nature genetics

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41588-018-0040-0

In the format provided by the authors and unedited.

Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*

**Joshua C. Stein[1], Yeisoo Yu[2,21], Dario Copetti[2,3], Derrick J. Zwickl[4], Li Zhang[5], Chengjun Zhang[5], Kapeel Chougule[1,2], Dongying Gao[6], Aiko Iwata[6], Jose Luis Goicoechea[2], Sharon Wei[1], Jun Wang[7], Yi Liao[8], Muhua Wang[2,22], Julie Jacquemin[2,23], Claude Becker[9], Dave Kudrna[2], Jianwei Zhang[2], Carlos E. M. Londono[2], Xiang Song[2], Seunghee Lee[2], Paul Sanchez[2,24], Andrea Zuccolo[2,25], Jetty S. S. Ammiraju[2,26], Jayson Talag[2], Ann Danowitz[2], Luis F. Rivera[2,27], Andrea R. Gschwend[5], Christos Noutsos[1], Cheng-chieh Wu[10,11], Shu-min Kao[10,28], Jhih-wun Zeng[10], Fu-jin Wei[10,29], Qiang Zhao[12], Qi Feng[12], Moaine El Baidouri[13], Marie-Christine Carpentier[13], Eric Lasserre[13], Richard Cooke[13], Daniel da Rosa Farias[14], Luciano Carlos da Maia[14], Railson S. dos Santos[14], Kevin G. Nyberg[15], Kenneth L. McNally[3], Ramil Mauleon[3], Nickolai Alexandrov[3], Jeremy Schmutz[16], Dave Flowers[16], Chuanzhu Fan[7], Detlef Weigel[9], Kshirod K. Jena[3], Thomas Wicker[17], Mingsheng Chen[8], Bin Han[12], Robert Henry[18], Yue-ie C. Hsing[10], Nori Kurata[19], Antonio Costa de Oliveira[14], Olivier Panaud[13], Scott A. Jackson[6], Carlos A. Machado[15], Michael J. Sanderson[4], Manyuan Long[5], Doreen Ware[1,20] and Rod A. Wing[2,3,4]*

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. [2]Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ, USA. [3]International Rice Research Institute, Los Baños, Philippines. [4]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA. [5]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA. [6]Center for Applied Genetic Technologies, University of Georgia, Athens, GA, USA. [7]Department of Biological Sciences, Wayne State University, Detroit, MI, USA. [8]State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China. [9]Max Planck Institute for Developmental Biology, Tübingen, Germany. [10]Institute of Plant and Microbial Biology, Academia Sinica, Taipei, Taiwan. [11]Institute of Botany, National Taiwan University, Taipei, Taiwan. [12]National Center for Gene Research, Chinese Academy of Sciences, Shanghai, China. [13]Laboratoire Génome et Développement des Plantes, UMR 5096 UPVD/CNRS, Université de Perpignan Via Domitia, Perpignan, France. [14]Plant Genomics and Breeding Center, Universidade Federal de Pelotas, Pelotas, Brazil. [15]Department of Biology, University of Maryland, College Park, MD, USA. [16]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA. [17]Institute of Plant Biology, University of Zurich, Zurich, Switzerland. [18]Queensland Alliance for Agriculture and Food Innovation, University of Queensland, Brisbane, Queensland, Australia. [19]National Institute of Genetics, Mishima, Japan. [20]Robert W. Holley Center for Agriculture and Health, US Department of Agriculture, Agricultural Research Service, Ithaca, NY, USA. Present addresses: [21]Phyzen Genomics Institute, Phyzen, Inc., Seoul, South Korea. [22]Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany. [23]Crop Biodiversity and Breeding Informatics Group, Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany. [24]Rice Experiment Station, Biggs, CA, USA. [25]Institute of Life Sciences, Scuola Superiore Sant'Anna, Pisa, Italy. [26]DuPont-Pioneer, Johnston, IA, USA. [27]BIOS–Parque Los Yarumos, Manizales, Colombia. [28]Department of Plant Systems Biology, VIB and Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium. [29]Department of Forest Molecular Genetics and Biotechnology, Forestry and Forest Products Research Institute, Tsukuba, Japan. *e-mail: rwing@email.arizona.edu

Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza

Supplementary Information

Table of Contents

Supplementary Note: The 13-genome Data Package: Sequencing, Assembly, Annotation, Assembly Validation.

RESULTS

Sequence and assembly.

Chromosome-level reference assemblies for seven wild species (*O. rufipogon*, *O. nivara*, *O. barthii*, *O. glumaepatula*, *O. meridionalis* and *O. punctata* and *L. perrieri*) (**Fig. 1, Table 1**) were generated using short-read technologies with extensive scaffold support from long-insert library reads, including BAC-ends. Each genome was shotgun sequenced to a minimum depth of 100X coverage using Illumina (San Diego, CA) technologies. Two species (*O. barthii* and *O. punctata*) received additional 10-20X sequence coverage using Roche 454 technology (Branford, CT) (**Supplementary Table 1**). Primary assemblies produced between 9,937 to 68,481 contigs, depending on species, with contig N50s ranging from 9.1 to 50.2 kb (**Supplementary Table 2**). High congruency was found after mapping paired-end reads from short-insert libraries back to the assembled contigs, with >95% of reads mapping as pairs within expected distances (range 187-430 bp) and 88% in the expected orientation (**Supplementary Table 3**).

Mate-pair reads were used to scaffold and orient adjacent contigs, resulting in scaffold N50s ranging from 137.9 kb (*O. rufipogon*) to 8.68 Mb (*L. perrieri*) (average = 1.65Mb). Final super-scaffolds representing chromosome-level pseudomolecules were built manually with the aid of Genome Puzzle Master (GMP) software[1] using paired BAC-end sequences ([*O. nivara, O. rufipogon, O. punctata*][2] and [*O. barthii, O. glumaepatula, O. meridionalis, L. perrieri*] described herein) and alignment to the *O. sativa* vg. japonica reference sequence[3] (herein referred to as the Nipponbare RefSeq) as guide information. Total lengths of the resulting pseudomolecules, constituting twelve chromosomes in each of the seven species, ranged between 267 Mb (*L. perrieri*) and 394 Mb (*O. punctata*) (**Table 1**), showing consistency with genome-size estimates using three independent methods (i.e. flow cytometry[4,5], K-mer and physical map length).

The quality of each wild reference assembly was assessed using metrics of gene and repeat space completeness, and accuracy at the levels of scaffold order/orientation and base-pair fidelity. Comparisons of assembly length to the consensus estimate of each species' genome size, provided estimates of completeness ranging from 77% (*O. meridionalis*) to 96% (*O. punctata*), with an average of 83% (**Supplementary Tables 4 and 5**). A more detailed accounting of sequence recovery was gained using a common ~5 Mb region within the short arm of chromosome 3, which had been independently sequenced and assembled from pooled BAC clones in each species.

Overall capture of this region (composed of ~48% genes and ~24% transposons depending on species) ranged from 94% (*O. glumaepatula*) to almost 99% (*O. niv*ara) within reference assemblies (**Supplementary Table 6**). However, recovered sequences tended to favor genes (both exons and introns) over intergenic and repetitive regions. In the worst case, *O. glumaepatula*, exons and introns were captured with 98.9% and 96.7% coverage respectively, while intergenic and transposon sequences showed coverages of 90.9% and 80.0%, respectively (**Supplementary Tables 6-9**). A similar pattern of biased recovery of genic regions over TEs was also evident among 44 finished BACs sequenced in four species (**Supplementary Tables 10-12**). Underrepresentation of repetitive sequences is an expected outcome of short-read sequencing, as many reads cannot be uniquely placed. Indeed, by directly annotating transposons within the raw reads, Copetti and Wing[6] found that between 11-25% (i.e. 37.5-110 Mb, average 17%) of repeat and TE content is missing among the seven wild short-read assemblies (**Supplementary Table 5**).

To confirm the order and orientation of sequence scaffolds, and thus the quality of each wild genome assembly at a wider scale, we mapped paired-BAC-end sequences (P-BESs)[2] to each assembly and quantified the number of P-BESs that supported each assembly *vs*. those in conflict. The results in **Supplementary Table 13** show that the vast majority (>96%) of P-BESs support each assembly, meaning that paired-ends mapped within the expected length of a BAC clone (25-300 kb) with correct orientation. Conflicting P-BES data represented between 1% (*L. perrieri*) to 4% (*O. barthii*) of each assembly (average of 2%) and were primarily attributed to sequence contigs located in the correct positions but with wrong orientations (e.g. see **Supplementary Fig.** 1). With one exception, the P-BES data (having 3.4-8.9 fold genome coverage) provided supporting information for 93% of each assembly, on average. Lower availability of P-BES for *O. meridionalis* (0.6 fold coverage) still enabled supporting information on 52% of this assembly (**Supplementary Table 13**). These results demonstrate that our assemblies faithfully represent the order and orientation of the majority of scaffolds across each genome.

Accurate placement of contigs within scaffolds is also supported in alignments to the chromosome 3 short arm and finished BAC sequences. However, we note that the *O. nivara* assembly exhibits misplacement of discrete segments of several BACs, and of about 12% of the chromosome 3 short arm assembly. These and other putative super-scaffolding artifacts are tractable within our gene syntelog sets and other supplementary data (**Supplementary Data 1**).

In six of the seven wild species, average base-pair accuracy, determined using finished BAC sequences and chromosome 3 short arm assemblies, ranged from 99.70% in *O. barthii* to 99.94%

in *L. perrieri* (**Supplementary Table 14**). *O. rufipogon* showed lower identity, averaging 98.40%, due to different biological accessions used in the present and previous studies[7].

To estimate gene space content in the seven new wild and four previously published reference assemblies, we first scanned each assembly for conserved sets of eukaryotic genes using both the CEGMA[8] (248 genes) and BUSCO[9] (956 genes) pipelines. The results (**Supplementary Table 15**) show an average gene space completeness of between 91% up to 96% across the seven wild genomes. To additionally evaluate gene space, we generated transcriptome data comprising 36 to 250 million paired-end, strand-specific RNA-seq reads (Illumina, USA) from leaves, roots, and panicles from each of the seven species, plus three additional species, *O. brachyantha*, *O. glaberrima*, and *O. sativa* vg. japonica (**Supplementary Table 16**). *De novo* assembly generated between 54,439 to 355,433 transcripts, with N50 lengths from 546 to 1,674 bp (**Supplementary Tables 17 and 18**). Using a conservative set of high-confidence transcripts to mitigate possible biological contamination, and accounting for multiple transcript isoforms, an average of 96.5% of transcribed loci were successfully mapped to the Nipponbare RefSeq[3] (**Supplementary Tables 19 and 20**). *O. nivara*, *O. glumaepatula*, *O. punctata*, and *L. perrieri* were in a similar range; *O. barthii*, *O. rufipogon*, and *O. brachyantha* were slightly lower; while the *O. meridionalis* and *O. glaberrima* assemblies had around 85% mapping success. Combined, these data support our conclusion that "gene space" in our assemblies is well represented, and in six out of seven cases (8 out of 10 overall) approaches or matches that of the Nipponbare RefSeq[3].

To validate our analyses in comparison to the Nipponbare RefSeq, we also generated reference genome assemblies for two cultivated accessions (*O. sativa* vg. indica cv. IR 8 [a.k.a. Miracle Rice] and the drought-tolerant *O. sativa* vg. aus cv. N 22) using PacBio RSII long-read sequencing technology. The IR 8 genome was sequenced to 73.4X coverage with a subread N50 of 23.4 kb, assembled into 67 contigs with Canu[22], polished with 66X 2X150bp Illumina reads, and edited with GPM[1] [REF] into a 12 chromosome assembly of 389.1 Mb. The N 22 genome was sequenced to 65X coverage with a subread N50 of 16.3 kb, assembled in to 912 contigs with FALCON, polished twice with whole-genome shotgun (WGS) raw PacBio reads, and edited with GPM in to a 12 chromosome assembly of 362.3 Mb. Gene space completeness of these two genomes with CEGMA[8] and BUSCO[9] gave results similar to our seven wild genome assemblies (**Supplementary Table 15**).

We therefore conclude that the nine new assemblies presented here, combined with four previously published *Oryza* genome sequences, constitute the highest quality within-genus data

set to date for any multicellular eukaryotic model system, as compared with twelve within-genus genome projects that have genomes sizes of 100 Mb or greater found in GenBank (seven animal [73 species], three plant [24 species], and two protist [19 species] genera, see **Supplementary Data 2**). In addition, since all but the PacBio assemblies have BAC-level support[10,11], virtually any region can be easily isolated, physically interrogated and functionally validated.

Annotation

To minimize bias associated with different methods of repeat and gene finding, we applied a uniform set of annotation protocols to all 13 genomes. Predominant classes of DNA transposons and retrotransposons were identified using both *de novo* and homology-based approaches[10] (**Supplementary Table 21**). These constituted between 27-50% of the assemblies (**Table 1**), showing a moderately positive correlation with genome size (Pearson's r = 0.64, p-value = 0.03). Long intergenic non-coding RNA (lincRNA) genes were identified as expressed loci with low protein-coding potential after eliminating other known classes of non-coding RNA (see below). Annotation of protein-coding genes integrated evidence from transcriptome assemblies (both *de novo* and reference-guided, **Supplementary Tables 17-22**) with homology and *ab initio* prediction. This yielded 24,208 to 38,550 annotated loci per genome (**Table 1**). While most of the differences in gene numbers could be attributed to lineage-specific loci (**Supplementary Fig. 2**), three of the genomes, *O. glaberrima*, *O. meridionalis*, and *O. brachyantha*, had notably lower counts within otherwise highly conserved genes. Among 13,397 highly conserved ortholog sets, 93-95% were present in these three species, compared to at least 98% for the other species. For the first two, and to a lesser extent for *O. brachyantha*, annotation deficiencies were attributed to genome assembly gaps, based on the identification of missing orthologs in the transcriptome data (**Supplementary Table Fig. 3**, **Supplementary Table 23**), that failed to map to the reference assemblies (**Supplementary Table 24**). In the end, transcriptome data enabled us to account for 94% of missed annotations in *O. glaberrima* and *O. meridionalis*, and 75% of missed annotations in *O. brachyantha*. Evidence for falsely split gene models, an artifact where a single gene is annotated as two or more genes, was found in fewer than 2% of all gene annotations (**Supplementary Table 25**).

METHODS

BAC-end sequencing, fingerprinting and FPC assembly.

All BAC-end sequencing, SNaPshot fingerprinting and physical map assembly methods have been described previously[2,11]. All BAC-end sequences were deposited at NCBI with the following accession numbers: *O. barthii* 67314 BESs (KS450671 - KS517984); *O. glumaepatula* 63194 BESs (JM144568-JM170463, JY086207-JY123504); *O. meridionalis* 30567 BES (JM114001-JM144567); *L. perrieri* 66421 BESs (JM429052- JM495472).

Genome assembly of the wild relatives of rice

Plant material and methods for genome sequencing and assembly. Voucher specimens of each species were obtained from the International Rice Research Institute (IRRI). Species names, cultivar accessions, and NCBI BioProject identifiers are listed in **Supplementary Note Table 1**. Sequence data generated for each species are shown in **Supplementary Table 1**, with NCBI SRA accessions given it **Supplementary Note Table 2**.

Illumina fragment library and sequencing: Young leaf tissue was collected and DNA extracted using DNEasy Plant mini kits (Qiagen, USA) following manufacturer's protocol. About 3-5 ug of DNA was sheared by nebulization and the fragmented DNA was used to construct Illumina sequencing libraries. Fragment libraries, with an insert size of 300-400bp, were constructed using the SPRIworks system I (Beckman, USA). Additionally, for *O. nivara,* an 800 bp insert library was constructed using an Illumina library kit. The fragment libraries were sequenced with 2x120 bp on an Illumina GAIIx for *O. barthii*. For *O. punctata*, *O. nivara*, *O. meridionalis*, *O. glumaepatula* and *L. perrieri,* the fragment libraries were 2x100 bp sequenced using a HiSeq2000.

Illumina large-insert mate pair library and sequencing: High molecular weight DNA was prepared from young leaf tissue and the DNA was sheared with a Hydroshear (Genemachine, USA). Size selection of 3, 10 and 20 kb fragments was performed by agarose gel electrophoresis. CHEF gel electrophoresis was used for the size-selection of 30-40 kb fragments, followed by DNA recovery with an Electro-Elutor (Bio-Rad, USA) in TE buffer. Mate pair (MP) libraries were constructed with the size-selected molecules following the Roche/454 Paired End library protocol (*Cre-lox* recombination) by ligating Illumina TruSeq indexed adapters for the final MP libraries (for *O. barthii*, *O. punctata*, *O. meridionalis*, *O. glumaepatula* and *L. perrieri*). Libraries from the same

species with different insert sizes and MID tags were pooled and 2x100 bp sequenced on an Illumina HiSeq2000 in one flowcell lane. For *O. rufipogon* and *O. nivara*, Illumina Mate Pair Library Sample Preparation kits were used to construct 2 and 5 kb MP libraries, which were then 2x100 bp sequenced on an Illumina HiSeq2000.

Roche/454 library and sequencing: High molecular weight DNA, 3-5 ug, was used to construct Roche 454 whole genome shotgun Titanium libraries for *O. barthii* and *O. punctata* using manufacturer's recommendations. Each library was sequenced up to 4.0 Gb for *O. barthii*, and 5.1 Gb for *O. punctata* using Roche/454 FLX+ chemistry. For *O. punctata*, two paired end libraries (8 and 10 kb insert) were constructed using the Roche/454 PE protocol and sequenced to a depth of 27x physical coverage.

Genome assembly and pseudomolecule construction: For genome assembly, low quality Illumina sequences (Q<20) were trimmed, and paired reads that had more than 40 bp of high quality sequences were selected using Trimmomatic software[12]. Illumina high-quality PE and MP sequences were assembled into contigs and scaffolds using de Bruijn assemblers (i.e. ALLPATHS-LG[13] for *O. barthii*, *O. nivara*, *O. glumaepatula*, *O. punctata* and *L. perrieri*; SOAPdenovo[14] for *O. rufipogon*; ABySS[15] for *O. meridionalis*). SSPACE[16] and GapFiller/GapCloser[17] software tools were used to scaffold contigs using MP data and to fill gaps with PE data, respectively. Roche 454 reads from *O. punctata* and *O. barthii* were assembled using Newbler[18]; the two independent assemblies (ALLPATHS-LG and Newbler) were then merged using Mimimus2[19]. Final super-scaffolds representing chromosome-level pseudomolecules were built manually with the aid of Genome Puzzle Master software[1] using paired BAC-end sequences ([*O. nivara, O. rufipogon, O. punctata*][2] and [*O. barthii, O. glumaepatula, O. meridionalis, L. perrieri*] described herein) and blastn alignment to the Nipponbare RefSeq[3] as guide information. Final assembly statistics are summarized in **Supplementary Table 2**. GenBank WGS and INSDC accession numbers of finished assemblies are given in **Supplementary Note Table 3**.

Genome assembly of N 22 and IR 8.

High molecular weight DNA was extracted from young leaves from purified genetic stocks for N 22 (IRGC 117534, "N 22::IRGC 19379-1" and IR 8 (IRGC 125776, "IR 8::IRGC 10320-2") adopting a published protocol[20] with minor modifications. PacBio library preparation used the 20 kb protocol

(http://www.pacb.com/wp-content/uploads/2015/09/User-Bulletin-Guidelines-for-Preparing-20-kb-SMRTbell-Templates.pdf) followed by sequencing on a PacBio RSII sequencing instrument with movie collection times of 6 hours. Raw N 22 and IR 8 reads were assembled with FALCON[21] and Canu[22], respectively. N 22 contigs were polished twice with all PacBio raw reads using Quiver (https://github.com/PacificBiosciences/GenomicConsensus), and the IR 8 assembly was polished once with Quiver and once with 66x WGS 2x150 bp Illumina data using Pilon[3]. Polished contigs were assigned to pseudomolecules using Genome Puzzle Master[23] using the Nipponbare RefSeq[3] as a guide. Assembly breakpoints were always overlapping with regions of low PacBio coverage. Final assembly statistics are summarized in **Supplementary Table 2**. GenBank WGS and INSDC accession numbers of finished assemblies are given in **Supplementary Note Table 3**.

Plant material and methods for transcriptome sequencing and assembly.

Strand-specific RNA-seq libraries were prepared from leaf, root, and panicle tissues of thirteen species (**Supplementary Note Table 4**). Leaf tissue for library construction was collected from plants at the four-leaf stage. Young root tissue was collected from plants growing in hydroponic conditions to ensure clean and disease free samples. The panicles were collected at different stages of flower development (booting and panicle initiation stage). All freshly collected leaf, root, and panicle samples were wrapped in labeled aluminum foil and immediately immersed in liquid nitrogen, followed by long-term storage at -80C until further use. Total RNA was prepared using a TRIzol method (Invitrogen, USA), followed by mRNA isolation using Dynabeads (Invitrogen) following manufacturer's recommended protocols. The mRNA was sheared (200-400 bp) and reverse transcribed to prepare strand-specific Illumina PE cDNA libraries, which were subjected to dUTP excision and library enrichment. Short-read sequences were produced on an Illumina HiSeq-2000 system, yielding 36 to 258 million paired-end reads per library, as summarized in **Supplementary Table 16**. Sequence reads were deposited in the NCBI Sequence Read Archive (**Supplementary Note Table 4**). Transcripts derived from each tissue were assembled using both "reference-guided" and *de novo* approaches. For reference-guided assemblies, reads were aligned to the corresponding species genomic reference using TopHat, and transcripts were modeled using Cufflinks software[24]. *De novo* assemblies were performed using Trinity software version 2012-05-18 with default parameters[25]. Both assemblies were used as expression evidence to inform gene-annotation as described below. The *de novo* "Trinity assemblies" were used to

evaluate the completeness of the genome assemblies with respect to gene-space, as described below.

