Global Genomic Diversity of *Oryza sativa* Varieties Revealed by Comparative Physical Mapping

Xiaoming Wang,* David A. Kudrna,† Yonglong Pan,* Hao Wang,* Lin Liu,* Haiyan Lin,* Jianwei Zhang,† Xiang Song,† Jose Luis Goicoechea,† Rod A. Wing,† Qifa Zhang,* and Meizhong Luo*,[1]

*National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China, and †Arizona Genomics Institute, University of Arizona, Tucson, Arizona 85721

ABSTRACT Bacterial artificial chromosome (BAC) physical maps embedding a large number of BAC end sequences (BESs) were generated for *Oryza sativa* ssp. *indica* varieties Minghui 63 (MH63) and Zhenshan 97 (ZS97) and were compared with the genome sequences of *O. sativa* spp. *japonica* cv. Nipponbare and *O. sativa* ssp. *indica* cv. 93-11. The comparisons exhibited substantial diversities in terms of large structural variations and small substitutions and indels. Genome-wide BAC-sized and contig-sized structural variations were detected, and the shared variations were analyzed. In the expansion regions of the Nipponbare reference sequence, in comparison to the MH63 and ZS97 physical maps, as well as to the previously constructed 93-11 physical map, the amounts and types of the repeat contents, and the outputs of gene ontology analysis, were significantly different from those of the whole genome. Using the physical maps of four wild *Oryza* species from OMAP (http://www.omap.org) as a control, we detected many conserved and divergent regions related to the evolution process of *O. sativa*. Between the BESs of MH63 and ZS97 and the two reference sequences, a total of 1532 polymorphic simple sequence repeats (SSRs), 71,383 SNPs, 1767 multiple nucleotide polymorphisms, 6340 insertions, and 9137 deletions were identified. This study provides independent whole-genome resources for intra- and intersubspecies comparisons and functional genomics studies in *O. sativa*. Both the comparative physical maps and the GBrowse, which integrated the QTL and molecular markers from GRAMENE (http://www.gramene.org) with our physical maps and analysis results, are open to the public through our Web site (http://gresource.hzau.edu.cn/resource/resource.html).

RICE is a staple food crop worldwide and a model organism for the study of monocots (Li *et al.* 2007). The genus *Oryza* contains 21 wild and 2 cultivated species (*Oryza sativa* and *O. glaberrima*) with 10 genome types (Ammiraju *et al.* 2006; Jacquemin *et al.* 2013). *O. sativa* is composed of two subspecies: *japonica* and *indica* (Han and Xue 2003). The *indica* varieties Minghui 63 (MH63) and Zhenshan 97 (ZS97) contain a number of important agronomic traits and are the parents of Shanyou 63, which is the most widely cultivated hybrid rice in China (Xie *et al.* 2010). Recombinant inbred line (RIL) populations, derived from a cross between MH63 and ZS97, have made great contributions for identifying and analyzing yield-related quantitative trait loci (QTL) (Li *et al.* 2000; Xing *et al.* 2002; Fan *et al.* 2006; Xue *et al.* 2008).

Comparative analysis is an important tool for structural, functional, and evolutionary genomics studies. Whole-genome sequences of *japonica* variety Nipponbare and *indica* variety 93-11 have been released and updated (Yu *et al.* 2002, 2005; International Rice Genome Sequencing Project 2005; Gao *et al.* 2013; Kawahara *et al.* 2013), and massive comparisons between these two genomes have been carried out to reveal the genetic variations (Han and Xue 2003; Feltus *et al.* 2004; Ma and Bennetzen 2004; Yu *et al.* 2005; Ding *et al.* 2007; Huang *et al.* 2008). The *Oryza* Map Alignment Project (OMAP) has constructed bacterial artificial chromosome (BAC) libraries and BAC-based physical maps for 17 of 23 *Oryza* species representing all the 10 genome types (Ammiraju *et al.* 2006, 2010b; Kim *et al.* 2008; Jacquemin *et al.* 2013). These BAC-based resources make whole or targeted genomic comparisons available and shed light

doi: 10.1534/genetics.113.159970
Manuscript received November 20, 2013; accepted for publication January 7, 2014; published Early Online January 14, 2014.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.159970/-/DC1.
The BAC end sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. KG702200–KG737748 for Minghui 63 and KG737749–KG771717 for Zhenshan 97.
[1]Corresponding author: National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China. E-mail: mzluo@mail.hzau.edu.cn

on the genomic variation and evolution (Kim *et al.* 2007; Zhang *et al.* 2007; Ammiraju *et al.* 2008; Feng *et al.* 2009; Lu *et al.* 2009; Ammiraju *et al.* 2010a; Hurwitz *et al.* 2010). With the development of next generation sequencing technology, it is now cost effective and possible to extensively detect small nucleotide variations, such as SNPs and indels (Huang *et al.* 2010, 2012).

Although many studies have been made for comparisons within *O. sativa*, those comparative genomics studies were mainly focused on the small nucleotide variations. Genome-wide comparisons at the structural level (such as expansion/contraction, inversion, etc.) were limited due to the lack of suitable resources. Previously we made a physical mapping comparison of two *japonica* varieties, Nipponbare and Zhonghua 11 (ZH11), that revealed substantial variations (SNPs, indels, and expansions/contractions) (Lin *et al.* 2012). Subsequently we constructed a BAC-based physical map of 93-11 that provided a resource for completion of the 93-11 reference sequence and robust intersubspecies comparisons (Pan *et al.* 2013). Although the varieties in both subspecies display high genetic diversity, especially in *indica* groups (Garris *et al.* 2005; Huang *et al.* 2012), to date, no suitable genomic resources of other *indica* varieties are available to carry out intra-subspecies comparisons with 93-11 and differentiate the real intersubspecies variations from the variety-specific variations between 93-11 and Nipponbare.

MH63 and ZS97 are typical *indica* varieties and important materials in functional genomics studies. Here we report the generation and application of the BAC libraries, BAC end sequences (BESs), and BAC-based physical maps of MH63 and ZS97. With these new resources, we detected intersubspecies BAC-sized and contig-sized structural variations by comparing these two physical maps with the Nipponbare reference sequence, and intersubspecies and intra-*indica* subspecies polymorphic simple sequence repeats (SSRs), SNPs, multiple nucleotide polymorphisms (MNPs), and indels by comparing these two BES data sets with the Nipponbare and 93-11 reference sequences. Using the physical maps of *O. glaberrima* (accession no. International Rice Germplasm Center (IRGC) 96717), *O. rufipogon* (*OR105491*) (accession no. IRGC 105491), *O. rufipogon* (*OR106424*) (accession no. IRGC 106424), and *O. nivara* (accession no. IRGC 100897, W0106) from OMAP (http://www.omap.org) as a control, we identified many conserved and divergent regions related to the evolution and domestication processes. Both the comparative physical maps and the GBrowse, which integrated the QTL and molecular markers from GRAMENE (http://www.gramene.org) with our physical maps and analysis results, are open to the public through our Web site (http://gresource.hzau.edu.cn/resource/resource.html).

Materials and Methods

BAC library construction, end sequencing, fingerprinting, and contig assembly

The BAC libraries of MH63 and ZS97 were constructed as previously described (Luo *et al.* 2001; Luo and Wing 2003; C. Wang *et al.* 2013). The megabase genomic DNA was isolated from young seedlings. The vector, pAGIBAC1 with *Hin*dIII, was prepared as previously described (Luo *et al.* 2006). The *Escherichia coli* DH10B T1-resistant competent cells were used as a host, and the clones were arranged into 384-well plates. Copies of the two BAC libraries were stored at both the Arizona Genomics Institute (www.genome.arizona.edu) and the Genome Resource Laboratory of Huazhong Agricultural University (http://gresource.hzau.edu.cn). For insert sizing, 220 clones of each library were randomly selected, and the plasmids were restricted with *Not*I and analyzed on 1% agarose CHEF gels (Bio-Rad, Hercules, CA) with a 5- to 15-sec linear ramp time at 6 V/cm and 14° in 0.5× TBE buffer for 16 hr. BAC end sequencing and fingerprinting were performed following previous protocols (Luo *et al.* 2003; Kim *et al.* 2007; Lin *et al.* 2012; X. Wang *et al.* 2013). BESs with a quality score less than Phred 16 or total length <100 bp were removed. All of the BESs are deposited in GenBank (KG702200–KG737748 for MH63 and KG737749–KG771717 for ZS97). The clones that produced <25 or >180 fingerprint bands were excluded in contig assembly. The clones were assembled into contigs with the program FPC (Soderlund *et al.* 2000).

Contig alignment and manual editing

The Os-Nipponbare-Reference-IRGSP-1.0 and updated 93-11 whole-genome sequences were used as reference sequences (Gao *et al.* 2013; Kawahara *et al.* 2013). The physical maps of MH63 and ZS97 were aligned to reference sequences and displayed by SyMAP (Soderlund *et al.* 2006). The contigs and alignments were manually edited and improved against the Nipponbare reference sequence as described by Kim *et al.* (2007) and Lin *et al.* (2012).

Comparative analysis of physical maps and integration of resources

The BAC-sized structural variations were detected as described by Hurwitz *et al.* (2010) with some modifications. We used a more conservative cutoff value (insert size ranges 75–190 kb for MH63 and ZS97 and 70–190 kb for 93-11) to eliminate false positives. Also, we added the following criteria to acquire more precise results:

1. The expansion/contraction clones, with differences of <20 kb between the length on the contig according to average consensus band (CB) size and its hit length on the reference sequence, were excluded.
2. The hit location of a discordant clone on the reference sequence was restricted to within the hit location of the corresponding contig on the reference sequence (defined by SyMAP on comparative physical maps). The two independent alignments (BESs alignment and SyMAP alignment) ensured that the BES could be mapped to the correct location on the reference sequence.
3. The clones whose hit regions spanned the gaps on the reference sequence were removed from the results, considering the imprecise sizes of the gaps.

