

Comparative Genomic and Transcriptomic Analysis of Tandemly and Segmentally Duplicated Genes in Rice

Shu-Ye Jiang¹, José M. González², Srinivasan Ramachandran^{1*}

1 Temasek Life Sciences Laboratory, The National University of Singapore, Singapore, Singapore, **2** Department of Microbiology, University of La Laguna, La Laguna, Tenerife, Spain

Abstract

Tandem and segmental duplications significantly contribute to gene family expansion and genome evolution. Genome-wide identification of tandem and segmental genes has been analyzed before in several plant genomes. However, comparative studies in functional bias, expression divergence and their roles in species domestication are still lacking. We have carried out a genome-wide identification and comparative analysis of tandem and segmental genes in the rice genome. A total of 3,646 and 3,633 pairs of tandem and segmental genes, respectively, were identified in the genome. They made up around 30% of total annotated rice genes (excluding transposon-coding genes). Both tandem and segmental duplicates showed different physical locations and exhibited a biased subset of functions. These two types of duplicated genes were also under different functional constraints as shown by nonsynonymous substitutions per site (K_a) and synonymous substitutions per site (K_s) analysis. They are also differently regulated depending on the tissues and abiotic and biotic stresses based on transcriptomics data. The expression divergence might be related to promoter differentiation and DNA methylation status after tandem or segmental duplications. Both tandem and segmental duplications differ in their contribution to genetic novelty but evidence suggests that they play their role in species domestication and genome evolution.

Citation: Jiang S-Y, González JM, Ramachandran S (2013) Comparative Genomic and Transcriptomic Analysis of Tandemly and Segmentally Duplicated Genes in Rice. PLoS ONE 8(5): e63551. doi:10.1371/journal.pone.0063551

Editor: Dorian Q. Fuller, University College London, United Kingdom

Received: November 22, 2012; **Accepted:** April 3, 2013; **Published:** May 16, 2013

Copyright: © 2013 Jiang et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This research was supported by Temasek Life Sciences Laboratory (Singapore) and by the Singapore National Research Foundation under the Competitive Research Programme Funding Scheme (CRP Award No. NRF-CRP7-2010-02). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sri@tll.org.sg

Introduction

Gene duplication is prominent in eukaryotes. More than one-third of protein-coding genes belong to multigene families in model organisms [1,2]. In rice, based on our preliminary study, around 6,000 gene families were detected to encode more than two thirds of the total annotated non-transposon proteins. Both tandem and segmental duplications significantly contribute to the origin, expansion and evolution of multigene families. Tandemly duplicated genes are located next to the original copy or are separated by several un-related genes. They are presumed to originate through unequal crossing over or transposon activities [3,4]. Segmentally duplicated genes result from duplications of chromosomal regions ranging from 1 to 400 Kb [5,6]. They arise from the genomic restructure caused by aberrant inter- or intra-chromosome recombination [7].

The genome-wide identification of tandemly duplicated genes has been carried out in both *Arabidopsis* and rice genomes. In both genomes, tandemly duplicated genes are enriched in genes encoding membrane proteins that function under “abiotic and biotic stress” [8]. In *Arabidopsis*, genome-wide identification of segmentally duplicated genes has been also studied [9,10]. More than 3,000 pairs of segmentally duplicated genes in rice have been identified [11]. Both tandem and segmental duplicates have significantly contributed to the evolution of large gene families in *Arabidopsis* [10]. The investigation from Hanada et al (2008) gives

evidence for the importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli [12]. Besides *Arabidopsis* and rice, genome-wide identification of tandem or segmental genes was also investigated in a few other genomes [13,14]. Tandem and/or segmental duplications have significantly contributed to the expansion of some gene families in organisms other than *Arabidopsis* [15–19]. Recently, Wang et al (2011) concluded that gene duplication modes contribute differently to genetic novelty and redundancy [20]. However, relatively little has been reported on the comparative analysis of these two duplicate modes in their roles in biological function and species evolution.

What is of particular interest is the fate of duplicated genes. In the classic model a duplicated gene has either lost or gained its new function, which is referred to as pseudogenization or neofunctionalization, respectively [21]. However, based on genomic and transcriptomic data, a much more complex model “duplication–degeneration–complementation (DDC)” was reported [22]. Indeed, the retention mechanisms of duplicated genes were quite diverse [23–27]. After gene duplication, one copy might be silenced due to the absence of any selective constraint within the genome [28]. Sometimes gene conversion might play a role in the survival of paralogous genes [29]. Another possibility is that one of the two copies gradually developed a similar new function (subfunctionalization) [30]. A third possibility is that one of the two copies acquires a new function [31]. Tandemly and segmentally

duplicated gene pairs provide an excellent genetic collection to study the retention mechanisms of duplicated genes. Both tandem and segmental duplications are originated through completely different mechanisms and the comparison in their functional divergence provides clues to understand their roles in biological evolution and species divergence.

In this study, we first carried out a genome wide identification of all tandemly or segmentally duplicated genes based on the latest version of annotated rice genes. We then examined and compared functional specificities of both tandem and segmental duplicates by gene function enrichment analysis. We also compared and evaluated protein divergence of these two modes of duplicated genes by *Ka/Ks* analysis (where *Ka* = nonsynonymous substitutions per site, and *Ks* = synonymous substitutions per site). In addition, expression divergence among different tissues and genotypes as well as under different abiotic and biotic stresses was investigated and compared to further evaluate their functional divergence after duplication. Finally, we analyzed the effect of promoter similarity and DNA methylation on expression divergence of duplicated genes. Our data showed that the rice genome encodes considerable tandemly or segmentally duplicated genes. Both tandem and segmental duplicates exhibited a biased subset of molecular functions. Both modes of duplicated genes were also under different functional constrains as shown by *Ka/Ks* analysis. Our data imply that these duplicated genes play a role in sub-species diversity in rice.

Results

Genome-wide Identification of Tandemly or Segmentally Duplicated Genes in Rice

In the latest version (release 7) of the MSU rice genome annotation (<http://rice.plantbiology.msu.edu>) [32], a total of 55,986 loci were predicted, including 16,941 loci encoding transposon/retrotransposon elements (TEs). Thus, 39,045 genes were predicted to encode non-TE proteins. These genes and their predicted proteins were used for the genome-wide identification of tandemly and segmentally duplicated genes according to the description in the Methods. Based on our searches, we have identified 3,646 pairs of tandem duplicates consisting of 5,888 (15.1%) annotated genes (Figure 1 and Table S1). In addition, a total of 3,633 pairs of segmental duplicates were detected, which consisted of 6,231 (16.0%) annotated genes (Figure 1 and Table S2). Some of the duplicated genes originated from both tandem and segmental duplications. Thus, a total of 11,500 genes were detected, which were involved in either tandem or segmental duplication, accounting for 29.5% of total annotated non-TE genes (Figure 1).

Distribution of Tandem and Segmental Duplicates on the Chromosomes

On each chromosome, tandem genes ranged from 316 (5.6%) to 838 (14.2%) and segmental genes were from 250 (4.0%) to 844 (13.5%). On chromosomes 1, 4, 6, 8, 9, and 12, similar numbers of tandem duplicates were detected when compared with segmental duplicates (Figure S1). On chromosomes 2, 3 and 5, significantly higher numbers of segmental genes have been detected. However, on chromosomes 7, 10 and 11, a greater ratio of tandem genes were detected. In general, these tandemly and segmentally duplicated genes were not evenly distributed on the 12 rice chromosomes and they exhibited variability in their location on the chromosome (Figure 2). Tandem genes on chromosomes 11 and 12 were distributed evenly except for centromere regions. For other tandem genes, they exhibited uneven distributions with a

