference genes exist but have not been identifiable from microarray data in Arabidopsis, which is the major plant model. Furthermore, non-model and minor model plants are less well supported by microarray data because of the poor provision of commercial microarray chips.

RNA sequencing (RNA-seq) is a method for transcriptomic analyses with several advantages over microarray analysis, such as a broader dynamic range, and no dependence on either a probe set or the availability of a commercial microarray chip. In particular, RNA-seq data should be an ideal source for reference gene exploration, since all expressed genes, including any currently unknown, can be detected in principle. Recently, the advantageous features of RNA-seq are driving the accumulation of large-scale gene expression data for various species, including crop species. Public RNA-seq data are therefore becoming useful for exploring reference genes, not only in the major model species but also in crops. Furthermore, the accumulation of comprehensive gene expression data is making it possible to compare gene expression profiles across experimental conditions and species. For instance, a web-based omics database, the Plant Omics Data Center (PODC, http://bioinf.mind.meiji.ac.jp/podc/), has been established with a search function for gene expression network constructed by correspondence analysis (Yano et al., 2006) using the global RNA-

seq data of eight plant species: Arabidopsis, *Glycine max* (soybean), *Medicago truncatula* (medicago), *Oryza sativa* (rice), *Solanum lycopersicum* (tomato), *S. tuberosum* (potato), *Sorghum bicolor* (sorghum) and *Vitis vinifera* (grapevine) (Ohyanagi et al., 2015). When one attempts to perform RT-qPCR analysis to validate the expression pattern found in such global profiling, a reference gene(s) that is usable in broad experimental conditions will be required. The existence of such a broadly available reference gene(s) has been implied by the genome-wide exploration of Arabidopsis reference genes: several genes including AT1G13320 were selected as superior reference genes in multiple experimental series (Czechowski et al., 2005).

Another potential advantage of using RNA-seq data to select reference genes is that RNA-seq data do not require reference gene-based normalization, thus avoiding the circularity problem. An expression level measured by RNA-seq is represented as a value called 'fragments per kilobase of exon per million mapped fragments' (FPKM), which is a value normalized by total fragment number and length of each transcript. Therefore, RNA-seq data should be free of the technical errors to be eliminated in RT-qPCR analysis by the use of reference genes (Zhan et al., 2014).

In this study, we initially attempted to explore superior reference genes that can be used under broad experimental conditions in the eight plant species supported by PODC. After characterization of RNA-seq data collected from a public database, we focused on Arabidopsis, soybean, tomato and rice for the exploration, and propose a set of reference genes for each species. Each of the proposed reference genes covered 24–50% of the experimental conditions under which the RNA-seq data were obtained and showed similar or better stability compared with known reference genes.

MATERIALS AND METHODS

RNA-seq data All available RNA-seq short-read data of *Arabidopsis thaliana* (Arabidopsis), *Glycine max* (soybean), *Medicago truncatula* (medicago), *Oryza sativa* (rice), *Solanum lycopersicum* (tomato), *Solanum tuberosum* (potato), *Sorghum bicolor* (sorghum) and *Vitis vinifera* (grapevine) were obtained from the Sequence Read Archive database (SRA, http://www.ncbi.nlm.nih.gov/sra) (Leinonen et al., 2011). Referring to experimental descriptions available in the SRA database, RNA-seq data obtained from technically biased RNA and small RNA libraries were roughly excluded manually. Quality control and mapping of reads, and calculation of gene expression levels, were carried out as previously described (Ohyanagi et al., 2015) with some version updates of tools and genome files: cutadapt (version 1.4.1) (Martin, 2011), tophat2 (version 2.0.12) (Kim et al., 2013) applying the '--no-novel-juncs' option, bowtie2 (version 2.2.3) (Langmead and Salzberg, 2012), and cufflinks software (version 2.2.1) (Trapnell et al., 2013) applying the '–u' option; and potato reference genome and genome annotation (version 4.03) (The Potato Genome Sequencing Consortium, 2011; Sharma et al., 2013). The RNA-seq data were further filtered by mapping rate after the quality control. Gene expression levels were calculated as FPKM values for each gene model (i.e., unique transcript).

After the calculations of FPKM values, RNA-seq data obtained from biased RNA and small RNA libraries were again removed by the following method. First, such libraries were explored by text mining against the experimental descriptions using an in-house PERL script. The PERL script was designed to detect RNA-seq data with a description containing keywords related to small RNA and fractionated RNA: 'miRNA', 'siRNA', 'smRNA', 'sRNA', 'ncRNA', 'tRNA', 'snRNA', 'exRNA' and 'piRNA'. The script also detects co-occurrence of 'RNA' and 'short', 'micro', 'interference', 'tasi', 'non-coding', 'transfer' or 'degradome'. Experimental descriptions where the terms were detected by the PERL script were again manually checked to exclude RNA-seq data obtained from biased RNA and small RNA libraries.

Selection of reference gene candidates and the possibility of designing gene-specific primers The selection of reference gene candidates was performed using an in-house PERL script. To predict possible gene-specific primers, sequences similar to the transcript of selected reference genes were searched for in the transcript database of each species using the blastn function of the NCBI Basic Local Alignment Search Tool (BLAST, version 2.2.30+) (Camacho et al., 2009). If a highly similar sequence was detected by blastn, the possibility of designing gene-specific primer(s) was evaluated. We designed primers against the transcript of the reference gene and the highly similar transcript using the Primer3 web tool (version 0.4.0, http://bioinfo.ut.ee/primer3-0.4.0/) (Untergasser et al., 2012), and tested their gene specificity.

Microarray data Microarray expression data obtained using the Arabidopsis ATH1 Genome Array and Soybean Genome Array (Affymetrix) were downloaded from the Gene Expression Omnibus database (GEO, http://www.ncbi.nlm.nih.gov/geo/) (Barrett et al., 2013) and the Arabidopsis Information Resource database (TAIR, https://www.arabidopsis.org/) (Lamesch et al., 2012) (see legends to Figs. 5 and 6 for accession numbers of the data sets). Normalization of the data was performed by the MAS5 method using R software (version 3.0.1, https://www.r-project.org/) with the software package 'affy' (Gautier et al., 2004). Matching between microarray probes and genes was performed based on blastn searches

for two rice microarrays, the Affymetrix Rice Genome Array and the Agilent Rice Gene Expression 4x44K Microarray. Affymetrix probe sets were considered to be gene-specific when all 11 of the consensus probes were perfectly matched (100% identity and 100% coverage) with a single gene. Agilent probes were considered to be gene-specific when there was a single gene showing a perfect match. For Affymetrix Genome Arrays of Arabidopsis, soybean and tomato, probe-gene matches were searched for in annotation lists provided by genome databases: TAIR for Arabidopsis, SoyBase (http://www.soybase.org) for soybean (Grant et al., 2010) and Sol Genomics Network (http://solgenomics.net/) for tomato (The Tomato Genome Consortium, 2012).

RESULTS

Acquisition, selection and validation of RNA-seq data To explore candidates for reference genes using as many and as varied gene expression data as possible, all RNA-seq data currently available were downloaded from the SRA database. In the repository database, RNA-seq data were arranged in a hierarchical structure with 'Study' as the largest structure corresponding to a project including 'Run(s)', which is the smallest structure typically corresponding to the individual library prepared from a single biological replicate. The Runs were filtered by RNA type sequenced to select data from an ordinary messenger RNA library, but not technically biased or small RNA libraries, and by mapping rate (> 70%) to control reliability on expression levels; next, Studies where only a single Run remained were excluded. At the step of filtering by mapping rate, 1.8–28.5% of Runs (714 of 2,506 in Arabidopsis, 34 of 915 in soybean, 63 of 995 in tomato, 150 of 878 in rice, 47 of 199 in grapevine, 29 of 128 in medicago, 7 of 130 in potato and 2 of 114 in sorghum) were removed.

