Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Control of gametophytic self-incompatibility in the African wild rice

Yanan Wang
Institute of Genetics and Developmental Biology
Hong Zhao
Institute of Genetics and Developmental Biology, Chinese Academy of Sciences
Sihui Zhu
Beijing Institute of Genomics https://orcid.org/0000-0002-1610-2234
Hongkui Zhang
Beijing Institute of Genomics, Chinese Academy of Sciences, and National Centre for Bioinformation, Beijing 100101 China
Yingying Chen
College of Agriculture, Yangzhou University
Xiang Tao
College of Agriculture, Yangzhou University
Wenzhi Tong
College of Agriculture, Yangzhou University
Huayang Tian
Institute of Genetics and Developmental Biology
Yu Guan
Institute of Genetics and Developmental Biology
Huaqiu Huang
Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China
Qianqian Han
Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China
Zhukuan Cheng
Institute of Genetics and Developmental Biology https://orcid.org/0000-0001-8428-8010
Yijing Zhang
Fudan University https://orcid.org/0000-0001-9568-9389
Chuandeng Yi
College of Agriculture, Yangzhou University
Yue Zhang
State Key Laboratory of Molecular Developmental Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences and National Center for Plant Gene Research, Beijing 100101, Chi
https://orcid.org/0000-0003-3618-8706

Yongbiao Xue (ybxue@genetics.ac.cn)
Institute of Genetics and Developmental Biology https://orcid.org/0000-0002-6895-8472

Article

Keywords:

Posted Date: December 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2121145/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Control of gametophytic self-incompatibility in the African wild rice

Authors: Yanan Wang[1,4†], Hong Zhao[1†*], Sihui Zhu[2,4†], Hongkui Zhang[2,4†], Yingying Chen[3], Xiang Tao[3], Wenzhi Tong[3], Huayang Tian[1,4], Yu Guan[1,4], Huaqiu Huang[1], Qianqian Han[1,4], Zhukuan Cheng[3], Yijing Zhang[6], Chuandeng Yi[3,5*], Yu'e Zhang[1,4*], Yongbiao Xue[1,2,4*]

[1]The State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Beijing 100101, China

[2]Beijing Institute of Genomics, Chinese Academy of Sciences, and National Centre for Bioinformation, Beijing 100101 China

[3]Jiangsu Key Laboratory of Crop Genomics and Molecular Breeding/Key Laboratory of Plant Functional Genomics of the Ministry of Education, College of Agriculture, Yangzhou University, Yangzhou 225009, China

[4]University of the Chinese Academy of Sciences, Beijing, 100049, China

[5]Jiangsu Key Laboratory of Crop Genetics and Physiology/Jiangsu Co-Innovation Center for Modern Production Technology of Grain Crops, College of Agriculture, Yangzhou University, Yangzhou 225009, China.

[6]State Key Laboratory of Genetic Engineering, Collaborative Innovation Center of Genetics and Development, Department of Biochemistry, Institute of Plant Biology, School of Life Sciences, Fudan University, Shanghai 200438, China

[†]**Authors contributed equally to this work.**

[*]**Authors for correspondence**

E-mail: zhhong@genetics.ac.cn (Hong Zhao)
cdyi@yzu.edu.cn (Chuandeng Yi)
zhangyue@genetics.ac.cn (Yu'e Zhang)
ybxue@genetics.ac.cn (Yongbiao Xue)

Phone: +86-01-6480-1181 (Yongbiao Xue)

Self-incompatibility (SI) in the grass family (Poaceae) is controlled by a gametophytic two-locus system (*S* and *Z*)[1-6]. Recent studies have revealed several genes linked to the *S* and *Z* loci[7-13], but their roles remain largely unclear. Here, we discover that *OlS$_1$-HPS10* and *OlS$_1$-DUF247-1* control self-incompatibility of African wild rice (*Oryza longistaminata*), a Poaceae species. *OlS$_1$-HPS10* and *OlS$_1$-DUF247-1* encode a small extracellular pistil peptide and a pollen-specific transmembrane protein, respectively, which can physically interact with each other to form a potential membrane-based signaling module. Multiple independent knockouts of *OlS$_1$-HPS10* or *OlS$_1$-DUF247-1* by CRISPR-Cas9 result in the respective breakdowns of the pistil and the pollen function of the *S$_1$* haplotype in *O. longistaminata*, suggesting that they control SI in a self-recognition manner. Furthermore, the orthologs of both *OlS$_1$-HPS10* and *OlS$_1$-DUF247-1* are the Oryzoideae- and Pooideae-specific and their linkage occurred around 50 million years ago, showing that the *S*-locus originated in the early Oryzoideae lineage of the grass family. Taken together, our results demonstrate that the pistil *OlS$_1$-HPS10* and the pollen *OlS$_1$-DUF247-1* together form a self-recognition system to control the gametophytic SI in *O. longistaminata,* uncovering a Poaceae-lineage-specific innovation of SI mechanism in angiosperms.

Introduction

Self-incompatibility (SI) is a widely occurring outcrossing mechanism adopted by approximately 40% of all angiosperms to avoid inbreeding[4,6]. In many species, SI is controlled by a single polymorphic *S* locus comprising at least two tightly-linked genes separately determining the pollen and pistil *S* specificity[6, 14-17]. Recent studies reveal that several molecular types of SI have evolved in eudicots, including Type 1 defined by linked *S-RNase* and *SLF* (*S-locus F-box*) genes[18-22] found in Plantaginaceae, Solanaceae, Rosaceae. and Rutaceae, Type 2 *SRK* and *SCR* in Brassicaceae[23,24], Type 3 *PrsS* and *PrpS* in Papaveraceae[25,26], heterostyly Type 4

CYP-GLO2-KFB-CCM-PUB in Primulaceae[27] and another heterostyly type of SI *BAHD-SPH1-YUC6* in Turneraceae[28], showing a highly dynamic evolutionary process of eudicot SI systems[29]. Consistently, SI is widespread in the grass family but is controlled by two independent *S* and *Z* loci, and self pollen rejection only occurs when their alleles are both matched between the pollen and pistil[1-3]. Recently, a pistil-specific *HPS10* was shown to be linked to the *S* locus and likely encodes a pistil *S* candidate in *Hordeum bulbosum*[9]. In perennial ryegrass (*Lolium perenne*) and rye (*Secale cereale*), the *S* and *Z* loci were mapped on chromosomes 1 and 2, respectively[7,8,30,31], and two polymorphic genes both encoding a polypeptide containing a domain of unknown function 247 (DUF247) appear to be the candidate pollen *S* and *Z* genes (*LpS*- and *Z-DUF247s*)[10,11]. In *Oryza longistaminata*, a self-incompatible African wild rice from the Poaceae[32,33], two *DUF247* genes (*OlSS1* and *OlSS2*) similar to *LpSDUF247* and a predicted pistil *S* named *OlSP* in a region syntenic to chromosome 5 of *O. sativa*[12] have also been characterized and the latter seems to be orthologous to *HPS10* (refs 9,13). However, it remains largely mysterious whether these candidate *S* and *Z* genes control SI in the grass family.

Haplotype-resolved genomes of *O. longistaminata*

To identify the candidate *S* genes from *O. longistaminata*, we first assembled its high-quality haplotype-resolved genomes by taking an integrated sequencing and assembling approach by generating 28.8 Gb Hi-Fi reads, 49.8 Gb Illumina short reads, and 42.4 Gb high-throughput chromatin conformation capture (Hi-C) data, and achieved two haplotype-resolved assemblies comprising 12 pseudo-chromosomes with the estimated genome sizes of 330.2 Mb and 373.1 Mb, respectively (Fig. 1, Extended Data Table 1). To annotate protein-coding genes, several lines of evidence from homologous proteins, RNA-seq data from *O. longistaminata* and *ab initio* prediction were combined and a total of 84,510 protein-coding genes predicted, of which 90.1% can be annotated in at least one functional database (Extended Data Table 1). The BUSCO scores of two haplotypes based on the Liliopsida_odb10

datasets were 97.7% and 94.6%, respectively, which are better than the Ensembl version (91.5%) (Extended Data Fig. 1a). Moreover, we detected the synteny of the candidate *S*-locus and its flanking regions between two assembled haplotypes (Extended Data Fig. 1b). Taken together, these results demonstrate the high accuracy and completeness of our genome assemblies of *O.longistaminata* significantly improved over three previous versions[34-36].

Identification of *OlS$_1$-HPS10* and *OlS$_1$-DUF247-1* as the candidate *S* genes

To search for the candidate pistil *S* genes in *O. longistaminata*, we first identified homologs to a candidate pistil *S* gene *HPS10* from *H. bulbosum*[9] and found two allelic copies on its chromosome 5, showing 53.3% amino acid identity to each other and named *OlS$_1$*- and *S$_2$-HPS10* (Fig. 2a, Extended Data Fig. 2). To examine their expression pattern, we performed RNA-seq analysis and found that they are highly expressed in style (Fig. 2a, Extended Data Fig. 3), similar to that in *H. bulbosum*[9]. To further examine if they are polymorphic, we aligned amino acid sequences of HPS10s from *H. bulbosum* (Hb), two *S* haplotypes of *O. longistaminata* (OlS$_1$ and S$_2$) and two previously reported OlSP-1 and -2 (ref. 12) and found that they are highly polymorphic with amino acid identities ranging from 35.6-64.8% (Extended Data Fig. 2a, b). In addition, HPS10s were predicted to contain a signal peptide (SP) at N termini and a hypervariable region (HV) in the center regions (Extended Data Fig. 2a, c, d), indicating that they are likely secreted extracellular polypeptides. To examine their subcellular locations, we expressed the OlS$_1$- and S$_2$-HPS10-GFP fusion proteins and found that they are localized in the intercellular space (Fig. 2b), confirming that they are extracellular proteins.

To look for the candidate pollen *S* gene, we searched homologs to *LpSDUF247* from *L. perenne*[11] and found two homologs *OlS$_1$-DUF247-1* and *OlS$_1$-DUF247-2* located 802 bp upstream and 4391 bp downstream of *OlS$_1$-HPS10*, respectively (Extended Data Fig. 4), indicating that they likely form a single locus. Further analysis showed their synteny among five Poaceae species examined (Extended Data

Fig. 4), suggesting that this locus is conserved in Poaceae. To examine the expression patterns of *OlS_1-* and *S_2-DUF247-1* and *-2*, we performed RNA-seq analysis and found that *OlS_1-* and *S_2-DUF247-1* are highly expressed in anther, consistent with their proposed roles in the pollen *S* (refs 10,11), and *OlS_1*-and *S_2-DUF247-2* are expressed in both anther and style (Fig. 2a, Extended Data Fig. 3), indicating that they likely play different roles not related to SI. To further explore their gene polymorphism, we aligned amino acid sequences of DUF247-1s from *L. perenne* (Lp), those from two *S* haplotypes of *O. longistaminata* (OlS_1 and S_2) as well as OlSS2-1 and OlSS2-2 published previously[11] and found that they show amino acid identities ranging from 69.6-88.2% (Extended Data Fig. 2a, b). In addition, OlS_1- and S_2-DUF247-1 and LpSDUF247 were predicted to contain a predicted transmembrane domain (TMD) and three hypervariable regions (HV1-HV3), and all three hypervariable regions reside on the surface of the protein (Extended Data Fig. 2a, c, d), indicating that they are membrane proteins. To examine their subcellular locations, we expressed OlS_1- and S_2-DUF247-1-GFP fusion proteins and found that they are localized on the plasma membrane in *N. benthamiana* (Fig. 2b), confirming their membrane locations.

