scientific reports

OPEN

Determination of reliable reference genes for gene expression studies in Chinese chive (*Allium tuberosum*) based on the transcriptome profiling

Jing Tong[1,2,3], Manman Hu[1,2,3], Beibei Han[1,2], Yanhai Ji[1,2], Baoju Wang[1,2], Hao Liang[1,2], Mingchi Liu[1,2], Zhanhui Wu[1,2]✉ & Ning Liu[1,2]✉

Chinese chive (*Allium tuberosum*) is widely cultivated around the world for its unique flavor, nutrient, and medicinal values, yet its molecular mechanism on flavor formation and other metabolic pathways remains intangible. The elucidation of these complex processes begins with investigating the expression of the genes of interest, however the appropriate reference genes (RGs) for normalizing the gene expression are still unavailable in *A. tuberosum*. To fill this lacuna, transcriptome-wide screening was undertaken to identify the most stable genes according to the analysis of their FPKM values. The expression stability of the RGs was further evaluated using geNorm, NormFinder, BestKeeper, and RefFinder algorithms. The comprehensive analysis showed that *GLY1* and *SKP1*, instead of two traditionally used RGs (*eIF1α* and *ACT2*), were the most stable genes across diverse *A. tuberosum* tissues, indicating the necessity to carefully validate the stability of RGs prior to their use for normalizations. As indicated by geNorm, the normalizations with at least two RGs could give more accurate results. qRT-PCR experiments were conducted with randomly selected genes, demonstrating that normalization with a combination of *GLY1* and *SKP1* resulted in reliable normalization results. Our finding represents the first attempt toward establishing a standardized qRT-PCR analysis in this economically important vegetable.

Chinese chive (*Allium tuberosum* Rottler ex Spr), a perennial herb plant native to North China, has become more and more popular in the food inventory owing to its garlicky flavor, abundant nutrient, and great medicinal properties[1–3]. The first description of Chinese chive used as a vegetable plant is found in the 'Classic of Poetry (Shijing)', the oldest existing collection of Chinese poetry written in the Zhou dynasty (1046–256 BC) of ancient China, suggesting its long cultivation history in China[4]. In 2018, the planting area of Chinese chive reached nearly 400,000 hectares, with an estimated annual yield of 30 million tons in China[5]. Moreover, the popularity of this vegetable extends far beyond China. Japanese, Korean, Indian, and other Asian countries. For example, Chinese chive is frequently used in meat and seafood recipes in Japan.

Next-generation sequencing technology identified a number of genes linked with flavor biosynthesis as well as other metabolic processes in *A. tuberosum*[6,7]. Understanding the precise roles of these genes requires investigations on their expression profiles across different tissues and organs. However, only a few genes in *A. tuberosum* were examined by Northern blot analysis where ribosomal RNA was served as loading controls[8,9]. Unlike the many agriculturally important plants like rice, cabbage, tomato[10–12], no report is available concerning the RG selection in *A. tuberosum*. Consequently, the absence of reliable RG(s) has dragged the application of the qRT-PCR method, which hampered the research on flavor formation as well as other metabolic processes in *A. tuberosum*.

[1]Key Laboratory of Urban Agriculture (North) of Minstry of Agriculture and Rural Affairs, Beijing Vegetable Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China. [2]National Engineering Research Center for Vegetables, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China. [3]These authors contributed equally: Jing Tong and Manman Hu. ✉email: wuzhanhui@nercv.org; liuning@nercv.org

The emergence of quantitative real-time reverse transcription-polymerase chain reaction (qRT-PCR) addresses the evident requirement for quantitative analysis in gene expressions. Compared to the Northern blotting, qRT-PCR provides a more sensitive, reproducible, and precise approach for the detection of gene expressions using a range of fluorescent report dyes that correlate the yield of PCR product with fluorescence intensity[13–15]. Gene expression data generated from qRT-PCR can be analyzed by two different approaches, absolute quantification, and relative quantification[16–18]. In absolute quantification, the expression data are determined using a standard curve generated usually based on the serially diluted standards of known concentrations[19]. However, it is labor-intensive to generate a reliable standard curve for each target gene and to include these standards in each PCR, which limits the usability of the absolute quantification in gene expression analysis[19,20]. In practice, the relative quantification method is extensively employed for the calculation of gene expressions in most laboratories. During relative quantification, the expression of target genes is calculated by the inclusion of RGs as internal controls[15,19]. Therefore, it is a prerequisite to select optimal internal control gene(s) for the normalizations under given experimental conditions.

The ideal RGs, in the relative quantifications, should be expressed constantly and stably in all examined samples regardless of experimental conditions, such as different developmental stages, biological processes, treatments, and even different organs or tissues[19,20]. However, it is a huge challenge, in reality, to identify such an RG that meets the criterion. Historically, a myriad of so-called housekeeping genes such as actins (*ACT*), glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*), and ubiquitin (*UBQ*) are commonly used as internal control genes in many experiments, but recent studies have shown that their expressions are also affected by specific treatments, tissue difference, or other experimental conditions[16,17]. Besides, the veracity of qRT-PCR easily fluctuated with factors such as the quality of RNA isolation, the efficiency of cDNA biosynthesis, and PCR amplification efficiency[13]. As a result, it becomes indispensable to validate the expression stabilities and amplification efficiency of the RGs before the qRT-PCR analysis, even if housekeeping genes are employed for normalizations.

In this study, ten genes with stable expression in diverse tissues of *A. tuberosum* were screened as candidates according to the systematic survey of multiple organ-specific transcriptome analysis. Two commonly used traditional RGs, *eIF1α*, and *ACT2* in Chinese chive were also included in the comprehensive analysis by algorithms such as geNorm, NormFinder, and BestKeeper that assess the stability of genes based on the variance of quantification cycle values in each tissue samples. Computational analysis of twelve candidate genes identified that *GLY1* (DN374_c0_g1) and *SKP1* (DN253_c0_g1) displayed the most stable expression levels throughout the tissues and organs of Chinese chive. The two stable RGs were further validated by qRT-PCR, supporting that both or in combination could produce similar expression patterns as revealed by transcriptome analysis. Based on the analysis, we propose *GLY1* and *SKP1* are the preferred internal control genes for normalizing gene expression across diverse tissues of Chinese chive. To our knowledge, this study is the first analysis of the validation of RGs for the accurate determination of transcriptional patterns in Chinese chive and will facilitate expression studies on the genes associated development in this *Allium* crop.

Results

In silico screening of candidate reference genes from transcriptome data.

In a previous study, 18 RNA-Seq libraries representing leaf, root, rhizome, mature flower, inflorescence stalk with its associated flower buds, and seed tissues of *A. tuberosum* were sequenced with Illumina Hiseq4000 platform. The raw data of all RNA-Seq samples obtained in this study were deposited in the NCBI Sequence Read Archive under the project with identification number PRJNA67394520[7]. After removal of low-quality reads, ambiguous reads, and adaptor sequences, de novo assembled transcriptome was annotated using Nr (NCBI non-redundant protein) database, which enables the identification of 223,529 tentative transcripts in *A. tuberosum*[7].

The entire workflow of the screening process is illustrated in Fig. 1. As the housekeeping genes, commonly used as RGs, are expressed at relatively high levels, the first step was attempted to identify the genes with high expressions in all *A. tuberosum* tissues. In the annotated dataset, we noted that the FPKM values of housekeeping genes such as *ACTIN, eIFs,* and *UBQ*, were high-expressing genes (average FPKM $\geq$ 100) in tissue samples (Supplementary Table 1). Thus, the cut-off value of median FPKM was set to 100. Accordingly, low-expressing tentative transcripts (median FPKM $<$ 100) were eliminated, and the remaining 197 genes showed comparable expression levels as housekeeping gens (Supplementary Table 2).

The coefficient of variations (CV), the normalized median absolute deviations (NMADs), and the p-values measured by the *Shapiro-Wilks* were determined on the basis of the FPKM values of 197 genes. Then these genes were further ranked according to the Euclidean distance (*d*) calculated with the three parameters aforementioned (Supplementary Table 2), and primers were designed based on the sequences of the top 12 genes with the least values of Euclidean distance. DN47561_c0_g1 and DN5072_c0_g1 could not be amplified by PCR, while others were confirmed by the sequencing of their PCR products. Finally, the 10 stable genes with abundant expressions were further evaluated by using comprehensive bioinformatics analyses (Fig. 2). Additionally, based on the homologous blast search against *A. tuberosum* transcriptome, two traditional RGs (*ACT2* and *eIF1α*) which were frequently used as internal controls in Arabidopsis[21,22], were also included in the evaluation analysis.

