Divergence of duplicate genes in exon–intron structure

Guixia Xu[a,1], Chunce Guo[a,b,1], Hongyan Shan[a], and Hongzhi Kong[a,2]

[a]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China; and [b]Graduate University, Chinese Academy of Sciences, Beijing 100049, China

Edited by Masatoshi Nei, Pennsylvania State University, University Park, PA, and approved December 14, 2011 (received for review June 8, 2011)

Gene duplication plays key roles in organismal evolution. Duplicate genes, if they survive, tend to diverge in regulatory and coding regions. Divergences in coding regions, especially those that can change the function of the gene, can be caused by amino acid-altering substitutions and/or alterations in exon–intron structure. Much has been learned about the mode, tempo, and consequences of nucleotide substitutions, yet relatively little is known about structural divergences. In this study, by analyzing 612 pairs of sibling paralogs from seven representative gene families and 300 pairs of one-to-one orthologs from different species, we investigated the occurrence and relative importance of structural divergences during the evolution of duplicate and nonduplicate genes. We found that structural divergences have been very prevalent in duplicate genes and, in many cases, have led to the generation of functionally distinct paralogs. Comparisons of the genomic sequences of these genes further indicated that the differences in exon–intron structure were actually accomplished by three main types of mechanisms (exon/intron gain/loss, exonization/pseudoexonization, and insertion/deletion), each of which contributed differently to structural divergence. Like nucleotide substitutions, insertion/deletion and exonization/pseudoexonization occurred more or less randomly, with the number of observable mutational events per gene pair being largely proportional to evolutionary time. Notably, however, compared with paralogs with similar evolutionary times, orthologs have accumulated significantly fewer structural changes, whereas the amounts of amino acid replacements accumulated did not show clear differences. This finding suggests that structural divergences have played a more important role during the evolution of duplicate than nonduplicate genes.

alternative splicing | coding-sequence evolution | exon shuffling | frame-shift mutation | regulatory divergence

Gene duplication plays important roles in organismal evolution. Paralogous genes, the products of gene duplication, initially have identical sequences and functions but tend to diverge in regulatory and coding regions. Divergence in regulatory regions can result in shifts in expression pattern, whereas changes in coding regions may lead to the acquisition of new functions. In the past few decades, owing to the availability of nucleotide, protein, and genomic sequences, as well as the accumulation of expressional and functional data, much has been learned about the mode, tempo, and consequences of duplicate gene evolution in coding and regulatory regions (1–15). However, there are still important issues that remain largely unexplored. For example, several recent studies have suggested that, although point mutation and insertion/deletion were generally believed to play overwhelming roles in coding-sequence evolution, the contributions of other mechanisms, such as exonization (a process in which an intronic or intergenic sequence becomes exonic) and pseudoexonization (the opposite process of exonization), should not be neglected (13–17). Yet, so far it is still unclear how and to what extent these and other less-well-known mechanisms for changes in exon–intron structure have contributed to the generation of functionally distinct duplicate genes.

To appreciate the contributions of structural divergence to functional innovations, we tried to investigate the evolutionary changes of a large number of duplicate and nonduplicate genes. However, because such investigations are extremely laborious and time consuming, we focused instead on a few hundred randomly sampled gene pairs. For example, 612 pairs of duplicate genes were sampled from the MADS-box, F-box, *AP2*, Cyclin, Homeodomain, Proteasome, and PP2C gene families for three reasons. First, these families code for proteins with diverse domain structures and functional properties (Fig. S1) and, therefore, the results obtained may well reflect the general patterns of structural divergence in duplicate genes. Second, all these families have experienced extensive gene duplication events during evolution, making it possible to identify plenty of paralogs for comparison. Third, members of these families play key roles in plant development and thus have been the focuses of functional studies; this suggests that the annotations for these families may be more reliable than others, especially in the species (such as *Arabidopsis thaliana*, hereafter called *Arabidopsis*; and *Oryza sativa* ssp. *japonica*, hereafter called rice) whose nuclear genomes have been completely sequenced and carefully annotated. For the analyses of nonduplicate genes, 300 pairs of orthologous genes from different species were used.

Results

Structural Divergences Were Widespread in Duplicate Genes. The *Arabidopsis* genome contains 106, 689, 145, 51, 104, 24, and 76 MADS-box, F-box, *AP2*, Cyclin, Homeodomain, Proteasome, and PP2C genes, respectively, and the corresponding numbers in rice are 71, 771, 167, 53, 101, 24, and 85. To create a dataset for this study, we conducted reciprocal BLAST and molecular phylogenetic analyses (*Methods*) and identified 612 pairs of closely related duplicate genes (hereafter called sibling paralogs) (Dataset S1). Comparison of these gene pairs indicated that in 180 cases (29.4% of 612), sibling paralogs had different numbers of exons, suggestive of severe divergences in gene structure (Fig. 1*A* and Fig. S2). In 402 other cases (65.7% of 612), the numbers of exons remained identical between sibling paralogs, whereas the lengths of one or more homologous exons were different, suggestive of relatively trivial structural divergences. In the remaining 30 cases (4.9% of 612), sibling paralogs possessed identical numbers and lengths of exons, and structural divergences could not be inferred at the first glance. Close inspections of their genomic sequences, however, revealed that in five cases, the apparently identical exon–intron structures were masked by the independent insertions or deletions of nucleotides. Note that in 182 cases (29.7% of 612), where alternatively spliced transcripts were produced by one or both genes, sibling paralogs were regarded as structurally divergent only if none of the splicing choices was shared; otherwise, they were considered as not yet diverged structurally. Using such a conservative criterion, we identified 587 pairs (95.9% of 612) of structurally diverged sibling paralogs (Fig. 1*A*), suggesting that structural divergences have played important roles in duplicate gene evolution.