After this, while 126, 39, 34 and 33 Studies remained

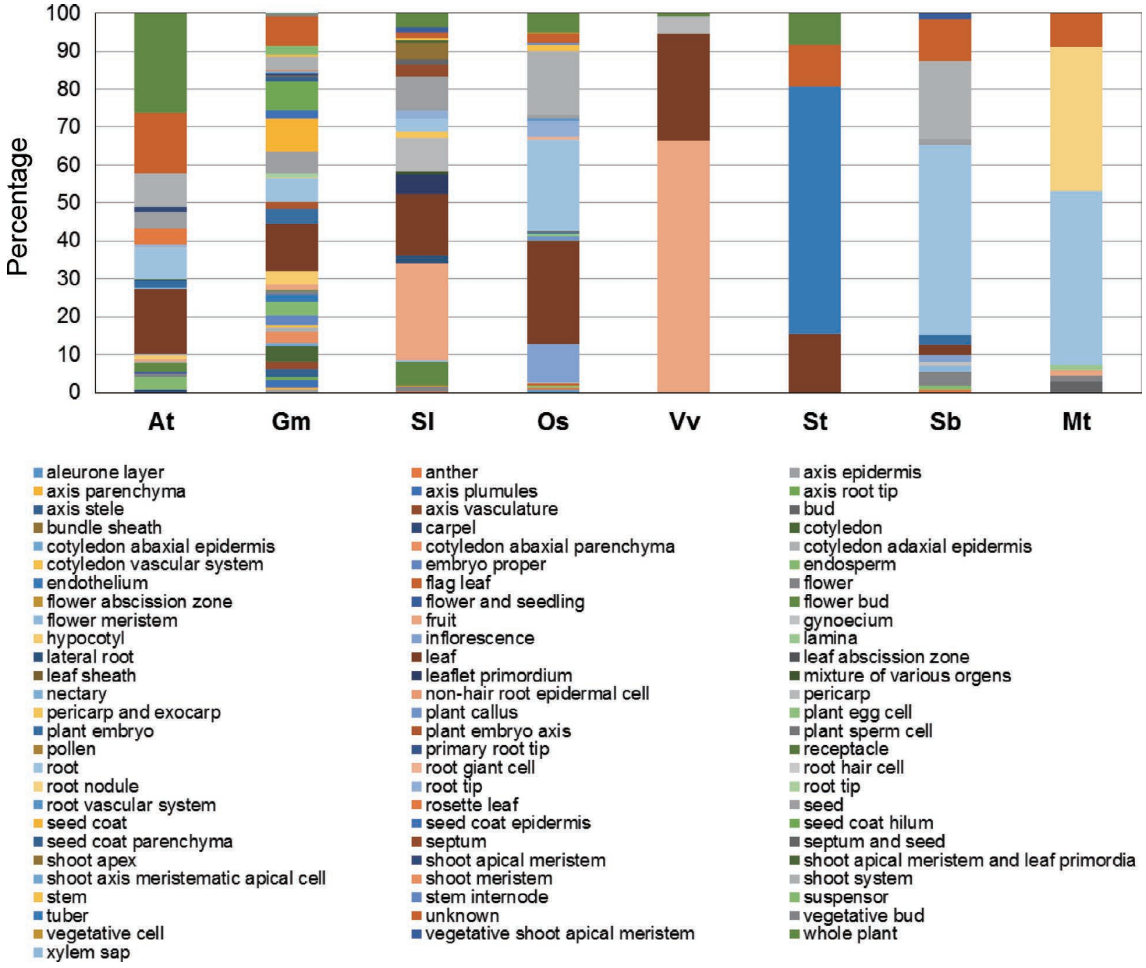

Fig. 1. Composition of anatomic origins in RNA-seq data used in this study. RNA-seq data were classified by anatomic origin, and the composition is represented in a 100% stacked column chart. The color legend for the classes is shown below the chart. At, *Arabidopsis thaliana*; Gm, *Glycine max*; Mt, *Medicago truncatula*; Os, *Oryza sativa*; Sb, *Sorghum bicolor*; Sl, *Solanum lycopersicum;* St, *Solanum tuberosum*; Vv, *Vitis vinifera*.

for Arabidopsis, soybean, tomato and rice, only seven or fewer Studies remained for grapevine, potato, sorghum and medicago, suggesting that biological variation is more biased in species with fewer numbers of Studies. Classification of Runs by anatomic origin (e.g., whole plant, root, stem, leaf, flower, aleurone layer and shoot apical meristem) indicated an over-concentration of one or two classes of origin in grapevine (fruits and leaves), potato (tuber), sorghum (roots) and medicago (roots and root nodules) (Fig. 1). Furthermore, the CV values for each gene model of these species were likely to be distributed in a narrower range compared with those of the species with larger numbers of Studies (Fig. 2). Therefore, we focused hereafter on Arabidopsis, soybean, tomato and rice.

Selection of reference gene candidates in Arabidopsis, tomato, soybean and rice Using the gene expression data selected as described above and CV as an index to evaluate variation in gene expression, reference gene candidates were explored through four steps: (1) if a minimum value of FPKM in a Study was less than 10, the Study was excluded from further analysis for the gene; (2) if CV among all Runs in each Study (CV_e) was larger than 0.1, the Study was excluded from further analysis for the gene; (3) if CV among all Runs in all Studies that remained after the second step (CV_a) was larger than 0.15, the Study showing the maximum deviation (absolute value calculated by subtraction between the mean of FPKMs among all Runs in the Study and the median of FPKMs among all Runs in all Studies that remained) was excluded and this step was repeated until the CV_a became 0.15 or less; and (4) genes were ranked in order of number of Studies retained (Fig. 3).

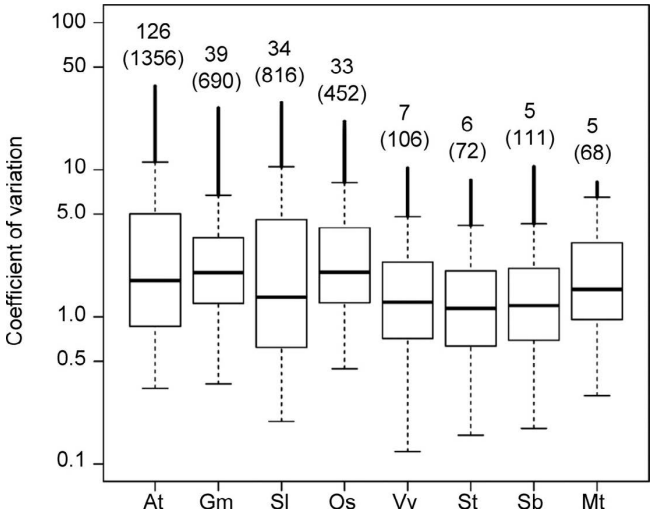

Fig. 2. Distribution of coefficients of variation. The distribution of the CVs of all gene models is represented for each species in this boxplot. The CV for each gene model was calculated from FPKM values of all RNA-seq data used in this study. The y-axis is shown in logarithmic scale. Numbers above the boxes indicate the number of Studies (upper) and Runs (lower, in parentheses). Species abbreviations are as in Fig. 1.

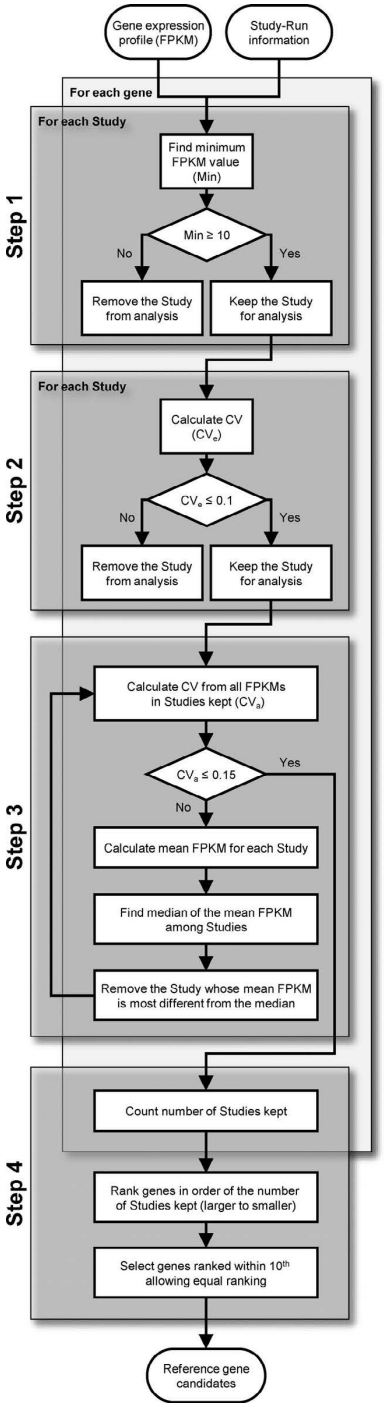

Fig. 3. Flowchart representing the procedure used to select top reference gene candidates. RNA-seq data for gene expression profiles and sample information, such as inclusive relationships between Studies and Runs, were obtained from the Sequencing Read Archive database (http://www.ncbi.nlm.nih.gov/sra). Rounded-corner rectangles, rectangles, and rhombi indicate input and output, processes, and conditional expression.

