the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

*The Plant Journal* (2010) doi: 10.1111/j.1365-313X.2010.04293.x

# Rice structural variation: a comparative analysis of structural variation between rice and three of its closest relatives in the genus *Oryza*

**Bonnie L. Hurwitz[1,2,†], Dave Kudrna[3,†], Yeisoo Yu[3], Aswathy Sebastian[3], Andrea Zuccolo[3], Scott A. Jackson[4], Doreen Ware[2,5], Rod A. Wing[1,3,6,*] and Lincoln Stein[2,7,*]**

[1]*Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA,*
[2]*Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA,*
[3]*Arizona Genomics Institute and School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA,*
[4]*Department of Agronomy, Purdue University, West Lafayette, IN 47907, USA,*
[5]*Robert W. Holley Center for Agriculture and Health, United States Department of Agriculture – Agricultural Research Service, Ithaca, NY 14853, USA,*
[6]*BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA, and*
[7]*Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada*

Received 16 April 2010; revised 29 June 2010; accepted 2 July 2010.

*For correspondence (fax 520 621 1259; e-mail rwing@ag.arizona.edu or fax 416 673 6664; e-mail lincoln.stein@gmail.com).
†Co-first authors, contributed equally to this work.

## SUMMARY

**Rapid progress in comparative genomics among the grasses has revealed similar gene content and order despite exceptional differences in chromosome size and number. Large- and small-scale genomic variations are of particular interest, especially among cultivated and wild species, as they encode rapidly evolving features that may be important in adaptation to particular environments. We present a genome-wide study of intermediate-sized structural variation (SV) among rice (*Oryza sativa*) and three of its closest relatives in the genus *Oryza* (*Oryza nivara*, *Oryza rufipogon* and *Oryza glaberrima*). We computationally identified regional expansions, contractions and inversions in the *Oryza* species genomes relative to *O. sativa* by combining data from paired-end clone alignments to the *O. sativa* reference genome and physical maps. A subset of the computational predictions was validated using a new approach for BAC size determination. The result was a confirmed catalog of 674 expansions (25–38 Mb) and 611 (4–19 Mb) contractions, and 140 putative inversions (14–19 Mb) between the three *Oryza* species and *O. sativa*. In the expanded regions unique to *O. sativa* we found enrichment in transposable elements (TEs): long terminal repeats (LTRs) were randomly located across the chromosomes, and their insertion times corresponded to the date of the A genome radiation. Also, rice-expanded regions contained an over-representation of single-copy genes related to defense factors in the environment. This catalog of confirmed SV in reference to *O. sativa* provides an entry point for future research in genome evolution, speciation, domestication and novel gene discovery.**

**Keywords: rice, *Oryza*, domestication, genome, evolution, comparative genomics.**

## INTRODUCTION

Rice (*Oryza sativa* L.) is the world's most important food crop, providing over one-fifth of the calories consumed by humans worldwide (Vaughan *et al.*, 2003). Efforts to improve yield, drought and disease resistance in rice are therefore of utmost importance. To meet these challenges and keep pace with increasing population needs, plant breeders have turned to molecular-based techniques to identify new sources of genetic variation, find markers linked to useful traits, and confirm the introgression of favorable genes from suitable plant donors (Fitzgerald *et al.*, 2009; Varshney *et al.*, 2009). These molecular approaches, however, depend on identifying genes in other plant species that impart new or optimized functions in the crop of interest. Most agronomically important genes are still largely uncharacterized among major crops of the *Poaceae* (Sallaud *et al.*, 2004; Doebley *et al.*, 2006; Itoh *et al.*, 2007; Salse *et al.*, 2008). As a result, comparative analysis of genome content within the grasses has become a fundamental component to understanding the unique physiological properties of each species, and how they contribute to the fitness and function of the organism in a given ecological niche.

Comparative studies among grasses have demonstrated that despite a relatively recent origin of 50–80 Mya from a common ancestor, grass genomes differ tremendously in chromosome number and genome size (Paterson *et al*., 2004; Prasad *et al*., 2005). These large-scale differences, however, cannot be explained by the presence or absence of genes among the grasses (Moore *et al*., 1995; Chen *et al*., 1997; Feuillet and Keller, 1999; Tikhonov *et al*., 1999; Bennetzen, 2007), which have been shown to be largely collinear and conserved. Instead, both ancient and recent polyploidy (Bennetzen, 2002; Paterson *et al*., 2004; Adams and Wendel, 2005), followed by the differential loss of DNA (Ge *et al*., 1999; Shimamoto and Kyozuka, 2002), as well as differential amplification/removal of transposable elements, appeared to be the dominant mechanisms regulating genome size variation (SanMiguel *et al*., 1998; Kazazian, 2004; Ma and Bennetzen, 2004; Bennetzen *et al*., 2005; Piegu *et al*., 2006). This suggested that functional differences among the grasses may lie in a handful of species-specific genes, or in differences in copy number or expression patterns in shared genes (Bennetzen, 2007). In maize, it was recently shown that inbred lines have significant copy number variation (CNV), which may impact phenotypic diversity, in particular disease response and heterosis (Springer *et al*., 2009; Belo *et al*., 2010). Despite recent strides in comparative sequence analysis among an increasing number of cereal genomes, little is known about the mechanisms involved in genomic rearrangements, and their effect on genes and functional selection in the genome. The difficulty lies in the fact that recent rearrangement events may only be detectable among closely related species or from recent polyploidization events, as molecular signatures degrade over evolutionary time: thus analyses of well-qualified genomic rearrangements may point the way towards understanding their impact.

The genus *Oryza* is an ideal model for studying genome evolution and domestication because it includes two cultivated and 22 wild species, with 10 genome types including several polyploids (Ge *et al*., 1999; Shimamoto and Kyozuka, 2002; Vaughan *et al*., 2003; Kellogg, 2009). Moreover, wild species in the genus *Oryza* include several AA genome types that are closely related to cultivated rice (*Oryza sativa*), and are mostly able to hybridize. This broad range of variation among species may be useful in distinguishing regions of the genome that are species- or lineage-specific, and have evolved to meet unique requirements for a given environment. In order to initiate robust studies in genome evolution in rice, a comprehensive and species-rich catalog of differences in genome content among the *Oryza* genomes is needed.

Here, we describe a genome-wide analysis of structural variation among three closely related *Oryza* genomes relative to the fully sequenced genome of *Oryza sativa* ssp. *japonica* (Asian rice). The three closely related *Oryza* species are of particular interest to crop improvement because *Oryza nivara* and *Oryza rufipogon* (genome sizes of 448 and 439 Mb respectively; Ammiraju *et al*., 2006) are predicted to be direct progenitors of *O. sativa* (Brar and Khush, 1997; Vaughan *et al*., 2003), and *Oryza glaberrima* (348–411 Mb; Martinez *et al*., 1994; Wing, 2010 unpublished data) is a domesticated rice native to Africa that is currently targeted for yield improvement through introgression with *O. sativa* (Kishine *et al*., 2008). Genomic regions of *O. nivara*, *O. rufipogon* and *O. glaberrima* were identified as expansions or contractions relative to *O. sativa* using BAC end sequences (BESs) and physical maps, and were then validated by the accurate size determination of BAC clones. Gene and transposable element (TE)/repeat contents in the expanded regions of the *O. sativa* genome that were not present in the other species were examined to further our understanding of structural variation, its implications toward speciation and the effects of domestication specific to *O. sativa.* This research provides a glimpse into genome-wide structural variation between domesticated and wild species within the genus *Oryza*, and lays the foundation for future studies in genome evolution that can be translated into functional applications for crop production.

## RESULTS

### A catalog of structural variation in rice

To quantitatively describe structural variation (expansions, contractions and inversions) in *O. nivara*, *O. rufipogon* and *O. glaberrima* relative to the *O. sativa* reference genome, an algorithm was developed to detect discordances between BAC clones that were aligned, using paired BAC end sequences, to the *O. sativa* reference sequence in conjunction with data from the associated physical maps (see Experimental procedures). Here, we define expansions, contractions and inversions relative to *O. sativa*: contractions refer to regions present in one or more *Oryza* species but not in *O. sativa*; expansions refer to regions present in *O. sativa* but not in the other *Oryza* species; and inversions refer to regions that have been inverted in an *Oryza* species as compared with *O. sativa*. Of the 117 452 BAC clones with paired end sequences, 66 996 BAC clones could be unambiguously aligned to the reference genome: *O. nivara* BES covered 16% of the *O. sativa* genome sequence, and *O. rufipogon* and *O. glaberrima* BESs each covered 11% of the *O. sativa* genome (Table 1). Of the 45 666 remaining BACs, 92% have just one BES of two that aligned to the *O. sativa* genome, and 8% have insufficient data in Genbank,

**Table 1** Alignment of BAC end sequences (BESs) and clones to the *Oryza sativa* reference sequence

| Species | Total BESs | Total clones | Total clones with paired BESs | BES aligned | Clones aligned | Clones aligned and in the FPC map | Clones aligned and in the FPC map with paired BESs | % of *O. sativa* covered by BES |
|---|---|---|---|---|---|---|---|---|
| *Oryza nivara* | 106 124 | 54 304 | 51 820 | 77 123 | 48 320 | 47 718 | 28 803 | 16 |
| *Oryza rufipogon* | 70 982 | 36 235 | 34 747 | 53 333 | 32 828 | 31 012 | 20 505 | 11 |
| *Oryza glaberrima* | 66 821 | 35 936 | 30 885 | 49 202 | 31 514 | 29 643 | 17 688 | 11 |
| Total | 243 927 | 126 475 | 117 452 | 179 658 | 112 662 | 108 373 | 66 996 | |

most probably because of low sequence quality. Solo alignments can result from: (i) BES alignments to the *O. sativa* genome that failed to meet our criterion, and were therefore filtered out because of low identity and coverage; (ii) being in a repetitive region; or (iii) lack of alignment because of the BES being located in a region of the genome that is unique to the other species. Repeat analysis of the BESs from clones with solo alignments revealed that 90, 90 and 80% of the clones for *O. rufipogon*, *O. nivara* and *O. glaberrima*, respectively, were composed primarily of repetitive sequences (see Experimental procedures). Therefore, the majority of clones were filtered out because of BES matches to repetitive regions. The remaining clones are expected to be in unique or variable regions of the other genomes.