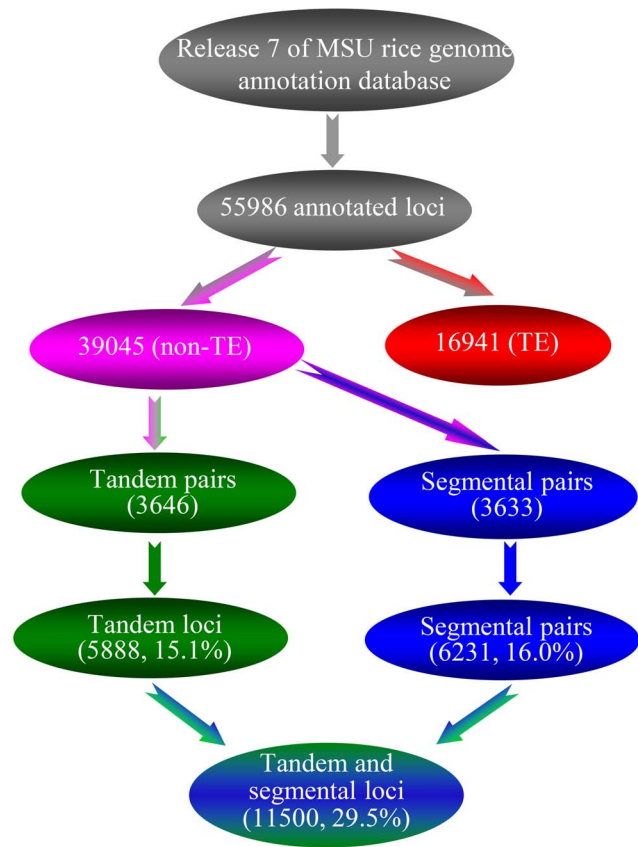


Figure 1. Schematic diagram for identification of tandemly and segmentally duplicated genes. The release 7 of the MSU Rice Genome Annotation Project Database (<http://rice.plantbiology.msu.edu/index.shtml>) was used for the identification of tandem and segmental genes. Genes encoding TEs were excluded for the identification.

doi:10.1371/journal.pone.0063551.g001

tendency to cluster near one end or either end of the chromosome (blue curves in Figure 2). Such a tendency was observed more frequently for segmental genes (pink curves in Figure 2). Higher frequencies of segmental genes were observed on the long arm or the end of chromosomes 4, 5, 6, 7, 8, and 10, while the majority of segmental genes were located at the first 5 Mb of chromosomes 11 and 12. Furthermore, overlap distribution of tandem and segmental genes was observed for all chromosomes. However, only 619 genes were detected to have undergone both tandem and segmental duplications. In addition, some chromosomal regions with low frequency of tandem duplication usually showed high frequency of segmental duplication.

Comparative Analysis of Functional Specificities of Tandemly and Segmentally Duplicated Genes

A protein domain/motif, which is usually highly conserved, play important roles in determining protein functions. To investigate functional specificities of tandem and segmental genes, we ran Pfam searches (<http://pfam.sanger.ac.uk/>) using all duplicated proteins to predict possible domains/motifs. We selected 10 domains/motifs, which were most frequently detected among tandemly or segmentally duplicated proteins for further analysis (Figure 3 A and B). We detected three domains/motifs that were commonly presented in both tandem and segmental proteins. These domain IDs were PF00069 (Protein kinase domain),

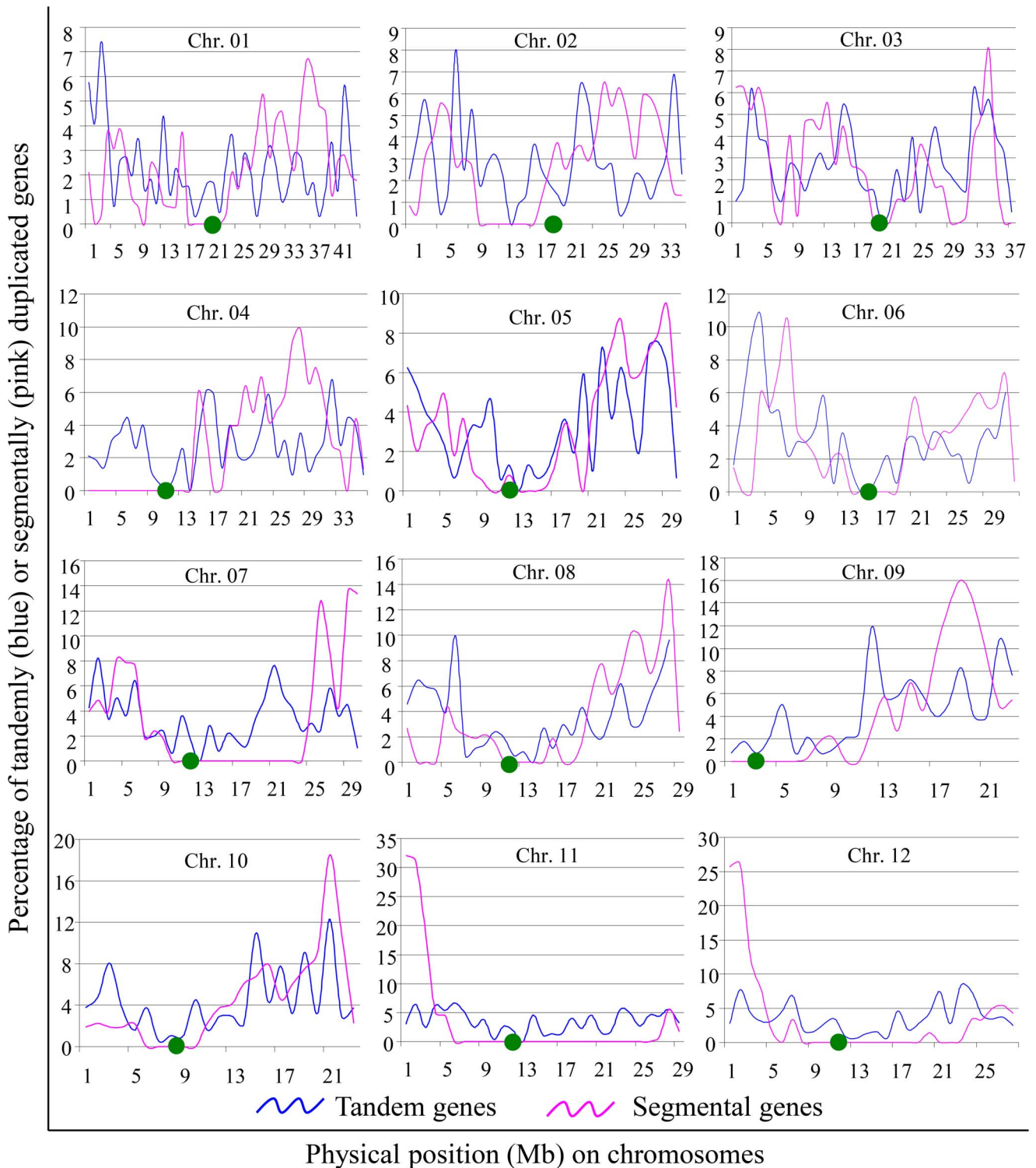


Figure 2. Genomic organization and physical distribution of tandemly and segmentally duplicated genes in the rice genome. Density is indicated by the percentage of the number of tandem/segmental genes in the total number of annotated genes. X-axis indicates chromosomal positions (Mb). Y-axis indicates gene density (the percentage of total number of genes). The distributions of tandem and segmental genes are marked in blue and pink, respectively. Centromere positions are shown with green dots on each chromosome. doi:10.1371/journal.pone.0063551.g002

PF00560 (Leucine Rich Repeat) and PF01535 (Pentatricopeptide repeat). The first two domains were over-represented in both tandem and segmental proteins. The data suggest that these genes

encoding protein kinase and leucine-rich repeat might have undergone large expansion mainly by tandem and segmental duplication during evolution. Pentatricopeptide repeat (PPR) is a

35-amino acid sequence motif, which was commonly found in the plant kingdom. Significantly reduced PPR members were duplicated by segmental duplication when compared with tandem duplication (Figure 3 A and B). The data also suggest that the motif containing family members were mainly expanded by other models of duplication or by transposition.

