International Journal of
Molecular Sciences

MDPI

Article

Bioinformatics Analysis of *MSH1* Genes of Green Plants: Multiple Parallel Length Expansions, Intron Gains and Losses, Partial Gene Duplications, and Alternative Splicing

Ming-Zhu Bai and Yan-Yan Guo *

College of Plant Protection, Henan Agricultural University, Zhengzhou 450046, China
* Correspondence: guoyy@henau.edu.cn

Abstract: *MutS homolog 1* (*MSH1*) is involved in the recombining and repairing of organelle genomes and is essential for maintaining their stability. Previous studies indicated that the length of the gene varied greatly among species and detected species-specific partial gene duplications in *Physcomitrella patens*. However, there are critical gaps in the understanding of the gene size expansion, and the extent of the partial gene duplication of *MSH1* remains unclear. Here, we screened *MSH1* genes in 85 selected species with genome sequences representing the main clades of green plants (Viridiplantae). We identified the *MSH1* gene in all lineages of green plants, except for nine incomplete species, for bioinformatics analysis. The gene is a singleton gene in most of the selected species with conserved amino acids and protein domains. Gene length varies greatly among the species, ranging from 3234 bp in *Ostreococcus tauri* to 805,861 bp in *Cycas panzhihuaensis*. The expansion of *MSH1* repeatedly occurred in multiple clades, especially in Gymnosperms, Orchidaceae, and *Chloranthus spicatus*. *MSH1* has exceptionally long introns in certain species due to the gene length expansion, and the longest intron even reaches 101,025 bp. And the gene length is positively correlated with the proportion of the transposable elements (TEs) in the introns. In addition, gene structure analysis indicated that the *MSH1* of green plants had undergone parallel intron gains and losses in all major lineages. However, the intron number of seed plants (gymnosperm and angiosperm) is relatively stable. All the selected gymnosperms contain 22 introns except for *Gnetum montanum* and *Welwitschia mirabilis*, while all the selected angiosperm species preserve 21 introns except for the ANA grade. Notably, the coding region of *MSH1* in algae presents an exceptionally high GC content (47.7% to 75.5%). Moreover, over one-third of the selected species contain species-specific partial gene duplications of *MSH1*, except for the conserved mosses-specific partial gene duplication. Additionally, we found conserved alternatively spliced *MSH1* transcripts in five species. The study of *MSH1* sheds light on the evolution of the long genes of green plants.

Keywords: *MutS homolog 1* (*MSH1*); bioinformatics analysis; gene expansion; long gene; intron gain and loss; partial gene duplication; transposable element (TE); alternative splicing

Citation: Bai, M.-Z.; Guo, Y.-Y. Bioinformatics Analysis of *MSH1* Genes of Green Plants: Multiple Parallel Length Expansions, Intron Gains and Losses, Partial Gene Duplications, and Alternative Splicing. *Int. J. Mol. Sci.* **2023**, *24*, 13620. https://doi.org/10.3390/ijms241713620

Academic Editors: Andrea Cavallini, Flavia Mascagni and Gabriele Usai

Received: 23 July 2023
Revised: 28 August 2023
Accepted: 29 August 2023
Published: 3 September 2023

1. Introduction

Plant organelle genomes (chloroplast and mitochondria) are derived from endosymbionts with cyanobacteria and α-proteobacterium-like ancestors, respectively [1]. The two genomes encode genes that are essential for photosynthesis and respiration. Furthermore, the plastome and mitogenome evolution of green plants is extremely complex. They present various variations, including size, structure, and gene content [2]. For example, the plastome size ranges from 11,348 bp in *Pilostyles aethiopica* [3] to 242,575 bp in *Pelargonium transvaalense* [4]; Pinaceae and Cupressophytes have lost one copy of the IR (Inverted Repeat) region [5]; the plastome of *Paphiopedilum* has undergone IR expansion and SSC (Small Single Copy) contraction [6]. Moreover, the mitogenome size ranges from 66 kb in *Viscum scurruloideum* [7] to 11.3 Mb in *Silene conica* [8]. Though most of the sequenced mitogenomes

had a single ring, a series of lineages found non-canonical mitogenome structures [9]. For example, all the sequenced mitogenomes of Orchidaceae showed a multichromosomal structure [10–13]. The mitogenome of *Gastrodia eleta* consisted of 19 contigs with a total length of 1340 kb [10], while the mitogenome of *Paphiopedilum micranthum* consisted of 26 contigs with a total length of 447 kb [13].

Many genes related to DNA repair and homologous recombination regulate the stability of the organellar genomes, such as *MSH1*, *POL1A*, *POL1B*, *RECA2*, *RECA3*, *SSB1*, and *SSB2* [14]. The *MSH1* gene regulates the organelle genome stability and alters the plant phenotype [15–19]. The *MSH1* gene was first cloned from the *Arabidopsis* mutant [19], and the gene contains six conserved domains, three of them (DNA binding domain, ATPase domain, and GIY-YIG domain) including recognizable features, and the C-terminus GIY-YIG domain differentiates *MSH1* from other *MutS* homologs (*MSH2*–*MSH6*) and the *MSH1* of yeast [20]. The disruption of *MSH1* increases the repeat-mediated homeologous recombination in *Arabidopsis thaliana* and *Physcomitrella patens* organelle genomes [21,22], and the gene is required to maintain the low mutation rates of the organelle genomes [23]. The mutant of *MSH1* impacts plant growth and induces phenotypic defects, such as variegation, variable growth rate, and delayed maturity [18,24–26]. *MSH1* even enhances plant phenotypic plasticity [27]. Moreover, *MSH1* accelerates the sorting of mutations in plant mitochondrial and plastid genomes [15]. All previous studies mainly focused on the function of *MSH1*.

Abdelnoor et al. [20] identified and compared the *MSH1* gene of six plant species, and the gene in these species has 22 exons and 21 introns with canonical splice sites and similar size coding regions. While the gene length ranged from 6.3 kb in *Arabidopsis* to 22 kb in common bean, the extreme length variation provided a unique opportunity to investigate the evolution of long genes. Then, Lin et al. [28] conducted a systematic phylogenetic analysis and inferred that the *MSH1* gene in eukaryotes horizontally transferred from bacteria. Furthermore, Odahara et al. [22] found the partial gene duplication (incomplete gene duplication) of *MSH1* in *P. patens*, and the two copies present a functional differentiation. However, whether the partial gene duplication is species-specific or clade-specific is unknown. Additionally, Wu et al. [23] reconstructed the phylogeny of *MSH1* with sparse sampling, and they found the disjunct distribution of *MSH1* across the tree of life. However, the species of green plants are poorly represented, and there are critical gaps in the extent of the partial gene duplication and the gene size expansion.

Gene length and expression level shape the novelties in the genome. The presence of introns enabled some genes of extraordinary size and the expansion of introns through the insertion of transposable elements (TEs) [29]. Guo et al. [30] defined genes over 20 kb as long genes in *Chloranthus spicatus*. A genome-wide analysis showed that some plant species have exceptionally long genes with a high TE content [30,31]. For instance, the average gene length of *A. thaliana* is 2070 bp (TAIR10), while the average gene length of Chinese pine reaches 25,170 bp [31], which means there is a dynamic evolution of gene length among species. However, the evolution of genes with an extreme length variation in green plants is poorly known. Furthermore, longer genes are less likely to produce duplicates and more likely to exhibit alternative splicing [29], and alternative splicing is widespread in multi-exonic genes [32].

Few studies have combined partial gene duplications, transposable elements, and alternative splicing analyses in gene evolution, and the application of these approaches in the study of gene evolution will expand our knowledge of long genes. Benefiting from the recent progress in sequencing technology, more and more high-quality genomes are available; e.g., more than 1031 genomes representing 788 plant species have been released in the last two decades [33], which provides an excellent opportunity to investigate the evolution of long genes. In this study, we intend to explore the evolution pattern of the *MSH1* gene in green plants. First, we will identify the *MSH1* gene in 85 sequenced genomes covering the main clades of green plants, and we will examine the gene length, gene structure, GC content, splice sites, motif and domain organization, and intron gain and loss;

secondly, we will survey the partial gene duplications of *MSH1*; thirdly, we will identify the TEs in the *MSH1* gene and determine the contribution of TEs to the gene length; finally, we will analyze the alternative splicing in the *MSH1* gene based on the annotations and the transcriptome data in the public databases.

2. Results

2.1. The General Features of MSH1 in the Selected Species

The complete *MSH1* gene in 75 species of green plants was identified, with only transcript sequences available in *Picea abies*, and the other nine species (*Abies alba*, *Ceratopteris richardii*, *Chara braunii*, *Penium margaritaceum*, *Pinus taeda*, *Sequoia sempervirens*, *S. noctiflora*, *Chlorokybus atmophyticus*, and *Monoraphidium neglectum*) with incomplete *MSH1* genes (Table S1). And 33 green plants were newly annotated in this study. The annotations of *A. alba*, *C. richardii*, *P. taeda*, and *S. sempervirens* failed due to a potential *MSH1* gene fragmentation across multiple scaffolds; the annotation of *S. noctiflora* failed due to the absence of exon 19; the annotations of *C. braunii* and *P. margaritaceum* failed due the extremely high GC content; and the annotations of *C. atmophyticus* and *M. neglectum* failed due to the lack complete domains. The nine incomplete sequences were excluded for further analysis. By contrast, we detected no *MSH1* orthologs in four non-green plants, including *Cyanophora paradoxa* of glaucophyte and three species (*Chondrus crispus*, *Galdieria sulphuraria*, and *Cyanidioschyzon merolae*) of Rhodophyta, and the disjunct gene distribution in the outgroup is consistent with Wu et al. [23].

