Supplementary Information

Whole Genome Sequencing of *Oryza brachyantha* Reveals Mechanisms Underlying *Oryza* Genome Evolution

Jinfeng Chen[1], Quanfei Huang[2], Dongying Gao[3], Junyi Wang[2], Yongshan Lang[2], Tieyan Liu[1], Bo Li[1], Jose Luis Goicoechea[4], Chengzhi Liang[1], Chengbin Chen[5], Wenli Zhang[6], Shouhong Sun[1], Yi Liao[1], Zetao Bai[1], Xuemei Zhang[1], Lu Yang[1], Chengli Song[1], Meijiao Wang[1], Jinfeng Shi[1], Geng Liu[2], Junjie Liu[2], Heling Zhou[2], Weili Zhou[2], Qiulin Yu[2], Na An[2], Yan Chen[2], Qingle Cai[2], Bo Wang[2], Binhang Liu[2], Jiumeng Min[2], Ying Huang[2], Honglong Wu[2], Zhenyu Li[2],Yong Zhang[2], Ye Yin[2], Wenqin Song[5], Jiming Jiang[6], Scott A. Jackson[3], Rod A. Wing[4], Jun Wang[2] & Mingsheng Chen[1]

[1]State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China;

[2]BGI-Shenzhen, Shenzhen 518000, China;

[3]Institute of Plant Breeding, Genetics & Genomics, University of Georgia, Athens, GA 30602, USA;

[4]Arizona Genomics Institute, School of Plant Sciences, BIO5 Institute, University of Arizona, Tucson, Arizona 85721, USA;

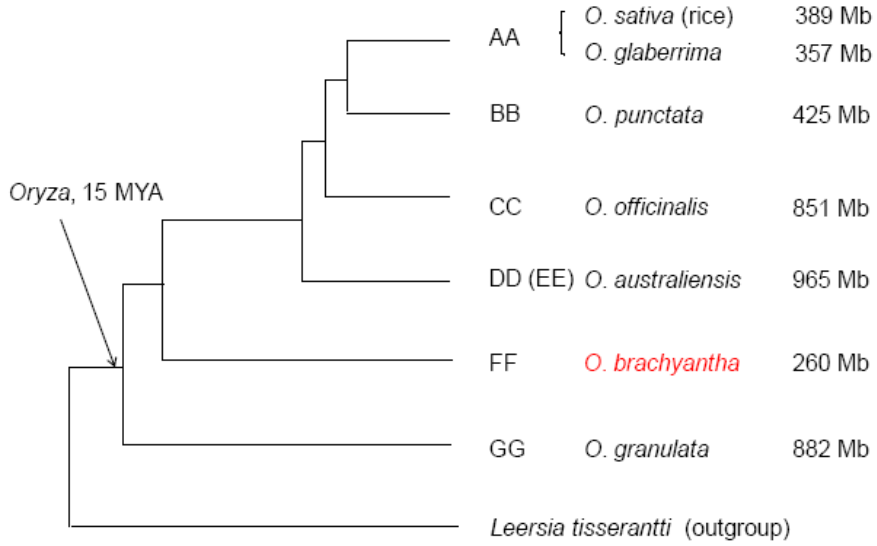
[5]Department of Genetics and Cell Biology, Nankai University, Tianjin 300071, China;

[6]Department of Horticulture, University of Wisconsin-Madison, Madison, WI 53706, USA;

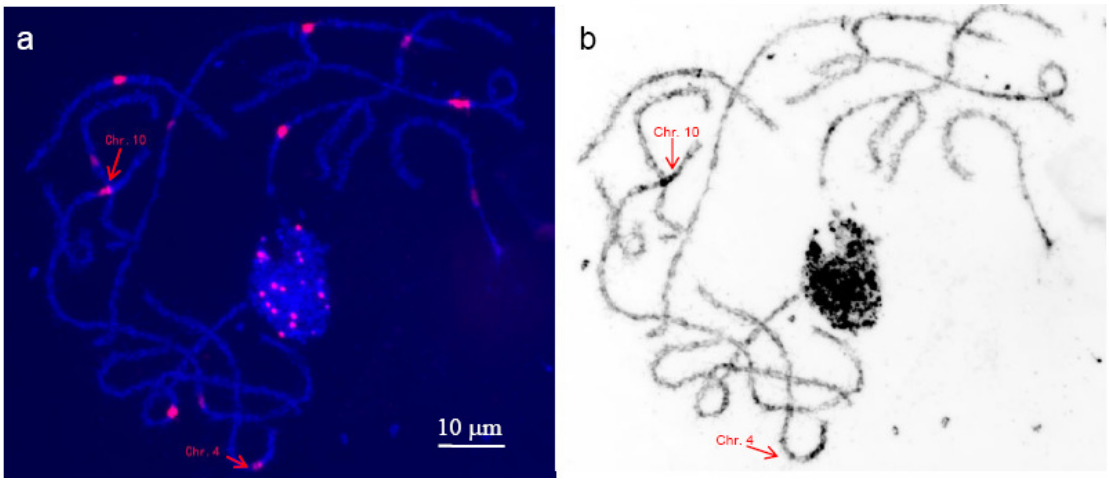
Correspondence and requests for materials should be addressed to M.-S.C. (mschen@genetics.ac.cn), or to J.W. (wangj@genomics.org.cn), or to R.A.W. (rwing@Ag.arizona.edu).

Table of content

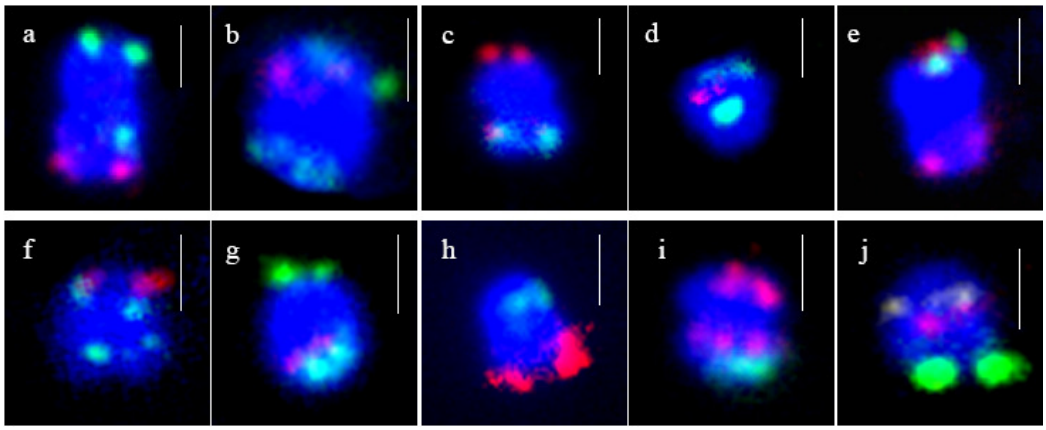
Supplementary Figures



Supplementary Figure S1 Phylogenetic tree of diploid genome types in the genus *Oryza*[61]. The deepest split of *Oryza* species was estimated to be ~ 15 million years ago (MYA)[62]. Genome sizes were estimated by flow cytometry[63]; the genome sizes of *O. sativa* and *O. brachyantha* used the sizes of the genome sequence.



Supplementary Figure S2 Karyotyping pachytene chromosomes of *O. brachyantha*. (**a**), Chromosomes in the pachytene cell of *O. brachyantha* were hybridized with centromere-specific satellite repeat CentO-F[64] (red). The short arms and pericentromeric part of the long arms of chromosomes 4 and10 are relatively condensed or relatively brighter after DAPI staining than the rest of the chromosomes. However, the differences between heterochromatin and euchromatin are not as distinct as observed in *O. sativa* ssp. *japonica* cv. Nipponbare[65]. *O. brachyantha* shows an overall lack of pericentromeric heterochromatin domains compared to Nipponbare[65], consistent with its compact genome size. (**b**), DAPI stained chromosomes in Supplementary Figure S2a were converted into black-and-white image to enhance the visualization of euchromatin and heterochromatin on chromosomes. Bar, 10 μm.



Supplementary Figure S3 Anchoring sequence blocks onto chromosomes of *O. brachyantha*.
Sequence blocks described in Figure 1 were anchored onto chromosomes by fluorescence in situ hybridization (FISH). Candidate BAC clones of the sequence blocks were used as probes to hybridize the somatic metaphase cells of *O. brachyantha*. Reference clones were developed for each chromosome in *O. brachyantha*, except for chromosome 6 and 9 due to experimental failure. BAC clones representing the sequence blocks were selected by integrating the sequence blocks with the physical map. The candidate clones should have BAC end sequences (BES) properly aligned to the sequence blocks, and have a lower proportion of repetitive sequences than the genome average (29%). All bars, 1 μm.

(**a**), a0014B22 (block1, green-top), a0054K20 (block2, green-down), a0016F05 (chr01, red).

(**b**), a0054A06 (chr02, green-top), a0077H19 (block1, red), a0039G13 (block4, green-down).

(**c**), a0001K24 (block1, red-top), a0054J06 (block2, red-down), a0002D06 (chr03, green).

(**d**), a0027N15 (block1, green-top), a0056E10 (chr04, red), a0043D12 (block3, green-down).

(**e**), a0063E20 (chr05, red-top), a0044E01 (block1, red-down), a0044D12 (block3, green).

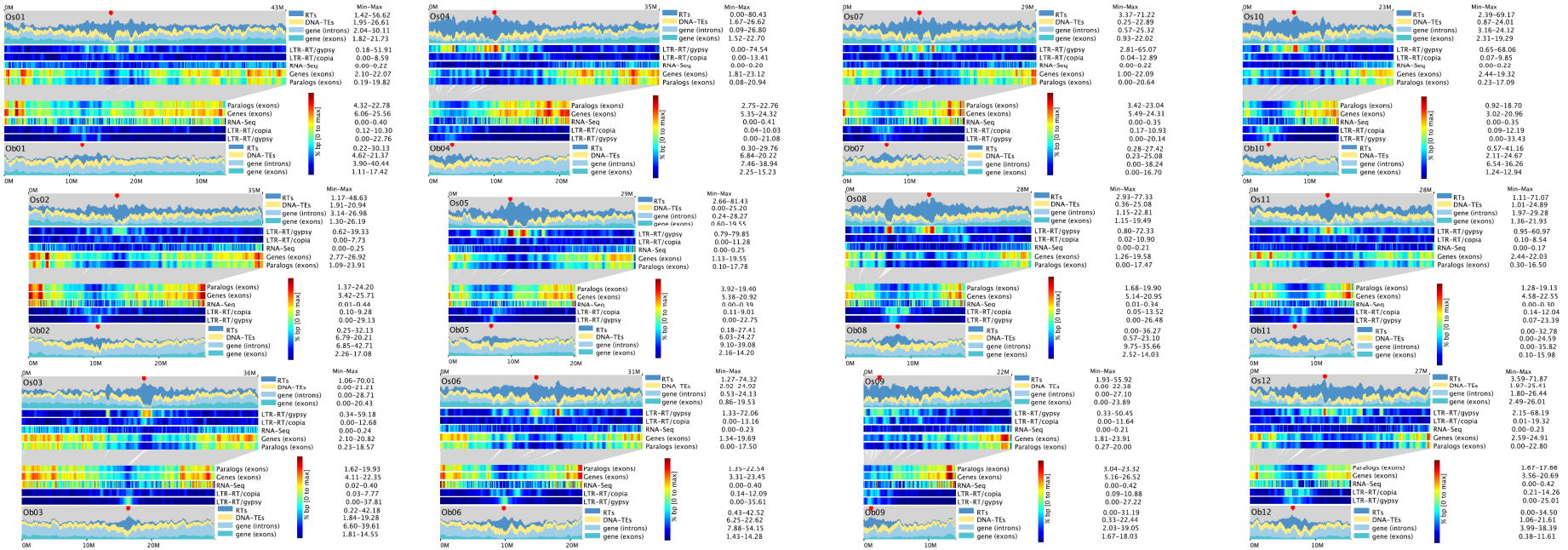
(**f**), a0039O02 (chr07, red), a0037L06 (block1, green-top), a0025B11 (block2, green-down).

(**g**), a0024A18 (block1, green-top), a0033B06 (chr08, red), a0040O15 (block2, green-down).

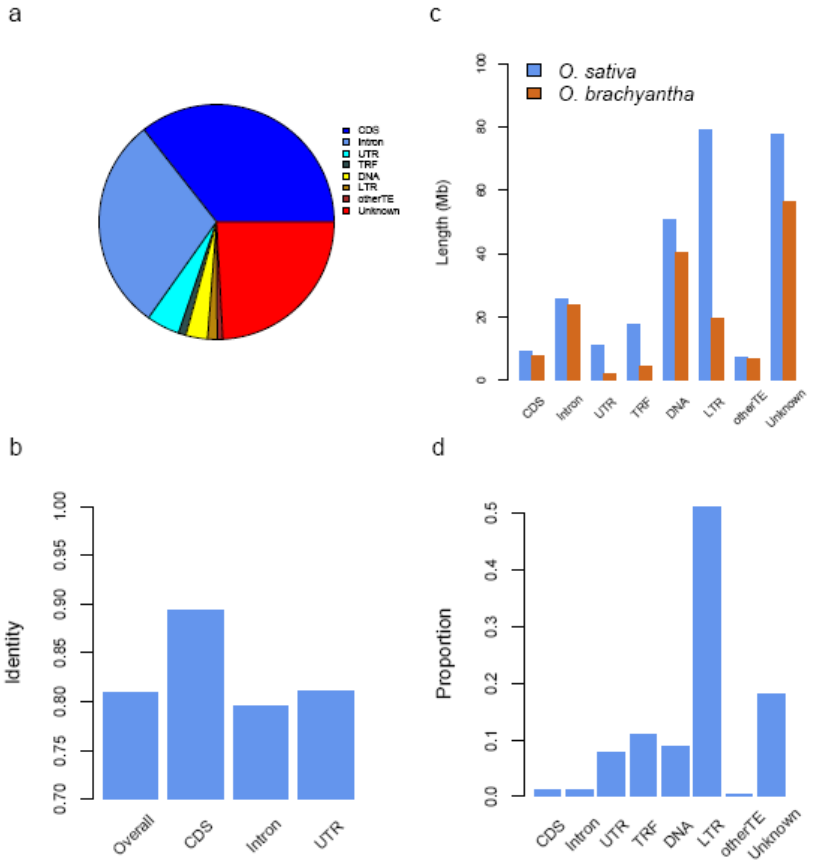
(**h**), a0027A23 (block1, green), a0090D06 (chr10, red).

(**i**), a0008N06 (block1, red-top), a0032H18 (chr11, red), a0074E19 (block3, green).

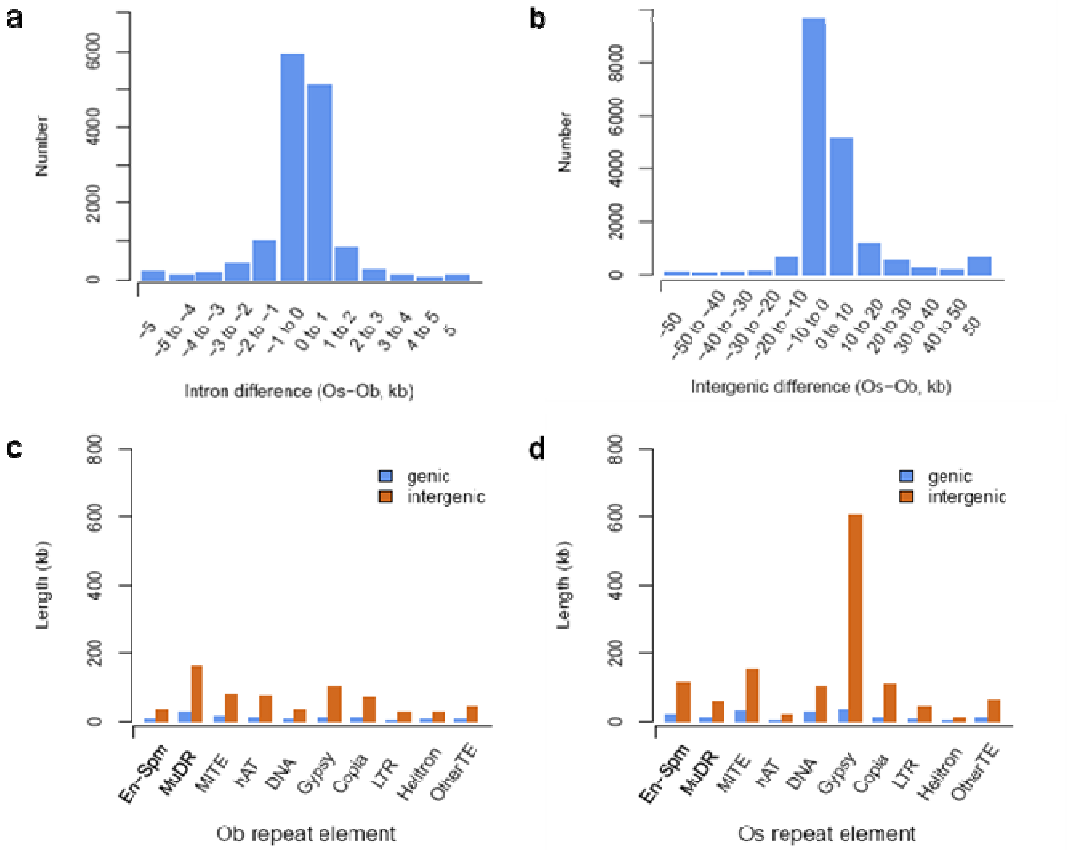
(**j**), a0071C04 (chr12, yellow), a0094I04 (block1, red), a0052P11 (block4, green).