Evaluation of the wild reference assemblies for accuracy and completeness.

The completeness and quality of the seven wild genome assemblies were evaluated using multiple different approaches. As a first approximation of completeness, assembly length was measured against expectations of genome size, which we estimated using a combination k-mer frequency analysis within sequence read data, analysis of available physical map data, and from flow cytometry studies[4,5]. K-mer genome size estimations were calculated using the KmerFreq_AR tool within the SOAPec genome assembly software package[14] adopting a k-mer length of 17. Input data were single reads from paired-end Illumina reads deposited in NCBI for our 7 new genomes (**Supplementary Note Table 2**), plus the following additional NCBI data sets from previously published genomes: Nipponbare RefSeq[3] - DRX049066; *O. sativa* vg. indica - SRX321813, SRX321814; *O. brachyantha* - SRX099350, SRX099349, SRX099343 and *O. glaberrima* (R. Wing, unpublished). Genome size estimations based on physical maps were estimated by summing the consensus band units[26] (CB units) for each heavily manually edited physical map where 1 CB unit was the equivalent of between 1156-1359 bp (**Supplementary Table 4**).

To evaluate the integrity of each assembly we first assessed the quality of individual assembled contigs prior to scaffolding by mapping Illumina short-read PE reads to these contigs using BWA-mem (v0.7.15)[27]. SAMtools[28] was then used to calculate mapping metrics for each data set (**Supplementary Table** 3).

Next, we evaluated each genome assembly by mapping all paired-BESs to their corresponding assembled pseudomolecules using BLASTN with minimum alignment length >300 bp and sequence identity >95%. Paired-end mapping was scored as valid for those that aligned in opposite orientation and within a target range of 25-300 kb, and used to determine BAC locations and fold genome coverage. For each species, we produced an informative BAC list that includes BAC location (pseudomolecule & physical map), BES orientation, alignment strand etc. Visualized coverage data is provided as SVG files (**Supplementary Data 1**).

The assemblies were then evaluated for gene space and transposable element/repeat completeness. Gene space was first tested by scanning each genome assembly for the presence of core eukaryotic genes we using both the CEGMA[8] (v2.5) and BUSCO[9] pipelines. Second, we assessed mapping frequency of the Trinity-assembled transcripts to corresponding genome

assemblies. Expressed transposable elements were screened using RepeatMasker[29] with custom repeat libraries[6] and transcripts with >50% masking were eliminated. To exclude contaminating biological sources of transcripts, the sequences were screened against the complete NCBI RefSeq collection of protein and RNA sequences (RefSeq FTP release 69), which encompassed 32,606 organisms[30]. Transcripts were aligned to the complete protein database using the DIAMOND blastx command[31] with an e-value threshold of 1e-5. Sequences were also aligned to the complete RNA database using NCBI+ BLASTN (megablast method)[32] with an e-value threshold of 1e-5. Non-contaminating sequences were positively identified as those aligning to an *Oryza* RefSeq sequence as the top significant hit in either screen. Among the *Oryza*-positive transcripts, those that aligned to their own species' reference genome was determined using GMAP[107] with thresholds of 90% coverage and 90% identity. Multiple transcript isoforms assigned to the same Trinity cluster were counted as a single locus. The above protocol was performed individually with each tissue assembly (i.e. leaf, root, and panicle) for each species. Screens were also performed after collapsing the transcripts across the three tissues within each species to generate "unigenes". In this approach, centroid clustering of the pooled transcripts was performed using the USEARCH (v8.0.1616_i86linux32)[33] cluster_fast algorithm with "-sort length" and "-id 0.9" parameters (minimum 90% sequence match). To extend these analyses to consistency of gene annotation, the Trinity-assembled transcripts were also aligned to the entire set of predicted proteins across the eleven species using blastx (v2.2.28) (E-value < 1e-10) and assigned to 13,397 highly conserved ortholog sets (see below).

Transposable element and repeat abundance values were taken from Copetti and Wing 2016[6]. Briefly, two sets of low coverage (1.5 genome equivalents) single reads (of length between 76 and 100 bp) were generated. One set was composed of raw Illumina reads to represent the content of the native nuclear genome. The second set was obtained by producing simulated Illumina reads (of the same length as the raw reads) from the genome assembly to mirror the sequence content of the assembly. Both datasets were independently aligned to a curated repeat and TE library and the type and abundance of each repeat/TE class was determined by counting the hits to the library sequences. The differences in hit counts between the alignments to the native genome and to the assembly represented the unassembled fractions of repeats and TEs. The amounts of Mb missing were determined by multiplying such percentages for the estimated genome size.

To obtain finer measures of genome assembly completeness with respect to gene and repeat space we compared independently sequenced regions (~5 Mb) of the short arms of chromosome

3, as well as a collection of 40 finished BAC sequences as follows: Chromosome 3 short arm (chr3-sa) assemblies, available in eight species, provided an independently-derived data set with which to evaluate the whole genome assemblies. The chr3-sa assemblies were built from high-depth sequencing of bacterial artificial chromosomes (BACs) that were pooled along a minimum tiling path, as described by Rounsley *et al*., 2009[34] (see also Zwickl *et al*. 2014[35]). GenBank (INSDC) assembly accession numbers for the chr3-sa assemblies are as follows: GCA_000325765.2 (*L. perrieri*), GCA_000182155.1 (*O. barthii*), GCA_000710535.1 (*O. nivara*), GCA_000700045.1 (*O. rufipogon*), GCA_000710525.1 (*O. punctata*), GCA_000710545.1 (*O. brachyantha*), GCA_000338915.1 (*O. glumaepatula*), and GCA_000338895.1 (*O. meridionalis*). The chr3-sa assemblies were annotated for protein-coding genes and transposons using the same methods as described below for whole-genome assemblies. Non-masked chr3-sa sequences were aligned to respective whole-genome assemblies using LAST[36]. Whole genome assemblies were indexed using the command "lastdb -cR11 –uNEAR", which optimizes for the comparison of closely related species. Alignments were performed with the command "lastal -m50 -E0.05 | last-split -m1", which identifies the single best alignment at each position in the query relative to the whole-genome database. In addition to the minimum e-value score of 0.05, alignments were filtered for minimum match identity of 98% for all species except for *O. rufipogon*, which used 96% match identity cutoff. To control for the variable quality of assembly within each chr3-sa sequence, we selected a common region that aligned consistently well over a 5 Mb span (chromosome 3 position 1-6 Mb) of the Nipponbare RefSeq[3]. The corresponding regions subjected to coverage analysis in each of the chr3-sa assemblies are as follows: 797,149-5,484,804 (*L. perrieri*), 794,017-5,053,691 (*O. barthii*), 519,869-4,573,476 (*O. brachyantha*), 741,307-5,262,481 (*O. glumaepatula*), 744,011-5,541,804 (*O. meridionalis*), 814,226-5,265,299 (*O. nivara*), 827,154-6,276,324 (*O. punctata*), and 469,396-4,852,874 (*O. rufipogon*). The BEDTools suite was used to merge overlapping annotated features of each class (gene, exon, CDS, intron, 5'-UTR, 3'-UTR, intergenic region, and transposon) in the chr3-sa assemblies. The BEDTools intersect[37] command was used to find overlaps between annotated features and regions that aligned to the whole genome assembly, and alignment coverages were calculated as the fractions of summed annotated features over aligned features.

Forty-four finished BAC sequences from four species were downloaded from NCBI (one *O. barthii*, five *O. nivara*, six *O. rufipogon*, and 32 *O. punctata*). Genes and repetitive sequences were annotated using the same methods as described for the whole-genome sequences (see below). Non-masked BAC sequences were aligned to respective whole-genome assemblies using LAST[36],

as described above for chr3-sa assemblies. In addition to the minimum e-value score of 0.05, alignments were filtered for minimum match identity of 98% for all BACs except those of *O. rufipogon*, which required lower thresholds because the BACs and whole-genome sequences were derived from different biological accessions of this species. For five of the six *O. rufipogon* BACs, a minimum match identity of 96% was used, and for one BAC (GenBank accession FJ581045.1, the Hd1 locus), it was necessary to reduce the match identity to 90%. The BEDtools suite[37] was used to find the intersection of merged features with aligned regions within each BAC. Alignment coverage were calculated as the fractions of the summed feature lengths that aligned to the whole genome assemblies.

Annotation of protein-coding and lincRNA genes of the wild *Oryza* and *L. perrieri* genome assemblies

Annotation of protein-coding loci: Protein-coding genes were annotated using the MAKER-P v2.30 annotation engine[38], incorporating expression evidence, homology, and *ab initio* prediction methods. The pipeline performed repeat masking[39] of genome assemblies using curated DNA and protein libraries of transposable elements that were annotated in each species (see methods and Copetti *et al.*[10]). Expression evidence included RNA-seq data from three tissues in each species, assembled as transcripts using both *de novo* and reference-guided methods (see above; **Supplementary Tables 17, 18 and 22**). Additional expression evidence included *Oryza* full-length cDNA and mRNA sequences downloaded from NCBI using the following two queries: 1) txid4527[organism] AND biomol_mrna[prop], and 2) txid4527[organism] AND FLI_cDNA[keyword]. From these, we excluded NCBI RefSeq accessions, as they were derived from annotated genomic sequences, leaving 61,203 unique sequences from cloned mRNA. Homology-based evidence included annotated gene models (CDS and protein) of rice, which combined non-redundant RAP-DB and MSU7.0 annotations of the Nipponbare RefSeq (IRGSP1-0 reference assembly)[40]. We also included annotated gene models of *O. glaberrima* (AGI1.1 MIPS)[41] and *Brachypodium distachyon* (v1.0 JGI)[42]. The MAKER-P automated pipeline performed Exonerate[43] and NCBI+ BLAST[44], applying default thresholds for various alignment parameters as specified in the control file, maker_bopts.ctl. The reference-guided RNA-seq assemblies were inputted as GFF3 files. *Ab initio* predictions were performed using FGENESH software with the monocot-trained model[45], and provided as evidence to MAKER-P as an external GFF3 file. MAKER-P performed hint-based predictions using SNAP software with the O.sativa.hmm model[46], producing optimized

gene models scored by annotation edit distance. Predicted loci were screened for transposon sequences by alignment to curated transposon libraries. Protein sequences were annotated using InterProScan 5 software to find protein functional domains and to assign Gene Ontology (GO) terms[47]. This enabled additional removal of 967 transposon genes on the basis of signature InterPro domains[48], which included, IPR000477, IPR001207, IPR001584, IPR002559, IPR004242, IPR004252, IPR004264, IPR004330, IPR004332, IPR005063, IPR005162, IPR006912, IPR007321, IPR013103, IPR013242, IPR014736, IPR015401, IPR018289, IPR026103, IPR026960, IPR027806.

Consistency of protein-coding gene annotations across species was evaluated using 13,397 highly conserved ortholog sets expected to have membership in all *Oryza* species. To construct these sets, all pairwise ortholog assignments from Compara gene trees (see below) were grouped by single-linkage clustering. Highly conserved sets were selected as having representatives in both of *L. perrieri* and *O. sativa* vg. japonica, and as belonging to gene families conserved in *Arabidopsis*.

Evaluation of gene expression in annotated loci: To evaluate annotations with respect to gene expression RNA-seq reads from three tissues (**Supplementary Note Table 4**) were aligned to the longest representative CDS using Bowtie 2[49], with parameters '-a -X 1000 --rdg 6,5 --rfg 6,5 --score-min L,-.6,-.4 --no-discordant --no-mixed'. Fragments aligned to each CDS were counted using eXpress software[50] with default parameters. Expression evidence was indicated if the lower bound of the 95% confidence interval for the FPKM was greater than zero.

Long intergenic non-coding RNA (lincRNA) loci: RNA sequence reads (**Supplementary Note Table 4**) were mapped to the genomes of individual species using TopHat v2.0[51] (mate inner distance = 60, minimum intron length = 15, maximum multihits = 1, minimum segment intron =15, mate distance standard deviation = 50, all other options default). Mapped sequence fragments were assembled into leaf and panicle transcriptomes using Cufflinks v2.2.1 (min-intron-length = 15, overlap-radius = 20), and a single merged transcriptome was generated for each species using Cuffmerge[24]. Transcripts were classified into classes using Cuffmerge[24] relative to the MAKER-P[38] annotated protein-coding genes. The sets of unannotated, intergenic transcripts (Cuffmerge class code "u") in each species were first purged of tRNAs and rRNAs identified from the Rfam database using cmscan in Infernal v1.1[52,53] (E-value < 0.01). Protein-coding ability in all transcripts was assessed using: 1) blastx v2.2.28 to the NCBI nr database (E-value < 1e-10) and 2) CPAT v1.2.1 (protein-coding training set = *O. sativa.* vg. japonica mRNAs as identified by MAKER-P[38]

annotation, non-coding training set = *O. sativa* vg. japonica UTRs longer than 300 nt, executed only on sense strand for transcripts with known orientation and both strands for transcripts with unknown orientation, 98 % sensitivity coding potential cutoff = 0.522)[54,55]. Loci with only transcripts that lack matches to rRNAs, tRNAs, or any protein-coding hit from blastx or CPAT were classified as putative lincRNA loci. Chi-square tests were performed in R[56], with fdr post-hoc analyses performed using the chisqPostHoc utility[57].

Genome annotation of N 22 and IR 8.

Gene models were predicted with MAKER-P (v2.31.8)[38], using RNA-Seq data from[58] and PacBio Iso-Seq data for *Oryza sativa* vg. indica cultivars Minghui 63 and Zenshan 97 (J. Zhang and R.A. Wing et al, unpublished data). *Ab initio* prediction of gene models was carried out with SNAP (v. 2006-07-28)[46] and Augustus (v. 3.1)[59]. Gene models containing TE domains or overlapping for more than 40% of their length to known TEs were filtered. The start of the CDS of genes not starting with an ATG were edited to the first in-frame ATG, genes without Rfam domains were removed. The fraction of the conserved core genes was assessed with CEGMA[8] and BUSCO[9], for the latter selecting the plant gene dataset and rice as the species for the Augustus gene models. Infernal[52,53] was adopted to identify non-coding RNAs (ncRNAs) using the Rfam library Rfam.cm.1_1. Hits above the e-value threshold of 1e-5 were filtered, as well as results with score lower than the family-specific gathering threshold. When loci on both strands were predicted, only the hit with the highest score was kept. Transfer RNAs were also predicted using tRNAscan-SE[60] at default parameters. The baseline repeat annotation of the assembly was obtained by merging the output of RepeatMasker[29] (http://www.repeatmasker.org/, v. 3.3.0) and Blaster (a component of the REPET[61] package). The two software packages were run using nucleotidic libraries (PReDa and RepeatExplorer) from RiTE-db[10] and an in-house curated collection of transposable element (TE) proteins, respectively. Reconciliation of the masked repeats was carried out using custom Perl scripts and formatted in .gff3 files. Gene and repeat annotations are provided in **Supplementary Data 1**.

Supplementary Note: Phylogenetic Inference

RESULTS

Ortholog clusters and alignments: Applying the BOS protocol to each chromosome resulted in a total of 6015 single-copy clusters containing single sequences from all 11 genomes. The number of clusters per chromosome varied roughly in proportion to chromosome size, e.g., ranging from 1042 alignments on chromosome 1 to only 142 on chromosome 11 (**Supplemental Table 26**). Lengths of the per-chromosome concatenated supermatrices ranged from ~2 million bp for chromosome 1 to ~280,000 for chromosome 11.

Supermatrix analyses: Supermatrix analyses inferred the same phylogeny for all chromosomes except chromosomes 6 and 12 (**Fig. 1** and **Supplemental Fig**. **4**; individual supermatrix trees appear in **Supplemental Fig. 5**). With a few exceptions, bootstrap support values were 100% for all clades for each chromosome. Two relationships received lower support on some chromosomes: the sister relationship between *O. sativa* vg. indica [93-11] and *O. nivara* (67% support from chromosome 6) and the monophyly of the African+Asian clade within the AA genome group (support 50%, 60% and 0% from chromosomes 6, 10 and 12, respectively). When the African+Asian clade did not receive 100% support, the alternate resolution placed *O. glumaepatula* sister to the pair of African AA species, *O. barthii* and *O. glaberrima*. On chromosome 12 this alternative received 100% bootstrap support.

Species tree analyses: MP-EST analyses gave high support for all branches of the species tree in **Figure 1** for almost all chromosomes. The only exceptions were nodes that were also poorly supported by the supermatrix analyses: low support for the *O. sativa* vg. indica [93-11] and *O. nivara* sister relationship from chromosome 6 (21%), and low support for the African-Asian clade from chromosome 12 (21%).

Oryza divergence time estimation: Mean divergence times estimated from the ML supermatrix trees using PATHd8 appear in **Supplemental Fig. 4,** as well as the range in estimates across chromosomes. Estimates were generally consistent across chromosomes, with a few outliers (**Supplementary Table 27**). In particular, chromosomes 6 and 11 resulted in much older

divergence time estimates for several nodes within the AA genome group. Most of the nodes with extreme dates are near nodes having lower support in the supermatrix bootstrap analyses.

Supplementary Note: Concerted Evolution

RESULTS

The duplicated nature of the rice genome (*Oryza* sativa vg. japonica and vg. indica) was deduced from analysis of the first genome sequences[62] and dated to 60-70MYR, suggesting that this duplication is common to all cereal genomes. However, one limited region in the subtelomeric region of chromosomes 11 and 12 is so highly conserved that it attracted attention and speculation ever since the first high-density genetic maps were established[63]. Genome sequence analysis confirmed its existence and allowed more detailed studies[62,64]. Depending on the approach used and the region considered (2.5 Mb-6.5 Mb), authors dated the duplication between 5 MYR and 25 MYR, but always placed it within the *Oryza* genus[37,62,65-67], suggesting that it would be absent from the basal species in the genus. The presence of a similarly highly-conserved duplication located in orthologous positions on chromosomes 5 and 8 of the sorghum genome[68] led the authors to conclude that this duplication in fact derives from the ancestral cereal WGD and proposed that such high sequence identity was maintained by concerted evolution of gene sequences. Jacquemin *et al*. [69] analyzed selected gene pairs from representative species within the *Oryza* genus and in two closely-related outgroups, concluding that recurrent concerted evolution had occurred throughout the *Oryza* genus and in the outgroup species. They further demonstrated the presence of conserved gene pairs in the orthologous regions of the *Brachypodium distachyon* genome and suggested that the lower conservation in sorghum and *B. distachyon* is probably due to genome rearrangements (a large inversion in sorghum chromosome 8 and nested chromosome fusion in Brachypodium), which have made further large-scale conversion impossible. In the maize genome, which has undergone an independent whole genome duplication, the four regions orthologous to the rice chromosome 11/12 blocks are no longer subtelomeric and have been rearranged to the extent that they are barely detectable. In contrast, in foxtail millet (*Setaria italica*), where the duplicated blocks are still in subtelomeric positions, we find evidence for a recent conversion event, supporting the hypothesis that a subtelomeric location is necessary for the large-scale conversion to occur[70] and our unpublished results. Jacquemin *et al*.[69] further demonstrated that sequence conservation between the two chromosomes in this region is not limited to the genic sequences but concerns both coding and non-coding sequences. They proposed double strand break repair or break induced repair, both of which have been described

as putative models of formation of segmental duplications[71], as mechanisms potentially leading to a large-scale conversion event extending from ~2 Mb to the end of the chromosome.

The availability of complete genome sequences for 9 *Oryza* species and *Leersia perrieri* allowed us to study the occurrence and frequency of conversion events over a period of ~15 MY. Preliminary analyses using a modified version of the Jacquemin *et al*.[69] analysis pipeline demonstrated that no regions other than the ~2 Mbp of chromosomes 11 and 12 showed similar conservation. We therefore concentrated our analyses on these two chromosomes. Analysis was carried out for AA (*O. sativa*, *O. rufipogon*, *O. glaberrima* and *O. barthii)*, BB, FF and *L. perrieri* genomes. Sequences from chromosomes 11 and 12 were aligned and analyzed (**Supplemental Fig. 6**). Gaps in the distribution of Bayesian distances correspond either to missing sequence (probably mostly at the beginning of the sequence), insertions/deletions in one or several sequences or unassembled regions (Ns). Particularly high values for Bayesian distances may be due to poor quality of alignment or short sequences for one or several chromosomes in these blocks. Analysis of sequence alignments >300 bp gave 5782 phylogenetic trees containing both sequences from at least five of the seven species and these were used for calculation of intra- and inter-species Bayesian distances. *O. sativa* chromosome 11 was used as a reference. 3378 trees (61% of all trees) correspond to blocks within the first 2.1 Mb (only ~7% of the complete chromosome sequence), emphasizing the highly-conserved nature of this duplication. This is further demonstrated by analysis of the alignments for *O. sativa* chromosomes 11 and 12 (which have the highest quality sequence) and which show that ~70% of each of the sequences are found in aligned blocks >300 bp in the first 2 Mb, although the ancestral duplication occurred some 60-70 MYR. Surprisingly, and despite the independent conversion events in different lineages, the breakpoint between the highly-conserved region and the rest of the chromosome is found in the same region in all species (**Supplemental Fig. 6b**), possibly corresponding to a region of high recombination frequency[69]. Whereas calculated inter- and intra-species divergence times for the converted region are in good agreement with previously-published estimates, those calculated for the rest of the chromosome are lower than expected. This could simply reflect the much lower number of trees available for the analysis, but is probably due to the fact that the conserved sequences mostly correspond to coding sequences in non-conserved regions and our method of dating will therefore underestimate divergence times. However, we cannot exclude the possibility

that longer conversion events occurred in the past, although there is no evidence of this from the distribution of Bayesian distances for *O. sativa* in **Figure 6a**.