The first step was performed to control the lower limit of a gene expression level. To determine the lower limit, two matters were taken into account. First, reference genes must show an expression level which is high enough to be quantified easily, or at least without difficulty, in applications, particularly RT-qPCR. To determine a practical threshold to evaluate the ease of quantification, distributions of FPKM values were investigated in 16 genes that have been practically used in RT-qPCR as reference genes by researchers (Abiko et al., 2005; Benschop et al., 2007; Papdi et al., 2008; Streitner et al., 2008; Li et al., 2009; Li et al., 2010; Fontaine et al., 2012; Kudo et al., 2012; Zhang et al., 2013; Wang et al., 2014; Yin et al., 2014; Gonzalez-Cabanelas et al., 2015; Kamada-Nobusada et al., 2015; Patil et al., 2015; Zhai et al., 2015). In 15 of the 16 genes, the 25th percentile of the FPKM values was above 10; the exception was the Arabidopsis *PPR* gene, all of whose FPKM values were below 10 (Supplementary Fig. S1). This result suggests that around 10 is appropriate as the threshold FPKM value. Second, an FPKM value representing the gene expression level must be reliable. Mortazavi et al. (2008) reported that 1.0 RPKM (equivalent to 1.0 FPKM in this study) or above of a typical transcript is stochastically robust when calculated with 40 million mapped reads. In most of the RNA-seq data used in this study, mapped read number was less than 40 million but more than 4 million. Therefore, instead of the threshold 1.0 given in the previous report (Mortazavi et al., 2008), 10 FPKM is again considered to be an appropriate threshold value. Taken together with the range of FPKM values of practical reference genes, we set the FPKM threshold to 10. After selection with this threshold in the first step, the distributions of CV_e and CV_a values were narrowed down, although it was not the purpose of this step (Fig. 4, Supplementary Fig. S2), indicating that the low expression levels made a considerable contribution to the large variation in gene expression.

The second and third steps were both implemented to control variation in gene expression levels: the second step judged if a gene is eligible to be used as a reference gene by evaluating its expression stability in the experimental conditions of each Study; and the third step detected Studies where a gene showed nearly the same expression level. Thus, the gene would be usable as a reference gene across the experimental conditions of all Studies remaining after these steps. In a previous report exploring new reference genes using CV as an index, CV values of proposed reference genes ranged from 0.031 to 0.194, with a median of 0.078, in each experimental series (Czechowski et al., 2005). Given these empirical CV values, 0.1 was employed as a threshold CV_e for evaluating the expression stability of a gene in a single Study. On the other hand, when evaluating the expression stability of the gene across multiple Studies, a slightly relaxed value of 0.15 was used as the threshold CV_a for all FPKM values in the Studies. After the second step, there still remained high CV_a values, and the distribution of CV_e was not affected by the third step (Fig. 4, Supplementary Fig. S2). These results indicate the importance of the third step to control variation in gene expression across multiple experimental conditions.

All genes remaining after the third step should be usable under the experimental conditions of the remaining Studies. However, to provide information about versatile reference genes, genes were ranked by number of remaining Studies at the fourth step. The top 10 ranked genes, shown in Table 1, are particularly strong candi-

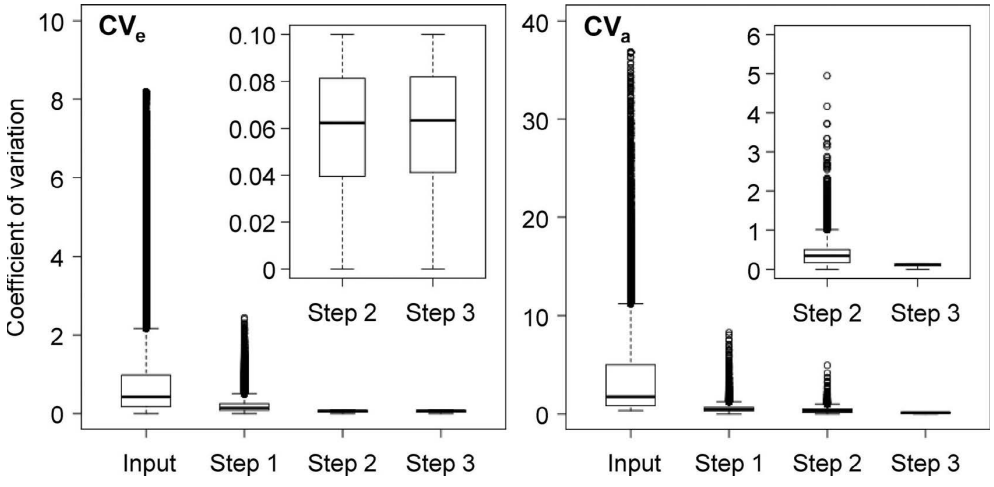

Fig. 4. Effectiveness of each selection step for Arabidopsis reference genes. Distributions of coefficient of variation (CV) values of expression levels of a gene model among all Runs in each Study (CV_e, left panel) and among all Runs in all Studies (CV_a, right panel) remaining after each selection step in Arabidopsis. Enlarged plots for step 2 and step 3 are also shown in the plot area of each panel.

Table 1. Selected reference genes stably expressed under broad conditions

Rank	Gene model ID	No. of Study (%)	Short description	Primer[†]	Probe[‡]
Arabidopsis thaliana (Arabidopsis)		Total 126			
1	AT3G42050.1	63 (50.0)	Vacuolar ATP synthase subunit H family protein	**	Available
2	AT1G11650.2	55 (43.7)	RNA-binding (RRM/RBD/RNP motifs) family protein	**	Available
3	AT5G56290.1	53 (42.1)	Peroxin 5	***	Available
4	AT1G13320.1	52 (41.3)	Protein phosphatase 2A subunit A3	*	Available
5	AT2G20990.1	51 (40.5)	Synaptotagmin A	**	N/A
5	AT5G40810.1	51 (40.5)	Cytochrome C1 family	**	Available
5	AT4G10790.1	51 (40.5)	UBX domain-containing protein	**	Available
8	AT4G26410.1	50 (39.7)	Uncharacterised conserved protein UCP022280	***	Available
8	AT2G45810.1	50 (39.7)	DEA(D/H)-box RNA helicase family protein	**	Available
10	AT5G13030.1	49 (38.9)	Unknown	**	Available
10	AT2G47210.1	49 (38.9)	Myb-like transcription factor family protein	***	N/A
10	AT5G58270.1	49 (38.9)	ABC transporter of the mitochondrion 3	**	Available
10	AT5G08530.1	49 (38.9)	51 kDa subunit of complex I	***	Available
Glycine max (soybean)		Total 39			
1	Glyma.13G041500.1	13 (33.3)	Polypyrimidine tract-binding protein 2 (blast-top *A.thaliana*)	**	Available
2	Glyma.08G152600.1	12 (30.8)	SWAP (Suppressor-of-White-APricot)/ surp domain-containing protein/D111/G-patch domain-containing protein (blast-top *A.thaliana*)	**	Available
2	Glyma.08G092100.1	12 (30.8)	Endomembrane protein 70 protein family (blast-top *A.thaliana*)	*	Available
2	Glyma.20G244000.1	12 (30.8)	Protein kinase family protein (blast-top *A.thaliana*)	**	Available
5	Glyma.19G232800.1	11 (30.8)	RING/U-box superfamily protein (blast-top *A.thaliana*)	**	N/A
5	Glyma.04G076700.1	11 (30.8)	Oligouridylate binding protein 1B (blast-top *A.thaliana*)	–	N/A
5	Glyma.03G150800.1	11 (30.8)	BTB-POZ and MATH domain 2 (blast-top *A.thaliana*)	**	N/A
5	Glyma.17G163700.1	11 (30.8)	casein kinase alpha 1 (blast-top *A.thaliana*)	**	N/A
5	Glyma.09G075400.1	11 (30.8)	DnaJ/Sec63 Brl domains-containing protein (blast-top *A.thaliana*)	*	N/A
5	Glyma.09G046500.1	11 (30.8)	Cullin 1 (blast-top *A.thaliana*)	**	N/A
Solanum lycopersicum (tomato)		Total 34			
1	Solyc03g044140.2.1	17 (50.0)	Protein kinase 2	**	N/A
2	Solyc05g018590.2.1	16 (47.1)	26S protease regulatory subunit 8 homolog A	**	Available
2	Solyc01g109940.2.1	16 (47.1)	26S protease regulatory subunit	**	Available
2	Solyc02g064700.2.1	16 (47.1)	Protein serine/threonine kinase	***	Available
2	Solyc08g067910.2.1	16 (47.1)	Syntaxin 32	***	Available
2	Solyc02g063130.2.1	16 (47.1)	UV excision repair protein RAD23	**	Available
2	Solyc07g062610.2.1	16 (47.1)	4-methyl-5(B-hydroxyethyl)-thiazole monophosphate biosynthesis enzyme ThiJ	***	Available
8	Solyc05g055770.2.1	15 (44.1)	Basic leucine zipper and W2 domain-containing protein 2	***	Available
8	Solyc04g010070.2.1	15 (44.1)	Protein phosphatase 2A regulatory subunit B & apos-like protein	**	Available
8	Solyc03g005300.2.1	15 (44.1)	ADP-ribosylation factor C1	***	Available
8	Solyc01g101000.2.1	15 (44.1)	Kinase family protein	**	Available
8	Solyc01g099760.2.1	15 (44.1)	26S protease regulatory subunit 6A homolog	**	Available

