Insights on the process of reciprocal gene loss in the duplicate DPL genes of rice

Xun Xu
Institute of Botany Chinese Academy of Sciences
Song Ge
Institute of Botany Chinese Academy of Sciences
Fu-min Zhang (zhangfm@ibcas.ac.cn)
Institute of Botany Chinese Academy of Sciences https://orcid.org/0000-0003-4766-8078

Research article

Keywords: Gene duplication, reciprocal gene loss, evolution history, rice

Posted Date: December 19th, 2019

DOI: https://doi.org/10.21203/rs.2.19306/v1

Abstract

Background: Reciprocal gene loss (RGL) of duplicate genes is an important genetic resource of reproductive isolation, which is essential for speciation. In the past decades, various RGL patterns have been revealed, but RGL process is still poorly understood. The RGL of the duplicate *DOPPELGANGER1* (*DPL1*) and *DOPPELGANGER2* (*DPL2*) gene can lead to BDM-type hybrid incompatibility between two rice subspecies. The evolutionary history of the duplicate genes, including their origin and mechanism of duplication as well as their evolutionary divergence after the duplication, remains unclear. In this study, we investigated the evolutionary history of the duplicate genes for gaining insights into the process of RGL.

Results: We reconstructed phylogenetic relationships of *DPL* copies from all 15 diploid species representing six genome types of rice genus and then found that all the *DPL* copies from the latest diverged A- and B-genome gather into one monophyletic clade. Southern blot analysis also detected definitely two *DPL* copies only in A- and B-genome. High conserved collinearity can be observed between A- and B-genomic segments containing *DPL1* and *DPL2* respectively but not between *DPL1* and *DPL2* segments. Investigations of transposon elements indicated that *DPL* duplication is related to DNA transposons. Likelihood-based analyses with branch models showed a relaxation of selective constraint in *DPL1* lineage but an enhancement in *DPL2* lineage after *DPL* duplication. Sequence analysis also indicated that quite a few defective *DPL1* can be found in 6 wild and cultivated species out of all 8 species of A-genome but only one defective *DPL2* occurs in a cultivated rice subspecies.

Conclusions: *DPL* duplication of rice originated in the recent common ancestor of A- and B-genome about 6.76 million years ago and the duplication was possibly caused by DNA transposons. The *DPL1* is a redundant copy and has being in the process of pseudogenization, suggesting that artificial selection may play an important role in forming the RGL of *DPLs* between two rice subspecies during the domestication.

Background

Gene duplication is a major source of new genetic material that is essential for the origin of evolutionary novelties [1]. Generally, gene duplication can arise from a number of independent mechanisms, including tandem, segmental and whole-genome duplication (WGD) as well as TE (Transposable element)-associated mechanisms involving retroposition and transduplication [2-4]. Under different mechanisms, duplicate genes exhibit different patterns in terms of their distribution across the genome. The tandem duplicate genes caused by unequal crossing-over events are always linked on a chromosome and consequently form a cluster with or without intervening genes [2]. Segmental duplication, firstly found in human genome [5], might be caused by TEs in mammals [6, 7] and is generally referred to the rearranged genomic regions derived from WGD in plants [4]. The genomic rearrangements can also account for the syntenic blocks in which the duplicate genes of WGD could be found. Retroposition is a process that mRNAs can be reverse transcribed into cDNA and then inserted into the genome, resulting in intron-free

gene copies [2]. Transduplication is created by transposable elements through capturing and transporting gene, so the duplicated genes are frequently found within the borders of some transposable elements [8].

Three major evolutionary fates of duplicate genes, nonfunctionalization, neofunctionalization and subfunctionalization, have been described by different authors [1, 9, 10]. Nonfunctionalization means that one copy of a pair of duplicates loses function by accumulating degenerative mutations, neofunctionalization refers to that one copy acquires a beneficial function while the other retains the original function, and subfunctionalization is that both copies adopt part of the original functions of the ancient gene. Nowadays, theoretical and empirical investigations have been mainly undertaken on neofunctionalization and subfunctionalization for duplicate genes [8, 11, 12]. In contrast, the evolutionary significance of nonfunctionalization has not been well documented in spite of its prevalence in duplicate genes [10]. Through comparing the genomic databases of several eukaryotic species. Lynch and Conery [10] found that the most of duplicate genes would be silenced within a few million years and the stochastic reciprocal gene loss (RGL) after gene duplication in isolated populations may play an important role in the origin of genomic incompatibilities, which conforms to the Bateson-Dobzhansky-Muller (BDM) model of speciation [10, 13]. In other words, RGL following the gene duplication can lead to postzygotic reproductive isolation between the recent diverged species [14-17].

Currently, it has been commonly accepted that RGL is an important genetic source of postzygotic reproductive isolation, which is essential for the process of speciation. For example, the postzygotic reproductive isolation between tetraodon and zebrafish was found to be created by RGL [17], so was the isolation between polyploid yeasts [16]. Maclean and Greig [15] induced artificial tetraploid hybrid yeasts then found divergent function loss at multiple duplicate loci among populations, indicating that RGL could happen rapidly after gene duplication. In addition, RGL can also occur after the duplication of a small genomic region [14, 18]. A pair of duplicated loci *S27* and *S28* was proved to lead to reproductive isolation between *Oryza sativa* and *O. glumaepatula* [19]. RGL in this pair of loci can also result in reproductive isolation between *O. sativa* and *O. nivara* despite the independent mechanism of mutation of S27 in *O. nivara* [20].