Since OlS_1-HPS10 and OlS_1-DUF247-1 are extracellular and membrane proteins, respectively, we hypothesized that they should interact with each other. To examine this possibility, we did BiFC and SFLC assays and found that cognate OlS_1-HPS10 and OlS_1-DUF247-1 or OlS_2-HPS10 and OlS_2-DUF247-1 interactions occurred on the membrane (Fig. 2c), and, notably, we found that the intensities of cognate interactions were stronger than that of non-self combinations (Fig. 2d), suggesting that the cognate proteins likely form a partnership.

Taken together, these results suggest that *OlS_1-HPS10* and *OlS_1-DUF247-1* encode the candidate *S* genes in *O. longistaminata* and their cognate products form a membrane-localized signaling module.

OlS_1-HPS10 **controls the pistil function of SI**

To examine if *OlS_1-HPS10* functions in SI, we used a CRISPR-Cas9-mutagenesis approach and obtained three independent transgenic lines (*H1*, *H2*, and *H3*) with 1 bp or 5 bp deletions in different sites, all resulting in premature translational terminations (Fig. 3a), indicating they are the loss of function deletions. We observed that all three edited lines set seeds after self-pollination while WT failed to do so (Fig. 3b), suggesting that the loss of function of *OlS_1-HPS10* breaks down SI. To further examine this finding, we performed the reciprocal crosses between the edited line *H1* and wild type and found that the pistils but not its pollen are compatible with wild type (Fig. 3b), confirming that the *OlS_1-HPS10* knockouts specifically lead to the breakdown of the pistil function.

To examine if the pistil breakdown of SI is inheritable, we firstly identified the edited sequences of the transgenic T_0 and T_1 progeny (T_1-X) in three independent CRISPR-Cas9 edited lines and then analyzed their genotypes using allele-specific PCR primers (Extended Data Fig. 5, Supplementary Table 1 and 2), and found that all T_1 progeny in three edited lines were segregated into two classes (*S_1S_1* and *S_1S_2*), and no other genotypes were observed and *S_1S_1*: *S_1S_2* ratio close to 1:1 (Fig. 3b, Supplementary Table 2), indicating that the SI breakdown occurs specifically for *S_1* haplotype in all three *OlS_1-HPS10* edited lines. To confirm this, we used other allele-specific PCR primers of *OlS_1*-and *S_2-DUF247-1* as well as *OlS_1*- and *S_2-DUF247-2* and found that they segregated in 1:1 (*S_1S_1*: *S_1S_2*) ratio and no *S_2S_2* genotype detected (Fig. 3b, Supplementary Table 2), demonstrating that the loss of function of *OlS_1-HPS10* renders its pistil compatible with *S_1* but not *S_2* pollen. Taken together, our results indicate that OlS_1-HPS10 controls the pistil function of the *S_1* haplotype in *O. longistaminata* SI in a self-recognition manner.

OlS_1-DUF247-1 controls the pollen function of SI

To examine the role of *OlS_1-DUF247-1* in SI, we also obtained three independent gene-edited lines using CRISPR-Cas9, resulting from an 7 bp deletion in line *D1*, an 11 bp deletion in *D2*, and a 1 bp deletion in *D3* (Fig. 3a), all leading to premature

translational terminations, indicating that they are the loss of function mutations. As a consequence, the edited lines all set seeds compared with WT (Fig. 3c), indicating a role in SI, consistent with the results of the reciprocal crosses between the *D1* line and wild type showing that its pollen but not pistils are compatible with wild type (Fig. 3c). To further confirm this, we did the genotype analysis of the *OlS$_1$-DUF247-1* CRISPR-Cas9-edited lines using PCR with allele-specific primers and found that all T_1 progeny (T_1-X) had either S_1S_1 or S_1S_2 genotypes (Fig. 3c, Extended Data Fig. 5), and the genotypes segregated close to 1:1 ratio, and no S_2S_2 genotype detected (Supplementary Table 1 and 2), indicating that the loss of function of *OlS$_1$-DUF247-1* made S_1 pollen but not S_2 pollen compatible with S_1S_2. Taken together, these results demonstrate that *OlS$_1$-DUF247-1* controls the pollen SI of *O. longistaminata* also in a self-recognition manner.

The origin of the *O. longistaminata S*-locus

Since the *S* genes of *O. longistaminata* are different from those already identified in eudicots[28,29], their origin and evolution remain unclear. Thus, we first annotated OlS-HPS10-like polypeptides in seed plants and found their exclusive presence in the Oryzoideae and Pooideae (Fig. 4a), suggesting that they are the Oryzoideae- and Pooideae-specific. By contrast, OlS-DUF247-1/-2-likes with a conserved DUF247 domain showed a universal existence in eudicots, monocots, basal angiosperms and the gymnosperm species *Ginkgo biloba* and mainly grouped into three clusters (Fig. 4a), suggesting their presence since the origin of the seed plants. Among them, the Oryzoideae and Pooideae members syntenic to *OlS-DUF247-1* formed Cluster I, whereas those to *OlS-DUF247-2* clustered with the members from monocots, eudicots, basal angiosperms and *Ginkgo biloba* (Extended Data Fig. 6), indicating that OlS-DUF247-1 rather than -2 belongs to the Oryzoideae- and Pooideae-specific DUF247 cluster, showing its single origin in the Poaceae family.

To further examine the evolutionary relationship of the *O. longistaminata S*-locus with Type 1-4 (ref. 29) and the recently characterized Turneraceae *S*-locus (Type 5)[28],

we annotated its presence in angiosperm families and Ginkgoaceae and found that it is the subfamily specific (Type 6) (Fig. 4b, Extended Data Fig. 7 and 8). Ancestral state reconstructions showed that Type 6 *S* is the most likely state in the most recent common ancestor (MRCA) of the Bambusoideae, Oryzoideae and Pooideae (BOP) clade around 50 million years ago (MYA) (Fig. 4b), consistent with the single origin of the ancestral *OlS-HPS10* and *OlS-DUF247-1* linkage. Taken together, our results show that the ancestral *O. longistaminata S* locus occurred as an early Poaceae lineage-specific innovation in angiosperms.

Discussion

The Poaceae SI system is known to be gametophytically controlled by two independently segregating polymorphic loci, *S* and *Z* (refs 1-5), but their molecular identities have remained elusive. In this study, we found that the *S* locus of *O. longistaminata* encodes the polymorphic pistil-specific small secreted polypeptide OlS-HPS10s and pollen-specific transmembrane protein OlS-DUF247-1s, which interact in a cognate manner to elicit self-pollen rejection. Based on our results, we propose a self-recognition model for control of gametophytic SI in *O. longistaminata* (Fig. 4c). For self pollen, the extracellular domain of OlS_1-DUF247-1 or OlS_2-DUF247-1 located on the pollen membrane can recognize and interact with its cognate OlS_1-HPS10 or OlS_2-HPS10 secreted by the stigma leading to self-pollen incompatibility (SPI), whereas, for non-self pollen, the absence of interaction of OlS_3-DUF247-1 with either OlS_1- or OlS_2-HPS10 results in cross-pollen compatibility (CPC). Nevertheless, we know little about the molecular identify of the *Z* locus. In perennial ryegrass, a conserved DUF247 domain encoding gene also appears to be localized in the *Z* locus[10]. It would be interesting to identify the *Z* locus gene (s) in *O. longistaminata* to examine how the *S* and *Z* loci function together to control SI.

In conclusion, our results have revealed the molecular identifies of the *S* locus of *O. longistaminata*, representing an early Poaceae lineage-specific innovation of SI in

angiosperms. Our findings also provide a molecular means to alter its SI phenotype to benefit rice breeding and improvement. Further experiments especially the structural analyses of the *S* locus products as well as the molecular identification of the *Z* locus are expected to provide further insights into the molecular mechanism of the grass SI.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available online.

References

1 Lundqvist, A. Self-incompatibility in rye. *Hereditas* **40**, 278-294 (1954).

2 Hayman, D. The genetical control of incompatibility in *Phalaris Coerulescens* Desf. *Aust. J. Biol. Sci.* **9**, 321-331 (1956).

3 Cornish, M.A., Hayward, M.D. & Lawrence, M.J. Self-incompatibility in ryegrass. I Genetic control in diploid *Lolium perenne* L. *Heredity* **50**, 35-45 (1979).

4 de Nettancourt, D. *Incompatibility and Incongruity in Wild and Cultivated Plants. Springer*, *Berlin Heidelberg*, *Germany* (2001).

5 Yang, B., Thorogood, D., Armstead, I. & Barth, S. How far are we from unravelling self-incompatibility in grasses? *New Phytol.* **178**, 740-753 (2008).

6 Franklin-Tong, V.E. *Self-Incompatibility in Flowering Plants. Springer*, *Berlin Heidelberg*, *Germany* (2008).

7 Wricke, G. & Wehling, P. Linkage between an incompatibility locus and a peroxidase isozyme locus (Prx 7) in rye. *Theor. Appl. Genet.* **71**, 289-291 (1985).

8 Gertz, A. & Wricke, G. Linkage between the incompatibility locus *Z* and a β-glucosidase locus in Rye. *Plant Breed.* **102**, 255-259 (1989).

9 Kakeda, K. *et al.* Molecular and genetic characterization of the *S* locus in *Hordeum bulbosum* L., a wild self-incompatible species related to cultivated barley. *Mol. Genet. Genomics* **280**, 509-519 (2008).

10 Shinozuka, H., Cogan, N.O., Smith, K.F., Spangenberg, G.C. & Forster, J.W. Fine-scale comparative genetic and physical mapping supports map-based cloning strategies for the self-incompatibility loci of perennial ryegrass (*Lolium perenne* L.). *Plant Mol. Biol.* **72**, 343-355 (2010).

11 Manzanares, C. *et al.* A gene encoding a DUF247 domain protein cosegregates with the *S* self-incompatibility locus in perennial ryegrass. *Mol. Biol. Evol.* **33**, 870-884 (2016).

12 Lian, X. *et al.* Confirmation of a gametophytic self-incompatibility in *Oryza longistaminata*. *Front. Plant Sci.* **12**, 576340 (2021).

13 Rohner, M. *et al.* Fine-mapping and comparative genomic analysis reveal the gene composition at the *S* and *Z* self-incompatibility loci in grasses. *bioRxiv*, doi:10.1101/2022.07.18.499170 (2022).

14 Takayama, S. & Isogai, A. Self-incompatibility in plants. *Annu. Rev. Plant Biol.* **56**, 467-489 (2005).

15 Zhang, Y., Zhao, Z. & Xue, Y. Roles of proteolysis in plant self-incompatibility. *Annu. Rev. Plant Biol.* **60**, 21-42 (2009).