Author contributions: H.K. designed research; G.X., C.G., H.S., and H.K. performed research; G.X., C.G., H.S., and H.K. analyzed data; and G.X. and H.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1] G.X. and C.G. contributed equally to this work.

[2] To whom correspondence should be addressed. E-mail: hzkong@ibcas.ac.cn.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1109047109/-/DCSupplemental.

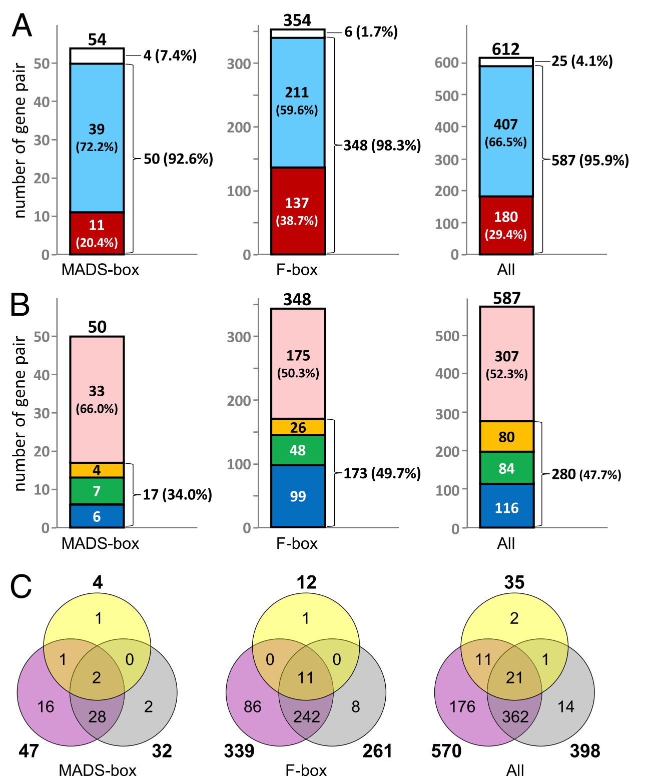

Fig. 1. Prevalence, consequences, and the underlying mechanisms for structural divergences. (*A*) Stacked bar charts showing the numbers and proportions of sibling paralogs that have diverged in exon–intron structure. Red boxes represent the gene pairs in which sibling paralogs possess different numbers of exons; blue boxes stand for those that have the same numbers of exons but have experienced insertion/deletion and/or exonization/pseudoexonization events. (*B*) Stacked bar charts showing the numbers and proportions of structurally diverged sibling paralogs that code for proteins with distinct domain organizations and/or sequence features. Blue boxes represent those that have different numbers or types of domains; green boxes represent those that have identical numbers and types of domains but show clear differences in sequence lengths; orange boxes represent those that are indistinguishable in domain organization or sequence length but possess relatively long, unalignable regions; and pink boxes represent those that do not show clear difference in protein sequences. (*C*) Venn diagrams depicting the numbers of sibling paralogs that have experienced insertion/deletion (purple), exonization/pseudoexonization (gray), and exon/intron gain/loss (yellow) events. For details, see Fig. S2.

The prevalence of structural divergences in duplicate genes raised the question of whether they can lead to the generation of functionally distinct proteins. To answer this question, we compared the protein sequences of the structurally diverged sibling paralogs. By searching against the SMART and Pfam databases (*Methods*), we found that in 116 cases (19.8% of 587 or 19.0% of 612), sibling paralogs contained distinct numbers and/or types of domains, suggestive of rather dramatic divergences in protein structure. In 84 cases (14.3% of 587 or 13.7% of 612), no difference could be detected in domain organization, yet sibling paralogs showed clear (>20%) differences in the lengths of their proteins. In 80 cases (13.6% of 587 or 13.1% of 612), sibling paralogs were indistinguishable in either domain organization or sequence length but possessed considerably large unalignable regions (Fig. 1*B*, Fig. S2, and Dataset S1). Taken together, these results suggest that nearly half (280; 47.7% of 587 or 45.8% of 612) of the structurally diverged sibling paralogs also code for proteins with distinct domain organizations and/or sequence features and that structural divergences do have the potential to generate proteins with distinct biochemical functions.

