Gapless *indica* rice genome reveals synergistic effects of active transposable elements and segmental duplications that promote rice genome evolution

Kui Li¹, Wenkai Jiang², Yuanyuan Hui¹, Mengjuan Kong¹, Li-Zhi Gao³, Pengfu Li^{1,*}, Shan Lu^{1,*}

1. State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing 210023, China

2. Novogene Bioinformatics Institute, Building 301, Zone A10 Jiuxianqiao North Road, Chaoyang District, Beijing 100083, China

3. Institution of Genomics and Bioinformatics, South China Agricultural University, Guangzhou 510642, China.

*Correspondence: shanlu@nju.edu.cn, pengfuli@nju.edu.cn

Running title: gapless rice genome

Abstract

The ultimate goal of genome assembly is a high-accuracy gapless genome. Here we report a new assembly pipeline which we have used to produce a gapless genome for the *indica* rice cultivar Minghui 63. The 395.82 Mb final assembly is composed of 12 contigs with a contig N50 size of 31.82 Mb. All chromosomes are now gapless, with each chromosome represented by a single contig. This is the first gapless genome assembly achieved for higher plants or animals. BUSCO evaluation showed that gene regions of our assembly have higher completeness than the current rice reference genome (IRGSP-1.0). Compared with *japonica* rice, *indica* has more transposable elements (TEs) and segmental duplications (SDs), the latter of which produce many duplicated genes that can affect plant traits through dose effect or sub-/neo-functionalization. The insertion of TEs can also affect the expression of duplicated genes, which may drive evolution of these genes. We also found the expansion of *NBS-LRR* disease resistance genes and *cZOGT* growth-related genes in SDs, suggesting that SDs contribute to the adaptative evolution of rice disease resistance and developmental processes. Our findings suggest that active TEs and SDs together provide synergistic effects to promote rice genome evolution.

Keywords

gapless genomes, genome evolution, rice genome, segmental duplications, transposable elements

1 Introduction

2 Repeat sequences, such as transposable elements (TEs), centromeres, and segmental
3 duplications (SDs), represent the most challenging genomic regions to assemble
4 (Alkan et al., 2011; Bailey et al., 2001; Elert, 2014; Liu et al., 2020a; Michael and
5 VanBuren, 2015; Numanagic et al., 2018). However, these regions can play major
6 roles in the genome evolution of a lineage. For example, TEs participate in shaping
7 gene expression patterns and the formation of genome structure. This is especially
8 true for long terminal repeat retrotransposons (LTR-RTs), which are the most
9 prevalent repeats in plant genomes, and their proliferation could result in genome
10 bloating (Elert, 2014). Other repeat regions, such as centromeres, are of central
11 importance to chromosome stability (Mackinnon et al., 2013). SDs are similarly
12 influential on genome structure due to their production of numerous repeat genes, and
13 thus serve as hotbeds for genomic rearrangements followed by gene innovation and
14 rapid gain-of-function adaptations (Han et al., 2009; Marques-Bonet et al., 2009;
15 Zhang et al., 1998).

16 With the development of long-read sequencing technology, the continuity of
17 genome assembly has been greatly increased by orders of magnitude. Some recently
18 developed technologies such as binano (Jiao et al., 2017), HIC (Dudchenko et al.,
19 2017), and 10x Genomics (Mostovoy et al., 2016; Yeo et al., 2018) have further
20 improved the continuity of assembly. The genomes of rice (Du et al., 2017), maize
21 (Liu et al., 2020a), and rose (Raymond et al., 2018), for example, have only a few
22 unresolved gaps. However, the ultimate goal of assembly, *i.e.*, a highly accurate
23 gapless genome, has yet to be achieved. High-fidelity (HiFi) sequence reads,
24 generated via circular consensus sequencing (CCS), are the first data type that has
25 advantages in both read lengths (greater than 10 kb) and accuracy (greater than 99%).
26 In this method, subreads from a single polymerase read are computationally combined
27 via the CCS algorithm to create a HiFi consensus read (Logsdon et al., 2020). This
28 produces high-quality reads without overlapping and error correction, which may

introduce errors caused by incorrect overlaps (especially for repeat sequence regions), making it possible to obtain high-quality genome assemblies.

Since rice (*Oryza sativa* L.) is a staple crop consumed by more than half of the global population (Elert, 2014), its genome is of broad interest for studies from the relationship and impacts of evolution on plant traits to the selection for high yield or disease resistance to maximize production. *Indica* rice is the most widely produced rice subspecies due to several desirable traits, and interestingly, has a substantially larger genome than another cultivated subspecies *japonica* (Du et al., 2017). Recent studies have shown that TEs, centromeres, and SDs all have strong evolutionary effects on the genome (Elert, 2014; Han et al., 2009; Mackinnon et al., 2013; Marques-Bonet et al., 2009; Zhang et al., 1998). However, the prevalence of these repeat sequences in plant genomes also hinders the assembly of high-quality genomes, for which current sequencing technologies and algorithms have considerable difficulties in chromosome-level assemblies without gaps. Here, with the help of other genome research methods, such as genetic maps (Catchen et al., 2020), HIC (Dudchenko et al., 2017), or alignment with genome sequences of related species, we developed a feasible approach to achieve the gapless assembly of the *indica* rice cultivar Minghui 63 based on HiFi reads. Further, we used our gapless genome to prove the hypothesis that recent TEs and duplicated genes in *indica* rice provide synergistic effects on genome evolution that impact adaptative traits. Our results indicate that *indica* rice has many SDs that result in gene duplications necessary for disease resistance and biomass accumulation. We also found a preponderance of active TE insertions in the proximity of the duplicated genes with low expression levels, while a few highly expressed duplicated genes appeared to have undergone sub- or neo-functionalization.

Results

56 **Gapless assembly of the rice genome and feature annotation**

57 We used PacBio HiFi reads from *indica* rice variety Minghui 63 to reassemble the
58 genome. PacBio HiFi reads provide long fragment reads with 99% accuracy, which
59 avoid the necessity of error correction before assembly. This is particularly useful for
60 the assembly of high-repeat areas such as centromeres, telomeres, TEs, and SDs, in
61 which read alignments are often collapsed during error corrections. Briefly, 25.3 Gb
62 HiFi reads from the National Center for Biotechnology Information (NCBI) Sequence
63 Read Archive database (SRX6957825) were used for preliminary assembly using pb-
64 assembly (Wenger et al., 2019) (Fig 1A, Table S1). Then, the reference genome of
65 *indica* rice (MH63RS2) (Zhang et al., 2016) was used to anchor contigs onto
66 chromosomes (Fig. 1B). With the mapping for each contig location (Table S2), every
67 chromosome was represented by a list of contig paths (encoding the graph for each
68 contig and represented by a list of unitigs) and gaps between them. If two contigs
69 were adjacent, they were connected without a gap. After this step, there were only 25
70 gaps remained between contig paths. We further reconstructed the string graph using
71 edges (sg_edges_list) and overlap information (pread.m4) generated by pb-assembly,
72 found all the paths between each gap, and then selected the path with the most
73 overlaps for each gap. Most of the paths were simple paths, except for a few complex
74 paths spanning multiple repeat regions (Fig. 1C). Through the above steps, 24 gap
75 areas were filled and one gap was removed for two flanking contigs becoming
76 adjacent after removing possible assembly errors. We obtained a path from the start
77 point to the end point for each chromosome, thus resulting in a gapless assembly of
78 the *indica* rice genome (Fig. 1D). Here we found MH63RS2 served as a good
79 reference for locating and orienting contigs to the 12 chromosomes to facilitate
80 assembly. Alternatively, genetic maps (Catchen et al., 2020), HIC (Dudchenko et al.,
81 2017), or other biological methods can also be used to anchor contigs.

Figure 1. Schematic depiction of the assembly process.

(A) Whole-genome assembly from High-fidelity (HiFi) reads. (B) Anchoring WGS contigs onto chromosomes. (C) String graphs were reconstructed using edges (sg_edges_list) and overlap information (pread.m4) generated by pb-assembly. Contigs were linked following their order on each chromosome and a unitig (typically a repeat sequence) for each gap between contigs can be found in the string graph. Red blocks indicate gaps; solid lines indicate contig paths; dashed lines indicate gap locations. (D) The final assembly graph. Each color represents one chromosome.

To assess the quality of the 24 filled gap areas, we compared the gap region sequences in our new assembly with MH63RS2. Among these gap region sequences, 18 were aligned to the correct positions and showed good collinearity (Fig. 2, Fig. S1, Table S3). One gap region was aligned to a previously known gap area in MH63RS2. For the remaining five gap region sequences, we checked the assembly graph and found that these gap sequences all spanned multiple repeat regions, which may require longer sequencing reads to resolve precisely (Fig. S2). Subsequently, we further used Racon (Vaser et al., 2017) to perform three rounds of correction. The final genome, namely MH63KL1, was assembled into 12 contigs corresponding to 12 chromosomes, resulting in a total length of 395.82 Mb with a contig N50 size of 31.82 Mb, which is both longer and more continuous than presently published rice genomes (Table 1, Table S1).

Table 1. Summary of different versions of the rice genome assemblies.

	MH63KL1	MH63RS1 ^a	MH63RS2 ^b	IRGSP-1.0 ^c
Total length	395,822,112	359,937,991	387,424,359	374,422,835
Contig number	12	201	26	971
Contig length	395,822,112	359,918,883	387,422,159	374,304,580
Largest contig	44,998,099	7,413,790	39,351,290	17,510,758
Contig N50	31,823,770	3,032,387	27,020,111	7,711,345
Contig L50	6	40	6	17
Length of Ns	0	19,108	2,200	118,255
Rate of Ns	0	0.0053%	0.0006%	0.0316%

^{a,b} Assemblies of the *indica* rice Minghui 63 genome in NCBI.

^c Assembly of the *japonica* rice Nipponbare genome in NCBI.

Figure 2. Distribution of repeat sequences and gap positions in MH63KL1.

The overlaid heatmap shows the TE density in a 1-Mb window. The tack labels refer to segmental duplications (SDs, green boxes) and the original gap locations (orange triangles).