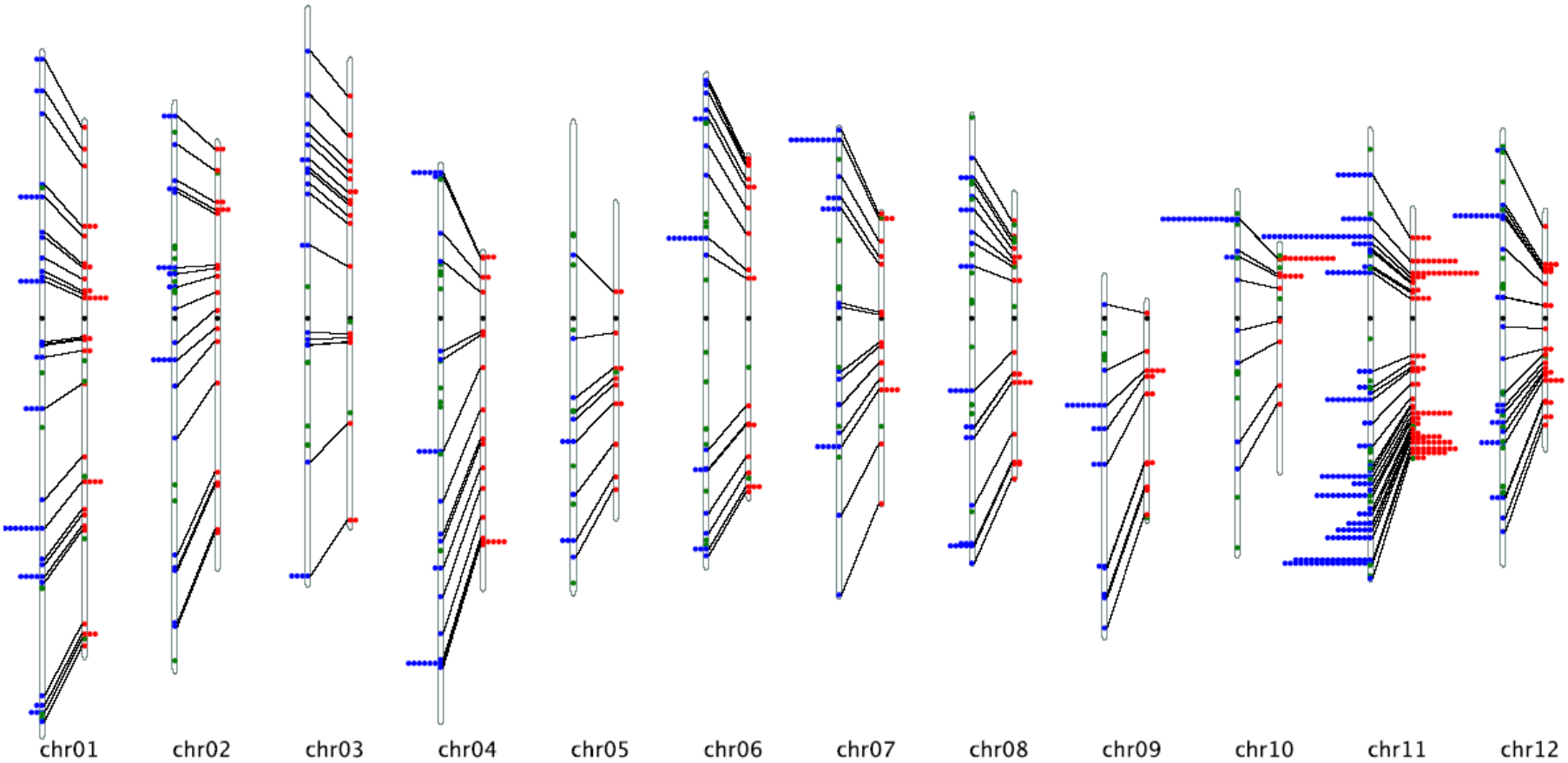
Supplementary Figure S4 Comparisons of genomic features between *O. brachyantha* (down) and *O. sativa* (up). The density of genomic features, including genes, transposable elements, RNA-Seq transcriptomes, were calculated for each 200 kb window. The stacked bar plots shows proportional distribution of retrotransposons (RTs), DNA-TEs (DNA transposons), introns and exons of genes. The heatmap tracks shows the distribution of *gypsy* and *copia* type LTR retrotransposons, RNA-Seq transcriptional level, all exons and exons with homologous sequence in the compared genome. The scale is different for each track, ranging from 0 to max value as indicated at the right of each chromosome. Centromeres are indicated by red arrows. Homologous exons between *O. brachyantha* and *O. sativa* are connected by gray lines. For all chromosomes, retrotransposons are concentrated at pericentromeric regions, whereas genes and DNA transposons are co-localized on chromosome arms. Exceptions were found on chromosome 4 and chromosome 10, which show a spread of a high proportion of retrotransposons along the short arms and the proximal regions of the long arms. The density of retrotransposons is much higher in *O. sativa* than *O. brachyantha*, largely due to the high proportion of *gypsy* type retrotransposons in *O. sativa*. These highly repetitive regions show low levels of transcription and homologous exons between the genomes of *O. brachyantha* and *O. sativa*.



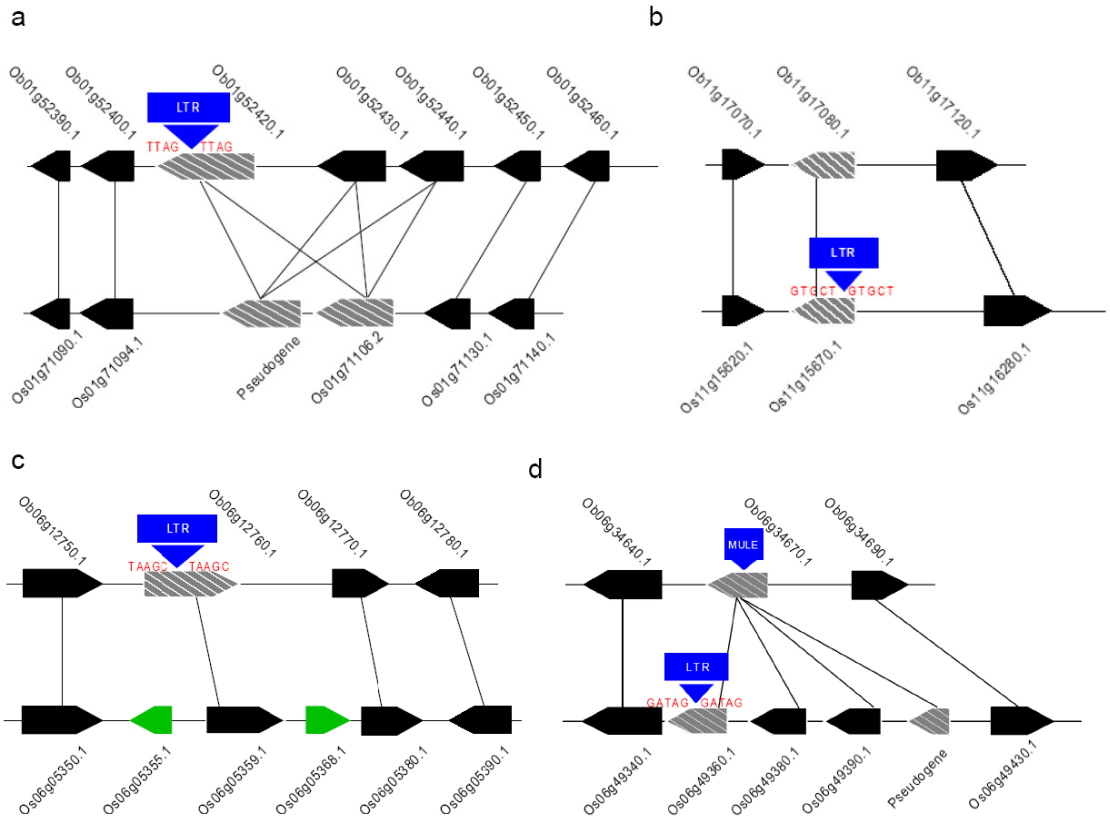
Supplementary Figure S5 Comparison of genome alignments between *O. brachyantha* and *O. sativa*. The genome sequences of *O. brachyantha* and *O. sativa* were aligned by LASTZ as described in Method. Gene and repeat annotations of *O. brachyantha* were used to calculate the distributions of genomic features in the alignable sequences. Most of the alignable sequences are consisted of coding sequences (CDS, 35%), introns (29%), unknown sequences (Unknown, 24%) and untranslated regions (UTR, 4%). The overall sequence identity between *O. brachyantha* and *O. sativa* is 80% with higher identity (89%) in CDS regions. The composition of unalignable sequences is different between *O. brachyantha* and *O. sativa*. Unknown sequences (35%) and DNA transposons (25%) are the most abundance unalignable sequences of *O. brachyantha*, whereas Unknown sequences (28%) and LTR retrotransposons (28%) are dominant in *O. sativa*. In total, the unalignable sequences contribute ~ 116 Mb sequence difference between *O. brachyantha* and *O. sativa*, in which LTR retrotransposons account for ~ 60 Mb. (**a**), Distribution of genomic component in the alignable sequences. (**b**), Sequence identity comparisons among genomic components. (**c**), Length distribution of genomic components in the unalignable sequences. (**d**), Contribution of genomic components to genome size variation between *O. brachyantha* and *O. sativa*.



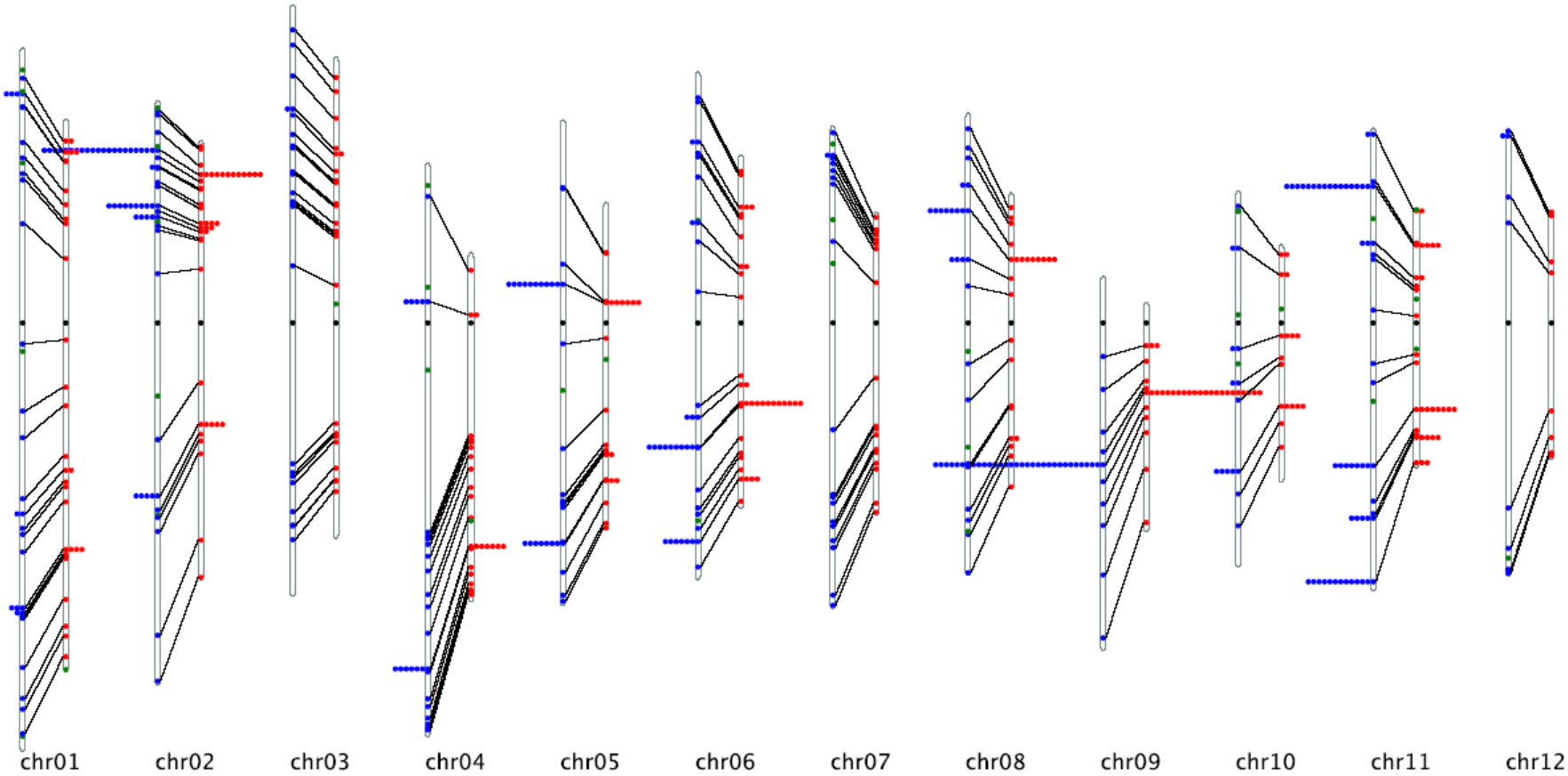
Supplementary Figure S6 Influence of introns and intergenic regions to genome size variation. (**a**). Differences on intron size of orthologous genes between *O. brachyantha* and *O. sativa*; (**b**). Differences on the size of orthologous intergenic regions between *O. brachyantha* and *O. sativa*; (**c**). Length of transposable elements located in genic and intergenic regions of *O. brachyantha*; (**d**). Length of transposable elements located in genic and intergenic regions of *O. sativa*.



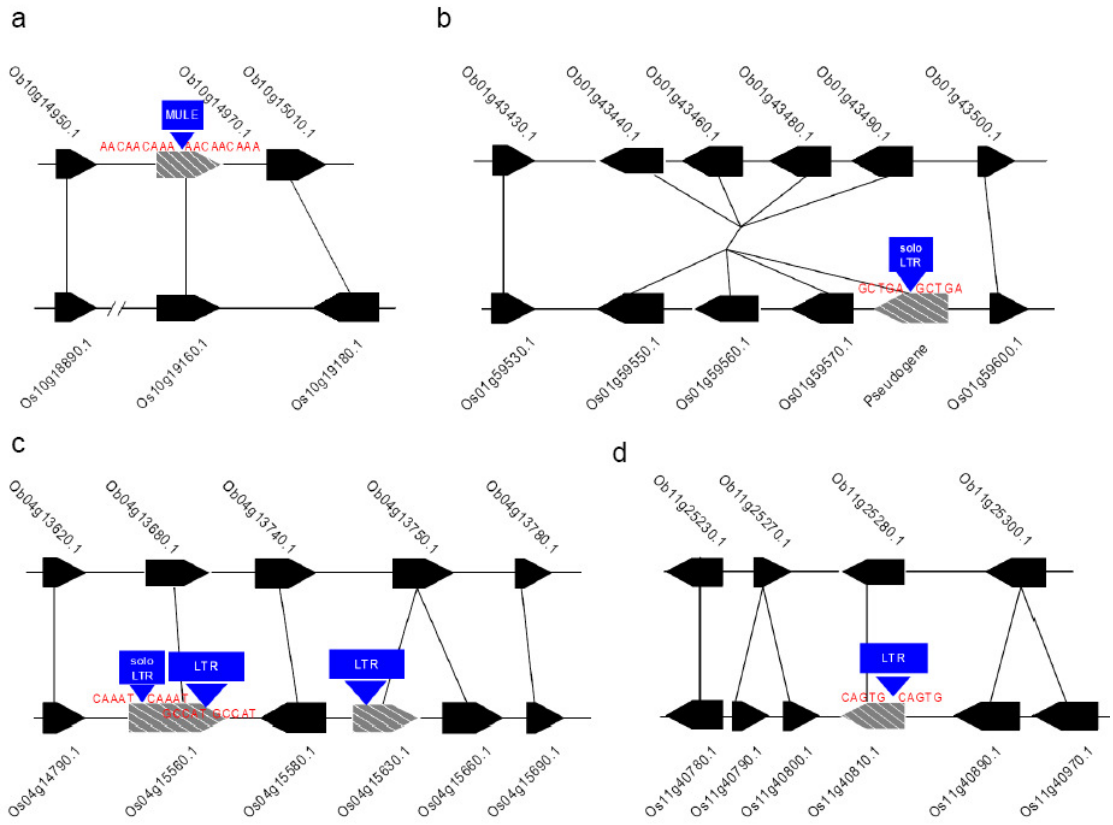
Supplementary Figure S7 Comparisons of chromosome distribution of NBS-LRR genes between *O. brachyantha* and *O. sativa*. For each chromosome pair, the left one represents the chromosome of *O. sativa* and the right one represents the chromosome of *O. brachyantha*. Centromeres are indicated by solid black circles. Genes are indicated by colored circles; Tandemly duplicated genes are clustered to one representive position; Blue circles stand for genes of *O. sativa*; Red circles stand for genes of *O. brachyantha*; Green circles stand for non-collinear genes; Orthologous genes or gene clusters are connected by black lines.



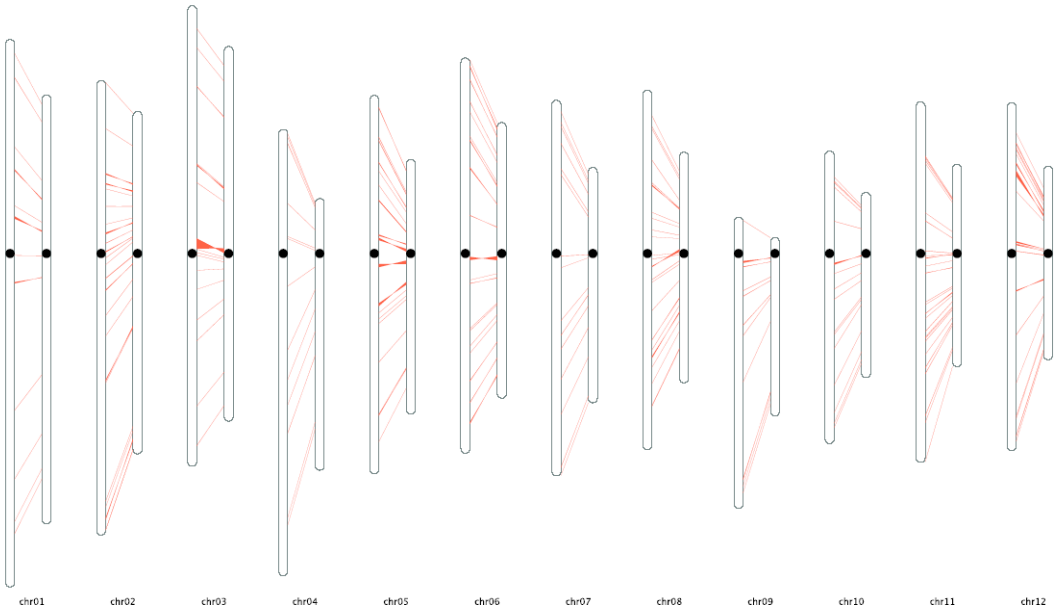
Supplementary Figure S8 Examples of NBS-LRR type pseudogenes in *O. brachyantha* (up) and *O. sativa* (down). We compared the gene structures of orthologous gene clusters to identify pseudogenes in both genomes. Only genes with exons interrunpted by frameshift mutations or transposable element insertions were defined as pseudogenes. The pseudogenezation may cause lose of function of resistense genes in these loci in one or both genomes, which contributed to the divergence of *Oryza* species. Black arrowhead boxes, genes; Shaded arrowhead boxes, pseudogenes; Green arrowhead boxes, non-collinear genes; blue boxes, transposable elements; MULE, Mutator-like element; LTR, long terminal repeat retrotransposon; Target site duplication (TSD) are in red; Orthologous genes are connected by black lines.



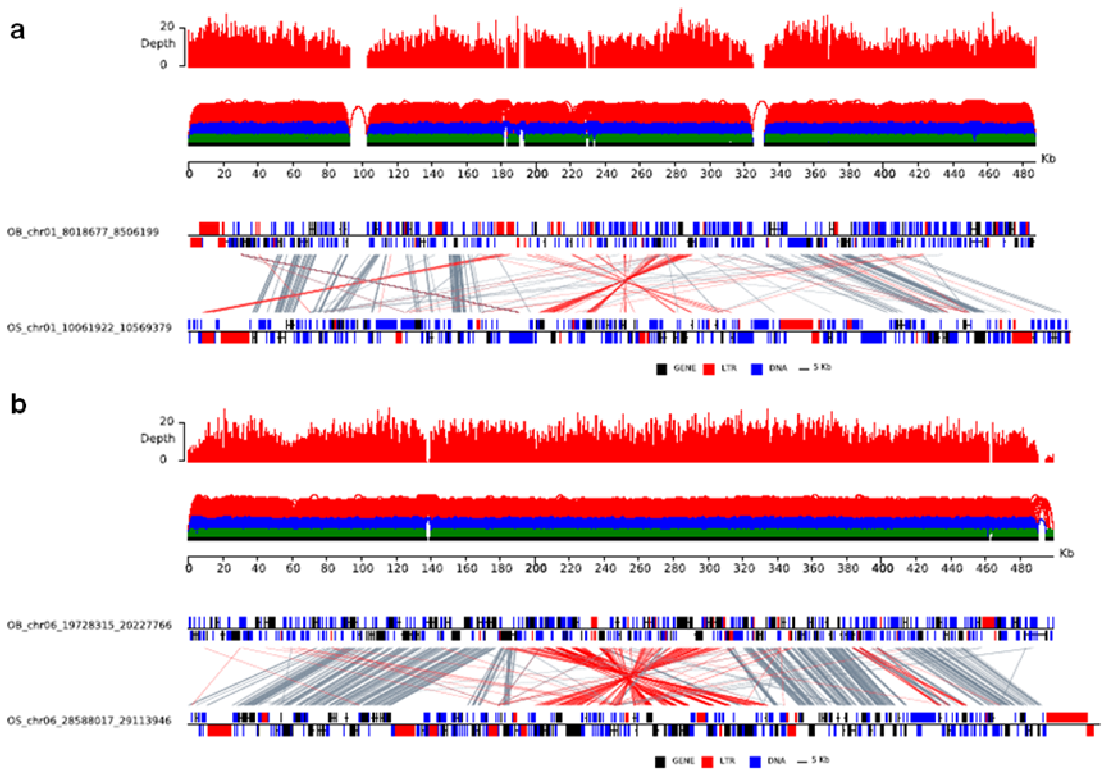
Supplementary Figure S9 Comparisons of chromosomal distribution of RLK-LRR genes between *O. brachyantha* and *O. sativa*. For each chromosome pair, the left one represents the chromosome of *O. sativa* and the right one represents the chromosome of *O. brachyantha*. Centromeres are indicated by solid black circles. Genes are indicated by colored circles; Tandemly duplicated genes are clustered to one representive position; Blue circles stand for genes of *O. sativa*; Red circles stand for genes of *O. brachyantha*; Green circles stand for non-collinear genes; Orthologous genes or gene clusters are connected by black lines.



