Deep evaluation of the evolutionary history of the Heat Shock Factor (HSF) gene family and its expansion pattern in seed plants

Yiying Liao[1], Zhiming Liu[2], Andrew W. Gichira[2], Min Yang[1], Ruth Wambui Mbichi[2], Linping Meng[1] and Tao Wan[2,3]

[1] Key Laboratory of Southern Subtropical Plant Diversity, Fairy Lake Botanical Garden, Shenzhen & Chinese Academy of Science, Shenzhen, China

[2] Core Botanical Gardens/Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, China

[3] Sino-Africa Joint Research Centre, Chinese Academy of Sciences, Wuhan, China

ABSTRACT

Heat shock factor (HSF) genes are essential in some of the basic developmental pathways in plants. Despite extensive studies on the structure, functional diversification, and evolution of HSF genes, their divergence history and gene duplication pattern remain unknown. To further illustrate the probable divergence patterns in these subfamilies, we analyzed the evolutionary history of HSF genes using phylogenetic reconstruction and genomic syntenic analyses, taking advantage of the increased sampling of genomic data from pteridophytes, gymnosperms and basal angiosperms. We identified a novel clade that includes HSFA2, HSFA6, HSFA7, and HSFA9 with a complex relationship, which is very likely due to orthologous or paralogous genes retained after frequent gene duplication events. We hypothesized that HSFA9 derives from HSFA2 through gene duplication in eudicots at the ancestral state, and then expanded in a lineage-specific way. Our findings indicate that HSFB3 and HSFB5 emerged before the divergence of ancestral angiosperms, but were lost in the most recent common ancestors of monocots. We also presumed that HSFC2 derives from HSFC1 in ancestral monocots. This work proposes that during the radiation of flowering plants, an era during which there was a differentiation of angiosperms, the size of the HSF gene family was also being adjusted with considerable sub- or neo-functionalization. The independent evolution of HSFs in eudicots and monocots, including lineage-specific gene duplication, gave rise to a new gene in ancestral eudicots and monocots, and lineage-specific gene loss in ancestral monocots. Our analyses provide essential insights for studying the evolutionary history of this multigene family.

Subjects Bioinformatics, Genetics, Genomics, Molecular Biology, Plant Science

Keywords Heat shock factor, Gene family evolution, Diversification, Lineage-specific expansions, Whole genome duplication

Submitted 7 January 2021
Accepted 26 May 2022
Published 9 August 2022

Corresponding author
Tao Wan, 87418331@qq.com

Academic editor
Adriana Basile

Additional Information and Declarations can be found on page 16

DOI **10.7717/peerj.13603**

INTRODUCTION

Heat shock factors (HSFs) play an important role in improving the thermotolerance of plants. They function as the central regulators of heat shock protein expression and other heat shock-induced gene expression. HSFs are the direct transcriptional activators of

genes regulated by thermal stress. They encode heat shock proteins to protect cells against proteotoxic damage under heat stress (HS; *Hu, Hu & Han, 2009*; *Ahuja et al., 2010*; *Ohama et al., 2017*). HSFs have been identified in most eukaryotes and non-plant organisms. HSFs also participate in the growth and development of cells (*Åkerfelt, Morimoto & Sistonen, 2010*; *Scharf et al., 2012*). HSFs have been widely studied in plants, especially angiosperms, and have been found to be critical under various environmental stressors (*Scharf et al., 2012*). The number of HSF genes varies widely among plants with green algae having just one or two HSF genes and angiosperms having more than 50 (*Wang et al., 2018*). HSFs generally contain the DNA binding domain (DBD), the oligomerization domain (OD), the nuclear localization signal (NES), and the C-transcriptional activation domain (CTD) (*Scharf et al., 2012*; *Guo et al., 2016*). Based on the topology of these domains, HSFs are normally classified into three groups: HSFA, HSFB and HSFC. These three groups of HSFs are further divided into 16 subfamilies which are distinguished in angiosperms, including the HSFA group (A1-A9), HSFB group (B1-B5) and HSFC group (C1-C2) (*Nover et al., 2001*; *Hu, Hu & Han, 2009*; *Scharf et al., 2012*; *Qiao et al., 2015*). HSFC was identified for the first time in the first overview of HSFs, presented in *Arabidopsis thaliana* by *Nover et al. (2001)*. Many valuable summaries followed, including compiled data from nine angiosperm species, and over 50 plant species showing the structure, function and evolution of HSFs (*Scharf et al., 2012*; *Wang et al., 2018*). These reports point out that HSF family members and their functions are greatly diverged among higher plant lineages in response to environmental stressors. However, the evolutionary relationships among these subfamilies are still unknown; some of the deepest nodes of the HSF phylogeny tree, such as the positions of HSFB5 and HSFA9, also remain unclear. In previous studies, HSFB5 was either placed with HSFA5 or other HSFB members, and HSFA9 may be clustered with HSFA2 or HSFA7 (*Scharf et al., 2012*; *Wang et al., 2018*). This is likely due to the limited HSF data available in representative seed plant lineages including gymnosperms and basal angiosperms. It is also partially attributed to the unpredictable gene copy turnover after recurring gene duplication events at tandem or genome-wide level.

In this study, we expanded the data collection to basal angiosperms, gymnosperms, and pteridophytes to reconstruct the diversification history of HSFs during seed plant evolution. We also detected the syntenic relationships of HSFs across a wide range of species, thus providing crucial information to address fundamental questions on their evolutionary history. We then estimated the divergence time of the derived genes from their ancestors based on a reliable gene orthology. Our results present critical evidence that help to explain the expansion of HSF subfamilies in seed plant lineages.

METHODS

Identification of HSFs and phylogenetic analysis

Here, we sampled 23 species representing three main taxa for pteridophytes, gymnosperms and basal angiosperms including seven genomes and 17 transcriptomes, (Table S1). Most of the transcriptome data was obtained from the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov) (Data from *Ran*

et al., 2018). Multiple databases were screened for the genome assemblies including: ConGenIE (http://congenie.org/), GigaDB (http://gigadb.org/dataset/100209), Dryad (https://datadryad.org/stash/dataset/), WaterlilyPond (http://eplant.njau.edu.cn/waterlily/), FernBase (https://www.fernbase.org/), and the *Liriodendron chinense* database (http://120.78.193.56:8000/). To increase the reliability of the data, we analyzed both the genomes and transcriptomes of *Ginkgo biloba* in this study. The methods of our RNA-seq dataset analysis were drawn from a study by *Ran et al. (2018)*. We used the predicted proteome of each genome as a query to search for HSF-type DBD domains (HSF_DNA_bind_PF00447) from Pfam-A.hmm (Pfam release 32.0) using PfamScan software (https://www.ebi.ac.uk/Tools/pfa/pfamscan/), which were considered as candidate genes. We then extracted the amino acid sequences of the HSFs. We also downloaded 537 HSF sequences extracted from 23 plant species (Table S2) representing seven main taxa in the Heatster database (http://www.cibiv.at/services/hsf) and used them in BLAST searches for analyzed species to further identify candidate HSF proteins. For those candidate sequences, we examined the facticity of DBDs and ODs using the SMART 7 software (*Letunic, Doerks & Bork, 2012*) (http://smart.embl-heidelberg.de/) and the HEATSTER website (https://applbio.biologie.uni-frankfurt.de/hsf/heatster/). The candidate proteins without an integrated DBD domain or HR-A/B domain were removed.

For the phylogenetic reconstruction in this study, we used MUSCLE (http://www.drive5.com/muscle) to conduct the alignment of the candidate genes. Phylogenetic trees were generated using both the NJ and the ML methods. The NJ tree was constructed by TreeBeST (version 1.9.2, http://treesoft.sourceforge.net/treebest.shtml, parameters: -t mm –b 100). Approximately maximum-likelihood (ML) phylogenetic trees were constructed using FastTree (version 2.1.11, http://www.microbesonline.org/fasttree/treecmp.html, with default parameters). Then, phylogenetic analyses were conducted using RaxML version 8.0.19 (*Stamatakis, 2014*) with 100 bootstraps, the PROTGAMMAAUTO model, and maximum likelihood reconstruction using rapid hill-climbing and rapid bootstrap analyses (-f ad). Phylogenetic trees were examined and manipulated with Evolview v2 (*He et al., 2016*). We classified the HSF subfamilies using both the phylogenetic tree and the annotation from the HEATSTER website. Some results from the HEATSTER website were inconsistent with the phylogenetic tree, so we performed follow-up checks to confirm the subfamily classification. The final results were based on the appearance of domain characteristic motifs.

We used all HSFA, HSFB, and HFSC genes identified for phylogenetic tree reconstruction in order to better understand the evolutionary relationship within subfamilies and for in depth phylogenetic analyses of the HSFB clade and HSFA-HSFC clade. To understand the complicated evolutionary relationship of the HSFA2, HSFA6, HSFA7, and HSFA9 clades of subfamilies and the HSFC clade, we extracted those two group genes for phylogenetic tree reconstruction, with *Chlamydomonas reinhardtii* used as an outgroup. Our methods for protein sequence alignments and phylogenetic analyses followed the same steps as previously outlined.

