Selection and validation of reference genes for quantitative gene expression studies in *Erythroxylum coca* [version 1; peer review: 2 approved]

Teresa Docimo¹,²*, Gregor W Schmidt¹,³*, Katrin Luck¹, Sven K Delaney⁴, John C D’Auria¹

¹Max-Planck-Institut für Chemische Ökologie, Jena, D-07745, Germany
²Current address: Instituto Biologia e Biotecnologia Agraria (CNR), Milan, 20133, Italy
³Current address: Department of Biosystems Science and Engineering, Eidgenössische Technische Hochschule Zurich, Basel, 5048, Switzerland
⁴School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, 2052, Australia

* Equal contributors

**Abstract**

Real-time quantitative PCR is a powerful technique for the investigation of comparative gene expression, but its accuracy and reliability depend on the reference genes used as internal standards. Only genes that show a high level of expression stability are suitable for use as reference genes, and these must be identified on a case-by-case basis. *Erythroxylum coca* produces and accumulates high amounts of the pharmacologically active tropane alkaloid cocaine (especially in the leaves), and is an emerging model for the investigation of tropane alkaloid biosynthesis. The identification of stable internal reference genes for this species is important for its development as a model species, and would enable comparative analysis of candidate biosynthetic genes in the different tissues of the coca plant. In this study, we evaluated the expression stability of nine candidate reference genes in *E. coca* (*Ec6409, Ec10131, Ec11142, Actin, APT2, EF1α, TPB1, Pex4, Pp2aa3*). The expression of these genes was measured in seven tissues (flowers, stems, roots and four developmental leaf stages) and the stability of expression was assessed using three algorithms (geNorm, NormFinder and BestKeeper). From our results we conclude that *Ec10131* and *TPB1* are the most appropriate internal reference genes in leaves (where the majority of cocaine is produced), while *Ec10131* and *Ec6409* are the most suitable internal reference genes across all of the tissues tested.

**Keywords**

qRT-PCR; qPCR; E. coca; Reference genes; Normalization
Corresponding author: John C D'Auria (dauria@ice.mpg.de)

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Max Planck Society and an Alexander von Humboldt Foundation postdoctoral fellowship to JCD.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2013 Docimo T et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

How to cite this article: Docimo T, Schmidt GW, Luck K et al. Selection and validation of reference genes for quantitative gene expression studies in *Erythroxylum coca* [version 1; peer review: 2 approved] F1000Research 2013, 2:37 (https://doi.org/10.12688/f1000research.2-37.v1)

First published: 08 Feb 2013, 2:37 (https://doi.org/10.12688/f1000research.2-37.v1)
Introduction

*Erythroxylum coca* has been cultivated by humans for more than 8000 years and has been selected for high-level production of cocaine, a pharmacologically active tropane alkaloid. Cocaine and other tropane alkaloids such as atropine and scopoline act on the nervous system, and their activity is largely due to their common chemical backbone (the tropane nucleus). Despite the socioeconomic importance of cocaine and other tropane alkaloids, the molecular basis for the biosynthesis of the tropane nucleus remains unknown. *E. coca* is emerging as a model for the investigation of tropane alkaloid synthesis\(^4\), and shows high-level, localized tropane alkaloid production and storage in its leaf tissue\(^4\).

We have performed metabolic and enzymatic studies to identify the molecular and biochemical basis of tropane alkaloid biosynthesis in *E. coca*, and have developed a number of genomic tools such as expressed sequence tag (EST) libraries and 454 sequence databases\(^3\). Quantitative real-time reverse-transcription PCR (qRT-PCR) would be a further source of information on candidate tropane alkaloid biosynthesis genes in the different tissues of the coca plant.

qRT-PCR is widely used to quantify and compare levels of gene transcription\(^7\). Variables such as RNA quality and the efficiencies of reverse transcription and PCR may compromise the accuracy and reliability of qRT-PCR, and so results are typically ‘normalized’ by comparison with one or more internal reference genes. The internal reference genes must be stably expressed, and the most stable reference genes vary widely in different species, tissues and sets of experimental conditions. Therefore, the identification of stable reference genes is a crucial step in the design of qRT-PCR experiments.