In order to evaluate the quality of our assembly, Illumina paired-end reads were mapped to our assembly with a mapping rate of 99.34% and a coverage of 98.98% (Table S4). We also mapped 2,045 full-length cDNA sequences to the assembly, among which 1,971 (96.38%) could be mapped with greater than 50% coverage (Table S5). We used NUCmer (Marcais et al., 2018) and MCScan (Tang et al., 2008) to analyze collinearity between our assembly and MH63RS2, and found a high degree of synteny across the whole genome (Fig. S3). The completeness of gene regions of our assembly was evaluated using Benchmarking Universal Single-Copy Orthologs

(BUSCO) (Simao et al., 2015). Of the 1614 single-copy orthologs identified in embryophytes, 98.6% were complete in our assembly, which indicates a higher degree of completeness than MH63RS1 (93.1%), MH63RS2 (98.0%), and IRGSP-1.0 (98.3%) (International Rice Genome Sequencing, 2005). The results also indicate that our assembly contains the fewest missing (1.1%) and fragmented (0.3%) single-copy orthologs (Fig. S4, Table S6). LTR-RT annotation of our assembly revealed an LTR assembly index (LAI) score of 22.72, which matches the current gold standard (Fig. S5) (Ou et al., 2018). Finally, we compared our assembly with the sequences of 12 BACs obtained from GenBank, each containing one of the centromeres from rice (http://rice.plantbiology.msu.edu/annotation_pseudo_centromeres.shtml). All the BAC sequences were successfully aligned to our assembly in the correct order and with more than 90% coverage. Since most of the BAC sequences were incomplete, their corresponding areas in our assembly were actually longer (Fig. S6, Table S7). This implies that our assembly of the centromere regions was likely more reliable than sequences of the corresponding BACs.

From our assembly, a total of 36,149 PCGs were predicted, with a mean coding sequence size of 1,146.49 bp and an average of 4.36 exons per gene, similar to the predicted gene sizes in the genomes of other grass species (Table S10, Table S11). Of the PCGs, 99.8% were functionally annotated based on information from the NR (non-redundant, NCBI), SwissProt (Bairoch and Apweiler, 1999), InterPro (Hunter et al., 2012), Pfam (Punta et al., 2012), and KEGG databases (Kanehisa et al., 2014) (Table S12).

We also determined the locations of the centromeres in our assembly by searching for sequences homologous to the highly repetitive 155–165 bp CentO satellite DNA (Cheng et al., 2002). We found that the locations were consistent with previously published results (Zhang et al., 2016) (Table S13). Taken together, these results illustrate the high quality, reliability, and accuracy of our assembly.

SDs and functional diversification of paralogous gene pairs

SDs are genomic segments >1 Kb that are repeated within the genome with at least 90% sequence identity (Bailey et al., 2001). SDs often carry many duplicated genes and can thus serve as variable hotbeds of gene functional innovation. Due to limitations in the assembly technology, SD regions are often incorrectly assembled, collapsed into a single representative sequence, or completely missed. Consequently, this missing or incorrect information imposes limits on our operability in understanding the structure and evolution of a genome. Our gapless MH63KL1 rice genome provides an opportunity to accurately characterize SDs. To this end, we used SEDEF in this study to analyze the SDs in the rice genome (Bailey et al., 2001), which resulted in the identification of a total of 22,447 SDs, equivalent to 89.68 Mb in length, and accounting for 22.66% of the genome (Fig. 2, 3A, Table S14). We found that SDs are not randomly distributed across the rice genome, but rather, are preferentially located at higher frequencies on chr4, chr10, chr11, and chr12, and at lower frequencies on chr1, chr2, and chr3, suggesting that chr4, chr10, chr11, and chr12 may have a previously unrecognized contribution to the evolution of rice (Table S14).

We then identified duplicated genes in the SD regions. We initially performed an all-vs-all comparison to find the potential paralogous gene pairs using BLASTP (version 2.2.26) (Altschul et al., 1997) with an E-value threshold of $1e-5$. A total of 2,646 pairs were identified in SD regions. For each pair of SD regions harboring paralogous gene pairs, we calculated the Ks values of the gene pairs as proxies for the generation time of their corresponding SDs (Fig. S7). Our result revealed that most of the SDs are relatively recently produced.

The creation of SDs results in a large number of duplicated genes, which can accelerate genome evolution (Conant and Wolfe, 2008). Although most young duplicated genes tend to be functionally redundant and thus are susceptible to loss-of-function mutations that degrade one of the copies into a pseudogene, some duplicated

pairs retain their encoded function and continue to express in the long term (Lan and Pritchard, 2016). To determine the fates of these duplicated genes in the SD regions, we analyzed their expression profiles and classified these duplicates into three categories. A gene pair is defined as potentially sub- or neo-functionalized if one gene expresses at a significantly higher level than the other in at least one tested tissue. We refer to those pairs as asymmetrically expressed duplicates (AEDs) when the expression level of one of the pair is higher in at least two tissues and not lower in other tested tissues than its sister gene. The remaining duplicates are classified as having no difference (non-difference) (Lan and Pritchard, 2016).

From the 2,646 gene pairs, 150 were identified as sub-/neo-functional gene pairs, 1,505 as non-difference pairs, and 991 as AEDs (Fig. S8, S9, S10). Then, we analyzed the proportion of each expression pattern category across different Ks values. With increasing time, the proportions of sub-/neo-functional gene pairs and AEDs also gradually increase, while the proportion of no-difference pairs decreases. This finding suggests that, although it is unable to predict which copy of a duplicated gene pair would be preferentially expressed after SD creation, one of them is likely to be subsequently pseudogenized or sub-/neo-functionalized (Fig. S11). Analysis of the positions of duplicated gene pairs showed that more no-difference pairs and sub-/neo-functional pairs are in tandem SDs, and more AEDs are in interspersed SDs (Fig. 3C, 3D). One possible explanation for such a difference is that tandemly repeated genes in the genome may be co-regulated, which facilitates their co-evolution, whereas chromosomally distant duplicated genes may evolve independently.

Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses (FDR <0.05) (Kanehisa et al., 2014) showed that sub-/neo-functionalized genes are mainly enriched in metabolic pathways which may contribute to stress resistance, such as glycine, serine and threonine metabolism (map00260); cutin, suberine and wax biosynthesis (map00073); glutathione metabolism (map00480); anthocyanin biosynthesis (map00942); phenylalanine metabolism (map00360); tyrosine

metabolism (map00350); and brassinosteroid metabolism (map00905). Further analysis revealed that the genes encoding 9-*cis*-epoxycarotenoid dioxygenase, glutathione *S*-transferase, primary-amine oxidase, very-long-chain (3*R*)-3-hydroxyacyl-CoA dehydratase, and catalase are duplicated, suggesting that the abundance of functional genes related to the environmental adaptability of rice was expanded by the production of SDs (Fig. 3B, Table S15).

Figure 3. Identification of SDs and their evolutionary impact on the rice genome. (A) Segmental duplicated regions in Minghui 63. Only SDs blocks greater than 5000 bp are shown. (B) Scatterplot of KEGG enrichment for the sub-/neo-functionalized genes in Minghui 63. (C, D) The proportion of each category of duplicated gene pair expression patterns located within 1 Mb in the same chromosomes (C) and on different chromosomes (D) across different Ks values. The X-axis indicates Ks boundaries. The Y-axis is the proportion of duplicates.

221 **Active TEs promote rice evolution**

222 We identified a total of 157.16 Mb (39.70%) of TEs in our assembled genome, which
 223 was close to the percentage of TEs reported in other published rice genomes (Table
 224 S8, Table S9). The predominant TE type is LTR elements, which represents
 225 approximately 71.36% (112.13 Mb) of the total length of TEs. We analyzed the
 226 genomes of Minghui 63 (MH63KL1) and Nipponbare (IRGSP-1.0) and identified
 227 5,516 and 4,706 LTR-RTs, with a total length of 53.65 Mb and 43.50 Mb,
 228 respectively (Table S16). Based on the differences between the LTRs at the ends of
 229 each LTR-RT, we calculated the insertion time of the LTR-RTs in Minghui 63 and
 230 Nipponbare genomes. Both genomes were found to undergo a dramatic expansion in
 231 LTR-RTs approximately 1.5 million years ago (myr). However, within 0.5 myr,
 232 Minghui 63 genome began expanding at a significantly higher rate and the expansion
 233 of LTR-RTs may still be ongoing (Fig. 4A, 4B). We next classified LTR-RTs
 234 according to RT domains and found that the *Del* family in Minghui 63, in particular,
 235 exhibits significant expansion (Table S17). Analysis of the divergence between TEs
 236 and consensus sequences showed an increase in recent insertion events of TEs such as
 237 *En-Spm*, *MuDR*, *Tc1*, *Copia*, and *ERV* (Fig. S12). The above results indicate that TEs
 238 in Minghui 63 have a lineage-specific expansion, which may partially explain the
 239 larger genome of Minghui 63 compared with Nipponbare. In addition, these findings
 240 also suggest that TEs in Minghui 63 are relatively active.

241 TEs can also substantially contribute to shaping gene expression. TE insertions
 242 within the coding regions of a gene can result in loss of function or an aberrant
 243 protein. Insertions in introns can change the splicing pattern of transcripts and thus the
 244 function of genes. TEs can also affect gene expression by providing novel promoters
 245 or regulatory regions. The epigenetic profile of the TEs can alter gene expression as
 246 well (Hirsch and Springer, 2017). Given these possibilities, we analyzed the
 247 occurrence of TEs around AEDs. We found that, between the duplicated genes, there

are higher abundances of all the types of TEs near the copies with lower expression levels (Fig. 4C, 4D).

Figure 4. LTR insertions and their effect on gene expression.

(A, B) The insertion time of LTR-RTs in Nipponbare (A) and Minghui 63 (B). (C, D) Distributions of Copia, Gypsy, LTR, and DNA transposons across higher (C) and lower (D) expression genes and their flanking regions (± 5 kb).

SDs can enhance rice disease resistance traits

Plants have evolved a series of defensive mechanisms in the process of resisting pathogen infection, primarily through various disease resistance genes that recognize pathogens and block their invasion (Wan et al., 2012). For example, the nucleotide-

binding site with leucine-rich repeat (*NBS-LRR*) genes comprise one of the largest gene families in plants and play a central role in plant resistance to biotic stress (McHale et al., 2006).

Compared with *japonica* rice, *indica* rice has a higher resistance to biotic stresses such as rice blast (Wu et al., 2017). We identified the *NBS-LRR* genes in Minghui 63 and Nipponbare, and found that they are significantly expanded in Minghui 63 compared to Nipponbare (487 versus 349) (Fig. S13, Table S18). We then counted the distribution of *NBS-LRR* genes on each chromosome, and found that they were expanded significantly on chr11 in particular (Table S19). This finding suggests that chr11 may play a major role in disease resistance in rice.

