Evolutional analysis of heat shock transcription factors in wild and cultivated rice (*Oryza sativa* L.)

Huan Tao

Fujian Agriculture and Forestry University <https://orcid.org/0000-0003-2440-1063>

Weifeng Wan

Fujian Agriculture and Forestry University

Jian Huang

Fujian Agriculture and Forestry University

Samuel Tareke Woldegiorgis

Fujian Agriculture and Forestry University

Yifan Xiong

Fujian Agriculture and Forestry University

Xiaoling Wei

Fujian Agriculture and Forestry University

Luoyao Wan

Fujian Agriculture and Forestry University

Zaiquan Xu

Fujian Agriculture and Forestry University

Lina Zhang

Fujian Agriculture and Forestry University

Wei Liu

Fujian Agriculture and Forestry University

Huaqin He (✉ hehq3@fafu.edu.cn)

Research article

Keywords: Hsfs, Wild rice, Cultivated rice, Evolution, Expression profile

Posted Date: January 28th, 2020

DOI: <https://doi.org/10.21203/rs.2.22058/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background Heat shock transcription factors (Hsfs) take part in many physiological and biochemical pathways in plants by regulating the expression of various stress-responsive genes, such as heat shock proteins (Hsps). With the development of rice genome re-sequencing projects, some researches had been carried out to identify Hsf gene family members in rice at the whole genomic scale. However, Hsfs in cultivated and wild rice genomes has not been fully studied and compared, although genetic diversity in cultivated rice is limited compared to wild rice.

Results In this research work, Hsfs genes were screened and evolutionally compared in the genomes of 6 wild rice and 1 cultivated rice varieties, including *O. barthii*, *O. glumaepatula*, *O. meridionalis*, *O. nivara*, *O. punctata*, *O. rufipogon* and *O. sativa* & Nipponbare. Total 22, 23, 24, 24, 25, 25 and 25 Hsf genes were identified in the tested 7 rice genomes, respectively. The different number of Hsf genes between wild and cultivated rice genotypes was due to dispersed duplication and whole genome duplication (WGD) events, reversely contributed to different stress-tolerant ability between wild and cultivated rice. The evolutionary analysis on the Hsf genes confirmed that *O. rufipogon* was the immediate ancestral progenitors of *O. sativa*. The expression profile of Hsf genes in Nipponbare and *O. rufipogon* under different stage of salinity stress showed that 4 root Hsf genes, including *HsfA3a*, *HsfA4d*, *HsfC2a* and *HsfC2b*, were simultaneously up-regulated by salinity stress in cultivated rice and its ancestral progenitor, implying that these 4 Hsf genes played conserved roles in rice in response to salinity stress. However, a substantial number of Hsf genes were exclusively regulated only in *Oryza rufipogon* rice seedling, suggesting that some of genuine salinity stress tolerance genes might be missing in cultivated rice.

Conclusion The results of this study would give insight into the evolution and function of Hsf gene members in rice, and hint to the use of wild relative genes to improve rice performance.

Background

Rice is one of the most important crops in the world, more than half of the population in the world use rice as the main source of food. However, the rice plants are growing under various biotic and abiotic stresses. For a case, abiotic stresses is serious threats to agricultural production and have become the primary cause of crop loss worldwide, reducing crop productivity by an estimated 50% annually[1]. Unlike animals, plants could not escape from the unfavorable environment because of immobility. During long-term evolution, plants have developed certain ability to survive under environmental stresses, such as the dominance of sporophyte that encloses the sensitive gametophyte, the presence of leaf epidermis with stomata for gas exchange, the formation of stress resistant dormant organs, and the presence of conducting tissues in long-lived and big plants for long-distance nutrient and water transport[2, 3]. A network of interconnected cellular stress response systems is a prerequisite for plant survival and productivity under environmental stresses[4]. Among them, the interaction between Heat shock proteins (Hsps) and heat shock factors (Hsfs) plays a critical role. Hsps possess molecular chaperone activities and contribute to cellular homeostasis in cells under both normal and adverse growth conditions[4, 5],

while Hsf proteins regulate the expression of various stress-responsive genes and play a key role in providing tolerance to multifarious abiotic stresses[6]. Hsf proteins contain a DNA-binding domain and bind to Heat shock elements (Hses) with a consensus sequence of GAAnnTTCnnGAA[7–9], which is presented at the upstream of the TATA box of HS-inducible genes[10], such as Hsp genes[7–9]. As transcriptional activators of Hsps, Hsfs cooperated with Hsps to form a network responding to multiple environmental stress treatments apart from heat stress[11, 12].

With the complement of genome DNA sequencing project, Hsf family members had been identified in more than 20 plant species at genome-wide scale[7], including *Oryza sativa* L.[8], *Pyrus bretschneideri*[13], *Sorghum bicolor*[14] and *Populus*[15]. The maximum of Hsf genes among monocots is 56 which was identified in wheat, while the maximum of Hsf genes among dicots were found in soybean[4]. Although plant Hsfs share highly conserved structure, their remarkable diversification across plants reflects their numerous functions as well as their integration into the complex stress signaling and response networks. In another word, there is functional divergency between Hsf orthologs among and within plant species. Therefore, it is necessary to adjust the research direction of Hsfs function from model plant to economic plant, including those with desired trait.

