Small-Scale Duplications Play a Significant Role in Rice Genome Evolution

GUO Xin-yi, XU Guo-hua, ZHANG Yang, HU Wei-min, FAN Long-jiang

(*Institute of Crop Science / Institute of Bioinformatics, Zhejiang University, Hangzhou 310029, China*)

Abstract: Genes are continually being created by the processes of genome duplication (ohnolog) and gene duplication (paralog). Whole-genome duplications have been found to be widespread in plant species and play an important role in plant evolution. Clearly un-overlapping duplicated blocks of whole-genome duplications can be detected in the genome of sequenced rice (*Oryza sativa*). Syntenic ohnolog pairs (ohnologues) of the whole-genome duplications in rice were identified based on their syntenic duplicate lines. The paralogs of ohnologues were further scanned using multi-round reciprocal BLAST best-hit searching ($E<e^{-14}$). The results indicated that an average of 0.55 sister paralogs could be found for every ohnologue in rice. These results suggest that small-scale duplications, as well as whole-genome duplications, play a significant role in the two duplicated rice genomes.

Key words: small-scale duplication; ohnologue; genome evolution; *Oryza sativa*; *Arabidopsis*

Whole-genome duplication or polyploidy is common in flowering plants [1, 2] and other organisms [3–5]. Genomic sequence analyses performed on rice (*Oryza sativa*) and *Arabidopsis* have provided strong evidence for ancient occurrences of polyploidy [2, 6-15]. Many un-overlapping duplicate regions with a conserved gene order and orientation, or duplicate blocks, cover most if not all of the chromosomes in the two species [6, 8-9, 14, 15]. Genes are continually being created by the processes of genome duplication (ohnolog) and gene duplication (paralog) [16]. Large-scale, or whole-genome, duplication events lead to a dramatic increase in the number of duplicate genes. Pairs of paralogs of a particular age that correspond to a large-scale duplication are expected to give rise to a secondary peak in the age distribution. An initial peak that accounts for the most recently duplicated genes is also expected [2]. In brief, their footprints should be left behind in a genome after large- and small-scale duplications during its evolution. A theoretical evolutionary scenario of large- and small-scale duplication events in a genome is illustrated in Fig. 1. A potentially diagonal (if the duplicated event is not too old) line with a conserved gene order and orientation, or a syntenic line of duplicated blocks (Fig. 1-B), and initial and secondary peaks of the age distributions of duplicated gene pairs (Fig. 1-C) will be observed. To date, evidence from the analysis of both duplicate blocks and age distribution has all supported the notion that at least one whole-genome duplication has occurred in rice and *Arabidopsis*, respectively, within the past 70 million years [2, 15, 17–20]. These two whole-genome duplications are also the latest ones, and are detectable in both species.

Most genes in a genome belong to a gene family because of continual duplication events. Rice and *Arabidopsis* are two species of plants to have had their genomes sequenced. An early glimpse of the genomes of these two species showed that duplicate genes make up significant proportions of the genomes in rice and *Arabidopsis* and some of them have many members per gene family or contig [6, 11]. For example, 40% of the genes in the *Arabidopsis* genome belong to gene families that contain ≥5 members, and 65% are in families that contain ≥2 members. In addition to large-scale duplication, small-scale duplication (tandem gene duplications) also accounts for a significant proportion of the increased family size [6]. However, the extent or the relative composition of the two kinds of duplication in plant genomes is not yet clear. In animals, such as vertebrates, both large- and

Received: 4 April 2005; **Accepted:** 20 July 2005

Corresponding author: FAN Long-jiang (fanlj@zju.edu.cn)

small-scale duplications have played significant roles in their evolution. Based on findings regarding human gene families, the number of duplication events arising from the ‘continuous mode’ (i.e. small-scale duplication) is 30–52% of the total number of duplication events [5].

In this study, the surviving ohnologues (a pair of duplicate genes produced by genome duplication) from whole-genome duplications in rice and *Arabidopsis* were identified and their duplicate genes (sister paralogs) produced by small-scale duplications were further searched using an exhaustive BLAST searching method. The comparison of the proportion of ohnologs and their sister paralogs showed that small-scale duplications, as well as whole-genome duplication, make a significant contribution to the genetic composition of rice genome.

MATERIALS AND METHODS

Sequence data source

The rice (osa1, version 2.0) and *Arabidopsis thaliana* (ath1, version 5.0) genome annotation databases were downloaded from The Institute for Genomic Research (TIGR) (www.tigr.org), USA.