4. The latest edition of Os-Nipponbare-Reference-IRGSP-1.0 (Kawahara *et al.* 2013) was used for the Nipponbare reference sequence.

The clones selected for expansion/contraction verifications were digested by *Not*I and separated on CHEF gels. The reference sequences of Maize, Sorghum, and *Brachypodium distachyon* were downloaded from the NCBI database (http://www.ncbi.nlm.nih.gov), and the contigs of MH63 and ZS97 were aligned to these reference sequences by SyMAP (Soderlund *et al.* 2006). GBrowse 2.0 (Stein *et al.* 2002) was used to integrate the genomic data.

BES analysis

SSRs in the BESs were identified by SciRoko3.4 (Kofler *et al.* 2007), with the criteria of a minimum repeat number of 3 and a minimum total length of 15 bp. The primers for SSR were designed by Primer3 (Rozen and Skaletsky 2000). Bowtie2 (Langmead and Salzberg 2012) was used to map the SSR primers to reference sequences, and the primers that could be mapped to their original locations in BESs were considered effective ones. At first, a pair of primers, designed for a SSR locus, was mapped to the reference sequence and BES dataset, respectively. If the mapped region in reference sequence and the mapped region in BES contained the same SSR motif but had a different length, this SSR locus were defined as a polymorphism between the reference sequence and the BES. To verify the SSR loci, a total of 21 pairs of primers were randomly selected from the above Primer3-designed primers and used for PCR with the MH63 and ZS97 genomic DNA as templates. Before performing the comparisons of SSR contents and distributions, the BESs of each species were clustered by CAP3 (Huang and Madan 1999) to reduce the redundancy.

The known repetitive sequences (homologous with the identified rice repetitive sequences) within BESs were identified by RepeatMasker (http://repeatmasker.org), using the Repbase of rice (Jurka *et al.* 2005) combined with the rice repetitive sequences downloaded from the Michigan State University Rice Genome Annotation Project (MSU: ftp://ftp.plantbiology.msu.edu/pub/data/TIGR_Plant_Repeats/). As described by Lin *et al.* (2012) and Aggarwal *et al.* (2009), we manually detected the variety-specific repeat sequences in MH63 and ZS97 genomes and annotated them with Blast2GO (Gotz *et al.* 2008). The BESs and physical maps of wild rice were downloaded from OMAP (http://www.omap.org). The small variations (SNP, MNP, insertion, and deletion) between the repeat/organelle sequences-masked BESs and reference sequences were detected with the method and standard of Feltus *et al.* (2004).

Results

Generation of MH63 and ZS97 BAC libraries, BAC end sequences, and fingerprints

To obtain basic genomic resources of MH63 and ZS97, we constructed high-quality and deep-coverage BAC libraries for both varieties. Each library contained 36,864 clones arrayed in ninety-six 384-well plates, and the average insert sizes for the MH63 and ZS97 libraries were 125 and 117 kb, covering ~10.7-fold and ~10.0-fold genome, respectively, based on the 430 Mb genome size.

The 18,432 clones from the first forty-eight 384-well plates of each library were end sequenced bidirectionally and fingerprinted. After quality trimming, 35,549 MH63 BESs with an average length of 659 bp and a cumulative length of 23,421,124 bp and 33,969 ZS97 BESs with an average length of 666 bp and a cumulative length of 22,623,600 bp were successfully generated. In terms of fingerprinting, we successfully generated 17,310 BAC fingerprints for MH63 with an average CB unit size of 1078 bp and 16,580 BAC fingerprints for ZS97 with an average CB unit size of 1114 bp (Supporting Information, Table S1).

Construction of the MH63 and ZS97 physical maps

With the tolerance of 4 and the cutoff of 10^{-25}, 16,580 and 12,965 clones were assembled into 1230 and 1929 contigs, and 730 and 3615 clones were left as singletons for MH63 and ZS97, respectively. Based on the average CB unit size, the total and average contig sizes are 353,085 kb and 287 kb and 393,765 kb and 204 kb for MH63 and ZS97, respectively (Table S1). The primary FPC assembly for each map was referred to as the phase I physical map.

The two phase I physical maps were manually edited, referring to the Nipponbare reference sequence. The resulting phase II physical maps of MH63 and ZS97 consisted of 574 and 1228 contigs, containing 16,735 and 13,972 clones, and leave 575 and 2608 clones as singletons, respectively. The total and average contig sizes are 304,227 kb and 530 kb for MH63 and 353,177 kb and 288 kb for ZS97. In contigs, a total of 15,823 clones for MH63 and 12,118 clones for ZS97 have paired-end sequences that are invaluable resources for comparative genomics (Table S1). The phase II physical maps of MH63 and ZS97 can be downloaded from our Web site.

Comparison of physical maps with the rice reference sequences

Genome-wide BAC-sized structural variations: The genome structural variations could be detected based on the information of paired-end BES hits and the physical maps (Hurwitz *et al.* 2010). We aligned 9563 clones of MH63, 7243 clones of ZS97, and 15,142 clones of 93-11 to the Nipponbare reference sequence through BESs. In Nipponbare, 33 contractions (83 clones) and 140 expansions (464 clones), 69 contractions (201 clones) and 31 expansions (84 clones), and 46 contractions (124 clones) and 180 expansions (726 clones) were identified when MH63, ZS97, and 93-11 BAC clones were aligned to the Nipponbare reference sequence, respectively (Table 1). The structural variations were widely dispersed across the Nipponbare genome, except for the middle region of chromosome 5 in which no

Table 1 Computationally detected structural variations in Nipponbare relative to three *indica* varieties

Species	Total BESs	Total clones with paired-end BESs	Total BES hits[a]	Clones aligned with paired-end BES hits	Clones aligned and in FPC contig (with paired-end BES hits)	Structural variations in the Nipponbare reference sequence					
						Contraction		Expansion		Inversion	
						Total clones	Cluster (clones in cluster)	Total clones	Cluster (clones in cluster)	Total clones	Cluster (clones in cluster)
MH63	35,550	17,336	26,796	10,332	15,185 (9,563)	402	33 (83)	983	140 (464)	199	18 (45)
ZS97	33,970	15,953	24,965	9,256	12,223 (7,243)	623	69 (201)	393	31 (84)	171	18 (44)
93-11	65,140	29,315	47,716	16,906	27,244 (15,142)	481	46 (124)	1,219	180 (726)	551	50 (143)

[a] The BESs were aligned to reference sequences by blat and processed by the PslReps program. The matches were removed if the hit had a gap >40 bp, had <90% coverage, or mapped to three or more locations.

expansion was found (Figure 1). The detailed information of structural variations is shown in File S1.

To validate the putative expansions and contractions, we compared the hit distance of paired-end BESs on the Nipponbare reference sequence with the actual size of each BAC clone measured on the CHEF gel. A total of 186 MH63 clones (from 14 contractions and 50 expansions) and 77 ZS97 clones (from 12 contractions and 18 expansions) were randomly selected and experimentally analyzed. The differences between the actual sizes of all clones and their hit lengths on Nipponbare were confirmed and were >20 kb, except for the 4 clones of one MH63 contraction (File S2).

The regions, which were identified as expansions in the Nipponbare genome relative to one or more of the three *indica* varieties, have higher than average repeat contents (48% in expansions *vs.* 42.73% in the whole genome). The repeats in the expansion regions were mainly LTR Gypsy/DIRS1 (55.98% of expansion repeats *vs.* 48.98% of whole-genome repeats). However, the ratio of DNA transposons in expansion (29.40% of expansion repeats) was lower than the average of the whole genome (34.64% of whole-genome repeats). It is interesting that the gene density of the expansion regions is approximately equal to the value of the whole genome. By analyzing the gene ontology (GO) terms in these expansions, we found a statistically significant ($P < 0.05$) overrepresentation of genes related to response to stimulus in biological processes, especially the genes related to response to stress and abiotic stimulus. We also found an overrepresentation of GO terms of macromolecule biosynthetic process, gene expression, and photosynthesis of metabolic process, under biological processes. The nucleic acid binding and transcription regulator activity categories under molecular function and the nitrogen compound metabolic process under biological processes were found to be underrepresented (File S3). A total of 14 expansion regions and 1 contraction region were shared when the three *indica* varieties were compared with the Nipponbare reference sequence. Of the above shared expansion regions, most were also shared in comparisons of *OR105491*, *OR106424*, and *O. nivara* to the Nipponbare reference sequence (File S4). In these shared expansion regions, 54.93% of the sequences were classified as repeat elements, in which the ratio of Gypsy/DIRS1 was 58.40% and that of DNA transposons was 25.69%, although the gene density was still approximately equal to the value of the whole genome.