Continued

Table 1. Continued

Rank	Gene model ID	No. of Study (%)	Short description	Primer[†]	Probe[‡]
Oryza sativa (rice)		Total 33		Available	
1	Os11t0296800-01	9 (27.3)	Chromatin SPT2 family protein.	***	Available
1	Os06t0183900-01	9 (27.3)	Protein of unknown function DUF602 family protein.	***	Available
1	Os01t0813500-01	9 (27.3)	EAP30 domain containing protein.	**	Available
1	Os05t0295800-01	9 (27.3)	Similar to Glyoxalase I (EC 4.4.1.5).	***	Available
1	Os07t0626600-01	9 (27.3)	Similar to Embryogenic callus protein-like.	***	Available
1	Os07t0543000-01	9 (27.3)	Similar to Helix-loop-helix-like protein (Fragment).	**	Available
1	Os07t0101400-01	9 (27.3)	Kelch related domain containing protein.	**	Available
8	Os02t0739600-01	8 (24.2)	Similar to Pyruvate dehydrogenase E1 component alpha subunit, mitochondrial precursor (EC 1.2.4.1) (PDHE1-A).	**	Available
8	Os12t0163400-01	8 (24.2)	Tubby family protein.	**	Available
8	Os08t0533700-01	8 (24.2)	Similar to cation antiporter.	**	Available
8	Os04t0617800-01	8 (24.2)	Similar to Imidazoleglycerol-phosphate dehydratase 1 (EC 4.2.1.19) (IGPD 1).	**	Available
8	Os06t0608500-01	8 (24.2)	Snf7 family protein.	**	Available
8	Os08t0384900-01	8 (24.2)	Zinc finger, RING/FYVE/PHD-type domain containing protein.	***	Available
8	Os01t0945800-01	8 (24.2)	Nucleotide-binding, alpha-beta plait domain containing protein.	***	Available
8	Os01t0260700-01	8 (24.2)	Similar to PUR ALPHA-1.	***	Available
8	Os03t0833300-04	8 (24.2)	Similar to Squamosa promoter-binding-like protein 6.	–	Available
8	Os05t0160100-01	8 (24.2)	CT11-RanBPM domain containing protein.	**	Available
8	Os05t0418000-01	8 (24.2)	GDP dissociation inhibitor protein OsGDI1.	**	Available
8	Os11t0157100-01	8 (24.2)	Similar to Cyclin T1 (Fragment).	***	Available
8	Os03t0259300-01	8 (24.2)	Tetratricopeptide-like helical domain containing protein.	**	Available
8	Os07t0295200-01	8 (24.2)	Protein of unknown function DUF167 family protein.	**	Available

[†] Possibility of gene-specific primer(s):
*** Would be easy to design gene-specific primers.
** Would be possible to design gene-specific forward and reverse primers with consideration to avoid mis-annealing on transcripts containing similar sequence.
* Would be possible to design a gene-specific forward or reverse primer with consideration to avoid mis-annealing on transcripts containing similar sequence.
– Would be hard/impossible to design a gene-specific primer.
[‡] Availability of gene-specific probes on the major commercial microarray chips (NCBI accession numbers: GPL198 for Arabidopsis, GPL4592 for soybean, GPL4741 for tomato, GPL2025 or GPL2025 for rice). N/A, not available.

dates for a reference gene that can be used under broader experimental conditions. These 13, 10, 12 and 21 genes of Arabidopsis, soybean, tomato and rice, respectively, include equally ranked genes (e.g., four genes were ranked equal 10th in Arabidopsis). Each of these candidates covered 24%–50% of Studies (Table 1). Collectively, 95 of 126 (75.4%), 23 of 39 (59.0%), 25 of 34 (73.5%) and 16 of 33 (48.5%) Studies of Arabidopsis, soybean, tomato and rice, respectively, were supported by the reference gene candidates listed in Table 1 (Supplementary Tables S1–S4). Except for soybean Glyma.04G076700 and rice Os03g0833300, it would be possible to design a gene-specific primer for the proposed reference genes (Table 1).

Validation of the reference gene candidates in Arabidopsis To validate our method, we compared the variation in expression levels between our reference gene candidates and existing reference genes that have been previously used or proposed. Public microarray data, which are independent from the RNA-seq data used to select the candidate genes, were used as gene expression data for the validation.

Two of our candidate genes, AT4G26410 and AT1G13320 (Table 1), were examined for the validation in Arabidopsis. Interestingly, these genes had been proposed as reference

genes in a previous study: AT4G26410 was recommended for biotic stress conditions and AT1G13320 for developmental, abiotic stress and light conditions, but neither were recommended for hormone treatments (Czechowski et al., 2005). In our analysis, AT4G26410 and AT1G13320 were stably expressed in 50 and 52 of 126 Studies, respectively. Contrary to the previous report, both genes showed stable expression under abiotic stress conditions, namely cold (SRA accession no. RP029896), heat (SRP032366), salt (SRP029598) and oxidative stresses (SRP047297), and also under hormone treatments with brassinosteroids (SRP012153, SRP031626, SRP032274, SRP010642) and gibberellin (SRP010642) (Supplementary Table S1). Public microarray data obtained from similar experiments were collected from the GEO database, and subjected to calculation of relative expression levels and CV values to evaluate the variation among the abiotic stress and hormone treatment conditions.

As Arabidopsis reference genes, *GPC2* (AT1G13440), *ACT2* (AT3G18780), *UBQ10* (AT4G05320), *TUB6* (AT5G12250) and *EF-1α* (AT5G60390) have been traditionally used; and 18 genes (AT1G47770, AT1G58050, AT1G62930, AT2G07190, AT2G28390, AT2G32170, AT3G01150, AT3G32260, AT3G53090, AT4G27960, AT4G33380, AT4G34270, AT4G38070, AT5G08290, AT5G12240, AT5G15710, AT5G46630, AT5G55840) were previously proposed in addition to AT1G13320 and AT4G26410 (Czechowski et al., 2005). Twenty of these 23 existing reference genes were employed in the comparison; *EF-1α*, AT1G58050 and AT4G27960 were omitted because gene-specific probes for these three are available on the Affymetrix Arabidopsis ATH1 Genome Array (Fig. 5).

The CV values of AT4G26410 (0.149) and AT1G13320 (0.155) were lower than those of the other existing reference genes (0.165–0.514) (Fig. 5A). Moreover, a boxplot visualizing the distribution of expression levels showed narrower distribution for these two genes than for the others (Fig. 5B, red). Thus, the stable expression of AT4G26410 and AT1G13320 under abiotic stress and hormone treatment conditions was confirmed with data sets derived from different platforms, supporting the adequacy of our method. In addition to AT4G26410 and AT1G13320, our results also proposed AT3G42050 as a reference gene under the same experimental conditions (Supplementary Table S1). However, this gene could not be tested since there is no gene-specific probe for AT3G42050 on the Affymetrix Arabidopsis ATH1 Genome Array, as mentioned above (Table 1).

Validation of a reference gene candidate in soybean To further validate our method, the highest-ranked soybean gene model Glyma.13G041500.1 was compared with known reference genes as described above for the Arabidopsis reference genes.

As reference genes in soybean, *EF1b* (Glyma.02g276600), *TUB4* (Glyma.03g124400), *ACT2* (Glyma.04g215900), *TUA5* (Glyma.05g157300), *UBQ10* (Glyma.07g199900), *CYP2* (Glyma.12g024700) and *ACT11* (Glyma.18g290800) have commonly been used (Hu et al., 2009); and *UKN2* (Glyma.06g038500), *HDC* (Glyma.08g050200), *UKN1* (Glyma.12g020500), *SKIP16* (Glyma.12g051100), *TIP41* (Glyma.20g130700), *ACTB* (Glyma.15g050200) and *TUA2* (Glyma.20g136000) were previously proposed (Hu et al., 2009; Nakayama et al., 2014). Among these existing reference genes, *TUB4*, *TUA5*, *UKN2*, *SKIP16* and *TUA2* were employed in the comparison, because gene-specific probes are available for them on the Affymetrix Soybean Genome Array.

The transcript level of Glyma.13G041500.1 satisfied the criteria on the CV values in 15 Studies, including research on the response to salt stress (SRP042248) and profiling among different tissues, including developmental series of seeds (SRP006767) (Supplementary Table S2). Thus, microarray expression data obtained from samples under a salt stress condition and a control condition (GSE41125), different organs (GSE18822), and developing seeds (GSE21598 and GSE26443) were collected for the comparison. Using these data, variation in expression levels was compared for Glyma.13G041500 and the existing reference genes by CV and boxplotting. Glyma.13G041500 showed a CV value of 0.123, which was the lowest among the genes tested (Fig. 6A), and the distribution of expression levels of Glyma.13G041500 appeared to be narrower than those of other genes (Fig. 6B, red). This observation was confirmed by plotting the expression levels of Glyma.13G041500, *TUA5* and *SKIP16* in each biological sample (Supplementary Fig. S3). These results showed that expression of Glyma.13G041500 is more stable than that of other genes under the tested experimental conditions. In conjunction with the result of validation in Arabidopsis, we concluded that our method successfully selected reference gene candidates that are stably expressed under broad experimental conditions.

