DNA Research **17**, 343–352, (2010) doi:10.1093/dnares/dsq022
Advance Access publication on September 23, 2010

# Conservation, Rearrangement, and Deletion of Gene Pairs During the Evolution of Four Grass Genomes

Nicholas Krom, and Wusirika Ramakrishna*

*Department of Biological Sciences, Michigan Technological University, Houghton, MI 49931, USA*

*To whom correspondence should be addressed. Tel. +1 906-487-3068. Fax. +1 906-487-3167.
E-mail: wusirika@mtu.edu

Edited by Satoshi Tabata
(Received 24 March 2010; accepted 1 September 2010; published online 23 September 2010)

**Abstract**

**Gene order and content differ among homologous regions of closely related genomes. Similarities in the expression profiles of physically adjacent genes suggest that the proper functioning of these genes depends on maintaining a specific position relative to each other. To better understand the results of the interaction of these two genomic forces, convergent, divergent, and tandem gene pairs in rice and sorghum, as well as their homologs in rice, sorghum, maize, and *Brachypodium* were analyzed. The status of each pair in all four species: whether it was conserved, inverted, rearranged, or missing homologs was determined. We observed that divergent gene pairs had lower rates of conservation than convergent or tandem pairs, but higher rates of rearranged pairs and missing homologs in maize than in any other species. We also discovered species-specific gene pairs in rice and sorghum. In rice, gene pairs with strongly correlated expression levels were conserved significantly more often than those with little or no correlation. We assigned three types of gene pair to one of 14 possible evolutionary history categories to uncover their evolutionary dynamics during the evolution of grass genomes.**

**Key words:** gene pairs; grass genomes; conservation; rearrangement; coexpression

## 1. Introduction

One of the primary areas of investigation in comparative genomics is the identification and characterization of homologous regions in closely related genomes. The subjects of these investigations range in scale from multi-megabase syntenic regions covering most of a chromosome to small loci containing just a few genes. Studying the syntenic regions can uncover large-scale events in the evolutionary history of a genome, such as segmental duplications or polyploidization; however, these regions in different species can differ significantly. Such variation results from a large number of genomic alterations occurring over time while maintaining sufficient collinearity to define regions of synteny. On the other hand, comparative analysis of small loci can produce detailed evolutionary histories of groups of neighboring genes and provide examples of the types of changes possible in a genome. However, it is difficult to expand these studies to a genome-wide scale due to the number of genes involved and the problem of generalizing these types of changes to allow their quantification.

In this study, we conduct an intermediate form of comparative analysis. By examining pairs of adjacent genes, we can detect changes at the level of single genes and still observe relationships between genes. Owing to the simplicity and small scale of our subjects, all changes can be assigned to a manageable number of classes, thereby producing results that are easily interpreted in genome-wide studies of this type. Our previous investigation compared gene pairs[1] in three plant species (rice, *Arabidopsis*, and *Populus*) which diverged 130–200 million years ago (mya).[2–4] This study, on the other hand, compares four members of the Poaceae family (rice, sorghum, maize, and *Brachypodium*) whose last common

ancestor dates to 50–70 mya.[2,5] Genomes of the four grass species used in this study have been sequenced.[6–9] The shorter evolutionary distances separating these species simplifies the interpretation of any observed genomic rearrangements, due to the reduced probability of multiple independent events affecting the same region. However, many small rearrangements have been identified in earlier comparative studies of Poaceae genomes,[10,11] providing sufficient variation among genomes to identify any trends regarding selection for or against disruption of ancestral gene pairs. It has been hypothesized that gene order is not entirely random, but rather is connected to gene function and regulation,[12] and that genomic rearrangements can alter the function of genes or even lead to the creation of new gene families. Thus, gene order possibly contributes to phenotypic differences between species, even when individual genes are conserved.[13]

We previously reported[1] that their strand-wise arrangement has a significant influence on many characteristics of gene pairs. For this study, we classified all pairs of adjacent genes as convergent (→ ←), divergent (← →), or tandem (→ → or ← ←), we identified homologous genes in other species and determined the status of each pair (conserved, inverted, moved, or missing homologs). We also estimated the effect of correlated expression on these types of gene-pair rearrangements. To gain an understanding of the evolutionary timing of the rearrangements we observed, a putative evolutionary history was created for each gene pair, based on its status in each of the four species. Overall, this study provides an overview of the frequencies and types of genomic rearrangements within a subset of the Poaceae, as well as many other properties of the genomes being studied.

## 2. Materials and methods

### 2.1. Identification of gene pairs

Genome sequence and annotation data were downloaded for rice (*Oryza sativa* subsp. *japonica*; http://rice.plantbiology.msu.edu, MSU rice pseudomolecules release 6), sorghum (*Sorghum bicolor*; http://www.phytozome.net/sorghum, sequence assembly v1.0, gene set v1.4), maize (*Zea mays*; http://www.maizesequence.org, release 3a.50), and *Brachypodium* (*Brachypodium distachyon*; http://www.brachypodium.org, 8X coverage release). A second set of rice sequence and annotation data was obtained from the Rice Annotation Project (RAP; http://rapdb.dna.affrc.go.jp/, sequence build 5) for comparison with the MSU rice gene set and was processed and analyzed using the same methods as the MSU rice and sorghum gene sets. Annotated genes in the rice and sorghum genomes were sorted by chromosome and position and then, based on which strand the gene is transcribed from, all pairs of adjacent genes were classified as either convergent (→ ←), divergent (← →), or tandem (→ → or ← ←) pairs. Pairs containing hypothetical- or transposon-related genes, as determined by annotation and RepeatMasker (www.repeatmasker.org; 50% or greater transposon content of unspliced sequence), were excluded from all analyses.