Evaluation of primers' specificity and amplification efficiency.

Before experiments, gene-specific primers designed based on the assembled transcriptomes were firstly evaluated using the regular PCR method, and the resulting amplicons which contained unique bands with predicted sizes were confirmed by sequencing (Supplementary Fig. 1). Standard curves of ten candidate RGs and two controls were generated from a tenfold series dilution of at least five cDNA template concentrations and non-template control. As the R^2 value of the standard curve represents how well the experimental data fit the regression line, the R^2 value $>$ 0.980 is recom-

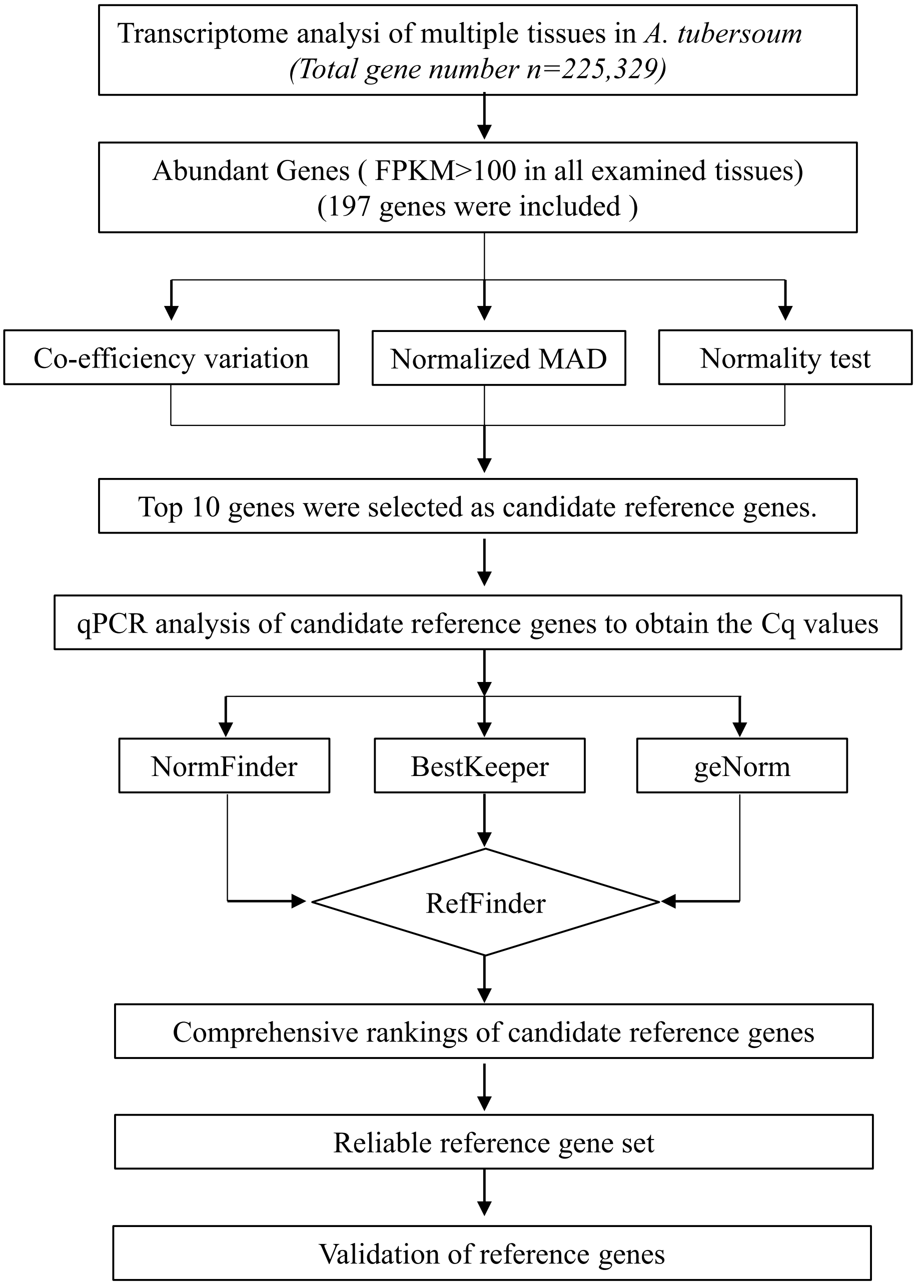
Transcriptome analysi of multiple tissues in A. tubersoum
(Total gene number n=225,329)
Abundant Genes (FPKM>100 in all examined tissues)
(197 genes were included)
Co-efficiency variation
Normalized MAD
Normality test
Top 10 genes were selected as candidate reference genes.
qPCR analysis of candidate reference genes to obtain the Cq values
NormFinder
BestKeeper
geNorm
RefFinder
Comprehensive rankings of candidate reference genes
Reliable reference gene set
Validation of reference genes

Figure 1. Working flow for the identification of reference genes in the experiment.

mended for the qRT-PCR reactions (Table 1 and Supplementary Fig. 2). Ideally, if the amount of PCR product perfectly doubled during each cycle, the theoretical amplification efficiency (E) should equal 95.4%, which indicates all templates were amplified during PCR reaction. However, in practice, an amplification efficiency of 90–105% is preferred, because qPCR was prone to be affected by suboptimal reaction conditions. As shown in Table 1 R^2 values of other candidates and control genes, except *CBS_X3* (DN6928_c0_g1), ranged from 0.983 to 0.999, suggesting that the equations of the linear standard curves could be used for further analysis; the E values ranged from 91.57 to 109.74, which indicated the high amplification efficiency of qPCR primers and optimal reaction conditions.

Expression analysis of candidate reference and control genes. Based on the Ct values obtained from qRT-PCR, the expression levels of the selected candidate and control genes were analyzed across various tissue of *A. tuberosum*. As shown in Fig. 3, the Ct values of all examined candidate genes ranged from 11.19 to 28.6, while the average Ct value (18.51) and the median Ct value (18.96) were fairly close, suggesting that the expression data of these selected genes nearly has a symmetrical distribution. Among these candidates, it seems that *HMG_B1* was the most abundantly transcribed genes, *Cyclase_X1* was the least abundant one according to the comparison of average Ct values.

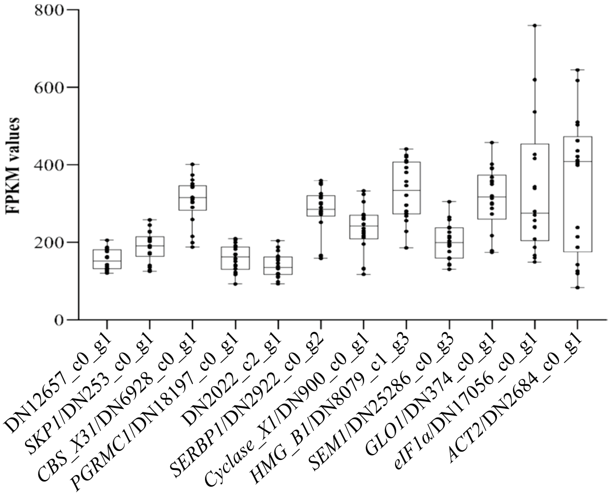

Figure 2. Expression levels of candidate RG and controls across various tissues based on the RNA-Seq data. The mean (horizontal line), upper and lower quartiles (white boxes), maximum and minimum values (whiskers), and each sample (black dots) are shown for individual gene.