16 Fujii, S., Kubo, K. & Takayama, S. Non-self- and self-recognition models in plant self-incompatibility. *Nat. Plants* **2**, 16130 (2016).

17 Bedinger, P.A., Broz, A.K., Tovar-Mendez, A. & McClure, B. Pollen-pistil interactions and their role in mate selection. *Plant Physiol.* **173**, 79-90 (2017).

18 Anderson, M.A. *et al.* Cloning of cDNA for a stylar glycoprotein associated with expression of self-incompatibility in. *Nature* **321**, 38-44 (1986).

19 McClure, B.A. *et al*. Style self-incompatibility gene products of *Nicotianaalata* are ribonucleases. *Nature* **342**, 955-957 (1989).

20 Lai, Z. *et al*. An F-box gene linked to the self-incompatibility (*S*) locus of *Antirrhinum* is expressed specifically in pollen and tapetum. *Plant Mol. Biol*.

50, 29-42 (2002).

21 Sijacic, P. *et al*. Identification of the pollen determinant of S-RNase-mediated self-incompatibility. *Nature* **429**, 302-305 (2004).

22 Qiao, H. *et al*. The F-box protein AhSLF-S_2 controls the pollen function of S-RNase-based self-incompatibility. *Plant Cell* **16**, 2307-2322 (2004).

23 Takasaki, T. *et al*. The *S* receptor kinase determines self-incompatibility in *Brassica* stigma. *Nature* **403**, 913-916 (2000).

24 Schopfer, C.R., Nasrallah, M.E. and Nasrallah, J.B. The male determinant of self-incompatibility in *Brassica*. *Science* **286**, 1697-1700 (1999).

25 Foote, H.C. *et al*. Cloning and expression of a distinctive class of self-incompatibility (*S*) gene from *Papaver rhoeas* L. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 2265-2269 (1994).

26 Wheeler, M.J. *et al*. Identification of the pollen self-incompatibility determinant in *Papaver rhoeas*. *Nature* **459**, 992-995 (2009).

27 Li, J. *et al*. Genetic architecture and evolution of the *S* locus supergene in *Primula vulgaris*. *Nat. Plants* **2**, 16188 (2016).

28 Shore, J.S., Hamam, H.J., Chafe, P.D.J., Labonne, J.D.J., Henning, P.M., and McCubbin, A.G. The long and short of the *S*-locus in *Turnera* (Passifloraceae). *New Phytol.* **224**, 1316-1329 (2019).

29 Zhao, H. *et al.* Origin, loss, and regain of self-incompatibility in angiosperms. *Plant Cell* **34**, 579-596 (2022).

30 Thorogood, D., Kaiser, W.J., Jones, J.G. & Armstead, I.P. Self-incompatibility in ryegrass 12. Genotyping and mapping the *S* and *Z* loci of *Lolium perenne* L. *Heredity* **88**, 385-390 (2002).

31 Philipp, U., Wehling, P. & Wricke, G. A linkage map of rye. *Theor. Appl. Genet.* **88**, 243-248 (1994).

32 Nayar, N. M. Prevalence of self-incompatibility in *Oryza barthii* Cheval.: Its bearing on the evolution of rice and related taxa. *Genetica* **38**, 521-527 (1967).

33 Ghesquiere, A. Evolution of *Oryza lngistaminata*. *Rice Genet*. 15-27 (1986).

34 Zhang, Y. *et al.* Genome and comparative transcriptomics of African wild rice *Oryza longistaminata* provide insights into molecular mechanism of rhizomatousness and self-Incompatibility. *Mol. Plant* **8**, 1683-1686 (2015).

35 Reuscher, S. *et al*. Assembling the genome of the African wild rice *Oryza longistaminata* by exploiting synteny in closely related *Oryza* species. *Commun. Biol.* **1**, 162 (2018).

36 Li, W. *et al*. Improved hybrid *de novo* genome assembly and annotation of African wild rice, *Oryza longistaminata*, from Illumina and PacBio sequencing reads. *Plant Genome* **13**, e20001 (2020).

Figures

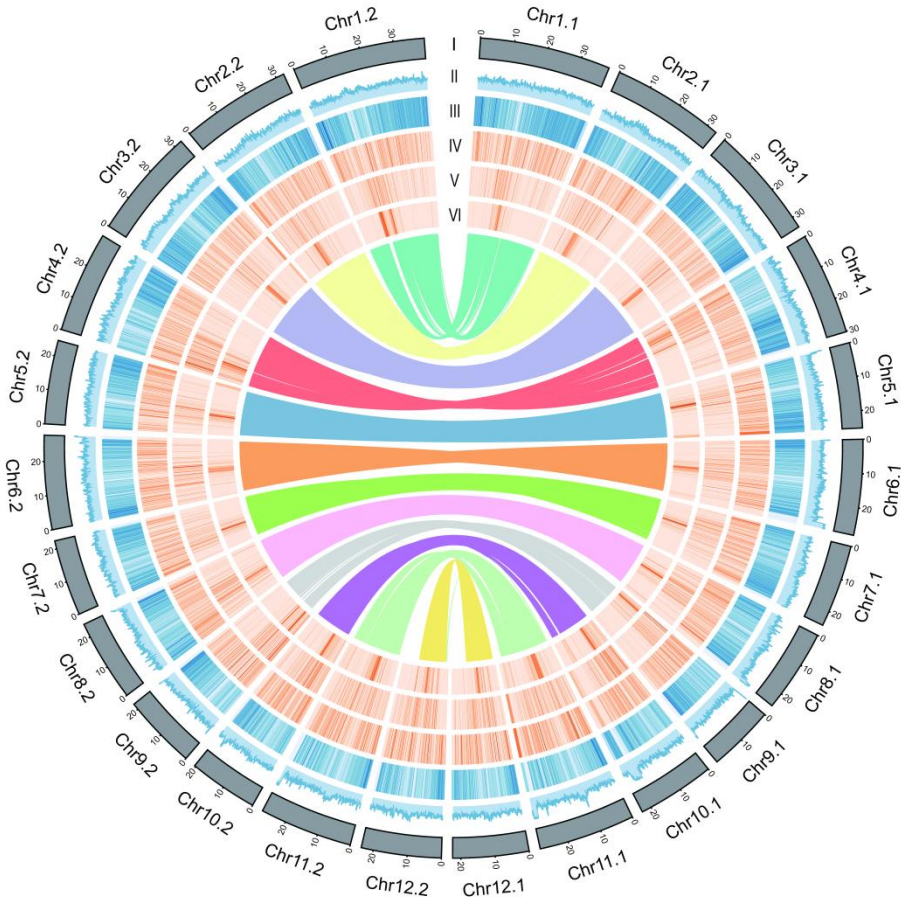


Fig. 1 | Genome landscape of *O. longistaminata.* Chr1.1-Chr12.1 and Chr1.2-Chr12.2 represent two haplotype-resolved genomes, respectively. Concentric circles show its structural and functional genomic features. I-VI show chromosome number, GC content, gene density, LINE density, Gypsy density, and Copia density, respectively. The central part is chord diagrams showing the syntenic relationship between two haplotype genomes.

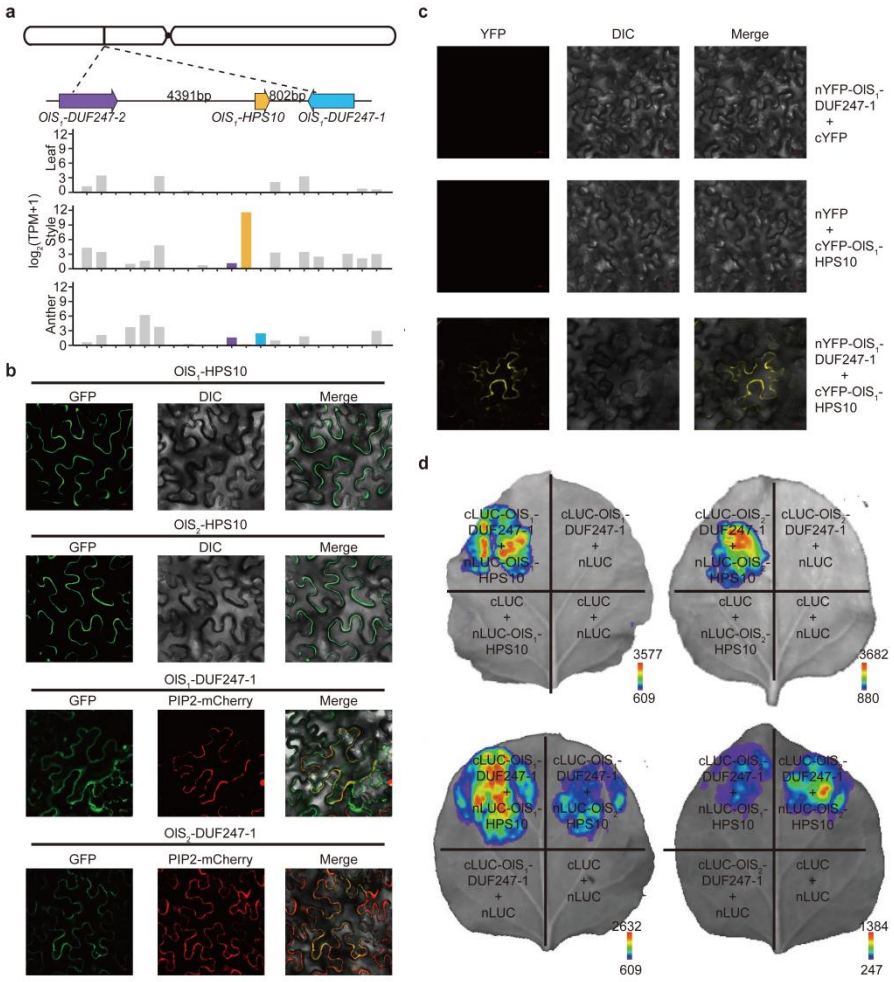


Fig. 2 | *OlS1-HPS10* and *OlS1-DUF247-1* encode the candidate *S* genes in *O. longistaminata*. a, Identification of *OlS1-HPS10* and *OlS1-DUF247-1*. Top, schematic diagram showing the chromosomal locations of *OlS1-HPS10* and *OlS1-DUF247s*. The arrows indicate transcriptional orientations. Purple, orange, and blue blocks represent *OlS1-DUF247-2*, *OlS1-HPS10*, and *OlS1-DUF247-1*, respectively. Bottom, the expression patterns of the *S1*-locus and its flanking regions as determined by RNA-seq of anther, style, and leaf. The grey blocks indicate other annotated genes. TPM, Transcripts Per kilobase per Million mapped reads. **b,** Subcellular locations of OlS_1- and S_2-HPS10-GFP and OlS_1- and S_2-DUF247-1-GFP fusion proteins in *N.benthamiana*. Scale bar, 20 μm. **c,** BiFC assays of OlS_1-HPS10 and OlS_1-DUF247-1. Scale bar, 20 μm. **d,** Physical interactions of OlS_1- or S_2-HPS10 with OlS_1- or S_2-DUF247-1 by SFLC assays in *N.benthamiana.* OlS_1- and S_2-HPS10 are fused with C-LUC while OlS_1- and S_2-DUF247-1 with N-LUC.