We identified the *MSH1* genes of 96 species (76 green plants and 20 non-green plants) in the whole genome data. Sixty of the ninety-six were extracted from the annotations in public databases, while the other 30 were annotated in this study (Table S1). The gene length of green plants ranges from 3234 bp in *Ostreococcus tauri* to 805,861 bp in *Cycas panzhihuaensis*. The coding region of green plants ranges from 1040 amino acids in Putative Chlorophyta to 1584 amino acids in *Mesotigma viride* (Table S1). The gene length of 34 species is over 50 kb, and the gene length of three species is even over 500 kb, while the gene length of non-green species does not exceed 10 kb (Figure 1, Table S1). Expanded *MSH1* genes (over 50 kb) were distributed in multiple clades, including Orchidaceae, *Nelumbo nucifera*, *C. spicatus*, *Liriodendron chinense*, *Magnolia officinalis*, *Vitis vinifera*, *Glycine max*, *Zostera marina*, *Hemerocallis citrina*, gymnosperms, and ferns (*Adiantum capillus-veneris* and *Alsophila spinulosa*) (Table S1). The species in these lineages have large gene sizes and relatively large introns compared with the other selected species, and the gene length positively correlated with the genome size ($r = 0.60$, $p < 0.01$). The copy number of *MSH1* ranged from one to three; the gene was preserved as a singleton in most of the selected species, except for two copies in *Euryale ferox*, *G. max*, *Selaginella moellendorffii*, *Vanilla planifolia*, and mosses, and three copies in *Spirogloea muscicola* (Table S1). Apart from the relatively low similarity in *E. ferox* (86.7%) and mosses, the sequence similarity of other non-singleton species is over 90%; e.g., the two copies of *V. planifolia* are almost identical, and the similarity of two copies is 97.2% in *S. moellendorffii*.

We identified twenty conserved motifs (Motifs 1–20) in the MSH1 proteins of green plants, and motif composition varies among species. For instance, Orchidaceae, *Asparagus officinalis*, *A. setaceus*, and *Spirodela polyrhiza* contained all of the 20 conserved motifs, while other species lack Motif 14. Furthermore, *G. max* lacks Motif 13 and Motif 18, *Welwitschia mirabilis* lacks Motif 19 and Motif 3, mosses lack Motif 7, and outgroup species contain less conserved motifs (Figure S1). All MSH1 proteins of green plants have the three domains (MutS_I, MutS_V, and GIY-YIG) that were detected in previous studies (Figure S1). According to HMM searches, Motif 4, Motif 11, and Motif 17 encoded the MutS_I domain (70 to 105 aa, 69.9% identity), Motif 1, Motif 6, Motif 12, and Motif 18 encoded the MutS_V domain (153 to 209 aa, 64.6% identity), and Motif 8 and Motif 13 encoded the GIY-YIG domain (28 to 78 aa, 53.7% identity).

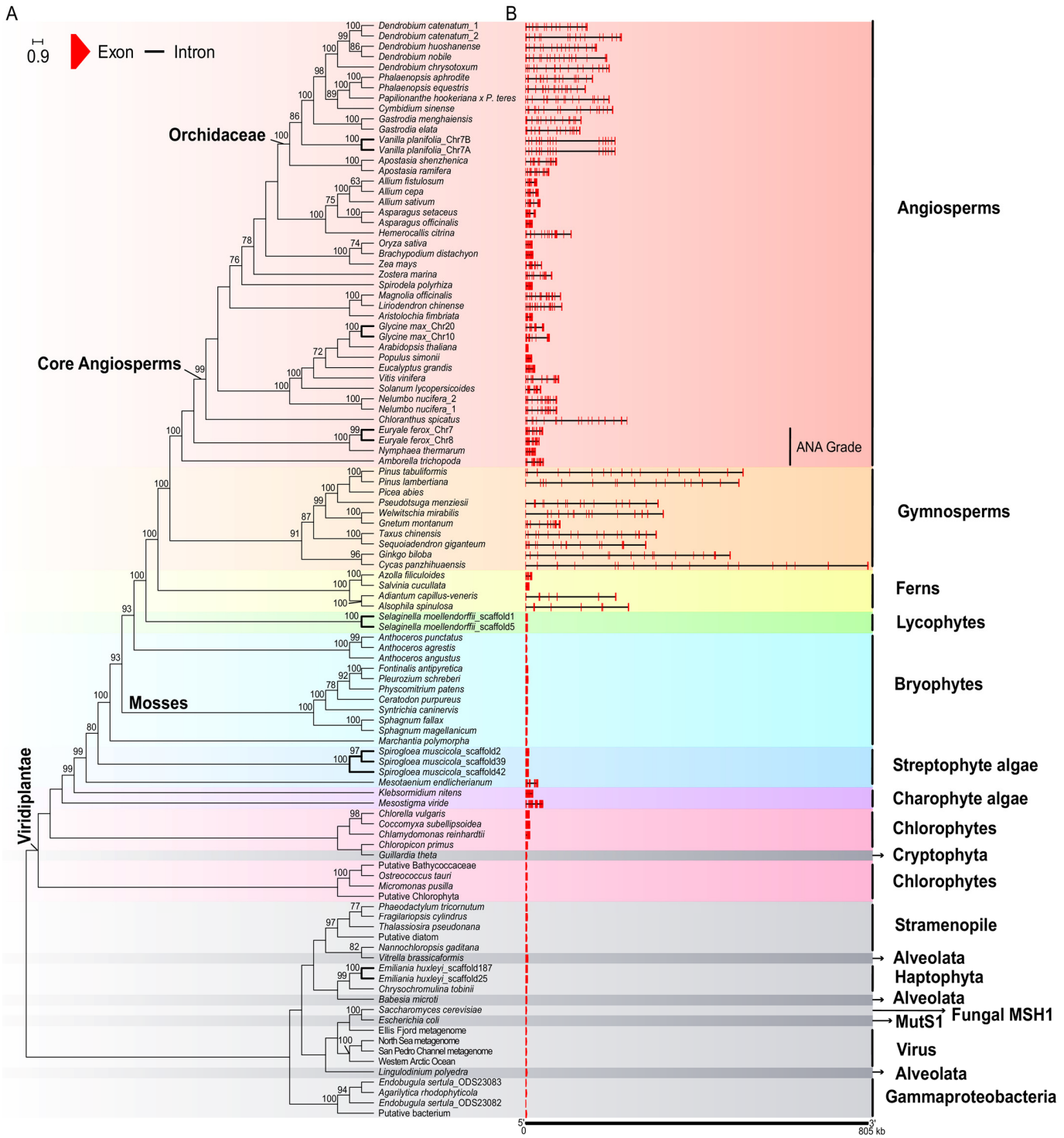

Figure 1. Species phylogeny and gene structures of *MSH1* genes in each species. (**A**) The gene tree was an ML tree constructed in RAxML v8.2.12 based on the coding sequences of *MSH1* in representative green plants, and species names in bold represent non-singleton species; (**B**) The gene structure of *MSH1* were obtained using GSDS 2.0; red boxes represent exons, and black lines represent introns.

Notably, the GC content and gene length varied greatly among species (Table S1). The GC content of the genes ranges from 32% in *Cymbidium sinense* to 71% in *Micromonas pusilla*, and the GC content of the coding regions ranges from 37.8% in *Z. marina* to 75.5% in *Chlamydomonas reinhardtii*. The GC content of the three codon positions is 46.68% to 79.33%, 38.92% to 56.95%, and 27.92% to 91.52%, respectively (Table S1). The codon usages of the three codon positions in most species follow the order of $GC_1 > GC_2 > GC_3$. The GC content at the third position (GC_3) drove the high GC content variation. Remarkably, the third position of most bryophytes (46.74% to 77.77%), streptophyte algae (70.06% to 75.59%), charophyte algae (71.45% to 77.90%), and chlorophytes (44.11% to 91.52%) with exceptionally high GC content strongly diverged among species (Figure S2, Table S1).

After removing the eight intronless and two incomplete *MSH1* genes, we calculated 1479 splice sites of 81 sequences representing 66 species of green plants. The canonical splice sites (GT-AG) account for 94.05% (1391 splice sites), while non-canonical splice sites account for 5.95% (88 splice sites), of which GG-CA is the dominant type of non-canonical splice site, with 23 splice sites accounting for 1.56% (Table S2). The non-canonical splice sites are mainly found in the basal clades of green plants (82 of 88 splice sites); e.g., 19 of the 22 splice sites in *Azolla filiculoides* are non-canonical, and 25 of the 27 splice sites in *Chlorella vulgaris* are non-canonical.