No	Gene_id	Gene_name	NR description	Sequence (5′ → 3′)	Amplicon size (bp)	R^2	Amplification efficiency (%)
1	DN12657_c0_g1	N/A	N/A	CTCAACGCTCCACCGTTACT AGCCAACCAAAAATTCCACT	176	0.992	105.74
2	DN253_c0_g1	*SKP1*	SKP1-like protein 1A	GGGATGCCGATTTTGTTAAA TTTCCGGATCTCTTCTGGTG	149	0.998	92.88
3	DN6928_c0_g1	*CBS_X3*	CBS domain-containing protein, CBSX3	GGTGCTTTCAAATCCATGCT ATGCAAGCAATGCTATCACG	152	0.969	92.20
4	DN18197_c0_g1	*PGRMC1_2*	Membrane steroid-binding protein 2-like, PGRMC1_2	TTCTAGCTCGAAGGGTCCAA GATGTCACCCAGAGCAGGAT	150	0.995	94.59
5	DN2022_c2_g1	N/A	Uncharacterized protein A4U43_C06F2590	GGTAGCTGGTTCACCTTGGA AATTCCCCTGATGGTGATGA	159	0.998	107.86
6	DN2292_c0_g2	*SERBP1*	RGG repeats nuclear RNA binding protein A	CAACCGAAACGAGAATGGTT CGGCATCTCCTTCAACATTT	152	0.998	90.41
7	DN900_c0_g1	*Cyclase_X1*	Putative cyclase LOC109841113 isoform X1	AACAATGCAGGGCCATTAAG CCGAAATCGATGAAGAATGG	167	0.999	107.10
8	DN8079_c1_g3	*HMG_B1*	HMG transcription factor B1	AAAAGCAAAAGCTGCAAAGG CACTTATCACCAGCGGCTTT	150	0.983	99.13
9	DN25286_c0_g3	*SEM1*	Putative 26S proteasome complex subunit sem1-1	GGATCTTTTCGAGGATGACG TTCCTTAGCTGCAGGGAGAA	150	0.995	91.57
10	DN374_c0_g1	*GLY1*	Glyoxalase I homolog 1	TCAAGGAAAGCGGTGATCTT AGGGCAATGCTTATGGACAG	154	0.994	93.67
11	DN17056_c0_g1	*eIF1α*	Eukaryotic translation initiation factor 5A	TCTGACGAGGAGCATTTTGA ATGCTTGCCTGTTTTGGAAG	153	0.997	109.74
12	DN2684_c0_g1	*ACT2*	ACTIN 2	GGGCATCTGAATCTCTCAGC TCGTCCGTGACATCAAAGAA	151	0.997	98.48

Table 1. Details of primers and amplicons of 12 candidate reference genes.

Evaluation of expression stability of new candidate reference genes.

Several software options are widely utilized to assess the variability of RGs, including GeNorm, NormFinder, BestKeeper, and RefFinder[23–26]. Thus, we employed the four programs to analyze the expression stability of candidate RGs in our qPCR experiment.

geNorm M analysis.

The geNorm is the most commonly used software package that helps in selecting the best RGs. The geNorm *M* value indicates the average expression stability value of remaining RGs at each step during stepwise exclusion of the least stable RG. The lower geNorm *M* values represent the more stability of

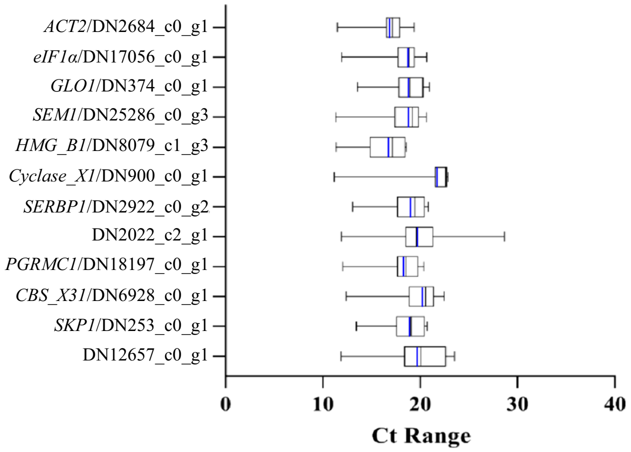

Figure 3. Variation of qRT-PCR Ct values for candidate RG and controls across diverse tissues in *A. tuberosum*. The vertical blue and black lines in the box show the mean and median values, respectively. The lower and upper boxes indicate the first and third quartile.

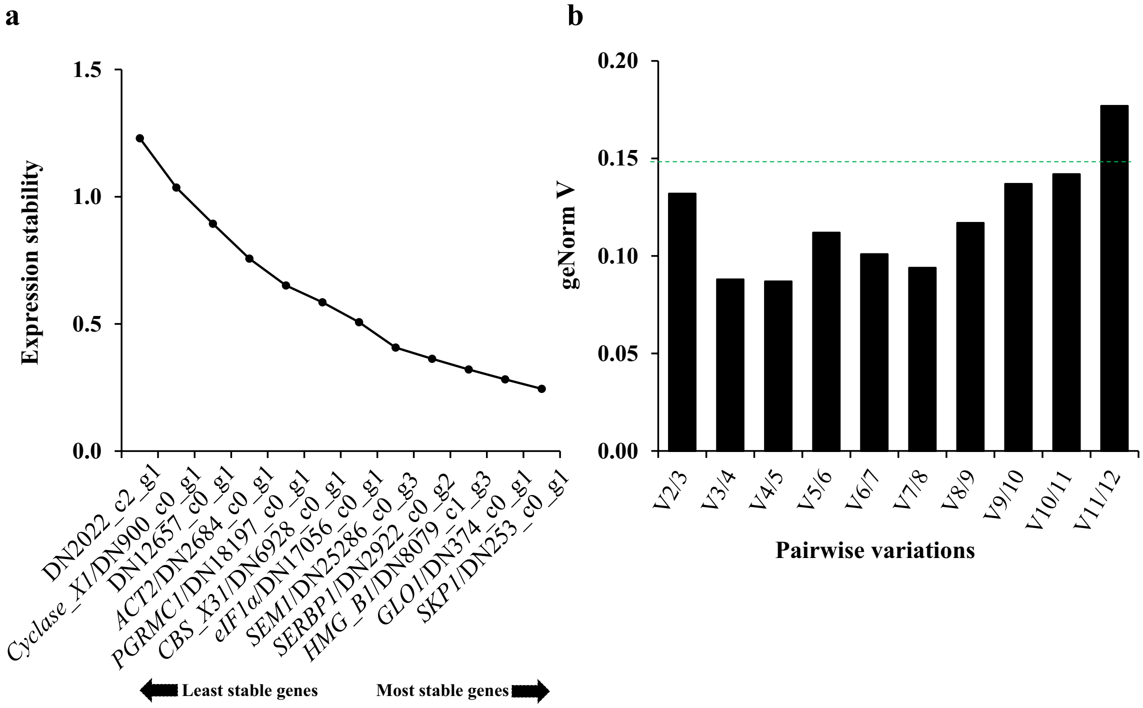

Figure 4. geNorm analysis of candidate RGs and controls. (**a**) Stability rankings of candidate reference and control genes based on the M values calculated by the geNorm algorithm. (**b**) Determination of the optimal number of reference genes for qRT-PCR normalization in different tissues. The variation value V was calculated by the geNorm algorithm. The default value of 0.15 (blue dash line) was taken as the threshold.

the RGs, and it was recommended that the *M* value of stable RGs should below 0.5. The geNorm *M* analysis for 12 candidate genes identified that five candidates, *SKP1*, *GLY1*/DN374_c0_g1, *HMG_B1*, *SERBP1,* and *SEM1*, displayed *M* values below 0.5 (Fig. 4a), suggesting that they might be housekeeping genes across various tissues.