The genetic diversity in cultivated rice is lower than that of wild rice species, and the extent of DNA sequence diversity and phenotypic variation in wild *Oryza* is still being established[16]. Recent progress in sequencing the genomes of wild and cultivated rice genus is a good step in identifying their genetic differences associated with abiotic and biotic stresses-resistant ability. With increasing climate variability, selecting for traits allowing for the mitigation of abiotic stresses has now moved to the forefront of rice improvement. Wild rice species demonstrated abiotic stress tolerance, in which Hsfs played key role. However, to the best of our knowledge, Hsf evolution in cultivated and wild rice genomes had not been fully studied. In this paper, we would identify and compare the Hsfs in 6 wild rice accessions and 1 cultivated rice variety, and analyze Hsf evolution between wild and cultivated rice.

Results

Identification and classification of Hsf genes

Total 25 Hsfs genes were identified in Nipponbare genome, including 13 class A Hsfs, 8 class B Hsfs and 4 class C Hsfs. In *O. barthii*, *O. glumaepatula*, *O. meridionalis*, *O. nivara*, *O. punctata* and *O. rufipogon* genomes, 22, 23, 24, 24, 25 and 25 Hsfs genes were identified respectively. They all had 13 class A Hsfs. However, the different number of class B and C Hsfs were found in the 6 wild rice genomes. There were 6 class B Hsfs in *Oryza barthii*, 7 in *Oryza glumaepatula* and *Oryza meridionalis*, while 8 in *Oryza nivara*, *Oryza punctata* and *Oryza rufipogon*. *Oryza barthii*, *Oryza glumaepatula* and *Oryza nivara* genomes had 3 class C Hsfs, while *Oryza meridionalis*, *Oryza punctata* and *Oryza rufipogon* genomes had 4 class C Hsfs.

As shown in Supplementary Tables 1 and 2, HsfB2b, HsfB4d and HsfC2b genes were present in the Nipponbare genome but absent in *O. barthii* genome. HsfB2c and HsfC1b were not present in *O. glumaepatula* genome. *O. meridionalis* genome did not have HsfB4a gene, while *O. nivara* did not have

HsfC1b gene. And two Hsf genes were found to be translocated among the 6 tested rice varieties. The HsfA8a gene was located on chromosome 3 in Nipponbare genomes, but the gene was located on chromosome 6 in *O. nivara* genome. HsfB1a gene was located on chromosome 9 in Nipponbare genome, but it was located on chromosome 3 in *O. nivara* genome.

Evolutionary analysis of Hsf genes

The full-length amino acid sequences of Hsf genes in the 6 wild and 1 cultivated rice genomes were aligned to evaluate the evolutionary relationship of Hsf proteins, and the orthologous pair of Hsf genes among the test 7 rice varieties were also identified by BLASTp. ObHsfA7b, ObHsfB2c and ObHsfC2a were removed because their alignment length was less than 200 bp. The numbers of orthologous pairs between Nipponbare and the six wild rice, *O. barthii*, *O. glumaepatula*, *O. meridionalis*, *O. nivara*, *O. punctata* and *O. rufipogon*, were 19, 23, 24, 24, 25 and 25, respectively.

As shown in Fig. 1, all the Hsf proteins were divided into 3 different clades corresponding to the main Hsf classes A, B, and C. Within the Hsf gene A clade, eight distinct subclades (A1, A2, A3, A4, A5, A6, A7 and A8) were resolved and contained all of the Hsf A genes. The class A Hsf genes were divided into 13 orthologous gene clusters (OGCs). The C-type Hsf genes from the tested 7 rice varieties also constituted one distinct clade, had 4 OGCs and appeared to be closely related to the Hsf A group. Correspondingly, the B type Hsf genes were grouped into a separate clade subdivided into three groups (B1, B2 and B4). Notably, the B2 sub-clade was obviously distinct from the other two subclades and formed earlier than other subgroups, hinting that the B2 subgroup was the ancestor of other subgroups. Particularly, OrHsfB2b and ObHsfB2c genes were separated, suggesting that these two genes might be the ancestors of the rice Hsf genes, and other Hsf genes might have evolved through random replication and mutation of these two genes.