In addition, the intron number of *MSH1* varied greatly among species, ranging from 0 to 27 introns in *M. viride* and *C. vulgaris* (Table S1). The exon/intron number variation is owing to the intron gains and losses, which occurred multiple times in green plants (Figure 2, Table S1). The intron number in the seed plant is relatively stable. All the selected gymnosperm consisted of 23 exons and 22 introns, except for *Gnetum montanu* and *W. mirabilis*, which consisted of 24 exons and 23 introns. In contrast, all the angiosperms consisted of 22 exons and 21 introns, except for species belonging to the ANA grade (*Amborella trichopoda*, *Nymphaea thermarum*, and *E. ferox*), which consisted of 23 exons and 22 introns (Table S1). The crown clade of the core angiosperms (Mesangiospermae) lost intron 21, while the ancestors of *G. montanu* and *W. mirabilis* gained intron 21′ (Figure 2). We inferred that the ancestors of seed plants consist of 23 exons and 22 introns, and the intron gains and losses of *MSH1* in seed plants are all at the 3′ end of the gene. On the contrary, the basal clades of green plants underwent more frequent intron gains and losses. For example, the intron number of the bryophytes ranged from zero to eight, while the intron number of chlorophytes ranged from 0 to 27. *MSH1* is intronless in *Anthoceros agrestis*, *A. angustus*, *A. punctatus*, *Chloropicon primus*, *Marchantia polymorpha*, Putative Bathycoccaceae, Putative Chlorophyta, and *O. tauri*, and these species are distributed in different clades of the tree. In contrast, the *MSH1* in most other species is intron-rich (Table S1). Notably, most outgroup species (11 of 20) are intronless. Considering the sparse sampling and the poorly resolved species tree, the accurate inference of the intron gains and losses events in the basal group of green plants is unlikely.

2.2. Partial Gene Duplications in the MSH1 Gene

Notably, all the examined mosses have a lineage-specific partial gene duplicate, and the two copies differ in length and domain. The copy with the GIY-YIG domain is the normal one, and the other copy lacks the GIY-YIG domain derived from partial gene duplication, which was named *MSH1L* in this study (Figure 3). The phylogenetic tree of mosses indicated that all the *MSH1L* in the mosses were grouped into clade I, while all the *MSH1* in the mosses were grouped into clade II; the two copies of the mosses duplicated before the diversification of the mosses, and they correspond to the two copies named *MSH1A* and *MSH1B* in *P. patens* (Figure 3) [22]. Furthermore, we found species-specific partial gene duplication in the other 39 species, such as *Dendrobium huoshanense*, *D. catenatum*, *and H. citrina* (Table S3). Notably, there are 25 seed plants with species-specific partial gene duplications, and the length of the *MSH1* gene in these species is over 20 kb. These partial gene duplications are inserted in the coding regions, or located upstream or downstream of the gene; for example, the duplicated exon 4 to exon 16 in *V. planifolia* is

situated in intron 14 of the gene, while the duplicated exon 1 to exon 14 in *D. huoshanense* is located 341 kb upstream of the gene (Table S3, Figure 3).

Figure 2. The schematic diagram of the gene structure illustrates the intron gains and losses. The boxes represent exons, and the horizontal lines represent introns; the exon lengths are drawn to scale, and the intron lengths are not drawn to scale. Pluses indicate the number of gained introns, and minuses indicate the number of lost introns. Red numbers and red pluses indicate the position of the newly gained introns.

2.3. Transposable Elements in the MSH1 Gene

The intron length ranges from 43 bp in *S. moellendorffii* to 101,025 bp in *V. planifolia* (Table S4). There are 359 introns larger than 5 kb in total, mainly in the seed plants (Figure 4). Interestingly, for Intron 2, Intron 5, Intron 11, Intron 17, and Intron 21 of the seed plants with conserved length, the length of the five introns is shorter than 5 kb in most selected species (Table S4). To further explore the gene expansion of the *MSH1* gene, we counted the TEs of the gene with introns in 39 species representing the main groups (Figure 5). Among them, eight species did not contain transposable elements (*A. thaliana*, *Populus simonii*,

Salvinia cucullata, *S. moellendorffii*, *P. patens*, *S. muscicola*, *Guillardia theta*, and *Nannochloropsis gaditana*), the TEs ranged from 1.16% (*Vitrella brassicaformis*) to 81.11% (*A. spinulosa*) in the remaining 31 species, and the TEs ranged from 30.49% to 68.59% in Orchidaceae. The gene length positively correlated with the proportion of the TEs (r = 0.81, $p < 0.01$). Among them, 23 species had the highest proportion of retrotransposons, three had the highest proportion of DNA transposons, five had the highest proportion of unclassified transposons, and only three had helitrons. The Gypsy and Copia long terminal repeat (LTR) retrotransposon elements were the dominant components of 20 species (>85%), with a prevalence of Gypsy over the Copia superfamily in 17 species.

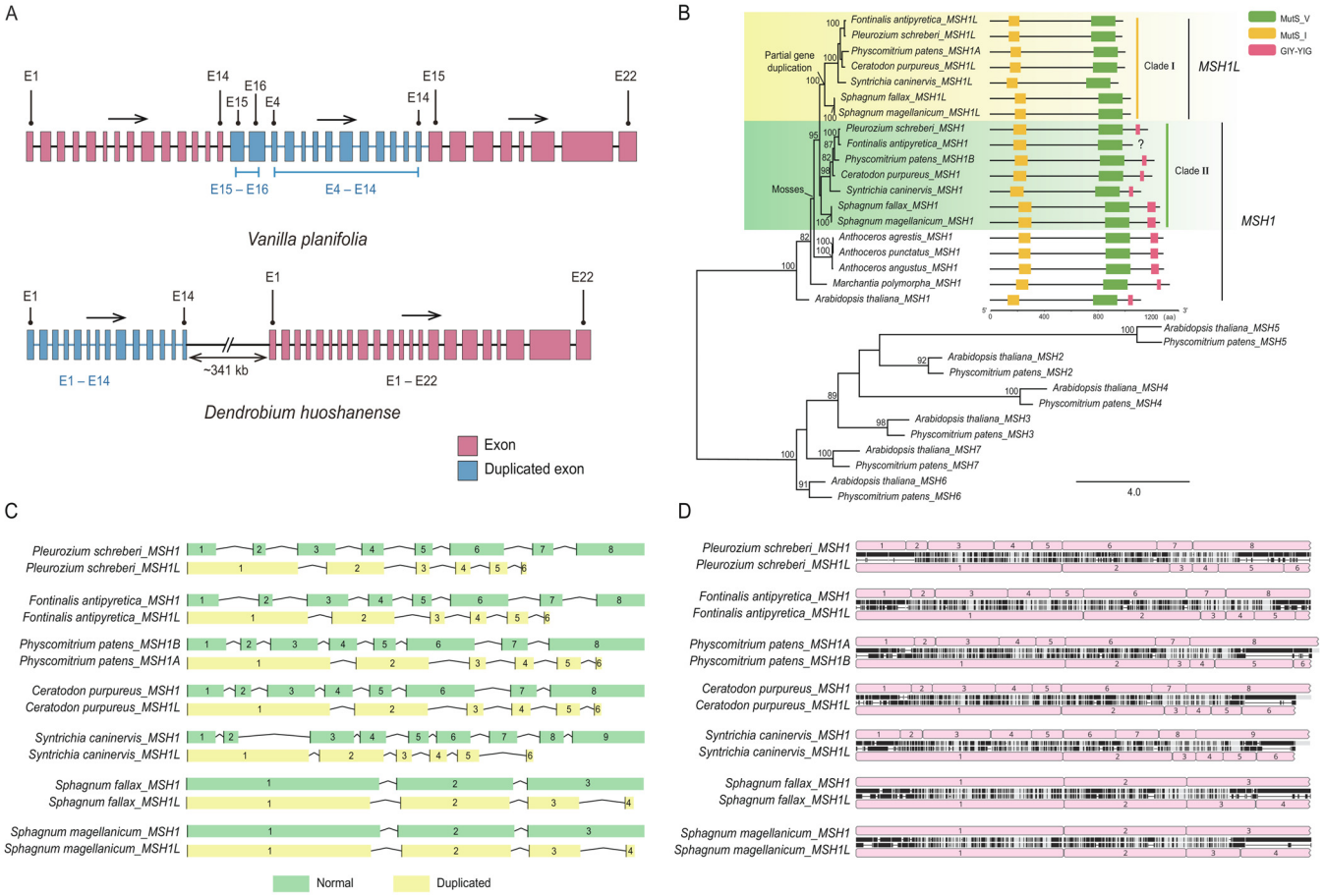

Figure 3. Cases of partial gene duplication detected in this study. (**A**) Diagram of the internal partial gene duplication detected in *Vanilla planifolia* and the external partial gene duplication detected in *Dendrobium huoshanense*, generated in RStudio v4.2.1; (**B**) phylogeny of *MSH1* and *MSH1L* in selected species and other *MSH* genes in *Arabidopsis thaliana* and *Physcomitrella patens*, constructed in RAxML v8.2.12. Clade I of mosses lacking the GIY-YIG domains are the partial gene duplication of *MSH1*, and Clade II of mosses preserve the GIY-YIG domains; (**C**) gene structure comparison of *MSH1* and *MSH1L* in mosses, plotted in RStudio v4.2.1; (**D**) sequence comparison of *MSH1* and *MSH1L* in mosses, obtained from Geneious Prime v2021.2.2. Pink boxes indicate exons; black vertical lines indicate unmatched amino acids.

