Expression stability of housekeeping genes: a proposal for gene quantification studies in pea

José V. Die*, Belén Román, Salvador Nadal, Clara I. González-Verdejo

IFAPA "Alameda del Obispo", Mejora y Biotecnología, 14004 Córdoba, Spain

* Corresponding author.
E-mail address: josev.die@juntadeandalucia.es
Tel.: +34 957 01 60 61
Fax: +34 957 01 60 43

Abstract

Real-time reverse transcriptase quantitative polymerase chain reaction is the most accurate measure of gene expression in biological systems. The data is analyzed through a process called normalization. Internal standards are essential for determining the relative gene expression in different samples. For this purpose, reference or housekeeping genes are selected based on their constitutive expression across samples. At present, there has not yet been any reference gene identified in any organism that is universally optimal across different tissue types or disease situations. Our goal was to test the regulation of eight housekeeping genes (protein phosphatase 2A, helicase, glyceraldehyse-3-phosphate dehydrogenase, α-tubulin, β-tubulin, actin, elongation factor 1α and 18S ribosomal RNA) in pea plants using the geNorm algorithm. Thirteen samples, including different tissues, treatments and genotypes, were included in this analysis. To validate the determined measure of gene-stability, the gene-specific variation was calculated using different normalization factors. The most non-specific variation was removed when the most stable genes were used for normalization, highlighting the importance of the choice of internal controls in gene expression experiments. The set of reference genes presented here will enable better normalization of transcript levels in pea studies.

Keywords: geNorm; housekeeping genes; normalization; pea; real-time PCR

Introduction

Post-genomic technologies have enabled a major breakthrough in gene profiling studies and have been rapidly integrated into the plant science field. These technologies can be a valuable resource for gaining insight into complex regulatory networks, improving our knowledge of the mechanisms underlying different plant-pathogen pathosystems and identifying new genes that are relevant to biological processes. Real-time reverse transcription PCR (real-time RT-PCR) is presently the most sensitive method for detecting mRNAs (reviewed in [Bustin, 2000; Wong and Medrano, 2005]) and is also often used to validate gene expression data obtained from high-throughput array experiments. Although real-time RT-PCR is widely used to quantitate biologically relevant changes in mRNA levels, a number of problems are associated with its use, including the variability in RNA samples, extraction protocols (particularly due to the

77 co-purification of inhibitors), and efficiencies of the reverse transcription and PCR
78 (Bustin et al. 2004). Consequently, it is important that an accurate method of
79 normalization is chosen. Unfortunately, normalization remains one of the most difficult
80 tasks in real-time RT-PCR (Dheda et al. 2004). Quantitative analysis of gene expression
81 using real-time PCR typically requires the use of a constitutively expressed
82 housekeeping gene, whose expression is unaffected by the experimental conditions, as
83 an internal control to normalize for differences in starting cDNA template between
84 samples (Bustin et al. 2002). Housekeeping gene-based normalization corrects for
85 variable starting amounts of RNA and for differences in RT efficiency; however, as
86 there are no universally applicable genes with invariant expression, it is necessary to
87 carefully evaluate the expression of candidate reference genes for every particular
88 experimental system. Normalization with suboptimal internal controls may result in
89 different estimated values and lead to erroneous interpretations (Dheda et al. 2005). To
90 avoid the bias caused by a fluctuation in expression level of a single reference gene,
91 Vandesompele et al. (2002) proposed that at least three proper control genes be used.
92 Thus, there have been a number of studies in plants aimed at validating the presumed
93 stability in the expression of certain reference genes recently (Kim et al. 2003; Brunner
94 et al. 2004; Iskandar et al. 2004; Czechowski et al. 2005; Gonçalves et al. 2005; Nicot et
95 al. 2005; Jain et al. 2006; Reid et al. 2006; Tu et al. 2007; González-Verdejo et al. 2008;
96 Hong et al. 2008; Libault et al. 2008; Remans et al. 2008); so far, however, no such
97 quantification of gene expression has been reported for *Pisum sativum,* the most widely
98 grown grain legume in Europe and the fourth-most in the world (FAOSTAT data,
99 2005). In the frame of different international networks, many powerful genomics
100 resources have been developed for the model legume *Medicago truncatula* over the past
101 few years (Samuel Roberts Noble Foundation, 2006). Translating the knowledge gained