Besides the three domains, both tandem and segmental duplication exhibited differences in the expansion of the other seven domains/motifs. Tandem duplication played a role in the expansion of p450, peroxidase, UDPGT (UDP-glucuronosyltransferase), F-box, NB-ARC (nucleotide-binding adaptor shared by APAF-1, certain *R* gene products and CED-4), DUF295 (Domain of unknown function) and LRRNT_2 (Leucine rich repeat N-terminal) domains/motifs. Many of these domain/motif containing members function under abiotic and biotic stress related biological processes. Segmental duplication significantly contributed to the expansion of Myb_DNA-binding, zf-C3HC4 (Zinc finger, C3HC4 RING-type), WD40 repeat (short ~40 amino acid motifs, often terminating in a Trp-Asp dipeptide), ehand (a helix-loop-helix structural domain), Pkinase_Tyr (protein tyrosine kinase), RRM_1 (RNA recognition motif) and AP2 (APE-TALA2/ethylene-responsive element-binding protein) domain/motif containing members. Many of these members encode transcription factors or regulatory proteins. Thus, these over-represented segmental genes might play a role in transcription regulation and signal transduction.

To further examine the difference in functional specificities between these tandemly or segmentally duplicated genes in rice, we investigated Gene Ontology (GO) terms to identify overrep-

resented GO terms by GSEA (see Methods). For each term, we identified GO terms in three categories: biological process (P), molecular function (F), and cellular component (C) [33]. Our primary motivation for this analysis was to evaluate whether duplicated genes from different duplication modes were biased toward particular functions. Our data set showed that only one GO term (signaling; GO:0023052; highlighted with red fonts in Figure 4 A and B) was commonly detected in both tandem and segmental genes. Further investigation showed that duplicated tandem genes might play roles in response to stimulus and death (category P), exhibited catalytic activity (category F) and were located in the extracellular region (category C) (Figure 4A). For segmental genes, a total of ten GO terms were detected with over-representation in the category P involved in biological regulation, growth, signaling, localization etc. (Figure 4B). They were also over-represented in transcription regulator and binding activities (category F). The data suggest that different duplication modes produce duplicated genes with biased subsets of biological functions, which might provide the basis for species domestication and genome evolution.

Protein Divergence after Tandem or Segmental Duplication

Tandemly or segmentally duplicated genes accounted for 29.5% of the total annotated non-TE genes. Thus, it would be of interest to know if these duplicated descendants are still functional or have become pseudogenes. The K_a/K_s ratios of these duplicated pairs were estimated and were tested statistically. Most K_a/K_s values for tandemly duplicated pairs were

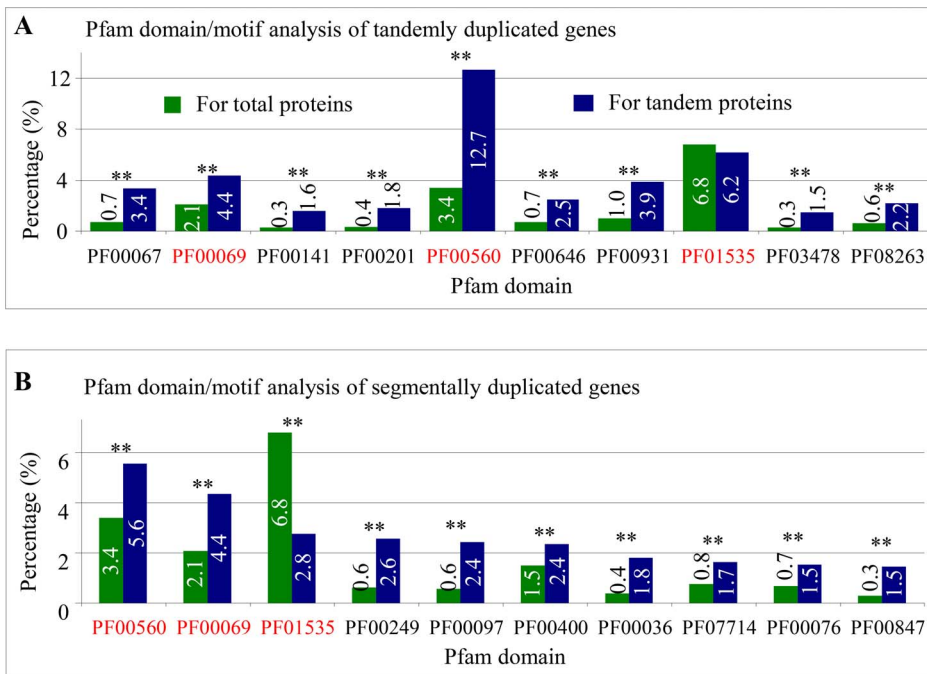


Figure 3. Protein domain analysis in tandemly and segmentally duplicated genes. (A) Pfam domain analysis of tandemly duplicated genes. Green and blue columns indicate the percentages of this domain in the total and tandem proteins, respectively. (B) Pfam domain analysis of segmentally duplicated genes. Green and blue columns indicate the percentages of this domain in the total and segmental proteins, respectively. Commonly detected domains in both tandem and segmental proteins are highlighted in red. Two stars indicate statistically significant differences at P value <0.01. PF00067, p450 domain; PF00069, Pkinase domain; PF00141, peroxidase family; PF00201, UDPGT family; PF00560, LRR_1 repeat; PF00646, F-box domain; PF00931, NB-ARC domain; PF01535, PPR motif; PF03478, DUF295 family; PF08263, LRRNT_2 family. Percentage was calculated as the frequency of the Pfam domain in tandemly duplicated proteins and the domain percentage in all proteins was used as the control. PF00249, Myb_DNA-binding; zf-C3HC4; PF00097, WD40 repeat; PF00400, ehand; PF00036, Pkinase_Tyr; PF07714, RRM_1; PF00076, RRM_1; PF00847, AP2. doi:10.1371/journal.pone.0063551.g003