Reference genes selected by using RNA-seq data but never by microarray data To evaluate the impact of using RNA-seq data to explore reference genes, the availability of gene-specific probes was searched on the major commercial microarray chips: Affymetrix Arabidopsis ATH1 Genome Array for Arabidopsis, Affymetrix Soybean Genome Array for soybean (NCBI accession number GPL4592), Affymetrix Tomato Genome Array for tomato (GPL4741), and Affymetrix Rice Genome Array and Agilent Rice Gene Expression 4x44K Microarray for rice (GPL2025 and GPL6864, respectively). Probes matching the reference gene candidates were first searched, and then cross-matching with other genes was checked to see

A

AGI code	Probe ID	CV	Status	Description
AT4G26410	253959_at	0.149	Proposed previously and in this study	Uncharacterised conserved protein UCP022280
AT1G13320	259407_at	0.155	Proposed previously and in this study	Protein phosphatase 2A subunit A3
AT2G28390	265256_at	0.165	Previously proposed	SAND family protein
AT2G32170	265697_at	0.187	Previously proposed	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
AT3G01150	259280_at	0.206	Previously proposed	Polypyrimidine tract-binding protein 1
AT5G15710	246515_at	0.206	Previously proposed	Galactose oxidase/kelch repeat superfamily protein
AT5G46630	248858_at	0.214	Previously proposed	Clathrin adaptor complexes medium subunit family protein
AT4G34270	253287_at	0.229	Previously proposed	TIP41-like family protein
AT4G05320	255220_at	0.242	Classically used	Polyubiquitin 10
AT3G53090	251998_at	0.244	Previously proposed	Ubiquitin-protein ligase 7
AT5G08290	246006_at	0.286	Previously proposed	mRNA splicing factor, thioredoxin-like U5 snRNP
AT1G47770	261732_at	0.303	Previously proposed	Beta-galactosidase related protein
AT5G12250	250317_at	0.306	Classically used	Beta-6 tubulin
AT3G32260	256640_at	0.312	Previously proposed	Nucleic acid-binding proteins superfamily
AT1G13440	259361_at	0.316	Classically used	Glyceraldehyde-3-phosphate dehydrogenase C2
AT5G55840	248024_at	0.356	Previously proposed	Pentatricopeptide repeat (PPR) superfamily protein
AT3G18780	257749_at	0.384	Classically used	Actin 2
AT2G07190	266429_at	0.403	Previously proposed	Domain of unknown function (DUF1985)
AT4G38070	253023_at	0.424	Previously proposed	Basic helix-loop-helix (bHLH) DNA-binding superfamily protein
AT5G12240	250311_at	0.427	Previously proposed	Unknown
AT1G62930	261095_at	0.466	Previously proposed	Tetratricopeptide repeat (TPR)-like superfamily protein
AT4G33380	253355_at	0.514	Previously proposed	Unknown

B

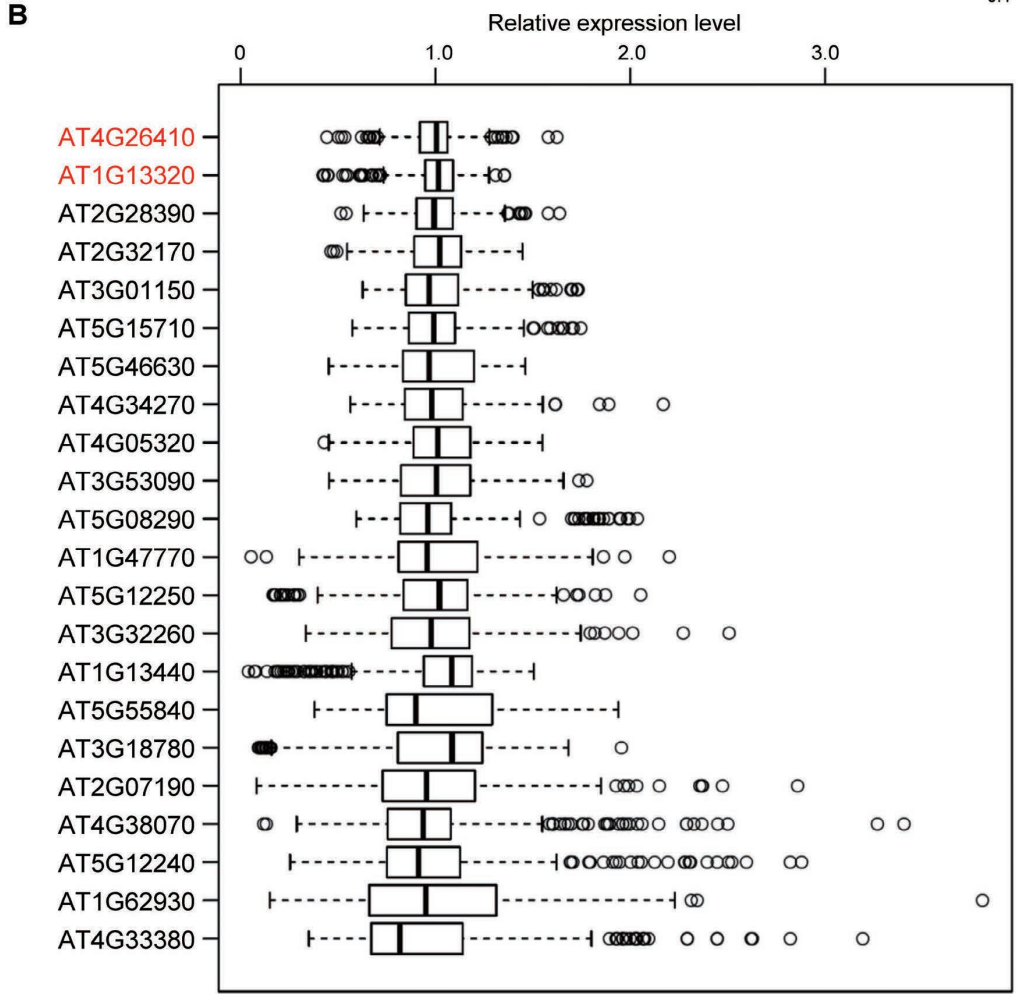

Fig. 5. Comparison of expression stability between Arabidopsis reference genes proposed in this study and other known reference genes. Microarray expression data investigating responses to cold (GEO accession no. GSE3326 and TAIR submission no. ME00325), heat (ME00339), salt (ME00328) and oxidative stress (ME00340), as well as to treatment with brassinosteroids (ME00335 and ME00352) and gibberellin (ME00343 and ME00350), were used to evaluate the reference genes proposed in this study (Lee et al., 2005; Kilian et al., 2007; Goda et al., 2008). **A** Coefficients of variation (CVs) calculated across the microarray data are represented in a heatmap with probe identifiers (IDs) and gene annotations. **B** A boxplot drawn with relative expression levels, which were scaled by division with the mean expression level of each gene (probe). The reference genes proposed in this study are shown in red.

A

Gene ID	Probe ID	CV	Status	Description
Glyma.13G041500	Gma.13295.1.S1_at	0.126	Proposed in this study	Polypyrimidine tract-binding protein 2 (blast-top A.thaliana)
Glyma.05G157300	Gma.13580.2.S1_x_at	0.146	Classically used	TUA5, tubulin alpha-5 (blast-top A.thaliana)
Glyma.12G051100	Gma.1181.1.S1_s_at	0.154	Previously proposed	SKIP16, SKP1/ASK-interacting protein 16 (blast-top A.thaliana)
	GmaAffx.91137.1.S1_at	0.216		
Glyma.06G038500	GmaAffx.82901.1.S1_at	0.230	Previously proposed	UKN2
Glyma.20G136000	GmaAffx.90508.1.S1_s_at	0.307	Previously proposed	TUA2, tubulin alpha-4 chain (blast-top A.thaliana)
	GmaAffx.90437.1.S1_at	0.372		
Glyma.03G124400	Gma.2242.1.S1_at	0.385	Classically used	TUB4, beta-6 tubulin (blast-top A.thaliana)