We then identified the orthologous genes between Minghui 63 and Nipponbare and found that 246 *NBS-LRR* genes are shared by Minghui 63 and Nipponbare, 241 are specific to Minghui 63, and 103 are specific to Nipponbare (Fig. S14). We then analyzed the 241 Minghui 63 specific *NBS-LRR* genes, and found that 71 were located within collinear block pairs of SDs. Most of these SDs are tandem repeats except for a few that are located on different chromosomes (Table S20). These results indicate that the tandem SDs likely participate in the expansion of *NBS-LRR* genes and especially in driving evolutionary adaptations to resist pathogen infection in rice. Next, we analyzed the expression profiles of the paralogous genes in the 71 collinear blocks and found that most of the paralogous gene pairs either exhibit no difference in expression or express asymmetrically in different tissues (Fig. S15). This finding is similar to the expression trend we observed for all paralogous genes in SDs in Minghui 63. Furthermore, a large number of *NBS-LRR* genes located in SDs probably suggest a dose effect to successfully resist pathogen infection. A small number of *NBS-LRR* genes may undergo sub-/neo-functionalization and have been co-opted into performing new functions.

Rice blast disease, caused by *Magnaporthe oryzae* (Ascomycota), is one of the most severe and destructive diseases of rice (Valent and Khang, 2010), occurring in

almost all rice production regions and commonly reducing yield by 10%-30% (Hossain and Fischer, 1995). We identified the previously cloned and characterized rice blast resistance genes (http://www.ricedata.cn/gene/gene_pi.htm) in Minghui 63 and found that *Pit*, *Pbl*, *Piz*, and *Pib* (and their corresponding paralogous genes) located in four collinearity block pairs in SDs. These SDs are all tandem repeats locate on the same chromosome (Table S21). Further analysis revealed that the *Piz* gene had experienced multiple tandem repeat events, two of which occurred in the SD region (Fig. S16, Table S21). Altogether, these findings suggest that the tandem SDs promote the expansion of rice blast resistance genes, which consequently drive new resistance mechanisms against rice blast disease.

SDs drive evolutionary changes in rice growth and development

Our KEGG pathway analyses (FDR <0.05) of sub-/neo-functionalized genes revealed an enrichment in the zeatin biosynthesis pathway (map00908). Our detailed analysis showed that *cis*-zeatin-*O*-glucosyltransferases (*cZOGTs*) are enriched and sub-functionalized. *cZOGTs*, which catalyzes the *O*-glucosylation of *cis*-zeatin (cZ) and *cis*-zeatin-riboside (Fig. 5A), can result in strong physiological impacts on rice growth and development (Kudo et al., 2012). For example, transgenic rice lines ectopically overexpressing the *cZOGT* genes exhibited short-shoot traits, delayed leaf senescence, and decreased crown root number (Kudo et al., 2012).

We identified *cZOGTs* in Minghui 63 and Nipponbare and found four copies in both varieties, compared with only two copies in corn (Fig. 5B, C). This indicates that *cZOGTs* copy numbers were already increased prior to the divergence of the two varieties of rice. Further analysis revealed that these four genes are distributed on chr4 in two collinearity block pairs of tandem SDs (Table S22), suggesting that *cZOGTs* in rice were doubled by two independent tandem duplication events (Fig. 5B). To further investigate this possibility, we constructed ML trees for *cZOGTs* in Minghui 63 and

Nipponbare and found that each group of the tandemly repeated genes are clustered together, which also supports the occurrence of two independent tandem duplication events (Fig. 5C). The gene expression profiles of the *cZOGTs* at different developmental stages showed that two genes are highly expressed in both young and mature leaves, while the third gene is expressed at low levels in all tissues, suggesting its pseudogenization, and the fourth gene is highly expressed in roots and young leaves, suggesting its sub-/neo-functionalization (Fig. 5D). The generation of these genes through segmental duplication strongly suggests that SDs can serve as a source for new genes, some of which may contribute to adaptative or other desirable traits.

Figure 5. The duplication and diversification of *cZOGTs*.

(A) *cZOGTs* catalyze *O*-glucosylation of *cis*-zeatin and *cis*-zeatin-riboside. (B) Schematic depiction of *cZOGTs* in collinearity blocks in tandem SDs on chr4. Arrows indicate duplication events. (C) Phylogenetic analysis of *cZOGTs*. Red points indicate corn; blue points indicate Minghui 63; yellow points indicate Nipponbare. (D) Heat map of transcriptional expression for *cZOGTs* in Minghui 63.

Discussion

Repeat sequences historically represent the most challenging genomic regions to assemble, although they play important roles in the evolution of the genome. With the development of long-read sequencing technology, the continuity of genome assembly has been greatly improved. However, highly accurate, gapless genome assemblies have remained unattainable. In our study, we demonstrated the first strategy to obtain the telomere-to-telomere gapless rice genome, in which the chromosomal position information for contigs is introduced into a string graph. Du et al. (2017) also used chromosomal contig positions to assist assembly, but did not incorporate chromosomal order information of contigs, which is a vital step in our approach. Although their assembly contains only five gaps, there are 14 insertions and deletions (indels) longer than 10 kb, as well as 17 potential indel errors of 1-10 kb (Du et al., 2017). These indels may be introduced by repeat sequences, which cannot be solved by PacBio Continuous long reads (CLRs). Our telomere-to-telomere gapless rice genome is the first gapless genome in higher plants and animals to the best of our knowledge.

Segmental duplications can produce a large number of duplicated genes, thus serving as a source of new genes and subsequently driving rapid adaptive evolution. *Indica* rice has a larger genome than *japonica* rice, likely due to TEs and SDs. The relatively active TEs and newly duplicated genes in Minghui 63 may have an important impact on its agronomic traits. We systematically studied the SDs and TEs in Minghui 63 and identified a total of 89.68 Mb of SDs, accounting for 22.66% of the genome. We also found that SDs in Minghui 63 contain many duplicated genes, a large number of which exhibit no difference in expression, while a small number of which are sub-/neo-functionalized, suggesting that the former might affect plant traits through dose effect and the latter contribute through the gain of new functions.

KEGG enrichment analysis for sub-/neo-functionalized duplicated genes revealed that genes related to disease resistance have expanded, suggesting that the

duplicated genes produced by SDs can improve the environmental adaptability of rice (e.g., adaptation to biotic stress or pathogen pressure). TEs can also play an influential role in modulating gene expression. Our analyses showed that TEs are markedly amplified in Minghui 63, and notably, occurred at a significantly higher frequency in the proximity of the genes with relatively lower expression levels in AEDs. We propose that active TEs and SDs work synergistically to promote rice evolution.

Compared with *japonica* rice, *indica* rice has a higher resistance to biotic stresses such as rice blast infection. We found that *NBS-LRR* genes in Minghui 63 have been considerably amplified through duplication, and many of these genes have undergone specialization for pathogen resistance. Notably, 71 *NBS-LRR* genes are located in SDs and most of them are tandem repeat genes. We analyzed the genes related to rice blast resistance and found that four blast resistance genes are located within tandem SDs, suggesting that tandem SDs might drive specific disease resistance adaptations in Minghui 63. In addition to pathogen resistance, we investigated *cZOGTs*, which catalyze *O*-glucosylation of *cis*-zeatin and *cis*-zeatin-riboside, and play important roles in biomass accumulation during the late stage of rice growth and development. Four copies of *cZOGT* genes in rice located on chr4 are found to be derived from two independent tandem repeat events, suggesting that SDs could be potentially engineered through closer study to drive advances in specific desirable agronomic traits.

In our MH63KL1 genome, there are still several overlaps between different chromosomes in the assembly graph. In order to resolve these overlaps, technologies that enable longer sequencing reads but also maintain high fidelity are needed. The development of high-fidelity sequencing technology and the improvement of the genome assembly quality will afford researchers the opportunities to accurately study TEs, SDs, structural variations, and the evolution and expression regulation of duplicated genes.

388 **Materials and methods**

389 **Genome assembly with PacBio HiFi reads**

390 About 25.3 Gb of Minghui 63 HiFi reads from PacBio Sequel II were downloaded
 391 from the NCBI Sequence Read Archive database (SRX6957825). *De novo* assembly
 392 of these long reads was performed using pb-assembly
 393 (<https://github.com/PacificBiosciences/pb-assembly>) (Wenger et al., 2019) with
 394 parameters ‘-k21 -h850 -e.99 -l2000 -s100 --max-diff 400 --max-cov 400 --min-cov 2
 395 --n-core 24 --min-idt 99.7 --ignore-indels’.

396

397 **Gapless assembly**

398 Here we used a new assembly pipeline to obtain a gapless rice genome. The pipeline
 399 consists of three steps: (1) anchoring contigs onto chromosomes; (2) linking contig
 400 paths to chromosome paths; (3) finding an unitig for each gap in the string graph.

401 (1) anchoring contigs onto chromosomes

402 Here we proposed a method of reference-based contig anchoring. NUCmer (Marcais
 403 et al., 2018) or MCSScan (Tang et al., 2008) was used to map contigs to chromosomes
 404 of closely related species (Fig. S3). The positions and orientation of blocks mapped
 405 on chromosomes were retrieved. If a contig was matched to multiple chromosomes,
 406 we assigned the contig to the chromosome with the largest number of matching
 407 blocks. Then contigs were ordered and oriented on each chromosome based on the
 408 positions and orientation of their longest blocks on the chromosomes. Then adjacent
 409 contigs were linked by 100 Ns. Of course, genetic maps (Catchen et al., 2020), HIC
 410 (Dudchenko et al., 2017), or other biological methods can also be used to anchor
 411 contigs. Here, the requirement of anchoring accuracy is relatively high and some short
 412 contigs can be removed to improve accuracy.

413 (2) linking contig paths to chromosome paths

First, we extracted the position and orientation of contigs on each chromosome according to the AGP file generated in the previous step to link contig paths generated by pb-assembly (ctg_paths, encoding the graph for each contig and represented by a list of unitigs). Then every chromosome was represented by a list of contig paths and gaps between them. If two contigs were adjacent, they were connected without gaps.

(3) finding an unitig for each gap in the string graph

We reconstructed the string graph using edges (sg_edges_list) and overlap information (pread.m4) generated by pb-assembly. For each gap, we used the starting point of the gap as the center to construct an ego graph and found the paths from the starting point to the ending point of the gap in the ego graph. If there were multiple paths for a gap, we selected the shortest path with the most overlaps. Because there might be assembly errors in the upstream and downstream of gaps, we deleted unitigs near one gap and found the shortest path again if the shortest path for the gap was not found. Finally, if the shortest path was still not found, we kept 100Ns in the corresponding position. The whole assembly graph was shown by Bandage (Wick et al., 2015) (Fig. 1D).

Assessment of assembly quality

To evaluate the completeness of the assembly and the uniformity of the sequencing, the clean sequencing reads (SRR3234369) of Minghui 63 from short-insert size library were mapped to our assembly using BWA (version 0.7.8) (Li and Durbin, 2009). The mapping rate was 99.96% and the genome coverage was 98.25%.