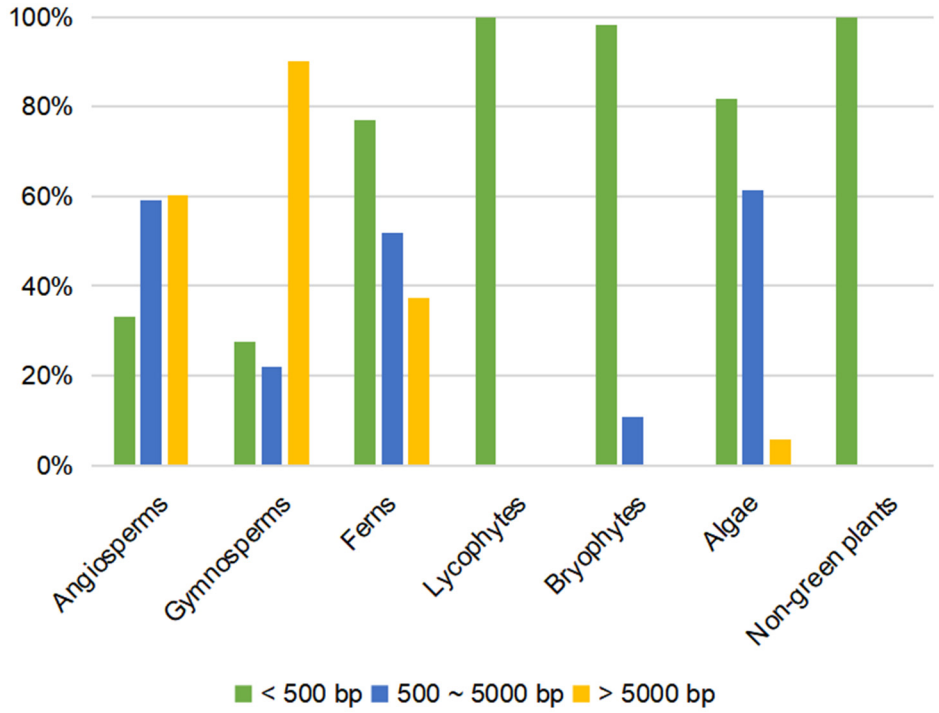

Figure 4. The proportion of different length introns in *MSH1* genes.

2.4. Alternative Splicing in the MSH1 Gene

In addition, we found five types of alternative splicing in the nine species, including two alternative acceptor sites, three alternative donor sites, five exon skippings, one mutually exclusive exon, and one other alternative type (Figure S3, Table S5). Besides the constitutive isoform, most of these alternative isoforms are species-specific. We found a shared splice variant in five species (*A. thaliana*, *Brachypodium distachyon*, *D. catenatum*, *Oryza sativa*, and *V. vinifera*). The isoform originated from exon skipping, and starts from 13 bp at the 3′ end of exon 8 and extends to exon 22, with the length ranging from 2622 bp in *A. thaliana* to 2700 bp in *V. vinifera,* and it has lost the MutS_I domain (Figure 6).

2.5. The Gene Tree of MSH1

In the gene tree of *MSH1* (Figure 1), the relationship between the main clades is consistent with previous studies [34,35]. However, the inner relationships in most groups are unresolved except for in gymnosperms, and the *G. theta* clusters of green plants with weak support. The copies in the same species cluster together, which means the duplication occurred after the speciation events.

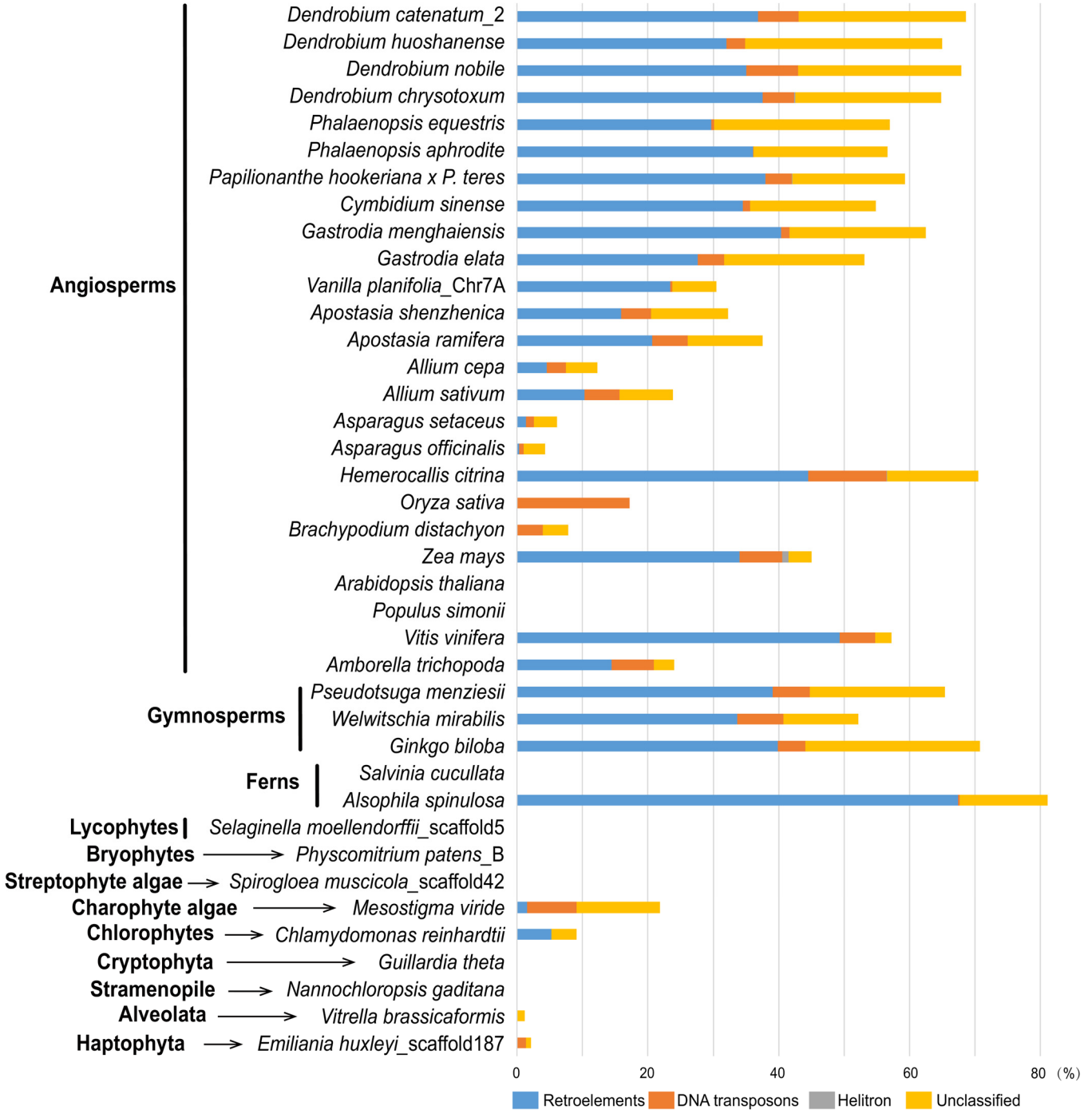

Figure 5. The proportion of the transposable elements of the *MSH1* genes in the 39 representative species.

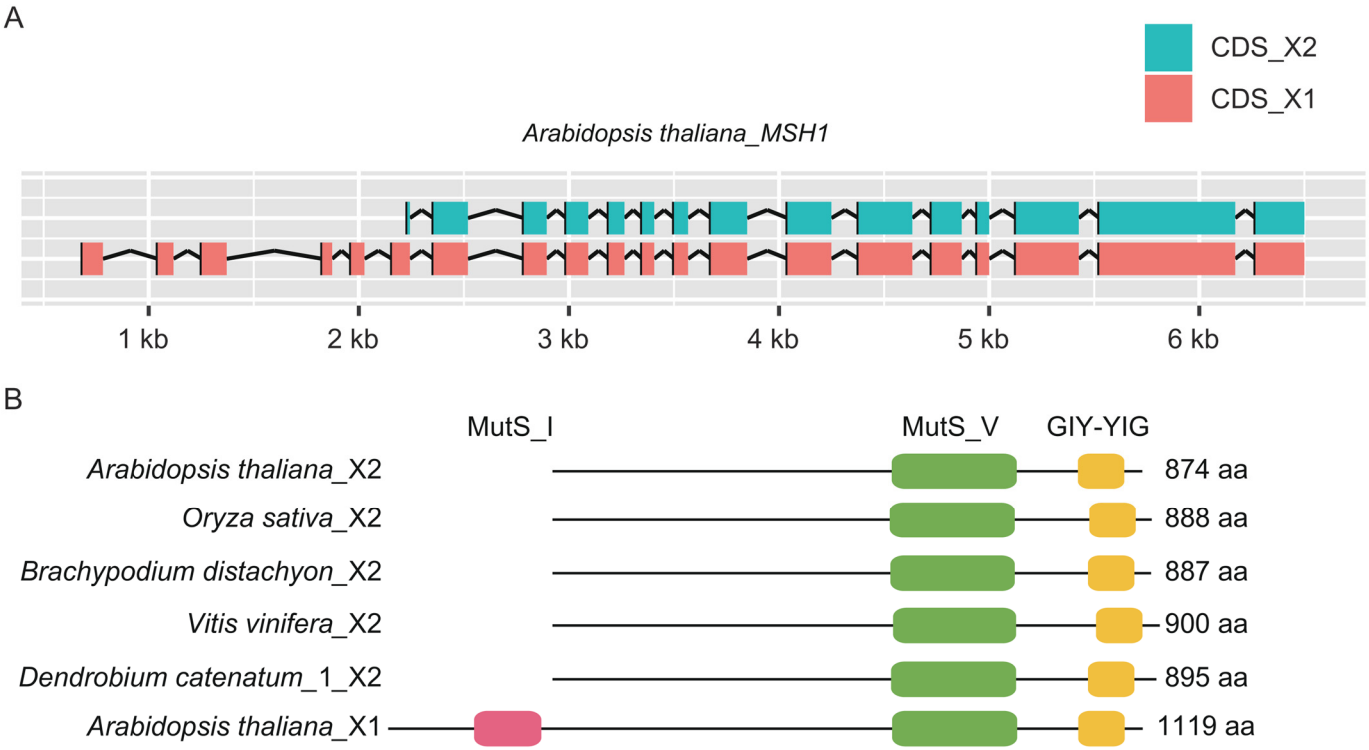

Figure 6. The shared alternative splicing of *MSH1* genes detected in five species. (**A**) Diagram of the shared alternative splicing based on the data of *Arabidopsis*, generated in RStudio v4.2.1; (**B**) diagram of the conserved domains of the shared alternative isoforms, plotted in TBtools v1.098685. X1 represents the constitutive isoform, and X2 represents the shared alternative isoform.

