Partida *et al. Plant Methods* (2025) 21:32
https://doi.org/10.1186/s13007-025-01329-0

Plant Methods

RESEARCH

Open Access

The genetic puzzle of multicopy genes: challenges and troubleshooting

Vania Gabriela Sedano Partida[1†], Henrique Moura Dias[1†], Maria Teresa Portes[1] and Marie-Anne Van Sluys[1*]

Abstract

Background Studies of multicopy genes impose challenges related to gene redundancy and sequence similarity among copies. However, recent advances in molecular biology and genomics tools associated with dedicated databases facilitate their study. Thus, the present work emphasizes the need for rigorous methodologies and standardized approaches to interpret RT-qPCR results accurately.

Results Data from *Physcomitrium patens* provides a comprehensive five-step protocol, using thiamine thiazole synthase (*THI1*) and sucrose 6-phosphate phosphohydrolase (*S6PP)* genes as proof of concept to showcase a systematic workflow for studying multicopy genes. Beyond examining genes of interest, we highlight the critical role of choosing appropriate internal controls in the analytical process for interpreting gene expression patterns accurately. We emphasize the importance of identifying the relevant orthologous gene, recognizing the inherent challenges in determining the most functional copy for subsequent studies. Our objective is to enhance comprehension of gene redundancy by dissecting multicopy genes' genomic landscape and its characteristics. Furthermore, we address the decision-making process surrounding the expression level quantification of multicopy genes.

Conclusions The study of multicopy genes discloses early events for functional adaptation. Here, we discuss the significance of multicopy genes in plant biology and provide an experimental protocol to analyze them. As plant systems are strongly influenced by light/dark cycles, challenges inherent to circadian processes are also acknowledged. Therefore, our comprehensive approach aims to advance the understanding of multicopy gene dynamics, offering practical methodologies and contributing with valuable insights to the scientific community.

Keywords Gene families, Gene expression, Paralogs, Primer design, Transcriptomes, Genomes, Life cycle

Background

Multicopy genes, or gene families, represent a fundamental aspect of plant genomes, contributing to the genomic complexity underlying species' adaptability and resilience. These genes often arise through various duplication events, including whole-genome duplications, tandem duplications and transposable elements activity, also playing pivotal roles in shaping the evolutionary trajectory of plant lineages [1, 2]. In plant genomes, repetitive sequences, including multicopy genes, are particularly abundant, constituting a significant portion of the genetic material [3]. Furthermore, multicopy genes frequently exhibit functional redundancy, which contributes to plant robustness to cope with environmental challenges. The evolutionary forces acting on duplicated genes may also lead to functional diversity, resulting in sub-functionalization or neofunctionalization, further enhancing the adaptability of plant

[†]Vania Gabriela Sedano Partida and Henrique Moura Dias have contributed equally to this work.

*Correspondence:
Marie-Anne Van Sluys
mavsluys@usp.br
[1] Departamento de Botânica, Instituto de Biociências, Universidade de São Paulo, São Paulo, SP 05508-090, Brazil

species [4]. Additionally, multicopy genes play pivotal roles in plant biology, influencing development, stress responses, and adaptation to changing environments [5]. These genes are essential for shaping plant morphology and regulating growth in response to external stimuli [6]. Examples of multicopy gene families include those involved in defense mechanisms against pathogens, response to abiotic stresses and biosynthesis of secondary metabolites [7, 8]. Thus, the existence of multiple copies allows plants to fine-tune their responses to a dynamic and challenging environment.

The retention of duplicated genes is described to promote neofunctionalization or sub-functionalization, which involves the acquisition of a new function or partitioning biochemical functions at the expression pattern level. Thus, without adverse environmental changes, the model theorizes similar expression levels among the paralogs. On the other hand, the absolute dosage and dosage-balance model also explains the retention of duplicated genes without a change in function [4, 9].

Despite their significance, studying multicopy genes remain challenging due to inherent complexities such as gene redundancy, where multiple copies perform similar functions, demanding technical efforts to decipher the specific contribution of individual copies (reviewed in [10]). Additionally, an elevated sequence similarity among copies impair an accurate characterization and analysis [11]. Distinguishing between individual gene copies and understanding their distinct functions becomes highly complex, hindering a comprehensive understanding of their biological roles. Recent advances in molecular biology and genomics tools, such as next-generation sequencing technologies allowing high-throughput sequencing have significantly enhanced our ability to study multicopy genes, enabling the characterization of entire gene families [12, 13]. Such technical tools provide means to explore the complexity of multicopy gene families in details. In this context, databases dedicated to multicopy gene identification facilitate the discovery and retrieval of relevant information [14].

The study of multicopy genes requires rigorous protocols to ensure accuracy and reproducibility. Methodological choices, including experimental design and data analysis, significantly influence results interpretation, showing the importance of standardized approaches. These protocols address challenges related to sequence similarity and gene redundancy, since established methodologies ensure reliable findings and advance our understanding of multicopy gene contributions to plant biology. Previous studies on multicopy genes in plants, such as Das and Bansal [15], has provided valuable insights in this context. However, significant pitfalls persist particularly in deciphering the functional nuances of individual gene copies and their contribution to specific biological pathways and for plant adaptation [16, 17]. Further research in this area is essential for unlocking the full potential of investigating multicopy genes for biological applications such as plant breeding, biotechnology, and crop improvement. Understanding the functional implications of multicopy genes variability could revolutionize the strategies for enhancing crop resilience, productivity, and adaptability to a changing environment. Addressing these knowledge gaps holds promise for a practical application of multicopy gene research in shaping the future of agriculture and plant biology.

Here, we highlight the importance of multicopy genes in plant genomes. For instance, in model organisms, approximately 65% of the *Arabidopsis thaliana* genome is represented by multigene families (with more than three genes), 75% of the *Oryza sativa* genome, and about 70% of the *Physcomitrium patens* genome also consists of multigene families [18]. Recognizing its inherent complexity, we developed a streamlined and curated protocol providing a practical framework for researchers to explore and navigate challenges such as gene redundancy and sequence similarity. Taking advantage of recent advances in molecular biology and genomics, coupled with dedicated databases and tools, our protocol aims to unravel the intricate expression pattern of individual paralogs amid multicopy genes. By emphasizing the importance of rigorous methodologies and standardized approaches through a case study of *S6PP* and *THI1* genes [16, 19, 20] in the model plant *Physcomitrium patens*, the proposed protocol would contribute to the reliability and reproducibility of experimental findings in this field.

Case study in *Physcomitrium patens*

Higher plants have a biological complexity that makes their study challenging at the genetic and metabolic level [21]. The need of understanding biological and genetic functions has led researchers to seek for model plants [22–24]. Thus, the study of bryophytes emerges to contribute to understand how land plants evolved and adapted to live on soil. Mosses are structurally less complex while sharing fundamental metabolic and physiological processes with flowering plants. Hence, relevant information on biological and genetic functions are common across many groups of Embryophytes [25].

Physcomitrium patens, formerly *Physcomitrella patens*, is a model organism due to its simple body-plant organization, large cells in both protonema and gametophore, as well as its small haploid genome 470 Mbp organized into 27 chromosomes [26], and a life cycle of approximately 12 weeks [27] under controlled conditions. The adoption of model organisms ensures their extensive study and facilitate the investigation of biological phenomena,

but this is not always the case [24]. *P. patens* imposes its challenges, mostly due to genetic redundancy as a result of ancestral events of whole-genome duplication (WGD) between 30 and 60 million years ago [28], resulting in the retention of a large number of duplicated genes, mainly involved in metabolism [28].