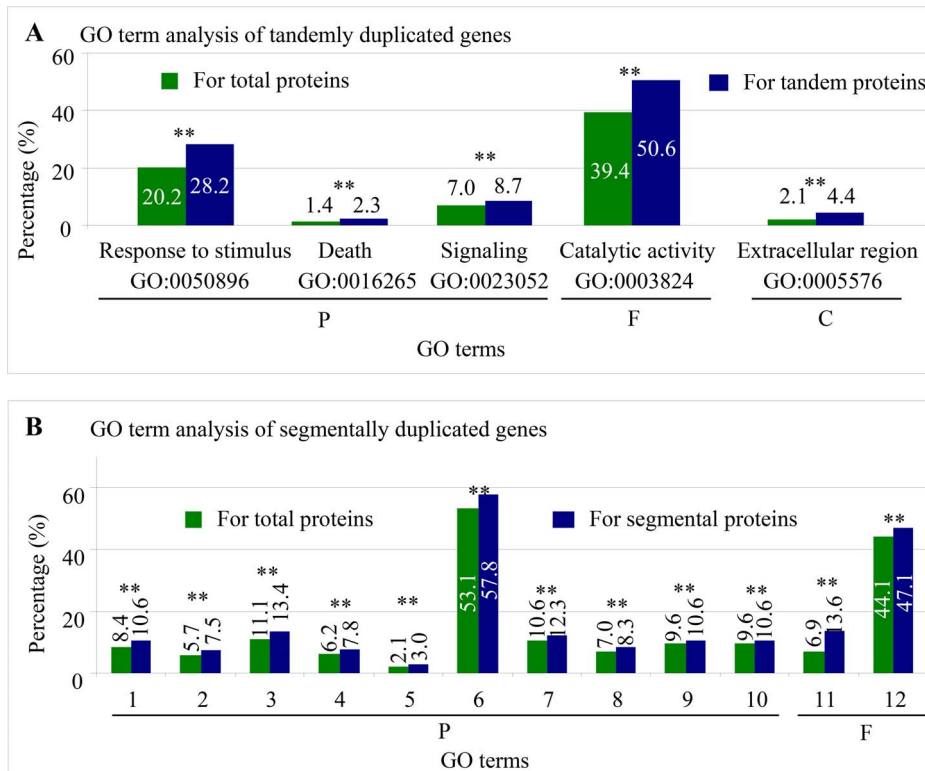


Figure 4. Gene set enrichment analysis of tandemly and segmentally duplicated genes. (A) GO term analysis of tandemly duplicated genes. (B) GO term analysis of segmentally duplicated proteins. Green and blue columns in (A) and (B) indicate the percentages of this GO term in total and tandem proteins, respectively. The percentage is calculated as the frequency of the total numbers of each GO term in all proteins with GO term assigned or in all segmentally duplicated proteins with GO term assigned. The two stars in (A) and (B) represent statistically significant differences at P value <0.01. "P," "F" and "C" in (A) and (B) indicate the GO categories biological process, molecular function and cellular component, respectively. GO term annotation in (B) refers to the following: 1, GO:0065007 (biological regulation); 2, GO:0023046 (signaling process); 3, GO:0032502 (developmental process); 4, GO:0050789 (regulation of biological process); 5, GO:0040007 (growth); 6, GO:0009987 (cellular process); 7, GO:0032501 (multicellular organismal process); 8, GO:0023052 (signaling); 9, GO:0051234 (establishment of localization); 10, GO:0051179 (localization); 11, GO:0030528 (transcription regulator activity); 12, GO:0005488 (binding). doi:10.1371/journal.pone.0063551.g004

centred near 0.3 (blue line in Figure 5A) with an average Ka/Ks ratio of 0.287. For segmentally duplicated genes, most were close to a $Ka/Ks = 0.2$ (pink line in Figure 5A) with an average Ka/Ks ratio of 0.141. These data suggest that most of segmental genes have been subject to stronger functional constraints when compared to tandem genes. To further assess the extend of the selective pressure between tandem and segmental duplicates, these genes were submitted to another set of Ka/Ks analysis. Such an analysis showed similar results (Figure 5C), further confirming that segmental genes have undergone a higher selective constraint. After tandem/segmental duplication, the Ka/Ks ratio in a pair could be as low as 0.5 if one gene maintains its original function and the other copy is a pseudogene [34]. Therefore, the ratio of 0.5 was taken as conservative criterion to test the null hypothesis that the ratios are equal to or smaller than 0.5 or greater than 0.5 by C-value test ($C = (X - 0.5N) / (0.5 \times \sqrt{N})$), where X is the number of pairs with $Ka/Ks < 0.5$ and N is the total number of pairs [34]. We further calculated the C-value to test the null hypothesis that one gene maintains its original function and the other copy is a pseudogene. The calculation showed that C-values of both tandem and segmental pairs are 43.9 and 57.8, respectively (Figure 5D). The results suggested that the probability of the null hypothesis should be very low ($P < 0.001$) for both tandem and segmental pairs (Figure 5D). Up to 84.9% and 97.4% of

tandem and segmental genes showed functional constraints (Figure 5D). Thus, most duplicated members are generally under strong selective constraints and both members in each pair should be functional.

Expression Divergence of Paralogs from Tandem or Segmental Duplication within a Variety

Our data showed that most tandem or segmental genes were under selective constraints. We further analyzed whether these duplicates showed the difference in their expression patterns. We investigated their expression patterns by employing both MPSS and microarray data as described in the Methods. In Nipponbare, only 50.9% of tandem genes were detected with expression signaling by MPSS while up to 83.4% of segmental genes showed expression using the same database (Figure 6A). Similar results were observed when microarray expression data were analyzed in Nipponbare (Figure 6A). Thus, significantly higher percentage of segmental genes was expressed when compared with tandem genes. We then examined expression divergence between tandem or segmental gene pairs in different tissues or their abundance. Around 3.1% or 17.9% of tandem pairs showed difference in their expression patterns among different tissues or their expression abundance (Figure 6B). Similar percentages of pairs were detected with divergence in different tissues or expression abundance for segmental pairs (Figure 6B).

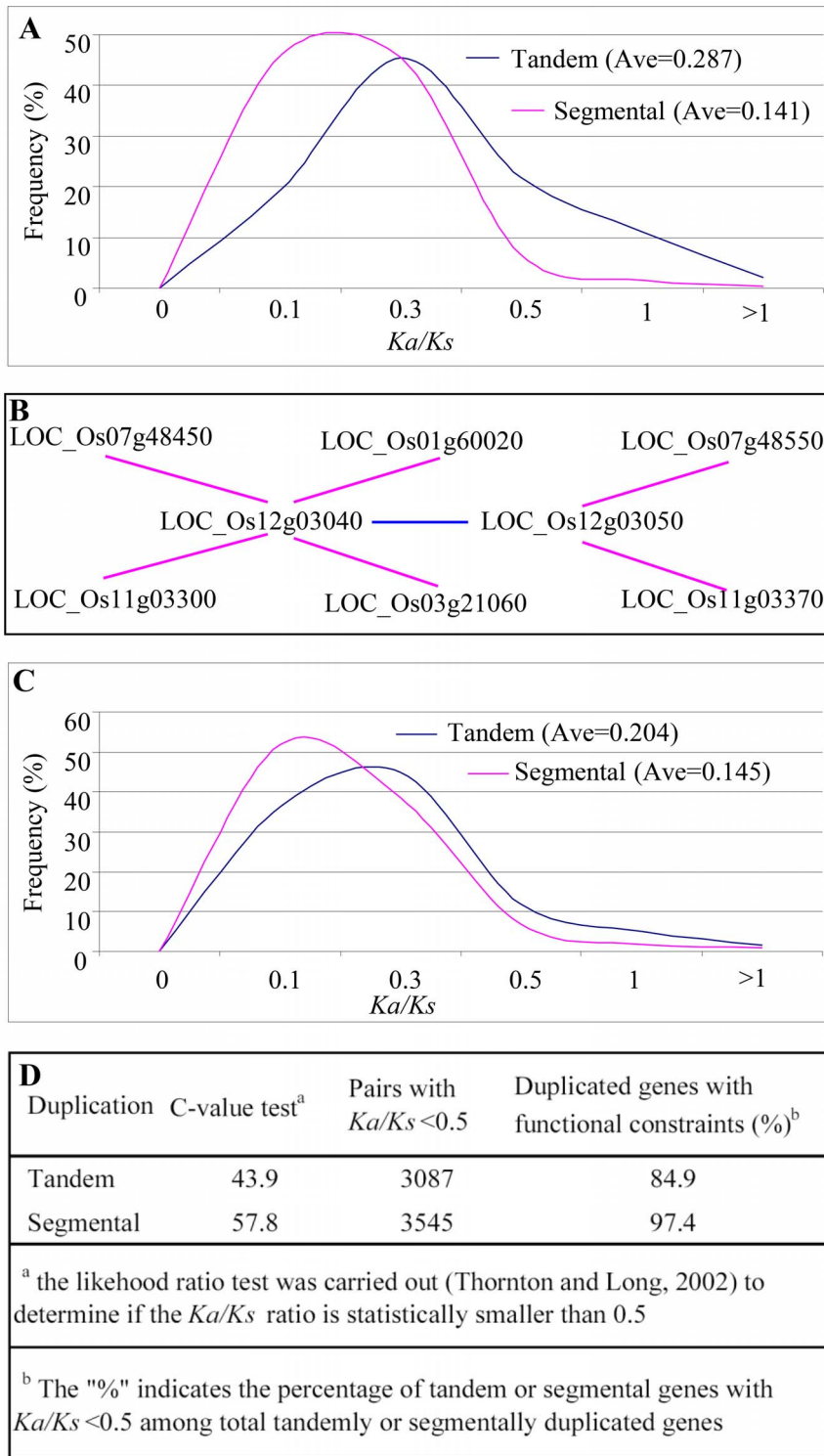


Figure 5. Frequency distributions and C-value test of Ka/Ks ratios between gene pairs duplicated from tandem (blue line) or segmental (red line) duplication. (A) Ka/Ks analysis was carried out using all pairs of tandemly or segmentally duplicated genes. (B) An example of genes that have undergone both tandem and segmental duplications. (C) Ka/Ks analysis using gene pairs from both tandem and segmental duplications. (D) Analysis of functional constraints by C-value test. doi:10.1371/journal.pone.0063551.g005

We further investigated their expression regulation of tandem and segmental genes under various biotic and abiotic stresses. For biotic stresses, we analyzed the effects of compatible/incompatible fungi and bacteria on expression. They were labeled as FungusR/

FungusS for compatible/incompatible fungus and BacteriumR/BacteriumS for compatible/incompatible bacterium, respectively. On the other hand, we also analyzed expression profiles under various abiotic stresses including cold, drought and high salinity.