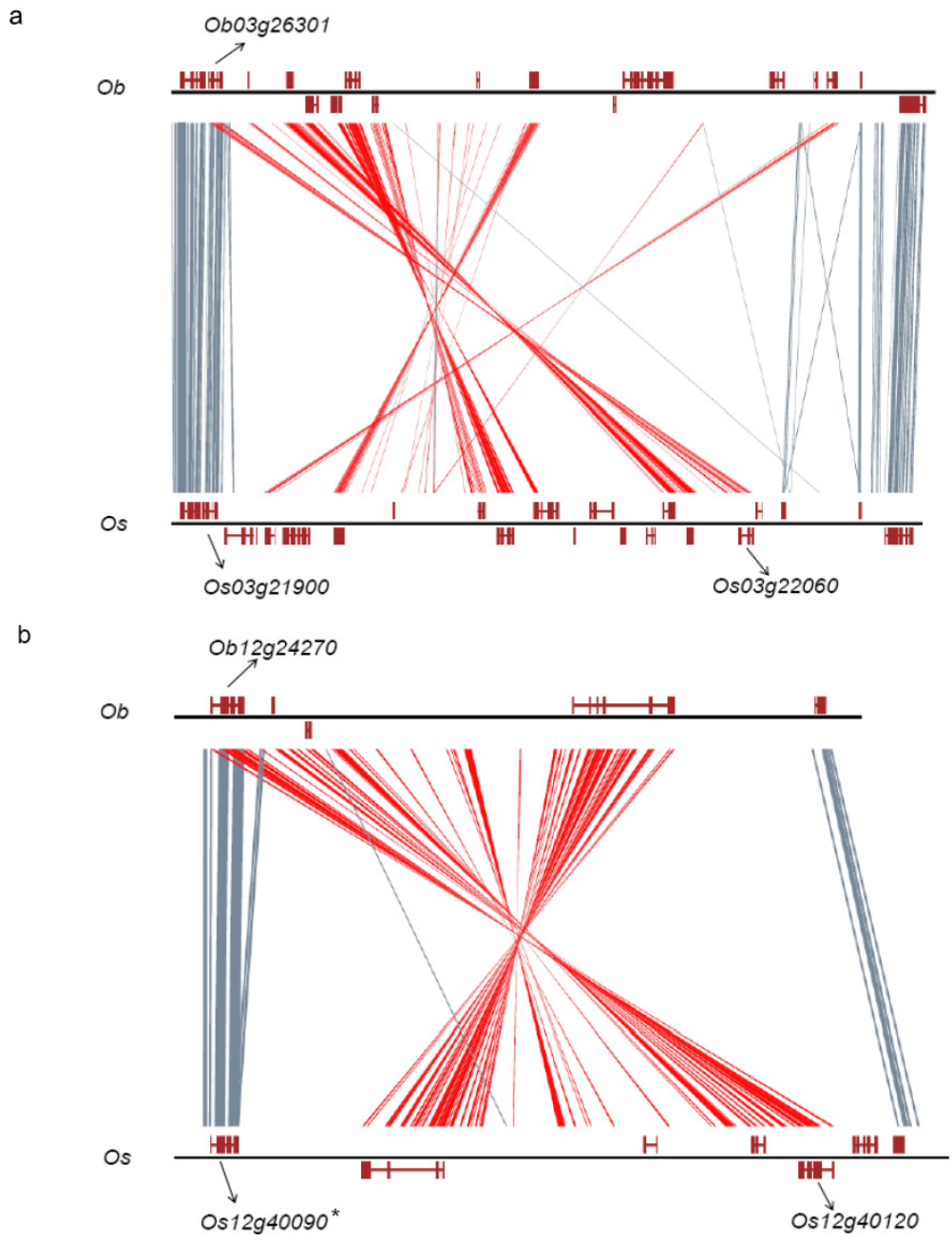
Supplementary Figure S10 Examples of RLK-LRR type pseudogenes in *O. brachyantha* (up) and *O. sativa* (down). We compared the gene structures of orthologous gene clusters to identify pseudogenes in both genomes. Only genes with exons interrunpted by frameshift mutations or transposable element insertions were defined as pseudogenes. The pseudogenezation may cause lose of function of resistense genes in these loci in one or both genomes, which contributed to the divergence of *Oryza* species. Black arrowhead boxes, genes; Shaded arrowhead boxes, pseudogenes; Green arrowhead boxes, non-collinear genes; blue boxes, transposable elements; MULE, Mutator-like element; LTR, long terminal repeat retrotransposon; Target site duplication (TSD) are in red; Orthologous genes are connected by black lines.



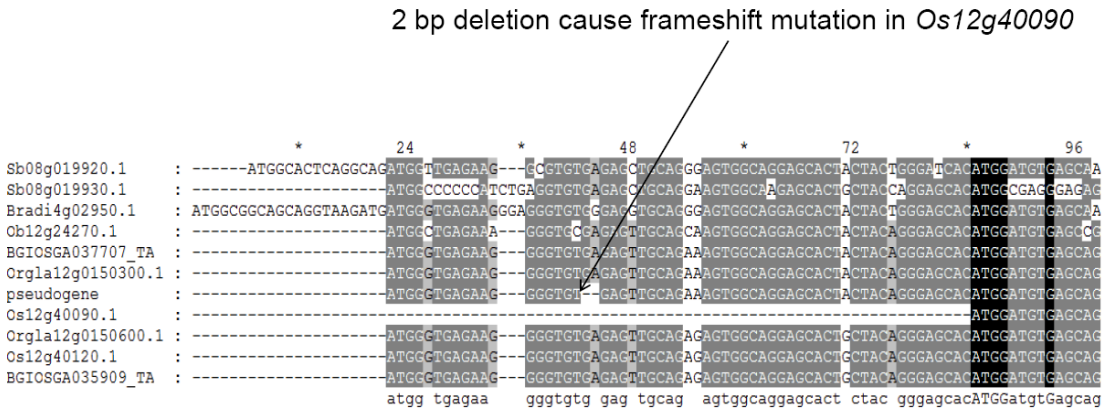
Supplementary Figure S11 Distribution of inversions between *O. brachyantha* and *O. sativa*. For each chromosome pair, the left one represents the chromosome of *O. sativa* and the right one represents *O. brachyantha*. Centromeres are indicated by solid black circles. Inversions are defined as described in Method and shaded by organge blocks.



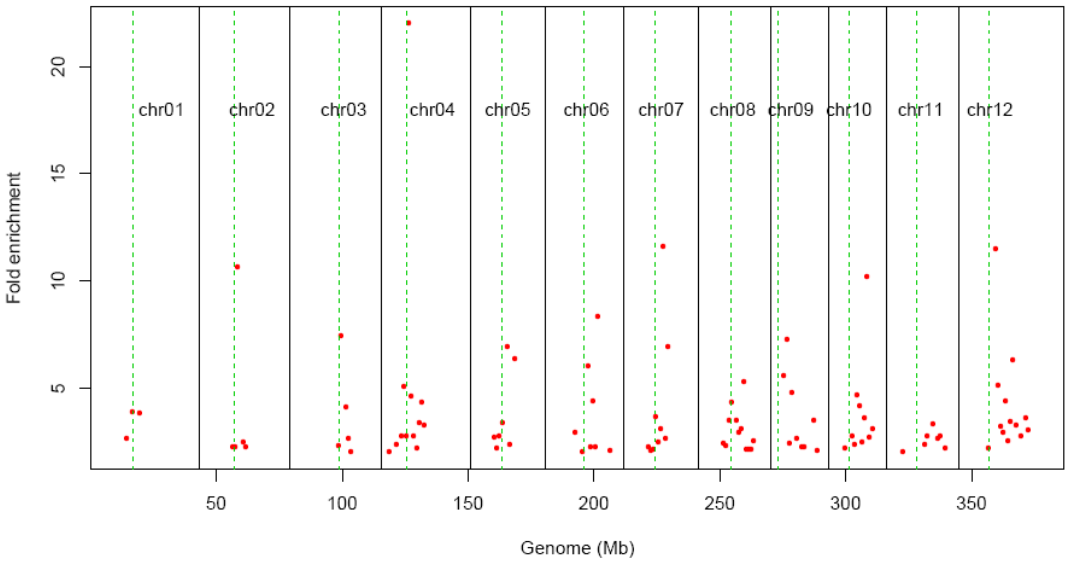
Supplementary Figure S12 The structure variations (inversions) in *O. brachyantha* supported by paired-end mapping and read depth. (**a**). OB_chr01_8018677_8506199 (top) vs OS_chr01_10061922_10569379 (bottom); (**b**). OB_chr06_19728315_20227766 (top) vs OS_chr06_28588017_29113946 (bottom). Colored arcs represented paired-end libraries of 10 kb (red), 5 kb (blue) and 2 kb (green). Read depth was calculated by 100 bp windows.



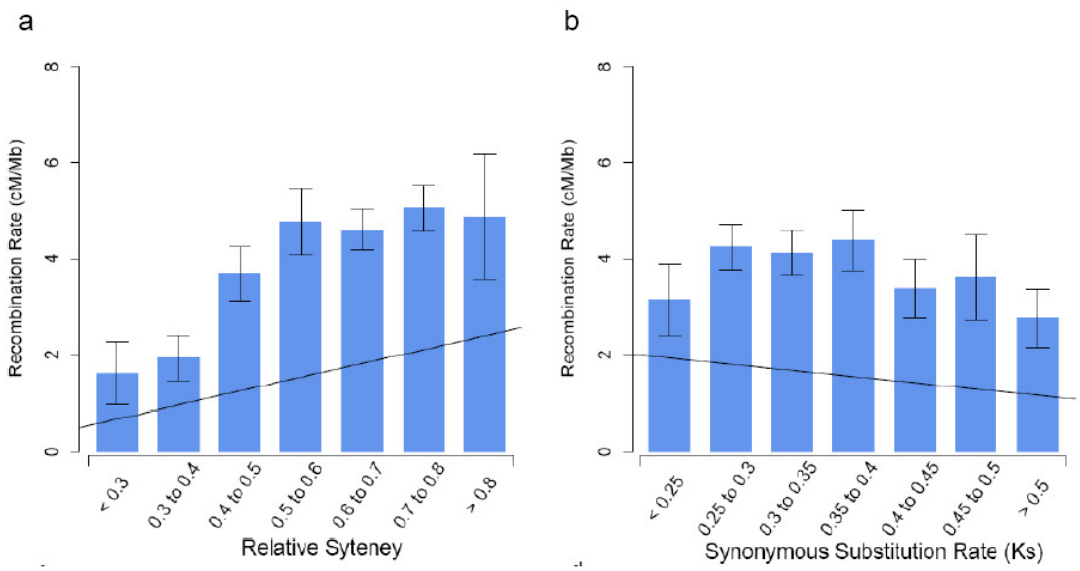
Supplementary Figure S13 Gene duplications associated with inversions in *O. sativa*. Inversions and the flanking collinear genes were compared between *O. brachyantha* (*Ob*) and *O. sativa* (*Os*). Homologous sequences are connected by grey lines if the matching sequences are in direct orientation, or by red lines if the matching sequences are in reverse orientation. Pseudogenes are indicated by stars following the locus name. (**a**), The inversion is associated with the duplication of gene *Os03g21900* in *O. sativa*. Although the duplicated genes are conserved in gene structure (data not shown), their expression patterns were diverged (Supplementary Table S10). (**b**), The inversion is associated with the duplication of gene *Os12g40090* in *O. sativa*. By comparing the gene duplications with their orthologs in related species, we found gene *Os12g40090*, which is the orginal copy of the duplicated gene *Os12g40120*, was pseudogenized due to a 2 bp deletion (Supplementary Fig. S14). The gene expression data also supports the pseudogenization of *Os12g40090* (Supplementary Table S10).



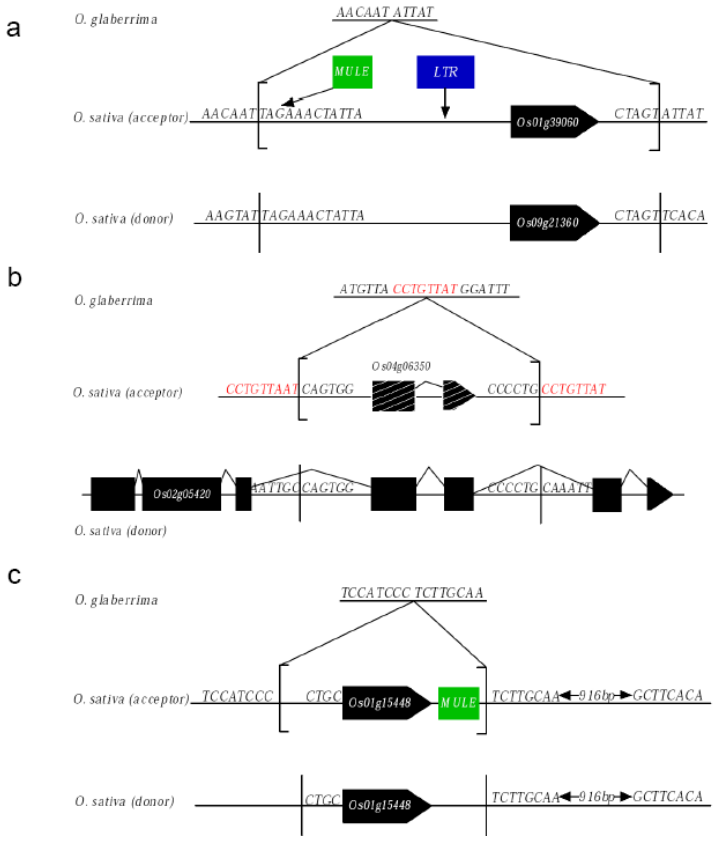
Supplementary Figure S14 Pseudogenization of the parent gene of the inversion-associated duplicated gene in *O. sativa*. Protein-coding sequences of the inversion-associated duplicated gene *Os12g40120* and the parent gene *Os12g40090* (pseudogene) were aligned with their orthologs in *indica* rice (*BGIOSGA037707_TA*, *BGIOSGA035909_TA*), *O. glaberrima* (*Orgla12g0150300.1*, *Orgla12g0150600.1*), *O. brachyantha* (*Ob12g24270.1*), *Brachypodium distachyon* (*Bradi4g02950.1*) and *Sorghun bicolor* (*Sb08019920.1*, *Sb08g019930.1*). The translation start site of orthologous genes are conserved except for *Os12g40090.1*, which have a 2 bp deletion leading to the pseudogenization of this gene.



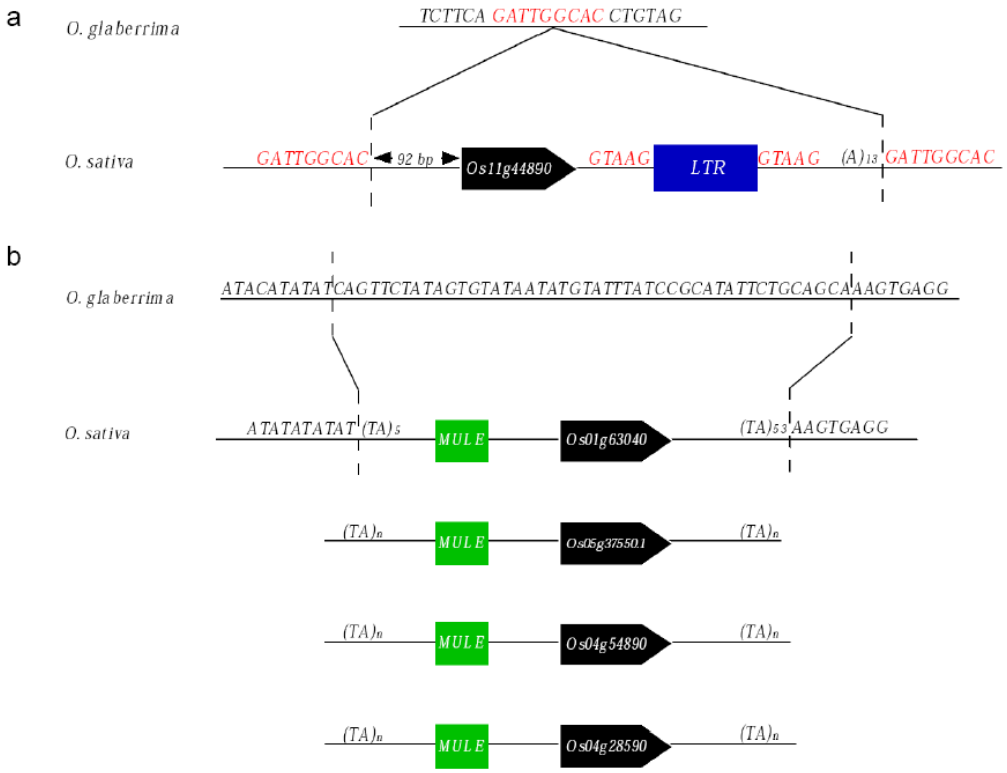
Supplementary Figure S15 Fold enrichment of non-collinear genes in *O. sativa*. The number of collinear genes and evidence-based non-collinear genes were counted in every 1 Mb window along the chromosomes of *O. sativa*. The enrichment of non-collinear genes in each 1 Mb window was tested against the genome average number of collinear genes and evidenced non-collinear genes in 1 Mb by Chi-squared test. Fold enrichments were calculated by the ratio between the proportion of non-collinear genes in each 1 Mb window and the genome average proportion of non-collinear genes in 1 Mb. Only significant enriched regions are shown in this figure (P-value ≤ 0.05). Dashed green lines vertical to the x-axis indicate the position of centromere for each chromosome. The chromosomes are seperated by black lines vertical to the x-axis.