Benchmarking Universal Single-Copy Orthologs (BUSCO, 4.0.5) (Simao et al., 2015) was used to assess the completeness of gene regions of our assembly using the dataset embryophyta_odb10 (2019-11-20). BUSCO consists of single-copy orthologs derived from OrthoDB in each evolutionary lineage. Of the 1614 single-copy orthologs presented in the embryophytes, 98.6% was complete in MH63KL1, better

than the extent in the Nipponbare reference genome (IRGSP-1.0, 98.3%),
MH63RS1(93.1%), and MH63RS2(98.0%) (Fig. S4).

In addition to BUSCO, we mapped 2,045 full-length cDNA sequences from
Oryza rufipogon W1943 (<http://server.ncgr.ac.cn/ricd/dym/ftp.php>) to our assembly
using BLAT (Kent, 2002). 1,971 (96.38%) of these could be mapped with greater
than 50% coverage.

LTR Assembly Index (LAI) uses LTR-RTs to evaluate assembly continuity. The
LAI were calculated using LTR_retriever with default parameters (Ou et al., 2018).

Finally, 12 BAC sequences of centromere regions obtained from the NCBI
GenBank and 24 gap region sequences were aligned to our assembly using
LASTZ (<http://www.bx.psu.edu/~rsharris/lastz/>). The detailed information of sequence
alignments was shown in Fig. S1, Fig. S6, Table S3, and Table S7.

Collinearity analysis

MCSan (Tang et al., 2008) was used to identify the collinearity of genes between
MH63KL1 and MH63RS2 with parameters (--cscore=0.99) (Fig. S3B). NUCmer was
used to align MH63KL1 and MH63RS2 with parameters (--mum, --mincluster=200, -
-minmatch=100). Delta-filter was then used to get the 1-to-1 alignments with
parameters (-i 95, -l 100, -1). The dot plot was showed using mummerplot (Fig. S3A)
(Marcais et al., 2018).

Repeats annotation

Transposable elements of Minghui 63 genome were annotated through a combination
of *ab initio* prediction and homology searching. We first used LTR_FINDER (Xu and
Wang, 2007), RepeatScout (Price et al., 2005), and RepeatModeler to build a
consensus sequence library. Repeatmasker

(<http://www.repeatmasker.org/RepeatModeler.html>) was used to annotate the repeat regions based on this library as well as the known Rice TE library (RiTE database, <https://www.genome.arizona.edu/cgi-bin/rite/index.cgi>) (Copetti et al., 2015). RepeatProteinMask was then used to identify repeat sequences at the protein level. Tandem Repeats Finder (TRF) (Benson, 1999) was used to find Tandem repeats. The divergences of TEs were retrieved from the GFF file produced by RepeatMasker.

Gene annotation

Protein-coding genes were conducted using a combination of homology-based prediction, *de novo* prediction, and transcriptome-based prediction methods. Proteins sequences from seven plant genomes (*Oryza sativa* L. *japonica*, *Oryza sativa* L. *indica*, *Brachypodium distachyon*, *Setaria italica*, *Zea mays*, *Arabidopsis thaliana*, and *Sorghum bicolor*) were downloaded from Ensemble (release-47). They were aligned to the Minghui 63 genome using the TBLASTN program (Altschul et al., 1990) with an E-value cutoff of 1e-5. The BLAST hits were conjoined using Solar software (Yu et al., 2006). GeneWise (Birney et al., 2004) was then utilized to predict the exact gene structure of the corresponding genomic regions for each BLAST hit. FL-cDNAs and ESTs (RICD, <http://www.ncgr.ac.cn/ricd>) of *Oryza sativa* L. *indica* and FL-cDNAs of *Oryza sativa* L. *japonica* (Rice Full-Length c et al., 2003) were directly mapped to the Minghui 63 genome and assembled by PASA (Campbell et al., 2006). Gene models created by PASA were used to train the *ab initio* gene prediction programs. Augustus (version 2.5.5) (Urnov et al., 2005), Genscan (version 1.0) (Burge and Karlin, 1997), GlimmerHMM (version 3.0.1) (Majoros et al., 2004), Geneid (Blanco et al., 2007), and SNAP (Korf, 2004) were used for *ab initio* gene predictions in the repeat-masked genome. RNA-Seq reads downloaded from the NCBI Sequence Read Archive database (SRR10751892-SRR10751899) were mapped to the assembly using Tophat (version 2.0.8) (Trapnell et al., 2009) with default parameters. Cufflinks (version 2.1.1) (Trapnell et al., 2010) was then used to assemble the transcripts into

gene models. All evidence of gene models were combined by Evidence Modeler (EVM) (Haas et al., 2008) into a non-redundant set of gene structures. UTR regions were then identified by integrating transcript evidence using PASA. Finally, gene models were filtered by removing the genes having 20% of their CDS sharing overlaps with TEs and coding region lengths less than 150 bp. The function of protein-coding genes was achieved by searching their protein sequences against the SwissProt (<http://web.expasy.org/docs/swiss-prot/guideline.html>) and NR databases using BLASTP (Altschul et al., 1997) (E-value 1e-05). Protein domains were identified by searching against the InterPro (<http://www.ebi.ac.uk/interpro/>, V32.0) and Pfam databases (<http://pfam.xfam.org/>, V27.0), using InterProScan (V4.8) and HMMER (Finn et al., 2011), respectively. The Gene Ontology (GO, <http://www.geneontology.org/page/go-database>) terms for each gene was obtained from the corresponding InterPro or Pfam entry. KEGG metabolic pathways were further assigned by BLAST (E-value 1e-05) against the KEGG databases (release 53) (Kanehisa et al., 2014).

Centromeres identification

HMMER (version 3.1b1) (Finn et al., 2011) was used to identify centromeres. Firstly, we built a hmm file through hmmbuild using a file of 155-165 bp CentO satellite DNA sequences in rice. Hmmssearch was then used to search the location of centromeres of our assembly.

SDs detection

Briefly, our genome assembly was first soft-masked converting all common and tandem repeats to lower-case letters. SEDEF was then used to detect SDs with the default parameters (Numanagic et al., 2018).

521

522 **Ks distribution of the duplicated gene pairs**

523 The protein sequences of the duplicated pairs in SDs were retrieved, and aligned using
524 the MUSCLE program (Edgar, 2004). Subsequently, the protein sequences alignments
525 were converted into CDS alignments based on the coding sequences. We then
526 calculated the numbers of synonymous substitutions per synonymous site (Ks) for
527 these gene pairs based on the MYN method implemented in the PAML program
528 (version 4.9b) (Yang, 2007). The Ks distribution was plotted and displayed using R
529 (version 3.0.1).

530

531 **Classification of expression patterns of sister duplicates**

532 RNA-Seq reads of roots, young leaves, mature leaves, and panicles from the NCBI
533 Sequence Read Archive database were mapped to MH63KL1 using TopHat (version
534 2.0.8) (Trapnell et al., 2009). HTSeq-count was used to calculate the numbers of
535 aligned reads (read counts) for each gene (Anders et al., 2015). Differential gene
536 expression (DEG) analyses among the duplicated gene pairs for each tissue were
537 performed using DESeq2 with FDR cut-off of 0.05 and log2 fold change (FC) cut-off
538 of 1 (Love et al., 2014). We divided the duplicated pairs into three classes according
539 to their expression in four tissues: 1) Sub- or neo-functionalized pairs in which each
540 of the two duplicates was significantly more highly expressed than the other in at least
541 one tissue; 2) Asymmetrically Expressed Duplicates (AEDs) in which one duplicate
542 was expressed higher than its sister in at least two tissues and its expression was not
543 lower than its partner in other tested tissues; 3) The remaining duplicates were
544 classified as no difference pairs (Lan and Pritchard, 2016).

545

546 **KEGG enrichment analysis**

KOBAS software was used to test the statistical enrichment of differential expression genes in KEGG Pathways with the cutoff set at $P < 0.05$ (Xie et al., 2011).

Identification and classification of LTR retrotransposons

We performed *de novo* searches for LTR retrotransposons (LTR-RTs) against our assembly using LTRharvest (Ellinghaus et al., 2008) and LTRfinder. Profile HMM models for domain detection were downloaded from the Gypsy Database (GyDB) (Llorens et al., 2011). LTRdigest (Steinbiss et al., 2009) was used for structure annotation (e.g., PBS, PPT, protein) of LTR-RTs (Liu et al., 2020b; Zhang et al., 2020). The RT domains were used to classify LTR-RT. The nucleotide substitution (λ) between the 5' and 3' terminals of intact LTR-RTs were estimated by MUSCLE (Edgar, 2004). The genetic distances (K) were calculated by $K = -0.75 \ln(1 - 4\lambda/3)$. Finally, the insertion time of LTR-RTs was calculated based on the following formula: $T = K/2r$ ($r = 1.3 \times 10^{-8}$ per site and per year) (Ma and Bennetzen, 2004).

Identification of NBS-LRR genes

For each species, the protein sequences were retrieved and searched against the raw Hidden Markov Model (HMM) of NB-ARC family (PF00931) using HMMER (version 3.1b1) (Finn et al., 2011) with default parameters. The high-quality NBS protein sets ($< 1E-60$) from these initial searches were then self-aligned using mafft (version 7.305) (Kato and Standley, 2013). The NBS HMMs were then construct using the program of hmmbuild implemented in HMMER (Finn et al., 2011) and were used to search NBS-encoding proteins (Xia et al., 2017). Finally, we identified 487 and 349 NBS-encoding proteins in Minghui 63 and Nipponbare, respectively. PfamScan was used to search these proteins against Pfam-A database (Pfam31.0) (Punta et al., 2012) to identify domains. CC domains were identified using ncoils (Lupas et al., 1991) with the default parameters.

574

575 **Identification and evolutionary analysis of rice blast resistance genes and *cis*-**
576 **zeatin-*O*-glucosyltransferases (cZOGTs) genes**

577 Based on the previously cloned amino acid sequences(<http://www.ricedata.cn/gene/>),
578 we identified rice blast resistance and *cZOGTs* genes (Kudo et al., 2012) using the
579 BLASTP program (version 2.2.26) (Altschul et al., 1997) with an E-value of $\leq 1E-10$.
580 Only those hits with $\geq 30\%$ identity and $\geq 50\%$ coverage of both query and target
581 were retained. The conserved domains were further examined in the retained
582 sequences. The protein sequences for each gene were individually retrieved and
583 aligned using MUSCLE (version 3.8.31) (Edgar, 2004). Maximum-Likelihood (ML)
584 trees were constructed using RAxML(version 8.0.19) (Stamatakis, 2006) with the
585 PROTGAMMAAUTO model and 100 bootstrap replicates for each alignment. We
586 further checked each evolutionary tree to further filter the identified genes. RNA-seq
587 reads were mapped to our assembly using TopHat (version 2.0.8) (Trapnell et al.,
588 2009). Cuffdiff(version 2.2.1) (Trapnell et al., 2013) was used to compute the FPKM
589 (Fragments Per Kilobase Of Exon Per Million Fragments Mapped) values for each
590 gene. We visualized the gene expression levels using pheatmap ([https://cran.r-](https://cran.r-project.org/web/packages/pheatmap/index.html)
591 [project.org/web/packages/pheatmap/index.html](https://cran.r-project.org/web/packages/pheatmap/index.html)) for the identified genes.