from a model species into improvements in crop growth has always been a major challenge for the comparative genomics field. There is no doubt that because of its economical importance and the high degree of synteny between *M. truncatula* and *P. sativum* (Choi et al. 2004), additional studies on pea germplasm will be conducted in next few years.

In this work, we aimed to evaluate different housekeeping genes for their potential use as internal normalization controls in order to more accurately measure the expression of genes of interest in pea. Eight reference genes were chosen based on their previous use as internal controls in plant gene expression studies, the availability of their gene sequences in *P. sativum* and their cellular function. We chose a group with varied roles in cellular processes (regulation of phosphorylation [*PDF2*], DNA replication (*Heli*), glucose metabolism [*GAPDH*], cytoskeletal structure [*TUA*, *TUB*, and *ACT*], protein biosynthesis [*EF-1α*] and ribosomal structure [*18S rRNA*]) in order to reduce the likelihood that they exhibited regulated covariation.

We show that some of the housekeeping genes that have been typically used in expression studies in the pre-genomic era have expression that is differentially regulated across different tissues, treatments and genotypes. Another set of control genes were identified as being the most stable. Our assessment of the validity of using these reference genes as internal controls is likely to be applicable to gene expression studies using real-time PCR techniques in *P. sativum*.

Materials and methods

Plant material

Seeds of *P. sativum* were germinated in filter paper and kept in the dark at 20ºC for 5 days. Seedlings with roots between 5-7 cm were placed in square Petri dishes containing a sheet of glass-fibre filter paper and perlite as a substrate. Plants were grown vertically on Hoagland nutrient solution under long-day conditions (16h day/8h night) at 23ºC. Fifteen day-old *P. sativum* cv. Athos plants were osmotically stressed upon adding 100mM NaCl to the medium. After 24h, roots from stressed plants were harvested. Hormone-treated *P. sativum* cv. Athos plants were produced by adding 5µM 2,4-dichlorophenoxyacetic acid (Sigma) to 15-day-old plants. Root and leaf tissues were harvested 24h after the hormone addition. To biotically stress the plants, two different genotypes with different sensitivities to the parasitic plant *Orobanche crenata* were selected: the susceptible cv. Messire and the incomplete resistant accession Ps624 (Rubiales et al. 2005). *O. crenata* seeds were inoculated at a density of 50 seeds cm^{-2}. Root samples were harvested before (15 days-post-inoculation) and after (21 days-post-inoculation) the attachment of *O. crenata* to the host plant. Unstressed "control" plants were harvested in parallel to obtain the same tissues at the same time intervals. Two serial experiments were performed.

RNA isolation

Total RNA from all the samples was isolated using TRIZOL reagent (Invitrogen, USA) according to manufacturer's protocols from different pools of five plants in order to minimize the individual plant variation in gene expression. The integrity of the total RNA was checked on formaldehyde 2% (w/v) agarose gels, and its quantity as well as purity was determined by measuring the optical density at 260nm and the A_{260}/A_{280}

absorption ratio using a NanoDrop ND-1000 spectrophotometer (Thermo Scientific). Only the RNA samples with A_{260}/A_{280} ratio between 1.9 and 2.1 and A_{260}/A_{230} greater than 2.0 were used in the analysis.