Genome-wide contig-sized structural variations: We constructed the intersubspecies comparative physical maps by aligning the MH63 and ZS97 physical map contigs onto the Nipponbare reference sequence, using the BESs embedded in the contigs as bridges (Figure 2). The anchored contigs accounted for 85% and 75% of the total contigs, and the percentage of CB units of anchored contigs to total contigs was 95% and 86% in the comparisons of MH63 and ZS97 to the Nipponbare reference sequence, respectively. A total of 16,356 MH63 clones and 12,845 ZS97 clones were anchored

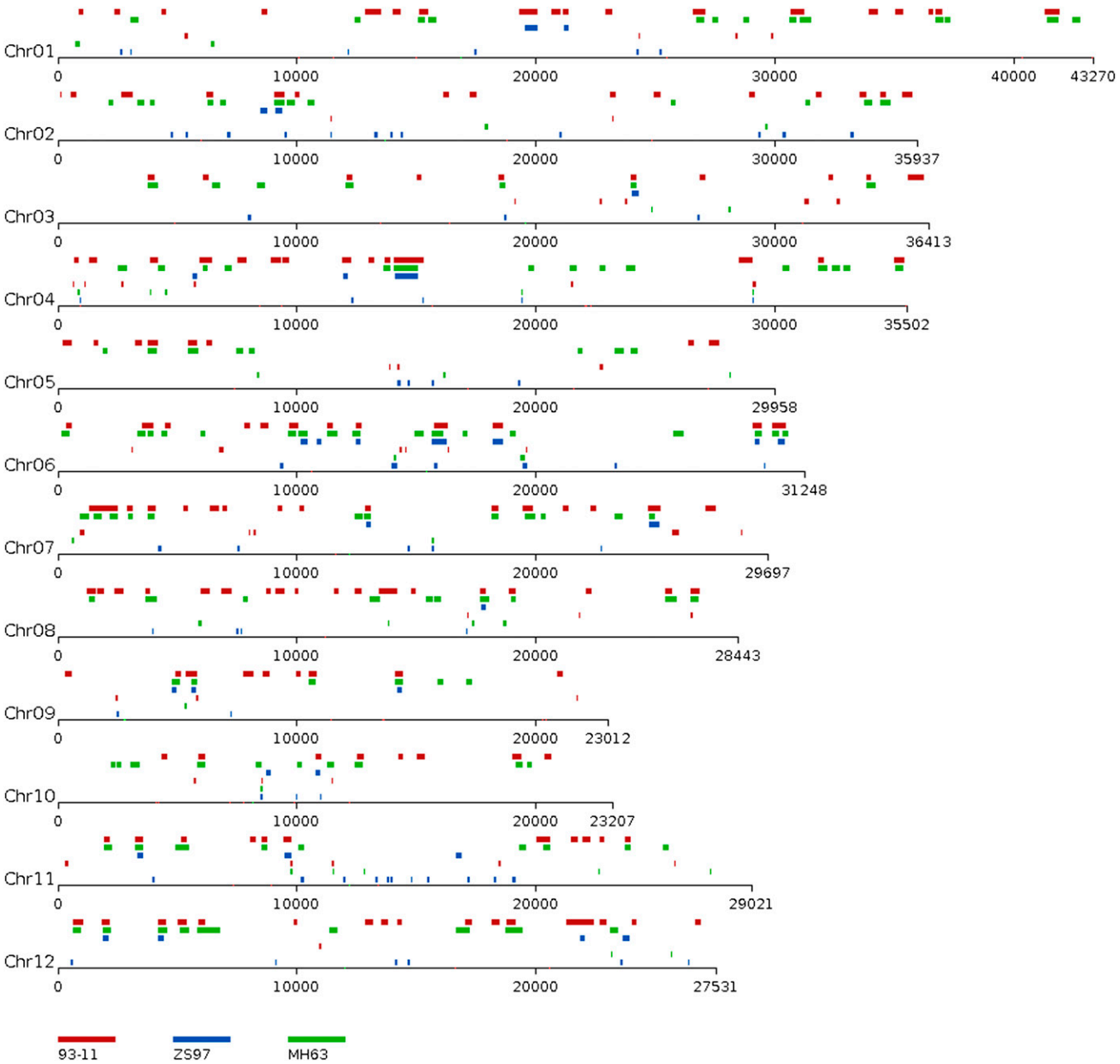

Figure 1 The expansions and contractions in *japonica* variety Nipponbare detected by the genome-wide comparisons with *indica* varieties MH63, ZS97, and 93-11. For each chromosome, lines 1–3 from top to bottom show expansions, and lines 4–6 show contractions.

to the reference sequence, of which 13,456 and 10,654 clones were directly anchored by BES hits (6258 and 4707 clones with paired-end BES hits), respectively (Table 2). The unanchored contigs either were unique regions in MH63 and ZS97 genomes, relative to reference sequences, or corresponded to the gap regions of the reference sequence. We also aligned the MH63 and ZS97 physical map contigs onto the 93-11 reference sequence. However, considering that the 93-11 reference sequence contains many gaps and assembly errors (Pan *et al.* 2013) that would affect the outputs of the comparisons, we displayed only the comparative results on our Web site and do not describe them here.

In anchored contigs, the single-end BES hits were resources that were as invaluable as paired-end BES hits for detecting the structural variations. Based on the CB unit size and distance of BES hits, the expansions and contractions could be visualized by nonparallel alignment lines on comparative physical maps (Lin *et al.* 2012). Figure 3 presents two expansions and two contractions on the Nipponbare reference sequence relative to MH63 and ZS97 contigs. The estimated sizes of the two expansion regions were 213 kb and 203 kb, and those of the two contraction regions were 38 kb and 90 kb, respectively. A user-friendly visualization of the expansions/contractions, with zoom

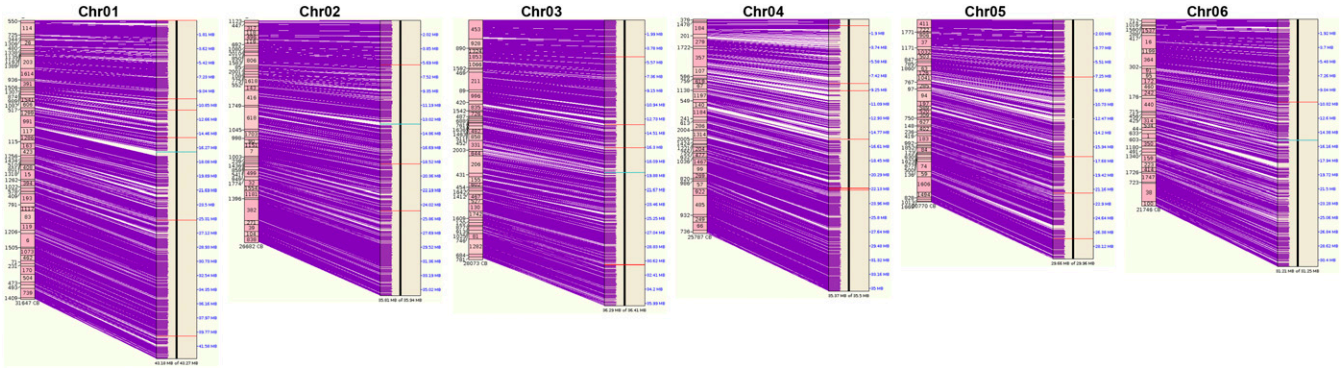

Figure 2 The SyMAP alignments of the MH63 phase II physical contigs to the Nipponbare reference sequence Chr1–Chr6. For each panel, the left boxes represent the contigs of MH63 and the right bar represents the Nipponbare chromosome. The green and red lines on the Nipponbare chromosomes indicate the centromere and gap locations, and the vertical black lines represent the gene locations. The centromere locations of chromosomes 4 and 5 were not available. The purple lines represent the alignments between BESs and reference sequences. The alignments for the remaining six chromosomes are available through our Web site (http://gresource.hzau.edu.cn/resource/resource.html).

features, may be browsed from our Web site. We also found many inversely matched BESs in the two alignment projects. Using the alignment of MH63 contigs to the Nipponbare reference sequence as an example, we detected 1225 inversely matched BESs, and 705 were clustered into 260 groups of ≥2 BES hits (Table 2). The clusters of inversely matched BESs on comparative maps (viewed as cross alignment lines) may reflect genetic or spurious inversions (Lin *et al.* 2012; Deng *et al.* 2013; Pan *et al.* 2013). Figure 4 illustrates a 6844-bp region on the Nipponbare reference sequence [chromosome (Chr)5: 3,550,117–3,556,961] inversely matched by 5 BESs of MH63 contig37. Because the same region was normally matched by BESs of Nipponbare physical contig41 (downloaded from OMAP), and MH63 contig37 has a high quality (with deep coverage and high stringency), this region could most probably contain a genetic inversion between MH63 and Nipponbare. We did PCR experiments to verify this speculation and the results showed that the OSIABa0011B13.r was on the left of the OSIABa0041A04.r in the MH63 contig37. Because the hit location of the OSIABa0011B13.r on the Nipponbare reference sequence was on the right of the hit location of the OSIABa0041A04.r, the genetic inversion between MH63 and Nipponbare was confirmed (Figure 4). All detailed information of the inversely matched BESs is listed in File S5.

Contigs spanning the remaining physical gaps of the Nipponbare reference sequence: Although the Nipponbare reference sequence is well known for its high quality, it still contains 44 physical gaps in the latest version (Kawahara *et al.* 2013). Of these, a total of 41 physical gaps can be spanned by contigs from one to all of our four physical maps (ZH11, 93-11, MH63, and ZS97 physical maps). One gap only could be spanned by MH63 contig and two gaps only could be spanned by ZS97 contigs. The BAC sequences of the gap-corresponding contigs could provide an initial insight into the contents of the gaps and help to close these gaps. Three gaps cannot be spanned by any contigs of the four physical maps (File S6). These three gap-corresponding regions may have complex structures and seem to be conserved in *O. sativa*. Of these three gaps, the seventh physical gap on chromosome 4 also cannot be spanned by any physical map contigs of three ancestors of *O. sativa* (*O. nivara*, *OR105491*, and *OR106424*). However, this gap can be spanned by a physical map contig of *O. glaberrima* that has a farther evolutionary distance with *O. sativa* and the above three mentioned *O. sativa* ancestors (Kovach *et al.* 2007; Huang *et al.* 2012). This result indicates that this gap-corresponding region was conserved in evolution and domestication of *O. sativa*. The other two gaps (the sixth physical gap on chromosome 3 and the eighth physical gap on chromosome 4) can be spanned by physical map contigs of all three wild rice except for the sixth physical gap on chromosome 3 that cannot be spanned by the physical map contig of *OR106424*.

