Deep Evaluation to the Evolution History of Heat Shock Factor (HSF) Gene Family and Its Expansion Pattern in Seed Plants

Yiying Liao
Fairy Lake Botanic Garden
Zhiming Liu
Fairy Lake Botanic Garden
Andrew W. Gichira
Chinese Academy of Sciences
Min Yang
Fairy Lake Botanic Garden
Ruth Wambui Mbichi
Chinese Academy of Sciences
Dan Yang
Capitalbio Genomics Ltd
Qingfeng Wang
Chinese Academy of Sciences
Tao Wan (wantao1983@gmail.com)
Key Laboratory of Southern Subtropical Plant Diversity, Fairy Lake Botanical Garden, Shenzhen & Chinese Academy of Science, Shenzhen 518000, China.

Research article

Keywords: Heat shock factor, HSFs, basal angiosperms, eudicots, monocots, gene duplication, diversification

DOI: https://doi.org/10.21203/rs.3.rs-70764/v1

Abstract

Background

HSF (Heat shock factor) genes are essential in the irreplaceable functions in some of the basic developmental pathways in plants. Despite the extensive studies on the structure, function diversification, and evolution of HSF, their divergent history and gene duplication pattern remain unsolved. To further illustrate the probable divergent patterns in these subfamilies, we visited the evolutionary history of the HSF via phylogenetic reconstruction and genomic syntenic analyses by taking advantage of the increased sampling of genomic data for pteridophyta, gymnosperms and basal angiosperms.

Results

We identified a novel clade including HSFA2, HSFA6, HSFA7, HSFA9 with complex relationship, very likely due to orthologous or paralogous genes retained after frequent gene duplication events. We suggested that HSFA9 was derived from HSFA2 through gene duplication in eudicots at ancestral state, and then expanded in a lineage-specific way. Our findings indicated that HSFB3 and HSFB5 emerged before the divergence of ancestral angiosperms, but were lost in common ancestors of monocots. We also presumed that HSFC2 was derived from HSFC1 in ancestral monocots.

Conclusion

This work proposes that in the era of early differentiation of angiosperms during the radiation of flowering plants, the member size of HSF gene family was also being adjusted, accompanied 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 evolution history of multigene family.

Background

Heat shock factors (HSFs), as the central regulators of the expression of heat shock proteins and other heat shock-induced genes, play crucial roles in the enhancement of thermo tolerance in plants. They function as molecular chaperones in protein folding and assembly, to protect cells against proteotoxic damage under heat stress (HS) ([1–3]. Besides their involvement in HS (Heat Shock) response, HSFs had been identified in most eukaryotes and non-plant organisms, where they participate in growth and development [4,5]. In plants, especially in angiosperms, HSFs have also been widely studied as essential elements to cope with various environmental stresses [5]. This gene family has expanded greatly such that the number of HSF genes ranges from one or two in green algae, to more than 50 in angiosperms [6]. The HSFs generally contain the DNA binding domain (DBD), the oligomerization domain (OD), and a flexible linker between DBD and OD regions [5,7]. Based on the topology of these domains, HSFs are classified into three groups: HSFA, HSFB and HSFC. These groups are further divided into 16 subfamilies

which are distinguished in angiosperms, including HSFA (A1-A9) group, HSFB (B1-B5) group and HSFC (C1-C2) group [1,5,8,9]. The first overview of HSFs was presented in *Arabidopsis thaliana* by Nover et al. [8], in which HSFC was discovered. Afterwards, valuable summaries including data from nine angiosperms species and over 50 species containing all lineages of land plants retrospect the structure, function and evolution of HSFs have been compiled [5,6]. These reports pointed out that the HSF family members and their functions diverged greatly among the higher plant lineages in response to environment stresses. However, the evolutionary relationships among the subfamilies are still obscure as some of the deepest nodes of the HSFs phylogeny tree, such as position of HSFB5 and HSFA9 remain elusive. These would be attributed to the limited access to complete HSFs data in representative seed plant lineages, including gymnosperms and basal angiosperms. It may also be influenced by the unpredictable gene copy turnover after recurring gene duplication events in tandem or genome-wide.

Here, we expanded the data collection to basal angiosperms, gymnosperms, and pteridophyte to reconstruct the diversification history of HSFs during seed plants evolution. Moreover, we detected syntenic relationships of HSFs across a wide range of species, thus providing crucial information to address fundamental questions on the evolution of gene families. We also predicted the divergence time of typical genes derived from their ancestors, based on the reliable gene orthology. Our results present critical evidence for explaining the expansion of the HSF subfamilies in seed plant lineages.

Results

The phylogeny and evolution of HSFs in land plants

We identified 287 new candidate HSF sequences from 24 species on the HEATSTER website, of which 228 belonged to known subfamilies (A1-A9, B1-B5, C1-C2), 59 belonged to HSF like (Table 1). In the data sets download from Heaster, 442 sequences were classified. A total of 670 HSF sequences (Table S4) from 44 species, were used for phylogenetic analysis. Across the studied comprehensive samples, the identified number of HSF gene subfamilies is greatly variable, ranging from 2 in chlorophyta to 30 in angiosperms. The unrooted phylogenetic tree inferred from amino acid sequences resolved three main clades, HSFA-HSFC, HSFB, and HSFC (Fig. 1). The newly identified HSF genes were reconfirmed on a phylogeny tree. Most subfamilies of clades (A3, A4, A5, A8, A9, B2, B3, B5, C1, C2) were strongly supported, while the relationships between these clades were weakly supported. The HSF subfamilies display strong diversification in structure, composition and function [5–7], thus, significant genetic differentiation between clades especially for HSFA and HSFB probably resulted in the unstable topology. Group HSFA was observed in all sampled taxa, however group HSFB was absent in chlorophyta, while group HSFC was only present in angiosperms.