592

593 **Availability of data and materials**

594 The genome assemblies have been deposited in GenBank under accession number
595 CP059874-CP059885. The code used in this study is available at the GitHub
596 repository <https://github.com/likui345/PGA>.

597

598 **Funding**

599 This study was supported by the National Natural Science Foundation of China
600 (31770331, 31970318).

601

602 **Author Contributions**

603 KL, PFL and SL designed the project. KL, WJ, YH, and MK worked on data analyses
604 and developed the gapless assembly pipeline. KL, LZG, PFL and SL wrote the
605 manuscript. All authors read and approved the final manuscript.

606

607 **Acknowledgments**

608 We thank Evan E. Eichler (Univ. Washington) for his help with SD analysis.

609

610 **Declaration of interests**

611 The authors declare that they have no competing interests.

Figure Legends

Figure 1. Schematic depiction of the assembly process.

(A) Whole-genome assembly from High-fidelity (HiFi) reads. (B) Anchoring WGS contigs onto chromosomes. (C) String graphs were reconstructed using edges (sg_edges_list) and overlap information (pread.m4) generated by pb-assembly. Contigs were linked following their order on each chromosome and a unitig (typically a repeat sequence) for each gap between contigs can be found in the string graph. Red blocks indicate gaps; solid lines indicate contig paths; dashed lines indicate gap locations. (D) The final assembly graph. Each color represents one chromosome.

Figure 2. Distribution of repeat sequences and gap positions in MH63KL1.

The overlaid heatmap shows the TE density in a 1-Mb window. The tack labels refer to segmental duplications (SDs, green boxes) and the original gap locations (orange triangles).

Figure 3. Identification of SDs and their evolutionary impact on the rice genome.

(A) Segmental duplicated regions in Minghui 63. Only SDs blocks greater than 5000 bp are shown. (B) Scatterplot of KEGG enrichment for the sub-/neo-functionalized genes in Minghui 63. (C, D) The proportion of each category of duplicated gene pair expression patterns located within 1 Mb in the same chromosomes (C) and on different chromosomes (D) across different Ks values. The X-axis indicates Ks boundaries. The Y-axis is the proportion of duplicates.

Figure 4. LTR insertions and their effect on gene expression.

(A, B) The insertion time of LTR-RTs in Nipponbare (A) and Minghui 63 (B). (C, D) Distributions of Copia, Gypsy, LTR, and DNA transposons across higher (C) and lower (D) expression genes and their flanking regions (± 5 kb).

Figure 5. The duplication and diversification of *cZOGTs*.

(A) *cZOGTs* catalyze *O*-glucosylation of *cis*-zeatin and *cis*-zeatin-riboside. (B) Schematic depiction *cZOGTs* in collinearity blocks in tandem SDs on chr4. Arrows indicate duplication events. (C) Phylogenetic analysis of *cZOGTs*. Red points indicate corn; blue points indicate Minghui 63; yellow points indicate Nipponbare. (D) Heat map of transcriptional expression for *cZOGTs* in Minghui 63.

647 **References**

- 648 Alkan, C., Sajjadian, S., and Eichler, E.E. (2011). Limitations of next-generation
649 genome sequence assembly. *Nat Methods* 8:61-65.
- 650 Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic
651 local alignment search tool. *J Mol Biol* 215:403-410.
- 652 Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and
653 Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of
654 protein database search programs. *Nucleic Acids Res* 25:3389-3402.
- 655 Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work
656 with high-throughput sequencing data. *Bioinformatics* 31:166-169.
- 657 Bailey, J.A., Yavor, A.M., Massa, H., Trask, B.J., and Eichler, E.E. (2001).
658 Segmental duplications: organization and impact within the current human
659 genome project assembly. *Genome Res* 11:1005-1017.
- 660 Bairoch, A., and Apweiler, R. (1999). The SWISS-PROT protein sequence data bank
661 and its supplement TrEMBL in 1999. *Nucleic Acids Res* 27:49-54.
- 662 Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences.
663 *Nucleic Acids Res* 27:573-580.
- 664 Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. *Genome*
665 *Res* 14:988-995.
- 666 Blanco, E., Parra, G., and Guigó, R. (2007). Using geneid to identify genes. *Curr*
667 *Protoc Bioinformatics* Chapter 4: Unit 4.3.
- 668 Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human
669 genomic DNA. *J Mol Biol* 268:78-94.
- 670 Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M., and Buell, C.R. (2006).
671 Comprehensive analysis of alternative splicing in rice and comparative
672 analyses with Arabidopsis. *BMC Genomics* 7:327.
- 673 Catchen, J., Amores, A., and Bassham, S. (2020). Chromonomer: A tool set for
674 repairing and enhancing assembled genomes through integration of genetic
675 maps and conserved synteny. *G3 (Bethesda)* 10:4115-4128.
- 676 Cheng, Z.K., Dong, F.G., Langdon, T., Ouyang, S., Buell, C.R., Gu, M., Blattner,
677 F.R., and Jiang, J. (2002). functional rice centromeres are marked by a satellite
678 repeat and a centromere-specific retrotransposon. *Plant Cell* 14:1691-1704.
- 679 Conant, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job: How duplicated
680 genes find new functions. *Nature Rev Genet* 9:938-950.
- 681 Copetti, D., Zhang, J., Baidouri, M.E., Gao, D., Wang, J., Barghini, E., Cossu, R.M.,
682 Angelova, A., Maldonado, E.L.C., and Roffler, S. (2015). RiTE database: A
683 resource database for genus-wide rice genomics and evolutionary biology.
684 *BMC Genomics* 16:538-538.
- 685 Du, H., Yu, Y., Ma, Y., Gao, Q., Cao, Y., Chen, Z., Ma, B., Qi, M., Li, Y., Zhao, X.,
686 et al. (2017). Sequencing and de novo assembly of a near complete indica rice
687 genome. *Nat Commun* 8:15324.

- Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., and Aiden, A.P. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* 356:92-95.
- Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792-1797.
- Elert, E. (2014). Rice by the numbers: A good grain. *Nature* 514:S50-51.
- Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* 9:18.
- Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 39:W29-37.
- Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EvidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* 9:1-22.
- Han, M.V., Demuth, J.P., McGrath, C.L., Casola, C., and Hahn, M.W. (2009). Adaptive evolution of young gene duplicates in mammals. *Genome Res* 19:859-867.
- Hirsch, C.D., and Springer, N.M. (2017). Transposable element influences on gene expression in plants. *Biochim Biophys Acta* 1860:157-165.
- Hossain, M., and Fischer, K.S. (1995). Rice research for food security and sustainable agricultural development in Asia: Achievements and future challenges. *GeoJournal* 35:286-298.
- Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., et al. (2012). InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res* 40:D306-312.
- International Rice Genome Sequencing, P. (2005). The map-based sequence of the rice genome. *Nature* 436:793-800.
- Jiao, Y., Peluso, P., Shi, J., Liang, T.Y., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X., and Chin, C. (2017). Improved maize reference genome with single-molecule technologies. *Nature* 546:524-527.
- Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 42:D199-205.
- Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30:772-780.
- Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res* 12:656-664.
- Korf, I.F. (2004). Gene finding in novel genomes. *BMC Bioinformatics* 5:59-59.
- Kudo, T., Makita, N., Kojima, M., Tokunaga, H., and Sakakibara, H. (2012). Cytokinin activity of cis-zeatin and phenotypic alterations induced by

731 overexpression of putative cis-zeatin-*O*-glucosyltransferase in rice. *Plant*
732 *Physiol* 160:319-331.

733 Lan, X., and Pritchard, J.K. (2016). Coregulation of tandem duplicate genes slows
734 evolution of subfunctionalization in mammals. *Science* 352:1009-1013.

735 Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-
736 Wheeler transform. *Bioinformatics* 25:1754-1760.

737 Liu, J., Seetharam, A.S., Chougule, K., Ou, S., Swentowsky, K.W., Gent, J.I., Llaca,
738 V., Woodhouse, M.R., Manchanda, N., Presting, G.G., et al. (2020a). Gapless
739 assembly of maize chromosomes using long-read technologies. *Genome Biol*
740 21:121.

741 Liu, J., Shi, C., Shi, C.C., Li, W., Zhang, Q.J., Zhang, Y., Li, K., Lu, H.F., Shi, C.,
742 Zhu, S.T., et al. (2020b). the chromosome-based rubber tree genome provides
743 new insights into spurge genome evolution and rubber biosynthesis. *Mol Plant*
744 13:336-350.

745 Llorens, C., Futami, R., Covelli, L., Dominguez-Escriba, L., Viu, J.M., Tamarit, D.,
746 Aguilar-Rodriguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G.P., et al.
747 (2011). The Gypsy Database (GyDB) of mobile genetic elements: release 2.0.
748 *Nucleic Acids Res* 39:D70-74.

749 Logsdon, G.A., Vollger, M.R., and Eichler, E.E. (2020). Long-read human genome
750 sequencing and its applications. *Nat Rev Genet* 21:597-614.

751 Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change
752 and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15:550.

753 Lupas, A., Van Dyke, M., and Stock, J. (1991). Predicting coiled coils from protein
754 sequences. *Science* 252:1162-1164.

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

757 Mackinnon, R.N., Wall, M., Zordan, A., Nutalapati, S., Mercer, B., Peverall, J., and
758 Campbell, L.J. (2013). Genome organization and the role of centromeres in
759 evolution of the erythroleukaemia cell line HEL. *Evolution, Medicine, and*
760 *Public Health* 2013:225-240.

761 Majoros, W.H., Pertea, M., and Salzberg, S.L. (2004). TigrScan and GlimmerHMM:
762 two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20:2878-
763 2879.

764 Marcais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A.
765 (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS*
766 *Comput Biol* 14:e1005944.

767 Marques-Bonet, T., Kidd, J.M., Ventura, M., Graves, T.A., Cheng, Z., Hillier, L.W.,
768 Jiang, Z., Baker, C., Malfavon-Borja, R., Fulton, L.A., et al. (2009). A burst of
769 segmental duplications in the genome of the African great ape ancestor.
770 *Nature* 457:877-881.

771 McHale, L., Tan, X., Koehl, P., and Michelmore, R.W. (2006). Plant NBS-LRR
772 proteins: adaptable guards. *Genome Biol* 7:212.

773 Michael, T.P., and VanBuren, R. (2015). Progress, challenges and the future of crop
774 genomes. *Curr Opin Plant Biol* 24:71-81.