Analysis of 7 AA genome sequences (*O. sativa* [Nipponbare RefSeq], *O. rufipogon*, *O. barthii*, *O. glaberrima*, *O. nivara*, *O. glumaepatula* and *O. meridionalis*) using the *O. punctata* BB sequence as outgroup gave 5504 trees derived from the 300 bp aligned blocks. Among these, 3121 (57%) fall within the first 2 Mb of the chromosomes. The calculated conversion dates except for *O. meridionalis* (**Supplementary Table 29**), where the apparent intra-species divergence date (3.69 MYR) is close to the calculated average divergence for this region of chromosomes 11 and 12 between *O. meridionalis* and *O. sativa* (3.62 MY), are all anterior to the calculated date of divergence of the AA genome group from the BB genome. This suggests that this common event may have occurred almost concomitantly with the divergence of the AA genome species. It is therefore logical to find no evidence of large-scale conversion in specific AA genome species. In agreement with this, the inter-species divergence dates (bottom of **Supplementary Table 29**) correspond well to those calculated for the various speciation events. We also looked without success for evidence of smaller-scale conversion tracts within individual AA genome species, although our approach was not developed to carry out analysis at such a small-scale level. Taken together, these results are consistent with concerted evolution of a 2.2 Mb region throughout the *Oryza* genus and in the outgroup *L. perrieri,* and suggest that large-scale conversion has been a frequent and recurrent event in the cereal lineage, as accumulated sequence divergence would in time decrease the probability of conversion.

Our previous results strongly suggested that a subtelomeric chromosomal arrangement is necessary for the conversion event to occur. They also demonstrated that this genome organization is not essential for survival of species, as the majority of cereals have undergone large-scale genome rearrangements. In sorghum, collinearity between chromosomes 5 and 8 has been interrupted by a 0.8 Mb inversion on chromosome 8. Analysis of gene pairs in the regions orthologous to the rice duplication showed higher sequence conservation between the telomere and the beginning of the inversion than within or beyond it, suggesting that further conversion was blocked by this rearrangement, although the number of genes analyzed is insufficient to obtain statistically significant results (our unpublished data). In *Brachypodium,* ancestral chromosomes 11 and 12 have fused with ancestral chromosome 9 to form *Brachypodium*

chromosome 4, on which the ancestral, duplicated subtelomeric regions now form an internal, inverted repeat, which certainly prevents large-scale conversion. The only species in which the two blocks are conserved in a subtelomeric location is foxtail millet, although in this species the region from the millet chromosome orthologous to rice chromosome 12 (chromosome 3) has translocated to the end of chromosome 7[70]. The translocation breakpoint appears to coincide with the conserved 7 observed in all the *Oryza* species. A conversion event can be detected in these orthologous regions of millet which we have dated to ~4.5 MYA (our unpublished data), although we cannot determine if the translocation or the conversion event occurred first. It should be noted, however, that rice and its relatives are the only cereals to have conserved the ancestral cereal genome structure of 12 chromosome pairs derived from the WGD.

If the recurrent conversion of this region is unnecessary, why does it occur with such frequency in the *Oryza* genus and in closely-related species? Wang *et al.*[72] suggested that the presence of NBS-LRR resistance genes on chromosomes 11 and 12 may play a role. However, the majority of these genes are not in the conserved regions. Neither is there any obvious selection for particular functions or GO categories[64,70]. One possibility would be that the region harbors one or several genes for which certain modifications of coding or regulatory sequences could be deleterious. This would lead to a selective pressure to fix a genome structure in which these gene sequences remain identical. In this case, the genome rearrangements which have led to the loss of one copy of conserved pairs in other species may have occurred almost concurrently with speciation events, precluding further conversion. Intriguingly, the DMC1 gene, which is present as a conserved pair at the breakpoint in rice, is present in only one copy in all other sequenced cereal genomes except sorghum, in which the two copies are located on chromosomes 4 and 8. However, the simplest explanation may be that this is a random process and that the *Oryza* and millet genomes have not yet undergone rearrangements that would make the conversion impossible. Finally, the question may be wider than this, as the *Oryza* species and their close relatives are also those that have conserved the 12 chromosome pair genome complement since the cereal ancestor some 60-70 MYA. As yet unknown mechanisms may function to maintain this genome structure, which facilitates frequent gene conversion.

Supplementary Notes REFERENCES

1 Zhang, J. *et al.* Genome Puzzle Master (GPN) - An integrated pipeline for building and editing pseudomoleclues from fragmented sequences. *Bioinfo.* **32**, 3058-3064 (2016).

2 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**, R25 (2008).

3 Matsumoto, T. *et al.* The map-based sequence of the rice genome. *Nature* **436**, 793-800 (2005).

4 Ammiraju, J. S. S. *et al.* The *Oryza* BAC resource: A genus-wide and genome scale tool for exploring rice genome evolution and leveraging useful genetic diversity from wild relatives. *Breeding Sci.* **60**, 536-543 (2010).

5 Ammiraju, J. S. 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**, 140-147 (2006).

6 Copetti, D. & Wing, R. A. The dark side of the genome: Revealing the native transposable element/repeat content of eukaryotic genomes. *Mol. Plant* **9**, 1664-1666 (2016).

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

8 Parra, G., Bradnam, K., Ning, Z., Keane, T. & Korf, I. Assessing the gene space in draft genomes. *Nucleic Acids Res.* **37**, 289-297 (2009).

9 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinfo,* **31**, 3210-3212 (2015).

10 Copetti, D. *et al.* RiTE database: A resource database for genus-wide rice genomics and evolutionary biology. *BMC Genomics* **16**, 538 (2015).

11 Kim, H. *et al.* Comparative physical mapping between *Oryza sativa* (AA genome type) and *O. punctata* (BB genome type). *Genetics* **176**, 379-390 (2007).

12 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinfo.* **30**, 2114-2120 (2014).

13 Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* **108**, 1513-1518 (2011).

14 Luo, R. *et al.* SOAPdenovo2: An empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1**, 18 (2012).

15 Simpson, J. T. *et al.* ABySS: A parallel assembler for short read sequence data. *Genome Res.* **19**, 1117-1123 (2009).

16 Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinfo.* **27**, 578-579 (2011).

17 Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome Biol.* **13**, R56 (2012).

18 Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380 (2005).

19 Sommer, D. D., Delcher, A. L., Salzberg, S. L. & Pop, M. Minimus: a fast, lightweight genome assembler. *BMC Bioinfo.* **8**, 64 (2007).

20 Doyle, J. J. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11-15 (1987).

21 Chin, C. S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050-1054 (2016).

22 Koren, S. *et al.* Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722-736 (2017).

23 Zhang, J. *et al.* Genome puzzle master (GPM): An integrated pipeline for building and editing pseudomolecules from fragmented sequences. *Bioinfo.* **32**, 3058-3064 (2016).

24 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotech.* **28**, 511-515 (2010).

25 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotech.* **29**, 644-652 (2011).

26 Soderlund, C., Humphray, S., Dunham, A. & French, L. Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res.* **10**, 1772-1787 (2000).

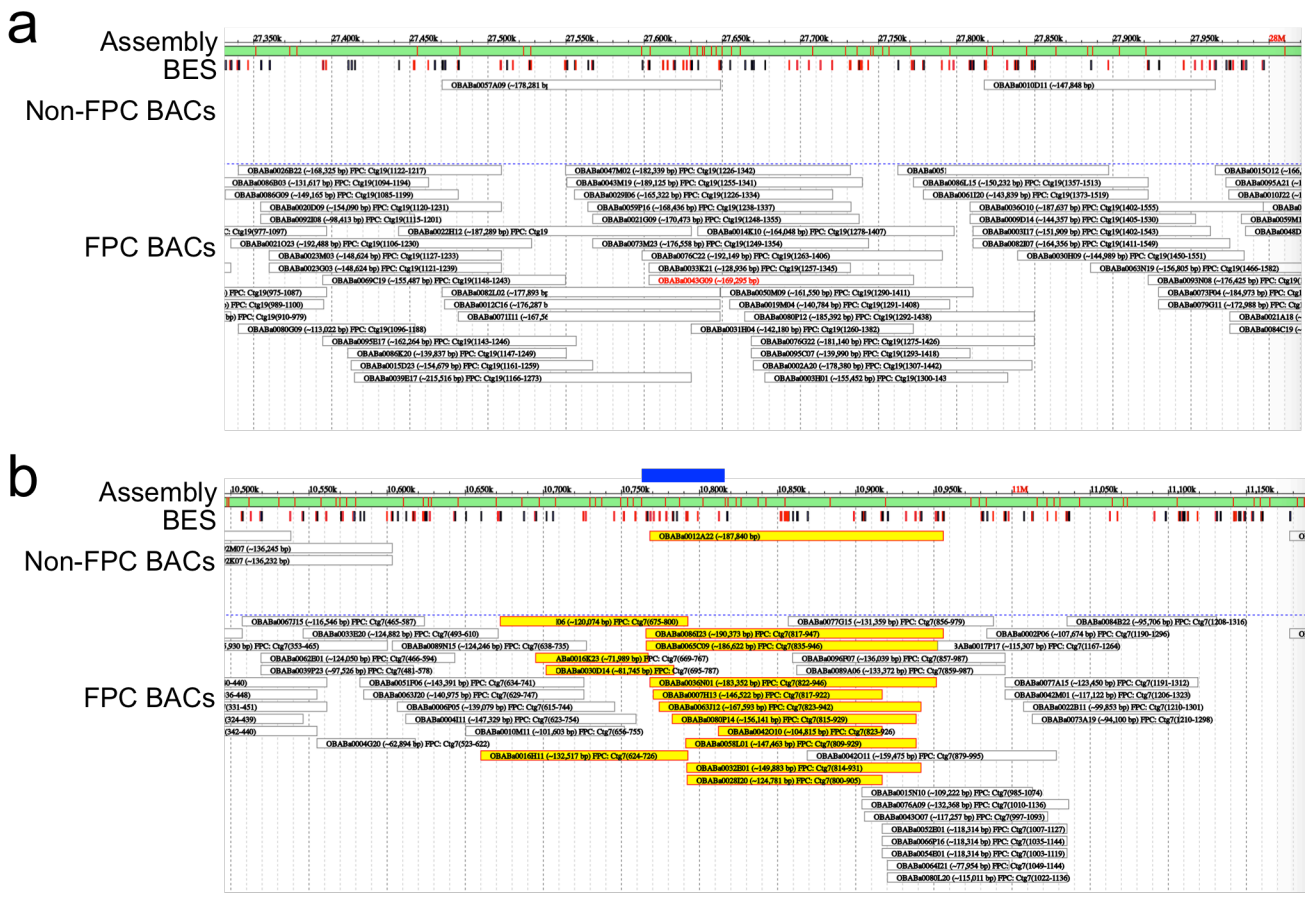
27 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinfo.* **26**, 589-595 (2010).

28 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinfo.* **25**, 2078-2079 (2009).

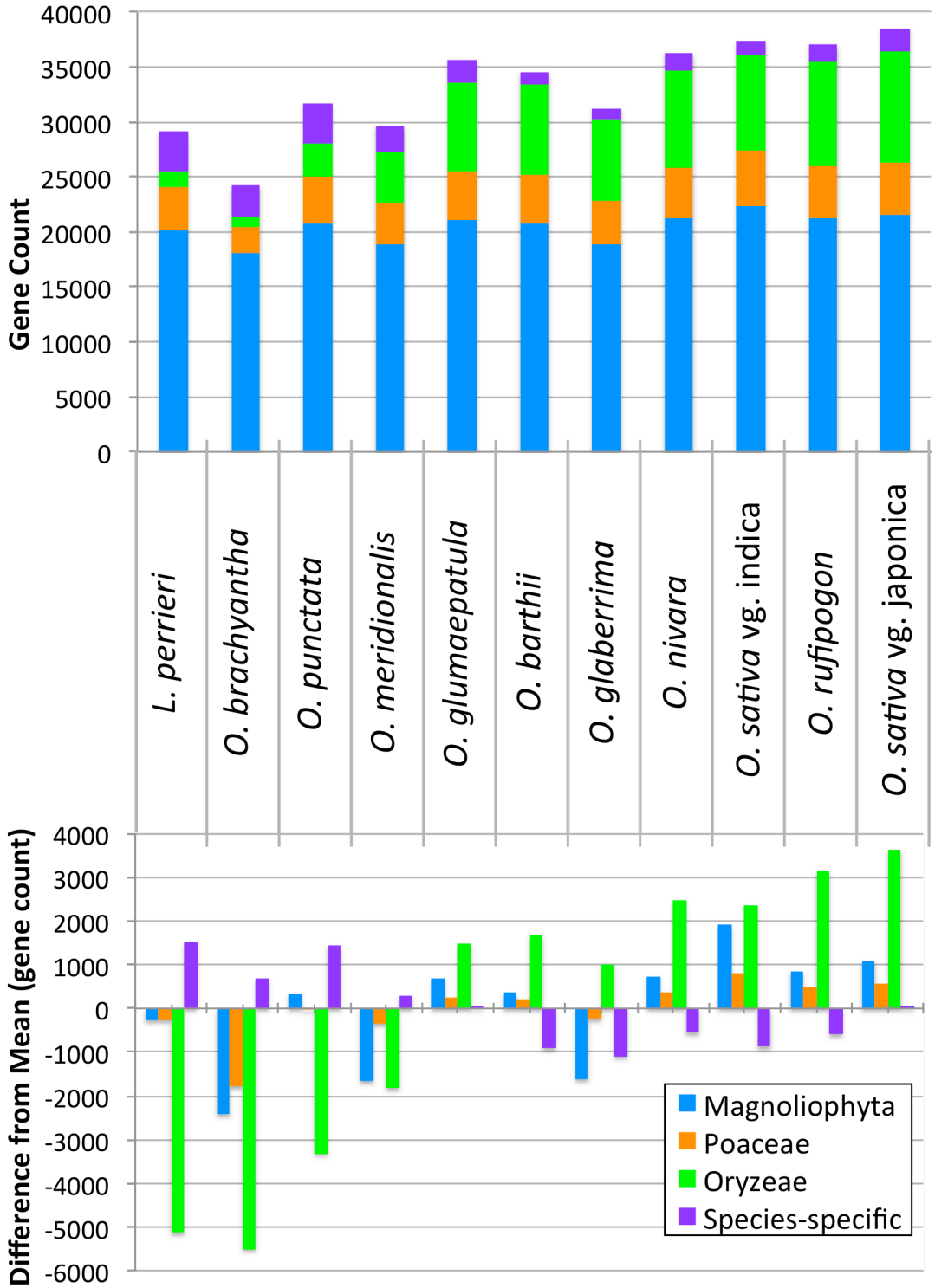
29 Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-3.0RepeatMasker Open-3.0*, http://repeatmasker.org (1996-2010).

30 McEntyre, J. & Ostell, J. *The NCBI Handbook [Internet]*, https://www.ncbi.nlm.nih.gov/books/NBK21101/?redirect-on-error=__HOME__&depth=2 (2002).
31 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59-60 (2015).
32 Morgulis, A. *et al.* Database indexing for production MegaBLAST searches. *Bioinfo.* **24**, 1757-1764 (2008).
33 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinfo.* **26**, 2460-2461 (2010).
34 Rounsley, S. *et al. De novo* next generation sequencing of plant genomes. *Rice* **2**, 35-43 (2009).
35 Zwickl, D., Stein, J., Wing, R., Ware, D. & Sanderson, M. Disentangling Methodological and Biological Sources of Gene Tree Discordance on Oryza (Poaceae) Chromosome 3. *Systematic Bio.* **63**, 645-659 (2014).
36 Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487-493 (2011).
37 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinfo.* **26**, 841-842 (2010).
38 Campbell, M. S. *et al.* MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Phys.* **164**, 513-524 (2014).
39 Smith, C. D. *et al.* Improved repeat identification and masking in Dipterans. *Gene* **389**, 1-9, (2007).
40 Kawahara, Y. *et al.* Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice (N Y)* **6** (2013).
41 Wang, M. *et al.* The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. *Nat. Genet.* **46**, 982-988 (2014).
42 Initiative, I. B. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763-768 (2010).
43 Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinfo.* **6**, 31 (2005).
44 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinfo.* **10**, 421 (2009).
45 Salamov, A. A. & Solovyev, V. V. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res* **10**, 516-522 (2000).
46 Korf, I. Gene finding in novel genomes. *BMC Bioinfo.* **5**, 59 (2004).
47 Jones, P. *et al.* InterProScan 5: Genome-scale protein function classification. *Bioinfo.* **30**, 1236-1240 (2014).
48 Schatz, M. C. *et al.* Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica. *Genome Biol.* **15**, 506 (2014).
49 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).
50 Roberts, A. & Pachter, L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods* **10**, 71-73 (2013).
51 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
52 Nawrocki, E. P. *et al.* Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* **43**, D130-137 (2015).
53 Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinfo.* **29**, 2933-2935 (2013).
54 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).
55 Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**, e74 (2013).
56 Carvalho, A. & Clark, A. Efficient identification of Y chromosome sequences in the human and *Drosophila* genomes. *Genome Res.* **23**, 1894-1907 (2013).
57 Carvalho, A. *et al.* Y chromosome and other heterochromatic sequences of the Drosophila melanogaster genome: how far can we go? *Genetica* **117**, 227-237 (2003).
58 Zhang, J. *et al.* Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. *Proc. Natl. Acad. Sci. USA* **113**, E5163-5171 (2016).
59 Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinfo.* **19**, 215-225 (2003).
60 Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955-964 (1997).
61 Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in *de novo* annotation approaches. *PLoS One* **6**, e16526 (2011).
62 Yu, J. *et al.* The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol* **3**, e38 (2005).
63 Nagamura, Y. *et al.* Conservation of duplicated segments between rice chromosomes 11 and 12. *Breeding Sci.* **45**, 373-376 (1995).

