

RESEARCH

Open Access



Evolution and functional diversification of catalase genes in the green lineage

Luzhao Pan^{1,2,3,4}, Yin Luo⁵, Jin Wang¹, Xiumin Li⁵, Bingqian Tang⁵, Huiping Yang⁵, Xilin Hou¹, Feng Liu^{2,3,4*} and Xuexiao Zou^{1,2,3,4*}

Abstract

Background: Catalases (CATs) break down hydrogen peroxide into water and oxygen to prevent cellular oxidative damage, and play key roles in the development, biotic and abiotic stresses of plants. However, the evolutionary relationships of the plant CAT gene family have not been systematically reported.

Results: Here, we conducted genome-wide comparative, phylogenetic, and structural analyses of CAT orthologs from 29 out of 31 representative green lineage species to characterize the evolution and functional diversity of CATs. We found that CAT genes in land plants were derived from core chlorophytes and detected a lineage-specific loss of CAT genes in Fabaceae, suggesting that the CAT genes in this group possess divergent functions. All CAT genes were split into three major groups (group α , β 1, and β 2) based on the phylogeny. CAT genes were transferred from bacteria to core chlorophytes and charophytes by lateral gene transfer, and this led to the independent evolution of two types of CAT genes: α and β types. Ten common motifs were detected in both α and β groups, and β CAT genes had five unique motifs, respectively. The findings of our study are inconsistent with two previous hypotheses proposing that (i) new CAT genes are acquired through intron loss and that (ii) the Cys-343 residue is highly conserved in plants. We found that new CAT genes in most higher plants were produced through intron acquisition and that the Cys-343 residue was only present in monocots, Brassicaceae and *Pp_CatX7* in *P. patens*, which indicates the functional specificity of the CATs in these three lineages. Finally, our finding that CAT genes show high overall sequence identity but that individual CAT genes showed developmental stage and organ-specific expression patterns suggests that CAT genes have functionally diverged independently.

Conclusions: Overall, our analyses of the CAT gene family provide new insights into their evolution and functional diversification in green lineage species.

Keywords: Catalases, Green plants, Phylogeny, Functional diversification, Functionally conserved

Background

The key components of the reactive oxygen species (ROS) gene network are thought to have originated as early as 4.1–3.5 billion years ago [1]. ROS-related proteins, such as catalases, are thought to have originated approximately

2.5 billion years ago, and the origin of these proteins was likely critically important to the survival of organisms during the Great Oxidation Event, which is thought to have occurred from 2.4 to 2.0 billion years ago [1]. After this planet-changing event, CATs evolved in aerobic organisms [2], which led to the appearance of three metalloenzyme families: (i) typical (monofunctional) heme CATs, (ii) (bifunctional) heme Catalase-Peroxidase, and (iii) (non-heme) manganese CATs [3–5]. The typical (monofunctional) heme CATs are most widespread in living organisms [6], they are the most extensively

*Correspondence: liufengrich@126.com; zouxuexiao428@163.com

¹ College of Horticulture, Nanjing Agricultural University, Nanjing, China

² College of Horticulture, Hunan Agricultural University, Changsha, China

Full list of author information is available at the end of the article



studied. Below, we summarize current research on typical CATs. Non-heme manganese CATs are widely distributed in microbial life and play an important role in antioxidant defense [7]. The typical heme CATs (hydrogen peroxide oxidoreductase E.C. 1.11.1.6) are critically important antioxidant enzymes that catalyze the degradation of hydrogen peroxide to produce water and oxygen [8]. CATs are mainly present in peroxisomes, but they have also been detected in the cytoplasm, mitochondria, and chloroplasts [9]. Typical heme catalases are the only CATs present in plants [2, 3, 8]. But in some plants (such as *Adiantum capillus-veneris* and *Marchantia polymorpha*), those CATs were generally classified into a distinct phylogenetic clade [8]. Unlike animals which only possess a single CAT, plants generally have multiple CAT isozymes [10], which suggests that the diversity in the structure and function of CATs among plants might be particularly rich [6, 10–13].

CATs play a key role in the development, defense, and senescence of plants, and several factors, such as hydrogen peroxide, temperature, plant hormones, wounding, and circadian rhythm, affect the expression of *CAT* genes [10, 13]. Plant *CATs* have been classified into three classes according to their expression properties, class I, class II, and class III, which are expressed in photosynthetic, vascular, and reproductive tissues, respectively [6]. In maize, *CAT-1* is transcribed and translated following tetrad formation and is expressed in mature pollen [14]; the expression of *Cat3* is regulated by circadian rhythm [15]. Recent studies have shown that maize *CATs* can facilitate the replication of maize chlorotic mottle virus [16]; catalase 1 in particular can promote viral multiplication and infection [17]. In *Arabidopsis*, *CATs* play key roles in the responses to irradiance and pathogens [18], signal transduction [19], plant hormones, plant senescence, and reproduction [20, 21]. The *Arabidopsis* CPK8 can phosphorylate the Ser-261 residue of *CAT3* to regulate ABA-mediated stomatal regulation in response to drought stress [19]. The *cat1/2/3* triple mutant generated using CRISPR/Cas9 technology displays severe redox perturbation and growth defects [21]. In pepper, the expression of *CaCat1* and *CaCat2* is differentially regulated by circadian rhythm, and the expression of *CaCat1* responds to wounding and paraquat treatment [12].

A phylogenetic analysis of prokaryotes and eukaryotes revealed that *CATs* comprise three main clades: clade 1, clade 2, and clade 3 [2, 3, 8]. Clade 1-type *CATs* occur in eubacteria, algae, and plants; clade 2-type *CATs* occur in eubacteria and fungi; and clade 3-type *CATs* are the most abundant and have been detected in archaeobacteria, fungi, protists, plants, and animals [2, 3, 8]. Clade 3-type *CATs* have been extensively studied in humans and other animals for their scientific and medical importance [2].

Clade 1-type *CATs* and clade 3-type *CATs* evolved from the older clade 2-type *CATs* [3, 8]. Plants generally contain both clade 1 and clade 3-type *CATs* [8]. Clade 1-type *CATs* are abundantly distributed in plants, ranging from unicellular green alga (*C. reinhardtii*) [22] to various land plants [11–13, 19]. A phylogenetic tree of 200 typical catalases were classified into three main evolutionary clades, and clade 1 contained plant catalases major group, Firmicutes group A and Proteobacterial minor group [8]. Trees of 70 typical catalases from all main living kingdoms classify bacterial *CATs* and major plant *CAT* members as clade 1 [2]. By contrast, clade 3-type *CATs* have only been detected in a few plant species to date [8]. Whether clade 3-type *CATs* are more widespread in plants and how these *CATs* evolved require further investigation. Plants also often possess multiple copies of *CAT* genes, but the evolutionary relationships among these genes have not yet been fully clarified.

Here, we conducted genome-wide comparative, phylogenetic, and structural analyses of 82 *CAT* sequences from major green lineage (green algae and land plants) taxa to evaluate the origin, distribution and duplication patterns, and functional features of *CAT* genes in plants. The results of our analyses revealed gene number variation in *CAT* genes among species and the lineage-specific loss of *CAT* genes in Fabaceae. *CAT* genes were divided into three subfamilies, and lateral gene transfer (LGT) led to the evolution of two clades of *CAT* genes. Paralogs were conserved among species, and changes in key amino acid residues might have led to the acquisition of new functions. The results of our study provide new insights into the evolution and functional diversification of *CATs* in plants.

Results

Homolog searches and verification of *CAT* genes

A BLASTP search was performed using the HMMER model to identify *CAT* genes among species with 31 fully sequenced genomes (Table 1). No *CAT* sequences were searched in *M. pusilla* and *O. lucimarinus*, so these two species were excluded. A total, 83 homolog sequences were identified from four algae (including three core chlorophytes and three charophytes), three bryophytes (liverworts, mosses and hornworts), one lycophyte, one gymnosperm, two ferns and 16 angiosperms (Table 1, Table S1). The Pfam database was then used to verify the presence of two conserved *CAT* domains in the putative *CAT* candidates: Catalase (PF00199.19) and Catalase-related immune-responsive (PF06628.12). These two domains were present in all 83 sequences and were thus used in subsequent analyses.