### 2.2. Comparative sequence analysis

The coding region sequences of all rice and sorghum gene pairs were aligned with the genome assemblies of the other three species using BLASTN. For each gene, individual hits (presumably corresponding to single exons) with *e*-values of 0.00001 or less were grouped with other nearby hits on the same strand and contig to produce a putative homolog. The locations of each pair's homologs were then used to determine the pair's status in that species. Pairs were considered 'conserved' if both genes had homologs in the original strand-wise arrangement (convergent, divergent, or tandem) within 50 kb of each other and had no other genes inserted between them. Fifty kilobases were chosen to allow for insertions of repetitive DNA such as nested retrotransposons. 'Inverted' pairs also possessed homologs within the cutoff distance, but had different strand-wise arrangements than the original pair. Pairs were considered 'rearranged' if homologs for both genes were found but were >50 kb apart, separated by other genes, or located on different contigs. Those pairs in which one or both genes were missing homologs in a given species were also identified. In addition to 50 kb, genes residing within 20 and 75 kb were used to perform the above analysis.

### 2.3. Expression analysis

Two types of quantitative expression data were collected for all rice genes (MSU data): microarray and massively parallel signature sequencing (MPSS). First, MPSS[14] data were downloaded from the Rice MPSS Database (http://mpss.udel.edu/rice/). Only 17-bp signatures of classes 1, 2, 5, and 7 that mapped to a single gene were used, and abundance values <5 were ignored as background interference. When multiple signatures had significant abundance values in the same library, their average abundance was used. Correlated expression between genes in convergent, divergent, and tandem pairs was examined by calculating the Pearson correlation coefficient using each gene's average abundance values across 72 libraries.

Microarray data were downloaded from the Yale rice project (http://bioinformatics.med.yale.edu/rc/overview.jspx) for a total of 446 hybridizations. Correlated expression was again tested with the Pearson correlation coefficient, this time pairing data points for each gene from the same hybridization and channel, where data were available for both genes for that hybridization and channel.

Correlation coefficients were calculated separately for each expression data type. Thus, if expression data were present for both genes in both sets, a gene pair would have both an MPSS-derived and microarray-derived correlation coefficient. For those pairs with both coefficients, the pair was considered strongly correlated if at least one was greater than the cut-off value.

The fraction of strongly correlated gene pairs falling into each of the four categories (conserved, inverted, rearranged, or missing homologs) was then compared with the fractions of uncorrelated or weakly correlated pairs in those same categories. The statistical significance of their differences was evaluated using the normal approximation of the binomial test, with a significance level of $P < 0.05$ ($Z > 1.6449$). Two definitions of 'strong' correlation were used: the first one using a correlation coefficient cut-off of 0.5 used in other studies,[1,15] and the second one based on a $P < 0.05$ significance level for a normal distribution with a mean of 0.187 and a standard deviation of 0.247, the sample statistics from the set of all calculated correlation coefficients, giving a minimum 'strong' correlation of 0.594.

### 2.4. *Evolutionary analysis of gene pairs*

The evolutionary history of each gene pair was constructed by comparing the status of the pair in each of the four species in this study. The likelihood of a given scenario was based on the number of gene rearrangements, deletions, and conservation using the fewest possible changes to the homologous regions of the last common ancestor to arrive at the present state. Gene pairs were then assigned to one of 14 groups based on their putative histories that could produce the observed results of the comparative analysis. Rice MSU data were used for this analysis.

## 3. Results and discussion

### 3.1. *Conservation and rearrangements*

Convergent, divergent, and tandem gene pairs in the rice and sorghum genomes were identified as described in methods. Four thousand and eight hundred convergent pairs, 3711 divergent pairs, and 9428 tandem pairs from the MSU rice gene set and 5059 convergent, 4913 divergent, and 11847 tandem pairs from the sorghum gene set were identified.

The primary goal of this analysis was to determine how frequently the exact arrangement of a pair of adjacent genes is conserved in the genomes of other grass species and what changes have taken place when the pair is not conserved. Out of the four grass species selected for this study, rice and sorghum were chosen as starting points for comparisons because their sequence and annotation data sets were considerably better than those of maize and *Brachypodium*. Our comparative sequence analysis placed each rice or sorghum gene pair into one of four categories based on the presence or the absence of homologous genes and their locations in the genome. A pair of adjacent genes in rice/sorghum was considered to be conserved in other genomes when both the individual sequences of genes and the strand-wise arrangement of the pair were conserved. If a pair's homologs were found to be still adjacent but with a different strand-wise arrangement, then the pair was designated 'inverted'. Homologs falling on different contigs, separated by other genes, or >50 kb apart were considered 'rearranged'. Finally, one or both genes in the original pair may be lacking homologs because they were deleted in that species' lineage or they arose in the ancestors of rice or sorghum after diverging from their last common ancestor. Together, these categories include all the major events of genomic evolution at this scale, and the relative frequencies of these events provide insight into the importance of proximity and strand-wise arrangement to proper gene function and regulation. To estimate the statistical significance of the variation observed among pair types, we compared the fraction of pairs of each type that fell into each category with the corresponding fraction of all pairs in that category.

Conservation rates for both rice and sorghum gene pairs (Figs 1 and 2, Supplementary Tables S1 and S2) followed quite similar patterns. In every comparison, divergent gene pairs were conserved least often, with conservation frequencies significantly below the all pair average. Similarly, convergent pairs were conserved significantly more often than the all pair average. Tandem pair conservation rates were generally close to the all pair average. The conservation frequency differed a great deal among pair types in maize, where divergent pairs are conserved at rates close to half that of convergent or tandem pairs. Further, in *Brachypodium*, rice gene pairs are more frequently conserved than sorghum pairs. Although *Brachypodium* appears to be closer to rice than sorghum based on the divergence times of 40–53 Myr for *Brachypodium*–rice and 45–60 Myr for *Brachypodium*–sorghum based on synonymous