Structural Divergences Were Accomplished by Three Types of Mechanisms. To determine how the differences in exon–intron structure were generated, we compared the genomic sequences of the structurally diverged sibling paralogs (*Methods*). We found that at least three types of mechanisms contributed to structural divergences, with exon/intron gain/loss being the most apparent but least frequent ones (Fig. 1*C* and Fig. S2; for more information, see Dataset S1 and *SI Appendix*). By definition, exon gain is the process through which an entire (or occasionally partial) exon is obtained, either by duplication of a local exon (i.e., exon repetition/duplication) or by recruitment of an exotic one (i.e., exon shuffling in its strict sense), with exon loss being its opposite process. Similarly, intron gain is the process through which a piece of unrelated, exotic nucleotide sequence is inserted into an exon and causes exon fission, whereas intron loss refers to the removal of a preexisting intron and the fusion of two neighboring exons. In practice, however, it is not always easy to determine whether an orphan exon or intron was gained by one paralog or lost from the other unless the ancestral state is known; for this reason, we collectively regarded these processes as exon/intron gain/loss. Of the 587 pairs of structurally diverged sibling paralogs, exon gains/losses could be inferred in 18 cases (3.1%) and intron gains/losses in 19 cases (3.2%) (Fig. 2 *A*–*C*). In two cases (At1g22130 and At1g77980, and Os04g47580 and Os06g51110), both mechanisms have likely occurred (Fig. 2*C*). Notably, however, although gains/losses of introns never caused a shift in reading frame, gains/losses of exons sometimes did, especially when the numbers of nucleotides involved were not multiples of 3. In fact, of the 18 pairs that have experienced exon gain/loss events, a total of 38 events were inferred, 16 of which (42.1% of 38) led to shifts in reading frame. This result suggests that, although it occurred rather rarely, the contribution of exon/intron gain/loss to structural divergence and functional differentiation was substantial.

The second and most noteworthy type of mechanisms for structural divergence concerns exonization and pseudoexonization, two processes that can lead to the interchanges between exonic and nonexonic sequences. By comparing the genomic sequences of duplicate genes, we found that exonization/pseudoexonization occurred in 398 pairs (67.8% of 587 or 65.0% of 612) of sibling paralogs (Figs. 1*C* and 2 *B*–*F*). In 14 cases, exonization/pseudoexonization was the sole mechanism for structural divergence, whereas in all other cases it occurred together with other mechanisms (Fig. 1*C* and Fig. S2). When counted, a total of 932 exonization/pseudoexonization events were deduced, and thus the average number of events per gene pair was 2.34 (932/398). When divided by the total number of the investigated gene pairs, the number became 1.52 (932/612), suggesting that, on average, one-and-a-half exonization/pseudoexonization events were identified when a pair of duplicate genes was compared. This, together with the fact that 434 (46.6%) of the 932 observed exonization/pseudoexonization events involved nucleotides that were not multiples of 3, suggests that exonization and pseudoexonization were two important, but largely underestimated, mechanisms for structural divergence and functional innovation. Interestingly, shifts between exonic and nonexonic sequences can happen in the 5′ or 3′ part of the genes and, in 275 cases (29.5% of 932), were associated with the generation of novel initiation/stop codons. In 158 other cases (17.0% of 932), they caused the appearances or disappearances of the entire exons and, in these cases, the corresponding exonic and intronic/intergenic sequences could still be aligned with confidence. This, in fact, is one of the most important features of exonization/pseudoexonization, by which it can be distinguished from exon/intron gain/loss.

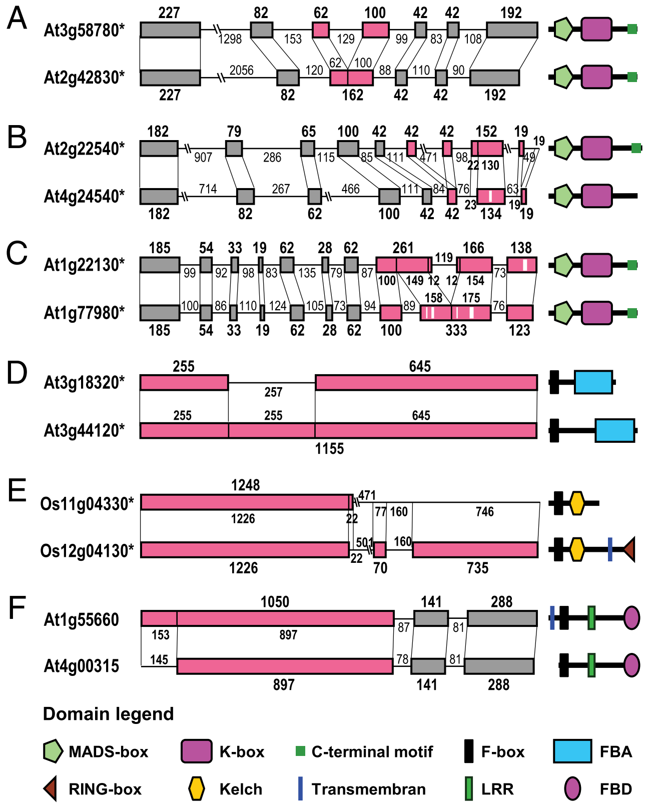

Fig. 2. The exon–intron structures of six pairs of representative sibling paralogs and the domain organization of their proteins, showing the three types of underlying mechanisms for structural divergences. Exons that have experienced exon/intron gain/loss (*A*–*C*), exonization/pseudoexonization (*B*–*F*), and insertion/deletion (*B* and *C*) events are highlighted with pink; those without structural difference are in gray. Small white bars in *B* and *C* depict the indels that have resulted from insertion/deletion events.