3. Discussion

3.1. Multiple Parallel Gene Length Expansion of MSH1 in Green Plants

Only six plant species were selected in the previous study [20]. In contrast, we sampled 85 species covering the main clades of green plants and 24 non-green plants. We identified the complete *MSH1* gene in 75 green plants, the transcript sequence in one green plant, and the incomplete *MSH1* gene in the other nine green plants. However, *MSH1* was not detected in four species of Rhodophyta and Glaucophyte, consistent with previous studies [23]. The disjunct distribution of *MSH1* suggests the complex origin of the gene in the ancestors of green plants. The *MSH1* gene is a single-copy gene in most selected species except for the five species with two to three copies. Three species (*E. ferox*, *G. max*, and *S. moellendorffii*) with two copies have undergone paleo-polyploidization events [36–38], *V. planifolia* is a phased genome [39], and the three-copy *S. muscicola* experienced a recent whole-genome triplication event [40], suggesting a strong selection for the singleton of the gene in green plants.

The coding region of *MSH1* is relatively conserved in all the selected species (3120 bp to 4752 bp), especially in the seed plants (3330 bp to 3795 bp). However, *MSH1* varied greatly in gene length (3234 bp to 805,861 bp) and intron number (0 to 27) (Table S1). The *MSH1* gene greatly expanded in multiple lineages, especially in the Orchidaceae and Gymnosperms, with the gene length of all the selected species in the two clades over 50 kb. The *MSH1* of Orchidaceae experiences different extents of expansion, with the gene length ranging from 55,035 bp in *Apostasia ramifera* to 225,727 bp in *D. catenatum* (Table S1). All the selected species of gymnosperms have ultra-long *MSH1* genes, ranging from 82,282 bp in *G. montanum* to 805,861 bp in *C. panzhihuaensis*. The most interesting aspect is that the two lineages (Orchidaceae and Gymnosperms) are renowned for their large genome sizes. The genome size of Orchidaceae ranged from 0.33 pg to 55.4 pg [41], and the modal genome size value of the 57 gymnosperm species is 30.0 pg [42]; e.g., the genome size of Chinese pine reaches 25.4 Gb [31]. Moreover, the average intron size of the two clades is much

longer than other clades; e.g., the average intron size of *G. elata* is 3252 bp [43], while the average intron size of Chinese pine is 10,034 bp [31]. In contrast, the average intron sizes of *A. thaliana* and *O. sativa* are 161 bp and 469 bp, respectively.

Considering the low length variation of the coding regions, the gene length expansion is mainly induced by the intron size expansion. Most of the introns of gymnosperms and Orchidaceae are over 5000 bp, with 97 introns (48.5%) and 85 introns (26.98%) that are longer than 10 kb, respectively; e.g., the longest intron is the 101,025 bp-long Intron 14 of *V. planifolia* (Table S1). In comparison, all the introns in *Arabidopsis* are shorter than 500 bp. The first intron is the longest in most genes [44]. However, the first intron of *MSH1* is not the longest in most selected species (Table S4). Owing to the extreme length expansion, the gene annotation of *A. alba*, *P. abies*, and *S. sempervirens* failed due to the potential *MSH1* gene being fragmented across multiple scaffolds. Besides the difficulties in assembling, the long genes with multiple introns pose great challenges to gene annotation and identification [31]. Notably, the two samples of *N. nucifera* have similar lengths (73,992 bp and 73,601 bp), while the two samples of *D. catenatum* have distinct sizes (145,440 bp and 225,727 bp) (Table S1). The two samples of *D. catenatum* were sequenced based on different sequencing technologies. The earlier one was sequenced using the Illumina platform [45], while the latter was sequenced based on PacBio long-reads, Illumina short-reads, and Hi-C data [46]. The long repeat sequences in Intron 8 induced the length variation in the two samples. The sequencing platform-induced sequence length variation indicated that the gene length of the short-read sequenced samples might be underestimated, especially the genes with long repeat regions. Moreover, previous studies showed a negative correlation between GC content and intron length, which means that short introns tend to have a higher GC content, while long introns have a lower GC content [47,48]. However, the GC content has no correlation with the intron length in *MSH1*. The GC content of the short introns (<5 kb) ranged from 17.40% to 56.90%, while the GC content of the long introns (≥5 kb) ranged from 27.90% to 56.90%.

Moreover, *MSH1* varied from being intronless in liverworts and hornworts, to having 27 introns in *M. viride* and *C. vulgaris* (Table S1). Notably, intronless, intron-poor, and intron-rich members coappear in the *MSH1* gene, while previous studies found intronless, intron-poor, and intron-rich genes in the same gene family [49]. The intron number variation revealed that the gene had undergone recurrent intron gains and losses (Figure 2). The ancestors of seed plants preserve 23 exons and 22 introns, while the core angiosperms lose one intron, which means that the other 21 introns have existed for over 400 Mya. However, the intron gain and loss events in the basal clades of green plants are more complex, and the coding region of seed-free species exhibited a GC-biased nucleotide composition (Table S1). Furthermore, Gozashti et al. [50] found that intron gains correlated with TEs named Introners; aquatic organisms were 6.5 times more likely to contain Introners than terrestrial organisms, and Introners exist towards insertion into the GC-rich regions.

3.2. Partial Gene Duplications in the MSH1 Gene

Notably, we found a lineage-specific partial duplication in the mosses (Figure 2). Based on the sequence comparison and phylogenetic analysis, we inferred that *MSH1L* originated from an ancient partial gene duplication of *MSH1* specific to mosses, and the two copies duplicated before the diversification of mosses, which means the partial duplication copy lasted more than 400 million years in this group. *MSH1L* and *MSH1* correspond to the two *MSH1* genes found in *P. patens* (*MSH1A* and *MSH1B*) (Figure 2) [22]. Odahara et al. [22] found that the function of *MSH1A* on the suppression of organelle recombination is minor, and *MSH1A* might be redundant with *MSH1B*. However, *MSH1L* (also *MSH1A* in *P. patens*) might gain other unknown new functions, which need further verification. Besides the mosses-specific partial gene duplicates, the other partial gene duplicates are all species-specific, and species-specific partial gene duplication tends to appear in long genes (Table S3), which is consistent with previous studies [29]. Furthermore, partial gene duplication followed by neo-functionalization might contribute to the evolutionary

innovation reported in other species [51,52]. For example, the species-specific *EXOV* is a partial gene duplicate of *EXOVL* in *A. thaliana*, and the *EXOV* acquired novel direct and indirect interactions with other genes and induced significant morphological effects [52].

3.3. The MSH1 Gene Length Is Positively Correlated with the Proportion of Transposable Elements

One of the surprising results of this study was the great length variation of *MSH1*. Apart from the intronless and incomplete genes, the intron content accounts for 4.97% to 99.55% of the gene (Table S4). The extraordinary length of *MSH1* is enabled by the TEs, especially the LTR retrotransposons, and the gene length is positively correlated with the proportion of TEs ($R = 0.81$, $p < 0.01$). The TEs were also found in the long genes of Chinese pine [31] and *C. spicatus* [30]. TEs insert throughout the genome and contribute to phenotype variation and evolution [53]; e.g., TE insertions at the *FLC* of *Capsella rubella* affect the natural variation in flowering time [54]. On the other hand, TE insertions might regulate the neighbor gene expression; e.g., a TE insertion named *redTE* upstream of the *MdMYB1* is linked to the red skin color of apples [55]. Genes with long introns tend to have a higher expression [30,31]. Furthermore, the suppression of *MSH1* changes the mitogenome conformation [56]. The widespread insertion of TEs in the *MSH1* gene hints at a correlation between TE insertion and organelle genome stability.

3.4. Alternative Splicing Detected in the MSH1 Gene

Long introns are associated with high rates of alternative splicing [57]. Notably, five species share a 15-exon alternative isoform in *MSH1*, and the isoform originated from changes in the alternative first exon usage (splicing out the first seven exons and part of exon 8) lacks the MutS_I domain and is relatively shorter than the constitutive splicing (Figure 6). Notably, we found an evolutionarily conserved upstream open reading frames (uORFs) range of 19 aa to 26 aa in the shared alternative isoform. Considering that uORFs potentially regulate stress-related alternative splicing events [32], we inferred that the shared alternative isoform might be stress-related and essential for plant development. Hazra and Mahadani [58] found exon skipping events in *D. officinale* leaves under cold acclimation. Additionally, the alternative splicing analysis indicated that the conventional 5′ splicing sites were not conserved, and generated novel proteins in response to abiotic stress [59,60]. Alternative splicing increases protein versatility and plays a vital role in adaptive evolution, phenotypic novelty, protein diversity, and organism complexity [61,62].