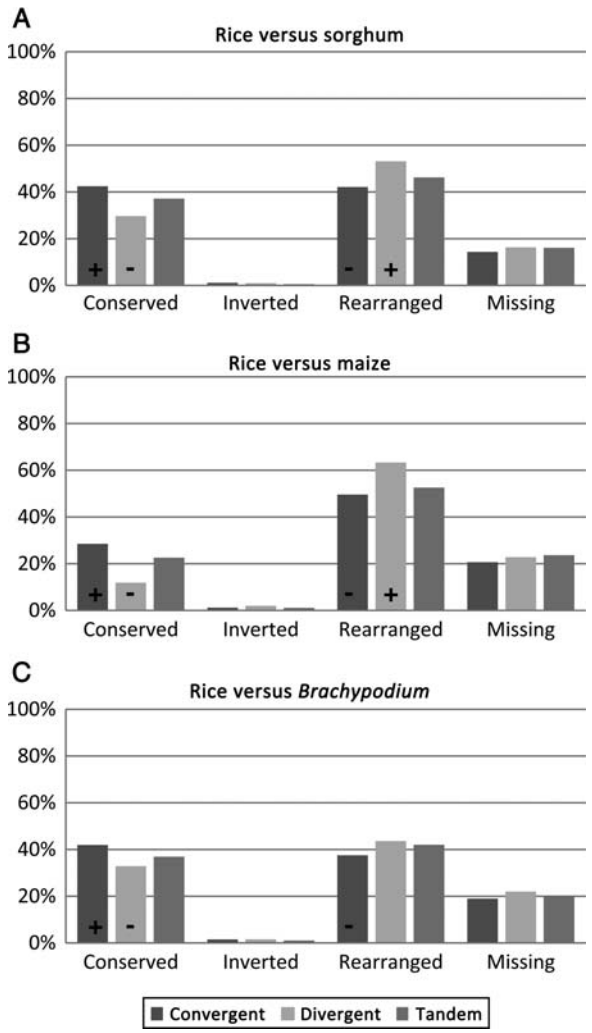


**Figure 1.** Conservation and rearrangement of rice gene pairs (MSU data) in sorghum, maize, and *Brachypodium*. A '+' or '−' at the base of a column indicates that pairs in that class are significantly ($P < 0.01$) overrepresented or underrepresented, respectively, compared with the general population of gene pairs using *Z*-test.

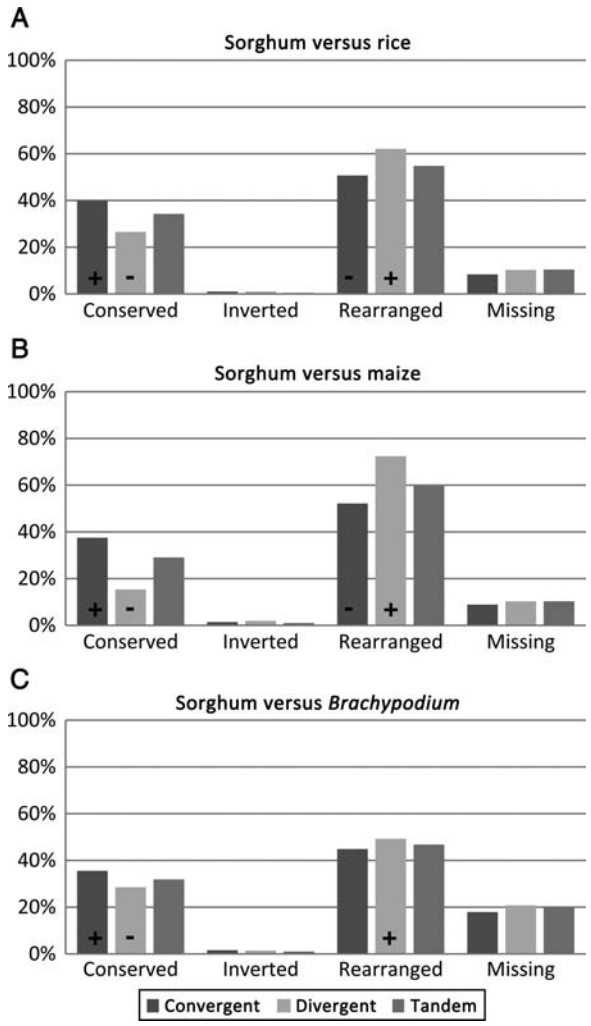


**Figure 2.** Conservation and rearrangement of sorghum gene pairs in rice (MSU data), maize, and *Brachypodium*. A '+' or '−' at the base of a column indicates that pairs in that class are significantly ($P < 0.01$) overrepresented or underrepresented, respectively, compared with the general population of gene pairs using *Z*-test.

substitution rates (Ks) of orthologous gene pairs,[9] these evolutionary relationships are not well resolved. Likewise, maize is more closely related to sorghum than to rice[11,16]; thus, sorghum pairs have higher rates of conservation in maize than do rice pairs.

Inversion of one or both genes was quite a rare phenomenon, ranging from 0.5% of sorghum tandem pairs inverted in rice (Fig. 1A) to 2% of sorghum divergent pairs in maize (Fig. 2B). We observed some variation among different pair types and comparison species, but the variation was not statistically significant.

Among rearranged pairs, rice and sorghum gene pairs differ more than conserved and inverted pairs. Overall, sorghum pairs are more likely to be rearranged than rice pairs, with the rearrangement rates of sorghum pairs being ~10–20% higher than those of rice pairs (Figs 1 and 2). For both rice and sorghum pairs, rearrangement is most common in maize (Figs 1B and 2B) by a substantial margin and is least common in *Brachypodium* (Figs 1C and 2C). Divergent pairs are consistently the most commonly rearranged type, whereas convergent and tandem pairs are generally rearranged with similar frequencies in all comparison species.

Considerable differences between rice and sorghum were also noted among gene pairs lacking homologs for one or both genes. Rice pairs missing homologs in sorghum are roughly one-half more common than sorghum pairs missing homologs in rice (15.7% versus 9.9%). In maize, rice genes are missing homologs more than twice as often as sorghum genes (22.7% versus 10%). The fractions of rice and sorghum pairs without homologs in

*Brachypodium* are almost equal. Pair type does not seem to affect the presence or the absence of homologs because none of the pair types in any comparison species deviated significantly from the all pair average.