Physcomitrium patens follows a typical moss life cycle and many other non-vascular plants, referred to as "alternation of generations," and involves two distinct phases: the generation of multicellular haploid gametophytes alternating with a generation of morphologically distinct diploid sporophytes). The mature stage of the gametophyte, known as the gametophore, exhibits a higher structural complexity featuring phyllodes (resembling leaves), stems, and rhizoids (root-like structures) (see Fig. 1A). The progression from the juvenile to adult gametophyte is instigated by the differentiation of initial cells within the protonema filament, leading to the development of buds. The gametophore represents the adult haploid phase, supporting sexual reproduction. This process involves initiating a structure, the sporophyte, at the apex of the gametophore, marking the transition to the diploid phase (2n), localized within the sporophyte. When mature, the sporophyte releases spores (see Fig. 1A), which, upon germination, give rise to new protonemata, thereby initiating a new cycle of the haploid phase in the plant's life cycle.

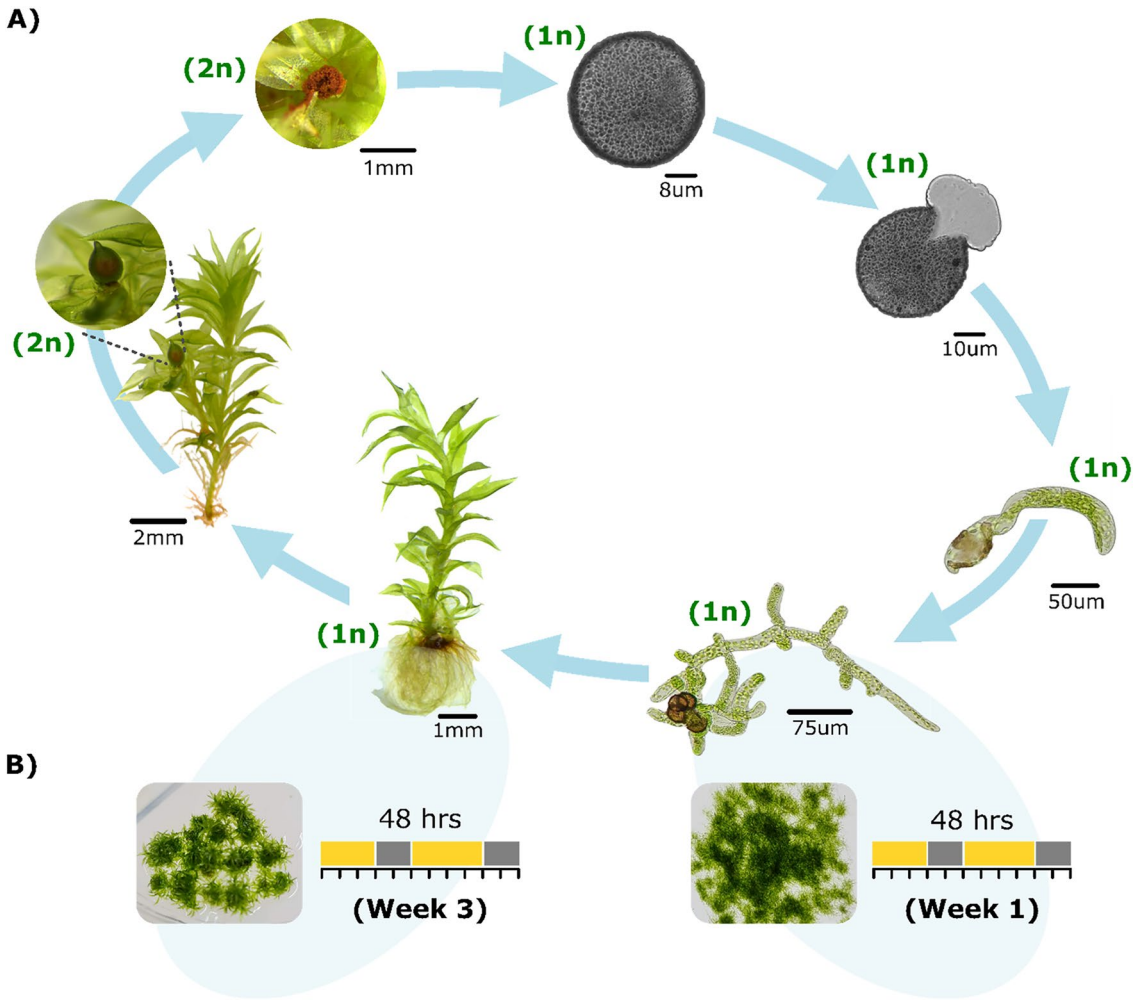

Fig. 1 *Physcomitrium patens* life cycle. The life cycle initiates with the germination of a spore in its haploid phase (1n). A filamentous structure emerges, producing protonemata (1n) composed of caulonema and chloronema cells. This juvenile phase is completed within a week of spore germination. Subsequently, more intricate structures develop, forming the gametophore (1n), comprising phyllodes, stems, and rhizoids. This stage is achieved in three weeks. The adult gametophore enters its reproductive phase, giving rise to a structure known as a sporophyte at the apex, marking the onset of the diploid phase (2n). Once mature, the sporophyte ruptures, releasing spores that will subsequently germinate, initiating a new cycle

Thus, the adoption of *P. patens* as a study case is proposed here to demonstrate the workflow rationale for multicopy gene expression studies of eight thiamine thiazole synthase (*THI1)* [16, 19] and five sucrose 6-phosphate phosphohydrolase (*S6PP)* [20] paralogs as a proof of concept. It highlights the significance and implications of identifying appropriate orthologous genes among the multicopy gene family, which is considered a critical decision. Challenges with multicopy genes may include determining which copy is functionally significant and understanding their physiological role. The experimental design focused on two distinct *P. patens* developmental stages: protonema and gametophore combined in the light/dark cycle over 48 h, 24 samples were collected in triplicates (Fig. 1B). In the first stage, 1-week-old protonema cells are present, and in the second stage, 3-week-old adult gametophore phyllodes, stems, and rhizoids predominate.

Standardized approach to workflow definition

As mentioned above, a streamlined and curated protocol enables the understanding of intricate expression patterns of individual paralogs amid multicopy genes with reliability and reproducibility (Fig. 2). Combining bioinformatics tools and experimental techniques ensure a thorough understanding of the genomic context and expression of multicopy gene families in diverse biological systems. Researchers can adapt and modify this protocol to suit specific organisms and research objectives, facilitating in-depth investigations into the role of multicopy genes in biological processes.

STEP 1: Understanding the features of multicopy genes

Background: Identifying multigene family members in a genome of interest.

Determining the significance of individual gene copies imposes a challenge, but our systematic investigation contributes substantially to unraveling this complexity. This step is crucial to enhance comprehension of gene redundancy by dissecting the genomic landscape and characteristics of multicopy genes, proposing a protocol for RT-qPCR validation assays using *S6PP* and *THI1* as target gene examples.

In the protocol's initial phase, computational analyses play a central role in the meticulous identification of candidate genes. The Basic Local Alignment Search Tool (BLAST) [29] was employed for a hypothesis-driven query, aligning it with distinctive features of multicopy genes under scrutiny. We chose the BLASTp algorithm from the Phytozome database, which maintains the *P. patens* v3.3 genome and relevant metadata annotation for putative subjects. Parameters were set rigorously to balance sensitivity and specificity based on E-value (> 1e−5), identity (> 70%), and coverage (> 70%).