775 Mostovoy, Y., Levysakin, M., Lam, J., Lam, E.T., Hastie, A.R., Marks, P., Lee, J.V.,
776 Chu, C.J., Lin, C., and Džakula, Ž. (2016). A hybrid approach for de novo
777 human genome sequence assembly and phasing. *Nat Methods* 13:587-590.

778 Numanagic, I., Gokkaya, A.S., Zhang, L., Berger, B., Alkan, C., and Hach, F. (2018).
779 Fast characterization of segmental duplications in genome assemblies.
780 *Bioinformatics* 34:i706-i714.

781 Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the
782 LTR Assembly Index (LAI). *Nucleic Acids Res* 46:e126.

783 Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat
784 families in large genomes. *Bioinformatics* 21:351-358.

785 Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N.,
786 Forslund, K., Ceric, G., Clements, J., et al. (2012). The Pfam protein families
787 database. *Nucleic Acids Res* 40:D290-301.

788 Raymond, O., Gouzy, J., Just, J., Badouin, H., Verdenaud, M., Lemainque, A.,
789 Vergne, P., Moja, S., Choisine, N., and Pont, C. (2018). The *Rosa* genome
790 provides new insights into the domestication of modern roses. *Nat Genet*
791 50:772-777.

792 Rice Full-Length c, D.N.A.C., National Institute of Agrobiological Sciences Rice
793 Full-Length c, D.N.A.P.T., Kikuchi, S., Satoh, K., Nagata, T., Kawagashira,
794 N., Doi, K., Kishimoto, N., Yazaki, J., Ishikawa, M., et al. (2003). Collection,
795 mapping, and annotation of over 28,000 cDNA clones from japonica rice.
796 *Science* 301:376-379.

797 Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M.
798 (2015). BUSCO: assessing genome assembly and annotation completeness
799 with single-copy orthologs. *Bioinformatics* 31:3210-3212.

800 Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic
801 analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688-
802 2690.

803 Steinbiss, S., Willhoeft, U., Gremme, G., and Kurtz, S. (2009). Fine-grained
804 annotation and classification of de novo predicted LTR retrotransposons.
805 *Nucleic Acids Res* 37:7002-7013.

806 Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008).
807 Synteny and collinearity in plant genomes. *Science* 320:486-488.

808 Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L.
809 (2013). Differential analysis of gene regulation at transcript resolution with
810 RNA-seq. *Nature Biotechnol* 31:46-53.

811 Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice
812 junctions with RNA-Seq. *Bioinformatics* 25:1105-1111.

813 Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J.,
814 Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and

815 quantification by RNA-Seq reveals unannotated transcripts and isoform
816 switching during cell differentiation. *Nat Biotechnol* 28:511-515.

817 Urnov, F.D., Miller, J.C., Lee, Y., Beausejour, C., Rock, J.M., Augustus, S.,
818 Jamieson, A., Porteus, M.H., Gregory, P.D., and Holmes, M.C. (2005). Highly
819 efficient endogenous human gene correction using designed zinc-finger
820 nucleases. *Nature* 435:646-651.

821 Valent, B., and Khang, C.H. (2010). Recent advances in rice blast effector research.
822 *Curr Opin Plant Biol* 13:434-441.

823 Vaser, R., Sovic, I., Nagarajan, N., and Sikic, M. (2017). Fast and accurate de novo
824 genome assembly from long uncorrected reads. *Genome Res* 27:737-746.

825 Wan, H., Yuan, W., Ye, Q., Wang, R., Ruan, M., Li, Z., Zhou, G., Yao, Z., Zhao, J.,
826 Liu, S., et al. (2012). Analysis of TIR- and non-TIR-NBS-LRR disease
827 resistance gene analogous in pepper: characterization, genetic variation,
828 functional divergence and expression patterns. *BMC Genomics* 13:502.

829 Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T.,
830 Ebler, J., Functammasan, A., Kolesnikov, A., Olson, N.D., et al. (2019).
831 Accurate circular consensus long-read sequencing improves variant detection
832 and assembly of a human genome. *Nat Biotechnol* 37:1155-1162.

833 Wick, R.R., Schultz, M.B., Zobel, J., and Holt, K.E. (2015). Bandage: interactive
834 visualization of de novo genome assemblies. *Bioinformatics* 31:3350-3352.

835 Wu, Y., Yu, L., Xiao, N., Dai, Z., Li, Y., Pan, C., Zhang, X., Liu, G., and Li, A.
836 (2017). Characterization and evaluation of rice blast resistance of Chinese
837 indica hybrid rice parental lines. *Crop J* 5:509-517.

838 Xia, E.H., Zhang, H.B., Sheng, J., Li, K., Zhang, Q.J., Kim, C., Zhang, Y., Liu, Y.,
839 Zhu, T., Li, W., et al. (2017). The Tea Tree Genome Provides Insights into
840 Tea Flavor and Independent Evolution of Caffeine Biosynthesis. *Mol Plant*
841 10:866-877.

842 Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., Kong, L., Gao, G., Li, C.Y.,
843 and Wei, L. (2011). KOBAS 2.0: a web server for annotation and
844 identification of enriched pathways and diseases. *Nucleic Acids Res* 39:W316-
845 322.

846 Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of
847 full-length LTR retrotransposons. *Nucleic Acids Res* 35:W265-268.

848 Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol*
849 *Evol* 24:1586-1591.

850 Yeo, S., Coombe, L., Warren, R.L., Chu, J., and Birol, I. (2018). ARCS: scaffolding
851 genome drafts with linked reads. *Bioinformatics* 34:725-731.

852 Yu, X., Zheng, H., Wang, J., Wang, W., and Su, B. (2006). Detecting lineage-specific
853 adaptive evolution of brain-expressed genes in human using rhesus macaque
854 as outgroup. *Genomics* 88:745-751.

855 Zhang, J., Chen, L., Xing, F., Kudrna, D.A., Yao, W., Copetti, D., Mu, T., Li, W.,
856 Song, J.M., and Xie, W. (2016). Extensive sequence divergence between the

857 reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui
858 63. Proc Natl Acad Sci USA 113:201611012.
859 Zhang, J., Rosenberg, H.F., and Nei, M. (1998). Positive Darwinian selection after
860 gene duplication in primate ribonuclease genes. Proc Natl Acad Sci USA
861 95:3708-3713.
862 Zhang, Q.J., Li, W., Li, K., Nan, H., Shi, C., Zhang, Y., Dai, Z.Y., Lin, Y.L., Yang,
863 X.L., Tong, Y., et al. (2020). The chromosome-level reference genome of tea
864 tree unveils recent bursts of non-autonomous ltr retrotransposons in driving
865 genome size evolution. Mol Plant 13:935-938.

SUPPLEMENTARY FIGURES

Supplemental Figure 1. Alignments between 18 gaps of Minghui 63 and MH63RS2.

The red line indicates each gap sequence and the blue line represents our assembly. The blocks in white indicate the unfilled gaps. The green blocks show that gap sequences are aligned to the reverse complementary sequence of our assembly.

Supplemental Figure 2. Assembly graphs around 24 gaps.

The gaps were marked in purple.

Supplemental Figure 3. The synteny between Minghui 63 and MH63RS2 across the whole genome.

(A) Dot plots of Minghui 63 versus MH63RS2 detected by NUCmer. (B) Syntenic regions between Minghui 63 and MH63RS2 detected by MCSanx.

A

B

Supplemental Figure 4. Assessment of the gene coverage rate using BUSCO.

Supplemental Figure 5. Comparison of LAI scores among four genome assemblies.

Supplemental Figure 6. Assessment of the Minghui 63 genome assembly by 12 BACs.

The red line indicates each BAC sequence and the blue line represents our assembly. The blocks in white indicate the unfilled gaps. The green blocks show that BAC sequences are aligned to the reverse complementary sequence of our assembly.

Supplemental Figure 7. Numbers of different type of duplicate gene pairs across different values of Ks.

Supplemental Figure 8. Heat map of expression ratios for AEDs.

For each duplicate pair, the ratios show the tissue-specific expression level of each gene relative to its duplicate.

Supplemental Figure 9. Heat map of expression ratios for no significant difference duplicate pairs. For each duplicate pair, the ratios show the tissue-specific expression level of each gene relative to its duplicate.

Supplemental Figure 10. Heat map of expression ratios for subfunctionalization duplicate pairs.

For each duplicate pair, the ratios show the tissue-specific expression level of each gene relative to its duplicate.

Supplemental Figure 11. Expression patterns of all the duplicate gene pairs in SDs across different values of Ks.

Supplemental Figure 12. Divergence distribution of classified families of TEs around AEDs.

Red lines indicate divergence distribution of TEs around lower expression genes. Green lines indicate divergence distribution of TEs around higher expression genes.

Supplemental Figure 13. Distribution of *NBS-LRR* genes.

A. Distribution of *NBS-LRR* genes in Minghui 63. B. Distribution of *NBS-LRR* genes in Nipponbare.

Supplemental Figure 14. The collinearity of *NBS-LRR* genes in genomes of Minghui63 (MH63KL1) and Nipponbare (IRGSP-1) rice.

Supplemental Figure 15. Heatmap of expression ratios for Minghui 63-specific *NBS-LRR* gene pairs.

For each duplicate pair, the ratios show the tissue-specific expression level of each gene relative to its duplicate.

Supplemental Figure 16. Phylogenetic analysis of rice blast-resistance *Piz* genes.

The maximum-likelihood phylogenetic trees were constructed using RAxML (version 8.0.19).

SUPPLEMENTARY TABLES

Supplemental Table 1. Statistic of Minghui 63 genome assembly.

Sample ID	Initial assembly		Final assembly	
	Contig(bp)	Number	Contig(bp)	Number
Total	400,454,019	93	395,822,112	12
Max	33,118,852	-	44,998,099	-
Number ≥ 100	-	93	-	12
Number ≥ 2000	-	93	-	12
N50	18,188,886	8	31,823,770	6
N60	15,442,314	11	31,239,580	7
N70	12,815,998	13	30,883,769	8
N80	9,777,320	17	30,306,463	9
N90	4,527,663	23	25,682,544	11

Supplemental Table 2. Statistic of Minghui 63 genome chromosome anchoring.

Chromosome	Anchored contigs	Length(bp)
chr1	4	45,070,763
chr2	7	36,788,985
chr3	2	39,838,936
chr4	2	37,246,311
chr5	4	33,005,269
chr6	3	31,884,747
chr7	1	30,837,837
chr8	4	30,394,236
chr9	2	24,931,883
chr10	4	23,894,053
chr11	3	35,128,029
chr12	3	26,761,116
Total	38	395,782,165

Supplemental Table 3. The alignment between 24 gaps and MH63RS2.