Conservation and rearrangement rates differ between maize and the other comparison species most likely as a result of three primary sources. First, the genomic sequence of maize used in our study is in the form of individual BAC sequences, most of which are shorter than ~250 kb, rather than in the form of assembled sequences of near-chromosome length. Smaller sequences are, of course, less likely to contain complete gene pairs, especially if their intergenic regions have accumulated other genes or transposons over time. Second, transposons make up a much larger fraction of the maize genome than that of rice, sorghum, or *Brachypodium* genomes. In addition to physically disrupting the region into which they insert themselves, transposons can also increase the likelihood of recombinations, deletions, and other alterations in any area they inhabit. Third, the ancestors of maize quite likely suffered large-scale gene loss.[16] If the first gene in a pair is deleted from one copy and the second gene was deleted in the other copy of the pair, both genes would still exist in the genome, but would no longer be paired. This type of occurrence would explain why more rice and sorghum gene pairs had more physically distant homologs in maize than in any other comparison species. All these factors would reduce the frequency of gene-pair conservation and correspondingly increase rearranged pairs, as we observed.

In addition to 50 kb, genes residing within 20 and 75 kb were used to perform the above analysis. Although the number of gene pairs varied, the results showed same trends that were observed with 50 kb limit (Supplementary Tables S3 and S4).

Comparative analysis was performed using a second set of rice genome annotation data, from the RAP (http://rapdb.dna.affrc.go.jp),[17] to evaluate the effect of different methods of genome annotation on our results. We found that the analysis of the MSU rice gene set as well as the RAP gene set exhibited similar major trends regarding differences among the three pair types and comparison species. Although many of the same conclusions about conservation and rearrangement of gene pairs could be inferred from both the MSU and RAP results, significant differences exist between the two data sets (Table 1 and Supplementary Table S5). In sorghum, gene pairs identified using the RAP annotation data were conserved significantly less frequently than the MSU rice gene pairs, whereas the fractions of rearranged pairs were significantly higher among all pair types in sorghum and *Brachypodium*, and among tandem gene pairs in maize. Last, pairs missing homologs were less common overall when using the RAP annotation data, with the largest differences being found in maize and *Brachypodium*. However, comparing Fig. 1, Supplementary Table S1, and Table 1, nine statistically significant comparisons of conserved and rearranged gene pairs were common when MSU and RAP gene sets were used as starting points. The remaining two statistically significant comparisons (rice convergent pairs conserved and rearranged in *Brachypodium*) using the MSU gene set were close to being significant using the RAP gene set. These variations are most likely due to

**Table 1.** RAP rice gene-pair conservation and rearrangement

| | Pairs | Conserved | *Z*-value | Inverted | *Z*-value | Rearranged | *Z*-value | Missing homologs | *Z*-value |
|---|---|---|---|---|---|---|---|---|---|
| Rice versus sorghum | | | | | | | | | |
| Convergent | 3641 | 1169 (32.1%) | **4.91** | 112 (3.1%) | 0.78 | 1875 (51.5%) | **−6.50** | 485 (13.3%) | −0.31 |
| Divergent | 3376 | 648 (19.2%) | **−4.01** | 119 (3.5%) | 1.02 | 2160 (64.0%) | **4.82** | 449 (13.3%) | −0.31 |
| Tandem | 7582 | 1887 (24.9%) | −0.51 | 39 (0.5%) | −1.12 | 4580 (60.4%) | 1.95 | 1076 (14.2%) | 0.37 |
| Total | 14 599 | 3704 (25.4%) | | 270 (1.8%) | | 8615 (59.0%) | | 2010 (13.8%) | |
| Rice versus maize | | | | | | | | | |
| Convergent | 3641 | 987 (27.1%) | **3.96** | 103 (2.8%) | 0.32 | 1880 (51.6%) | **−4.92** | 671 (18.4%) | −0.31 |
| Divergent | 3376 | 472 (14.0%) | **−4.71** | 146 (4.3%) | 1.20 | 2137 (63.3%) | **5.75** | 621 (18.4%) | −0.33 |
| Tandem | 7582 | 1678 (22.1%) | 0.62 | 92 (1.2%) | −0.95 | 4347 (57.3%) | 0.04 | 1465 (19.3%) | 0.41 |
| Total | 14 599 | 3137 (21.5%) | | 341 (2.3%) | | 8364 (57.3%) | | 2757 (18.9%) | |
| Rice versus *Brachypodium* | | | | | | | | | |
| Convergent | 3641 | 1429 (39.2%) | 1.82 | 117 (3.2%) | 0.50 | 1569 (43.1%) | −2.25 | 526 (14.4%) | −0.23 |
| Divergent | 3376 | 1128 (33.4%) | **−2.48** | 154 (4.6%) | 1.29 | 1587 (47.0%) | 0.88 | 507 (15.0%) | 0.14 |
| Tandem | 7582 | 2830 (37.3%) | 0.47 | 81 (1.1%) | −1.17 | 3541 (46.7%) | 0.96 | 1130 (14.9%) | 0.10 |
| Total | 14 599 | 5387 (36.9%) | | 352 (2.4%) | | 6697 (45.9%) | | 2163 (14.8%) | |

Test used is the binomial test (normal approximation) with cutoff of $P < 0.01$ ($Z < 2.3267$) shown in bold.

differences in RAP and MSU annotations. The MSU gene set contains considerably more genes than the RAP set (57 840 versus 34 780), a distinction that remains even after excluding hypothetical- and transposon-related genes (29 686 versus 22 308). Therefore, it is likely that either the MSU set contains a large number of incorrectly predicted genes or the RAP set is missing a similar number of real genes. The lower rates of missing homologs among the RAP set suggests that at least some of the genes found only in the MSU set are either false positives or unique to rice. These unique genes could also be low copy number transposable elements which are transcribed but not annotated as transposable elements.[18]

### 3.2. *Effects of correlated expression on rearrangements*

The Pearson correlation coefficient of all rice gene pairs was calculated using microarray and MPSS expression data. Those pairs with coefficients of 0.5 or greater were considered significantly correlated as described earlier.[1] The full set of rice gene pairs (MSU data set) was divided into correlated and uncorrelated sets, and difference in the frequencies of each type of rearrangement within these sets was tested for significance using the binomial test. The purpose of this test was to determine whether gene pairs with correlated expression levels were subject to any of the various types of rearrangements at a significantly different rate than uncorrelated pairs.