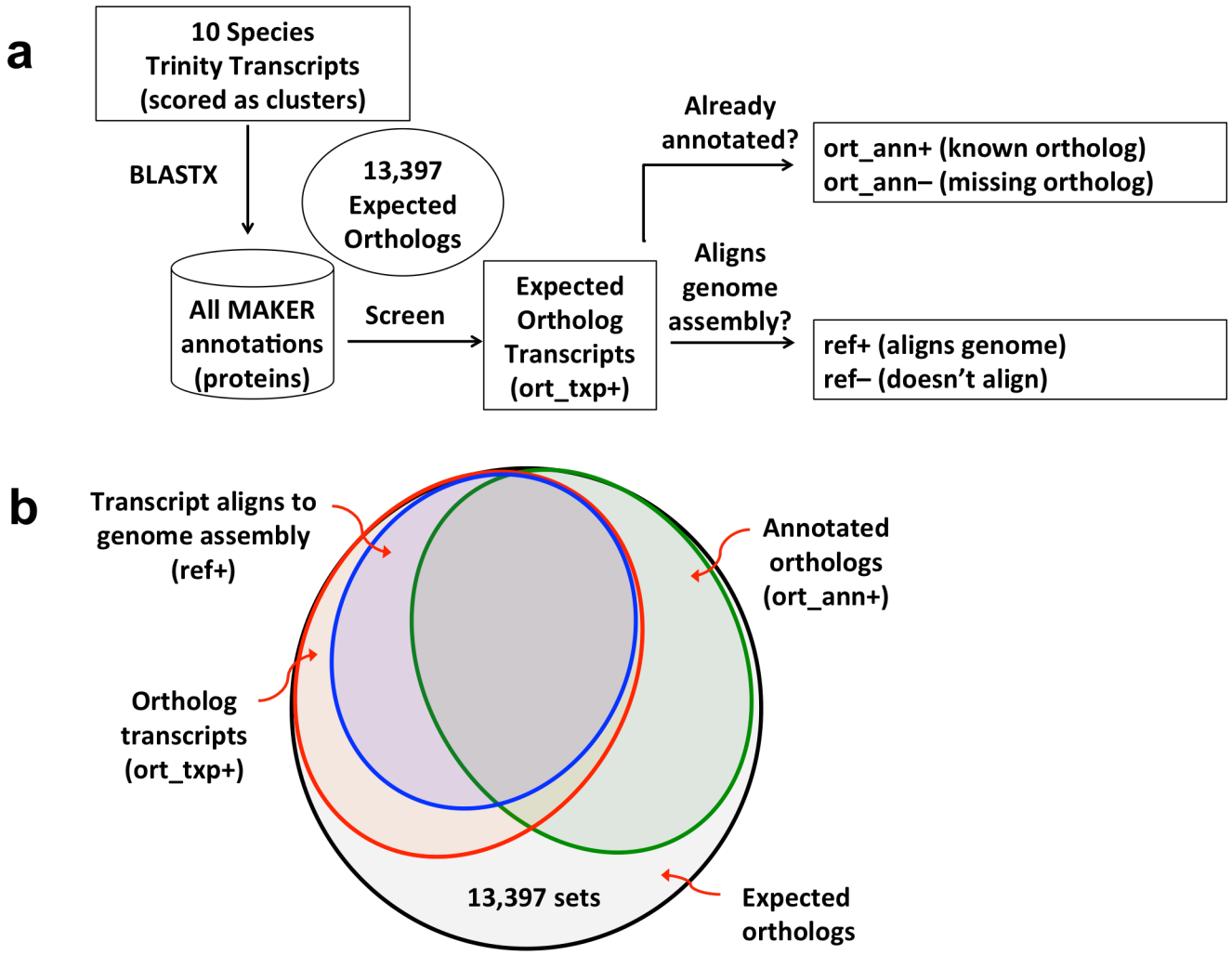
64 The Rice Chromosome 11 and 12 Sequencing Consortia. The sequence of rice chromosomes 11 and 12, rich in disease resistance genes and recent gene duplications. *BMC Biol.* **3**, 20 (2005).
65 Goff, S. A. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**, 92-100 (2002).
66 Wang, X., Shi, X., Hao, B., Ge, S. & Luo, J. Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytol.* **165**, 937-946 (2005).
67 Salse, J. *et al.* Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* **20**, 11-24 (2008).
68 Paterson, A. H. *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551-556 (2009).
69 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**, 3139-3150 (2011).
70 Murat, F. *et al.* Shared subgenome dominance following polyploidization explains grass genome evolutionary plasticity from a seven protochromosome ancestor with 16K protogenes. *Genome Biol. Evol.* **6**, 12-33 (2014).
71 Koszul, R. & Fischer, G. A prominent role for segmental duplications in modeling eukaryotic genomes. *C. R. Biologies.* **332**, 254-266 (2009).
72 Wang, X., Tang, H. & Paterson, A. H. Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major Poaceae lineages. *Plant Cell* **23**, 27-37 (2011).



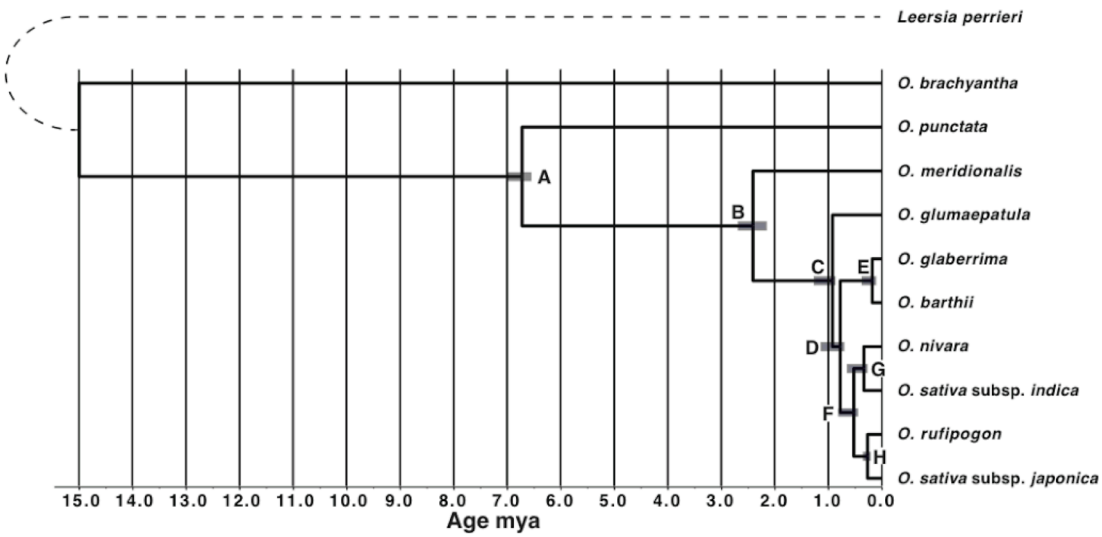
Supplementary Figure 1. Evaluation of contig order and orientation within reference assemblies using paired BAC-end sequences (BES). Figure illustrates two 650 kb regions on chromosome 1 of the *O. barthii* reference assembly. **a,** No conflicts found. **b,** Region showing conflicting paired BES alignments, suggesting incorrect orientation of four contigs under the blue bar as a possible interpretation. For each panel, tracks show the reference assembly at top with red bars indicating gaps between green-colored contigs, followed by the position of BES alignments, with black lines indicating forward and red lines indicating reverse orientation, and the bottom tracks showing BACs whose sequenced ends aligned within a 25-300 kb range. Both fingerprinted and non-fingerprinted BACs were used. BACs are highlighted in yellow when their alignment conflicts with the reference assembly. BACs with red labels are singletons within the fingerprinted contig (FPC) map. Images were excerpted from supplementary data files providing data and images for alignments of BES against seven reference assemblies (**Supplementary Data 1**).



Supplementary Figure 2. Counts of protein coding genes annotated in 11 *Oryzeae* species and categorized according to species-specificity at the level of gene family. "Magnoliophyta" are most conserved, having homologs in Arabidopsis thaliana; "Poaceae" are conserved in *Brachypodium distachyon* and/or *Sorghum bicolor*; "Oryzeae" are conserved between two or more species reported here; "Species-specific" have within-species homology or show no homology. Top panel: stacked bar chart shows counts in each category. Bottom panel: difference in gene count for a species compared to the average count across all species, with negative values indicating fewer genes in the species compared to average.



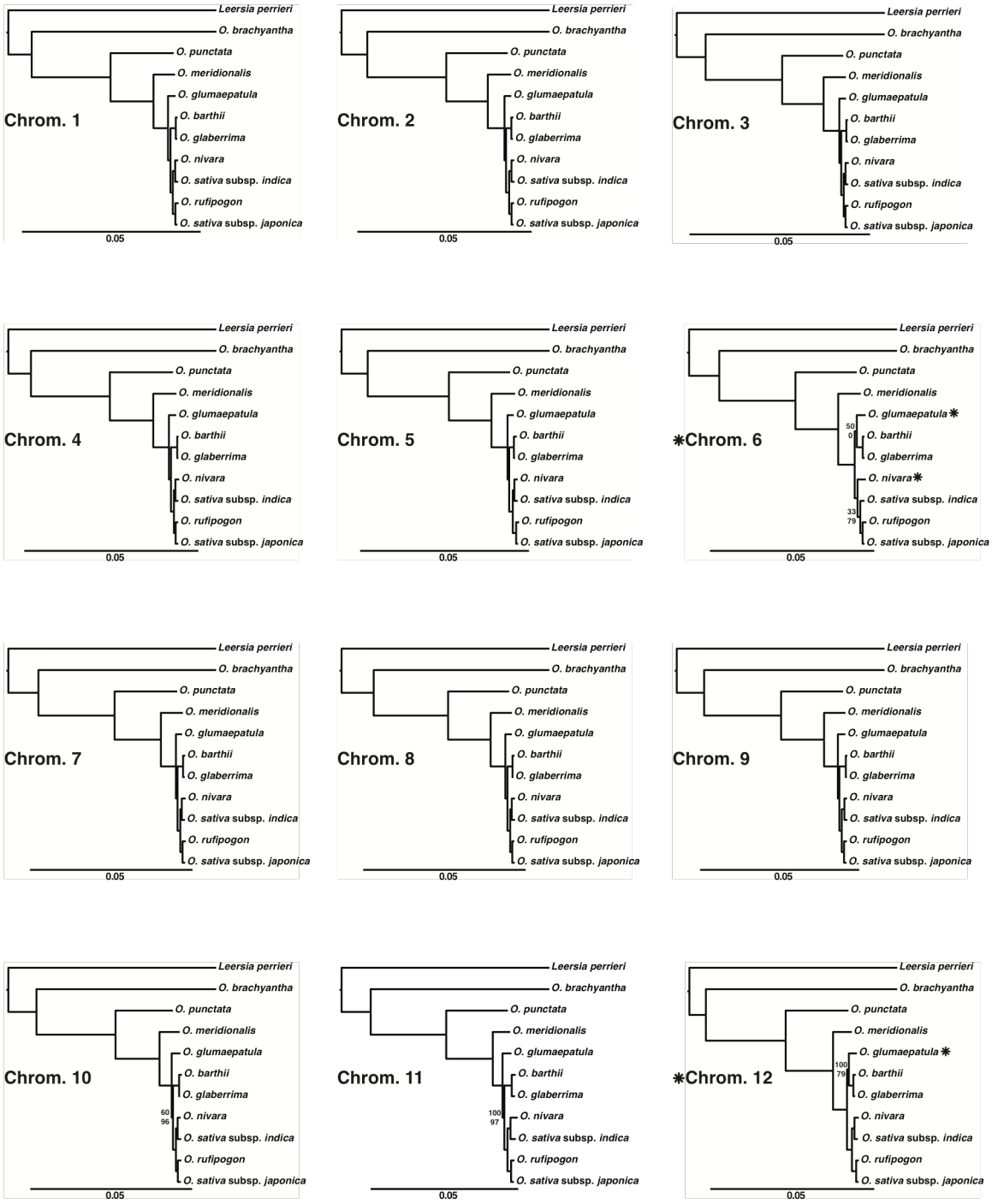
Supplementary Figure 3. **Screen to detect expected orthologous genes within transcriptome data in ten species and their classification with respect to gene annotation and presence in genome assemblies. a,** Schematic flow-chart of protocol: 13,397 sets of orthologous genes were identified as conserved in Arabidopsis, and having representatives in annotated genes of *L. perrieri* and *O. sativa* vg. japonica, and thus expected to be present in the other *Oryza* species. For each species, each ortholog set was scored as present in the gene annotations (ort_ann+) or absent (ort_ann-). Transcripts were assigned to ortholog sets by BLASTX alignment to annotated proteins, and if found the ortholog set was scored as ort_txp+. Transcripts were also scored as either aligning to the reference genome assembly (ref+) or not (ref-). **b,** Venn diagram illustrating classification of orthologous sets with respect to annotation, presence in transcriptome data, and presence in genome reference assemblies. An ortholog identified in transcriptome data but not in annotations or the genome assembly implies a gap in the assembly. An ortholog detected in transcriptome data and in the genome assembly, but not in gene annotations implies the possibility of false-negative annotation.



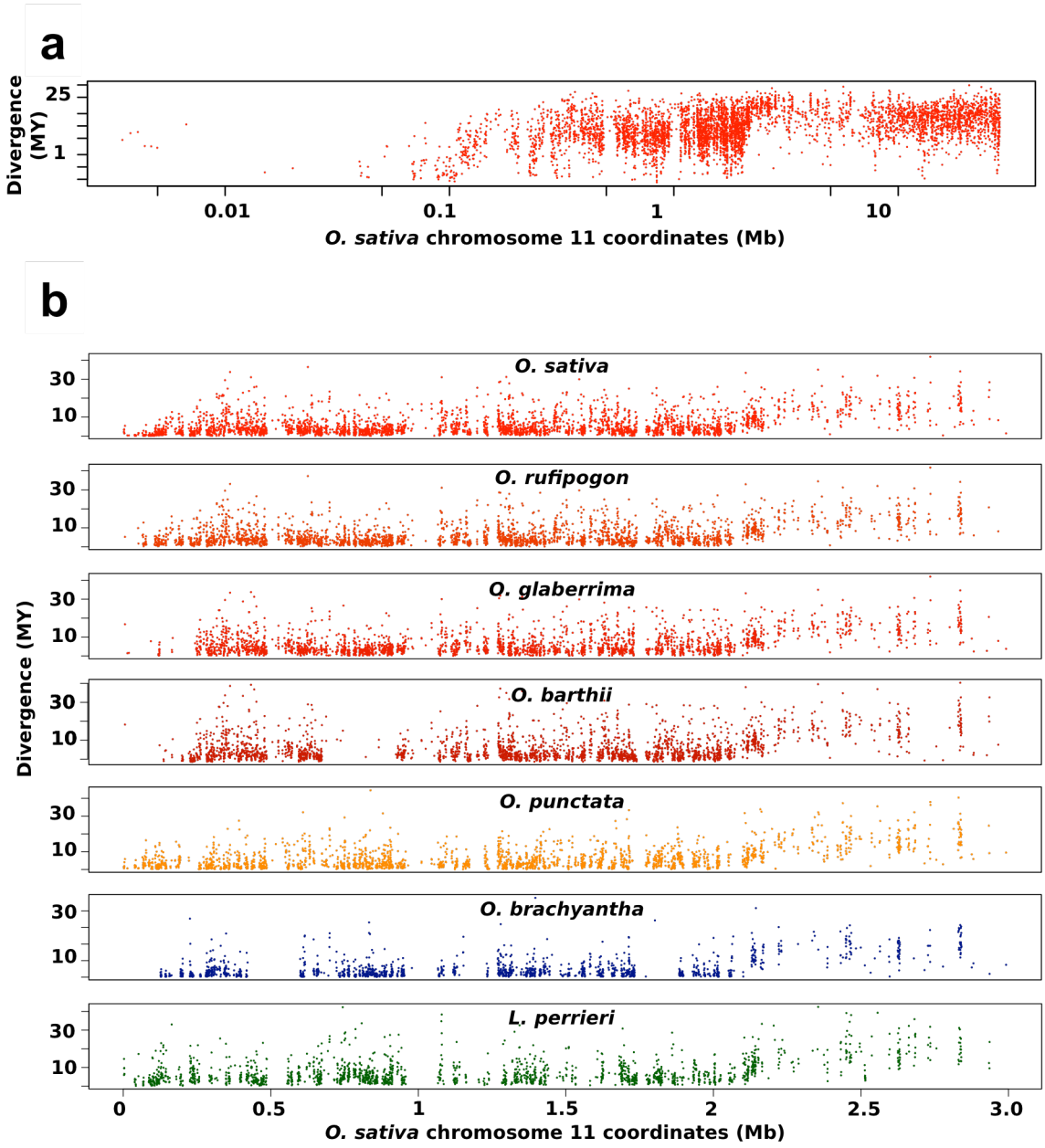
Chromosome	Node							
	A	B	C	D	E	F	G	H
1	6.67	2.39	0.97	0.75	0.17	0.54	0.31	0.27
2	6.68	2.56	0.90	0.78	0.19	0.50	0.30	0.23
3	6.71	2.48	0.89	0.72	0.17	0.48	0.31	0.27
4	6.74	2.47	0.92	0.75	0.12	0.46	0.34	0.26
5	6.63	2.49	0.93	0.82	0.20	0.53	0.33	0.24
6	6.93	2.68	1.06	NA	0.18	NA	0.79	0.28
7	6.98	2.39	0.90	0.78	0.20	0.52	0.36	0.30
8	6.83	2.43	0.93	0.74	0.16	0.48	0.34	0.24
9	6.70	2.32	0.92	0.79	0.13	0.53	0.29	0.28
10	6.57	2.17	0.95	0.82	0.19	0.52	0.34	0.34
11	6.90	2.20	1.26	1.14	0.37	0.81	0.64	0.32
12	6.83	2.30	1.02	NA	0.21	0.64	0.38	0.25
mean (standard error)	6.76 (0.13)	2.41 (0.15)	0.97 (0.11)	0.81 (0.12)	0.19 (0.06)	0.55 (0.10)	0.39 (0.16)	0.27 (0.03)

"NA" indicates a node not appearing in the ML supermatrix phylogeny inferred for a given chromosome.

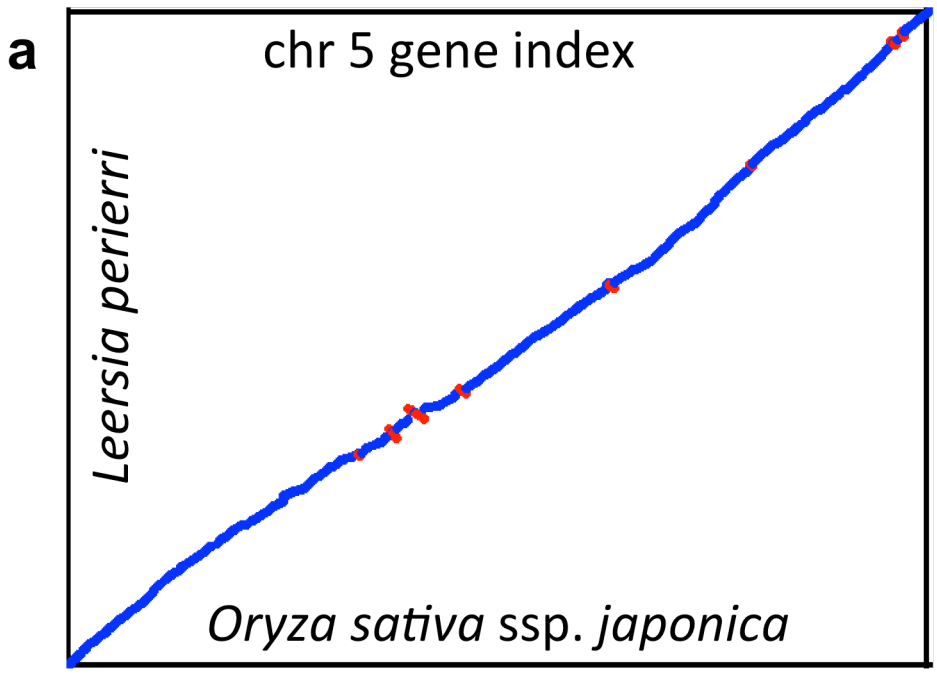
Supplementary Figure 4. Species phylogeny estimated by both supermatrix and MP-EST analyses of each chromosome, with divergence times estimated by PATHd8. Bars on nodes indicate range of PATHd8 age estimates across all 12 chromosomes. Phylogenies estimated from chromosomes 6 and 12 supermatrices did not contain exactly the same set of nodes to be dated.

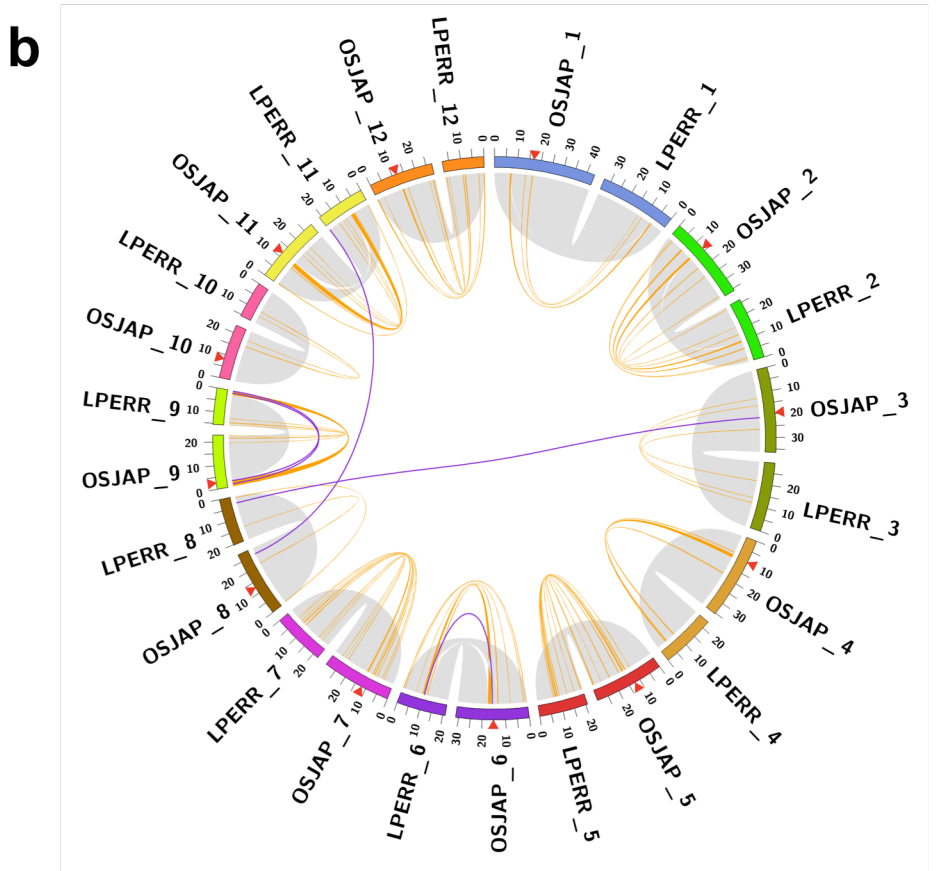


Supplementary Figure 5. **Maximum likelihood supermatrix phylogenies for each chromosome, with branch lengths proportional to the number of substitutions per site.** All supermatrix tree topologies identical to that in Fig. 1 except those denoted with ∗ (offending taxa also indicated by ∗). MP-EST tree topologies were identical to supermatrix topologies for all chromosomes except 6. Support values shown for clades that did not receive 100% support from both bootstrap and MP-EST analyses (bootstrap support above, MP-EST support below).