One gene from *M. truncatula* (*Medtr1386s0010*) showed low sequence identity, and its position in the

Table 1 The detail information of plant genome

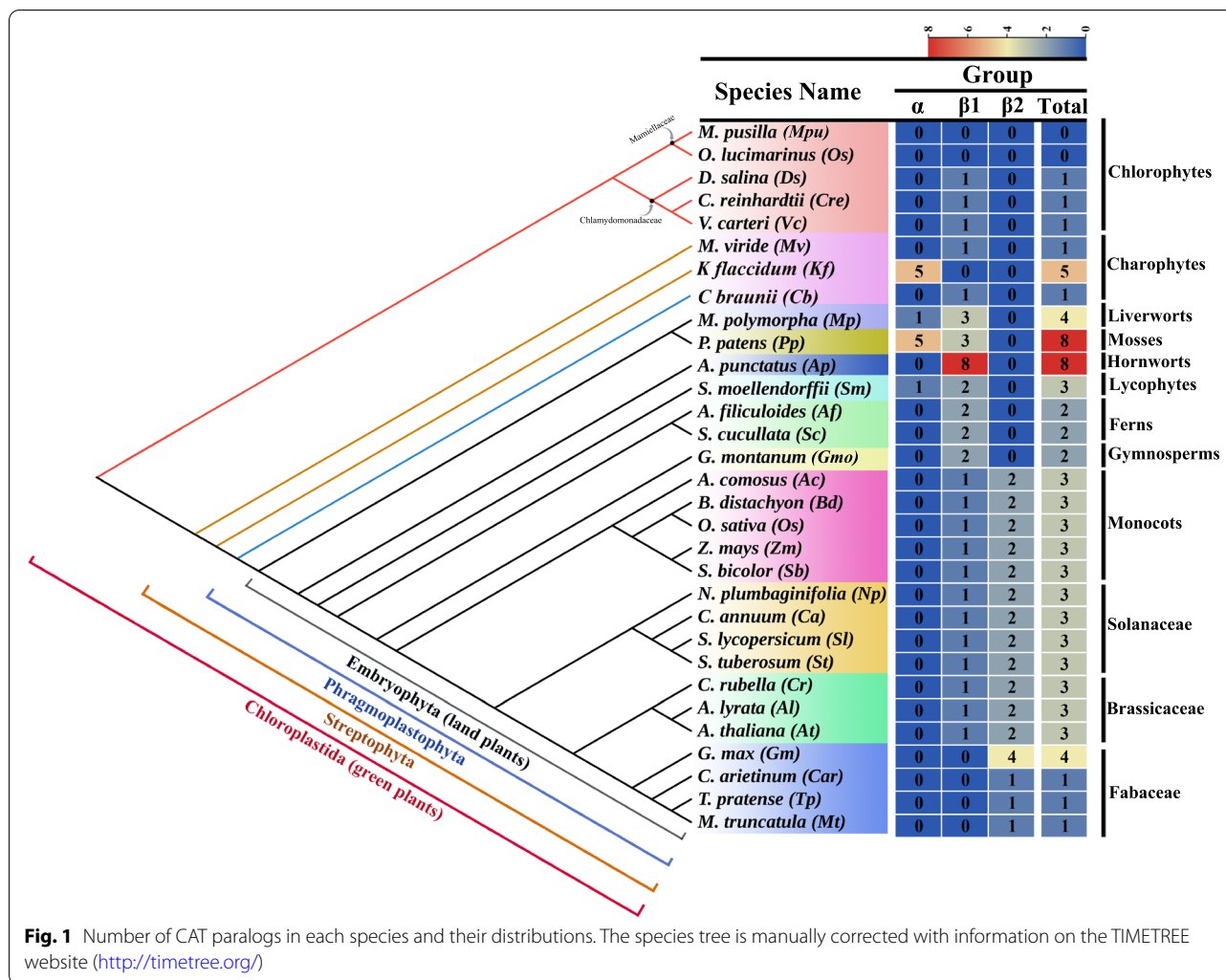
Species name	Genome version	Gene number	Genome size	Gene number vs. genome size
<i>Micromonas pusilla</i>	CCMP1545 v3.0	0	22 Mbp	0
<i>Ostreococcus lucimarinus</i>	v2.0	0	13.2 Mbp	0
<i>Dunaliella salina</i>	v1.0	1	343.7 Mbp	0.29%
<i>Chlamydomonas reinhardtii</i>	v5.6	1	121 Mbp	0.83%
<i>Volvox carteri</i>	v2.1	1	131.2 Mbp	0.76%
<i>Mesostigma viride</i>	CCAC 1140	1	329 Mbp	0.30%
<i>Klebsormidium flaccidum</i>	v1.1	5	117.1 ± 21.8 Mbp	4.27%
<i>Chara braunii</i>	Cbr_1.0	1	1430 Mbp	0.069%
<i>Physcomitrella patens</i>	v3.3	8	473 Mbp	1.69%
<i>Marchantia polymorpha</i>	v3.1	4	225.8 Mbp	1.77%
<i>Anthoceros punctatus</i>		8	132.8 Mbp	6.02%
<i>Selaginella moellendorffii</i>	v1.0	3	212.5 Mbp	1.41%
<i>Azolla filiculoides</i>	v1.2	2	759 Mbp	0.27%
<i>Salvinia cucullata</i>	v1.2	2	250 Mbp	0.80%
<i>Gnetum montanum</i>	v1.1	2	4200 Mbp	0.05%
<i>Ananas comosus</i>	v3	3	526 Mbp	0.57%
<i>Brachypodium distachyon</i>	v3.1	3	272 Mbp	1.10%
<i>Oryza sativa</i>	v7.0	3	430 Mbp	0.70%
<i>Zea mays</i>	RefGen_V4	3	2300 Mbp	0.13%
<i>Sorghum bicolor</i>	v3.1.1	3	818 Mbp	0.37%
<i>Nicotiana glauca</i>	NT	3	2000 Mbp	0.15%
<i>Capsicum annuum</i>	cv CM334_1.55	3	3070~3480 Mbp	0.10%
<i>Solanum lycopersicum</i>	ITAG4.0	3	900 Mbp	0.33%
<i>Solanum tuberosum</i>	v4.03	3	850 Mbp	0.35%
<i>Capsella rubella</i>	v1.1	3	219 Mbp	1.37%
<i>Arabidopsis lyrata</i>	v2.1	3	207 Mbp	1.45%
<i>Arabidopsis thaliana</i>	TAIR10	3	125 Mbp	2.40%
<i>Glycine max</i>	Wm82.a2. v1	4	1115 Mbp	0.36%
<i>Cicer arietinum</i>	v1.0	1	738 Mbp	0.14%
<i>Trifolium pratense</i>	v2	1	420 Mbp	0.24%
<i>Medicago truncatula</i>	Mt4.0v1	1	454 Mbp	0.22%

NT Unknown

phylogeny, its intron phase, and motif elements were inconsistent with the general evolutionary patterns revealed by phylogenetic and structural analyses. Moreover, *C. arietinum* and *T. pratense*, which are from the same family as *M. truncatula*, only possessed a single CAT member (Fig. 1, Table S1). We speculate that *Medtr1386s0010*, which is categorized as encoding a CAT heme-binding enzyme in the NCBI database, was incorrectly annotated; this gene was thus excluded from subsequent analyses. The remaining 82 genes were renamed using the same nomenclature that has been used for CAT genes in previous studies of various taxa, such as *Arabidopsis*, rice, and maize. Published gene names were used for CAT genes that have been previously characterized

and the remaining genes were sequentially arranged based on the abbreviation of the species name.

Information on the renamed CATs is shown in Table S1. The protein length of CATs varied substantially among core chlorophytes, charophytes, and lower land plants. Core chlorophytes CAT proteins all possessed 493 amino acid residues. *M. viride* and *C. braunii* all possessed 492 amino acids, whereas the length of *K. flaccidum* varied from 503 to 651 amino acids. The length of CAT proteins of some land plants (including bryophytes, lycophytes and ferns) was very unstable, ranging from 329 to 601 amino acid residues. On the contrary, the length of the CATs in the remaining plants was generally 492 amino acid residues; some minor sequences deviating from this



pattern were likely caused by genome annotation errors. Some protein sequences were successfully annotated using online tools (Table S2).

Gene number variation of CAT genes among various species

In total, orthologous CAT genes were identified from 29 out of 31 organisms, including six algae and 23 land plants (Fig. 1). Gene number variation (GNV) of CAT genes varied substantially among all organisms sampled, including core chlorophytes (one gene), *M. viride* and *C. braunii* (one), *K. flaccidum* (five), monocots (three), dicots (one to four), *A. punctatus* and *P. patens* (eight) (Table 1 and Fig. 1). These results suggest that there was no correlation between CAT gene number and genome size (Table 1).