Correlated divergent and tandem gene pairs were more common among conserved pairs in all three species (Table 2), although the increase in conservation was statistically significant only for tandem pairs in sorghum and for divergent and tandem pairs in *Brachypodium*. The difference in the conservation rates between correlated and uncorrelated pairs was highest for tandem pairs, followed closely by divergent pairs, whereas the effect of correlation on convergent pairs was considerably weaker. Conservation of coexpressed tandem gene pairs could be due to tandemly duplicated genes which are likely to have similar expression patterns. Conservation of coexpressed divergent gene pairs could be due to bidirectional promoters regulating these gene pairs.[19]

The small sample size used in the examination of the effect of correlated expression on gene-pair inversion reduced the effectiveness of the binomial test. In contrast, correlated expression in rice gene pairs appears to select against the disruption of a gene pair's physical arrangement in the other three grass genomes, with tandem pairs being significantly underrepresented in all three comparison species among pairs whose homologs are physically distant (Table 2). Divergent pairs showing correlated expression are also underrepresented in all three species, although the difference is not statistically significant. Correlated convergent pairs showed no significant difference compared with non-correlated pairs.

Rice-correlated convergent pairs were strongly underrepresented among pairs lacking homologs in sorghum and maize.

A second analysis was performed, this time using a statistically significant (see 'Materials and methods' section for details) cut-off value for 'strong' correlation of $R > 0.594$ (Supplementary Table S6). The greatest difference between this analysis and the first one using a cut-off of $R > 0.5$ is the number of strongly correlated pairs. The reduced sample size substantially affects the ability of the binomial test in determining significant variation between correlated and non-correlated pairs. As a result, statistical significance was observed only in convergent gene pairs missing homologs in maize, which were significantly underrepresented among strongly correlated pairs. Otherwise, most trends noted in the first analysis were also observed in the second analysis, with modest increases in conservation and inversion frequency and decreases in rearrangement frequency. The fractions of pairs missing homologs, either unchanged or reduced with the $R > 0.5$ definition of strong correlation, were more often increased with the alternative definition. These differences most likely result from the smaller sample size and its effect on the statistical test rather than any real biological differences between the set of gene pairs with $R > 0.5$ and those with $R > 0.594$.

Rice gene pairs displaying strongly correlated expression levels were more likely to be conserved in sorghum, maize, and *Brachypodium*. These results lend further support to the hypothesis that the strand-wise arrangement of pairs of adjacent genes may be essential to the regulatory schemes of some strongly correlated gene pairs, such that rearrangements disturbing the pair would be selected against. Correlated expression levels have also been found to increase the likelihood of conservation among fungi.[20] The largest increases in the frequency of conservation as a result of correlated expression was observed among divergent and tandem pairs, a pattern that has been observed before in a comparison of human, mouse, and chicken gene clusters.[21,22] Therefore, correlated expression levels increasing the likelihood of conservation appears to be a universal phenomenon in eukaryotes including plant genomes.

### 3.3. *Evolutionary history of gene pairs*

The estimated evolutionary history of each rice or sorghum gene pair was arrived at by examining the status of each gene pair in its three comparison species. For this analysis, a pair could be in one of

**Table 2.** Conservation and rearrangement of correlated rice gene pairs with $r > 0.5$

| | | Total | Conserved | | | Inverted | | | Rearranged | | | Missing homologs | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Number | Percentage | *Z*-value | Number | Percentage | Z-value | Number | Percentage | Z-value | Number | Percentage | Z-value |
| Rice versus sorghum | | | | | | | | | | | | | | |
| Convergent | Correlated | 329 | 153 | 46.5 | 1.61 | 7 | 2.1 | 1.91 | 135 | 41.0 | −0.42 | 34 | 10.3 | **−2.21** |
| | Uncorrelated | 4471 | 1883 | 42.1 | | 47 | 1.1 | | 1886 | 42.2 | | 655 | 14.6 | |
| Divergent | Correlated | 296 | 97 | 32.8 | 1.28 | 4 | 1.4 | 0.80 | 153 | 51.7 | −0.53 | 42 | 14.2 | −1.07 |
| | Uncorrelated | 3415 | 1003 | 29.4 | | 31 | 0.9 | | 1818 | 53.2 | | 563 | 16.5 | |
| Tandem | Correlated | 651 | 276 | 42.4 | **2.98** | 4 | 0.6 | 0.23 | 263 | 40.4 | **−3.21** | 108 | 16.6 | 0.40 |
| | Uncorrelated | 8777 | 3227 | 36.8 | | 48 | 0.5 | | 4096 | 46.7 | | 1406 | 16.0 | |
| Rice versus maize | | | | | | | | | | | | | | |
| Convergent | Correlated | 329 | 101 | 30.7 | 0.94 | 5 | 1.5 | 0.56 | 171 | 52.0 | 0.92 | 52 | 15.8 | **−2.32** |
| | Uncorrelated | 4471 | 1268 | 28.4 | | 53 | 1.2 | | 2210 | 49.4 | | 940 | 21.0 | |
| Divergent | Correlated | 296 | 42 | 14.2 | 1.34 | 4 | 1.4 | −0.85 | 181 | 61.1 | −0.85 | 69 | 23.3 | 0.23 |
| | Uncorrelated | 3415 | 399 | 11.7 | | 70 | 2.0 | | 2169 | 63.5 | | 777 | 22.8 | |
| Tandem | Correlated | 651 | 166 | 25.5 | 1.90 | 6 | 0.9 | −0.52 | 316 | 48.5 | **−2.22** | 163 | 25.0 | 0.87 |
| | Uncorrelated | 8777 | 1966 | 22.4 | | 100 | 1.1 | | 4641 | 52.9 | | 2070 | 23.6 | |
| Rice versus Brachypodium | | | | | | | | | | | | | | |
| Convergent | Correlated | 329 | 141 | 42.9 | 0.35 | 4 | 1.2 | −0.48 | 124 | 37.7 | 0.06 | 60 | 18.2 | −0.37 |
| | Uncorrelated | 4471 | 1873 | 41.9 | | 69 | 1.5 | | 1678 | 37.5 | | 851 | 19.0 | |
| Divergent | Correlated | 296 | 112 | 37.8 | **1.98** | 5 | 1.7 | 0.19 | 118 | 39.9 | −1.41 | 61 | 20.6 | −0.61 |
| | Uncorrelated | 3415 | 1108 | 32.4 | | 53 | 1.6 | | 1500 | 43.9 | | 754 | 22.1 | |
| Tandem | Correlated | 651 | 266 | 40.9 | **2.23** | 8 | 1.2 | 0.45 | 247 | 37.9 | **−2.28** | 130 | 20.0 | 0.01 |
| | Uncorrelated | 8777 | 3217 | 36.7 | | 92 | 1.0 | | 3717 | 42.3 | | 1751 | 19.9 | |