Selective constraints on OGCs of Hsf genes

To test the selective constraints on the evolution of each Hsf-family OGC, we detected the variations in dN/dS ratio among codon positions within each cluster, which was shown in Table 1. Overall, the average dN/dS ratio for the 25 clusters under model M0 was 0.217 (with a standard deviation of 0.112), which was statistically lower than 1 but greater than 0 (one-sample t-test, $p < 0.01$), illustrating that purifying selection was the predominant constraint on the evolution of Hsf gene family in rice. However, the log-likelihood differences between M3 and M0 were statistically significant for 15 OGCs, suggesting that the selective-constraint levels differed across these 15 Hsf OGCs. To evaluate whether positive selection facilitated the evolution of each Hsf OGC in rice, we compared M8 and M7 models to determine whether they had undergone positive selection. According to Yang et al. (2014), the following 3 criteria was used, including an estimate of $\omega > 1$ under M8, any sites found to be under positive selection and a statistically significant likelihood ratio test (LRT) [17]. Finally, we found 8 Hsf OGCs had undergone positive selection during the evolution of rice because they satisfied the above 3 criteria (Table 1).

Table 1

Parameter of positive selection under site-specific models for each Hsf OGC in rice.

Group	No.	Hsf	dN/dS (ω) under M0	2 Δ I, M3 vs. M0	2 Δ I, M8 vs. M7	Parameter estimates under M8	Positively selected sites
OGC1	6	HsfB2b	0.14049	14.604**	0.02	p1 = 0.13271, w = 1.00000 p = 3.87480, q = 98.99983	NAN
OGC2	5	HsfB2c	0.18177	144.38**	40.388**	p = 0.01317, q = 0.09377 p1 = 0.04901, w = 10.27485	11
OGC3	7	HsfB2a	0.038	0	0	p = 3.96425, q = 99.00000 p1 = 0.00001, w = 1.00000	NAN
OGC4	7	HsfB1a	0.10086	0.279	0	p = 0.89716, q = 7.69407 p1 = 0.00001, w = 1.00000	NAN
OGC5	7	HsfB4c	0.1591	32.029**	0.067	p = 1.37462, q = 99.00000 p1 = 0.16360, w = 1.11258	NAN
OGC6	6	HsfB4a	0.22777	2.057	0	p = 0.27556, q = 0.85859 p1 = 0.00001, w = 1.00000	NAN
OGC7	6	HsfB4d	0.30055	140.628**	22.899**	p = 0.01708, q = 0.05438 p1 = 0.05865, w = 6.91555	10
OGC8	7	HsfB4b	0.24141	48.148**	1.617	p = 5.97909, q = 99.00000 p1 = 0.17175, w = 1.70970	1
OGC9	7	HsfA4d	0.26651	3.821	0.375	p = 0.00500, q = 0.75866 p1 = 0.20236, w = 1.47764	NAN

* NAN means there is no positively selected site

Group	No.	Hsf	dN/dS (ω) under M0	2 Δ I, M3 vs. M0	2 Δ I, M8 vs. M7	Parameter estimates under M8	Positively selected sites
OGC10	7	HsfA4a	0.22383	0	0	p = 28.64176, q = 99.00000 p1 = 0.00001, w = 33.18236	NAN
OGC11	7	HsfA5a	0.17942	3.047	0	p = 0.05818, q = 0.24971 p1 = 0.00674, w = 1.00000	NAN
OGC12	7	HsfA7a	0.47716	53.843**	22.701**	p = 31.49855, q = 99.00000 p1 = 0.05797, w = 7.63926	5
OGC13	6	HsfC2a	0.24124	27.25**	0.376	p = 5.81163, q = 99.00000 p1 = 0.18562, w = 1.29356	NAN
OGC14	6	HsfC2b	0.17291	51.512**	5.467	p = 0.01413, q = 0.35141 p1 = 0.10764, w = 2.40297	1
OGC15	7	HsfC1a	0.08375	17.029**	0.794	p = 0.00500, q = 3.01839 p1 = 0.06962, w = 1.47290	NAN
OGC16	5	HsfC1b	0.12084	18.875**	2.87	p = 0.26773, q = 2.63611 p1 = 0.01325, w = 11.10354	1
OGC17	6	HsfA7b	0.17444	6.771	0.47	p = 13.23116, q = 99.00000 p1 = 0.04866, w = 1.79039	NAN
OGC18	7	HsfA2e	0.06398	0	0	p = 6.85736, q = 99.00000 p1 = 0.00001, w = 1.00000	NAN
OGC19	7	HsfA8a	0.44664	104.571**	54.387**	p = 14.66857, q = 99.00000 p1 = 0.06573, w = 16.87959	14

* NAN means there is no positively selected site

Group	No.	Hsf	dN/dS (ω) under M0	2 Δ I, M3 vs. M0	2 Δ I, M8 vs. M7	Parameter estimates under M8	Positively selected sites
OGC20	7	HsfA1a	0.32209	46.417**	18.311**	p = 7.36998 ,q = 99.00000 p1 = 0.05482 ,w = 7.16638	7
OGC21	7	HsfA3a	0.42448	40.071**	16.113**	p = 29.51001 ,q = 99.00000 p1 = 0.04751 ,w = 8.53702	4
OGC22	7	HsfA6b	0.26424	1.109	0	p = 0.29028 ,q = 0.77927 p1 = 0.00001 ,w = 1.00000	NAN
OGC23	7	HsfA6a	0.20438	20.264**	7.07*	p = 8.27014 ,q = 98.95912 p1 = 0.03015 ,w = 6.96406	3
OGC24	7	HsfA2b	0.18087	29.57**	6.052*	p = 1.78956 ,q = 99.00000 p1 = 0.08047 ,w = 2.96326	2
OGC25	7	HsfA2a	0.18992	8.449	2.631	p = 16.44481 ,q = 99.00000 p1 = 0.01004 ,w = 8.66996	NAN
* NAN means there is no positively selected site							