Traditionally, ‘Housekeeping’ genes such as actin, glyceraldehyde 3-phosphate dehydrogenase (GAPDH) and ubiquitin were used for data normalization\(^6\). These genes were widely assumed to have a uniform level of expression due to their involvement in fundamental cellular processes. However, evaluation of the expression stability of classical housekeeping genes in many species including *Arabidopsis thaliana*, *Oryza sativa*, *Zea mays* and *Linum usitatissimum*\(^4\) has revealed unstable expression of these genes under a range of experimental conditions. In addition, several novel reference genes have been shown to be more stably expressed than classical housekeeping genes\(^9\). Hence there is a need for systematic validation of internal reference genes in each organism and experiment\(^9,14\).

The stability of candidate internal reference genes may be assessed using a number of models, including geNorm\(^1\), NormFinder\(^6\) and BestKeeper\(^7\). These models differ significantly in their assumptions, and so candidate genes are often assessed with several of these algorithms\(^15\). geNorm iteratively calculates an expression stability value (M) for each candidate gene. This is based on the mean pairwise variation between the gene and the other candidate genes across all samples. Genes with lower M values are more stably expressed, and less stable genes (with higher M) are progressively excluded from the analysis. The optimal number of reference genes for qRT-PCR normalization may also be determined by identifying the smallest number of genes needed to minimize mean variation. By contrast, NormFinder estimates the standard deviation for each gene relative to the global expression of all genes included in the analysis, and genes with lower standard deviations are considered better reference genes. BestKeeper uses a third approach involving the calculation of a stability index (the ‘BestKeeper index’ or BKI), which is assumed to represent the highest level of stability because it includes all genes across all samples. The stability of each reference gene is assessed by its correlation with the BKI, with a high correlation indicating a more stable reference gene\(^16\). In this study, we evaluate the stability of nine candidate reference genes (*Ec6409*, *Ec10131*, *Eci11142*, *Actin*, *APT2*, *EF1a*, *TPB1*, *Pex4* and *Pp2aa3*) in a variety of *E. coca* tissues (four developmental leaf stages, stems, roots and flowers). We then identify the most stable internal reference genes using the geNorm, NormFinder and BestKeeper algorithms and present guidelines for transcript analysis in different tissues of *E. coca* by qRT-PCR.

Materials and methods

Plant material

*Erythroxylum coca* was obtained from the Bonn Botanical Garden. Plants were grown at 22°C under a photoperiod of 12 h light/12 h dark with relative humidities of 65% and 70% for light and dark conditions respectively (and fertilized once a week with Ferty 3 (15-10-15) and Wuxal Top N (Planta, Regenstauf, Germany).

The organs used for RNA extraction and qRT-PCR analysis were obtained from four-month old *E. coca* plants grown from rooted cuttings. Leaves in four developmental stages, roots, stems and flowers were analysed. The leaf developmental stages were: leaf buds; young expanding leaves in a rolled state (Stage 1); young expanded (unrolled) leaves (Stage 2); and fully mature leaves (Stage 3) (see Figure 1).

RNA extraction and cDNA synthesis

Total RNA was extracted from 100 mg of fresh plant tissue using a total RNA extraction kit (Invitrek, Berlin, Germany). Genomic DNA was removed by treatment with RNase-free DNase I (Qiagen, Hilden, Germany). RNA quality was assessed on an Agilent Bioanalyzer 2100 using a RNA 6000 Nano Kit (Agilent, Böblingen, Germany). RNA concentration was determined using a NanoDrop 2000 c spectrophotometer (NanoDrop Technologies, Wilmington, USA). cDNA was synthesized using a Super Script III First Strand Kit (Invitrogen, Karlsruhe, Germany) according to the manufacturer’s instructions. In brief, random hexamer primers and deoxynucleoside5’-triphosphates (dNTPs) were added to 5 µg total RNA and the mixture was incubated at 65°C for 5 min before brief chilling on ice. The first strand was then reverse transcribed by adding First Strand Buffer, 20 mM dithiothreitol and Super Script III reverse transcriptase to a final volume of 20 µl and incubating the mixture at 42°C for 1h. The resulting cDNA was diluted 1:20 (vol:vol) with deionized water and stored at -20°C.