RGL is an important mechanism to creating reproductive isolation, but we still know little about its process. Evolutionary history of duplicate genes in which RGL took place is crucial for understanding the process of RGL. Unfortunately, the evolutionary history on such duplicate genes has not been investigated yet very well [21-23]. Here, we offer a solid example of the evolutionary history for gaining insights into the process of RGL. We chose *DOPPELGANGER1* (*DPL1*) and *DOPPELGANGER2* (*DPL2*), a pair of hybrid sterility genes in rice (*O. sativa*), as our studying duplicated loci. The RGL of *DPL1* and *DPL2* can lead to BDM-type hybrid incompatibility between *O. sativa* ssp. *japonica* and *indica* [24]. *DPL1* and *DPL2* that were located on chromosomes 1 and 6, respectively, encode highly conserved proteins of 94 and 95 amino acids and have same gene structure although their lengths of introns are remarkable different [24]. Functional *DPLs* were found to highly express in mature anther of rice and the pollens carrying defective alleles at both loci of *DPL1* and *DPL2* failed to germinate. The defective allele of *DPL1* in *indica* aroused from a 518-bp insertion in the coding region of the second exon while the defective

DPL2 allele found in *japonica* was caused by a splicing site mutation in the second intron that resulted in a readthrough protein. Therefore, one fourth of pollens in the F_1 hybrid between *indica* and *japonica* has defective alleles at both *DPL1* and *DPL2* loci, and thus fail to germinate. The recent research confirmed that *DPL2* is the ancient copy and the duplication event occurred after the divergence of *O.sativa* and *Brachypodium distachyon* [24], but the accurate origin of the gene duplication remains unknown. As we know, there have been roughly 50 million years since *B. distachyon* and *O. sativa* diverged [25], but functional loss of duplicated genes usually happens in a few million years after the duplication [10]. For this reason, we speculate that *DPLs* were produced by a more recent duplication within a few millions years and their functions may be redundant initially. Therefore, in this research, we are to explore the evolutionary history of *DPLs*, including origin and mechanism of the duplication, and the evolutionary divergence of two *DPL* copies following duplication, which, we believe, can provide insights into the process of RGL.

Results

Sequences of *DPLs*

According to previous phylogenetic analyses [26], there are six diploid genome types in *Oryza*, including A-, B-, C-, E-, F- and G-genomes, and *O. sativa* belongs to A-genome. We sampled all 15 diploid species representing the six genome types in *Oryza* and the diploid *Leerisia perrieri* as an outgroup and obtained all the sequences of their *DPLs* (Additional File 1: Table S1). Because there is a striking insertion in the second intron of *DPL1* in A-genome species (Additional File 2: Figure S1) that causes the second intron in *DPL1* is longer than that in *DPL2* [24], we could distinguish *DPL1* from *DPL2* in A-genome species depending on the length of their second intron. According to the previous research [24], two functional *DPL* copies could be found in all species of A-genome except in *O. barthii* and *O. glaberrima* in which a large deletion in their *DPL1* genes occurs and consequently the genes' function lost. In our sequences, we confirmed pseudo alleles of *DPL1* have fixed in Africa rice *O. glaberrima* and its wild ancestor *O. barthii*. No remarkable difference was found on the length of the second intron in the two *DPL* copies of B-genome, but our results of whole genome analysis showed that the two copies located at chromosome 1 and 6 respectively. The genomic locations of two *DPL* copies in B-genome are similar to those in *O. sativa*, so we assumed that B-genome contains both *DPL1* and *DPL2* like A-genome. In C- and E-genome, highly similar copies were found even though we tried various PCR strategies. In F- and G-genome, only one *DPL*-like sequence was isolated separately. We used *L. perrieri* from a closely related genus of *Oryza* as an outgroup and only one *DPL-like* gene was obtained by BLAST searches against the whole genome of *L. perrieri*.

Southern blot analysis

In the light of previous phylogenetic analyses of *Oryza* [26, 29], we redraw a simplified rooted phylogenetic tree of the 6 genome types in *Oryza* (Additional File 3: Figure S2). The tree indicates that the latest diverged genome types are A- and B-genome within *Oryza* while the earliest is G-genome and

following by F-genome. A recent study on 13 genome types of *Oryza* showed that F-genome and the ancestor of A-, B-, C- and E-genome diverged approximately 15 million years ago [28]. Hence, we firstly used Southern blot to detect copy number of *DPL* in the five diploid genome types for determining whether the duplication of *DPLs* originated in this time scale.

In our Southern blot analysis, we used three endonucleases including BamHI, ECoRV and HindIII. The results with different endonuclease were showed in Figure 1. As a positive control, A-genome shows two bands respectively in ECoRV and HindIII in line with our expectation. B-genome also shows two bands respectively in BamHI and ECoRV, indicating two *DPL* copies. Among all the five genomes, only E-genome has bands in all three endonucleases and numbers of bands are two and three, suggesting at least two *DPL* copies. Unlike A-, B- and E-genome that show two or more bands, C- and F-genome was detected only one band separately in one endonuclease thus there may be only one copy. Therefore, we speculate that duplication of *DPLs* happened within *Oryza* and hence our phylogenetic analysis on *DPLs* was conducted within the genus for identifying the accurate origin of the duplicate genes.

Phylogenetic analysis

To reconstruct a phylogenetic tree of *DPLs* of *Oryza* covering all diploid genome types, we used both sequenced and downloaded *DPLs*. All genes used in this study contain coding sequence (CDS) and intron sequences except F-genome in which only CDS were used because the sequences of its intron sequences cannot be aligned with the others.

Using the aligned nucleotide sequences with a length of 1423 bp and 122 parsimony-informative sites in *DPLs* of diploid *Oryza* species and *L. perrieri*, we constructed a bootstrap consensus ML tree to illustrate their phylogenetic relationships (Figure 2). The phylogenetic relationships of the six diploid genome types in our tree are approximately consistent with those revealed by previous studies [26, 30]. In our ML tree, the earliest diverged lineage of *Oryza* is also G-genome and following by F-genome. *DPL* copies of E-, C-, B- and A-genome form a large monophyletic group with 94% bootstrap support, which could be regarded as the sister lineage of F-genome. This group consists of three monophyletic clades, including E-genome clade with 88% bootstrap support, C-genome clade with 100% support and the clade of A- and B-genome with 88% support. Note-worthily, *DPL* copies form a monophyletic clade in E- and C-genome respectively, but not in A- and B-genome separately. In the monophyletic clade of A- and B-genome, *DPL1* copies gather into one monophyletic branch with 99% bootstrap support but *DPL2* gather into three monophyletic branches, involving B-genome, *O. meridionalis* of A-genome, and other A-genome species.

Collinearity analyses and investigations of transposon elements

Using five whole genome database [28, 31], three species of A-genome (*O. sativa*, *O. rufipogon* and *O. glumaepatula*), one of B-genome (*O. punctata*) and one outgroup (*L. perrieri*) were obtained with online database EnsemblPlant (http://plants.ensembl.org). We conducted collinearity analyses between A- and B-genome on *DPL* segments containing *DPLs* and five genes flanking each side of *DPLs* (Figure 3, Additional File 4: Table S2).