NormFinder analysis. NormFinder offers a method of RGs selection that takes into account intra- and inter-group variability, while geNorm does not differentiate between groups of samples or treatments. Thus, NormFinder can prevent co-regulated genes from being selected from the candidate list. The stability value calculated by NormFinder reflects the variability of gene expression, and thus the lower value indicates more stability of RGs. Using NormFinder, *GLY1* and *SKP1* turned out to be the most stable genes, with expression stabilities below the cut-off value of 0.15; DN2022_c2_g1 was the most unstable gene across various tissues

	NormFinder		BestKeeper			Comparative Delta-Ct		RefFinder	
Rank	Genes	Stability	Genes	SD	CV	Genes	SV	Gene name	Ranking values
1	DN374_c0_g1	0.064	DN2684_c0_g1	1.43	8.47	DN2292_c0_g2	1.03	DN374_c0_g1	1.57
2	DN253_c0_g1	0.117	DN374_c0_g1	1.79	9.67	DN253_c0_g1	1.07	DN253_c0_g1	1.86
3	DN2292_c0_g2	0.183	DN253_c0_g1	1.84	9.98	DN374_c0_g1	1.07	DN2292_c0_g2	3.03
4	DN8079_c1_g3	0.364	DN18197_c0_g1	1.89	10.58	DN25286_c0_g3	1.12	DN8079_c1_g3	4.68
5	DN17056_c0_g1	0.489	DN17056_c0_g1	1.98	10.98	DN8079_c1_g3	1.15	DN2684_c0_g1	5.20
6	DN25286_c0_g3	0.499	DN2022_c2_g1	1.99	10.07	DN17056_c0_g1	1.19	DN17056_c0_g1	5.48
7	DN6928_c0_g1	0.679	DN2292_c0_g2	2.03	10.93	DN6928_c0_g1	1.22	DN25286_c0_g3	5.73
8	DN18197_c0_g1	0.690	DN8079_c1_g3	2.03	12.49	DN18197_c0_g1	1.31	DN18197_c0_g1	6.73
9	DN2684_c0_g1	1.071	DN25286_c0_g3	2.10	11.54	DN2684_c0_g1	1.57	DN6928_c0_g1	7.65
10	DN12657_c0_g1	1.296	DN6928_c0_g1	2.23	11.5	DN12657_c0_g1	1.65	DN2022_c2_g1	10.09
11	DN900_c0_g1	1.700	DN12657_c0_g1	2.69	13.82	DN900_c0_g1	1.92	DN12657_c0_g1	10.24
12	DN2022_c2_g1	3.732	DN900_c0_g1	2.70	12.98	DN2022_c2_g1	3.80	DN900_c0_g1	11.24

Table 2. Ranking of the candidate reference genes using NormFinder, BestKeeper, Comparative Delta Ct, and RefFinder methods.

(Table 2). Despite the algorithm difference, the two aforementioned programs predicted that *GLY*1 and *SKP1* were the most stable genes while DN2022_c2_g1 was the least stable gene.

BestKeeper analysis. BestKeeper assesses the stability of gene expression by comparing the standard deviation (SD) and the coefficient of variation (CV) of Ct values. Thereby, the most stable RG should have the lowest SD and CV according to the BestKeeper algorithm. The ranking result suggested that the top three stable genes were *ACT2*, *GLY1*, and *SKP1* (Table 2), indicating that they were the best RGs for normalization.

RefFinder analysis. Depending on the input of raw Ct values, RefFinder is a web-based computational program that integrates geNorm, Normfinder, BestKeeper as well as the comparative Delta-Ct method to evaluate the RGs. Based on the rankings from four programs, it generates the geomean of the ranking values of the tested RGs. The rankings by RefFinder suggested that *GLY1*, *SKP1*, and *SERBP1* were the most stably expressed genes (Table 2), which was largely in agreement with the other three algorithms used.

Analysis of the optimal number of reference genes for normalization. In addition to the calculation of gene stability, geNorm also calculates the pairwise variation between two sequential RGs used for normalization via geNorm V analysis. As previously proposed, the cut-off value of $V_{n/(n+1)}$ is above 0.15, the inclusion of an additional RG is required. As illustrated in Fig. 4b, the value of $V_{2/3}$ was 0.132, indicating that two references were sufficient for accurate qRT-PCR normalization. Considering the expression stabilities of candidate RGs, we thus concluded that the most stable combination was *GLY1* and *SKP1*, which could be used as reliable normalization factors for investigating the gene expression across different tissues in *A. tuberosum*.

Validation of the reference genes. To validate the applicability of the selected RGs, we conducted further qRT-PCR to analyze the expression profiles of several randomly selected genes from the tissue-specific transcriptome in *A. tuberosum*. Their expressions of these genes were normalized against either the most stable or unstable RGs. In general, qRT-PCR results are largely consistent with the expression patterns revealed by transcriptome analysis, when *GLY1* and *SKP1*, alone or in combination, were used as internal control genes (Fig. 5a–d). For instance, DN4756_c1_g1 shared 94% sequence similarity with the Arabidopsis *LHCB1* (Light-harvesting Chlorophyll A/B-protein 1) protein. As revealed by transcriptome data, in *A. tuberosum*, DN4756_c1_g1 was expressed in photosynthetic tissues such as the leaf, flower (sepal included), and inflorescence (with stalk). In the qRT-PCR experiments normalized against the stable RGs, we confirmed that DN4756_c1_g1 exhibited the highest expression in leaf tissues, with weak expression in flower, inflorescence, and seed tissues. Likewise, seed-abundant gene DN7391_c1_g1 encoded a putative vicilin-like seed storage protein homologous to At3g22640. Accordingly, we observed that DN7391_c1_g1 was highly expressed in seeds when using *GLY1* and *SKP1* as internal controls. Conversely, when the least stable gene, *Cyclase_X1*, was used as an internal control, DN4756_c1_g1 and DN7391_c1_g1 displayed expression patterns that differ from that of transcriptome analysis (Fig. 5e). A similar phenomenon was observed for the other three tested genes in *A. tuberosum*. Hence, the results demonstrated that, to a large extent, the RG determined the accuracy of qRT-PCR normalization. Furthermore, the results also confirmed that *GLY*1 and *SKP1* could be used as appropriate RGs in analyzing gene expression across various *A. tuberosum* tissues.

Discussion

As one of *Allium* crops with great economic importance, Chinese chive is well-known for its unique flavor which combines garlicky and sweet odors, and for the ability to enhance the flavor of other food[2,27,28]. Several genes encoding the key enzymes in flavor production pathways have been isolated in *A. tuberosum*, but the molecular

