PCP
PLANT & CELL PHYSIOLOGY

Special Issue – Regular Paper

Identification of Reference Genes for RT–qPCR Expression Analysis in Arabidopsis and Tomato Seeds

Bas J. W. Dekkers[1,2,*], Leo Willems[1], George W. Bassel[3], R. P. (Marieke) van Bolderen-Veldkamp[1,2], Wilco Ligterink[1], Henk W. M. Hilhorst[1] and Leónie Bentsink[1,2]

[1]Wageningen Seed Lab, Laboratory of Plant Physiology, Wageningen University, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands
[2]Utrecht University, Molecular Plant Physiology, Padualaan 8, 3584 CH, Utrecht, The Netherlands
[3]University of Nottingham, Division of Plant & Crop Sciences, Sutton Bonington Campus, Loughborough, Leics LE12 5RD, UK
*Corresponding author: E-mail, bas.dekkers@wur.nl; Fax, +31-317-48-47 40.

(Received June 17, 2011; Accepted August 12, 2011)

Quantifying gene expression levels is an important research tool to understand biological systems. Reverse transcription–quantitative real-time PCR (RT–qPCR) is the preferred method for targeted gene expression measurements because of its sensitivity and reproducibility. However, normalization, necessary to correct for sample input and reverse transcriptase efficiency, is a crucial step to obtain reliable RT–qPCR results. Stably expressed genes (i.e. genes whose expression is not affected by the treatment or developmental stage under study) are indispensable for accurate normalization of RT–qPCR experiments. Lack of accurate normalization could affect the results and may lead to false conclusions. Since transcriptomes of seeds are different from other plant tissues, we aimed to identify reference genes specifically for RT–qPCR analyses in seeds of two important seed model species, i.e. Arabidopsis and tomato. We mined Arabidopsis seed microarray data to identify stably expressed genes and analyzed these together with putative reference genes from other sources. In total, the expression stability of 24 putative reference genes was validated by RT–qPCR in Arabidopsis seed samples. For tomato, we lacked transcriptome data sets of seeds and therefore we tested the tomato homologs of the reference genes found for Arabidopsis seeds. In conclusion, we identified 14 Arabidopsis and nine tomato reference genes. This provides a valuable resource for accurate normalization of gene expression experiments in seed research for two important seed model species.

Keywords: Arabidopsis • Gene expression • Normalization • RT–qPCR • Seed • Tomato.

Abbreviations: Cq, quantification cycle; CV, coefficient of variation; HKG, housekeeping gene; RT–qPCR, reverse transcription–quantitative real-time PCR.

Introduction

Quantifying gene expression levels is an important research tool to unravel complex regulatory gene networks. Reverse transcription–quantitative real time PCR (RT–qPCR) is a widely used method for gene expression measurements because of its sensitivity, reproducibility and dynamic quantification range (Pfaffl, 2004). RT–qPCR is employed to quantify relative levels of expression based on normalization using a stably expressed reference gene (Pfaffl 2004, Huggett et al. 2005). Accurate normalization is fundamental for reliable RT–qPCR results. The use of unvalidated or unstable reference genes can have significant impact on the results obtained and could lead to erroneous conclusions (Huggett et al. 2005, Gutierrez et al. 2008, Guénin et al. 2009). Despite the importance of systemic validation of reference genes, this is still under-utilized in plant sciences (Guénin et al. 2009).

For normalization, often so-called housekeeping genes (HKGs) are used. These include genes such as *ACTIN*, *TUBULIN*, *GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE* and *UBIQUITIN* that play essential cellular roles and that are therefore thought to be stably expressed. However, HKGs were found to be subject to considerable regulation in certain conditions (for a review, see Huggett et al. 2005, and references therein). Also for the model plant *Arabidopsis thaliana* (Arabidopsis) it has been shown that such genes are not necessarily stably expressed (Czechowski et al. 2005). In the last decade relevant tools for selecting genes for normalization have become available. Several research groups have developed software tools to identify the most stably expressed genes across a set of samples. These tools include geNORM (Vandesompele et al. 2002), Bestkeeper (Pfaffl et al. 2004) and NormFinder (Anderson et al. 2004) which are freely available on the web and allow researchers to find the best reference gene for their

Plant Cell Physiol. 53(1): 28–37 (2012) doi:10.1093/pcp/pcr113, available online at www.pcp.oxfordjournals.org

experiments. In addition, these programs allow the calculation of a normalization factor over multiple reference genes, which improves the robustness of normalization even further.

As indicated previously, traditional HKGs are not always stably expressed. That is why such genes may not provide a suitable candidate for normalization. Therefore, it is pivotal to identify the best potential reference genes for the experimental system under study. One resource that has been exploited is data from gene expression studies using microarrays in a wide range of developmental stages and environmental conditions. Microarray data sets can be mined for genes that are stably expressed over a diverse set of conditions and have been employed in the medical field (e.g. Eisenberg and Levanon 2003, Kidd et al. 2007, Monaco et al. 2010). This strategy has also been successfully employed in several plant species such as Arabidopsis (Czechowski et al. 2005), soybean (Libault et al. 2008) and rice (Narsai et al. 2010). These studies have identified and validated approximately 20 novel reference genes in Arabidopsis, four in soybean and another 12 in rice. The above studies are important since they have led to the identification of better reference genes compared with those that were available previously.

Different plant parts and tissues are expected to have distinct transcriptomes, which was shown by Principle Component Analysis (PCA) (Schmid et al. 2005). In the case where one is interested in a particular developmental stage or organ (such as seeds in our case) one might expect to find better reference genes by evaluating microarray data sets of the particular stage or organ under study. Seeds are unique plant structures which link successive generations. During orthodox seed development seeds are filled with reserve food and they undergo a developmental program that induces dormancy and desiccation tolerance. This results in dehydrated, quiescent and stress-tolerant structures at the end of seed development. Due to their special characteristics, seeds show distinct transcriptomes as compared with other plant tissues (Schmid et al. 2005). Moreover, Czechowski and co-workers (2005) observed that the inclusion of pollen and seed samples in the developmental series samples led to higher coefficient of variation (CV) values (a measure of variation for gene expression). This also shows that the transcriptomes of pollen and seeds are deviating from those of the other tissues, with the implication that the reference genes identified to date are not necessarily the best references for expression quantification in seed experiments. This is further corroborated by a recent study of Graeber et al. (2011). These authors identified stably expressed genes based on microarray data of *Lepidium sativum* (Lepidium) seeds. Homologous Arabidopsis genes also showed stable expression in seed samples. The expression stability of the reference genes tested was shown to be even higher between both species in a single process (seed germination) compared with two different processes within a single species (Graeber et al. 2011).