Synteny analysis of Hsf genes

The 3 main types of gene duplication result from chromosomal rearrangements, including retrotranspositions, tandem duplications or segmental duplications, and whole genome duplication (WGD), might drive the evolution of protein-coding gene families[18]. In this research work, MCScanX package was used to detect the origins of duplicate genes of Hsf gene family in the tested 7 rice variety genomes. As shown in Table 2, each member of Hsf gene family was grouped into one of five different categories: singleton, WGD, tandem, proximal or dispersed, while percentage of total Hsf family member in each category was calculated.

Table 2
Numbers of Hsf genes from different origins in the test 7 rice variety genomes*.

Species	No. of Hsf genes	No. and percentage of Hsf genes from different origins				
		Singleton	WGD	Tandem	Proximal	Dispersed
Nipponbare	25	9(36.0)	5(20.0)	0	0	11(44.0)
O.meridionalis	24	3(12.5)	5(20.8)	1(4.2)	1(4.2)	14(58.3)
O.glumaepatula	23	0	11(47.8)	0	0	12(52.2)
O.barthi	22	0	10(45.5)	0	0	12(54.5)
O.nivara	24	0	11(45.8)	0	0	13(54.2)
O.punctata	25	0	16(64.0)	0	0	9(36.0)
O.rufipogon	25	0	14(56.0)	0	0	11(44.0)
* Data in () indicated the percentage of Hsf genes.						

All the Hsf genes in *O.barthi*, *O.glumaepatula*, *O.nivara*, *O.punctata* and *O.rufipogon* genomes were duplicated and retained from dispersed duplication and WGD (Table 2). This result indicated that dispersed duplication and WGD played an important role in the expansion of Hsfs family genes for wild rice. Compared to the 6 wild rice varieties, the Hsf genes in Nipponbare had considerably the lowest proportions of WGD duplication (20.0%) and the highest proportion of singleton (36.0%), showing that Nipponbare experienced singleton since it divergenced from the wild rice. Only Hsf genes in *O. meridionalis* had tandem and proximal duplication event.

Meanwhile, we identified the conservation blocks both in intra- and inter-genomic pairwise to analyze the collinearity of Hsf genes as introduced by Qiao et al. [12]. According to the definition of fragment replication, if there are two or more common linear regions in one species, these regions are considered as fragment replication regions. Totally, 3, 4, 7, 0, 6, 12 and 10 collinear regions were found in Hsf genes in Nipponbare, *O. barthii*, *O. glumaepatula*, *O. meridionalis*, *O. nivara*, *O. punctate* and *O. rufipogon*, respectively (Fig. 2).

Expression profile of Hsf genes

Studying the gene-expression patterns of all members of a gene family would provide insight into their functional divergence[19]. To analyze the expression profile of Hsf genes in cultivated rice and their immediate ancestral progenitors in response to same stress condition, we searched and downloaded two transcription data from the previous studies, one was Nipponbare rice seedling under 1 h salinity stress while the other was *Oryza rufipogon* rice seedlings under 12 days salinity stress. In early response of Nipponbare rice seedling to salinity stress, three up-regulated shoot Hsf genes, 7 up-regulated and 7 down-regulated root Hsf genes were found. Whilst, in the late response of *Oryza rufipogon* rice seedling to

salinity stress, six up-regulated and 5 down-regulated shoot Hsf genes, 4 up-regulated and 1 down-regulated root Hsf genes were detected (Table 3). This result indicated that more Hsf genes were regulated in the shoot of rice seedling in response to long-term than short-term salinity stress, the reverse was true in the root. However, the same tissue of rice seedling under short and long-term salinity stress did not have the common regulated Hsf genes, except 4 root Hsf genes (Table 3). This result implied that different Hsf family members were function to response to different stage of salinity stress, or that cultivated rice and its immediate ancestral progenitor employed different Hsf genes to response to salinity stress. Four root Hsf genes, including HsfA3a, HsfA4d, HsfC2a and HsfC2b, were simultaneously up-regulated in Nipponbare under 1 h and *Oryza rufipogon* under 12d salinity stress, showing that they might function critically in the early and late response to salinity stress.

Table 3

Differentially expressing Hsf genes in the shoots and roots of *O. rufipogon* and Nipponbare rice seedling under salinity stress.

