PNAS PLUS

Extensive sequence divergence between the reference genomes of two elite *indica* rice varieties Zhenshan 97 and Minghui 63

Jianwei Zhang (张建伟)[a,b,c,1], Ling-Ling Chen (陈玲玲)[a,1], Feng Xing (邢锋)[a,1], David A. Kudrna[b,c], Wen Yao (姚文)[a], Dario Copetti[b,c,d], Ting Mu (穆婷)[a], Weiming Li (李伟明)[a], Jia-Ming Song (宋佳明)[a], Weibo Xie (谢为博)[a], Seunghee Lee[b,c], Jayson Talag[b,c], Lin Shao (邵林)[a], Yue An (安玥)[a], Chun-Liu Zhang (张春柳)[a], Yidan Ouyang (欧阳亦聃)[a], Shuai Sun (孙帅)[a], Wen-Biao Jiao (焦文标)[a], Fang Lv (吕芳)[a], Bogu Du (杜博贾)[a], Meizhong Luo (罗美中)[a], Carlos Ernesto Maldonado[b,c], Jose Luis Goicoechea[b,c], Lizhong Xiong (熊立仲)[a], Changyin Wu (吴昌银)[a], Yongzhong Xing (邢永忠)[a], Dao-Xiu Zhou (周道绣)[a], Sibin Yu (余四斌)[a], Yu Zhao (赵毓)[a], Gongwei Wang (王功伟)[a], Yeisoo Yu[b,c,2], Yijie Luo (罗艺洁)[a], Zhi-Wei Zhou (周智伟)[a], Beatriz Elena Padilla Hurtado[b,c], Ann Danowitz[b], Rod A. Wing[b,c,d,3], and Qifa Zhang (张启发)[a,3]

[a]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China; [b]Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721; [c]BIO5 Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721; and [d]International Rice Research Institute, Genetic Resource Center, Los Baños, Laguna, Philippines

Contributed by Qifa Zhang, July 12, 2016 (sent for review May 12, 2016; reviewed by James J. Giovannoni and Yaoguang Liu)

Asian cultivated rice consists of two subspecies: *Oryza sativa* subsp. *indica* and *O. sativa* subsp. *japonica*. Despite the fact that *indica* rice accounts for over 70% of total rice production worldwide and is genetically much more diverse, a high-quality reference genome for *indica* rice has yet to be published. We conducted map-based sequencing of two *indica* rice lines, Zhenshan 97 (ZS97) and Minghui 63 (MH63), which represent the two major varietal groups of the *indica* subspecies and are the parents of an elite Chinese hybrid. The genome sequences were assembled into 237 (ZS97) and 181 (MH63) contigs, with an accuracy >99.99%, and covered 90.6% and 93.2% of their estimated genome sizes. Comparative analyses of these two *indica* genomes uncovered surprising structural differences, especially with respect to inversions, translocations, presence/absence variations, and segmental duplications. Approximately 42% of non-transposable element related genes were identical between the two genomes. Transcriptome analysis of three tissues showed that 1,059–2,217 more genes were expressed in the hybrid than in the parents and that the expressed genes in the hybrid were much more diverse due to their divergence between the parental genomes. The public availability of two high-quality reference genomes for the *indica* subspecies of rice will have large-ranging implications for plant biology and crop genetic improvement.

Oryza sativa | reference genomes | BAC-by-BAC strategy | transcriptome

Rice is one of the most important food crops in the world and provides more than 20% of the caloric intake for one-half of the world's population. Asian cultivated rice can be divided into two subspecies—that is, *Oryza sativa* subsp. *indica* and *O. sativa* subsp. *japonica*—which are highly distinctive in geographical distribution, reproductively isolated, and have been shown to have extensive differentiation in genome structure and gene content (1). *Indica* rice accounts for more than 70% of world rice production (2) and is genetically much more diverse than *japonica* rice (3). Genomic studies have established that *indica* rice can be further subdivided into two major varietal groups, *indica I* and *indica II*, which have been independently bred and widely cultivated in China and Southeast Asia, respectively (4). Hybrids between these groups usually show strong heterosis, which provided the basis for the great success of hybrid rice in several countries, including China and the United States. For example, Zhenshan 97 (ZS97, *indica I*) and Minghui 63 (MH63, *indica II*) are the parents of the elite hybrid Shanyou 63 (SY63) (*SI Appendix*, Fig. S1 *A* and *B*), which exhibits superiority for a large array of agronomic traits including yield, resistance to multiple diseases, wide adaptability, and good eating quality, and thus has been the most widely cultivated hybrid in China over the past three decades (*SI Appendix*, Fig. S1*C*).

Because of the importance of hybrid rice in helping to ensure a stable and secure food supply for generations, a series of attempts have been made to gain a fundamental understanding of the biological basis of heterosis, a mystery that has puzzled the scientific community for more than a century, using the ZS97, MH63, and SY63 hybrid system as a model (5–10). Although the genetic components governing heterosis have been identified for yield and yield component traits using this model system, based upon which many heterosis hypotheses were proposed (11), a mechanistic understanding of these components is not possible without

Significance

Indica rice accounts for >70% of total rice production worldwide, is genetically highly diverse, and can be divided into two major varietal groups independently bred and widely cultivated in China and Southeast Asia. Here, we generated high-quality genome sequences for two elite rice varieties, Zhenshan 97 and Minghui 63, representing the two groups of *indica* rice and the parents of a leading rice hybrid. Comparative analyses uncovered extensive structural differences between the two genomes and complementarity in their hybrid transcriptome. These findings have general implications for understanding intraspecific variations of organisms with complex genomes. The availability of the two genomes will serve as a foundation for future genome-based explorations in rice toward both basic and applied goals.

Author contributions: J.Z., L.-L.C., R.A.W., and Q.Z. designed research; J.Z., L.-L.C., D.A.K., D.C., S.L., J.T., L.S., M.L., C.E.M., J.L.G., Y.Y., B.E.P.H., and A.D. performed research; J.Z., L.-L.C., W.X., L.X., C.W., Y.X., D.-X.Z., S.Y., Y.Z., G.W., R.A.W., and Q.Z. contributed new reagents/analytic tools; J.Z., L.-L.C., F.X., W.Y., D.C., T.M., W.L., J.-M.S., W.X., Y.A., C.-L.Z., Y.O., S.S., W.-B.J., F.L., B.D., Y.L., and Z.-W.Z. analyzed data; and J.Z., L.-L.C., D.A.K., R.A.W., and Q.Z. wrote the paper.

Reviewers: J.J.G., USDA-ARS Robert W. Holley Center and Boyce Thompson Institute for Plant Research; and Y.L., South China Agricultural University.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The genome assemblies have been deposited in the GenBank database [accession nos. LNNJ00000000 (ZS97RS1) and LNNK00000000 (MH63RS1)].

[1]J.Z., L.-L.C., and F.X. contributed equally to this work.

[2]Present address: Phyzen Genomics Institute, Phyzen Inc., Seoul 151-836, Korea.

[3]To whom correspondence may be addressed. Email: qifazh@mail.hzau.edu.cn or rwing@mail.arizona.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1611012113/-/DCSupplemental.

the availability of high-quality reference genome sequences of the parents—ZS97 and MH63.

In this paper, we presented the generation of a pair of high-quality reference genome sequences for ZS97 and MH63 using a map-based sequencing approach and detailed comparative annotation and analysis of the two genomes and transcriptomes. Of note, these are the highest quality genome sequences for *indica* rice published to date and are expected to have a lasting impact on cereal genomics research and rice crop improvement.

Results

Generation of Map-Based Reference Genomes for ZS97 and MH63. The genomes of ZS97 and MH63 were sequenced using a bacterial artificial chromosome (BAC)-by-BAC approach, supplemented with Illumina whole genome shotgun (WGS) data. Previously, two individual BAC libraries (~10× coverage) (12) as well as improved physical maps (PMs, covering ~90% of each genome) (13) with whole genome profiling (14), were constructed for each variety. Minimum tiling paths (MTPs) of BAC clones were selected for each genome (i.e., 3,862 for ZS97; 3,254 for MH63) and sequenced in pools with PacBio single molecule, real-time (SMRT) sequencing technology. The average amount of raw sequence per BAC was over 110× coverage in depth (13). The sequences of each BAC were assembled from the pooled sequence data and then assigned to groups to generate PM-guided BAC sequence contigs using our Genome Puzzle Master pipeline (15). The final reference genomes, named ZS97RS1 and MH63RS1, were completed by gap-filling between BAC sequence contigs with a few contigs derived from assembled WGS Illumina data (13) (over 200× base coverage). The hybrid assemblies resulted in a total of 237 contigs (largest, 10,264,344 bp; smallest, 75,758 bp; N50, 2,339,070 bp) for ZS97 and 181 contigs (largest, 9,849,077 bp; smallest, 62,739 bp; N50, 3,097,358 bp) for MH63 (Table 1). The estimated amount of missing sequence from each genome assembly was ~37 Mb for ZS97 and ~26 Mb for MH63 (*SI Appendix* and Dataset S1, sections 1 and 2). Hence, the total contig lengths of ZS97 (346.86 Mb) and MH63 (359.92 Mb) covered ~90.6% and 93.2% of the estimated sizes of the two genomes and included 6 out of 24 complete centromere sequences (i.e., centromeres 8 and 10 from ZS97RS1 and centromeres 6, 8, 9, and 12 from MH63RS1), identified by sequence homology to the highly repetitive 155–165 bp CentO satellite DNA and centromere-specific retrotransposons in rice (16) (*SI Appendix*, Table S1).

To assess the sequence accuracy and completeness of each genome, we measured the nucleotide identity of all overlapping BAC sequences (~133 Mb for ZS97 and ~141 Mb for MH63) in each assembly (*SI Appendix* and Dataset S1, sections 3 and 4) and determined the presence/absence of a highly conserved set of 248 genes that are expected to reside in most, if not all, eukaryotic genomes using the core eukaryotic genes mapping approach (CEGMA) pipeline (17). Both measures showed that the sequence accuracy of the two genomes was high (i.e., 99.99% accurate) and that the genomes contained ~92.7% (ZS97) and ~94.8% (MH63) of the CEGMA repertoire of genes (*SI Appendix*, Table S2).

Characteristics of the genome assemblies and final sequences are summarized in Table 1.

Structural Variations Among the ZS97RS1, MH63RS1, and Nipponbare Genomes. To understand the genomic differences and commonalities between representatives of the two *indica* varietal groups and *japonica* rice, we compared the ZS97RS1 and MH63RS1 genomes between one another and to the *O. sativa* subsp. *japonica* cv. Nipponbare reference genome (henceforth termed Nipponbare RefSeq) for single nucleotide polymorphisms (SNPs), small (length, <100 bp) insertion/deletions (InDels), inversions, translocations, and presence/absence variations (PAVs).

The densities of SNPs and InDels between ZS97RS1 and MH63RS1 were 3.65 SNPs per kilobase and 0.70 InDels per kilobase, compared with 7.14 SNPs per kilobase and 1.31 InDels per kilobase between ZS97RS1 and the Nipponbare RefSeq and 7.36 SNPs per kilobase and 1.34 InDels per kilobase between MH63RS1 and the Nipponbare RefSeq (*SI Appendix*, Table S3), confirming that intersubspecies (*indica* vs. *japonica*) variation was much larger than intrasubspecies (*indica* vs. *indica*) variation. SNPs in both inter- and intrasubspecies comparisons showed that G→A and C→T transitions (Tss) were the most abundant, whereas G→ C and C→G transversions (Tvs) were the least abundant, with a Ts/Tv ratio of about 2.4 (*SI Appendix*, Table S4).

The distributions of SNPs and InDels between ZS97RS1 and MH63RS1 varied along the chromosomes (Fig. 1 and *SI Appendix*, Fig. S2). Some regions (e.g., 10.0–15.6 Mb and 20.5–28.0 Mb on chromosome 2, 10.5–15.0 Mb on chromosome 5, and 8.5–15.2 Mb on chromosome 7) had very low densities of SNPs and InDels, whereas other regions (e.g., 15.8–17.5 Mb on chromosome 4, 13.5–18.0 Mb

Table 1. Characteristics of the ZS97RS1 and MH63RS1 genomes

Genomic feature	ZS97RS1	MH63RS1	Nipponbare RefSeq*
Estimated genome size, Mbp	~384	~386	~389
Total size of assembled contigs, bp	346,854,256	359,918,891	373,245,519
Estimated gap size, Mbp	36.66	26.24	13.75
Completeness†	90.6%	93.2%	95.3%
Number of contigs	237	181	85 (251)‡
Largest contig, bp	10,264,344	9,849,077	17,269,798
Smallest contig, bp	75,758	62,739	503
Contig N50, bp	2,339,070	3,097,358	7,711,345
GC content	43.59%	43.63%	43.57%
Numbers of gene models/transcripts	54,831/78,033	57,174/80,581	55,986/66,338
Number of non-TE gene loci	34,610	37,324	39,045
Estimated gene models in gap regions /non-TE gene models	5,136/2,957	3,481/1,891	—
Mean transcript length, bp	1,955	1,962	1,708
Mean CDS length, bp	1,348	1,368	1,342
Total size of transposable elements, bp	143,193,235	149,662,360	148,139,763

*Obtained from International Rice Genome Sequencing Project (IRGSP) release 7.
†Based on estimated genome size.
‡Seventy-three gaps (85 contigs) are listed in IRGSP release 7, but there are 239 gaps (251 contigs) with more than 100 continuous unsequenced N bases in this release.