The HSFA group, which contains major regulators in the HS response of plants [6], has undergone diversification during plant evolution hence the result of classification displayed variations in different taxa. Strikingly, subfamilies clades (A4, A5, A6, A7, A8, A9) were only identified in angiosperms, whereas A9 was merely identified in Eudicots. In detail, some subfamilies clustered on a single branch, such as A3,

A5, A9, while others were clustered on several branches (Fig. S1, Fig. S2). HSFA1, as a master regulator which cannot be replaced by any other HSF [5], is likely to be considered as the most ancient group in HSFA. Although all HSFA1 and HSFA8 of angiosperms clustered as a clade, most of HSFA1 from pteridophyte and gymnospermae were dispersed into several clades. The deep divergence of HSFA1 in pteridophyta and gymnospermae, indicated that it diversified before the radiation of seed plants. HSFA2, HSFA6, HSFA7, and HSFA9 were blended into a complex clade, and while HSFA9 formed monophyletic groups, the others remained unclear. It is interesting to observe that HSFA2 gene and HSFA6 gene clustered together with very little genetic difference in some angiosperm species such as *O. sativa*, *Phoenix dactylifera*, and *Citrullus lanatus*, the same as HSFA6 gene and HSFA7 gene in *C. lanatus*. Two HSFA5 genes clustered with HSFA4 genes, indicating a close relationship between the two clades. It had been previously suggested that subclass HSFA3 and group HSFC form a cluster. However, due to increased number of ferns and gymnosperm used in this study, the HSFA1 of the gymnosperms and HSFC, rather than HSFA3, formed a single cluster (Fig. S1, Fig. S2). It is assumed that a duplication event occurred in the ancestral angiosperms which could have contributed to the rise of HSFC.

The numbers of genes in HSFC1, as a common subfamily, varied between monocots, which had typically two members, and the eudicots with only one member (Table 1, Fig. S2). The results indicated that the HSFC experienced a steady expansion during the evolution of monocots, and may be involved in important developmental pathways [6]. Notably, HSFC2 was only present in monocots, but HSFC1 was in all angiosperm species except for A. trichopoda. In monocots, both HSFC1 and HSFC2 formed a strongly supported cluster. HSFC1-HSFC2 clade of monocots and HSFC1 of eudicots formed a branch, based with HSFC1 of basal angiosperms. This suggested that HSFC2 are a result of recent duplications occurring in the early stages of divergence between monocots and eudicots.

Contrary to a previous study [6], the results presented here suggested that HSFB subfamily (B1, B2, B3, B4, B5) were moderately supported as a monophyletic group (Fig. S1, Fig. S3). While HSFB1, HSFB2, and HSFB4 have widely been observed across land plants, both HSFB3 and HSFB5 are only present in eudicots, and basal angiosperms. Unlike other subfamily members of the HSFB groups, the HSFB5 has a conserved tetrapeptide LFGV in the C-terminal domain, thus 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 average number reduced from 3.4 in gymnosperms to 1.2 in angiosperms (Table 1). In particular, the gene number of HSFB1 in conifers (*Picea abies*, *Pinus taeda*, *Picea glauca*) is significantly higher than that of other seed plants. Multiple copies of HSFB1 gene of *P. abies*, *P. taeda*, and *P. glauca* clustered and formed a strongly supported monophyletic group. This result indicated that the evolution of those three conifers probably involved both polyploidy and repetitive element activity [10–12]. The multi-copy genes may be attributed to two whole genome duplication (WGD) events in the ancestry of major conifer clades [11]. Though many angiosperm lineages have experienced additional rounds of genome duplication [11,13,14], the number of HSFB1 in angiosperms displayed no significant increase which is consistent with speculation that WGD in angiosperms did not give rise to remarkable expansion of HSFB1 genes. The HSFB1 group in angiosperms, gymnosperms and pteridophyta clustered independently, suggesting that HSFB1 was an ancient group that diverged during the evolutionary history of the different taxa. In

gymnosperms, the HSFB1 group experienced several expansions including ancient duplication that were generally rare in angiosperms except for few recent duplicates. All HSFB2 of gymnosperms and angiosperms formed different clusters. We could not trace any 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*, it was observed that some genes identified as different subfamilies such as HSFB1 and HSFB4 have high genetic similarity with highly supported clades. The complicated relationship of those two subfamilies may be as a result of involvement of recent duplication events. In this study, subfamilies HSFB3 and HSFB5 were only present in eudicots and basal angiosperms probably as a result of duplication events occurring in ancestral angiosperms, but the paralogue genes were lost in monocots.

Gene duplication analysis

To examine expansion patterns, genetic divergences, and identify gene duplication events that affected the evolution of genomes in the HSF gene family, synteny analysis was performed across twenty-one species (Table S3). The synteny analysis between different species was also conducted on pairwise species which were closely related taxa.

Gene duplication events were identified in eleven species among the pteridophyta, basal angiosperms, monocots and eudicots (Table 2, Table S5, Table S6, Table S7). We did not detect any synteny blocks of the HSFs gene among the green alga, moss, and gymnosperm. This result indicate that the ancient HSF gene duplications were not easy to detect, because most duplicates had been lost. In *S. moellendorffii*, the only fern, we identified one pair of duplicated genes. The two genes belong to different subclasses of HSF gene family;'SelmoHSFB1b' and'SelmoHSFB4' which were calculated as syntenic to each other. It is speculated that they may be derived from an ancient tandem duplication and evolved a certain degree of difference at the gene sequence level. In *L. chinense*, the only basal angiosperm, we identified five pairs of duplicated genes out of which members of each of four pairs belonged to the same gene subclass (HSFA2, HSFB1, HSFB2, HSFC1) while genes in the other pair belonged to different subclasses (HSFA4-HSFA5). Gene duplication events were detected in all sampled eudicot and monocot species (Table S7). In five eudicots (*A. thaliana*, *Populus trichocarpa*, *Prunus persica*, *S. lycopersicum*, *Mimulus guttatus*), we identified a total of thirty-three pairs of duplicated genes out of which members of each of twenty-nine pairs belonged to the same gene subclasses (HSFA1, HSFA4, HSFA5, HSFA6, HSFA8, HSFB2, HSFB3, HSFB4, HSFB5) while genes in each of the remaining pairs belonged to different subclasses (HSFA2-HSFA9, HSFA6-HSFA7). The same number of gene duplications were identified in four monocots (*O. sativa*, *Sorghum bicolor*, *Z. mays*, *Brachypodium distachyon*), however, members of each of the twenty-nine pairs belonged to the same gene subclasses (HSFA1, HSFA2, HSFA4, HSFA6, HSFB1, HSFB2, HSFB4, HSFC1, HSFC2) while genes in each of the remaining four pairs belonged to different subclasses (HSFA2-HSFA6, HSFB1-HSFB2, HSFB2-HSFB4). In general, all subfamilies of HSF genes except HSFA3, were involved in duplication events. The results demonstrated that, pairs of genes from different subfamilies such as HSFA2-HSFA6, HSFA2-HSFA9, HSFA4-HSFA5, HSFA6-HSFA7, HSFB1-HSFB4, HSFB1-HSFB2, and HSFB2-HSFB4, were paralogous.