Here we describe the identification of novel reference genes for the quantification of gene expression in seeds of Arabidopsis and tomato. To create a set of genes for the analyses we made use of the extensive amount of seed microarray data that is publicly available on the Bio Array Resource (BAR) website (www.bar.utoronto.ca; Winter et al. 2007, Bassel et al. 2008). In total we validated 24 genes in an RT–qPCR experiment using a diverse set of 16 Arabidopsis seed samples. This resulted in a set of 14 validated reference genes. Since for tomato, which is another important seed model, such a public resource is not available, we have used the tomato homologs of the identified Arabidopsis genes. This proved a successful strategy and identified nine reference genes for tomato seeds. We describe the identification of two sets of reference genes for two important seed models, which provides a good starting point for accurate normalization of seed experiments.

Results

Identification of putative reference genes for seed experiments in Arabidopsis

To identify the most stably expressed genes in Arabidopsis seeds we mined publicly available microarray data. In total, data of 151 seed arrays were used that are available on the BAR website (http://www.bar.utoronto.ca; Winter et al. 2007, Bassel et al. 2008; see also **Supplementary Table S1**). These include, among others, a seed imbibition time course (Nakabayashi et al. 2005), the effect of stratification (Yamauchi et al. 2004), dormancy cycling in the Cvi accession (Cadman et al. 2006, Finch-Savage et al. 2007), differences in expression in dormant and after-ripened seeds (Carrera et al. 2007), response to gibberellic acid (Ogawa et al. 2003), expression in endosperm and embryo (Penfield et al. 2006), effect of chemical inhibitors of germination (Bassel et al. 2008) and expression in imbibed ABA signaling mutants (Nakabayashi et al. 2005). We used a similar approach to that described by Czechowski et al. (2005) and Nasai et al. (2010) for the identification of novel reference genes in Arabidopsis and rice, respectively. For each gene we calculated the mean expression and the SD over all experiments. Next, the CV was calculated by dividing the SD by mean expression. Genes with a low CV value are more stably expressed. A list of the 50 genes with the lowest CV values is presented in **Supplementary Table S2**.

In order to visualize the expression stability of this seed-specific gene set we compared it with two known sets of reference genes, i.e. the set of Arabidopsis references (Czechowski et al. 2005, which we will refer to as the Czechowski set) and the 'classic' HKGs. From each of these three sets of reference genes 5–6 genes were selected. From our seed-specific list, genes were picked which were ranked 1, 3, 4, 20 and 21, based on their CV value (i.e. at1g16970, at4g12590, at2g43770, at3g59990 and at4g02080, respectively; see also **Supplementary Table S2**). Of the Czechowski set, at1g13320, at1g58050, at4g26410, at4g34270, at5g12240 and at5g46630 were included, and as representatives of the 'classic' HKGs we show *ACT8* (at1g49240), *ACT2* (at3g18780), *UBC* (at5g25760), *TUB4* (at5g44340) and

EF-1α (at5g60390). For all these genes, expression data from 50 seed microarray experiments were obtained from the BAR website (http://www.bar.utoronto.ca; Winter et al. 2007, Bassel et al. 2008). These are the group 1 seed microarrays of which we used the averaged expression levels per experiment (for overview of seed experiments, see **Supplementary Table S1**). For each gene the average expression level over the entire set of 50 seed experiments was calculated. To obtain relative expression levels between the individual seed experiments we calculated the expression ratio (the expression per experiment divided by the average expression level) for each gene (plotted in **Fig. 1**). This analysis shows that traditional HKGs have considerable variation in expression over these 50 seed experiments. In comparison, the Czechowski set which was identified over a wide range of Arabidopsis microarray data shows more stable expression. However, an even higher level of expression stability was shown by the seed-specific genes, indicating that this set contains some interesting candidates for normalization of gene expression experiments in seeds.

To confirm further the expression stability, 13 genes of the seed-specific set (also indicated in **Supplementary Table S2**) were compared with a set of seven 'classic' HKGs (including *ACTIN*, *UBIQUITIN* and *TUBULIN* genes) and 18 references of the Czechowski set. Further, we added four seed reference genes originally identified based on microarray experiments using Lepidium seeds (Linkies et al. 2009, Graeber et al. 2011). Arabidopsis homologs of three of these Lepidium reference genes were more stably expressed compared with traditional HKGs. Gene expression information of this whole gene set (in total 42 genes) of 50 seed transcriptomics experiments was obtained from the BAR website (same set as used in the previous experiment, for details see **Supplementary Table S1**). The microarray expression data of these putative seed reference genes were analyzed using geNORM (Vandesompele et al. 2002). The geNORM software package uses pairwise comparisons to identify the most stably expressed genes within a set of reference genes across a given set of samples. For each gene a stability value M was calculated; the lower the M value the more stably the gene is expressed. Repetitive analysis combined with stepwise elimination of the least stable gene ranks the gene set from the least to the most stably expressed genes (Vandesompele et al. 2002). This analysis showed that there was a substantial set of genes from the seed-specific list that outperformed the 'classic' HKGs and known reference genes (**Fig. 2**). Interestingly, some genes originally identified by Czechowski et al. (2005) also performed well, considering their relatively low M value which positions them in between the most stably expressed seed-specific reference genes.

Validation of seed reference genes by RT–qPCR in Arabidopsis

The analysis described above indicated that this seed-specific list included some strong candidates for reference genes in Arabidopsis seed gene expression experiments. For further confirmation, 24 genes [11 genes from the seed-specific list, six 'classic' HKGs, four references from the Czechowski set and three reference genes from Lepidium (for gene and primer information see **Supplementary Table S3**)] were validated in an RT–qPCR experiment. All primer pairs produced a specific PCR product indicated by melting curve and agarose gel analysis (**Supplementary Fig. S1**). In total, 16 different Arabidopsis seed samples were used. This diverse set of seed samples included dry seeds of three accessions, imbibed seeds, stress/hormone-treated seeds, germinated seeds, seedlings and different seed tissues which largely overlapped with the seed samples present in the microarray data set. For a detailed description of the Arabidopsis seed samples see **Supplementary Table S4**. The quality of the RNA was thoroughly checked before further processing (**Supplementary Fig. S2**).