Reference gene selection

Candidate reference genes were selected from an *E. coca* 454 sequence library\(^1\) based on their homology to previously reported reference genes in *A. thaliana*. Nine candidate reference genes with an E-value higher than 2e\(^{-2}\) were identified by BlastN comparison as orthologues to *Arabidopsis* genes: Expressed protein (Ec6409),...
Table 2

Efficiencies and regression coefficients were calculated in qBASE and regression analysis of the standard curve and ranged between 79% (Ec10131) and 97% (Actin). The coefficient of correlation ($r^2$) of the linear regression analysis was always greater than 0.986 as shown in Table 1, indicating a linear relationship between C values and log-transformed transcript quantities in the range of the standard curve.

To ensure that the primer pairs are specific for the desired sequence in all samples and do not target homologous transcripts in some sample subsets, a melting curve analysis of each sample was performed after PCR amplification (Supplementary Figure 2). A single peak in the melting curve specific for each primer pair was obtained for all samples, and no peak could be observed in the melting curves of the control reactions (NTC and negative control reverse transcriptase reactions).

Expression stability of candidate reference genes

The expression stability of the candidate genes were evaluated with the geNorm, NormFinder and BestKeeper algorithms (Table 2). C values were transformed to relative quantities using qBASE prior to analysis with geNorm and NormFinder, while C values and PCR efficiencies were used in BestKeeper. The cDNA samples were considered as either a single, diverse set derived from all organ
The optimal number of reference genes required for accurate normalization in the respective sample sets (all samples, leaf bud/leaf and mature tissues) was determined by calculating the mean variation in each normalization factor (V) and then observing the effect of iterative addition of the next most stable reference gene ranked according to variance within a defined group of samples. 

The ‘housekeeping’ genes Actin and EF1α were relatively unstable and were ranked at positions six and seven (respectively) in all sample sets. 

The calculation of the normalization factor (V)

BestKeeper ranks gene stability by calculating the correlation coefficient (r) between the expression of each candidate gene and the BestKeeper index (BKI; calculated using all genes across all samples). Across all of the samples tested, BestKeeper indicated that Actin (r = 0.784) and Ec6409 (r = 0.768) were the most stable, while Ec10131 was ranked as the least stable (r = 0.638). In the leaf bud/leaf sample subset, Actin (r = 0.869) and APT2 (r = 0.868) had the highest correlation with the BKI, and Ec10131 again showed the lowest correlation (r = 0.385). In the mature organs sample subset, Pex4 and APT2 were strongly correlated with the BestKeeper index (r = 0.767 and r = 0.724, respectively), whereas Ec10131 showed low correlation (r = 0.309) (Supplementary Table 1 and Supplementary Table 2).

To provide a further ranking of gene stability, the results were also evaluated with NormFinder, in which candidate reference genes are ranked according the variance of their expression relative to the expression variance within a defined group of samples. Pp2aa3 was the most stably expressed gene with the lowest expression variance (stability value of 0.291), followed by $E_c$6409 and $E_c$11142, when all samples were included in the calculation. When the leaf bud/leaf and mature organ subsets of samples were considered, the rankings varied considerably (Supplementary Table 1 and Supplementary Table 2). Actin, APT2 and Pex4 were always ranked as the seventh, eighth and ninth most stable reference genes (respectively), but there was no consistent order of ranking for the other reference genes. The NormFinder rankings were also distinct from the geNorm rankings, although both algo-
Raw Ct values and relative quantities for Erythroxylum coca reference genes