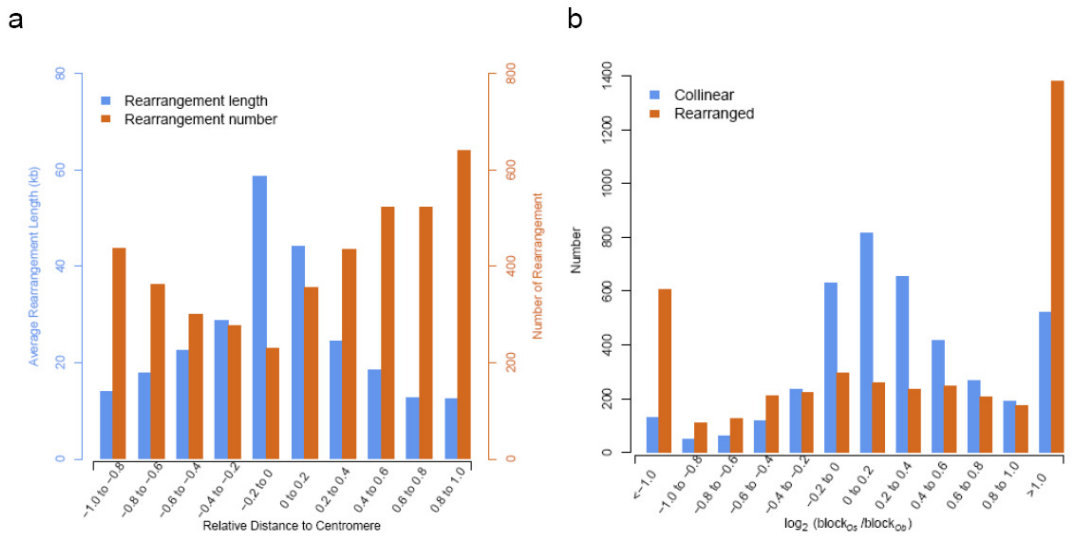
Supplementary Figure S16 Synteny along the chromosomes between *O. brachyantha* and *O. sativa*. (**a**), Correlation of relative synteny with recombination rate (R = 0.39, *P*-value $\leq 2.2 \times 10^{-16}$). (**b**), Correlation of synonymous substitution rate (Ks) with recombination rate (R = -0.13, *P*-value ≤ 0.0007). The genetic and physical positions of genetic markers from Japanese Rice Genome Program (http://rgp.dna.affrc.go.jp) were obtained from Rice Genome Annotation Project (http://rice.plantbiology.msu.edu). Recombination rates along each chromosome were estimated by the ratio of genetic distance to physical distance in 1 Mb window with 0.5 Mb shift. Relative synteny was estimated by the ratio of non-collinear gene number to total gene number in 1 Mb window with 0.5 Mb shift. Average synonymous substitution rate (Ks) of collinear gene pairs in 1 Mb window with 0.5 Mb shift was estimated using Nei-Gojibori Method as implemented in PAML (http://abacus.gene.ucl.ac.uk/software/paml.html). The Pearson's correlation was estimated using R language (www.r-project.org).



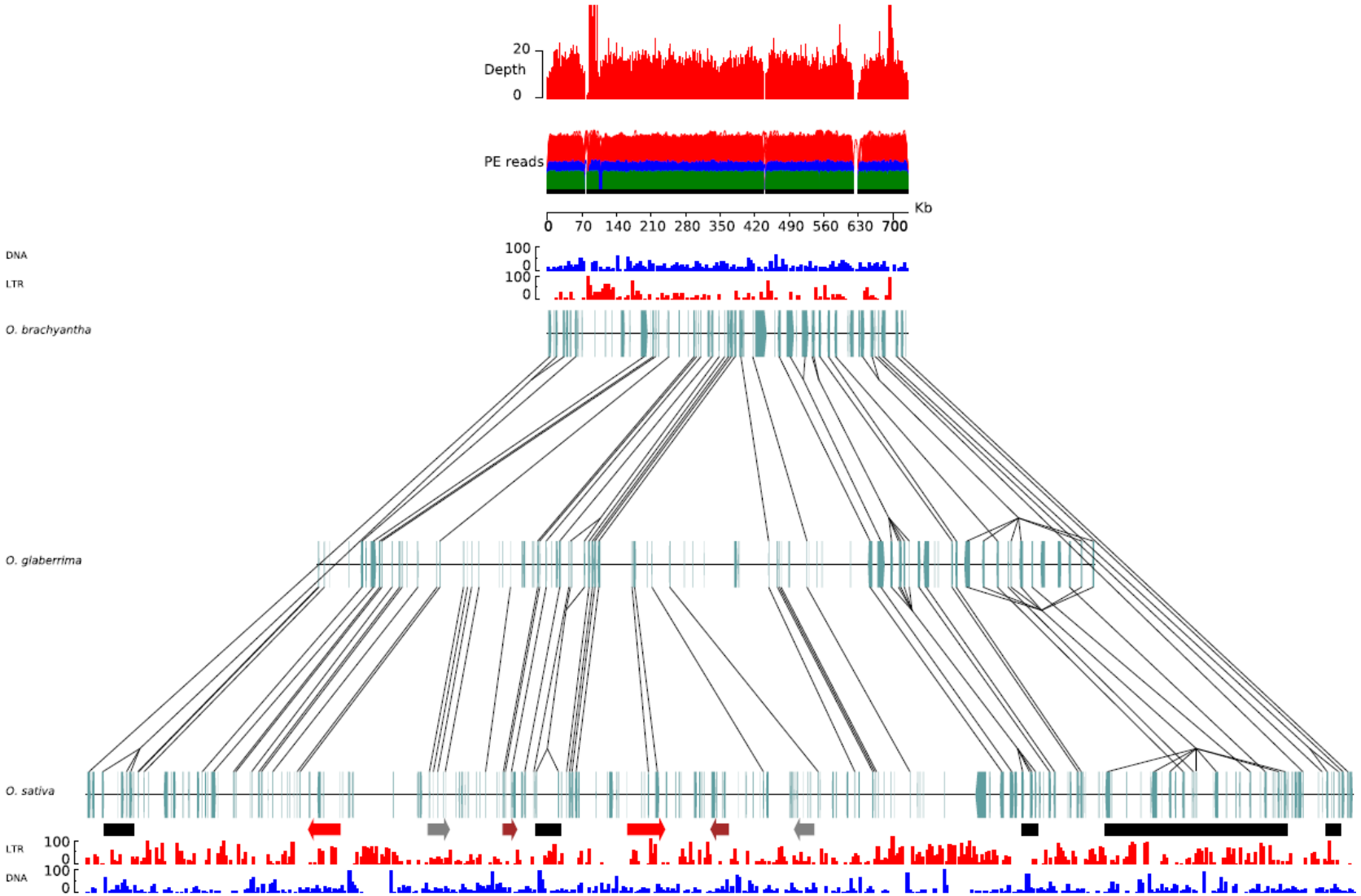
Supplementary Figure S17 Breakpoint signatures of non-collinear genes formed by sequence duplications. (**a**), Signatures of locus *Os01g39060* (8-AMINO-7-OXONONANOATE SYNTHASE), which is shown in Figure 5a and Figure 5b I. Transposable elements were inserted after the duplication. (**b**), Signatures of locus *Os04g06350*, which is shown in Figure 5b II. Only two exons were duplicated to the acceptor site, which may form a pseudogene at the duplicated locus. (**c**), Signatures of locus *Os01g15448*, which is shown in Figure 5b III. The 932-base-pair sequence downstream of the duplicated region on the acceptor site is highly homologous to the donor site and possibly serves as the template in the repair of double-strand breaks. Red characters, target site duplication (TSD); black arrowhead boxes, genes; shaded black arrowhead boxes, pseudogenes; blue boxes, LTR retrotransposons; green boxes, DNA transposons; orange boxes, non-LTR retrotransposons; sequences between square brackets on acceptor sequences indicate the duplicated sequences. Sequences between black lines on donor sequences show the candidate for donors of duplicated sequences.



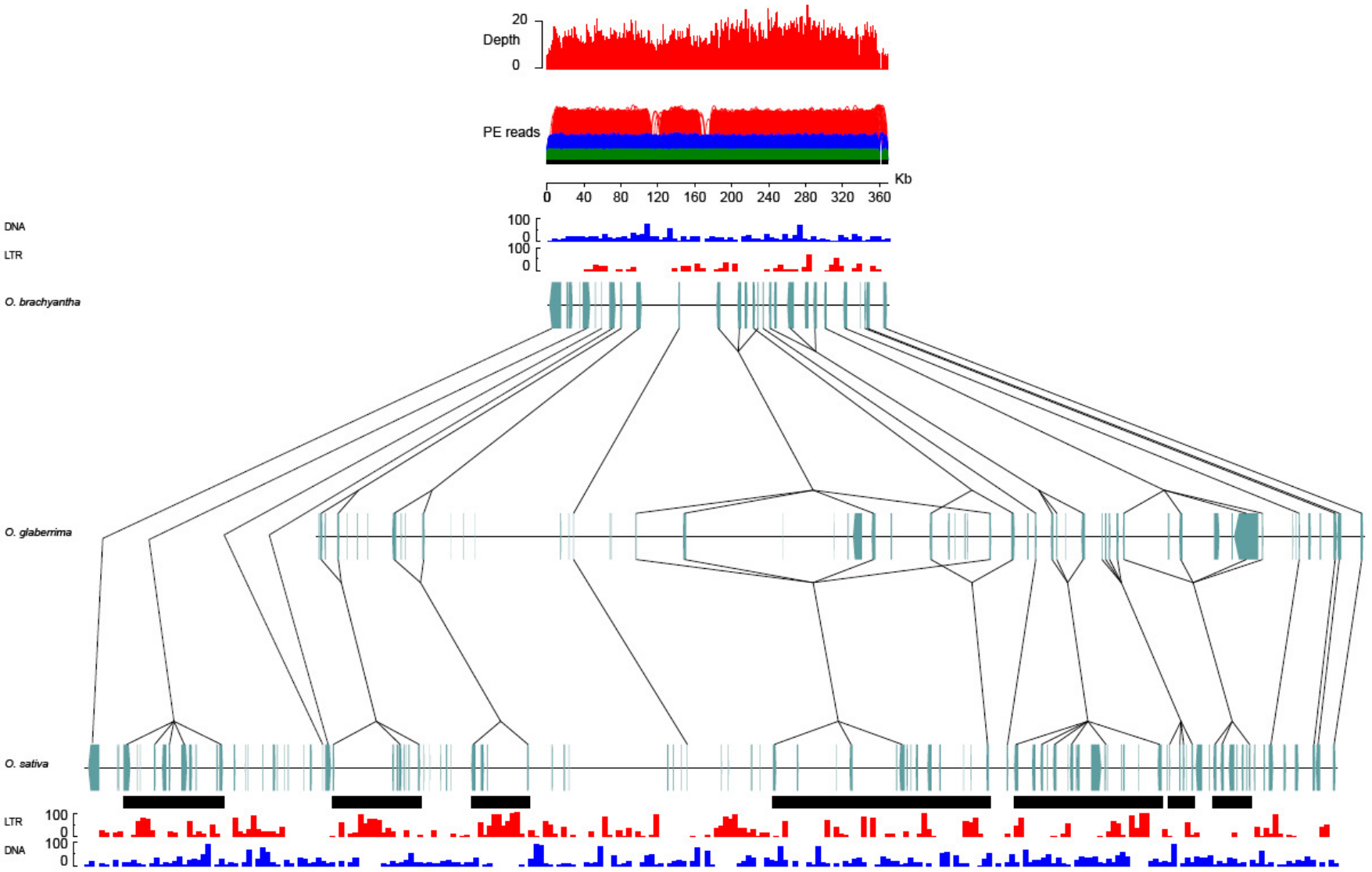
Supplementary Figure S18 Breakpoint signatures of non-collinear genes formed by sequence duplications without identified donor sequences. (**a**), Signatures of locus *Os11g44890* (Zinc knuckle family protein), which is shown in Figure 5b V. The insertion contains a single-exon gene and have poly-A tracts immediately adjacent to the target site duplication (TSD), suggesting the duplication was possibly formed by retroposition. (**b**), Signatures of locus *Os01g63040* (Ankyrin-like protein), which is shown in Figure 5b IV. Multi-copies of this locus were found in the rice genome. The present of TA repeats flanking the insertions suggests that the duplication was possibly induced by the repair of double-strand breaks, which occurred frequently in tandemly repeated sequences.



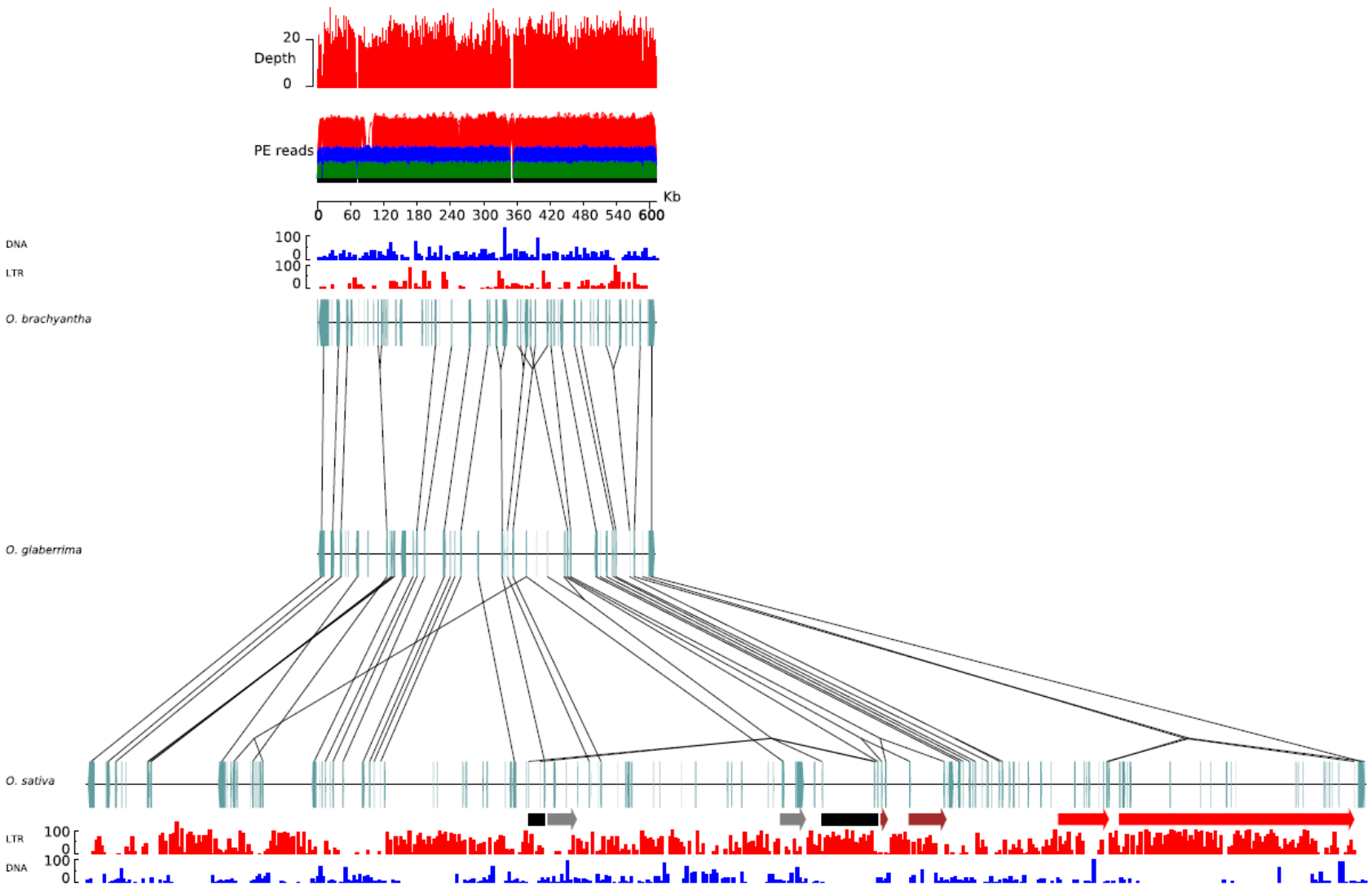
Supplementary Figure S19 The distribution of sequence rearrangements and their size differences compared with collinear regions between *O. brachyantha* and *O. sativa*. (**a**), The number and length of sequence rearrangements along the chromosomes. (**b**), Size comparison of collinear and rearranged regions between *O. brachyantha* and *O. sativa*. The collinear and rearrangement regions are defined as described in Method. The positions of sequence rearrangements on the chromosomes are scaled as relative distance to centromere, where negative values stand for regions on the short arms and positive values stand for the long arms. The size variations were calculated by log-ratio of block sizes between *O. brachyantha* and *O. sativa*. A positive value suggests expansion in *O. sativa*, whereas negative value suggests expansion in *O. brachyantha*.


20
Depth
0
PE reads
Kb
0 70 140 210 280 350 420 490 560 630 700
DNA
LTR
100
0
100
0
O. brachyantha
O. glaberrima
O. sativa
LTR
DNA
100
0
100
0

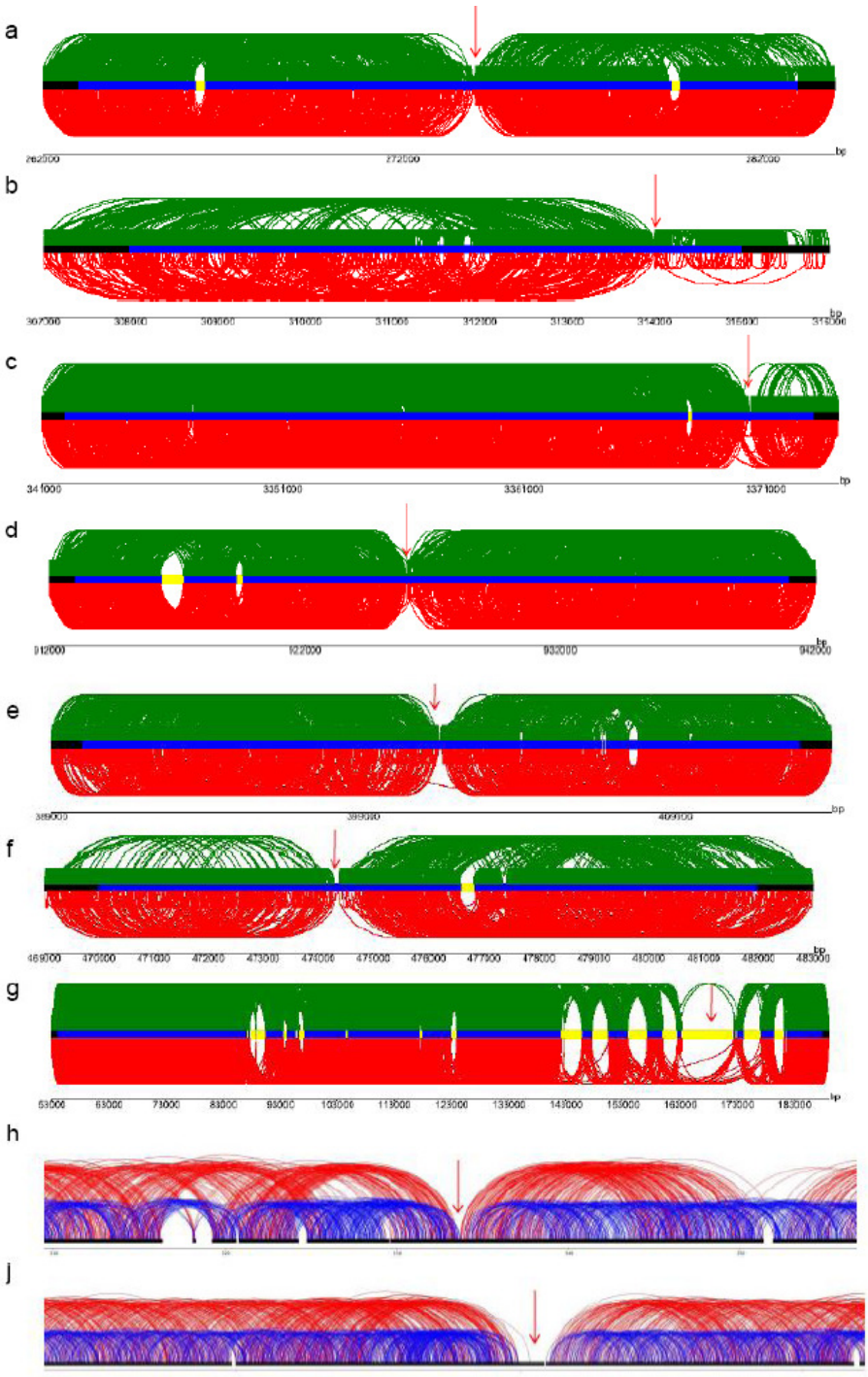
Supplementary Figure S20 Comparative analysis of the H7 heterochromatic domains among *O. brachyantha*, *O. glaberrima* and *O. sativa*. Genes are indicated by cadetblue boxes; orthologous genes or gene clusters are connected by black lines; Tandem gene clusters are highlighted by black rectangles; Organellar insertions are highlighted by yellow rectangles; Segmental duplications (SD) are highlighted by colored rectangles, each color representing one pair of SD. Red and black bars along the sequences illustrate the density of LTR retrotransposons and DNA transposons in 5 kb window. Colored arcs represented paired-end libraries of 10 kb (red), 5 kb (blue) and 2 kb (green). Read depth was calculated by 100 bp windows.