The RT–qPCR data of all 24 putative reference genes were analyzed by geNORM. The stability value M was calculated and, by stepwise elimination of the least stable gene, the genes were ranked (**Fig. 3**). We used a cut-off value of ≤0.5 for the M value, which is typical for stably expressed genes (Hellemans et al. 2007). This analysis identified 14 genes stably expressed in our set of Arabidopsis seed samples (**Fig. 3**). More than half of this stably expressed set (eight genes) were genes from the seed-specific list. Furthermore, the set included four previously identified Arabidopsis references and one gene that was previously identified as stably expressed in Lepidium (at2g20000). Only one 'classic' HKG (UBC, at5g25760) was among the 14 most stably expressed genes.

Identification of tomato reference genes based on Arabidopsis homologs

For seeds of tomato, which is another important seed model, a public microarray resource is not available. Therefore, we used tomato homologs of the identified Arabidopsis genes to identify reference genes for experiments in tomato seeds. For 20 of the 24 validated genes in Arabidopsis, homologous genes were identified in tomato. This set was complemented with four known tomato references, i.e. *ACTIN* (*ACT*, SGN-U580422), *PROTEIN PHOSPHATASE 2A catalytic subunit* (*PP2Acs*, SGN-U567355), *RIBOSOMAL PROTEIN L2* (*RPL2*, SGN-U581377) and *UBIQUITIN* (*UBI*, SGN-U593552) (Lovdal and Lillo 2009). For details of the gene and primer information see **Supplementary Table S5**. Thus, in total 24 putative tomato seed reference genes were tested and all primer pairs used in this study produced a single product as shown by both melting curve and agarose gel analysis of RT–qPCR products (see **Supplementary Fig. S3**). For the RT–qPCR experiment, 15 different tomato seed samples were used which included dry seeds of three different tomato accessions, imbibed seeds, stressed seeds, primed seeds, different seed tissues and germinated seedlings (see **Supplementary Table S6** for a detailed description of the tomato seed samples). The quality of isolated RNA from the different tomato seed samples was thoroughly checked before

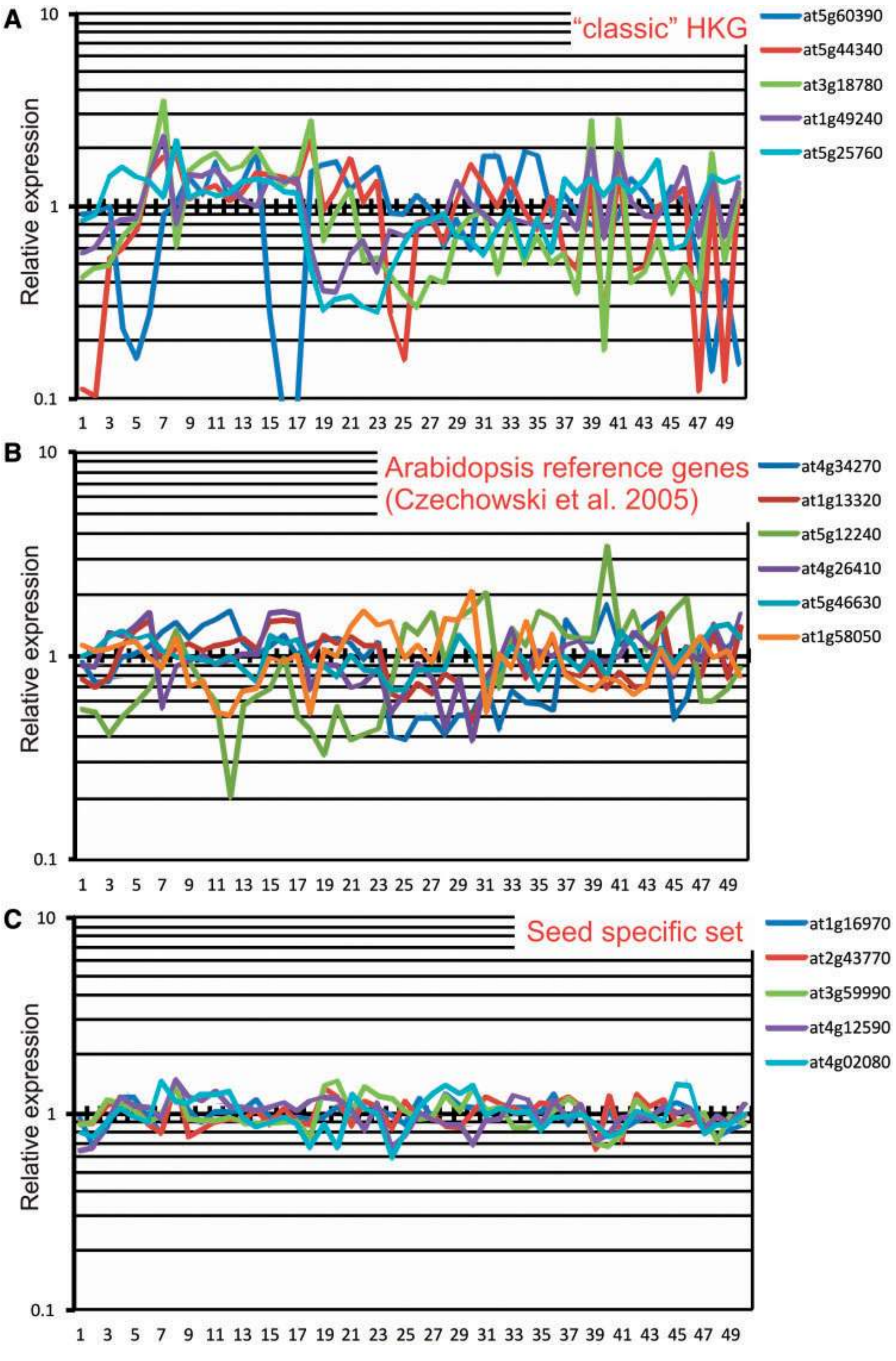

Fig. 1 Relative expression of traditional HKGs, known Arabidopsis references and putative novel seed reference genes over 50 different seed microarray experiments. Relative expression levels per gene were obtained by dividing the expression value per experiment by the average expression level calculated across all 50 seed microarray experiments. For each class [‘classic’ HKGs (A), Czechowski set (B) and seed-specific set (C)], relative expression values were calculated for 5–6 example genes.

the cDNA synthesis step and use in the RT–qPCR experiment (**Supplementary Fig. S4**).