### Rice expansions, contractions and inversions

Overall, we identified 674, 611 and 140 putative contractions, expansions and inversions, respectively, in the *O. sativa* reference sequence, based on discordant clone clusters from the other *Oryza* species (Table 2). *O. sativa* had the fewest contracted regions when compared with *O. glaberrima* (79), whereas comparisons with *O. nivara* showed the most contracted regions (389), followed by *O. rufipogon* (206). In contrast, *O. sativa* had the most expansions when compared with *O. glaberrima* (257), followed by *O. nivara* (183) and *O. rufipogon* (171). *O. sativa* had a similar number of putative inverted regions as compared with the closely related *Oryza* species: *O. rufipogon*, 40; *O. nivara*, 53; and *O. glaberrima*, 47. However, *O. nivara* had the greatest length of sequence in putative inverted regions (19.4 Mb), followed by *O. glaberrima* (14.7 Mb) and *O. rufipogon* (14.1 Mb). A complete list of all computationally predicted regions of structural variation in *O. sativa* is provided (Table S1).

### Validation of the expansions and contractions

To validate the putative expansions and contractions, we compared the distances in base pairs between mapped pairs of BESs with the actual size of each BAC clone measured on CHEF gels. The differences between the computational and empirical measurements were then quantified to obtain an overall measure of the magnitude of the expansions and contractions in each species relative to the *O. sativa* reference genome.

A total of 1285 expansion/contraction regions were computationally identified across the non-*O. sativa* genomes (572 from *O. nivara*, 377 from *O. rufipogon* and 336 from *O. glaberrima*). We analyzed 1125 of these regions experimentally (553 contractions and 572 expansions; 160 regions were not analyzed because of repeated DNA extraction failures and/or clone non-validation steps), and confirmed 938 regions (68% of the contractions and 98% of the expansions), resulting in an overall confirmation rate of 83% (Table 3). For *O. sativa* as compared with *O. nivara*, *O. rufipogon* and *O. glaberrima*, 162 (92%), 156 (99%) and 244 (99%) of the expansions, and 161 (54%), 155 (83%) and 60 (88%) of the contractions, were confirmed, respectively. The algorithm for detecting structural variation depends on a normal clone size distribution in the clone library in order to establish a range of concordant clone sizes. Only in the case of *O. nivara* were clone sizes not normally distributed (http://www.genome.arizona.edu/BAC_special_projects/img/OR_BBa.png), which may have caused the lower rate of

**Table 2** Computationally derived contractions, expansions and inversions in the *Oryza sativa* genome as identified from three closely related *Oryza* species

| | *Oryza sativa* | | | | | |
|---|---|---|---|---|---|---|
| | Contraction | | Expansion | | Inversion | |
| Compared species | Regions | Clones | Regions | Clones | Regions | Clones |
| *Oryza nivara* | 389 | 896 | 183 | 406 | 53 | 375 |
| *Oryza rufipogon* | 206 | 445 | 171 | 363 | 40 | 170 |
| *Oryza glaberrima* | 79 | 166 | 257 | 620 | 47 | 199 |
| Total | 674 | 1507 | 611 | 1389 | 140 | 744 |

doi: 10.1111/j.1365-313X.2010.04293.x

**Table 3** Confirmed expanded or contracted regions in the *Oryza sativa* genome, as identified from three closely related *Oryza* species. BAC clones, from comparative species that were informatically identified to be discordant in alignment with *O. sativa,* were analyzed for size to confirm the structural variation between the species

| | | *Oryza sativa* | | | |
|---|---|---|---|---|---|
| | | Expansion/contraction regions | | | Total genome |
| Compared species | Type | Identified | Analyzed | Confirmed | (Mb) |
| *Oryza nivara* | Contractions | 389 | 299 | 161 | 19.3 |
| | Expansions | 183 | 169 | 162 | 27.4 |
| *Oryza rufipogon* | Contractions | 206 | 186 | 155 | 15.4 |
| | Expansions | 171 | 158 | 156 | 25 |
| *Oryza glaberrima* | Contractions | 79 | 68 | 60 | 4.1 |
| | Expansions | 257 | 245 | 244 | 37.9 |

confirmation of contractions because of a broader range of clone sizes biased towards smaller sized clones. The overall sums of the expanded regions were 27.4, 25.0 and 37.9 Mb for *O. nivara*, *O. rufipogon* and *O. glaberrima*, respectively, and the overall sums of the contracted regions were 19.3, 15.4 and 4.1 Mb, for the same species. The average size, size range and standard deviation for expansions and contractions in *O. sativa* are shown in Table 4. A complete list of confirmed expansions and contractions (Table S2) is provided online, as are distributions of the sizes of the contractions (Figure S1) and expansions (Figure S2).

### Genome-wide distribution of expansions and contractions in rice

Expansions and contractions from *O. nivara*, *O. rufipogon* and *O. glaberrima* were displayed relative to the *O. sativa* reference sequence (Figure 1) to look for patterns in distribution across the rice genome. For each species, expansions and contractions were widely dispersed across all *O. sativa* chromosomes, and were not clustered at telomeres or near centromeric areas. Overall, we found three regions in the *O. sativa* genome that were contracted as compared with all three species, and 50 that were expanded (Table 5). Similarly, we found 50 contracted and 111 expanded regions that were shared between two species, and 257 contracted and 187 expanded regions that were found in one species. Contractions in the *O. sativa* reference sequence were more highly represented when compared with *O. nivara* and *O. rufipogon* species than with *O. glaberrima*, although expansions appeared to be, overall, more uniformly represented for all three species.

### Genes, TEs and repeats in the expanded regions of the rice genome

Regions from *O. sativa* that were identified as being expanded relative to the other three *Oryza* species were analyzed for gene and repeat content (including transposable elements) for each chromosome (Table 6). As a control, we calculated the expected gene, TE and repeat content for a region of the same size based on the genome-wide average across the *O. sativa* genome (International Rice Genome Sequencing Project, 2005). In eleven chromosomes the TE and repeat content was significantly greater in the expanded regions than the genome-wide average ($P < 0.001$). The average combined TE and repeat content for the expanded regions was 43.5%, whereas the genome-wide average was 34.8% ($P < 0.001$). In contrast, gene content was significantly lower in the expanded regions as compared with the genome-wide average in 10 chromosomes (1, 3–5 and 7–12) ($P < 0.005$). The average gene content in the expanded regions was 24.0%, in contrast to 27.3% genome-wide (International Rice Genome Sequencing Project, 2005). The majority of chromosomes (1, 3–5 and 7–12) showed a pattern where TE/repeat content was increased in expanded regions relative to the genome-wide average, and gene content was decreased.

To test whether the higher TE/repeat content found in the expanded regions of *O. sativa* resulted from non-uniform repeat content across the genome, we compared the TE/repeat content of a set of ‘pseudo BESs’ derived from five randomly selected expanded regions of *O. sativa*, from each of 12 pseudomolecules, with *O. nivara* BESs that defined these expanded regions (see details in Table S3; *O. nivara* BES provided in Table S4). The comparisons revealed that on average the *O. sativa* pseudo-BESs contained ~15% more TE/repeat sequences than the *O. nivara* BESs, thereby

**Table 4** Sizes (bp) of confirmed expansions and contractions in the *Oryza sativa* genome

| | *Oryza sativa* | | | | |
|---|---|---|---|---|---|
| Compared species | SV type | Mean | Min | Max | Standard deviation |
| *Oryza nivara* | Contraction | 120 279 | 17 411 | 434 156 | 83 505 |
| *Oryza rufipogon* | Contraction | 99 334 | 10 774 | 330 808 | 60 139 |
| *Oryza glaberrima* | Contraction | 73 558 | 8204 | 323 057 | 62 112 |
| *Oryza nivara* | Expansion | 169 479 | 8648 | 1 126 348 | 170 186 |
| *Oryza rufipogon* | Expansion | 159 310 | 15 630 | 1 163 409 | 182 972 |
| *Oryza glaberrima* | Expansion | 152 274 | 12 666 | 791 689 | 120 606 |