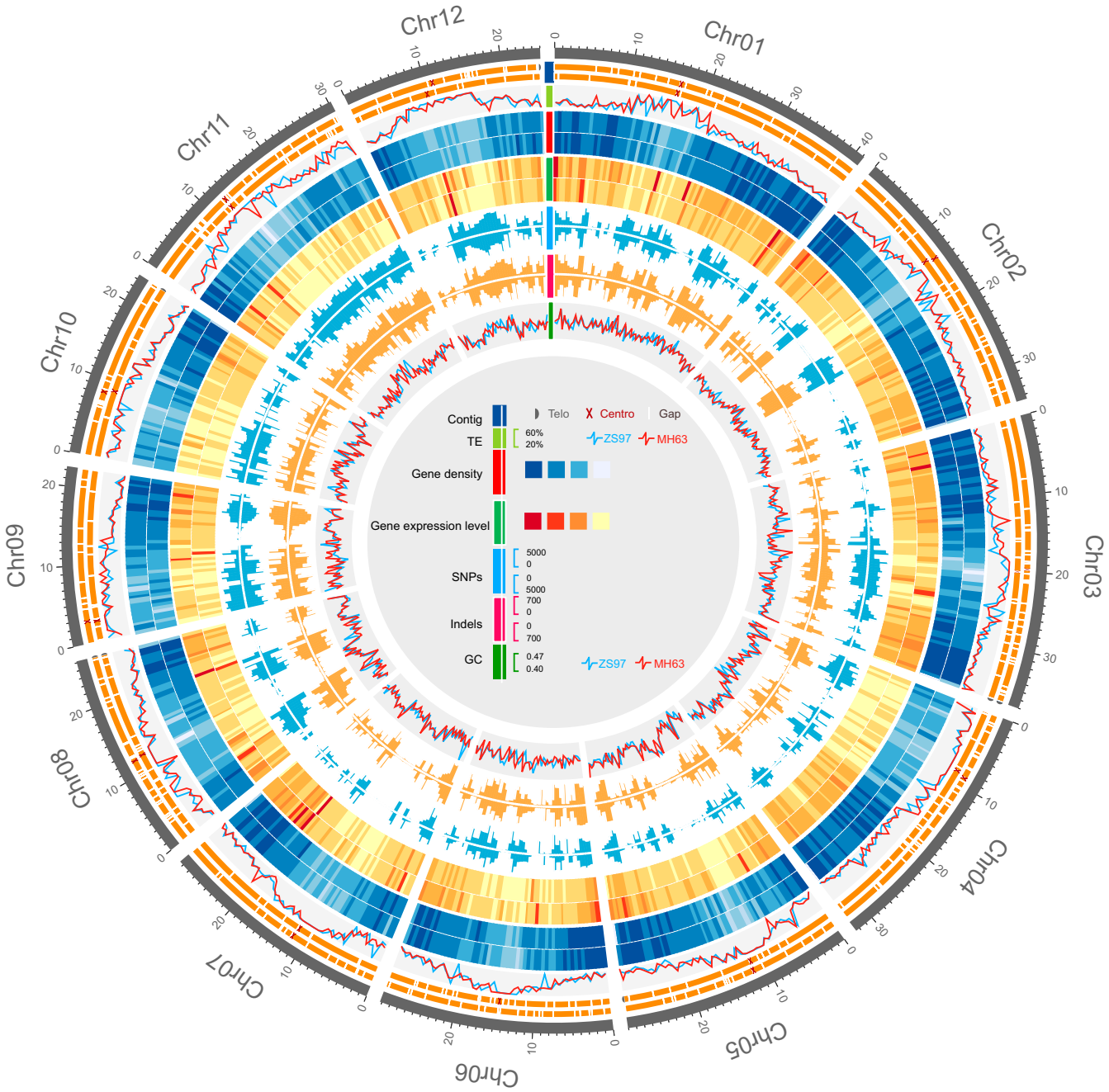

Fig. 1. Overview of the ZS97RS1 and MH63RS1 reference genomes. Tracks from outer to inner circles indicate: contigs and gaps, TE content (window size of 500 kb), gene density (darker color indicates more genes), gene expression level (yellow being the lowest and dark red the highest), SNPs, Indels (mapping to ZS97RS1 and MH63RS1, respectively), and GC content (window size of 100 kb). For each track, the outer and inner layers indicate ZS97RS1 and MH63RS1 data, respectively.

on chromosome 8, and 0.5–12.5 Mb on chromosome 9) had high densities.

We detected 131 large inversions, ranging in size from 512 bp to 362,444 bp, between the ZS97RS1 and MH63RS1 genomes totaling ~1.96 Mb in ZS97RS1. These inversions were unevenly distributed and occurred more frequently on chromosome 11 and less frequently on chromosome 5. The largest inversion (i.e., IV12010) was 362,444 bp in ZS97RS1 on chromosome 12. The orthologous region in MH63RS1 had a ~81-kb gap, which was confirmed by BAC-clone SMRT sequencing, PCR assays, and Illumina reads from 5-kb and 10-kb mate-pair read data (*SI Appendix*, Fig. S3 *A–C*). All other inversions occurred in gap-free regions (*SI Appendix* and Dataset S1, section 5). We validated the 10 largest inversions using Illumina mate-pair read data and PCR, all of which were confirmed. We also detected 357 large inversions (ranging from 128 bp to 100,422 bp) totaling ~4.66 Mb between the ZS97RS1 and Nipponbare RefSeq genomes (*SI Appendix* and Dataset S1, section 6) and 402 large inversions (ranging from 119 bp to 457,227 bp) totaling ~5.64 Mb between the MH63RS1 and Nipponbare RefSeq genomes (*SI Appendix* and Dataset S1, section 7).

Over 5,000 regions (ranging from 100 bp to 54.6 kb, totaling ~8.9 Mb) were detected as translocations (*SI Appendix* and

Dataset S1, section 8) including translocations within the same chromosomes and ones between different chromosomes (*SI Appendix*, Fig. S4 *A* and *B*). These translocations were unevenly distributed across the 12 chromosomes of each rice genome analyzed.

Genome comparisons for PAVs showed that 4,509 genomic segments (length, >100 bp, excluding gaps) totaling ∼21.5 Mb were present in ZS97RS1 but absent in MH63RS1, whereas 4,566 segments totaling ∼23.3 Mb were present in MH63RS1 but not in ZS97RS1 (*SI Appendix* and Dataset S1, sections 9 and 10). The largest unique presence in ZS97RS1 was a segment of 106,206 bp on chromosome 4, and in MH63RS1 a unique segment of 281,512 bp was found on chromosome 5. An example of a unique presence region in MH63 is shown in *SI Appendix*, Fig. S5. PAVs were abundant on chromosomes 1, 4, and 11 and rare on chromosomes 5 and 7. Compared with the Nipponbare RefSeq, ZS97RS1 contained 6,464 unique presence regions (length, >100 bp) totaling ∼34.2 Mb, and MH63RS1 contained 6,906 unique presence regions totaling ∼40.3 Mb (*SI Appendix* and Dataset S1, sections 11 and 12), further confirming that variations between the *indica* and *japonica* genomes were much more abundant than between the *indica I* and *indica II* varietal group genomes.

All of the comparisons above also indicated that ZS97 is more closely related to Nipponbare than is MH63.

Genome Annotation of ZS97RS1 and MH63RS1. To determine the transposable element (TE) content of both *indica* genomes, we used RepeatMasker (18), loaded with the Repbase (19) and plant miniature inverted-repeat transposable elements (P-MITE) (20) databases, and identified 143.2 Mb of TE sequence in ZS97RS1 (i.e., 41.28% of the estimated genome size) and 149.7 Mb of TE sequence in MH63RS1 (i.e., 41.58% of the genome size) (Fig. 1), both of which were about 2% greater than that found in the Nipponbare RefSeq. (21) (*SI Appendix*, Table S5). The largest class of TEs were the retrotransposons, accounting for >25% of both genomes, and consisted mostly of *Gypsy* and *Copia* retroelement families. *Gypsy* alone represented about half of the total amount of repeats, and its content in both *indica* genomes (i.e., ZS97 and MH63) was higher than that found in the Nipponbare RefSeq. (21). DNA transposons accounted for slightly over 15% of each genome assembly, with miniature inverted-repeat-TEs (MITEs: e.g., hAT, CACTA, Mariner, Mutator, Mim, and Harbinger) accounting for ∼60% of all DNA transposons in both the *indica* and *japonica* rice genomes (*SI Appendix*, Table S6). The majority (>90%) of translocations between ZS97RS1 and MH63RS1 were found to be associated with TEs, suggesting that transposon activities played an important role in structural changes of these genomes.

Homology-based annotation of noncoding RNA sequences predicted 592 and 589 transfer RNA (tRNA) genes, 449 and 457 small nucleolar RNA (snoRNA) genes, and 92 and 97 spliceosomal RNA (snRNA) genes. Most of the tRNA, snoRNA, and snRNA genes were found in syntenic positions between the two genomes. However, only 40 and 60 ribosomal RNA (rRNA) sequences, including 5S, 5.8S, 18S, and 28S rRNA coding units scattered on several chromosomes, were identified in the genome assemblies of ZS97RS1 and MH63RS1, respectively (*SI Appendix*, Table S7). In addition, 341 and 363 of the microRNA-coding sequences (CDSs) in the miRBase database (22) were identified in ZS97RS1 and MH63RS1, 287 of which were syntenic (*SI Appendix*, Table S7), whereas the remaining were in either the unique presence or gap regions of their respective genomes.

A comprehensive strategy combining ab initio gene prediction, protein-based homology searches, expressed sequence tag (EST) alignment, and RNA sequencing of three tissues (seedling shoot, developing panicle, and flag leaf) was used to annotate protein-coding genes (*SI Appendix*, Fig. S6). We identified 54,831 gene models with 78,033 transcripts in ZS97RS1 and 57,174 gene models with 80,581 transcripts in MH63RS1 (Table 1). Of these gene models, 34,610 in ZS97RS1 and 37,324 in MH63RS1 were classified as non-TE gene loci (Table 1 and *SI Appendix*, Table S8). The protein-coding non-TE genes were unevenly distributed across each chromosome with gene density increasing toward the chromosome ends (Fig. 1). We compared the features of TE-related and non-TE–related gene loci and observed that TE-related gene loci had on average fewer alternative transcription isoforms per gene locus, fewer exons per gene, smaller gene size, but longer CDS lengths (Fig. 2*A* and *SI Appendix*, Table S9). Moreover, we estimated that 5,136 (including 2,179 TE and 2,957 non-TE) and 3,481 (including 1,590 TE and 1,891 non-TE) gene loci were located in ZS97RS1 and MH63RS1 gap regions, respectively (Table 1 and Fig. 2*B*).

Differences Between ZS97 and MH63 Caused by Genomic Variations. To investigate genome complementarity on a genome-wide scale, we examined the distributions of SNPs and InDels in different genomic regions including intergenic regions, introns, 5′- and 3′-untranslated regions (UTRs), protein CDSs, and TEs. For intersubspecific comparisons (*indica–japonica*), SNPs per kilobase in these regions ranged from 4.17 to 10.34 (*SI Appendix*, Tables S10 and S11), and InDels per kilobase ranged from 0.32 to 1.71. As expected, the InDels per kilobase values were the lowest in CDS regions. Intrasubspecific comparisons (*indica–indica*) revealed a similar pattern, although with lower values; SNPs per kilobase varied from 1.96 to 5.19 and InDels per kilobase ranged from 0.17 (in CDS regions) to 0.84 (*SI Appendix*, Table S12). In CDS regions, we observed that single-base InDels were the most abundant, and InDels in lengths of multiples of 3 bp (without frame shift) were much more abundant than others, suggesting that these differences may have resulted as a consequence of functional selection (Fig. 2*C*).

Potential gene structure alterations as a consequence of SNPs or InDels between ZS97RS1 and MH63RS1 were predicted. In total, 128,547 SNPs and 5,010 InDels resulted in in-frame InDels of predicted proteins affecting 18,131 gene loci in ZS97RS1 and 18,420 gene loci in MH63RS1. Large alterations included CDS frame shifts (5,536 SNPs and 5,373 InDels), alterations of splice-site acceptors (447 SNPs and 379 InDels) and splice-site donors (450 SNPs and 350 InDels), start codon losses (210 SNPs and 249 InDels), and stop codon gains (2,193 SNPs and 1,918 InDels) and stop codon losses (390 SNPs and 659 InDels) (Fig. 2*D*).