Very strong conservations of collinearity between A- and B- genome were found in *DPL1* segments of chromosome 1 and *DPL2* segments of chromosome 6 in *Oryza* separately, but no paralogs were identified between *DPL1* and *DPL2* segments except *DPLs*. We also found two segments in *L. perrieri* that show good collinearity with *DPL1* and *DPL2* segments but *DPL-like* gene only occurs in the segment that has collinearity with *DPL2* segments.

Transposon elements (TEs) were identified in sequences of upstream and downstream intergenic region of *DPL1* of A- and B-genome. TEs of several DNA Transposon families were found around *DPL1*, including Helitron, PIF, TcMar-Stowaway, CMC-EnSpm, hAT-Ac, MULE-MuDR (Additional File 5: Table S3). We found a 9 bp (GAKCTGCCA) repeat sequence at the upstream and downstream of *DPL1*, and the regions between the repeat sequences were orthologous between A- and B-genome. However, we failed to identify any target site of duplication between TEs and *DPL1s*.

Test for selection

We performed the program codeml of PAML to detect significant difference of selective pressure on *DPLs* under branch models. In the analysis, we focused on the ω ratios (ω = dN/dS) of four branches containing A-, B- and C genomes. $\omega 1$ indicates selective pressure on *DPLs* in the branch of the most recent common ancestor of A-, B- and C-genomes. $\omega 2$ refers to selective pressure on *DPLs* in the branch of the most common recent ancestor of A- and B- genome, representing the lineage before the duplication. $\omega 3$ and $\omega 4$ show selective pressure in *DPL2* and *DPL1* lineages respectively, representing the two branches after the duplication. At first, we used M0 Model as a null hypothesis in which one single ω ratio was assumed for all branches. There is no any restriction on ω ratio for any branch in the assumption of Model M1 while ω ratios are set to be different between any two branches in that of Model 2.0. We found significant differences between M1 and M0 Models and between M2.0 and M0 Models (Table 1), rejecting the null hypothesis M0 Model. M2.1-M2.6 assume that there are at least two equal ω ratios, but none of the models is significantly different from M2.0, supporting the M2.0 Model that $\omega 1 \sim \omega 4$ are different from each other. Hence we accepted M2.0 Model as the most suitable model to describe the selective pressures on *DPLs*. In this Model, we found $\omega 1$ is obviously greater than one (10.716) that may caused by nucleotide substitution saturation because it is mainly estimated using the outgroup *L. perrieri* that has a long divergence time from the most recent common ancestor of A-, B- and C-genome [28]. The $\omega 2$ ratio before the duplication is 0.314 that is lower than the ratio of the *DPL1* lineage ($\omega 4$=0.556) but higher than the *DPL2* lineage ($\omega 3$=0.186), suggesting a relaxation of selective pressure in *DPL1* lineage and an enhancement of selective constraint in *DPL2* lineage after the duplication.

Discussion

Origin of the *DPL* duplication

RGL has been confirm to be an important source to reproductive isolation [14, 18, 19, 24, 32-34], but their evolution processes have not been clarified very well. It has been reported that the duplication of the *DPLs*

in rice occurred after the divergence of *O. sativa* and *Brachypodium distachyon* [24], but there have been roughly 50 million years since the divergence of the two species [25] and functional loss of duplicated genes usually happens in a few million years after the duplication [10]. Therefore, we conducted the analyses of Southern blot and phylogeny to reveal the accurate origin of the *DPL* duplication.

Our results suggest that the duplication of *DPLs* in rice happened in the most recent common ancestor of A- and B-genome. At first, our analysis of Southern blot, whole genome BLAST and sequencing implicated that the duplication happened within *Oryza*, supporting the speculation in former research [24]. A previous research, however, believes that F-genome, the second most ancient lineage in *Oryza*, has two *DPLs* and one exists on chromosome 1 and pseudogenized right after the species-specific duplication by double-strand break repair between non-allelic homologous chromosomes [27]. The Southern blot experiment showed F-genome has only one *DPL* copy. In the whole genome of *O. brachyantha* we also found only one copy on chromosome 6 whose location is similar to *DPL2* in rice genome. Furthermore, we tried various PCR strategies to sequence *DPLs* in the most ancient lineage of G-genome, and also obtained only one copy. Therefore, both ancient G- and F-genome have only one *DPL* copy, suggesting that the duplication originated within *Oryza*.

The phylogenetic relationships of diploid *Oryza* species in our ML tree are almost same to those in previous studies [26, 29], offering us a good phylogenetic framework to explore the origin of *DPLs* in rice. All *DPL* copies of A- and B-genome formed a monophyletic branch with high bootstrap support, indicating that the duplication of *DPLs* in rice originated in the common ancestor of A- and B-genome. The estimated divergence time of A- and B-genome is about 6.76 million years ago [28], thus the duplication of rice *DPLs* might happened much later than that estimated in the previous research in which the duplication was thought to occurred after the divergence of *O. sativa* and *B. distachyon* [24]. The *DPL* copies of C- and E-genome used in the phylogenetic analysis did not occur in the clade of A- and B-genome, indicating that they are independent from the duplication of *DPLs* in rice. The Southern blot analysis indicates that C-genome has only one *DPL* copy, which is not in consistence with the former research [24], so we tried more than 10 pairs of primers in all three species of C-genome (Additional File 6: Table S4) and obtained still only one copy though we retained two or more clones for each species in phylogenetic analysis. In contrast to C-genome, the Southern blot analysis indicates that E-genome may have two or more copies, but we obtained two highly similar sequences with various PCR strategies and both of them in phylogenetic analysis clustered in one branch. Like A-genome, B- has two copies locating on chromosome 1 and 6 respectively. All *DPL* copies of A- and B-genome formed a monophyletic clade, indicating that duplication of *DPL* originated in the most recent common ancestor of A- and B-genome.

Mechanism of the *DPL* duplication

In a recent research, it was thought that chromosome 1 and 6 went through an event of double-strand break repair, resulting in the duplicated *DPL1* in *O. brachyantha* genome [27]. However, by this way new copy losses functional structure and pseudogenizes immediately, thus double-strand break repair can

hardly be the cause of the duplication event of *DPLs* in rice in that there are both functional copies in *DPL1* and *DPL2* in A- and B-genome.