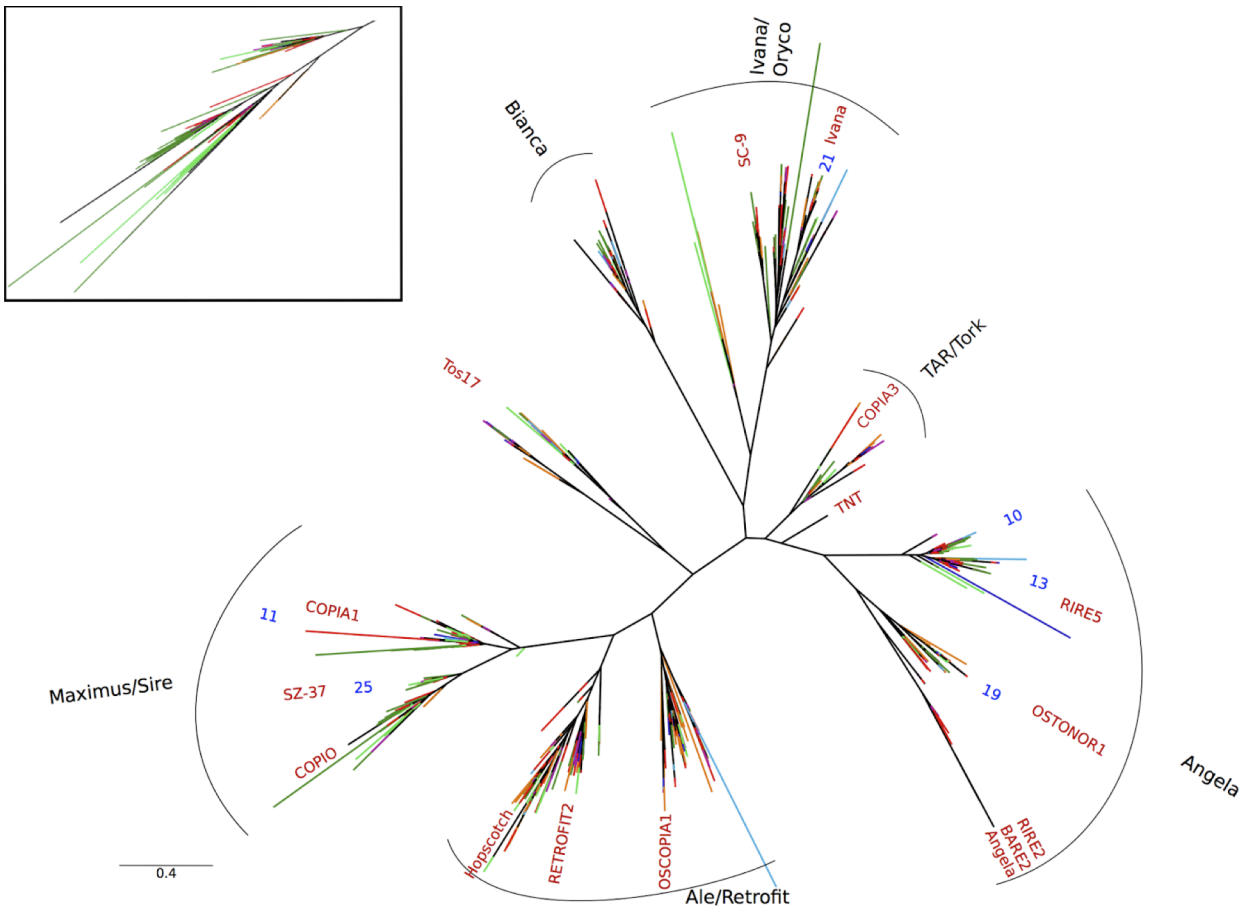


Supplementary Figure 6. **Concerted evolution in the distal region of chromosomes 11 and 12.** Analyses were carried out on aligned blocks of 300bp. **a,** Double logarithmic plot of distances between sequences of *O. sativa* along chromosomes 11 and 12, showing the break-point between converted and unconverted regions. **b,** Linear plots of distances for the first 3Mb of chromosomes 11 and 12 for six rice species using *O. sativa* chromosome 11 as reference.

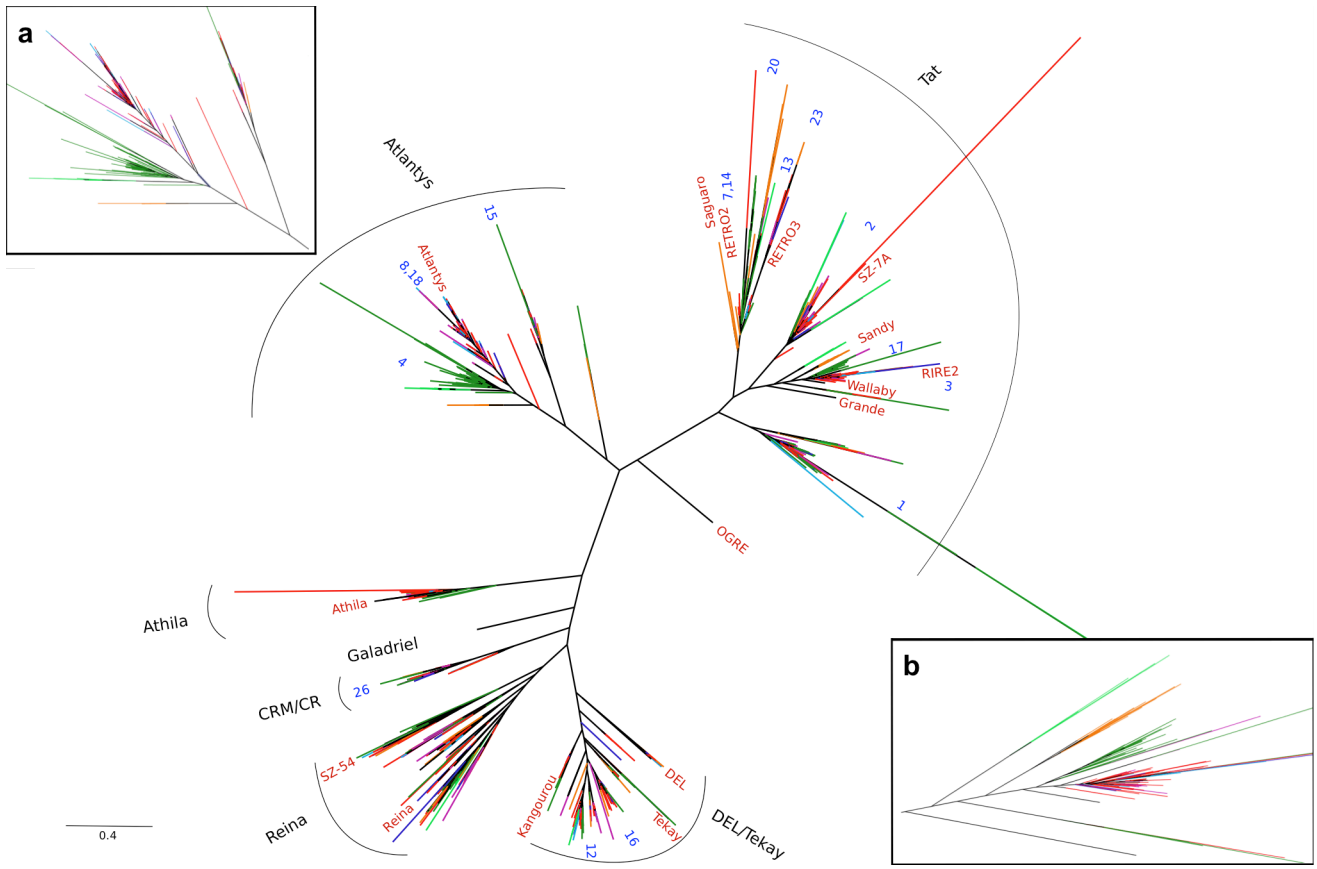




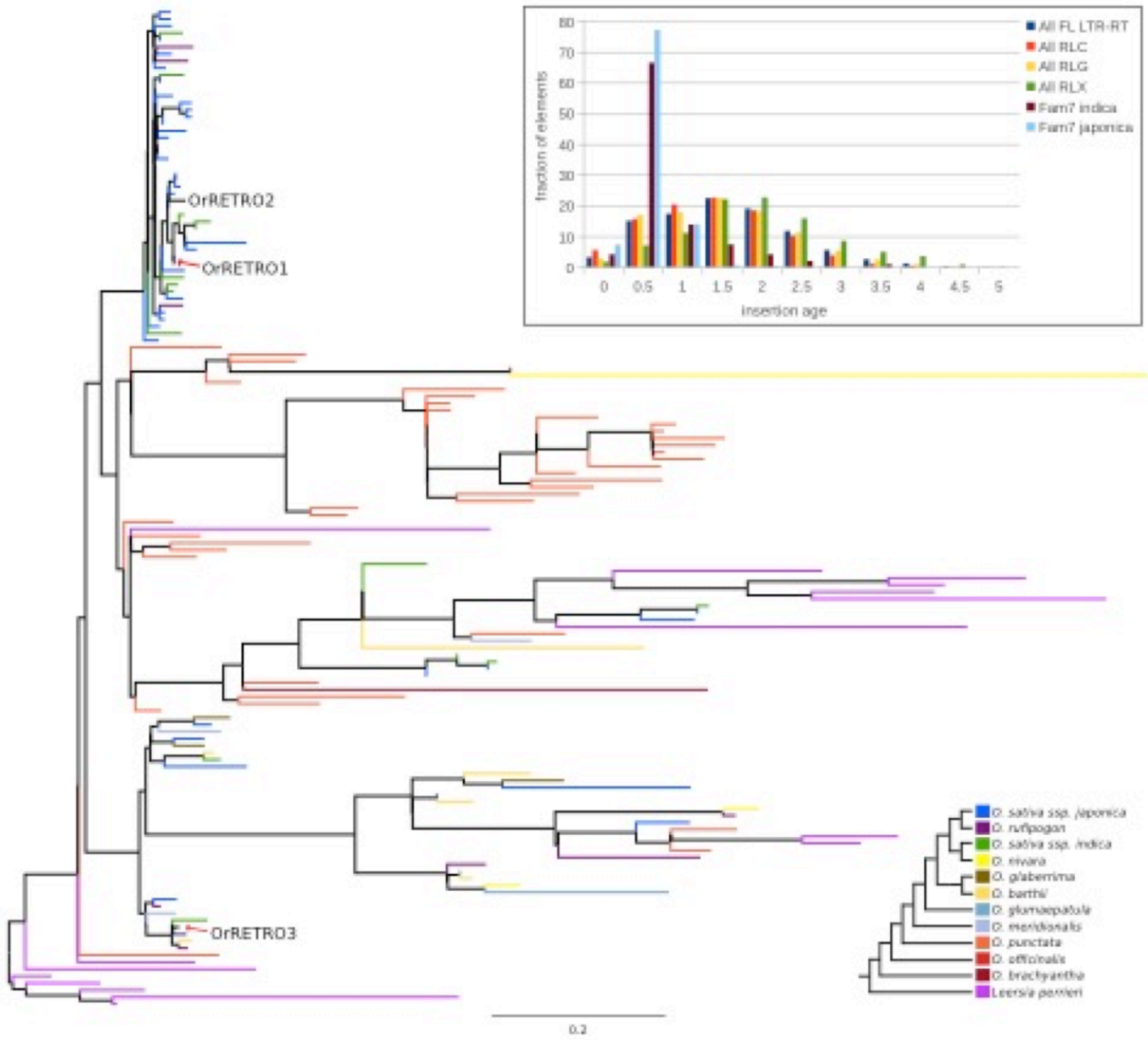
Supplementary Figure 7. Synteny mapping between *L. perrieri* and *O. sativa* vg. japonica. a, Dot plot of collinear orthologs, here shown for chromosome 5, exhibits typically small numbers of short in-place inversions along a largely conserved karyotype. **b,** Comparative map shows linkages between synteny blocks; grey, conserved arrangement; orange, in-place inversions; purple, transpositions.



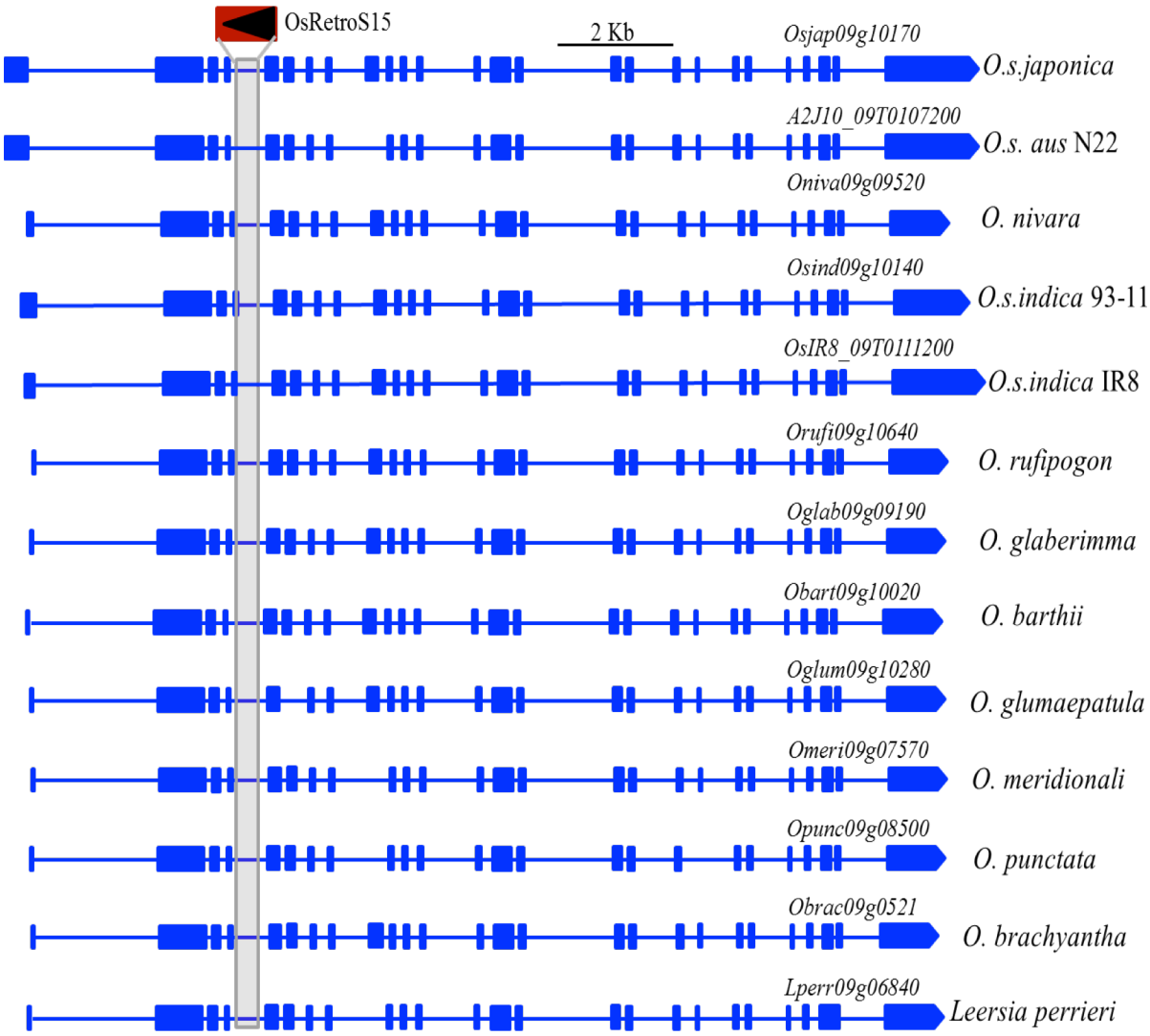
Supplementary Figure 8. Maximum-likelihood tree of 1500 randomly-selected *Copia* reverse transcriptase sequences. The major lineages (see Online Methods) are reported in black and/or enclosed by black arches; known families are in red; blue numbers indicate the location and abundance of the families with more than 100 complete elements isolated from the 11 genomes. Branch colors represent different species or groups of species: *O. sativa* vg. japonica, *O. rufipogon*, *O. sativa* vg. indica [93-11], and *O. nivara* are in red; *O. glaberrima* and *O. barthii* are in purple; *O. glumaepatula* in blue; *O. meridionalis* in pale blue; *O. punctata* in dark green, *O. brachyantha* in light green, and *L. perrieri* in orange. Black branch tips denote previously-characterized elements used to assign the branches to known families. The inset depicts in detail the independent *O. punctata* and *O. brachyantha* proliferations of elements related to the *O. sativa COPIO*.



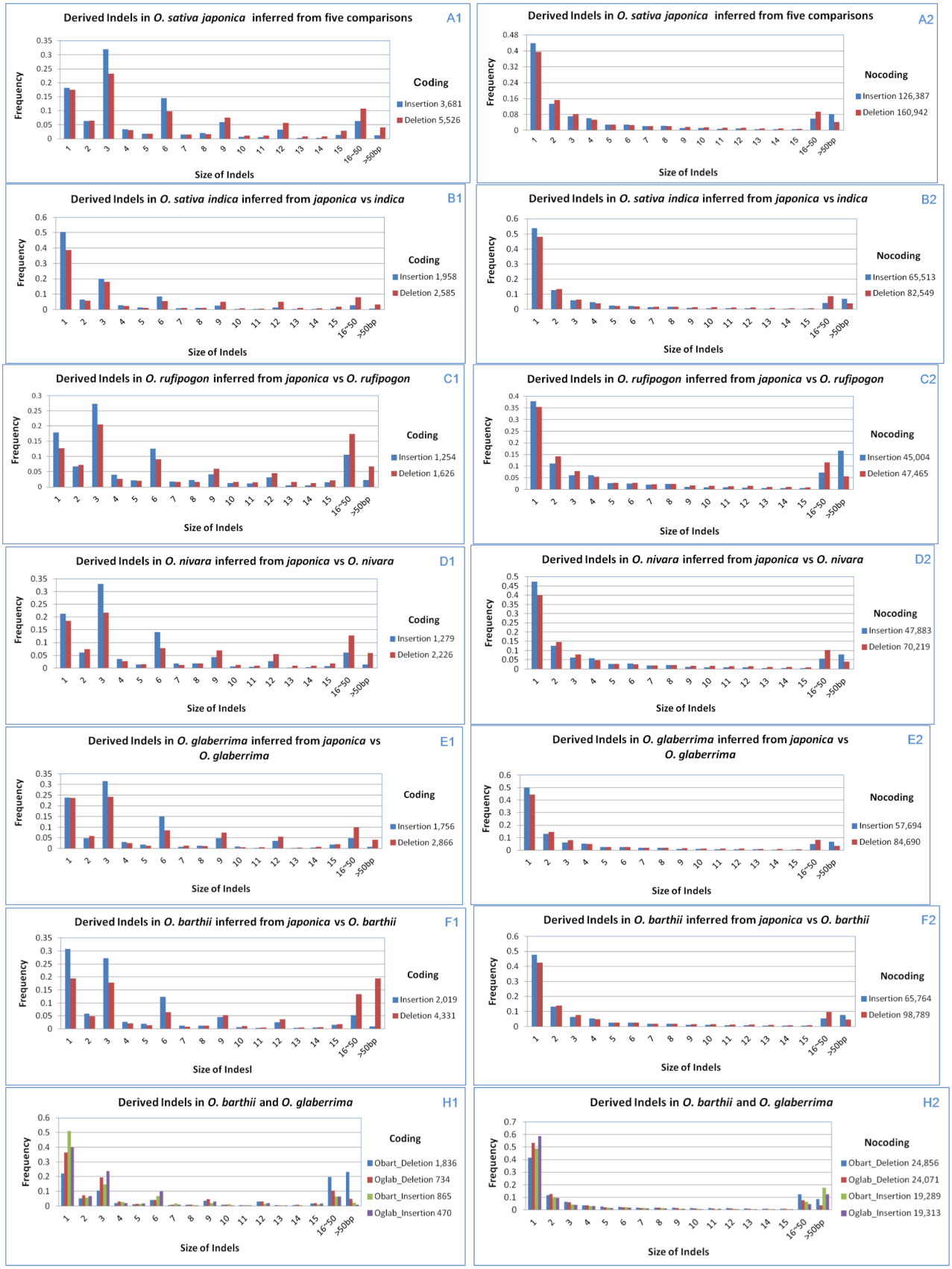
Supplementary Figure 9. Maximum-likelihood tree of 1500 randomly-selected *Gypsy* reverse transcriptase sequences. Names and colors have the same legend as in Supplementary Fig. 8. The *Atlantys* proliferation in *O. punctata* and *O. brachyantha* is depicted in inset **a**, the expansion of elements similar to *Grande* and *Wallaby* are represented in inset **b**.