B

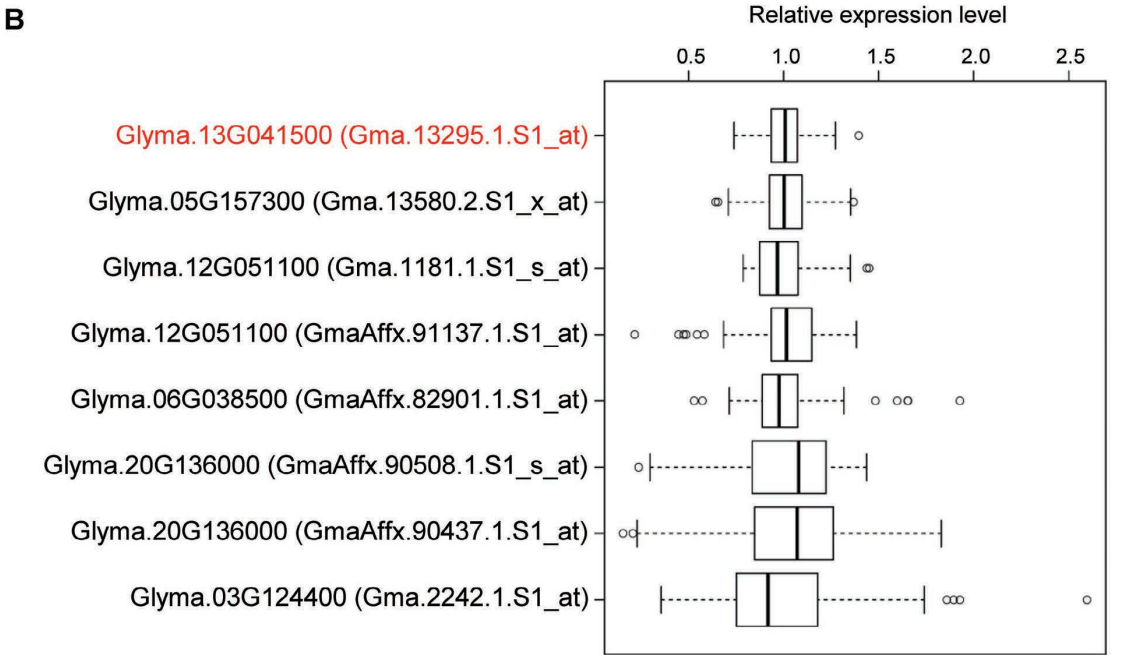

Fig. 6. Comparison of expression stability between a novel soybean reference gene and known reference genes. Microarray expression data obtained from different tissues, including leaves, roots and nodules (GSE18822), seeds during development (GSE21598 and GSE26443) and seedlings treated with salt stress (GSE41125), were used to evaluate a novel reference gene of soybean, Glyma.13G041500 (Yang et al., 2010; Asakura et al., 2012). **A** CVs calculated across the microarray data are represented in a heatmap with probe IDs and gene annotations. **B** A boxplot drawn with relative expression levels, which were scaled by division with mean expression level of each gene (probe). The reference gene proposed in this study is shown in red.

if the probes worked in a gene-specific manner, as described in MATERIALS AND METHODS. This search revealed that among the 53 reference gene candidates, nine (two from Arabidopsis, six from soybean and one from tomato) were found by using RNA-seq, but not by the major microarrays because of the lack of a gene-specific probe (Table 1). Notably, tomato Solyc03g044140.2 was the highest-ranked reference gene candidate, showing stable expression in 50% of all Studies (Table 1), emphasizing the importance of using RNA-seq data in a comprehensive survey for reference genes.

DISCUSSION

In this study, using RNA-seq data, we propose a total of 53 genes as candidates for reference genes that show stable expression in a broad spectrum of experimental conditions in Arabidopsis, soybean, tomato and rice (Table 1). Among them, nine genes were found never to be selected by using microarray data obtained with the major commercial chips due to the lack of corresponding gene-specific probes (Table 1). Notably, a tomato gene, Solyc03g044140.2, showed stable expression in half of the Studies examined, and ranked highest in tomato (Table 1). These results demonstrate the benefit of using RNA-seq data for an exhaustive survey of reference genes.

Because of the method employed in this study, our reference gene candidates are varied in their applicable experimental conditions (Supplementary Tables S1–S4). Therefore, it is hard to evaluate the stability of all candidate genes under the various conditions. Nevertheless, all three genes employed for the validation tests are more stably expressed in multiple conditions compared with other existing reference genes even in microarray

data sets (Figs. 5 and 6). Since RNA-seq is generally accurate and less affected by technical variation in transcript quantification (Marioni et al., 2008; Fu et al., 2009), we expect that not only the three evaluated genes but also the other highly ranked genes are strong candidates for a utility reference gene that is usable under broad experimental conditions.

In plant species including Arabidopsis, soybean, tomato and rice, efforts have been devoted to exploring and validating reference gene candidates (Czechowski et al., 2005; Jain et al., 2006; Gutierrez et al., 2008; Hu et al., 2009; Hong et al., 2010; Narsai et al., 2010; Dekkers et al., 2012; Ji et al., 2014; Nakayama et al., 2014). Among approximately 90 reference gene candidates mentioned in these previous studies, only two Arabidopsis genes were selected as our reference gene candidates (Table 1, Fig. 5) and the Arabidopsis *YLS8* gene (AT5G08290; Czechowski et al., 2005), the tomato TIP41-like protein coding gene (Solyc10g049850.1; Dekkers et al., 2012) and the rice *PPP6* gene (Os01g0691700; Ji et al., 2014) were found in the 100 top-ranked genes in this study (data not shown). These results strongly suggest that the reference genes proposed through conventional methods and classically used are not always the best choices, especially in an experiment with broad conditions. Our method is useful to explore reference genes that are suitable for such experiments but rarely selected by conventional methods.

We employed four filtering steps in the selection of reference genes: the first step was to control basal expression level; the second step was to control expression stability in each experiment; the third step was to control the stability over the experiments; and the last step was to rank genes by the number of experiments in which the gene was expressed stably. Among these, the second step removed most of the Studies in Arabidopsis and tomato reference genes, but the first step typically showed comparable contributions to the second step in soybean, as well as a predominant contribution in rice (data not shown). This difference among the species may be caused by biological characteristics of the species in gene expression profile, redundancy of gene models in reference genomes used for short-read mapping, and a variety of experimental conditions in the RNA-seq data.

Among the RNA-seq Studies used in this study, 31, 16, nine and 17 Studies of Arabidopsis, soybean, tomato and rice, respectively, were assigned no reference gene candidate (Supplementary Tables S1–S4). This result shows that there exist cases of experimental conditions for which a suitable reference gene should be explored by focusing on a particular condition. When lower-ranked genes were overlooked, 23, eight, six and four Studies, respectively, were assigned at least one gene satisfying the criteria of CV values used in this study. Furthermore, if 0.2 was applied as the threshold for CV_e, which is further relaxed but still within an empirically acceptable range, most of the remaining Studies were assigned a potential reference gene, except for a single Study of Arabidopsis (ERP002233) and soybean (SRP021098), and four Studies of rice (SRP002106, SRP007395, SRP010960 and SRP015433). It seems that these six Studies can be categorized into three classes by features of their biological samples: tissues spatially finely dissected, tissues inoculated with pathogens, and varied organs. For instance, three Studies of Arabidopsis (ERP002233), soybean (SRP021098) and rice (SRP007395) include Runs for libraries prepared from root or seed tissues dissected by laser capture microdissection, manual dissection or fluorescence-activated cell sorting. These observations indicate that a suitable reference gene can be found in most cases, if exploration is conducted in an experimental condition-specific manner, but also that there are cases where no stably expressed gene is found in the experimental condition.

Interestingly, we found that a soybean gene (Glyma.17G163700.1) and two tomato genes (Solyc03g044140.2 and Solyc02g064700.2) (see Table 1) encode the catalytic α subunit of casein kinase II (CK2A), which is an essential housekeeping protein kinase (Mulekar and Huq, 2015). Also, in Arabidopsis and rice, one of the CK2A genes was moderately highly ranked in our analysis: AT2G23070 ranked 70th, supporting 41 of 126 Studies, and Os07g0114400 ranked 206th, supporting five of 33 Studies. Among these CK2A genes, AT2G23070 encodes a chloroplast-localized CK2A (αcp) (Salinas et al., 2006), and deduced polypeptides encoded by Glyma.17G163700.1 and Solyc02g064700.2 were predicted to be plastid-localized proteins by the WoLF PSORT program (Horton et al., 2007). These results indicate that CK2A genes, including αcp genes, are potentially useful as a reference gene across multiple plant species. Here, it should again be noted that the CK2A-encoding Solyc03g044140.2 is not supported by a gene-specific probe on the Affymetrix Tomato Genome Array.