Synteny analysis and molecular dating analyses

We used MCScanX (*Wang et al., 2012*) to detect the gene replication events and included a total of 21 plant genomes in a synteny analysis covering green algae, mosses, ferns, gymnosperms, basal angiosperms and angiosperms (Table S3). We analyzed all protein models from these genomes for all possible intra- and inter-species genome-wide comparisons and downloaded all genome annotation and corresponding protein sequences for those species. Homologous genes are classified as either orthologous in different species if they are separated by a speciation event, or paralogous in the same species if they are separated by a gene duplication event. We identified the paralogous and orthologous genes in or between those genomes through synteny detection using MCScanX with default parameters (minimum match size for a collinear block = five genes, max gaps allowed = 25 genes). The output files from all the intra- and inter-species comparisons were integrated into a single file named "Total_Synteny_Blocks", including the headers "Block_Index", "Locus_1", "Locus_2", and "Block_Score", which served as the database file. We performed the all-against-all protein sequence comparisons necessary for MCScanX using DIAMOND v 0.8.25 (*Buchfink, Xie & Huson, 2015*). The gene list containing all candidate HSF genes was queried against the "Total_Synteny_Blocks" file. We used these results to identify whether or not HSF genes exist in a syntenic block. We chose eight representative species for gymnosperms (*Gnetum montanum, Ginkgo biloba*), basal angiosperms (*L. chinense, Amborella trichopoda*), monocots (*Oryza sativa, Zea mays*), and eudicots (*A. thaliana, Solanum lycopersicum*) to do a synteny analysis between species on close taxa. The methods and procedures used were the same as those previously outlined.

The HSFC1-C2 subfamily genes and the HSFA2 and HSFA9 subfamily genes were extracted from the database and used to estimate divergence time. We calibrated a relaxed molecular clock on the node and found the divergence time of monocots and eudicots to be between 140 Mya (a minimum age) and 200 Mya (a maximum age) (represented by the divergence of *A. thaliana* and *O. sativa*, *Gensel & Andrews, 1984*). We performed a Bayesian dating analysis in the Markov chain Monte Carlo (MCMC) tree (*Yang, 2007*) using an approximate likelihood calculation for the branch lengths, an auto-correlated model of among-lineage rate variation, the GTR substitution model, and a uniform prior on the relative node times. We used Markov chain Monte Carlo sampling to estimate posterior distributions of node ages, with samples drawn every two steps over 200,000 steps following a burn-in of 10,000 steps. We could then trace the gene duplication time based on the resulting gene divergence times.

RESULTS

The phylogeny and evolution of HSFs in land plants

A total of 670 HSF sequences from 44 species were used for the phylogenetic analysis (Tables S1 and S2). We identified 287 new candidate HSF sequences from 24 species, with 228 of those divided into known subfamilies (A1-A9, B1-B5, C1-C2) (Table 1) on the HEATSTER website. Across the comprehensive samples studied, the number of HSF gene subfamilies identified varied greatly, ranging from two in chlorophyta to 30 in angiosperms.

Table 1 The species used for phylogenetic tree construction, and the category of HSFs.

			Category of HSFs																
Taxonomy	**Species**	**Abbreviation**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**	**B1**	**B2**	**B3**	**B4**	**B5**	**C1**	**C2**	**HSF like (N.C.)**	**Total**
Chlorophyta	*Chlamydomonas reinhardtii*	Chlre	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	4
Chlorophyta	*Volvox carteri*	Volca	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Bryophyta	*Physcomitrella patens*	Phypa	4	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	8
Pteridophyta	*Selaginella moellendorffii*	Selmo	4	0	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0	7
Pteridophyta	*Azolla filiculoides*	Azofi	7	0	0	0	0	0	0	0	0	3	0	0	3	0	0	0	1	14
Pteridophyta	*Ceratopteris gametophytes* [a]	Cerga	3	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	3	9
Pteridophyta	*Lygodium japonicum* [a]	Lygja	5	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	6	13
Pteridophyta	*Pteridium aquilinum* [a]	Pteaq	1	1	0	0	0	0	0	0	0	3	0	0	0	0	0	0	5	10
Pteridophyta	*Salvinia cucullata*	Salcu	7	1	0	0	0	0	0	0	0	4	0	0	3	0	0	0	0	15
Gymnosperm	*Abies firma* [a]	Abifi	2	1	0	0	0	0	0	0	0	2	1	0	2	0	0	0	3	11
Gymnosperm	*Araucaria cunninghamii* [a]	Aracu	4	0	0	0	0	0	0	0	0	4	1	0	1	0	0	0	3	13
Gymnosperm	*Cephalotaxus sinensis* [a]	Cepsi	3	0	0	0	0	0	0	0	0	2	1	0	1	0	0	0	1	8
Gymnosperm	*Cycas revoluta* [a]	Cycre	4	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	1	8
Gymnosperm	*Ephedra equisetina* [a]	Epheq	4	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	7
Gymnosperm	*Ginkgo biloba*	Ginbi	5	0	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0	8
Gymnosperm	*Ginkgo biloba* [a]	GinbiR	4	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	2	9
Gymnosperm	*Gnetum montanum*	Gnemo	6	0	0	0	0	0	0	0	0	1	0	0	4	0	0	0	13	24
Gymnosperm	*Metasequoia glyptostroboides* [a]	Metgl	6	0	0	0	0	0	0	0	0	2	1	0	1	0	0	0	2	12
Gymnosperm	*Picea abies*	Picab	4	0	0	0	0	0	0	0	0	7	1	0	2	0	0	0	5	19
Gymnosperm	*Picea abies* [a]	PicabR	3	1	0	0	0	0	0	0	0	2	1	0	0	0	0	0	2	9
Gymnosperm	*Picea glauca*	Picgl	4	0	0	0	0	0	0	0	0	11	1	0	2	0	0	0	0	18
Gymnosperm	*Pinus taeda*	Pinta	7	1	0	0	0	0	0	0	0	15	1	0	4	0	0	0	20	48
Gymnosperm	*Pinus taeda* [a]	PintaR	4	0	0	0	0	0	0	0	0	2	1	0	2	0	0	0	1	10
Gymnosperm	*Podocarpus macrophyllus* [a]	Podma	3	0	0	0	0	0	0	0	0	3	1	0	0	0	0	0	3	10
Gymnosperm	*Sciadopitys verticillata* [a]	Scive	2	0	0	0	0	0	0	0	0	3	1	0	1	0	0	0	3	10
Gymnosperm	*Taxus chinensis* [a]	Taxch	5	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	1	9
Gymnosperm	*Welwitschia mirabilis* [a]	Welmi	8	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	2	12
Gymnosperm	*Zamia furfuracea* [a]	Zamfu	4	0	0	0	0	0	0	0	0	2	1	0	1	0	0	0	1	9
Basal angiosperms	*Amborella trichopoda*	Ambtr	2	1	1	0	1	1	0	0	0	1	1	0	2	1	0	0	2	13
Basal angiosperms	*Liriodendron chinense*	Lirch	2	2	1	1	2	1	1	0	0	2	2	1	1	1	1	0	0	18
Basal angiosperms	*Nymphaea colorata*	Nymco	3	1	1	1	1	0	1	0	0	1	2	0	3	1	2	0	4	21

(*continued on next page*)

Table 1 (*continued*)

			Category of HSFs																
Taxonomy	**Species**	**Abbreviation**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**	**B1**	**B2**	**B3**	**B4**	**B5**	**C1**	**C2**	**HSF like (N.C.)**	**Total**
Eudicots	*Arabidopsis thaliana*	Arath	4	1	1	2	1	2	2	1	0	1	2	1	1	0	1	0	4	24
Eudicots	*Cajanus cajan*	Cajca	2	1	1	2	1	2	1	1	1	2	2	1	4	1	1	0	4	27
Eudicots	*Citrullus lanatus*	Citla	1	2	1	3	1	2	0	2	1	1	2	2	3	1	2	0	0	24
Eudicots	*Mimulus guttatus*	Mimgu	2	1	1	2	0	2	0	1	0	0	2	1	2	1	1	0	5	21
Eudicots	*Nelumbo nucifera*	Nelnu	4	2	1	2	1	1	0	1	0	2	2	2	3	2	2	0	0	25
Eudicots	*Populus trichocarpa*	Poptr	3	1	1	3	2	4	0	2	1	1	3	2	4	2	1	0	4	34
Eudicots	*Prunus persica*	Prupe	2	1	1	2	0	2	0	1	1	1	2	1	1	1	1	0	3	20
Eudicots	*Solanum lycopersicum*	Solly	4	1	1	3	1	2	0	1	3	1	2	2	2	1	1	0	1	26
Monocots	*Brachypodium distachyon*	Bradi	1	3	1	2	1	2	2	1	0	1	3	0	3	0	2	2	2	26
Monocots	*Oryza brachyantha*	Orybr	0	3	0	2	1	2	2	1	0	1	1	0	3	0	2	1	3	22
Monocots	*Oryza sativa*	Orysa	1	4	1	2	1	2	2	1	0	1	3	0	4	0	2	2	3	29
Monocots	*Phoenix dactylifera*	Phoda	7	3	2	2	2	2	0	0	0	2	4	0	3	0	1	2	1	31
Monocots	*Phyllostachys heterocycla*	Phyhe	1	4	2	3	2	3	1	0	0	2	2	0	4	0	2	1	14	41
Monocots	*Sorghum bicolor*	Sorbi	1	3	1	1	1	2	2	1	0	1	3	0	3	0	2	2	3	26
Monocots	*Triticum urartu*	Triur	1	5	1	1	1	0	0	1	0	1	0	0	0	0	2	1	6	20
Monocots	*Zea mays*	Zeama	2	2	1	3	1	2	2	2	0	2	4	0	2	0	2	2	13	40

Notes.