Based on our PAV analysis, we detected 3,984 (ZS97RS1) and 4,308 (MH63RS1) gene loci that were located in unique presence regions, 1,389 (ZS97RS1) and 1,713 (MH63RS1) of them were non-TE genes (Fig. 2*B*). Pfam analysis showed that protein domains, such as NB-ARC domain, protein kinase, leucine-rich repeat (LRR), protein tyrosine kinase, wall-associated receptor kinase galacturonan-binding, calcium-binding EGF domain, ankyrin repeats, salt stress response/antifungal, and many proteins of unknown functions, were differentially enriched in these ZS97-specific and MH63-specific genes (Fig. 2 *E* and *F*), thereby indicating that many disease and stress resistance-like genes are complementary between the two genomes. Interestingly 66.7% (ZS97RS1) and 65.8% (MH63RS1) of the TE content of each genome was localized to PAV regions of each genome (*SI Appendix*, Table S13), indicating that the ZS97RS1 and MH63RS1 unique presence regions were correlated with TE activity.

A comparison of non-TE gene loci between ZS97RS1 (i.e., 34,610) and MH63RS1 (i.e., 37,324) allowed us to determine the level of gene complementarity between the two genomes (*SI Appendix*, Table S14). As expected, a large proportion of non-TE genes (i.e., 15,214: 44% ZS97RS1 and 40.8% MH63RS1) were found to be identical in both sequence and position, with allowance for SNPs that did not cause amino acid substitutions. For the remaining non-TE genes, we divided them into four major categories: (*i*) 4,174 gene pairs had the same lengths and syntenic positions but contained nonsynonymous substitutions;

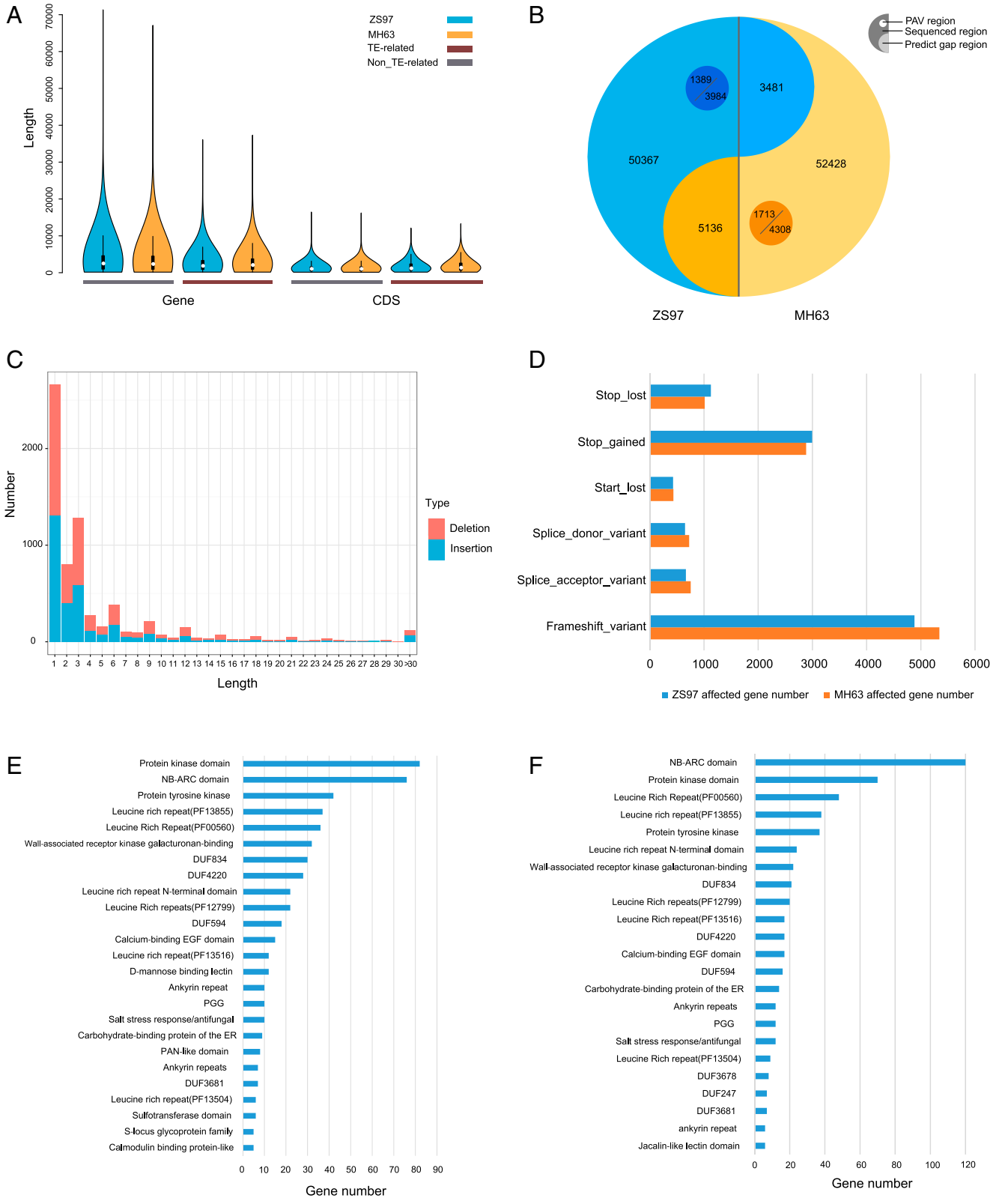

Fig. 2. Gene features of ZS97RS1 and MH63RS1. (*A*) Features of TE-related and non-TE–related genes in the ZS97RS1 and MH63RS1 genomes (see *SI Appendix*, Table S9 for details). (*B*) Gene comparisons between the ZS97RS1 and MH63RS1 genomes. A total of 3,984 (1,389 non-TE) and 4,308 (1,713 non-TE) gene loci are in the unique present region of ZS97RS1 and MH63RS1, and 5,136 and 3,481 gene loci are located in ZS97RS1 and MH63RS1 gap regions. (*C*) Length distribution of InDels in protein-coding regions. The distribution indicates nonsingle base InDels with a length of 3 bp (and/or multiples of 3) were much more abundant than the others. (*D*) SNPs and InDels that caused high-impact gene variations between ZS97RS1 and MH63RS1. (*E*) Gene enrichment in ZS97RS1 unique present regions. (*F*) Gene enrichment in MH63RS1 unique present regions.

(*ii*) 5,932 gene pairs had "good collinearity," by having syntenic chromosomal locations, and protein sequences of >80% sequence identity and >50% coverage; (*iii*) 6,010 (ZS97RS1) and 7,334 (MH63RS1) non-TE genes were classified as "divergent genes," which resulted from large genomic variations between the two genomes; and (*iv*) a total of 1,389 genes were identified as unique presence in ZS97RS1, and 1,713 were only found in MH63RS1. Lastly, there were 1,891 non-TE genes that were present in the ZS97RS1 genome but not sequenced in the MH63RS1 genome, and conversely 2,957 non-TE genes were present in the MH63RS1 genome but not sequenced in ZS97RS1.

To assess the possible impacts of structural differences on gene function, we compared the sequences of 1,931 rice genes whose functions have been characterized previously in rice (www.ricedata.cn/gene/). The comparison revealed that 1,001 protein CDSs were identical and collinear between the two genomes, 267 were of the same length and collinear with nonsynonymous substitutions in the protein sequences, 276 were collinear but with variable lengths in the two genomes, 75 genes were in the divergent category, 11 were TE-related genes, 62 were in PAV regions, 218 were in gap regions, and the remaining 21 genes were not found in the two genomes (*SI Appendix* and Dataset S1, section 13), presumably due to the use of an automated annotation pipeline.

More specifically, three genes for yield traits were cloned based on populations derived from a ZS97/MH63 cross (23–26). For example, *Ghd7*, a gene having large effects simultaneously on the number of grains, plant height, and heading date (23), belongs to the PAV category, located in a presence fragment (MH07p00800) in MH63 but absent in ZS97 (*SI Appendix*, Fig. S5*A*). The MH63 allele in a ZS97 background can delay heading by 21 days and increase plant height by 33 cm, resulting in a 66% increase in the number of spikelets per panicle (23). *GS3*, a major quantitative trait locus (QTL) for grain size, has a SNP in the predicted protein that caused a loss-of-function mutation of the MH63 allele and thus was classified as divergent by our comparative annotation. The mutant MH63 allele produces long grains, whereas the wild-type ZS97 allele results in a short grain phenotype (24). Another QTL for grain size, *GS5*, with a minor effect on grain width, was annotated as collinear between ZS97 and MH63, as the causal mutation was two SNPs in the promoter region with little alteration in their CDSs (25, 26). Genetic analyses showed that all these genes contribute complementary effects to the performance of the F_1 hybrid (23–26). A similar situation can also be found with a number of disease-resistance genes (27, 28).

Segmental Duplications of ZS97RS1 and MH63RS1. Previous studies have shown that the rice genome is featured by extensive segmental duplications (29, 30). Based on the sequences of non-TE genes, we detected 3,504 (241 blocks) and 3,429 (238 blocks) pairs of collinear genes in the ZS97 and MH63 genomes, respectively (Fig. 3 *A* and *B*, *SI Appendix*, and Dataset S1, sections 14 and 15). Most of the previously identified segmentally duplicated regions were detected in both the ZS97RS1 and MH63RS1 assemblies, including the large duplicated segments between chromosomes 1 and 5, 2 and 4, 2 and 6, 3 and 7, 3 and 10, 8 and 9, and 11 and 12. However, we also detected 20 and 19 different collinearity blocks (containing three or more PAV/divergent genes) in the ZS97 and MH63 genomes (Fig. 3 *A* and *B*, *SI Appendix*, and Dataset S1, section 16). In addition, 3,065 and 3,134 tandem duplicated genes were identified in ZS97RS1 and MH63RS1, respectively (*SI Appendix* and Dataset S1, sections 17 and 18). The comparison revealed good collinearity of non-TE–related genes in nonduplicated regions between the ZS97RS1 and MH63RS1 genomes on their corresponding chromosomes (Fig. 3*C*).

Hybrid Transcriptome Complementation Between ZS97 and MH63. To investigate differences in gene expression between the inbred lines ZS97 and MH63 and their hybrid, we sequenced mRNA from seedling shoot, developing panicle, and flag leaf samples from all three lines, with two biological replicates per tissue, under identical growth conditions. The number of transcripts detected at fragments per kilobase of transcript per million mapped reads (FPKM) values ≥0.2 in the three tissues are shown in Table 2 and ranged from 18,696 in flag leaf (ZS97) to 24,818 in the developing panicle (hybrid). The largest number of genes was expressed in panicle tissues followed by seedling shoot and flag leaf. As expected, more genes were found to be expressed in the MH63 tissue than in ZS97, which is consistent with the total number of annotated genes. Of more significance was the finding that a much larger number of transcripts (>1,700 on average) was detected in the hybrid than for either parent (Table 2). Additionally, detailed expression levels of 1,931 reported genes in rice were achieved (*SI Appendix* and Dataset S1, section 13).

The identical genes between MH63RS1 and ZS97RS1 accounted for 44.4%, 43.9%, and 44.0% of the total number of expressed genes (transcripts) in the hybrid genome in the three tissues, which were about 4% lower than in the corresponding tissues of the parents. Approximately 45% of the total transcripts in the hybrid were contributed by genes that differed in one way or another between the two genomes including those classified as nonsynonymous substitution, collinear with slight difference, and divergent (Table 2). In addition, ∼11% of the transcripts were from genes located in the unsequenced regions of either parent (i.e., gaps between contigs). When the latter category was discounted, approximately half of the transcripts in the hybrid resulted from the genes that were different between the two parental genomes. Thus, the genes expressed in the hybrid were not only larger in number but also more diverse due to the complementarity between the parental genomes, which may explain the transcriptomic basis for the superior performance of the hybrid.

Discussion

As the human population increases by more than 3 billion by 2050, the agricultural sector needs to develop the tools necessary to increase crop yields in a more sustainable fashion. The exploitation of heterosis is a key component of the solution for the future of food security worldwide. Our work lays a solid foundation for a deeper understanding of heterosis by producing two of the highest quality *indica* rice genomes to date from the parents of a leading hybrid rice that has been widely grown in China for the past three decades.

We chose to sequence these two genomes using a conservative yet proven map-based strategy combined with the power of long-read PacBio sequencing chemistry. One significant advantage of our approach is that the community has immediate access to virtually any region of the sequenced genome for functional characterization through publicly available repositories of BAC resources in both the United States and China. Further, single seed decent seed of ZS97 and MH63 as well as a set of advanced mapping populations are freely available. The genomes, germplasm, and molecular resources make the ZS97/MH63 system an ideal platform to investigate the molecular basis of heterosis.