Primer design and real-time RT-PCR conditions

PCR primers were designed with the following criteria: T_m of 59ºC ± 1ºC and PCR amplicon lengths of 60-120 bp, yielding primer sequences with lengths of 18-25 nucleotides and GC contents of 50%-65% (Table 1). The PCR efficiency (E) of each primer pair in each individual reaction was estimated from the data obtained from the exponential phase of each individual amplification plot and the equation $(1+E)=10^{slope}$ (Ramakers et al. 2003). With this method, the E value is derived from the log slope of the fluorescence versus cycle number curve for each particular primer pair, does not require standard curves and yields very similar amplification efficiencies compared to methods based on series of template dilutions (Czechowski et al. 2005). To avoid any genomic contamination, total RNA (1µg) was reverse-transcribed with a blend of oligo-dT and random primers using the QuantiTec Reverse Transcription Kit (QuiagenGermany), according to the manufacturer's instructions. This kit ensures complete digestion of genomic DNA with a brief incubation of the samples at 42ºC in a specific Wipeout buffer before retrotranscription. Polymerase chain reactions were performed in a 96-well plate with a Mx3000P Real-Time PCR System (Stratagene, USA), using SYBR Green to monitor dsDNA synthesis. Reactions contained 0.5 µl 50x SYBR Green Solution, 7.5 µl 2x SensiMix (dT) (Quantace, London), 1.5 µl of cDNA (corresponding to 10ng of cDNA) and 200 nM of each gene-specific primer in a final volume of 15 µl. The following standard thermal profile was used for all PCR reactions: polymerase activation (95°C for 10 min), amplification and quantification cycles

repeated 40 times (95°C for 1 min, 60°C for 1 min). The specificity of the amplicons was checked by electrophoresis in 2% (w/v) agarose gel and a melting-curve analysis performed by the PCR machine after 40 amplification cycles (60 to 95ºC with one fluorescence read every 0.6ºC). All investigated RT-PCR products showed only single peaks and no primer-dimer peaks or artifacts. Two biological repetitions were used for the measurement, and two technical replicates were analyzed for each biological repetition.

Data analysis

Data were analyzed using the Mx3000P analysis software v4.0 (Stratagene). All amplification plots were analyzed with an R_n threshold of 0.03 to obtain C_T (threshold cycle) values for each gene-cDNA combination. To determine which reference genes were best suited for transcript normalization, we used the statistical algorithm *geNorm* (Vandesompele et al. 2002). The C_T values transformed into quantities using the PCR efficiencies (Ramakers et al. 2003) for the tested genes in 26 different samples in order to use geNorm are given in Supplemental Table 1.

Results

RNA quality and overall gene expression

To evaluate the stability of the expression of eight commonly-used housekeeping genes, their transcription profiles were assessed by real-time PCR in a set of 26 cDNA samples that included various tissues and treatment series. Within a biological repetition for a tissue sample, the same pool of cDNA from five plants was used to analyze each of the eight genes using gene-specific primers. Real-time PCRs were performed in

duplicate for each of the 26 cDNA pools. High quality total RNA was obtained and evaluated by denaturing formaldehyde 2% agarose gel electrophoresis and absorbance ratios. All samples were pure and free from protein and organic pollutants derived from the RNA extraction. The melting-curve analysis performed by the PCR machine after 40 cycles of amplification showed that each of the eight primer pairs amplified a single product (Supplemental Fig. 1). The studied housekeeping genes displayed a wide range of expression levels, with the lowest mean C_T value (15.48) in *18S rRNA* and the highest (22.23) in *PDF2*. The first analysis of the data showed that individual control genes had different expression levels across all studied samples. *PDF2* and *TUB* have the smallest variation in gene expression (below 2.5 cycles), while *18S rRNA* and *GAPDH* are the genes with the most variable levels of expression (over eight cycles; Fig.1).