[a]The data from transcriptomes. N.C. the sequence only contains some of the necessary domains for a heat shock transcription factor and therefore it could not be classified.

The unrooted phylogenetic tree inferred from amino acid sequences was well resovled to three main clades: HSFA, HSFB and HSFC (Fig. 1, Figs. S1–S4). The newly identified HSF genes were re-confirmed on a phylogenic tree. Most subfamilies of clades (A3, A4, A5, A8, A9, B2, B3, B5, C1, C2) were accordingly recovered, while the relationships between these clades were weakly supported. The HSF subfamilies displayed a strong diversification in structure, composition and function (*Scharf et al., 2012*; *Guo et al., 2016*; *Wang et al., 2018*), thus, significant genetic differentiation between clades, especially for HSFA and HSFB were likely resulted from the unstable topology observed. The HSFA group was found in all sampled taxa, while the HSFB group was absent in chlorophyta, and the HSFC group was only present in the angiosperms.

The HSFA group contains major regulators in the HS response of plants (*Wang et al., 2018*), and, as a result of diversification during plant evolution, displayed variations in different taxa. Interestingly, the A4-A9 subfamily clades are only occurred in angiosperms and A9 genes are only identified in Eudicots. Some subfamilies clustered as a branch, such as A3, A4, A5, and A9, while others were clustered as several branches (Figs. S1 and S5). HSFA1 is a master regulator which cannot be replaced by any other HSF (*Scharf et al., 2012*) and probably be the most ancient HSFA group. Although all the HSFA1 genes with HSFA8 in angiosperms clustered as a clade, most of the HSFA1 genes from pteridophytes and gymnosperms were dispersed into several clades. The deep divergence of HSFA1 in pteridophytes and gymnosperms indicates the early diversification of HSFA1 before the radiation of all seed plants. Meanwhile, HSFA2, HSFA6, HSFA7, and HSFA9 were blended into a complex clade, and HSFA9 formed a monophyletic group, but others remain unclear. We also noticed that the HSFA2 gene and the HSFA6 gene clustered together with very little genetic difference in some angiosperm species such as *O. sativa*, *Phoenix dactylifera, Citrullus lanatus*, as the HSFA6 and HSFA7 did in *C. lanatus*. The relationship between the HSFA4 clade and the HSFA5 clade was closer in the tree, with two HSFA5 genes sneaked into the HSFA4 clade. It has previously been suggested a close relationship between HSFA3 and HSFC, however, due to the increased number of ferns and gymnosperms, one HSFA1 clade of gymnosperms, rather than HSFA3, was clustered with HSFC.

HSFC only displayed the pattern as the angiosperms clade clustered to one clade of HSFA1 in gymnosperms (Figs. S1, S2, S3 and S5). It is assumed that a duplication event occurred in the ancestral angiosperms which could have contributed to the rise of HSFC. HSFC1 is a common gene subfamily and varies in gene numbers between monocots and eudicots (Table 1, Fig. S5); there are usually two members in most monocots and only one member in eudicots. These results indicate that HSFC experienced steady expansion during the evolution of monocots, and may be involved in important developmental pathways (*Wang et al., 2018*). Notably, HSFC2 was only present in monocots, but HSFC1 was present in all angiosperm species except for *A. trichopoda.* In monocots, HSFC1 and HSFC2 clustered together with strong support. HSFC1-HSFC2 clade of monocots group to HSFC1 of eudicots, based with HSFC1 of basal angiosperms. This suggests that HSFC2 is the result of recent duplication that occurred early in the divergence between monocots and eudicots.

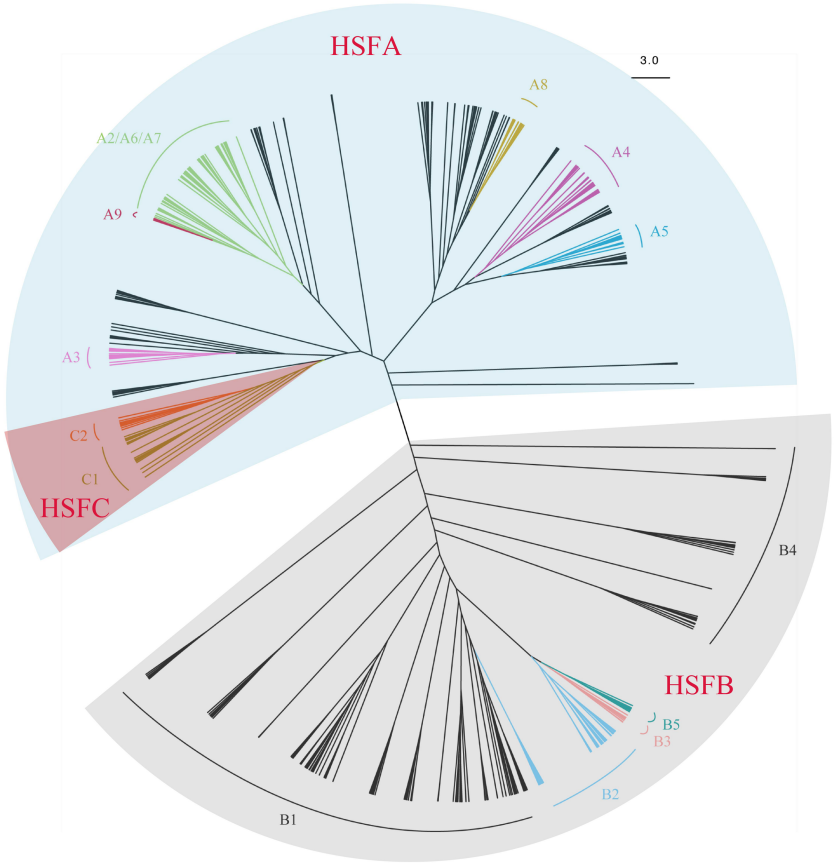

Figure 1 An unrooted Maximum-Likelihood tree showing the hylogeny and classification of 670 HSFs sequences from 44 species representing seven main taxa including chlorophyta, bryophyta, peridophyta, gymnospermae, basal angiosperms, eudicots and monocots. The information of species and sequences accession numbers used for the tree are listed in File S1. HSFA, HSFB and HSFC are clustered into three main clades. The clade of subfamilies HSFA2-7, HSFA8 and HSFA9, HSFB2-5, and HSFC1 and HSFC2, were shown over relevant branches with different colors. The three groups HSFA, HSFB, and HSFC were highlighted with shades of different colors. The scale bar represents amino acid substitutions per site.

Full-size DOI: 10.7717/peerj.13603/fig-1

Contrary to a previous study (*Wang et al., 2018*), the results of this study suggest that the HSFB subfamily (HSFB1-HSFB5) is moderately supported as a monophyletic group (Figs. S1 and S6). HSFB1, HSFB2, and HSFB4 have been widely observed across land plants, while both HSFB3 and HSFB5 are only present in the eudicots and basal angiosperms. Although HSFB5, unlike other subfamily members of the HSFB group, has a conserved tetrapeptide LFGV in the C-terminal domain, it is closely related to HSFB3 (Fig. 1). Additionally, the number of HSFB1 genes in gymnosperms is far more than that in angiosperms, with the common number reduced from 3 or 4 in gymnosperms to 1 or 2 in angiosperms (Table 1). In particular, the number of HSFB1 genes in conifers (*Picea abies, Pinus taeda, Picea glauca*) is significantly increased than that of other seed plants. Multiple copies of the HSFB1 gene in *P. abies, P. taeda,* and *P. glauca* clustered and formed a strongly supported monophyletic group. This result indicates that the evolution of these three conifers probably involved

both polyploidy and repetitive element activity (*Drewry, 1988*; *Ahuja, 2005*; *Li et al., 2015*). The multi-copy genes may be attributed to two whole genome duplication (WGD) events in the ancestry of major conifer clades (*Li et al., 2015*). Though many angiosperm lineages have experienced additional rounds of genome duplication (*Soltis, Visger & Soltis, 2014*; *Jiao et al., 2014*; *Li et al., 2015*), there is no obvious proliferation in the member. This result is consistent with the speculation that WGD in angiosperms did not give rise to a remarkable expansion of HSFB1 genes. The HSFB1 genes in angiosperms, gymnosperms and pteridophytes were independently found in different branches, which suggests that HSFB1 is an ancient group which diverged during the evolutionary history of different taxa. HSFB1 genes in gymnosperms experienced several expansions including ancient duplication, while HSFB1 genes in angiosperms rarely retained duplication except for a few recent duplicates. All HSFB2 genes in gymnosperms and angiosperms clustered as a group, respectively. We were unable to trace out a remarkable expansion in gymnosperms, but more than two genes in angiosperms were assumed to be the result of recent duplication. In some species, such as *Selaginella moellendorffii*, we observed that some genes identified as different subfamilies, such as HSFB1 and HSFB4, and have genetic similarities to highly supported clades. The complicated relationship of these two subfamilies may be a result of recent duplication events. In this study, the HSFB3 and HSFB5 subfamilies were only present in eudicots and basal angiosperms. This is likely the result of duplication events occurring in ancestral angiosperms with subsequent loss of paralogue genes in the monocots.