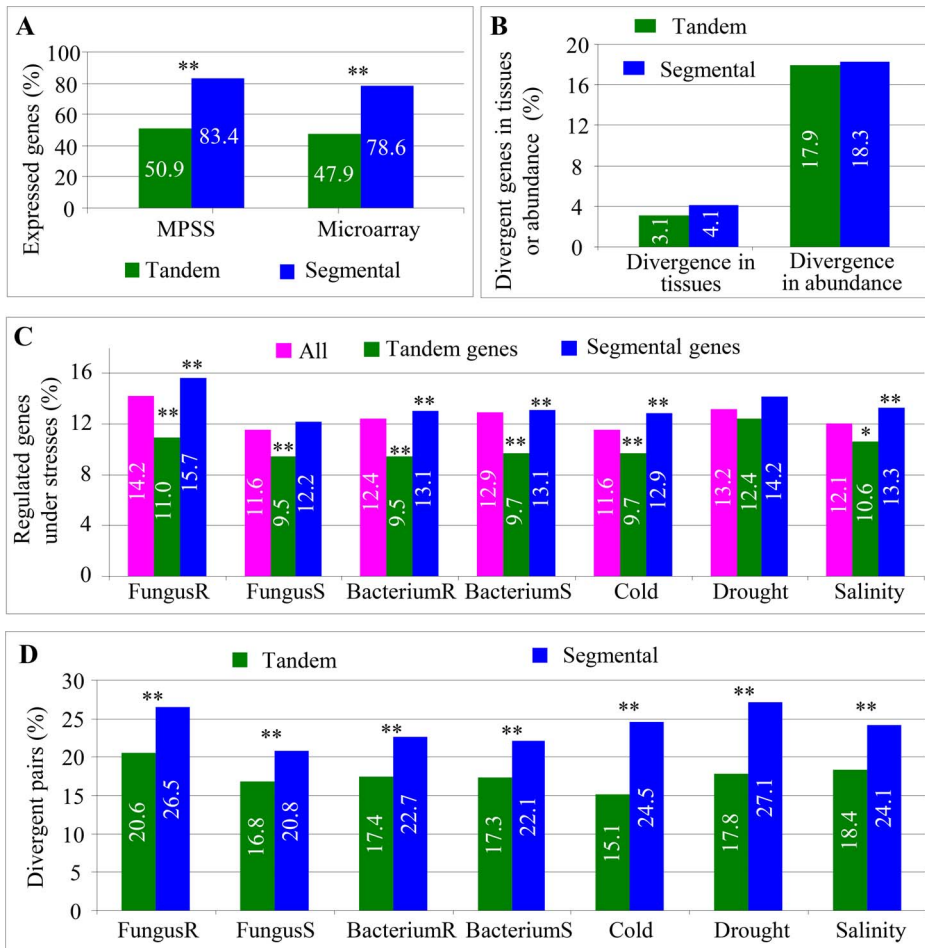


Figure 6. Expression divergence of tandem and segmental genes in Nipponbare. (A) Summary of expression profiles of tandem and segmental genes. The analysis was based on MPSS and Affymetrix microarray expression data. Some of the annotated genes were not probed in the array chips. (B) Expression divergence of tandem and segmental genes in tissues or abundance. (C) Up- or down-regulated tandem and segmental genes under various biotic and abiotic stresses. (D) Divergent tandem and segmental pairs in their expression under various biotic and abiotic stresses. In (C) and (D), FungusR, FungusS, BacteriumR and BacteriumS indicated the biotic stress treatments with incompatible fungus, compatible fungus, incompatible bacterium and compatible bacterium, respectively. doi:10.1371/journal.pone.0063551.g006

In general, a smaller percent of tandem genes (green columns) were detected with regulated expression patterns under various biotic and abiotic stresses; however, higher percentages were detected for segmental genes (blue columns) (Figure 6C). For example, under the FungusR stress, we have detected 14.2% of total tested genes with differential expression profiles (pink column in Figure 6C) and the percentage was reduced to only 11.0% for tandem genes (green column in Figure 6C). However, the percentage was increased to 15.7% for segmental genes (blue column in Figure 6C). On the contrary, no difference was detected for the expression regulation under the FungusS stress for segmental genes and under drought stress for both tandem and segmental genes (Figure 6C).

We also examined expression dissimilarities under various biotic and abiotic stresses between tandem or segmental pairs. Interestingly, significantly higher percentages of segmental pairs (blue columns in Figure 6D) exhibited expression divergence under either biotic or abiotic stresses when compared with tandem pairs (green columns in Figure 6D). For segmental pairs, 20.8–27.1% of pairs showed expression divergence while the percentages were reduced to only 15.1–20.6% for tandem pairs (Figure 6D).

Expression Regulation and Divergence among Genotypes

We have analyzed expression profiles of tandem and segmental genes under normal and stressed conditions in the rice variety Nipponbare. We further examined the expression profiles of these orthologs in other rice genotypes (IR29, FL478 and IR64). Slightly less or similar percentage of expressed genes was detected either for tandem or segmental genes when compared with the data for Nipponbare (Figure 6A and Figure 7A). In addition, a similar trend was observed, that is, significantly higher percentages of segmental genes were expressed when compared with tandem genes. Available data analysis also showed that different genotypes exhibited divergence in the numbers of tandem/segmental genes regulated by high salinity stress (Figure 7B). The highest percentage of segmental genes (16.7%) was observed to be regulated in their expression by the stress. Segmental duplication significantly contributed to expression divergence under high salinity stress in Nipponbare, IR29 and IR64, all of which were salinity-sensitive species. However, in the salinity-tolerance line FL478, no difference was observed (Figure 7B). Genome-wide comparative expression analysis provided a platform to analyze

allele-specific expression patterns [35]. We further investigated the expression pattern of tandem or segmental genes in japonica Nipponbare and indica 93-11 (Figure 7C). In total, we detected 2.7% of all genes, which were differentially expressed between these two varieties. However, among tandem or segmental genes, up to 5.3% or 3.3% of the genes, respectively, showed expression divergence (Figure 7C). These data imply a role of tandem and segmental duplication in species divergence.

Promoter Variation and DNA Methylation in Tandem and Segmental Genes

Both tandemly and segmentally duplicated genes exhibited significant transcriptional similarities as well as divergences (Figures 6 and 7). The observation prompted us to analyze further their promoter variation after duplication. We determined the promoter similarity by aligning the promoter sequence and comparing these alignments to randomly selected promoter pairs (see Methods). For randomly selected promoter pairs, more than 60% of them had promoter similarity smaller than 10% (pink curve in Figure 8A). For tandem promoter pairs, around 45% of them exhibited the promoter similarity at around 20% (green

curve in Figure 8A). For segmental promoter pairs, less than 50% of them had promoter similarity smaller than 20% (blue curve in Figure 8A). In general, the average promoter similarity is only 9.1% for randomly selected promoters, 21.1% for tandem promoters and 17.4% for segmental promoters (Figure 8B). Thus, tandemly duplicated promoters have higher similarity than segmentally duplicated promoters. During tandem and segmental duplication, not only the transcribed regions of genes but also their promoters were duplicated.