	Up-regulated		Down-regulated	
	Shoot	Root	Shoot	Root
<i>Oryza rufipogon</i> (12d salinity stress)	HsfA7a, HsfB2a, HsfB2b, HsfB2c, HsfC1b, HsfC2b	HsfA3a, HsfA4d, HsfC2a, HsfC2b	HsfA2b, HsfB4b, HsfB4c, HsfB4d, HsfC1a	HsfB4a
Nipponbare (1 h salinity stress)	HsfA3a, HsfC1a, HsfC2a	HsfA3a, HsfA4d, HsfC2a, HsfC2b, HsfA7a, HsfC1a, HsfC1b		HsfA6a, HsfA2b, HsfA6b, HsfA2e, HsfB2a, HsfB2b, HsfB2c

Discussion

Cultivated rice is considered to have been domesticated from wild rice thousands of years ago. However, in the long-term domestication, cultivated rice lost various valuable traits with regard to tolerance to cold, drought and salinity which derived from wild rice[19, 20]. Hsf proteins are regarded as key role in providing tolerance to multifarious abiotic stresses[6], the Hsf genes had therefore been identified and analyzed in various plants species. In this study, we achieved 25 Hsf genes in the genome of cultivated rice variety, Nipponbare. This was consistent with the research report of Guo et al. [8]. Meanwhile, we found 22 Hsf genes in *O. barthii* genome, the same number as the research of Guo et al. [6]. Duplication drives the evolution of gene in planta[21]. In this research work, different numbers of Hsf genes were found in 6 wild rice varieties. There were 22 Hsf genes in *O. barthii* genome, 23 in *O. glumaepatula* genome, 24 in *O. meridionalis* and *O. nivara* genomes, while 25 in *O. punctate* and *O. rufipogon* genomes.

This could be inferred from the results that dispersed duplication and WGD played an important role in the expansion of Hsf family genes in rice taxa, which lead to rice polyploid. Song et al. observed that Hsf genes were amplified by gene replication in Chinese cabbage and that the replication of Hsf genes promoted the polyploidy of Brassica[22]. Therefore, WGDs and additional recent lineage-specific WGDs had presumably resulted in varying numbers of Hsf genes within flowering plants^[18]. Meanwhile, in this study, *O. rufipogon* genomes, as well as Nipponbare, had 25 Hsf genes, implying that *O. rufipogon* was the immediate ancestral progenitors of *O. sativa*[23].

The evolutionary tree of Hsfs in the test 7 rice varieties was all divided into three categories, A, B and C. The class A was the largest one. Qiao et al.[13] constructed the Hsfs evolution tree of Chinese white pear, poplar and Arabidopsis, and found that Hsfs of three species were clustered to the corresponding A, B and C[13]. But the number of China pear Hsf A branch was 19, while B and C branch Hsfs were 8 and 2. Nagaraju et al. used the same strategy to build the evolutionary tree of rice, Arabidopsis and sorghum Hsfs[14]. They also found that the most Hsfs was clustered into A branches followed by B branches, and the least was C branches. Meanwhile, they concluded that the difference between the three species of Hsfs in the A and B branches lead to the different responses of different species to stress[14]. Therefore, the results in this research work suggested that the different stress-tolerant ability between wild and cultivated rice might be induced by the difference of Hsfs between 6 wild rice and cultivated rice on the evolutionary branches.

The determination of orthology is an essential part of comparative genomics, which had been carried out by using synteny analysis in many previous studies[24, 25]. In this research work, collinear genes and collinear regions were not found in Hsfs of *Oryza meridionalis*, which probably due to the serious loss of homologous genes in the Hsf regions or its imperfect genome assembly and annotation. The number of collinear regions of Hsf genes detected in the wild and cultivated rice genomes was lower, because some of collinear region was associated with the loss of Hsf genes. The collinear region plays an important role in understanding the replication, deletion and expansion of Hsf genes. In this research work, we found that different patterns of gene duplication contributed differentially to the expansion of the Hsf gene family in the investigated rice taxa (Table 1). This result further confirmed that divergence in Hsf gene number in these species was the result of Hsf gene duplication and/or loss in rice taxa.

As Nipponbare rice root was only exposed to 1 h salinity stress, less Hsf genes were regulated in the shoot than in the root. When only the root was exposed to environmental stresses, it might take time to transduce the signal and induce the expression of more genes in rice shoot [26]. OsTPP1 was found to be up-regulated in rice shoot after exposure to abiotic stress for 10h[27]. The reverse results were saw in *Oryza rufipogon* rice seedling under 12d salinity stress, the immediate ancestral progenitor of cultivated rice. After the root exposed to long-term stress, the rice shoot could sense the stress and induce more Hsf genes to response to the stress. Meanwhile, more regulated Hsf genes might hint the critical stress-response time stage for rice tissue. Ma et al.[28]found more genes in blast-resistant rice variety were regulated at 24 h than 48 h after blast fungus infection, and they revealed 24 h after infection was the critical response time for rice plants to resist blast fungus. Therefore, we could make a conclusion here

that the root and shoot Hsf genes in rice seedling were regulated to response to short-term and long-term salinity stress, respectively.