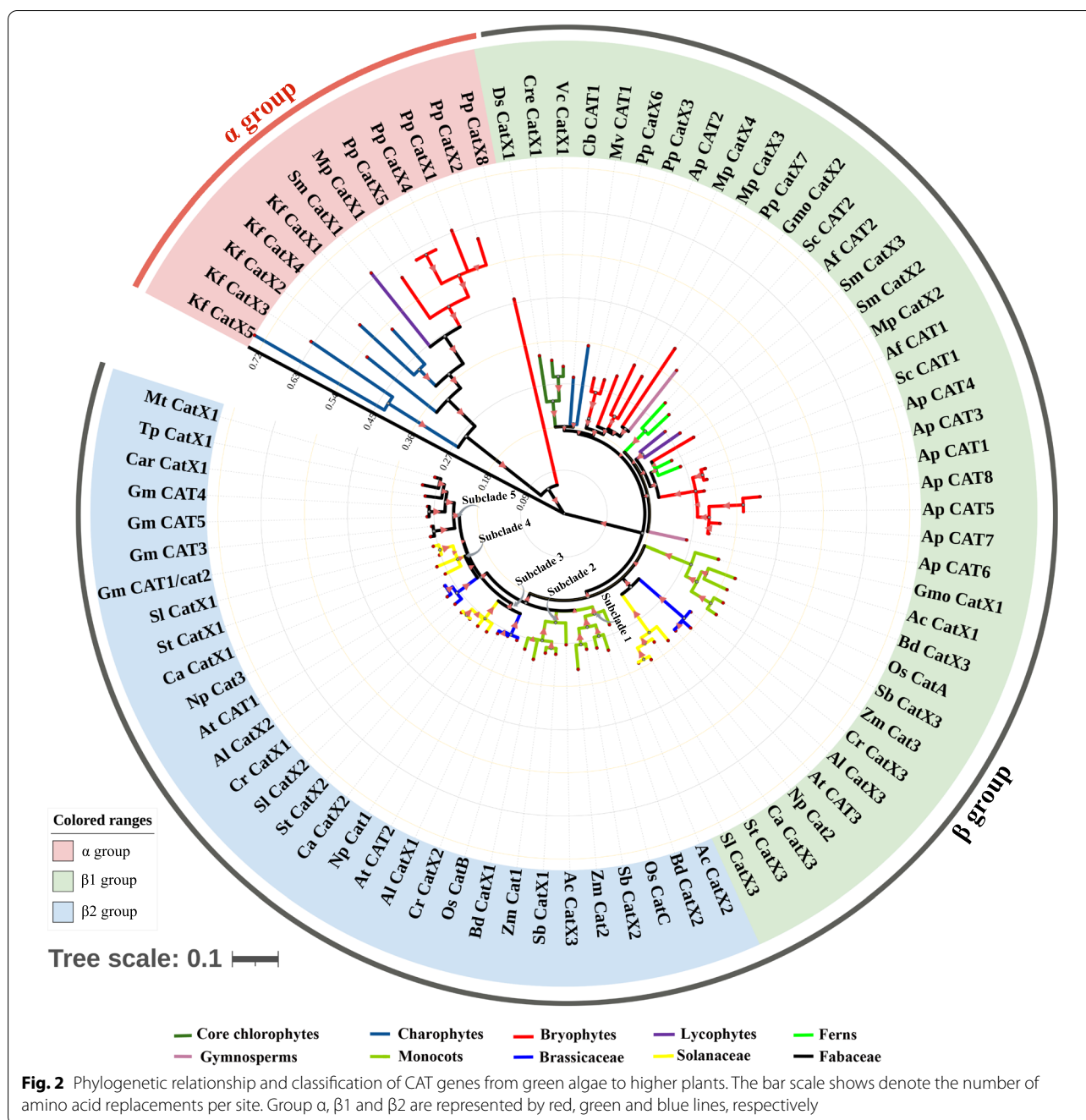
GNV among species was mainly driven by differences among groups. Multiple CAT members from *K. flaccidum* and *P. patens* were observed in group α, and

several CAT from bryophytes, lycophytes, ferns and gymnosperms were detected in group β1. Only CAT genes from angiosperms were included in group β2. Three CAT genes were present in all angiosperms, with the exception of members of Fabaceae, which only possessed a single CAT member; however, four CATs were observed in soybean.

No CAT genes were detected in Mamiellaceae (*M. pusilla* and *O. lucimarinus*), and only one CAT gene was detected in each member of the Chlamydomonadaceae (core chlorophytes) (Fig. 1), suggesting that CAT genes in plants might be derived from ancestral core chlorophyte genes. Multiple CATs were detected in *K. flaccidum*, suggesting that the first large-scale expansion of CAT genes occurred in Charophyte.

Classification of the CAT gene family in plants

A phylogenetic analysis was conducted using the full-length protein sequences from all organisms sampled



with the ML and NJ methods to clarify the evolutionary relationships among *CAT* genes. The overall topologies of the ML and NJ trees were similar. Thus, only the ML tree was shown. All *CAT* genes could be divided into three major phylogenetic lineages, which were referred to as group α, β1 and β2 based on phylogeny (Fig. 2, Fig. S1). Meanwhile, this phylogenetic separation was supported by their different exon–intron structure (Fig. S2). Group α only included *CAT* genes from Charophyte, Bryophyte,

and Lycophyte; Group β1 contained *CAT* genes from all green plants except for *K. flaccidum* and Fabaceae; and group β2 only has *CAT* genes from monocots and dicots.

All *CAT* genes from *K. flaccidum* were included in group α and have undergone at least two rounds of duplication (Fig. 2). A part of *CAT*s from bryophytes and lycophytes were also included in group α, but their evolutionary history was difficult to infer because of their high

levels of sequence divergence. No *CATs* from flowering plants were present in this group.

The core chlorophytes *CAT* genes were the earliest diverging lineage in group $\beta 1$ and formed a small monophyletic cluster (Fig. 2). Multiple *CAT* genes from bryophytes and a lycophyte successively diverged, and at least one duplication event has occurred in each species. A single gymnosperm *CAT* gene and multiple *CAT* orthologs from angiosperms formed three distinct clades (Fig. 2). The phylogeny, coupled with gene structure analysis (Fig. S2), revealed substantial differences in the structure of *CAT* genes within group $\beta 1$. By contrast, *CAT* orthologs within the same lineage, such as *CAT* orthologs within Poaceae, Brassicaceae, and Solanaceae, were relatively conserved in their intron phases, exon length, and number of exons (Fig. S2).

Group $\beta 2$ was an angiosperm-specific clade and consists of five subclades. *CAT* genes underwent several duplication events prior to the divergence of monocots and dicots, indicating that monocots and dicots possess the ancestral genes in this clade. Within monocots, all *CAT* orthologs underwent a round of duplication, which resulted in the formation of subclades 1 and 2. Within dicots, *CAT* orthologs from rosids and asterids underwent a round of duplication, which generated subclades 3 and 4. *CAT* genes from the Fabaceae family appeared to be differentiated from those of other dicots, which resulted in the formation of the Fabaceae-specific subclade 5 with 86% bootstrap support.

The evolutionary relationships of plant *CAT* gene family among the major lineages of Life

Phylogenetic and structural analysis provided many insights into the origin and evolution of *CAT* genes. In Fig. 3a, group α and β can be treated as two distinct evolutionary types of *CAT* genes, respectively. The α group *CAT* genes evolved more rapidly than β *CATs* according to the positions of their branch nodes and branch lengths. Furthermore, the α group was derived from *K. flaccidum* (charophyte), whereas the β group, which included all *CAT* genes from subfamilies $\beta 1$ and $\beta 2$, was derived from chlorophyte. In order to explore the evolutionary relationships of plant *CAT* genes among the major lineages of Life, we then constructed another phylogenetic tree using *CAT* sequences from representative bacteria, protists, fungi, plants, and animals to further trace the origin and evolution of *CAT* genes (Fig. 3b, Table S3). The phylogenetic tree revealed that the α and β groups of the *CAT* gene family corresponded exactly to clade 3 and clade 1 proposed by Zámocký [6], whereas clade 2 was a missing clade in the green lineage. Further analysis revealed that α and β *CAT* genes, both of which included *CAT* genes from bacteria with strong support, shared a common ancestral

gene (Fig. 3b). The clade2 catalases were detected only in bacteria, protists and fungi. Clade 3 (α group) was widely distributed in bacteria, protists, fungi, animals and partial green plants, and clade1 (group β) was present in almost all analyzed plants (Fig. 3b). Group α and β shared 10 conserved motifs, but showed highly divergence at C-terminus and N-terminus (Fig. 3c).

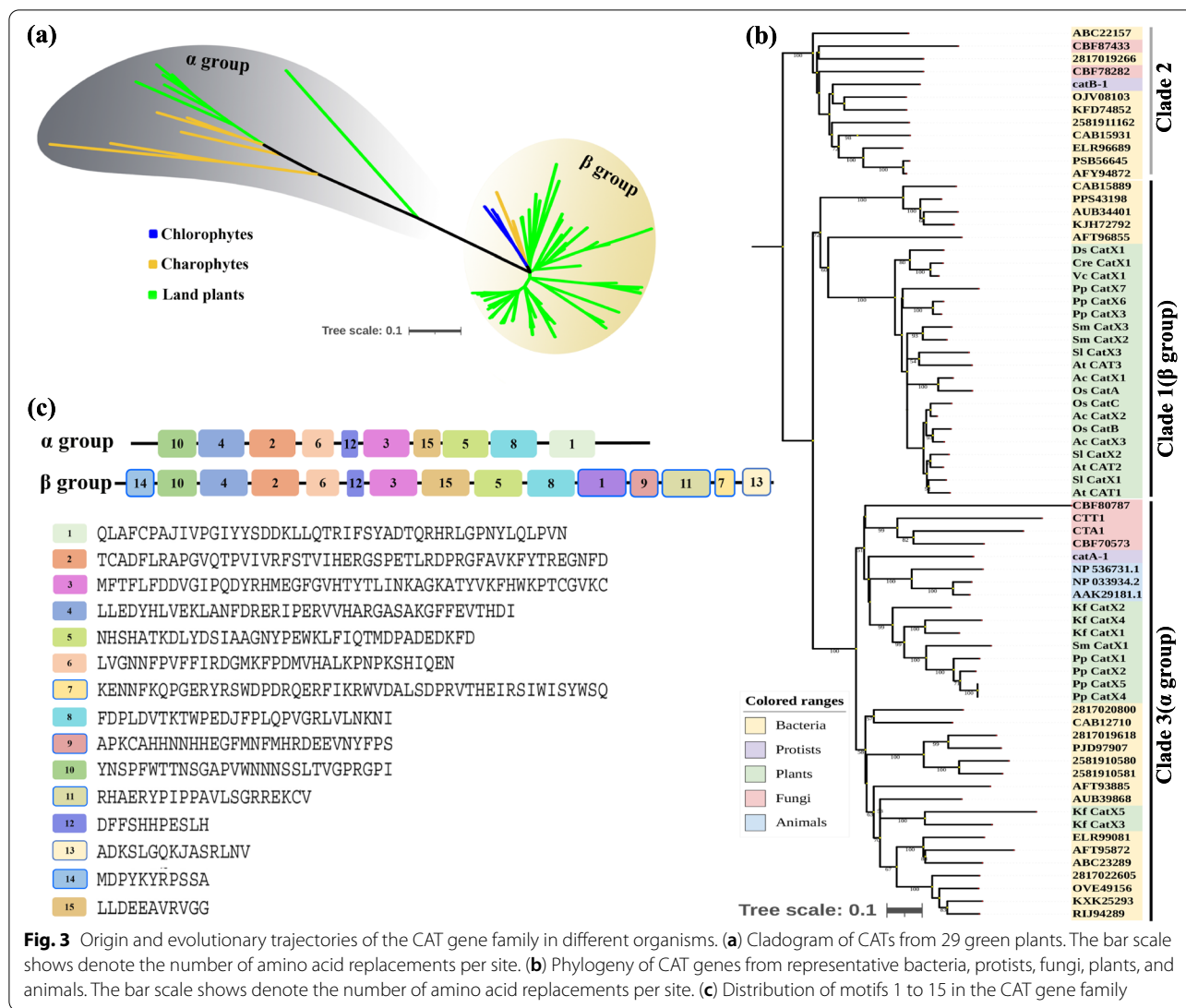
Key amino acid residues in Brassicaceae and monocots