Analyses of MH63 and ZS97 BESs

Simple sequence repeats in BESs: SSRs are effective and robust molecular markers in genetic and genomic analyses. We identified 3781 SSR loci and designed 3157 pairs of primers from 3165 MH63 BESs, and 3750 SSR loci and designed 3163 pairs of primers from 3129 ZS97 BESs. After mapping these primers to the Nipponbare and 93-11 reference sequences, we identified 415 (18%) and 383 (15.9%) polymorphic SSR loci from MH63 and 409 (17.3%) and 325 (12.9%) polymorphic SSR loci from ZS97, respectively (File S7). Ten pairs of MH63 and 11 pairs of ZS97 primers were randomly selected to experimentally verify the SSR loci and all of these SSR loci were confirmed (File S8). The SSRs from BESs of *japonica* variety ZH11, *O. glaberrima*, *OR105491*, *OR106424*, and *O. nivara* were also identified, and the ratios of polymorphic SSRs to the Nipponbare and 93-11 reference sequences were 6.7% and 20.6%, 20.9% and 26%, 19.8% and 22.3%, 19.5% and 22.6%, and 20.7% and 22.8%, respectively (File S7). The detailed

Table 2 Summary of the SyMAP alignments between phase II physical maps of MH63 and ZS97 and the Nipponbare reference sequence

Catalog	MH63	ZS97
Anchored contigs with one placement	489	925
Percentage to total contigs	85	75
Clones in anchored contigs	16,356	12,845
With paired-end BESs	15,477	11,178
With single-end BESs	758	1,415
Total BESs with hits	19,714	15,361
Percentage to total BESs	55.5	45.2
With inversely matched BES hits	1,225	1,589
Hits in cluster (cluster number)	705 (260)	949 (348)
Clones anchored by BESs	13,456	10,654
With paired-end BES hits	6,258	4,707
With single-end BES hits	7,198	5,947
Total CB units of anchored contigs	268,620	273,316
Percentage to total CB units	95	86
Total length of anchored contigs (kb)	289,572	304,474
Covered length of reference sequence (kb)	330,350 (88%)	285,800 (76%)
Unanchored contigs	85	303
Clones in unanchored contigs	379	1,127
With paired-end BESs	347	940
With single-end BESs	22	154
Total CB units of unanchored contigs	13,641	43,719
Percentage to total CB units	5	14
Total length of unanchored contigs (kb)	14,705	48,703

information of SSRs and primers is available from our Web site.

We compared the SSR contents and distributions among the above analyzed species. The results showed that all of the species have similar SSR densities and motif types. For example, the CCG, AG, and AT motifs were predominant in all of the species and accounted for between 27% in *O. glaberrima* and 32% in MH63 (Figure S1). The SSRs composed of AT or AAT motifs showed the most variable repeat length, with an average of 49.5 ± 46.1 bp and 31.6 ± 27.2 bp in MH63 and 56.9 ± 63 bp and 34 ± 24.5 bp in ZS97. Similar to the results for wild rice (Kim *et al.* 2008), A/T-rich dinucleotide and trinucleotide repeat motifs have longer average and variable repeat lengths in cultivated rice than other motifs.

Repeat sequences in BESs: A total of 8,781,060 bp (37.49%) of MH63 BESs and 8,936,689 bp (39.5%) of ZS97 BESs were homologous with the repeat sequences deposited in databases. In terms of repeat category, the retroelement Gypsy/DIRS1 was most common in these two species, comprising 19.93% and 20.95% of total BESs of MH63 and ZS97, respectively (Table S2).

To search variety-specific repeat sequences and reveal their functions, the known repeat/organelle sequences-masked BESs were self-blasted and the high-score pairings (HSPs) sequences were further analyzed (Aggarwal *et al.* 2009; Lin *et al.* 2012). A total of 119 sequences (49,197 bp) in MH63 BESs and 84 sequences (34,813 bp) in ZS97 had at least six HSPs in BESs (25.62 presentation times in 100-Mb MH63 sequence and 26.52 presentation times in 100-Mb ZS97 sequence) and were identified as new repeat sequences by this method. Of these sequences, 38 from MH63 and 29 from ZS97 were considered as specific repeats in *indica* varieties that have less than or equal to five HSPs in the whole Nipponbare reference sequence. With this standard, a total of 12 sequences from MH63 and 15 sequences from ZS97 were regarded as MH63- and ZS97-specific repeats, respectively. Of the specific repeats in *indica* varieties, 22 sequences from MH63 and 18 sequences from ZS97 contained coding regions, most of which were annotated as proteins containing the NBS-LRR domain, a typical domain related to disease resistance in plants. This result indicates that repeat sequences may play an important role in the evolution process of disease-resistance-related genes. The BESs of other *Oryza* species from OMAP were also used to estimate the frequencies of these specific sequences in wild rice (the presentation times per every 100 Mb of BESs were used because the wild rice have different numbers of BESs and genome size), and the result indicates that most of the specific sequences have high frequencies in *OR105491*, *OR106424*, *O. nivara*, and *O. glaberrima*, which have closer distances to *O. sativa* (Ma and Bennetzen 2004; Kovach *et al.* 2007) (File S9).

Substitutions and indels between BESs and the reference sequences: To identify small genetic variations, we compared the repeat-masked BESs of MH63 and ZS97 with their orthologous regions in Nipponbare and 93-11 genomes. A total of 46,349 SNPs, 1007 MNPs, 3806 insertions, and 5134 deletions were identified by the comparisons of MH63 and ZS97 BESs to the Nipponbare reference sequence, and 49.16% of SNPs, 44.59% of MNPs, 39.99% of insertions, and 45.62% deletions were located in Nipponbare gene

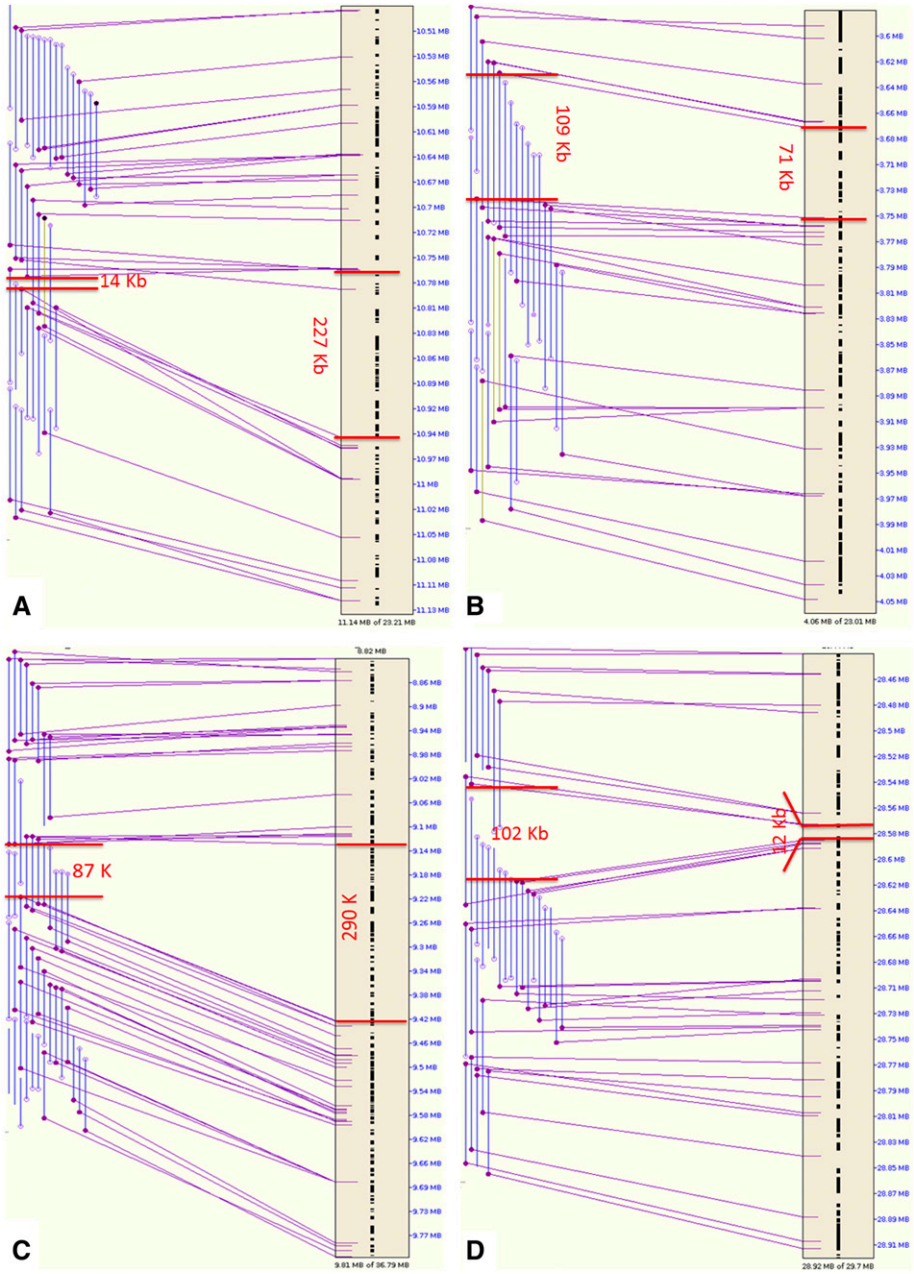

Figure 3 Alignments between MH63 contig1540 (A), MH63 contig514 (B), ZS97 contig244 (C), and ZS97 contig1079 (D) and the Nipponbare reference sequence. For each panel, the left vertical blue lines represent the BAC clones, and the purple lines in the center represent the alignments between the BESs and the reference sequence. A and C exhibit expansions in reference sequence, and B and D exhibit contractions in reference sequence. More structural variations can be browsed through our Web site (http://gresource.hzau.edu.cn/resource/resource.html).

regions. A total of 25,034 SNPs, 760 MNPs, 2534 insertions, and 4003 deletions were identified by the comparisons of MH63 and ZS97 BESs to the 93-11 reference sequence, and 34.35% of SNPs, 27.5% of MNPs, 29.08% of insertions, and 35.12% deletions were located in 93-11 gene regions (File S10). Compared to other variation types, the SNP, followed by the deletion, has the highest frequency of occurring in exon regions. The BES regions encompassed by the 21 pairs of primers, which were used to experimentally verify the SSR loci above, contained 14 SNPs, 5 insertions, and 8 deletions. All these variations were confirmed by sequencing of the PCR products (File S8). The masked BESs of wild species were also used to identify the genetic variations (File S10). The detailed variations are available from our Web site.