Some common root Hsf genes, including HsfA3a, HsfA4d, HsfC2a and HsfC2 were up-regulated by salinity stress in Nipponbare and its ancestral progenitor, indicating that these four Hsf genes played important roles in rice during salinity stress. However, a substantial number of Hsf genes were exclusively regulated only in *Oryza rufipogon* rice seedling, suggesting that some of genuine salinity stress tolerance genes might be missing in Nipponbare. Many of the genes in cultivated rice have been selected by humans under field conditions, not by environmental stress[26]. During rice domestication, some of stress-responsive genes might be omitted in cultivated rice in order to achieve high rice yield. However, the use of crop wild relative genes to improve crop performance was well established with important examples dating back more than 60 years, such as wheat and tomato. Therefore, these essentially missing genes in ancestral progenitor of the cultivated rice could serve as potential genetic resources for the improvement of cultivated crops.

Methods

Identification of Hsf genes in wild and cultivated rice genotype

The genome, gene and protein sequences of 6 wild rice varieties, including *Oryza barthii*, *Oryza glumaepatula*, *Oryza meridionalis*, *Oryza nivara*, *Oryza punctata* and *Oryza rufipogon*, were downloaded from FTP database (<ftp://ftp.ensemblgenomes.org/pub/release-23/plants/fasta/>). The protein sequences of the cultivated rice variety, Nipponbare, was downloaded from MSU database (<ftp://ftp.plantbiology.msu.edu/>), its genome and gene sequences was downloaded from RAPDB database (<http://rapdb.dna.affrc.go.jp/>). The Hsf domain model, PF00447[29] was used to search against the protein sequence dataset of the tested genome using HMMER software with a strict cut-off E-value of 10⁻⁵ and removing the repetitive sequence. All the candidate Hsf protein sequences were analyzed in the Pfam database to verify the presence of DBD domains with a strict cut-off E-value of 10⁻³. Furthermore, online tools, MARCOIL (<http://toolkit.tuebingen.mpg.de/marcoil/>), NLStradamus[30] (<http://www.moseslab.csb.utoronto.ca/NLStradamus/>) and NetNES[31] (<http://www.cbs.dtu.dk/services/NetNES/>) were applied to confirm the HR-A/B, NLS and NES domains, respectively. MEME (<http://meme-suite.org/tools/meme>) was finally employed to verify the exactness of these domains.

Phylogenetic analyses

ClustalW software was used to multiple align HSF domains from the investigated rice varieties with default settings. After that, MEGA6.06[32] was used to build up the maximum-parsimony phylogenetic trees with Neighbor-Joining method and 1000 bootstrap replicates, and displayed trees using TreeView (<http://www.taxonomy.zoology.gla.ac.uk/rod/treeview.html>).

Hsf duplication patterns and collinearity analysis

Local BLASTp was performed to search potential homologous gene pairs with a strict cut-off E-value of $1e-5$, Identity > 80%, length > 200 bp, and output the top five matches in each alignment results. Then the output gff file was used as input file for MCScanX to identify syntenic chains, as well as the duplication patterns in the Hsf gene family. At last, Hsf genes and their collinearity were labeled on the rice chromosomes by Circos[33] software.

Detection of positive selection

The program MEGA6 was used to obtain aligned sequence files and tree files, and DAMBE was used to obtain PML format which was input into the codeml program in PAML[34]. Using the program codeml, we detected variation in the ω parameter among sites by employing likelihood-ratio tests (LRTs) for M0 vs. M3 and M7 vs. M8. The LRT for the M0 vs. M3 comparison was used to test the heterogeneity in ω between codon positions, where M0 model assumes one ω among all sites and M8 model uses an unconstrained discrete distribution of ω with a set number of classes. The M7 vs. M8 comparison was used to detect the role of positive selection. M7 is the null model assumes a beta distribution of ω values between $\omega = 0$ and $\omega = 1$ among the sites, while the alternative model M8 adds a free parameter to the null model and allows positive selection to occur. In each LRT, twice the difference of the log likelihood of the two models was compared to the chi-squared (χ^2) statistic, with the degrees of freedom (DFs) being equal to the difference in the number of parameters. In our analyses, the DFs were 4 for the M0 vs. M3 test and 2 for the M7 vs. M8 test[35].

Transcript profiling of Hsf genes in wild and cultivated rice under high salt stress

The RNA-seq data of *Oryza rufipogon* and *Nipponbare* under high salt stress was derived from previous studies[26, 36]. Briefly, The Cuffdiff software was employed to evaluate the differential gene expression between any two samples. The gene abundance differences between sample pairs were calculated based on the ratio of the FPKM values. The significance of gene expression differences was calculated using the FDR (False Discovery Rate) control method to justify the p-value. Here, only genes with an absolute value of $\log_2$ ratio ≥ 0.76 (fold-change ≥ 1.7) and a FDR significance score ≤ 0.001 were used for subsequent analysis.