4. Materials and Methods

4.1. Data Sources

To explore the evolution of the *MSH1* in green plants, we used the genome sequences of 85 species representing the major lineages of green plants, including thirty-nine angiosperms, thirteen gymnosperms, five ferns, one lycophyte, eleven bryophytes, three streptophyte algae, four charophyte algae, and nine chlorophytes (Table S6). We also chose 24 species from the other nine clades as outgroups, following the sampling of Wu et al. [23] (Table S6).

4.2. Identification of MSH1 Genes

We downloaded the genome sequences, annotation files, protein sequences, and transcriptome sequences of these species from GenBank or other databases (Table S6). We used BlastP v2.9.0 ($E \leq 1 \times 10^{-6}$) (https://blast.ncbi.nlm.nih.gov/Blast.cgi, accessed on 1 March 2021) to identify the homology of the *MSH1* gene with the protein sequence from *A. thaliana* as a query. Then, the retrieved sequences were used as queries to blast against the species lacking the annotation of the *MSH1* gene. The sequence without annotation was annotated in Geneious Prime v2021.2.2 (Biomatters, Inc., Auckland, NewZealand) and refined manually. Furthermore, we performed a reference assembly in Trinity v2.10.7 [63] to verify the gene annotations. The partial gene duplications were identified via a repeated blast with each exon. Then, we inferred the intron losses and gains of *MSH1* based on

parsimony. Considering over half of the species in the outgroup are intronless, we suppose that being intronless is an ancestral form of the gene.

4.3. Motif and Gene Structure Analysis

Then, we use HMMER [64] and NCBI-CDD [65] to identify the conserved protein domains. We identified conserved motifs in *MSH1* using MEME version 5.5.4 [66] with the following settings: maximum number of motifs set at 20, and optimum motif width set to $\geq$6 and $\leq$100 residues. We visualized the results using TBtools v1.098685 [67]. Finally, we draw the gene structure of *MSH1* using GSDS v2.0 [68]. In addition, the GC content of the three positions was calculated in EMBOSS v6.5.7.0 [69].

4.4. Transposable Elements in the MSH1 Genes

We selected 39 species representing all the lineages to characterize the reason for gene expansion. We constructed the species-specific repeat library using RepeatModeler v2.0.2 [70]. Then, we used RepeatMasker v4.1.2 [71] to annotate the TEs in *MSH1* and analyzed the contributions of the four major classes of TEs.

4.5. Alternative Splicing in the MSH1 Genes

We selected nine species using the GenBank annotation files and multiple transcripts to analyze alternative splicing. Then, the putative alternative splicing events were identified using AStalavista v4.0 [72] through the GTF file obtained above.

4.6. Phylogenetic Analysis

We excluded the copy lacking the GIY-YIG domain in the mosses and the nine incomplete sequences from further analysis, and we included the transcript of *P. abies* for the tree construction. Finally, we preserved 106 sequences representing 96 species (76 green plants and 20 non-green plants) for phylogenetic analysis. The *MSH1* protein-coding sequences alignment was performed using MAFFT v7.407 [73] with the default parameters, and was refined manually. The unalignable regions were removed using Gblock v0.91b [74]. PartitionFinder v2.1.1 [75] was used to determine the optimal partitioning scheme and evolutionary model under the Akaike Information Criterion (AIC). We constructed a maximum likelihood (ML) phylogeny in RAxML v8.2.12 [76] under the GTR+G model with 1000 bootstrap interactions using the inferred alignment of the *MSH1* gene. The species phylogeny obtained was used to infer the intron gains and losses in *MSH1*. Then, we constructed the phylogenetic tree of the mosses using the above methods to clarify the origin of the duplicate copies in the group, and we selected the *MSH1* gene of 12 species and the other *MSH* genes of *A. thaliana* and *P. patens* for the tree construction. The generated trees were visualized using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/, accessed on 11 January 2021).

4.7. Statistical Analysis and Visualization

We tested for the correlation of the following two pairs of variables: gene length and genome size, and gene length and the proportion of TEs. The correlation tests were performed using the cor.test function in RStudio v4.2.1 [77], using a Pearson test. The diagrams of the partial gene duplication, alternative splicing, and gene structure comparison of *MSH1* and *MSH1L* in the mosses were plotted in RStudio v4.2.1 [77] with the following packages: GenomicRanges v1.49.0 [78], ggbio v1.46.0 [79], ggplot2 v3.4.1 [80], ggtranscript v0.99.0 [81], and magrittr v2.0.3 [82]. The Figures were arranged and polished in Adobe Illustrator 2020.

5. Conclusions

This study provides an overall picture of the evolutionary history of *MSH1* in green plants. We expanded the gene analysis of *MSH1* to 109 sequenced genomes. *MSH1* is universally available in green plants. The gene experienced multiple parallel expansions,

intron gains and losses, partial gene duplications, and alternative splicing. Gene length is positively correlated with TEs. Intron gain and loss are mainly reported at the genome scale with distantly related species (e.g., [83]). This study provides a typical example of rampant intron gain and loss in a particular gene with dense sampling, and the intron gains and losses are more complex than expected. The species-specific partial gene duplication in *MSH1* is widespread. However, the accurate annotation is incomplete or lacking, and its function is unknown. Moreover, the mosses-specific partial gene duplication and the alternative splicing shared by five species need further functional verification. In general, partial gene duplication, alternative splicing, and TEs in long introns might lead to neofunctionalization in *MSH1* and boost its adaptation. The expansions of *MSH1* might be potentially correlated to the aberrant mitogenomes and plastomes detected. However, there is no direct link between the *MSH1* gene and the pattern of the organelle genomes. This study suggested that we might underestimate long genes in the plant genome due to the assembling and annotation, and these genes provide unique opportunities to study gene evolution.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/ijms241713620/s1.

Author Contributions: Y.-Y.G. conceived and designed the study. M.-Z.B. and Y.-Y.G. analyzed the data. Y.-Y.G. and M.-Z.B. wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number U1804117.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All sequences in this study are openly available from the public database. Datasets for the phylogenetic tree construction are available from the corresponding author.

Acknowledgments: The authors thank the Editor and the anonymous reviewers for their insightful comments and suggestions on the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dyall, S.D.; Brown, M.T.; Johnson, P.J. Ancient invasions: From endosymbionts to organelles. *Science* **2004**, *304*, 253–257. [CrossRef] [PubMed]
2. Smith, D.R.; Keeling, P.J. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 10177–10184. [CrossRef] [PubMed]
3. Bellot, S.; Renner, S.S. The plastomes of two species in the endoparasite genus *Pilostyles* (Apodanthaceae) each retain just five or six possibly functional genes. *Genome Biol. Evol.* **2016**, *8*, 189–201. [CrossRef]
4. Weng, M.L.; Ruhlman, T.A.; Jansen, R.K. Expansion of inverted repeat does not decrease substitution rates in *Pelargonium* plastid genomes. *New Phytol.* **2017**, *214*, 842–851. [CrossRef] [PubMed]
5. Wu, C.S.; Wang, Y.N.; Hsu, C.Y.; Lin, C.P.; Chaw, S.M. Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and Cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. *Genome Biol. Evol.* **2011**, *3*, 1284–1295. [CrossRef]
6. Guo, Y.-Y.; Yang, J.-X.; Bai, M.-Z.; Zhang, G.-Q.; Liu, Z.-J. The chloroplast genome evolution of Venus slipper (*Paphiopedilum*): IR expansion, SSC contraction, and highly rearranged SSC regions. *BMC Plant Biol.* **2021**, *21*, 248. [CrossRef]
7. Skippington, E.; Barkman, T.J.; Rice, D.W.; Palmer, J.D. Miniaturized mitogenome of the parasitic plant *Viscum scurruloideum* is extremely divergent and dynamic and has lost all *nad* genes. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E3515–E3524. [CrossRef]
8. Sloan, D.B.; Alverson, A.J.; Chuckalovcak, J.P.; Wu, M.; McCauley, D.E.; Palmer, J.D.; Taylor, D.R. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biol.* **2012**, *10*, e1001241. [CrossRef]
9. Wu, Z.-Q.; Liao, X.-Z.; Zhang, X.-N.; Tembrock, L.R.; Broz, A. Genomic architectural variation of plant mitochondria—A review of multichromosomal structuring. *J. Syst. Evol.* **2022**, *60*, 160–168. [CrossRef]
10. Yuan, Y.; Jin, X.; Liu, J.; Zhao, X.; Zhou, J.; Wang, X.; Wang, D.; Lai, C.; Xu, W.; Huang, J. The *Gastrodia elata* genome provides insights into plant adaptation to heterotrophy. *Nat. Commun.* **2018**, *9*, 1615. [CrossRef]