20
Depth
0
PE reads
Kb
0 40 80 120 160 200 240 280 320 360
DNA
LTR
100
0
100
0
O. brachyantha
O. glaberrima
O. sativa
LTR
DNA
100
0
100
0

Supplementary Figure S21 Comparative analysis of the H8 heterochromatic domains among *O. brachyantha*, *O. glaberrima* and *O. sativa*. Genes are indicated by cadetblue boxes; orthologous genes or gene clusters are connected by black lines; Tandem gene clusters are highlighted by black rectangles; Organellar insertions are highlighted by yellow rectangles; Segmental duplications (SD) are highlighted by colored rectangles, each color representing one pair of SD. Red and black bars along the sequences illustrate the density of LTR retrotransposons and DNA transposons in 5 kb window. Colored arcs represented paired-end libraries of 10 kb (red), 5 kb (blue) and 2 kb (green). Read depth was calculated by 100 bp windows.

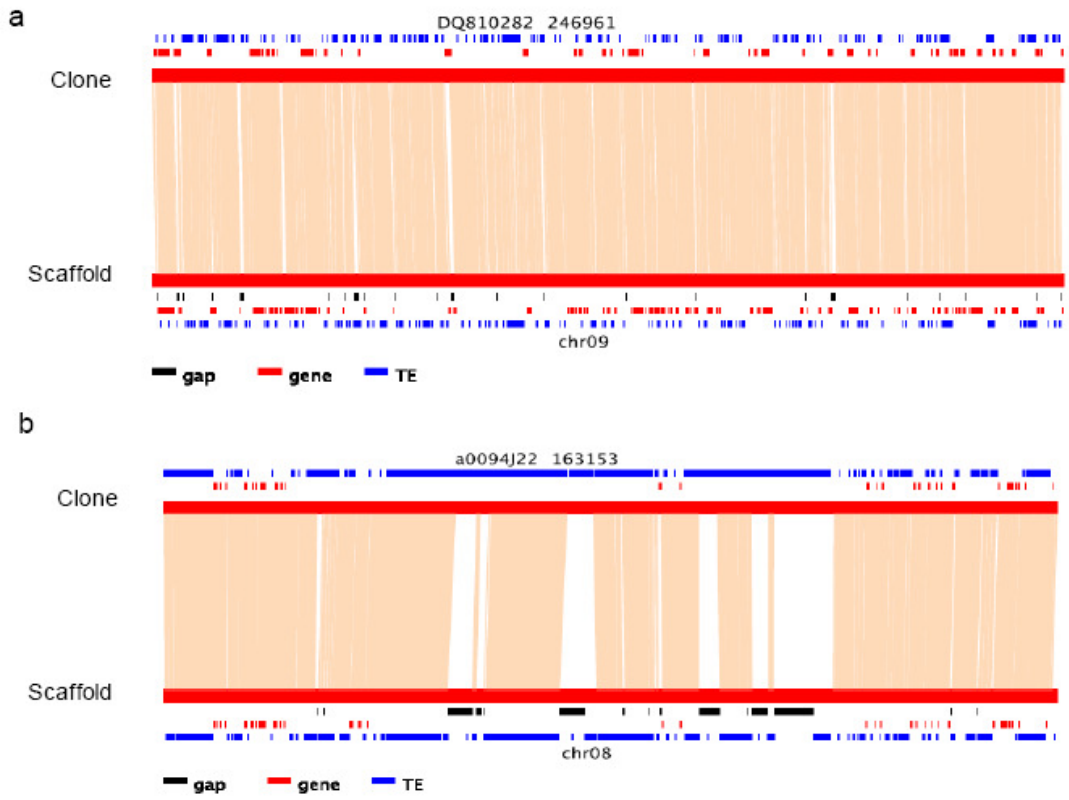

20
Depth
0
PE reads
Kb
0 60 120 180 240 300 360 420 480 540 600
DNA
LTR
O. brachyantha
O. glaberrima
O. sativa
LTR
DNA
100
0

Supplementary Figure S22 Comparative analysis of the H1 heterochromatic domains among *O. brachyantha*, *O. glaberrima* and *O. sativa*. Genes are indicated by cadetblue boxes; orthologous genes or gene clusters are connected by black lines; Tandem gene clusters are highlighted by black rectangles; Organellar insertions are highlighted by yellow rectangles; Segmental duplications (SD) are highlighted by colored rectangles, each color representing one pair of SD. Red and black bars along the sequences illustrate the density of LTR retrotransposons and DNA transposons in 5 kb window. Colored arcs represented paired-end libraries of 10 kb (red), 5 kb (blue) and 2 kb (green). Read depth was calculated by 100 bp windows.

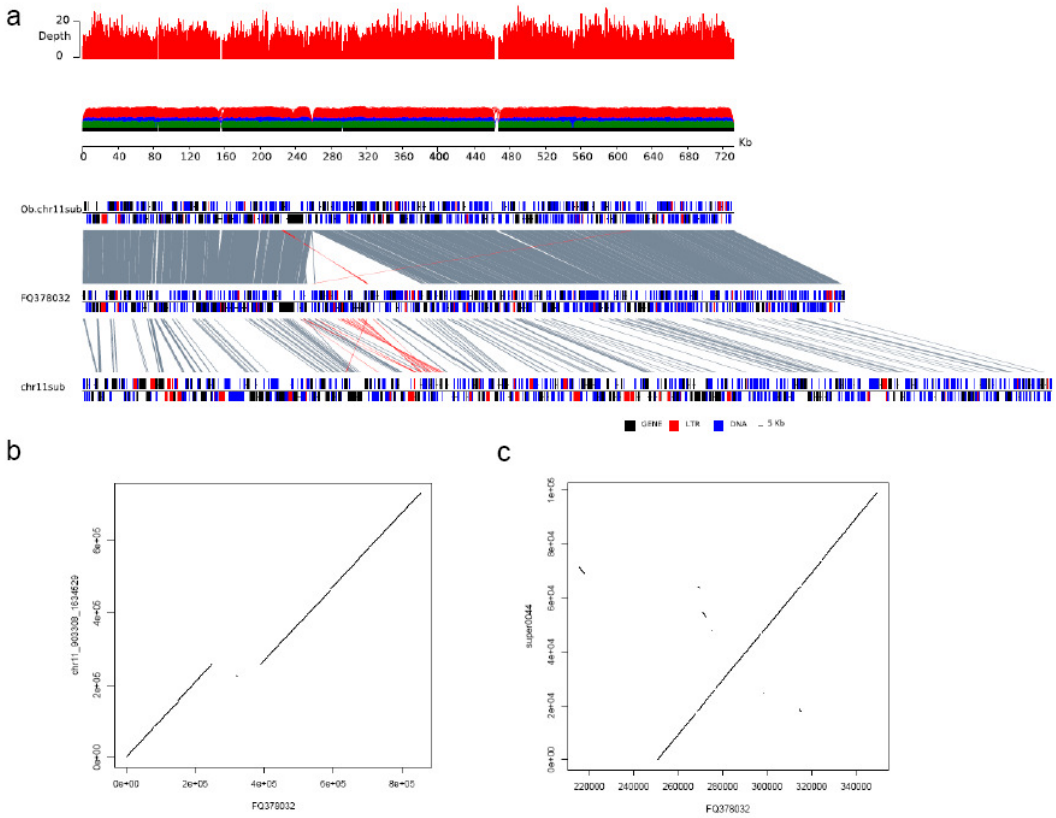


Supplementary Figure S23 Validation of sequence assemblies which are inconsistent with the physical map. The locations of uniquely mapped pair-end reads described in Supplementary Table S1 were drawn on scaffolds with inconsistency between sequence assemblies and physical map. Pair-end reads with unexpected orientation or distance were drawn in red, except for h-j which showed only correctly mapped pair-end reads. Misassembled points are indicated by red arrows. The gap regions on

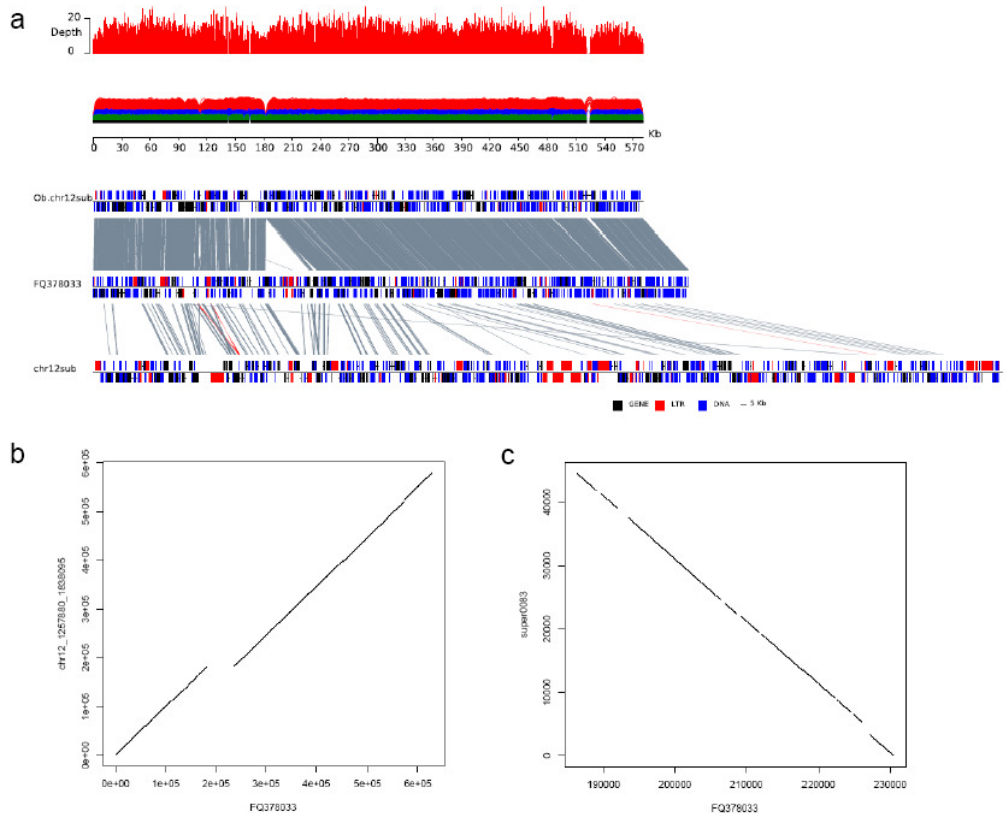
scaffolds are shown by yellow color (a-g) or left blank (h-j). The lack of pair-end reads covering these regions suggests that the nine inconsistencies were caused by misassemblies in sequence assembly.



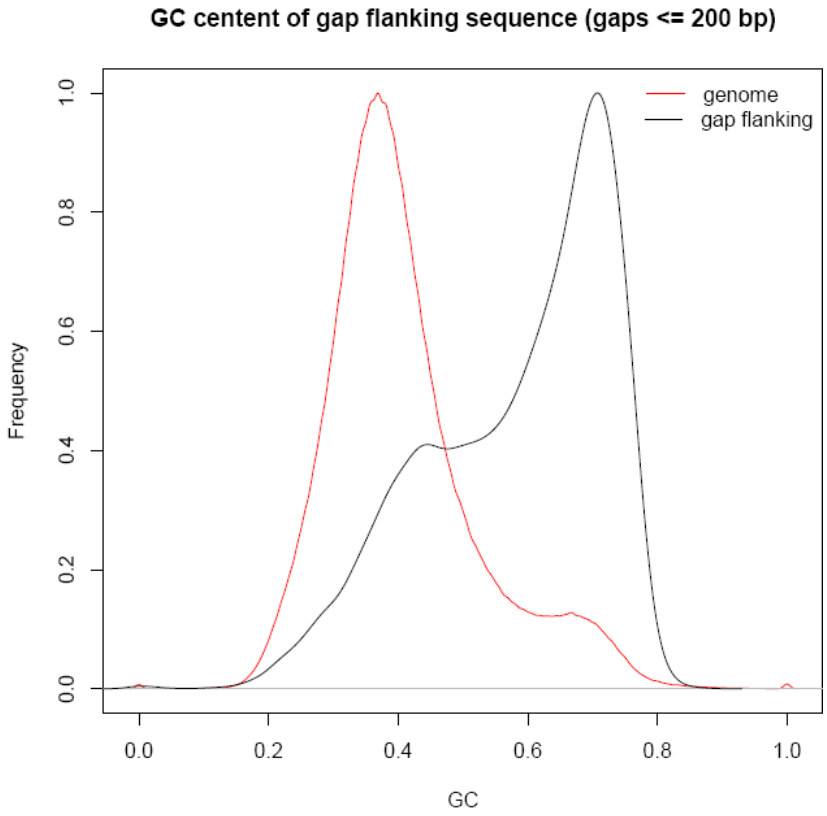
Supplementary Figure S24 Comparison of sequence assembly with BAC clones sequenced by Sanger technology. Scaffold sequences were aligned with BAC sequences using BLASTN with an E-value of 1×10^{-10}. High-scoring segment pairs (HSP) longer than 200 bp are connected by pink lines. Genes, transposable elements (TE), and sequence gaps are indicated by red, blue and black boxes, respectively. The sequence gaps in the euchromatic regions tend to be small and are located in high GC regions (Supplementary Fig. S27). In pericentromeric regions, these large gaps are located on repetitive sequences. (**a**), BAC clone a0002C07 located in the euchromatic region (Accession: DQ810282). (**b**), BAC clone a0094J22 located in the pericentromeric region.



Supplementary Figure S25 The quality of the assembled sequence corresponding to chromosome 11 segmental duplication region (FQ378032) assessed by paired-end mapping and read depth. (**a**), FQ378032 (middle) was compared with the corresponding sequence in *O. brachyantha* (Ob.chr11sub, top) and *O. sativa* (chr11sub, bottom). Colored arcs represented paired-end libraries of 10 kb (red), 5 kb (blue) and 2 kb (green). Read depth was calculated by 100 bp windows. (**b**), Dotplot of FQ378032 with corresponding sequence in *O. brachyantha*. (**c**), Dotplot of FQ378032 with the unanchored superscaffold (super0044).



Supplementary Figure S26 The quality of the assembled sequence corresponding to chromosome 12 segmental duplication region (FQ378033) assessed by paired-end mapping and read depth. (**a**), FQ378033 (middle) was compared with the corresponding sequence in *O. brachyantha* (Ob.chr12sub, top) and *O. sativa* (chr12sub, bottom). Colored arcs represented paired-end libraries of 10 kb (red), 5 kb (blue) and 2 kb (green). Read depth was calculated by 100 bp windows. (**b**), Dotplot of FQ378033 with corresponding sequence in *O. brachyantha*. (**c**), Dotplot of FQ378033 with the unanchored superscaffold (super0083).



Supplementary Figure S27 GC content of 200 bp flanking sequence of small gaps (≤200 bp).

Flanking sequences of 200 bp of the small gaps were retrieved from the scaffold sequences. The genome of *Oryza brachyantha* was shuffled into 200 bp fragments to use as control for genome average GC content. The high GC content peak of gap flanking sequences at 75-80% suggests that a large proportion of the small gaps were failed to close due to low sequence coverage for high GC regions.

Supplementary Tables

Supplementary Table S1 Libraries used for the sequence assembly

Sequence Data	Insert Size	Total Data (Gb)	Read Length (bp)	Sequence Coverage (X)	Physical Coverage (X)
Solexa Reads	200bp	7.56	75, 100	25.20	21.94
	350bp	8.67	75	28.88	75.65
	650bp	5.04	75	16.80	71.90
	2kb	3.65	44	12.17	300.48
	5kb	4.13	44	13.77	743.36
	10kb	2.16	44	7.20	854.08
BAC end sequences	130kb	0.045	672	0.15	14
Total		31.21		104.02	2067.42

Supplementary Table S2 Summary on statistics of the sequence assembly

	Contig		Scaffold		Super-Scaffold	
	Size(bp)	Number	Size(bp)	Number	Size(bp)	Number
N90	5992	11898	158856	332	263373	195
N80	9671	8752	308860	217	529280	126
N70	13041	6597	459159	149	888938	87
N60	16468	4933	683080	101	1263858	63
N50	20372	3603	1013071	70	1612996	45
Longest	172014		4900882		8861877	
Total Size	2.43E+08		2.63E+08		2.61E+08	
Total Number (>100bp)		28981		7999		7803
Total Number (>2kb)		16716		850		654

Supplementary Table S3 Summary of the transposable elements in the *O. brachyantha* genome

Superfamily	Copy number ($\times 10^3$)	Occupied size (bp)	Percentage of genome (%)[a]
Class I	**52.8**	**25469313**	**10.47**
LTR retrotransposon	32.8	21271582	8.74
Ty1-copia	12.0	7522651	3.09
Ty3-gypsy	10.8	11020417	4.53
Unclassified	3.6	738779	0.30
Other	6.4	1989735	0.82
Non-LTR retrotransposon	20.0	4197731	1.73
LINEs	19.1	4083483	1.69
SINES	0.9	114248	0.05
Class II	**167.8**	**45492659**	**18.70**
Mutator	70.4	18299216	7.52
hAT	26.8	7827987	3.22
CACTA	8.4	3883302	1.60
PIF/Harbinger	14.3	3190864	1.31
Tc1/Mariner	3.9	667655	0.27
Helitron	13.9	2817571	1.16
MITEs	44.0	8806064	3.62
Tourist	23.1	5200414	2.14
Stowaway	20.9	3605650	1.48
Total TEs	**220.6**	**70961972**	**29.17**

[a] Percentage was estimated based on the gap-free genome size of 243 Mb.