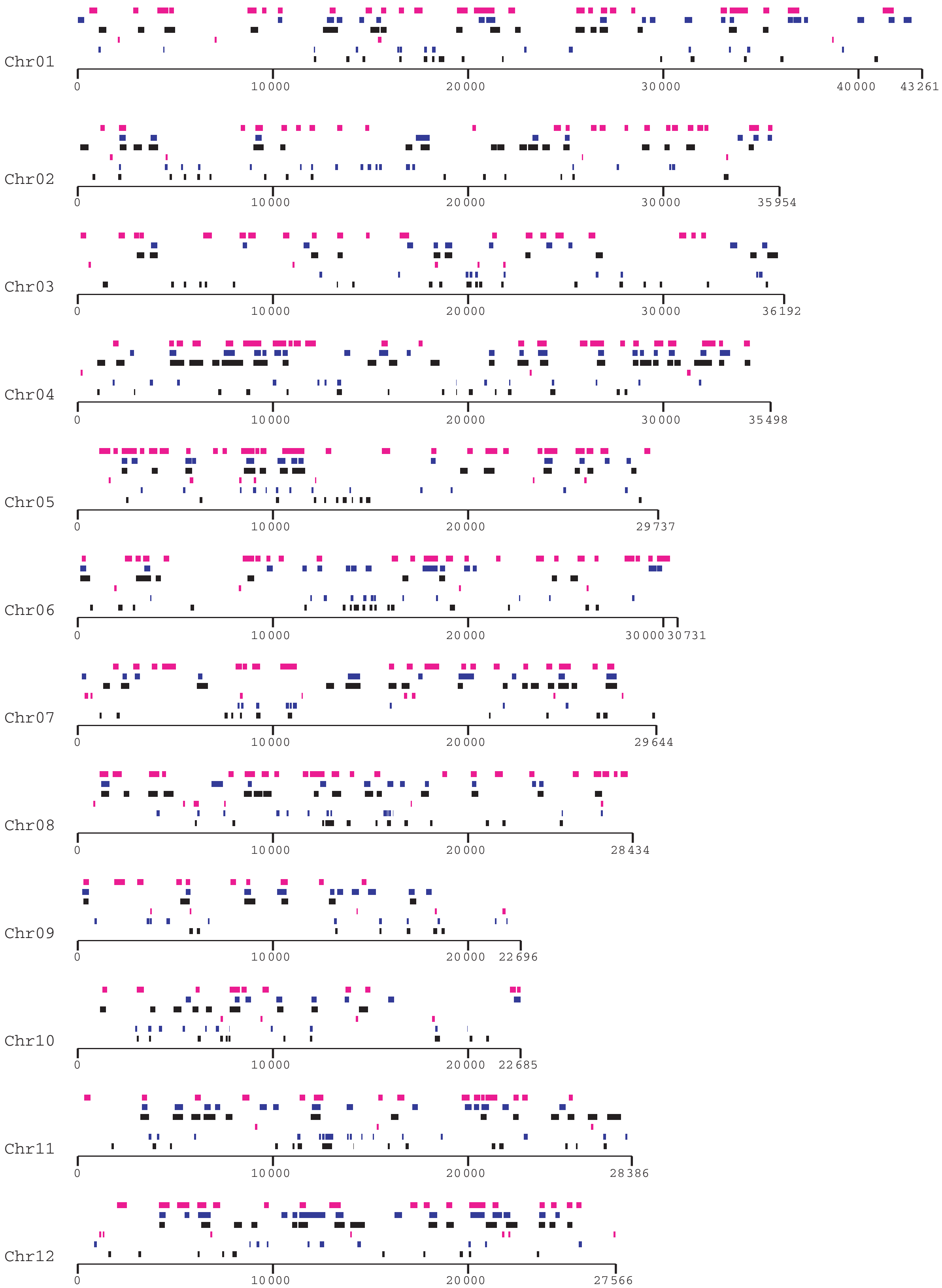

**Figure 1.** Confirmed expansions and contractions in *Oryza sativa* determined from comparisons with *Oryza nivara* (black), *Oryza rufipogon* (blue) and *Oryza glaberrima* (magenta). Lanes 1–3, from bottom to top, are contractions, and lanes 4–6 are expansions (Figure drawn with ''Parasight'' software: http://eichlerlab.gs.washington.edu/jeff/parasight/index.html).

doi: 10.1111/j.1365-313X.2010.04293.x

**Table 5** Number of confirmed shared structural variation regions in the *Oryza sativa* genome

| SV type | Species included in the SV | SV region count |
|---|---|---|
| Contraction | *Oryza nivara* | 113 |
| | *Oryza rufipogon* | 101 |
| | *Oryza glaberrima* | 43 |
| Expansion | *Oryza nivara* | 37 |
| | *Oryza rufipogon* | 43 |
| | *Oryza glaberrima* | 107 |
| Contraction | *Oryza nivara*, *Oryza rufipogon*, *Oryza glaberrima* | 3 |
| Expansion | *Oryza nivara*, *Oryza rufipogon*, *Oryza glaberrima* | 50 |
| Contraction | *Oryza nivara*, *Oryza glaberrima* | 8 |
| Expansion | *Oryza nivara*, *Oryza glaberrima* | 55 |
| Contraction | *Oryza rufipogon*, *Oryza glaberrima* | 6 |
| Expansion | *Oryza rufipogon*, *Oryza glaberrima* | 32 |
| Contraction | *Oryza nivara*, *Oryza rufipogon* | 36 |
| Expansion | *Oryza nivara*, *Oryza rufipogon* | 24 |

providing a second line of evidence pointing to the roles played by TE/repeats in the expansion events of *O. sativa*, as compared with *O. nivara*.

The most common classes of TEs (DNA transposons including helitrons, MITEs, LINEs, LTR retrotransposons and SINEs) and other repetitive sequences, such as ribosomal sequences and other unclassified repeats, were found in all the expanded regions (Table S5). All regions contained LTR TEs and MITEs, and only one region, though small (~17 Kb), contained no DNA TEs. LTR TEs, the largest type of class-I TEs, were identified in 30.7% of the total genome analyzed, and comprised over 67% of total TEs identified in the expanded regions. DNA TEs and MITEs, the most prevalent types of the class-II TEs, comprised almost 23% of the TEs identified in the expanded regions, as opposed to 9.9% found in the rice genome. Helitrons were identified on all chromosomes and in nearly half of all the expanded regions (44%), and the length of the sequences within the helitron signatures were as high as 9687 bp per region, with chromosome 8 having nearly 10-fold more helitron-attributed base pairs than in any other chromosome. Interestingly, all classes of TEs were in mixed positions along the chromosomes, without bias towards the centromere, although differences in the amounts within classes at specific locations varied widely, with LTRs being the predominant class followed by DNA transposons and MITEs.

Estimations of insertion times for intact complete LTR-RT elements of *O. sativa* expanded regions (in comparison with *O. nivara*) showed that 360 elements (64.9% of the 555 total) were inserted <0.58 Mya (see Figure S3); this correlates to the estimated date of the AA lineage diversification

**Table 6** TE/repeat and gene contents of confirmed expanded regions in the *Oryza sativa* genome

| Chr | No. expanded regions | Region sizes (kb) | Total TE/repeat content (kb) | Total non-repeat (kb) | % Repeat | $\chi^2$ *P* value | Total gene content (kb) | Total non-gene content (kb) | % Gene content | $\chi^2$ *P* value |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 20 | 4986 | 1966 | 3020 | 39.40 | 0 | 1169 | 3817 | 23.40 | 0 |
| | Expected | | 1735 | 3251 | 34.80 | | 1363 | 3623 | 27.30 | |
| 2 | 12 | 3294 | 1236 | 2058 | 37.50 | 0.001 | 962 | 2332 | 29.20 | 0.0157 |
| | Expected | | 1146 | 2148 | 34.80 | | 900 | 2394 | 27.30 | |
| 3 | 10 | 2130 | 988 | 1142 | 46.40 | 0 | 482 | 1648 | 22.60 | 0 |
| | Expected | | 741 | 1389 | 34.80 | | 582 | 1548 | 27.30 | |
| 4 | 20 | 6509 | 2909 | 3600 | 44.70 | 0 | 1610 | 4899 | 24.70 | 0 |
| | Expected | | 2264 | 4245 | 34.80 | | 1779 | 4730 | 27.30 | |
| 5 | 15 | 4964 | 2074 | 2890 | 41.80 | 0 | 1244 | 3720 | 25.10 | 0.0003 |
| | Expected | | 1727 | 3237 | 34.80 | | 1357 | 3607 | 27.30 | |
| 6 | 10 | 2302 | 1100 | 1202 | 47.80 | 0 | 743 | 1559 | 32.30 | 0 |
| | Expected | | 801 | 1501 | 34.80 | | 629 | 1673 | 27.30 | |
| 7 | 13 | 3689 | 1694 | 1995 | 45.90 | 0 | 850 | 2839 | 23.00 | 0 |
| | Expected | | 1283 | 2406 | 34.80 | | 1008 | 2681 | 27.30 | |
| 8 | 16 | 3955 | 1740 | 2215 | 44.00 | 0 | 885 | 3070 | 22.40 | 0 |
| | Expected | | 1376 | 2579 | 34.80 | | 1081 | 2874 | 27.30 | |
| 9 | 7 | 1698 | 788 | 910 | 46.40 | 0 | 306 | 1392 | 18.00 | 0 |
| | Expected | | 591 | 1107 | 34.80 | | 464 | 1234 | 27.30 | |
| 10 | 10 | 1981 | 1006 | 975 | 50.80 | 0 | 484 | 1497 | 24.50 | 0.0041 |
| | Expected | | 689 | 1292 | 34.80 | | 541 | 1440 | 27.30 | |
| 11 | 13 | 2994 | 1069 | 1925 | 35.70 | 0.292 | 721 | 2273 | 24.10 | 0.0001 |
| | Expected | | 1042 | 1952 | 34.80 | | 818 | 2176 | 27.30 | |
| 12 | 14 | 4818 | 2085 | 2733 | 43.30 | 0 | 916 | 3902 | 19.00 | 0 |
| | Expected | | 1676 | 3142 | 34.80 | | 1317 | 3501 | 27.30 | |
| All | All chrs. | 43 320 | 18 655 | 24 665 | 43.10 | 0 | 10372 | 32 948 | 23.90 | 0 |
| | Expected | | 15 071 | 28 249 | 34.80 | | 11 839 | 31 481 | 27.30 | |