The third and most predominant type of mechanisms for structural divergence were intraexonic insertions and deletions, which were observed in 570 pairs (97.1% of 587 or 93.1% of 612) of sibling paralogs (Fig. 1*C*). In total, 5,796 insertion/deletion events were inferred, and the average number of mutational events per gene pair was 9.47 (5,796/612). When individual exons were taken into consideration, insertion/deletion could explain the divergences of 948 (51.8%) of 1,829 pairs of homologous exons, and the numbers of nucleotides involved varied from 1 to 283, with the most common number being 3 (1,722, or 29.7% of 5,796). Notably, however, although indels with multiples of 3 nucleotides were predominant (3,586, or 61.9%), those with other numbers also occurred frequently (2,210, or 38.1%), suggesting that a considerable number of indels have caused shifts in reading frame and changes in biochemical function.

It should be pointed out that the three main types of mechanisms for structural divergences were not mutually exclusive. For example, in 21 pairs of sibling paralogs, all three types of mechanisms occurred, sometimes making it difficult to determine the exact processes through which two duplicate genes diverged structurally. Of all the possible combinations of the three mechanisms, however, those of exonization/pseudoexonization and insertion/deletion were by far the most common and were documented in 383 cases (65.2% of 587 or 62.6% of 612) (Fig. 1*C*).

Structural Divergences Occurred Largely Proportionally to Evolutionary Time. To gain more insight into the general patterns of structural divergence, we pursued to see whether their occurrences were correlated with evolutionary time. We adopted the proportion of synonymous substitutions (P_S) as a crude measure for evolutionary time because synonymous substitutions are generally believed to be evolutionarily neutral and therefore can approximately reflect the evolutionary time elapsed since gene duplication (6, 8). We found that when P_S values were <0.15, the proportions of structurally divergent sibling paralogs increased as P_S value increased, suggesting that, at the early stages of duplicate gene evolution, the occurrence of structural divergence was largely proportional to evolutionary time. Thereafter, however, a plateau was reached when P_S values became larger (Fig. 3*A*), suggesting that structural divergences became more or less saturated and that nearly all sibling paralogs accumulated differences in exon–intron structure. Interestingly, of the 35 gene pairs with $P_S < 0.05$, 19 (54.3%) had already diverged in exon–intron structure, and the underlying mechanisms included exon/intron gain/loss, exonization/pseudoexonization, and insertion/deletion (Dataset S1). This suggests that all three types of mechanisms for structural divergence have occurred at the very early stages of duplicate gene evolution, even when nucleotide substitutions were scarce.

We also noticed that, although both exonization/pseudoexonization and insertion/deletion were major contributors to structural divergence, their relative contributions were different. For example, while insertions/deletions were documented in 570 pairs (93.2% of 612) of sibling paralogs, exonization/pseudoexonization only occurred in 398 (65.0%) of them. When evolutionary time became sufficiently long, nearly all sibling paralogs accumulated indels in coding regions, whereas the proportions of gene pairs with exonization/pseudoexonization never exceeded 90% (Fig. 3*A*). When individual mutational events were considered, the average number (9.47) of insertion/deletion events per gene pair was ~10 times lower than that (61,734/612 = 100.87) of amino acid replacements but 6 times higher than that (1.52) of exonization/pseudoexonization (Fig. S3). The fact that no significant correlation could be detected between the occurrences of exonization/pseudoexonization event and P_S values (Fig. S3) further suggests that, compared with amino acid replacements and insertions/deletions, exonization/pseudoexonization may be a slow and more or less conditional process.

Structural Divergences Were Less Prevalent in Orthologs than in Paralogs. The prevalence of structural divergences in duplicate genes also raised the question of whether they were paralogous gene specific. To address this question, we randomly sampled 300 pairs of one-to-one orthologs, 100 from each of the *Arabidopsis*–*A. lyrata* (A-Aly), *Arabidopsis*–poplar (A-P), and *Arabidopsis*–rice (A-O) comparisons (*Methods*). We found that 219 pairs (or 73.0%) of orthologs had diverged in exon–intron structure, with the numbers of gene pairs that had experienced insertion/deletion, exonization/pseudoexonization, and exon/intron gain/loss events being 212 (70.7%), 48 (16.0%), and 30 (10.0%), respectively (Fig. S4 and Dataset S2). The fact that 49 pairs (or 16.3%) of orthologs also coded for proteins with distinct domain organizations and/or sequence features (Fig. S4) further suggests that many of these structural divergences have led to the generation of structurally and functionally divergent orthologs.