Gap ID	Chr ID	Gap length (bp)	Alignment coverage (%)	Alignment length
chr1-013-00	Chr01	26,449	100	26,449
chr1-015-00	Chr01	41,095	100	41,095
chr2-008-00	Chr02	227,150	100	227,150
chr2-009-00	Chr08	346,200	88.62	306,793
chr2-010-00	Chr02	332,102	100	332,102
chr2-011-00	Chr02	172,627	100	172,627
chr2-012-00	Chr02	56,472	100	56,472
chr3-007-00	Chr03	72,573	99.99	72,569
chr4-006-00	Chr04	66,912	100	66,912
chr5-003-00	Chr03	87,199	100	87,199
chr5-004-00	Chr05	84,405	100	84,405
chr5-005-00	Chr10	616,719	87.85	541,779
chr6-001-00	Chr06	608,305	100	608,304
chr6-002-00	Chr06	383,428	52.96	203,067
chr8-017-00	Chr08	97,121	100	97,121
chr8-018-00	Chr08	2,177	100	2,177
chr8-019-00	Chr08	21,090	100	21,090
chr9-016-00	Chr09	386,058	100	386,058
chr10-024-00	Chr10	648,325	100	648,321
chr10-025-00	Chr10	520,537	100	520,537
chr11-022-00	Chr11	36,960	100	36,959
chr11-023-00	Chr11	77,603	100	77,603
chr12-020-00	Chr12	127,827	100	127,827
chr12-021-00	Chr12	72,384	100	72,383

Supplementary Table 4. Statistics of paired-end reads mapping.

		Percent (%)
Reads	Mapping rate (%)	99.34
Genome	Average sequencing depth	67.81
	Coverage (%)	98.98
	Coverage at least 4X (%)	97.10
	Coverage at least 10X (%)	93.78

Supplementary Table 5. Assessment of Minghui 63 genome using 2045 *Oryza rufipogon* W1943 full-length cDNA sequences (<http://server.ncgr.ac.cn/ricd/dym/ftp.php>).

Dataset	Number	with >90% sequence in one scaffold		with >50% sequence in one scaffold	
		Number	Percent	Number	Percent
>0bp	2,045	1,880	91.93	1,971	96.38
>200bp	2,043	1,878	91.92	1,969	96.38
>500bp	1,872	1,724	92.09	1,806	96.47
>1000bp	620	583	94.03	606	97.74

Supplementary Table 6. Assessment the gene coverage rate using BUSCO.

Species	Size (Mb)	BUSCO notation assessment results
MH63KL1	395.82	C:98.6% [S:96.1%, D:2.5%], F:0.3%, M:1.1%
IRGSP-1.0	374.42	C:98.3% [S:95.7%, D:2.6%], F:0.6%, M:1.1%
MH63RS1	359.93	C:93.1% [S:90.8%, D:2.3%], F:0.7%, M:6.2%
MH63RS2	387.42	C:98.0% [S:95.7%, D:2.3%], F:0.3%, M:1.7%

Supplemental Table 7. Assessment of the Minghui 63 genome assembly using the whole sequence of 12 BACs of centromeric regions.

BAC ID	Chr ID	BAC length (bp)	Alignment coverage (%)	Alignment length
AC022352.5	chr10	139,398	90.53	126,203
AC137922.3	chr11	141,001	100	141,001
AC137925.2	chr3	140,507	99.61	139,962
AC137984.2	chr5	114,329	100	114,329
AP005763.3	chr6	140,893	100	140,893
AP006480.2	chr8	190,806	100	190,806
AP007145.2	chr2	69,673	100	69,673
AP007147.1	chr9	33,982	99.97	33,973
AP007228.1	chr1	104,999	98.84	103,778
AP007259.1	chr7	66,793	99.85	66,693
BX000556.3	chr12	137,071	99.78	136,767
BX890594.1	chr4	181,586	100	181,586

Supplemental Table 8. The prediction of repeat elements in Minghui 63 genome.

Type	Repeat Size (bp)	Percent (%)
TRF	17,612,333	4.45
RepeatMasker	148,491,480	37.51
RepeatProteinMask	48,979,668	12.37
Total	163,341,094	41.27

The RepeatMasker and RepeatProteinMask were used to identify TEs. The TRF was used to discover tandem repeats. The Total was integrated of all repeat elements without redundancy.

Supplemental Table 9. Categories of TEs predicted in Minghui 63 genome.

Type	Repeatmasker		TE Proteins		Combined TEs	
	Length (bp)	% in Genome	Length (bp)	% in Genome	Length (bp)	% in Genome
DNA	21,016,884	5.31	9,536,305	2.41	26,015,497	6.57
LINE	2,154,546	0.54	4,320,725	1.09	5,832,459	1.47
SINE	112,631	0.03	0	0	112,631	0.03
LTR	109,277,567	27.61	35,229,749	8.90	112,129,584	28.33
Unknown	2,166,311	0.55	25,047	0.01	2,191,358	0.55
Total	148,491,480	37.51	48,979,668	12.37	157,157,235	39.70

Supplemental Table 10. General statistics of predicted protein-coding genes.

	Gene set	Number	Average transcript length (bp)	Average CDS length (bp)	Average exons per gene	Average exon length (bp)	Average intron length (bp)
<i>De novo</i> [#]	Augustus	34,361	2,313.48	1,008.19	3.68	274.06	487.27
	GlimmerHMM	79,929	3,859.33	688.64	2.75	250.61	1,814.05
	SNAP	59,095	4,907.51	720.34	3.76	191.62	1,517.49
	Geneid	94,755	2,693.71	757.43	3.34	226.93	828.28
	Genscan	64,931	4,265.12	1,147.96	5.53	207.50	687.75
	<i>Oryza sativa</i> L. <i>japonica</i>	80,950	1,200.58	720.23	2.41	298.78	340.53
Homolog ^{\$}	<i>Brachypodium distachyon</i>	28,827	2,454.62	1,113.64	4.13	269.63	428.40
	<i>Setaria italica</i>	32,324	2,262.27	1,057.77	3.80	278.40	430.26
	<i>Oryza sativa</i> L. <i>indica</i>	41,142	2,181.00	1,092.37	3.56	307.01	425.57
	<i>Zea mays</i>	33,384	2,322.07	1,161.51	3.57	325.10	451.09
	<i>Arabidopsis thaliana</i>	27,618	2,020.20	975.16	3.52	277.41	415.48
RNA_Seq	<i>Sorghum bicolor</i>	28,969	2,589.89	1,169.29	4.18	279.65	446.56
	Cufflinks	48,512	4,675.67	2,241.90	6.13	365.65	474.30
	PASA	25,681	2,380.04	1,068.65	4.18	255.58	412.21
	EVM	45,019	2,483.56	1,043.42	3.95	264.04	487.89
	PASA update	44,977	2,452.91	1,042.40	3.95	263.97	478.31
	Final set	36,149	2,728.08	1,146.49	4.36	262.83	470.41

[#] Statistics calculated from the gene set predicted by ab initio methods.

^{\$} Statistics calculated from the gene set predicted by homolog proteins from each species.

Supplemental Table 11. General statistics of protein-coding genes of Minghui 63 and homolog species.

Species	Number	Average transcript length (bp)	Average CDS length (bp)	Average exons per gene	Average exon length (bp)	Average intron length (bp)
<i>Oryza sativa</i> L.indica	36,149	2,728.08	1,146.49	4.36	262.83	470.41
<i>Oryza sativa</i> L. japonica	35,311	2,200.75	1,008.79	3.83	263.56	421.56
<i>Brachypodium distachyon</i>	34,014	2,524.76	1,140.59	4.46	255.70	399.98
<i>Setaria italica</i>	33,664	2,308.30	1,165.43	4.54	256.45	322.44
<i>Zea mays</i>	39,235	3,485.39	1,183.48	4.93	239.91	585.29
<i>Sorghum bicolor</i>	33,878	2,771.31	1,169.67	4.53	257.97	453.19

Supplemental Table 12. General statistics of functional annotation of protein-coding genes.

	Number	Percent (%)
Total	36,149	-
Swissprot	22,729	62.9
NR	34,965	96.7
KEGG	21,520	59.5
InterPro	35,650	98.6
Pfam	24,194	66.9
GO	32,381	89.6
Annotated	36,062	99.8
Unannotated	87	0.2

Supplemental Table 13. CentO (155 bp satellite DNA) units in Minghui 63 genome.

Chromosome	CentO sequence location		Total no. of units	Total length (bp)
	Start	End		
chr1	17,510,384	18,355,493	1,111	845,110
chr2	13,502,971	13,829,567	683	326,597
chr3	21,134,162	21,646,956	124	512,795
chr4	9,549,585	9,697,276	307	147,692
chr5	12,537,089	12,656,508	707	119,420
chr6	15,460,307	15,580,164	322	119,858
chr7	12,448,237	12,450,250	13	2,014
chr7	13,139,270	13,143,534	27	4,265
chr8	13,481,133	13,583,132	183	102,000
chr8	14,466,603	14,536,492	13	69,890
chr9	2,612,761	2,654,031	56	41,271
chr9	3,105,486	3,289,248	149	183,763
chr10	9,456,089	9,495,536	5	39,448
chr11	13,789,153	13,806,601	59	17,449
chr12	10,570,208	11,003,912	549	433,705
Total	-	-	4,308	2,965,277

Supplemental Table 14. Statistic distribution of SDs on chromosomes.

The overlapped blocks are merged using bedtools.

Chromosome	SDs number	SD Length (bp)	Chr length (bp)	percentage (%)
chr1	2,452	7,819,396	44,998,099	17.38
chr2	1,992	6,831,623	36,790,212	18.57
chr3	2,025	6,775,221	39,918,125	16.97
chr4	1,934	9,505,882	37,315,983	25.47
chr5	1,762	6,279,285	31,239,580	20.10
chr6	1,808	6,999,350	31,823,770	21.99
chr7	1,933	6,942,480	30,883,769	22.48
chr8	1,778	7,157,808	30,306,463	23.62
chr9	1,467	5,540,613	24,951,677	22.21
chr10	1,622	6,720,233	25,682,544	26.17
chr11	2,103	11,236,816	35,116,109	32.00
chr12	1,571	7,875,604	26,795,781	29.39
Total	22,447	89,684,311	395,822,112	22.66

Supplemental Table 15. The KEGG enrichment analysis of the subfunctional gene pairs.