Accurate genomic sequence is critical for studies of genetic variation. We uncovered a surprisingly large number of structural variations including inversions, translocations, and PAVs between the ZS97 and MH63 high-quality reference genomes, which are difficult, if not impossible, to identify with short-read genome assemblies. In the future, we intend to add additional long-read WGS data to our assemblies to capture the remaining regions of each genome that are not present in the current assembly and to better understand genome methylation patterns of both the parents and their heterotic hybrid—SY63.

The availability of these two new reference quality genome sequences will greatly facilitate the understanding of the rice genome diversity, especially of *indica* rice, and unlock a rich

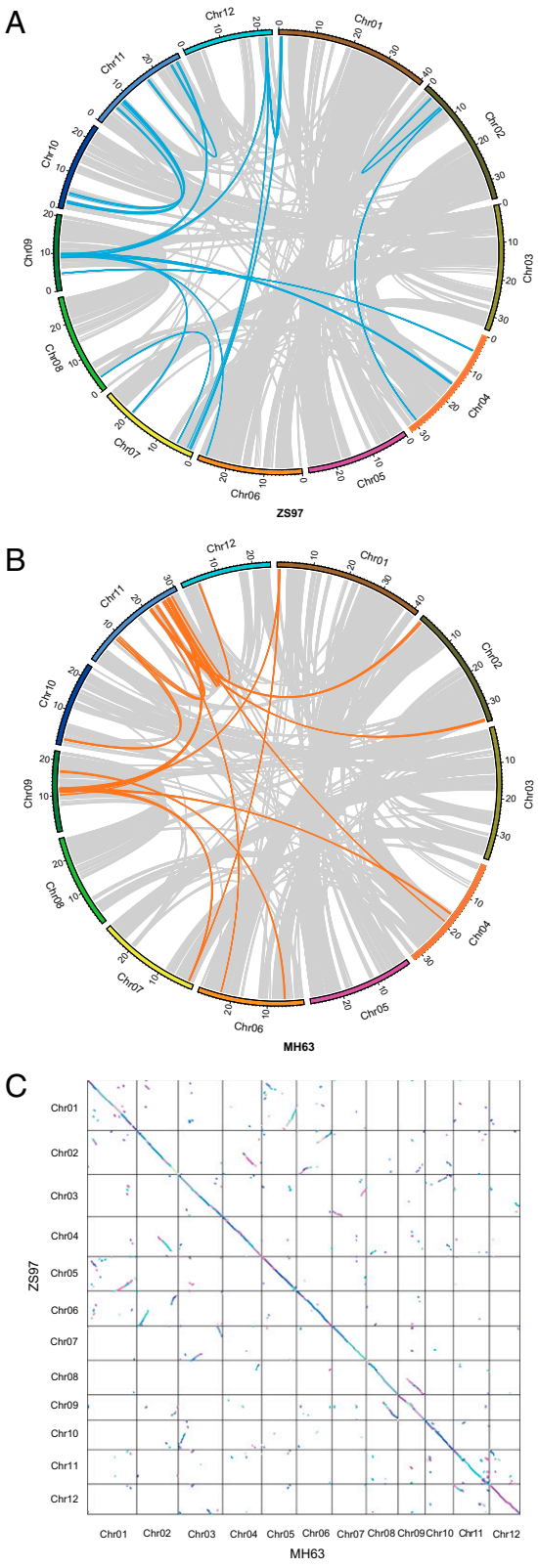

Fig. 3. Collinearity and segmental duplications in the ZS97RS1 and MH63RS1 genomes. Collinearity of the non-TE genes in the two *indica* genomes were identified by using MCScanX (60) with default parameters and a threshold E-value of 1e-20 for each collinearity gene pair. (*A*) Segmental duplicated regions in ZS97RS1. Gray lines indicate common collinearity regions in the two genomes; blue lines indicate different collinearity regions in ZS97RS1. (*B*) Segmental duplicated regions in MH63RS1. Gray lines indicate common collinearity regions in the two genomes; orange lines indicate different collinearity regions in MH63. (*C*) Collinearity of the non-TE–related genes in ZS97RS1 and MH63RS1.

reservoir of the genetic information based on studies of populations from the ZS97/MH63 cross for understanding a range of important questions in plant biology and agronomic performance, such as heterosis, adaptation, yield, quality, flowering, and disease resistances. These sequences will also be very useful for rice genetic improvement to help feed the future of mankind.

Methods

Genome Sequencing and Assembly. To obtain high-quality genome sequences of ZS97 and MH63, we produced and assembled sequences of (*i*) over 4,000 BAC clones selected for each variety from two improved PMs by using PacBio SMRT technology and (*ii*) over 200× WGS data by Illlumina technology. Details on data generation and processing were described in our data descriptor paper (13). The genome assemblies have been deposited in GenBank under accession nos. LNNJ00000000 (ZS97RS1) and LNNK00000000 (MH63RS1).

Assessment of BAC Sequence Quality and Assembly Completeness. We used the average identity of overlapping regions between two neighboring BAC sequences to define the overall accuracy rate of our assembled BAC sequence. We extracted the sequences of all neighboring BACs and ran MUMmer (version 3.23) (31) with the "–mum -p" parameter for aligning each pair of BAC sequences, followed by "delta-filter -1" and "show-coords –clrT" processing to identify overlapping regions between two BAC sequences. Then we detected base substations and InDels in those overlapping regions with "show-snps" using MUMmer (31). *SI Appendix* and Dataset S1, sections 3 and 4 showed the detailed sequence quality of all overlapping regions. Sequence accuracy based on base pair discrepancies was 99.9969% for ZS97 and 99.9931% for MH63, and the accuracy based on both substitutions and InDels was 99.9787% for ZS97 and 99.9749% for MH63.

We used the CEGMA pipeline (17) to assess the completeness of each genome assembly with regard to gene content by searching each assembly for a set of core eukaryotic genes (CEGs) that are highly conserved and present in nearly all eukaryotes. The proportion of complete and partial CEGs (out of 248 possible) is taken as a measure of the completeness of the gene content of an assembly.

The Completeness of Centromeres on ZS97RS1 and MH63RS1 Chromosomes. We built a fasta file of 155–165 bp CentO satellite DNA sequences in rice into a hmm file through HMMER (version 3.1b1) (32). Then we searched the ZS97RS1 and MH63RS1 genomes using hmmsearch to obtain the location of centromeres. We determined the completeness of the ZS97RS1 and MH63RS1 centromeres through comprehensively comparing the positions of centromeres, copy numbers, and their corresponding locations in the Nipponbare RefSeq.

Analysis of SNPs and InDels. We used MUMmer (version 3.23) (31) to align ZS97RS1 and MH63RS1 using the parameters –maxmatch –c 90 –l 40 and then used the delta –filter –1 parameter with the one-to-one alignment block option to filter the alignment results. We used show-snp to identify SNPs and Indels in the one-to-one alignment block (parameter –Clr TH). We used snpEff (33) software to annotate the effects of SNPs and InDels. We used a sliding window method (window size, 100 kb; step, 100 kb) to calculate the distribution of SNPs and InDels along each genome.

Identification of Inversions and Translocations. We first extracted the alignment blocks of MUMmer (version 3.23) (31) with inversions and filtered the blocks with low similarity in the two flanks. We manually checked the remaining inversion blocks and combined the neighboring blocks within 50 bp. Translocation refers to the situation when a DNA segment occurs in different locations in the two genomes and was detected by identifying noncollinear single-copy homologous blocks (length, >100 bp; identity, >90%) between the ZS97RS1 and MH63RS1 genomes.

Identification of PAVs. We used "show-diff" in MUMmer3 (version 3.23) (31) to select for unaligned regions and classified them into "link-inversion," "link-jump," and "gap" regions. We filtered the unaligned sequences in gap regions and retained 34.0 Mb of potential unique presence regions in ZS97RS1 and 35.8 Mb in MH63RS1. We then aligned the potential unique presence regions to the other genome using blastn (e < –5) and filtered

Table 2. Number of expressed genes detected in both replicates with an FPKM expression level ≥0.2

Variety	Category	Seedling shoot	Panicle	Flag leaf
ZS97	Total genes	21,842	23,005	18,696
	Identical genes	10,532 (48.2%)*	10,978 (47.7%)*	9,038 (48.3%)†
	Nonsynonymous genes	2,957	3,131	2,570
	Other collinear genes	3,569	3,781	3, 083
	Divergent genes from ZS97	1,816	1,979	1,517
	Unique presence genes from ZS97	521	537	445
	ZS97 gap region	1,402	1,469	1,165
	MH63 gap region	1,045	1,130	878
MH63	Total genes	21,834	23,240	19,854
	Identical genes	10,501 (48.1%)*	11,022 (47.4%)*	9,452 (47.6%)*
	Nonsynonymous genes	2,954	3,134	2,703
	Other collinear genes	3,569	3,830	3,276
	Divergent genes from MH63	1,722	1,982	1,610
	Unique presence genes from MH63	581	612	577
	ZS97 gap region	1,568	1,668	1,393
	MH63 gap region	939	992	843
Hybrid	Total genes	23,681	24,818	20,913
	Identical genes	10,520 (44.4%)*	10,890 (43.8%)*	9,209 (44.0%)*
	Nonsynonymous genes†	2,981	3,094	2,644
	Other collinear genes†	3,649	3,839	3,245
	Divergent genes from ZS97	1,549	1,670	1,393
	Divergent genes from MH63	1,447	1,596	1,273
	Unique presence genes from ZS97	432	471	406
	Unique presence genes from MH63	497	529	484
	ZS97 gap region	1,567 (6.6%)*	1,635 (6.6%)*	1,358 (6.5%)*
	MH63 gap region	1,039 (4.4%)*	1,094 (4.4%)*	901 (4.3%)*
HY – ZS97		1,839	1,813	2,217
HY– MH63		1,847	1,578	1,059

*Proportion of the genes in this category is the total number of expressed genes in its respective genome.
†Transcripts from the two alleles of each gene in these two categories were counted as from one gene.

regions with coverage >50% and identity >90% to obtain the final unique presence regions in the ZS97RS1 and MH63RS1 genomes.

TE Analysis. First we integrated rice MITEs into both the P-MITE database (20) and Repbase (19) and then executed the RepeatMasker program (18) to identify TEs across the ZS97RS1 and MH63RS1 genomes.

Annotation of Protein-Coding Genes. A comprehensive strategy that combined ab initio gene finding programs, protein-based homology searches, EST alignment, and assembly of RNA-seq reads (flag leaf, panicle, and seeding shoot) was used to annotate protein-coding genes in ZS97RS1 and MH63RS1. We used four ab initio gene-finding programs including Augustus (34), GeneMark (35), semi-hidden Markov model (HMM)-based nucleic acid parser (SNAP) (36), and Fgenesh (37), on the repeat-masked genome sequences. In addition, we collected protein sequences from the UniProt database (38) and several model plant species (i.e., rice, maize, *Arabidopsis*), ESTs from GenBank (39) and other databases [i.e., TriFLDB (40), RICD (41), PlantDB (42)]. Protein sequences were aligned to the ZS97RS1 and MH63RS1 genomes using exonerate (43) and genBlastG (44) to generate spliced alignments, and EST sequences were aligned to genomes using exonerate (43) and BLAT (45). In addition, we generated RNA-seq reads from three tissues, which were assembled with Trinity (46) and aligned to the genomes using exonerate (43) and BLAT (45). All gene structures predicted by the above methods were combined into consensus gene models using EVidenceModeler (EVM) (47). Gene models produced by EVM (47) were then updated by the Program to Assemble Spliced Alignments (PASA) (48).

Gene functions were assigned according to the best hit alignment, using BLASTP (E value < 10^{-5}), to the SwissProt and TrEMBL (49) databases. Gene models with no matches in these databases were identified as "hypothetical proteins." Pathway analysis for each annotated protein was derived from matched genes in the Kyoto encyclopedia of genes and genomes (KEGG) database (50). Gene Ontology (GO) (51) term assignments, motifs, and domains of genes were extracted with InterProScan (52), which analyzed peptide sequences against InterPro member databases (53), including ProDom, PROSITE, PRINTS, Pfam, PANTHER, and SMART.

Identification of TE-Related Genes. TE-related genes were identified by combing the following two approaches: (*i*) screening the annotated proteins containing "transposon," "transposase," "intergrase," or "reverse transcriptase" in the Pfam database (54), and (*ii*) aligning the annotated proteins to a TE library using TBLASTN and screening the proteins with an E value <1e-5.

Noncoding RNA Prediction. tRNA genes were identified by tRNAscan-SE (55) with default parameters. The RNAmmer (56) program was used to predict rRNAs. snRNAs and snoRNAs were searched against the Rfam database (57) using INFERNAL (58). miRNA genes were annotated in three steps. First, miRNA deep-sequencing reads were merged and mapped to each genome using bowtie (59) to generate a mapping. Then the file was mapped to rice miRNAs in mirbase to find all known miRNAs. Finally, we identified all of the miRNAs in ZS97RS1 and MH63RS1 using the known miRNA score as a threshold.