A recent study has shown that *CAT3* in *Arabidopsis* (i.e., *ROG1*) possesses a conserved Cys-343 residue that can decrease its catalase activity but increase its transnitrosylase activity [23]. Conversely, *CAT2* (*At_CAT2*), which possesses a Thr-343 residue, shows reduced *CAT* activity but increased transnitrosylase activity [23]. Multiple sequence alignment of *CAT* sequences from β group (Fig. 4a, Fig. S4) revealed that the Cys-343 residue was only present in $\beta 1$ group *CATs* from monocots and Brassicaceae. With the exception of *Gm_CatX2*, the remaining *CATs* were consistent with *At_CAT2* in *Arabidopsis* in possessing the highly conserved Thr-343 residue. It's worth noting that *Pp_CatX7* also contained a Cys-343 residue, but two *Mp_CatX3/4* orthologs, which were closely related to *Pp_CatX7* in the phylogenetic tree, had a Thr-343 residue instead of a Cys-343 residue. In addition, α *CATs* possessed neither a Cys-343 residue nor Thr343 residue (Fig. 4a, Fig. S4).

The three-dimensional structures of three *CATs* in *Arabidopsis* were predicted using Phyre2 Server to explore the structural characteristics of *CATs*. The high coverage and 100% confidence indicated that the modeling results were robust (Table S4). The prediction results indicated that these three genes have highly similar α -helices and β -strands (Fig. 4b). The Thr-343 residue of *At-CAT1/2* possesses more atoms capable of forming hydrogen bonds than the Cys-343 residue of *At-CAT3*. The Cys-343 residue lacks the oxygen atoms needed to form hydrogen bonds; it also has a sulfur atom in the place of the carbon atom on the Thr-343 residue (Fig. 4b).

High sequence identity among angiosperm *CAT* genes

Multiple sequence alignments of *CAT* genes were performed to determine the degree of conservation and divergence in *CAT* genes among angiosperms. There was a high average pairwise identity (83.01%) ranging from 66.50% to 99.40% among angiosperm *CAT* genes (Fig. 5, Table S5). Extremely high sequence identity was observed among paralogs in individual species. For example, the paralogs of two monocots, *A. comosus* and *B. distachyon*, showed 90.04% and 86.04% sequence identity, respectively, and the paralogs of the dicots tobacco and *Arabidopsis* showed 88.28% and 89.02% sequence identity, respectively (Table S6). The highest sequence



identity among paralogs was observed for soybean (94.82%) (Table S7).

Gene duplication promoted increased expression and functional divergence in dicots

Changes in the expression of genes can lead to changes in gene function, and these changes might in turn affect the growth and development of plants [24]. RNA-seq datasets from *Arabidopsis*, tomato, potato, and pepper were downloaded and analyzed to determine the possible functions of *CAT* genes in plants. Nearly all genes were expressed in all the tissues examined (Fig. 6). *CAT* genes exhibited developmental stage and organ-specific expression patterns in each species. Clade II-type *CAT* genes, such as *At_CAT3*, *Sl_CatX3*, and *St_CatX3*, were highly expressed in all tissues (Fig. 2, Fig. 6). *At_CAT3*, *Sl_CatX3*, and *St_CatX3* were most highly expressed in

the stem, fruit, and stamen, respectively. By contrast, duplicated *CAT* genes, including *At_CAT2*, *Sl_CatX2*, *St_CatX2*, and *Ca_CatX2*, showed high expression in both source and sink organs, such as the flower and stamen. *At_CAT1*, *Sl_CatX1*, *St_CatX1*, and *Ca_CatX1* were highly expressed in certain tissues (such as mature pollen, flower, and stamen), suggesting that these *CAT* genes have functionally differentiated following gene duplication.

Discussion

Lineage-specific loss of *CAT* genes in Fabaceae

The gene number of *CAT* genes varied substantially (0–8) among algae and higher plants; however, variation in the gene number of *CAT*s was low within angiosperms (Fig. 1). Only one *CAT* member was detected in all members of Fabaceae, except for soybean, which had four *CAT*

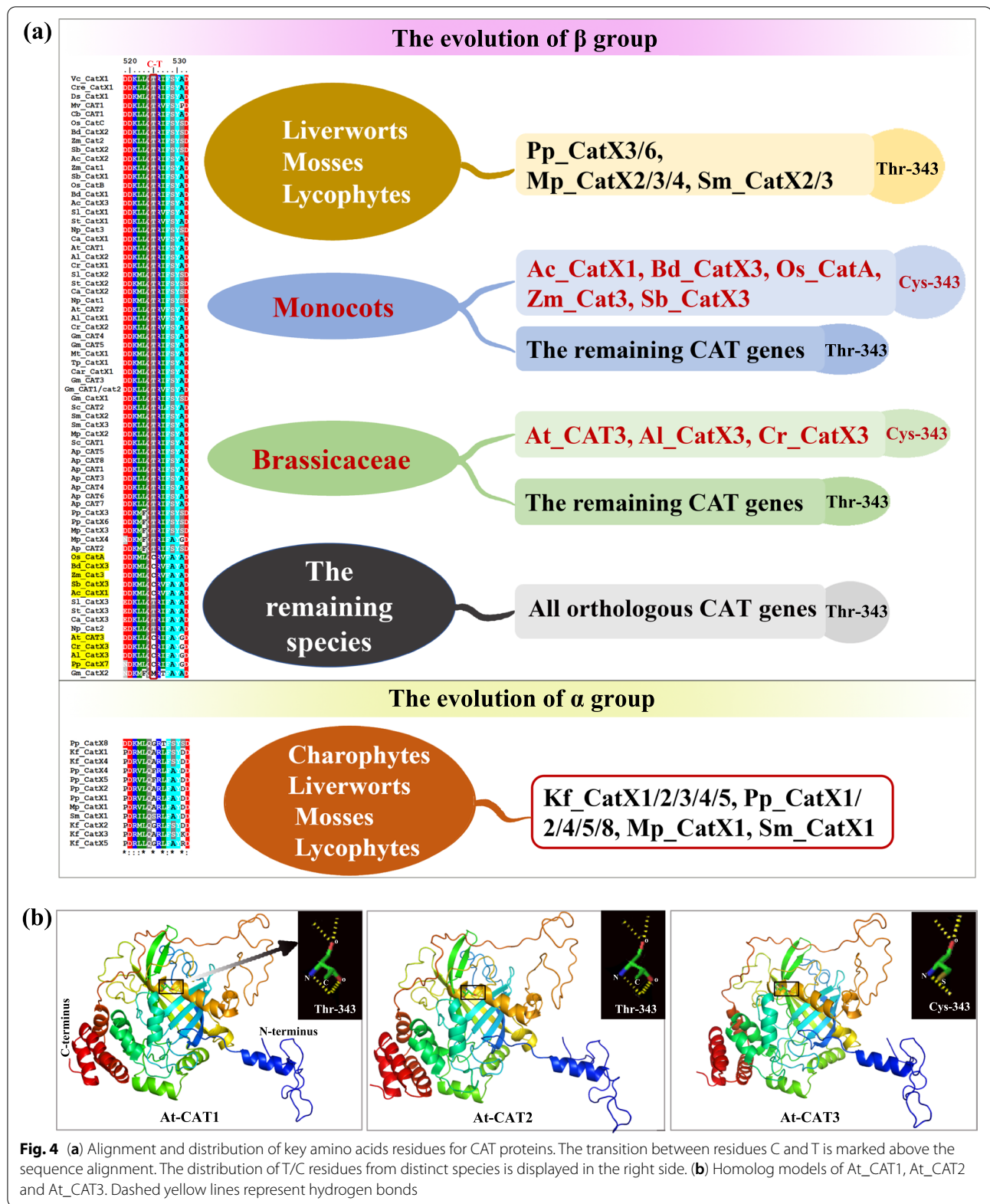


Fig. 4 (a) Alignment and distribution of key amino acids residues for CAT proteins. The transition between residues C and T is marked above the sequence alignment. The distribution of T/C residues from distinct species is displayed in the right side. (b) Homolog models of At_Cat1, At_Cat2 and At_Cat3. Dashed yellow lines represent hydrogen bonds