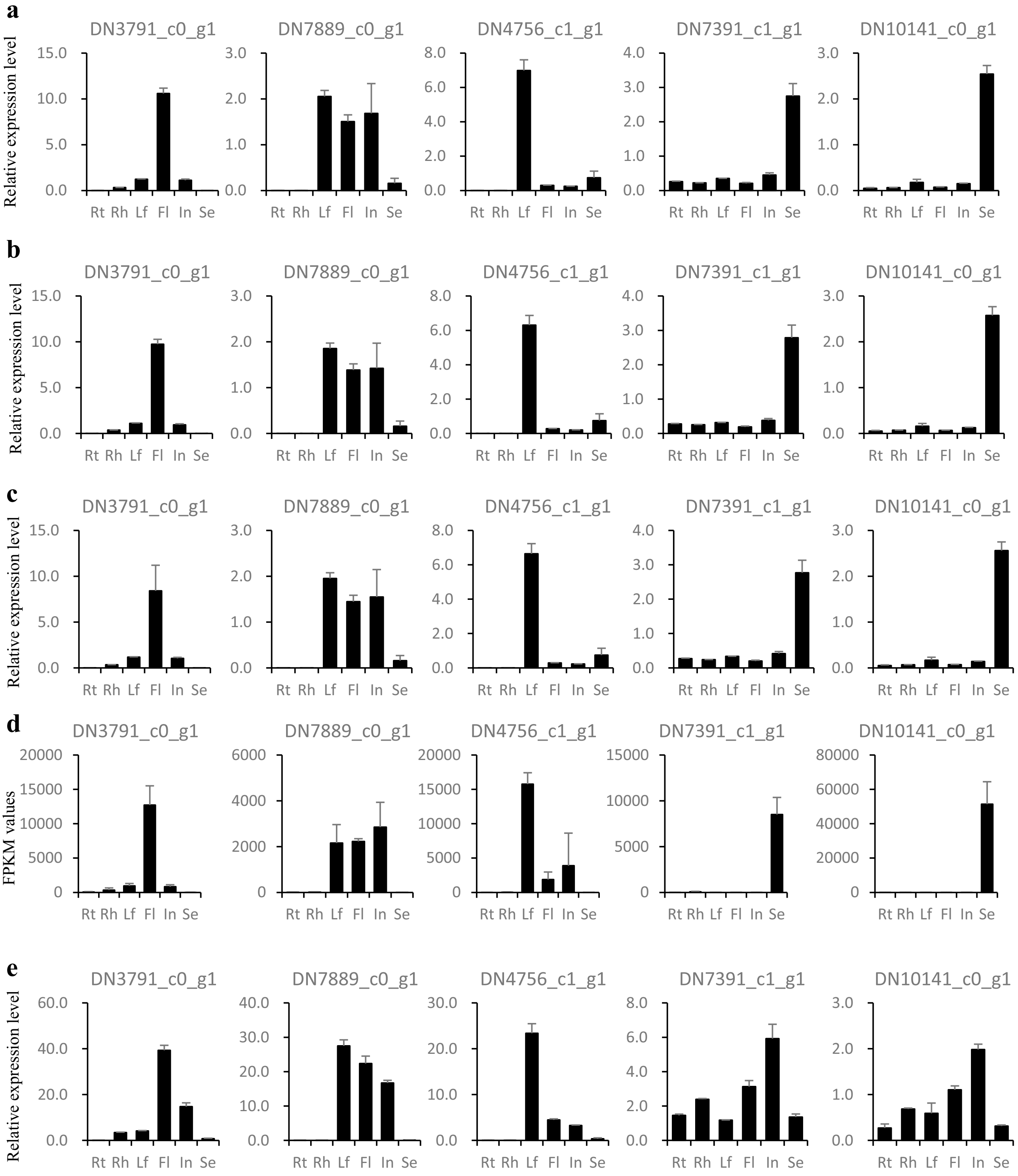

Figure 5. Expression analysis of five randomly selected genes across various *A. tuberosum* tissues. Expression patterns of selected genes were generated by qRT-PCR results normalized against the most stable genes, DN374_c0_g1 (**a**), DN253_c0_g1 (**b**), and in combination (**c**), or against the least stable gene DN900_c0_g1 (**e**). Expression level of selected genes in different tissues was presented according to the *A. tuberosum* transcriptome data (**d**). All these bars was represented by mean ± SD from three biological replicates.

basis of flavor production remains intangible[29]. Expression profiles of several genes encoding alliinase, cysteine synthase, and serine acetyltransferase were characterized by using the Northern blot method whereby ribosomal RNA was used as loading controls[8,9]. However, studies qRT-PCR to explore expression patterns of genes in

Chinese chive are still rare, and to date, no prior report on the reliable RGs has been published in *A. tuberosum* yet.

Since the relative quantification method has been widely used in the qRT-PCR analysis, reliable results are largely dependent on the selection of the appropriate RGs, expression of which is supposed to be stable in all samples, irrespective of all experimental conditions. Therefore, it is imperative to identify the stable genes that can be used as internal controls for qRT-PCR experiments. In earlier studies, the frequently used RGs in *Allium*s were usually derived from the identification of housekeeping genes that were associated with the basic cellular metabolism. For instance, *ACTIN* was used as RGs for normalizing the expression of waxy cuticle-related genes in welsh onion[30]. Similarly, an actin gene, *AsACT*, was served as internal control genes in the qRT-PCR analysis of GA-treated garlic plants[31]. However, no evidence about their constitutive expression patterns under the experimental conditions was available. In fact, the expression of the *ACT* gene could be influenced by pathogen attacks in Arabidopsis[32]. As reliable quantification of gene expression mainly depends on an accurate normalization, the selection of RGs is one of the crucial factors in investigating gene expressions under given experimental conditions.

High-throughput RNA-Seq technologies facilitate gene expression analysis especially for those species whose genomes are still unavailable. Apart from the sequence information of all the transcribed genes, the assembled transcriptome provides the landscape of gene expressions at the genome-scale, which permits the identification of the genes with stable expressions across different experimental conditions or samples. Depending on the calculation methods employed in the transcriptome analysis, gene expressions can be presented as FPKM (Fragments per kilobase per million mapped reads) value or others. Therefore, according to the FPKM values, researchers can examine the expression stability using the mean expression value, standard deviation, coefficient of variation (CV), and normalized median absolute deviation (NMAD). The CV reflects the extent of variability in relation to the mean of the expression levels, and it was successfully used to identify the best RGs from the transcriptome dataset[33–35]. However, the CV is simply calculated based on the means, which is less susceptible to the deviations caused by the outliers. Thus, the NMAD can be included to evaluate the spread of the distribution according to the medians of the expression levels, which provides a robust alternative criterion.

It was not surprising that different screening criteria might lead to contrasting results. As observed in our analysis, the rankings according to the CV values suggested that DN47561_c0_g1, DN12657_c0_g1, DN6928_c0_g1, DN253_c0_g1/*SKP1*, and DN18197_c0_g1 were the top 5 stable genes; however, DN5298_c0_g2, DN47561_c0_g1, DN6928_c0_g1, DN253_c0_g1/*SKP1*, and DN2292_c0_g2 were ranked as the top 5 stable genes when these RG candidates were scored by NMAD values. Thus, *DN47561_c0_g1, DN6928_c0_g1, DN253_c0_g1/SKP1* seemed to be the promising RG candidates because they could meet the two screening criteria. However, other candidates only met one of the criteria mentioned above, which made it difficult to pick up the ideal RG for further analysis. To solve the problem, CV and NMAD, together with the normality test of gene expressions (p), should be taken into account. Assuming that CV, NMAD, and p had equal contributions to the evaluation of RG's expression stability, the Euclidean distance (d) between the three parameters could be used as an effective measure in the RG selections from the huge RNA-Seq dataset.

A Venn diagram was generated with the top 20 stable RG candidates ranked by their CV, NMAD, or d values (Supplementary Fig. 3). As revealed by the Venn analysis, 12 RG candidates were shared by three rankings lists, suggesting that they might be the ideal RG candidates which met all three criteria. 19 genes were shared between d and CV ranking lists, while only one gene was unique for the CV ranking list. Likewise, 13 genes appeared on both NMAD and d ranking lists, whereas 7 genes formed a unique group that was specific to the NMAD ranking list. In our analysis with top 20 RG candidates, the outcomes with the CV and d criteria were similar to a large extent, though the ranking orders of some RGs were different; on the contrary, the NMAD might serve as a more restrictive criterion, and nearly half of genes appeared in the overlapping between both d and NMAD ranking list, implying that NMAD criterion alone is not sufficient to evaluate the stability of RG expressions. As the Euclidean distance can balance the effects from three parameters, we preferred to use d values instead of CV or NMAD to evaluate the first-round screenings, and in the subsequent bioinformatics analyses, the selected 10 RG candidates, which were further confirmed by PCR experiments, exhibited better expression stability compared to that of two traditional RGs.

By combining some bioinformatics tools, NormFinder, geNorm, BestKeeper, and RefFinder, a series of RG candidates were identified in non-model species, such as cotton, onion, soybean, and garlic[36–39]. In addition to traditionally used RGs—*eIF1α* and *ACT2*, the top 10 potential RGs identified from transcriptome analysis were quantitatively evaluated with the programs aforementioned. Surprisingly, the majority of RG candidates were associated with either ubiquitin or ribosomal pathways except that SEM1 might encode one of the components in the 26s proteasome[40]. Ubiquitin genes were often used to normalize the quantification of gene expression due to their abundance and universal expressions in diverse samples. Among the top 200 RGs, we found at least 13 genes that were associated with ubiquitin pathways, but the ranking list suggested that they did not appear to be the best RGs as some expression variations of these genes were detected (Supplementary Table 1). It has been generally accepted that, despite housekeeping roles in the intracellular protein degradation, the ubiquitin–proteasome system was involved in selective proteolysis, and therefore the process was tightly controlled and dynamically regulated[41]. Similarly, it becomes controversial whether the 18s or 28s ribosomal genes are perfect RGs. On one hand, like ubiquitin genes, transcription of some ribosomal genes was also affected by biological factors[26]; On the other hand, the ribosomal RNAs are not polyadenylated, which makes it difficult to investigate their expression when dealing with cDNA from total RNA with oligo-dT primers. It was reported that the use of ribosomal genes in expression studies could give erroneous normalizations compared with the use of other RGs[42,43]. Consistent with previous findings, in this study, the top 10 RG candidates could thus be used to normalize expression levels of genes of interest despite that most of the candidates might be not related to ubiquitin or ribosomal pathways.