In agreement with the findings of Nasu *et al.* (2002) and Feltus *et al.* (2004), the SNPs that we found were not evenly distributed across the Nipponbare chromosomes, and many low- and high-SNP frequency regions could be observed that represent the low- and high-divergence regions (Figure 5). The average SNP frequencies of each Nipponbare chromosome were obviously different and chromosome 4 had the highest SNP frequency (Table S3). A total of 3943 SNPs and 75 MNPs were shared when the three *indica* varieties were compared with the Nipponbare reference sequence, and these variations were putative intersubspecies variations. In these variations, 1111 (28.18%) SNPs and 10 MNPs were located in exon regions, 949 SNPs and 22 MNPs in intron regions, and

490 SNPs and 23 MNPs in 1-kb upstream regions of genes. Of the 3943 SNPs and 75 MNPs shared by the three *indica* varieties, the 1668 SNP and 32 MNP loci were also shared in comparison between *O. nivara* and Nipponbare and the 1230 SNP and 21 MNP loci in comparison between *OR106424* and Nipponbare. The results indicate that these variations may have occurred in Nipponbare or represent the genetic diversity between the ancestors of *japonica* and *indica*.

Resource integration and application

To facilitate functional and comparative genomics studies, we integrated physical maps (MH63, ZS97, 93-11, ZH11, *O. nivara*, *OR105491*, *OR106424*, and *O. glaberrima*) and analysis results with the information of QTL and molecular markers from GRAMENE into GBrowse. The contigs, clones, SNPs, MNPs, indels, and polymorphic SSRs were linked with molecular markers, QTL, and gene models and were aligned to the Nipponbare reference sequence. The discordantly matched clones indicated the high-divergence regions in the Nipponbare genome. To allow searching the collinear regions, we aligned the above eight physical maps to the genome sequences of Maize, Sorghum, and *B. distachyon* to allow searching the collinear regions and carrying out comparative genomic studies. The GBrowse and comparative physical maps are available to the public at our Web site.

Discussion

The *O. sativa* ssp. *indica* cv. MH63 line, which was bred from a cross between IR30 and Gui630, is an elite restoring line and has the following characters: strong ability and broad spectrum in restoration, resistance to rice blast, and high production yields. The *indica* cv. ZS97 is one of the most widely used sterile lines. Together, MH63 and ZS97 are the parents of Shanyou 63, which is the most widely cultivated hybrid rice in China (Xie *et al.* 2010). Based on the RIL populations of MH63 and ZS97, Xing *et al.* (2002) constructed a genetic linkage map and Xie *et al.* (2010) constructed an ultrahigh-density linkage map. With these RIL populations as materials, many important agronomic trait-related genes and QTL have been identified and analyzed, such as *GS3* (Fan *et al.* 2006; Mao *et al.* 2010), *Ghd7* (Xue *et al.* 2008), and *Xa26* (Sun *et al.* 2004). The integrated genomic and genetic resources reported here provided a platform for identifying and analyzing other important agronomic traits possessed by the two varieties.

Genome-wide comparative analysis at the structural level would facilitate the understanding of the evolution and domestication processes of cultivated rice. In other model systems, analyses of genome-wide structural variations have proved to be an effective strategy for uncovering the evolution process and relating the genetic diversity to the given phenotypes (Zhao *et al.* 2004; Tuzun *et al.* 2005; Kidd *et al.* 2008; Nicholas *et al.* 2009; Springer *et al.* 2009; Zhang *et al.* 2011). The BAC libraries, BESs, physical maps of MH63 and ZS97, and the integrated information provided here are invaluable resources for inter- and intra-subspecies genome-wide comparative analysis in *O. sativa*.

The studies from OMAP shed light on the genome-wide structural variations between Nipponbare and wild rice species (Kim *et al.* 2007, 2008; Hurwitz *et al.* 2010). Previously we constructed the physical maps of ZH11 and 93-11 and compared them with the Nipponbare and 93-11 reference sequences (Lin *et al.* 2012; Pan *et al.* 2013). Here we identified BAC-sized and contig-sized structural variations in *O. sativa* by comparisons of the physical maps of MH63 and ZS97 to the Nipponbare reference sequence. The shared structural variations between the Nipponbare genome and the three *indica* varieties were putative intersubspecies genetic diversities that may have occurred before divergence of these *indica* varieties. Huang *et al.* (2012) provided the model that *japonica* rice was first domesticated and then *indica* rice was generated from the cross between *japonica* and wild rice. Among the above shared regions, some could be conserved between the *indica* group and its wild ancestors (the regions that were also shared in one or more of three relative wild species) and not interrupted by the gene introgression of early *japonica*, whereas others could be diverged immediately after the cross between *japonica* and wild rice (the regions that were not shared with wild species).

Transposable elements (TEs) are one of the major contributing genomic features that relate to an organism's genome size and its course of evolution (Ma and Bennetzen 2004). In our analysis, the expansion regions contained more repeat content, especially the LTR element Gypsy/DIRS1, but the gene density in these regions was approximately the same as the average value of the whole genome. By skimming the gene annotations, we observed a number of LTR retrotransposon-related genes in the expansion regions and a statistically significant difference of GO term representations, such as the genes related to response to stress and abiotic stimulus. Most of the unique repeat sequences in the *indica* group contained the motif of NBS-LRR, which has been reported to be a typical domain related to disease resistance. These results further demonstrate that TEs, especially LTR retrotransposons, may contribute largely to gene formation, genome evolution, genetic diversity, and asymmetry in *O. sativa*.

Our results exhibited that the Nipponbare reference sequence contained more expansions than contractions of BAC and contig sizes relative to MH63 and 93-11 physical maps. Han and Xue (2003) and Ma and Bennetzen (2004) also found more insertions in Nipponbare than in *indica* variety GLA4 and proposed that the Nipponbare genome was expanded relative to *indica*. In the *japonica* group, Lin *et al.* (2012) reported that the Nipponbare genome is ~7% larger than ZH11. The larger genome size of Nipponbare could be mainly due to the amplification of LTR retrotransposons,