Identification of duplicate genes and duplicate blocks

A total of 59 712 annotated coding sequences of rice (version 2.0) and 26 207 of *Arabidopsis* (version 5.0) encoded by their chromosomal order were compared by reciprocal BLASTN searching ($E<e^{-14}$) for any two chromosomes. Two sequences were defined as one-to-one paralogous or pairs of a duplicate gene when each was the best hit of the other. Such first-round searching will return the latest pair of a duplicate gene (Fig. 1-A). To obtain all pairs of duplicate genes occurring before and after a large-scale duplication event (Fig. 1-A), multiple rounds of searching were done. In multi-round reciprocal BLAST best-hit searching (MRB), coding sequences on two chromosomes were used as queries to search against each other, respectively, and searching was performed continually after the coding sequence returned from the previous search was removed until no hit was returned. Coding sequences that show a BLASTN match ($<1\times e^{-10}$) with members of the rice and *Arabidopsis* repeat databases by TIGR should first be removed [11]. A pair of duplicate genes identified by this method is presented as a single dot in Fig. S1(Supplementary information see http: // ibi. zju. edu. cn / bioinplant/data/).

Identification of ohnologues and their sister paralogs

Pairs of duplicate genes on detected syntenic duplicate lines (Fig. S1) were considered ohnologues due to whole-genome duplication. Otherwise, they were considered paralogues due to small-scale (tandem or segmental) gene duplication. Wolfe [16, 21] suggested that duplicate genes formed by polyploidy should be called ‘ohnologs’, after Susumu Ohno [3], to distinguish them from other kinds of paralogs, because they are all the same age.

Ohnologues were identified further, as follows: 1) Remove redundancy of ohnologues. There was a discernible redundancy in the ohnologues identified in the above step. For example, if the paralog pair *ac* was on or near a syntenic line, like ohnologue *ad* (Fig. 1-A, B), it would all be considered an ohnologue in the above step. However, only pair *ad* is actually an ohnologue. Only those with synonymous substitution rate (Ks) values closest to the peak values of age distributions of the ohnologues of rice (1.0) and *Arabidopsis* (0.8) (Fig. 2) were retained and used for the next analysis. A total of 1800 and 1574 unique ohnologues were identified for rice and *Arabidopsis*, respectively (Table S1, http: // ibi. zju. edu. cn / bioinplant/data/); 2) Based on the age distribution of unique ohnologues (Fig. 2), 1176 and 1479 ohnologues with Ks values within 0.4–2.0 (for rice) or 0.4–1.6 (for *Arabidopsis*) were selected and used to search for sister paralogs of ohnologues.

Based on the results of MRB searching, all pairs of duplicate genes that involved one of the homologs of the ohnologues identified above were identified as sister paralogs of the ohnologues for rice and *Arabidopsis*, respectively.

Estimation of synonymous substitution rate (Ks)

Translated protein sequences were aligned using the water program (Smith-Waterman algorithm) of

EMBOSS (www.uk.embnet.org/Software/EMBOSS/), and the resulting alignment was used as a guide to align the nucleotide sequences. After gaps were removed, the level of synonymous substitutions was estimated using the maximum likelihood method implemented in codeml [22] under the F3 × 4 model [23].

RESULTS

Whole-genome duplications in rice and *Arabidopsis*

Rice and *Arabidopsis* each experienced at least one round of whole-genome duplication within the past 70 million years [2, 15, 17–20]. Clearly unoverlapping duplicate blocks of two whole-genome duplications, including the latest duplication, were detected in these genomes (Fig. S1). In rice lineage, except for a recent ~5 Mb duplicated block from the segmental duplication of chromosomes 11 and 12, duplicated blocks all originated from a whole-genome duplication event (Fig. S1-A). The duplicate blocks we detected in rice and *Arabidopsis* using updated data sets were basically the same as those in previous reports [15, 17].

Based on the above dot-plots (Fig. S1) of pairs of duplicate genes, duplicate blocks with clear syntenic duplicate lines in rice and *Arabidopsis* genome were collected and used for the next analysis (for the detailed genomic positions of the duplicate blocks and the number of unique ohnologues forming the syntenic lines, see online supplementary information Table S1, http://ibi.zju.edu.cn/ bioinplant/data/).