Gene duplication analysis

To examine the expansion patterns and genetic divergences of the HSF family, a synteny analysis was performed to identify gene duplication events across 21 species (Table S3). We also conducted a synteny analysis between different species on the closely related taxa.

Gene duplication events were identified in 11 species including pteridophytes, basal angiosperms, monocots and eudicots (Table 2). In green algae, moss, and gymnosperms, we did not detect any HSF genes in synteny blocks. In *S. moellendorffii,* the only non-seed plant analyzed, we identified one pair of duplication genes. These two genes, 'SelmoHSFB1b' and 'SelmoHSFB4,' belong to different subclasses of the HSF gene family which were observed as being syntenic to each other. We speculate that these genes may be derived from a duplication event and have evolved with differences at the gene sequence level. In *L. chinense*, the only basal angiosperm analyzed, we identified five pairs of duplication genes with four of those five pairs from the same gene subclass (HSFA2, HSFB1, HSFB2, HSFC1) and the remaining pair from a different gene subclass (HSFA4-HSFA5). Gene duplication events were detected in all sampled eudicot and monocot species. In five eudicots (*A. thaliana, Populus trichocarpa, Prunus persica, S. lycopersicum, Mimulus guttatus*), we identified 29 pairs of duplication genes out of which 33 pairs belonged to the same gene subclasses (HSFA1, HSFA4, HSFA5, HSFA6, HSFA8, HSFB2, HSFB3, HSFB4, HSFB5) and four pairs belonged to different gene subclasses (HSFA2-HSFA9, HSFA6-HSFA7). In four monocots (*O. sativa, Sorghum bicolor, Z. mays, Brachypodium distachyon*), we also identified 29 pairs of duplication genes out of which 33 pairs belonged to the same gene

Table 2 The detected paralogous genes within different species.

Order	Species	Paralogous genes Types
Pteridophyta	*Selaginella moellendorffii*	HSFB1-HSFB4
Basal angiosperms	*Liriodendron chinense*	HSFC1-HSFC1, HSFA2-HSFA2, HSFA4-HSFA5, HSFB1-HSFB1, HSFB2-HSFB2
Eudicots	*Arabidopsis thaliana*	HSFA1-HSFA1, HSFA4-HSFA4, HSFA6-HSFA6, HSFA6-HSFA7
	Populus trichocarpa	HSFA1-HSFA1, HSFA4-HSFA4, HSFA5-HSFA5, HSFA6-HSFA6, HSFA8-HSFA8, HSFA9-HSFA2, HSFB2-HSFB2, HSFB3-HSFB3, HSFB4-HSFB4, HSFB5-HSFB5
	Prunus persica	HSFA2-HSFA9, HSFA6-HSFA6, HSFB2-HSFB2
	Solanum lycopersicum	HSFA1-HSFA1, HSFA4-HSFA4, HSFA6-HSFA6, HSFA9-HSFA2, HSFB2-HSFB2, HSFB3-HSFB3
	Mimulus guttatus	HSFB4-HSFB4
Monocots	*Oryza sativa*	HSFA2-HSFA2, HSFA6-HSFA2, HSFB2-HSFB2, HSFB4-HSFB4, HSFC2-HSFC2
	Sorghum bicolor	HSFA2-HSFA2, HSFA2-HSFA6, HSFA6-HSFA6, HSFB2-HSFB2, HSFC2-HSFC2
	Zea mays	HSFA1-HSFA1, HSFA2-HSFA2, HSFA4-HSFA4, HSFB1-HSFB1, HSFB2-HSFB1, HSFB2-HSFB2, HSFB2-HSFB4, HSFC1-HSFC1, HSFC2-HSFC2
	Brachypodium distachyon	HSFA2-HSFA2, HSFA6-HSFA6, HSFB2-HSFB2, HSFB4-HSFB4, HSFC2-HSFC2

subclasses (HSFA1, HSFA2, HSFA4, HSFA6, HSFB1, HSFB2, HSFB4, HSFC1, HSFC2) and four pairs belonged to different gene subclasses (HSFA2-HSFA6, HSFB1-HSFB2, HSFB2-HSFB4). In general, all HSF gene subfamilies except HSFA3 showed the signature of gene duplication. These results also demonstrated that gene pairs from different subfamilies, such as HSFA2-HSFA6, HSFA2-HSFA9, HSFA4-HSFA5, HSFA6-HSFA7, HSFB1-HSFB4, HSFB1-HSFB2, and HSFB2-HSFB4, were paralogous gene pairs.

Beyond that, synteny analysis among different species identified the orthologous genes in different taxa (Table 3). In detail, only HSFA1 genes from different sources were found as orhologous genes between two gymonosperms (*G. montanum, G.biloba*). As a result of the analysis of gymnosperms (*G. biloba*) and basal angiosperms (*L. chinense*), HSFA1, HSFA4, and HSFA5 were detected as orthologous genes. Though the analysis among basal angiosperms (*A. trichopoda, L. chinense*) and eudicots (*S. lycopersicum, A. thaliana*) found several orthologous genes, such as HSFA6-HSFA7, HSFA4-HSFA5, HSFA2-HSFA9, and HSFB2-HSFB5, among basal angiosperms and monocots (*O. sativa*, *Z. mays*), we only identified out HSFA2-HSFA6 and HSFA2-HSFA7 as orthologous genes. Interestingly, the analysis of eudicots-monocots reveal a consistent pattern as basal angiosperms-monocots with HSFA1-HSFA5 and HSFA2-HSFA7 being identified as orthologous genes in monocots and HSFA6-HSFA7 as orthologous genes in eudicots.

Our results indicate that gene duplication in HSF genes has been a frequent event during the evolution of plants, significantly contributing to their expansion and functional diversification (Fig. 2). Our results also suggest that HSFA4 and HSFA5 have a close genetic relationship, the origin of which may be related to the ancient duplication of HSFA1. It is possible that HSFA6 and HSFA7 originated from gene duplication, most probably derived from HSFA2. HSFA9 was proven to be derived from HSFA2 after the divergence of ancestral angiosperms. HSFB1 is considered to be the most ancient among the HSFB

Table 3 The ortologous gene clusters detected between different species.