MapID	MapTitle	Pvalue	Number
map00908	Zeatin biosynthesis	9.01E-07	7
map00460	Cyanoamino acid metabolism	7.40E-05	9
map00960	Tropane, piperidine and pyridine alkaloid biosynthesis	3.27E-04	6
map00260	Glycine, serine and threonine metabolism	4.82E-04	8
map00950	Isoquinoline alkaloid biosynthesis	6.22E-04	6
map00073	Cutin, suberine and wax biosynthesis	1.30E-03	6
map00480	Glutathione metabolism	1.31E-03	8
map00942	Anthocyanin biosynthesis	1.48E-03	4
map00410	beta-Alanine metabolism	1.53E-03	6
map00906	Carotenoid biosynthesis	1.89E-03	6
map00360	Phenylalanine metabolism	2.09E-03	6
map00350	Tyrosine metabolism	2.80E-03	6
map00903	Limonene and pinene degradation	4.70E-03	4
map00905	Brassinosteroid biosynthesis	8.27E-03	3
map00565	Ether lipid metabolism	1.27E-02	4
map00062	Fatty acid elongation	1.33E-02	4
map00630	Glyoxylate and dicarboxylate metabolism	4.48E-02	4
map00790	Folate biosynthesis	4.82E-02	2

Supplemental Table 16. General statistics of intact LTR-RTs.

Species	Total Length (bp)	Total Number	Max length (bp)	Min length (bp)	Average length (bp)
MH63KL1	53,645,453	5,516	25,673	1,165	9,725.43
IRGSP-1.0	43,495,559	4,706	24,598	1,217	9,242.58

Supplemental Table 17. The classification of LTR-RTs.

MH63KL1		IRGSP-1.0	
family	number	family	number
Del	1,911	Tat	1,223
Tat	1,057	Del	923
Tork	373	Tork	455
Crm	319	Crm	238
Sire	187	Sire	231
Reina	139	Reina	143
Oryco	131	Retrofit	129
Retrofit	117	Oryco	127
Athila	91	Athila	114
Copia	31	Copia	34

Supplemental Table 18. General statistics of *NBS-LRR* genes.

NBS type	MH63KL1	IRGSP-1.0
TIR-NBS	2	1
CC_NBS_LRR	25	8
CC_NBS	92	67
NBS_LRR	77	51
NBS	291	222
Total	487	349

Supplemental Table 19. The distribution of *NBS-LRR* genes on each chromosome.

chr	MH63KL1	IRGSP-1.0
chr1	44	35
chr2	30	20
chr3	19	11
chr4	27	15
chr5	23	12
chr6	31	25
chr7	39	23
chr8	47	35
chr9	26	16
chr10	33	24
chr11	123	98
chr12	45	35

Supplemental Table 20. SDs collinear block pairs of Minghui 63-specific *NBS-LRR* genes.

Gene	Paralog	Block 1			Block 2		
		Chr	Begin	End	Chr	Begin	End
chr1.1950	chr1.1948	chr1	15,206,081	15,208,811	chr1	15,191,933	15,194,653
chr1.2351	chr1.2353	chr1	20,048,633	20,050,118	chr1	20,061,794	20,063,208
chr10.196	chr10.195	chr10	2,272,713	2,274,174	chr10	2,260,604	2,262,123
chr10.198	chr10.207	chr10	2,278,777	2,281,186	chr10	2,459,314	2,461,701
chr10.214	chr10.212	chr10	2,542,414	2,543,660	chr10	2,517,918	2,519,224
chr11.1929	chr11.1933	chr11	20,161,866	20,180,486	chr11	20,199,831	20,223,919
chr11.2559	chr11.2569	chr11	26,199,357	26,204,183	chr11	26,268,592	26,279,836
chr11.2563	chr11.2569	chr11	26,230,469	26,240,673	chr11	26,258,836	26,276,867
chr11.2564	chr11.2569	chr11	26,230,469	26,240,673	chr11	26,258,836	26,276,867
chr11.2569	chr11.2562	chr11	26,268,593	26,273,632	chr11	26,216,546	26,221,632
chr11.2627	chr11.2631	chr11	26,772,540	26,786,250	chr11	26,791,102	26,798,583
chr11.2630	chr11.2650	chr11	26,794,026	26,798,833	chr11	26,995,293	26,999,044
chr11.2650	chr11.2627	chr11	26,993,091	26,999,037	chr11	26,782,281	26,788,587
chr11.2758	chr11.2762	chr11	27,960,453	27,962,164	chr11	28,010,487	28,012,132
chr11.504	chr11.502	chr11	3,031,061	3,032,647	chr11	3,016,824	3,018,344
chr11.3080	chr11.3159	chr11	30,908,616	30,921,124	chr11	32,232,486	32,236,510
chr11.3098	chr11.3109	chr11	31,234,226	31,250,174	chr11	31,382,126	31,391,580
chr11.3171	chr11.3169	chr11	32,431,917	32,448,521	chr11	32,347,057	32,357,213
chr11.3171	chr11.3186	chr11	32,429,933	32,450,434	chr11	32,678,699	32,692,357
chr11.3202	chr11.3204	chr11	32,824,032	32,827,580	chr11	32,852,436	32,856,050
chr11.3202	chr11.3212	chr11	32,818,595	32,827,146	chr11	32,928,981	32,937,482
chr11.3202	chr5.2660	chr11	32,818,595	32,827,497	chr5	24,529,983	24,538,245
chr11.3204	chr11.3216	chr11	32,852,780	32,855,888	chr11	32,973,570	32,977,182
chr11.3210	chr11.3201	chr11	32,928,981	32,937,482	chr11	32,818,595	32,827,146
chr11.3210	chr5.2659	chr11	32,928,978	32,937,927	chr5	24,529,980	24,538,332
chr11.3211	chr11.3216	chr11	32,935,401	32,937,697	chr11	32,975,159	32,977,274
chr11.900	chr11.865	chr11	6,172,902	6,182,055	chr11	5,929,229	5,935,737
chr12.273	chr12.255	chr12	1,574,743	1,590,466	chr12	1,493,993	1,504,897
chr12.274	chr12.255	chr12	1,574,743	1,590,466	chr12	1,493,993	1,504,897
chr2.1330	chr2.1328	chr2	10,269,885	10,271,823	chr2	10,247,026	10,249,033
chr2.1501	chr8.3011	chr2	11,824,604	11,827,719	chr8	28,818,155	28,820,600
chr2.1947	chr2.1933	chr2	17,099,867	17,117,072	chr2	16,978,349	16,997,586
chr2.1949	chr2.1947	chr2	17,123,173	17,124,435	chr2	17,115,622	17,116,894
chr2.508	chr2.725	chr2	3,332,086	3,335,267	chr2	5,022,234	5,025,338
chr2.4450	chr2.4451	chr2	35,944,050	35,951,317	chr2	35,969,028	35,977,131
chr2.804	chr2.806	chr2	5,564,333	5,565,856	chr2	5,575,822	5,577,354
chr4.1216	chr4.1135	chr4	15,750,050	15,817,779	chr4	14,916,357	14,982,666
chr4.1495	chr4.1503	chr4	19,335,420	19,339,595	chr4	19,415,063	19,419,297
chr4.3440	chr4.3434	chr4	33,578,106	33,579,345	chr4	33,472,397	33,473,651
chr4.3440	chr4.3438	chr4	33,545,152	33,579,345	chr4	33,529,517	33,563,204
chr4.67	chr4.71	chr4	671,931	673,114	chr4	687,954	689,256
chr4.68	chr4.71	chr4	671,931	673,114	chr4	687,954	689,256

Gene	Paralog	Block 1		Block 2	
		Chr	Begin	Chr	Begin
chr5.1948	chr5.1950	chr5	19,117,554	chr5	19,129,760
chr5.1949	chr5.1952	chr5	19,120,093	chr5	19,191,927
chr5.1951	chr5.1950	chr5	19,155,626	chr5	19,130,356
chr5.1952	chr5.1948	chr5	19,191,927	chr5	19,120,093
chr5.906	chr6.1235	chr5	7,295,742	chr6	8,786,255
chr5.1107	chr11.3290	chr5	9,179,105	chr11	33,665,103
chr5.1107	chr5.1108	chr5	9,184,088	chr5	9,190,981
chr6.1349	chr6.1350	chr6	10,024,579	chr6	10,032,923
chr6.1353	chr6.1350	chr6	10,054,200	chr6	10,033,864
chr6.2803	chr6.2797	chr6	25,718,166	chr6	25,693,277
chr6.530	chr6.531	chr6	3,108,499	chr6	3,115,474
chr6.532	chr6.530	chr6	3,115,474	chr6	3,108,499
chr6.1235	chr4.1126	chr6	8,786,359	chr4	14,874,827
chr7.154	chr7.199	chr7	1,015,554	chr7	1,277,302
chr7.154	chr7.201	chr7	1,015,554	chr7	1,277,302
chr7.1315	chr1.1400	chr7	10,415,662	chr1	10,056,197
chr7.1410	chr5.599	chr7	11,696,749	chr5	4,608,356
chr7.1410	chr7.1411	chr7	11,696,747	chr7	11,744,738
chr7.198	chr7.155	chr7	1,277,302	chr7	1,015,554
chr7.199	chr7.155	chr7	1,277,302	chr7	1,015,554
chr7.200	chr7.154	chr7	1,277,302	chr7	1,015,554
chr7.1684	chr7.1662	chr7	15,758,135	chr7	15,437,951
chr7.2245	chr7.2242	chr7	21,198,468	chr7	21,141,097
chr7.151	chr7.153	chr7	995,036	chr7	1,012,593
chr7.151	chr7.154	chr7	995,036	chr7	1,012,593
chr8.3001	chr8.3013	chr8	28,759,495	chr8	28,826,185
chr8.3001	chr8.3018	chr8	28,759,131	chr8	28,863,960
chr9.1897	chr9.1895	chr9	20,037,644	chr9	20,027,647
chr9.739	chr9.742	chr9	9,277,792	chr9	9,325,725

Supplemental Table 21. SDs collinear block pairs of four rice blast resistance genes.

Gene Name	Gene ID	Paralog ID	Chr	Block1		Block2	
				Begin	End	Begin	End
<i>Pit</i>	chr1.398	chr1.396	chr1	2,724,084	2,729,780	2,695,549	2,700,641
<i>Pbl</i>	chr11.2569	chr11.2562	chr11	26,268,593	26,273,632	26,216,546	26,221,632
<i>Piz</i>	chr6.1350	chr6.1349	chr6	10,032,923	10,036,249	10,024,579	10,027,886
<i>Pib</i>	chr2.4450	chr2.4451	chr2	35,944,050	35,951,317	35,969,028	35,977,131

Supplemental Table 22. SDs collinear block pairs of *cis*-zeatin-*O*-glucosyltransferase genes.

Gene ID	Paralog ID	Chr	Block1		Block2	
			Begin	End	Begin	End
chr4.2981	chr4.2983	chr4	30,071,801	30,073,214	30,093,424	30,094,837
chr4.2909	chr4.2910	chr4	29,601,285	29,603,206	29,606,146	29,607,947