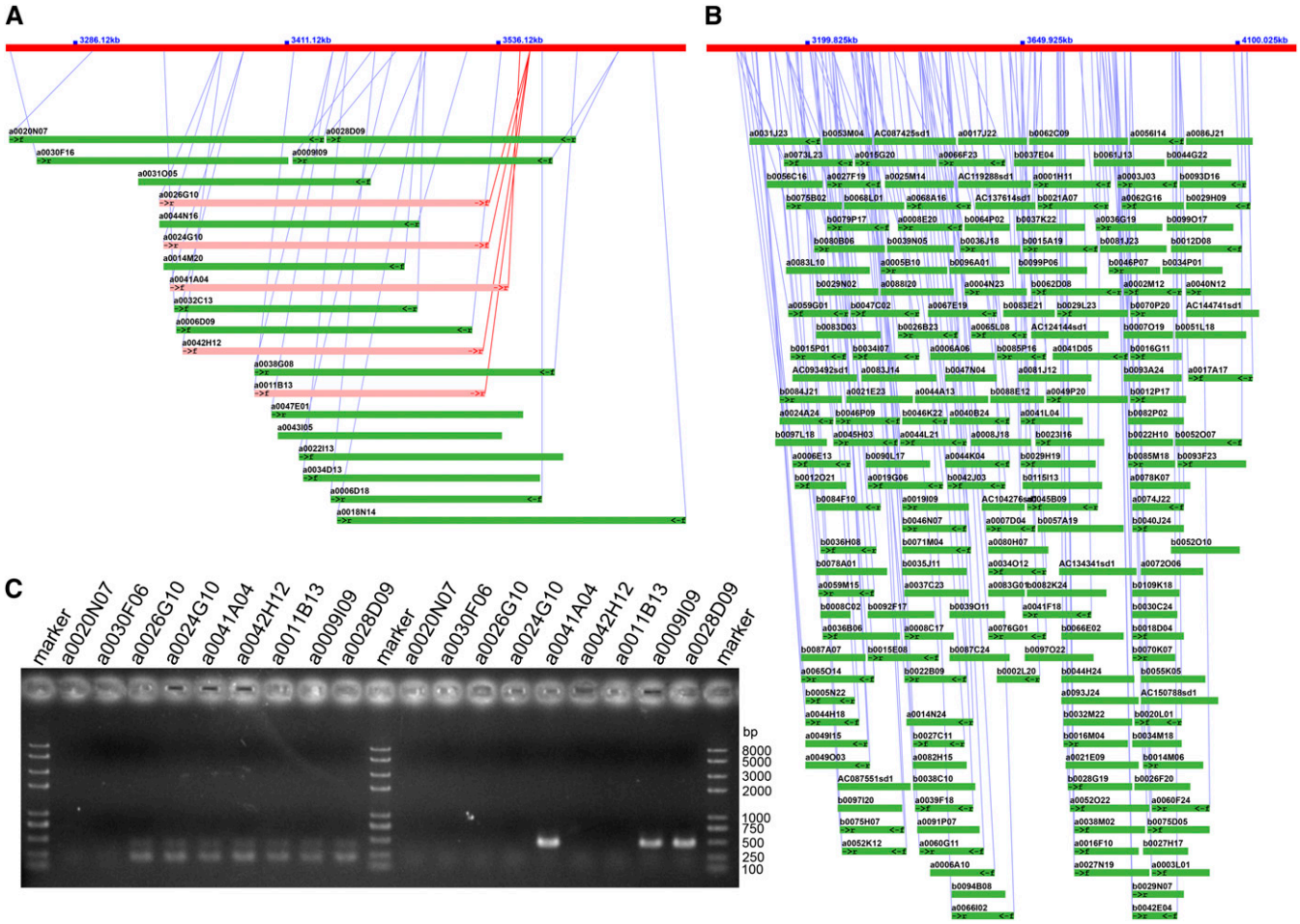

Figure 4 Discovery and confirmation of the inversions by BES alignments. (A) The alignments between the BESs of partial MH63 contig37 and the Nipponbare reference sequence. (B) The alignments between the BESs of the Nipponbare contig41 and the same region of the Nipponbare reference sequence. The top red bars represent the Nipponbare reference sequence, and the blue lines represent the alignments of BESs to the reference sequence. The green bars represent the normally matched clones and the pink bars represent the inversely matched clones. (C) Experimental validation of the genetic inversion. Left, the pair of primers designed from the BES OSIABa0011B13.r. of the clone a0011B13 in the MH63 contig37. Right, the pair of primers designed from the BES OSIABa0041A04.r. of the clone a0041A04 in the MH63 contig37.

especially the Gypsy/DIRS1 elements. However, more contractions of BAC size were detected on the Nipponbare reference sequence relative to the ZS97 physical map. This could be caused by the smaller average contig size of ZS97 than MH63 and 93-11 physical maps because the structural variations were restricted to contig hit regions.

Small nucleotide variations played an important role in the rice domestication process (Fan *et al.* 2006; Konishi *et al.* 2006; Li *et al.* 2006; Sweeney *et al.* 2006; Jin *et al.* 2008; Shomura *et al.* 2008). Feltus *et al.* (2004) identified many SNPs and indels between 93-11 and Nipponbare. Xie *et al.* (2010) identified many SNPs by resequencing 238 RILs derived from a cross between MH63 and ZS97, and Huang *et al.* (2010) identified many SNPs by sequencing 517 rice landraces at low coverage. Our SNPs, MNPs, and indels increased the nucleotide variation densities and provided the materials for identifying and analyzing important agronomic trait-related genes in MH63 and ZS97. Similar to the distribution of SNPs between Nipponbare and 93-11 (Feltus *et al.* 2004), the SNPs between Nipponbare and MH63 or ZS97 were also not evenly distributed. According to the domestication model of Huang *et al.* (2012) as mentioned above, the low-frequency regions could be from the exchanged regions between *japonica* and the wild rice. The shared variations in *indica* varieties relative to Nipponbare could be *indica*-specific gene candidates, possibly for domestication and adaptation. The repeat-masked BESs of *O. nivara*, *OR106424*, and *OR105491* were also used to identify variations relative to Nipponbare and 93-11, and the shared variations relative to *O. sativa* could be putative domestication loci.

In conclusion, we generated the BAC libraries, BESs, and physical maps for MH63 and ZS97. Compared with the reference sequences of Nipponbare and 93-11, we detected a number of polymorphic SSRs, SNPs, indels, and structural variations. The physical maps, comparative analysis results, molecular markers, and QTL from

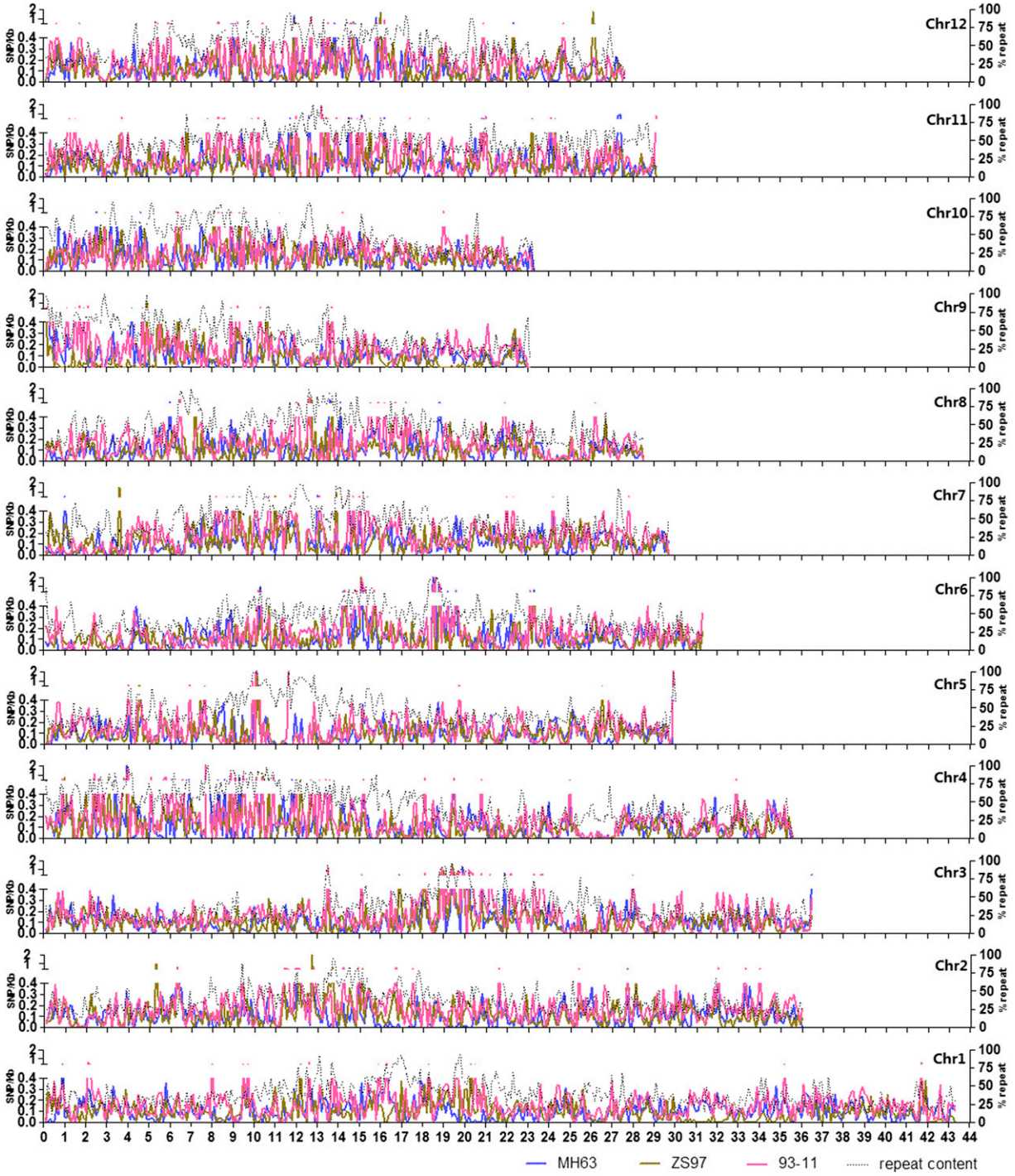

Figure 5 SNP distributions across the Nipponbare chromosomes. The SNPs were identified from the comparisons of repeat-masked BESs of three *indica* varieties: MH63, ZS97, and 93-11 with the Nipponbare reference sequence. The frequency was defined as the number of SNPs divided by the length of the repeat-masked reference sequence within a 100-kb window. The *x*-axis shows the chromosome position in which each unit is a megabase. The left *y*-axis shows SNP frequency. The right *y*-axis shows the percentage of repeat sequences in a 100-kb window (dotted black lines).

GRAMENE and gene models were integrated and aligned to the Nipponbare genome in Gbrowse. These invaluable resources provide a platform for comparative genomics, positional cloning, whole-genome sequencing, functional genomics research, and crop improvement in *O. sativa*. Both the comparative physical maps and the GBrowse are open to the public through our Web site. The BAC-based whole-genome sequencing of MH63 and ZS97 is now underway. Once whole-genome reference sequences become available, more genetic diversity and important agronomic traits-related genes will be revealed and understood clearly.

Acknowledgments

We thank So-Jeong Lee, Nicholas B. Sisneros, Fusheng Wei, HyeRan Kim, Yeisoo Yu, Jetty S. S. Ammiraju, and other members of the Arizona Genomics Institute for production of the BAC library, fingerprints, and BESs. This work was supported by the National Natural Science Foundation of China (grant 30971748) and the Chinese 111 Project (grant B07041).

Literature Cited

Aggarwal, R., T. R. Benatti, N. Gill, C. Zhao, M. S. Chen *et al.*, 2009 A BAC-based physical map of the Hessian fly genome anchored to polytene chromosomes. BMC Genomics 10: 293.

Ammiraju, J. S., M. Luo, J. L. Goicoechea, W. Wang, D. Kudrna *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. S., F. Lu, A. Sanyal, Y. Yu, X. Song *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.

Ammiraju, J. S., C. Fan, Y. Yu, X. Song, K. A. Cranston *et al.*, 2010a Spatio-temporal patterns of genome evolution in allo-tetraploid species of the genus Oryza. Plant J. 63: 430–442.