The comprehensive rankings suggested that, of the 12 genes, *GLY1*/DN374_c0_g1 was the most stable gene, followed by SKP1/DN253_c0_g1; on the contrary, Cyclase_X1 was thought to be the least stable one. Interestingly, we found that the housekeeping genes, *eIF1α* and *ACT2,* were not constantly expressed despite that their Arabidopsis homologs were widely used as internal controls in molecular biological experiments[43,44]. As revealed herein, the most stable genes identified in our analysis also included genes involved in signaling pathways. For example, DN253_c0_g1 encoded a putative S-phase kinase SKP1, which might mediate the ubiquitination of proteins involved in cell cycle progression[45]. DN8079_c1_g3 was ranked as the fourth stable gene by RefFinder, though it might encode a putative a high mobility group protein (HMG) which could regulate chromatin remodeling and ultimately gene transcription[46]. Therefore, the results reminded us that the empirically selected 'housekeeping' genes could be unsuitable as the internal controls, and their stabilities should be carefully evaluated before experiments.

Compared to the other algorithms, only geNorm provides a method to define the optimal number of RGs for accurate normalization of gene expression via the pairwise variation V analysis. The cut-off Vn/n + 1 value of 0.15 is recommended, indicating that the preferred primer number is n; otherwise, an additional RG should be included. In our experiments, the V2/3 was below the threshold value of 0.15, suggesting that two RGs were sufficient for data normalization. When using two RGs alone and in combination, the qRT-PCR analysis of randomly selected genes showed similar expression patterns in our experiments. However, it is still recommended to always use at least two RGs to avoid substantial errors, because we could not deny the possibility that these stable genes might participate in other biochemical pathways other than plant development.

We analyzed the expression data of five randomly selected genes in *A. tuberosum* normalized against either the most or the least stable genes according to the comprehensive rankings given by the RefFinder algorithms, demonstrating that transcriptional patterns analyzed with the stable RG were almost the same as observed in transcriptome analysis. On the contrary, when the raw data normalized with the least stable gene, the qRT-PCR analysis displayed distinct expression patterns of target genes, which could lead to biased interpretations of gene functions. Thus, our results suggested that the importance of appropriate RGs, which will enable more accurate and reliable normalizations in gene expression analysis.

In conclusion, we finally demonstrated the fidelity of *GLY1* and *SKP1* as the optimal candidate RGs in qRT-PCR experiments with diverse *A. tuberosum* tissues. This is the first report on the selection of RGs in *A. tuberosum*, which will benefit the study of gene expressions and other related subjects in Chinese chive. Nevertheless, we should be cautious when including the two RGs into qRT-PCR experiments in *A. tuberosum*, because their expression stability was only verified at the developmental aspect and could be affected by other experimental conditions.

Materials and methods

Plant materials.

The Seeds of *A. tuberosum* cultivar '791' were bought from the Jingyan Yinong (Beijing) Seed Sci-Tech Co. Ltd, and all subsequent experiments were conducted according to the guidelines and regulations of the Beijing Vegetable Research Center (BVRC), Beijing Academy of Agriculture and Forestry Sciences. The seeds of Chinese chive were germinated on a 32-cell tray with granulated rockwool, and seedlings were irrigated with ¼ strength of Hoagland's medium once a day. After 2-month growth, seedlings were transferred to a hydroponic cultivation system in a controlled greenhouse located in the BVRC. The plants were supplied with the nutrient solution, and the composition formula as previously described[7]. Tissue samples of 3-year-old seedlings were harvested from the mature leaves, roots, rhizome, mature flowers, inflorescences, and seeds. All tissue samples were fleshly collected from at least 10 individual plants and were frozen immediately in liquid nitrogen until use. Sample collections were performed on separate days for three replicates. In the course of experiments, we complied with the IUCN Policy Statement on Research Involving Species at Risk of Extinction and the Convention on the Trade in Endangered Species of Wild Fauna and Flora. The ethical approval was deemed unnecessary according to the Decree of the State Council of China No. 204.

RNA-Seq and transcriptome de novo assembly.

The 18 RNA-Seq libraries representing the six tissues were prepared were sequenced at the Illumina NovaSeq 6000 platform to an average depth of 50 million reads per sample. Sequence reads were filtered using SeqPrep (https://github.com/jstjohn/SeqPrep) and Sickle (https://github.com/najoshi/sickle) to remove the low-quality and adaptor sequences. Clean reads were assembled via the Trinity de novo assembly program (https://github.com/trinityrnaseq/trinityrnaseq)[47] and TransRate (http://hibberdlab.com/tran-srate/)[48]. For annotation purposes, sequences were handled with CD-HIT (http://weizhongli-lab.org/cd-hit/) program to reduce the transcript redundancy[49], and finally, the assembly quality was evaluated using BUSCO (Benchmarking Universal Single-Copy Orthologs, http://busco.ezlab.org) program with default configurations[50]. All raw data were deposited NCBI Sequence Read Archive under the project with identification number PRJNA67394520, and the assemble transcriptome is accessible at a public server (https://doi.org/10.6084/m9.figshare.14820201).

Screening of candidate reference genes from transcriptome data.

Candidate RGs were screened according to the previously described method with modifications[37]. Briefly, based on the transcriptome data from the six tissues, four parameters were adopted to evaluate the stability of gene expressions: (1) The coefficient of variation (CV), which detects the extent of variability in relation to the mean of the expression levels; (2) The normalized median absolute deviation (NMAD). In statistics, median absolute deviation (MAD) is commonly used for assessing the spread of the distribution according to the medians and is less susceptible to the deviations by outliers. The MAD was further normalized against the median, which could reflect the variability of gene expressions based on the medians. Compared to the CV that is often affected by outliers, and the NMAD provides a robust alternative; (3) Normality test. The p-value measured by the *Shapiro–Wilks* hypothesis indicates whether the expression data does fit the normal distribution, and in the test, p-value < 0.05 means a significant departure from normality. Most importantly, FPKM values were also taken into account. Because genes with low expression abundance are unsuitable as RGs, we used FPKM ≥ 100 as the cut-off value for candidate selection.

The ideal RGs should have lower or similar statistical variations across samples, which was indicated by low CV and NMAD values. Moreover, their expression data also should pass the normality test, which allows that with 95% confidence they fit the normal distribution. To meet the criteria proposed herein, RGs were finally ranked according to the calculation of Euclidean distance (*d*), and the formula was as follows: $d = \sqrt{CV^2 + NMAD^2 + (1-p)^2}$. After comparisons, the top 10 genes with the least values of Euclidean distance were selected for further analysis.

RNA extraction, the cDNA synthesis. Total RNA isolation was performed as described elsewhere[51]. Briefly, total RNA was extracted with Transzol (Transgen Biotech, China) according to the manufacturer's instructions. Briefly, tissues of *A. tuberosum* were homogenized in liquid nitrogen and the frozen powder was lysed using Transzol extraction buffer. 1/5 volume of chloroform was added before 5-min centrifugation at 12,000*g*. The aqueous containing nucleic acid was precipitated with an equal volume of isopropanol by centrifuging at 12,000*g* for 10 min. The resultant pellets were washed in 75% ethanol and resuspended in RNase-free water. After DNase I digestion, the clean total RNAs were stored at − 80 °C freezer. RNA quantity and quality were examined using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, USA) to ensure structural integrity for further experiments.

First-strand cDNA synthesis was conducted by using the EasyScript First-Strand cDNA Synthesis SuperMix kit (Transgen Biotech, China) according to the user's manual. In summary, 1 µg of total RNA was mixed with 1 µl oligo $(dT)_{18}$ primers (10 mM), and RNase-free water. The mixture was incubated for 5 min at 65 °C before being chilled on ice for 3 min. Subsequently, 4 µl of the first-strand buffer, 1 µl of gDNA remover, and 1 µl of reverse transcriptase were added to finalize the reaction. After 1-h incubation at 42 ℃, the reaction was inactivated for 10 s at 70 °C. The cDNA diluted by fivefold with PCR-grade water was ready for use.