Notably, compared to paralogs with similar evolutionary times, orthologs accumulated much fewer structural differences. For example, in all three comparisons, the proportions of orthologous gene pairs that experienced insertion/deletion and exonization/pseudoexonization events were always significantly smaller than in paralogs with similar P_S values (Fig. S5). When the average numbers of mutational events per gene pair were taken into consideration, the differences became more distinct

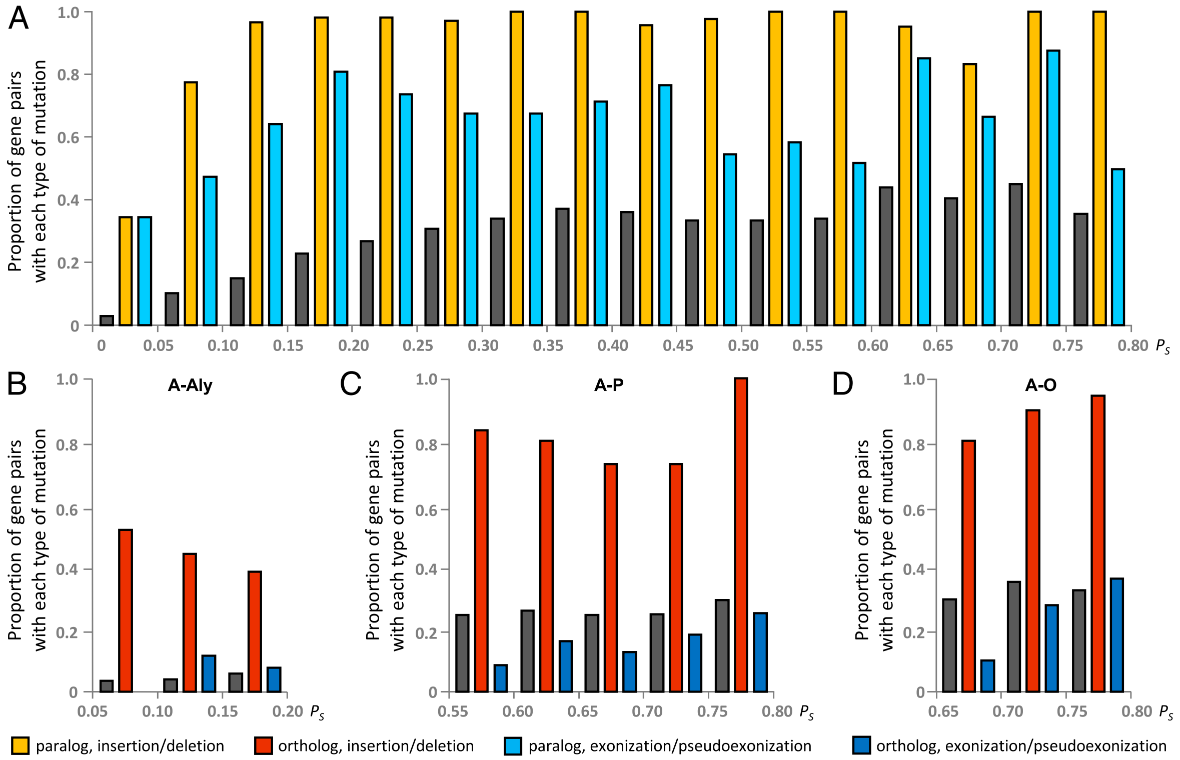

Fig. 3. Proportions of paralogous (*A*) and orthologous (*B–D*) gene pairs that have experienced insertion/deletion and exonization/pseudoexonization event (s). For simplicity, proportions of synonymous changes (P_S) are used to roughly measure the evolutionary times that have elapsed since the divergence of paralogous or orthologous genes. Gray bars show the proportions of amino acid replacements (d_A) between genes.

(especially for exonization/pseudoexonization; Fig. S5). Similar phenomena were observed when genes were grouped according to their P_S values (Fig. 3 and Fig. S3), suggesting that orthologs indeed accumulated much fewer structural differences than paralogs. This, together with the fact that only a small number of orthologs coded for proteins with distinct domain organization and/or sequence features, suggests that, unlike duplicate genes, nonduplicate genes have been very conserved in exon–intron structure and biochemical function. The exact and relative numbers of amino acid replacements per gene pair, however, were rather small and did not show clear difference between orthologs and paralogs (Figs. S3 and S5), suggesting that the rates of amino acid-altering substitutions did not change very much during the evolution of duplicate and nonduplicate genes.