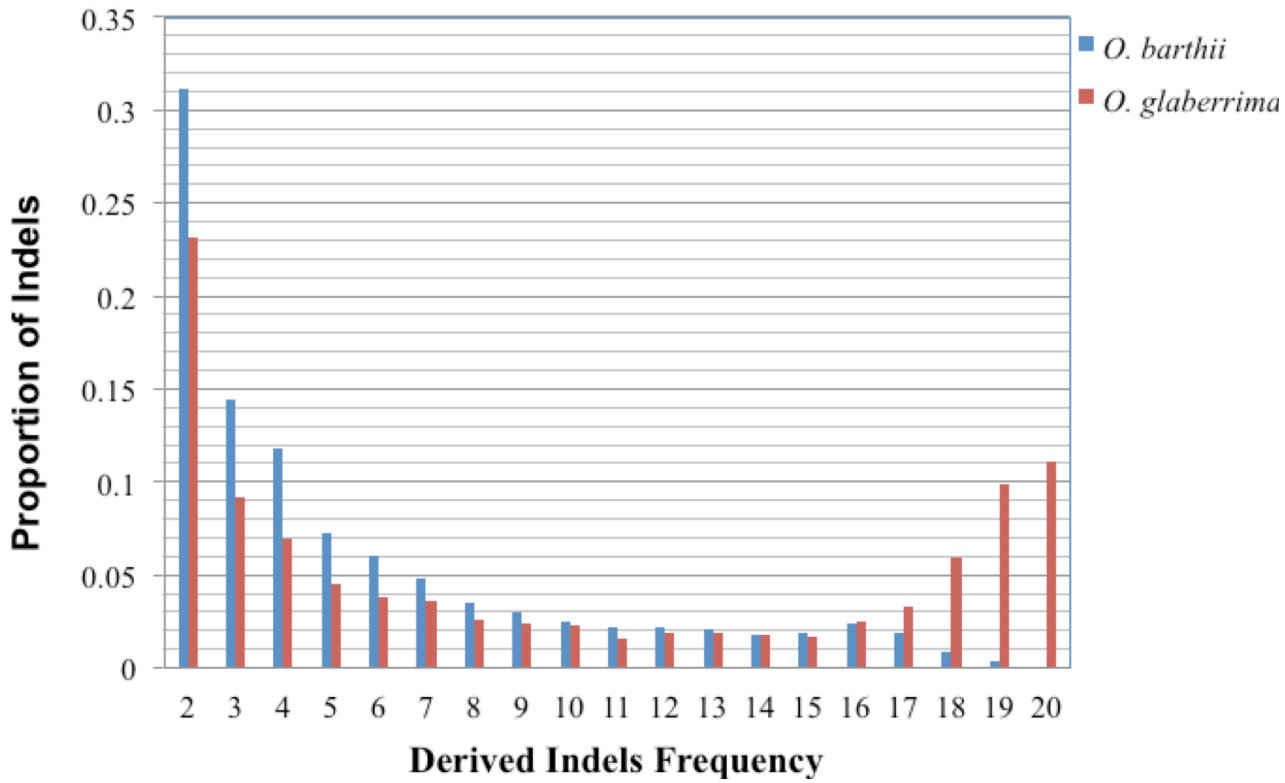
Supplementary Figure 10. Proliferation of the RETRO1 and 2 elements in *O. sativa*. This subtree is a magnification of the RETRO subclade from within the *Gypsy* Tat clade in Supplementary Fig. 9. The three *O. sativa* RETRO elements are differently distributed in the subtree: the RETRO1 and 2 families form a clade distinct from that of RETRO3, have shorter branches and are present only in *O. rufipogon* (4 elements) and *O. sativa* (18 and 64 elements in vgs. indica [93-11] and japonica, respectively), while other elements including RETRO3 are more diversified, older and present in all investigated species. The inset shows that most of the RETRO1/2 elements inserted between 0.5 and 1 MYR ago, a clearly different activity pattern than the rest of the LTR-RTs.



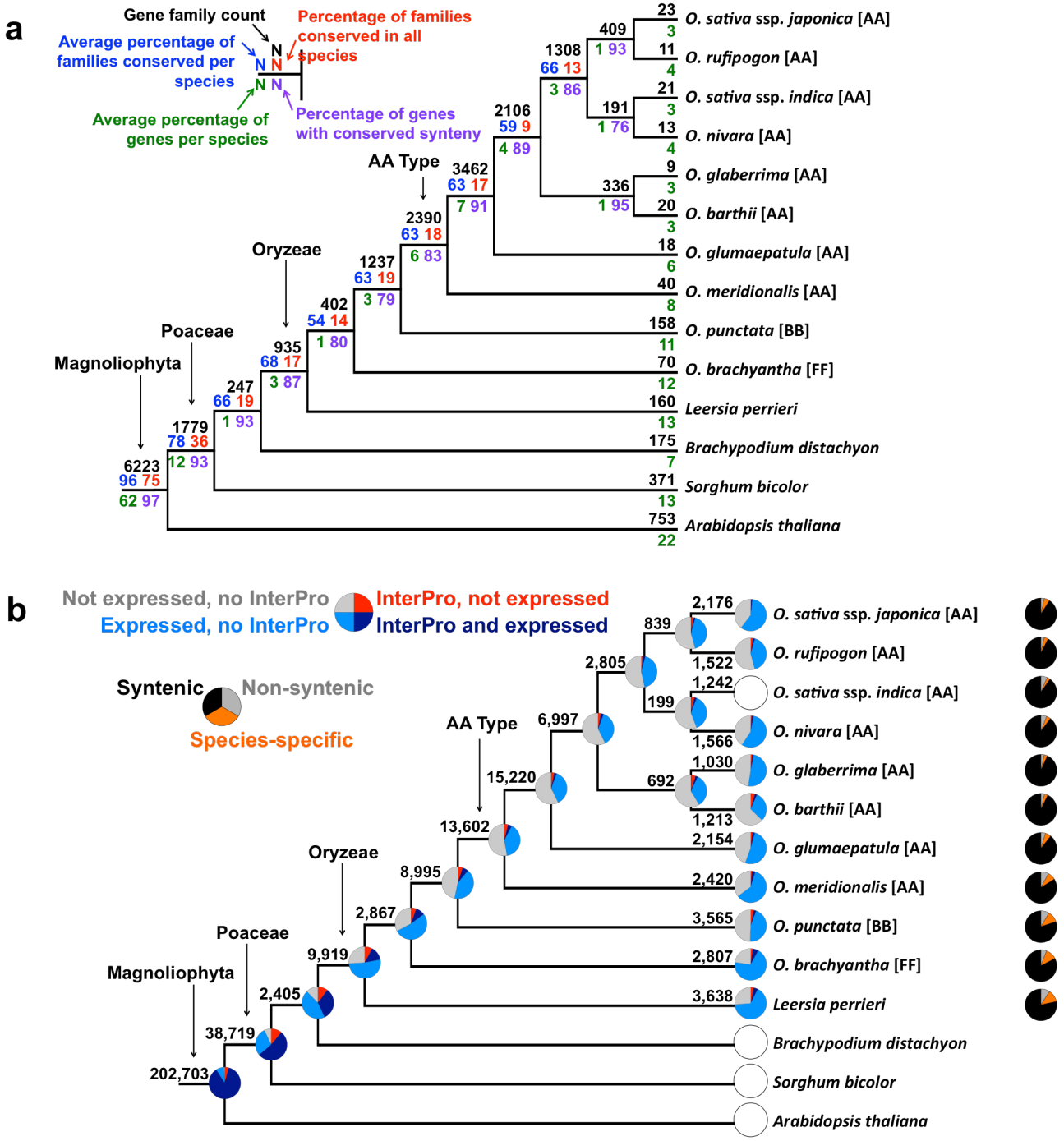
Supplementary Figure 11. A solo-LTR TRIM of OsRetroS15 located in an intron of an orthologous gene and shared among all 13 *Oryzeae* genomes. The blue boxes and lines represent exons and introns. The gene is expressed in all 13 genomes.



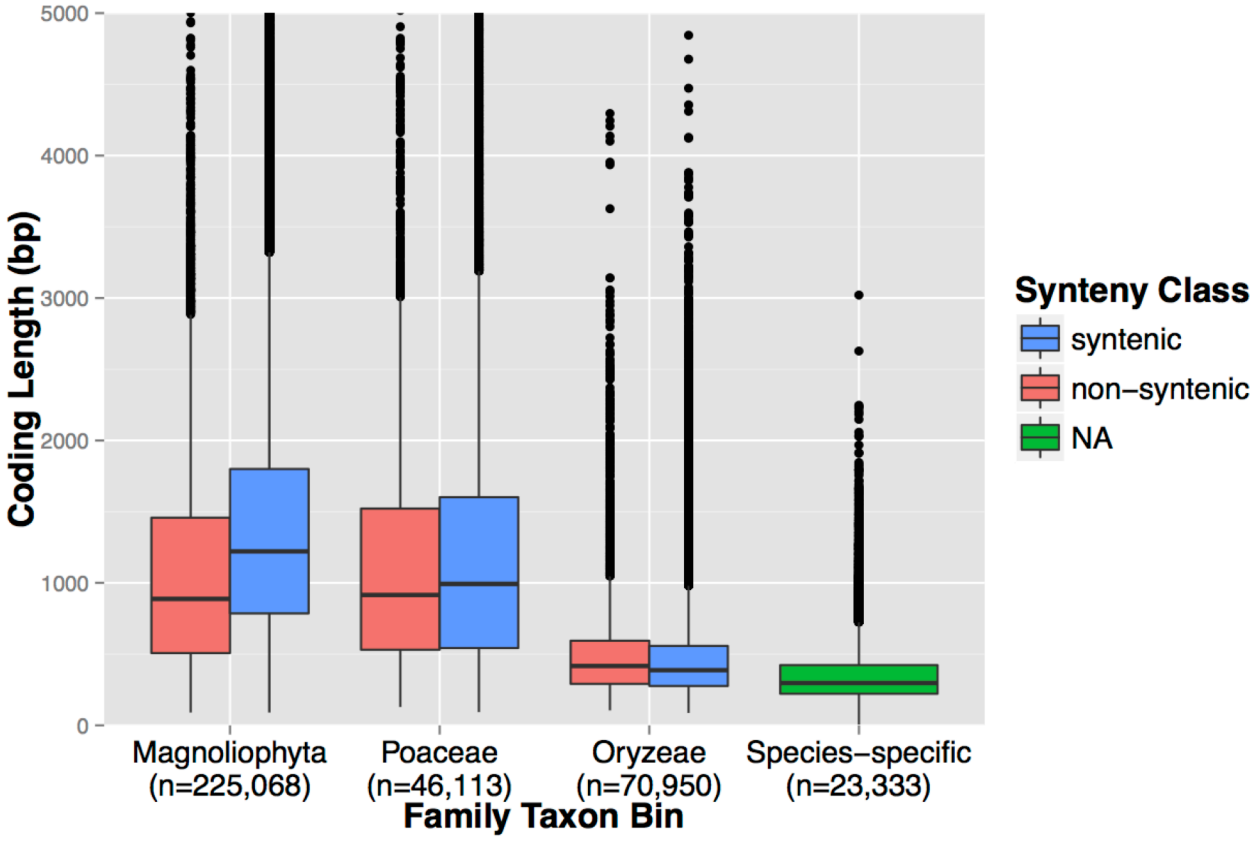
Supplementary Figure 12. Size frequency distribution of insertions and deletions. Left panels show indels within or overlapping with coding regions, in which multiples of three bases are overrepresented. Right panels show indels in noncoding regions, with a trend of declining frequency with increased indel size.



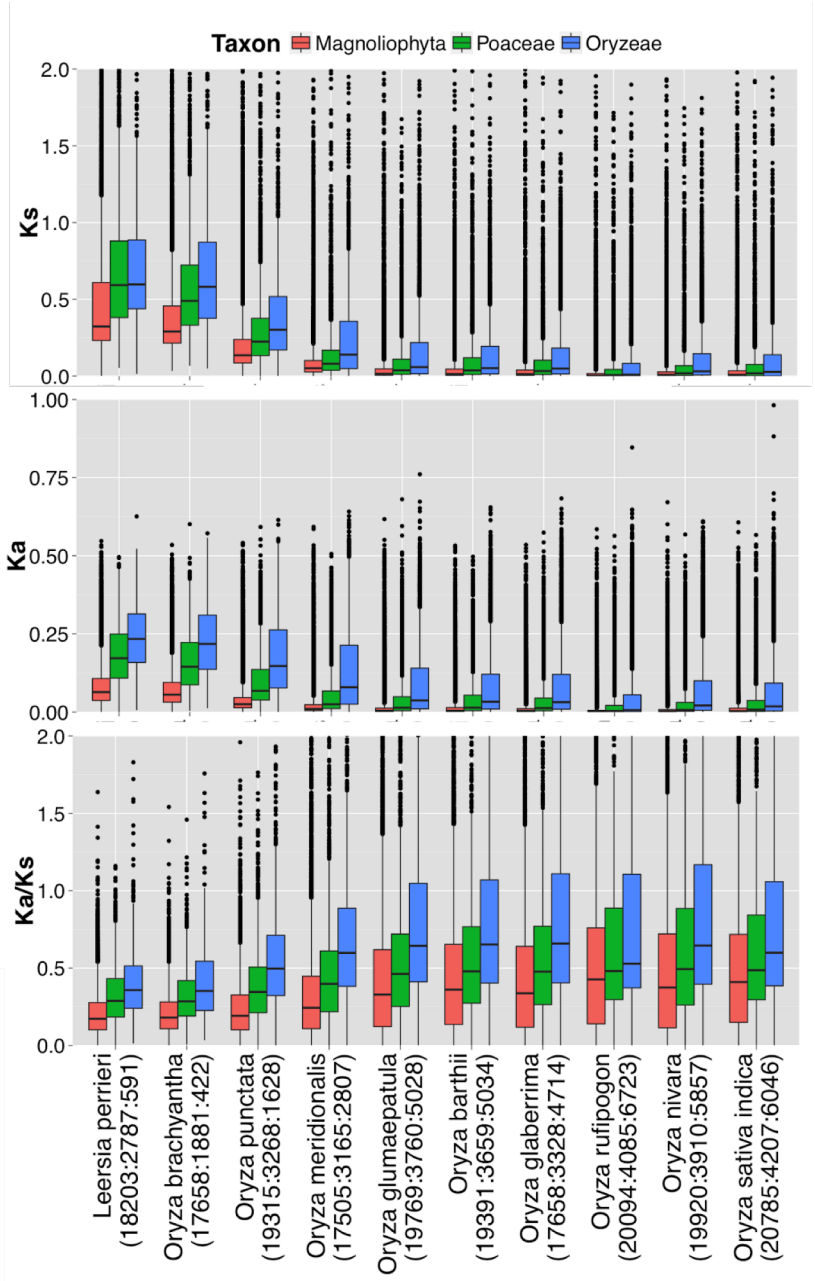
Supplementary Figure 13: The observed derived site-frequency spectra of indels for *O. barthii* and *O. glaberrima* populations. A total of 80,507 and 41,947 indels (>1 bp) from *O. barthii* and *O. glaberrima* populations were examined and the ancestral states were determined using both *O. sativa* vg. japonica and O. *glumaepatula* as outgroups.



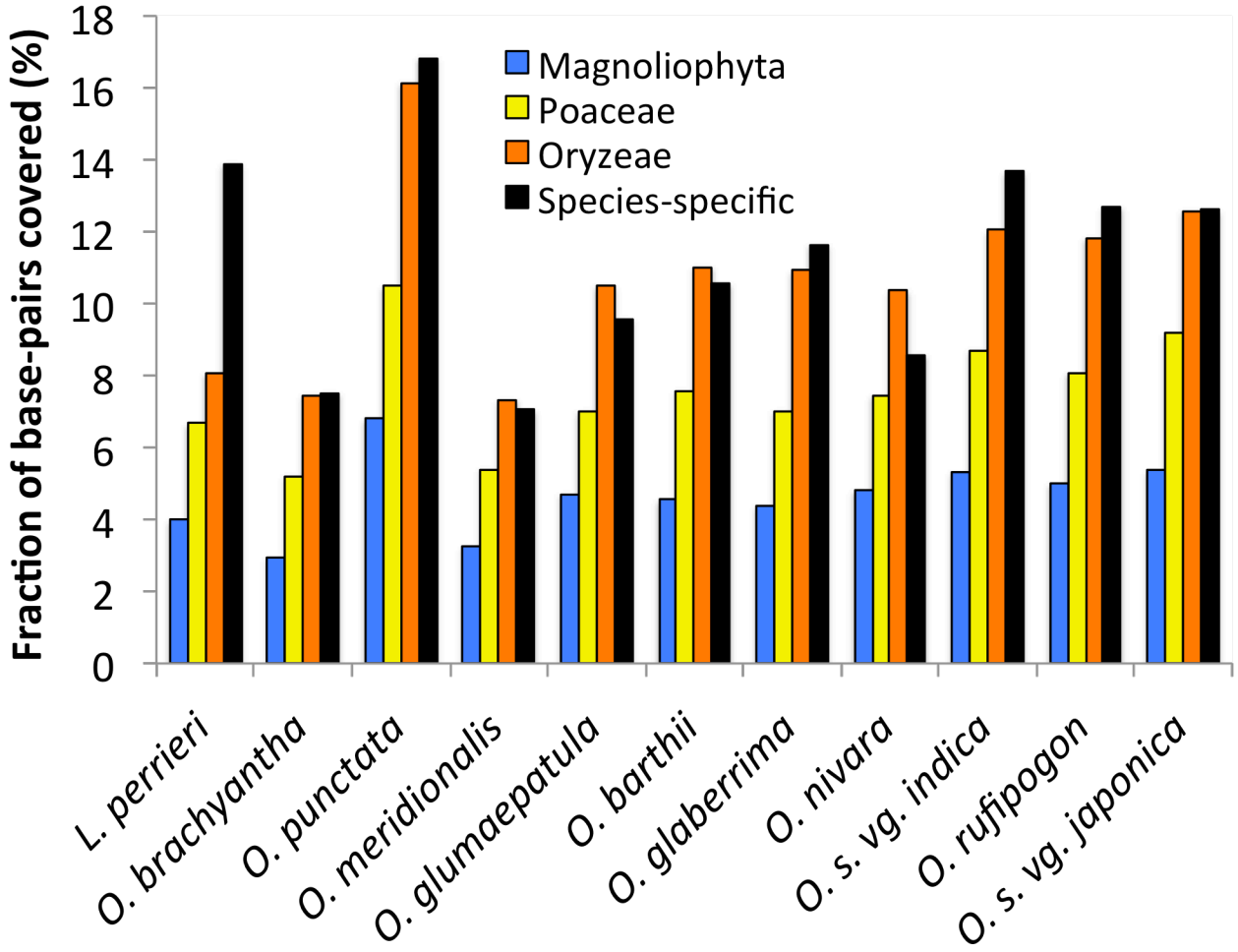
Supplementary Figure 14. Origin, conservation, and expression of putative genes and gene families distributed across taxa within the *Oryzeae* and flowering plant progenitors. Putative families, clustered by homology using the Ensembl Compara method (see Online Methods), were assigned to nodes of the species tree to give an inferred emergence time that corresponds with the last common ancestor of species represented in the family. **a,** Summary statistics among 22,867 clustered families. Shown at each node of the species tree: family counts (black text), average percentage of families conserved per descendent species (blue text), percentage of families conserved in all descendent species (red text), average percentage of putative genes per descendant species (green text), and percentage of putative genes exhibiting conserved synteny (purple text). **b,** Pie charts show prevalence of expression from RNA-seq data and InterPro domains in genes (black text) at each level of the species tree. At right pie charts give overall proportions of syntenic, non-syntenic, and species-specific loci.