Fig. 5 Multiple sequence alignments for CATs in angiosperms. (a) Amino acid conservation patterns of CATs. (b) Sequence features of web logo in the CAT gene family

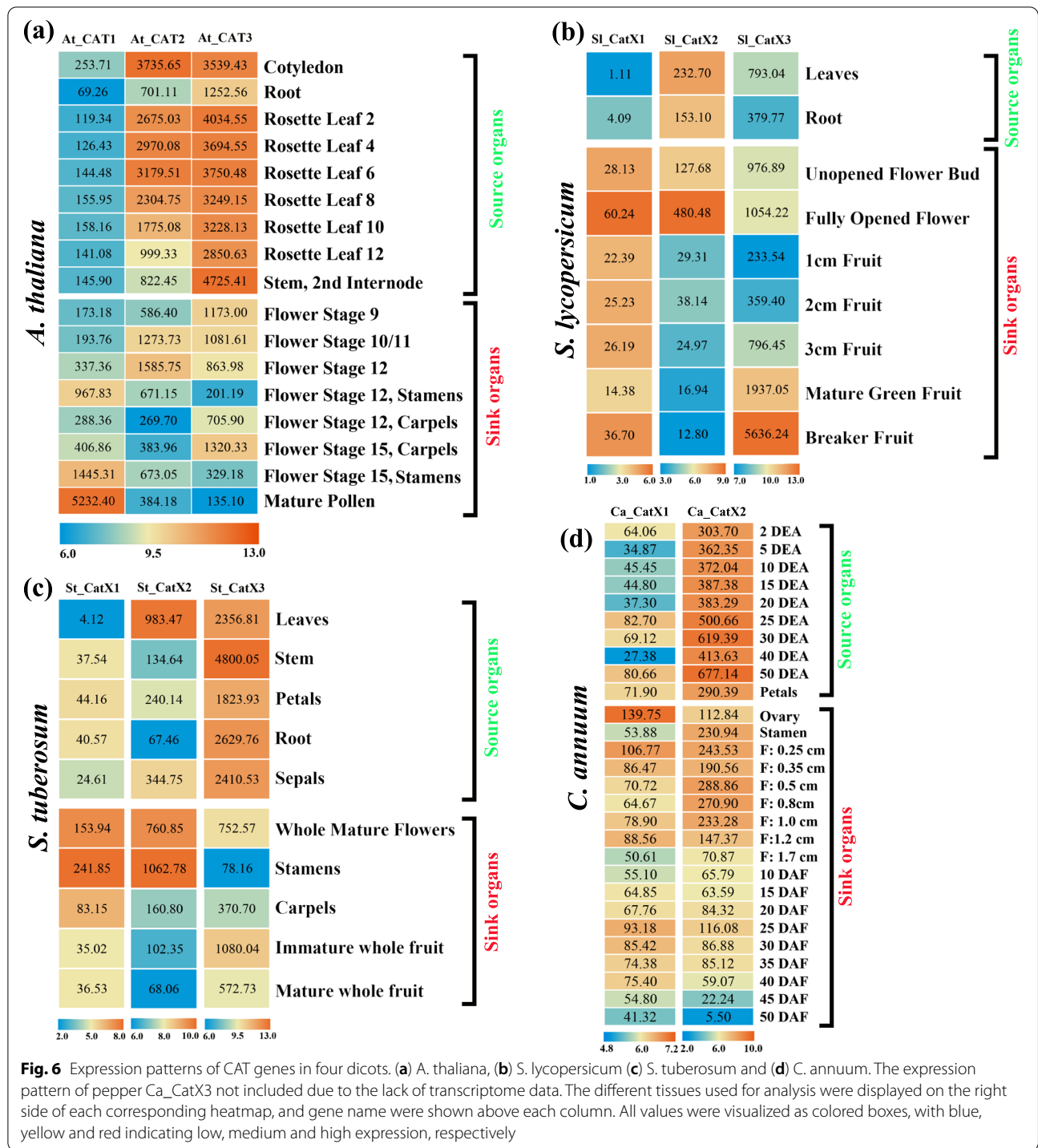


Fig. 6 Expression patterns of CAT genes in four dicots. (a) *A. thaliana*, (b) *S. lycopersicum*, (c) *S. tuberosum* and (d) *C. annuum*. The expression pattern of pepper *Ca_CatX3* not included due to the lack of transcriptome data. The different tissues used for analysis were displayed on the right side of each corresponding heatmap, and gene name were shown above each column. All values were visualized as colored boxes, with blue, yellow and red indicating low, medium and high expression, respectively

copies due to whole-genome duplication [25]. As the retention of multiple copies of the same gene has often been observed in the palaeopolyploid genome of soybean [26], the four *CAT* copies in soybean likely belong to a single gene.

We found that *Medtr1386s0010*, a *CAT* gene of *M. truncatula*, was incorrectly annotated in the NCBI database. Thus, members of Fabaceae likely only possessed one *CAT* member. Phylogenetic analysis revealed that all *CAT* genes from Fabaceae formed a Fabaceae-specific clade within group β 2 (Fig. 2). Given that both group α 1

and $\beta 1$ were the oldest clades in plants, the presence of a single *CAT* member in Fabaceae likely reflects a lineage-specific loss of *CAT* members. Gene loss might have a greater impact on organisms than most amino acid substitutions; it is thus one of the major drivers of gene family evolution, organogenesis, morphological diversity, and adaptation [27–29]. The presence of a single *CAT* gene in Fabaceae might be related to the special organogenesis and morphological characteristics of members of this group.

Single-copy genes tend to show higher expression levels and sequence identity in more tissues than non-single-copy genes in a species [28]. Consistent with this expectation, *CAT* sequences of Fabaceae were more similar (94.77%) than those of Solanaceae (87.13%) and Brassicaceae (87.33%) (Table S7).

Intron acquisition promoted the evolution of multiple *CAT* copies in most angiosperms

Gene duplication is one of the main engines driving evolutionary novelties, as it can often lead to neofunctionalization and subfunctionalization through alterations in expression levels or coding sequences [30, 31]. *CAT* genes underwent independent duplication events following the divergence of monocots and eudicots, which resulted in different evolutionary patterns between the two lineages. Multiple copies in a plant genome may be produced by a single copy of a *CAT* gene [3]. New copies of *CAT* genes were obtained through the loss of introns from intron-rich ancestral *CAT* genes in plants [11].

Our results, coupled with the findings of a previous study [3], indicated that new copies of *CAT* genes in plants can be generated through the acquisition of introns from older genes with fewer introns in the same species (Fig. 2, Fig. S2). This was a general pattern among the most angiosperms examined. Except for Fabaceae, the number of introns of group $\beta 2$ *CAT* genes (5–7) was always higher than that of group $\beta 1$ *CAT* genes (2–6) (Fig. S2). These findings suggest that the evolution of angiosperm *CAT* genes was mainly driven by the acquisition of introns. Prokaryotic *CAT* genes naturally lack introns [3], and some basal bacterial lineages possess intronless *CAT* alleles [32]; both of these findings support our interpretation. In Fabaceae, *Gm_CAT1/cat2* and *Gm_CAT3* possessed six introns, whereas the remaining *CAT* genes contained seven introns. Due to the Fabaceae *CAT* genes lacking in group $\beta 1$, its evolutionary pattern was uncertain.

The intron phases of some gene families, such as the *PDAT* [33], *rhombooid* [34], and *RNAP* [35] families, are conserved among orthologs in higher plants. Introns have a wide range of functions in contemporary species and are involved in almost every step of mRNA

processing [36]. Some highly expressed genes typically have higher intron gain rates [37]. We observed large differences in the intron phase of *CAT* genes within the same subfamily, suggesting that the *CAT* gene may affect gene expression or function between different species.

Acquisition of *CAT* genes by eukaryotic genomes via LGT

Several DNA fragments containing functional genes can be transferred from bacteria to eukaryotes, and this process is referred to as LGT [38, 39]. Generally, sequences transferred to eukaryotes retain their functionality and acquire eukaryotic characteristics [39]. A previous analysis of the *CAT* gene family has shown that several LGT events have occurred not only between bacteria and fungi but also between bacterial and protist ancestors of the green lineage [3].

Our analysis suggested that *CAT* genes from bacteria or protists were transferred to green lineage taxa by LGT events and led to the formation of two types: α and β (Fig. 3a). α -type *CAT* genes comprised all *CAT* genes from charophytes and part of the *CAT* genes from lower land plants (including *P. patens*, *M. polymorpha*, and *S. moellendorffii*). Previous studies have indicated that typical *CAT* genes comprise three clades: clade 1, 2, and 3 [2, 3]. An ancestral clade 2-type *CAT* that consists of large subunit catalases (~750 residues per subunit) gave rise to small subunit catalases (~500 residues per subunit) in clade 1 and 3 [2, 8, 40]. These three main clades of the *CAT* family are shown in Fig. 3b and were consistent with the results of previous studies. Clade 1 (referred to as β type in our study) was more closely related to clade 2 than to clade 3 (referred to as α type in our study) according to the phylogenetic tree.