(Ammiraju *et al.*, 2008). Of interest, 75 elements (∼13%) have identical sequences, indicting that they have been inserted extremely recently, thus testifying to their ongoing retrotranspositional activity. It was found that ∼25% of the intact LTR elements were within regions that mapped to identified rice centromeric domains (Yan *et al.*, 2008), whereas the remaining elements were mostly randomly distributed across chromosomes (Table S6). We also detected complete elements with older insertion times: 58 (0.5–1 Mya), 59 (1–1.5 Mya), 34 (1.5–2 Mya), 17 (2–2.5 Mya), 19 (2.5–3 Mya), 11 (>3 Mya) (Figure S3). Overall, these results provide evidence pointing to the possible roles played by TEs and repeats in the expanded regions of *O. sativa*.

### Gene function in expanded regions of the rice genome

To better understand gene function in the expanded regions in *O. sativa*, we looked for enrichment in these regions of gene ontology (GO) terms involved in biological processes, cellular components, and molecular function using a hypergeometric test with the Benjamini and Hochberg correction in the software plug-in BiNGO (Maere *et al.*, 2005) (see Experimental procedures). These analyses were further broken down into sets of single copy genes, multiple copy genes and total genes in the genome, to examine the functions, relevant to copy number, that were present in the expanded regions as compared with the rice genome. Overall we found that 85% (11 025/12 981) of the genes identified in the *O. sativa* expanded regions were single copy in the *O. sativa* reference genome, which is nearly identical to that found genome-wide (86% single copy; 32 461/37 544). The most broadly-based GO terms that were statistically over- or under-represented ($P < 0.05$) in the expanded regions, as compared with the complete genome, are shown in Table 7 (the complete data set is provided in Table S7). In both the single copy and total gene set, we found a statistically significant ($P < 0.05$) over-representation of genes related to phosphate metabolic processes, and in particular genes related to phosphorylation (95% total genes; 95% single copy genes; Table S7) in biological processes. Genes under the developmental processes category, under biological processes, were mostly related to cell death for both the single copy and total gene sets (95% single copy genes; 96% total genes; Table S7). GO terms for responses to stimulus under biological processes were over-represented in both the total and single copy gene sets, and were primarily involved in stress (92% of total and single copy gene sets) and defense (58% of total single gene sets) responses (Table S7). In particular, these genes contained pathogen-resistance and antimicrobial properties. The aromatic compound catabolic processes category under biological processes was found to be over-represented in the total and multiple copy gene sets, and in particular for L-phenylalanine catabolic process (100% total genes; 100% multiple copy genes; Table S7). The biopolymer metabolic processes category, under biological processes, was only found to be significant in the total gene set. DNA recombination and amine catabolic processes under biological processes were found to be over-represented in the multiple

**Table 7** Statistically over- and under-represented gene ontology (GO) categories for genes in the expanded regions of the *Oryza sativa* genome, compared with the complete genome

| Statistically over- or under-represented GO terms ($P$ value < 0.05) | GO category | GO term | Total genes (expanded regions) | Total genes (complete genome) | Single copy genes (expanded regions) | Single copy genes (complete genome) | Multiple copy genes (expanded regions) | Multiple copy genes (complete genome) |
|---|---|---|---|---|---|---|---|---|
| Over-represented | Biological processes | Phosphorus metabolic processes | 671 | 1719 | 583 | 1510 | 0 | 0 |
| | Biological processes | Responses to stimulus | 385 | 947 | 337 | 834 | 0 | 0 |
| | Biological processes | Developmental processes | 235 | 517 | 208 | 456 | 0 | 0 |
| | Biological processes | Biopolymer metabolic processes | 1016 | 2724 | 0 | 0 | 0 | 0 |
| | Biological processes | Aromatic compound catabolic processes | 9 | 10 | 0 | 0 | 7 | 7 |
| | Biological processes | Amine catabolic process | 0 | 0 | 0 | 0 | 8 | 8 |
| | Biological processes | DNA recombination | 0 | 0 | 0 | 0 | 21 | 32 |
| | Molecular function | Transferase activity | 974 | 2609 | 837 | 2262 | 0 | 0 |
| | Molecular function | Binding | 2599 | 7320 | 2224 | 6342 | 0 | 0 |
| | Molecular function | Protease inhibitor activity | 43 | 80 | 0 | 0 | 10 | 11 |
| | Molecular function | Ammonia ligase activity | 0 | 0 | 0 | 0 | 7 | 7 |
| Under-represented | Biological processes | Localization | 489 | 1621 | 0 | 0 | 64 | 242 |

**Table 8** Ranking scheme of BAC end sequence (BES) alignments to the *Oryza sativa* genome

| BES in NCBI per clone | BES aligned | BES orientation | BES distance | BES location | Rank | Type | *O. nivara* clone count | *O. rufipogon* clone count | *O. glaberrima* clone count |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 2 | Opposite | Within 2 SD of mean clone size | Same chr. | 1 | Concordant | 18 094 | 13 060 | 10 274 |
| 2 | 2 | Opposite | Short | Same chr. | 2 | Discordant | 2272 | 1412 | 605 |
| 2 | 2 | Opposite | Long | Same chr. | 3 | Discordant | 102 | 678 | 550 |
| 2 | 2 | Same | Within 2 SD of mean clone size | Same chr. | 4 | Discordant | 187 | 103 | 134 |
| 2 | 2 | Same | Short | Same chr. | 5 | Discordant | 53 | 42 | 79 |
| 2 | 2 | Same | Long | Same chr. | 6 | Discordant | 3 | 18 | 5 |
| 2 | 2 | Opposite | Greater than max. clone size | Same chr. | 7 | Discordant | 1785 | 553 | 1731 |
| 2 | 2 | Same | Greater than max. clone size | Same chr. | 8 | Discordant | 547 | 274 | 343 |
| 2 | 2 | N/A | N/A | Different chr. | 9 | Discordant | 3784 | 2159 | 1800 |
| 2 | 1 | N/A | N/A | N/A | 10 | N/A | 19 350 | 11 755 | 10 842 |
| 1 | 1 | N/A | N/A | N/A | 11 | N/A | 1541 | 958 | 3280 |
| 2 | 0 | N/A | N/A | N/A | 12 | N/A | 0 | 0 | 0 |
| 1 | 0 | N/A | N/A | N/A | 13 | N/A | 495 | 293 | 1068 |
| 0 | 0 | N/A | N/A | N/A | 14 | N/A | 307 | 288 | 544 |

copy gene set only. We also found an over-representation of GO terms related to molecular function in transferase activity and binding (Table S7) for both the single copy and total gene sets. In the multiple copy and total gene sets, we found an over-representation of molecular function GO terms related to protease inhibitor activity and ammonia ligase activity. When we looked for an under-representation of genes in the expanded regions of *O. sativa*, we found that genes related to localization, and in particular transport (99% total genes, 100% multiple copy genes), in the biological processes category were under-represented ($P < 0.05$) in the total and multiple copy gene sets. We found no significant over- or under-representation of genes related to cellular processes for any of the gene sets.

## DISCUSSION

We report a genome-wide analysis of structural variation (SV) between Asian rice (*O. sativa* spp. *japonica*) and three closely-related species from the genus *Oryza*. These data provide a genome-wide comparative analysis of closely related plant species within a genus, similar in magnitude to other such studies as the fully sequenced *Drosophila* and yeast genomes (Kellis *et al.*, 2003; Bhutkar *et al.*, 2008). We show that despite lacking whole genome sequence data for *O. nivara*, *O. rufipogon* and *O. glaberrima*, BES alignments from high-resolution physical maps to the *O. sativa* reference genome are an exceptional resource for detecting BAC-sized structural changes in these genomes (Lewin *et al.*, 2009). Because of the size and complexity of plant genomes (Soltis and Soltis, 2003) and bioinformatic challenges associated with whole-genome sequencing and assembly, this BAC-based strategy is useful for the investigation of genome dynamics and evolution among closely related species.