The synteny analysis between different species detected orthologous genes in different taxa (Table 3, Table S8, Table S9). Between two gymnosperm species (*G. montanum*, *G. biloba*), only HSFA1 genes from different sources were detected as orthologous genes. The analysis detected that genes HSFA1, HSFA4, and HSFA5 were detected as orthologous in *G. biloba* (gymnosperms) and L. chinense (basal angiosperms). Though we found several orthologous genes, such as HSFA6-HSFA7, HSFA4-HSFA5, HSFA2-HSFA9, and HSFB2-HSFB5, among the basal angiosperms (*A. trichopoda*, *L. chinense*) and eudicots (*S. lycopersicum*, *A. thaliana*), only orthologous genes HSFA2-HSFA6, and HSFA2-HSFA7 were detected among basal angiosperms and monocots (*O. sativa*, *Z. mays*). Interestingly, the analysis of eudicots-monocots got the same results as basal angiosperms-monocots. Orthologous genes HSFA1-HSFA5, and HSFA2-HSFA7 were found in monocots, while HSFA6-HSFA7 was detected in eudicots.

On the whole, the results indicate that duplication of the HSF genes has been a common event during the evolution of plants, significantly contributing to the expansion and functional diversification (Fig. 2). Thus, it is suggested that HSFA4 and HSFA5 have a close genetic relationship, and their origin may be related to ancient duplication of HSFA1. It is possible that HSFA6 and HSFA7 originated from gene duplication, most probably involving HSFA2. HSFA9 was proved to be derived from HSFA2 after the divergence of ancestral angiosperms. Moreover, HSFB1 is considered as the most ancient among HSFB, and we predict that HSFB2 and HSFB4 were derived from HSFB1 considering the close relationship between them.

Divergence time analysis

The estimated divergence time of subfamilies HSFA2 and HSFA9 in eudicots ranges between 131 Mya and 155.2 Mya, which is within the Late Jurassic and Lower Cretaceous periods (Fig. 3). The estimated split time of clade HSFC2 and HSFC1 in monocots ranges between 125 Mya and 190.4 Mya which is within the Jurassic and Lower Cretaceous (Fig. 4). The occurrence time of those gene duplication events were consistent with the most recent common ancestor of all living angiosperms, which likely existed ~ 140–250 Mya [15,16]. Although uncertainty remains for other characters, our reconstruction of 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 [17], and research in the evolution of gene families (such as the origin, expansion and loss of genes) provides an unprecedented opportunity to explore remarkable long-standing questions that probably hold important clues to understand present-day biodiversity and adaption to environment.

Discussion

Previous phylogenetic studies of HSF gene family in plants, have provided valuable insights into their evolutionary history [5,6]. However, the limited sampling in pteridophyta, gymnosperms and basal angiosperms left unresolved questions relating to the origin of subclasses in HSF gene family, their phylogenetic relationship, and gene expansion patterns in different taxa. HSFs play a key role in plants adaptation to the changing habitat and overcoming stresses. However, our understanding of land plants

evolution at a genetic level and in relation to environmental changes is also obscure [5,6,18–23]. Although ongoing plant genome projects will certainly uncover additional species or family-specific deletions and duplications, the general features are unlikely to change [24]. In this study, the number and diversity of plants examined allowed us to raise the question of the evolutionary history of this gene family in a broader taxonomic context. Our phylogenetic analyses revealed divergence of subfamilies of HSFs and independent evolution in plants, especially in angiosperms. With the increased number of simultaneously analyzed genomes, it is becoming more difficult to organize and display such syntenic relationships. This is due to the ubiquity of ancient and recent polyploidy events, as well as smaller scale events that derive from tandem and transposition duplications [25–28]. However, thanks to a combination of phylogeny and synteny analyses in this study, the results had scratched the surface of just how gene expansion in different land plant taxa occurred. It proved that puzzling clades (HSFA2, HSFA6, HSFA7, HSFA4, HSFA5) with members from other group snuck in caused by recent gene duplication events.

Our studies on different members of the HSF gene family, from pteridophyta and gymnosperm, reveal that this gene family is quite complex in terms of gene number and sequence diversity. We identified four subfamilies of HSFs (HSFA1, HSFA2, HSFB1, HSFB4) across candidates of six species in pteridophyta, and five subfamilies (HSFA1, HSFA2, HSFB1, HSFB2, HSFB4) from sixteen species of gymnosperms. Though the number of HSFs in pteridophyta and gymnosperm is significantly less than angiosperms, the number of HSFA1 and HSFB1 in those taxa were higher than angiosperms. It was assumed that pteridophyta and gymnosperm preferred to reserve more ancient members in HSFs subfamily. Both subfamily HSFA1 and HSFB1, in pteridophyta and gymnosperm, clustered on multiple clades on the phylogeny tree with low support, which is consistent with findings that more ancient duplication events affect more distant taxonomic comparisons [26]. The syntenic analysis detected only two genes (SelmoHSFB1b and SelmoHSFB4) in *S. moellendorffii*, which appeared to be a result of duplication events. These findings indicate that HSFA1, HSFA2, HSFB1 and HSFB4 which are already present in the ancestor of all land plants, were more ancient groups.

Gymnosperms lineages varied from each other during the Late Carboniferous to the Late Triassic, and were dominant through most of the Mesozoic [29,30]. However, major gymnosperm extinctions occurred in the Cenozoic, and in contrast with angiosperms, the surviving gymnosperm genera have diversified slower than angiosperms [31]. Ancient gene subfamilies such as HSFB1 and HSFA1 accompanying plants experienced the long time differentiation and the successional variation process, which may explain the molecular phylogenetic uncertainty within gymnosperm. Ancient WGD is found in the ancestry of all extant seed plants, and angiosperm and gymnosperm lineages have experienced additional rounds of genome duplication [11,13,14,32–34]. Although no syntenic gene was detected in gymnosperms, two or more genes from different subclasses formed strongly supported clades (such as PintaHSFA1a and PintaHSFA2, AbifiRHSFB1a and AbifiRsfB4a), so the absence of syntenic gene in gymnosperm may be a result of incomplete genome information, or assembly and annotation problems. The ancient interspersed segmental duplication of those genes in recent times could be detected by phylogenetic analysis and synteny analysis.