Clade 3 contains a small paralog group of CATs that evolved from genes in Chlorophyte [8]. However, our findings indicate that this small group (α group) was actually derived from Charophyte, not Chlorophyte. The paralog group was present in multiple species, including all *CAT* genes in *K. flaccidum* and some of the *CAT* genes from *P. patens*, *M. polymorpha*, and *S. moellendorffii*. Thus, a small paralog group may be selectively retained by certain species. The eukaryotic genomes hold most of the genes of bacteria [41]. The catalases travel frequently laterally between the eukaryotic host and bacteria [3]. In order to adapt to the eukaryotic intracellular environment in eukaryotic host, they may eventually abandon the metabolic process of bacterial genes [41]. In the evolution of the *CAT* gene family, frequent LGT between eukaryotic hosts and bacteria may lead to the formation of two clades of *CAT* genes with different functions in plants.

The retention of clade 3-type *CAT* genes in Charophyte, *M. polymorpha*, *P. patens*, and *S. moellendorffii*

possibly facilitated adaptation to semi-aquatic environments; by contrast, the loss of clade 3-type *CAT* genes in higher plants indicates that these genes are no longer beneficial in a completely terrestrial environment. The loss of several *CAT* genes has been documented to promote adaptation in several species [42]. Nevertheless, clade 1, 2, and 3-type *CAT* genes share a common ancestor (Fig. 3b).

Changes in key amino acid residues alter the functions of *CAT* genes

A recent study has proposed that the functional features of plants can be altered when a crucial amino acid residue of *Arabidopsis* ROG1/CAT3 is changed from Cys-343 to Thr-343 or when an important amino acid residue of CAT2 is converted from Thr-343 to Cys-343 [23]. Similar observations have been made in rice OsCATA and OsCATC, which possess Cys-343 and Thr-343, respectively, and are orthologs of *Arabidopsis* At_CAT3 and At_CAT2, respectively. A previous sequence analysis of ROG1-like proteins revealed that the Cys-343 residue was distributed among 12 plants (including one in *P. patens*, seven in the grass family, and three in Brassicaceae), indicating that ROG1-like proteins are structurally and functionally conserved in plants [23]. However, our analysis showed that Cys-343 was a species-specific residue that was only present in Brassicaceae, monocots and Pp_CatX7 in *P. patens* (Fig. 4a).

Changes in specific amino acid residues in proteins have been shown to induce functional diversification in various enzymes. A pair of homologous but functionally different enzymes, the C-glycosyltransferases CGTa and CGTb, sequentially catalyzes the biosynthesis of (iso) schaftosides in plants [43]. The functions of SbCGTb and SbCGTa can be switched through structural analysis and mutagenesis of key amino acids. Based on our findings and the results of previous studies, there might be some key residues in CATs with substantial functional implications when altered that have yet to be discovered.

Conclusion

A total of 82 *CAT*s were identified from 29 species and could be clustered into three groups. The transfer of plant *CAT* genes from bacteria to core chlorophytes and charophytes by LGT led to the independent evolution of two types of *CAT* genes: α and β . The newer *CAT* genes were produced through intron acquisition in higher plants, and the Cys-343 residue was only present in monocots and Brassicaceae. Furthermore, all *CAT* genes show high overall sequence identity that individual *CAT* genes showed developmental stage and organ-specific expression patterns.

Materials and methods

Data retrieval and identification of *CAT* homologs in plants

All recognizable *CAT* genes were obtained via three steps. First, a search was conducted on the NCBI using “catalase” as a keyword, and the *Arabidopsis* *CAT* sequence (GenBank: CAA45564.1) was obtained. Second, the Hidden Markov Model (HMM) profiles of the conserved *CAT* domains PF00199 and PF06628 in the Pfam database [44] were downloaded and used as query sequences with an e-value threshold of $<10^{-5}$. Third, the sequences obtained in the second step were used to retrieve homologous proteins; related sequences from the *Ostreococcus lucimarinus* [45], *Micromonas pusilla* [46], *Volvox carteri* [47], *Chlamydomonas reinhardtii* [48], *Dunaliella salina* [49], *Mesostigma viride* [50], *Physcomitrella patens* [51], *Marchantia polymorpha* [52], *Selaginella moellendorffii* [53], *Ananas comosus* [54], *Brachypodium distachyon* [55], *Oryza sativa* [56], *Zea mays* [57], *Sorghum bicolor* [58], *Solanum lycopersicum* [59], *Solanum tuberosum* [60], *Capsella rubella* [61], *Arabidopsis lyrata* [62], *Arabidopsis thaliana* [63], *Glycine max* [64], *Cicer arietinum* [65], *Trifolium pratense* [66], and *Medicago truncatula* [67] genomes were searched using the Phytozome 13 website (<https://phytozome-next.jgi.doe.gov/>). *Klebsormidium flaccidum* [68], *Gnetum montanum* [69], and *Capsicum annuum* [70] sequences were downloaded from the *Klebsormidium* genome project (http://www.plantmorphogenesis.bio.titech.ac.jp/~algae_genome_project/klebsormidium/index.html), the Dryad database (<https://datadryad.org/search?utf8=%E2%9C%93&q=gnetophytes>), and the Sol Genomics Network (<https://solgenomics.net/>), respectively. *Nicotiana plumbaginifolia* [71] sequences were obtained from the NCBI database. *Aspergillus nidulans* [72], *Saccharomyces cerevisiae* [73], *Escherichia coli* [74], *Bacillus subtilis* [75], *Chara braunii* [76], and *Dictyostelium discoideum* [77] sequences were obtained from the Ensembl Genomes database (<https://ensemblgenomes.org/>). The protein sequences of *Anthoceros punctatus* [78] were downloaded from the Hornwort genomes (<https://www.hornworts.uzh.ch/en.html>). The *CAT* sequences of *Azolla filiculoides* and *Salvinia cucullata* [79] were searched from the FernBase genome database (<https://www.fernbase.org/>). Finally, all sequences were submitted to the Pfam database to verify the presence of conserved domains.

Sequence annotation and genome size

Sequences that were too long or short or affected by obvious errors in the genome assembly were reannotated using Softberry (FGENESH-HMM-based gene structure prediction (<http://www.softberry.com>)) [80].

Genome sizes of *M. pusilla*, *O. lucimarinus*, *D. salina*, *C. reinhardtii*, *V. carteri*, *P. patens*, *M. polymorpha*, and *S. moellendorffii* were obtained using the Phytozome 13 database. The size of the *K. flaccidum* genome was obtained from a previous study [81]. The genome sizes of the remaining plants were obtained from the Published Plante Genomes website (https://plabipd.de/plant_genomes_pa.ep).

Species tree and gene tree construction

Data for each species from the evolutionary TimeTree of life (<http://timetree.org/about>) were used to construct species and gene trees [82]. Trees of *CAT* sequences were built using the maximum likelihood (ML) and neighbor-joining (NJ) methods. MEGA5, MEGA-X software and the online tool LIRMM (<http://www.phylogeny.fr/index.cgi>) were used to construct phylogenetic trees [83, 84]. The newly produced species and gene trees were displayed using the Interactive Tree of Life online tool (<https://itol.embl.de/>).

Sequence alignment and tertiary structure prediction

Multiple sequence alignments for *CATs* were conducted using ClustalX and BioEdit to identify key amino acid residues. The online Phyre2 Server (<http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index>) was used to predict the tertiary structure of *CAT* proteins [85]. The generated protein models were visualized in cartoon mode using the PyMOL tool.

Tissue-specific expression of *CAT* genes

Expression data of *CAT* genes in different tissues of *Arabidopsis*, tomato, and potato were downloaded from the Bio-Analytic Resource for Plant Biology website (<http://bar.utoronto.ca/>). Previously published RNA sequencing (RNA-seq) data [86] were used to analyze the expression profiles of *CAT* genes in pepper (the elite *Capsicum* line 6421). Expression profiles were determined in the following tissues: leaves at 2, 5, 10, 15, 20, 25, 30, 40, and 50 days after emergence; floral buds at seven different stages (0.25, 0.35, 0.5, 0.8, 1.0, 1.2, and 1.7 cm); petals, stamens, and ovaries with stigmas in fully blossomed flowers; and fruits on 10, 15, 20, 25, 30, 35, 40, 45, and 50 days after flowering. All data were normalized ($\log_2(\text{FPKM} + 1)$), and heat maps were built using TBtools [87].

Abbreviations

CATs: Catalases; ROS: Reactive oxygen species; GNV: Gene number variation; LGT: Lateral gene transfer; MYA: Million years ago.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-022-08621-6>.

Additional file 1: Figure S1. Phylogenetic relationship of *CATs* in 29 green plants. *Synechocystis* sp. PCC 6803 as an outgroup.

Additional file 2: Figure S2. Exon-intron structure of the *CAT* gene family among different species. Exon-intron structure of some species are not shown due to lacked related sequences. In order to visually show the results, intron phases of several genes are manually simulated and framed with dashed lines.

Additional file 3: Figure S3. The motif arrangement of *CATs* in 29 green plants.