Stability analysis

To analyze the stability of expression and identify the most suitable reference genes, we used the statistical algorithm geNORM v3.5 (http://medgen.ugent.be/~jvdesomp/genorm/). The stability measure relies on the principle that the expression ratio of two ideal reference genes is identical in all samples, regardless of the experimental condition and cell-type (Vandesompele et al. 2002). The program defines a stability measure (*M)* as the average pairwise variation between a gene and all other reference genes in a given set of samples. Genes with the lowest *M* values have the most stable expression. We analyzed our data, and the average expression stability values *M* are shown in Fig. 2a. The *M* value for *18S rRNA* and *GAPDH* was considerably higher (over the default software limit of *M*=1.5) than for the rest of the control genes. *TUB* and *ACT* had the lowest *M* values (0.48), corresponding to the most stable expression. Subsequently, the optimal number of internal control

genes required for reliable normalization of real-time PCR data was determined. To obtain this, normalization factors (NF) are calculated for the most stable genes and then for the next one by stepwise inclusion of the control gene that remains most stable. Vandesompele et al. (2002), proposed 0.15 as a cutoff value for the pairwise variation (*V*); below this level, the inclusion of an additional control gene is not required. In our experiment, the pairwise variation was significantly decreased with the inclusion of the fifth gene (Fig. 2b).

Validation of stability data

To asses the validity of the established gene-stability measure (that is, that genes with the lowest *M* values have the most stable expression), we determined the gene-specific variation for the three most stable genes as the coefficient of variation of the expression levels after normalization, following the approach outlined by Vandesompele et al. (2002). Three different normalization factors were calculated based on the geometric mean of the genes with the lowest ($NF_{4(1-4)}$, $NF_{5(1-5)}$) and highest ($NF_{2(7-8)}$) *M* values (as determined by geNorm). The gene-specific variation is higher when the data are normalized to $NF_{2(7-8)}$. In contrast, the smallest variation is detected when a normalization factor based on the reference genes with the lowest *M* values is used (Fig. 3).

Discussion

Normalization of the expression to a reference gene is a simple method that is frequently used to internally control for errors in real-time RT-PCR. The most commonly used reference genes include those involved in basic cellular processes, such

252 as *18S rRNA*, *ACT*, *TUB* and *GAPDH,* due to the fact that they were used for many
253 years as references in Northern blots and conventional RT-PCR assays. The advent of
254 post-genomic technologies, however, has raised the question of whether such genes are
255 actually suitable for normalization purpose. Numerous studies have shown that the
256 expression of these 'classic' genes can be regulated under various situations
257 (Czechowski et al. 2005; Dheda et al. 2005). Furthermore, a number of studies have
258 included reference genes for normalization without any prior validation of their stable
259 expression. As a result, it has been suggested that at least three proper control genes be
260 used for normalization, and statistical algorithms such as geNORM have been
261 developed for reliable normalization. Thus, plant gene expression studies have been
262 conducted to validate the use of particular internal controls and have shown that
263 housekeeping genes are regulated differently in different plant species. A reference gene
264 with stable expression in one organism may exhibit a different expression pattern in
265 another organism. Here, we describe an assessment of eight housekeeping genes for
266 their use as internal controls in gene expression studies in a given set of cDNA samples
267 containing different cell-types and treatments in pea. Several factors that could affect
268 the reliability of the data were carefully controlled during the experiments. RNA quality
269 control, DNAse I treatment, two-step RT-PCR and dissociation analysis by melting
270 curves were the critical points considered leading to a robust strategy to the analysis.
271 We further analyzed the stability of expression in this dataset. In order to consider any
272 variation in expression between biological replicates that is not due to the treatments,
273 we evaluated the biological replicates separated in the input panel in geNorm as has
274 been recently shown by Remans et al. (2008). Our results show that the housekeeping
275 genes were differentially expressed in the analyzed samples. Both *TUB* and *ACT* were
276 the most stably expressed genes, whereas *18S rRNA* and *GAPDH* were the least stable