Discussion

In this study, by comparing the genomic sequences of 612 pairs of sibling paralogs and 300 pairs of one-to-one orthologs, we established the general patterns of structural divergences in coding-sequence evolution. We found that: (*i*) divergences in exon–intron structure have been widespread in duplicate gene evolution; (*ii*) structural divergences can lead to the generation of proteins with distinct domain organization and sequence features, suggestive of the acquisition of new biochemical functions; (*iii*) structural divergences were caused by three types of mechanisms (i.e., insertion/deletion, exonization/pseudoexonization, and exon/intron gain/loss); (*iv*) the relative contributions of the underlying mechanisms for structural divergences were by no means the same; (*v*) structural divergences can occur at the very early stages of duplicate gene evolution, even when nucleotide substitutions were very scarce; (*vi*) like point mutations, insertion/deletion and exonization/pseudoexonization (and possibly exon/intron gain/loss) occurred more or less randomly, with the number of mutational events per gene pair being largely proportional to evolutionary time; (*vii*) structural divergences also occurred in orthologous genes, although the rates were generally much lower; and (*viii*) structural divergences have played a more important role in the evolution of duplicate rather than nonduplicate genes. Clearly, with these findings, a general picture has emerged for the mode, tempo, and consequence of structural divergences, and several relevant controversies can now be clarified.

Class I vs. Class II Mutations. As mentioned, changes in the function of a gene can be achieved by amino acid-altering substitutions (hereafter called Class I mutations) and/or structural changes in exon–intron organization (hereafter called Class II mutations, which include insertion/deletion, exonization/pseudoexonization, and exon/intron gain/loss). Class I mutations only lead to the replacements of amino acids at the same or homologous positions, so the length and homology of the proteins will not change. Class II mutations, however, usually cause additions or removals of nonhomologous amino acids and, as a result, can break the homology of the site(s) or region(s) concerned. In the past, Class I mutations have gained overwhelming attention, based upon which some important principles/theories of molecular evolution were uncovered or established (3, 6, 8). However, relatively little is known about Class II mutations, although several studies have suggested that their contributions to coding-sequence evolution were noteworthy (16–20). Therefore, our results, which highlight the prevalence and importance of structural divergences, will help clarify many important issues regarding to coding-sequence evolution. At least, in the future, when two or more genes are compared, special attention should be paid to their genomic sequences. Without the knowledge of exon–intron organization, it is impossible to guarantee the reliability of the alignments of genes if structural divergences, especially those that can cause shifts of reading frame, have occurred.

Structural Divergences and Shifts of Reading Frames. It is intriguing that, compared with orthologs, paralogs with similar evolutionary times have accumulated much more structural changes. This suggests that duplicate genes may have evolved in a way different from nonduplicate genes; at least, changes in exon–intron structure may have played a more important role in duplicate genes than in nonduplicate genes. In addition, because the numbers of nucleotides involved in many (~40%) structural changes were not multiples of 3, a considerably large number of duplicate genes have accumulated frameshift-inducing mutations in their coding regions. This is surprising, because shifts in reading frames were generally believed to be disastrous and can cause defects in gene function (1, 4, 6, 7). In the case of duplicate genes, however, it becomes understandable because when two duplicates coexist in the genome, the seemingly disastrous mutations (such as those that cause shifts in reading frames) in one paralog may no longer be disastrous; the normally expressed and properly functioning copy may be able to compensate for the negative effects caused by the defects in its paralog (11, 13–15, 21). Indeed, as has been illustrated in several recent studies, this kind of compensation sometimes can allow the generation of genes with novel biochemical functions, and that is why frameshift mutations have been more common in paralogs than in orthologs (13, 15). Nevertheless, the fact that some orthologs have also diverged in exon–intron structure suggests that, in addition to compensation, some other mechanisms may have also worked to preserve structurally diverged or even defective genes.

Exonization/Pseudoexonization and Alternative Splicing. One of the most striking findings of this study is that exonization/pseudoexonization has been very common in coding-sequence evolution. In the literature, exonization/pseudoexonization was sometimes regarded as a synonym of alternative splicing, likely because both processes can lead to the interchanges between exonic and nonexonic sequences (22, 23). However, it should be pointed out that they are two distinct processes for three reasons. First, alternative splicing generates different types of transcripts from a single gene, whereas exonization/pseudoexonization occurs to different genes and leads to the generation of structurally distinct paralogs or orthologs. Second, alternative splicing does not necessarily need the mutation or modification of the genomic sequences, whereas exonization/pseudoexonization requires the creation and/or fixation of a novel start/stop codon (of a gene) or donor/acceptor site (of an intron) (13, 22, 23). Third, alternative splicing is not always the prerequisite of exonization/pseudoexonization, as is intuitively believed; rather, in many cases, when exonization/pseudoexonization occurs, two genes did not partition the splicing choices used by their ancestor (13–15, 17). Nevertheless, there is no doubt that these are two similar and closely related processes and, working complementarily, they generate proteins with distinct sequence features and biochemical functions.

Possible Reasons for the Differences in Contribution Patterns. We have also shown that the relative contributions, as well as their contribution patterns, of the mechanisms underlying structural divergences were markedly different. Interestingly, these differences may be attributed to the processes through which these mutations were generated. Like point mutations, insertions and deletions occurred exclusively at the genome level and were the results of mistakes in meiosis; that is why they were so widespread. Exonization and pseudoexonization, however, occur less frequently because transcription itself is a precisely and rigidly regulated process, and shifts in splicing choices would never succeed unless they were allowed by the transcriptional machinery. Likewise, exon/intron gains/losses are the results of more unusual processes (e.g., exon duplication/repetition, exon scrambling, recombination, transposition, or retroposition) and, therefore, occur more or less circumstance-dependently. However, because the number of gene pairs investigated is still not very large, we were unable to estimate the exact rates of occurrence for these mechanisms; additional studies are needed to solve this problem.