Supplementary Table S4 Summary of the solo-LTR families in *O. brachyantha* and *O. sativa*

Retroelements in *O. brachyantha*							Homologous elements in *O. sativa*				
Name	Copy number			Coverage (bp)	Name	Superfamilies	Copy number				Coverage (bp)
	Solo-LTR	Fragment	All				Element	Solo-LTR	Fragment	All	
FRetro143	1	112	113	31448	NRetro143	Ty3-gypsy	42	85	341	468	245282
FRetro144	1	46	47	25766	NRetro144	Ty3-gypsy	24	14	1109	1147	585867
FRetro146	1	100	101	34792	NRetro146	unclassified	9	330	1003	1342	2390127
FRetro147	10	154	164	84201	NRetro147	Ty3-gypsy	3	102	3467	3572	5261482
FRetro149	3	9	12	3216	NRetro149	Ty1-copia	9	2	302	313	372090
FRetro150	16	172	168	50477	NRetro150	Ty3-gypsy	27	108	713	848	526232
FRetro151	20	803	823	378278	NRetro151			5	634	639	205098
FRetro152	1	45	46	19579	NRetro152			13	561	574	430086
FRetro153	4	9	13	4234	NRetro153	Ty1-copia	1		26	27	16667
FRetro155	2	15	17	4167	NRetro155	Ty1-copia	24	53	233	310	138149
FRetro156	1	9	10	930	NRetro156	Ty1-copia	4	2	287	293	296999
FRetro158	4	27	31	11336	NRetro158			10	48	58	28955
FRetro159	7	214	221	47875	NRetro159-houba	Ty1-copia	147	31	823	1001	1644839
FRetro160	1	52	53	20260	NRetro160	Ty3-gypsy	1	6	105	112	64466
FRetro161	5	11	16	2996	NRetro161	Ty3-gypsy	3		478	481	124049
FRetro162	10	17	27	12842	NRetro162	Ty1-copia	1	6	37	44	28868
FRetro163	4	39	43	8106	NRetro163	Ty1-copia	8	16	418	442	304177
FRetro164	1	45	46	9359	NRetro164	Ty1-copia	56	15	255	326	709720
FRetro165	7	59	66	12593	NRetro165	Ty1-copia	1		95	96	36160
FRetro166	1	1430	1431	138249	NRetro166	Ty3-gypsy	24	225	4035	4284	2959968
FRetro168	5	91	96	43151	NRetro168	Ty1-copia	1	16	1333	1350	800201

FRetro170	7	60	67	20089	NRetro170	Ty3-gypsy	18	20	3266	3304	1706885
FRetro171	5	11	16	2511	NRetro171	Ty1-copia	1	7	14	22	10154
FRetro172	2	309	311	101602	NRetro172	Ty1-copia	3	355	2367	2725	2366958
FRetro173	20	131	151	91150	NRetro173			12	40	52	66215
FRetro174	19	216	235	154361	NRetro174			2	71	73	24677
FRetro175	1	55	56	19415	NRetro175-SZ5	Ty1-copia	102	282	1703	2087	2692753
FRetro176	20	222	242	130743	NRetro176			4	49	53	32604
FRetro177	7	33	40	8173	NRetro177	unclassified	1	12	31	44	11433
FRetro178	1	10	11	3671	NRetro178	Ty1-copia	1		26	27	17777
FRetro179	2	20	22	5188	NRetro179	Ty1-copia	1	9	5494	5504	1262851
FRetro180	5	21	26	6274	NRetro180	Ty3-gypsy	9	1	1579	1589	1672733
FRetro181	67	124	191	39628	NRetro181	Ty1-copia	20	105	340	465	376793
Total	261	4671	4912	1526660	Total		541	1848	31283	33672	27411315

Note: 1. No element was identified from the NRetro151,152,158,173, 174 and 176 families; 2. Each read was counted as one fragment, thus the copy number of the fragments and all the TE families were overestimated, because one TE copy can appear as multiple reads in the RepeatMasker program (http://www.repeatmasker.org/).

Supplementary Table S5 Number of R-genes characterized in *O. brachyantha* and *O. sativa*

Class	*O. brachyantha*	Tandem	Transpose	*O. sativa*	Tandem	Transpose
CC-NBS	26	13	1	42	20	10
CC-NBS-LRR	136	83	12	259	156	56
NBS	71	25	3	106	51	18
NBS-LRR	165	75	20	249	118	72
TIR-NBS	1	0	0	1	0	0
Total	399	196	36	657	345	156

Supplementary Table S6 Number of pseudogenes characterized in R-genes of *O. brachyantha* and *O. sativa*

	Locus	Class	Ortholog	Note
O. brachyantha	Ob01g52420.1	CC-NBS	NA	intact LTR (TSD)
	NA	NBS	LOC_Os02g39884.1	diverged
	Ob04g10550.1	NBS-LRR	LOC_Os04g02030.1	LINE
	NA	NBS	LOC_Os05g37260.1	diverged
	NA	NBS	LOC_Os05g45670.1	diverged
	Ob06g12760.1	NBS-LRR	LOC_Os06g05359.1	intact LTR (TSD)
	Ob06g33310.1	NBS	LOC_Os06g47800.1	solo LTR
	Ob06g34670.1	NBS	LOC_Os06g49390.1	MULE
	Ob11g17080.1	NBS	LOC_Os11g15670.1	frameshift
	Ob11g24380.1	CC-NBS-LRR	LOC_Os11g38520.1	DNA
	Ob12g20170.1	CC-NBS-LRR	LOC_Os12g29710.1	LTR
O. sativa	LOC_Os01g71106.2	CC-NBS	Ob01g52430.1	frameshift
	LOC_Os06g49360.1	NBS	Ob06g34670.1	intact LTR (TSD)
	LOC_Os11g15670.1	NBS	Ob11g17080.1	intact LTR (TSD)
	LOC_Os11g16510.1	NBS	Ob11g17240.1	diverged

Supplementary Table S7 Number of RLK-LRR genes characterized in *O. brachyantha* and *O. sativa*

Class	*O. brachyantha*	Tandem	Transpose	*O. sativa*	Tandem	Transpose
LRR-I	38	32	1	52	42	3
LRR-II	12	3	0	13	3	0
LRR-III	37	2	0	40	1	1
LRR-IV	3	0	0	3	0	0
LRR-IX	3	0	0	3	0	0
LRR-V	13	0	1	12	0	0
LRR-VI-1	3	1	0	3	0	0
LRR-VI-2	12	0	3	11	0	2
LRR-VII	9	1	0	9	1	0
LRR-VIII-1	10	2	1	10	2	1
LRR-VIII-2	34	21	0	38	24	1
LRR-XI	43	15	1	42	18	1
LRR-XII	72	58	5	125	88	24
LRR-XIIIa	3	0	0	3	0	0
LRR-XIIIb	3	0	0	3	0	0
LRR-XIV	3	1	0	3	1	0
LRR-XV	2	0	0	2	0	0
LRR-Xa	3	0	0	3	0	0
LRR-Xb	31	17	0	43	30	0
LRR_XVI	3	0	0	3	0	0
Total	337	153	12	421	210	33

Supplementary Table S8 Number of pseudogenes characterized in RLK-LRR genes of *O. brachyantha* and *O. sativa*

	Locus	Class	Ortholog	Note
O. brachyantha	Ob10g14970.1	LRR-XII	LOC_Os10g19160.1	MULE (TSD)
	NA	LRR-XI	LOC_Os02g12910.1	frameshift
	NA	LRR-XI	LOC_Os03g12730.1	frameshift
	Ob10g10640.1	LRR-Xb	LOC_Os10g02500.1	frameshift
	Ob10g20010.1	LRR-XII	LOC_Os10g32990.1	frameshift
	Ob11g13710.1	LRR-XII	LOC_Os11g07160.1	frameshift
	NA	LRR-XII	LOC_Os02g30540.1	frameshift
	Ob11g13750.1	LRR-XII	NA	frameshift
O. sativa	NA	LRR-I	Ob01g43490.1	solo LTR(TSD)
	LOC_Os04g15630.1	LRR-XII	Ob04g13750.1	intact LTR
	LOC_Os11g40810.1	LRR-XII	Ob11g25280.1	intact LTR (TSD)
	LOC_Os04g15560.1	LRR-XII	Ob04g13680.1	solo/intact LTR (TSD)
	LOC_Os06g44430.1	LRR-VIII-2	Ob06g30980.1	DNA
	LOC_Os06g38640.1	LRR-XII	NA	frameshift
	LOC_Os06g38730.1	LRR-XII	NA	frameshift
	LOC_Os11g40890.1	LRR-XII	NA	frameshift
	LOC_Os11g36200.1	LRR-XII	NA	frameshift
	LOC_Os11g07124.1	LRR-XII	NA	frameshift
	NA	LRR-VI-2	Ob04g29710.1	frameshift

Supplementary Table S9 Functional enrichment of GO category in tandemly duplicated genes of *O. brachyantha* and *O. sativa*

Genome	GO-ID	*P*-value	Corrected *P*-value	Annotation
O. sativa	GO:0009607	4.70E-05	8.71E-04	response to biotic stimulus
	GO:0007154	5.07E-16	1.48E-14	cell communication
	GO:0048610	1.16E-18	3.72E-17	cellular process involved in reproduction
	GO:0002252	1.70E-04	2.07E-03	immune effector process
	GO:0008037	1.16E-18	3.72E-17	cell recognition
	GO:0051707	2.99E-05	5.71E-04	response to other organism
	GO:0006950	3.37E-33	3.44E-31	response to stress
	GO:0009856	1.16E-18	3.72E-17	pollination
	GO:0008219	1.31E-52	2.01E-50	cell death
O. brachyantha	GO:0071554	8.54E-04	3.38E-02	cell wall organization or biogenesis
	GO:0008219	4.21E-05	2.69E-03	cell death
	GO:0055114	7.24E-07	1.39E-04	oxidation-reduction process

Supplementary Table S10 Expression divergence of inversion-associated duplicated genes

Rice MPSS[a]	Os12g40090	Os12g40120	Description
NGD	0	17	10 days - Germinating seedlings grown in dark
XS06	0	4	Leaves collected from 2 months old Nipponbare Xa21 plants 6hr after Xanthamonas Oryzae inoculation showing susceptible reaction
MS48	0	7	Leaves collected from 3 weeks old M.grisea treated Nipponbare at 48hr
Rice MPSS[a]	Os03g21900	Os03g22060	Description
NSL	5	0	Young leaves stressed in 250 mM NaCl for 24h
MS03	14	0	Leaves collected from 3 weeks old M.grisea treated Nipponbare at 3hr
MC00	33	0	Leaves collected from 3 weeks old water treated Nipponbare at 0hr
NGD	0	2	10 days - Germinating seedlings grown in dark
NYL	0	10	Young leaves
NOS	0	2	Ovary and mature stigma
NDR	0	36	Young roots stressed in drought for 5 days
XC00	0	4	Leaves collected from 2 months old untreated Control-Nipponbare-Xa21 at 0hr
XR24	0	5	Leaves collected from 2 months old Nipponbare Xa21 plants 24hr after Xanthomonas oryzae
XR48	0	2	Leaves collected from 2 months old X.oryzae Nipponbare-Xa21 at 48hr
XS03	0	19	Leaves collected from 2 months old Nipponbare Xa21 plants 3hr after Xanthamonas Oryzae
MR24	0	1	Leaves collected from 3 weeks old M.grisea treated Nipponbare-Pi9 at 24hr
MR48	0	25	Leaves collected from 3 weeks old M.grisea treated Nipponbare-Pi9 at 48hr
MS12	0	2	Leaves collected from 3 weeks old M.grisea treated Nipponbare at 12hr
MS48	0	2	Leaves collected from 3 weeks old M.grisea treated Nipponbare at 48hr
PSI	0	3	6 days old developing seeds from Ilpumbyeo (high taste quality)
PSN	0	6	6 days old developing seeds from Nipponbare (control)

PLC	0	12	Nipponbare leaves collected 24 hrs after mechanical damage (control)

[a] Gene expressions of duplicated genes were obtained from Rice MPSS database (http://mpss.udel.edu)

Supplementary Table S11 Identification of non-collinear genes in *O. sativa* and *O. brachyantha*

	O. sativa	*O. brachyantha*
Protein-coding genes	41404	32038
Genes without synteny status	215	2484
Collinear genes	24103	22405
Non-collinear genes	17086	7149
-Not evidence based	7315	1736
--Collinear in grasses	76	15
-Evidence based	9771	5413
--Collinear in grasses	1213	339
--Evidence based non-collinear genes	8558	5074

Supplementary Table S12 Sequence signatures of breakpoints of duplicated sequence pairs

Acceptor	Donor	TSD	5'homology	3'homology	mechanism	Annotation
LOC_Os01g05770.1	LOC_Os06g40420	15bp	/	8bp	NHEJ	expressed protein
LOC_Os01g10160.1	LOC_Os04g57250	8bp	5bp	0bp	NHEJ	latency associated nuclear antigen,putative
LOC_Os01g10240.1	LOC_Os07g37530	8bp	/	/	TE flanking	urate anion exchanger,putative,expressed
LOC_Os01g15448.1	LOC_Os06g08510	0bp	0bp	>200bp	NAHR	expressed protein (DPL)
LOC_Os01g38100.1	LOC_Os05g49164	2bp	1bp	0bp	NHEJ	expressed protein
LOC_Os01g39060.1	LOC_Os09g21360	0bp	0bp	0bp	NHEJ	expressed protein (8-AMINO-7-OXONONANOATE SYNTHASE)
LOC_Os01g39120.1	LOC_Os10g32780	8bp	/	/	TE/TIR	RING zinc finger ankyrin protein,putative
LOC_Os01g42170.1	LOC_Os09g35530	(AT)n	(AT)n	0bp	NHEJ	zinc knuckle family protein
LOC_Os01g57250.1	LOC_Os03g07270	/	(AG)3	(T)5	NHEJ	expressed protein
LOC_Os01g65902.1	LOC_Os10g21290	0bp	3bp	2bp	NHEJ	apocytochrome f precursor,putative,expressed
LOC_Os03g07270.1	LOC_Os01g57250	13bp	1bp	2bp	NHEJ	glycine-rich cell wall protein,putative,expressed
LOC_Os03g21310.1	LOC_Os08g33280	0bp	>200 bp	0bp	NAHR	ulp1 protease family,C-terminal catalytic domain containing protein
LOC_Os03g32526.1	LOC_Os06g20500	0bp	0bp	3bp	NHEJ	tRNA-splicing endonuclease positive effector-related,putative
LOC_Os03g52230.1	LOC_Os10g10580	4bp	0bp	0bp	TE/TIR	dynamin-2B,putative
LOC_Os04g06350.1	LOC_Os02g05320				NHEJ	ANKYRIN REPEAT
LOC_Os04g09790.1	LOC_Os12g36960	0bp	0bp	2bp	TE flanking	spidroin-1,putative
LOC_Os04g38914.1	LOC_Os03g60639	0bp	1bp	2bp	NHEJ	expressed protein
LOC_Os04g54890.1	LOC_Os01g63040				NHEJ	
LOC_Os04g57250.1	LOC_Os01g10160	10bp	3bp	0bp	NHEJ	latency associated nuclear antigen,putative
LOC_Os05g01675.1	LOC_Os04g16760	0bp	0bp	3bp	NHEJ	photosystem I P700 chlorophyll a apoprotein A1,putative
LOC_Os05g03810.1	LOC_Os12g32130	8bp	0bp	2bp	NHEJ	trehalose phosphatase,putative
LOC_Os05g10570.1	LOC_Os07g15980	0bp	4bp	3bp	NHEJ	expressed protein
LOC_Os05g22712.1	LOC_Os10g21332	0bp	3bp	/	NHEJ	chloroplast S ribosomal protein S11,putative
LOC_Os05g22716.1	LOC_Os10g21336					translation initiation factor IF-1,chloroplast,putative
LOC_Os05g22718.1	LOC_Os10g21338					chloroplast S ribosomal protein S8,putative
LOC_Os05g22722.1	LOC_Os10g21342					chloroplast S ribosomal protein L14,putative
LOC_Os05g22724.1	LOC_Os10g21344					chloroplast S ribosomal protein L16,putative

LOC_Os05g50650.1	LOC_Os01g55010	11bp	0bp	1bp	TE/TIR	ligA,putative,expressed
LOC_Os06g06970.1	LOC_Os08g43200	7bp	/	/	TE flanking	AP2 domain containing protein,expressed
LOC_Os06g13780.1	LOC_Os04g32610	0bp	2bp	0bp	NHEJ	expressed protein
LOC_Os06g39700.1	LOC_Os10g21328	0bp	0bp	/	NHEJ	DNA-directed RNA polymerase subunit alpha,putative
LOC_Os06g39704.1	LOC_Os10g21322					photosystem II reaction center protein H,putative
LOC_Os06g39708.1	LOC_Os10g21310					photosystem II P680 chlorophyll A apoprotein,putative
LOC_Os06g39712.1	LOC_Os10g21300					OsClp10 - Putative Clp protease homologue,expressed
LOC_Os06g39716.1	LOC_Os10g21314					chloroplast S ribosomal protein S18,putative,expressed
LOC_Os06g39718.1	LOC_Os10g21312					chloroplast S ribosomal protein L33,putative,expressed
LOC_Os06g39722.1	LOC_Os10g21306					expressed protein
LOC_Os06g39728.1	LOC_Os10g21298					cytochrome b559 subunit alpha,putative
LOC_Os06g39738.1	LOC_Os10g21290					apocytochrome f precursor,putative
LOC_Os06g39740.1	LOC_Os10g21266					ATP synthase subunit beta,putative
LOC_Os06g39744.1	LOC_Os10g21282					photosystem I assembly protein ycf4,putative
LOC_Os06g39754.1	LOC_Os10g21268					expressed protein
LOC_Os06g39756.1	LOC_Os10g21264					ATP synthase epsilon chain,putative
LOC_Os07g14470.1	LOC_Os10g01410	0bp	/	2bp	NHEJ	OsWAK67 - OsWAK short gene
LOC_Os07g14490.1	LOC_Os10g01390					OsWAK68 - OsWAK pseudogene
LOC_Os07g20170.1	LOC_Os10g37670	0bp	/	90bp	NAHR	actin-depolymerizing factor,putative
LOC_Os07g22504.1	LOC_Os06g46435	0bp	/	2bp	NHEJ	NADPH-dependent oxidoreductase,putative
LOC_Os07g23810.1	LOC_Os10g29240	0bp	/	1bp	NHEJ	expressed protein
LOC_Os07g29440.1	LOC_Os01g10940	200bp	3bp	160bp	TE/TIR	S-adenosylmethionine synthetase ,putative
LOC_Os07g37780.1	LOC_Os04g52354	6bp	3bp	2bp	NHEJ	ribosomal protein S17,putative
LOC_Os07g37790.1	LOC_Os04g52340					expressed protein (Cell differentiation protein rcd1)
LOC_Os08g23280.1	LOC_Os06g06310	0bp	/	/	TE flanking	expressed protein (Porin-like)
LOC_Os08g24160.1	LOC_Os06g49760	9bp	/	/	NHEJ	invertase/pectin methylesterase inhibitor family protein,putative
LOC_Os08g28200.1	LOC_Os09g36690	0bp	3bp	0bp	NHEJ	expressed protein (Ankyrin repeat)
LOC_Os08g33280.1	LOC_Os08g40300	9bp	3bp	3bp	NHEJ	ulp1 protease family,C-terminal catalytic domain containing protein
LOC_Os08g39580.1	LOC_Os01g42170	(AT)n	/	/	NHEJ	zinc knuckle family protein