Quantitative real-time PCR. The amounts of individual genes were estimated with gene-specific primers by quantitative real-time PCR analysis with a real-time PCR instrument Roche LightCycler 480 and SYBR Green mixture (Toyobo, Japan) as previously described[52]. Briefly, primer pairs showing a single amplified product with the predicted sizes were chosen for further qRT-PCR experiments. Each PCR reaction mix consisted of 10 µl of SYBR Green Supermix, 0.5 µl of forward and reverse primers (10 mM), and 2.5 µl fivefold diluted template cDNA. Finally, the resulting reaction volume was made up to 20 µl by adding 7 µl PCR-grade water. The PCR cycling was performed as follows: 5 min at 94 ℃ followed by 40 cycles of 15 s at 94 ℃, 10 s at 58 ℃, 30 s at 72 ℃, and finally, 1 round of 60 s at 60 ℃, and the melting curve cycling consisted of 15 s at 95 ℃, 1 min at 60 ℃, 30 s at 95 ℃, and 15 s at 60 ℃. For qPCR efficiency, Based on the standard curve generated with serial dilutions of the pooled cDNA template, the amplification efficiency (E) of each primer pair was calculated using the equation: $E\% = (10^{-1/\text{slope}} - 1) \times 100$. For the validation experiment, the relative expression level was calculated by the $2^{-\Delta ct}$ method with selected RGs, where ΔCt is the difference in the threshold cycles between the target and the RG. All qRT-PCR reactions were performed in technical triplicate. The specific primers used in this experiment are shown in Supplementary Table 3.

Expression stability analysis of potential reference genes. The expression stability of the candidate RGs was evaluated by using geNorm (https://genorm.cmgg.be/), NormFinder (https://moma.dk/normfinder-software/), BestKeeper (https://www.gene-quantification.de/bestkeeper.html), and RefFinder (https://heartcure.com.au/) statistical algorithms[23–26,53]. The qBase+ software with geNorm analysis was employed[26], while NormFinder and BestKeeper analyses were performed with free add-ins for Microsoft excel[23,24]. RefFinder integrated the above three algorithms and the Delta Ct method is a free, web-based service for comprehensive ranking of RGs[25]. Based on the sequential pair-wise comparisons, geNorm calculates the stability value M and the variation value V, by which the stability of each gene and the number of optimal RGs is determined. NormFinder ranks individual candidates according to their stability value. BestKeeper determines the optimal references by employing pair-wise correlation analysis of all pairs of candidate genes. Using the LightCycler 480 Relative Quantification Software module (Roche, USA), quantification cycle (Cq) values were obtained and converted into relative quantities by using standard curves, then applied for evaluation. The relative expression levels were imported into geNorm and NormFinder, whereas BestKeeper and RefFinder analyses were conducted with the raw non-transformed Cq values.

Received: 17 February 2021; Accepted: 22 July 2021
Published online: 16 August 2021

References

1. Park, K. W. *et al.* Cytotoxic and antitumor activities of thiosulfinates from *Allium tuberosum* L. *J. Agric. Food Chem.* **55**, 7957–7961. https://doi.org/10.1021/jf0713051 (2007).
2. Saito, S. In *Onions and Allied Crops Volume III: Biochemistry, Food Science, and Minor Crops* (eds Brewster, J. L. & Rabinowitch H. D.) Chapter 11, (CRC Press, 1990).

3. Tang, X., Olatunji, O. J., Zhou, Y. & Hou, X. *Allium tuberosum*: Antidiabetic and hepatoprotective activities. *Food Res. Int.* **102**, 681–689. https://doi.org/10.1016/j.foodres.2017.08.034 (2017).
4. Karlgren, B. *The Book of Odes* (Chinese text, Transcription and Translation). 1st edn, 96–101, https://archive.org/details/Bulletin22 (Museum of Far Eastern Antiquities, 1950).
5. Fang, Z. Y. (ed.) *Vegetable Breeding in China* (Chinese Agriculture Press, 2017).
6. Zhou, S. M., Chen, L. M., Liu, S. Q., Wang, X. F. & De Sun, X. D. novo assembly and annotation of the Chinese chive (*Allium tuberosum* Rottler ex Spr.) transcriptome using the Illumina platform. *PLoS ONE* **10**, e0133312. https://doi.org/10.1371/journal.pone.0133312 (2015).
7. Liu, N. *et al.* Transcriptome landscapes of multiple tissues highlight the genes involved in the flavor metabolic pathway in Chinese chive (*Allium tuberosum*). *Genomics* **113**, 2145–2157. https://doi.org/10.1016/j.ygeno.2021.05.005 (2021).
8. Manabe, T., Hasumi, A., Sugiyama, M., Yamazaki, M. & Saito, K. Alliinase [S-alk(en)yl-L-cysteine sulfoxide lyase] from *Allium tuberosum* (Chinese chive)—purification, localization, cDNA cloning and heterologous functional expression. *Eur. J. Biochem.* **257**, 21–30. https://doi.org/10.1046/j.1432-1327.1998.2570021.x (1998).
9. Urano, Y., Manabe, T., Noji, M. & Saito, K. Molecular cloning and functional characterization of cDNAs encoding cysteine synthase and serine acetyltransferase that may be responsible for high cellular cysteine content in *Allium tuberosum*. *Gene* **257**, 269–277. https://doi.org/10.1016/s0378-1119(00)00399-1 (2000).
10. Lovdal, T. & Lillo, C. Reference gene selection for quantitative real-time PCR normalization in tomato subjected to nitrogen, cold, and light stress. *Anal. Biochem.* **387**, 238–242. https://doi.org/10.1016/j.ab.2009.01.024 (2009).
11. Pabuayon, I. M. *et al.* Reference genes for accurate gene expression analyses across different tissues, developmental stages and genotypes in rice for drought tolerance. *Rice* **9**, 32. https://doi.org/10.1186/s12284-016-0104-7 (2016).
12. Wang, C. *et al.* Identification and validation of reference genes for RT-qPCR analysis in non-heading Chinese cabbage flowers. *Front Plant. Sci.* **7**, 811. https://doi.org/10.3389/fpls.2016.00811 (2016).
13. Heid, C. A., Stevens, J., Livak, K. J. & Williams, P. M. Real-time quantitative PCR. *Genome Res.* **6**, 986–994. https://doi.org/10.1101/gr.6.10.986 (1996).
14. Navarro, E., Serrano-Heras, G., Castaño, M. J. & Solera, J. Real-time PCR detection chemistry. *Clin. Chim. Acta* **439**, 231–250. https://doi.org/10.1016/j.cca.2014.10.017 (2015).
15. Wong, M. L. & Medrano, J. F. Real-time PCR for mRNA quantitation. *Biotechniques* **39**, 75–85. https://doi.org/10.2144/05391RV01 (2005).
16. Bustin, S. A. Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): Trends and problems. *J. Mol. Endocrinol.* **29**, 23–39. https://doi.org/10.1677/jme.0.0290023 (2002).
17. Bustin, S. A., Benes, V., Nolan, T. & Pfaffl, M. W. Quantitative real-time RT-PCR—a perspective. *J. Mol. Endocrinol.* **34**, 597–601. https://doi.org/10.1677/jme.1.01755 (2005).
18. Liu, N. & Avramova, Z. Molecular mechanism of the priming by jasmonic acid of specific dehydration stress response genes in Arabidopsis. *Epigenetics Chromatin* **9**, 1–23. https://doi.org/10.1186/s13072-016-0057-5 (2016).
19. Bustin, S. A. Absolute quantification of mRNA using real-time reverse transcription-polymerase chain reaction assays. *J. Mol. Endocrinol.* **25**, 169–193. https://doi.org/10.1677/jme.0.0250169 (2000).
20. Doak, S. H. & Zaïr, Z. M. Real-time reverse-transcription polymerase chain reaction: Technical considerations for gene expression analysis. *Methods Mol. Biol.* **817**, 251–270. https://doi.org/10.1007/978-1-61779-421-6_13 (2012).
21. Liu, N., Staswick, P. E. & Avramova, Z. Memory responses of jasmonic acid-associated Arabidopsis genes to a repeated dehydration stress. *Plant Cell Environ.* **39**, 2515–2529. https://doi.org/10.1111/pce.12806 (2016).
22. Schröder, F., Lisso, J. & Müssig, C. Expression pattern and putative function of *EXL1* and homologous genes in Arabidopsis. *Plant Signal Behav.* **7**, 22–27. https://doi.org/10.4161/psb.7.1.18369 (2012).
23. Andersen, C. L., Jensen, J. L. & Ørntoft, T. F. Normalization of real-time quantitative reverse transcription-PCR data: A model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res.* **64**, 5245–5250. https://doi.org/10.1158/0008-5472.Can-04-0496 (2004).
24. Pfaffl, M. W., Tichopad, A., Prgomet, C. & Neuvians, T. P. 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. https://doi.org/10.1023/b:bile.0000019559.84305.47 (2004).
25. Silver, N., Best, S., Jiang, J. & Thein, S. L. Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC Mol. Biol.* **7**, 33. https://doi.org/10.1186/1471-2199-7-33 (2006).
26. Vandesompele, J. *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, research0034.0031. https://doi.org/10.1186/gb-2002-3-7-research0034 (2002).
27. Kamenetsky, R. & Rabinowitch, H. D. The genus Allium: A developmental and horticultural analysis. *Hortic. Rev.* **32**, 329–378. https://doi.org/10.1002/9780470767986.ch7 (2006).
28. Rabinowitch, H. D. & Currah, L. (eds) *Allium Crop Science: Recent Advances* (CABI Publishing, 2002).
29. Shigyo, M., Khar, A. & Abdelrahman, M. *The Allium Genomes* (Springer International Publishing, Basel, 2018).
30. Liu, Q. *et al.* RNA-Seq reveals leaf cuticular wax-related genes in welsh onion. *PLoS ONE* **9**, e113290. https://doi.org/10.1371/journal.pone.0113290 (2014).
31. Liu, H. *et al.* Histological, physiological and transcriptomic analysis reveal gibberellin-induced axillary meristem formation in garlic (*Allium sativum*). *Plants* **9**, 970 (2020).
32. Jin, S., Xu, R., Wei, Y. & Goodwin, P. H. Increased expression of a plant actin gene during a biotrophic interaction between round-leaved mallow, *Malva pusilla*, and *Colletotrichum gloeosporioides* f. sp. malvae. *Planta* **209**, 487–494. https://doi.org/10.1007/s004250050752 (1999).
33. Yang, C. L. *et al.* Comprehensive transcriptome analysis of reference genes for fruit development of *Euscaphis konishii*. *PeerJ* **8**, e8474. https://doi.org/10.7717/peerj.8474 (2020).
34. Yi, S. *et al.* Selection and validation of appropriate reference genes for quantitative RT-PCR analysis in *Rubia yunnanensis* diels based on transcriptome data. *Biomed. Res. Int.* **2020**, 5824841. https://doi.org/10.1155/2020/5824841 (2020).
35. Liang, L. *et al.* Selection and validation of reference genes for gene expression studies in *Codonopsis pilosula* based on transcriptome sequence data. *Sci. Rep.* **10**, 1362. https://doi.org/10.1038/s41598-020-58328-5 (2020).
36. Kim, S. *et al.* Integrative structural annotation of de novo RNA-Seq provides an accurate reference gene set of the enormous genome of the onion (*Allium cepa* L.). *DNA Res.* **22**, 19–27. https://doi.org/10.1093/dnares/dsu035 (2015).
37. Smitha, P. K. *et al.* Genome wide search to identify reference genes candidates for gene expression analysis in *Gossypium hirsutum*. *BMC Plant Biol.* **19**, 405. https://doi.org/10.1186/s12870-019-1988-3 (2019).
38. Wang, G. *et al.* Selection of reliable reference genes for quantitative RT-PCR in garlic under salt stress. *PeerJ* **7**, e7319. https://doi.org/10.7717/peerj.7319 (2019).
39. Yim, A.K.-Y. *et al.* Using RNA-Seq data to evaluate reference genes suitable for gene expression studies in soybean. *PLoS ONE* **10**, e0136343. https://doi.org/10.1371/journal.pone.0136343 (2015).
40. Wang, X. *et al.* Mass spectrometric characterization of the affinity-purified human 26S proteasome complex. *Biochemistry* **46**, 3553–3565. https://doi.org/10.1021/bi061994u (2007).
41. Sadanandom, A., Bailey, M., Ewan, R., Lee, J. & Nelis, S. The ubiquitin–proteasome system: Central modifier of plant signalling. *New Phytol.* **196**, 13–28. https://doi.org/10.1111/j.1469-8137.2012.04266.x (2012).