An integral aspect of our strategy involved iterative refinement, wherein filtering strategies are applied to eliminate false positives and negatives, encompassing domain prediction (optimized by Yang et al. [30]). Domains representing conserved functional or structural units provide crucial insights into gene copy differentiation, evolutionary dynamics, and potential functional roles. The inclusion of domain annotation validates and improves the accuracy of homologs identification

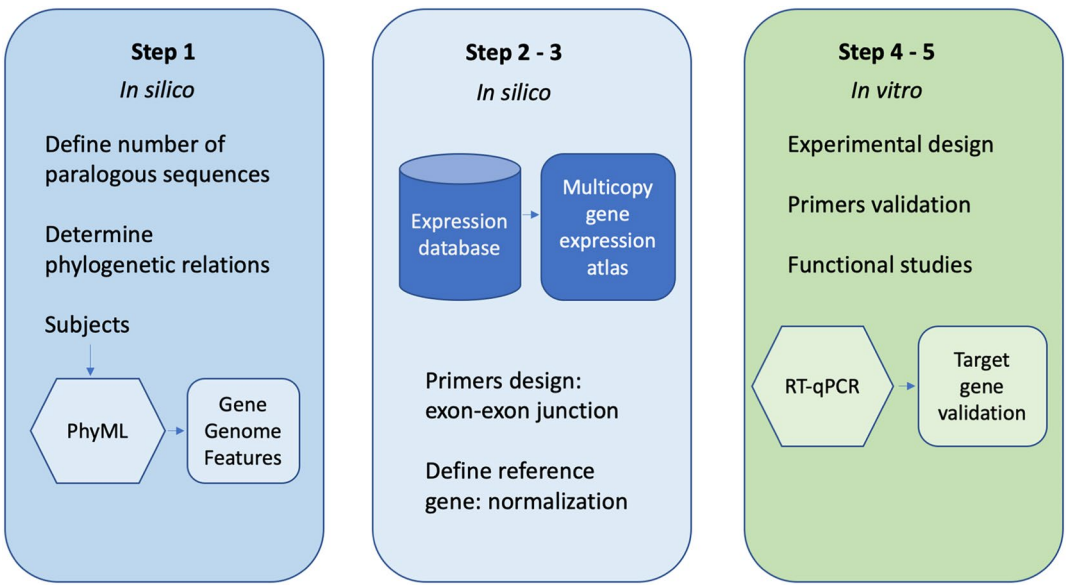

Fig. 2 Workflow to multi-copy gene studies and definition

(Figure S1), introducing a new layer of information (see references [31, 32]). Through this process, we achieved the complete identification and validation of S6PP (IPR006380) and S6PP-C domains (IPR013679) in S6PP orthologs, as well as the THI4 domain (IPR002922) in THI1 (Fig. 3). These parameters are generally sufficient for determining homologs and were applied to THI1 and S6PP. Following this parameter setup, we confirmed six THI1 homologs copies (Pp3c20_13540, Pp3c20_13770, Pp3c23_6510, Pp3c23_6600, Pp3c23_6580, Pp3c24_10800), located on chromosomes 20, 23, and 24, including tandem duplications of THI1 copies on chromosomes 20 and 23. Two excluded copies do not conform to features (Pp3c8_11240 and Pp3c22_8930) with coverages below 70% (Fig. 3A). While this approach generated reliable results, it is essential to highlight that its efficacy depends on the most conserved domains of the sequences [33, 34]. In searching for S6PP homologs in *P. patens*, adjusted criteria were employed, considering an identity below 52% but with coverage exceeding 90% (Fig. 3B). The S6PP homologs Pp3c10_9450, Pp3c14_5810, Pp3c22_1840, and Pp3c24_1340 contain S6PP-like and S6PP-C domains, supporting the presence on chromosomes 10, 14, 22, and 24. Pp3c19_6350 presents two incomplete domains; presumably, this copy has lost its phosphohydrolytic function.

STEP 2: Expression profile by RNAseq brings first clues about multicopy genes

Background: Using available RNAseq data to identify within the multicopy gene family which paralogs are expressed (also relevant for primer design).

After identifying potential homologs in Step 1, evaluating the coverage and expression profiles becomes essential to distinguish authentic and biologically relevant copies from pseudogenes. Authentic genes typically display distinct expression patterns that hold biological significance. Thus, examining the expression profile allows the evaluation of the observed expression to align with known biological functions. A consistent and contextually relevant expression profile supports the validity of a gene, while erratic or inconsistent patterns may indicate artifacts. To gain a preliminary view of paralogous genes transcription, we explored the expression profile, which refers to the gene expression pattern across different conditions or samples using the available PEATmoss database [35]. Evaluating a gene's behavior under various circumstances provides insights into its temporal and tissue-specific expression conditions (Fig. 4).

Each of these datasets likely involves different experimental designs and conditions. This approach is beneficial once allows the determination of expression patterns of target genes in various contexts, potentially uncovering biological scenarios. Tissues or cell types may have specific gene expression patterns, and exploring these patterns can shed light on the spatial expression pattern of paralog genes. They often have similar functions but can present distinct expression patterns, being relevant the expression analysis of paralogs that can provide insights into their functional divergence or redundancy. The initial analysis of gene expression using RNA-seq can serve as a foundation for designing future experiments, since it provides a preliminary understanding of the gene's behavior, essential to guide more specific and hypothesis-driven studies.

While the expression profile provides valuable biological insights into a gene´s temporal and spatial profile, technically, distinguishing genuine expression from sequencing errors or artifacts becomes challenging when a gene exhibits low read coverage. Sufficient coverage is imperative to mitigate the probability of false positives or negatives [36]. For instance, higher coverage enhances the confidence that the observed expression faithfully reflects the gene's actual activity (Figure S2).

Our study revealed the importance of using RNAseq datasets to explore the expression patterns of paralogous genes (*S6PP* and *THI1*) under varying experimental conditions and tissue types. The results showed that some paralogs are not expressed, revealing that not all copies are good candidates for studying their biological function. In addition, this approach can provide a comprehensive view of how these genes are regulated, supporting meaningful biological insights from the data acquired and novel experimental designs. These findings highlight the dynamic nature of *THI1* and *S6PP* multicopy gene expression in response to different developmental

(See figure on next page.)

Fig. 3 Integrated analysis of gene structure, phylogeny, BLAST parameters, and domain protein prediction. The panel displays a schematic phylogenetic tree, demonstrating the evolutionary relationships among paralog genes. Additionally, it provides a visual representation of gene structures, highlighting exons (blue boxes), introns (connecting lines), and untranslated regions (UTR) (green boxes). This depiction allows for a quick assessment of the genomic organization of the gene of interest. In the right panel, a schematic outlines the parameters used in the BLAST (Basic Local Alignment Search Tool) analysis. Parameters such as sequence identity cutoffs, e-values, and coverage significantly impact the outcomes of homology searches, influencing the identification of related gene copies. Also included was the domain protein prediction, which illustrates identified protein domains