Our collinearity analysis indicates that there is conserved collinearity of *DPL1* and *DPL2* segment respectively but not between them, indicating that they don't belong to a pair of syntenic blocks derived from WGD. This means *DPLs* in *O. sativa* are not produced by whole genome duplication. In ancestor of Poaceae, a whole genome duplication [25,35,36] has been confirmed, but no shared ancestral region of genome was found between chromosome 1 and 6 [37]. *DPL* segment of *L. perrieri* showed good collinearity with *DPL2* segments, suggesting *DPL2* is an original copy in consistence with the form study [24]. Unequal cross-over happens between homologous chromosome during meiosis, and results tandem duplication [2]. Therefore, the duplication of *DPLs* was not produced by unequal cross-over as they locate on different chromosomes. Retrotransposons cause increase of copy number in wide range of organisms [38-40], but retrotransposons are unable to bring intron into new copies. Transposons can capture and transport gene copies and result in duplication [8, 41]. Rice *DPLs* share paralog introns, so we thought that the duplication was produced by DNA transposons.

Because *DPL1* is the new copy of *DPLs*, we identified transposons in sequences of upstream and downstream intergenic region of *DPL1*. The results showed that the intergenic regions of *DPL1* in A- and B-genome contained lots of DNA transposons, including superfamilies causing copy number increase. Unfortunately, no target sites of duplication was found in region between each transposon and *DPL1.* We consider that the trace may already loss because of other species-specific DNA movements.

Evolutionary divergence of *DPLs*

The hypothesis of asymmetric evolution ratio in duplicate genes has been commonly accepted [10, 21, 42, 43]. The hypothesis assumes that the evolutionary ratios after gene duplication are more likely to be different between the two duplicate copies. This bias is common and can occur in any sort of gene duplication [3, 44]. An example concerning the bias is found in wild rice. The duplicated loci of *S27* and *S28* caused hybrid pollen sterility between *O. sativa* and *O. glumaepatula* [19]. *O. glumaepatula* and *O. sativa* carry incompatible alleles at *S27* and *S28* locus, respectively. The incompatible *S27* in *O. glumaepatula* is hypothetical loss-of-function, caused by the lack of a specific duplicated segment. *S28* in *O. sativa* fails to expression. Another type of loss-of-function in *S27* locus was found in *O. nivara*, which is also incompatible with the unexpressed *S28* in *O. sativa* [20]. The *O. nivara* contains the specific duplicated segment in *S27* locus, but several mutations at coding and promoter regions may lead to the inactivity of *S27*. Therefore, it is reasonable to assume *S27* is more likely to accumulate mutations than *S28*.

Our study on *DPLs* can offer a comprehensive example with not only the difference of selective pressures between duplicate copies, but also the bias in losing function of the copies. Our results indicated that the selective pressures of the two *DPL* copies are different after the duplication in that the selective constraint on *DPL1* relaxed and that on *DPL2* strengthened. It is in accord with the phenomenon that the pseudo copies are much more likely to occur in *DPL1* than *DPL2* in A-genome since the lack of functional *DPL1*

was found in 6 out of the 8 species in A-genome. *DPL1* defective allele caused by a transposable elements in *O. sativa* and *O. rufipogon*, and functional mRNA of *DPL1* is absent in *O. glumaepatula* [24]. Besides, in our study (unpublished), we found that the same *DPL1* defective allele as *O. sativa* occurs in *O. nivara*. In *O. barthii* and *O. glaberrima*, *DPL1* has lost the first CDS. On the contrary, the defective *DPL2* allele caused by a splicing site mutation in the second intron only occurs in *japonica* rice. These evidences suggest that the new copy *DPL1* is more likely to accumulate mutations and be in the process of losing function while the original copy *DPL2* is more conservative and more likely to be retained. However, functional *DPL1* were retained while *DPL2* defective allele takes up high proportion in *japonica* rice [24, 45], which caused RGL within rice. Therefore, we believe that the retaining of *DPL2* defective allele in rice is caused by artificial selection during the domestication.

Conclusion

In summary, the duplication of *DPLs* in rice originated in the common ancestor of A- and B-genome about 6.76 million years ago, which is much later than that estimated in the previous research. *DPL1* was duplicated from *DPL2*, which might be caused by DNA transposons. After the duplication, the selective pressures are obviously different between the *DPL1* and *DPL2* lineage. The *DPL1* underwent a relaxation of selective pressure while the *DPL2* experienced a stronger selective constraint. More pseudo copies of *DPL1* in A-genome indicated that *DPL1* is a redundant and may be in the process of pseudogenization. On the contrary, the defective *DPL2* allele occurs in *japonica* rice with a high frequency, suggesting that artificial selection may play an important role in forming the RGL in rice during the domestication.

Methods

Species samples, Acquiring online data, DNA isolation and Sequencing

To investigate variation of sequences in *DPLs* within *Oryza*, we sampled all 15 diploid species covering 6 genome types (A-, B-, C-, E-, F- and G- genome) in *Oryza* and a diploid species, *L. perrieri*, from a closely related genus of *Oryza* [26]. One to three accessions were sampled for each species and their sequences of *DPLs* were obtained either by clone sequencing or from online database (Additional File 1: Table S1). We obtained considerable sequences of *DPLs* from online database. Besides those of A- and C-genome used in the previous study [24], we obtained sequences of *DPLs* of B-, F-genome and *L. perrieri* by whole genome BLAST searches. The genomic sequences of B- and F-genome were downloaded from The National Center for Biotechnology Information Center (https://www. ncbi. nm. nih. gov/) [27, 28]. We searched *DPL* against their whole genomic sequences using BLAST in BioEdit [46]. For sequences of homologous gene of *DPLs* in *L. perrieri*, we obtained them by using online BLAST on EnsemblPlant Database (http://plants. ensembl. org).

We used the DNA secure plant kit (TIANGEN, Beijing, China) for our DNA isolation. Genomic DNA for Southern blot was only isolated from fresh leaves while DNA for PCR amplification was isolated from fresh or silica-gel desiccated leaves. DNA for Southern blot was evaluated by NanoDrop (Gene Company

Limited, Hong Kong, China) and 0.8% agarose gel electrophoresis. DNA for PCR amplification was evaluated by 1.5% agarose gel electrophoresis.