The RT–qPCR data of all 24 putative tomato reference genes for the 15 different samples were analyzed by geNORM. The stability value M was calculated and, by stepwise exclusion of the least stable gene, the genes were ranked (**Fig. 4**). Nine stably expressed genes were identified (with a cut-off value of M ≤0.5) in our set of tomato seed samples. Seven of

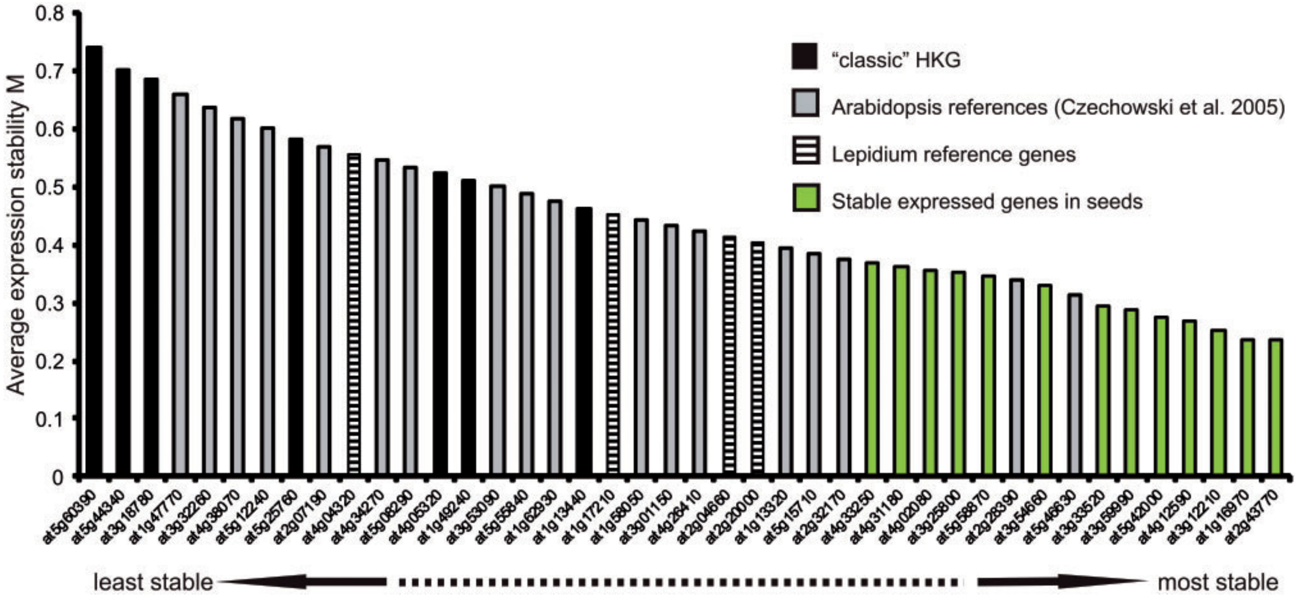

Fig. 2 Expression stability of 42 putative reference genes for seed research based on microarray expression data analyzed by geNORM. Gene expression values of 42 genes across 50 different microarrays of Arabidopsis seed experiments were obtained and analyzed using the geNORM software package. The genes were ranked by stepwise exclusion of the least stable gene. Genes with a high M value are less stably expressed compared with genes with a low M value.

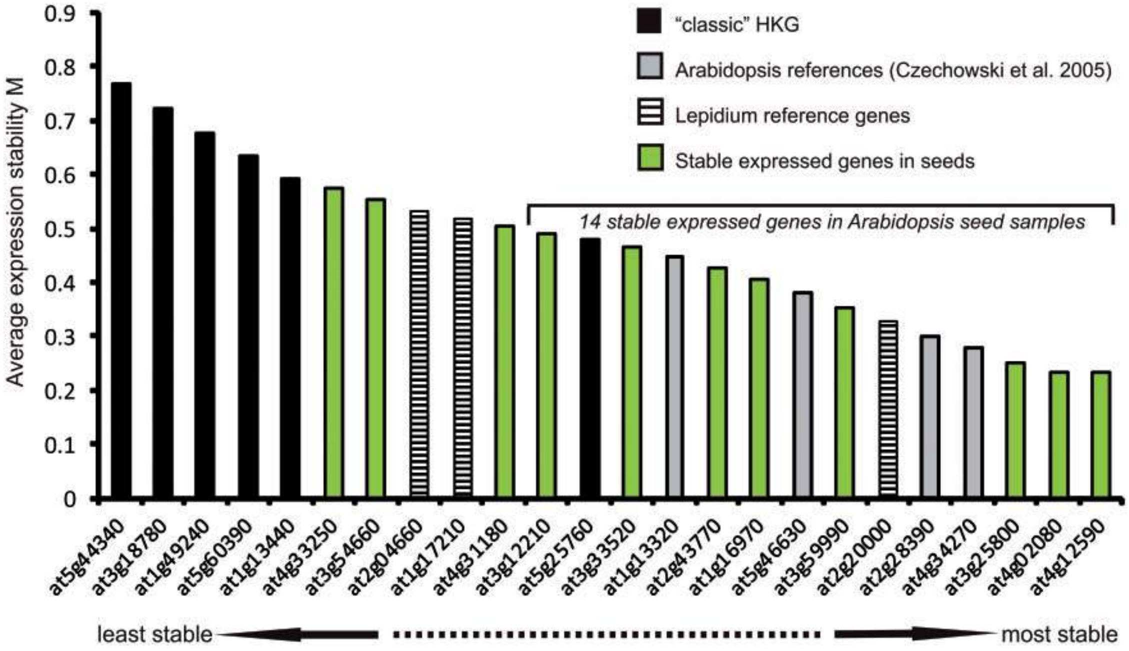

Fig. 3 Expression stability of 24 putative reference genes based on RT–qPCR analysis of Arabidopsis seed samples. Expression values of the 24 putative reference genes over 16 different Arabidopsis seed samples obtained in the RT–qPCR experiment were analyzed in geNORM. geNORM analysis allowed ranking of the genes based on their average expression stability value M. Fourteen genes had an M value $\leq$0.5, and these were considered as stably expressed in Arabidopsis seed samples.

these nine genes were also present in the stably expressed set of Arabidopsis. The other two genes that were stably expressed are a known tomato reference gene (*PP2Acs*) and the homologous gene of the Arabidopsis *ACT2* gene. In conclusion, the use of homologous tomato genes of stably expressed Arabidopsis genes enabled us to identify a set of references for normalization of gene expression experiments of tomato seed samples.