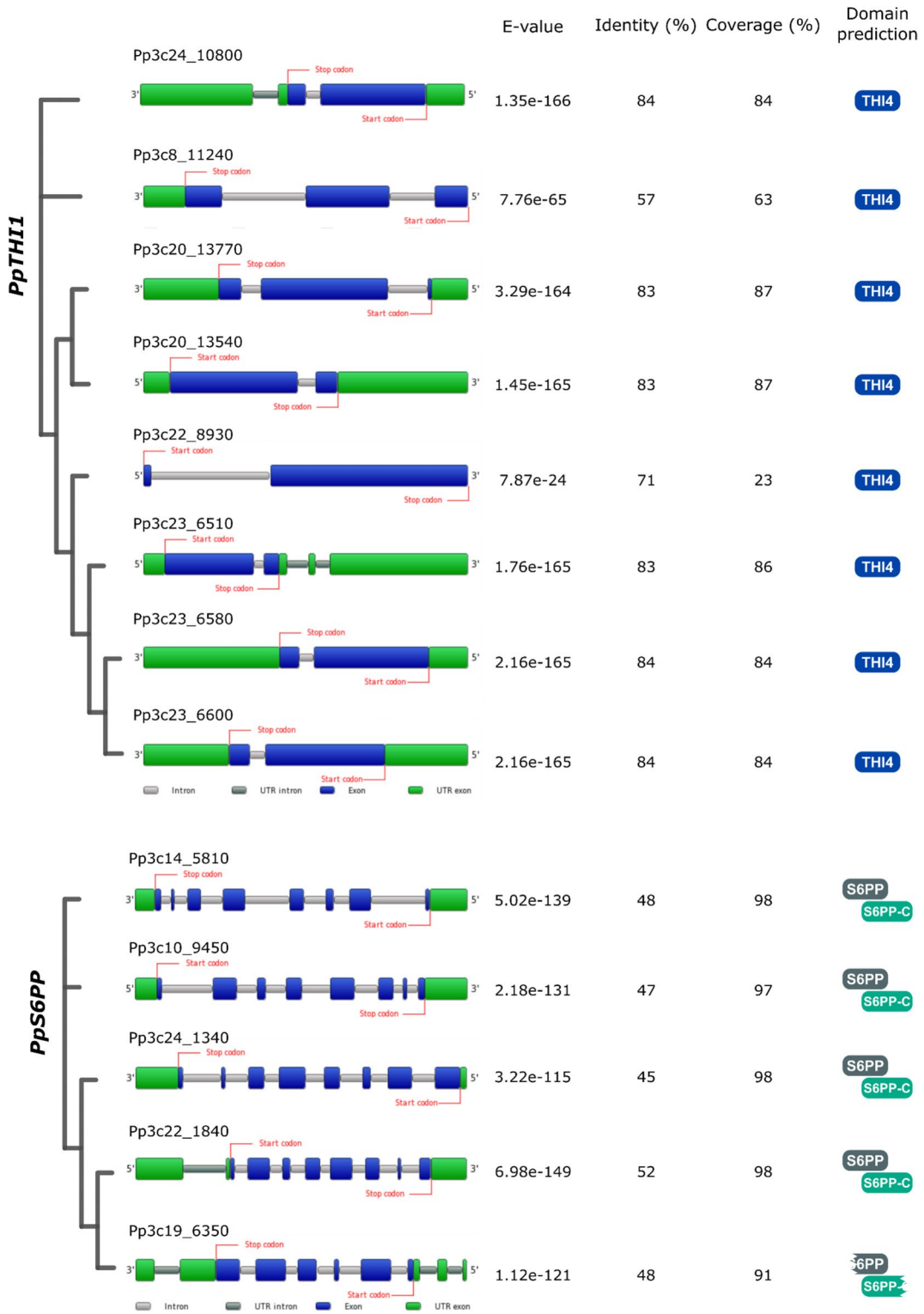

Fig. 3 (See legend on previous page.)

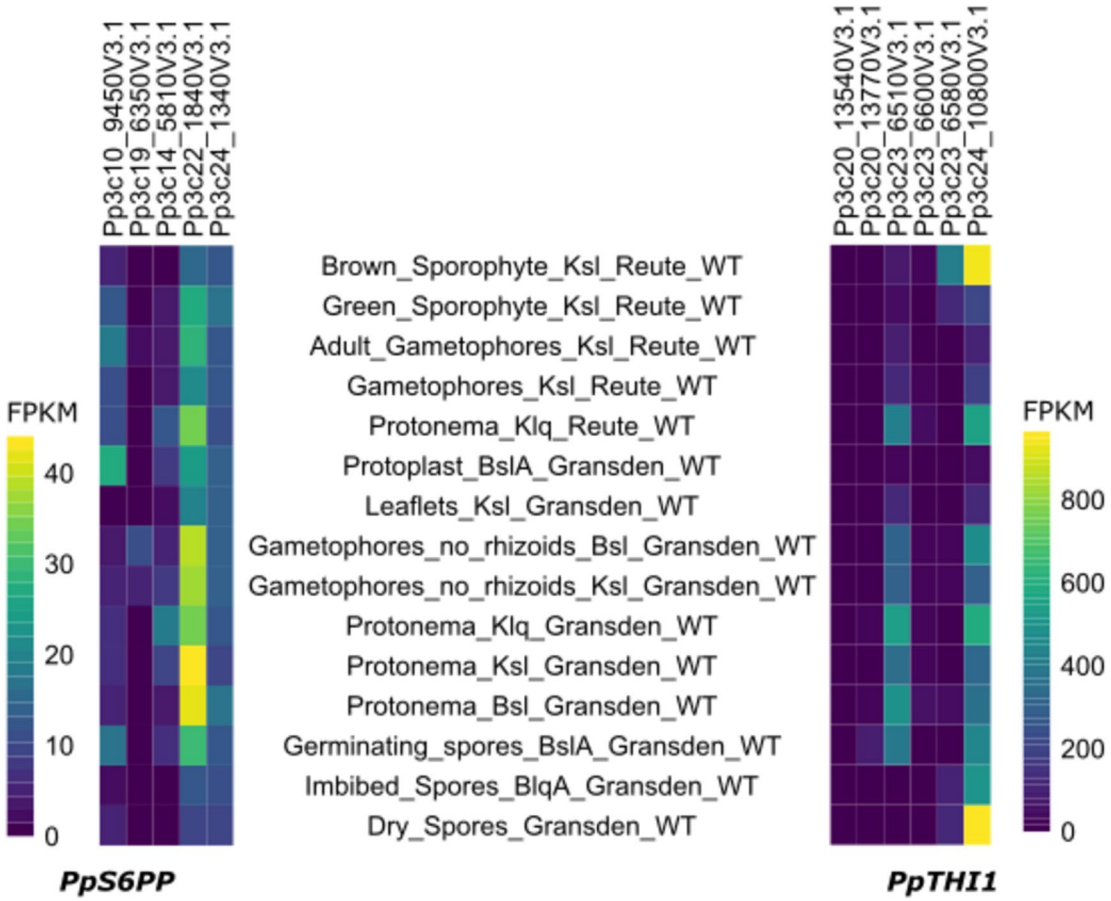

Fig. 4 *PpS6PP* and *PpTHI1* Expression Atlas from PEATmoss database. Experiment treatments are displayed for two *P. patens* lineage (Gransden and Reute), including stages/tissues and media conditions. The expression values are presented by FPKM (Fragments Per Kilobase of transcript per Million mapped reads)

stages and environmental conditions. RNA-seq reveals that each paralog appears to have a specific expression pattern, with some being more prevalent in certain experimental conditions or tissues. This information can be valuable for understanding the functional diversity of such gene copies and their relevance in different biological processes. Also, it is key to defining the primer design, as discussed below (STEP 3), as a fundamental step for the RT-qPCR experiments that provides an overview of the selected tissues and stages for analysis.

STEP 3: Primer design and reference definition

Background: Uncovering the reference (housekeeping) gene for the experimental design chosen.

RT-qPCR is a widely used technique for measuring mRNA expression levels due to its sensitivity and specificity. However, the success and accuracy of this method depend on various factors, including the sample quality, primer specificity, sample reaction efficiency, and data analysis methods [37]. Normalization is a fundamental step in RT-qPCR, and it is one of the most critical challenges. It involves selecting a housekeeping gene (also known as reference gene) that plays a critical role for normalizing gene expression data ensuring accurate and reliable results.

To ensure reliable normalization, it is essential to choose a reference gene that exhibits stable expression across the conditions tested for proper data normalization [38, 39]. Researchers often underestimate the importance of rigorous reference gene selection, and pre-existing reference genes from the literature are sometimes used without validation in specific experimental conditions. Consequently, addressing the expression stability of a collection of reference genes to a given experimental design before applying RT-qPCR techniques for target gene analysis is fundamental as suggested in other studies to avoid erroneous interpretation [40–43].