Collinearity and Segmental Duplication in ZS97RS1 and MH63RS1. MCScanX (60) was used to identify the collinearity of non-TE genes in ZS93RS1 and MH63RS1 with default parameters (match score, 50; match size, 5; gap penalty, –1; overlap window, 5; E-value, 1e-5). The alignment threshold for each collinear gene pair was set to an E-value of <1e-20.

RNA Sequencing and Data Analysis. For RNA sequencing, ZS97, MH63, and the hybrid were grown in a phytotron with the day/night cycle set at 14 h/10 h and a temperature of 32 °C/28 °C. Tissues of seedling shoot at the four-leaf stage, developing panicle at stage III, and flag leaf at heading were collected for RNA sequencing following the procedures and protocols described previously (61).

Although diverse programs were developed to quantify gene expression levels using mRNA sequencing, alignment of sequencing data of a hybrid is still challenging, as the hybrid contains the genomes of its two parents (62). A previous study showed that constructing individualized diploid genomes and transcriptomes by merging the parental genomes and transcriptomes for the hybrid improved the accuracy of quantification of gene expression levels compared with alignments to a single reference genome (61). This approach used RSEM (63), which has claimed to be capable of fully handling reads that map ambiguously between both isoforms and genes, to measure

transcript abundance after aligning reads using Bowtie (59). RSEM (63) was used to process all ZS97 and MH63 mRNA sequencing data with no more than one mismatch. The output of RSEM (63) included the read count and expression level of each gene and transcript. The mRNA sequencing data of ZS97 and MH63 were aligned to ZS97RS1 and MH63RS1 plus the gap regions, respectively, and RNA-seq data of the hybrid were aligned to the merged genome of the two parents. We excluded TE-related genes in the expression analyses, as only a few of them were expressed in ZS97, MH63, and their hybrid.

For the transcripts of non-TE genes that were determined to be derived from allelic genes in the parental genomes, the expression levels of parental alleles were added up to determine the expression levels in the hybrid genome. Expression levels for nonallelic genes determined by RSEM (63) were used directly.

ACKNOWLEDGMENTS. This work was supported by Grant 31330039 from the National Natural Science Foundation; the National Key Research and Development Program of China (2016YFD0100804); the 111 Project of China (to Q.Z.); the Project 2662015PY223 by the Fundamental Research Funds for the Central Universities (to J.Z.); the Bud Antle Endowed Chair of Excellence in Agriculture and Life Sciences; and the AXA Chair in Genome Biology and Evolutionary Genomics (to R.A.W.).

1. Ouyang Y, Zhang Q (2013) Understanding reproductive isolation based on the rice model. *Annu Rev Plant Biol* 64:111–135.
2. IRRI (1991) *World Rice Statistics 1990* (International Rice Research Institute, Manila, Philippines).
3. Huang X, et al. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* 42(11):961–967.
4. Xie W, et al. (2015) Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. *Proc Natl Acad Sci USA* 112(39): E5411–E5419.
5. Yu SB, et al. (1997) Importance of epistasis as the genetic basis of heterosis in an elite rice hybrid. *Proc Natl Acad Sci USA* 94(17):9226–9231.
6. Hua JP, et al. (2002) Genetic dissection of an elite rice hybrid revealed that heterozygotes are not always advantageous for performance. *Genetics* 162(4):1885–1895.
7. Hua J, et al. (2003) Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. *Proc Natl Acad Sci USA* 100(5):2574–2579.
8. Huang Y, et al. (2006) Heterosis and polymorphisms of gene expression in an elite rice hybrid as revealed by a microarray analysis of 9198 unique ESTs. *Plant Mol Biol* 62(4-5):579–591.
9. Zhou G, et al. (2012) Genetic composition of yield heterosis in an elite rice hybrid. *Proc Natl Acad Sci USA* 109(39):15847–15852.
10. Huang X, et al. (2015) Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. *Nat Commun* 6:6258.
11. Goff SA, Zhang Q (2013) Heterosis in elite hybrid rice: Speculation on the genetic and biochemical mechanisms. *Curr Opin Plant Biol* 16(2):221–227.
12. Wang X, et al. (2014) Global genomic diversity of *Oryza sativa* varieties revealed by comparative physical mapping. *Genetics* 196(4):937–949.
13. Zhang J, et al. (2016) Building two *indica* rice reference genomes with PacBio long-read and Illumina paired-end sequencing data. *Sci Data*, 10.1038/sdata.2016.76.
14. van Oeveren J, et al. (2011) Sequence-based physical mapping of complex genomes by whole genome profiling. *Genome Res* 21(4):618–625.
15. Zhang J, et al. (2016) Genome puzzle master (GPM): An integrated pipeline for building and editing pseudomolecules from fragmented sequences. *Bioinformatics*, 10.1093/bioinformatics/btw370.
16. Cheng Z, et al. (2002) Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* 14(8):1691–1704.
17. Parra G, Bradnam K, Ning Z, Keane T, Korf I (2009) Assessing the gene space in draft genomes. *Nucleic Acids Res* 37(1):289–297.
18. Zhi D, Raphael BJ, Price AL, Tang H, Pevzner PA (2006) Identifying repeat domains in large genomes. *Genome Biol* 7(1):R7.
19. Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* 6:11.
20. Chen J, Hu Q, Zhang Y, Lu C, Kuang H (2014) P-MITE: A database for plant miniature inverted-repeat transposable elements. *Nucleic Acids Res* 42(Database Issue): D1176–D1181.
21. International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436(7052):793–800.
22. Kozomara A, Griffiths-Jones S (2014) miRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* 42(Database Issue):D68–D73.
23. Xue W, et al. (2008) Natural variation in *Ghd7* is an important regulator of heading date and yield potential in rice. *Nat Genet* 40(6):761–767.
24. Fan C, 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(6):1164–1171.
25. Li Y, et al. (2011) Natural variation in *GS5* plays an important role in regulating grain size and yield in rice. *Nat Genet* 43(12):1266–1269.
26. Xu C, et al. (2015) Differential expression of *GS5* regulates grain size in rice. *J Exp Bot* 66(9):2611–2623.
27. Sun X, 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(4):517–527.
28. Liu Q, et al. (2011) A paralog of the MtN3/saliva family recessively confers race-specific resistance to *Xanthomonas oryzae* in rice. *Plant Cell Environ* 34(11):1958–1969.
29. Thiel T, et al. (2009) Evidence and evolutionary analysis of ancient whole-genome duplication in barley predating the divergence from rice. *BMC Evol Biol* 9:209.
30. Guyot R, Keller B (2004) Ancestral genome duplication in rice. *Genome* 47(3):610–614.
31. Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* 30(11):2478–2483.
32. Finn RD, Clements J, Eddy SR (2011) HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res* 39(Web Server Issue):W29–W37.
33. Cingolani P, et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6(2):80–92.
34. Hoff KJ, Stanke M (2013) WebAUGUSTUS–A web service for training AUGUSTUS and predicting genes in eukaryotes. *Nucleic Acids Res* 41(Web Server Issue):W123–W128.
35. Besemer J, Borodovsky M (2005) GeneMark: Web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res* 33(Web Server Issue):W451–W454.
36. Korf I (2004) Gene finding in novel genomes. *BMC Bioinformatics* 5:59.
37. Solovyev V, Kosarev P, Seledsov I, Vorobyev D (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol* 7(Suppl 1):S10.1–S10.12.
38. Dimmer EC, et al. (2012) The UniProt-GO annotation database in 2011. *Nucleic Acids Res* 40(Database Issue):D565–D570.
39. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2011) GenBank. *Nucleic Acids Res* 39(Database Issue):D32–D37.
40. Mochida K, Yoshida T, Sakurai T, Ogihara Y, Shinozaki K (2009) TriFLDB: A database of clustered full-length coding sequences from *Triticeae* with applications to comparative grass genomics. *Plant Physiol* 150(3):1135–1146.
41. Lu T, et al. (2008) RICD: A rice *indica* cDNA database resource for rice functional genomics. *BMC Plant Biol* 8:118.
42. Exner V, Hirsch-Hoffmann M, Gruissem W, Hennig L (2008) PlantDB - A versatile database for managing plant research. *Plant Methods* 4:1.
43. Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31.
44. She R, et al. (2011) genBlastG: Using BLAST searches to build homologous gene models. *Bioinformatics* 27(15):2141–2143.
45. Kent WJ (2002) BLAT–The BLAST-like alignment tool. *Genome Res* 12(4):656–664.
46. Grabherr MG, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29(7):644–652.
47. Haas BJ, et al. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* 9(1):R7.
48. Haas BJ, et al. (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* 31(19):5654–5666.
49. Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 28(1):45–48.
50. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28(1):27–30.
51. Dutkowski J, et al. (2013) A gene ontology inferred from molecular networks. *Nat Biotechnol* 31(1):38–45.
52. Zdobnov EM, Apweiler R (2001) InterProScan–An integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17(9):847–848.
53. Mitchell A, et al. (2015) The InterPro protein families database: The classification resource after 15 years. *Nucleic Acids Res* 43(Database Issue):D213–D221.
54. Punta M, et al. (2012) The Pfam protein families database. *Nucleic Acids Res* 40(Database Issue):D290–D301.
55. Lowe TM, Eddy SR (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25(5):955–964.
56. Lagesen K, et al. (2007) RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 35(9):3100–3108.
57. Gardner PP, et al. (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res* 39(Database Issue):D141–D145.
58. Nawrocki EP, Eddy SR (2007) Query-dependent banding (QDB) for faster RNA similarity searches. *PLOS Comput Biol* 3(3):e56.
59. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.
60. Wang Y, et al. (2012) MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* 40(7):e49.
61. Wang J, Yao W, Zhu D, Xie W, Zhang Q (2015) Genetic basis of sRNA quantitative variation analyzed using an experimental population derived from an elite rice hybrid. *eLife* 4:e04250.
62. Munger SC, et al. (2014) RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. *Genetics* 198(1):59–73.
63. Li B, Dewey CN (2011) RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323.

SI Appendix

Figures

Fig. S1. Some features of ZS97, MH63 and their hybrid SY63. **(A)** Plants of ZS97, MH63 and the hybrid, and the seeds produced per plant. **(B)** Positions of ZS97 and MH63 in the *indica I* (brown) and *indica II* (green) groups by genome sequence analysis[1]. **(C)** Planting area of SY63 (the hybrid) in the last 30 years. Total rice area in China is from State Statistic Bureau (http://www.stats.gov.cn/) and the area for SY63 is from National Agricultural Technology Extension and Service Center (NATESC) of China.

Fig. S2. Distributions of SNPs and InDels between ZS97RS1 and MH63RS1 genomes. The x axis indicates the coordinate of MH63RS1genome.

Fig. S3. The largest inversion (IV12010) between ZS97RS1 and MH63RS1 genomes detected on chromosome 12. **(A)** Illustration of the inversion and BAC-clone SMRT sequencing reads across the breakpoints of IV12010. **(B)** PCR and Sanger sequencing reads across the breakpoints of IV12010. **(C)** Distribution of NGS meta-pair reads across the breakpoints of IV12010.

Fig. S4. Distributions of translocations (> 1 kb) between ZS97RS1 and MH63RS1 genomes. **(A)** Intrachromosomal translocations between the two genomes. **(B)** Interchromosomal translocations between the two genomes.

Fig. S5. Example of a unique presence region (MH07p00800) in MH63RS1 genome. **(A)** Illustration of the PAV region. **(B)** SMRT sequencing and NGS reads in the ZS97RS1. **(C)** SMRT sequencing and NGS reads in the left breakpoint of MH63RS1. **(D)** SMRT sequencing and NGS reads in the right breakpoint of MH63RS1.

Fig. S6. Integrated annotation pipeline for the ZS97RS1 and MH63RS1 genomes. This pipeline included genome anchoring, repetitive element identification, non-coding RNA prediction and protein-coding gene annotation. The annotation workflow is shown by arrows. Related software, data and results for each step are indicated in round-cornered boxes.