Ammiraju, J. S., X. Song, M. Luo, N. Sisneros, A. Angelova *et al.*, 2010b The Oryza BAC resource: a genus-wide and genome scale tool for exploring rice genome evolution and leveraging useful genetic diversity from wild relatives. Breed. Sci. 60: 536–543.

Deng, Y., Y. Pan, and M. Luo, 2013 Detection and correction of assembly errors of rice Nipponbare reference sequence. Plant Biol. DOI: 10.1111/plb.12090

Ding, J., H. Araki, Q. Wang, P. Zhang, S. Yang *et al.*, 2007 Highly asymmetric rice genomes. BMC Genomics 8: 154.

Fan, C., Y. Xing, H. Mao, T. Lu, B. Han *et al.*, 2006 GS3, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. Theor. Appl. Genet. 112: 1164–1171.

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

Feng, Q., T. Huang, Q. Zhao, J. Zhu, Z. Lin *et al.*, 2009 Analysis of collinear regions of Oryza AA and CC genomes. J. Genet. Genomics 36: 667–677.

Gao, Z. Y., S. C. Zhao, W. M. He, L. B. Guo, Y. L. Peng *et al.*, 2013 Dissecting yield-associated loci in super hybrid rice by re-sequencing recombinant inbred lines and improving parental genome sequences. Proc. Natl. Acad. Sci. USA 110: 14492–14497.

Garris, A. J., T. H. Tai, J. Coburn, and S. Kresovich, and S. McCouch, 2005 Genetic structure and diversity in *Oryza sativa* L. Genetics 169: 1631–1638.

Gotz, S., J. M. Garcia-Gomez, J. Terol, T. D. Williams, S. H. Nagaraj *et al.*, 2008 High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36: 3420–3435.

Han, B., and Y. Xue, 2003 Genome-wide intraspecific DNA-sequence variations in rice. Curr. Opin. Plant Biol. 6: 134–138.

Huang, X., and A. Madan, 1999 CAP3: a DNA sequence assembly program. Genome Res. 9: 868–877.

Huang, X., X. Wei, T. Sang, Q. Zhao, Q. Feng *et al.*, 2010 Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. 42: 961–967.

Huang, X. H., G. J. Lu, Q. Zhao, X. H. Liu, and B. Han, 2008 Genome-wide analysis of transposon insertion polymorphisms reveals intraspecific variation in cultivated rice. Plant Physiol. 148: 25–40.

Huang, X. H., N. Kurata, X. H. Wei, Z. X. Wang, A. Wang *et al.*, 2012 A map of rice genome variation reveals the origin of cultivated rice. Nature 490: 497–501.

Hurwitz, B. L., D. Kudrna, Y. Yu, A. Sebastian, A. Zuccolo *et al.*, 2010 Rice structural variation: a comparative analysis of structural variation between rice and three of its closest relatives in the genus Oryza. Plant J. 63: 990–1003.

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

Jacquemin, J., D. Bhatia, K. Singh, and R. A. Wing, 2013 The International Oryza Map Alignment Project: development of a genus-wide comparative genomics platform to help solve the 9 billion-people question. Curr. Opin. Plant Biol. 16: 147–156.

Jin, J., W. Huang, J.-P. Gao, J. Yang, M. Shi *et al.*, 2008 Genetic control of rice plant architecture under domestication. Nat. Genet. 40: 1365–1369.

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany *et al.*, 2005 Repbase update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110: 462–467.

Kawahara, Y., M. de la Bastide, J. P. Hamilton, H. Kanamori, W. R. McCombie *et al.*, 2013 Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice 6: 1–10.

Kidd, J. M., G. M. Cooper, W. F. Donahue, H. S. Hayden, N. Sampas *et al.*, 2008 Mapping and sequencing of structural variation from eight human genomes. Nature 453: 56–64.

Kim, H., P. San Miguel, W. Nelson, K. Collura, M. Wissotski *et al.*, 2007 Comparative physical mapping between Oryza sativa (AA genome type) and O. punctata (BB genome type). Genetics 176: 379–390.

Kim, H., B. Hurwitz, Y. Yu, K. Collura, N. Gill *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.

Kofler, R., C. Schlotterer, and T. Lelley, 2007 SciRoKo: a new tool for whole genome microsatellite search and investigation. Bioinformatics 23: 1683–1685.

Konishi, S., T. Izawa, S. Y. Lin, K. Ebana, Y. Fukuta *et al.*, 2006 An SNP caused loss of seed shattering during rice domestication. Science 312: 1392–1396.

Kovach, M. J., M. T. Sweeney, and S. R. McCouch, 2007 New insights into the history of rice domestication. Trends Genet. 23: 578–587.

Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357–359.

Li, C., A. Zhou, and T. Sang, 2006 Rice domestication by reducing shattering. Science 311: 1936–1939.

Li, J. X., S. B. Yu, C. G. Xu, Y. F. Tan, Y. J. Gao *et al.*, 2000 Analyzing quantitative trait loci for yield using a vegetatively replicated F2 population from a cross between the parents of an elite rice hybrid. Theor. Appl. Genet. 101: 248–254.

Li, Y., T. Uhm, C. Ren, C. Wu, T. S. Santos *et al.*, 2007 A plant-transformation-competent BIBAC/BAC-based map of rice for functional analysis and genetic engineering of its genomic sequence. Genome 50: 278–288.

Lin, H., P. Xia, R. A. Wing, Q. Zhang, and M. Luo, 2012 Dynamic intra-japonica subspecies variation and resource application. Mol. Plant 5: 218–230.

Lu, F., J. S. Ammiraju, A. Sanyal, S. Zhang, R. Song *et al.*, 2009 Comparative sequence analysis of MONOCULM1-orthologous regions in 14 Oryza genomes. Proc. Natl. Acad. Sci. USA 106: 2071–2076.

Luo, M., and R. A. Wing, 2003 An improved method for plant BAC library construction. Methods Mol. Biol. 236: 3–20.

Luo, M., Y. H. Wang, D. Frisch, T. Joobeur, R. A. Wing *et al.*, 2001 Melon bacterial artificial chromosome (BAC) library construction using improved methods and identification of clones linked to the locus conferring resistance to melon Fusarium wilt (Fom-2). Genome 44: 154–162.

Luo, M., H. Kim, D. Kudrna, N. B. Sisneros, S. J. Lee *et al.*, 2006 Construction of a nurse shark (Ginglymostoma cirratum) bacterial artificial chromosome (BAC) library and a preliminary genome survey. BMC Genomics 7: 106.

Luo, M. C., C. Thomas, F. M. You, J. Hsiao, O. Y. Shu *et al.*, 2003 High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics 82: 378–389.

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

Mao, H., S. Sun, J. Yao, C. Wang, S. Yu *et al.*, 2010 Linking differential domain functions of the GS3 protein to natural variation of grain size in rice. Proc. Natl. Acad. Sci. USA 107: 19579–19584.

Nasu, S., J. Suzuki, R. Ohta, K. Hasegawa, R. Yui *et al.*, 2002 Search for and analysis of single nucleotide polymorphisms (SNPs) in rice (Oryza sativa, Oryza rufipogon) and establishment of SNP markers. DNA Res. 9: 163–171.

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

Pan, Y., Y. Deng, H. Lin, D. A. Kudrna, R. A. Wing *et al.*, 2013 Comparative BAC-based physical mapping of Oryza sativa ssp. indica var. 93–11 and evaluation of the two rice reference sequence assemblies. Plant J. DOI: 10.1111/tpj.12412.

Rozen, S., and H. Skaletsky, 2000 Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. 132: 365–386.

Shomura, A., T. Izawa, K. Ebana, T. Ebitani, H. Kanegae *et al.*, 2008 Deletion in a gene associated with grain size increased yields during rice domestication. Nat. Genet. 40: 1023–1028.

Soderlund, C., S. Humphray, A. Dunham, and L. French, 2000 Contigs built with fingerprints, markers, and FPC V4.7. Genome Res. 10: 1772–1787.

Soderlund, C., W. Nelson, A. Shoemaker, and A. Paterson, 2006 SyMAP: a system for discovering and viewing syntenic regions of FPC maps. Genome Res. 16: 1159–1168.

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

Stein, L. D., C. Mungall, S. Shu, M. Caudy, M. Mangone *et al.*, 2002 The generic genome browser: a building block for a model organism system database. Genome Res. 12: 1599–1610.

Sun, X., Y. Cao, Z. Yang, C. Xu, X. Li *et al.*, 2004 Xa26, a gene conferring resistance to Xanthomonas oryzae pv. oryzae in rice, encodes an LRR receptor kinase-like protein. Plant J. 37: 517–527.

Sweeney, M. T., M. J. Thomson, B. E. Pfeil, and S. McCouch, 2006 Caught red-handed: Rc encodes a basic helix-loop-helix protein conditioning red pericarp in rice. Plant Cell 18: 283–294.

Tuzun, E., A. J. Sharp, J. A. Bailey, R. Kaul, V. A. Morrison *et al.*, 2005 Fine-scale structural variation of the human genome. Nat. Genet. 37: 727–732.

Wang, C., X. Shi, L. Liu, H. Li, J. S. Ammiraju *et al.*, 2013 Genomic resources for gene discovery, functional genome annotation, and evolutionary studies of maize and its close relatives. Genetics 195: 723–737.

Wang, X., Q. Liu, H. Wang, C. X. Luo, G. Wang *et al.*, 2013 A BAC based physical map and genome survey of the rice false smut fungus Villosiclava virens. BMC Genomics 14: 883.

Xie, W., Q. Feng, H. Yu, X. Huang, Q. Zhao *et al.*, 2010 Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. Proc. Natl. Acad. Sci. USA 107: 10578–10583.