Based on the above premise, we analyzed four commonly used housekeeping genes: PpE2, PpEf1α, PpST-P2α, and PpVH + PP [44], to define which one would be

more adequate. The expression level of each gene was analyzed at least at six-time points throughout the 24-h cycle (light and dark) in the protonema development phase. The statistical test of the 25th and 75th percentiles of the CT values for each gene indicated that the PpE2 gene had the most stable expression among the four candidates evaluated (Fig. 5). Therefore, PpE2 is the most suitable housekeeping gene for this experimental design. The results support that there is no universal reference gene for a given organism under a given experimental design.

STEP 4: Crafting experimental design

Background: Considering the organism's biology and gene primary function to define the experimental design.

Different developmental stages, environmental conditions, and tissue types can significantly impact the expression of the target gene, yielding different results [45]. The absence of expression doesn't necessarily mean that the gene is non-functional. Genes can be silent under specific conditions or time points, even if they play crucial roles in other conditions. This implies that researchers must carefully consider these factors when designing their experiments and hypotheses [38]. Under this scenario and according to previous results [16, 20], an experiment was set to investigate the transcriptional profile of *S6PP* and *THI1* paralogs in two synchronized developmental stages of *P. patens* (protonemata and gametophyte) and during the light/dark cycle. Plant material was sampled for 48 h (every 4 h) along the light/dark cycle on weeks 1 and 3 (Fig. 1B). For this study, all copies of *THI1* and copies *Pp3c10_9450, Pp3c14_5810, Pp3c22_1840* and *Pp3c24_1340* of *S6PP* were selected based on the expression results from STEP 2—RNAseq analysis. Total RNA was extracted, and cDNA was prepared as described in the methods section.

Among the most common approaches to study gene expression patterns, the quantification of expression by the qPCR method is the most used [41], which often uses the selection of primers in conserved regions. However, in the case of multicopy genes with no previous knowledge of evolving functions, this strategy can lead to potentially incongruent conclusions. Thus, an essential step is the selection of specific primers that can distinguish distinct copies.

STEP 5: Selection of the modeling data

Background: Quantification of the expression level of multicopy *THI1* and *S6PP* genes by RT-qPCR. Which to use: $2^{-\Delta CT}$ or $2^{-\Delta\Delta CT}$ equation that is a relevant decision!

There are two types of expression quantification: absolute and relative. In the present work, we addressed the comparative method using relative quantification [46], which assesses variations in gene expression (targets) within a specific sample (condition). This is achieved through normalization with a constitutive gene (housekeeping) or by examining changes relative to another reference sample, such as an internal control sample, guided by the $2^{-\Delta CT}$ or $2^{-\Delta\Delta CT}$ equations, respectively.

After normalization with housekeeping values, the $2^{-\Delta CT}$ equation allows comparisons among samples or between gene copies within a given sample. It is used for relative gene expression analysis within a single sample and does not provide information about the modulation of gene expression relative to other samples. It is a valuable tool for understanding gene expression levels in a specific context. The $2^{-\Delta\Delta CT}$ calculation is employed to assess alterations in gene expression concerning a control

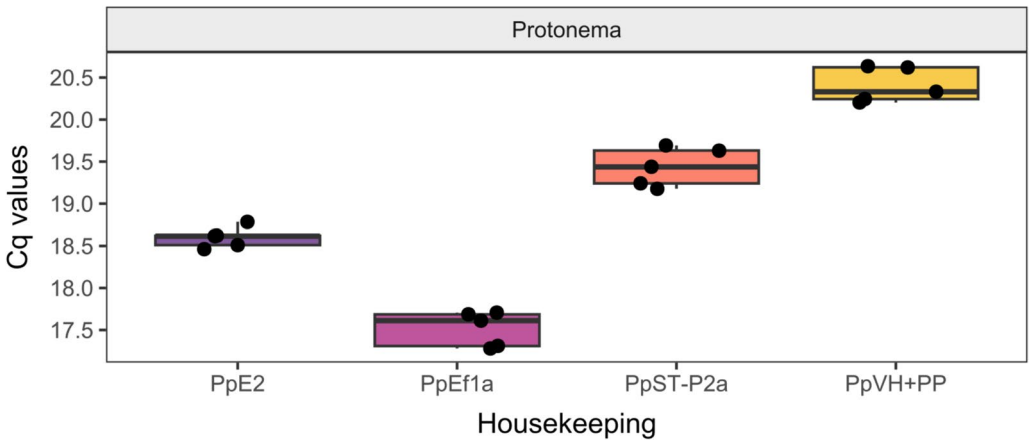

Fig. 5 Variation in CT values of the housekeeping genes across samples/conditions. Box plot graph showing variation in CT values for each housekeeping gene across the 24 h diurnal/circadian samples in protonemata. The median values are represented as lines across the box. The lower and the upper boxes represent the 25th and 75th percentile, respectively. Whiskers represent the maximum and minimum values

group. This method quantifies how often a particular gene has been either upregulated or downregulated, expressed as the fold change compared to a given reference, the internal reaction control. Each equation provides specific information related to the expression of the target gene. To illustrate these approaches, we explored our case study genes, THI1 and S6PP, and used $2^{-\Delta CT}$ or $2^{-\Delta\Delta CT}$ equations to analyze their expression level under varying conditions. RT-qPCR assays were performed in both week 1 and week 3, during 48 h and in a light/dark cycle (16h/8h), as described above.

The $2^{-\Delta CT}$ equation evidences that all copies of THI1 paralogs in protonemata tissues and adult gametophores reached their maximum expression peak between 6 and 10 PM, corresponding to the evening and early night. Conversely, the lowest expression is observed between 6 and 10 AM, corresponding to early morning, suggesting a diurnal expression pattern. Moreover, it was noted that Pp3c23_6600 is highly expressed in gametophore adult tissues, indicating that this specific paralog might play a prominent role in the development or function of adult gametophores (Fig. 6A). Thus, the method doesn't provide information about how the target gene expression is modulated, but it compares the expression of a gene across different conditions.

In the $2^{-\Delta\Delta CT}$ approach, we studied the oscillation of gene expression at five time points; 2 PM, 6 PM, 10 PM, 2 AM and 6 AM, during 48 h, relative to 10 AM that was used as an internal control in both protonema and gametophore for THI1. The analysis revealed that all THI1 paralogs exhibited rhythmic expression patterns over the 48-h cycle in response to light/dark conditions compared to the 10 AM control, with transcription levels increasing from 2 PM and reaching their maximum peak of positive regulation between 2 and 10 PM. The modulation of expression levels differed between paralogs and developmental stages (protonema and gametophore). The maximum positive regulation for some *THI1* paralogs (Pp3c20_13540, Pp3c20_13770, Pp3c24_10800) occurred at 6 PM in protonemata, with significant fold changes, 75-fold, 200-fold, and 20-fold, respectively (Fig. 6B). In gametophores, Pp3c23_6510 and Pp3c23_6600 had the highest expression levels at approximately 400-fold and 25-fold, but their modulation did not vary in protonemata. Expression modulation was not uniform across developmental stages, indicating distinct regulation profiles in protonemata and gametophores. Pp3c20_13540, Pp3c20_13770, and Pp3c24_10800 copies displayed rhythmic expression in both developmental stages. All five THI1 paralogs respond to the light/dark cycle in adult gametophore tissues, while four THI1 paralogs (Pp3c20_13770, Pp3c20_13540, Pp3c23_6600, and Pp3c24_10800) showed such regulation in protonemata (Fig. 6B). This suggests that their expression level is influenced by environmental cues, possibly related to light conditions, and the regulation of THI1 paralogs can vary between different tissues, with adult gametophores presenting distinct paralog patterns. The two equations support THI1 paralogs are regulated by the light/dark cycle and that their expression patterns may be tissue-specific.