Pairwise_Taxa	Pairwise_Species	Ortologous gene Types
Gymnosperm-Gymnosperm	*Gnetum montanum-Ginkgo biloba*	HSFA1-HSFA1
Gymnosperm-Basal angiosperms	*Amborella trichopoda-Ginkgo biloba*	HSFA1-HSFA1
	Liriodendron chinense-Ginkgo biloba	HSFA4-HSFA1, HSFA5-HSFA1
Basal angiosperms-Basal angiosperms	*Liriodendron chinense-Amborella trichopoda*	HSFA1-HSFA1,HSFA2-HSFA2,HSFA3-HSFA3,HSFA5-HSFA5,HSFA6-HSFA6,HSFB2-HSFB2,HSFB5-HSFB5
Basal angiosperms-Eudicots	*Arabidopsis thaliana-Amborella trichopoda*	HSFA1-HSFA1, HSFA5-HSFA5, HSFA6-HSFA6, HSFA6-HSFA7, HSFB2-HSFB2,
	Arabidopsis thaliana-Liriodendron chinense	HSFA1-HSFA1, HSFA2-HSFA2, HSFA4-HSFA4, HSFA4-HSFA5, HSFB1-HSFB1, HSFB2-HSFB2, HSFB3-HSFB3, HSFC1-HSFC1, HSFC1-HSFC1
	Liriodendron chinense-Solanum lycopersicum	HSFA1-HSFA1, HSFA2-HSFA2, HSFA2-HSFA9, HSFA4-HSFA4, HSFA4-HSFA5, HSFB1-HSFB1, HSFB2-HSFB2, HSFB3-HSFB3, HSFB4-HSFB4, HSFC1-HSFC1, HSFC1-HSFC1
	Amborella trichopoda-Solanum lycopersicum	HSFA1-HSFA1, HSFA2-HSFA2, HSFA2-HSFA9, HSFA5-HSFA5, HSFA6-HSFA6, HSFB5-HSFB2, HSFB5-HSFB5,
	Zea mays-Amborella trichopoda	HSFA2-HSFA6, HSFA3-HSFA3, HSFA6-HSFA6, HSFB2-HSFB2
	Zea mays-Liriodendron chinense	HSFB1-HSFB1,HSFB2-HSFB2,HSFB4-HSFB4
Basal angiosperms-Monocots	*Oryza sativa-Amborella trichopoda*	HSFA2-HSFA7, HSFA3-HSFA3, HSFA6-HSFA2, HSFA6-HSFA6, HSFB2-HSFB2
	Oryza sativa-Liriodendron chinense	HSFA4-HSFA4, HSFA7-HSFA2, HSFB1-HSFB1, HSFB2-HSFB2, HSFB4-HSFB4
	Oryza sativa-Arabidopsis thaliana	HSFA6-HSFA2, HSFA6-HSFA6, HSFA7-HSFA2
Eudicots-Monocots	*Oryza sativa-Solanum lycopersicum*	HSFA2-HSFA6, HSFA4-HSFA4, HSFA6-HSFA6, HSFA7-HSFA2, HSFB1-HSFB1, HSFB2-HSFB2, HSFB4-HSFB4,
	Zea mays-Arabidopsis thaliana	HSFA2-HSFA6
	Zea mays-Solanum lycopersicum	HSFA2-HSFA6, HSFA6-HSFA6, HSFB1-HSFB1, HSFB2-HSFB2,
Monocots-Monocots	*Oryza sativa-Zea mays*	HSFA1-HSFA1, HSFA1-HSFA5, HSFA2-HSFA2, HSFA3-HSFA3, HSFA4-HSFA4, HSFA6-HSFA2, HSFA6-HSFA6, HSFA7-HSFA7, HSFA8-HSFA8, HSFB1-HSFB1, HSFB2-HSFB2, HSFB4-HSFB4, HSFC1-HSFC1, HSFC2-HSFC2
Eudicots-Eudicots	*Arabidopsis thaliana-Solanum lycopersicum*	HSFA1-HSFA1, HSFA2-HSFA2, HSFA3-HSFA3, HSFA4-HSFA4, HSFA5-HSFA5, HSFA6-HSFA6, HSFA6-HSFA7, HSFB1-HSFB1, HSFB2-HSFB2, HSFB3-HSFB3, HSFC1-HSFC1

genes, and we predict that HSFB2 and HSFB4 derived from HSFB1 considering the close relationship between them.

Divergence time analysis

The estimated divergence dates of HSFA2 and HSFA9 in eudicots are indicated in Fig. 3. The divergence time of those two gene subfamilies in this study ranges from 131 Mya to 155.2 Mya, which is within the period of Late Jurassic to Lower Cretacous. The estimated split time of the HSFC2 clade and HSFC1 in monocots is indicated in Fig. 4, and ranges from 125 Mya to 190.4 Mya, which is within the Jurassic and Lower Cretaceous periods.

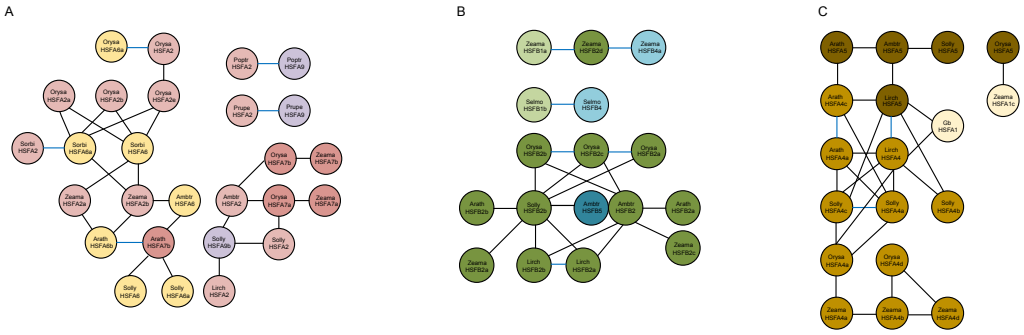

Figure 2 **(A) Synteny analysis between the subfamilies HSFA2, HSFA6, HSFA7, HSFA9 of seven representative plant species (*Amborella trichopoda, Liriodendron chinense, Arabidopsis thaliana, Solanum lycopersicum, Oryza sativa, Sorghum bicolor, Zea mays*). (B) Synteny analysis between the subfamilies HSFB1, HSFB2, HSFB4, HSFB5 of seven representative plant species (*Selaginella moellendorffii, Amborella trichopoda, Liriodendron chinense, Arabidopsis thaliana, Solanum lycopersicum, Oryza sativa, Zea mays*). (C) Synteny analysis between the subfamilies HSFA1, HSFA4, HSFA5 of eight representative plant species (*Ginkgo biloba, Amborella trichopoda, Liriodendron chinense, Arabidopsis thaliana, Solanum lycopersicum, Oryza sativa, Sorghum bicolor, Zea mays*).** Black, and blue lines indicate orthologous, and paralogous gene pairs respectively. The different colored circle represent HSF genes from different subfamilies. The name of the genes is inside the circle.

Full-size DOI: 10.7717/peerj.13603/fig-2

The time of the occurence of these gene duplications are consistent with the origin of the most recent common acestors of all living angiosperm, which likely be around 140–250 Mya (*Magallón et al., 2015*; *Foster et al., 2017*). Although uncertainty remains for other characters, our reconstruction of the differentiation time scale between gene subfamilies allows us to propose a new plausible scenario for the early diversification of angiosperms at genomic level. The origin and rapid diversification of angiosperms represent one of the most intriguing topics in evolutionary biology (*Sauquet & Magallón, 2018*), and the evolution research of this gene family (such as the origin, expansion and loss of genes) provides an unprecedented opportunity to explore remarkable long-standing questions that may hold important clues toward understanding present-day biodiversity and adaption to different environments.

DISCUSSION

Previous phylogenetic studies of the HSF gene family in plants have provided valuable insights into its evolutionary history (*Scharf et al., 2012*; *Wang et al., 2018*). However, the limited sampling of pteridophytes, gymnosperms and basal angiosperms have left unresolved questions regarding the origin of subclasses in the HSF gene family and their phylogenetic relationship and gene expansion patterns in different taxa. HSFs play a key role in the adaptation of plants to changing habitats and environmental stressors. Our understanding of land plant evolution at a genetic level in relation to environmental changes has also been hindered by sampling limitations (*Rensing et al., 2008*; *Banks et al., 2011*; *Scharf et al., 2012*; *Nystedt et al., 2013*; *Lin et al., 2014*; *Wang et al., 2018*; *Lohani et al., 2019*). Although ongoing plant genome projects will certainly uncover additional species or family-specific deletions and duplications, the general features are likely not to change