http://dx.doi.org/10.6084/m9.figshare.154973

Discussion

Real time RT-PCR has become a central technique for the evaluation of quantitative changes in gene expression. Reliable and accurate expression data can only be obtained by normalization with stably expressed reference genes. Normalization is an essential prerequisite for the correct measurement of gene expression changes in different plant tissues, organs, developmental stages or treatments of a given plant species and is highly influenced by the choice of reference genes. Traditional reference genes (e.g. actin and ubiquitin) are useful as stable reference genes in some experiments, but their expression is often highly variable and is often inferior to the stability of less-commonly used genes. Therefore it is important to assess the expression stability of several candidate reference genes before gene expression studies are performed. Several models including geNorm, NormFinder and BestKeeper have been developed to rank candidate reference genes on the basis of their expression stability. These methods often vary in their stability rankings and so expression data is commonly analysed using several approaches.

In this study, we report the identification and validation of nine candidate reference genes in E. coca (Ec6409, Ec10131, Ec11142, Actin, APT2, EF1α, TPB1, Pex4 and Pp2aa3). These genes were identified by analysing a 454 E. coca sequence library for sequences with homology to the top 100 reference genes of Arabidopsis, on the assumption that homologous genes are likely to have similar expression patterns. Primer pairs specifically targeting the E. coca transcripts were successfully developed and evaluated: all primer pairs produced only the expected amplicon and were highly efficient (Table 1 and Supplementary Figure 1). The relative stabilities of the candidate reference genes were then assessed using geNorm, BestKeeper and NormFinder (Table 2 and Supplementary Table 1 and Supplementary Table 2).

geNorm produced similar results in all sample sets. Ec6409 and Ec10131 were always identified as two of the three most stably expressed reference genes (although Ec10131 and TPB1 were most stable in the leaf bud/leaf sample subset), and Actin, APT2 and Pex4 were always identified as the least stable. geNorm may identify co-regulated genes as stable reference genes. However, exclusion of either Ec10131 or Ec6409 did not change the gene rankings (not shown), suggesting that their high ranking is not attributable to co-regulation.

BestKeeper yielded very different rankings to geNorm, and these varied according to the sample subset. The inconsistent results with BestKeeper may be explained by several features of the BestKeeper algorithm. Calculation of the BestKeeper index excludes genes with a standard deviation of more than one C_{t} value, which results in the exclusion of different genes in different sample sets. Extensive variation in C_{t} values is to be expected in a non-normalized data set, and so the algorithm may not be able to effectively distinguish between stable and unstable reference genes. In our experiments, the candidate E. coca reference genes showed very similar correlations with the BestKeeper index, suggesting that the algorithm could not distinguish between the genes to produce useful stability rankings. NormFinder produced a third ranking of gene stability that differed from both BestKeeper and geNorm. Pp2aa3 and Ec6409 were ranked as the most stably expressed genes when all samples were considered (Table 2). geNorm also identified Ec6409 as one of the most stable genes in the entire sample set. However, only Pp2aa3 was consistently ranked by NormFinder, geNorm and BestKeeper as one of the most stable genes in the leaf bud/leaf and mature organs sample sets, and there was no consistency between the algorithms in the order of ranking for the most stable genes (Table 2). The ranking of the least stable genes was more consistent: NormFinder identified Actin, APT2 and Pex4 as the least stable genes in all of the sample sets, and geNorm ranked these genes in the same order.

**Table 2.** Ranking of Erythroxylum coca reference gene stability in all Erythroxylum coca tissues according to the geNorm, BestKeeper and NormFinder algorithms.