Our expression data also showed that some of tandem or segmental gene pairs were detected with expression divergence even though their promoter regions exhibited very high similarity or with 100% homology. To explore further the underlining mechanisms, we investigated the DNA methylation status between tandemly or segmentally duplicated pairs. We determined if a promoter region was methylated according to the description [36]. For randomly selected promoter pairs, up to 43.8% of them showed methylation variation (pink column in Figure 8C). For tandem promoter pairs, the percentage is 42.5% and no significant difference was observed (green column in Figure 8C). However, for segmental promoter pairs, significantly less pairs (36.0%) showed the change in their methylation status (blue column in Figure 8C). In addition, we also analyzed genes that have undergone epimutation according to the method as described [35]. In total, around 34.4% of rice genes were detected to have epimutation (pink column in Figure 8D). However, up to 63.3% and 40.5% of tandemly and segmentally duplicated genes were detected with epimutation, respectively (green and blue columns in Figure 8D). The percentages were significantly higher than the control (34.4%). These data suggest that tandem and segmental genes might have undergone epimutation more frequently.

Discussion

Functional Bias of Genes by Tandem and Segmental Duplication and Functional Complementation

In *Arabidopsis* and rice, tandem gene density was positively correlated with the recombination rate [8,37]. This might be partially due to recombination-mediated processes being involved in tandem duplication [3]. However, our analysis showed that no correlation was observed between segmental gene distribution and recombination rate in rice. Thus, tandem and segmental duplication do not occur at random. Such a duplication mechanism might partially contribute to functional bias of tandem and segmental genes. Tandem arrays were enriched for genes that encoded proteins related to stimulus and death (Figure 4A) but under-represented for genes involved in transcription and DNA/RNA binding [8]. However, segmental duplicates were enriched for genes encoding transcription factors or regulatory proteins (Figure 4B). Thus, our data provided some evidence that shows that tandem and segmental genes might encode genes with functional complementation. The functional bias with a complementation between tandem and segmental genes might be partially determined by the duplication modes. Tandem duplication typically copies one gene each time. Thus, the evolutionarily successful tandem duplication events are most likely to target genes at the end of a pathway, or genes representing flexible steps, such as those involved in environmental response [8]. However, segmental duplication allows for multiple genes to be copied each time, which permits the retention, evolution and divergence of redundant networks.

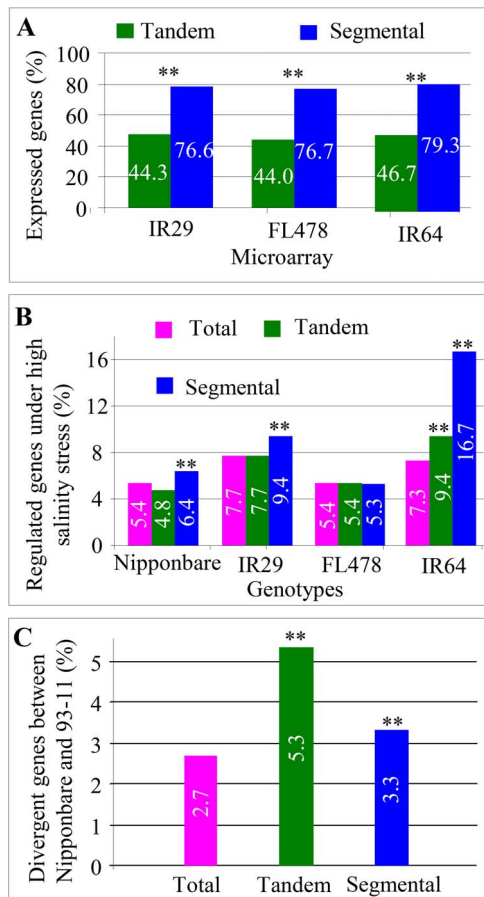


Figure 7. Expression divergence of tandem and segmental genes among different rice genotypes. (A) Summary of expressed tandem and segmental genes among three different rice lines including IR29, FL478 and IR64. (B) A comparison between tandem and segmental genes in their expression regulation under high salinity stress among four different rice genotypes including Nipponbare, IR29, FL478 and IR64. (C) Expression divergence of tandem and segmental genes between Japonica variety Nipponbare and indica variety 93-11. doi:10.1371/journal.pone.0063551.g007

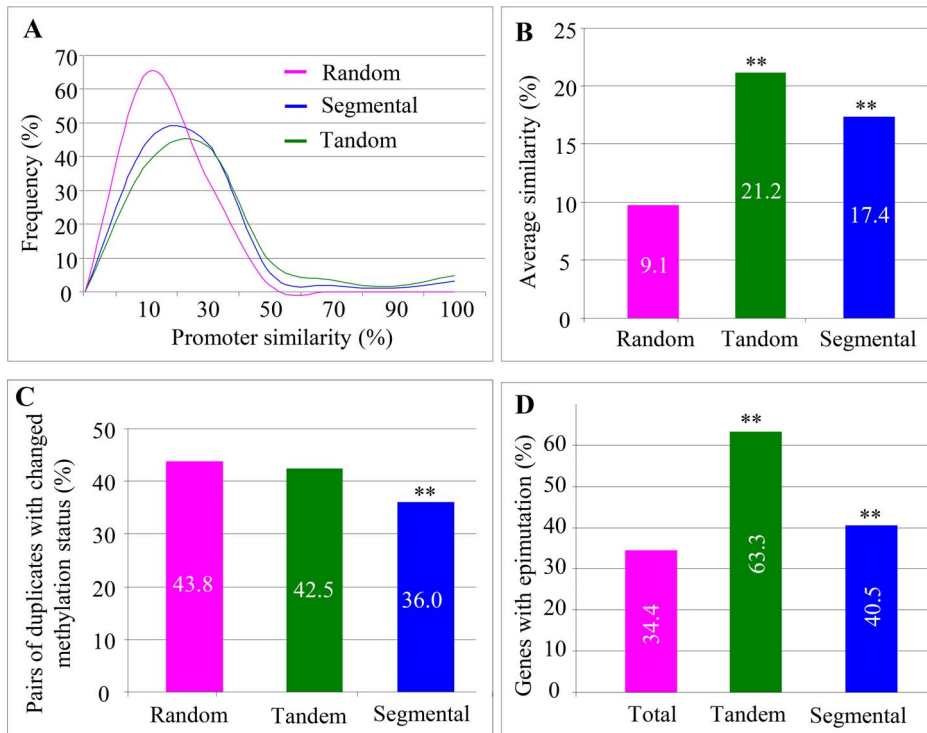


Figure 8. Sequence similarities and DNA methylation status of promoters from tandem and segmental genes. (A) Frequency distribution of tandemly and segmentally duplicated promoters by comparing with randomly selected promoters. (B) Average similarities of randomly, tandemly and segmentally selected promoters. (C) The percentages of promoters with changed DNA methylation status among random, tandem and segmental promoters. (D) A comparison in the percentages of genes with epimutation among total annotated, tandem and segmental genes.

doi:10.1371/journal.pone.0063551.g008

Selective Constraints and Mechanisms of Tandem and Segmental Genes

After tandem or segmental duplication, one copy might be silenced or evolve into a pseudogene [28]. Alternatively, both copies might survive under certain selection pressures [38]. Segmentally duplicated genes were generally subject to more stringent functional constraints with an average Ka/Ks ratio of 0.141 when compared to tandemly duplicated genes (Figure 5). C-value test further showed that more than 84% and 97% of tandem and segmental genes, respectively, were under functional constraints (Figure 5). The data provided evidence that a limited number of tandemly or segmentally duplicated genes gained novel functions. In this case, it is of interest for us to explain how duplicated genes with similar protein function could have been retained during long evolution. Our data revealed that segmentally duplicated gene pairs showed higher level of expression divergence (Figure 6). This might be due to the promoters from segmental gene pairs having lower level of similarity, thereby, higher promoter divergence when compared with tandem gene pairs (Figure 8). In addition, both tandem and segmental genes also showed the difference in DNA methylation and epimutation (Figure 8). In general, the divergence of protein sequences, transcriptional patterns and abundance as well as DNA methylation status significantly contributed to the retention and evolution of tandemly or segmentally duplicated genes. Higher rate of protein divergence was observed for tandem genes, which might contribute to the retention of tandem genes. In contrast, higher percentage of segmental genes was retained after evolving new transcriptional patterns or abundance with relatively lower protein divergence. In some cases, domain combination was observed in

some tandem or segmental genes (data not shown), which might also contribute to the retention of duplicated genes [39,40]. More expression divergence between tandem or segmental pairs would be expected if expression data from other stress conditions had been available for the analysis.