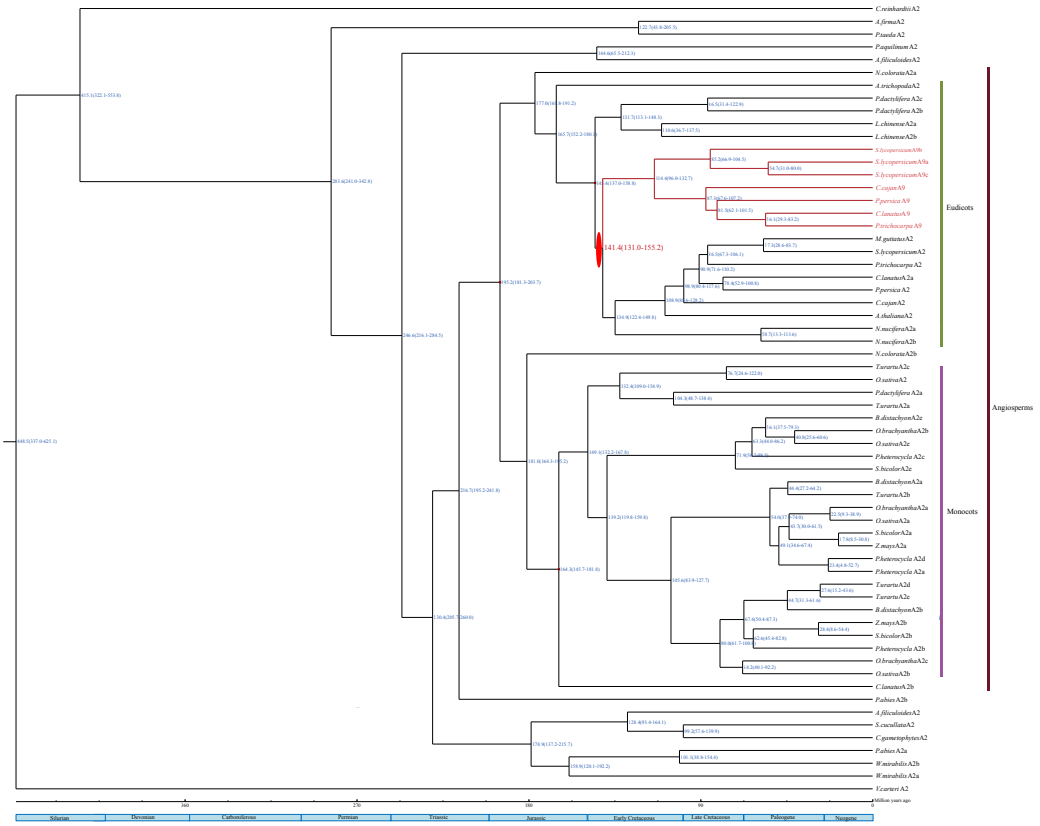

Figure 3 **A dated phylogenetic reconstruction for the subfamilies HSFA2 and HSFA9.** Red ovals indicate gene duplication events. The divergence time of HSFA2 and HSFA9 are marked with red. The blue numbers on each node refer to the mean time to MRCA estimates; the blue numbers in parentheses on each node refer to the 95% highest posterior density intervals.

Full-size DOI: 10.7717/peerj.13603/fig-3

(*Thalmann et al., 2019*). In this study, the diversity and number of plants examined allowed us to examine the evolutionary history of this gene family in a broader taxonomic context. Our phylogenetic analyses revealed a divergence of HSF subfamilies and independent evolution in plants, especially in angiosperms. It is still a big challege for multi-alignment of genomes to recognize the potential syntenic relationships due to the ubiquity of ancient and recent polyploidy events, as well as smaller scale events that derive from tandem and transposition duplications (*Lynch & Conery, 2000*; *Bowers et al., 2003*; *Tang et al., 2008*; *Schranz, Mohammadin & Edger, 2012*). However, thanks to a combination of phylogenic analyses and synteny analysis in this study, our results have scratched the surface of just how gene expansion in different land plant taxa occurred. Our results show that recent duplication events are mostly contributed to the puzzle clades (HSFA2, HSFA6, HSFA7, HSFA4, HSFA5) with members from other groups snuck in.

Our studies on different members of the HSF gene family from pteridophytes and gymnosperms reveal that this gene family is quite complex in terms of gene numbers and sequence diversity. We identified four subfamilies of HSFs (HSFA1, HSFA2, HSFB1,

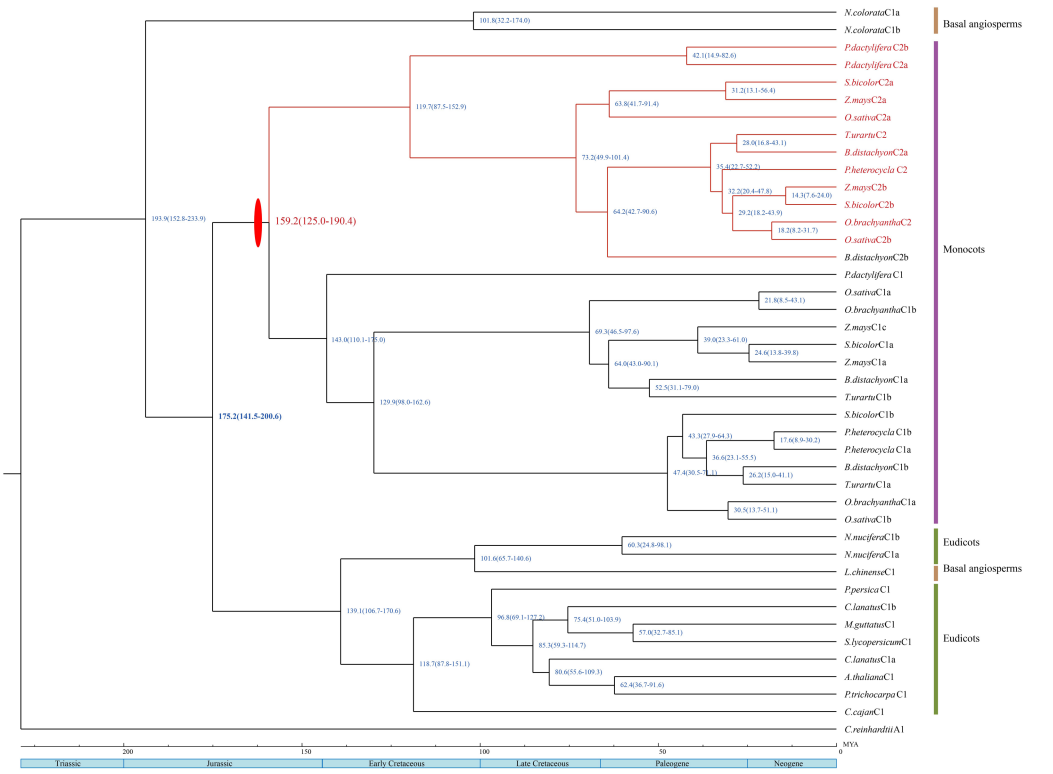

Figure 4 A dated phylogenetic reconstruction were done for the subfamilies HSFC1 and HSFc2. Red ovals indicate gene duplication events. The divergence time of HSFC1 and HSFc2 are marked with red. The blue numbers on each node refer to the mean time to MRCA estimates; the blue numbers in parentheses on each node refer to the 95% highest posterior density intervals.

Full-size DOI: 10.7717/peerj.13603/fig-4

HSFB4) across candidate HSFs in six species of pteridophyte, and five subfamilies of HSFs (HSFA1, HSFA2, HSFB1, HSFB2, HSFB4) from 16 species of gymnosperm. Though the number of HSFs in pteridophytes and gymnosperms is significantly less than in angiosperms, the number of HSFA1 and HSFB1 genes in those taxa was higher than in angiosperms. It is assumed that pteridophytes and gymnosperms preferred to reserve more ancient members in HSFs subfamily. The HSFA1 and HSFB1 subfamilies in pteridophytes and gymnosperms separately formed more than one clade on a phylogenic tree with low support without clustering together, consistent with the findings that more ancient duplication events affect more distant taxonomic comparisons (*Bowers et al., 2003*). Only two genes (SelmoHSFB1b and SelmoHSFB4) in *S. moellendorffii* appeared to be the result of duplication events detected in a syntenic analysis. These findings indicate that HSFA1, HSFA2, HSFB1 and HSFB4, which were already commenced in the ancestor of all land plants, are ancestral gene groups.

Gymnosperm lineages were considerably diverged during the Late Carboniferous to the Late Triassic periods, and were dominant through most of the Mesozoic period (*Bowe, Coat & DePamphilis, 2000*; *Chaw et al., 2000*). However, massive extinction occurred in the Cenozoic period caused gymnosperm genera to diversify slower than angiosperms

(*Crisp & Cook, 2011*). Ancient gene subfamilies, such as HSFB1 and HSFA1, experienced differentiation and variation over a long period of time, which may explain the molecular phylogenetic uncertainty within gymnosperms. Ancient WGDs have been probably inferred in the ancestry of all extant seed plants, and angiosperm and gymnosperm lineages have experienced additional rounds of WGD (*Cui et al., 2006*; *Barker et al., 2008*; *Soltis, Visger & Soltis, 2014*; *Jiao et al., 2014*; *Li et al., 2015*; *Cannon et al., 2015*). Although no syntenic gene was detected in gymnosperms, two or more genes from different subclasses form strongly supported clades (such as PintaHSFA1a and PintaHSFA2, AbifiRHSFB1a and AbifiRsfB4a), so the absence of syntenic genes in gymnosperms may be a result of the incomplete data sampling, or relatively lower quality of currently available assembly in gymnosperms. Alternatively, ancient interspersed segmental duplication of those genes occurred recently could be detected though phylogenetic and synteny analyses.