Additional file 4: Figure S4. Amino acid alignment of *CATs* from 29 green plants. The critical amino acid residues (Thr343/Cys-343) are framed with black line.

Additional file 5.

Additional file 6: Table S1. The detail information of *CAT* genes in each species.

Additional file 7: Table S2. The detail information of reannotated *CAT* proteins.

Additional file 8: Table S3. The details information of bacteria, protists, fungi and animals

Additional file 9: Table S4. The 3D structure prediction for three *CATs* in *Arabidopsis*.

Additional file 10: Table S5. Homology matrix of 43 *CAT* sequences in angiosperms.

Additional file 11: Table S6. The identity level of *CATs* in specific higher plants.

Additional file 12: Table S7. The sequence identity of *CATs* in Solanaceae, Brassicaceae and Fabaceae.

Acknowledgements

We thank TopEdit (www.topedit.com) for linguistic assistance during the preparation of this manuscript.

Authors' contributions

X.Z., and F.L. conceived and supervised the study; L.P., Y.L., and X.L. performed the experiments and analyzed the data; J. W., B. Q., and H.Y. assisted with bioinformatic analysis; L.P. edited the manuscript; X.H., X.Z., and F.L. revised the final version of the paper. All authors have read and approved the final manuscript.

Funding

This work was supported by the National Key and Development Program of China (Grant No. 2019YFD1000300).

Availability of data and materials

All methods using plant material were carried out in accordance with relevant guidelines and regulations in this paper. The data used and/or analyzed during the current study are obtained from the Phytozome 13 website (<https://phytozome-next.jgi.doe.gov/>); the Ensembl Genomes database (<https://ensemblgenomes.org/>); the Softberry (FGENESH-HMM-based gene structure prediction (<http://www.softberry.com>)); the Published Plante Genomes website (https://plabipd.de/plant_genomes_pa.ep); TimeTree of life (<http://timetree.org/about>); the Interactive Tree of Life online tool (<https://itol.embl.de/>); the Phyre2 Server (<http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index>); and the Bio-Analytic Resource for Plant Biology website (<http://bar.utoronto.ca/>).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors have no conflict of interest to declare.

Author details

¹College of Horticulture, Nanjing Agricultural University, Nanjing, China. ²College of Horticulture, Hunan Agricultural University, Changsha, China. ³ERC for Germplasm Innovation and New Variety Breeding of Horticultural Crops, Changsha, China. ⁴Key Laboratory for Vegetable Biology of Hunan Province, Changsha, China. ⁵Longping Branch, Graduate School of Hunan University, Changsha, China.

Received: 24 January 2022 Accepted: 9 May 2022

Published online: 01 June 2022

References

- Lenton TM. The coupled evolution of life and atmospheric oxygen. *Evol Planet Earth*. 2003;3:35–53.
- Zamocky M, Furtmueller PG, Obinger C. Evolution of catalases from bacteria to humans. *Antioxid Redox Signal*. 2008;10(9):1527–47.
- Klotz MG, Loewen PC. The molecular evolution of catalytic hydroperoxidases: Evidence for multiple lateral transfer of genes between prokaryota and from bacteria into Eukaryota. *Mol Biol Evol*. 2003;20(7):1098–112.
- Ballal A, Chakravarty D, Bihani SC, Banerjee M. Gazing into the remarkable world of non-heme catalases through the window of the cyanobacterial Mn-catalase “KatB.” *Free Radic Biol Med*. 2020;20(160):480–7.
- Chakravarty D, Bihani SC, Banerjee M, Kalwani P, Ballal A. Unique functional insights into the antioxidant response of the cyanobacterial Mn-catalase (KatB). *Free Radic Biol Med*. 2022;179:266–76.
- Sharma I, Ahmad P. Chapter 4 - Catalase: A Versatile Antioxidant in Plants: Oxidative Damage to Plants: Antioxidant Networks and Signaling. 2014;4:131–48.
- Whittaker JW. Non-heme manganese catalase-The “other” catalase. *Arch Biochem Biophys*. 2012;525(2):111–20.
- Zámocky M, Gasselhuber B, Furtmüller P, Obinger C. Molecular evolution of hydrogen peroxide degrading enzymes. *Arch Biochem Biophys*. 2012;525(2):131–44.
- Mullen RT, Lee MS, Trelease RN. Identification of the peroxisomal targeting signal for cottonseed catalase. *Plant J*. 1997;12(2):313–22.
- Mhamdi A, Noctor G, Baker A. Plant catalases: Peroxisomal redox guardians. *Arch Biochem Biophys*. 2012;525(2):181–94.
- Frugoli JA, McPeck MA, Thomas TL, McClung CR. Intron loss and gain during evolution of the catalase gene family in Angiosperms. *Genet*. 1998;149(1):355–65.
- Lee SH, An CS. Differential expression of three catalase genes in hot pepper (*Capsicum annuum* L.). *Mol Cells*. 2005;20(2):247–55.
- Lin A, Wang Y, Tang J, Xue P, Li C, Liu L, Hu B, Yang F, Loake GJ, Chu C. Nitric oxide and protein S-nitrosylation are integral to hydrogen peroxide-induced leaf cell death in rice. *Plant Physiol*. 2011;158(1):451–64.
- Acevedo A, Scandalios JG. Expression of the catalase and superoxide dismutase genes in mature pollen in maize. *Theor Appl Genet*. 1990;80(5):705–11.
- Willekens H, Langebartels C, Tiré C, Van Montagu M, Inzé D, Van Camp W. Differential expression of catalase genes in *Nicotiana glauca* (L.). *Proc Natl Acad Sci U S A*. 1994;91(22):10450–4.
- Jiao Z, Wang J, Tian Y, Wang S, Sun X, Li S, Ma W, Zhou T, Fan Z. Maize catalases localized in peroxisomes support the replication of maize chlorotic mottle virus. *Phytopathology Res*. 2021;3(1):17 Article No.: 17.
- Jiao Z, Tian Y, Cao Y, Wang J, Zhan B, Zhao Z, Sun B, Guo C, Ma W, Liao Z, Zhang H, Zhou T, Xia Y, Fan Z. A novel pathogenicity determinant hijacks maize catalase 1 to enhance viral multiplication and infection. *New Phytol*. 2021;230(3):1126–41.
- Mhamdi A, Queval G, Chaouch S, Vanderauwera S, Van Breusegem F, Noctor G. Catalase function in plants: a focus on Arabidopsis mutants as stress-mimic models. *J Exp Bot*. 2010;61(15):4197–220.
- Zou JJ, Li XD, Ratnasekera D, Wang C, Liu WX, Song LF, Zhang WZ, Wu WH. Arabidopsis CALCIUM-DEPENDENT PROTEIN KINASE8 and CATALASE3 function in abscisic acid-mediated signaling and H₂O₂ homeostasis in stomatal guard cells under drought stress (vol 27, pg 1445, 2015). *Plant Cell*. 2016;28(4):990–1.
- Du YY, Wang PC, Chen J, Song CP. Comprehensive functional analysis of the catalase gene family in *Arabidopsis thaliana*. *J Integr Plant Biol*. 2008;50(10):1318–26.
- Su T, Wang P, Li H, Zhao Y, Lu Y, Dai P, Ren T, Wang X, Li X, Shao Q, Zhao D, Zhao Y, Ma C. The Arabidopsis catalase triple mutant reveals important roles of catalases and peroxisome-derived signaling in plant development. *J Integr Plant Biol*. 2018;60(7):591–607.
- Stepanov SS, Zolotareva EK, Belyavskaya NA. The role of catalase in assimilation of exogenous methanol by *Chlamydomonas reinhardtii* cells. *J Appl Phycol*. 2020;32:1053–62.
- Chen L, Wu R, Feng J, Feng T, Wang C, Hu J, Zhan N, Li Y, Ma X, Ren B, Zhang J, Song CP, Li J, Zhou JM, Zuo J. Transnitrosylation Mediated by the Non-canonical Catalase ROG1 Regulates Nitric Oxide Signaling in Plants *Dev Cell*. 2020;53(4):444–457.e445.
- Ng M, Yanofsky MF. Function and evolution of the plant MADS-box gene family. *Nat Rev Genet*. 2001;2(3):186–95.
- Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, et al. Genome sequence of the palaeopolyploid soybean. *Nat*. 2010;463(7278):178–83.
- Tate JA, Joshi P, Soltis KA, Soltis PS, Soltis DE. On the road to diploidization? Homoeolog loss in independently formed populations of the allopolyploid *Tragopogon miscellus* (Asteraceae). *BMC Plant Bio*. 2009;9(1):80.
- Aravind L, Watanabe Hidemi. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc Natl Acad Sci U S A*. 2009;97(21):11319–24.
- De Smet R, Adams KL, Vandepoele K, van Montagu MCE, Maere S, Van de Peer Y. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci U S A*. 2013;110(8):2898–903.
- Smith SD, Rausher MD. Gene loss and parallel evolution contribute to species difference in flower color. *Mol Biol Evol*. 2011;28(10):2799–810.
- Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D, Liberles DA. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *J Exp Zool B Mol Dev Evol*. 2007;308(1):58–73.
- Wang Y, Wang X, Paterson AH. Genome and gene duplications and gene expression divergence: a view from plants. *Ann NY Acad Sci*. 2012;1256:1–14.
- Hausner G, Hafez M, Edgell DR. Bacterial group I introns: mobile RNA catalysts. *Mob DNA*. 2014;5(1):8.
- Pan X, Peng FY, Weselake RJ. Genome-wide analysis of PHOSPHOLIPID:DIACYLGLYCEROL ACYLTRANSFERASE (PDAT) genes in plants reveals the eudicot-wide PDAT gene expansion and altered selective pressures acting on the core eudicot PDAT paralogs. *Plant Physiol*. 2015;167(3):887–904.
- Li Q, Zhang N, Zhang L, Ma H. Differential evolution of members of the rhomboid gene family with conservative and divergent patterns. *New Phytol*. 2015;206(1):368–80.
- Wang Y, Ma H. Step-wise and lineage-specific diversification of plant RNA polymerase genes and origin of the largest plant-specific subunits. *New Phytol*. 2015;207(4):1198–212.
- Chorev M, Carmel L. The function of introns *Front Genet*. 2012;3:55.
- Carmel L, Rogozin IB, Wolf YI, Koonin EV. Evolutionarily conserved genes preferentially accumulate introns. *Genome Res*. 2007;17(7):1045–50.
- Lacroix B, Citovsky V. Transfer of DNA from Bacteria to Eukaryotes. *mBio*. 2016;7(4):e00863–16.
- Husnik F, McCutcheon JP. Functional horizontal gene transfer from bacteria to eukaryotes. *Nat Rev Microbiol*. 2018;16(2):67–79.
- Klotz MG, Klassen GR, Loewen PC. Phylogenetic relationships among prokaryotic and eukaryotic catalases. *Mol Biol Evol*. 1997;14(9):951–8.
- Julia B, Martin WF. Bacterial genes outnumber archaeal genes in eukaryotic genomes. *Genome Biol Evol*. 2020;4:4.
- Cavaliere-Smith T. The neomuran origin of archaeobacteria, the negibacterial root of the universal tree and bacterial mega-classification. *Int J Syst Evol Microbiol*. 2002;52(Pt 1):7–76.
- Wang ZL, Gao HM, Wang S, Zhang M, Chen K, Zhang YQ, Wang HD, Han BY, Xu LL, Song TQ, Yun CH, Qiao X, Ye M. Dissection of the general