Structural Divergence and Exon Shuffling. It has also been suggested that exon shuffling plays an important role in coding-sequence evolution. Yet, it should be pointed out that the term "exon shuffling" was used by different authors for different meanings. Initially, it referred to the process through which an exotic DNA fragment (usually a mobile element) was introduced and became an exon (6, 19, 20, 24). Later, however, it was used for whatever process (such as insertion/deletion, exon elongation/abridgement, exonization/pseudoexonization, or even point mutation) that has led to the rearrangement of exons (13–15, 23). In this study, we have shown that shuffling of exons can actually be accomplished by various mechanisms and, therefore, it is no longer appropriate to use the term exon shuffling for any specific, narrowly defined process. For this reason, we propose that the classic, mobile element-mediated process of exon shuffling (i.e., exon shuffling in its strict sense), together with exon duplication/repetition and several other processes, should better be included into the more broadly circumscribed "exon/intron gain/loss" category.

Coding vs. Regulatory Divergences. Another issue regarding to the evolution of genes concerns whether coding divergences were more important than regulatory divergences during evolution. To some extent, this has been one of the most controversial issues in the research field of evolutionary developmental biology and has gained increasing interest in recent years (1–3, 25–30). In particular, several recent studies have shown that genes affecting physiological traits (i.e., physiogenes) tend to evolve through changes in coding regions, whereas those controlling morphological traits (i.e., morphogenes) tend to accumulate more changes in regulatory regions (27, 31, 32). However, we argue that, although this is an interesting point of view, the real situation may be much more complex than we can imagine, for three reasons. First, the boundaries between morphogenes and physiogenes are not always clear (especially in plants) and, therefore, it may be arbitrary to classify genes according to their functions. Members of the F-box gene family, for example, code for proteins that function as hormone receptors, together with those that are involved in protein degradation. Because both groups of genes can affect the formation of morphological and physiological traits, it has been difficult to say whether they are morphogenes or physiogenes. Second, based solely on the limited expression and functional data that we have, it has been impossible to evaluate the exact consequences of regulatory and coding changes. For instance, although shifts in expression pattern are usually good reflection of changes in regulatory regions, the effects of amino acid-altering substitutions and structural divergences in coding regions cannot be measured accurately unless comprehensive functional analyses have been conducted. However, up to now only a very small number of genes have been carefully studied, and thus our understanding of gene function is still far from complete. Third, although each gene has its own expression domains and biochemical functions, it is the interactions between genes that matter (27, 33–36). In other words, although a gene may be arbitrarily divided into its regulatory and coding regions for simplicity, it is the combined contributions of the two portions that determine its functions in the pathway or network in which it is involved. This implies that it may be difficult to determine whether changes in regulatory or coding regions have contributed to the formation of morphological or physiological characters.

Methods

Identification of Sibling Paralogs and One-to-One Orthologs. Members of the MADS-box, F-box, *AP2*, Cyclin, Homeodomain, Proteasome, and PP2C gene families were retrieved from The *Arabidopsis* Information Resource (TAIR; http://www.arabidopsis.org/) and Rice Genome Annotation Project (RGAP; http://rice.plantbiology.msu.edu/) Web sites. To assure the reliability of the data, only the newest versions of annotation were used: TAIR10 for *Arabidopsis* and RGAP6.1 for rice. Sibling paralogs were first identified by reciprocal BLAST searches against the retrieved dataset and then confirmed by phylogenetic analyses of each gene family. One-to-one orthologs were selected randomly from the gene pairs identified by previous studies (37, 38) or retrieved by reciprocal BLAST and phylogenetic analyses. Domain organization of each protein was checked and visualized on the SMART and Pfam Web sites.

Determination of Structural Differences and Their Underlying Mechanisms. To determine whether paralogous or orthologous genes have diverged in exon–intron structure, we compared their genomic sequences. Two paralogs or orthologs were regarded as structurally divergent if they had different numbers of exons or if they had the same number of exons but the lengths of at least one pair of homologous exons were different. To understand the underlying mechanisms for structural divergence, we generated pairwise alignments for each gene pair, using the corresponding mRNAs as guidance. Intraexonic insertion/deletion was deduced when an indel was found within the aligned homologous exons. Exon/intron gain/loss was inferred if an orphan exon/intron was the result of exon duplication, exon shuffling, exon scrambling, intron insertion, or intron deletion. Exonization/pseudoexonization was identified when the corresponding exonic and nonexonic sequences could be aligned with confidence.

Calculation of P_S Values. P_S values between paralogous and orthologous genes were calculated in MEGA 5 (39) by using the Nei-Gojobori method (Proportion). For each gene pair, all alignable regions, excluding those that suffered frameshift mutations, were used to ensure the reliability.