277 genes. The clear decrease of *M* in the remaining reference controls during the stepwise
278 exclusion of these two worst-scoring genes demonstrates that cannot be used to
279 normalize gene expression data in pea.
280
281 To determine the optimal number of control genes for normalization, pairwise
282 variations, *V*, were assessed. After the stepwise inclusion of five control genes, there
283 was no significant effect on *V* value. Vandesompele et al. (2002), stated that the 0.15
284 value must not be taken as a strict cutoff, however. There are multiple factors—such as
285 time, resources, and accuracy requirements—that must be taken into account in order to
286 properly identify the number of most stable genes to be included in the experiment.
287 There was not a largely significant effect upon the inclusion of the fifth control gene
288 (Fig. 2B). The data clearly shows that the inclusion of four of the most stable control
289 genes may be adequate for the accurate normalization of pea gene expression data.
290
291 In order to identify real gene-specific variation, we determined the average gene-
292 specific variation for the three control genes with the smallest *M* value as the variation
293 coefficient of normalized expression levels. This coefficient should be minimal for
294 proper housekeeping genes. The raw expression data were standardized to different
295 normalization factors, which were calculated as the geomean of the four and five most
296 stable and the two least stable genes. Unstable reference genes cannot completely
297 remove variation; instead they add more, resulting in larger gene-specific variations for
298 the tested genes. This analysis clearly demonstrates that most specific variation is
299 removed when the four or five most stable control genes are used for normalization.
300 This result has serious implications for studies that have used unsuitable reference

genes. The choice of reference genes whose expression is regulated under the experimental conditions can result in unreliable data.

In conclusion, we have evaluated eight commonly-used housekeeping genes for their use as reference standards to normalize gene expression data. Our analysis revealed that *TUB*, *ACT*, *PDF2* and *Heli* are the most stably expressed genes. We recommend the use of these internal standards as a starting point to evaluate their expression stability in individual experimental systems to accurately normalize and quantify gene expression data in pea.

Acknowledgments

The authors acknowledge the USDA-ARS (Washington State University) for the supply of Ps624 seeds. Financial support was provided by the Spanish project INIA2007-0009.

References

Brunner AM, Yakovlev IA, Strauss SH (2004) Validating internal controls for quantitative plant gene expression studies. BMC Plant Biol 4:14-20.

Bustin SA (2000) Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. J Mol Endocrinol 25:169-193.

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

326 Bustin SA, Nolan T (2004) Pitfalls of quantitative real-time reverse transcription
327 polymerase chain reaction. J Biomol Tech 15: 155–166.
328
329 Choi HK, Mun JH, Kim DJ, Uhm T, Zhu H, Baek JM, Mudge J, Roe B, Ellis N, Doyle
330 J, Kiss GB, Young ND, Cook DR (2004) Estimating genome conservation between
331 crop and model legume species. PNAS 101:15289-15294.
332
333 Czechowski T, Stitt M, Altman T, Udvardi MK, Scheible WR (2005) Genome-wide
334 identification and testing of superior reference genes for transcript normalization in
335 *Arabidopsis.* Plant Physiol 139:5-17.
336
337 Dheda K, Huggett JF, Bustin SA, Johnson MA, Rook G, Zumla A (2004) Validation of
338 housekeeping genes for normalizing RNA expression in real-time PCR. BioTechniques
339 37:112–119.
340
341 Dheda K, Huggett JF, Chang JS, Kim LU, Bustin SA, Johnson MA, Rook GAW, Zumla
342 A (2005) The implications of using an inappropriate reference gene for real-time
343 reverse transcription PCR data normalization. Anal Biochem 344:141-143.
344
345 Gonçalves S, Cairney J, Moroco, Margarida M, Miguel C (2005) Evaluation of control
346 transcripts in real-time RT-PCR expression analysis during maritime pine
347 embryogenesis. Planta (2005) 222:556-563.
348
349 Gonzalez-Verdejo CI, Die JV, Nadal S, Jimenez-Marin A, Moreno MT, Román B
350 (2008) Selection of housekeeping genes for normalization by real-time RT–PCR:

Analysis of *Or-MYB1* gene expression in *Orobanche ramosa* development. Anal Biochem 379:176-181

Hong SY, Seo PJ, Yang MS, Xiang F, Park CM (2008) Exploring valid reference genes for gene expression studies in *Brachypodium distacyon* by real-time PCR. BMC Plant Biol 8:112

Iskandar HM, Simpson RS, Casu RE, Bonnet GD, Maclean DJ, Manners JM (2004) Comparison of reference genes for quantitative real-time polymerase chain reaction analysis of gene in sugarcane. Plant Mol Biol Rep 22:325-337.