The use of the $2^{-\Delta CT}$ equation to study the *S6PP* paralogs expression supports its higher expression in protonema, with the maximum peak of expression between 10 AM and 2 PM. We found that Pp3c22_1840 is the highest expressed copy among S6PP homologs, as verified by RNAseq analysis, showing its maximum expression between 6 AM and 6 PM with light present in both protonema and gametophore. Conversely, the Pp3c10_9450 and Pp3c14_5810 copies show lower expression levels throughout the developmental cycle (Fig. 7A).

To analyze *S6PP* homologs, the $2^{-\Delta\Delta CT}$ equation was applied with internal controls at 2 AM in protonema and 10 PM in gametophore; the same experimental setup used for THI1 over a 48-h light/dark cycle. The results showed, unlike *THI1*, there is no uniform modulation throughout the day among S6PP homologs. Pp3c10_9450, Pp3c14_5810, Pp3c22_1840 and Pp3c24_1340 homologs exhibited oscillation in both protonema and gametophore developmental stages (Fig. 7B).

All *S6PP* homologs are expressed during the light period, from 6 AM to 6 PM, and their lowest fold change was at 2 PM. Overall, the modulation of S6PP expression is more pronounced in the adult gametophore phase when compared to protonema. Pp3c24_1340 homolog exhibited higher fold-change values than the control, indicating a greater change in the gametophore. Pp3c24_1340 expression was higher in the gametophore compared to protonemata, reaching a significant peak of approximately 70-fold at 6 PM. Pp3c10_9450 homolog had low fold-change values (ranging from 1 to sixfold) in both gametophore and protonema. Pp3c14_5810 and Pp3c22_1840 copies showed the highest modulation peak at 2 PM, with approximately 13-fold modulation and Pp3c10_9450 had a sevenfold increase at 10 AM, all in protonema (Fig. 7B).

While the $2^{-\Delta\Delta CT}$ quantification methodology revealed that homologs show a higher modulation in the gametophore stage, we noticed that expression levels are higher in young protonema tissues based on $2^{-\Delta CT}$ quantification. These findings show how different *S6PP* homologs are expressed and modulated over a 48-h light/dark cycle in various developmental stages. The data indicates that expression patterns vary between homologs and between developmental stages, with some homologs exhibiting significant modulation during specific periods of the

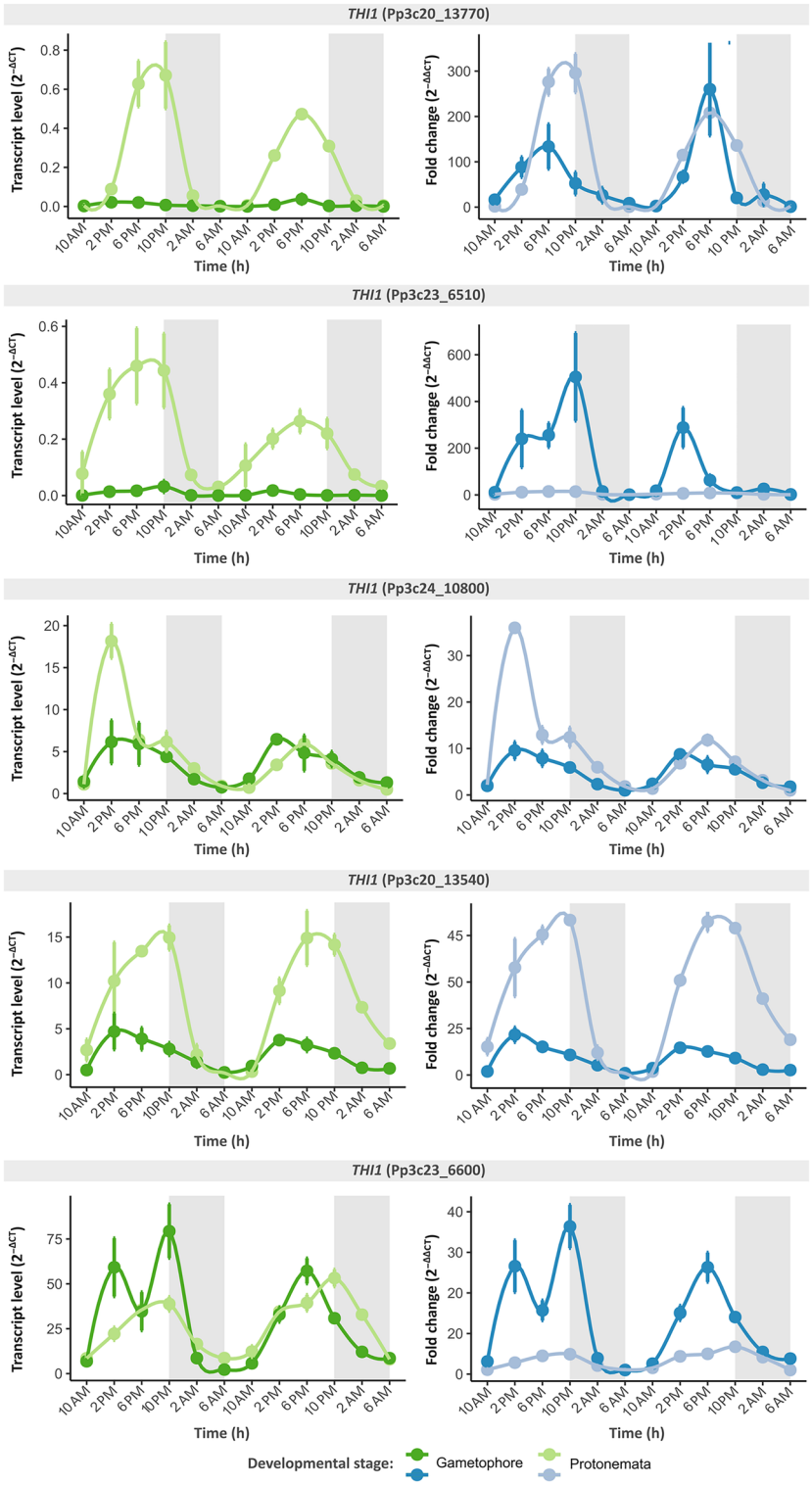

Fig. 6 THI1 RT-qPCR data modeling and visualization. **A** Transcript level ($2^{-\Delta CT}$) plot of THI1 paralogs in *P. patens* samples/conditions. **B** Fold change ($2^{-\Delta\Delta CT}$) plot of THI1 paralogs in *P. patens* samples/conditions