Abbreviations

OGCs:orthologous gene clusters; WGD:whole genome duplication; Hsps:Heat shock proteins; Hsfs:heat shock factors; Hses:Heat shock elements; LRTs:likelihood-ratio tests; DFs:degrees of freedom; FDR:False Discovery Rate

Declarations

Acknowledgments: The authors thank the anonymous referees whose constructive comments were helpful in improving the quality of this work.

Authors' Contributions

HH conceived the study and revised the manuscript. HT and WW carried out the experiment and wrote the manuscript. JH, SW and YX analyzed the evolution of Hsf genes. XW, LW and ZX analyzed the RNA-seq data. LZ and WL helped to revise the manuscript. The authors have read and approved of the final version.

Funding

This work was supported by Natural Science Foundation of China and Fujian (No. 31270454; 2019J01381 and JB14016), Innovative Foundation of FAFU (No. CXZX2017132 and CXZX2017303), Program for Innovative Research Team in College of Life Sciences, FAFU (No. 712018009). The funding body did not participate in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript

Availability of data and materials

The data sets supporting the results of this article are included within the article and its additional files.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing Interests

The authors declare that they have no competing interests

Author details

1 College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, China

References

1. Wang W, Vinocur B, Shoseyov O, Altman A: Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. *Trends in plant science* 2004, 9(5):244-252.
2. Baniwal SK, Bharti K, Chan KY, Fauth M, Ganguli A, Kotak S, Mishra SK, Nover L, Port M, Scharf K-D: Heat stress response in plants: a complex game with chaperones and more than twenty heat stress transcription factors. *Journal of biosciences* 2004, 29(4):471-487.
3. Al-Whaibi MH: Plant heat-shock proteins: a mini review. *Journal of King Saud University-Science* 2011, 23(2):139-150.

4. Scharf K-D, Berberich T, Ebersberger I, Nover L: The plant heat stress transcription factor (Hsf) family: structure, function and evolution. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 2012, 1819(2):104-119.
5. Kotak S, Vierling E, Bäumlein H, von Koskull-Döring P: A novel transcriptional cascade regulating expression of heat stress proteins during seed development of Arabidopsis. *The Plant Cell* 2007, 19(1):182-195.
6. Guo M, Liu J-H, Ma X, Luo D-X, Gong Z-H, Lu M-H: The plant heat stress transcription factors (HSFs): structure, regulation, and function in response to abiotic stresses. *Frontiers in plant science* 2016, 7:114.
7. Nover L, Bharti K, Döring P, Mishra SK, Ganguli A, Scharf K-D: Arabidopsis and the heat stress transcription factor world: how many heat stress transcription factors do we need? *Cell stress & chaperones* 2001, 6(3):177.
8. Guo J, Wu J, Ji Q, Wang C, Luo L, Yuan Y, Wang Y, Wang J: Genome-wide analysis of heat shock transcription factor families in rice and Arabidopsis. *Journal of Genetics and Genomics* 2008, 35(2):105-118.
9. Xue G-P, Sadat S, Drenth J, McIntyre CL: The heat shock factor family from Triticum aestivum in response to heat and other major abiotic stresses and their role in regulation of heat shock protein genes. *Journal of experimental botany* 2013, 65(2):539-557.
10. Scharf K-D, Siddique M, Vierling E: The expanding family of Arabidopsis thaliana small heat stress proteins and a new family of proteins containing α -crystallin domains (Acd proteins). *Cell stress & chaperones* 2001, 6(3):225.
11. Shao H-B, Guo Q-J, Chu L-Y, Zhao X-N, Su Z-L, Hu Y-C, Cheng J-F: Understanding molecular mechanism of higher plant plasticity under abiotic stress. *Colloids and Surfaces B: Biointerfaces* 2007, 54(1):37-45.
12. Hahn A, Bublak D, Schleiff E, Scharf K-D: Crosstalk between Hsp90 and Hsp70 chaperones and heat stress transcription factors in tomato. *The Plant Cell* 2011, 23(2):741-755.
13. Qiao X, Li M, Li L, Yin H, Wu J, Zhang S: 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* 2015, 15(1):12.
14. Nagaraju M, Sudhakar Reddy P, Anil Kumar S, K Srivastava R, Kavi Kishor P, Rao DM: Genome-wide scanning and characterization of Sorghum bicolor L. heat shock transcription factors. *Current genomics* 2015, 16(4):279-291.
15. Zhang J, Liu B, Li J, Zhang L, Wang Y, Zheng H, Lu M, Chen J: Hsf and Hsp gene families in Populus: genome-wide identification, organization and correlated expression during development and in stress responses. *BMC genomics* 2015, 16(1):181.
16. Atwell BJ, Wang H, Scafaro AP: Could abiotic stress tolerance in wild relatives of rice be used to improve Oryza sativa? *Plant Science* 2014, 215:48-58.