To date, studies on structural variation in rice have focused on small-scale variation or limited regions and features of the genome. A comparison of the two sequenced cultivated Asian rice genomes, *japonica* and *indica*, showed that 5% of the genes were asymmetrically located between the genomes, and that 5% were present in only one genome (Ding *et al.*, 2007). Most asymmetric genes were classified as disease resistance or RLK kinase genes, indicating that selection may be playing a role in maintaining asymmetry. By comparing these same genome sequences, Feltus *et al.* identified 408 898 putative DNA polymorphisms (SNPs/INDELs), and discovered both polymorphism-poor and -rich regions that may be due to hotspots of recombination (Feltus *et al.*, 2004). More surprisingly, Ma *et al.* found that centromeres in rice were hotspots for unequal homologous recombination (Ma and Bennetzen, 2006). The dynamics of genome evolution between *O. sativa* and other species in the genus *Oryza* has also begun to be explored, and shows substantial diversity. For example, Piegu *et al.* demonstrated that amplifications of three LTR retrotransposons were responsible for an over two-fold increase in the genome size of *Oryza australiensis* in 2.5 My, as compared with *O. sativa*, indicating that LTRs play a fundamental role in genome size evolution in the genus *Oryza* (Piegu *et al.*, 2006). A recent study of 46 genes in the ADH1–ADH2 region of the *O. sativa* genome revealed lineage-specific gain and loss of genes in gene families among six diploid genomes separated by an evolutionary timescale of ~15 My (Ammiraju *et al.*, 2008). Although results from all of these studies indicate a rich and dynamic evolutionary history in rice, they neglect to analyze genome-wide structural differences among diverse species that may be of greater consequence in adaptation, speciation and crop improvement.

Studies in other model systems have proven effective in elucidating both the size and spectrum of genomic structural variation within one species, or between multiple species, and by relating structural variants to given phenotypes (Kellis *et al.*, 2003; Zhao *et al.*, 2004; Newman *et al.*, 2005; Tuzun *et al.*, 2005; Chen *et al.*, 2007; Bhutkar *et al.*, 2008; Kidd *et al.*, 2008; Mefford and Eichler, 2009; Nicholas *et al.*, 2009; Springer *et al.*, 2009). In humans, these studies have offered insights into human health, and have initiated studies in connecting phenotypes and genotypes associated with complex disease (Hannes *et al.*, 2009; Helbig *et al.*, 2009). In dogs, structural variation has been linked to olfaction, immunity and gene regulation, and may be important in distinguishing phenotypic differences and traits from domestication (Nicholas *et al.*, 2009). Studies of structural variation among closely related *Drosophila* and yeast species have yielded insights into genome evolution, and have pinpointed regions that are conserved in terms of gene order and content (Kellis *et al.*, 2003; Bhutkar *et al.*, 2008). Because of their large and complex genomes, similar studies on structural variation in plant species are largely understudied, but could have tremendous utility in differentiating genomic regions associated with complex traits, domestication and adaptation. Recently published studies in maize used comparative genome hybridization (CGH) methods to investigate SV: as CNV and presence/absence variation in two inbred lines (Springer *et al.*, 2009), and CNV comparison of 14 inbred lines. Given the extreme heterozygosity of maize, as expected these studies revealed variation among comparative genome segments with frequency differences and SV that suggest links toward understanding heterosis and identifying ancient origins of phenotype and disease response. In order to initiate robust studies in genome speciation and evolution, the effects of domestication and adaptation to extreme environments in *Oryza*, we compiled a genome-wide and species-rich catalog of differences in genome content among three *Oryza* species and domesticated rice, *O. sativa* ssp. *japonica*.

The four *Oryza* genomes studied were selected with careful consideration to their functional utility towards crop improvement. *O. sativa* is an Asian rice that is the major food source for the world's populations, and is cultivated worldwide (Vaughan *et al.*, 2003). *O. nivara* and *O. rufipogon* are thought to be progenitors of *O. sativa* (Brar and Khush, 1997; Vaughan *et al.*, 2003; Fitzgerald *et al.*, 2009); hence, novel genes or alleles in these species that have arisen since their divergence could be introgressed into *O. sativa* for crop improvement. The African cultivated rice, *O. glaberrima*, is of particular interest in crop improvement because it is resistant to drought, pests and diseases, but is low in yield (Kishine *et al.*, 2008). Agencies such as the Africa Rice Center (http://www.WARDA.org) and the International Rice Research Institute (http://www.IRRI.org) are working towards the development of new interspecific cultivars of rice, such as the New Rice for Africa (NERICA), derived from crosses between *O. glaberrima* and *O. sativa*, to enable sustainable crop production in West Africa. Understanding the exact nature of these crosses and the introgression of genes between these species can further crop development and self-sufficiency for rice production.

Because each of these *Oryza* species are closely related to *O. sativa*, and are estimated to have diverged <1 Mya, sequence conservation and gene collinearity were expected to be high (Ge *et al.*, 1999; Soderlund *et al.*, 2006). As a result, we were able to accurately map paired BESs from the *Oryza* genomes to the *O. sativa* reference sequence, and find optimal pairing using a graph theory-based algorithm to efficiently walk through all combinations of alignments. We identified clones with substantial differences in size and/or orientation similar to methods employed for SV studies of human and chimp (Newman *et al.*, 2005). A BAC clone end-sequence filtering rational (paired reads, multiple hit, hit distances) was used to provide an excellent resource for producing high-confidence comparative alignments, similar to those used for the identification of syntenic promoter regions in Arabidopsis relatives (Windsor *et al.*, 2006). Regions of the rice reference genome with discordant clones that indicated a similar type of SV were further validated through the use of manually edited physical maps and a new method for BAC clone size determination. By simultaneously using evidence from these disparate resources, we were able to make high-confidence predictions of intermediate BAC-sized SV in the *O. sativa* genome. However, additional work will be required to examine SV on a larger scale, and differentiate local duplications from translocation events and investigate CNV.

Overall, 674 putative contractions, 611 expansions and 140 inversions between *O. nivara*, *O. rufipogon*, *O. glaberrima* and the *O. sativa* reference genome were identified. We found a considerable level of contraction relative to *O. sativa*, with 15–19 Mb of chromosomal material present in the wild *Oryza* species that is absent in *O. sativa*, whereas the African cultivated rice, *O. glaberrima,* differs by only 4.1 Mb from *O. sativa*. This suggests that artificial selection may be a factor in regulating genome size. Expansions ranged from 25 to 38 Mb of unique sequence in *O. sativa*, i.e. not present in the other *Oryza* species. The putative inverted regions of the *O. sativa* genome ranged from 14 to 19 Mb. The contraction and expansion regions were validated at a high threshold of 68 and 98%, respectively. We found that one of the species in our analysis, *O. nivara,* had an abnormal distribution of clone sizes that contributed to our lower rate of confirmation for contractions (see the Results). The numbers of observed expansion or contraction regions may be a reflection of genome size differences. Currently, the sequenced size of *O. sativa* (389 Mb, after factoring in sequencing and physical gaps by IRGSP) is 26–50 Mb smaller than flow cytometry estimations of 415–439 Mb.

As flow-cytometry size estimations are variable because of many factors, it can be suggested that non-*O. sativa* genome sizes may be more accurate after adjustment based upon differences found between the cytometry size estimations and whole-genome sequencing of *O. sativa*. Thus, the genome species size for *O. glaberrima,* when adjusted based upon *O. sativa* cytometry sequencing differences, becomes ~316–369 Mb, and this size reduction would support our data that revealed a greater number of expansions in *O. sativa*, with fewer contractions because of the genome size difference between the two species.

Of the confirmed contractions, three regions were shared by all three *Oryza* species, 50 were shared by two species and 257 were found in just one species. The regions of shared contraction indicate that a region in the most recent common ancestor was maintained in the other *Oryza* species, but was lost in *O. sativa*, and may be related to specific speciation events in *O. sativa*. Likewise, 50 regions of confirmed expansion were shared among all three *Oryza* species, 111 were shared in two species and 187 were shared in one species. The shared expansions indicate that *O. sativa* contains sections of unique sequence that are not present in the other species, and could contain features that are functionally important to *O. sativa* specifically.

We closely examined the unique regions of the *O. sativa* genome that were not present in one or all three of the other *Oryza* species. These regions are potentially important because they may contain clues about speciation and domestication in *O. sativa* following its divergence from these close relatives (Izawa *et al.*, 2009). Overall, we found a high level of TE/repeat content and diminished gene content, as compared with the rest of the genome. The expanded regions were enriched for LTR TEs (71% of the total), DNA transposons (13%) and MITEs (10%), indicating that these mobile elements contributed to the dynamic gain and loss of sequence in the *O. sativa* genome. More interesting was the total length of sequence represented by the LTR TEs in the expanded regions. Class-I LTRs comprised over 30% of the sequence (>13 Mb) of the confirmed expanded regions, as opposed to 14% indentified in the rice genome sequence (International Rice Genome Sequencing Project, 2005). Additionally, we found 555 intact LTR-RT elements, attesting to the recent activity of this class of TEs. This provides evidence to suggest that major genome expansions within *O. sativa* may be attributed to class-I LTR TE proliferation. TEs have been previously shown to be major contributors to genome size evolution based on analyses of smaller regions of cereal genomes (SanMiguel *et al.*, 1996; Shirasu *et al.*, 2000; Wicker *et al.*, 2001; Ma and Bennetzen, 2004), as well as in other eukaryotes (Wessler, 2006). Although cereal genomes display exceptional gene collinearity and conservation, regions surrounding TEs vary extensively, and represent the unstable and species-specific regions of genomes (SanMiguel *et al.*, 1996; Shirasu *et al.*, 2000; Wicker *et al.*, 2001). Initial analyses of our data revealed a relatively even distribution of expansions and contractions across the *O. sativa* genome, conflicting with the hypothesis that these regions would be concentrated in unstable regions of the genome (SanMiguel *et al.*, 1996; Shirasu *et al.*, 2000; Wicker *et al.*, 2001). However, a recent finding in sorghum suggests that 'young' LTR insertions were randomly distributed across the genome, whereas 'older' LTR insertions were concentrated in centromeric regions (Paterson *et al.*, 2009). This result implies that young LTRs are purged quickly from regions of the genome that impact fitness, but remain in centromeric regions. Given the relatively short evolutionary timescale between the species used in our study, many of the expansions we detected may be 'young' in origin, and therefore more widely distributed. To answer this, we found that >64% of the intact *O. sativa* LTR elements had insertion times that corresponded to the predicted radiation of the A-genome lineages, suggesting that these elements are indeed 'young' and may be either more stable or yet to be purged. Although we detected older insertions, it is difficult to draw firm conclusions concerning the timing of bursts or rates of proliferation because the removal rate of older elements is unknown. Furthermore, our data revealed that 13% of the completely intact elements had identical sequences, which suggests the retrotransposon activity in *O. sativa* is recent or ongoing. Mapping of the intact LTR elements revealed that ~75% were in non-centromeric locations, thus providing evidence to support the hypothesis that younger LTRs may be found randomly distributed across the genome (Paterson *et al.*, 2009).