Age distributions of gene duplication events

In addition to the expected initial peak, one clear secondary peak was observed in the frequency distribution of Ks values of pairs of duplicate genes from duplicated blocks in rice and *Arabidopsis*, respectively (Fig. 2). The initial peak represents the most recently duplicated genes and secondary peaks reflect an expected polyploidy event [2]. The pairs of duplicate genes on syntenic lines of duplicate blocks (ohnologues) were further separated from the overall distribution. The results showed that the secondary peak indeed corresponded to the whole-genome duplications detected above in rice and *Arabidopsis*, respectively. The peak values of the two secondary

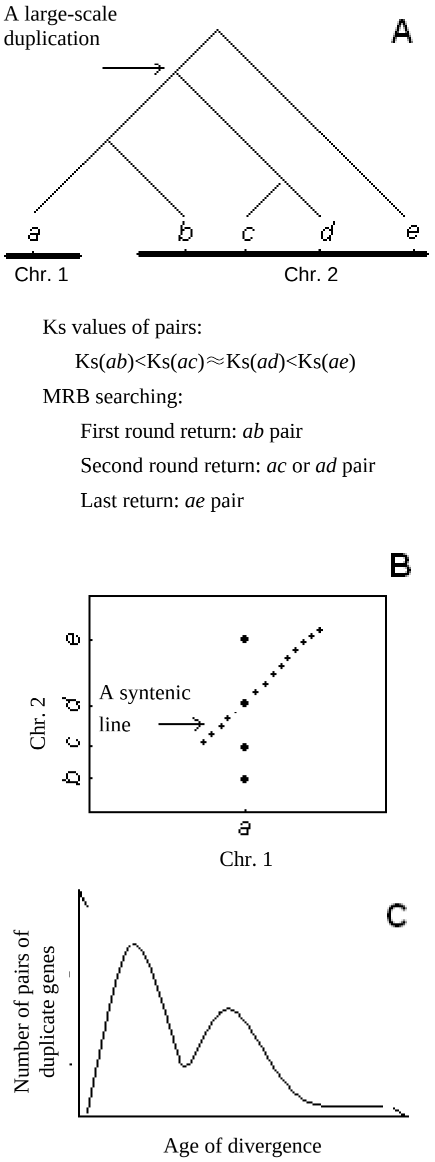
A large-scale duplication
A
a
b
c
d
e
Chr. 1
Chr. 2
Ks values of pairs:
Ks(ab)<Ks(ac)≈Ks(ad)<Ks(ae)
MRB searching:
First round return: ab pair
Second round return: ac or ad pair
Last return: ae pair
B
e
d
c
b
Chr. 2
A syntenic line
a
Chr. 1
C
Number of pairs of duplicate genes
Age of divergence

Fig. 1. Theoretical scenarios of duplicate genes in a genome.

(A) Five duplicate genes (*a*–*e*) in two chromosomes originated from four duplication events: an ancient large-scale (genome) duplication and three small-scale duplications (one of them is older than the large-scale duplication). The pair of duplicate genes "*ad*" was thought to be the original pair from the large-scale duplication event (ohnologue) and the other three genes are their sister duplicate genes (paralogs). The evolutionary distances (Ks, synonymous substitution rate) among the five genes and their best-hit returns from multi-round reciprocal BLAST best-hit (MRB) searching are expected to be as follows.

(B) Theoretical dot-plot between two chromosomes. Only the pairs of the five duplicate genes and a syntenic duplicate line of pairs of duplicate genes produced by the large-scale duplication are shown. Each dot in the plot represents one duplicate gene pair.

(C) Theoretical age distributions of pairs of duplicate genes in the genome. The secondary peak is the result of the large-scale duplication.

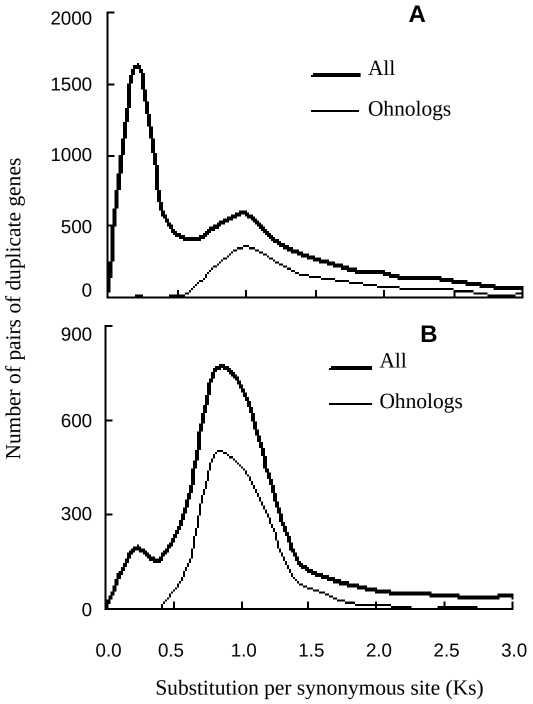

Fig. 2. Frequency distributions of synonymous substitution rates (Ks) obtained from pairs of duplicate genes on duplicate blocks of whole-genome duplication in the rice (A) and *Arabidopsis* (B) genomes.