Tables

Table S1. CentO (155 bp satellite DNA) units in ZS97RS1 and MH63RS1 genomes

Table S2. Statistics of the completeness of the ZS97RS1, MH63RS1 and Nipponbare genomes based on 248 core eukaryotic genes (CEGs)

Table S3. Comparison of SNPs and InDels between three rice genomes

Table S4. Comparison of base transition and transversion between three rice genomes

Table S5. Summary of transposable elements in ZS97, MH63 and Nipponbare genomes

Table S6. The ratio of MITEs in the ZS97, MH63 and Nipponbare genomes

Table S7. Numbers of annotated non-coding RNAs in the ZS97 and MH63 genomes

Table S8. Comparison of rice gene annotation among the ZS97, MH63 and Nipponbare genomes

Table S9. Features of TE-related and non-TE-related gene loci in the ZS97, MH63 and Nipponbare genomes

Table S10. Distribution of SNPs and InDels in different genome features in a comparison between Nipponbare and ZS97

Table S11. Distribution of SNPs and InDels in different genome features in a comparison of Nipponbare and MH63

Table S12. Distribution of SNP and InDel in different genome features in a comparison of ZS97 and MH63

Table S13. TE content in PAV regions of the ZS97RS1 and MH63RS1 genomes

Table S14. Numbers of non-TE related genes in different categories in the ZS97RS1 and MH63RS1 genomes

Dataset S1

Section 1. Gap estimation in the ZS97RS1 genome

Section 2. Gap estimation in the MH63RS1 genome

Section 3. Sequence quality based upon BAC overlapping regions in the ZS97RS1 genome

Section 4. Sequence quality based upon BAC overlapping regions in the MH63RS1 genome

Section 5. Inversions between the ZS97RS1 and MH63RS1 genomes

Section 6. Inversions between the ZS97RS1 and Nipponbare genomes

Section 7. Inversions between the MH63RS1 and Nipponbare genomes

Section 8. Translocations between the ZS97RS1 and MH63RS1 genomes

Section 9. Unique presence regions in the ZS97RS1 genome compared with the MH63RS1 genome

Section 10. Unique presence regions in the MH63RS1 genome compared with the ZS97RS1 genome

Section 11. Unique presence regions in the ZS97RS1 genome compared with the Nipponbare genome

Section 12. Unique presence regions in the MH63RS1 genome compared with the Nipponbare genome

Section 13. Cloned genes in rice and their similarity and expression in ZS97 and MH63

Section 14. Collinearity blocks and gene pairs identified in the ZS97RS1 genome

Section 15. Collinearity blocks and gene pairs identified in the MH63RS1 genome

Section 16. Different collinearity blocks (contain 3 or more PAV/divergent genes) identified in the ZS97RS1 and MH63RS1 genomes

Section 17. Tandem duplicated genes in the ZS97RS1 genome

Section 18. Tandem duplicated genes in the MH63RS1 genome

References

1. Xie W, et al. (2015) Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. *Proc Natl Acad Sci USA* 112:E5411–E5419.

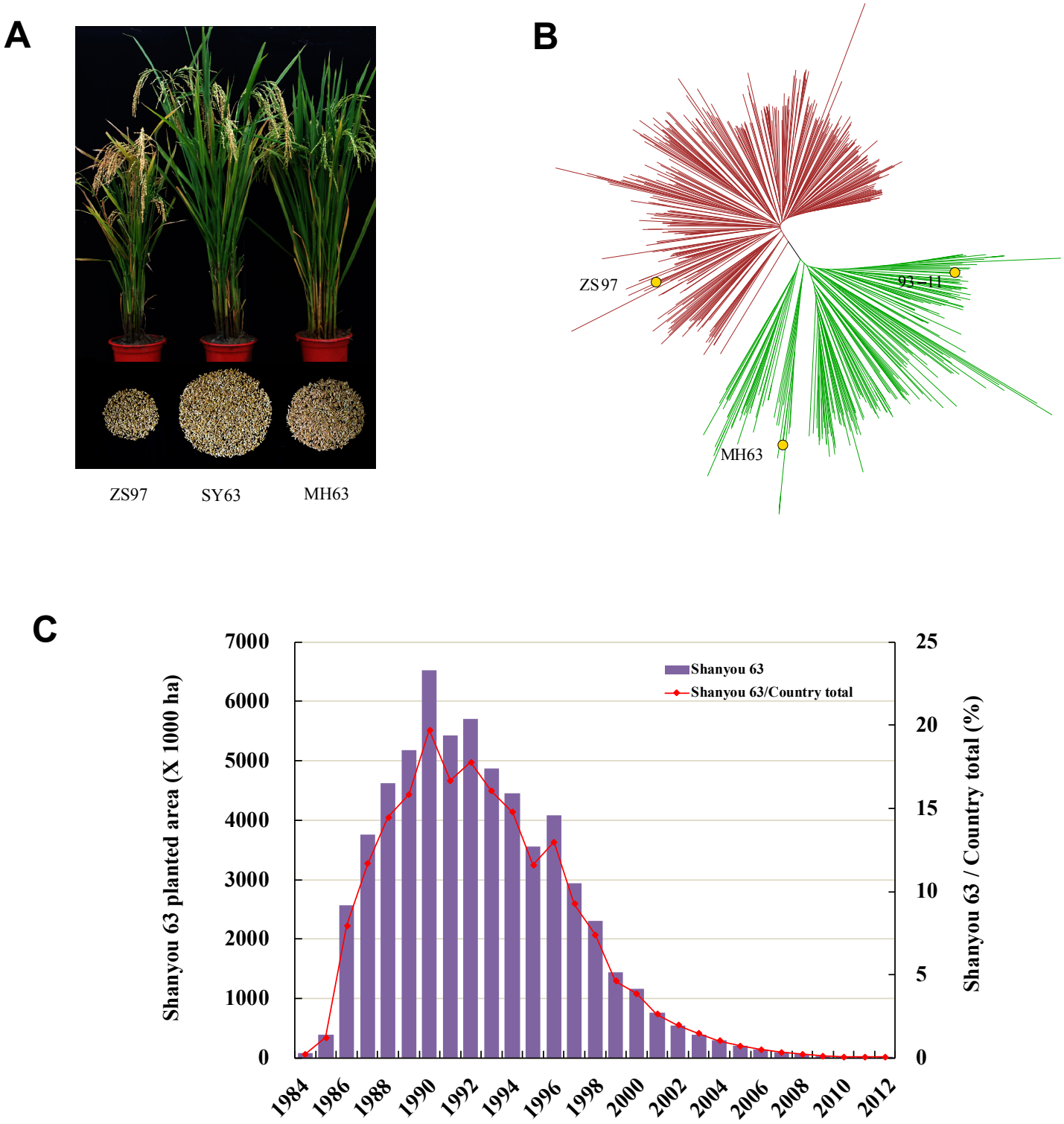

Fig. S1. Some features of ZS97, MH63 and the hybrid. **(A)** Plants of ZS97, MH63 and the hybrid, and the seeds produced per plant. **(B)** Positions of ZS97 and MH63 in the indica I (brown) and indica II (green) groups by genome sequence analysis (1). **(C)** Planting area of SY63 (the hybrid) in the last 30 years. Total rice area in China is from State Statistic Bureau (http://www.stats.gov.cn/) and the area for SY63 is from National Agricultural Technology Extension and Service Center (NATESC) of China.

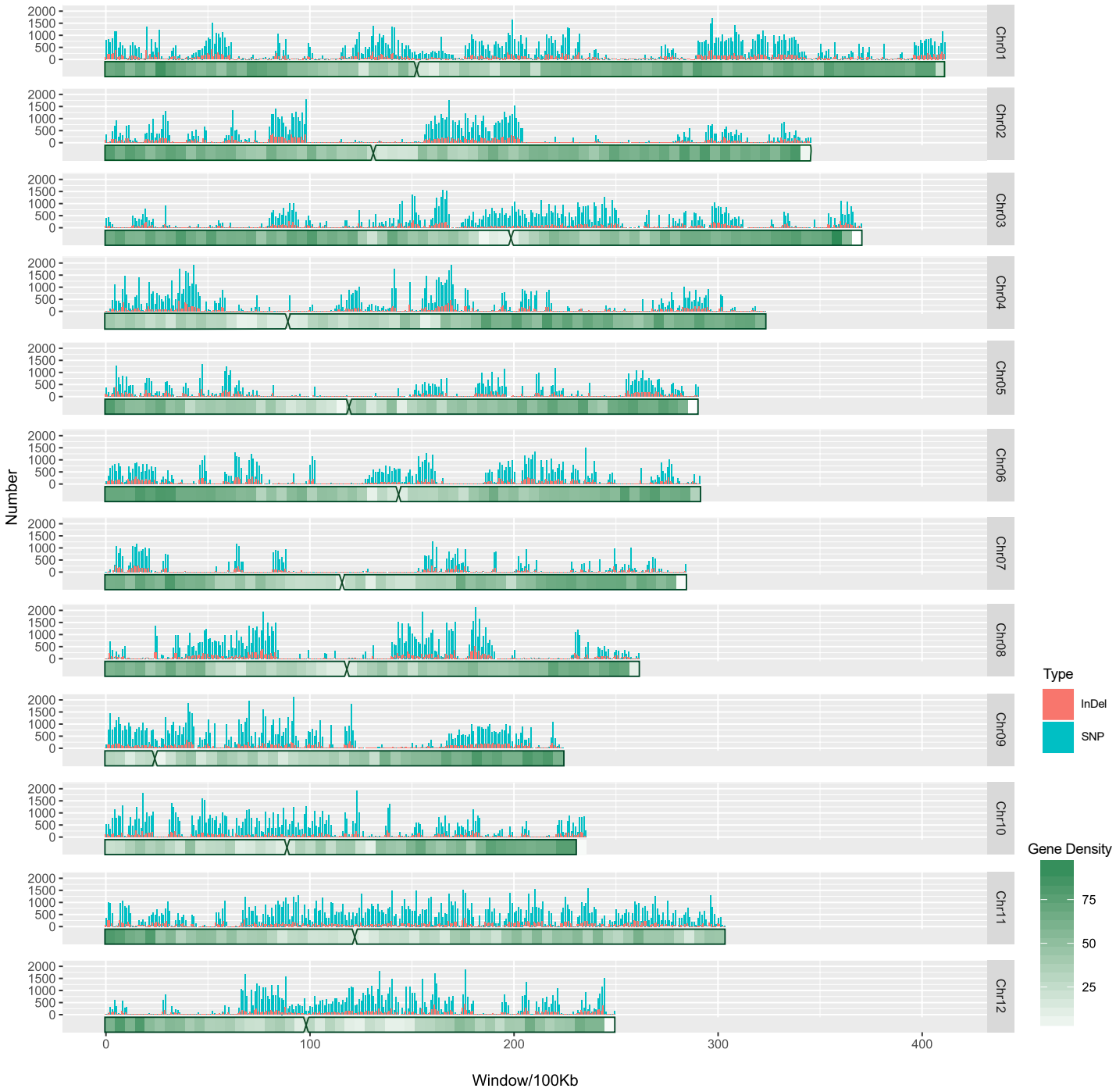

Fig. S2. Distributions of SNPs and InDels between ZS97RS1 and MH63RS1 genomes. The x axis indicates the coordinate of MH63RS1 genome.

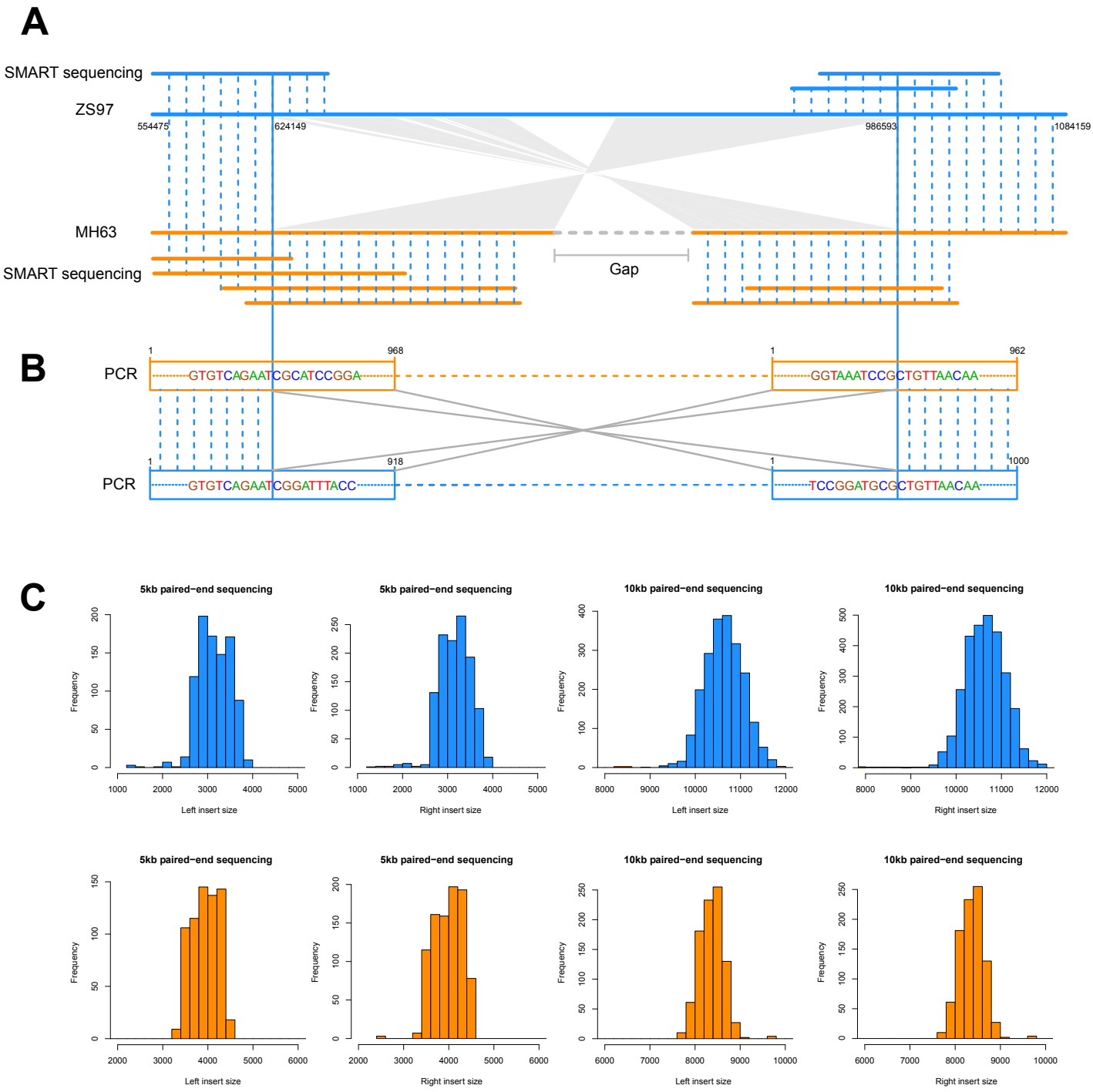

Fig. S3. The largest inversion (IV12010) between ZS97RS1 and MH63RS1 genomes detected on chromosome 12. **(A)** Illustration of the inversion and BAC-clone SMRT sequencing reads across the breakpoints of IV12010. **(B)** PCR and Sanger sequencing reads across the breakpoints of IV12010. **(C)** Distribution of NGS meta-pair reads across the breakpoints of IV12010.