PCR primers were designed with Primer Premier 6 (Premier Biosoft Interpairs, CA, USA) or obtained from the previous research [24, 45]. All primers were listed in Additional Files 6 and 7 (Tables S4 and S5). PCR amplification was prepared in a volume of 25 μl reaction using 2×Taq PCR MasterMix (TIANGEN, Beijing, China). All the PCR products were cloned into pEASY-T1 vectors (TransGen Biotech, Beijing, China) and at least 5 independent clones were sanger sequenced (Sangon Biotech, Shanghai, China). For samples from C-, E-, F- and G-genome, we used various PCR strategies to amplify *DPL* copies. The DNA sequences of 12 diploid species in *Oryza* obtained by our clone sequencing were submitted to GenBank (MK569018-MK569047).

Southern blot analysis

To detect the copy number of *DPL* in *Oryza*, we conducted a Southern blot analysis. We designed a probe at length of 320 bp based on the most conservative region of *DPL1* and *DPL2* in rice (Additional File 2: Figure S1). Five diploid genome types, including A-, B-, C-, E- and F- genome, were used in the experiments. For each of them, we sample one species (Additional File 1: Table S1). Samples with DNA concentration higher than 500 ng/μl and no degradation were chosen, and all samples used in the same experience have similar concentration.

In each experience, genomic DNA was divided into two equal parts. Each part was digested by a restriction endonuclease for 18 h at 37°C. Altogether we used three endonucleases including BamHI, ECoRV, or HindIII (Promega, Madison, USA). Every single enzyme digestion contained 4 mg DNA and 50 units of restriction enzyme. The digested genome DNAs were fractionated by the 0.8% agarose gel-electrophoresis in TAE buffer at 60 V for 2 h and then at 40 V for 6 h. After electrophoresis blotted onto Biodyne Plus nylon membrane with a capillary blotting system using 20×SSC (3 M NaCl, 300 mM sodium citrate, pH 7.0) as transfer buffer, prepared membranes were hybridized with a digoxigenin-labelled DNA probe produced by the DIG-High Prime DNA Labeling and Detection Starter Kit II (Roche, German). Membranes were covered evenly with prepared probe under 37°C for 24 h in hybridization oven (HL-2000 HybriLinker, LABRepCo, UK). Bands were visualized by the kit mentioned above and the fluorescence signals were caught by X-ray film.

Phylogenetic analysis

To determine the origin of duplicate *DPL* copies in rice, we conduct a phylogenetic analysis. We used all the 15 diploid species [26] and an outgroup *L. perrieri* to reconstruct a phylogenetic tree with all available sequences of *DPL* copies. All sequences of *DPLs* were aligned by a combination of CLUSTAL X [47] and Muscle [48] initially and then the alignments were manually refined. We employed the final alignments to conduct a phylogenetic analysis using the Maximum Likelihood (ML) method under the Jukes-Cantor model [49] built in MEGA 6 [50]. The reliability of the phylogenetic tree was evaluated by 3,000 bootstraps.

Collinearity analyses and investigations of transposon elements

To explore the mechanism of the duplication of *DPLs* in rice, collinearity analyses were conducted for two genomic segments containing *DPL1* and *DPL2* separately. Each segment contained around five flanking genes on both sides of *DPL1* or *DPL2*. Using BLAST, we searched sequences of *DPLs* and their flanking genes of *O. sativa* (A) (based on Rice annotation project database, Sakai, et al. [31]) against whole-genome database of *O. rufipogon* (A) and *O. glumaepatula* (A), *O. punctata* (B) and *L. perrieri* with online database EnsemblPlant (http://plants.ensembl.org) (annotation based on annotation of Stein, et al. [28]).

According the results of collinearity analyses, the sequences in upstream and downstream intergenic region of *DPL1* gene were used to identified transposon elements with RepeatMasker (http://www.repeatmasker.org) choosing RMBlast as searching engine, then verified simple repeat with Tandem Repeats Finder (http://tandem.bu.edu/trf/trf.submit.options.html). We aligned intergenic region of A- (*O. sativa*) and B-genome (*O. punctata*) with BioEdit [46] to identify orthologous target sites of duplication. The assumed target sites of duplication was searched in the whole intergenic region by Local BLAST of BioEdit and checked manually.

Test for selection

We used the ratio (ω or d_N/d_S) of the number of nonsynonymous substitutions (d_N) to that of synonymous substitutions (ds) for measuring selective pressure on *DPLs*. The program codeml of PAML X [51, 52] was performed to detect significant difference of ω ratios before and after the duplication of *DPLs* under Branch Models. In this approach, a simplified phylogenetic tree (Additional File 8: Figure S6) of *DPL* copies and 285 bp aligned CDS (coding sequence) of *DPL* copies in 5 A-genome species, 1 B-genome species, 3 C-genome species and the outgroup *L. perrieri* were adopted. A likelihood ratio test (LRT) was conducted to detect the significant difference between different models [53, 54]. The one ratio model (M0) assumes single ω ratio for all branches and all sites while the other models allow different ω ratios between branches.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its supplementary information files. All sequence data used in this article are available in GenBank

(http://www.ncbi.nlm.nih.gov) under the accessions MK569018-MK569047.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (31470332; 91731301; 91231201), the grants from the Ministry of Science and Technology of China (2013CB835201) and the Chinese Academy of Sciences (XDB31000000; XDA08020103).

Authors' contributions

F-MZ, SG and XX design the research. XX performed all experiments. XX and F-MZ analyzed the data. F-MZ and XX wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Dr. Yu-su Du who help us with Southern blot and PAML and Dr. Xin-hui Zou who help us with investigations of transposon elements. We also thank the International Rice Research Institute (Los Banos, Philippines) for providing samples.

Abbreviations

CDS: coding sequence; *DPL*: *DOPPELGANGER*; LRT: likelihood ratio test; ML: Maximum Likelihood; RGL: reciprocal gene loss; TE: Transposon element; WGD: whole-genome duplication.