Pairs of duplicate genes from the whole-genome duplication (ohnologues) are shown separately from the overall distribution. For detailed genomic positions of the duplicate blocks, see online supplementary information Fig. S1 and Table S1.

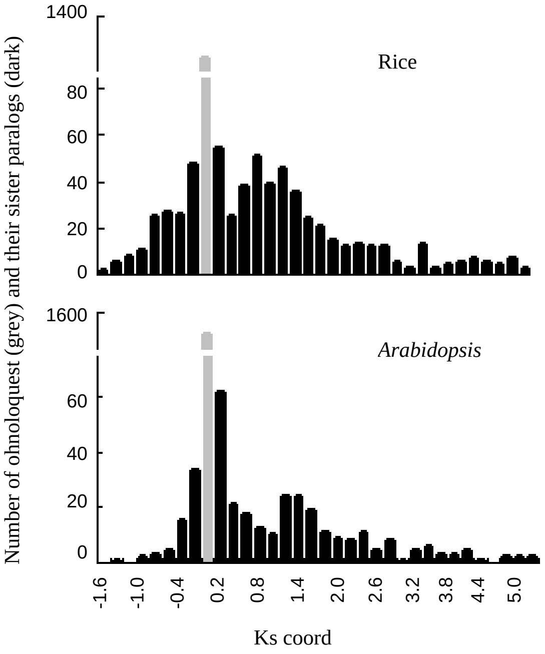

Fig. 3. Frequency distribution of ohnologues and their sister duplicate genes (paralogs) as a function of their synonymous substitution rates (Ks), relative to the Ks values of ohnologues, for the rice and *Arabidopsis* genomes.

For the two homologs of an ohnologue, their Ks values with sister paralogs were normalized by minimizing their ohnologue's Ks value. Ohnologues are arranged at position "0" and their sister paralogs, which were created before (positive values) or after (negative values) the whole-genome duplication, are shown on either side. Also see Fig. 1-A.

peaks for rice and *Arabidopsis* were about 0.8 and 1.0, respectively (Fig. 2). Meanwhile, the initial peaks of rice and *Arabidopsis* differed, in that a sharp initial peak was observed for rice, which suggested that there have been several recent gene duplications in rice, compared to the *Arabidopsis* genome.

Ohnologs and their sister paralogs

Since the true evolutionary scenario should reflect a combination of large- and small-scale duplications, an important question is their relative importance [5]. Ohnologs and their sister paralogs are the most important components of the genetic composition of a duplicated genome and their respective percentages may reflect their relative importance.

Based on the distributions of ohnologs and their sister paralogs that were created before or after whole-genome duplication, a significant proportion of duplicate genes due to small-scale duplication was observed in rice and *Arabidopsis* (Fig. 3). For a total of 1176 and 1479 ohnologues of rice and *Arabidopsis*, 648 (55.1% of 1176, or 0.55 sister paralogs per ohnologue) and 358 (24.2% of 1479, or 0.24 sister paralogs per ohnologue) sister paralogs were detected, respectively; 644 (54.8% of 1176) rice and 445 (30.1% of 1479) *Arabidopsis* ohnologues have at least one sister paralog. These results suggest that small-scale duplications played a considerable role, though not to the same extent as whole-genome duplication (0.24-0.55 small-scale duplication events are noted for each whole-genome duplication) in the duplicated genomes of the two plants. According to Fig. 3, small-scale duplications occurred continuously before and after whole-genome duplication events. Apparently, more duplicate genes were formed by small-scale duplications before whole-genome duplication than by small-scale duplications after whole-genome duplication, especially in rice (Fig. 3). These ancient paralogs, just like some ohnologues, survived long and brutal evolutionary selection. There were many sister paralogs near the Ks coord of their respective ohnologues, especially for *Arabidopsis* (Fig. 3). This should not be considered unusual. The Ks values of many ohnologs and their sister paralogs are approximately the same (Fig. 1-A). For example, the Ks values between all sister paralogs of ohnolog *d* (such as paralog *c*) and ohnolog *a* are approximately the same as the Ks value of ohnologue *ad* (Fig. 1-A). Meanwhile, pairs of paralogs *ab* and *ae* also have Ks values similar to that of ohnologue *ad*, if the relevant small-scale duplication events occurred not too far from their large-scale duplication (Fig. 1-A). In rice lineage, there appeared to be a region of recent sister paralogs (with greater negative values in Fig. 3). These results are consistent with the above observation of a sharp initial peak for rice. Meanwhile, there was a region of older sister paralogs, which suggested that ancient small-scale duplications occurred before the whole-genome duplication in rice.