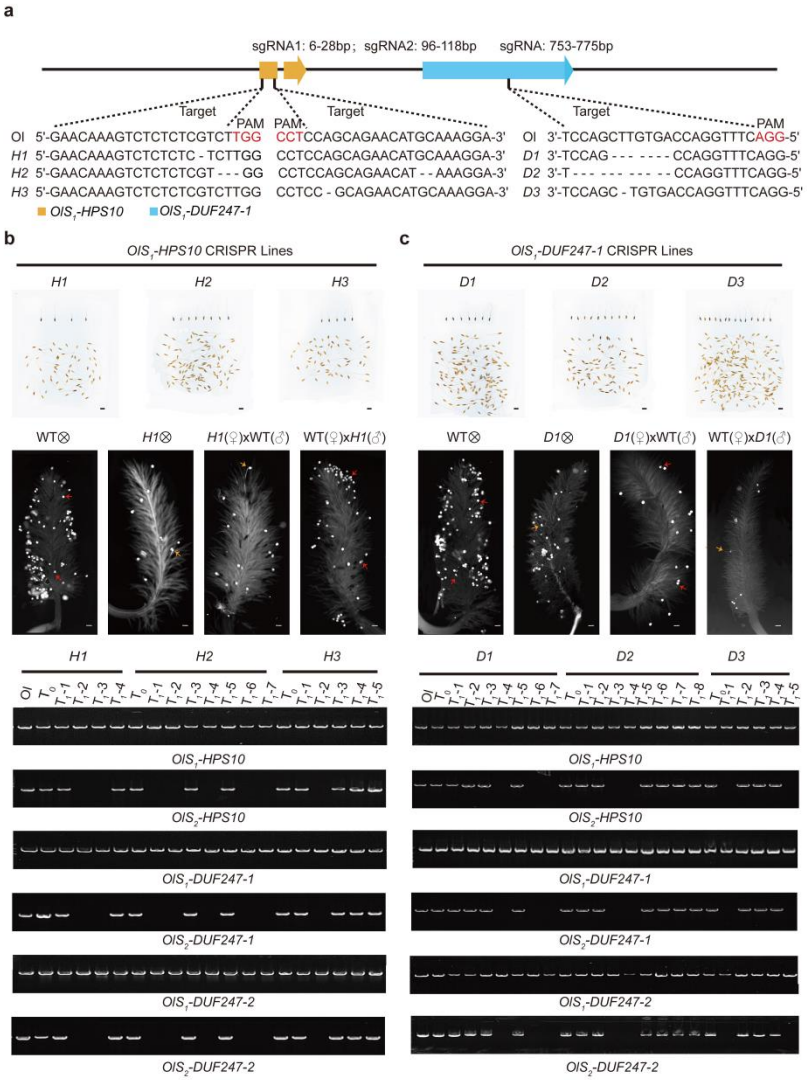


Fig. 3 | Disruptions of *OlS$_1$-HPS10* or *OlS$_1$-DUF247-1* break self-incompatibility of *O. longistaminata*. a, Schematic diagram showing CRISPR-Cas9 targets of *OlS$_1$-HPS10* and *OlS$_1$-DUF247-1*. Orange and blue boxes indicate *OlS$_1$-HPS10* and *OlS$_1$-DUF247-1*, respectively. Vertical black lines denote the sgRNAs, and the protospacer-adjacent motifs (PAM, NGG) are indicated in red. *Ol*, *H1-3* and *D1-3* indicate wild type, *OlS$_1$-HPS10*, and *OlS$_1$-DUF247-1* edited lines, respectively. Dashes represent introduced gaps. **b** and **c,** Seed sets, the reciprocal crosses and PCR genotype analysis of the transgenic CRISPR-Cas9-edited lines of *OlS$_1$-HPS10* (*H1-3*) and *OlS$_1$-DUF247-1* (*D1-3*). Top, seeds of the *O. longistaminata* CRISPR-Cas9-edited lines after self-pollination. The seeds from a single panicle of each line are shown on the top half together with the rest empty husks on the bottom half part. Scale bar, 1cm.

363 Middle, the reciprocal crosses of the *O. longistaminata* CRISPR-Cas9-edited lines
364 with wild type. Aniline blue staining of the pollen tubes eighteen hours after
365 pollination with red and orange arrows indicate the adherent pollen and pollen tubs,
366 respectively. Scale bar, 200 μm. Bottom, PCR genotyping of T_0 and T_1 progeny (T_1-
367 X) with allele-specific primers from three *S*-locus genes (*HPS10*, *DUF247-1*, and
368 *DUF247-2*).

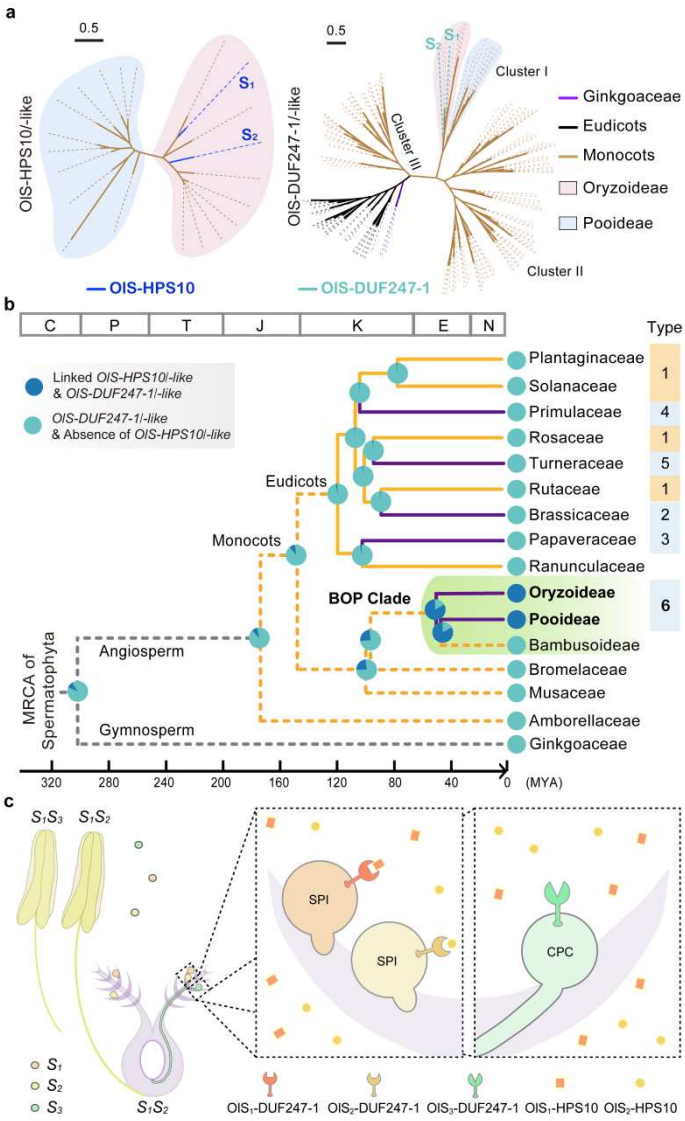


Fig. 4 | Origin and evolution of the *O. longistaminata S* locus in angiosperms. a, ML phylogenetic trees of the HPS10 family of Oryzoideae and Pooideae (left) and the DUF247 family of angiosperms (right). OlS_1-HPS10 and OlS_1-DUF247-1 are highlighted by blue and cyan branch lines, respectively. **b**, A phylogenetic tree showing the evolutionary relationship of six molecular types of SI among 15 angiosperm families and gymnosperms (Ginkgoaceae). Top, abbreviations of geological periods with N indicating Neogene, E Paleogene, K Cretaceous, J Jurassic, T Triassic, P Permian and C Carboniferous. The rectangle boundaries correspond to the bottom timeline (MYA, million years ago). Pie charts at each node indicate the probabilities of its ancestral states following analyses of two traits represented by color-filled circles, including linked *OlS-HPS10/-like* & *OlS-DUF247-1/-like* (dark

blue) and *OlS-DUF247-1*/*-like* & absence of *OlS-HPS10*/*-like* (cyan). The grey dotted lines indicate the absence of linked T2 *RNase* and *FBA*/*FBK* genes and the orange linked class I/II T2 *RNase* and *FBA*/*FBK* genes, while the solid orange and purple lines Type-1 and Type 2-6 *S*, respectively. The green oval background illustrates the BOP clade of Poaceae. MRCA: most recent common ancestor. **c**, Model of gametophytic SI in *O. longistaminata*. Schematic diagrams show anthers (yellow), self (circles filled with orange and yellow colors) and non-self (green) pollen, and pollinated pistil (purple) of *O. longistaminata*. The *S* haplotypes are indicated below the anthers and the pistil. Enlarged dotted box shows SPI (Self-Pollen Incompatibility) and CPC (Cross-Pollen Compatibility) occurring for self and cross pollen, respectively. Self-recognitions between OlS_1-DUF247 and OlS_1-HPS10 or OlS_2-DUF247 and OlS_2-HPS10 eventually result in the growth inhibition of self pollen tube, whereas the absence of non-self recognition for OlS_3-DUF247-1 leads to cross pollen tube growth.

Methods

Plant materials and growth conditions

One accession of *Oryza longistaminata* (AA genome, IRGC 104977) that originated from Kenya was kindly provided by the International Rice Research Institute (IRRI). To ensure normal heading under the ecological conditions of Yangzhou, Jiangsu Province, China (119º25'E, 32º23'N, at an altitude of 28 m), *O. longistaminata* and the transgenic CRISPR-Cas9-edited T_0 and T_1 lines of *OlS_1-HPS10* and *OlS_1-DUF247-1* were exposed to a short-day treatment (8 h of bright sunshine per day). The treatment began at the peak of tillering and continued for more than 20 days[37]. All plants were also grown in natural short-day conditions of Lingshui, Hainan Province, China (110º12'E, 18º32'N, at an altitude of 10 m). Field irrigation, fertilizer application and pest control followed essentially the normal agricultural practice.