Comparably in angiosperms, this gene family has undergone extensive duplications that gave rise to complicated relationships of orthology, paralogy, and functional heterology. The results showed that, besides having a remarkably higher number and diversity of HSF family than in older taxa, angiosperms had multiple paralogue and orthology genes. Most of the gene copies generated by WGD events are lost due to fractionation and subsequent “rediploidization” or non-functionalization [35]. Gene duplication is an important mechanism that contributes to genomic novelty [11], and the functional divergence of duplicate genes retained from WGD is thought to promote evolutionary diversification. The multiple recent WGDs occurring in angiosperms lineages allowed the expansion and variation of HSFs, as confirmed by previous studies in *Fagopyrum tataricum* [36], and genus *Brassica* [21]. The results of synteny analysis confirmed that subfamily HSFA9, only present in eudicots, was derived from HSFA2, and that HSFC2, only present in monocots was derived from HSFC1. New genes originated from divergence of paralogue genes which resulted from duplication events. The two duplication events occurred in the early stages of angiosperm divergence, which is consistent with angiosperm radiations that occurred in the Late Jurassic and Lower Cretaceous [37]. Approximately 132 mya ago, angiosperms underwent massive adaptive radiations to become the most diverse and successful plant group on land [38]. The coincidence of retained duplication events with key moments in the evolution of biological innovations and survival in the face of mass born underlines the importance of this crucial process [39]. Lineage-specific duplications will provide the keys to understand both common underlying regulatory mechanisms and the species-specific differences that generate diversity. The subfamily HSFB3 and HSFB5 were found to be absent in monocots, but present in most of basal angiosperms and eudicots. Consequently, we deduce that HSFB3 and HSFB5 were thoroughly lost in monocots, nevertheless, their origin and evolutionary history remain poorly understood. We speculate that those gene loss events took place during the early divergence time in the angiosperm history. The above results indicate that not only did the species experience early rapid radiation, diversification, and mass extinction [37, 40–43], their genes also went through expansion, diversification, and loss. After divergence of angiosperms, eudicots and monocots experienced different evolutionary processes.

The recent upward trend in number of completely sequenced genomes of plants in different phylogenetic lineages has advanced our evolutionary understanding of gene families with important functions. Our comprehensive analysis reveals that the diversification of HSFs in plants was as a result of extensive gene duplications, gene loss and sub- or neo-functionalization 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 plasticity and flexibility in complex and changing environments. The patterns of gene duplication and evolution history of HSFs in plants provide novel insight into their diversity which facilitate plant diversification, adaptation and evolution in various habitats. Our analyses provide essential insights for studying evolution history of multigene families.

Conclusions

The recent upward trend in number of completely sequenced genomes of plants in different phylogenetic lineages has advanced our evolutionary understanding of gene families with important functions. Our

comprehensive analysis reveals that the diversification of HSFs in plants was as a result of extensive gene duplications, gene loss and sub- or neo-functionalization 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 plasticity and flexibility in complex and changing environments. The patterns of gene duplication and evolution history of HSFs in plants provide novel insight into their diversity which facilitate plant diversification, adaptation and evolution in various habitats. Our analyses provide essential insights for studying evolution history of multigene families.

Methods

Identification of HSFs and phylogenetic analysis

We sampled 23 species representing three main taxa; peridophyta, gymnospermae and basal angiosperms (Table S1). The data included 7 genomes and 17 transcriptomes. 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., [44]). For genome assemblies, multiple databases were screened including ConGenIE (http://congenie.org/), GigaDB (http://gigadb.org/dataset/100209), Dryad (https://datadryad.org/stash/dataset/), WaterlilyPond (http://waterlily.eplant.org), FernBase (ftp://ftp.fernbase.org/), and *Liriodendron chinense* Database (http://120.78.193.56:8000/). To increase the reliability of the data, both genome and transcriptome data of *Ginkgo biloba* were analyzed in this study. We used the predicted proteome of each genome as a query to search for HSF-type DBD domain (HSF_DNA_bind_PF00447) from Pfam-A.hmm (Pfam release 32.0) using software pfamscan (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/Tools/), which were considered as candidate genes. The amino acid sequences of candidate HSFs were then extracted. We also downloaded 537 plant 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 search for analyzed species, to further identify candidate HSF proteins. For those candidate sequences, we examined the facticity of DBDs and ODs by using SMART 7 software [45] (http://smart.embl-heidelberg.de/), and also identified further using HEATSTER web site (https://applbio.biologie.uni-frankfurt.de/hsf/heatster/). Candidate proteins without integrated DBD and HR-A/B domains were removed.

The alignments of the identified and classified candidate genes excluded Hsf like sequences were performed by MUSCLE (http://www.drive5.com/muscle). Phylogenetic analyses were conducted by RAxML version 8.0.19 [46], with 100 bootstraps, PROTGAMMAAUTO model, and Maximum likelihood reconstruction using the new rapid hill-climbing and rapid Bootstrap analysis (-f ad). Phylogenetic trees were examined and manipulated with Evolview v2 [47]. To better understand the evolutionary relationship within the subfamilies, and for an in-depth phylogenetic analyses of the HSFB, and HSFA-HSFC clades, all the identified HSFB, HSFA and HSFC genes were respectively used for phylogenetic tree reconstruction. Also, for better understanding of the complicated evolutionary relationship between the clade of subfamilies HSFA2, HSFA6, HSFA7, HSFA9 and the HSFC clade, those two group of genes were respectively extracted for construction of phylogenetic trees, with *Chlamydomonas reinhardtii* as an

outgroup. The methods of multiple proteins sequence alignments and phylogenetic analyses followed the same steps as described above.

Synteny analysis and molecular dating analyses

Further investigations of the evolutionary relationship of HSFs between the main taxa of plants, were made using MCScanX [48] in order to detect the gene replication events. Twenty-one plant genomes were subjected to 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. Genome annotation and corresponding protein sequences were downloaded for each species. Paralogous and orthologous genes in or between those genomes were identified through synteny detection by using MCScanX with default parameters (minimum match size for a collinear block = 5 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. All -vs- all protein sequence comparisons necessary for MCScanX were performed using DIAMOND v 0.8.25 [49]. The gene list containing all candidate HSF genes was queried against the “Total_Synteny_Blocks” file. From the results, we checked whether or not HSFs genes existed in the syntenic block. For synteny analysis between species on close taxa, eight representative species were chosen as follows; gymnosperms (*Gnetum montanum*, *Ginkgo biloba*), basal angiosperms (*L. chinense*, *Amborella trichopoda*), Monocots (*Oryza sativa*, *Zea mays*), Eudicots (*A. thaliana*, *Solanum lycopersicum*). The methods and procedures were the same as previously stated.

The subfamily HSFC1 and HSFC2 genes, and subfamily HSFA2 and HSFA9 genes were extracted from database and used for estimating the divergence time respectively. We calibrated a relaxed molecular clock prior on the node with the divergence time of monocots and eudicots between 140 Mya (a minimum age) and 200 Mya (a maximum age) (represented by the divergence of *A. thaliana* and *O. sativa*; [50]. We performed a Bayesian dating analysis in MCMC tree [51], using 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 2 steps over 200, 000 steps following a burn-in of 10,000 steps.