The average numbers of sister paralogs per duplicate gene pair from other pairs of duplicate genes (ohnologues excluded) in the duplicate blocks used in our study were 0.67 and 0.32 for rice and *Arabidopsis*, respectively. These were both about 10 percentage points higher than the numbers of their respective ohnologues. Only ohnologues in which the two homologs have survived synchronously up to now after whole-genome duplication were detectable and used for our analysis.

DISCUSSION

In this study, the extent of small-scale duplications in genome evolution was estimated for two model plants. The results suggested that 0.24–0.55 small-scale duplication events per ohnologue corresponded to whole-genome duplication events in the monocot rice and dicot *Arabidopsis*. This estimation was based on syntenic duplicate genes (ohnologues) on well-documented duplicate blocks in whole-genome duplication in the two model plants. Ohnologues can be easily recognized based on syntenic duplicate lines and were therefore used in this study. The retained syntenic ohnologues comprised 12.7% and 22.5% of rice and *Arabidopsis* genes, respectively (Table S1), which are close to the estimates of Paterson et al [15] for rice (21.4%) and Bowers et al [14] for *Arabidopsis* (29.7%).

Only the latest whole-genome duplication, for which duplicate blocks are still discernible, was used in our study and other older polyploidies, which may have indeed occurred, were ignored. While this should overestimate the number of sister paralogs of ohnologues, this bias is limited. For example, two other potentially older large-scale duplications that predate monocot-dicot divergence have been suggested in *Arabidopsis* [14]. However, few sister paralogs correspond to these two ancient duplications (Ks<2.5 in Fig. 3). A Ks value of 2.5 in Fig. 3 is approximately the same as the absolute Ks values 3.3 (0.8+2.5) for *Arabidopsis*, corresponding to 220 million years for rice and *Arabidopsis*, assuming clock-like rates of synonymous substitution of 1.5×10^{-8} substitutions/synonymous site/year for dicots [24]. Monocot and dicots are believed to have diverged about 200 million years ago [25]. It is difficult to precisely target the actual sister paralogs of ohnologues from a mass of genes that are similar to ohnologs. A relatively conservative course was used in our study. Therefore, the true number of sister paralogs may be greater than that in the present estimation.

Many differences have been noted between the monocot rice and dicot *Arabidopsis* at both the gene and genome levels [26, 27]. We found an additional difference between them. Our results suggested that small-scale duplications are extensive or more common in rice, relative to *Arabidopsis*. Rice shows strong and continuous ‘meteor showers’ of duplicate genes from small-scale duplications in its genomic history. Further studies are needed to elucidate the mechanism that underlies this difference in the activity of small-scale duplication between these two model plants.

ACKNOWLEGEMENTS

This work was supported by the National Natural Science Foundation of China (90208022 and 30471067) and IBM Shared University Research (Life Science) .

REFERENCES

1 Wendel J F. Genome evolution in polyploids. *Plant Mol Biol*, 2000, **42**: 225–249.

2 Blanc G, Wolfe K H. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, 2004, **16**: 1667–1678.

3 Ohno S. Evolution by Gene Duplication. London: George Allen and Unwin, 1970.

4 Wolfe K H, Shields D C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 1997, **387**: 708–713.

5 Gu X, Wang Y, Gu J. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nature Genet*, 2002, **31**: 205–209.

6 The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 2000, **408**: 796–815.

7 Blanc G, Barakat A, Guyot R, Cooke R, Delseny M. Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell*, 2000, **12**: 1093–1101.