References

1. Ohno S: Evolution by gene duplication. Berlin Heidelberg: Springer; 1970.
2. Zhang J: Evolution by gene duplication: an update. Trends Ecol Evol. 2003; 18:292-298.
3. Freeling M: Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol. 2009; 60:433-453.
4. Panchy N, Lehti-Shiu M, Shiu SH: Evolution of gene duplication in plants. Plant Physiol. 2016; 171:2294-2316.
5. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE: Recent segmental duplications in the human genome. Science. 2002; 297:1003-1007.
6. Bailey JA, Liu G, Eichler EE: An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet. 2003; 73:823-834.
7. She X, Cheng Z, Zöllner S, Church DM, Eichler EE: Mouse segmental duplication and copy number variation. Nat Genet. 2008; 40:909-914.

8. Flagel LE, Wendel JF: Gene duplication and evolutionary novelty in plants. New Phytol. 2009; 183:557-564.
9. Hughes AL: The evolution of functionally novel proteins after gene duplication. Proc Biol Sci. 1994; 256:119-124.
10. Lynch M, Conery JS: The evolutionary fate and consequences of duplicate genes. Science. 2000; 290:1151-1155.
11. Conant GC, Wolfe KH: Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet. 2008; 9:938.
12. Freeling M, Scanlon MJ, Fowler JE: Fractionation and subfunctionalization following genome duplications: mechanisms that drive gene content and their consequences. Curr Opin Genet Dev. 2015; 35:110-118.
13. Orr AH: Dobzhansky, Bateson, and the genetics of speciation. Genetics. 1996; 144:1331-1335. .
14. Bikard D, Patel D, Le Metté C, Giorgi V, Camilleri C, Bennett MJ, Loudet O: Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. Science. 2009; 323:623-626.
15. Maclean CJ, Greig D: Reciprocal gene loss following experimental whole-genome duplication causes reproductive isolation in yeast. Evolution. 2011; 65:932-945.
16. Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH: Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature. 2006; 440:341-345.
17. Semon M, Wolfe KH: Reciprocal gene loss between tetraodon and zebrafish after whole genome duplication in their ancestor. Trends Genet. 2007; 23:108-112.
18. Zuellig MP, Sweigart AL: Gene duplicates cause hybrid lethality between sympatric species of *Mimulus*. PLoS Genet. 2018; 14:20.
19. Yamagata Y, Yamamoto E, Aya K, Win KT, Doi K, Sobrizal, Ito T, Kanamori H, Wu J, Matsumoto T et al: Mitochondrial gene in the nuclear genome induces reproductive barrier in rice. Proc Natl Acad Sci U S A. 2010; 107:1494-1499.
20. Win KT, Yamagata Y, Miyazaki Y, Doi K, Yasui H, Yoshimura A: Independent evolution of a new allele of F-1 pollen sterility gene *S27* encoding mitochondrial ribosomal protein L27 in *Oryza nivara*. Theor Appl Genet. 2011; 122:385-394.
21. Conant GC, Birchler JA, Pires JC: Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. Curr Opin Plant Biol. 2014; 19:91-98.
22. Hudson CM, Puckett EE, Bekaert M, Pires JC, Conant GC: Selection for higher gene copy number after different types of plant gene duplications. Genome Biol Evol. 2011; 3:1369-1380.
23. Rhee SY, Mutwil M: Towards revealing the functions of all genes in plants. Trends Plant Sci. 2014; 19:212-221.
24. Mizuta Y, Harushima Y, Kurata N: Rice pollen hybrid incompatibility caused by reciprocal gene loss of duplicated genes. Proc Natl Acad Sci U S A. 2010; 107:20417-20422.

25. Paterson AH, Bowers JE, Chapman BA: Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci U S A. 2004; 101:9903-9908.
26. Ge S, Sang T, Lu BR, Hong DY: Phylogeny of rice genomes with emphasis on origins of allotetraploid species. Proc Natl Acad Sci U S A. 1999; 96:14400-14405.
27. Chen J, Huang Q, Gao D, Wang J, Lang Y, Liu T, Li B, Bai Z, Luis Goicoechea J, Liang C et al: Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying *Oryza* genome evolution. Nat Commun. 2013; 4:1595.
28. Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, Chougule K, Gao D, Iwata A, Goicoechea JL et al: Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. Nat Genet. 2018; 50:285-296.
29. Zou XH, Zhang FM, Zhang JG, Zang LL, Tang L, Wang J, Sang T, Ge S: Analysis of 142 genes resolves the rapid diversification of the rice genus. Genome Biol. 2008; 9:R49.
30. Zou XH, Yang Z, Doyle JJ, Ge S: Multilocus estimation of divergence times and ancestral effective population sizes of *Oryza* species and implications for the rapid diversification of the genus. New Phytol. 2013; 198:1155-1164.
31. Sakai H, Lee SS, Tanaka T, Numa H, Kim J, Kawahara Y, Wakimoto H, Yang C, Iwamoto M, Abe T et al: Rice annotation project database (RAP-DB): an Integrative and interactive database for rice genomics. Plant Cell Physiol. 2013; 54:e6.
32. Kubo T, Takashi T, Ashikari M, Yoshimura A, Kurata N: Two tightly linked genes at the hsa1 locus cause both F1 and F2 hybrid sterility in rice. Mol Plant. 2016; 9:221-232.
33. Nguyen GN, Yamagata Y, Shigematsu Y, Watanabe M, Miyazaki Y, Doi K, Tashiro K, Kuhara S, Kanamori H, Wu J: Duplication and loss of function of genes encoding RNA polymerase III subunit C4 causes hybrid incompatibility in rice. G3: Genes, Genomes, Genetics. 2017:g3. 117.043943.
34. Sakata M, Yamagata Y, Doi K, Yoshimura A: Two linked genes on rice chromosome 2 for F1 pollen sterility in a hybrid between *Oryza sativa* and *O. glumaepatula*. Breed Sci. 2014; 64:309-320.
35. Wang XY, Shi XL, Hao BL, Ge S, Luo JC: Duplication and DNA segmental loss in the rice genome: implications for diploidization. New Phytol. 2005; 165:937-946.
36. Yu J, Wang J, Lin W, Li SG, Li H, Zhou J, Ni PX, Dong W, Hu SN, Zeng CQ et al: The genomes of *Oryza sativa*: a history of duplications. PLoS Biol. 2005; 3:266-281.
37. Wu Y, Zhu Z, Ma L, Chen M: The preferential retention of starch synthesis genes reveals the impact of whole-genome duplication on grass evolution. Mol Biol Evol. 2008; 25:1003-1006.
38. Cordaux R, Batzer MA: The impact of retrotransposons on human genome evolution. Nat Rev Genet. 2009; 10:691-703.
39. Hirochika H, Fukuchi A, Kikuchi F: Retrotransposon families in rice. Mol Gen Genet. 1992; 233:209-216.
40. Xiao H, Jiang N, Schaffner E, Stockinger EJ, van der Knaap E: A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Science. 2008; 319:1527-1530.