LOC_Os08g40300.1	LOC_Os08g33280	0bp	3bp	3bp	NHEJ	expressed protein
LOC_Os09g04680.1	LOC_Os10g21310	140bp	5bp	4bp	NHEJ	photosystem II P680 chlorophyll A apoprotein,putative
LOC_Os09g11850.1	LOC_Os04g25120	9bp	0bp	0bp	NHEJ	ulp1 protease family,C-terminal catalytic domain containing protein
LOC_Os09g11860.1	LOC_Os04g25110					ulp1 protease family,C-terminal catalytic domain containing protein
LOC_Os09g14570.1	LOC_Os03g38700	0bp	0bp	3bp	NHEJ	tropinone reductase,putative
LOC_Os09g14590.1	LOC_Os03g38720					proteasome maturation factor UMP1 family protein
LOC_Os09g14600.1	LOC_Os03g38730					peroxisomal membrane protein-related,putative,expressed
LOC_Os09g14610.1	LOC_Os03g38740					DCL2,putative,expressed
LOC_Os09g14614.1	LOC_Os03g38745					serine/arginine repetitive matrix protein ,putative,expressed
LOC_Os09g24402.1	LOC_Os10g21358	1.3kb	/	4bp	NHEJ	chloroplast S ribosomal protein L23,putative
LOC_Os09g24404.1	LOC_Os10g21354					chloroplast S ribosomal protein L2,putative
LOC_Os09g24406.1	LOC_Os10g21356					uncharacterized protein ycf72,putative
LOC_Os09g24412.1	LOC_Os10g21352					chloroplast S ribosomal protein S19,putative
LOC_Os09g24414.1	LOC_Os10g21348					chloroplast S ribosomal protein L22,putative
LOC_Os09g24416.1	LOC_Os10g21344					chloroplast S ribosomal protein S3,putative
LOC_Os09g27970.1	LOC_Os12g13290	0bp	2bp	4bp	NHEJ	expressed protein
LOC_Os09g36690.1	LOC_Os08g28200	0bp	/	2bp	NHEJ	expressed protein (ANKYRIN REPEAT)
LOC_Os10g05900.1	LOC_Os02g05140-130	0bp	/	/	unknown	expressed protein
LOC_Os10g12300.1	LOC_Os06g42060	1bp	3bp	5bp	NHEJ	expressed protein
LOC_Os10g22630.1	LOC_Os03g55750	0bp	/	2bp	NHEJ	expressed protein
LOC_Os10g24954.1	LOC_Os02g27830	0bp	/	/	unknown	ulp1 protease family,C-terminal catalytic domain containing protein
LOC_Os10g31360.1	LOC_Os08g35980	0bp	/	/	TE/TIR	GRF zinc finger family protein,expressed
LOC_Os10g31650.1	LOC_Os03g47530	0bp	/	/	TE flanking	glycosyl transferase domain containing protein,putative
LOC_Os11g05630.1	LOC_Os04g08170	10bp	2bp	4bp	TE/TIR	CBL-interacting protein kinase,putative
LOC_Os11g09979.1	LOC_Os04g26330	0bp	3bp	0bp	NHEJ	expressed protein
LOC_Os11g12500.1	LOC_Os09g08450	0bp	/	2bp	NHEJ	ulp1 protease family,C-terminal catalytic domain containing protein
LOC_Os11g38600.1	LOC_Os04g48700	0bp	4bp	4bp	NHEJ	TruB family pseudouridylate synthase,putative
LOC_Os12g10570.1	LOC_Os10g21266	2kb	4bp	/	NHEJ	ATP synthase subunit beta,putative
LOC_Os12g10580.1	LOC_Os10g21268					ribulose bisphosphate carboxylase large chain precursor,putative

LOC_Os12g10590.1	LOC_Os10g21300					OsClp13 - Putative Clp protease homologue
LOC_Os12g10600.1	LOC_Os10g21310					photosystem II P680 chlorophyll A apoprotein,putative
LOC_Os12g19420.1	LOC_Os01g58022	2bp	4bp	/	NHEJ	ubiquinone oxidoreductase,putative
LOC_Os12g19430.1	LOC_Os01g58000					ATP synthase epsilon chain,putative
LOC_Os12g28520.1	LOC_Os09g28860	10bp	0bp	0bp	TE/TIR	abscisic acid-inducible,putative

Supplementary Table S13 Sequence coverage evaluated by comparison with BAC clones sequenced by Sanger technology and Roche 454

BAC	Length	Coverage	DNA TE	Percent%	LOSS	Percent%	RNA TE	Percent%	LOSS	Percent%	GENE NUM	LOSS	Percent%	GENE LEN	LOSS	Percent%
Euchromatin	2698297	97%	605535	22%	19523	3.20%	168508	6%	10977	6.50%	301	16	5.30%	386147	41837	10.83%
a0061N13	125279	98%	43365	35%	885	2.00%	17068	14%	140	0.80%	6	0	0.00%	11877	1	0.00%
a0063A14	169028	100%	37411	22%	156	0.40%	38484	23%	726	1.90%	7	0	0.00%	12501	26	0.20%
a0066D17	161657	98%	50075	31%	800	1.60%	22312	14%	385	1.70%	11	0	0.00%	17118	1	0.00%
DQ810282	246961	95%	42009	17%	525	1.20%	2809	1%	5	0.20%	36	0	0.00%	40955	4893	11.90%
FJ032639	183535	96%	38441	21%	809	2.10%	2385	1%	3	0.10%	20	1	5.00%	28578	3208	11.20%
FJ266021	213323	99%	42877	20%	379	0.90%	21000	10%	31	0.10%	20	0	0.00%	37908	451	1.20%
FQ378032	857171	94%	173957	20%	11982	6.90%	31001	4%	3945	12.70%	116	11	9.50%	135723	17332	12.80%
FQ378033	628490	94%	155127	25%	3878	2.50%	25360	4%	2475	9.80%	74	4	5.40%	85452	15042	17.60%
GQ407207	112853	96%	22273	20%	109	0.50%	8089	7%	3267	40.40%	11	0	0.00%	16035	883	5.50%
Pericentromere[a]	1041915	80%	229217	22%	38347	16.73%	475360	46%	161162	33.90%	46	1	2.17%	46950	3277	6.98%
a0003N12	146708	87%	29647	20%	3308	11.20%	63034	43%	13265	21.00%	7	0	0.00%	5829	61	1.00%
a0008B23	147522	93%	34252	23%	1826	5.30%	34955	24%	3615	10.30%	11	0	0.00%	10377	5	0.00%
a0025L07	184702	79%	43256	23%	9088	21.00%	72256	39%	29467	40.80%	9	0	0.00%	9549	125	1.30%
a0073D22	142790	75%	46589	33%	5546	11.90%	72569	51%	31435	43.30%	6	0	0.00%	3039	82	2.70%
a0079B10	126061	61%	12714	10%	1357	10.70%	75016	60%	38005	50.70%	4	1	25.00%	8520	2464	28.90%
a0094J22	163153	87%	33450	21%	9418	28.20%	86104	53%	16786	19.50%	5	0	0.00%	5760	194	3.40%
a0095J07	130979	75%	29309	22%	7804	26.60%	71426	55%	28589	40.00%	4	0	0.00%	3876	346	8.90%

[a] These seven clones were overlapping BAC clones located on chromosome 8. This region is orthologous to the centromeric region of rice chromosome 8 and covers the centromere of chromosome 8 in *O. brachyantha*. These BAC clones were sequenced by Sanger dye-terminator chemistry (unpublished data from Dr. Jiming Jiang & Dr. Rod Wing)

Supplementary Table S14 Evidences used in gene prediction

Evidence	Rice (Tigr6.1; RAP3)	*Brachypodium* (v1)	*Sorghum* (v1.4)	Maize (v5b)	*Poplar* (v2.0)	*Arabidopsis* (TAIR9)	Total
Protein	51729; 50318	32255	34496	63540	45778	33410	311526
cDNA	54970	-	-	38869	-	-	109017[a]
EST	1226650	-	-	-	-	-	1226650
RNAseq transcript	92349	-	-	-	-	-	92349

[a] including 15178 other monocot cDNAs

Supplementary Table S15 Exonerate parameters used in gene prediction

Evidence	Model	Coverage (%)	Identity (%)
Protein	protein2genome	20	30
rice cDNA	est2genome	30	50
other cDNA	coding2genome	20	30
EST	est2genome	30	50
RNAseq transcripts	est2genome	80	90

Supplementary Table S16 Gene statistical data from selected grass species

	Oryza brachyantha (v1.4)	*O. sativa* (TIGR6.1)	*O. sativa* (RAP3)	*Brachypodium* (v1.0)	*Sorghum* (v1.4)
Gene Number	32038	41404	33265	25532	34496
Gene Length	2629/1873	2774/2093	3024/2422	3336/2643	2850/2038
Gene GC	0.49/0.45	0.50/0.47	0.49/0.46	0.50/0.47	0.50/0.47
Coding GC	0.55/0.53	0.57/0.57	0.57/0.57	0.56/0.55	0.56/0.54
Coding Length	1087/855	1041/819	1004/828	1300/1101	1159/987
Exon Number	4.7/3.0	4.1/2.0	3.9/2.0	5.2/3.0	4.3/3.0
Exon Length	230/131	253/136	260/140	250/134	270/143
Exon GC	0.50/0.47	0.51/0.48	0.51/0.48	0.51/0.47	0.51/0.48
Intron Length	416/164	407/173	412/150	397/155	444/145
Intron GC	0.38/0.36	0.37/0.36	0.37/0.35	0.38/0.37	0.38/0.37
Single Exon	8110	11201	12134	5690	9357

Supplementary Table S17 Comparison of the gene prediction with the published data

Query ID	Hit ID	Length	Coverage	Identity	Query Annotation
Adh1.1	Ob11g15170.1	3723	0.88	1.00	NBS-LRR disease resistance protein family-1
Adh1.10	Ob11g15050.1	282	0.98	1.00	expressed protein-RZ53 family-2
Adh1.11	Ob11g15040.1	1008	1.00	1.00	peroxidase
Adh1.12	Ob11g15030.1	1215	0.65	0.99	NifS-like protein
Adh1.13	Ob11g15020.1	1839	1	1.00	phosphatidylinositol kinase
Adh1.14	Ob11g15010.1	2481	1.00	1.00	S-domain receptor-like protein kinase
Adh1.15	Ob11g14990.1	1335	1.00	1.00	F-box family-1
Adh1.16	Ob11g14970.1	909	0.79	0.98	expressed protein
Adh1.17	Ob11g14950.1	344	0.47	1.00	F-box family-2
Adh1.18	Ob11g14940.1	1164	1.00	1.00	flavin monoxygenase family-1
Adh1.19	Ob11g14930.1	1044	1.00	1.00	flavin monoxygenase family-2
Adh1.2	Ob11g15140.1	2003	0.97	0.99	protein kinase domain containing protein
Adh1.20	Ob11g14920.1	277	0.33	1.00	myb-like DNA-binding domain containing protein
Adh1.3	Ob11g15130.1	3894	1.00	0.99	NBS-LRR disease resistance protein family-2
Adh1.4	Ob11g15120.1	2802	0.65	1.00	NBS-LRR disease resistance protein family-3
Adh1.5	Ob11g15100.1	4058	0.93	1.00	NBS-LRR disease resistance protein family-4
Adh1.6	Ob11g15090.1	1143	1.00	0.99	alcohol dehydrogenase family-3
Adh1.7	Ob11g15080.1	1140	1.00	0.99	alcohol dehydrogenase family-2
Adh1.8	Ob11g15070.1	923	0.81	1.00	alcohol dehydrogenase family-1
Adh1.9	Ob11g15060.1	291	1.00	1.00	expressed protein-RZ53 family-1
Hd1.1	Ob06g19700.1	453	1.00	1.00	putative nucleolar protein family a member 2
Hd1.10	Ob06g19850.1	1350	1.00	1.00	putative Na+ dependent neutral amino acid transporter
Hd1.11	Ob06g19870.1	2307	1.00	0.99	putative adaptor protein kanadaptin
Hd1.2	Ob06g19710.1	651	1.00	1.00	putative somatic embryogenesis receptor kinase 1
Hd1.3	Ob06g19720.1	261	1.00	1.00	hypothetical protein
Hd1.4	Ob06g19740.1	1171	0.61	1.00	putative somatic embryogenesis protein kinase 1
Hd1.5	Ob06g19760.1	1011	1.00	1.00	putative peroxidase 49 precursor
Hd1.6	Ob06g19770.1	666	0.56	1.00	putative heading date 1 protein
Hd1.7	Ob06g19800.1	2682	1.00	1.00	putative polycomb protein EZ1
Hd1.8	Ob06g19820.1	1281	1.00	0.98	putative TA1 protein
Hd1.9	Ob06g19830.1	2565	0.87	1.00	putative armadillo/beta-catenin repeat related protein
Moc1.1	Ob06g28430.1	2977	1.00	1.00	U5 small nuclear ribonucleoprotein component
Moc1.10	Ob06g28570.1	1212	0.68	1.00	unknown
Moc1.11	Ob06g28590.1	528	0.61	1.00	unknown
Moc1.12	Ob06g28600.1	247	0.19	0.99	Monoculm1
Moc1.13	Ob06g28610.1	1254	1.00	1.00	Mlo family protein
Moc1.14	Ob06g28620.1	712	0.82	0.99	aspartic proteinase nepenthesin-1 precursor
Moc1.15	Ob06g28630.1	1743	1.00	1.00	microtubule-associated protein MAP65-1a
Moc1.16	Ob06g28640.1	190	0.46	0.95	IQ calmodulin-binding motif family protein
Moc1.17	Ob06g28660.1	1527	0.81	1.00	EMB2261 putative
Moc1.18	Ob06g28680.1	1218	1.00	1.00	polygalacturonase precursor
Moc1.19	Ob06g28690.1	1210	1.00	1.00	exopolygalacturonase precursor
Moc1.2	Ob06g28440.1	666	1.00	1.00	charged multivesicular body protein 4b
Moc1.20	Ob06g28700.1	3396	0.99	1.00	putative RNA polymerase A(I) large subunit

Moc1.3	Ob06g28450.1	580	0.84	1.00	putative SFT2 protein
Moc1.4	Ob06g28450.1	955	0.97	1.00	fructose-bisphosphate aldolase cytoplasmic isozyme
Moc1.5	Ob06g28460.1	1038	1	0.99	copine-1
Moc1.6	Ob06g28530.1	920	0.39	0.99	subtilisin-like protease precursor
Moc1.7	Ob06g28530.1	966	1.00	1.00	ZAC
Moc1.8	Ob06g28540.1	999	1.00	1.00	MYB-CC type transfactor
Moc1.9	Ob06g28560.1	1509	1.00	1.00	unknown

Supplementary Table S18 Determination of orthologous genes in *O. brachyantha* for the collected genes of *O. sativa*

	Total Gene	Ortholog	Miss Gene
Flower pathway	96	93	3
Starch pathway	15	14	1
Agriculture gene	30	25	5
Total	141	132	9

Supplementary Table S19 Orthologous genes not found by the automatic strategy

Gene ID	Name	Position	Ortholog	Note
LOC_Os01g11940	OsFTL1	6493516	Ob01g17510.1	fine
LOC_Os03g42900	OsFLKb	23923751		have ortholog in maize/sorghum
LOC_Os10g41100	OsP	21998104	Ob10g25540.1	sequence gap
LOC_Os01g44220	AGPls	25352725		miss in prediction
LOC_Os03g29380	GS3	16728081	Ob03g30090.1	incorrect prediction in rice
LOC_Os07g47330	FZP	28298599	Ob07g31130.1	sequence gap
LOC_Os05g09520	qSW5	5365100	Ob05g15380.1	sequence gap
LOC_Os06g40780	MOC1	24310311	Ob06g28600.1	sequence gap
LOC_Os04g57530	sh4	34046082	Ob04g36080.1	sequence gap

Supplementary Note

Supplementary Note 1. Divergence time between *O. brachyantha* and *O. sativa* (rice)

The divergence time of *Oryza* species was initially estimated to be at ~9 million years ago (MYA)[66]. However, two recent studies suggested older divergence time for the split of *Oryza* species (14 MYA[67] or 15 MYA[62]) using more sequences and/or more reasonable approaches. Phylogenetic analysis using 142 gene fragments placed *O. brachyantha* between the basal lineage G-genome and the other genome types, although the authors indicated that the divergence of F-genome and G-genome from the rest of *Oryza* genome was rapid and being the first diversification in the genus *Oryza*[61]. Therefore, the divergence time between *O. brachyantha* and rice should be consistent with the split of *Oryza* species at ~15 MYA.

Supplementary Methods

The plant materials and background

Sequencing materials. *Oryza brachyantha* A. Chev. et Roehr. is an annual diploid wild rice distributed in central African. It grows in an open habitat, with slender culms highly branched. It is the only species assigned to the F genome type. This wild rice possesses potential useful traits for rice breeding, such as resistance to yellow stem borer, leaf-folder, whorl maggot and bacterial blight (BB)[68]. The material used for genome sequencing is the same accession (IRGC101232) as used for the BAC library construction in the *Oryza* Map Alignment Project (OMAP). The plants were kindly provided by Dr. Dashan Brar of the International Rice Research Institute (IRRI). Young leafs from a few plants were harvested to reduce heterozygosity.

Genome size estimation of *O. brachyantha*. The genome size of *O. brachyantha* was estimated to be 342 Mb[69] or 362 Mb[70] using flow cytometry. The genome sequencing project of *O. brachyantha* allows us to give a more accurate estimate of the genome size. Firstly, we employed K-mer depth distribution of sequence reads to estimate the genome size[71]. A total of 166 million reads from small libraries (200-500 bp insertion size) were analyzed to exact 17 K-mer. We obtained 11,330 million of 17 K-mer with the peak of depth at 38, which lead to the estimation of genome size to be ~ 298 Mb (11,330/38). Secondly, we used the integrated BAC-based physical map with the sequence assembly to estimate the genome size, which is the total number of bands covered by the physical map (total CB units) multiplied by the average band size (CB unit size)[72]. The CB unit size was calculated to be 1.38 kb/CB unit by calculating 83 non-overlapping BAC-based fingerprinted contigs covered by a single scaffold sequence (mean contig size = 568.4 CB unit, mean sequence size = 777.6 kb). This leads to an estimation of the genome size to be ~ 297 Mb (Total CB unit × CB unit size = 215,208 × 1.38).

Genome sequencing and assembly

Genome sequencing. Nuclear DNA of *O. brachyantha* was isolated from young leaves using a modified CTAB protocol, followed by purification using phenol-chloroform. The genomic DNA was fragmented into different sizes to prepare pair-end libraries using standard Illumina protocols. Sequencing was performed on an Illumina Genome Analyzer II. We generated more than 30 Gb of clean sequence data, which is ~ 104-fold coverage of the genome of *O. brachyantha*. The BAC library of *O. brachyantha* was constructed at Arizona Genomics Institute[70]. BAC end sequences (BES) were obtained from the *Oryza* Map Alignment Project (OMAP, www.omap.org)[73]. The details on the libraries and sequence coverage are provided in Supplementary Table S1.

Genome assembly. The sequence assembly was performed with SOAPdenovo[74] at Beijing Genomics Institute at Shenzhen (BGI). We used only reads from a small insert size library (< 2,000 bp) to assemble the contigs. These 60-fold sequences were assembled into contigs with an N50 of 1.5 kb. Scaffolds were constructed by adding pair-end mapping reads step by step, ranging from small insertion size libraries to large insertion size libraries. The gaps in the scaffolds were filled by *de novo* assembly of the reads covering the gap regions. The assembly, which was solely based on short reads, produced a genome of 262 Mb with a contig N50 of 20.4 kb and a scaffold N50 of 1 Mb. The scaffold N50 was further improved to 1.6 Mb by adding the BAC end sequences to the scaffold reconstruction. The final assembly,

which is referred to as the "super-scaffold" in Supplementary Table S2, contains 654 super-scaffolds longer than 2 kb, with the longest one reaching 8.8 Mb.

Anchoring the scaffolds onto chromosomes. By integration with the BAC-based physical map, we further promoted the super-scaffold sequences into 36 "big-scaffold" sequences which made up 96% of the total assembly (Fig. 1). These big-scaffold sequences were mapped onto chromosomes based on the gene collinearity in the genus *Oryza* and confirmed with FISH experiments (Supplementary Fig. S3). The unmapped sequences are made up of 4% of the genome and contain 661 predicted gene models.

Evaluation of the accuracy of the sequence assembly. The super-scaffold sequences were integrated with the physical map of *O. brachyantha* to validate the assembly and make further scaffolding[75]. We identified and confirmed nine inconsistencies between the sequence assembly and the physical map (Supplementary Fig. S23). We also compared the assembled sequence with 14 BAC sequences and two large sequence segments that generated by Sanger technology and Roche 454 (9 located in the euchromatic regions and 7 located in the pericentromeric regions. Examples of comparisons were shown in Supplementary Fig. S24). Higher sequence coverage in the euchromatic regions (97%) and lower coverage in the pericentromeric regions (80%) were observed. The sequence coverage is ranging from 61% to 93% in the pericentromeric regions depending on the proportion of LTR retrotransposons (Supplementary Table S13). In the two large sequence segments compared (FQ378032 and FQ378033)[76], although the overall coverage is high (94%), one large sequence gap in each sequence was found to be caused by an unanchored super-scaffold (Supplementary Figs S25 and S26). Considering that we have ~ 661 gene models in the unanchored super-scaffolds, we incorporated the collinear genes on super-scaffolds into synteny analysis by comparing the super-scaffolds with the rice genome using MCscan[77] (Method). This will make the influence of assemble artifacts minimal.