Genome sequencing, assembly and annotation

To obtain haplotype-resolved genomes of *O.longistaminata*, we first sequenced the leaf DNA from an individual with HiFi PacBio technology at ~40-fold coverage per haplotype. Chromosome conformation capture (Hi-C) library was constructed from the same individual and then sequenced on the Illumina platform. The HiFi reads were assembled using hifiasm (0.16.1-r375) along with Hi-C reads[38]. Two resulting sets of contigs were considered as fully resolved haploid genome assemblies. Hi-C reads were mapped to the haplotigs by Bwa[39], and the Hi-C links were used for scaffolding by 3d-DNA[40]. Hi-C interaction maps were then inspected by Juicebox[41] and a few mis-assemblies were manually curated followed by generating chromosome-level scaffolds. Syntenic relationships of the scaffolds with *O.sativa* chromosomes were checked by whole genome alignment using MUMmer4 (ref. 42) and visualized by SyRI[43]. Orthologs in liliopsida_odb10 were used to compute the BUSCO[44] scores and assess assembly completeness. Repeat sequences were analyzed and annotated using RepeatMasker[45] with curated *Oryza* repeat database download from http://rice.uga.edu/index.shtml, parameters setting as "-s -nolow -norna -gff -engine ncbi". Gene prediction of *O. longistaminata* genome was implemented with combination of *ab initio*, RNA-seq-based, and homology-based approaches on the repeat-masked assembly. The RNA-seq datasets and *Oryza* protein were fed into the BRAKER2 (ref. 46) pipeline to derive final gene models. The functional annotation was performed using InterProscan[47] for protein-domain identification.

Identification of the *S*-locus in *O. longistaminata*

The *S*-locus was determined to locate on chromosome 5 based on sequence structure and gene homologous search in *Hordeum bulbosum*[9] and perennial ryegrass[11]. Therefore, two intronless genes annotated with the term PF03140 were considered as *OlS-DUF247* genes. Furthermore, *OlS-HPS10* gene was found between two *OlS-DUF247-1* and *-2* and encodes a protein product size of 90 amino acids and one intron. Moreover, the linkage relationship of three genes was confirmed by continuous HiFi reads stack for both two haplotypes to avoid heterozygous region assembly

collapse. The reads mapping was performed using minimap2 (ref. 48) with parameters "-ax map-hifi". Jalview[49] was used to analyze and plot multiple sequence alignment.

Synteny analysis

The synteny between two haplotypes of *O. longistaminata* was analyzed and plotted by MCscan via jcvi[50], which also used to plot the synteny of the *S*-locus among *O. longistaminata*, *O. sativa*, *L. perenne*, *H. bulbosum* and *S. cereale*. The *S*-locus genes were verified by TBLASTN[51] and FGENESH[52].

RNA-seq analysis

Total RNA was isolated from three tissues (leaf, anther, and style) using the TRIzol reagent (Invitrogen) according to the manufacturer's instructions. RNA was qualified and quantified using NanoDrop 2000 and Agilent 2100/4200 system. Strand-specific RNA libraries were constructed and sequenced using the Illumina PE150 platform (Berry Genomics; Beijing, China) (https://www.berrygenomics.com). Raw reads were firstly filtered by fastp v0.23.0 (ref. 53) with default parameters. Clean reads were aligned to the genome using HISAT2 v2.2.1 (ref. 54). FeatureCounts v2.0.1 (ref. 55) was used to count the reads numbers mapped to each gene.

RT-PCR assays

Total RNAs were extracted from anthers and pistils of *O. longistaminata* plants using the TRIzol reagent (Invitrogen) according to the manufacturer's protocol and treated with RNase-free DNase I (Invitrogen) to remove contaminating genomic DNAs. The cDNAs were then reverse-transcribed using a cDNA synthesis kit (TRANSGEN, AE311-02). The full-length cDNAs of *OlS_1-HPS10*, *OlS_1-DUF247-1, OlS_2-HPS10* and *OlS_2-DUF247-1* were used as templates for RT-PCR assays. Relevant primer sequences are given in Supplementary Table 1.

Subcellular localization

Full-length cDNAs of *OlS_1-HPS10*, *OlS_1-DUF247-1, OlS_2-HPS10* and *OlS_2-DUF247-1* were amplified from *O. longistaminata* and then inserted into *pCAMBIA1300-35S-GFP* vectors (GFP, green fluorescent protein) to generate fusion constructs. The *pCAMBIA1300-PIP2-mCherry* construct was used as a plasma membrane localization marker. Co-transfection of constructs into tobacco leaf epidermal cells by Agrobacterium-mediated infiltration to test the subcellular localization. After 48 hours of incubation in the dark, the GFP signal was captured using a confocal microscope (Zeiss LSM780). Each assay was repeated at least three times. Relevant primer sequences are given in Supplementary Table 1.

Split firefly luciferase complementation (SFLC) assays

Full-length cDNAs of *OlS_1-HPS10*, *OlS_1-DUF247-1, OlS_2-HPS10*, and *OlS_2-DUF247-1* were inserted into *pCAMBIA1300-35SCluc-RBS* or *pCAMBIA1300-35S-HA-Nluc-RBS* vectors[56] to generate fusion constructs. Two different vectors (e.g., nLUC-OlS_1-DUF247-1 and cLUC-OlS_1-HPS10) can test a protein-protein interaction, together with the p19 silencing plasmid, were co-transfected into tobacco leaf epidermal cells by Agrobacterium-mediated infiltration. After 48 hours of incubation in the dark, the injected leaves were sprayed with 1 mM luciferin (Promega, E1605) and the LUC signal was captured using a cooled CCD imaging apparatus (Berthold, LB985). Each assay was repeated at least three times. Relevant primer sequences are given in Supplementary Table 1.

CRISPR-Cas9 vector construction and *O. longistaminata* transformation

Single guide RNAs (sgRNAs) were designed by the CRISPR-GE tool (http://skl.scau.edu.cn/)[57] and cloned into the binary vectors derived from *pCAMBIA1300*. The plasmids were then separately electroporated into *Agrobacterium tumefaciens* EHA105 and selected calli induced from young panicles of *O. longistaminata* were infected with Agrobacterium. The infected calli were screened under hygromycin selection followed by resistant callus differentiation and 1 cm

regenerated buds were subsequently transferred to rooting medium and cultured at 30ºC with light to induce root development as described (Patent No: ZL200910273352.3 Biorun Biosciences Co., Ltd.) (CN101717749A , https://pss-system.cponline.cnipa.gov.cn/retrieveList?prevPageTit=chagngui).

PCR genotyping

For genotyping of the T_0 and T_1 transgenic lines, their leaves were collected from each plant and genomic DNA was extracted by using the CTAB method[58]. PCR amplification was performed by using primers flanking the sgRNA target sites and Sanger sequencing was used to identify mutations in the target genes of T_0 and T_1 plants. The sgRNAs used in this study are given in Supplementary Table 1.

Identity and similarity analyses of amino acid sequences

The sequences were aligned by MAFFT and their identities and similarities were calculated with default settings at SIAS (http://imed.med.ucm.es/Tools/sias.html).

Amino acid sequence motif detection

To measure the variability at each site precisely, a Normed Variability Index was adapted as previously described[59], which was defined as the number of different amino acids at a site divided by the frequency of the most common one (s). The Normed Variability Index at a site is defined as follows:

$$\text{Normed Variability Index} = \frac{\log (\text{Variability Index})}{\log (\text{Number of residues at the site})}$$

Analysis of the six sequences of HPS10 and DUF247-1 family members displayed one hypervariable region (the peak, HV) for HPS10 and three hypervariable regions (the peaks, HV1-HV3) for DUF247-1. The Normed Variability Index is defined to be 0 and -0.3, respectively.

Signal peptide and transmembrane region predictions

The signal peptide prediction was performed by SignalP 6.0 (ref. 60) and the transmembrane region DeepTMHMM[61].

3D structure prediction

3D structures for OlS_1-HPS10 and OlS_1-DUF247-1 proteins were generated by ab initio modeling using a local copy of AlphaFold2 (ref. 58). The predicted structures of both proteins were ranked by a per-residue confidence score (pLDDT) and one with the highest score as the structural models. All structural models were visually shown and edited using PyMol (https://pymol.org/2/).

***In vivo* pollen germination and pollen tube growth**

Decolorized Aniline Blue solution was prepared by making a 0.1% (w/v) solution of Aniline Blue (Acros Organics), adding 1 M NaOH drop-wise, and incubating the solution overnight at 48 °C until it became a transparent yellow color. Following in vivo pollination, mature flowers in *O. longistaminata* and transgenic lines were pollinated with their pollen, respectively. After 18 hours, the pollinated pistils were removed from the plants and incubated in a fixing solution containing ethanol: acetic acid (3:1) for 2 h at room temperature. The fixed pistils were then washed with distilled water three times for 5 min each followed by incubation in 1 M NaOH softening solution overnight. Pistils were placed into decolorized Aniline Blue solution and allowed to stain for 12-24 h before being mounted on slides and observed using 405/440-480 nm setting or fluorescent microscopy. Images were acquired using a Zeiss 780 two-photon laser scanning confocal microscope (CLSM; Carl Zeiss).

Family-/species-level phylogenetic tree construction

Family-/species-level phylogenetic trees were constructed by phylomatic (version 3) and TimeTree (http://www.timetree.org/) with the divergence times derived from the latter. iTOL (https://itol.embl.de) was used for visualization, annotation, and management of the trees.

554

555 **Annotations of the Type 5 *S*- locus supergene**

556 To annotate the genomic structures of Type 5 *S*-locus in seed plants, the three linked
557 genes defining the Type 5 *S*-locus in *Turnera subulata*, including *TsBAHD*, *TsSPH1,*
558 and *TsYUC6* (ref. 28) (Supplementary Table 3) were separately performed TBLASTN
559 with the expected threshold cut-off of 0.05 and a 1 Mb genomic window was
560 restricted to search for their homologs as described[29].

561

562 **Phylogenetic analyses of the *S* or *S*-like genes**

563 S-HPS10/-like and S-DUF247/-like amino acid sequences were obtained by
564 performing BLASTP or TBLASTN with BLAST version 2.2.29 against the protein or
565 genome databases listed in Supplementary Table 4 using amino acid sequences of
566 OlS_1- and OlS_2-HPS10 or OlS_1- and OlS_2-DUF247-1/2 proteins as queries,
567 respectively. BLAST was run using default settings with the expected threshold cut-
568 off of 0.05 and sequences containing both an N-terminal signaling peptide and PCAS
569 motif (for S-HPS10/-like) or DUF247 domain were retained. The full-length
570 sequences of S-HPS10/-like and S-DUF247/-like proteins were aligned through the L-
571 INS-i method in MAFFT version 7.407 (ref. 63) and curated by MEGAX to delete
572 gaps and sequences without conserved motifs. Then maximum likelihood (ML) trees
573 were constructed using IQ-TREE based on the best-fitting amino acid substitution
574 models determined by ModelFinder including a JTT + G4 model for S-HPS10/-like,
575 JTT + F + R6 for S-DUF247-1/-like families and JTT + F +R7 for S-DUF247-1/-2/-
576 like families and the branch supports were measured as described previously[29].

577

578 **Ancestral state reconstruction**

579 To examine the evolution of type-6 *S*-locus, we totally selected three discrete traits
580 and codified them as possessing linked *S-HPS10/-like* and *S-DUF247/-like* and *S-*
581 *DUF247/-like* & absence of *S-HPS10/-like*. To further evaluate the evolutionary
582 relationship of the six types of *S*-loci, a total of nine traits were selected, including

possessing type-1, -2, -3, -4, -5, -6 *S*, ancestral type-1 *S* consisting of linked class I/II T2 *RNase*s & *FBA*/*FBK* genes, class III T2/*S-like-RNase*s & absence of linked *SLF*s/*S-like SLF*s and Class I/II T2 *RNase*s & absence of linked *FBA*/*FBK* genes. Ancestral state reconstruction and the result annotation were conducted as previously described[29].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The whole genome sequencing and RNA-seq data reported in this study is available in National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation under project number PRJCA010642 that is publicly accessible at https://ngdc.cncb.ac.cn. The genome datasets and the accession numbers of the genes used in this study were listed in Supplementary Table 3 and 4.