42. Stürzenbaum, S. R. & Kille, P. Control genes in quantitative molecular biological techniques: The variability of invariance. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **130**, 281–289. https://doi.org/10.1016/S1096-4959(01)00440-7 (2001).
43. Nicot, N., Hausman, J. F., Hoffmann, L. & Evers, D. Housekeeping gene selection for real-time RT-PCR normalization in potato during biotic and abiotic stress. *J. Exp. Bot.* **56**, 2907–2914. https://doi.org/10.1093/jxb/eri285 (2005).
44. Liu, N., Ding, Y., Fromm, M. & Avramova, Z. Different gene-specific mechanisms determine the "revised-response" memory transcription patterns of a subset of *A. thaliana* dehydration stress responding genes. *Nucleic Acids Res.* **42**, 5556–5566. https://doi.org/10.1093/nar/gku220 (2014).
45. Schulman, B. A. *et al.* Insights into SCF ubiquitin ligases from the structure of the Skp1-Skp2 complex. *Nature* **408**, 381–386. https://doi.org/10.1038/35042620 (2000).
46. Lildballe, D. L. *et al.* The expression level of the chromatin-associated HMGB1 protein influences growth, stress tolerance, and transcriptome in Arabidopsis. *J. Mol. Biol.* **384**, 9–21. https://doi.org/10.1016/j.jmb.2008.09.014 (2008).
47. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652. https://doi.org/10.1038/nbt.1883 (2011).
48. Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M. & Kelly, S. TransRate: Reference-free quality assessment of de novo transcriptome assemblies. *Genome Res.* **26**, 1134–1144. https://doi.org/10.1101/gr.196469.115 (2016).
49. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565 (2012).
50. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282–283. https://doi.org/10.1093/bioinformatics/17.3.282 (2001).
51. Xu, Y. *et al.* Genome-wide identification, expression profile and evolution analysis of karyopherin β gene family in *Solanum tuberosum* group phureja DM1–3 reveals its roles in abiotic stresses. *Int. J. Mol. Sci.* **21**, 931. https://doi.org/10.3390/ijms21030931 (2020).
52. Liu, N., Fromm, M. & Avramova, Z. H3K27me3 and H3K4me3 chromatin environment at super-induced dehydration stress memory genes of *Arabidopsis thaliana*. *Mol. Plant* **7**, 502–513. https://doi.org/10.1093/mp/ssu001 (2014).
53. Xie, F., Xiao, P., Chen, D., Xu, L. & Zhang, B. miRDeepFinder: A miRNA analysis tool for deep sequencing of plant small RNAs. *Plant Mol. Biol.* https://doi.org/10.1007/s11103-012-9885-2 (2012).

Acknowledgements

We thank Prof. Changlong Wen (Beijing Vegetable Research Center, Beijing Academy of Agricultural and Forestry Sciences) for kindly providing the technical assistance for the qRT-PCR experiments, and also thanks to Prof. Wei Liu (Beijing Vegetable Research Center, Beijing Academy of Agricultural and Forestry Sciences) for fruitful discussions on this manuscript.

Author contributions

N.L. conceived and designed the study; M.H. and N.L. conducted the in vitro experiments; J.T., B.H., Y.J., B.W., H.L., and N.L. were involved in bioinformatics analysis, including transcriptome sequencing, assembly, annotations, and gene stability evaluation; N.L. and M.H. analyzed the data; N.L. supervised the whole research work; Z.W. and M.L. were responsible for project administration; N.L. drafted the manuscript with inputs from T.J. and H.M. All authors read and approved the final version of the manuscript.

Funding

The work was supported by the National Engineering Research Center for Vegetables "Science Innovation Program (KYCX202001-09)" and the Beijing Agricultural and Forestry Sciences "Young Talent Award (YCXTD00002-10)" to NL, and by the Special Innovation Ability Construction Fund of Beijing Academy of Agriculture and Forestry Sciences and the China Agriculture Research System of MOF and MARA to ZW.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-95849-z.

Correspondence and requests for materials should be addressed to Z.W. or N.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.