41. Volff JN: Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays. 2006; 28:913-922.
42. Adler M, Anjum M, Berg OG, Andersson DI, Sandegren L: High fitness costs and instability of gene duplications reduce rates of evolution of new genes by duplication-divergence mechanisms. Mol Biol Evol. 2014; 31:1526-1535.
43. Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y: Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A. 2005; 102:5454-5459.
44. Pegueroles C, Laurie S, Mar Alba M: Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. Mol Biol Evol. 2013; 30:1830-1842.
45. Craig SM, Reagon M, Resnick LE, Caicedo AL: Allele distributions at hybrid incompatibility loci facilitate the potential for gene flow between cultivated and weedy rice in the US. PloS One. 2014; 9:e86647.
46. Hall TA: BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser. 1999; 41:95-98.
47. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 1997; 25:4876-4882.
48. Edgar RC: MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinform. 2004; 5:1-19.
49. Munro HN: Mammalian protein metabolism, vol. Evolution of protein molecules. New York: Academic Press; 1969.
50. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013; 30:2725-2729.
51. Xu B, Yang ZH: PAMLX: a graphical user interface for PAML. Mol Biol Evol. 2013; 30:2723-2724.
52. Yang ZH: PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007; 24:1586-1591.
53. Bielawski JP, Yang Z: Maximum likelihood methods for detecting adaptive evolution after gene duplication. J Struct Funct Genomics. 2003; 3:201-212.
54. Yang ZH, Nielsen R: Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol. 1998; 46:409-418.

Table

Table 1. Parameter estimates and values of logarithm likelihood under different branch models and tests of hypotheses

Model	p	Ln	Parameters for Branches	Models Compared	$\|2\triangle L\|$
M1: Free Ratio Model	24	-787.89	ω=0~0.187		
M0: One ratio	1	-803.54	$\omega1=\omega2=\omega3=\omega4$= 0.32503	M1 Vs. M0	31.31***
M2.0: $\omega1\neq\omega2\neq\omega3\neq\omega4$	29	-802.02	$\omega1$=10.716 $\omega2$=0.314 $\omega3$=0.186 $\omega4$=0.556	M2.0 Vs. M0	3.06***
M2.1: $\omega1=\omega2\neq\omega3\neq\omega4$	28	-802.04	$\omega1=\omega2$=0.366 $\omega3$=0.186 $\omega4$=0.556	M2.0 Vs. M2.1	0.05
M2.2: $\omega1=\omega2\neq\omega3=\omega4$	27	-803.46	$\omega1=\omega2$=0.362 $\omega3=\omega4$=0.297	M2.0 Vs. M2.2	2.89
M2.3: $\omega1=\omega3=\omega4\neq\omega2$	27	-803.52	$\omega1=\omega3=\omega4$=0.331 $\omega2$=0.282	M2.0 Vs. M2.3	3.01
M2.4: $\omega1=\omega2=\omega3\neq\omega4$	27	-802.78	$\omega1=\omega2=\omega3$=0.279 $\omega4$=0.556	M2.0 Vs. M2.4	1.52
M2.5: $\omega1=\omega2=\omega4\neq\omega3$	27	-803.05	$\omega1=\omega2=\omega4$=0.375 $\omega3$=0.224	M2.0 Vs. M2.5	2.07
M2.6: $\omega1\neq\omega2=\omega3=\omega4$	27	-803.39	$\omega1$=0.390 $\omega2=\omega3=\omega4$=0.296	M2.0 Vs. M2.6	2.74

*** Significant at the P<0.001 level.

p, number of parameters.