Method references

37 Hu, Z. *et al.* Comparative analysis morphology, anatomical structure and transcriptional regulatory network of chlorophyll biosynthesis in *Oryza longistaminata*, *O. sativa* and their F1 generation. *PeerJ* **9**, e12099 (2021).

38 Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170-175 (2021).

39 Li, H., & Wren, J. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).

40 Dudchenko, O. *et al. De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92-95 (2017).

612 41 Durand, N.C. *et al.* Juicebox provides a visualization system for Hi-C contact
613 maps with unlimited zoom. *Cell Syst.* **3**, 99-101 (2016).
614 42 Marcais, G. *et al.* MUMmer4: A fast and versatile genome alignment system.
615 *PLoS Comput. Biol.* **14**, e1005944 (2018).
616 43 Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic
617 rearrangements and local sequence differences from whole-genome
618 assemblies. *Genome Biol.* **20**, 277 (2019).
619 44 Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov,
620 E.M. BUSCO: assessing genome assembly and annotation completeness with
621 single-copy orthologs. *Bioinformatics* **31**, 3210-3212 (2015).
622 45 Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0
623 http://www.repeatmasker.org (2013).
624 46 Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-genome
625 annotation with BRAKER. *Methods Mol. Biol.* **1962**, 65-95 (2019).
626 47 Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids*
627 *Res.* **33**, W116-120 (2005).
628 48 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*
629 **34**, 3094-3100 (2018).
630 49 Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. & Barton, G.J.
631 Jalview Version 2--a multiple sequence alignment editor and analysis
632 workbench. *Bioinformatics* **25**, 1189-1191 (2009).
633 50 Tang, H. *et al.* Synteny and collinearity in plant genomes. *Science* **320**, 486-
634 488 (2009).
635 51 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local
636 alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).
637 52 Solovyev, V., Kosarev, P., Seledsov, I. & Vorobyev, D. Automatic annotation
638 of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* **7 Suppl 1**,
639 S10.11-12 (2006).
640 53 Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ

preprocessor. *Bioinformatics* **34**, i884-i890 (2018).

54 Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907-915 (2019).

55 Liao, Y., Smyth, G.K. & Shi, W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).

56 Chen, H. *et al.* Firefly luciferase complementation imaging assay for protein-protein interactions in plants. *Plant Physiol.* **146**, 368-376 (2008).

57 Xie, X. *et al.* CRISPR-GE: a convenient software toolkit for CRISPR-based genome editing. *Mol. Plant* **10**, 1246-1249 (2017).

58 Allen, G.C., Flores-Vergara, M.A., Krasynanski, S., Kumar, S. & Thompson, W.F. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat. Protoc.* **1**, 2320-2325 (2006).

59 Kheyr-Pour, A. *et al.* Sequence diversity of pistil S-proteins associated with gametophytic self-incompatibility in *Nicotiana alata*. *Sex. Plant Reprod.* **3**, 88-97 (1990).

60 Teufel, F. *et al.* SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat Biotechnol.* **40**, 1023-1025 (2022).

61 Hallgren, J. *et al.* DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. *bioRxiv* (2022).

62 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).

63 Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).

Acknowledgements

We thank Dr. Yangwen Qian from Wimi Biotechnology Co., Ltd. (Changzhou, China)

for the plant maintenance of *Oryza longistaminata.* This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA26030302), the National Natural Science Foundation of China (32270352 and 32030007) and the China Postdoctoral Science Foundation (2022T150696).

Author contributions

Y.X. conceived and designed the project. C.Y. provided the plant materials and assisted genetic analyses of *O. longistaminata*. Y.W. and Y.E.Z. performed the experiments. H.Z. performed the phylogenetic analyses of HPS10s and DUF247s, annotation of Type 5 *S*-locus of seed plants and ancestral state reconstruction of SI systems. S.Z. assembled and annotated the haplotype-resolved genome of *O. longistaminata*. H.K.Z. annotated HPS10s and S-DUF247s and performed RNA-seq and synteny analyses. Y.C., W.T. and X.T. assisted PCR genotyping. H.T. assisted confocal microscopy experiments. Y.G., H.H., Q.H., Y.J.Z. and Z.C. provided technical support. Y.X., Y.W., H.Z., S.Z. and Y.E.Z. analyzed the data and wrote the manuscript with the input of all other authors.

Competing interest declaration

The other authors declare no competing interests.

Correspondence and requests for materials should be addressed to Yongbiao Xue, Yu'e Zhang, Chuandeng Yi or Hong Zhao.

Supplementary Table 1

A list of primers used in this study.

Supplementary Table 2

The genotype analysis of the CRISPR-Cas9-edited lines.

Supplementary Table 3

Accession numbers of *Turnera subulata* (*Ts*) BAHD, SPH1 and YUC6 sequences used

699 in this study.

700 **Supplementary Table 4**

701 Accession numbers of DUF247 sequences used in this study.

702

703 **Supplementary File 1**

704 Fasta files of the sequence alignments corresponding to the phylogenetic analyses of

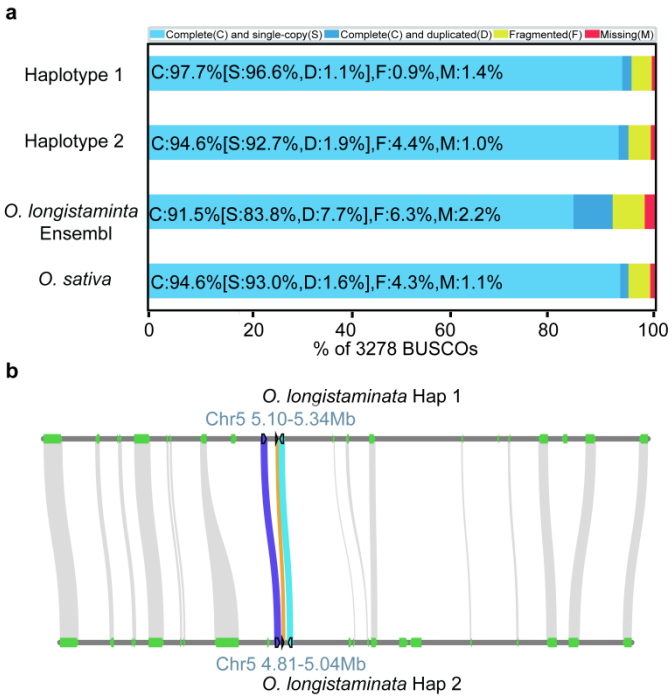
705 the Fig. 4a, Extended Data Fig. 6.

706 **Supplementary File 2**

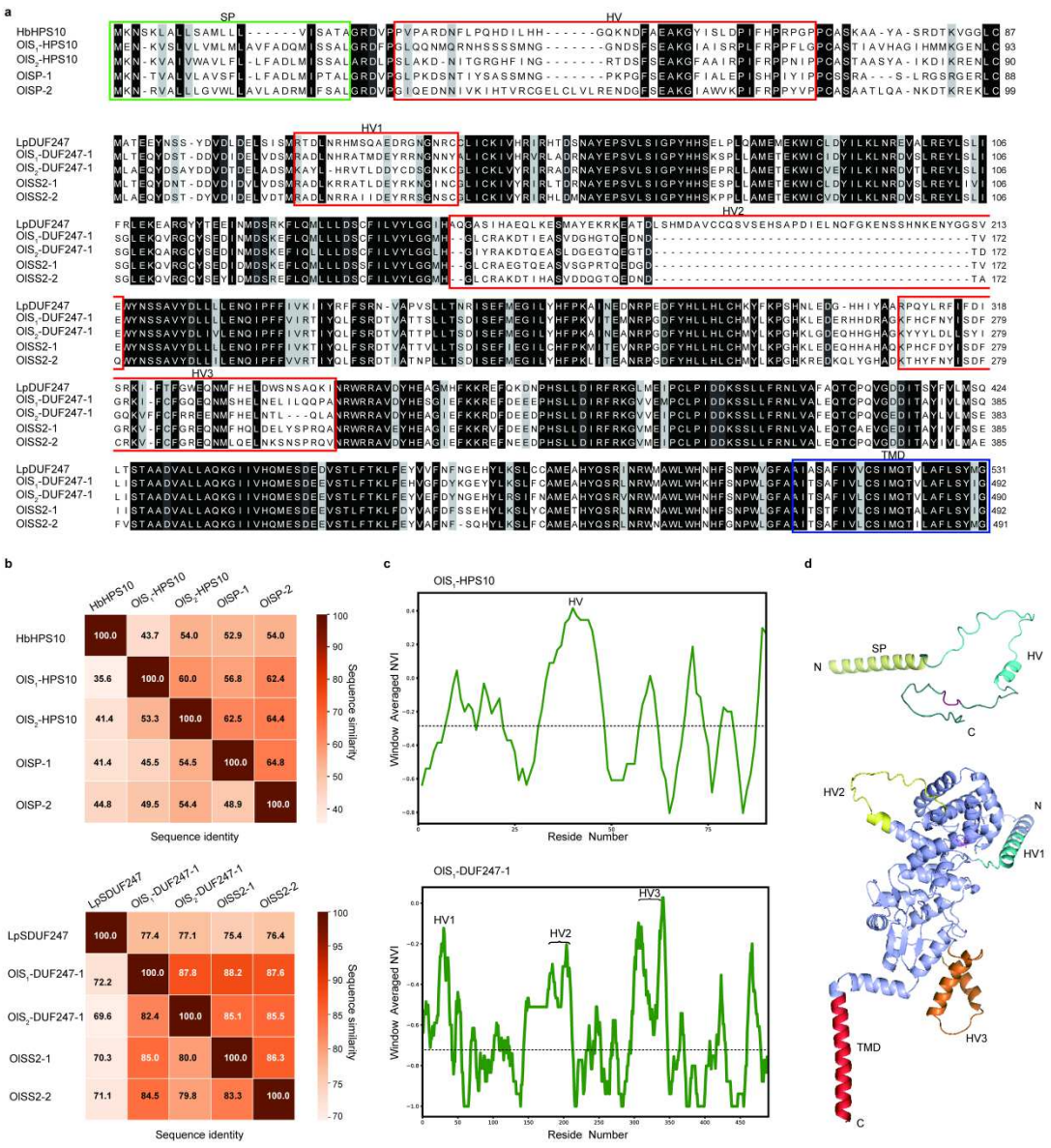
707 Tree text files corresponding to the phylogenetic analyses in Fig. 4a and

708 Extended Data Fig. 6.

Extended Data



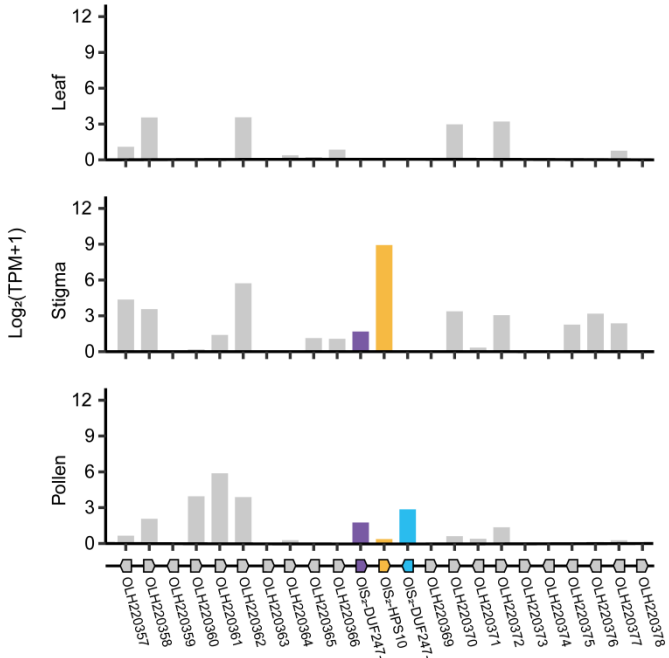
Extended Data Fig. 1 | Intergenomic comparisons of *O. longistaminata.* a, BUSCO scores of two haplotype-resolved genome assemblies obtained in this study, and *O.longistaminata* and *O.sativa* genomes from Ensembl database. **b,** The synteny of the *S*-locus and its flanking regions between two *S* haplotypes.