Xing, Y. Z., Y. F. Tan, J. P. Hua, X. L. Sun, C. G. Xu *et al.*, 2002 Characterization of the main effects, epistatic effects and their environmental interactions of QTLs on the genetic basis of yield traits in rice. Theor. Appl. Genet. 105: 248–257.

Xue, W. Y., Y. Z. Xing, X. Y. Weng, Y. Zhao, W. J. Tang *et al.*, 2008 Natural variation in Ghd7 is an important regulator of heading date and yield potential in rice. Nat. Genet. 40: 761–767.

Yu, J., S. Hu, J. Wang, G. K. Wong, S. Li *et al.*, 2002 A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science 296: 79–92.

Yu, J., J. Wang, W. Lin, S. Li, H. Li *et al.*, 2005 The genomes of Oryza sativa: a history of duplications. PLoS Biol. 3: e38.

Zhang, S., Y. Q. Gu, J. Singh, D. Coleman-Derr, D. S. Brar *et al.*, 2007 New insights into Oryza genome evolution: high gene colinearity and differential retrotransposon amplification. Plant Mol. Biol. 64: 589–600.

Zhang, Y., X. Zhang, T. H. O'Hare, W. S. Payne, J. J. Dong *et al.*, 2011 A comparative physical map reveals the pattern of chromosomal evolution between the turkey (Meleagris gallopavo) and chicken (Gallus gallus) genomes. BMC Genomics 12: 447.

Zhao, S., J. Shetty, L. Hou, A. Delcher, B. Zhu *et al.*, 2004 Human, mouse, and rat genome large-scale rearrangements: stability *vs.* speciation. Genome Res. 14: 1851–1860.

Communicating editor: J. A. Birchler

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.159970/-/DC1

Global Genomic Diversity of *Oryza sativa* Varieties Revealed by Comparative Physical Mapping

Xiaoming Wang, David A. Kudrna, Yonglong Pan, Hao Wang, Lin Liu, Haiyan Lin, Jianwei Zhang, Xiang Song, Jose Luis Goicoechea, Rod A. Wing, Qifa Zhang, and Meizhong Luo

DOI: 10.1534/genetics.113.159970

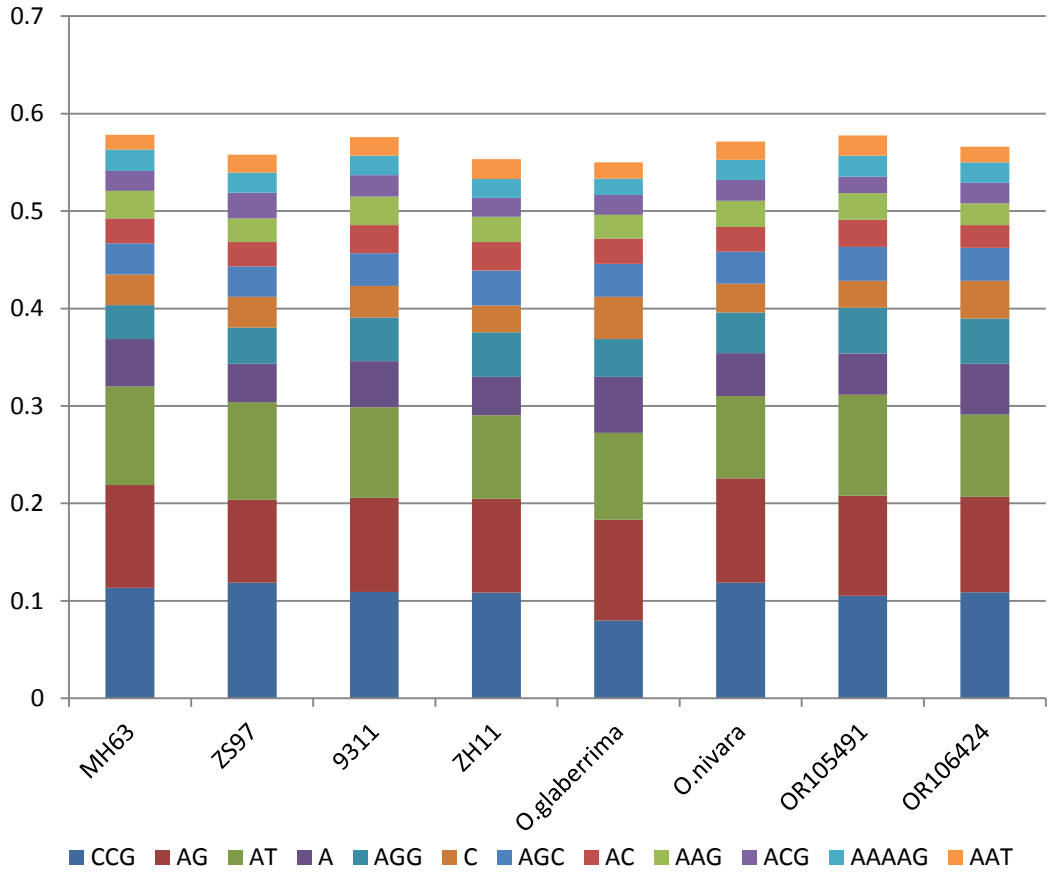

Figure S1 The compositions of the ten most frequent SSR motifs in eight *Oryza* species. The x-axis present the varieties, and the y-axis present the ratio of relative SSRs to total SSRs.

Files S1-S10

Available for download as Excel files at

http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.159970/-/DC1

File S1 Detailed information of all expansion and contraction regions in the Nipponbare genome.

File S2 Validation of expansions and contractions.

File S3 Over- and under-representation of gene ontology (GO) terms in the expanded regions of the Nipponbare.

File S4 The shared expansion and contraction regions in the Nipponbare.

File S5 The clusters of inversely matched BESs between the Nipponbare reference sequence and six physical maps (MH63, ZS97, *O.nivara, OR105491, OR106424 and O.glaberrima*).

File S6 The spanned gaps on the Nipponbare reference sequence.

File S7 The detailed information of the verified SSRs and variations.

File S8 Statistics of SSR information in BESs of eight *Oryza* species.

File S9 The frequencies of specific repeat sequences in BESs of *Oryza* species.

File S10 Summary of the small variations identified by the comparisons between repeat-masked BESs and reference sequences.

Table S1 Statistics of the BESs and fingerprinting information.

Category	MH63	ZS97
Clone number	36864	36864
Average insert size (Kb)	125	117
Genome coverage	10.7X	10X
End sequenced and fingerprinted clones	18432	18432
Clones with BESs	18214	18017
With paired-end BESs	17335	15952
With single-end BESs	879	2065
BESs number	35549	33969
Matched to organelle genomes	1636	1042
Paired-end BESs	1482	872
Single-end BESs	154	170
Only Match to chloroplast	974	708
Only Match to mitochondrion	76	82
Matched to both	586	252
Average BES length (bp)	659	666
GC content	42.08%	42.34%
Clones with Fingerprints	17310	16580
Average bands per clone	116	105
In phase I physical map		
Assembled into contigs	16580	12965
Clones with paired-end BESs	15695	11256
Clones with single-end BESs	764	1446
Left as singletons	730	3615
Clones with paired-end BESs	642	3118
Clones with single-end BESs	39	395
In phase II physical map		
Assembled into contigs	16735	13972
Clones with paired-end BESs	15823	12118
Clones with single-end BESs	781	1569
Left as singletons	575	2608
Clones with paired-end BESs	514	2256
Clones with single-end BESs	22	272

Table S2 Repeat contents in the BESs of *O. sativa* and related *Oryza* species.

Species	Repeat content by RepeatMasker				Total
	RepeatMasker rice repeat database			MSU repeat database[a]	
	Retro TEs	DNA TEs	Total		
MH63	24.93%	8.72%	34.91%	2.58%	37.49%
ZS97	26.49%	8.74%	36.66%	2.84%	39.50%
93-11	23.01%	8.05%	32.39%	2.85%	35.24%
ZH11	22.79%	8.57%	32.80%	3.26%	36.06%
O.glaberrima	18.69%	8.38%	28.33%	2.64%	30.97%
OR105491	25.99%	8.68%	36.03%	2.80%	38.83%
OR106424	27.83%	8.83%	37.95%	2.75%	40.70%
O.nivara	26.88%	8.26%	36.38%	2.86%	39.24%

[a]The repeat sequences which were homologous with the rice repeat database of RepeatMasker were masked, and then the remaining sequences were analyzed by RepeatMasker with the MSU rice repeat data as database.

Table S3 The SNP frequencies on each Nipponbare chromosome.

Chr	MH63	ZS97	93-11	ZH11	*O.nivara*	*O.glaberrima*	*OR106424*
Chr1	0.1094	0.0997	0.1532	0.0441	0.3143	0.2595	0.2125
Chr2	0.1090	0.1301	0.1719	0.0300	0.3394	0.2808	0.2166
Chr3	0.1235	0.1125	0.1639	0.0370	0.3342	0.2714	0.2259
Chr4	0.2780	0.2750	0.3298	0.1570	0.4616	0.2416	0.7928
Chr5	0.1219	0.1118	0.1718	0.0630	0.3967	0.2641	0.2646
Chr6	0.1181	0.1165	0.1785	0.1086	0.3562	0.2457	0.2785
Chr7	0.1164	0.1265	0.1657	0.0425	0.3270	0.2535	0.2173
Chr8	0.1075	0.1059	0.1552	0.0401	0.3349	0.2700	0.2258
Chr9	0.1178	0.0869	0.1660	0.0350	0.2931	0.2680	0.2155
Chr10	0.1299	0.1349	0.1667	0.0349	0.3001	0.2649	0.2156
Chr11	0.1236	0.1261	0.1836	0.0461	0.3307	0.2592	0.2286
Chr12	0.1217	0.1232	0.1734	0.0539	0.3509	0.2724	0.2323