Figures

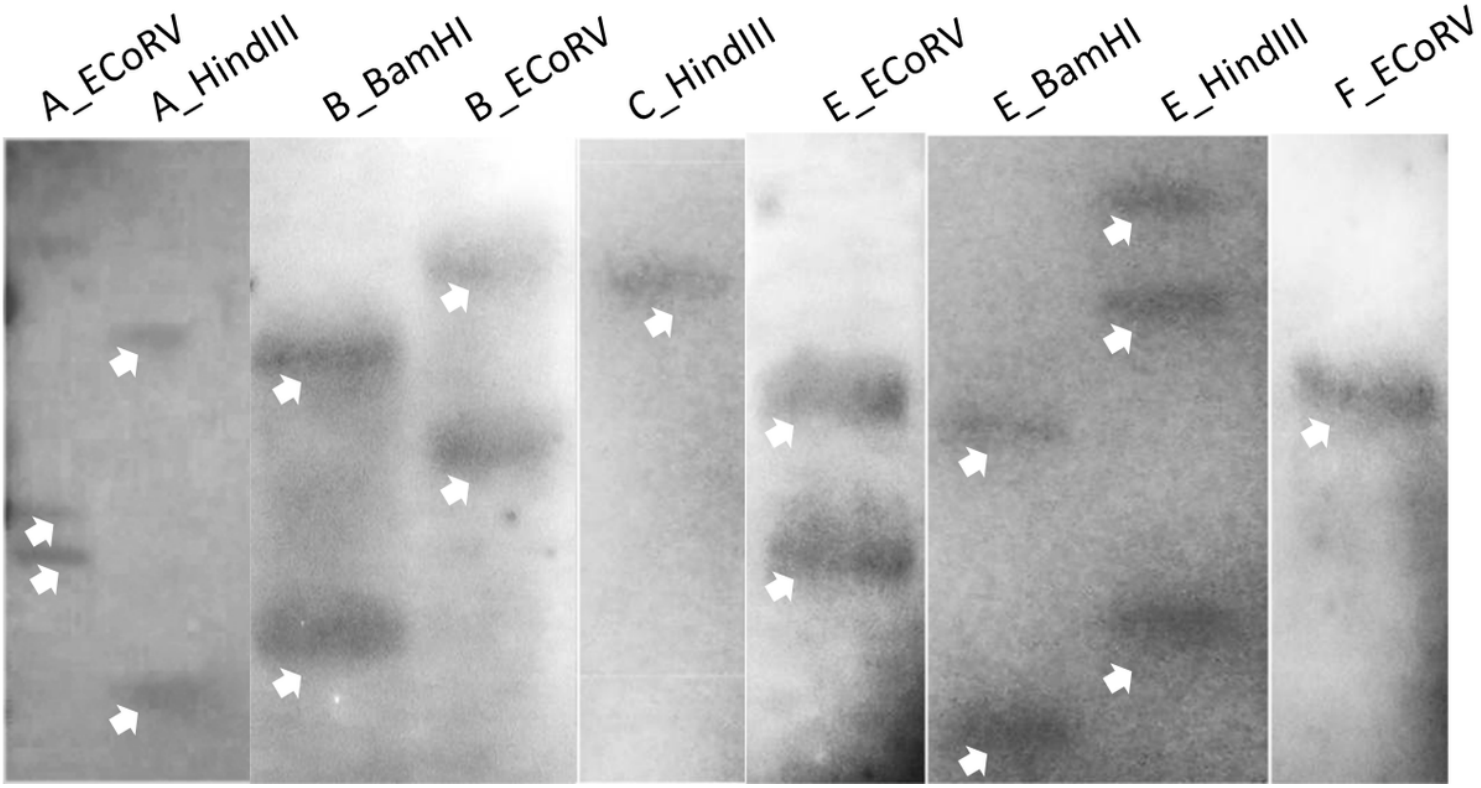

Figure 1

Southern blot results of 5 diploid genome types in Oryza. The results of A-genome (O. rufipogon), B-genome (O. punctata), C-genome (O. officinalis), E-genome (O. australiensis) and F-genome (O. brachyantha) were marked with the genome type and the used restriction endonuclease, respectively.

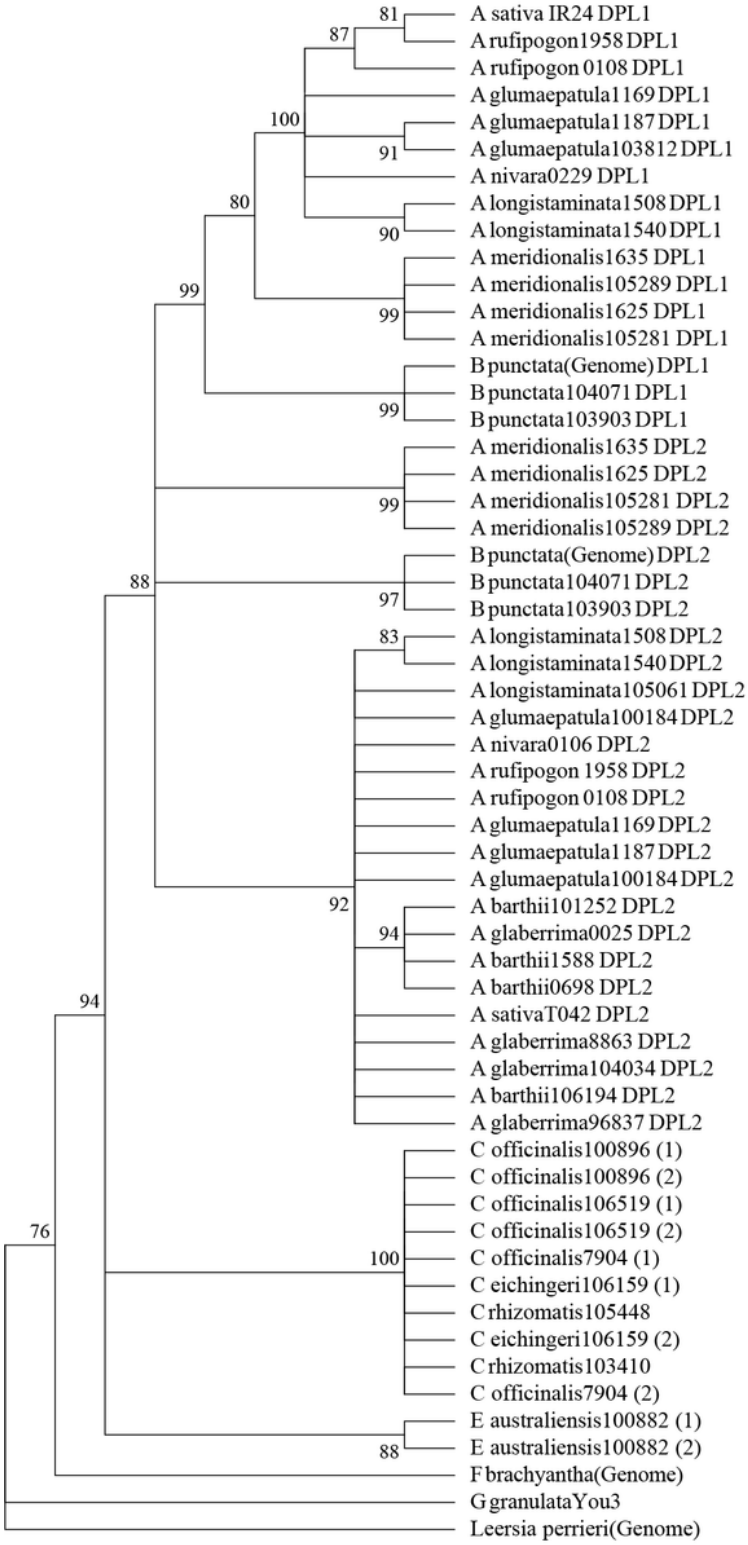

Figure 2

The ML phylogeny of Oryza reconstructed using the sequences of DPLs. The numbers below or above branches show bootstrap percentages.

Figure 3

The collinearity between DPLs flanking regions. Flanking genes are indicated with hollow arrows and DPLs with solid arrows. Double arrows indicate predicted tandem repeats. Homological regions among genomes are indicated with green shadow.

Supplementary Files

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

- Additionalfile1.xlsx
- Additionalfile3.png
- Additionalfile4.xlsx
- Additionalfile6.xlsx
- Additionalfile5.xlsx
- Additionalfile7.xlsx
- Additionalfile8.png
- Additionalfile2.png