715

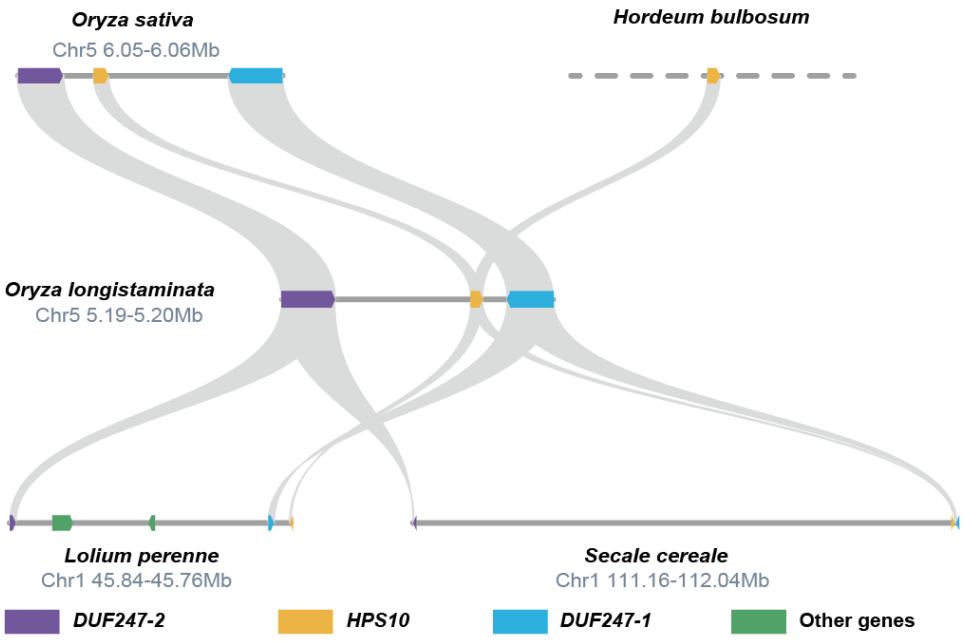
716 **Extended Data Fig. 2 | Structural features of OlS-HPS10s and OlS-DUF247s. a,**
717 Amino acid sequence alignments of HPS10s from *Hordeum bulbosum* (Hb), two *S*
718 haplotypes of *O. longistaminata* (OlS_1 and S_2) and two previously reported OlSP-1
719 and -2 (ref. 11) (Top) and DUF247-1s from *Lolium perenne* (Lp), two *S* haplotypes of
720 *O. longistaminata* (OlS_1 and S_2), and two published OlSS2-1 and -2 (ref. 11)
721 (Bottom). The conserved amino acids are indicated by shaded boxes with black
722 showing 100% identify. The signal peptide (SP), hypervariable regions (HVs), and
723 transmembrane domain (TMD) are highlighted in green, red, and purple boxes,
724 respectively. **b,** The amino acid sequence identities and similarities of HPS10s (top)
725 and DUF247-1s (bottom). **c,** Normed Variability Index of the HPS10 and DUF247-1

polypeptides. The peaks over the dotted lines represent predicted hypervariable regions (HVs). **d,** Predicted 3D structures of OlS_1-HPS10 and OlS_1-DUF247-1. Top, the predicted 3D structure of OlS_1-HPS10 with a signal peptide (SP, in yellow) and a hypervariable region (HV, in blue). Bottom**,** the predicted 3D structure of OlS_1-DUF247-1 with a transmembrane domain (TMD, in red) and three hypervariable regions (HVs, in blue, yellow and orange, respectively). The structural predictions were carried out using the Alpha Fold server (https://www.alphafold.ebi.ac.uk/).

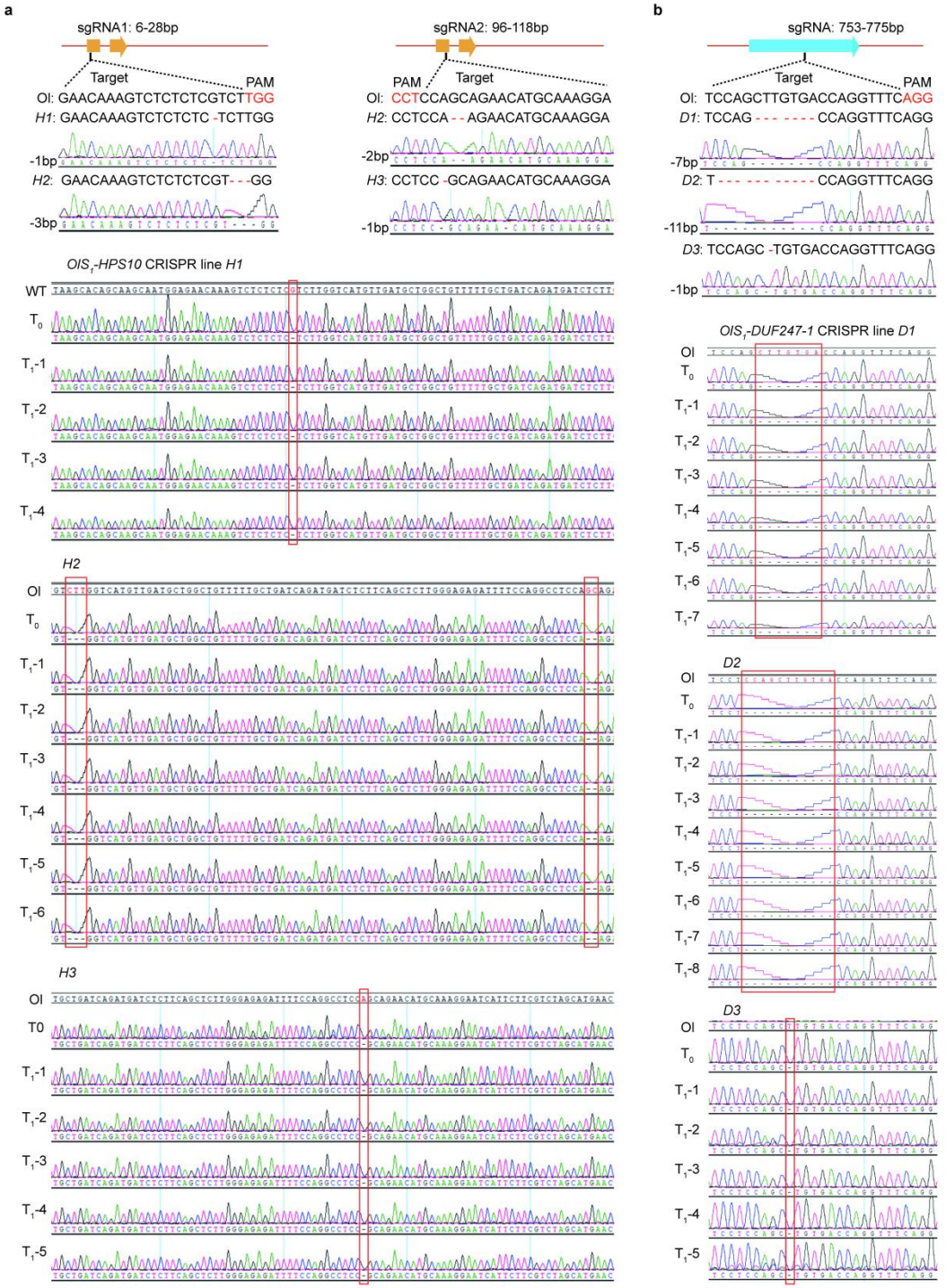


733

734 **Extended Data Fig. 3 | Transcriptional profiles of the *S_2*-locus and its flanking**
735 **regions in *O. longistaminata*.** The bottom boxed line is a schematic representation of
736 annotated genes. The purple, orange, blue, and grey blocks represent *OlS_2-DUF247-2,*
737 *OlS_2-HPS10*, *OlS_2-DUF247-1*, and other annotated genes, respectively. TPM,
738 Transcripts Per kilobase per Million mapped reads.