ACKNOWLEDGMENTS. We thank Zhenguo Lin, Manyuan Long, Hong Ma, Masatoshi Nei, Kenneth Wolfe, Jianzhi Zhang, Zhenguo Zhang, and two anonymous reviewers for helpful suggestions. This work was supported by National Basic Research Program of China Grants 2009CB941500 and 2011CB944600 and National Natural Science Foundation of China Grants 30970210 and 30990240.

1. Ohno S (1970) *Evolution by Gene Duplication* (Springer, New York).
2. King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. *Science* 188:107–116.
3. Nei M (1987) *Molecular Evolutionary Genetics* (Columbia Univ Press, New York).
4. Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* 256:119–124.
5. Force A, et al. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545.
6. Graur D, Li WH (2000) *Fundamentals of Molecular Evolution* (Sinauer, Sunderland, MA).
7. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155.
8. Nei M, Kumar S (2000) *Molecular Evolution and Phylogenetics* (Oxford Univ Press, Oxford).
9. Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* 16:1679–1691.
10. Li WH, Yang J, Gu X (2005) Expression divergence between duplicate genes. *Trends Genet* 21:602–607.
11. Moore RC, Purugganan MD (2005) The evolutionary dynamics of plant duplicate genes. *Curr Opin Plant Biol* 8:122–128.
12. Zhang J (2006) Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nat Genet* 38:819–823.
13. Conant GC, Wolfe KH (2008) Turning a hobby into a job: How duplicated genes find new functions. *Nat Rev Genet* 9:938–950.
14. Innan H, Kondrashov F (2010) The evolution of gene duplications: Classifying and distinguishing between models. *Nat Rev Genet* 11:97–108.
15. Soskine M, Tawfik DS (2010) Mutational effects and the evolution of new protein functions. *Nat Rev Genet* 11:572–582.
16. Xu G, Kong H (2007) Duplication and divergence of floral MADS-box genes in grasses: Evidence for the generation and modification of novel regulators. *J Integr Plant Biol* 49:927–939.
17. Xu G, Ma H, Nei M, Kong H (2009) Evolution of F-box genes in plants: Different modes of sequence divergence and their relationships with functional diversification. *Proc Natl Acad Sci USA* 106:835–840.
18. Long M, Langley CH (1993) Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. *Science* 260:91–95.
19. Long M, de Souza SJ, Rosenberg C, Gilbert W (1996) Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor. *Proc Natl Acad Sci USA* 93:7727–7731.
20. Long M (2001) Evolution of novel genes. *Curr Opin Genet Dev* 11:673–680.
21. Raes J, Van de Peer Y (2003) Gene duplication, the evolution of novel gene functions, and detecting functional divergence of duplicates *in silico*. *Appl Bioinformatics* 2: 91–101.
22. Ast G (2004) How did alternative splicing evolve? *Nat Rev Genet* 5:773–782.
23. Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: Diversification, exon definition and function. *Nat Rev Genet* 11:345–355.
24. Patthy L (1999) Genome evolution and the evolution of exon-shuffling—a review. *Gene* 238:103–114.
25. Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet* 39:121–152.
26. Hoekstra HE, Coyne JA (2007) The locus of evolution: evo devo and the genetics of adaptation. *Evolution* 61:995–1016.
27. Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell* 134:25–36.
28. Nei M, Niimura Y, Nozawa M (2008) The evolution of animal chemosensory receptor gene repertoires: Roles of chance and necessity. *Nat Rev Genet* 9:951–963.
29. Matute DR, Butler IA, Coyne JA (2009) Little effect of the *tan* locus on pigmentation in female hybrids between *Drosophila santomea* and *D. melanogaster*. *Cell* 139: 1180–1188.
30. Rebeiz M, et al. (2009) Evolution of the *tan* locus contributed to pigment loss in *Drosophila santomea*: A response to Matute et al. *Cell* 139:1189–1196.
31. Carroll SB (2005) Evolution at two levels: On genes and form. *PLoS Biol* 3:e245.
32. Liao BY, Weng MP, Zhang J (2010) Contrasting genetic paths to morphological and physiological evolution. *Proc Natl Acad Sci USA* 107:7353–7358.
33. Davidson EH, Erwin DH (2006) Gene regulatory networks and the evolution of animal body plans. *Science* 311:796–800.
34. Peter IS, Davidson EH (2011) Evolution of gene regulatory networks controlling body plan development. *Cell* 144:970–985.
35. Qian W, He X, Chan E, Xu H, Zhang J (2011) Measuring the evolutionary rate of protein-protein interaction. *Proc Natl Acad Sci USA* 108:8725–8730.
36. Wagner GP, Zhang J (2011) The pleiotropic structure of the genotype-phenotype map: The evolvability of complex organisms. *Nat Rev Genet* 12:204–213.
37. Duarte JM, et al. (2010) Identification of shared single copy nuclear genes in *Arabidopsis, Populus, Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evol Biol* 10:61.
38. Liu Y, Guo C, Xu G, Shan H, Kong H (2011) Evolutionary pattern of the regulatory network for flower development: Insights gained from a comparison of two *Arabidopsis* species. *J Syst Evol* 49:528–538.
39. Tamura K, et al. (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28:2731–2739.