Contribution of Tandem and Segmental Duplication to Gene Family Expansion and Species Divergence

Previously, we carried out a genome-wide identification and characterization of several gene families including Lectin [16], GST [18], GRAM [40], and WRKY [41]. The family size ranges from 17 to 267 in the rice genome and both tandem and segmental duplication significantly contributed to their expansion (Table S3). Besides the families mentioned above, both duplications also contributed to the expansion of other families. For example, a total of 687 genes were identified to encode the F-box domain [42]. Based on our analysis, 247 (36%) and 62 (9%) of them were related to tandem and segmental duplication, respectively. These data suggest that both tandem and segmental duplication contribute to the gene family expansion.

During long evolution, some families exhibited lineage-specific expansion through tandem or segmental duplication. Such an expansion formed the basis for adaptive evolution and provided important sources for organizational and regulatory diversity in plants. Shiu et al. (2004) reported a two fold larger RLK/Pelle family in rice than in *Arabidopsis* and tandem duplication seems to be the major mechanism for recent expansions in rice [15]. Their data showed that most of the recent expansions have involved defense/resistance-related genes [15]. Hanada et al. (2008) reported the importance of lineage-specific expansion of plant

tandem duplication in the adaptive response to environmental stimuli [12]. Our domain analysis and GSEA data suggest a functional bias of tandem and segmental genes after duplication (Figures 3 and 4). Thus, lineage-specific expansion of some gene families followed by functional bias might significantly contribute to genome evolution and diversity.

The Ka/Ks analysis of both tandem and segmental pairs showed obvious functional constraints in their protein sequences (Figure 5). However, functional divergence of small part of tandem or segmental pairs were also observed with $Ka/Ks > 1$ (Figure 5). In addition, we also observed considerable tandem and segmental genes with expression divergence within and between genotypes (Figures 6 and 7). Other studies provided evidence that tandem or segmental genes practiced substantial divergence in the expression abundance or tissue specificity or in the response to various abiotic and biotic stresses [12,18,19,41,43,44]. On the other hand, tandem or segmental genes also exhibited the divergence in promoter sequences and methylation status (Figure 8). All these data demonstrate the contribution of tandem and segmental genes in variety domestication and species diversity.

Materials and Methods

DNA and Protein Data

The release 7 of rice pseudomolecules and protein data were downloaded from the MSU Rice Genome Annotation Project Database (<http://rice.plantbiology.msu.edu/index.shtml>) [32]. These pseudomolecules are identical to those from the International Rice Genome Sequencing Project (IRGSP, <http://rgp.dna.affrc.go.jp/IRGSP/>) or the Rice Annotation Project (RAP, <http://rapdb.dna.affrc.go.jp>). In the release 7, a total of 66,433 gene models were predicted to encode 17,314 TEs and 49,119 non-TEs. A gene may have multiple gene models due to alternative splicing. These gene models were from 55,986 genes including 16,941 loci for TEs and 39,045 loci for non-TEs.

Identification of Tandemly and Segmentally Duplicated Genes and their Distribution on the Chromosomes

A total of 49,119 non-TE peptides were used for identification of tandemly duplicated genes. Only the longest peptide was retained if multiple peptides were annotated from a same gene locus. Protein sequences were screened in an all versus all BLAST searches using BLOSUM62 matrix and an E-value < 0.01 . A pair of matching peptides were retained when the identity was $\geq 30\%$ and the alignment covered $\geq 70\%$ of the protein length. Pairs of matching proteins were clustered into groups (families) using a transitive closure algorithm: if $A = B$ and $B = C$, then $A = C$. Two genes were regarded as tandem pairs if they belonged to the same family, were located on the same chromosome and were separated by no more than 10 unrelated genes.

Segmentally duplicated genes were identified according to the description by Lin et al. (2006) using the release 7 of rice genome dataset (<http://rice.plantbiology.msu.edu/index.shtml>) [11].

To study the density of tandemly and segmentally duplicated genes, chromosome sequences were split into 1 Mb partitions. Density was calculated for each partition by calculating the percentage of number of tandem/segmental genes among total annotated genes.

Domain Analysis

All annotated proteins were submitted to the Pfam family database (<http://pfam.sanger.ac.uk/>) [45] for domain detection. We studied all domains detected in tandem or segmental proteins.

For each domain, we calculated the percentage of the domains represented in the tandem and segmental proteins or among the total proteins. We determined whether these two proportions were equivalent by Pearson's χ^2 test.

GO Annotation and Gene Set Enrichment Analysis

GO assignments for rice genes were obtained from the MSU dataset (<http://rice.plantbiology.msu.edu/index.shtml>). Three top GO categories (B, F and C) [46] were analyzed. Gene Set Enrichment Analysis (GSEA) [47] was used to determine if a GO category was over-represented in tandem or segmental genes by comparing the partition of the GO category in all annotated rice genes with nominal p-value < 0.05 and false discovery rate (FDR) < 0.25 .

Ka/Ks Analysis and C-value Test

Firstly, amino acid sequences from tandem or segmental pairs were aligned using the "water" program (Smith-Waterman local alignment of sequences, <http://emboss.bioinformatics.nl/>). The aligned sequences were then transferred to the original coding sequences using the PAL2NAL program [48]. The aligned coding sequences were used for Ka and Ks estimation by the yn00 program of the PAML4.6 package (<http://abacus.gene.ucl.ac.uk/software/paml.html>) [49]. The Ka/Ks ratios were then used to evaluate the protein divergence by testing the C-value according to the description [34].

Expression and DNA Methylation Analysis

Both massively parallel signature sequencing (MPSS) [50] and Affymetrix rice microarray data were used for expression analysis of tandem and segmental genes. The MPSS expression data were downloaded from the website http://mpss.udel.edu/rice/mpss_index.php. The Affymetrix microarray data were downloaded from the GEO dataset (<http://www.ncbi.nlm.nih.gov/geo/>) with accession numbers GSE13735, GSE14300, GSE14403, GSE17002, GSE27064, GSE28124, GSE3053, GSE4438, GSE6893, GSE6901, GSE7951. The experiments covered all available Affymetrix microarray data under different tissues or under cold, drought and high salinity stresses. A total of 11 tissues were included for the expression divergence: crown vegetative meristematic tissue, germinating seed, germinating seedlings, immature panicle, mature leaves, mature pollens, mature roots, meristematic tissue, stem, young leaves and roots. A duplicated tandem or segmental pairs was regarded as expression divergence in tissues if they showed the difference in detectable tissue numbers. For the transcriptionally detectable tissues in duplicated pairs, if they showed at least two folds difference in their expression abundance in at least one tissue with statistic analysis, they were also regarded as divergent gene pairs. A similar procedure was also applied to the detection of duplicated pairs with down- or up-regulated duplicates. A total of two biotic and three abiotic stresses were analyzed. For biotic stresses, expression data from compatible (S)/incompatible (R) bacteria and fungi were employed including *X.oryzae*-S, *X.oryzae*-R, *M. grisea*-S and *M. grisea*-R. For abiotic stress, we analyzed the expression data under the treatment with cold, drought and high salinity stresses.

For DNA methylation analysis, data sets were downloaded from the GEO dataset (<http://www.ncbi.nlm.nih.gov/geo/>) with accession numbers GSE21152 and GSE38480 and were analyzed according to their description [35,51].

Promoter Similarity Analysis

A total of 3,647 and 3,634 pairs of tandemly and segmentally duplicated promoter sequences (1 Kb upstream of starting code ATG) were aligned using the “matcher” program (Waterman-Eggert local alignment of two sequences, <http://emboss.bioinformatics.nl/>). Similar pairs of randomly selected rice promoter sequences were also aligned using the same program as a control. For each pairwise alignment, the promoter similarity was calculated as the length of the alignment divided by their total length. Statistic analysis was carried out to evaluate the promoter similarities according to the description [52].