Abbreviations

HS
Heat stress;DBD:DNA binding domain□OD:Oligomerization domain□WGD:Whole genome duplication

Declarations

Ethics approval and consent to participate

Not applicable.

Consent to publish

Not applicable.

Availability of data and materials

The dataset supporting the results of this article is available as Additional files.

Competing interests

The authors declare that they have no competing interests.

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), 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, and preparation of the manuscript.

Author's contributions

TW and QFW designed the study and modified manuscript. YYL conducted the sequence analyses and drafted the manuscript. ZML, DY and MY performed the experiments and analyzed the data. RWM improved the language. All authors read and approved the final manuscript.

Acknowledgments

The authors sincerely thank Meiping Wang for the assistance with data collection.

References

1. Ahuja, & Raj. M. (2005). Polyploidy in gymnosperms: revisited. Silvae Genetica, 54(1-6), 59-69. https://doi.org/10.1515/sg-2005-0010
2. Ahuja, I., de Vos, R. C., Bones, A. M., & Hall, R. D. (2010). Plant molecular stress responses face climate change. Trends in plant science, 15(12), 664-674. https://doi.org/10.1016/j.tplants.2010.08.002
3. Airoldi, C.A., and Davies, B. (2012). Gene duplication and the evolution of plant MADS-box transcription factors. Journal of Genetics and Genomics, 39: 157–165. https://doi.org/10.1016/j.jgg.2012.02.008
4. Akerfelt, M., Morimoto, R. I., & Sistonen, L. (2010). Heat shock factors: integrators of cell stress, development and lifespan. Nature reviews Molecular cell biology, 11(8), 545-555. https://doi.org/

10.1038/nrm2938

5. Albert, V. A., Barbazuk, W. B., Der, J. P., Leebens-Mack, J., Ma, H., and Palmer, J. D. (2013). The amborella genome and the evolution of flowering plants. Science 342:1241089. https://doi.org/10.1126/science.1241089
6. Banks, J. A., Nishiyama, T., Hasebe, M., Bowman, J. L., Gribskov, M., dePamphilis, C., et al. (2011). The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. Science 332, 960–963. https://doi.org/ 10.1126/science.1203810
7. Barker, M. S., Kane, N. C., Matvienko, M., Kozik, A., Michelmore, R. W., Knapp, S. J., & Rieseberg, L. H. (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(11), 2445-2455. https://doi.org/10.1093/molbev/msn187
8. Bowe, L. M. , Coatt, G. , & Depamphilis, C. W. . (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, 97(8), 4092-4097. https://doi.org/10.1073/pnas.97.8.4092
9. Bowers, J., Chapman, B., Rong, J. et al. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422, 433-438 (2003). https://doi.org/10.1038/nature01521
10. Buchfink, B. , Xie, C. , & Huson, D. H. . (2014). Fast and sensitive protein alignment using diamond. Nature Methods, 12(1), 59-60. https://doi.org/10.1038/nmeth.3176
11. Cannon, S. B., McKain, M.R., Harkess, A., Nelson, M. N., Dash, S., Deyholos, M.K., & Kutchan, T. (2015) Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. Molecular Biology and Evolution. 32, 193-210. https://doi.org/10.1093/molbev/msu296
12. Chaw, S.-M., Parkinson, C. L., Cheng, Y., Vincent, T.M & Palmer, J. D. (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. 97, 8, 4086-4091. https://doi.org/10.1073/pnas.97.8.4086
13. Crisp, M. D., & Cook, L. G. (2011). Cenozoic extinctions account for the low diversity of extant gymnosperms compared with angiosperms. New Phytologist, 192(4), 997-1009. https://doi.org/10.1111/j.1469-8137.2011.03862.x
14. Cui, L., Wall, P. K., Leebens-Mack, J. H., Lindsay, B. G., Soltis, D. E., Doyle, J. J., Soltis, P. S., Carlson, J. E., Arumuganathan, K., Barakat, A., Albert, V. A., Ma, H., & dePamphilis, C. W. (2006) Widespread genome duplications throughout the history of flowering plants. Genome research, 16(6), 738-749. http://www.genome.org/cgi/doi/10.1101/gr.4825606
15. Deenen, M. H. L. , Ruhl, M. , Bonis, N. R. , Krijgsman, W. , Kuerschner, W. M. , & Reitsma, M. , et al. (2010). A new chronology for the end-triassic mass extinction. Earth & Planetary Science Letters, 291(1-4), 0-125. https://doi: 10.1016/j.epsl.2010.01.003

16. Drewry, A. (1988). The G-banded karyotype of Pinus resinosa Ait. Silvae Genet, 37, 218-221.
17. Foster, C. S. P. , Sauquet Hervé, Marlien, V. D. M. , Hannah, M. P. , Maurizio, R. , & Ho, S. Y. W. . (2017). Evaluating the impact of genomic data and priors on bayesian estimates of the angiosperm evolutionary timescale. Systematic Biology(3), 3. https://doi.org/ 10.1093/sysbio/syw086
18. Gensel, P. G. &H. N. Andrews. 1984. Plant life in the Devonian. Praeger, New York.
19. Guo, M., Liu, J. H., Ma, X., Luo, D. X., Gong, Z. H., & Lu, M. H. (2016). The plant heat stress transcription factors (HSFs): structure, regulation, and function in response to abiotic stresses. Frontiers in plant science, 7, 114. https://doi.org/10.3389/fpls.2016.00114
20. He, Z., Zhang, H., Gao, S., Lercher, M. J., Chen, W. H., & Hu, S. (2016). Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. Nucleic acids research, 44(W1), W236-W241. https://doi.org/10.1093/nar/gkw370
21. 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(4), 583-590.. https://doi. org/10.1016/j.plantsci.2009.01.016
22. Jiao, Y., Li, J., Tang, H., & Paterson A. H. (2014) Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. the Plant Cell, 26 (7), 2792-2802. https://doi.org/10.1105/tpc.114.127597
23. Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., Soltis, D.E., Clifton, S.W., Schlarbaum, S.E., Schuster, S.C., Ma, H., Leebens-Mack, J., & dePamphilis, C.W. (2011). Ancestral polyploidy in seed plants and angiosperms. Nature, 473(7345), 97-100. https://doi.org/10.1038/nature09916
24. Letunic I., Doerks T., and Bork P., 2012, SMART 7: recent updates to the protein domain annotation resource, NUCLEIC ACIDS RESEARCH, 40: 302-305. http://smart.emblheidelberg.de/
25. Li, H., Yi, T., Gao, L. et al. Origin of angiosperms and the puzzle of the Jurassic gap. Nature Plants 5, 461–470 (2019). https://doi.org/10.1038/s41477-019-0421-0
26. Li, Z. , Baniaga, A. E. , Sessa, E. B. , Scascitelli, M. , Graham, S. W. , & Rieseberg, L. H. , et al. (2015). Early genome duplications in conifers and other seed plants. Science Advances, 1(10), e1501084-e1501084. https://doi.org/10.1126/sciadv.1501084
27. 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(7). https://doi.org/10.1371/journal.pone.0102825
28. Lohani, N., Golicz, A. A., Singh, M. B., & Bhalla, P. L. (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(3), 515-531. https://doi.org/10.1007/s10142-018-0649-1
29. Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155. https://doi.org/10.1126/science.290.5494.1151