<table>
<thead>
<tr>
<th>Gene rank</th>
<th>geNorm (M*, V_{n+1})</th>
<th>BestKeeper (correlation coefficient, r)</th>
<th>NormFinder (stability value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ec10131 / 6409 (0.28)</td>
<td>Actin (0.784)</td>
<td>Pp2aa3 (0.291)</td>
</tr>
<tr>
<td>2</td>
<td>Ec6409 (0.768)</td>
<td>Ec11142 (0.300)</td>
<td>Ec12131 (0.750)</td>
</tr>
<tr>
<td>3</td>
<td>Pp2aa3 (0.30; 0.095)</td>
<td>APT2 (0.765)</td>
<td>Ec11142 (0.300)</td>
</tr>
<tr>
<td>4</td>
<td>TPB1 (0.34; 0.083)</td>
<td>Pp2aa3 (0.737)</td>
<td>EF1α (0.304)</td>
</tr>
<tr>
<td>5</td>
<td>Ec11142 (0.38; 0.080)</td>
<td>EF1α (0.73)</td>
<td>TPB1 (0.339)</td>
</tr>
<tr>
<td>6</td>
<td>EF1α (0.50; 0.115)</td>
<td>Pex4 (0.715)</td>
<td>Ec10131 (0.305)</td>
</tr>
<tr>
<td>7</td>
<td>Actin (0.62; 0.125)</td>
<td>TPB1 (0.688)</td>
<td>Actin (0.483)</td>
</tr>
<tr>
<td>8</td>
<td>APT2 (0.72; 0.120)</td>
<td>Ec11142 (0.661)</td>
<td>APT2 (0.596)</td>
</tr>
<tr>
<td>9</td>
<td>Pex4 (0.88; 0.147)</td>
<td>Ec10131 (0.638)</td>
<td>Pex4 (0.904)</td>
</tr>
</tbody>
</table>

*M indicates stability values listed from most stable to least stable.
The NormFinder, BestKeeper and geNorm models have been shown to produce conflicting stability rankings in many studies\(^\text{18,29}\). The rankings produced by one or more of the models may be combined to produce a hybrid ranking\(^\text{18}\), but this complicates the analysis by merging models with very different underlying assumptions. Hence, we favour using a single model when possible.

geNorm produced a consistent gene ranking across all of our samples, and provides a clear rationale for determining the minimum number of genes required for accurate normalization. We therefore recommend the use of Ec10131 and Ec6409 as internal reference genes for most *E. coca* sample sets. If leaves and leaf buds are the primary organs of interest, then we recommend the use of Ec10131 and TPB1. These results provide a foundation for qRT-PCR studies in *E. coca*, and will further its development as a model of tropane alkaloid biosynthesis.

**Author contributions**

TD, GWS and JCD designed the research; TD, GWS, KL, and JCD performed the research; TD, GWS, and JCD analyzed the data; TD, GWS, SKD and JCD wrote the paper. All authors have approved the final manuscript for publication.

**Competing interests**

No competing interests were disclosed.

**Grant information**

This work was supported by the Max Planck Society and an Alexander von Humboldt Foundation postdoctoral fellowship to JCD.

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

---

**Supplementary figures**

**Supplementary Figure 1.** Specificity of qRT-PCR primers and amplicon length. The lane marked M represents a 1 Kb ladder (Invitrogen, California) used for size comparisons.
Supplementary Figure 2. Melting curve analysis of RT-PCR products. NTC indicates: no template control.
# Supplementary tables

## Supplementary Table 1. Ranking of *Erythroxylum coca* reference gene stability in a sample subset containing only leaf tissues (Buds, Leaf Stage I-III). Analysis was performed using the geNorm, BestKeeper and NormFinder algorithms.