8 Paterson A H, Bowers J E, Burow M D, Draye X, Elsik C G, Jiang C X, Katsar C S K, Lan T H, Lin Y R, Ming R, Wright R J. Comparative genomics of plant chromosomes. *Plant Cell*, 2000, **12**: 1523–1540.

9 Paterson A H, Bowers J E, Peterson D G, Estill J C, Chapman B A. Structure and evolution of cereal genomes. *Curr Opin Genet Dev*, 2003, **13**: 644–650.

10 Vision T J, Brown D G, Tanksley S D. The origins of genomic duplications in *Arabidopsis*. *Science*, 2000, **290**: 2114–2117.

11 Goff S A, Ricke D, Lan T H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange B M, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun W L, Chen L, Cooper B, Park S, Wood T C, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller R M, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S. A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). *Science*, 2002, **296**: 92–100.

12 Simillion C, Vandepoele K, van Montagu M C E, Zabeau M, van de Peer Y. The hidden duplication past of *Arabidopsis thaliana*. *Proc Natl Acad Sci USA*, 2002, **99**: 13627–13632

13 Blanc G, Hokamp K, Wolfe K H. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res*, 2003, **13**: 137–144.

14 Bowers J E, Chapman B A, Rong J, Paterson A H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 2003, **422**: 433–438.

15 Paterson A H, Bowers J E, Chapman B A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci USA*, 2004, **101**: 9903–9908.

16 Wolfe K H. Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet*, 2001, **2**: 333–341.

17 Zhang Y, Xu G, Guo X, Fan L. Two ancient rounds of polyploidy in rice genome. *J Zhejiang Univ* (*Sci*), 2005, **6**: 87–90.

18 Guyot R, Keller B. Ancestral genome duplication in rice. *Genome*, 2004, **47**: 610–614.

19 Yu J, Wang J, Lin W, Li S, Li H, Zhou J, Ni P, Dong W, Hu S, Zeng C, Zhang J, Zhang Y, Li R, Xu Z, Li S, Li X, Zheng X, Cong L, Lin L,Yin J, Geng J, Li G, Shi J1, Liu J1, Lu H1, Li J, Wang J, Deng Y, Ran L, hi X1,3, Wang X, Wu Q, Li C, Ren X, Wang J, Wang X, Li D, Liu D, Zhang X, Ji Z, Zhao W, Sun Y, Zhang Z, Bao J, Han Y, Dong L, Ji J, Chen P, Wu S, Liu J, Xiao Y, Bu D, Tan J, Yang L, Ye C, Zhang J, Xu J, Zhou Y, Yu Y, Zhang B, Zhuang S, Wei H, Liu B, Lei M, Yu H, Li Y, Xu H, Wei S, He X, Fang L, Zhang Z, Zhang Y, Huang X, Su Z, Tong W, Li J, Tong Z, Li S, Ye J, Wang L, Fang L, Lei T, Chen C, Chen H, Xu Z, Li H, Huang H, Zhang F, Xu H, Li N, Zhao C, Li S, Dong L, Huang Y, Li L, Xi Y, Qi Q, Li W, Zhang B, Hu W, Zhang Y, Tian X, Jiao Y, Liang X, Jin J, Gao L, Zheng W, Hao B, Liu S, Wang W, Yuan L, Cao M, McDermott J, Samudrala R, Wang J, Wong G K, Yang H. The genome of *Oryza sativa*: A history of duplication. *PloS Biol*, 2005, **3**(2): e38.

20 Wang X, Shi X, Hao B, Ge S, Luo J. Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytol*, 2005, **165**: 937–946.

21 Wolfe K H. Evolutionary genomics: yeasts accelerate beyond BLAST. *Curr Biol*, 2004, **14**: R392–R394.

22 Yang Z. Phylogenetic Analysis by Maximum Likelihood (PAML). Ver. 2. London, U K: University College, 1999.

23 Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*, 1994, **11**: 725–736.

24 Koch M A, Haubold **B,** Mitchell-Olds T. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (*Brassicaceae*). *Mol Biol Evol*, 2000, **17**: 1483–1498.

25 Wolfe K H, Gouy M, Yang Y W, Sharp P M, Li W H. Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. *Proc Natl Acad Sci USA*, 1989, **86**: 6201–6205.

26 Wong G K, Wang J, Tao L, Tan J, Zhang J, Passey D A, Yu J. Compositional gradients in *Gramineae* genes. *Genome Res*, 2002, **12**: 851–856.

27 Yu J, Hu S, Wang J, Wong G K, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, 2002, **296**: 79–92.