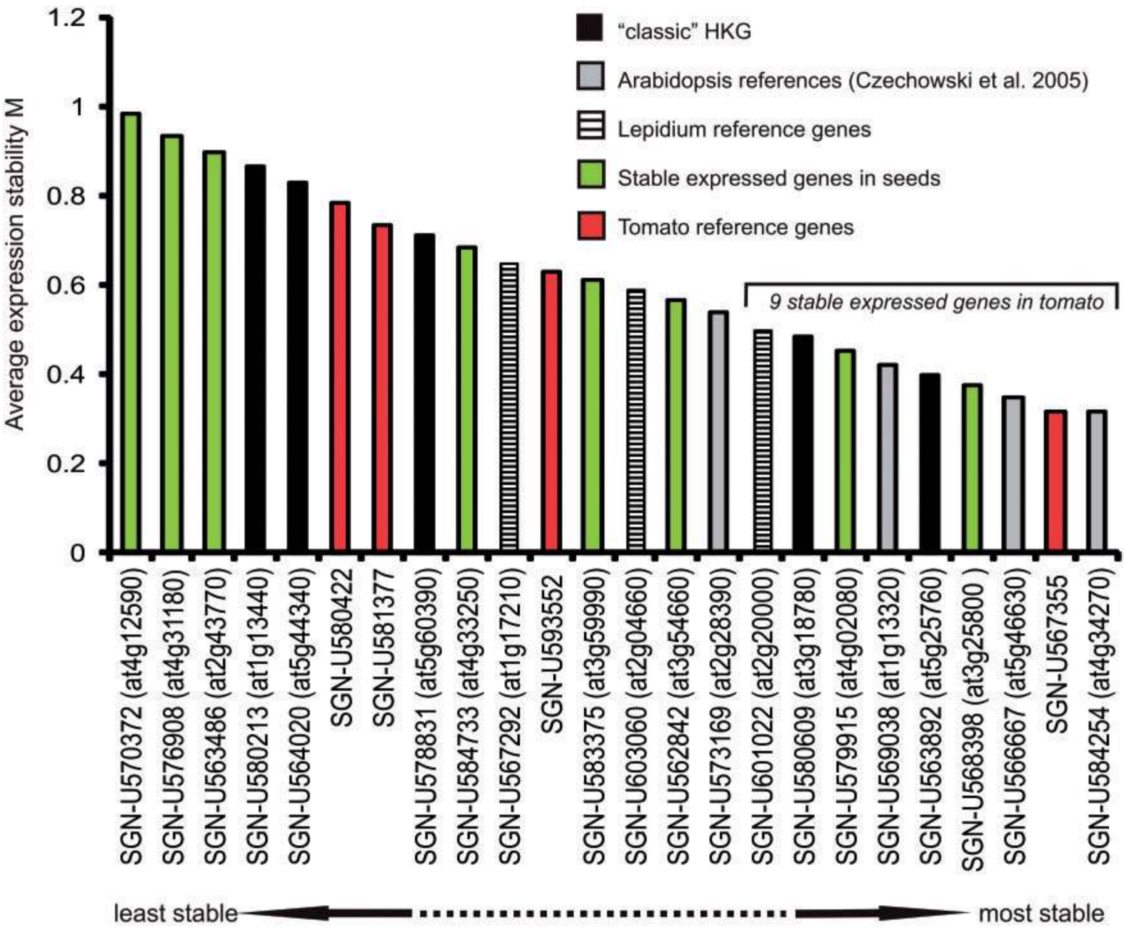

Fig. 4 Expression stability of 24 putative reference genes based on an RT–qPCR experiment of tomato seed samples. Expression values of the 24 putative reference genes over 15 different tomato seed samples obtained in the RT–qPCR experiment were analyzed using geNORM. geNORM analysis allowed ranking of the genes based on their average expression stability value M. Nine genes had an M value $\leq$0.5, and these were considered as stably expressed in tomato seed samples.

Discussion

Normalization is a key step to obtain reliable gene expression data by RT–qPCR. The use of inappropriate reference genes can impact the results obtained and may lead to erroneous conclusions (Gutierrez et al. 2008, Guénin et al. 2009). Therefore, the identification and validation of reference genes for gene expression studies is important. In this study we identified and validated 14 reference genes for gene expression experiments in Arabidopsis seeds and nine references for tomato seeds (**Figs. 3**, **4**). For gene expression studies in seeds, 'classic' HKGs such as *ACTIN* and *TUBULIN* genes are often used. This study shows that such 'classic' HKGs are generally among the least stable genes in both Arabidopsis and tomato seed samples. This indicates that one should be careful using such genes as reference genes, certainly without proper validation. Preferably, however, the use of such unstable HKGs should be avoided for gene expression studies in seeds, and this study provides ample alternative reference genes for use in seed biology research based on microarray data mining and experimental validation.

Czechowski et al. (2005) has already identified a set of reference genes in Arabidopsis. This set contains putative reference genes for use in seed biology, but our rationale was that potentially better reference genes could be identified for seed biology research when we mined microarray experiments of seed experiments specifically. Interestingly, recently, Hruz and co-workers (2011) have shown that no gene is universally stable and that a subset of stable genes from a specific biological context has a smaller variance compared with references identified based on their stability across all conditions, supporting our rationale. Our set of 14 references for Arabidopsis seed experiments was identified by mining a large set of seed microarrays and comparing it with known Arabidopsis references (Czechowski set), Lepidium reference genes and some 'classic' HKGs. More than half (eight genes) of this stably expressed gene set for Arabidopsis seeds included genes derived from the seed-specific list, and these are among the most stably expressed. We searched to determine whether these eight genes were identified previously in one of the microarray sets used by Czechowski et al. (2005). Indeed, at3g25800 was identified as stably expressed in the stressed roots set (rank 42) and the hormone series (rank 41), while at4g12590 was identified in the hormone series (rank 3). Therefore, these two references may be used in a wider range of gene expression studies. The other genes were not among the 100 most stably expressed genes in either of the Arabidopsis sets, indicating that these are more seed-specific references. The four genes (at1g13320, at2g28390, at4g34270 and at5g46630) from the Czechowski set were also validated as stably expressed genes in seeds. This suggests that these are useful for normalization of seed experiments as well. Another stably expressed gene among this validated set was originally identified as stably expressed in Lepidium seeds (at2g20000). The other two genes that originate from Lepidium performed generally better as compared with the traditional HKGs, which is in agreement with the results of Graeber et al. (2011); however, both genes were just above the cut-off. The order of these Lepidium-derived references is slightly different in the study by Graeber et al. (2011) and this study. These differences are probably due to the different samples that were used in both studies. Thus although these reference genes are found to be stably expressed in seeds in both studies, their performance might vary slightly based on the experimental conditions. Therefore, the use of multiple reference genes is better since this would reduce such variation and improve normalization (Vandesompele et al. 2002). Lastly, only a single traditional HKG was found in our set of 14 genes, which is a ubiquitin-conjugating enzyme (UBC, at5g25760). Interestingly, this gene was also stably expressed in the set tested and validated by Czechowski et al. (2005).