<table>
<thead>
<tr>
<th>Gene rank</th>
<th>geNorm (M*, V_{n/n+1})</th>
<th>BestKeeper (correlation coefficient, r)</th>
<th>NormFinder (stability value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ec10131/TPB1 (0.26)</td>
<td>Actin (0.869)</td>
<td>Ec6409 (0.176)</td>
</tr>
<tr>
<td>2</td>
<td>APT2 (0.868)</td>
<td>EF1α (0.264)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Ec6409 (0.30; 0.096)</td>
<td>Ec6409 (0.837)</td>
<td>Pp2aa3 (0.306)</td>
</tr>
<tr>
<td>4</td>
<td>Pp2aa3 (0.34; 0.089)</td>
<td>EF1α (0.805)</td>
<td>TPB1 (0.318)</td>
</tr>
<tr>
<td>5</td>
<td>Ec11142 (0.40; 0.088)</td>
<td>Pex4 (0.762)</td>
<td>Ec11142 (0.366)</td>
</tr>
<tr>
<td>6</td>
<td>EF1α (0.48; 0.100)</td>
<td>TPB1 (0.733)</td>
<td>Ec10131 (0.415)</td>
</tr>
<tr>
<td>7</td>
<td>Actin (0.68; 0.161)</td>
<td>Pp2aa3 (0.652)</td>
<td>Actin (0.642)</td>
</tr>
<tr>
<td>8</td>
<td>APT2 (0.78; 0.128)</td>
<td>Ec11142 (0.554)</td>
<td>APT2 (0.664)</td>
</tr>
<tr>
<td>9</td>
<td>Pex4 (0.90; 0.133)</td>
<td>Ec10131 (0.385)</td>
<td>Pex4 (0.815)</td>
</tr>
</tbody>
</table>

*M indicates stability values listed from most stable to least stable.

## Supplementary Table 2. Ranking of *Erythroxylum coca* reference gene stability in a sample subset containing only mature organs (Leaf stage III, Flowers, Roots, Stems). Analysis was performed using the geNorm, BestKeeper and NormFinder algorithms.

<table>
<thead>
<tr>
<th>Gene rank</th>
<th>geNorm (M*, V_{n/n+1})</th>
<th>BestKeeper (correlation coefficient, r)</th>
<th>NormFinder (stability value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ec10131/6409 (0.17)</td>
<td>Pex4 (0.767)</td>
<td>EF1α (0.226)</td>
</tr>
<tr>
<td>2</td>
<td>APT2 (0.724)</td>
<td>Pp2aa3 (0.250)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Ec11142 (0.24; 0.088)</td>
<td>EF1α (0.693)</td>
<td>Ec11142 (0.273)</td>
</tr>
<tr>
<td>4</td>
<td>Pp2aa3 (0.27; 0.066)</td>
<td>Actin (0.576)</td>
<td>Ec6409 (0.332)</td>
</tr>
<tr>
<td>5</td>
<td>TPB1 (0.30; 0.064)</td>
<td>Pp2aa3 (0.538)</td>
<td>Ec10131 (0.395)</td>
</tr>
<tr>
<td>6</td>
<td>EF1α (0.45; 0.121)</td>
<td>TPB1 (0.452)</td>
<td>TPB1 (0.436)</td>
</tr>
<tr>
<td>7</td>
<td>Actin (0.61; 0.136)</td>
<td>Ec6409 (0.442)</td>
<td>Actin (0.522)</td>
</tr>
<tr>
<td>8</td>
<td>APT2 (0.74; 0.133)</td>
<td>Ec11142 (0.441)</td>
<td>APT2 (0.638)</td>
</tr>
<tr>
<td>9</td>
<td>Pex4 (0.97; 0.189)</td>
<td>Ec10131 (0.309)</td>
<td>Pex4 (1.170)</td>
</tr>
</tbody>
</table>

*M indicates stability values listed from most stable to least stable.
References


Open Peer Review

Current Peer Review Status: ✓ ✓

Version 1

Reviewer Report 15 February 2013
https://doi.org/10.5256/f1000research.1225.r773

© 2013 O'Connor S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✓ Sarah O'Connor
The John Innes Centre, Norwich, UK

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 11 February 2013
https://doi.org/10.5256/f1000research.1225.r763

© 2013 McCormick S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✓ Sheila McCormick
Plant Gene Expression Center, University of California, Berkeley, Berkeley, CA, USA

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com