17. Yang Z, Wang Y, Gao Y, Zhou Y, Zhang E, Hu Y, Yuan Y, Liang G, Xu C: Adaptive evolution and divergent expression of heat stress transcription factors in grasses. *BMC evolutionary biology* 2014, 14(1):147.
18. Maher C, Stein L, Ware D: Evolution of Arabidopsis microRNA families through duplication events. *Genome research* 2006, 16(4):510-519.
19. Whittle CA, Krochko JE: Transcript profiling provides evidence of functional divergence and expression networks among ribosomal protein gene paralogs in *Brassica napus*. *The Plant cell* 2009, 21(8):2203-2219.
20. Zhang X, Zhou S, Fu Y, Su Z, Wang X, Sun C: Identification of a drought tolerant introgression line derived from Dongxiang common wild rice (*O. rufipogon* Griff.). *Plant molecular biology* 2006, 62(1-2):247-259.
21. Wendel JF: Genome evolution in polyploids. In: *Plant molecular evolution*. Springer; 2000: 225-249.
22. Song X, Liu G, Duan W, Liu T, Huang Z, Ren J, Li Y, Hou X: Genome-wide identification, classification and expression analysis of the heat shock transcription factor family in Chinese cabbage. *Molecular genetics and genomics* 2014, 289(4):541-551.
23. Zhu Y, Ellstrand NC, Lu BR: Sequence polymorphisms in wild, weedy, and cultivated rice suggest seed-shattering locus *sh4* played a minor role in Asian rice domestication. *Ecology and evolution* 2012, 2(9):2106-2113.
24. Jun J, Mandoiu II, Nelson CE: Identification of mammalian orthologs using local synteny. *BMC genomics* 2009, 10(1):630.
25. Du D, Cheng T, Pan H, Yang W, Wang J, Zhang Q: Genome-wide identification, molecular evolution and expression analyses of the phospholipase D gene family in three Rosaceae species. *Scientia horticulturae* 2013, 153:13-21.
26. Mizuno H, Kawahara Y, Sakai H, Kanamori H, Wakimoto H, Yamagata H, Oono Y, Wu J, Ikawa H, Itoh T: Massive parallel sequencing of mRNA in identification of unannotated salinity stress-inducible transcripts in rice (*Oryza sativa* L.). *BMC genomics* 2010, 11(1):683.
27. Pramanik MHR, Imai R: Functional identification of a trehalose 6-phosphate phosphatase gene that is involved in transient induction of trehalose biosynthesis during chilling stress in rice. *Plant molecular biology* 2005, 58(6):751-762.
28. Ma S, Lin S, Wang M, Zou Y, Tao H, Liu W, Zhang L, Liang K, Ai Y, He H: Differential Expression Proteins Contribute to Race-Specific Resistant Ability in Rice (*Oryza sativa* L.). *Plants* 2019, 8(2):29.
29. Finn RD, Mistry J, Tate J, Coghill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K: The Pfam protein families database. *Nucleic acids research* 2009, 38(suppl_1):D211-D222.
30. Ba ANN, Pogoutse A, Provart N, Moses AM: NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. *BMC bioinformatics* 2009, 10(1):202.
31. La Cour T, Kierner L, Mølgaard A, Gupta R, Skriver K, Brunak S: Analysis and prediction of leucine-rich nuclear export signals. *Protein Engineering Design and Selection* 2004, 17(6):527-536.

32. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular biology and evolution* 2013, 30(12):2725-2729.
33. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: Circos: an information aesthetic for comparative genomics. *Genome research* 2009, 19(9):1639-1645.
34. Yang Z: PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* 2007, 24(8):1586-1591.
35. Yang Z, Bielawski JP: Statistical methods for detecting molecular adaptation. *Trends in ecology & evolution* 2000, 15(12):496-503.
36. Zhou Y, Yang P, Cui F, Zhang F, Luo X, Xie J: Transcriptome analysis of salt stress responsiveness in the seedlings of Dongxiang wild rice (*Oryza rufipogon* Griff.). *PloS one* 2016, 11(1):e0146242.

Figures

Figure 1

Phylogenetic tree of Hsf genes in 6 wild and 1 cultivated rice varieties. (The phylogenetic tree was obtained using the MEGA 6.0 software on the basis of amino-acid sequences. Bootstrap analysis was conducted with 1000 replicates. The abbreviations of species names were showed in the following. Os, Nipponbare; Ob, *Oryza barthii*; Og, *Oryza glumaepatula*; Om, *Oryza meridionalis*; On, *Oryza nivara*; Op, *Oryza punctata*; Or, *Oryza rufipogon*. The brown-labeled genes do not belong to the OGC)

Figure 2

Localization of Hsf gene synteny in the 7 test rice genomes. (A to G show the localization of Hsf gene synteny in *Oryza barthii*, *Oryza glumaepatula*, *Oryza meridionalis*, *Oryza nivara*, *Oryza punctata*, *Oryza rufipogon* and *Nipponbare*, respectively).

Supplementary Files

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

- [SupplementTables.docx](#)