Like Arabidopsis, tomato is an important model for seed biology research. In contrast to Arabidopsis, a public microarray resource for seed experiments is not available. In the past an expressed sequence tag (EST) database was mined with the purpose of finding reference genes for tomato (Coker and Davies 2003). However, this study did not identify genes that were stably expressed under all conditions and did not include seed experiments. Therefore, we used tomato homologs of stably expressed genes in Arabidopsis. The strategy of using homologous genes was previously shown to be a powerful approach (Expósito-Rodríguez et al. 2008, Graeber et al. 2011). For example, Expósito-Rodríguez and co-workers (2008) analyzed seven known 'classic' HKGs in tomato together with four novel genes that were originally identified in Arabidopsis. Interestingly, these four tomato homologs of Arabidopsis reference genes outperformed the other seven traditional references in a tomato developmental series. In this study, we identified nine stably expressed genes in tomato seeds (**Fig. 4**). Seven of these nine genes were also present in the stably expressed gene set in Arabidopsis. This shows that this approach using homologs was successful and, interestingly, indicates that the reference genes identified in this study are potential candidates for reference genes for gene expression experiments in seeds of other species. The other two genes that were stably expressed in tomato included a known tomato reference gene (*PP2Acs*) and the homologous gene of Arabidopsis *ACT2*. A remarkable result was the presence of at2g20000 in the tomato set among the most stable genes. This gene was originally identified in Lepidium and is also among the most stably expressed genes in Arabidopsis and tomato. This gene was not identified among the 100 most stable expressed genes in any other microarray sets tested by Czechowski et al. (2005) and therefore appears to be a seed-specific reference gene.

Generally, the tomato data concerning the stably expressed genes were rather similar to those of Arabidopsis. A large overlapping set was identified, and in both species the traditional HKGs were generally among the least stably expressed genes. However, there were some obvious differences. For example, the presence of the homologous gene *ACT2* in the stably expressed gene set in tomato is surprising since it is one of the most unstable genes tested in Arabidopsis seeds. Further, it is remarkable that the most unstable genes in tomato originate from the seed-specific set of Arabidopsis. These results suggest that either the homologous gene has a deviating expression/function between Arabidopsis and tomato or that we did not identify the true homologous gene. In conclusion, this study shows that it is possible to identify good reference genes using a homologous gene approach over different species and, no matter what the reason for the discrepancies, they simply indicate that a proper validation of reference genes across different species is essential. This study has identified and validated 14 novel reference genes for gene expression studies in Arabidopsis seeds and nine reference genes for tomato seeds. These two sets of reference genes create a good starting set for accurate normalization of gene expression experiments in two important seed model species.

Materials and Methods

Identification of stably expressed genes in seed microarray data sets

For mining of stably expressed genes, MAS5 normalized data of 151 microarrays of seed experiments were used (from the BAR website www. bar.utoronto.ca; Winter et al. 2007, Bassel et al. 2008; for details, see **Supplementary Table S1**). Over the entire set of arrays the average expression and SD were calculated per gene. For each gene, the CV value was calculated by dividing the SD by the mean expression. Genes with low CV values are more stably expressed compared with genes with high CV values. The gene set was filtered for an average expression level >100, to obtain genes with a reasonable expression level. A list of 50 genes with the lowest CV values in seed experiments is shown in **Supplementary Table S2**.

Plant material

In these experiments, four different genotypes of Arabidopsis were used, i.e. Columbia-0 (Col-0, N60000), Cape Verde Islands (Cvi, N8580), Landsberg *erecta* (L*er*-0, NW20) and the NIL*DOG17-1* (near isogenic line containing the *DOG1* locus of

Cvi in the L*er*-0 background) (Bentsink et al. 2006). For most treatments the Col-0 accession was used.

Arabidopsis plants were grown on rockwool in a climate cell at 22°C and 70% humidity in a 16 h light/8 h dark cycle for seed production. Plants were watered with a Hyponex nutrient solution (1 g l^{-1}, www.hyponex.co.jp). For germination, seeds were sown on 0.7% water agarose (Eurogentec) (tissue dissections) or on wetted filter paper (Sartorius-Stedim Biotech) (other treatments) supplemented with additives as indicated (**Supplementary Table S4**).

Three different tomato species were used, i.e. *Solanum lycopersicum* cv. Moneymaker (LA2706), *Solanum pimpinellifolium* (CGN 15528) and *Solanum pennellii* (LA716). For most treatments, Moneymaker was used. Tomato plants were grown in a greenhouse under a 16 h light/8 h dark regime. The day and night temperatures were maintained at 25 and 15°C, respectively. Plants were supplied with standard nutrient solution. Seeds were collected from fruits and treated with a HCl solution (2.3% HCl/1.5 h). After this treatment, seeds were washed and separated from the remnants of the fruits. Next, seeds were disinfected using tri-sodium phosphate (100 g of $Na_3PO_4 \cdot 12H_2O$ in 1 liter of water for 1 h) followed by a rinsing step with water. Afterwards seeds were dried on filter papers at room temperature for 3 d. The dried seeds were polished to remove hairs using a seed brushing machine (type 4100.10.00, Seed Processing Holland BV). Seeds were stored at 12°C at 30% relative humidity. For germination experiments, seeds were germinated on filter paper (Allpaper BV) with the indicated additives (**Supplementary Table S6**).

RNA isolation

For Arabidopsis, RNA was isolated as described by Dekkers et al. (2008) except for the tissue dissections. For the dissected samples, radicle and cotyledon samples (approximately 500 seeds) and endosperm samples (approximately 2000 seeds) were dissected using forceps and a scalpel knife. Frozen material was ground in a dismembrator (Mikro-dismembrator U, B. Braun Biotech International) using stainless steel beads. Since the sample sizes of the seed tissues were very small, we were not able to use the above protocol but used a commercial kit (Nanoprep kit, Stratagene) for RNA extraction according to the manual. The only modification was the addition of polyvinyl polypyrrolidone (PVPP; 60 mg ml^{-1}) to the extraction buffer for RNA extraction of endosperm samples, to inactivate phenolic compounds present in the seed coat.