In angiosperms, the HSF gene family has undergone extensive duplications that have given rise to complicated orthology, paralogy, and functional heterology relationships. Our results showed that the diversity and number of HSF genes in angiosperms is remarkably higher than in other much earlier diverged taxa. We also observed a higher diversity and number of multiple paralogous and ortholog genes in angiosperms. Most of the gene copies generated by WGD events have been lost due to fractionation and subsequent "postdiploidization" or malfunctionalization (*Jiao et al., 2011*). Gene duplication is an important mechanism for genomic innovation (*Li et al., 2016*), and the functional divergence of duplicate genes retained from whole genome duplication (WGD) is thought to promote evolutionary diversification. Recent WGDs occurring in angiosperms, especially lineage-specific WGDs, have allowed the expansion and variation of HSFs, which supported by previous studies in *Fagopyrum tataricum* (*Liu et al., 2019*) and genus *Brassica* (*Lohani et al., 2019*). The results of a synteny analysis confirmed that the HSFA9 subfamily was only present in eudicots which derived from HSFA2, and HSFC2 genes were only present in monocots which derived from HSFC1. New genes originated from the divergence of paralogue genes, which resulted from duplication events. These two duplication events occurred early in angiosperm divergence, consistent with angiosperm radiations occurring in the Late Jurassic and Lower Cretaceous periods (*Li et al., 2019*). Approximately 132 Mya ago, angiosperms underwent rapid radiation to become the most diverse and successful plant group on land (*Sanderson & Donoghue, 1994*). The co-occurrence of retained duplication events with key processes in biological innovations underlines the importance of this crucial mechanism (*Airoldi & Davies, 2012*). The HSFB3 and HSFB5 subfamilies were found to be absent in monocots, but present in most basal angiosperms and eudicots. We hypothesize that HSFB3 and HSFB5 were thoroughly lost in the most recent common acestors of monocots, yet, their origin and evolutionary history remain poorly understood. We speculate that those gene loss events occurred from divergence early in angiosperm history. The above results indicate that species not only experienced rapid early radiation, diversification and mass extinction (*Deenen et al., 2010*; *Meredith et al., 2011*; *Wickett et al., 2014*; *Zeng et al., 2014*; *Li et al., 2019*), but also that genes went through expansion, diversification, and loss. After the divergence of angiosperms, eudicots and monocots experienced independent evolutionary processes.

CONCLUSIONS

The progressive increased data of whole geome assembly from different phylogenetic lineages has advanced our evolutionary understanding of gene families. Our comprehensive analysis reveals that the diversification of HSFs in plants resulted from extensive gene duplications and gene loss during the evolution and diversification of land plants. Lineage-specific expansions in angiosperms, especially in eudicots and monocots, may reflect the potential evolutionary advantage of flexibility in complex environments. The patterns of gene duplication and the evolutionary history of HSFs in plants provide novel insights into their diversity which facilitates the plant diversification, adaptation and evolution in various habitats. Our analyses provide essential insights for studying the evolutionary history of multigene families.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by National Natural Scientific Foundation of China (Grant No. 31870206, 31670369), the Innovation of Science and Technology Commission of Shenzhen Foundation (Grant No. JCYJ201206151530054), the Scientific Research Foundation of Fairy Lake Botanical Garden and the Scientific Research Program of Sino-Africa Joint Research Center (Grant No. SAJL201607). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Natural Scientific Foundation of China: 31870206, 31670369.
Innovation of Science and Technology Commission of Shenzhen Foundation: JCYJ201206151530054.
The Scientific Research Foundation of Fairy Lake Botanical Garden.
Scientific Research Program of Sino-Africa Joint Research Center: SAJL201607.

Competing Interests

The authors declare there are no competing interests. Dan Yang and Linping Meng all have been analyzed data. Because of her transfer, Dan Yang suggested the authorship change to Linping Meng.

Author Contributions

- Yiying Liao conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Zhiming Liu performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Andrew W. Gichira analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

- Min Yang analyzed the data, prepared figures and/or tables, and approved the final draft.
- Ruth Wambui Mbichi analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Linping Meng analyzed the data, prepared figures and/or tables, and approved the final draft.
- Tao Wan conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The raw data is available in the Supplemental File.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.13603#supplemental-information.

REFERENCES

Åkerfelt M, Morimoto RI, Sistonen L. 2010. Heat shock factors: integrators of cell stress, development and lifespan. *Nature Reviews Molecular Cell Biology* **11**:545–555 DOI 10.1038/nrm2938.

Ahuja MR. 2005. Polyploidy in gymnosperms: revisited. *Silvae Genetica* **54**:59–69 DOI 10.1515/sg-2005-0010.

Ahuja I, DeVos RCH, Bones AM, Hall RD. 2010. Plant molecular stress responses face climate change. *Trends in Plant Science* **15**:664–674 DOI 10.1016/j.tplants.2010.08.002.

Airoldi CA, Davies B. 2012. Gene duplication and the evolution of plant MADS-box transcription factors. *Journal of Genetics and Genomics* **39**:157–165 DOI 10.1016/j.jgg.2012.02.008.

Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, DePamphilis C, Albert VA, Aono N, Aoyama T, Ambrose BA, Ashton NW, Axtell MJ, Barker E, Barker MS, Bennetzen JL, Bonawitz ND, Chapple C, Cheng C, Correa LGG, Dacre M, DeBarry J, Dreyer I, Elias M, Engstrom EM, Estelle M, Feng L, Finet C, Floyd SK, Frommer WB, Fujita T, Gramzow L, Gutensohn M, Harholt J, Hattori M, Heyl A, Hirai T, Hiwatashi Y, Ishikawa M, Iwata M, Karol KG, Koehler B, Kolukisaoglu U, Kubo M, Kurata T, Lalonde S, Li K, Li Y, Litt A, Lyons E, Manning G, Maruyama T, Michael TP, Mikami K, Miyazaki S, Morinaga S, Murata T, Mueller-Roeber B, Nelson DR, Obara M, Oguri Y, Olmstead RG, Onodera N, Petersen BL, Pils B, Prigge M, Rensing SA, Riaño Pachón DM, Roberts AW, Sato Y, Scheller HV, Schulz B, Schulz C, Shakirov EV, Shibagaki N, Shinohara N, Shippen DE, Sorensen I, Sotooka R, Sugimoto N, Sugita M, Sumikawa N, Tanurdzic M, Theissen G, Ulvskov P, Wakazuki S, Weng J-K, Willats WWGT, Wipf D, Wolf PG, Yang L, Zimmer AD, Zhu Q, Mitros T, Hellsten U, Loqué D, Otillar R, Salamov A, Schmutz

J, Shapiro H, Lindquist E, Lucas S, Rokhsar D, Grigoriev IV. 2011. The selaginella genome identifies genetic changes associated with the evolution of vascular plants. *Science* **332**:960–963 DOI 10.1126/science.1203810.

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH. 2008. Multiple paleopolyploidizations during the evolution of the compositae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution* **25**:2445–2455 DOI 10.1093/molbev/msn187.

Bowe LM, Coat G, DePamphilis CW. 2000. Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. *Proceedings of the National Academy of Sciences of the United States of America* **97**:4092–4097 DOI 10.1073/pnas.97.8.4092.

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**:433–438 DOI 10.1038/nature01521.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**:59–60 DOI 10.1038/nmeth.3176.

Cannon SB, McKain MR, Harkess A, Nelson MN, Dash S, Deyholos MK, Peng Y, Joyce B, Stewart CN, Rolf M, Kutchan T, Tan X, Chen C, Zhang Y, Carpenter E, Wong GK-S, Doyle JJ, Leebens-Mack J. 2015. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Molecular Biology and Evolution* **32**:193–210 DOI 10.1093/molbev/msu296.

Chaw SM, Parkinson CL, Cheng Y, Vincent TM, Palmer JD. 2000. Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. *Proceedings of the National Academy of Sciences of the United States of America* **97**:4086–4091 DOI 10.1073/pnas.97.8.4086.

Crisp MD, Cook LG. 2011. Cenozoic extinctions account for the low diversity of extant gymnosperms compared with angiosperms. *New Phytologist* **192**:997–1009 DOI 10.1111/j.1469-8137.2011.03862.x.

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert VA, Ma H, DePamphilis CW. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Research* **16**:738–749 DOI 10.1101/gr.4825606.

Deenen MHL, Ruhl M, Bonis NR, Krijgsman W, Kuerschner WM, Reitsma M, Van Bergen MJ. 2010. A new chronology for the end-Triassic mass extinction. *Earth and Planetary Science Letters* **291**:113–125 DOI 10.1016/j.epsl.2010.01.003.

Drewry A. 1988. The G-banded karyotype of Pinus resinosa Ait. *Silvae Genetica* **37**:218–221.

Foster CSP, Sauquet H, Van der Merwe M, McPherson H, Rossetto M, Ho SYW. 2017. Evaluating the impact of genomic data and priors on bayesian estimates of the angiosperm evolutionary timescale. *Systematic Biology* **66**:338–351 DOI 10.1093/sysbio/syw086.

Gensel PG, Andrews HN. 1984. *Plant life in the Devonian*. New York: Praeger.

Guo M, Liu JH, Ma X, Luo DX, Gong ZH, Lu MH. 2016. The Plant Heat Stress Transcription Factors (HSFs): structure, regulation, and function in response to abiotic stresses. *Frontiers in Plant Science* **7**:114–114 DOI 10.3389/fpls.2016.00114.

He Z, Zhang H, Gao S, Lercher MJ, Chen WH, Hu S. 2016. Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. *Nucleic Acids Research* **44**:W236–W241 DOI 10.1093/nar/gkw370.