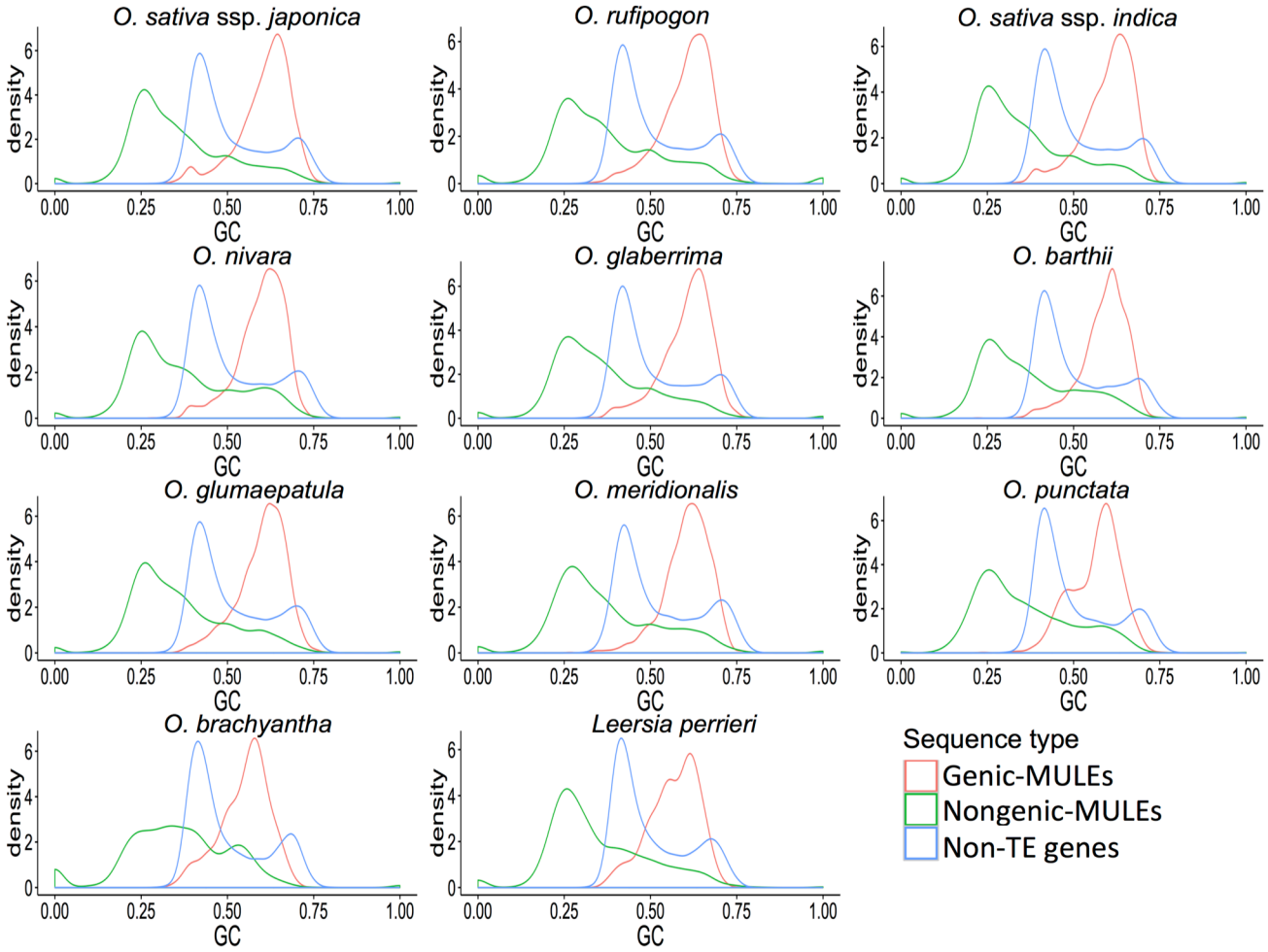
Supplementary Figure 15. Relationship between gene family age and predicted coding length in the *Oryzeae*. Gene annotations in all species were binned according to family root taxon, with the *Magnoliophyta* bin conserved in Arabidopsis, the *Poaceae* bin conserved in sorghum and/or Brachypodium, and the *Oryzeae* bin conserved in two or more species in the *Oryza* genus or in *L. perrieri*. Those classified as species-specific included both orphans and multi-gene families specific to a species. Syntenic genes were classified on the basis of collinear mapping of orthologous loci in all pairwise comparisons of the eleven species. In cases of multiple transcript isoforms, the one with the longest CDS was used. Significant differences in the distributions of CDS length in syntenic loci between taxon bins were found using the Wilcoxon rank sum test, with p-value < 2.2e-16 for all comparisons. Median CDS lengths for syntenic loci within the *Magnoliophyta*, *Poaceae*, and *Oryzeae* age groups are as follows: 1221 bp (n=218,523), 993 bp (n=42,489), 387 bp (n=60,991). The median length for species-specific loci is 297 bp (n=23,333).



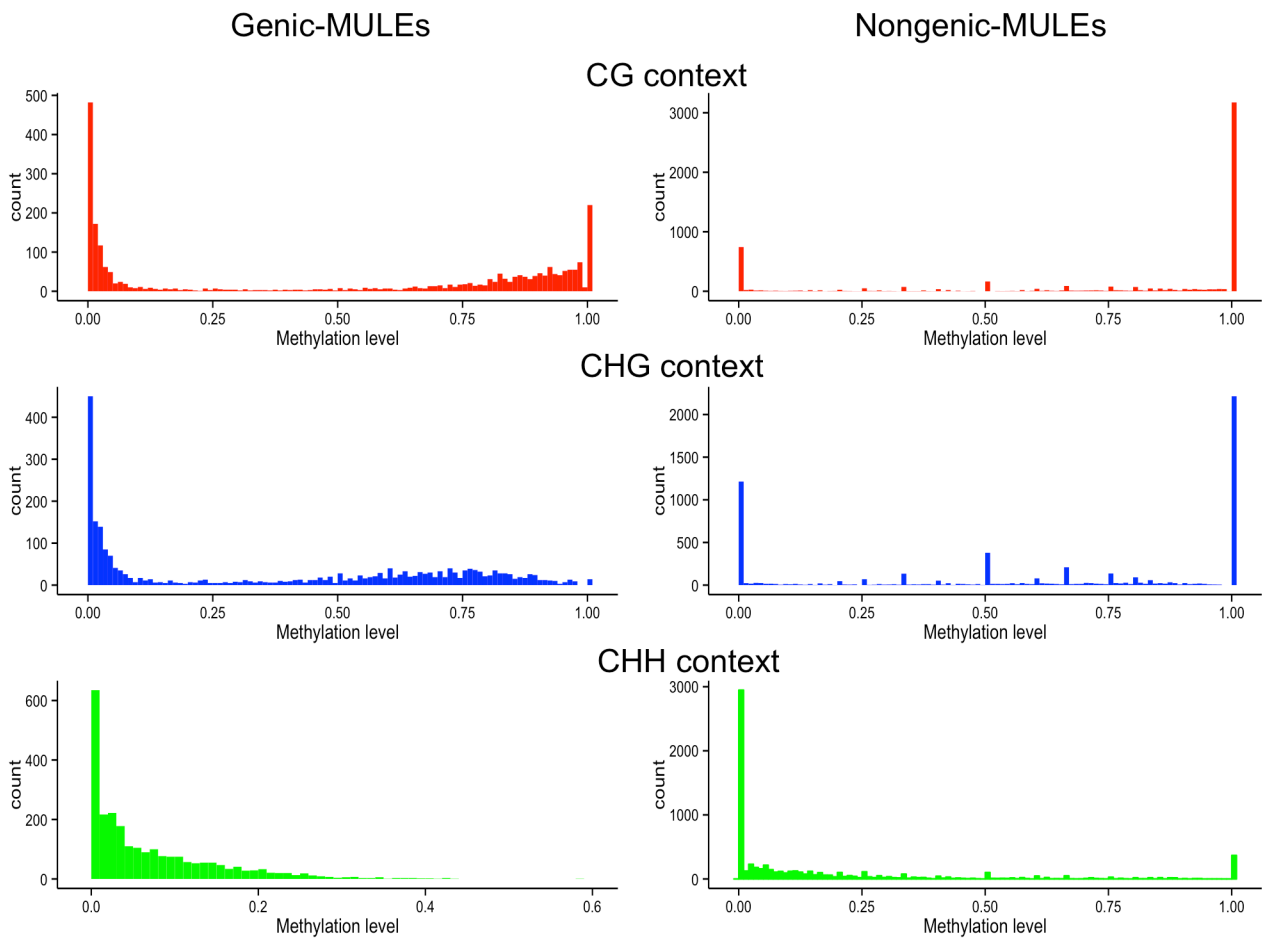
Supplementary Figure 16. Higher substitution rates and relaxed selection in recently emerged families of annotated loci compared to ancient families. Putative genes, binned into ancient (*Magnoliophyta*), intermediate age (*Poaceae*) and recently emerged (*Oryzeae*) families, were pairwise aligned to their predicted ortholog in *Oryza sativa* vg. japonica and analyzed with PAML software (Online Methods). Boxplots show distributions of synonymous substitution rates (Ks) in the top panel, non-synomous rates (Ka) in the middle panel, and their ratio (Ka/Ks) in the bottom panel. In cases of multiple orthologs, the one giving the lowest Ks was used. Results were filtered to exclude pairs with Ks > 2 or amino acid identity < 50% to minimize spurious alignments. The numbers of loci contributing to each taxon bin are given as colon-separated values beneath each species label, corresponding to *Magnoliophyta*, *Poaceae*, and *Oryzeae* taxon bins, respectively. The distributions of Ka, Ks, and Ka/Ks were shown to be significantly different between taxon bins in all species (Wilcoxon rank sum test with continuity correction: p-value < 0.05) except in *L. perrieri* comparing Ks distributions between the *Poaceae* and *Oryzeae* bins (p-value = 0.06945).



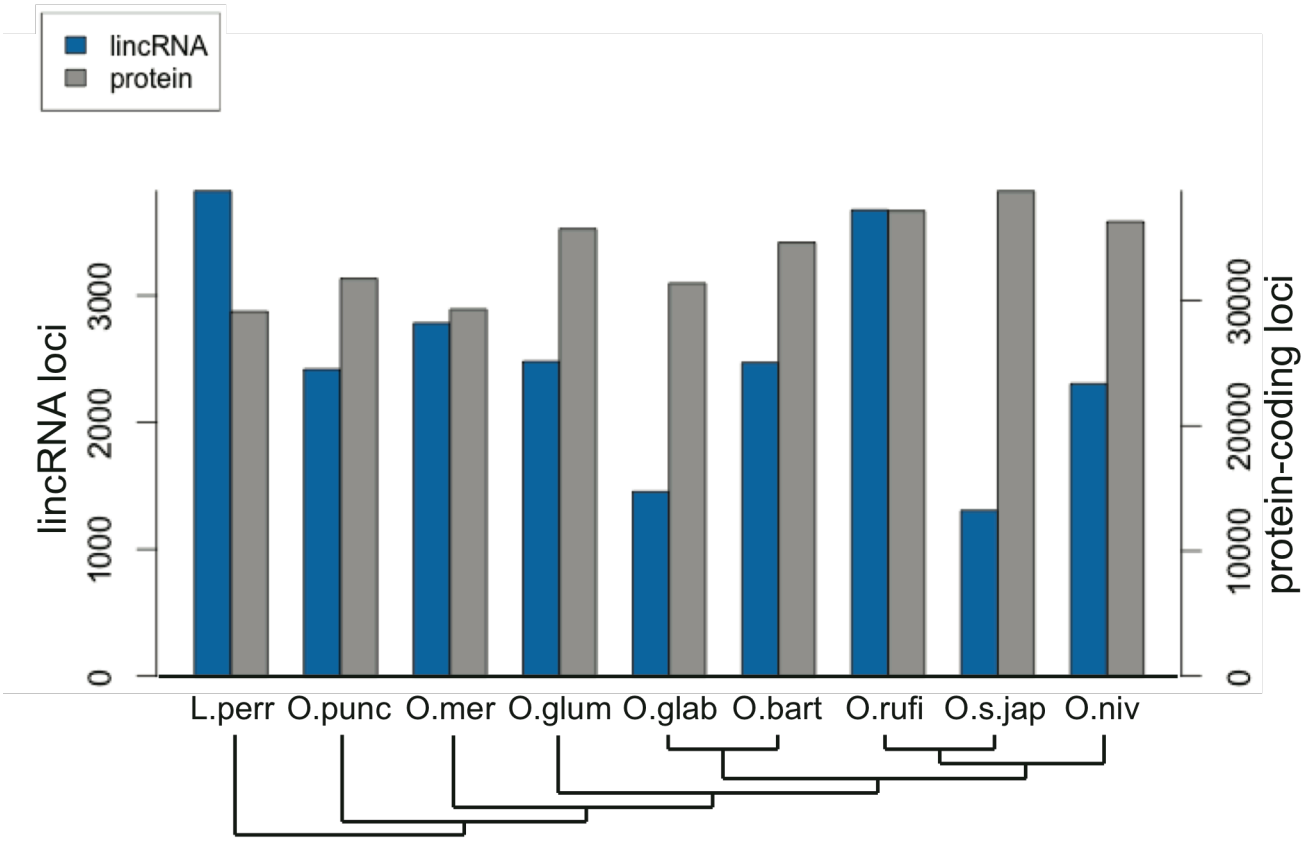
Supplementary Figure 17. Trend of higher LTR-retrotransposon repeat content flanking genes of recently-emerged gene families compared to older gene families. The percentage of base-pairs overlapping LTR-retrotransposon sequences is shown after averaging over all 2-kb upstream and downstream sequences flanking annotated genes. Repeat content was significantly different in angiosperm-derived loci (*Magnoliophyta*) compared to all other age classes (Welch's T-test, p-value < 2.2e-16), and in *Poaceae*-derived loci compared to all other age classes (Welch's T-test, p-value < 1e-03). Repeat content was significantly different in *Oryzeae*-derived and species-specific loci only in the cases of *L. perrieri*, *O. nivara*, and *O. sativa* vg. indica [93-11] (Welch's T-test, p-value < 1e-02).



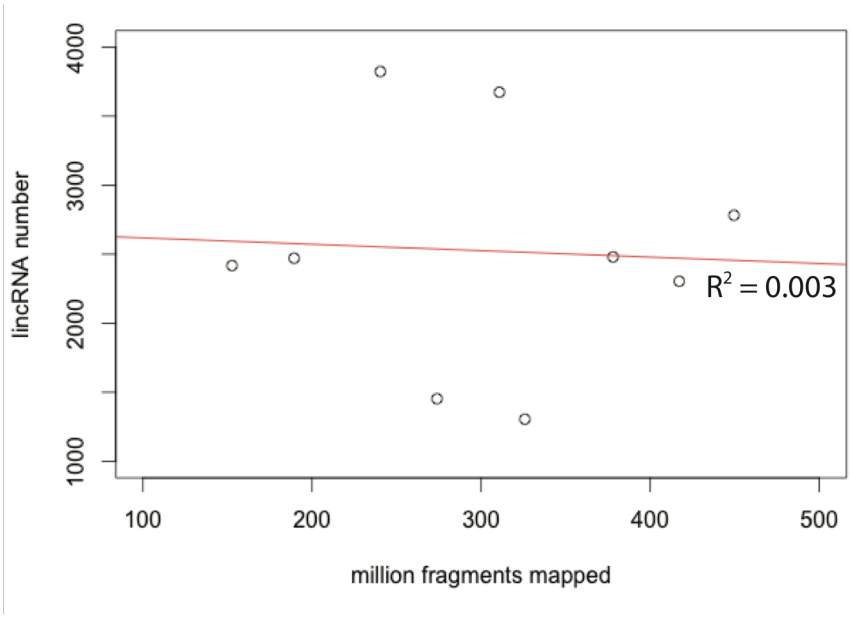
Supplementary Figure 18. The GC content distribution of genic-, nongenic-MULE internal sequences and non-TE genes. The number of non-TE genes, genic-MULEs, and non-genic MULEs in this figure are: 14003, 2557, 7103 (*O. sativa*. vg. japonica); 13164, 2275, 7479 (*O. rufipogon*); 14945, 2116, 6882 (*O. sativa*. vg. indica [93-11]); 12979, 2248, 7270 (*O. nivara*); 12840, 1604, 5352 (*O. glaberrima*); 13468, 2120, 7214 (*O. barthii*); 12439, 1936, 6643 (*O. glumaepatula*); 9608, 1938, 4972 (*O. meridionalis*); 15089, 1341, 5086 (*O. punctata*); 10226, 283, 3862 (*O. brachyantha*); 14480, 356, 2916 (*Leersia perrieri*). The density plot was generated with the geom_density function of ggplot2 package in R (H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009).



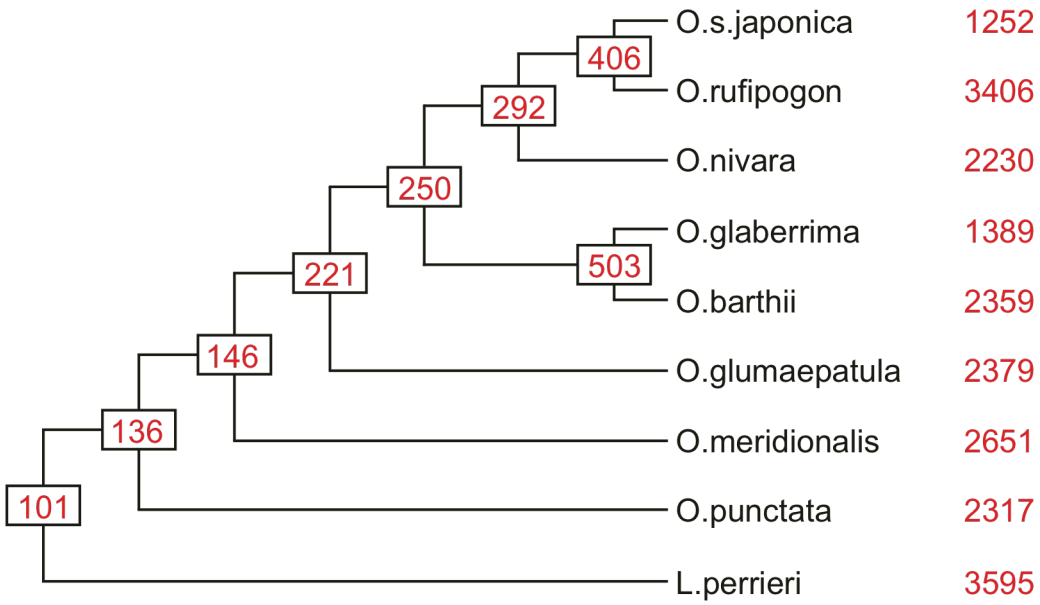
Supplementary Figure 19. Methylation levels within internal sequences of genic- and nongenic- MULEs in three cytosine contexts (CG, CHG, CHH) in the *O. sativa* vg. japonica genome. On the left, histograms are shown genic-MULEs (n=2,472, n=2,467, and n=2,486 for CG, CGH, and CHH methylation contexts, respectively). On the right, histograms are shown nongenic-MULEs (n=5,607, n=5,814, and n=6,930 for CG, CGH, and CHH methylation contexts, respectively).



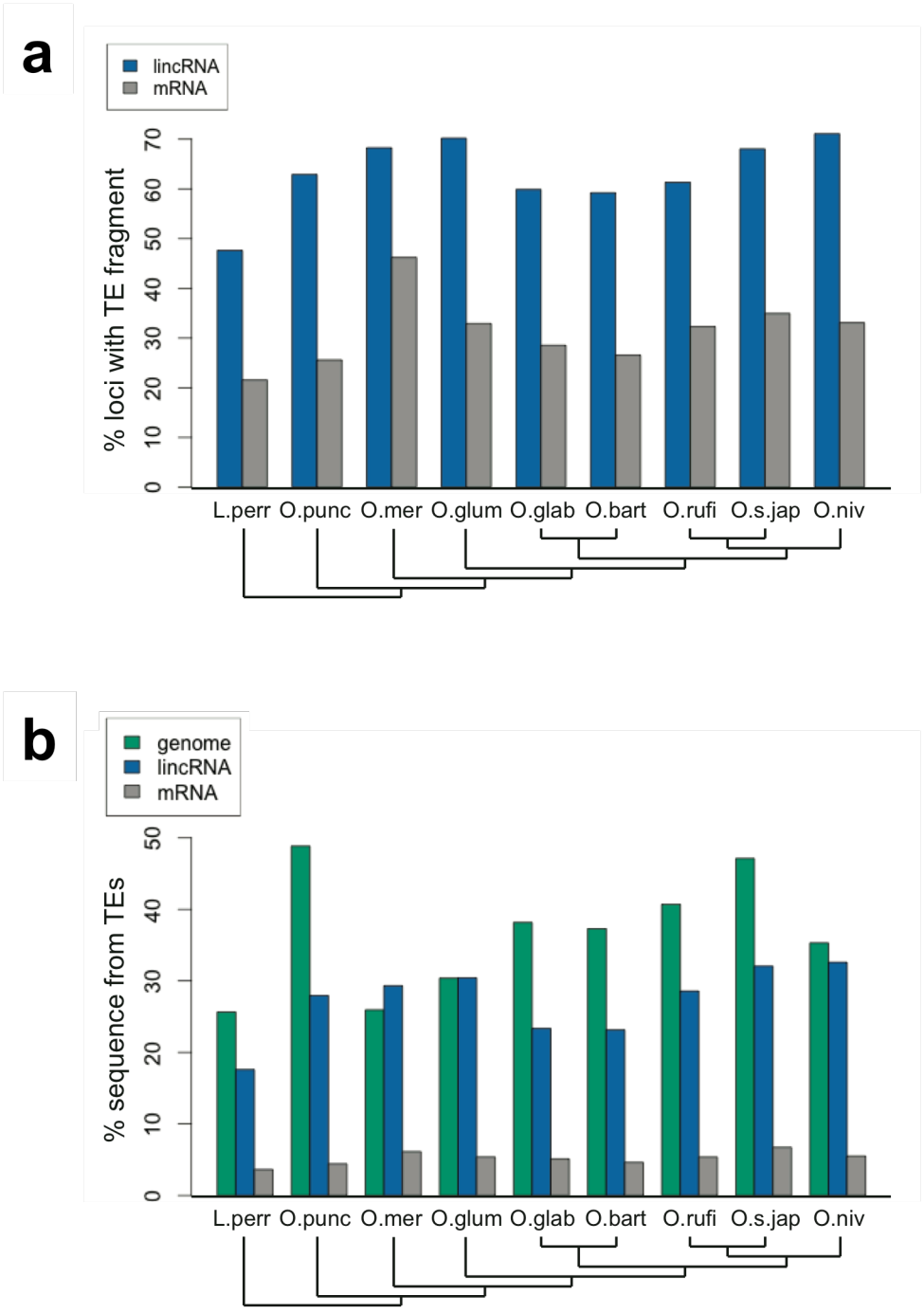
Supplementary Figure 20. Number of annotated lincRNA and protein-coding loci. Numbers of annotated lincRNA loci (left axis) and protein-coding loci (right axis) are displayed for each of 8 species of *Oryza* and the outgroup *Leersia perrieri*.