Tomato seeds were frozen in liquid nitrogen after the different seed germination treatments for RNA isolation. To obtain the four seed tissues of tomato seeds (micropylar endosperm cap, lateral endosperm, radicle and cotyledons), approximately 100 seeds were dissected after 24 h of imbibition on water using a scalpel knife and forceps. Frozen seed material was ground in a dismembrator (Mikro-dismembrator U, B. Braun Biotech International) using stainless steel beads, and tomato RNA was extracted using a phenol extraction method (Nonogaki et al. 2000).

RNA integrity of all samples was assessed by analysis on a 1% agarose gel. For all Arabidopsis and tomato samples, clear rRNA bands were visible, indicating that RNA was intact. Further, OD 260/280 ratios were determined using a Nanodrop ND-1000 (Nanodrop Technologies Inc.) and were close to 2.0 for all samples used in this experiment, indicating good RNA quality (data shown in **Supplementary Figs. S2 and S4**). Genomic DNA was removed using a DNase treatment (RNase-free DNase set, Qiagen). Absence of DNA was checked by comparing cDNA samples with RNA samples which were not reverse transcribed (minus RT control). In some samples, signals were detected in the minus RT control, but these were at least 7 Cq (quantification cycle) values higher compared with the cDNA samples. This is clearly above the limit of 5 Cq values as proposed by Nolan et al. (2006).

cDNA synthesis, RT–qPCR conditions and primer design

RNA was reverse transcribed using the iScript^{TM} cDNA synthesis kit (Bio-Rad). In total 1.5 µg of total RNA was reverse transcribed according to the protocol. cDNA samples were diluted 11× with sterile milliQ water.For each qPCR, 5 µl of sample, 12.5 µl of iQ SYBR Green Supermix (Bio-Rad) and 0.5 µl of primer (from a 10 µM working solution) was added and supplemented with water to a final volume 25 µl. The RT–qPCRs were run on a MyiQ (Bio-Rad). The qPCR program run consisted of a first step at 95°C for 3 min and afterwards 40 cycles alternating between 15 s at 95°C and 1 min at 60°C.

In this experiment we used primers available from the literature and also designed primers ourselves (see **Supplementary Tables S3** and **S5** for details). Primers were designed preferably in the 3′ part of the transcript. When possible the primer or primer pair was designed in such a way that it spanned an intron–exon border. The T_m of the primers was between 59 and 62°C. Only a few primer pairs showed slightly higher T_ms, but PCR efficiencies were similar to those of the other primers, indicating that this small difference in T_m did not affect our analysis. Arabidopsis primers that we designed were blasted against all Arabidopsis transcripts using WU-Blast2 on The Arabidopsis Information Resource (TAIR) website (www.arabidopsis.org). In all cases two or more mismatches were found in the most similar sequences found in the Blast search. Routinely a melting curve analysis was performed after the qPCR run (between 55 and 95°C with 0.5°C increments for 10 s each). For all primers, a single peak was observed, confirming the synthesis of a single product which was further confirmed by analysis of the RT–qPCR products on a 2.5% agarose gel (see **Supplementary Figs. 1** and **3**).

geNORM analysis

Microarray and RT–qPCR expression data were used for analysis with the geNORM program (Vandesompele et al. 2002).

The RT–qPCR data were normalized per reference gene. Normalization was carried out on the seed sample with the highest expression (lowest Cq value, also known as Ct, Cp or TOP; nomenclature as suggested by Bustin et al. 2009) to obtain relative expression data. In the calculation, we corrected for primer efficiency (shown in **Supplementary Tables S3** and **S5**) which was calculated from the amplification curve using LinReg PCR (Ramakers et al. 2003, Ruijter et al. 2009).

To assess the robustness of our analysis of the Arabidopsis and tomato samples, we also performed the analysis using different primer efficiencies (optimal efficiency of 1 or the one calculated from a dilution series). Furthermore, we randomly removed one or two samples from our sample set and performed the geNORM analysis to test whether this had an impact on the ranking of the gene list. Both analyses had little impact on the ranking by geNORM. Lastly, we analyzed our data using NormFinder which is another software tool that ranks genes based on the stability of expression (Anderson et al. 2004). A largely similar gene list was identified as most stably expressed (data not shown). These extra tests indicate that we followed a robust procedure for the analysis and identification of these reference genes.

Supplementary data

Supplementary data are available at PCP online.

Funding

This work was supported by the ERA-NET [Plant Genomics grant vSEED (to B.J.W.D., R.P.vB.-V. and L.B.)], Marie Curie IIF [to G.W.B.]; the Technology Foundation [STW grant No. 10027 (to L.W. and W.L.)].

Acknowledgments

We thank Ada Linkies, Kai Graeber and Gerhard Leubner-Metzger (University of Freiburg) for providing gene information prior to publication.

References

Anderson, C.L., Jensen, J.L. and Orntoft, T.F. (2004) Normalizaion 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.

Bassel, G.W., Fung, P., Chow, T.-f.F., Foong, J.A., Provart, N.J. and Cutler, S.R. (2008) Elucidating the germination transcriptional program using small molecules. *Plant Physiol.* 147: 143–155.

Bentsink, L., Jowett, J., Hanhart, C.J. and Koornneef, M. (2006) Cloning of *DOG1*, a quantitative trait locus controlling seed dormancy in *Arabidopsis*. *Proc. Natl Acad. Sci, USA* 103: 17042–17047.

Bentsink, L., Hanson, J., Hanhart, C.J., Blankenstijn-de Vries, H., Coltrane, C., Keizer, P. et al. (2010) Natural variation for seed dormancy in Arabidopsis is regulated by additive genetic and molecular pathways. *Proc. Natl Acad. Sci, USA* 107: 4264–4269.

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M. et al. (2009) The MIQE guidelines: minimum information for publications of quantitative real-time PCR experiments. *Clin. Chem.* 55: 611–622.

Cadman, C.C.S., Toorop, P.E., Hilhorst, H.W.M. and Finch-Savage, W.E. (2006) Gene expression profiles of Arabidopsis Cvi seeds during dormancy cycling indicate a common underlying dormancy control mechanism. *Plant J.* 46: 805–822.

Carrera, E., Holman, T., Medhurst, A., Peer, W., Schmuths, H., Footitt, S. et al. (2007) Gene expression profiling reveals defined functions of the ATP-binding cassette transporter COMATOSE late in phase II of germination. *Plant Physiol.* 143: 1669–1679.