11. Li, X.; Zhe, M.; Huang, Y.; Fan, W.; Yang, J.; Zhu, A. The evolution of mitochondrial genomes between two *Cymbidium* sister species: Dozens of circular chromosomes and the maintenance and deterioration of genome synteny. *Genes* **2023**, *14*, 864. [CrossRef] [PubMed]
12. Wang, M.-T.; Hou, Z.-Y.; Li, C.; Yang, J.-P.; Niu, Z.-T.; Xue, Q.-Y.; Liu, W.; Ding, X.-Y. Rapid structural evolution of *Dendrobium* mitogenomes and mito-nuclear phylogeny discordances in *Dendrobium* (Orchidaceae). *J. Syst. Evol.* 2023, *early view*.
13. Yang, J.-X.; Dierckxsens, N.; Bai, M.-Z.; Guo, Y.-Y. Multichromosomal mitochondrial genome of *Paphiopedilum micranthum*: Compact and fragmented genome, and rampant intracellular gene transfer. *Int. J. Mol. Sci.* **2023**, *24*, 3976. [CrossRef] [PubMed]
14. Chevigny, N.; Schatzdaas, D.; Lotfi, F.; Gualberto, J.M. DNA repair and the stability of the plant mitochondrial genome. *Int. J. Mol. Sci.* **2020**, *21*, 328. [CrossRef]
15. Broz, A.K.; Keene, A.; Fernandes Gyorfy, M.; Hodous, M.; Johnston, I.G.; Sloan, D.B. Sorting of mitochondrial and plastid heteroplasmy in *Arabidopsis* is extremely rapid and depends on MSH1 activity. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2206973119. [CrossRef]
16. Zou, Y.; Zhu, W.; Sloan, D.B.; Wu, Z. Long-read sequencing characterizes mitochondrial and plastid genome variants in Arabidopsis *msh1* mutants. *Plant J.* **2022**, *112*, 738–755. [CrossRef] [PubMed]
17. Lencina, F.; Landau, A.; Prina, A.R. The barley chloroplast mutator (*cpm*) mutant: All roads lead to the *Msh*1 gene. *Int. J. Mol. Sci.* **2022**, *23*, 1814. [CrossRef]
18. Xu, Y.-Z.; Arrieta-Montiel, M.P.; Virdi, K.S.; Paula, W.B.M.D.; Mackenzie, S.A. MutS HOMOLOG1 Is a nucleoid protein that alters mitochondrial and plastid properties and plant response to high light. *Plant Cell* **2011**, *23*, 3428–3441. [CrossRef]
19. Abdelnoor, R.V.; Yule, R.; Elo, A.; Christensen, A.C.; Meyer-Gauen, G.; Mackenzie, S.A. Substoichiometric shifting in the plant mitochondrial genome is influenced by a gene homologous to MutS. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 5968–5973. [CrossRef]
20. Abdelnoor, R.V.; Christensen, A.C.; Mohammed, S.; Munoz-Castillo, B.; Moriyama, H.; Mackenzie, S.A. Mitochondrial genome dynamics in plants and animals: Convergent gene fusions of a *MutS* homologue. *J. Mol. Evol.* **2006**, *63*, 165–173. [CrossRef]
21. Shedge, V.; Arrietamontiel, M.P.; Christensen, A.C.; Mackenzie, S.A. Plant mitochondrial recombination surveillance requires unusual *RecA* and *MutS* homologs. *Plant Cell* **2007**, *19*, 1251–1264. [CrossRef]
22. Odahara, M.; Kishita, Y.; Sekine, Y. *MSH1* maintains organelle genome stability and genetically interacts with *RECA* and *RECG* in the moss *Physcomitrella patens*. *Plant J.* **2017**, *91*, 455–465. [CrossRef]
23. Wu, Z.; Waneka, G.; Broz, A.K.; King, C.R.; Sloan, D.B. *MSH1* is required for maintenance of the low mutation rates in plant mitochondrial and plastid genomes. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 16448–16455. [CrossRef] [PubMed]
24. Virdi, K.S.; Wamboldt, Y.; Kundariya, H.; Laurie, J.D.; Keren, I.; Kumar, K.S.; Block, A.; Basset, G.; Luebker, S.; Elowsky, C. MSH1 is a plant organellar DNA binding and thylakoid protein under precise spatial regulation to alter development. *Mol. Plant* **2016**, *9*, 245–260. [CrossRef] [PubMed]
25. Xu, Y.-Z.; Santamaria, R.d.l.R.; Virdi, K.S.; Arrieta-Montiel, M.P.; Razvi, F.; Li, S.; Ren, G.; Yu, B.; Alexander, D.; Guo, L. The chloroplast triggers developmental reprogramming when MUTS HOMOLOG1 is suppressed in plants. *Plant Physiol.* **2012**, *159*, 710–720. [CrossRef]
26. Yang, X.; Kundariya, H.; Xu, Y.-Z.; Sandhu, A.; Yu, J.; Hutton, S.F.; Zhang, M.; Mackenzie, S.A. MutS HOMOLOG1-derived epigenetic breeding potential in tomato. *Plant Physiol.* **2015**, *168*, 222–232. [CrossRef]
27. Mackenzie, S.A.; Kundariya, H. Organellar protein multi-functionality and phenotypic plasticity in plants. *Philos. Trans. R. Soc. B* **2019**, *375*, 20190182. [CrossRef]
28. Lin, Z.; Nei, M.; Ma, H. The origins and early evolution of DNA mismatch repair genes—Multiple horizontal gene transfers and co-evolution. *Nucleic Acids Res.* **2007**, *35*, 7591–7603. [CrossRef] [PubMed]
29. Grishkevich, V.; Yanai, I. Gene length and expression level shape genomic novelties. *Genome Res.* **2014**, *24*, 1497–1503. [CrossRef]
30. Guo, X.; Fang, D.; Sahu, S.K.; Yang, S.; Guang, X.; Folk, R.; Smith, S.A.; Chanderbali, A.S.; Chen, S.; Liu, M.; et al. *Chloranthus* genome provides insights into the early diversification of angiosperms. *Nat. Commun.* **2021**, *12*, 6930. [CrossRef]
31. Niu, S.; Li, J.; Bo, W.; Yang, W.; Zuccolo, A.; Giacomello, S.; Chen, X.; Han, F.; Yang, J.; Song, Y.; et al. The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell* **2022**, *185*, 204–217.e14. [CrossRef]
32. Martín, G.; Márquez, Y.; Mantica, F.; Duque, P.; Irimia, M. Alternative splicing landscapes in *Arabidopsis thaliana* across tissues and stress conditions highlight major functional differences with animals. *Genome Biol.* **2021**, *22*, 35. [CrossRef]
33. Sun, Y.; Shang, L.; Zhu, Q.-H.; Fan, L.; Guo, L. Twenty years of plant genome sequencing: Achievements and challenges. *Trends Plant Sci.* **2022**, *27*, 391–401. [CrossRef]
34. Szövényi, P.; Gunadi, A.; Li, F.-W. Charting the genomic landscape of seed-free plants. *Nat. Plants* **2021**, *7*, 554–565. [CrossRef]
35. One Thousand Plant Transcriptomes Initiative, One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **2019**, *574*, 679–685. [CrossRef]
36. Wu, P.; Zhang, L.; Zhang, K.; Yin, Y.; Liu, A.; Zhu, Y.; Fu, Y.; Sun, F.; Zhao, S.; Feng, K.; et al. The adaptive evolution of *Euryale ferox* to the aquatic environment through paleo-hexaploidization. *Plant J.* **2022**, *110*, 627–645. [CrossRef]
37. Wang, J.; Yu, J.; Sun, P.; Li, C.; Song, X.; Lei, T.; Li, Y.; Yuan, J.; Sun, S.; Ding, H.; et al. Paleo-polyploidization in Lycophytes. *Genom. Proteom. Bioinform.* **2020**, *18*, 333–340. [CrossRef] [PubMed]
38. Wang, J.; Sun, P.; Li, Y.; Liu, Y.; Yu, J.; Ma, X.; Sun, S.; Yang, N.; Xia, R.; Lei, T.; et al. Hierarchically aligning 10 legume genomes establishes a family-level genomics platform. *Plant Physiol.* **2017**, *174*, 284–300. [CrossRef] [PubMed]