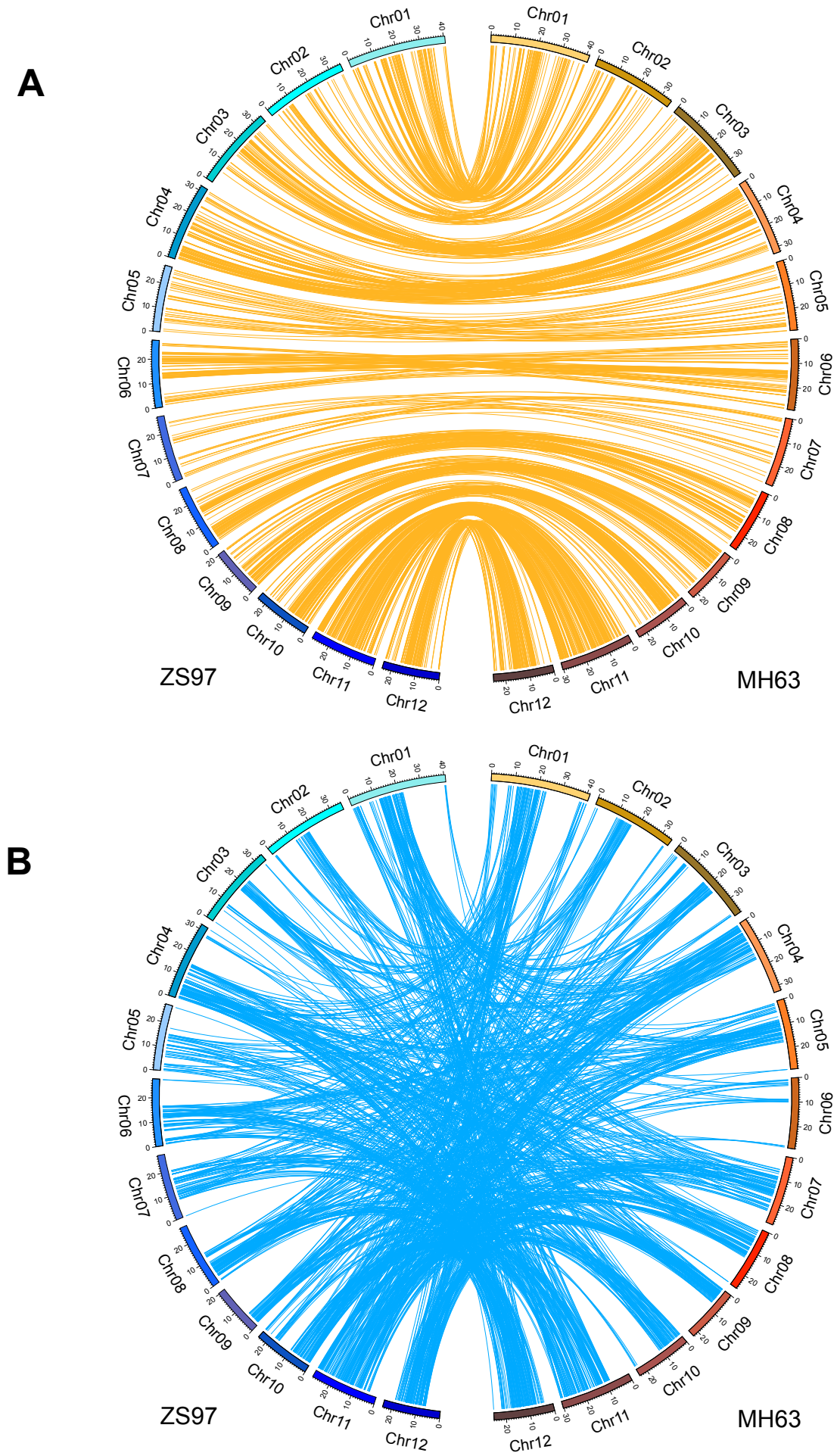

Fig. S4. Distributions of translocations (> 1 kb) between ZS97RS1 and MH63RS1 genomes. **(A)** Intrachromosomal translocations between the two genomes. **(B)** Interchromosomal translocations between the two genomes.

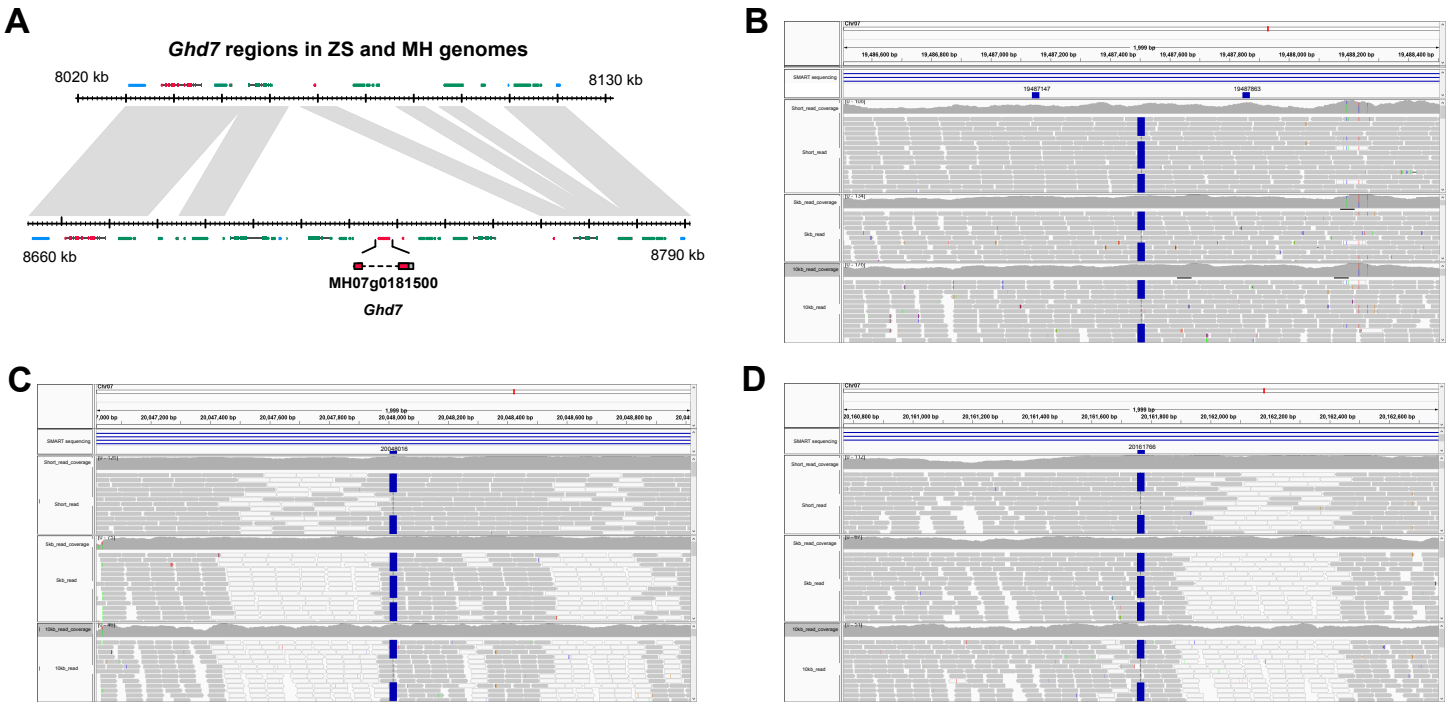

Fig. S5. Example of a unique presence region (MH07p00800) in MH63RS1 genome. **(A)** Illustration of the PAV region. **(B)** SMRT sequencing and NGS reads in the ZS97RS1. **(C)** SMRT sequencing and NGS reads in the left breakpoint of MH63RS1. **(D)** SMRT sequencing and NGS reads in the right breakpoint of MH63RS1.

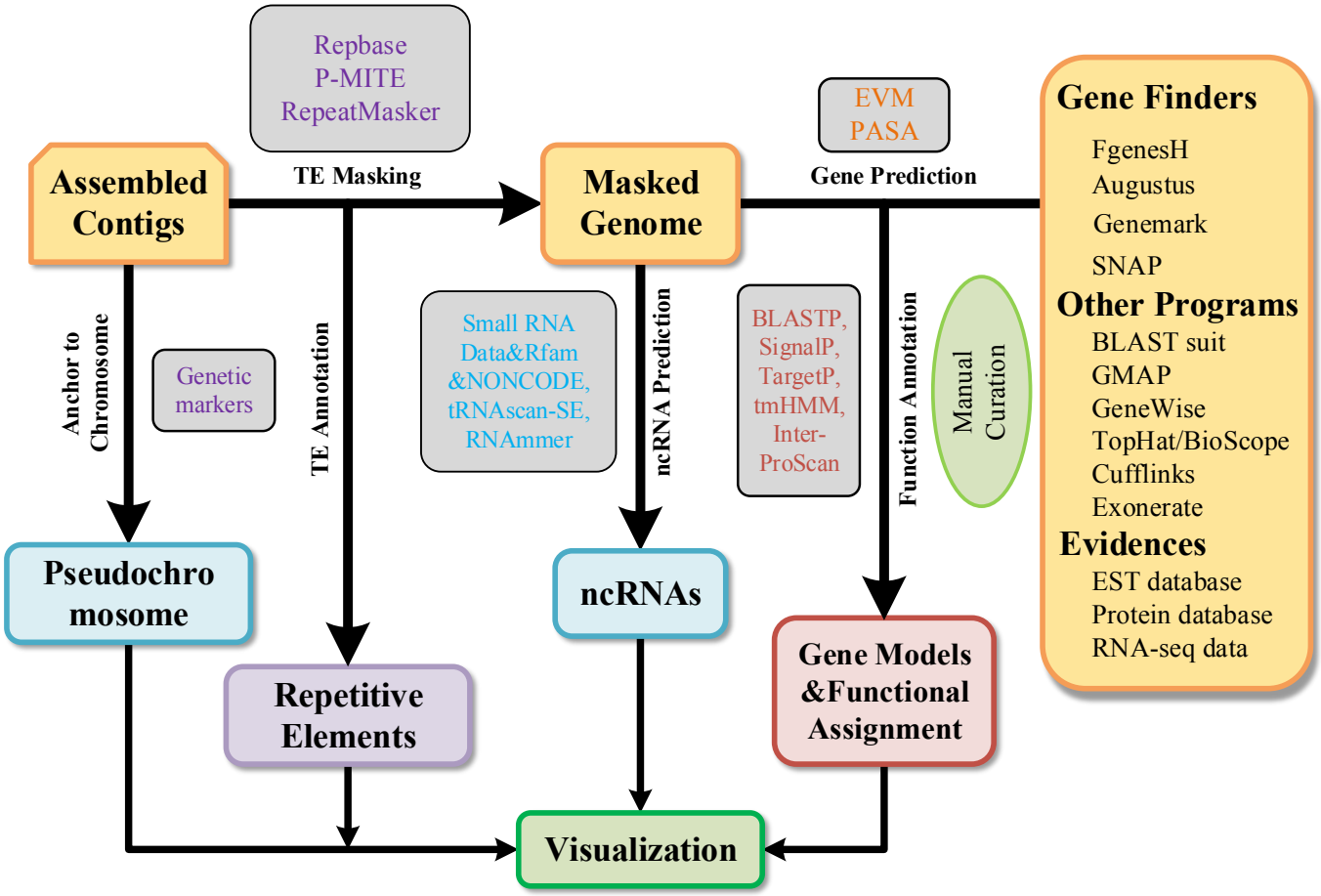

Fig. S6. Integrated annotation pipeline for the ZS97RS1 and MH63RS1 genomes. This pipeline included genome anchoring, repetitive element identification, non-coding RNA prediction and protein-coding gene annotation. The annotation workflow is shown by arrows. Related software, data and results for each step are indicated in round-cornered boxes.