Values in the 'Z' columns are test statistics of the binomial test. Values in bold denote significant differences ($P < 0.05$) in the frequency of strongly correlated pairs in each category compared with the frequency of uncorrelated pairs.

three states in each species: conserved, rearranged (physically distant homologs, any inversion, and insertions), or deleted (one or both homologs nonexistent). On the basis of the possible combinations of these states, 14 categories of evolutionary history were devised. The putative evolutionary tree for the four species consisted of two branches, one with rice and *Brachypodium*, the other with sorghum and maize. Similarities within branches, as well as differences between them, served as the basis for many of the 14 categories.

The first category consisted of those pairs whose exact arrangement was shared in all four species (Table 3, four species, and Fig. 3A). Results varied little between the rice- and sorghum-based analyses. Convergent pairs were the most common in this class, with over 18% of pairs falling into this category, followed by tandem pairs (~10%) and divergent pairs (~6%).

Pairs conserved in two of their three comparison species most likely underwent a single species-specific rearrangement or deletion (Table 3 three species, and Fig. 3B and C). The former event was by far the most common, comprising 16–19% of all pairs, compared with the ~1% or less of pairs with one or both homologs deleted in a single species. Rice and sorghum results differed by less than one percentage point across all pair types.

The six categories comprise pairs conserved in only one other species were divided into two groups (Table 3, two species), those in which the pair was conserved within one branch of the evolutionary tree (i.e. a rice pair conserved in *Brachypodium*), referred to here as a 'branch-specific' pair (Fig. 3D–F), and those in which the pair was conserved in one species in each branch, a state referred to as 'cross-branch conservation' (Fig. 3G–I). Among both branch-specific and cross-branch conserved pairs, it was far more common (7–12% of pairs in rice and 4–15% in sorghum) for the pair to be rearranged in the other two species than for it to be deleted in one (0.5–1.9%) or both species (0.1–1.4%). In rice, branch-specific pairs were slightly more common than cross-branch conserved pairs; in sorghum, the opposite was true. There were only two sets of genes in which rice and sorghum differed substantially. The first was branch-specific divergent pairs with the

**Table 3.** Evolutionary history of rice and sorghum gene pairs

| | Rice | | | Sorghum | | |
|---|---|---|---|---|---|---|
| | Convergent | Divergent | Tandem | Convergent | Divergent | Tandem |
| Pair with conserved orientation in four species | 874 (18.2%) | 207 (5.6%) | 934 (9.9%) | 921 (18.2%) | 258 (5.3%) | 1222 (10.3%) |
| Pair with conserved orientation in three species | | | | | | |
| One rearrangement | 878 (18.3%) | 591 (15.9%) | 1827 (19.4%) | 932 (18.4%) | 800 (16.3%) | 2245 (18.9%) |
| One deletion | 64 (1.3%) | 33 (0.9%) | 151 (1.6%) | 56 (1.1%) | 28 (0.6%) | 158 (1.3%) |
| Pair with conserved orientation in two species | | | | | | |
| Branch-specific pairs | | | | | | |
| Others rearranged | 396 (8.3%) | 434 (11.7%) | 945 (10.0%) | 408 (8.1%) | 215 (4.4%) | 840 (7.1%) |
| Others deleted | 53 (1.1%) | 30 (0.8%) | 129 (1.4%) | 20 (0.4%) | 14 (0.3%) | 101 (0.9%) |
| One deletion, one rearrangement | 32 (0.7%) | 25 (0.7%) | 78 (0.8%) | 45 (0.9%) | 24 (0.5%) | 110 (0.9%) |
| Cross-branch conserved pairs | | | | | | |
| Others rearranged | 340 (7.1%) | 346 (9.3%) | 974 (10.3%) | 438 (8.7%) | 718 (14.6%) | 1481 (12.5%) |
| Others deleted | 24 (0.5%) | 12 (0.3%) | 71 (0.8%) | 10 (0.2%) | 4 (0.1%) | 51 (0.4%) |
| One deletion, one rearrangement | 68 (1.4%) | 45 (1.2%) | 163 (1.7%) | 55 (1.1%) | 57 (1.2%) | 225 (1.9%) |
| Pair with unique orientation in one species | | | | | | |
| Species-specific gene(s) | 505 (10.5%) | 475 (12.8%) | 1130 (12.0%) | 197 (3.9%) | 218 (4.4%) | 529 (4.5%) |
| Common genes, species-specific pair | 1037 (21.6%) | 1047 (28.2%) | 2038 (21.6%) | 1293 (25.6%) | 1730 (35.2%) | 3189 (26.9%) |
| Branch-specific genes, species-specific pair | 73 (1.5%) | 51 (1.4%) | 170 (1.8%) | 158 (3.1%) | 219 (4.5%) | 457 (3.9%) |
| Two rearrangements, one deletion, mixed | 299 (6.2%) | 278 (7.5%) | 530 (5.6%) | 395 (7.8%) | 460 (9.4%) | 913 (7.7%) |
| One rearrangement, two deletions, mixed | 157 (3.3%) | 137 (3.7%) | 288 (3.1%) | 131 (2.6%) | 168 (3.4%) | 326 (2.8%) |