- two-step di-C-glycosylation pathway for the biosynthesis of (iso)schaftosides in higher plants. *Proc Natl Acad Sci U S A*. 2020;117(48):30816–23.
44. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. Pfam: The protein families database in 2021. *Nucleic Acids Res*. 2021;49(D1):D412–d419.
 45. Palenik B, Grimwood J, Aerts A, et al. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proc Natl Acad Sci U S A*. 2007;104(18):7705–10.
 46. Worden AZ, Lee JH, Mock T, et al. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science*. 2009;324(5924):268–72.
 47. Prochnik SE, Umen J, Nedelcu AM, et al. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carterii*. *Science*. 2010;329(5988):223–6.
 48. Merchant SS, Prochnik SE, Vallon O, et al. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science*. 2007;318(5848):245–50.
 49. Polle JEW, Barry K, Cushman J, et al. Draft Nuclear Genome Sequence of the Halophilic and Beta-Carotene-Accumulating Green Alga *Dunaliella salina* Strain CCAP19/18. *Genome Announc*. 2017;5(43):e01105–e1117.
 50. Wang S, Li L, Li H, et al. Genomes of early-diverging streptophyte algae shed light on plant terrestrialization. *Nat Plants*. 2020;6(2):95–106.
 51. Lang D, Ullrich KK, Murat F, et al. The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution. *Plant J*. 2018;93(3):515–33.
 52. Bowman JL, Kohchi T, Yamato KT, et al. Insights into Land Plant Evolution Garnered from the *Marchantia polymorpha* Genome. *Cell*. 2017;171(2):287–304.e15.
 53. Banks JA, Nishiyama T, Hasebe M. The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science*. 2011;332(6032):960–3.
 54. Ming R, VanBuren R, Wai CM, Tang H, et al. The pineapple genome and the evolution of CAM photosynthesis. *Nat Genet*. 2015;47(12):1435–42.
 55. Mockler TC, Schmutz J, Rokhsar D, et al. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*. 2010;463(7282):763.
 56. Goff Stephen A. A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *japonica*). *Science*. 2002;296(5565):92–100.
 57. Springer NM, Anderson SN, Andorf CM, Ahern KR, Bai F, Barad O, Barbazuk WB et al. The maize W22 genome provides a foundation for functional genomics and transposon biology. *Nat Genet*. 2018;50(9):1282–88.
 58. Deschamps S, Zhang Y, Llaca V, Ye L, Sanyal A, King M, May G, Lin H. A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nat Commun*. 2018;9(1):4844.
 59. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*. 2012;485(7400):635–41.
 60. Pham GM, Hamilton JP, Wood JC, Burke JT, Zhao H, Vaillancourt B, Ou S, Jiang J, Buell CR. Construction of a chromosome-scale long-read reference genome assembly for potato. *Gigascience*. 2020;9(9):giaa100.
 61. Slotte T, Hazzouri KM, Ågren JA, et al. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat Genet*. 2013;45(7):831–5.
 62. Akama S, Shimizu-Inatsugi R, Shimizu KK, Sese J. Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid *Arabidopsis*. *Nucleic Acids Res*. 2014;42(6): e46.
 63. Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *Nat Commun*. 2018;9(1):541.
 64. Xie M, Chung CY, Li MW, et al. A reference-grade wild soybean genome. *Nat Commun*. 2019;10(1):1216.
 65. Parveen S, Nawaz K, Roy R, Pole AK, Venkata Suresh B, Misra G, Jain M, Yadav G, Parida SK, Tyagi AK, Bhatia S, Chattopadhyay D. An advanced draft genome assembly of a desi type chickpea (*Cicer arietinum* L.). *Sci Rep*. 2015;5:12806.
 66. De Vega JJ, Ayling S, Hegarty M, et al. Red clover (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. *Sci Rep*. 2015;5:17394.
 67. Young ND, DeBellé F, Oldroyd GE, et al. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature*. 2011;480(7378):520–4.
 68. Hori K, Maruyama F, Fujisawa T, et al. *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nat Commun*. 2014;5:3978.
 69. Wan T, Liu ZM, Li LF, et al. A genome for gnetophytes and early evolution of seed plants. *Nat Plants*. 2018;4(2):82–9.
 70. Kim S, Park M, Yeom SI, et al. Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nat Genet*. 2014;46(3):270–8.
 71. Xu S, Brockmüller T, Navarro-Quezada A, Kuhl H, Gase K, Ling Z, Zhou W, Kreitzer C, Stanke M, Tang H, Lyons E, Pandey P, Pandey SP, Timmermann B, Gaquerel E, Baldwin IT. Wild tobacco genomes reveal the evolution of nicotine biosynthesis. *Proc Natl Acad Sci U S A*. 2017;114(23):6133–8.
 72. Galagan JE, Calvo SE, Cuomo C, et al. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. Nature*. 2005;438(7071):1105–15.
 73. Engel SR, Dietrich FS, Fisk DG, et al. The reference genome sequence of *Saccharomyces cerevisiae*: then and now. *G3 (Bethesda)*. 2014;4(3):389–98.
 74. Hayashi K, Morooka N, Yamamoto Y, Fujita K, Isono K, Choi S, Ohtsubo E, Baba T, Wanner BL, Mori H, Horiuchi T. Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. *Mol Syst Biol*. 2006;2006(2):0007.
 75. Schroeder JW, Simmons LA. Complete Genome Sequence of *Bacillus subtilis* Strain PY79. *Genome Announc*. 2013;1(6):e01085–e1113.
 76. Nishiyama T, Sakayama H, de Vries J, et al. The *Chara* Genome: Secondary Complexity and Implications for Plant Terrestrialization. *Cell*. 2018;174(2):448–464.e24.
 77. Eichinger L, Pachebat JA, Glöckner G, et al. The genome of the social amoeba *Dictyostelium discoideum*. *Nature*. 2005;435(7038):43–57.
 78. Li FW, Nishiyama T, Waller M, et al. *Anthoceros* genomes illuminate the origin of land plants and the unique biology of hornworts. *Nat Plants*. 2020;6(3):259–72.
 79. Li FW, Brouwer P, Carretero-Paulet L, et al. Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nat Plants*. 2018;4(7):460–72.
 80. Solovyev V, Kosarev P, Seledsov I, Vorobyev D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol*. 2006;7 Suppl 1(Suppl 1):S10.11–12.
 81. Hori K, Maruyama F, Fujisawa T, Togashi T, Yamamoto N, Seo M, et al. *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nat Commun*. 2014;5:3978.
 82. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Time-lines, Timetrees, and Divergence Times. *Mol Biol Evol*. 2017;34(7):1812–9.
 83. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*. 2011;28(10):2731–9.
 84. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res*. 2008;36(Web Server issue):W465–469.
 85. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc*. 2015;10(6):845–58.
 86. Liu F, Yu H, Deng Y, Zheng J, Liu M, Ou L, et al. PepperHub, an Informatics Hub for the Chili Pepper Research Community. *Mol Plant*. 2017;10(8):1129–32.
 87. Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R. TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol Plant*. 2020;13(8):1194–202.

Publisher's Note

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