30. Meredith, R. W. , Janecka, J. E. , Gatesy, J. , Ryder, O. A. , Fisher, C. A. , & Teeling, E. C. , et al. (2011). Impacts of the cretaceous terrestrial revolution and kpg extinction on mammal diversification. Science, 334(6055), 521-524. https://doi.org/10.1126/science.1211028
31. Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L. & Hernández-Hernández, T. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytologist. 207, 437-453 (2015). https://doi.org/10.1111/nph.13264
32. Nover, L., Bharti, K., Doring, P., Mishra, S. K., Ganguli, A., and Scharf, K. D. (2001). Arabidopsis and the heat stress transcription factor world: how many heat stress transcription factors do we need? Cell stress & chaperones, 6(3), 177. https://doi.org/10.2307/1601759
33. Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y. C., Scofield, D. G., et al. (2013). The Norway spruce genome sequence and conifer genome evolution. Nature 497, 579-584. https://doi.org/10.1038/nature12211
34. 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. https://doi.org/10.1016/j.tplants.2016.08.015
35. 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(1), 12. https://doi.org/10.1186/s12870-014-0401-5
36. Jin-Hua, R. , Ting-Ting, S. , Ming-Ming, W. , & Xiao-Quan, W. . (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(1881), 20181012-. https://doi.org/10.1098/rspb.2018.1012
37. Rensing, S. A., Lang, D., Zimmer, A. D., Terry, A., Salamov, A., Shapiro, H., et al. (2008). The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science 319, 64-69. https://doi.org/ 10.1126/science.1150646
38. Sanderson M.J., Donoghue M.J. 1994. Shifts in diversification rate with the origin of angiosperms. Science 264: 1590-1593. https://doi.org/10.1126/science.264.5165.1590
39. Sauquet, Hervé, & Magallón, Susana. (2018). Key questions and challenges in angiosperm macroevolution. New Phytologist. https://doi.org/10.1111/nph.15104
40. Scharf, K. D., Berberich, T., Ebersberger, I., & Nover, L. (2012). The plant heat stress transcription factor (HSF) family: structure, function and evolution. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms, 1819(2), 104-119. https://doi.org/10.1016/j.bbagrm.2011.10.002
41. Schranz, M.E., Mohammadin, S., and Edger, P.P. (2012). Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. Current Opinion in Plant Biology. 15: 147-153. https://doi.org/10.1016/j.pbi.2012.03.011

42. Soltis, D. E., Visger, C. J., & Soltis ,P. S. (2014) The polyploidy revolution then…and now: Stebbins revisited. American Journal of Botany, 101(7), 1057-1078. https://doi.org/10.3732/ajb.1400178
43. Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312-1313. https://doi.org/10.1093/bioinformatics/btu033. pmid:24451623
44. Tang, H., Wang, X., Bowers, J.E., Ming, R., Alam, M., and Paterson, A.H. (2008b). Unraveling ancient hexaploidy through multiplyaligned angiosperm gene maps. Genome Res. 18: 1944-1954. https://doi.org/10.1101/gr.080978.108
45. Thalmann, M., Coiro, M., Meier, T., Wicker, T., Zeeman, S. C., & Santelia, D. (2019). The evolution of functional complexity within the β-amylase gene family in land plants. BMC evolutionary biology, 19(1), 66. https://doi.org/10.1186/s12862-019-1395-2
46. Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J.P., Wang, X.Y., et al. (2012) MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Research, 40, e49. https://doi.org/10.1093/nar/gkr1293
47. Xiaoming, W. , Xue, S. , Siyuan, C. , Chuang, M. , & Shengbao, X. . (2018). Evolutionary origin, gradual accumulation and functional divergence of heat shock factor gene family with plant evolution. Frontiers in Plant Science, 9, 71-. https://doi.org/10.3389/fpls.2018.00071
48. Wickett, N. J. , Mirarab, S. , Nguyen, N. , Warnow, T. , Carpenter, E. , & Matasci, N. , et al. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. Proceedings of the National Academy of Sciences, 111(45), 4859-68. https://doi.org/10.1073/pnas.1323926111
49. Yuzhou, Z. , Yue, J. , Hengwu, J. , Huabin, Z. , & Yu-Xian, Z. . (0). Two-step functional innovation of the stem-cell factors WUS/WOX5 during plant evolution. Molecular Biology & Evolution (3), 3. https://doi.org/ 10.1093/molbev/msw263
50. Ziheng Yang (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution, 1586-1591, 24(8), https://doi.org/10.1093/molbev/msm088
51. 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. https://doi.org/10.1038/ncomms5956

Tables

Due to technical limitations, table 1-3 is only available as a download in the Supplemental Files section.