Interestingly, these mobile elements may also be important in new gene formation and genome evolution. Helitron-related transposable elements have been shown to carry pseudogenes in maize, and to participate in the expansion and evolution of the maize genome (Gupta *et al.*, 2005; Lai *et al.*, 2005; Morgante *et al.*, 2005; Yang and Bennetzen, 2009). Similarly, Pack-MULE transposable elements were shown to carry fragments of cellular genes (Jiang *et al.*, 2004), and recently MITEs were implicated in altering gene function and regulation through insertion into rice introns (Oki *et al.*, 2008). Our analysis showed that the expanded regions of the *O. sativa* genome were enriched in TEs and were lower in gene content, but the genes in these regions showed an over-representation of GO terms related to defense against stresses from the environment and pathogens in the total and single gene sets, but not the multiple copy gene set (Shadle *et al.*, 2003; Zhuang and Liu, 2004; Wang and Duan, 2009; Wilson and Talbot, 2009), indicating a possible tie between these rapidly evolving genes and transposable elements. The birth–death theory of gene evolution posits that new genes are created through duplication followed by random mutation (Ohno, 1970; Nei *et al.*, 1997; Michelmore and Meyers, 1998; Nei and Rooney, 2005), and are maintained by natural selection for function that confers a selective advantage. Under artificial selection the same theory may hold true with selection for agronomically important traits. Therefore, the genes that we found to be over-represented in the expanded regions of *O. sativa*, but not the other *Oryza* genomes, may reflect recent gene acquisition events, and are prime candidates for studying the underlying mechanisms of new gene origination because the molecular signatures are likely to have remained intact. Also, as these genes are in SV regions rich in TEs, then the role that TEs play in gene or gene fragment movement, and gene formation of these genes, can be examined in detail. Currently, we are comparing the identified expanded *O. sativa* ssp. *japonica* region sequences with orthologous regions from *O. sativa* ssp. *Indica*, which may shed some light on these theories.

In contrast, we found genes related to essential functions, like the mechanisms and control of nutrient uptake through transport (Miller *et al.*, 2009), were significantly under-represented in the multiple copy and total gene sets, and therefore may be under strong purifying selection. The 140 inversions that we identified in the three species may also be evolutionarily significant. Inversions can interfere with interspecific hybridization, and can act to maintain genetic linkage between genes across multiple generations. Hence one can imagine inversions being generated and then selected for during the process of cultivation and improvement. As full-genome sequences become available for these species, analyses of structural variations in combination with gene copy number will be important to distinguish whether high copy number genes have less selective disadvantage than low copy number genes in expanded regions, and their implications on function.

Although our method for detecting SV is robust and applies stringent criteria, our results were dependent on the quality of the *O. sativa* reference sequence assembly, BES alignment to the *O. sativa* genome and accurate BAC sizes. The *O. sativa* sequence assembly represents over 95.4% of the rice genome, with over 10 times coverage; therefore, the majority of the genome is finished and represents a low threshold of error. We were also able to unambiguously map 74% of the BESs representing 16, 11 and 11% coverage of the *O. sativa* genome for *O. nivara*, *O. rufipogon* and *O. glaberrima*, respectively, indicating a high level of sequence conservation among these close relatives. Additionally, BESs were evenly distributed across the *O. sativa* genome with no major bias in coverage, thus supporting the subsequent predictions of SV. However, as the BES pairs were filtered for highly repetitive matches, some repetitive regions of the genome were not represented. Likewise, BESs of *O. nivara*, *O. rufipogon* and *O. glaberrima* that were absent in the *O. sativa* genome sequence were not included because completed genome sequences for these *Oryza* genomes are not currently available. Finally, our analysis is dependent on accurate BAC sizing because the algorithm for determining structural variation uses a range based on two standard deviations from the mean BAC clone size to differentiate concordant and discordant BES pairs, and find regions of structural variation.

As population data and full genome sequence data become available for each of the species in the genus *Oryza,* we will be able to address detailed questions regarding the symmetry and asymmetry of SV within the genus. Two projects are currently underway to address these needs. The first is a vertical data set of chromosome 3 short-arm sequences for three AA genomes, one BB genome, one CC genome and a BBCC tetraploid using traditional Sanger-based sequencing methods. The second project will provide the full genome sequence for *O. glaberrima* using next generation sequencing technologies (Wing, unpublished data). These data will provide an exceptional resource for answering questions related to the introgression and recombination of cultivated rice and its wild relatives, and will allow for analysis into the mechanisms of genome evolution and domestication. Additionally, examination of different accessions from each *Oryza* species can provide insight into regional differences that may be fundamental to crop improvement and sustainable agriculture (Fitzgerald *et al.*, 2009; Izawa *et al.*, 2009; Belo *et al.*, 2010).

In summary, this detailed catalog of SV across the *O. sativa* genome, in conjunction with the associated physical maps and BAC clone libraries from the other *Oryza* species, are an invaluable resource for both rice breeders and evolutionary biologists. Future functional studies of these regions may include positional cloning and crop improvement through interspecific hybridization and gene transformation. From an evolutionary perspective, these regions are fodder for future analyses into the detailed mechanisms of speciation and domestication in rice, and for the roles of artificial and natural selection in these genomes.

## EXPERIMENTAL PROCEDURES

The BESs for three *Oryza* genomes (*O. rufipogon*, *O. nivara* and *O. glaberrima*; Kim *et al.*, 2008) were downloaded from the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov). This consisted of 243 927 BESs representing 117 452 distinct BAC clones with paired BESs. A rearrangement catalog for each species, compared with *O. sativa,* was constructed using a four-step process: (i) alignment of BESs to the *O. sativa* pseudomolecules; (ii) classification of clones based on paired BAC end sequence alignments; (iii) detection of putative rearrangements; and (iv) classification of rearrangements. Brief descriptions of these methods are detailed below; further details on these procedures are available in Appendix S1. The source code for the custom PERL software developed to perform this analysis is available at Google Projects (http://code.google.com/p/omap-structural-variation).

### Classification of clones based on paired BES alignments

Paired BESs from the three genomes were aligned to the *O. sativa* reference genome IRGSP V3, using BLAT (–minScore = 160), followed by PslReps (–minAli = 0.90 – nearTop = 0.01 – single hit; Table 1) (Kent, 2002). Paired BES alignments were classified (Table 8) using an algorithm based on graph theory followed by highest-ranking alignment for each clone. BES pairs were concordant if the distance between them was within two standard deviations of the average insert size of each BAC library, and were in the opposite orientation; BES pairs were discordant if they aligned in the same orientation, suggesting that there was an inversion in the sequence or that a pair of BESs mapped either closer or further than the expected size range on the reference genome.

To determine the expected distance between paired BESs for each species, BAC sizes were estimated by multiplying the number of bands from the physical map data by the estimated band size (1195 bp), and analyzed to determine an acceptable insert size range based on two standard deviations from the mean. Regions of each genome that were potentially part of a translocation or transposition event were avoided by removing maximum sized clones. We found that 62% of the paired BES reads aligned to the *O. sativa* genome were concordant, 29% were discordant and 9% were discarded because they were not located in the physical map contigs. BESs from clones with only a single BES alignment were analyzed for repeat content using data from Kim *et al.* (2008) to determine why the alignments were filtered out.

### Detection of putative rearrangements

A clustering algorithm was applied to clones with discordant BES pairs, and generated clusters of two or more clones that were required to be from the same FPC contig and aligned together in the *O. sativa* genome. The following formula was applied to check the aligned distance between the clones, where ($x_1$, $y_1$) are the coordinates of the first clone and ($x_2$, $y_2$) are the coordinates of the second clone:

$$|x_2 - x_1| + |y_2 - y_1| \leq 2\,(\text{mean insert size})$$

### Classification of rearrangements

Clusters were classified into rearrangement types: contraction, expansion or inversion, relative to the *O. sativa* reference genome. Clusters in which clones had BESs that were too close together relative to the *O. sativa* reference sequence were identified as contractions; clusters that were too far apart from each other were identified as expansions. Inversions were identified when clusters contained clones with BES pairs that were in the same orientation. All computationally identified rearrangements were listed in Table S1.