Hu W, Hu G, Han B. 2009. Genome-wide survey and expression profiling of heat shock proteins and heat shock factors revealed overlapped and stress specific response under abiotic stresses in rice. *Plant Science* **176**:583–590 DOI 10.1016/j.plantsci.2009.01.016.

Jiao Y, Li J, Tang H, Paterson AH. 2014. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *The Plant Cell* **26**:2792 DOI 10.1105/tpc.114.127597.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE, Clifton SW, Schlarbaum SE, Schuster SC, Ma H, Leebens-Mack J, DePamphilis CW. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**:97–100 DOI 10.1038/nature09916.

Letunic I, Doerks T, Bork P. 2012. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Research* **40**:D302–D305 DOI 10.1093/nar/gkr931.

Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, Barker MS. 2015. Early genome duplications in conifers and other seed plants. *Science Advances* **1**:e1501084 DOI 10.1126/sciadv.1501084.

Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. 2016. Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* **28(2)**:326–344 DOI 10.1105/tpc.15.00877.

Li HT, Yi TS, Gao LM, Ma PF, Zhang T, Yang JB, Gitzendanner MA, Fritsch PW, Cai J, Luo Y, Wang H, Van der Bank M, Zhang S-D, Wang Q-F, Wang J, Zhang Z-R, Fu C-N, Yang J, Hollingsworth PM, Chase MW, Soltis DE, Soltis PS, Li D-Z. 2019. Origin of angiosperms and the puzzle of the Jurassic gap. *Nature Plants* **5**:461–470 DOI 10.1038/s41477-019-0421-0.

Lin Y, Cheng Y, Jin J, Jin X, Jiang H, Yan H, Cheng B. 2014. Genome duplication and gene loss affect the evolution of heat shock transcription factor genes in legumes. *PLOS ONE* **9**:e102825 DOI 10.1371/journal.pone.0102825.

Liu M, Huang L, Ma Z, Sun W, Wu Q, Tang Z, Bu T, Li C, Chen H. 2019. Genome-wide identification, expression analysis and functional study of the GRAS gene family in Tartary buckwheat (*Fagopyrum tataricum*). *BMC Plant Biology* **19**:342 DOI 10.1186/s12870-019-1951-3.

Lohani N, Golicz AA, Singh MB, Bhalla PL. 2019. Genome-wide analysis of the Hsf gene family in *Brassica oleracea* and a comparative analysis of the Hsf gene family in *B. oleracea, B. rapa* and *B. napus*. *Functional & Integrative Genomics* **19**:515–531 DOI 10.1007/s10142-018-0649-1.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**:1151–1155 DOI 10.1126/science.290.5494.1151.

Magallón S, Gómez-Acevedo S, Sánchez-Reyes LL, Hernández-Hernández T. 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytologist* **207**:437–453 DOI 10.1111/nph.13264.

Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TLL, Stadler T, Rabosky DL, Honeycutt RL, Flynn JJ, Ingram CM, Steiner C, Williams TL, Robinson TJ, Burk-Herrick A, Westerman M, Ayoub NA, Springer MS, Murphy WJ. 2011. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science* **334**:521–524 DOI 10.1126/science.1211028.

Nover L, Bharti K, Döring P, Mishra SK, Ganguli A, Scharf KD. 2001. Arabidopsis and the heat stress transcription factor world: how many heat stress transcription factors do we need? *Cell Stress & Chaperones* **6**:177–189 DOI 10.1379/1466-1268(2001)006¡0177:aathst¿2.0.co;2.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, Vicedomini R, Sahlin K, Sherwood E, Elfstrand M, Gramzow L, Holmberg K, Hällman J, Keech O, Klasson L, Koriabine M, Kucukoglu M, Käller M, Luthman J, Lysholm F, Niittylä T, Olson A, Rilakovic N, Ritland C, Rosselló JA, Sena J, Svensson T, Talavera-López C, Theißen G, Tuominen H, Vanneste K, Wu Z-Q, Zhang B, Zerbe P, Arvestad L, Bhalerao R, Bohlmann J, Bousquet J, Gil RG, Hvidsten TR, De Jong P, MacKay J, Morgante M, Ritland K, Sundberg B, Thompson SL, Van de Peer Y, Andersson B, Nilsson O, Ingvarsson PK, Lundeberg J, Jansson S. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**:579–584 DOI 10.1038/nature12211.

Ohama N, Sato H, Shinozaki K, Yamaguchi-Shinozaki K. 2017. Transcriptional regulatory network of plant heat stress response. *Trends in Plant Science* **22**:53–65 DOI 10.1016/j.tplants.2016.08.015.

Qiao X, Li M, Li L, Yin H, Wu J, Zhang S. 2015. Genome-wide identification and comparative analysis of the heat shock transcription factor family in Chinese white pear (*Pyrus bretschneideri*) and five other Rosaceae species. *BMC Plant Biology* **15**:12 DOI 10.1186/s12870-014-0401-5.

Ran JH, Shen TT, Wang MM, Wang XQ. 2018. Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. *Proceedings of the Royal Society B: Biological Sciences* **285**:20181012 DOI 10.1098/rspb.2018.1012.

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y, Tanahashi T, Sakakibara K, Fujita T, Oishi K, Shin IT, Kuroki Y, Toyoda A, Suzuki Y, Hashimoto S-I, Yamaguchi K, Sugano S, Kohara Y, Fujiyama A, Anterola A, Aoki S, Ashton N, Barbazuk WB, Barker E, Bennetzen JL, Blankenship R, Cho SH, Dutcher SK, Estelle M, Fawcett JA, Gundlach H, Hanada K, Heyl A, Hicks KA, Hughes J, Lohr M, Mayer K, Melkozernov A, Murata T, Nelson DR, Pils B, Prigge M, Reiss B, Renner T, Rombauts S, Rushton

PJ, Sanderfoot A, Schween G, Shiu S-H, Stueber K, Theodoulou FL, Tu H, Van de Peer Y, Verrier PJ, Waters E, Wood A, Yang L, Cove D, Cuming AC, Hasebe M, Lucas S, Mishler BD, Reski R, Grigoriev IV, Quatrano RS, Boore JL. 2008. The Physcomitrella Genome reveals evolutionary insights into the conquest of land by plants. *Science* **319**:64–69 DOI 10.1126/science.1150646.

Sanderson MJ, Donoghue MJ. 1994. Shifts in diversification rate with the origin of angiosperms. *Science* **264**:1590–1593 DOI 10.1126/science.264.5165.1590.

Sauquet H, Magallón S. 2018. Key questions and challenges in angiosperm macroevolution. *New Phytologist* **219**:1170–1187 DOI 10.1111/nph.15104.

Scharf KD, Berberich T, Ebersberger I, Nover L. 2012. The plant heat stress transcription factor (Hsf) family: structure, function and evolution. *Biochimica et Biophysica Acta* **1819**:104–119 DOI 10.1016/j.bbagrm.2011.10.002.

Schranz ME, Mohammadin S, Edger PP. 2012. Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model. *Current Opinion In Plant Biology* **15**:147–153 DOI 10.1016/j.pbi.2012.03.011.

Soltis DE, Visger CJ, Soltis PS. 2014. The polyploidy revolution then…and now: stebbins revisited. *American Journal of Botany* **101**:1057–1078 DOI 10.3732/ajb.1400178.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313 DOI 10.1093/bioinformatics/btu033.

Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. 2008. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Research* **18**:1944–1954 DOI 10.1101/gr.080978.108.

Thalmann M, Coiro M, Meier T, Wicker T, Zeeman SC, Santelia D. 2019. The evolution of functional complexity within the β-amylase gene family in land plants. *BMC Evolutionary Biology* **19**:66 DOI 10.1186/s12862-019-1395-2.

Wang X, Shi X, Chen S, Ma C, Xu S. 2018. Evolutionary origin, gradual accumulation and functional divergence of heat shock factor gene family with plant evolution. *Frontiers in Plant Science* **9**:71 DOI 10.3389/fpls.2018.00071.

Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, Th Lee, Jin H, Marler B, Guo H, Kissinger JC, Paterson AH. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40**:e49-e49 DOI 10.1093/nar/gkr1293.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, Ruhfel BR, Wafula E, Der JP, Graham SW, Mathews S, Melkonian M, Soltis DE, Soltis PS, Miles NW, Rothfels CJ, Pokorny L, Shaw AJ, DeGironimo L, Stevenson DW, Surek B, Villarreal JC, Roure B, Philippe H, DePamphilis CW, Chen T, Deyholos MK, Baucom RS, Kutchan TM, Augustin MM, Wang J, Zhang Y, Tian Z, Yan Z, Wu X, Sun X, Wong GK-S, Leebens-Mack J. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences of the United States of America* **111**:E4859–E4868 DOI 10.1073/pnas.1323926111.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**:1586–1591 DOI 10.1093/molbev/msm088.

Zeng L, Zhang Q, Sun R, Kong H, Zhang N, Ma H. 2014. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nature Communications* **5**:4956 DOI 10.1038/ncomms5956.