Table S1. CentO (155 bp satellite DNA) units in ZS97RS1 and MH63RS1 genomes

	ZS97				MH63			
Chr	Completeness	CentO sequence location	Total no. of units	Total length (bp)	Completeness	CentO sequence location	Total no. of units	Total length (bp)
1	Partial	16,150,042– 16,170,949	117	20,908	Partial	16,113,268– 16,474,603	863	361,336
2	Partial	13,750,944 – 13,779,000	66	28,057	Partial	13,149,113– 13,175,261	54	26,149
3	Partial	17,836,719 – 19,520,098	1,192	1,683,380	Partial	19,843,069– 20,100,373	476	257,305
4	Partial	8,794,905 – 8,912,722	302	117,818	Partial	8,535,237– 8,616,246	192	81,010
5	Partial	11,275,182 – 11,278,964	24	3,783	Partial	11,849,549– 11,853,331	23	3,783
6	N/A	–	–	_	Complete	14,517,820– 14,641,670	741	123,851
7	Partial	11,215,832 – 11,441,556	262	225,725	Partial	11,583,370– 11,585,417	13	2,048
8	Complete	11,382,536 –12,437,644	495	1,055,109	Complete	11,257,642– 12,195,055	483	937,414
9	Partial	2,568,970 – 2,572,174	22	3,205	Complete	2,134,375– 2,911,387	735	777,013
10	Complete	8,249,417 – 9,528,056	208	1,278,640	Partial	8,811,120– 8,856,329	71	45,210
11	Partial	11,186,264 –15,446,386	641	4,260,123	Partial	12,256,999– 12,259,359	15	2,361
12	Partial	10,351,868 – 10,423,961	154	72,094	Complete	9,707,620– 10,131,713	1,138	424,094

Table S2. Statistics of the completeness of the ZS97RS1, MH63RS1 and Nipponbare genomes based on 248 core eukaryotic genes (CEGs)

		ZS97		MH63		Nipponbare	
	No. of CEGs	Number of proteins	Completeness	Number of proteins	Completeness	Number of proteins	Completeness
Group1	66	60	90.91%	60	90.91%	65	98.48%
Group2	56	53	94.64%	52	92.86%	54	96.43%
Group3	61	56	91.80%	59	96.72%	59	96.72%
Group4	65	61	93.85%	64	98.46%	64	98.46%
Total	248	231	92.74%	236	94.76%	242	97.58%

No. of proteins: number of 248 ultra-conserved CEGs present in each genome;
Completeness: percentage of 248 ultra-conserved CEGs present;
Total: total number of CEGs present including putative orthologs.

Table S3. Comparison of SNPs and InDels between three rice genomes

	Nipponbare *vs*. ZS97	Nipponbare *vs*. MH63	ZS97 *vs*. MH63
SNPs	2,665,280	2,746,894	1,300,802
SNPs/kb	7.14	7.36	3.65
InDels	486,015	501,741	251,837
InDels/kb	1.31	1.34	0.70

Table S4. Comparison of base transition and transversion between three rice genomes

Substitution	Nipponbare *vs*. ZS97		Nipponbare *vs*. MH63		ZS97 *vs*. MH63	
	Number	Ratio	Number	Ratio	Number	Ratio
A->G	447,240	0.1678	461,898	0.1682	229,335	0.1762
G->A	497,244	0.1866	509, 875	0.1856	230,827	0.1774
C->T	495,844	0.1860	509,774	0.1856	229,698	0.1765
T->C	449,693	0.1687	463,994	0.1689	228,140	0.1753
A->C	99,660	0.0374	103,591	0.0377	48,624	0.0374
C->A	103,398	0.0388	106, 212	0.0387	49, 599	0.0381
A->T	110,671	0.0415	115,421	0.0420	56,357	0.0433
T->A	112,538	0.0422	115,242	0.0420	57, 383	0.0441
C->G	72,939	0.0274	75, 744	0.0276	36,241	0.0279
G->C	73,645	0.0276	76,003	0.0277	36,753	0.0282
G->T	102,695	0.0385	106,212	0.0387	48,958	0.0376
T->G	99,713	0.0374	102,846	0.0374	49,338	0.0379
Transition	1,890,021	0.7091	1, 945,541	0.7083	917,991	0.7055
Transversion	775,259	0.2909	801,353	0.2917	383,253	0.2945
Ts/Tv	2.44		2.43		2.40	

Table S5. Summary of transposable elements in ZS97, MH63 and Nipponbare genomes

TE Classification	ZS97			MH63			Nipponbare		
	Copy no.	Length (kb)	Percentage (%)	Copy no.	Length (kb)	Percentage (%)	Copy no.	Length (kb)	Percentage (%)
Retrotransposons	**86,554**	**87,815**	**25.32**	**90,353**	**92,184**	**25.61**	**87,719**	**84,546**	**22.59**
SINEs	8,522	1,260	0.36	8,901	1,317	0.37	9,638	1,426	0.38
LINEs	5,759	2,932	0.85	6,035	3,102	0.86	6,502	3,350	0.90
Ty1/copia	11,184	9,462	2.73	11,486	9,730	2.70	12,829	11,346	3.03
Ty3/gypsy	58,743	72,938	21.03	61,441	76,695	21.31	55,901	66,689	17.82
Others	2,346	1,223	0.35	2,490	1,340	0.37	2,849	1,735	0.46
DNA transposons	**233,819**	**55,006**	**15.86**	**241,846**	**57,027**	**15.84**	**262,151**	**63,146**	**16.87**
hAT	19,995	3,854	1.11	20,621	3,958	1.10	22,353	4,337	1.16
CACTA	22,460	11,410	3.29	23,589	11,925	3.31	26,185	14,404	3.85
Mariner	53,629	8,638	2.49	55,519	8,940	2.48	60,297	9,689	2.59
Mutator	53,103	14,174	4.09	54,453	14,675	4.08	58,816	15,736	4.20
Helitron	8,080	2,943	0.85	8,618	3,081	0.86	9,438	3,375	0.90
Mim	363	107	0.03	387	117	0.03	585	182	0.05
Tourist/Harbinger	60,227	10,768	3.10	62,121	11,126	3.09	66,598	11,959	3.19
Others	15,962	3,110	0.90	16,538	3,205	0.89	17,879	3,461	0.92
Unclassified	**1,509**	**373**	**0.11**	**1,594**	**452**	**0.13**	**1,911**	**448**	**0.12**
Total based masked	**321,882**	**143,193**	**41.28**	**333,793**	**149,662**	**41.58**	**351,781**	**148,140**	**39.58**

Table S6. The ratio of MITEs in the ZS97, MH63 and Nipponbare genomes

	ZS97	MH63	Nipponbare
MITEs[a] (%)	9.58	9.54	9.96
DNA transposons (%)	15.86	15.84	16.87
MITE/ DNA transposons (%)	60.40	60.23	59.04

[a] MITEs include: hAT, CACTA, Mariner, Mutator, Mim and Harbinger DNA transposons.

Table S7. Numbers of annotated non-coding RNAs in the ZS97 and MH63 genomes

	ZS97	MH63	Syntenic
rRNA	40	60	13
tRNA	592	589	516
snoRNA	449	457	400
snRNA	92	97	79
miRNA	341	363	287

Table S8. Comparison of rice gene annotation among the ZS97, MH63 and Nipponbare genomes

	ZS97	MH63	Nipponbare
Gene number	54,831	57,174	55,968
Transcripts	78,033	80,581	66,338
Non-TE gene loci	34,610	37,324	39,045
Multiple exon gene loci	42,829	44,926	43,781
Exons/gene	5.19	5.23	4.71
Gene size (including intron)	3,111	3,137	2,965
Transcript size	1,955	1,962	1,708
CDS length	1,348	1,368	1,342
5’ UTR length	397	398	259
3’ UTR length	737	736	466

Table S9. Features of TE-related and non-TE-related gene loci in the ZS97, MH63 and Nipponbare genomes

	TE-related genes			Non-TE-related genes		
	ZS97	MH63	Nipponbare	ZS97	MH63	Nipponbare
Number of gene loci	20,221	19,850	16,941	34,610	37,324	39,045
Number of transcripts	20,766	20,378	17,272	57,267	60,203	46,051
Transcripts/gene locus	1.03	1.03	1.02	1.65	1.61	1.26
Exons/transcript	3.41	3.64	4.2	5.84	5.77	4.9
Gene size (bp)	2,566	2,759	3,224	3,431	3,338	2,853
Transcripts length (bp)	1,741	1,848	2,150	2,033	2,001	1,552
CDS length (bp)	1,677	1,783	2,081	1,229	1,227	1,082
Intron length bp)	377	377	345	441	442	416
5'-UTR length (bp)	537	507	511	743	744	464
3'-UTR length (bp)	495	454	248	394	396	254

Table S10. Distribution of SNPs and InDels in different genome features in a comparison between Nipponbare and ZS97

	SNP			InDel		
	Number	Proportion (%)	Density[a]	Number	Proportion (%)	Density[a]
Intergenic	649,444	24.4	6.25	155,178	31.92	1.57
5'-UTR	29,764	1.12	4.51	10,846	2.23	1.65
3'- UTR	49,317	1.85	4.17	14,654	3.02	1.24
Intron	23,3845	8.77	4.78	60,830	12.51	1.24
CDS	173,125	6.50	4.27	12,891	2.95	0.32
TE	1,529,785	57.40	10.34	231,616	47.66	1.57

[a] Density, number of SNP per kilo base or number of InDel per kilo base based on ZS97RS1.

Table S11. Distribution of SNPs and InDels in different genome features in a comparison of Nipponbare and MH63

	SNP			InDel		
	Number	Proportion (%)	Density[a]	Number	Proportion (%)	Density[a]
Intergenic	657,128	23.92	7.11	157,977	31.49	1.71
5'-UTR	30,399	1.11	4.62	10,936	2.18	1.66
3'- UTR	51,979	1.89	4.33	15,523	3.09	1.29
Intron	251,778	9.16	4.92	65,289	13.01	1.28
CDS	194,104	7.07	4.51	14,822	2.95	0.34
TE	1,561,506	56.85	10.09	237,194	47.27	1.53

[a] Density, number of SNP per kilo base or number of InDel per kilo base based on MH63RS1.

Table S12. Distribution of SNP and InDel in different genome features in a comparison of ZS97 and MH63

	SNP			InDel		
	Number	Proportion (%)	Density[a]	Number	Proportion (%)	Density[a]
Intergenic	294,387	22.62	3.24	76,034	30.18	0.84
5'- UTR	13,478	1.04	2.04	5,314	2.11	0.81
3'- UTR	23,172	1.78	1.96	7,317	2.90	0.62
Intron	113,324	8.71	2.32	31,861	12.65	0.65
CDS	89,320	6.86	2.20	6,814	2.70	0.17
TE	767,121	58.99	5.19	124,497	49.45	0.84

[a] Density, number of SNP per kilo base or number of InDel per kilo base based on MH63RS1.

Table S13. TE content in PAV regions of the ZS97RS1 and MH63RS1 genomes

	ZS97			MH63		
	PAV length	TE length	TE/PAV ratio (%)	PAV length	TE length	TE/PAV ratio (%)
Chr01	2,188,540	1,429,180	65.3	2,033,418	1,226,441	60.3
Chr02	1,562,465	1,055,362	67.5	1,060,488	618,575	58.3
Chr03	914,995	700,607	76.6	1,157,664	916,690	79.2
Chr04	2,826,888	1,809,408	64.0	2,885,613	2,049,310	71.0
Chr05	491,437	330,206	67.2	1,016,710	688,585	67.7
Chr06	1,496,421	1,022,089	68.3	1,289,736	892,588	69.2
Chr07	1,261,279	924,022	73.3	1,431,159	917,990	64.1
Chr08	2,012,183	1,428,388	71.0	1,629,303	1,070,910	65.7
Chr09	1,300,862	727,901	56.0	1,903,535	1,206,488	63.4
Chr10	2,033,377	1,425,390	70.1	1,945,173	1,378,923	70.9
Chr11	3,403,673	2,051,292	60.3	4,515,354	2,706,773	59.9
Chr12	1,986,088	1,411,701	71.1	2,546,706	1,725,385	67.7
Average	**21,478,208**	**14,315,526**	**66.7**	**23,414,859**	**15,398,658**	**65.8**

Table S14. Numbers of non-TE related genes in different categories in the ZS97RS1 and MH63RS1 genomes

	ZS97RS1	MH63RS1
Identical genes	15,214	15,214
Non-synonymous genes	4,174	4,174
Other collinear genes	5,932	5,932
Genes in unique present regions	1,389	1,713
Genes in gap regions	1,891	2,957
Divergent genes	6,010	7,334
Total non-TE genes	34,610	37,324