39. Hasing, T.; Tang, H.; Brym, M.; Khazi, F.; Huang, T.; Chambers, A.H. A phased *Vanilla planifolia* genome enables genetic improvement of flavour and production. *Nat. Food* **2020**, *1*, 811–819. [CrossRef]
40. Cheng, S.; Xian, W.; Fu, Y.; Marin, B.; Keller, J.; Wu, T.; Sun, W.; Li, X.; Xu, Y.; Zhang, Y.; et al. Genomes of subaerial Zygnematophyceae provide insights into land plant evolution. *Cell* **2019**, *179*, 1057–1067.e14. [CrossRef] [PubMed]
41. Leitch, I.J.; Kahandawala, I.; Suda, J.; Hanson, L.; Ingrouille, M.J.; Chase, M.W.; Fay, M.F. Genome size diversity in orchids: Consequences and evolution. *Ann. Bot.* **2009**, *104*, 469–481. [CrossRef] [PubMed]
42. Ohri, D.; Khoshoo, T. Genome size in gymnosperms. *Plant Syst. Evol.* **1986**, *153*, 119–132. [CrossRef]
43. Xu, Y.; Lei, Y.; Su, Z.; Zhao, M.; Zhang, J.; Shen, G.; Wang, L.; Li, J.; Qi, J.; Wu, J. A chromosome-scale *Gastrodia elata* genome and large-scale comparative genomic analysis indicate convergent evolution by gene loss in mycoheterotrophic and parasitic plants. *Plant J.* **2021**, *108*, 1609–1623. [CrossRef] [PubMed]
44. Bradnam, K.R.; Korf, I. Longer first introns are a general property of eukaryotic gene structure. *PLoS ONE* **2008**, *3*, e3093. [CrossRef] [PubMed]
45. Zhang, G.-Q.; Xu, Q.; Bian, C.; Tsai, W.-C.; Yeh, C.-M.; Liu, K.-W.; Yoshida, K.; Zhang, L.-S.; Chang, S.-B.; Chen, F.; et al. The *Dendrobium catenatum* Lindl. genome sequence provides insights into polysaccharide synthase, floral development and adaptive evolution. *Sci. Rep.* **2016**, *6*, srep19029. [CrossRef]
46. Niu, Z.; Zhu, F.; Fan, Y.; Li, C.; Zhang, B.; Zhu, S.; Hou, Z.; Wang, M.; Yang, J.; Xue, Q.; et al. The chromosome-level reference genome assembly for *Dendrobium officinale* and its utility of functional genomics research and molecular breeding study. *Acta Pharm. Sin. B* **2021**, *11*, 2080–2092. [CrossRef] [PubMed]
47. Gazave, E.; Marqués-Bonet, T.; Fernando, O.; Charlesworth, B.; Navarro, A. Patterns and rates of intron divergence between humans and chimpanzees. *Genome Biol.* **2007**, *8*, R21. [CrossRef]
48. Duret, L.; Mouchiroud, D.; Gautier, C. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J. Mol. Evol.* **1995**, *40*, 308–317. [CrossRef]
49. Liu, H.; Lyu, H.-M.; Zhu, K.; Van de Peer, Y.; Cheng, Z.-M. The emergence and evolution of intron-poor and intronless genes in intron-rich plant gene families. *Plant J.* **2021**, *105*, 1072–1082. [CrossRef]
50. Gozashti, L.; Roy, S.W.; Thornlow, B.; Kramer, A.; Ares, M.; Corbett-Detig, R. Transposable elements drive intron gain in diverse eukaryotes. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2209766119. [CrossRef]
51. Rajaraman, J.; Douchkov, D.; Lück, S.; Hensel, G.; Nowara, D.; Pogoda, M.; Rutten, T.; Meitzel, T.; Brassac, J.; Höfle, C. Evolutionarily conserved partial gene duplication in the Triticeae tribe of grasses confers pathogen resistance. *Genome Biol.* **2018**, *19*, 116. [CrossRef]
52. Huang, Y.; Chen, J.; Dong, C.; Sosa, D.; Xia, S.; Ouyang, Y.; Fan, C.; Li, D.; Mortola, E.; Long, M. Species-specific partial gene duplication in *Arabidopsis thaliana* evolved novel phenotypic effects on morphological traits under strong positive selection. *Plant Cell* **2022**, *34*, 802–817. [CrossRef] [PubMed]
53. Catlin, N.S.; Josephs, E.B. The important contribution of transposable elements to phenotypic variation and evolution. *Curr. Opin. Plant Biol.* **2022**, *65*, 102140. [CrossRef]
54. Niu, X.-M.; Xu, Y.-C.; Li, Z.-W.; Bian, Y.-T.; Hou, X.-H.; Chen, J.-F.; Zou, Y.-P.; Jiang, J.; Wu, Q.; Ge, S.; et al. Transposable elements drive rapid phenotypic variation in *Capsella rubella*. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 6908–6913. [CrossRef] [PubMed]
55. Zhang, L.; Hu, J.; Han, X.; Li, J.; Gao, Y.; Richards, C.M.; Zhang, C.; Tian, Y.; Liu, G.; Gul, H.; et al. A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. *Nat. Commun.* **2019**, *10*, 1494. [CrossRef] [PubMed]
56. Arrieta-Montiel, M.P.; Shedge, V.; Davila, J.; Christensen, A.C.; Mackenzie, S.A. Diversity of the *Arabidopsis* mitochondrial genome occurs via nuclear-controlled recombination activity. *Genetics* **2009**, *183*, 1261–1268. [CrossRef]
57. Fox-Walsh, K.L.; Dou, Y.; Lam, B.J.; Hung, S.-p.; Baldi, P.F.; Hertel, K.J. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 16176–16181. [CrossRef]
58. Hazra, A.; Mahadani, P. Delineating genome-wide alternative splicing landscapes and their functional significance in orchids. *S. Afr. J. Bot.* **2022**, *148*, 552–560. [CrossRef]
59. Chen, M.-X.; Zhu, F.-Y.; Wang, F.-Z.; Ye, N.-H.; Gao, B.; Chen, X.; Zhao, S.-S.; Fan, T.; Cao, Y.-Y.; Liu, T.-Y.; et al. Alternative splicing and translation play important roles in hypoxic germination in rice. *J. Exp. Bot.* **2019**, *70*, 817–833. [CrossRef]
60. Zhu, F.-Y.; Chen, M.-X.; Ye, N.-H.; Shi, L.; Ma, K.-L.; Yang, J.-F.; Cao, Y.-Y.; Zhang, Y.; Yoshida, T.; Fernie, A.R.; et al. Proteogenomic analysis reveals alternative splicing and translation as part of the abscisic acid response in Arabidopsis seedlings. *Plant J.* **2017**, *91*, 518–533. [CrossRef]
61. Verta, J.-P.; Jacobs, A. The role of alternative splicing in adaptation and evolution. *Trends Ecol. Evol.* **2022**, *37*, 299–308. [CrossRef]
62. Singh, P.; Ahi, E.P. The importance of alternative splicing in adaptive evolution. *Mol. Ecol.* **2022**, *31*, 1928–1938. [CrossRef] [PubMed]
63. Haas, B.J.; Papanicolaou, A.; Yassour, M.; Grabherr, M.; Blood, P.D.; Bowden, J.; Couger, M.B.; Eccles, D.; Li, B.; Lieber, M.; et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **2013**, *8*, 1494–1512. [CrossRef] [PubMed]
64. Eddy, S.R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **2011**, *7*, e1002195. [CrossRef]
65. Marchler-Bauer, A.; Derbyshire, M.K.; Gonzales, N.R.; Lu, S.; Chitsaz, F.; Geer, L.Y.; Geer, R.C.; He, J.; Gwadz, M.; Hurwitz, D.I.; et al. CDD: NCBI's conserved domain database. *Nucleic Acids Res.* **2015**, *43*, D222–D226. [CrossRef] [PubMed]
66. Bailey, T.L.; Johnson, J.; Grant, C.E.; Noble, W.S. The MEME Suite. *Nucleic Acids Res.* **2015**, *43*, W39–W49. [CrossRef] [PubMed]

67. Chen, C.; Chen, H.; Zhang, Y.; Thomas, H.R.; Frank, M.H.; He, Y.; Xia, R. TBtools: An integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* **2020**, *13*, 1194–1202. [CrossRef] [PubMed]
68. Hu, B.; Jin, J.; Guo, A.-Y.; Zhang, H.; Luo, J.; Gao, G. GSDS 2.0: An upgraded gene feature visualization server. *Bioinformatics* **2015**, *31*, 1296–1297. [CrossRef]
69. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European molecular biology open software suite. *Trends Genet* **2000**, *16*, 276–277. [CrossRef]
70. Flynn, J.M.; Hubley, R.; Goubert, C.; Rosen, J.; Clark, A.G.; Feschotte, C.; Smit, A.F. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 9451–9457. [CrossRef]
71. Tempel, S. Using and understanding RepeatMasker. In *Mobile Genetic Elements*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 29–51.
72. Foissac, S.; Sammeth, M. Analysis of alternative splicing events in custom gene datasets by AStalavista. In *RNA Bioinformatics*; Picardi, E., Ed.; Springer: New York, NY, USA, 2015; pp. 379–392.
73. Rozewicki, J.; Li, S.; Amada, K.M.; Standley, D.M.; Katoh, K. MAFFT-DASH: Integrated protein sequence and structural alignment. *Nucleic Acids Res.* **2019**, *47*, W5–W10. [CrossRef]
74. Talavera, G.; Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **2007**, *56*, 564–577. [CrossRef]
75. Lanfear, R.; Frandsen, P.B.; Wright, A.M.; Senfeld, T.; Calcott, B. PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* **2017**, *34*, 772–773. [CrossRef]
76. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [CrossRef] [PubMed]
77. Team, R. RStudio: Integrated Development for R. Available online: http://www.rstudio.com/ (accessed on 19 July 2022).
78. Lawrence, M.; Huber, W.; Pagès, H.; Aboyoun, P.; Carlson, M.; Gentleman, R.; Morgan, M.T.; Carey, V.J. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **2013**, *9*, e1003118. [CrossRef] [PubMed]
79. Yin, T.; Cook, D.; Lawrence, M. ggbio: An R package for extending the grammar of graphics for genomic data. *Genome Biol.* **2012**, *13*, R77. [CrossRef] [PubMed]
80. Wickam, H. *ggplot2: Elegant Graphics for Data Analysis*, 2nd ed.; Springer: New York, NY, USA, 2016; Volume 16, p. 2021.
81. Gustavsson, E.K.; Zhang, D.; Reynolds, R.H.; Garcia-Ruiz, S.; Ryten, M. ggtranscript: An R package for the visualization and interpretation of transcript isoforms using ggplot2. *Bioinformatics* **2022**, *38*, 3844–3846. [CrossRef]
82. Bache, S.; Wickham, H. Magrittr: A Forward-Pipe Operator for R. Available online: https://magrittr.tidyverse.org (accessed on 5 March 2023).
83. Rogozin, I.B.; Wolf, Y.I.; Sorokin, A.V.; Mirkin, B.G.; Koonin, E.V. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.* **2003**, *13*, 1512–1517. [CrossRef]