Figures

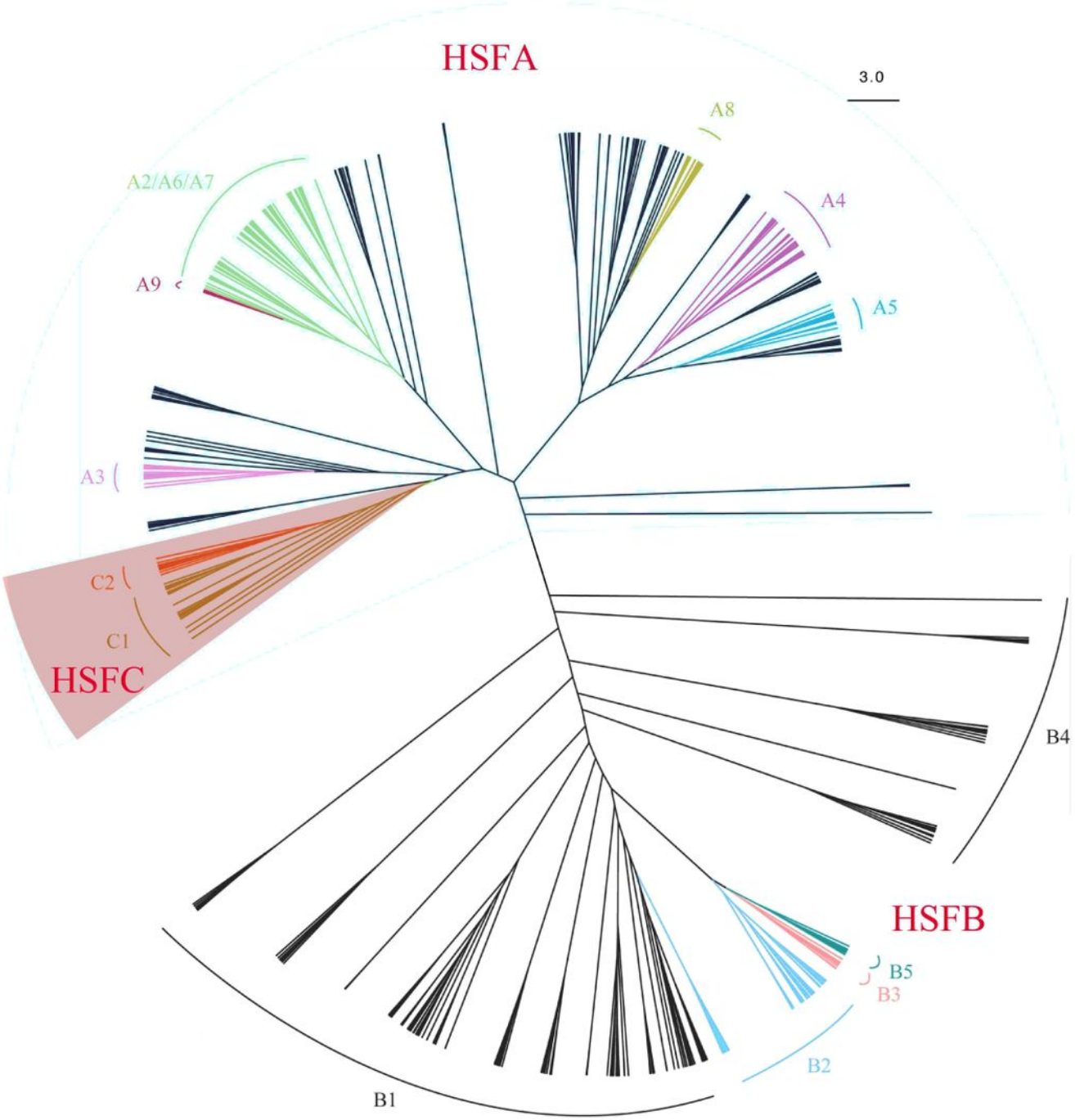

Figure 1

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 Additional files 1. 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.

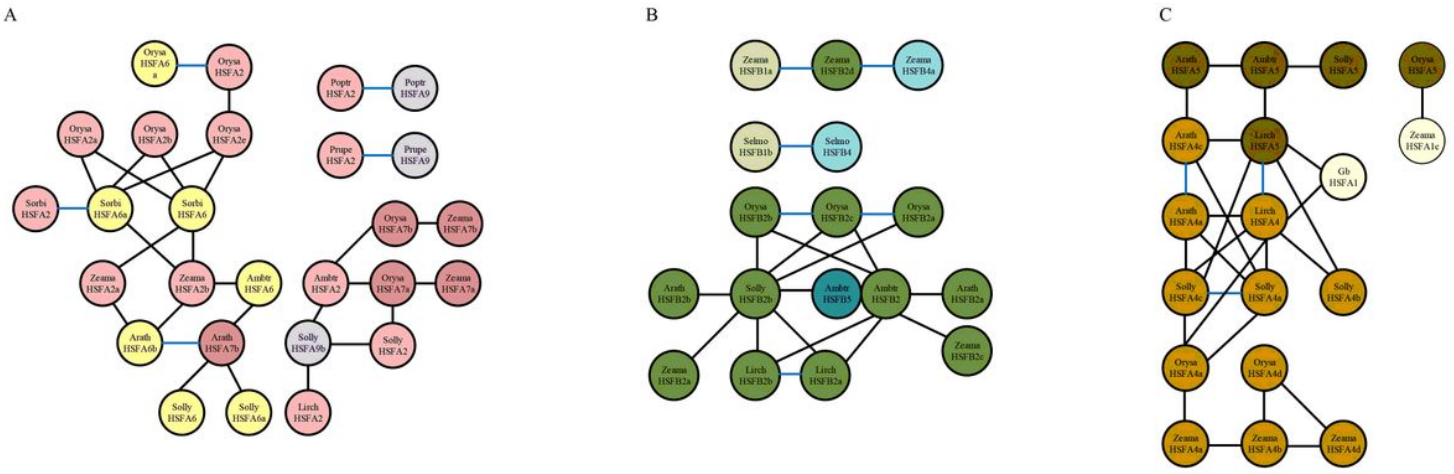

Figure 2

[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.