Gene prediction and evaluation

Evidence-based gene prediction. We employed Gramene GeneBuilder[78] to predict gene models in *O. brachyantha*. The data of the proteins and transcripts were aligned with the soft-masked genome sequence by Exonerate[79], using different model and filter stringency (Supplementary Table S14 and Supplementary Table S15). For each line of evidence, a set of gene models was constructed using Gramene GeneBuilder[78]. Gene models were also predicted on the soft-masked genome sequence by *ab initio* tool FGENESH (http://linux1.softberry.com). The final gene models were obtained by combining these gene models using Gramene GeneBuilder[78]. After filtering pseudogenes and short gene models within introns of other genes, 31,601 protein-coding gene models were obtained, including 39,680 transcripts. Another 8,470 FGENESH gene models that did not overlap with the current gene models were also included. We selected a single best transcript for each locus to represent a gene model. Transposable element-related gene models were filtered out if more than 70% of the coding sequences were overlapped with the transposable elements. Species-specific genes were filtered out if they did not have evidence support or code for short proteins ($\leq$ 50 amino acids). The final gene prediction includes 32,038 protein-coding gene models, of which 25,546 (80%) genes have homologous proteins in plant genomes or assignment with a Pfam domain annotation; 4,014 (13%) genes are solely supported by RNA-seq evidence and 2,478 genes have no supporting evidence. The characters of predicted genes of *O. brachyantha* are similar with gene predictions for rice or other monocot genomes (Supplementary

Table S16).

Evaluation of completeness and accuracy of the gene sets. To evaluate the comprehensiveness of the gene prediction efforts, we collected a set of RNA-seq transcripts that were assembled with SOAPdenovo[74], which did not rely on the reference genome. We selected only those transcripts that have ≤ 30% transposable element-related sequences, ≥ 500 bp in length and code a protein product ≥ 100 amino acids. The resulting 13,150 transcripts were used as a standard set of protein-coding genes in *O. brachyantha*. These transcripts were then aligned to the genome sequence by BLAT[80] with an identity ≥ 95%. Approximately 12,282 transcripts were properly aligned to the genome, which suggests 93% (12,282/13,150) of the protein-coding gene sequences of *O. brachyantha* are covered in the assembled genome. Among the aligned transcripts, nearly 91% (11,206/12,282) are overlapped with the predicted gene models (i.e. ≥ 95% identity and ≥ 30% coverage). We found approximately half (579/1,076) of the unaligned transcripts have no homologs upon BLAST search against plant protein database (E-value ≤ 10^{-5}). This suggests that the gene coverage might be underestimated due to false positives in the assembled RNA-seq transcripts.

We also compared the gene prediction with the gene models of three published BAC sequences, which were annotated with manual inspection (Supplementary Table S17). Of the 51 gene models, 32 genes were perfectly matched with the predicted gene models in this study. Ten genes have sequencing gaps in the coding sequences, which caused smaller gene sizes compared with the gene sizes in the BAC sequences. However, almost all of these gap-containing genes (9/10) covered more than 1/3 of the coding sequences.

In a collection of studies of 141 rice genes (Supplementary Table S18), we were able to find clear orthologs for 132 of them by reciprocal BLAST (E-value ≤ 10^{-5} and further confirmed by analysis the orthologous sequence regions). The nine genes without clear orthologs were further analyzed by comparisons of orthologous sequences between rice and *O. brachyantha*. Five genes do not exhibit any homologs in the orthologous regions of *O. brachyantha* because of too large sequence gap in the coding sequences. No homologous sequences were found in the orthologous region of gene *LOC_Os03g42900*. However, the orthologous genes of *LOC_Os03g42900* are detected in maize and sorghum. This suggested that the gene *LOC_Os03g42900* was deleted from *O. brachyantha* or entirely missed by the current assembly. The orthologous gene of *LOC_Os01g11940* is well conserved, but was not found by automatic script. The other two genes, *LOC_Os01g44220* and *LOC_Os03g29380*, were the result of incorrect prediction either in *O. brachyantha* or in rice (Supplementary Table S19). In total, approximately 5% (7/141) of the functional genes may have not been covered in the genome sequence, which is consistent with previous evaluation using RNA-seq transcripts.

As shown above, the sequence gaps in the gene regions may cause a partial or incorrectly predicted gene model. We evaluated the influence of the sequence gaps on all of the predicted gene models. Approximately 2,177 genes were found to have coding sequence gaps (with a mean gap length of 243 bp). We randomly selected 100 genes to check the quality of gene prediction on these gap genes (50 genes ≤ 100 bp gap and 50 genes ≥ 100 bp gap). For 85 cases, the gap-containing genes are homologous to rice genes, of which 17 genes cover less than 30% length of the homologous genes due to sequence gaps or frameshift in coding sequences. Seven genes are specific to *O. brachyantha*, and only

Ob04g15920.1 has a homolog in other plant genomes but not in the rice genome. These gene models may have been due to false positive predictions. Eight genes had an incorrect gene model caused by sequence gaps. In the selected set, approximately 26% (17+1+8) of the gap genes were shown to have a large impact on gene structure. Of the 1,993 genes with single sequence gaps in the coding sequence, we found that approximately 30% of genes were unable to cover ≥ 30% of the homologous genes, which means that approximately 720 (2,177 × 30%) genes are not covered or are presented as gene fragments due to sequences gaps in the genome (i.e. 2% of all the genes of *O. brachyantha*).

Transposable element annotation

LTR retrotransposons. First, the genome sequence was masked by the centromeric tandem repeats CentO-F[64] and retrotransposons[81] in *O. brachyantha* using the RepeatMasker program (http://www.repeatmasker.org). Then the masked sequence was screened by the LTR-Finder program[82] using default parameters with 2 exceptions in that we set 50 bp minimum LTR length and 100 bp minimum distance between the LTRs. All of the output was manually checked to determine the exact boundaries of the characterized LTR retrotransposons and to filter out the incorrect predicted sequences. In addition, the internal sequences of all of the LTR retrotransposons detected by the LTR-Finder were used as queries to search against GenBank and the Rice genome annotation project database (http://rice.plantbiology.msu.edu) in order to group these LTR retrotransposons into different families.

Non-LTR retrotransposons. To identify long interspersed nuclear elements (LINEs), the conserved reverse transcriptase domain of the LINEs superfamily[83] was screened for the *O. brachyantha* genome. Ten kilo base (kb) flanking sequences (5-kb on each side) for each hit were extracted and manually examined for polyA/polyT tails and TSDs. Short interspersed nuclear elements (SINEs) were annotated by polyA or polyT tails and TSDs. Additionally, the rice and maize LINEs and SINEs deposited in the GIRI (http://www.girinst.org) were used as queries to search against the *O. brachyantha* genome to find their homologs in the genome.

DNA transposons. In order to characterize *Mutator*-like elements (MULEs), the *mudrA* conserved domain[84] was used as a query to search against the *O. brachyantha* genome. Twenty kilo base (kb) flanking sequences (10 kb on each side) for each hit were aligned to determine terminal inverted repeats (TIRs) and TSDs. The autonomous MULEs database in *Nipponbare* (Dr Dongying Gao, unpublished) was also utilized to discover homologous MULEs in *O. brachyantha*. All of the MULEs detected by these two methods were combined and used to identify nonautonomous MULEs in the *O. brachyantha* genome. The hAT transposons were detected by hAT family dimerisation domain (pfam05699) and their exact boundaries were determined based on TIRs and 8-bp TSDs. The CACTA transposase conserved domain (pfam02992) was used to conduct tblastn searches and the boundaries of CACTA transposons were determined by the CMCWR terminal motif and 2-3 bp TSDs. Transposons belonging to PIF*/Harbinger* superfamily were annotated based on conserved domain of PIF-like TPases[85] and 3-bp TSD (TWA). Tc1*/mariner* like elements were identified by the transposase sequence[86] and the boundaries were decided by TIRs and 2-bp TSD (TA). Helitron elements were identified based on 5'TC-CTAG3' and the hairpin loop features using the *Helitron* Finder program[87]. The rice transposon library of Dr. Ning Jiang (unpublished data) also was used as a reference for the identification of DNA transposons in *O. brachyantha*.

Influence of sequence assembly on transposable element annotation and genome size evolution. Using the validation dataset, we estimated that the proportion of DNA transposons in euchromatin (EU) and heterochromatin (HE) are similar (22% EU vs. 22% HE). However, the proportion of RNA repeat elements in heterochromatin was much higher (6% EU vs. 45% HE). By comparing the validation dataset with the genome assembly, we found 3.2% and 16.7% of DNA transposons were missed in transposable element annotation in euchromatin and heterochromatin, respectively. The miss-annotation rate is higher for RNA repeat elements (6.5% in EU and 33.9% in HE).

We assume that the estimated genome size of *O. brachyantha* is 300 Mb. We used two schemes to evaluate how many repeat sequences were possibly miss-annotated in the current assembly. In the first scheme, we regard the pericentromeric regions (~2 Mb in the centromeric regions) of *O. brachyantha* as heterochromatic based on cytogenetic observations. The miss-annotated DNA transposons in euchromatin are estimated to be ~ 1.94 Mb (276 Mb * 0.22 * 0.032 = 1.94 Mb), heterochromatin ~ 0.88 Mb (24 Mb * 0.22 * 0.167 = 0.88 Mb). The miss-annotated RNA repeat elements in euchromatin are estimated to be ~ 1.1 Mb (276 Mb * 0.06* 0.065 = 1.1 Mb), heterochromatin ~ 3.7 Mb (24 Mb * 0.45 * 0.34 = 3.7 Mb). In total, we miss-annotate 4.74-Mb of repeat elements in this scheme. In the second scheme, we assume the proportion of heterochromatin in *O. brachyantha* was the same as in rice (15%). In this scheme, the miss-annotated transposable elements are ~ 11.3 Mb, including ~ 3.44 Mb DNA transposons and ~ 7.9 Mb RNA repeat elements.

Considering the genome size differences between *O. brachyantha* (262 Mb, including 70 Mb transposable elements) and rice genomes (372 Mb, including 138 Mb transposable elements), the differences of transposable elements were estimate to contribute ~ 56.6 Mb (138-(70+11.34) = 56.6) to genome size variation. And LTR retrotransposons alone may contribute 50.4 Mb (79.3-(21+7.9) = 50.4) to genome size variation, approximately 45% (50.4/(372-262) = 46%) of the genome size variation. This result is slightly lower but comparable to the estimation in the manuscript (~ 50%).

Resistance related gene family

Characterize resistance gene families NBS-LRR and RLK-LRR. Genes for disease resistance comprise the most dynamic gene family in plant genomes[88]. In *O. sativa*, NBS-LRR and RLK-LRR are two major classes of resistance genes involved in defense-related responses[89]. To characterize members of NBS-LRR family, we first searched protein-coding genes with hmmer3[90] using Pfam family NBS (NB-ARC, domain Pfam00931). Candidate genes were filtered and classified into subfamily by presence of specific domains (NBS, nucleotide-binding site; CC, Coiled-coil; TIR, Toll Interleukin Receptor; LRR, Leucine Rich-Repeat). CC domains were detected by ncoils with default parameters[91]. LRR domains were detected by hmmsearch[90] using LRR domains in Smart database[92] (E-value $\leq$ 0.1). NBS and TIR domains were detected by hmmsearch[90] using Pfam00931 and Pfam001582 with E-value $\leq$ 0.1. To characterized members of RLK-LRR gene family, we searched protein-coding gene with hmmer3[90] using kinase domain (Pfam00069) and LRR domains in Smart database[92] (E-value $\leq$ 0.1). Classification of subfamilies of RLK-LRRs were based the phylogenetic relationship with subfamilies in *O. sativa*[93]. For both NBS-LRR and RLK-LRR, additional members were retrieved by adding collinear genes that assigned as NBS-LRR or RLK-LRR in one species but missed in another.

Comparative analysis of resistance gene families NBS-LRR and RLK-LRR. In total, we

characterized 399 and 657 NBS-LRR genes in *O. brachyantha* and *O. sativa*, respectively (Supplementary Table S5). Ninety percent (363) of the NBS-LRR genes in *O. brachyantha* are located in orthologous positions compared to *O. sativa*, of which 196 are present as tandem duplications in *O. brachyantha*, compared to 345 in *O. sativa* (Supplementary Table S5 and Supplementary Fig. S7). Eleven and four pseudogenes were identified by comparison of the orthologous gene structures in *O. brachyantha* and *O. sativa*, respectively (Supplementary Table S6). Transposable element insertions were found to be involved in the pseudogenization of 7 and 2 genes in *O. brachyantha* and *O. sativa*, resulting in the absence of functional resistance genes in these loci (Supplementary Table S6 and Supplementary Fig. S8). In addition, 156 NBS-LRR genes were found to be present in non-collinear positions in *O. sativa*, thus contributing to the expansion of NBS-LRR gene family in *O. sativa* (Supplementary Table S5 and Supplementary Fig. S7). The RLK-LRR gene family is highly conserved between the *O. brachyantha* and *O. sativa*. Only the LRR-I, LRR-XII and LRR-Xb subfamilies were amplified in *O. sativa* (Supplementary Table S7). The LRR-XII subfamily have 30 genes amplified by tandem duplication and another 24 genes transposed to non-collinear positions (Supplementary Table S7 and Supplementary Fig. S9). Most of the pseudogenes of RLK-LRR are members of LRR-XII, demonstrating the dynamic evolution of this subfamily in the plant immune system (Supplementary Table S8 and Supplementary Fig. S10).

RNA-seq transcriptome

Shoots and roots were harvested from seedlings of *O. brachyantha* (IRGC101232) two weeks after germination. Total RNA was extracted using Trizol reagent (Invitrogen) with further purification. We then used Dynabeads with Oligo(dT) (Invitrogen) to enrich mRNA from the purified total RNA. The mRNAs were fragmented into 200-700 nts and used to synthesized cDNA with random hexamer primers. The products were purified with a QIAquick PCR Perification Kit (QIAGEN) and 200-250 bp fragments were used to construct the sequencing library using the mRNA-seq sample prep kit (Illumina). The sequencing was performed with an Illumina Genome Analyzer (75 bp PE). In total, we produced 30.5 million reads (16 million for roots and 14.5 million for shoots). The reads were aligned with the genome using Tophat[94], which can align junction reads to the exon-intron boundary. By allowing 2 mismatches per set of reads, approximately 26 million reads (88%) were uniquely mapped to the genome, in which 19 million reads (63%) with matching pairs were properly mapped. Only 1.7 million (5%) mapped reads with matching pair were unmapped. The short reads were assembled into transcripts by Cufflink[95] (92,349 transcripts with an N50 size of 1,022 bp), as well as SOAPdenovo[74] (113,340 transcripts with an N50 size of 447 bp).

Supplementary References

61. Zou, X.H. et al. Analysis of 142 genes resolves the rapid diversification of the rice genus. *Genome Biol* **9**, R49 (2008).
62. Tang, L. et al. Phylogeny and biogeography of the rice tribe (Oryzeae): evidence from combined analysis of 20 chloroplast fragments. *Mol Phylogenet Evol* **54**, 266-77 (2010).
63. Kim, H. et al. Construction, alignment and analysis of twelve framework physical maps that represent the ten genome types of the genus Oryza. *Genome Biol* **9**, R45 (2008).
64. Lee, H.R. et al. Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in Oryza species. *Proc Natl Acad Sci U S A* **102**, 11793-8 (2005).
65. Cheng, Z., Buell, C.R., Wing, R.A., Gu, M. & Jiang, J. Toward a cytological characterization of the rice genome. *Genome Res* **11**, 2133-41 (2001).
66. Guo, Y.L. & Ge, S. Molecular phylogeny of Oryzeae (Poaceae) based on DNA sequences from chloroplast, mitochondrial, and nuclear genomes. *Am J Bot* **92**, 1548-58 (2005).
67. Lu, F. et al. Comparative sequence analysis of MONOCULM1-orthologous regions in 14 Oryza genomes. *Proc Natl Acad Sci U S A* **106**, 2071-6 (2009).
68. Ali, M., Sanchez, P., Yu, S.B., Lorieux, M. & Eizenga, G. Chromosome Segment Substitution Lines: A Powerful Tool for the Introgression of Valuable Genes from Oryza Wild Species into Cultivated Rice (O. sativa). *Rice* **3**, 218-234 (2010).
69. Uozu, S. et al. Repetitive sequences: cause for variation in genome size and chromosome morphology in the genus Oryza. *Plant Mol Biol* **35**, 791-9 (1997).
70. Ammiraju, J.S. et al. The Oryza bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza. *Genome Res* **16**, 140-7 (2006).
71. Huang, S. et al. The genome of the cucumber, Cucumis sativus L. *Nat Genet* **41**, 1275-81 (2009).
72. Chen, M. et al. An integrated physical and genetic map of the rice genome. *Plant Cell* **14**, 537-45 (2002).
73. Wing, R.A. et al. The oryza map alignment project: the golden path to unlocking the genetic potential of wild rice species. *Plant Mol Biol* **59**, 53-62 (2005).
74. Li, R. et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20**, 265-72 (2009).
75. Nelson, W. & Soderlund, C. Integrating sequence with FPC fingerprint maps. *Nucleic Acids Res* **37**, e36 (2009).
76. Jacquemin, J. et al. Long-range and targeted ectopic recombination between the two homeologous chromosomes 11 and 12 in Oryza species. *Mol Biol Evol* **28**, 3139-50 (2011).
77. Tang, H. et al. Synteny and collinearity in plant genomes. *Science* **320**, 486-8 (2008).
78. Liang, C., Mao, L., Ware, D. & Stein, L. Evidence-based gene predictions in plant genomes. *Genome Res* **19**, 1912-23 (2009).
79. Slater, G.S. & Birney, E. Automated generation of heuristics for biological sequence

comparison. *BMC Bioinformatics* **6**, 31 (2005).

80. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-64 (2002).
81. Gao, D. et al. A lineage-specific centromere retrotransposon in Oryza brachyantha. *Plant J* **60**, 820-31 (2009).
82. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* **35**, W265-8 (2007).
83. Xiong, Y. & Eickbush, T.H. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *Embo J* **9**, 3353-62 (1990).
84. Lisch, D.R., Freeling, M., Langham, R.J. & Choy, M.Y. Mutator transposase is widespread in the grasses. *Plant Physiol* **125**, 1293-303 (2001).
85. Zhang, X., Jiang, N., Feschotte, C. & Wessler, S.R. PIF- and Pong-like transposable elements: distribution, evolution and relationship with Tourist-like miniature inverted-repeat transposable elements. *Genetics* **166**, 971-86 (2004).
86. Yang, G., Nagel, D.H., Feschotte, C., Hancock, C.N. & Wessler, S.R. Tuned for transposition: molecular determinants underlying the hyperactivity of a Stowaway MITE. *Science* **325**, 1391-4 (2009).
87. Du, C., Caronna, J., He, L. & Dooner, H.K. Computational prediction and molecular confirmation of Helitron transposons in the maize genome. *BMC Genomics* **9**, 51 (2008).
88. Meyers, B.C., Kaushik, S. & Nandety, R.S. Evolving disease resistance genes. *Curr Opin Plant Biol* **8**, 129-34 (2005).
89. Chen, X. & Ronald, P.C. Innate immunity in rice. *Trends Plant Sci* **16**, 451-9 (2011).
90. Eddy, S.R. A new generation of homology search tools based on probabilistic inference. *Genome Inform* **23**, 205-11 (2009).
91. Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162-4 (1991).
92. Letunic, I., Doerks, T. & Bork, P. SMART 6: recent updates and new developments. *Nucleic Acids Res* **37**, D229-32 (2009).
93. Shiu, S.H. et al. Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. *Plant Cell* **16**, 1220-34 (2004).
94. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-11 (2009).
95. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-5 (2010).