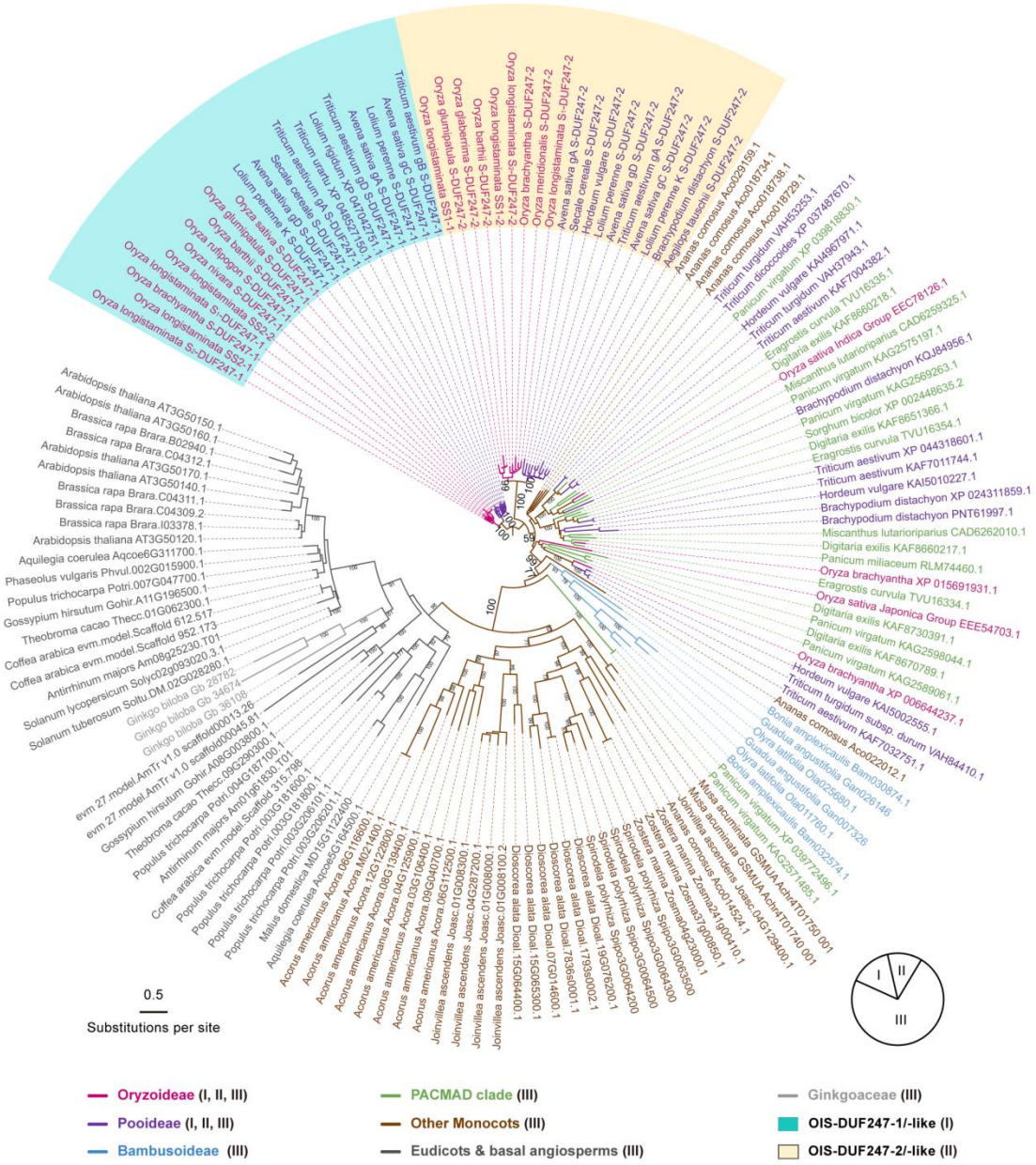


Extended Data Fig. 4 | Synteny of the *S*-locus regions from five Poaceae species. The grey ribbons show the synteny of the *S*-locus genes among *O. longistaminata*, *O. sativa*, *L. perenne*, *H. bulbosum*, and *S. cereale*. The purple, orange, blue and green blocks represent *DUF247-2*, *HPS10*, *DUF247-1*, and other genes, respectively.

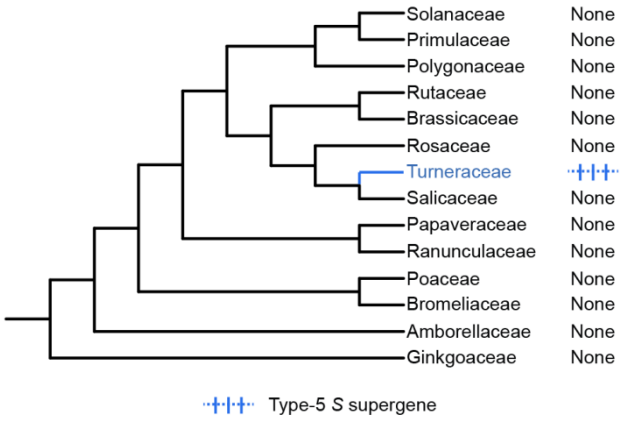


Extended Data Fig. 5 | DNA sequencing chromatograms of sgRNA target regions for *OlS1-HPS10* and *OlS1-DUF247-1* CRISPR-Cas9-edited lines. a and **b,** Sequences of *OlS1-HPS10* and *OlS1-DUF247-1* in three independent CRISPR-Cas9-edited lines. Vertical black lines denote sgRNA, and the protospacer-adjacent motifs (PAM, NGG) are indicated in red. Yellow and blue boxes indicate the coding

sequences. Dashes represent introduced gaps. DNA sequencing chromatograms are shown for T_0 and T_1 progeny (T_1-X) of *OlS_1-HPS10* and *OlS_1-DUF247-1* CRISPR-Cas9-edited lines. *Ol*, *H1-3*, and *D1-3* indicate wild type, *OlS_1-HPS10*, and *OlS_1-DUF247-1* edited lines, respectively. Edited regions are highlighted in red boxes.

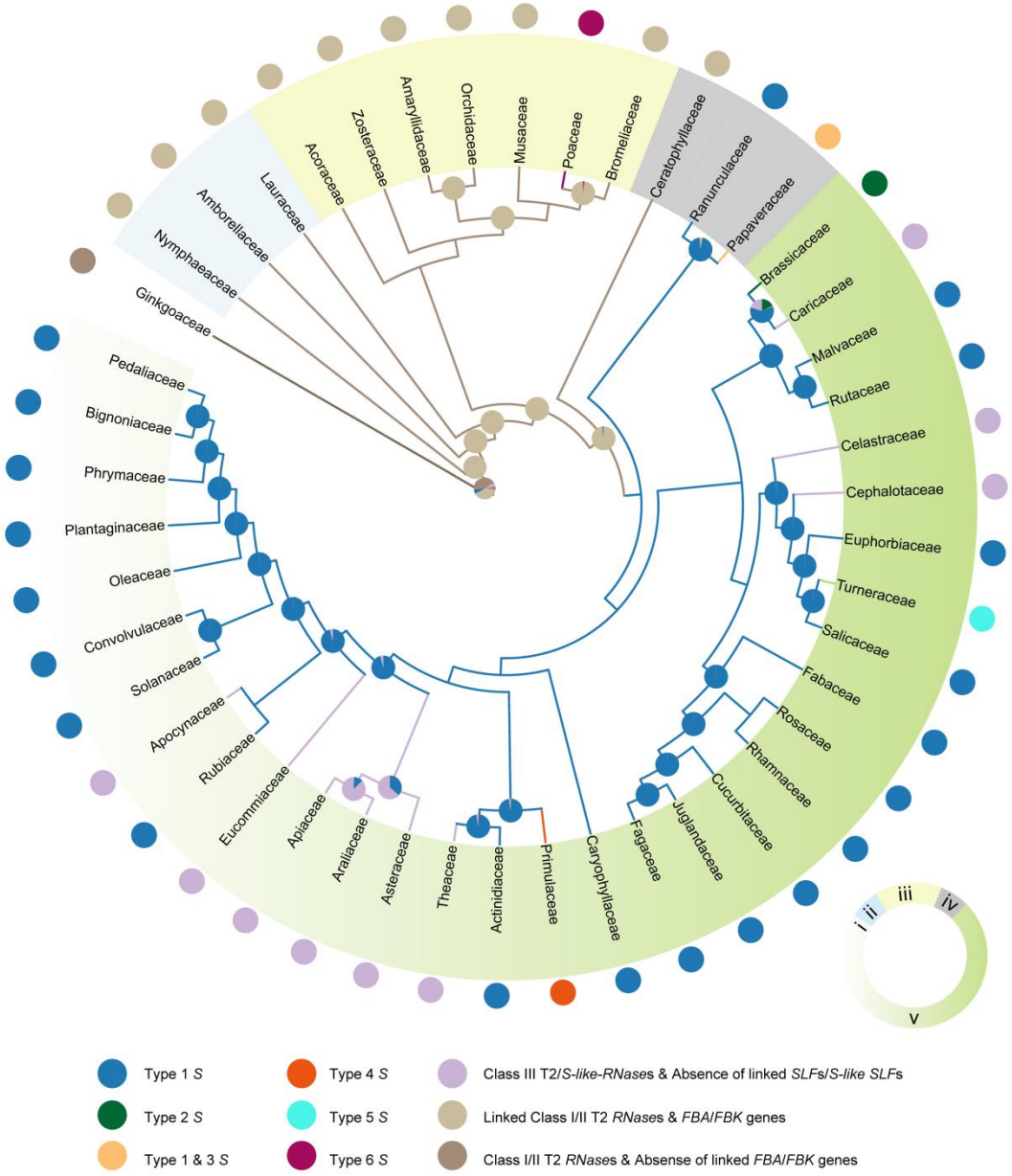


Extended Data Fig. 6 | Origin and evolution of OlS-DUF247-1/-2/-like in seed plants. ML phylogenetic tree of the DUF247 family of the seed plants with several bootstrap confidence values above 50% are indicated. OlS-DUF247-1/-like and -2/-like highlighted by cyan and light orange color ranges, respectively. I, II and III: Cluster I, II and III comprising different DUF247 families as shown under the tree.



760

761 **Extended Data Fig. 7 | Origin and evolution of Type 5 *S*-locus in seed plants.** Left,
762 a simplified phylogenetic tree of the seed plants with Turneraceae highlighted by blue
763 branch and fonts. Right, the *S*-locus supergene consisting of three genes of *BAHD*,
764 *SPH1*, and *YUC6* is shown by the blue comb line, while "None" indicates that no
765 Type 5 *S* supergene-like structures were detected.



Extended Data Fig. 8 | Origin and evolution of Type 1-6 *S* loci in seed plants. Ancestral state analyses of Type 1-6 *S* inferred by PastML based on a family-level tree of seed plants. Pie charts indicate the probabilities of the alternative ancestral states of each node with their colors corresponding to the nine different traits as illustrated by the color-filled circles. The branch colors are consistent with the trait colors of each family. i, ii, iii, iy, and v indicate gymnosperms, basal angiosperms and Lauraceae of Magnolia, monocots, basal dicots, and eudicots, respectively.

Extended Data Table 1 | Summary of genome assembly and annotation of *O. longistaminata*

Sequencing		***O. longistaminata***
PacBio HiFi	Raw data (Gb)	28.8
	Sequencing depth (×)	89
	Reads N_{50} (bp)	17359
Hi-C	Raw data (Gb)	42.4
	Sequencing depth (×)	132
Haplotype-resolved chromosomal-level assembly and annotation		
	Haplotype 1	**Haplotype 2**
Assembly size (Mb)	330.2	373.1
BUSCO completeness of assembly (%)	97.7	94.6
BUSCO completeness of annotation (%)	96.8	96.8
Longest sequence length (Mb)	38.5	39.1
GC content (%)	43	43.2
Numbers of genes	39329	41752
Percentage of gene length in genome (%)	33.1	29.7
Mean Transcript length (bp)	2849	2702
Gene density (per Mb)	119.0	111.9
Mean coding sequence length (bp)	1171	1147
Mean exon length (bp)	242	247
Mean intron length (bp)	439	429
Masked repeat sequence length (Mb)	132.3	144.9
Repeats percentage of genome size (%)	40.07	38.84
Contig N_{50} (Mb)	22.8	24.2

Supplementary Files

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

- SupplementaryFile1.txt
- SupplementaryFile2.txt
- SupplementaryTables14.xlsx
- re.pdf