Coker, J.S. and Davies, E. (2003) Selection of candidate housekeeping controls in tomato plants using EST data. *BioTechniques* 35: 740–748.

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., Schuurmans, J.A.M.J. and Smeekens, S.C.M. (2008) Interaction between sugar and abscisic acid signalling during early seedling development in Arabidopsis. *Plant Mol. Biol.* 67: 151–167.

Eisenberg, E. and Levanon, E.Y. (2003) Human housekeeping genes are compact. *Trends Genet.* 19: 362–365.

Expósito-Rodríguez, M., Borges, A.A., Borges-Pérez, A. and Pérez, J.A. (2008) Selection of internal control genes for quantitative real-time RT–PCR studies during tomato development process. *BMC Plant Biol.* 8: 131.

Finch-Savage, W.E., Cadman, C.S.C., Toorop, P.E., Lyn, J.R. and Hilhorst, H.W.M. (2007) Seed dormancy release in Arabidopsis Cvi by dry after-ripening, low temperature, nitrate and light shows common quantitative patterns of gene expression directed by environmentally specific sensing. *Plant J.* 51: 60–78.

Graeber, K., Linkies, A., Wood, A.T.A. and Leubner-Metzger, G. (2011) A guideline to family-wide comparative state-of-the-art qRT–PCR analysis exemplified with a Brassicaceae cross-species seed germination case study. *Plant Cell* 23: 2045–2063.

Guénin, S., Mauriat, M., Pelloux, J., van Wuytswinkel, O., Bellini, C. and Gutierrez, L. (2009) Normalization of qRT–PCR data: the necessity of adopting a systematic, experimental conditions-specific, validation of references. *J. Exp. Bot.* 60: 487–493.

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

Hellemans, J., Mortier, G., De Paepe, A., Speleman, F. and Vandesompele, J. (2007) qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biol.* 8: R19.

Hruz, T., Wyss, M., Docquier, M., Pfaffl, M.W., Masanetz, S., Borghi, L. et al. (2011) RefGenes: identification of reliable and condition specific reference genes for RT–qPCR data normalization. *BMC Genomics* 12: 156.

Huggett, J., Dheda, K., Bustin, S. and Zumla, A. (2005) Real-time RT–PCR normalisation: strategies and considerations. *Genes Immun.* 6: 279–284.

Kidd, M., Nadler, B., Mane, S., Eick, G., Malfertheiner, M., Champaneria, M. et al. (2007) GeneChip, geNORM, and gastrointestinal tumors: novel reference genes for real-time PCR. *Physiol. Genomics* 30: 363–370.

Libault, M., Thibivilliers, S., Radman, O., Clough, S.J. and Stacey, G. (2008) Identification of four soybean reference genes for gene expression normalization. *Plant Genome* 1: 44.

Linkies, A., Müller, K., Morris, K., Turecková, V., Wenk, M., Cadman, C.S.C. et al. (2009) Ethylene interacts with abscisic acid to regulate endosperm rupture during germination: a comparative approach using *Lepidium sativum* and *Arabidopsis thaliana*. *Plant Cell* 21: 3803–3822.

Lovdal, T. and Lillo, C. (2009) Reference gene selection for quantitative real-time PCR normalization in tomato subjected to nitrogen, cold, and light stress. *Anal. Biochem.* 237: 238–242.

Monaco, E., Bionaz, M., de Lima, A.S., Hurley, W.L., Loor, J.J. and Wheeler, M.B. (2010) Selection and reliability of internal reference genes for quantitative PCR verification of transcriptomics during the differentiation process of porcine adult mesenchymal stem cells. *Stem Cell Res. Ther.* 1: 7.

Nakabayashi, K., Okamoto, M., Koshiba, T., Kamiya, Y. and Nambara, E. (2005) Genome-wide profiling of stored mRNA in *Arabidopsis thaliana* seed germination: epigenetic and genetic regulation of transcription in seed. *Plant J.* 41: 697–709.

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.

Nolan, T., Hands, R.E. and Bustin, S.A. (2006) Quantification of mRNA using real-time RT–PCR. *Nat. Protoc.* 1: 1559–1582.

Nonogaki, H., Gee, O.H. and Bradford, K.J. (2000) A germination-specific endo-β-mannanase gene is expressed in the micropylar endosperm cap of tomato seeds. *Plant Physiol.* 123: 1235–1246.

Ogawa, M., Hanada, A., Yamauchi, Y., Kuwahara, A., Kamiya, Y. and Yamaguchi, S. (2003) Gibberellin biosynthesis and response during Arabidopsis seed germination. *Plant Cell* 15: 1591–1604.

Penfield, S., Li, Y., Gilday, A.D., Graham, S. and Graham, I.A. (2006) *Arabidopsis* ABA INSENSITIVE4 regulates lipid mobilization in the embryo and reveals repression of seed germination by the endosperm. *Plant Cell* 18: 1887–1899.

Pfaffl, M.W. (2004) Quantification strategies in real-time PCR. *In* A–Z of quantitative PCR. Edited by Bustin, S.A. pp. 87–112. International University Line (IUL), La Jolla, CA.

Pfaffl, M.W., Tichopad, A., Prgomet, C. and Neuvians, T.P. (2004) Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper—Excel-based tool using pair-wise correlations. *Biotechnol. Lett.* 26: 509–515.

Ramakers, C., Ruijter, J.M., Lekanne Deprez, R.H. and Moorman, A. (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neurosci. Lett.* 339: 62–66.

Ruijter, J.M., Ramakers, C., Hoogaars, W.M.H., Karlen, Y., Bakker, O., van den Hof, M.J.B. et al. (2009) Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. *Nucleic Acids Res.* 37: e45.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M. et al. (2005) A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet.* 37: 501–506.

Sugliani, M., Brambilla, V., Clerkx, E.J., Koornneef, M. and Soppe, W.J. (2010) The conserved splicing factor SUA controls alternative splicing of the developmental regulator ABI3 in Arabidopsis. *Plant Cell* 22: 1936–1946.

Vandesompele, J., de Preter, K., Pattyn, F., Poppe, B., van Roy, N., de Paepe, A. et al. (2002) Accurate normalization of real-time quantitative RT–PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* 3: RESEARCH0034.

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.

Yamauchi, Y., Ogawa, M., Kuwahara, A., Hanada, A., Kamiya, Y. and Yamaguchi, S. (2004) Activation of gibberellin biosynthesis and response pathways by low temperature during imbibition of *Arabidopsis thaliana* seeds. *Plant Cell* 16: 367–378.