### BAC clone size determination

All BAC clones and materials used in this study are available to the public from the AGI Resource Center (http://www.genome.arizona.edu/orders). A new BAC clone sizing method, using controlled nicking (see Appendix S2), was developed to quickly visualize open circular and linearized plasmids (the linearized lower band was used for size determination; Figure S4). The method was validated using BACs from *O. sativa* chromosome 3 (Appendix S2; Figures S5 and S6; Tables S8 and S9).

### Confirmation of expansion and contraction regions

Putative expansions or contractions and accurate sizes of each BAC clone were determined as described above. The distances between BES alignments in the *O. sativa* reference sequence were compared with the empirical BAC sizes to confirm the expansions and contractions. BAC clone size differences were classified as a 'contraction' if the gel-based sizes were larger than the end sequence alignment sizes; likewise, they were classified as an 'expansion' if the gel-based sizes were smaller than the end sequence alignment sizes. Regions on the *O. sativa* reference sequence, from the computational analysis, were adjusted to match regions from confirmed expansion and contraction BAC clones, and are shown in Table S2.

### Analysis of TE, repeats and genes in the expanded regions of *O. sativa* identified from comparisons with *O. nivara*, *O. rufipogon* and *O. glaberrima*

Confirmed expansions in *O. sativa* (Table S10) were analyzed for TEs, repeats and genes, and compared with expected values for *O. sativa* reference sequences of the same size based on average gene and repeat content (International Rice Genome Sequencing Project, 2005). Repeats were detected using RepeatMasker (Smit, Hubley, Green, RepeatMasker Open-3.0 1996–2010, http://www.repeatmasker.org) with the following settings: maximum divergence 25%; low complexity and contaminants skipped. Repeat content was expressed as the percentage of nucleotides masked versus the total.

We estimated the insertion time (SanMiguel *et al.*, 1998) for complete LTR-RT elements isolated in the expanded regions of *O. sativa* (in comparison with *O. nivara*) and analyzed the genes in these regions for single copy or duplication using BLAST set to 70% coverage and 50% identity. LTRs were identified using the software LTR finder (Xu and Wang, 2007), and the mutation rate was set at $1.3 \times 10^{-8}$ (Ma and Bennetzen, 2004).

The databases used were: TIGR *Oryza* repeats (ftp://ftp.plantbiology.msu.edu/pub/data/TIGR_Plant_Repeats/TIGR_Oryza_Repeats.v3.3), Repbase version 11_09 (Jurka *et al.*, 2005) and a proprietary collection of LTR-retroelements isolated in *O. sativa* and other *Oryza* species.

### Gene analysis

The *O. sativa* expanded regions were examined for gene content by calculating the overlap in base pairs between the putative expanded regions and the IRGSP V3 gene models (http://rgp.dna.affrc.go.jp/IRGSP/Build3/build3.html). Genes in gene families or duplicated across the genome were separated from a single copy gene set using an all-versus-all blast search of *O. sativa* genes (default parameters), and filtering hits to another gene that had >50% identity and 70% coverage. Over- or under-representation of GO terms, as compared with the rest of the genome, were identified using Cytoscape (Shannon *et al.*, 2003) with the plug-in BiNGO (Biological Networks Gene Ontology; Maere *et al.*, 2005). A hypergeometric distribution test was used to impose a *P*-value calculation (≥0.05) to ensure that random results were ignored. For multiple hypothesis testing, and to reduce the number of false positives, a false discovery rate (FDR) correction (Benjamini and Hochberg, 1995) was applied. Of the 37 544 IRGSP V3 gene models used, 14 997 (40%) had GO terms associated based on an InterProScan from RAP-DB (http://rapdb.dna.affrc.go.jp). Of the 12 981 gene models in the *O. sativa* expanded regions, 3088 (24%) had GO annotation used in our BiNGO analysis.

## ACKNOWLEDGEMENTS

We thank Noreen Lyle, Angelina Angelova, Elizabeth Ashley, Kristi Collura and Nick Sisneros for excellent technical assistance, and Jetty S.S. Ammiraju for helpful discussions. We thank Ken Youens-Clark for reviewing the Perl code and providing helpful computational advice. This work was supported by the National Science Foundation Plant Genome Program, Grant No. 0208329 to SAJ, RAW and LS.

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

**Figure S1.** Size ranges of confirmed contractions in *Oryza sativa* derived from alignments of BACs from other *Oryza* species.

**Figure S2.** Size ranges of confirmed expansions in *Oryza sativa* derived from alignments of BACs from other *Oryza* species.

**Figure S3.** Histogram showing estimations of the insertion times of intact complete LTR-RT elements *Oryza sativa* expanded regions (from comparison with *Oryza nivara*).

**Figure S4.** CHEF gel picture of separated BAC clone DNA bands produced by the control nicking method.

**Figure S5.** Validation of the control nicking BAC clone DNA sizing method using sequenced *Oryza sativa* BACs from chromosome 3.

**Figure S6.** High-throughput production CHEF gel showing control nicking sizing of *O. nivara* BAC clones.

**Table S1.** Summary table of computationally derived expansion, contraction and inversion regions of *Oryza sativa*.

**Table S2.** Summary table of all confirmed expansion and contraction regions of *Oryza sativa*.

**Table S3.** Summary of repeat analysis of identified expansion regions from *Oryza nivara* BAC end sequences (BESs) and pseudo-BESs of *Oryza sativa*.

**Table S4.** *Oryza nivara* BAC clones aligned to the identified expanded regions of *Oryza sativa* pseudomolecules that were analyzed for repeats.

**Table S5.** Locations and classifications of repeats found in the orthologous expanded regions of *Oryza sativa* pseudomolecules.

**Table S6.** Chromosomal mapping of complete long terminal repeat (LTR) elements found in expanded regions of *O. sativa.*

**Table S7.** Over- and under-representation of gene ontology (GO) terms in expanded regions of *Oryza sativa.*

**Table S8.** Empirically determined size data of *Oryza sativa* BAC clones from *O. sativa* chromosome 3.

**Table S9.** Comparison of BAC clone sizes determined by *Not*I digestion and control nicking methods.

**Table S10.** Expansion regions of pseudomolecules identified by confirmed contractions in the *Oryza* species.

**Appendix S1.** Construction of a rearrangement catalog.

**Appendix S2.** Control nicking method for sizing BAC clones and BAC clone size determination.

Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

## REFERENCES

**Adams, K.L. and Wendel, J.F.** (2005) Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* **8**, 135–141.

**Ammiraju, J.S., Luo, M., Goicoechea, J.L. *et al.*** (2006) The *Oryza* bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza. *Genome Res.* **16**, 140–147.

**Ammiraju, J.S., Lu, F., Sanyal, A. *et al.*** (2008) Dynamic evolution of *Oryza* genomes is revealed by comparative genomic analysis of a genus-wide vertical data set. *Plant Cell*, **20**, 3191–3209.

**Belo, A., Beatty, M.K., Hondred, D. *et al.*** (2010) Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theor. Appl. Genet.* **120**, 355–367.

**Benjamini, Y. and Hochberg, T.** (1995) Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **85**, 289–300.

**Bennetzen, J.L.** (2002) Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica*, **115**, 29–36.

**Bennetzen, J.L.** (2007) Patterns in grass genome evolution. *Curr. Opin. Plant Biol.* **10**, 176–181.

**Bennetzen, J.L., Ma, J. and Devos, K.M.** (2005) Mechanisms of recent genome size variation in flowering plants. *Ann. Bot. (Lond)* **95**, 127–132.

**Bhutkar, A., Schaeffer, S.W., Russo, S.M. *et al.*** (2008) Chromosomal rearrangement inferred from comparisons of 12 *Drosophila* genomes. *Genetics*, **179**, 1657–1680.

**Brar, D.S. and Khush, G.S.** (1997) Alien introgression in rice. *Plant Mol. Biol.* **35**, 35–47.

**Chen, M., SanMiguel, P., de Oliveira, A.C. *et al.*** (1997) Microcolinearity in sh2-homologous regions of the maize, rice, and sorghum genomes. *Proc. Natl. Acad. Sci. USA* **94**, 3431–3435.

**Chen, F.C., Chen, C.J., Li, W.H. *et al.*** (2007) Human-specific insertions and deletions inferred from mammalian genome sequences. *Genome Res.* **17**, 16–22.

**Ding, J., Araki, H., Wang, Q. *et al.*** (2007) Highly asymmetric rice genomes. *BMC Genomics*, **8**, 154.

**Doebley, J.F., Gaut, B. and Smith, B.D.** (2006) The molecular genetics of crop domestication. *Cell*, **127**, 1309–1321.

**Feltus, F.A., Wan, J., Schulze, S.R. *et al.*** (2004) An SNP resource for rice genetics and breeding based on subspecies *indica* and *japonica* genome alignments. *Genome Res.* **14**, 1812–1819.

**Feuillet, C. and Keller, B.** (1999) High gene density is conserved at syntenic loci of small and large grass genomes. *Proc. Natl. Acad. Sci. USA*, **96**, 8265–8270.

**Fitzgerald, M.A., McCouch, S.R. and Hall, R.D.** (2009) Not just a grain of rice: the quest for quality. *Trends Plant Sci.* **14**, 133–139.

**Ge, S., Sang, T., Lu, B.R. *et al.*** (1999) Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *Proc. Natl. Acad. Sci. USA*, **96**, 14400–14405.

**Gupta, S., Gallavotti, A., Stryker, G.A. *et al.*** (2005) A novel class of Helitron-related transposable elements in maize contain portions of multiple pseudogenes. *Plant Mol. Biol.* **57**, 115–127.