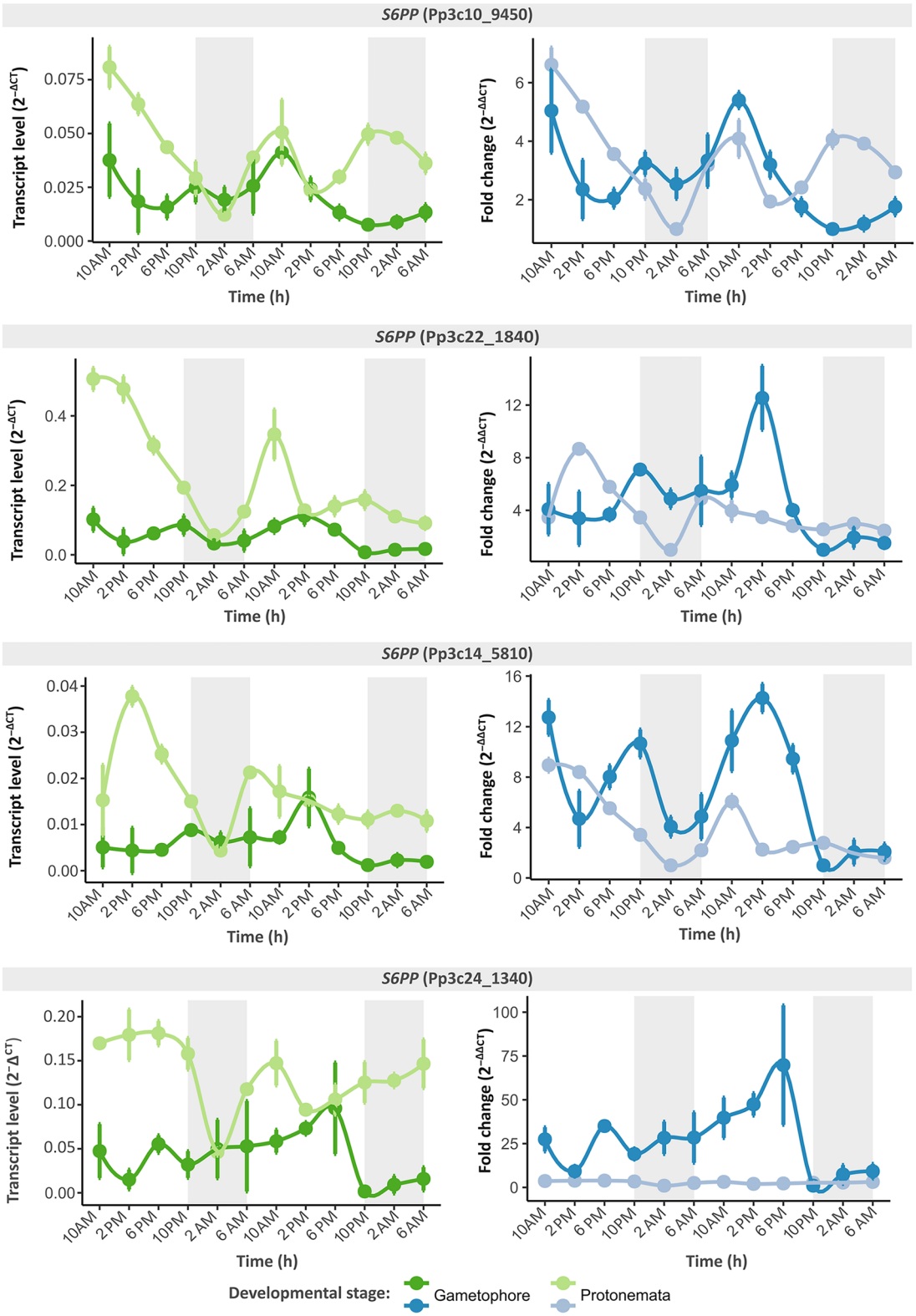

Fig. 7 S6PP RT-qPCR data modeling and visualization. **A** Transcript leves ($2^{-\Delta CT}$) plot of S6PP paralogs in *P. patens* samples/conditions. **B** Fold change ($2^{-\Delta\Delta CT}$) plot of S6PP paralogs in *P. patens* samples/conditions

day. The gametophore stage appears to have more pronounced modulation than protonemata (Fig. 7B).

Overall, this information provides valuable insights into the regulation of THI1 and S6PP paralogs, tissue-specific expression patterns, and responsiveness to the light/dark cycle. These findings could have implications for understanding the biological processes and functions associated with these genes in the context of the organism's development and environmental adaptation.

The two equations bring different biological information; the choice of calculation will always depend on the question to be addressed and how the information is approached. It is important to notice that $2^{-\Delta CT}$ analysis focuses on the expression levels of the target gene in the given experimental design, and it doesn't involve a comparison of modulation expression between different samples or conditions. This methodology assigns a value that provides biological information regarding the expression level of a gene in each condition and $2^{-\Delta\Delta CT}$ calculation furnishes biological insights into the modulation of gene expression, utilizing a specific sample as a reference or control to discern the extent of modulation relative to it. In addition, the take-home message of the present work is that the time of day when the sample is collected can interfere with the results obtained and eventually lead to a wrong biological interpretation.

Conclusions

Our study addresses the challenges associated with multicopy gene analysis, emphasizing the importance of rigorous methodologies and standardized approaches. The provided protocol, exemplified in the model moss *P. patens* with S6PP and THI1 genes, highlights the significance of identifying appropriate homologous genes. We focus on the variables that must be considered when using gene expression quantification techniques and the differences in evaluating gene expression using different methodologies.

The streamlined and curated protocol presented here serves as a practical framework for researchers, employing bioinformatics tools and experimental techniques to reliably and reproducibly understand the intricate expression patterns of individual paralogs among multicopy genes. Despite recent advances, significant gaps persist in deciphering the functional differences of individual gene copies and their contributions to plant adaptation. This approach can be adapted for specific organisms and research objectives, facilitating in-depth investigations into the role of multicopy genes in diverse biological processes. Further research in this area is crucial for unlocking the full potential of multicopy genes in diverse applications such as plant breeding, biotechnology, and crop improvement, ultimately shaping the future of agriculture and plant biology.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13007-025-01329-0.

Additional file 1
Additional file 2
Additional file 3
Additional file 4

Acknowledgements
Authors acknowledge Dr Tatiana C. Pisetta technical support to maintain GaTELab infrastructure.

Author contributions
VS and HMD conceived and outlined the experimental design, analyzed the data, prepared figures and tables, authored, reviewed and approved the final draft. MTP contributed with microscopy of P. patens life cycle, reviewed the manuscript, and approved the final draft. MAVS conceived the project, discussed results, authored, reviewed and approved the final draft.

Funding
Financial support was obtained from grants CNPq 310779/2017-0, FAPESP 2016/17545-8 to MAVS and 2019/26129-6 to MTP. VSP CAPES Financial code 001, HMD FAPESP 2022/16208-9. The funders had no role in study design, data collection, analysis, publication decision, or manuscript preparation.

Availability of data and materials
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 7 May 2024 Accepted: 17 January 2025
Published online: 07 March 2025

References

1. Fischer I, Dainat J, Ranwez V, Glémin S, Dufayard JF, Chantret N. Impact of recurrent gene duplication on adaptation of plant genomes. BMC Plant Biol. 2014;14(1):1–15.
2. Flagel LE, Wendel JF. Gene duplication and evolutionary novelty in plants. New Phytol. 2009;183(3):557–64.
3. Kong W, Wang Y, Zhang S, Yu J, Zhang X. Recent advances in assembly of plant complex genomes. Genom Proteom Bioinform. 2023;21(3):427–39.
4. Makarova KS, Wolf YI, Mekhedov SL, Mirkin BG, Koonin EV. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. Nucleic Acids Res. 2005;33(14):4626–38.
5. Seidel DS, Claudino PH, Sperotto G, Wendt SN, Shomo ZD, Mural RV, Dias HM. Comprehensive analysis of the Aquaporin genes in Eucalyptus grandis suggests potential targets for drought stress tolerance. Trop Plant Biol. 2023;17:233–48.