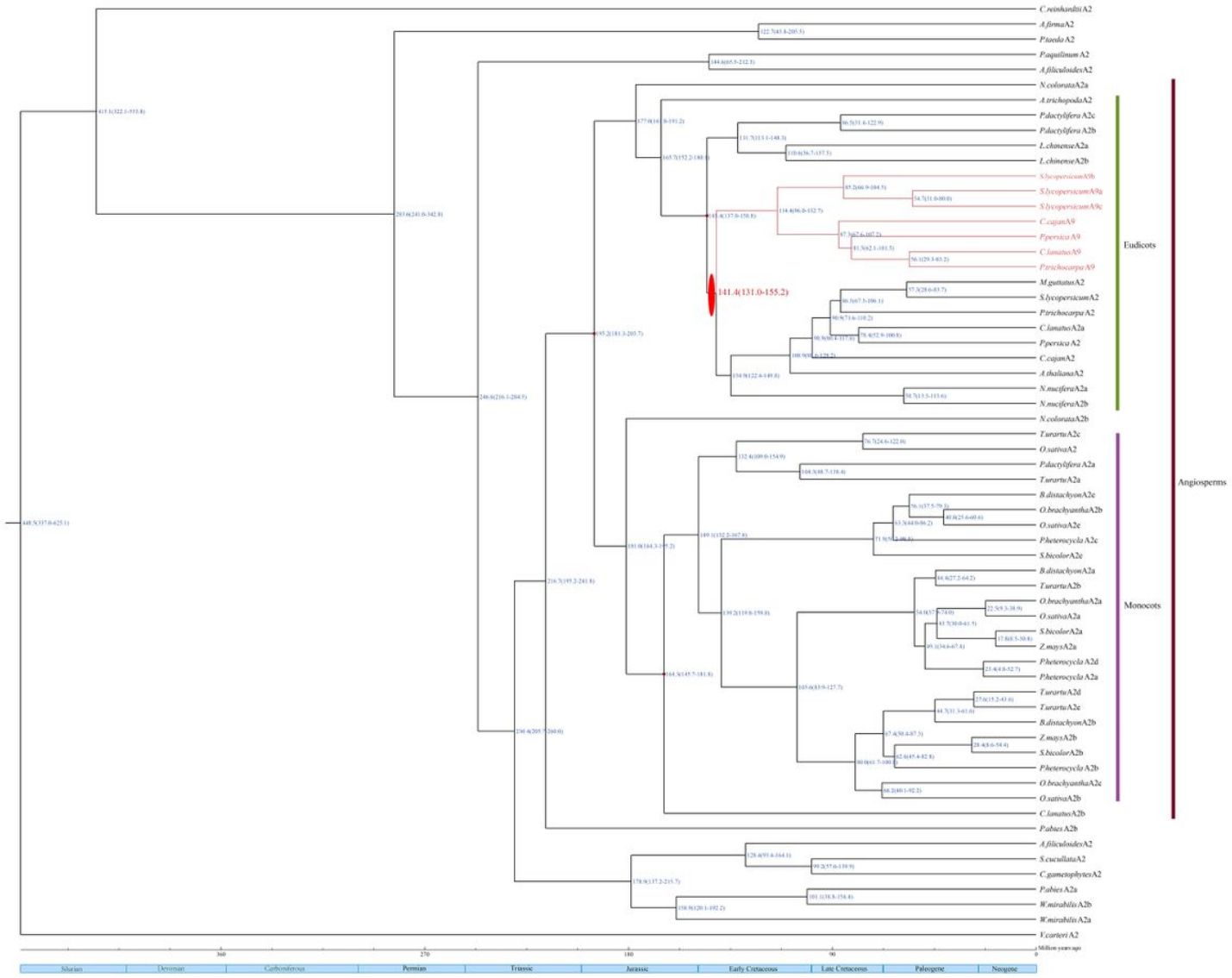

Figure 3

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.

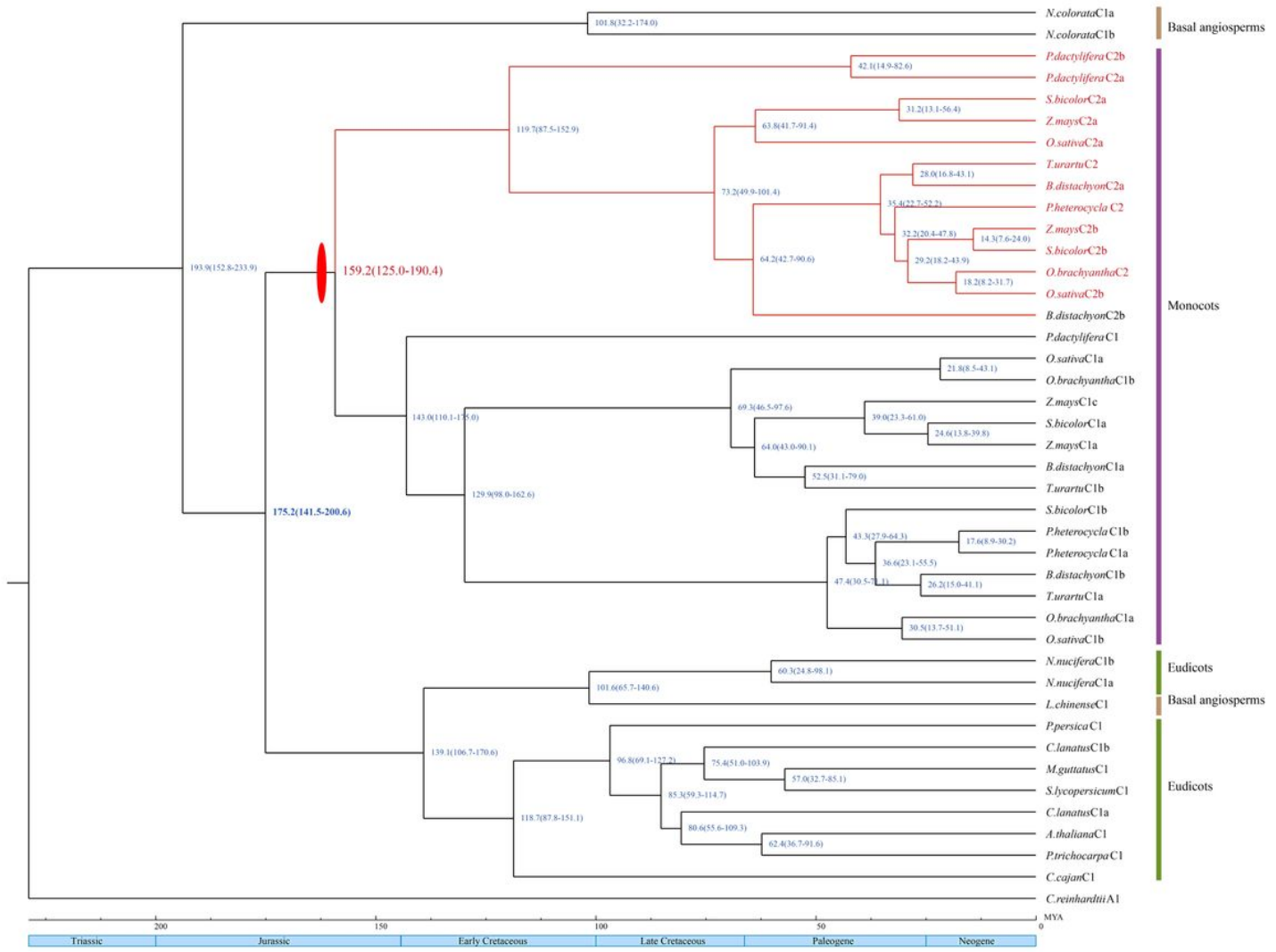

Figure 4

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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1S9.xlsx
- Additionalfile1.fasta
- FigS3.pdf
- FigS2.pdf
- FigS1.pdf
- Figurelegends.docx

- Table13.xlsx