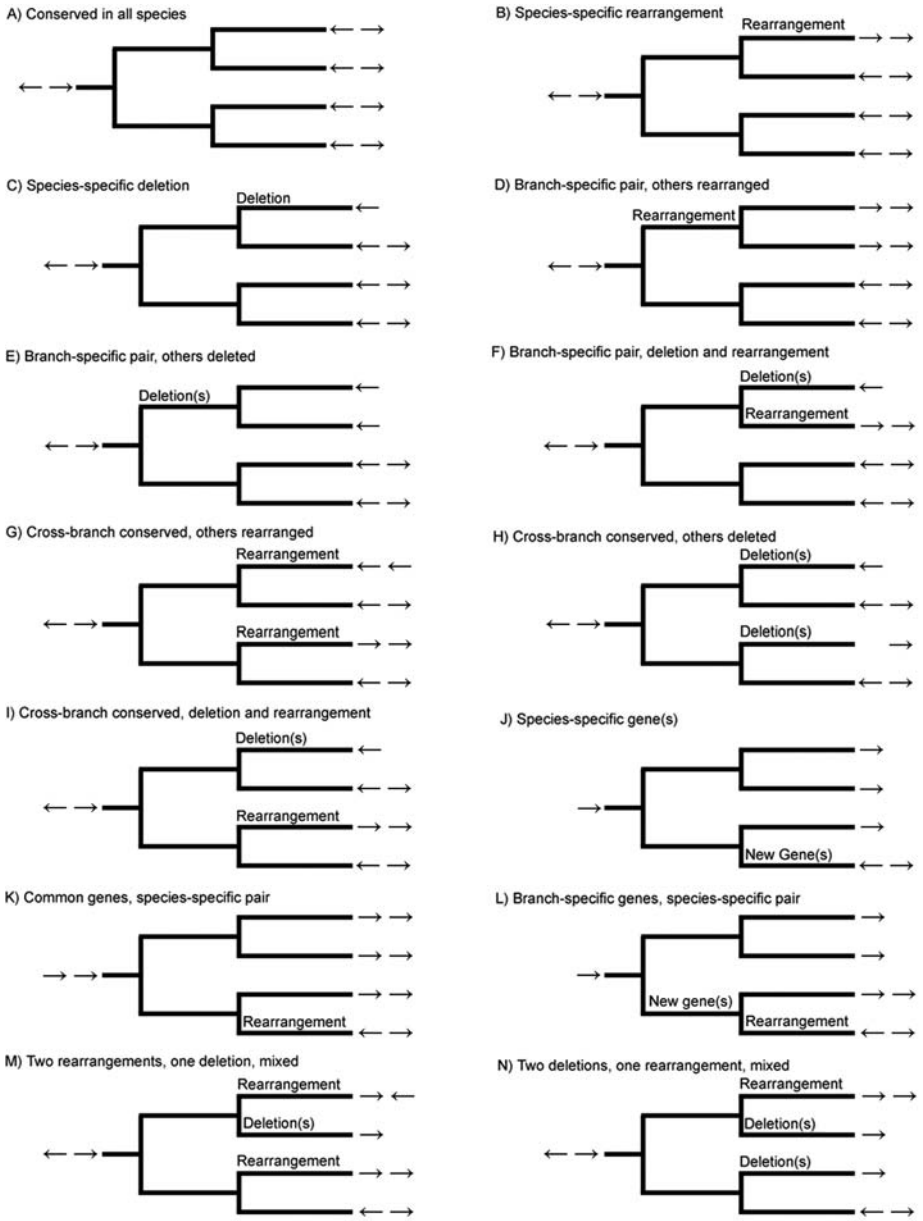


**Figure 3.** Categories of gene-pair evolution. Each image is a representative of the many specific scenarios that may be found in that category. The bottom branch of each tree represents the species in which the pair was first identified (i.e. either rice or sorghum), and the two genes in question are shown in a divergent pair in these examples. Rearrangements are represented by the inversion of one gene, inversion of both genes, insertions within the pair, or translocation to other regions or chromosomes. Likewise, deletions may involve one gene, as shown, or both genes in the pair. In some of the scenarios where the pair is conserved in two species (D, E, G, and H), the rearranged or deleted states are just as likely to be the ancestral state as the divergent pair shown. In scenario L, it is also possible that both genes existed in the common ancestor and a deletion took place in the top branch rather than new gene(s) being created.

pair being rearranged in the other two species, including 11.7% of rice divergent pairs but only 4.4% of sorghum pairs (Table 3, two species). The opposite situation was observed among cross-branch conserved divergent pairs, again with two rearrangements. These pairs made up 9.3% of rice divergent pairs, compared with 14.6% of sorghum pairs.