6. Stupar RM, Beaubien KA, Jin W, Song J, Lee MK, Wu C, et al. Structural diversity and differential transcription of the patatin multicopy gene family during potato tuber development. Genetics. 2006;172(2):1263–75.
7. Guitton AE, Berger F. Loss of function of MULTICOPY SUPPRESSOR OF IRA 1 produces nonviable parthenogenetic embryos in Arabidopsis. Curr Biol. 2005;15(8):750–4.
8. Weglöhner W, Subramanian AR. Multicopy GTPase center protein L12 of Arabidopsis chloroplast ribosome is encoded by a clustered nuclear gene family with the expressed members closely linked to tRNA (Pro) genes. J Biol Chem. 1994;269(10):7330–6.
9. Qian W, Zhang J. Genomic evidence for adaptation by gene duplication. Genome Res. 2014;24(8):1356–62.
10. Zeira R, Shamir R. Genome rearrangement problems with single and multiple gene copies: a review. In: Bioinformatics and phylogenetics: Seminal contributions of Bernard Moret. Cham: Springer; 2019. p. 205–41.
11. Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics. 2012;28(21):2711–8.
12. Namias A, Sahlin K, Makoundou P, Bonnici I, Sicard M, Belkhir K, Weill M. Nanopore sequencing of PCR products enables multicopy gene family reconstruction. Comput Struct Biotechnol J. 2023;21:3656–64.
13. Reis-Cunha JL, Coqueiro-dos-Santos A, Pimenta-Carvalho SA, Marques LP, Rodrigues-Luiz GF, Baptista RP, et al. Accessing the variability of multicopy genes in complex genomes using unassembled next-generation sequencing reads: The Case of Trypanosoma cruzi Multigene Families. Mbio. 2022;13(6):e02319-22.
14. Manni M, Berkeley MR, Seppey M, Zdobnov EM. BUSCO: assessing genomic data quality and beyond. Current Protocols. 2021;1(12): e323.
15. Das S, Bansal M. Variation of gene expression in plants is influenced by gene architecture and structural properties of promoters. PLoS ONE. 2019;14(3): e0212678.
16. Dias HM, Vieira AP, de Jesus EM, de Setta N, Barros G, Van Sluys MA. Functional and comparative analysis of THI1 gene in grasses with a focus on sugarcane. PeerJ. 2023;11: e14973.
17. Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. Gene duplicability of core genes is highly consistent across all angiosperms. Plant Cell. 2016;28(2):326–44.
18. Guo YL. Gene family evolution in green plants with emphasis on theorigination and evolution of Arabidopsis thaliana genes. Plant J. 2013;73(6):941-51. https://doi.org/10.1111/tpj.12089. Epub 2013Jan 15. PMID: 23216999
19. Dias HM, Toledo NDA, Mural RV, Schnable JC, Van Sluys MA. THI1 gene evolutionary trends: A comprehensive plant-focused assessment via data mining and large-scale analysis. Genome Biol Evol. 2024. https://doi.org/10.1093/gbe/evae212.
20. Partida VGS, Dias HM, Corcino DSM, Van Sluys MA. Sucrose-phosphate phosphatase from sugarcane reveals an ancestral tandem duplication. BMC Plant Biol. 2021;21:1–13.
21. Yuan JS, Galbraith DW, Dai SY, Griffin P, Stewart CN. Plant systems biology comes of age. Trends Plant Sci. 2008;13(4):165–71.
22. Izawa T, Shimamoto K. Becoming a model plant: the importance of rice to plant science. Trends Plant Sci. 1996;1(3):95–9.
23. Koornneef M, Meinke D. The development of Arabidopsis as a model plant. Plant J. 2010;61(6):909–21.
24. Rensing SA, Goffinet B, Meyberg R, Wu SZ, Bezanilla M. The moss Physcomitrium (Physcomitrella) patens: a model organism for non-seed plants. Plant Cell. 2020;32(5):1361–76.
25. Naramoto S, Hata Y, Fujita T, Kyozuka J. The bryophytes Physcomitrium patens and Marchantia polymorpha as model systems for studying evolutionary cell and developmental biology in plants. Plant Cell. 2022;34(1):228–46.
26. Lang D, Ullrich KK, Murat F, Fuchs J, Jenkins J, Haas FB, et al. The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. Plant J. 2018;93(3):515–33.
27. Schaefer DG, Zrÿd JP. The moss Physcomitrella patens, now and then. Plant Physiol. 2001;127(4):1430-8. PMID: 11743086; PMCID: PMC1540175.
28. Rensing SA, Ick J, Fawcett JA, Lang D, Zimmer A, Van de Peer Y, Reski R. An ancient genome duplication contributed to the abundance of metabolic genes in the moss Physcomitrella patens. BMC Evol Biol. 2007;7(1):1–10.
29. Altschul SF, GishW, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10. https://doi.org/10.1016/S0022-2836(05)80360-2. PMID: 2231712
30. Yang M, Derbyshire MK, Yamashita RA, Marchler-Bauer A. NCBI's conserved domain database and tools for protein domain analysis. Curr Protoc Bioinform. 2020;69(1): e90.
31. Basu MK, Poliakov E, Rogozin IB. Domain mobility in proteins: functional and evolutionary implications. Brief Bioinform. 2009;10(3):205–16.
32. Wang Y, Zhang H, Zhong H, Xue Z. Protein domain identification methods and online resources. Comput Struct Biotechnol J. 2021;19:1145–53.
33. Albà MM, Castresana J. On homology searches by protein Blast and the characterization of the age of genes. BMC Evol Biol. 2007;7:1–8.
34. Kerfeld CA, Scott KM. Using BLAST to teach "E-value-tionary" concepts. PLoS Biol. 2011;9(2): e1001014.
35. Fernandez-Pozo N, Haas FB, Meyberg R, Ullrich KK, Hiss M, Perroud PF, et al. PEATmoss (Physcomitrella Expression Atlas Tool): a unified gene expression atlas for the model plant Physcomitrella patens. Plant J. 2020;102(1):165–77.
36. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15(2):121–32.
37. Die JV, Román B, Nadal S, González-Verdejo CI. Evaluation of candidate reference genes for expression studies in Pisum sativum under different experimental conditions. Planta. 2010;232:145–53.
38. Huggett J, Dheda K, Bustin S, Zumla A. Real-time RT-PCR normalisation; strategies and considerations. Genes Immun. 2005;6(4):279–84.
39. Radonić A, Thulke S, Mackay IM, Landt O, Siegert W, Nitsche A. Guideline to reference gene selection for quantitative real-time PCR. Biochem Biophys Res Commun. 2004;313(4):856–62.
40. Artico S, Nardeli SM, Brilhante O, Grossi-de-Sa MF, Alves-Ferreira M. Identification and evaluation of new reference genes in Gossypium hirsutum for accurate normalization of real-time quantitative RT-PCR data. BMC Plant Biol. 2010;10:1–12.
41. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, et al. The MIQE Guidelines: Minimum information for publication of quantitative real-time PCR experiments. Clin Chem. 2009;55(4):611–22.
42. Czechowski T, Stitt M, Altmann T, Udvardi MK, Scheible WR. Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. Plant Physiol. 2005;139(1):5–17.
43. Gutierrez L, Mauriat M, Guénin S, Pelloux J, Lefebvre JF, Louvet R, et al. The lack of a systematic validation of reference genes: a serious pitfall undervalued in reverse transcription-polymerase chain reaction (RT-PCR) analysis in plants. Plant Biotechnol J. 2008;6(6):609–18.
44. Le Bail A, Scholz S, Kost B. Evaluation of reference genes for RT qPCR analyses of structure-specific and hormone regulated gene expression in Physcomitrella patens gametophytes. PLoS ONE. 2013;8(8): e70998.
45. Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. Cell. 2016;165(3):535–50.
46. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method. Methods. 2001;25(4):402–8.
47. Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, et al. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 2021;49(D1):D344–54.
48. Bustin S, Huggett J. qPCR primer design revisited. Biomol Detect Quantif. 2017;14:19–28.
49. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47(D1):D427–32.
50. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):307–21.
51. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
52. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417–9.

Publisher's Note