**Hannes, F.D., Sharp, A.J., Mefford, H.C. *et al.*** (2009) Recurrent reciprocal deletions and duplications of 16p13.11: the deletion is a risk factor for MR/MCA while the duplication may be a rare benign variant. *J. Med. Genet.* **46**, 223–232.

**Helbig, I., Mefford, H.C., Sharp, A.J. *et al.*** (2009) 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nat. Genet.* **41**, 160–162.

**International Rice Genome Sequencing Project** (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.

**Itoh, T., Tanaka, T., Barrero, R. *et al.*** (2007) Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with Arabidopsis thaliana. *Genome Res.* **2007**, 17.

**Izawa, T., Konishi, S., Shomura, A. *et al.*** (2009) DNA changes tell us about rice domestication. *Curr. Opin. Plant Biol.* **12**, 185–192.

**Jiang, N., Bao, Z., Zhang, X. *et al.*** (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, **431**, 569–573.

**Jurka, J., Kapitonov, V.V., Pavlicek, A. *et al.*** (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467.

**Kazazian, H.H. Jr** (2004) Mobile elements: drivers of genome evolution. *Science*, **303**, 1626–1632.

**Kellis, M., Patterson, N., Endrizzi, M. *et al.*** (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

**Kellogg, E.** (2009) The Evolutionary history of *Ehrhartoideae*, *Oryzeae*, and *Oryza*. *Rice*, **2**, 1–14.

**Kent, W.J.** (2002) BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664.

**Kidd, J.M., Cooper, G.M., Donahue, W.F. *et al.*** (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56–64.

**Kim, H., Hurwitz, B., Yu, Y. *et al.*** (2008) Construction, alignment and analysis of twelve framework physical maps that represent the ten genome types of the genus *Oryza*. *Genome Biol.* **9**, R45.

**Kishine, M., Suzuki, K., Nakamura, S. *et al.*** (2008) Grain qualities and their genetic derivation of 7 new rice for Africa (NERICA) varieties. *J. Agric. Food. Chem.* **56**, 4605–4610.

**Lai, J., Li, Y., Messing, J. *et al.*** (2005) Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proc. Natl. Acad. Sci. USA*, **102**, 9068–9073.

**Lewin, H.A., Larkin, D.M., Pontius, J. *et al.*** (2009) Every genome sequence needs a good map. *Genome Res.* **19**, 1925–1928.

**Ma, J. and Bennetzen, J.L.** (2004) Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci. USA*, **101**, 12404–12410.

**Ma, J. and Bennetzen, J.L.** (2006) Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. *Proc. Natl. Acad. Sci. USA*, **103**, 383–388.

**Maere, S., Heymans, K. and Kuiper, M.** (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.

**Martinez, C.P., Arumuganathan, K., Kikuchi, H. and Earle, E.D.** (1994) Nuclear DNA content of ten rice species as determined by flow cytometry. *Jpn. J. Genet.* **69**, 513–523.

**Mefford, H.C. and Eichler, E.E.** (2009) Duplication hotspots, rare genomic disorders, and common disease. *Curr. Opin. Genet. Dev.* **19**, 196–204.

**Michelmore, R.W. and Meyers, B.C.** (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.* **8**, 1113–1130.

**Miller, A.J., Shen, Q. and Xu, G.** (2009) Freeways in the plant: transporters for N, P and S and their regulation. *Curr. Opin. Plant Biol.* **12**, 284–290.

**Moore, G., Devos, K.M., Wang, Z. *et al.*** (1995) Cereal genome evolution. Grasses, line up and form a circle. *Curr. Biol.* **5**, 737–739.

**Morgante, M., Brunner, S., Pea, G. *et al.*** (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat. Genet.* **37**, 997–1002.

**Nei, M. and Rooney, A.P.** (2005) Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* **39**, 121–152.

**Nei, M., Gu, X. and Sitnikova, T.** (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc. Natl. Acad. Sci. USA*, **94**, 7799–7806.

**Newman, T.L., Tuzun, E., Morrison, V.A. *et al.*** (2005) A genome-wide survey of structural variation between human and chimpanzee. *Genome Res.* **15**, 1344–1356.

**Nicholas, T.J., Cheng, Z., Ventura, M. *et al.*** (2009) The genomic architecture of segmental duplications and associated copy number variants in dogs. *Genome Res.* **19**, 491–499.

**Ohno, S.** (1970) *Evolution by Gene Duplication*. Berlin: Springer-Verlag.

**Oki, N., Yano, K., Okumoto, Y. *et al.*** (2008) A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, *Oryza sativa* ssp. *japonica*. *Genes Genet. Syst.* **83**, 321–329.

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

**Paterson, A.H., Bowers, J.E., Bruggmann, R. *et al.*** (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, **457**, 551–556.

**Piegu, B., Guyot, R., Picault, N. *et al.*** (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* **16**, 1262–1269.

**Prasad, V., Stromberg, C.A., Alimohammadian, H. *et al.*** (2005) Dinosaur coprolites and the early evolution of grasses and grazers. *Science*, **310**, 1177–1180.

**Sallaud, C., Gay, C., Larmande, C. *et al.*** (2004) High throughput T-DNA insertion mutagenesis in rice: a first step towards in silico reverse genetics. *Plant J.* **39**, 450–464.

**Salse, J., Bolot, S., Throude, M. *et al.*** (2008) Identification and Characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell*, **20**, 11–24.

**SanMiguel, P., Tikhonov, A., Jin, Y.K. *et al.*** (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science*, **274**, 765–768.

**SanMiguel, P., Gaut, B.S., Tikhonov, A. *et al.*** (1998) The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43–45.

**Shadle, G.L., Wesley, S.V., Korth, K.L. *et al.*** (2003) Phenylpropanoid compounds and disease resistance in transgenic tobacco with altered expression of L-phenylalanine ammonia-lyase. *Phytochemistry*, **64**, 153–161.

**Shannon, P., Markiel, A., Ozier, O. *et al.*** (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504.

**Shimamoto, K. and Kyozuka, J.** (2002) Rice as a model for comparative genomics of plants. *Annu. Rev. Plant Biol.* **53**, 399–419.

**Shirasu, K., Schulman, A.H., Lahaye, T. *et al.*** (2000) A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res.* **10**, 908–915.

**Soderlund, C., Nelson, W., Shoemaker, A. *et al.*** (2006) SyMAP: A system for discovering and viewing syntenic regions of FPC maps. *Genome Res.* **16**, 1159–1168.

**Soltis, D.E. and Soltis, P.S.** (2003) The role of phylogenetics in comparative genetics. *Plant Physiol.* **132**, 1790–1800.

**Springer, N.M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T. *et al.*** (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.* **5**(11), e1000734. doi: 10.1371/journal.pgen.1000734

**Tikhonov, A.P., SanMiguel, P.J., Nakajima, Y. *et al.*** (1999) Colinearity and its exceptions in orthologous adh regions of maize and sorghum. *Proc. Natl. Acad. Sci. USA*, **96**, 7409–7414.

**Tuzun, E., Sharp, A.J., Bailey, J.A. *et al.*** (2005) Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732.

**Varshney, R.K., Hoisington, D.A., Nayak, S.N. *et al.*** (2009) Molecular plant breeding: methodology and achievements. *Methods Mol. Biol.* **513**, 283–304.

**Vaughan, D.A., Morishima, H. and Kadowaki, K.** (2003) Diversity in the *Oryza* genus. *Curr. Opin. Plant Biol.* **6**, 139–146.

**Wang, L. and Duan, G.** (2009) Effect of external and internal phosphate status on arsenic toxicity and accumulation in rice seedlings. *J. Environ. Sci. (China)* **21**, 346–351.

**Wessler, S.R.** (2006) Transposable elements and the evolution of eukaryotic genomes. *Proc. Natl. Acad. Sci. USA*, **103**, 17600–17601.

**Wicker, T., Stein, N., Albar, L. *et al.*** (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. *Plant J.* **26**, 307–316.

**Wilson, R.A. and Talbot, N.J.** (2009) Under pressure: investigating the biology of plant infection by *Magnaporthe oryzae*. *Nat. Rev. Microbiol.* **7**, 185–195.

**Windsor, A.J., Schranz, M.E., Formanova, N. *et al.*** (2006) Partial Shotgun Sequencing of the *Boechera stricta* Genome Reveals Extensive Microsynteny and Promoter Conservation with Arabidopsis. *J. Plant Physiol.* **140**, 1169–1182.

**Xu, Z. and Wang, H.** (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35** (Web Server issue): W265–W268.

**Yan, H., Talbert, P.B., Lee, H.-R. *et al.*** (2008) Intergenic locations of rice centromeric chromatin. *PLoS Biol.* **6**(11), e286. doi: 10.1371/journal.pbio.0060286.

**Yang, L. and Bennetzen, J.L.** (2009) Structure-based discovery and description of plant and animal Helitrons. *Proc. Natl. Acad. Sci. USA*, **106**, 12832–12837.

**Zhao, S., Shetty, J., Hou, L. *et al.*** (2004) Human, mouse, and rat genome large-scale rearrangements: stability versus speciation. *Genome Res.* **14**, 1851–1860.

**Zhuang, J. and Liu, Z.X.** (2004) The evolution of plant disease resistance gene. *Yi Chuan*, **26**, 962–968.