The last five categories consist of pairs with unique orientation in only one species (Table 3, one species). Pairs whose genes exist in all four species but whose pair-wise arrangement is found in a single species (Fig. 3K) were the most common category in both species (25–35% in sorghum and 21–28% in rice). Divergent pairs fell into this category substantially more often than convergent or tandem pairs in both rice and sorghum. In rice, the second most common category (10–12% of all pairs) is those pairs containing one or more genes unique to that

species (Fig. 3J); in sorghum, this category is approximately one-third that of rice. These observations are most likely due to differences in the gene annotation methods used by the two genome projects and/or a larger number of genes unique to rice than to sorghum reflecting their biology. This conclusion is supported by RAP data which identified 56, 28, and 172 species-specific genes that are part of convergent, divergent, and tandem gene pairs, respectively, in rice but not the other three grass genomes, respectively, compared with 505, 475, and 1130 species-specific rice genes in MSU data that are part of convergent, divergent, and tandem gene pairs, respectively. A few other studies have also identified several genes and gene families specific to rice and sorghum.[7,23] The third category (Fig. 3L) for all pair types were more than 2-fold in sorghum compared with rice (Table 3, one species). The distribution of pairs among the remaining two categories (Fig. 3M and N) showed little variation both between rice and sorghum and between pair types.

Overall, our study provides valuable insights into conservation and rearrangement of gene pairs during the evolution of the grasses serving as basis for future investigations on functional interactions between adjacent genes.

**Supplementary data:** Supplementary data are available at www.dnaresearch.oxfordjournals.org.

## Funding

This work was supported by the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service (2007-35301-18036).

## References

1. Krom, N. and Ramakrishna, W. 2008, Comparative analysis of divergent and convergent gene pairs and their expression patterns in rice, *Arabidopsis*, and *Populus*, *Plant Physiol.*, **147**, 1763–73.
2. Wolfe, K.H., Gouy, M., Yang, Y.W., Sharp, P.M. and Li, W.H. 1989, Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data, *Proc. Natl Acad. Sci. USA*, **86**, 6201–5.
3. Chaw, S.M., Chang, C.C., Chen, H.L. and Li, W.H. 2004, Dating the monocot–dicot divergence and the origin of core eudicots using whole chloroplast genomes, *J. Mol. Evol.*, **58**, 424–41.
4. Tuskan, G.A., Difazio, S., Jansson, S., et al. 2006, The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray), *Science*, **313**, 1596–604.
5. Buell, C.R. 2009, Poaceae genomes: going from unattainable to becoming a model clade for comparative plant genomics, *Plant Physiol.*, **149**, 111–6.
6. International Rice Genome Sequencing Project (IRGSP). 2005, The map-based sequence of the rice genome, *Nature*, **436**, 793–800.
7. Paterson, A.H., Bowers, J.E., Bruggmann, R., et al. 2009, The *Sorghum bicolor* genome and the diversification of grasses, *Nature*, **457**, 551–6.
8. Schnable, P.S., Ware, D., Fulton, R.S., et al. 2009, The B73 maize genome: complexity, diversity, and dynamics, *Science*, **326**, 1112–5.
9. International Brachypodium Initiative (IBI). 2010, Genome sequencing and analysis of the model grass *Brachypodium distachyon*, *Nature*, **463**, 763–8.
10. Bennetzen, J.L. and Ramakrishna, W. 2002, Numerous small rearrangements of gene content, order and orientation differentiate grass genomes, *Plant Mol. Biol.*, **48**, 821–7.
11. Ilic, K., Sanmiguel, P.J. and Bennetzen, J.L. 2003, A complex history of rearrangement in an orthologous region of the maize, sorghum and rice genomes, *Proc. Natl Acad. Sci. USA*, **100**, 12265–70.
12. Hurst, L.D., Pal, C. and Lercher, M.J. 2004, The evolutionary dynamics of eukaryotic gene order, *Nat. Rev. Genet.*, **5**, 299–310.
13. Ciccarelli, F.D., Von Mering, C., Suyuma, M., Harrington, E.D., Izaurralde, E. and Bork, P. 2005, Complex genomic rearrangements lead to novel primate gene function, *Genome Res.*, **15**, 343–51.
14. Meyers, B.C., Lee, D.K., Vu, T.H., et al. 2004, *Arabidopsis* MPSS. An online resource for quantitative expression analysis, *Plant Physiol.*, **135**, 801–13.
15. Lee, T-H., Kim, Y-K., Pham, T.T.M., et al. 2009, RiceArrayNet: a database for correlating gene expression from transcriptome profiling, and its application to the analysis of coexpressed genes in rice. *Plant Physiol.*, **151**, 16–33.
16. Lai, J., Ma, J., Swigonova, Z., et al. 2004, Gene loss and movement in the maize genome, *Genome Res.*, **14**, 1924–31.
17. The Rice Annotation Project Database (RAP-DB): 2008 update. 2008, Rice Annotation Project, *Nucleic Acids Res.*, **36**(Database issue), D1028–33.
18. Bennetzen, J.L., Coleman, C., Liu, R., Ma, J. and Ramakrishna, W. 2004, Consistent over-estimation of gene number in complex plant genomes, *Curr. Opin. Plant Biol.*, **7**, 732–6.
19. Dhadi, S.R., Krom, N. and Ramakrishna, W. 2009, Genome-wide comparative analysis of putative bidirectional promoters from rice, Arabidopsis, and Populus, *Gene*, **429**, 65–73.
20. Kensche, P.R., Oti, M., Dutilh, B.E. and Huynen, M.A. 2008, Conservation of divergent transcription in fungi, *Trends Genet.*, **24**, 207–11.
21. Singer, G.A., Lloyd, A.T., Huminiecki, L.B. and Wolfe, K.H. 2005, Clusters of co-expressed genes in mammalian genomes are conserved by natural selection, *Mol. Biol. Evol.*, **22**, 767–75.
22. Semon, M. and Duret, L. 2006, Evolutionary origin and maintenance of coexpressed gene clusters in mammals, *Mol. Biol. Evol.*, **23**, 1715–23.
23. Campbell, M.A., Zhu, W., Jiang, N., et al. 2007, Identification and characterization of lineage-specific genes within the Poaceae, *Plant Physiol.*, **145**, 1311–22.