We excluded grapevine, potato, sorghum and medicago from the comprehensive exploration of reference genes because of their lower sample variation in RNA-seq data (Figs. 1 and 2). An idea for exploring a reference gene in such species is to test genes orthologous to a reference gene used in other species, as previously conducted in several plant species including grapevine, tomato and potato (Reid et al., 2006; Dekkers et al., 2012; Mariot et al., 2015). To briefly check if this might work with our candidates, αcp and other CK2A genes were explored and evaluated as reference genes in the four excluded species. Putative αcp genes of grapevine (VIT_207s0129g00410), sorghum (Sobic.001G080700) and medicago (Medtr4g095400) ranked 13th, 3rd and 34th with three of seven, one of five and two of five Studies remaining, respectively. In potato, whereas no αcp gene was found, a putative cytosolic CK2A gene (PGSC0003DMG400024939) ranked 2nd,

with four of six Studies remaining. Thus, one of the CK2A genes may also be suitable as a reference gene in these plants, depending on experimental conditions, suggesting the potency of ortholog-based selection. While a number of plant databases, such as PODC (Ohyanagi et al., 2015), ATTED-II (Obayashi et al., 2014), OryzaExpress (Hamada et al., 2011), eFP Browsers (Winter et al., 2007), UniVIO (Kudo et al., 2013) and qTeller (http://www.qteller.com/), are dealing with plant transcriptomic data, direct information on reference gene candidates has not been provided in the databases. A database implemented with a search function for reference genes, combined with information on orthologs and experimental conditions, would be helpful to explore reference gene candidates in species with insufficient accumulation of transcriptomic data. Our reference genes should also be helpful for administrators of the databases to evaluate expression profiles used or to be used in the databases.

Recently, large-scale association/modeling studies using transcriptomic data have emerged (Nagano et al., 2012; Zou et al., 2012). It will not be surprising if such omics-based association studies reveal a key transcript whose accumulation level is correlated with an important agricultural trait of crop plants in the near future. Once this has been achieved, RT-qPCR may be an essential technique for rapid line selection in crop breeding, and utility reference genes should be useful for that. To improve the accuracy of selection of proper reference genes, and to provide useful information on reference genes for plant science and crop breeding, further accumulation of public RNA-seq data of model and non-model plants, including crops, is desired.

This study was supported in part by MEXT Grants-in-Aid for Scientific Research on Innovative Areas (No. 26113716 to K.Y., No. 23113002 to S.T., No. 23113006 to G.S., K.S. and M.W., No. 23113005 to M.M., No. 23113001 to S.T., G.S., M.W. and M.M.); Scientific Research (A) (No. 25252001 to M.W.) from the Japan Society for the Promotion of Science; MEXT-Supported Program for the Strategic Research Foundation at Private Universities (2014–2018); and Research Funding for Computational Software Supporting Program from Meiji University to K.Y.. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

REFERENCES

Abiko, T., Obara, M., Ushioda, A., Hayakawa, T., Hodges, M., and Yamaya, T. (2005) Localization of NAD-isocitrate dehydrogenase and glutamate dehydrogenase in rice roots: candidates for providing carbon skeletons to NADH-glutamate synthase. Plant Cell Physiol. **46**, 1724–1734.

Andersen, C. L., Jensen, J. L., and Ørntoft, T. F. (2004) Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer Res. **64**, 5245–5250.

Asakura, T., Tamura, T., Terauchi, K., Narikawa, T., Yagasaki, K., Ishimaru, Y., and Abe, K. (2012) Global gene expression profiles in developing soybean seeds. Plant Physiol. Biochem. **52**, 147–153.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., et al. (2013) NCBI GEO: Archive for functional genomics data sets - update. Nucleic Acids Res. **41 (D1)**, D991–D995.

Benschop, J. J., Millenaar, F. F., Smeets, M. E., Zanten, M. van, Voesenek, L. A. C. J., and Peeters, A. J. M. (2007) Abscisic acid qntagonizes ethylene-induced hyponastic growth in Arabidopsis. Plant Physiol. **143**, 1013–1023.

Bustin, S. A. (2002) Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. J. Mol. Endocrinol. **29**, 23–39.

Bustin, S. A., Benes, V., Nolan, T., and Pfaffl, M. W. (2005) Quantitative real-time RT-PCR - a perspective. J. Mol. Endocrinol. **34**, 597–601.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009) BLAST+: architecture and applications. BMC Bioinformatics **10**, 421.

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.

Dekkers, B. J. W., Willems, L., Bassel, G. W., Van Bolderen-Veldkamp, R. P. M., Ligterink, W., Hilhorst, H. W. M., and Bentsink, L. (2012) Identification of reference genes for RT-qPCR expression analysis in Arabidopsis and tomato seeds. Plant Cell Physiol. **53**, 28–37.

Fontaine, J. X., Terce-Laforgue, T., Armengaud, P., Clement, G., Renou, J. P., Pelletier, S., Catterou, M., Azzopardi, M., Gibon, Y., Lea, P. J., et al. (2012) Characterization of a NADH-dependent glutamate dehydrogenase mutant of *Arabidopsis* demonstrates the key role of this enzyme in root carbon and nitrogen metabolism. Plant Cell **24**, 4044–4065.

Fu, X., Fu, N., Guo, S., Yan, Z., Xu, Y., Hu, H., Menzel, C., Chen, W., Li, Y., Zeng, R., et al. (2009) Estimating accuracy of RNA-Seq and microarrays with proteomics. BMC Genomics **10**, 161.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004) affy - analysis of *Affymetrix GeneChip* data at the probe level. Bioinformatics **20**, 307–315.

Goda, H., Sasaki, E., Akiyama, K., Maruyama-Nakashita, A., Nakabayashi, K., Li, W., Ogawa, M., Yamauchi, Y., Preston, J., Aoki, K., et al (2008) The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. Plant J. **55**, 526–542.

González-Cabanelas, D., Wright, L. P., Paetz, C., Onkokesung, N., Gershenzon, J., Rodríguez-Concepción, M., and Phillips, M. A. (2015) The diversion of 2-*C*-methyl-D-erythritol-2,4-cyclodiphosphate from the 2-*C*-methyl-D-erythritol 4-phosphate pathway to hemiterpene glycosides mediates stress responses in *Arabidopsis thaliana*. Plant J. **82**, 122–137.

Grant, D., Nelson, R. T., Cannon, S. B., and Shoemaker, R. C. (2010) SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. **38 (suppl 1)**, D843–D846.

Gutierrez, L., Mauriat, M., Guénin, S., Pelloux, J., Lefebvre, J. F., Louvet, R., Rusterucci, C., Moritz, T., Guerineau, F., Bellini, C., et al. (2008) The lack of a systematic validation of reference genes: a serious pitfall undervalued in reverse transcription-polymerase chain reaction (RT-PCR) analysis in plants. Plant Biotechnol. J. **6**, 609–618.

Hamada, K., Hongo, K., Suwabe, K., Shimizu, A., Nagayama, T., Abe, R., Kikuchi, S., Yamamoto, N., Fujii, T., Yokoyama, K., et al. (2011) OryzaExpress: an integrated database of gene expression networks and omics annotations in rice. Plant Cell Physiol. **52**, 220–229.

Hoebeeck, J., Speleman, F., and Vandesompele, J. (2007) Real-time quantitative PCR as an alternative to Southern blot or fluorescence in situ hybridization for detection of gene copy number changes. Methods Mol. Biol. **353**, 205–226.

Hong, S. M., Bahn, S. C., Lyu, A., Jung, H. S., and Ahn, J. H. (2010) Identification and testing of superior reference genes for a starting pool of transcript normalization in Arabidopsis. Plant Cell Physiol. **51**, 1694–1706.

Horton, P., Park, K. J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., and Nakai, K. (2007) WoLF PSORT: protein localization predictor. Nucleic Acids Res. **35 (suppl 2)**, W585–W587.

Hu, R., Fan, C., Li, H., Zhang, Q., and Fu, Y. F. (2009) Evaluation of putative reference genes for gene expression normalization in soybean by quantitative real-time RT-PCR. BMC Mol. Biol. **10**, 93.

Jain, M., Nijhawan, A., Tyagi, A. K., and Khurana, J. P. (2006) Validation of housekeeping genes as internal control for studying gene expression in rice by quantitative real-time PCR. Biochem. Biophys. Res. Commun. **345**, 646–651.

Ji, Y., Tu, P., Wang, K., Gao, F., Yang, W., Zhu, Y., and Li, S. (2014) Defining reference genes for quantitative real-time PCR analysis of anther development in rice. Acta Biochim. Biophys. Sin. **46**, 305–312.

Kamada-Nobusada, T., Makita, N., Kojima, M., and Sakakibara, H. (2013) Nitrogen-dependent regulation of de novo cytokinin biosynthesis in rice: the role of glutamine metabolism as an additional signal. Plant Cell Physiol. **54**, 1881–1893.

Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'Angelo, C., Bornberg-Bauer, E., Kudla, J., and Harter, K. (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. Plant J. **50**, 347–363.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. **14**, R36.

Kudo, T., Makita, N., Kojima, M., Tokunaga, H., and Sakakibara, H. (2012) Cytokinin activity of cis-Zeatin and phenotypic alterations induced by overexpression of putative cis-Zeatin-*O*-glucosyltransferase in rice. Plant Physiol., **160**, 319–331.