Jain M, Nijhawan A, Tyagi AK, Khurana JP (2006) Validation of housekeeping genes as internal control for studying gene expression in rice by quantitative real-time PCR. Biochem Biophys Res Comm 345:646-651.

Kim B, Nam H, Kim S, Chang YJ (2003) Normalization of reverse transcription quantitative-PCR with housekeeping genes in rice. Biotechnol Lett 25:1869–1872.

Libault M, Thibivilliers S, Bilgin DD, Radwan O, Benitez M, Clough SJ, Stacey G (2008) Identification of four soybean reference genes for gene expression normalization. Plant Genome 1:44-54.

Nicot N, Hausman JF, Hoffmann L, Evers D (2005) Housekeeping gene selection for real-time RT-PCR normalization in potato during biotic and abiotic stress. J Exp Bot 56: 2907-2914.

376 Ramakers C, Ruijter JM, Deprez RH, Moorman AF (2003) Assumption-free analysis of
377 quantitative real-time polymerase chain reaction (PCR) data. Neurosc Lett 13:62-66.
378
379 Reid KE, Olsson N, Schlosser J, Peng F, Lund ST (2006) An optimized grapevine RNA
380 isolation procedure and statistical determination of reference genes for real-time RT-
381 PCR during berry development. BMC Plant Biol 6:27
382
383 Remans T, Smeets K, Opdenakker K, Mathijsen D, Vangronsveld J, Cuypers A (2008)
384 Normalisation of real-time RT-PCR gene expression measurements in *Arabidopsis*
385 *thaliana* exposed to increased metal concentrations. Planta 227:1343-1349.
386
387 Rubiales D, Moreno MT, Sillero JC (2005) Search for resistance to crenate broomrape
388 (*Orobanche crenata*) in pea germplasm. Gen Res Crop Evol 52:853–861.
389
390 Samuel Roberts Noble Foundation (2006) *Medicago truncatula* Handbook, U.
391 Mathesius, E.P. Journet, L.W. Sumner (eds.), www.noble.org/MedicagoHandbook/
392
393 Tu L, Zhang X, Liu D, Jin S, Cao J, Zhu L, Deng F, Tan J, Zhang C (2007) Suitable
394 internal control genes for qRT-PCR normalization in cotton fiber development and
395 somatic embryogenesis. Chin Sci Bull 52:3110-3117.
396
397 Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F
398 (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric
399 averaging of multiple internal control genes. Genome Biol 3:RESEARCH0034.
400

401 Wong ML, Medrano JF (2005) Real-time PCR for mRNA quantitation. BioTechniques
402 39:75-85.
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425

Figure legends

Figure 1. Expression levels of candidate housekeeping genes. Values are given as real-time RT-PCR cycle threshold numbers (C_T) in the 26 samples. Max. and Min. C_T values for each gene are shown.

Figure 2. geNorm output charts. (a) Expression stability and ranking of eight housekeeping genes in 26 cDNA samples. **(b)** Pairwise variation (*V*) to determine the optimal number of control genes. *V* values under the 0.15 threshold line indicate that there is no need to include further reference genes in the calculation of a reliable normalization factor.

Figure 3. Validation of the gene-stability measure. The average gene-specific variation (determined as the coefficient of variation, in percent) for the three reference genes with the smallest variation within the 26 tested tissues after normalization with three different factors calculated as the genomic mean of the reference genes with the lowest ($NF_{4(1-4)}$, $NF_{5(1-5)}$) and highest ($NF_{2(7-8)}$) gene stability values.