Supporting Information

Figure S1 Tandemly or segmentally duplicated genes in each rice chromosome.

(PPT)

References

- Rubin G, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, et al. (2000) Comparative genomics of the eukaryotes. *Science* 287: 2204–2215.
- Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci USA* 100: 11484–11489.
- Freeling M (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Ann Rev Plant Biol* 60: 433–453.
- Zhao XP, Si Y, Hanson RE, Crane CF, Price HJ, et al. (1998) Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. *Genome Res* 8: 479–492.
- IHGSC (International Human Genome Sequencing Consortium) (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.
- Shaikh TH, Kurahashi H, Emanuel BS (2001) Evolutionarily conserved duplications in 22q11 mediate deletions, duplications, translocations and genomic instability. *Genet Med* 3: 6–13.
- Emanuel BS, Shaikh TH (2001) Segmental duplications: an ‘expanding’ role in genomic instability and disease. *Nat Rev Genet* 2: 791–800.
- Rizzon C, Ponger L, Gaut BS (2006) Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. *PLoS Comput Biol* 2: e115.
- Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, Van de Peer Y (2002) The hidden duplication past of *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* 99: 13627–13632.
- Cannon SB, Mitra A, Baumgarten A, Young ND, May G (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol* 4: 10.
- Lin H, Zhu W, Silva JC, Gu X, Buell CR (2006) Intron gain and loss in segmentally duplicated genes in rice. *Genome Biol* 7: R41.
- Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol* 148: 993–1003.
- Han Y, Zheng D, Vimolmangkang S, Khan MA, Beever JE, et al. (2011) Integration of physical and genetic maps in apple confirms whole-genome and segmental duplications in the apple genome. *J Exp Bot* 62: 5117–5130.
- Giannuzzi G, D'Addabbo P, Gasparro M, Martinelli M, Carelli FN, et al. (2011) Analysis of high-identity segmental duplications in the grapevine genome. *BMC Genomics* 12: 436.
- Shiu SH, Karlowski WM, Pan R, Tzeng YH, Mayer KF, et al. (2004) Comparative analysis of the receptor-like kinase family in *Arabidopsis* and rice. *Plant Cell* 16: 1220–1234.
- Jiang SY, Ma Z, Ramachandran S (2010) Evolutionary history and stress regulation of the lectin superfamily in higher plants. *BMC Evol Biol* 10: 79.
- Nuruzzaman M, Manimekalai R, Sharoni AM, Satoh K, Kondoh H, et al. (2010) Genome-wide analysis of NAC transcription factor family in rice. *Gene* 465: 30–44.
- Chi YH, Cheng YS, Vanitha J, Kumar N, Ramamoorthy R, et al. (2011) Expansion mechanisms and functional divergence of the glutathione S-transferase family in sorghum and other higher plants. *DNA Res* 18: 1–16.
- Yang HL, Liu YJ, Wang CL, Zeng QY (2012) Molecular evolution of trehalose-6-phosphate synthase (TPS) gene family in *Populus*, *Arabidopsis* and rice. *PLoS One* 7: e42438.
- Wang Y, Wang X, Tang H, Tan X, Ficklin SP, et al. (2011) Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across divergent angiosperms. *PLoS One* 6: e28150.
- Ohno S (1970) Evolution by gene duplication. Springer-Verlag, New York.
- Moore R, Purugganan MD (2005) The evolutionary dynamics of plant duplicate genes. *Curr Opin Plant Biol* 8: 122–128.
- Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 30: 1151–1155.
- Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* 16: 1679–1691.
- Li WH, Yang J, Gu X (2005) Expression divergence between duplicate genes. *Trends Genet* 21: 602–607.
- Jackson S, Chen ZJ (2010) Genomic and expression plasticity of polyploidy. *Curr Opin Plant Biol* 13: 153–159.
- Wang Y, Wang X, Paterson AH (2012) Genome and gene duplications and gene expression divergence: a view from plants. *Ann N Y Acad Sci* 1256: 1–14.
- Seoighe C, Johnston CR, Shields DC (2003) Significantly different patterns of amino acid replacement after gene duplication as compared to after speciation. *Mol Biol Evol* 20: 484–490.
- Wang XY, Paterson AH (2011) Gene conversion in angiosperm genomes with an emphasis on genes duplicated by polyploidization. *Genes* 2: 1–20.
- Pinyopich A, Ditta GS, Savidge B, Liljegren SJ, Baumann E, et al. (2003) Assessing the redundancy of MADS-box genes during carpel and ovule development. *Nature* 424: 85–88.
- Franzke A, Lysak MA, Al-Shehbaz IA, Koch MA, Mummenhoff K (2010) Cabbage family affairs: the evolutionary history of *Brassicaceae*. *Trends Plant Sci* 16: 108–116.
- Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, et al. (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res* 35: D846–D851.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. *Nat Genet* 25: 25–29.
- Thornton K, Long M (2002) Rapid divergence of gene duplicates on the *Drosophila melanogaster* X chromosome. *Mol Biol Evol* 19: 918–925.
- Chodavarapu RK, Feng S, Ding B, Simon SA, Lopez D, et al. (2012) Transcriptome and methylome interactions in rice hybrids. *Proc Natl Acad Sci USA* 109: 12040–12045.
- Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S (2007) Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet* 39: 61–69.
- Zhang L, Gaut BS (2003) Does recombination shape the distribution and evolution of tandemly arrayed genes (TAGs) in the *Arabidopsis thaliana* genome? *Genome Res* 13: 2533–2540.
- Force A, Lynch M, Pickett FB, Amores A, Yan YL, et al. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151: 1531–1545.
- Apic G, Gough J, Teichmann SA (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol* 310: 311–325.
- Jiang SY, Ramamoorthy R, Ramachandran S (2008) Comparative transcriptional profiling and evolutionary analysis of the GRAM domain family in eukaryotes. *Dev Biol* 314: 418–432.
- Ramamoorthy R, Jiang SY, Kumar N, Venkatesh P, Ramachandran S (2008) A comprehensive transcriptional profiling of the WRKY gene family in rice under various abiotic and phytohormone treatments. *Plant Cell Physiol* 49: 865–879.
- Jain M, Nijhawan A, Arora R, Agarwal P, Ray S, et al. (2007) F-box proteins in rice: genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress. *Plant Physiol* 143: 1467–1483.
- Gingerich DJ, Hanada K, Shiu SH, Vierstra RD (2007) Large-scale, lineage-specific expansion of a bric-a-brac/tramtrack/broad complex ubiquitin-ligase gene family in rice. *Plant Cell* 19: 2329–2348.
- Zou C, Lehti-Shiu MD, Thomashow M, Shiu SH (2009) Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana*. *PLoS Genet* 5: e1000581.

Table S1 Genome-wide identification of tandemly duplicated genes in the rice genome.

(XLS)

Table S2 Genome-wide identification of segmentally duplicated genes in the rice genome.

(XLS)

Table S3 Contribution of tandem and segmental duplication to the gene family expansion.

(XLS)

Author Contributions

Conceived and designed the experiments: SYJ SR. Performed the experiments: SYJ. Analyzed the data: SYJ JMG. Contributed reagents/materials/analysis tools: SYJ. Wrote the paper: SYJ SR JMG.

45. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, et al. (2002) The Pfam protein families database. *Nucleic Acids Res* 30: 276–280.
46. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, et al. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 32: D258–D261.
47. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102: 15545–15550.
48. Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* 34: W609–W612.
49. Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under evolutionary models. *Mol Biol Evol* 17: 32–43.
50. Nobuta K, Venu RC, Lu C, Belo A, Vemaraju K, et al. (2007) An expression atlas of rice mRNAs and small RNAs. *Nat Biotechnol* 25: 473–477.
51. Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, et al. (2010) Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci USA* 107: 8689–8694.
52. Haberer G, Hindemitt T, Meyers BC, Mayer KF (2004) Transcriptional similarities, dissimilarities, and conservation of *cis*-elements in duplicated genes of *Arabidopsis*. *Plant Physiol* 136: 3009–3022.