Kudo, T., Akiyama, K., Kojima, M., Makita, N., Sakurai, T., and Sakakibara, H. (2013) UniVIO: A multiple omics database with hormonome and transcriptome data from rice. Plant Cell Physiol. **54**, e9.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., et al. (2012) The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. Nucleic Acids Res. **40 (D1)**, D1202–D1210.

Langmead, B., and Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**, 357–359.

Lee, B., Henderson, D. A., and Zhu, J. K. (2005) The *Arabidopsis* cold-responsive transcriptome and its regulation by ICE1. Plant Cell **17**, 3155–3175.

Leinonen, R., Sugawara, H., and Shumway, M. (2011) The sequence read archive. Nucleic Acids Res. **39 (suppl 1)**, D19–D21.

Li, J. Y., Fu, Y. L., Pike, S. M., Bao, J., Tian, W., Zhang, Y., Chen, C. Z., Zhang, Y., Li, H. M., Huang, J., et al. (2010) The *Arabidopsis* nitrate transporter NRT1.8 functions in nitrate removal from the xylem sap and mediates cadmium tolerance. Plant Cell **22**, 1633–1646.

Li, Q. F., Zhang, G. Y., Dong, Z. W., Yu, H. X., Gu, M. H., Sun, S. S. M., and Liu, Q. Q. (2009) Characterization of expression of the *OsPUL* gene encoding a pullulanase-type debranching enzyme during seed development and germination in rice. Plant Physiol. Biochem. **47**, 351–358.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. **18**, 1509–1517.

Mariot, R. F., de Oliveira, L. A., Voorhuijzen, M. M., Staats, M., Hutten, R. C. B., Van Dijk, J. P., Kok, E., and Frazzon, J. (2015) Selection of reference genes for transcriptional analysis of edible tubers of potato (*Solanum tuberosum* L.). PLoS One **10**, e0120854.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, **17 (1)**, 10–12.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods. **5**, 621–628.

Mulekar, J. J., and Huq, E. (2015) Arabidopsis casein kinase 2 α4 subunit regulates various developmental pathways in a functionally overlapping manner. Plant Sci. **236**, 295–303.

Nagano, A. J., Sato, Y., Mihara, M., Antonio, B. A., Motoyama, R., Itoh, H., Nagamura, Y., and Izawa, T. (2012) Deciphering and prediction of transcriptome dynamics under fluctuating field conditions. Cell **151**, 1358–1369.

Nakayama, T. J., Rodrigues, F. A., Neumaier, N., Marcelino-Guimarães, F. C., Farias, J. R. B., de Oliveira, M. C. N., Borém, A., de Oliveira, A. C. B., Emygdio, B. M., and Nepomuceno, A. L. (2014) Reference genes for quantitative real-time polymerase chain reaction studies in soybean plants under hypoxic conditions. Genet. Mol. Res. **13**, 860–871.

Narsai, R., Ivanova, A., Ng, S., and Whelan, J. (2010) Defining reference genes in *Oryza sativa* using organ, development, biotic and abiotic transcriptome datasets. BMC Plant Biol. **10**, 56.

Obayashi, T., Okamura, Y., Ito, S., Tadaka, S., Aoki, Y., Shirota, M., and Kinoshita, K. (2014) ATTED-II in 2014: evaluation of gene coexpression in agriculturally important plants. Plant Cell Physiol. **55**, e6.

Ohyanagi, H., Takano, T., Terashima, S., Kobayashi, M., Kanno, M., Morimoto, K., Kanegae, H., Sasaki, Y., Saito, M., Asano, S., et al. (2015) Plant Omics Data Center: An integrated web repository for interspecies gene expression networks with NLP-based curation. Plant Cell Physiol. **56**, e9.

Papdi, C., Ábrahám, E., Joseph, M. P., Popescu, C., Koncz, C., and Szabados, L. (2008) Functional identification of Arabidopsis stress regulatory genes using the controlled cDNA overexpression system. Plant Physiol. **147**, 528–542.

Patil, G., Valliyodan, B., Deshmukh, R., Prince, S., Nicander, B., Zhao, M., Sonah, H., Song, L., Lin, L., Chaudhary, J., et al. (2015) Soybean (*Glycine max*) SWEET gene family: insights through comparative genomics, transcriptome profiling and whole genome re-sequence analysis. BMC Genomics, **16**, 520.

Redman, J. C., Haas, B. J., Tanimoto, G., and Town, C. D. (2004) Development and evaluation of an *Arabidopsis* whole genome Affymetrix probe array. Plant J. **38**, 545–561.

Reid, K. E., Olsson, N., Schlosser, J., Peng, F., and Lund, S. T. (2006) An optimized grapevine RNA isolation procedure and statistical determination of reference genes for real-time RT-PCR during berry development. BMC Plant Biol. **6**, 27.

Salinas, P., Fuentes, D., Vidal, E., Jordana, X., Echeverria, M., and Holuigue, L. (2006) An extensive survey of CK2 α and β subunits in Arabidopsis: Multiple isoforms exhibit differential subcellular localization. Plant Cell Physiol. **47**, 1295–1308.

Schmittgen, T. D., Lee, E. J., Jiang, J., Sarkar, A., Yang, L., Elton, T. S., and Chen, C. (2008) Real-time PCR quantification of precursor and mature microRNA. Methods **44**, 31–38.

Sharma, S. K., Bolser, D., de Boer, J., Sønderkær, M., Amoros, W., Carboni, M. F., D'Ambrosio, J. M., de la Cruz, G., Di Genova, A., Douches, D. S., et al. (2013) Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. G3 (Bethesda) **3**, 2031–2047.

Streitner, C., Danisman, S., Wehrle, F., Schöning, J. C., Alfano, J. R. and Staiger, D. (2008) The small glycine-rich RNA binding protein *AtGRP7* promotes floral transition in *Arabidopsis thaliana*. Plant J. **56**, 239–250.

The Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. Nature **475**, 189–195.

The Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485**, 635–641.

Thellin, O., Zorzi, W., Lakaye, B., De Borman, B., Coumans, B., Hennen, G., Grisar, T., Igout, A., and Heinen, E. (1999) Housekeeping genes as internal standards: Use and limits. J. Biotechnol. **75**, 291–295.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. **31**, 46–53.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., and Rozen, S. G. (2012) Primer3--new capabilities and interfaces. Nucleic Acids Res. **40**, e115.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., and Speleman, F. (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. **3**, RESEARCH0034.

Wang, Y., Wang, H., Fan, R., Yang, Q., and Yu, D. (2014) Transcriptome analysis of soybean lines reveals transcript diversity and genes involved in the response to common cutworm (*Spodoptera litura* Fabricius) feeding. Plant Cell Environ. **37**, 2086–2101.

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G. V., and Provart, N. J. (2007) An "electronic fluorescent pictograph" Browser for exploring and analyzing large-scale biological data sets. PLoS One **2**, e718.

Yang, S. S., Valdés-López, O., Xu, W. W., Bucciarelli, B., Gronwald, J. W., Hernández, G., and Vance, C. P. (2010) Transcript profiling of common bean (*Phaseolus vulgaris* L.) using the GeneChip Soybean Genome Array: optimizing analysis by masking biased probes. BMC Plant Biol. **10**, 85.

Yano, K., Imai, K., Shimizu, A., and Hanashita, T. (2006) A new method for gene discovery in large-scale microarray data. Nucleic Acids Res. **34**, 1532–1539.

Yin, G., Xu, H., Liu, J., Gao, C., Sun, J., Yan, Y., and Hu, Y. (2014) Screening and identification of soybean seed-specific genes by using integrated bioinformatics of digital differential display, microarray, and RNA-seq data. Gene **546**, 177–186.

Zhai, H., Lü, S., Wu, H., Zhang, Y., Zhang, X., Yang, J., Wang, Y., Yang, G., Qiu, H., Cui, T., et al. (2015) Diurnal expression pattern, allelic variation, and association analysis reveal functional features of the *E1* gene in control of photoperiodic flowering in soybean. PLoS One **10**, e0135909.

Zhan, C., Zhang, Y., Ma, J., Wang, L., Jiang, W., Shi, Y., and Wang, Q. (2014) Identification of reference genes for qRT-PCR in human lung squamous-cell carcinoma by RNA-Seq. Acta Biochim. Biophys. Sin. **46**, 330–337.

Zhang, P., Mar, T. T., Liu, W., Li, L., and Wang, X. (2013) Simultaneous detection and differentiation of *Rice black streaked dwarf virus* (RBSDV) and Southern rice black streaked dwarf virus (SRBSDV) by duplex real time RT-PCR. Virol. J. **10**, 24.

Zou, F., Chai, H. S., Younkin, C. S., Allen, M., Crook, J., Pankratz, V. S., Carrasquillo, M. M., Rowley, C. N., Nair, A. A., Middha, S., et al. (2012) Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. PLoS Genet. **8**, e1002707.