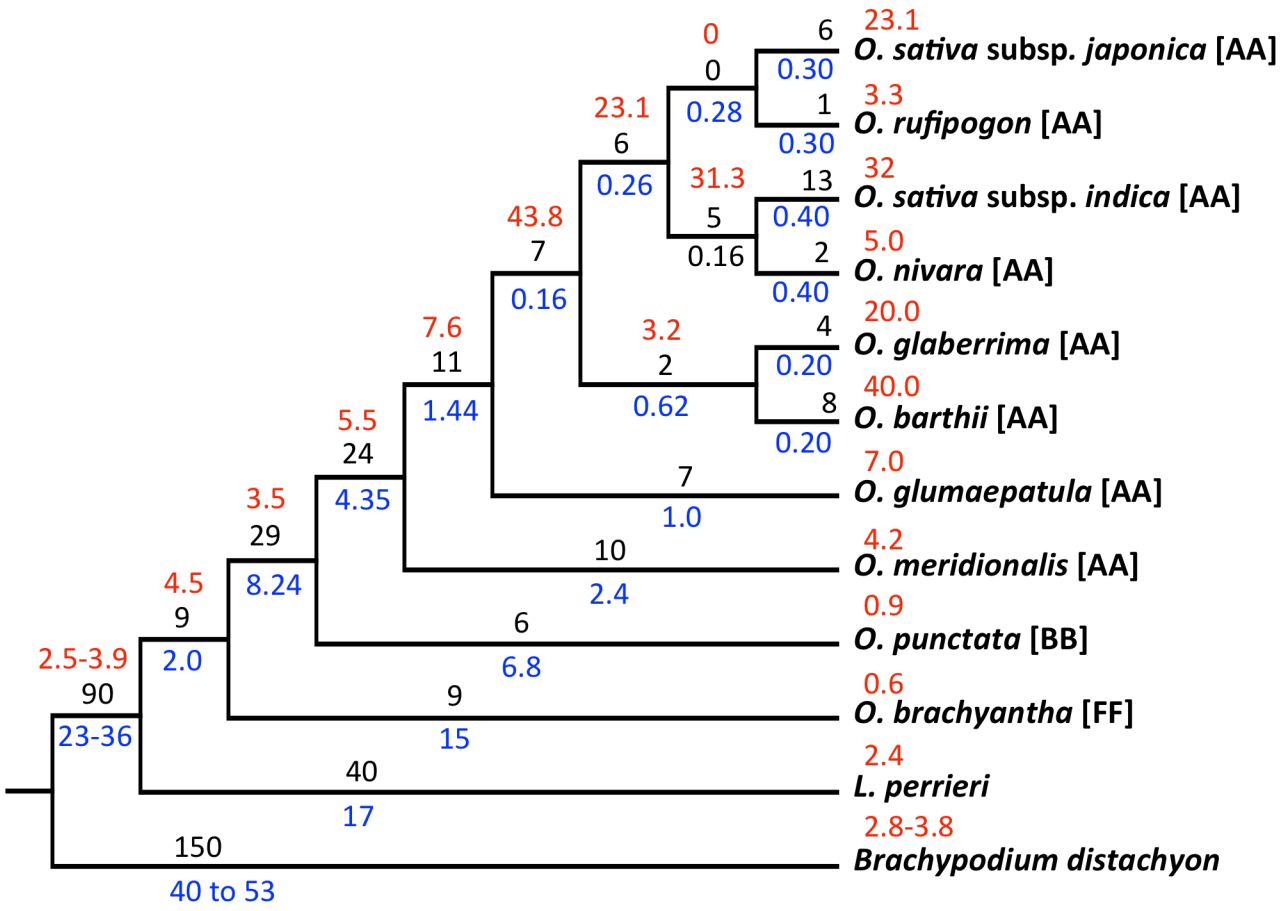
Supplementary Figure 21. Number of annotated lincRNAs versus RNA-Seq library size. Plot of numbers of annotated lincRNAs in each of 8 species of *Oryza* and the outgroup *Leersia perrieri* versus RNA-Seq library size suggests that lincRNA number is not an artifact of varying library sizes.



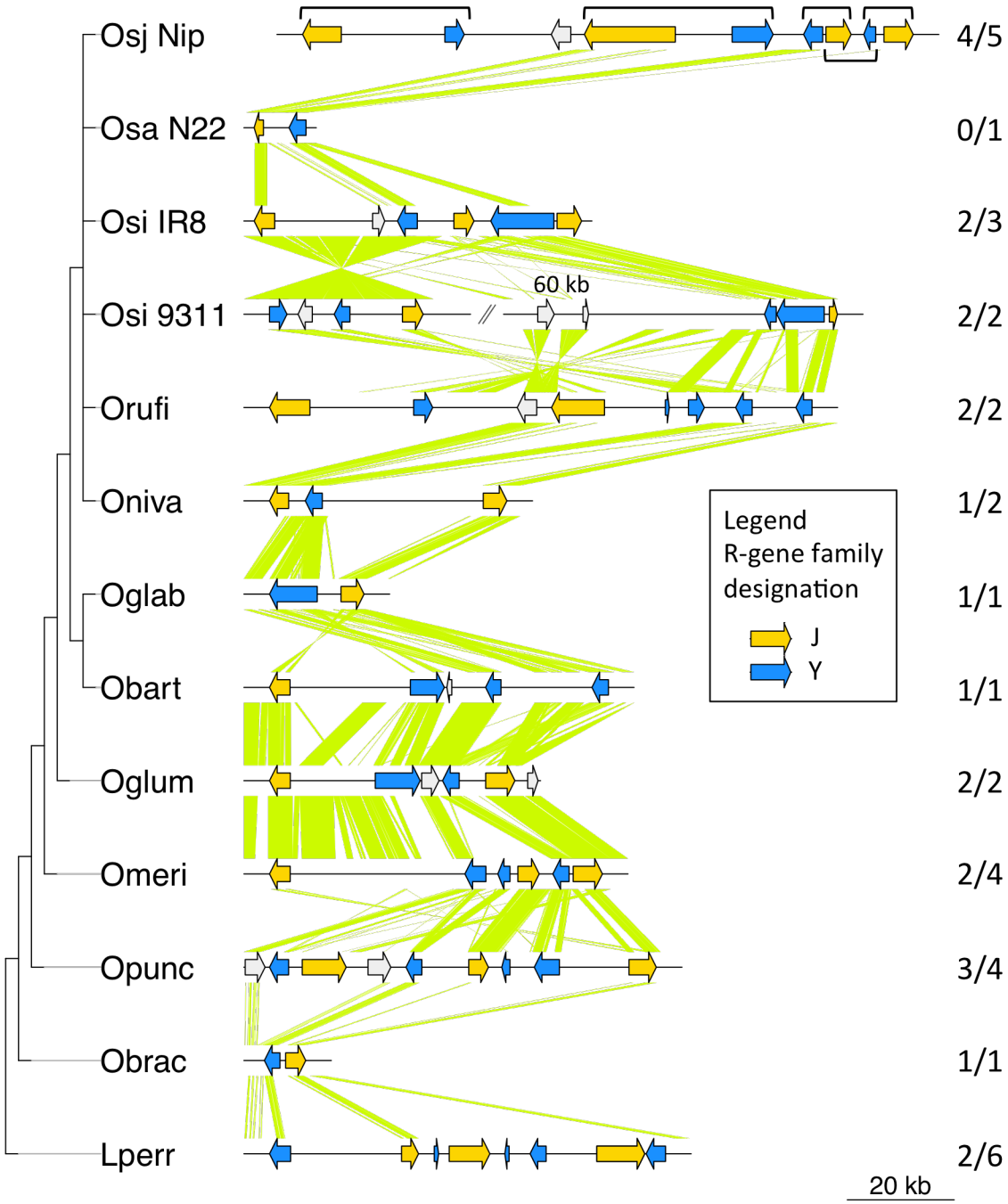
Supplementary Figure 22. Phylogenetic history of lincRNA families in *Oryza*. Total number of lincRNA families identified in each species is listed in red at right. Boxed numbers at each phylogenetic node list the number of lincRNA families detected in that common ancestor using a parsimony approach with best hit reciprocal blast matches between species.



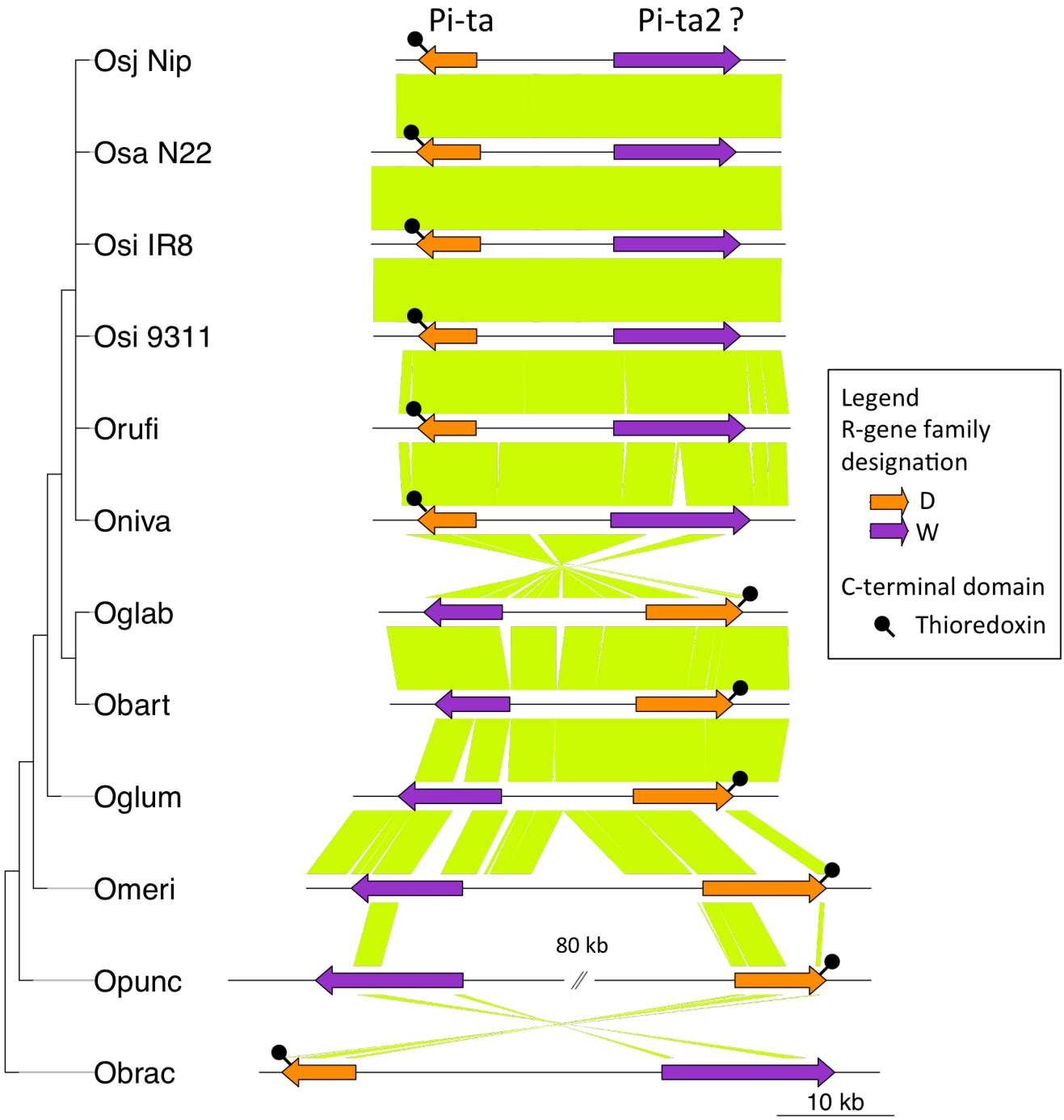
Supplementary Figure 23. Transposable element content of lincRNA and protein-coding loci in each of 8 species of *Oryza* and the outgroup *Leersia perrieri*. a, Fraction of lincRNA and protein-coding (mRNA) loci with detectable TE fragments. **b,** Fraction of total sequence within lincRNA loci, protein-coding loci (mRNA), and whole genome that are comprised of TE fragments.



Supplementary Figure 24. Duplication events giving rise to NLR disease resistance genes in the *Oryzeae* lineage. The number of inferred duplication nodes are shown in black above the branch preceding the node. The estimated branch length (MYR) is given in blue and the estimated rate (duplications/MYR) given in red. Inferred duplication nodes with consistency scores greater than 0.5 were summed across all trees within 28 gene families in the Compara gene tree database. Spikes in NLR gene proliferation appear at the node preceding the split of Asian and African rice, and in some subsequent lineages.



Supplementary Figure 25. **High prevalence of head-to-head configured heterologous pairs of R-genes within an orthologous region across 13 species of *Oryzeae*.** This complex cluster on chromosome 11 consists of genes from the "Y" and "J" families, forming putatively coupled gene pairs akin to *Pik1*/*Pik2* locus (see Fig. 7b, Online Methods). Fractions to the right of each segment shows the number of "J:Y" pairs in h2h configuration over the total number of "J:Y" pair combinations. For example osjap has four "J" genes alternating with four "Y" genes, interrupted by a single non-R-gene spacer at the third position. There are a total of five adjacent relationships between one "J" gene and one "Y" gene, shown in brackets, of which four are head-to-head and one is tail-to-tail. Region shown corresponds to 11:23,300,866,-23,430,153 on the *O. sativa* vg. japonica cv. Nipponbare RefSeq. See Fig. 7 for species key.



Supplementary Figure 26. Comparison of the rice Pi-ta resistance locus across ten *Oryza* species reveals hallmarks of a functionally coupled R-gene pair and offers a hypothesis on the long sought-after identity of the *Pi-ta2* gene. The proposed *Pi-ta2* gene (cited in Online Methods) corresponds to OS12G0281600 (LOC_Os12g18374) in the public Nipponbare annotation. Region shown corresponds to 12:10,604,560-10,639,373 on the *O. sativa* vg. japonica cv. Nipponbare RefSeq. See Fig. 7 legend for species key.

Supplementary Table 1. List of nine species sequenced, summarizing methods and participants in the IOMAP consortium.

Species	Cultivar	Paired reads (millions)	Long mate-pair reads (length: millions)	Sequence depth	Sequence technologies	Assembly/scaffold software	Institution(s)*
Oryza sativa vg.indica	IR 8	NA	NA	73.4X	PacBio	Canu[22]	AGI, IRRI
Oryza sativa vg. aus	N 22	NA	NA	65X	PacBio	FALCON[21]	AGI, IRRI
Oryza nivara	IRGC:100897	140.3	2k: 19.7 5k: 15.4	102X	Illumina	ALLPATHS-LG[13] SSPACE[16] GapFiller[17]	IPMB, AGI
Oryza rufipogon	W1943	327.4	3k: 50.1 8k: 48.2	201X	Illumina	SOAPdenovo[14] SSPACE[16] GapCloser[17]	NCGR
Oryza barthii	IRGC:105608	174.8	3k: 5.5 10k: 4.2	110X	Illumina 454	ALLPATHS-LG[13] Newbler[18] Minimus2[19]	AGI
Oryza glumaepatula	GEN1233	113.5	3k: 8.7 10k: 6.6	135X	Illumina	ALLPATHS-LG[13] SSPACE[16] GapFiller[17]	UPVD, UFPEL, AGI
Oryza meridionalis	OR44 (W2112)	185.3	3k: 37.9 10k: 25.9 20k: 11.4 40k: 3.9	166X	Illumina	ABySS[15] SSPACE[16] GapFiller[17]	UPVD, QAAFI, AGI
Oryza punctata	IRGC:105690	203.8	3k: 28.9 10k: 13.5 20k: 4.7 30k: 1.0	130X	Illumina 454	ALLPATHS-LG[13] Newbler[18] Minimus2[19]	AGI
Leersia perrieri	IRGC:105164	150.7	3k: 8.7 20k: 1.3	150X	Illumina	ALLPATHS-LG[13] SSPACE[16] GapFiller[17]	AGI

*AGI: Arizona Genome Institute, University of Arizona, USA; IRRI: International Rice Research Institute, Los Baños, Laguna, Philippines; IPMB: Institute of Plant and Microbial Biology of Academia Sinica, Taiwan; NCGR: National Center for Gene Research of Chinese Academy of Sciences, China; UPVD: University of Perpignan, Via Domitia, France; UFPEL: Universidade Federal de Pelotas, Brazil; QAAFI: Queensland Alliance for Agriculture & Food Innovation, University of Queensland, Australia.

Supplementary Table 2. Assembly statistics for 13 *Oryzeae* species, including seven not previously published.

Species	Estimated size (Mb)†	Assembly length (bp)	Contig count	Contig N50 (bp)	Scaffold count	Scaffold N50 (bp)	Super-scaffold count
O. sativa vg. japonica[30]	389	373,245,519	242	7,711,345	NA	NA	12
Oryza sativa vg. indica[31] [93-11]	466	374,545,499	35,416	27,122	NA	NA	12
Oryza sativa vg. indica [IR 8]	NA	389,088,367	67	14,564,657	66	14,564,657	12
Oryza sativa vg. aus [N 22]	NA	362,279,097	912	946,210	912	946,210	12
Oryza nivara	448	337,950,324	16,484	37,688	2,430	295,425	12
Oryza rufipogon	450	339,177,042	68,481	34,232	49,224	137,860	12
Oryza glaberrima[26]	372	285,037,524	21,269	24,838	NA	NA	12
Oryza barthii	411	308,272,304	25,427	18,926	3,001	443,744	12
Oryza glumaepatula	464	372,860,283	17,912	31,921	3,157	451,867	12
Oryza meridionalis	435	335,668,232	62,778	9,149	18,305	238,568	12
Oryza punctata	423	393,816,603	16,598	43,035	1,641	1,310,149	12
Oryza brachyantha[25]	362	250,927,218	19,463	21,984	NA	NA	12
Leersia perrieri	323	266,687,832	9,937	50,248	1,005	8,681,563	12

†Based on flow cytometry data[4,5]

Supplementary Table 3. Evaluation of individual contig assemblies prior to scaffolding.

				Reads mapped to individual contigs		
Species	Illumina PE-reads	Reads mapped as pairs	% of reads mapped as pairs	Reads mapped in correct orientation	% mapped in correct orientation*	Insert size ± standard deveation (bp)
Oryza nivara	300,996,439	296,585,508	98.5	275,918,451	91.7	325±61
Oryza rufipogon	64,520,291	62,772,255	97.3	52,853,279	81.9	430±41
Oryza barthii	436,399,646	418,401,630	95.9	407,863,306	93.5	239±69
Oryza glumaepatula	277,527,350	270,242,125	97.4	242,522,252	87.4	314±97
Oryza meridionalis	371,962,892	344,704,703	92.7	310,092,536	83.4	187±68
Oryza punctata	488,879,958	477,064,501	97.6	463,100,651	94.7	296±82
Leersia perrieri	454,192,609	406,810,702	89.6	393,257,463	86.6	258±72

*The low proportion of correctly placed PE mappings for *O. rufipogon* and *O. meridionalis* is likely associated with the large number of small contigs in these assemblies (Table S2).

Supplementary Table 4. Calibration of consensus band units from fingerprinted BAC clones to nucleotide length.

Species	Physical length of PE-BACs (bp)	Total CB units of PE-BACs	CB unit equivalent (bp)
Oryza rufipogon	1,726,090,320	1,449,337	1190
Oryza nivara	2,931,651,698	2,237,294	1310
Oryza barthii	2,368,232,618	1,970,442	1201
Oryza glumaepatula	1,190,686,584	878,504	1355
Oryza meridionalis	272,359,756	200,307	1359
Oryza punctata	3,098,757,790	2,535,546	1222
Leersia perrieri	2,328,945,413	2,013,689	1156

Supplementary Table 5. Genome size estimates and assembly completeness for seven newly sequenced wild *Oryzeae* genomes.

Species	Assembly length (Mb)	Genome size (Mb) based on:				% of Genome assembled	Repeat & TE deficit, Length (Mb) (%)*
		Flow cytometry[10,11]	Physical map	K-mer	Average estimate		
Oryza nivara	338	448	486	357	430	79	68.5 (16)
Oryza rufipogon	338	450	436	374	420	81	60.0 (14)
Oryza barthii	308	411	380	282	358	86	49.2 (14)
Oryza glumaepatula	373	464	515	379	453	82	110.2 (24)
Oryza meridionalis	336	435	450	415	433	77	107.1 (25)
Oryza punctata	394	423	403	401	409	96	60.0 (15)
Leersia perrieri	267	323	359	302	328	81	37.5 (11)

*Estimated from Copetti & Wing 2016[6] with percentage based on average estimated genome size.

Supplementary Table 6. Coverage of chromosome 3 short-arm assemblies of eight species aligned to respective whole-genome assemblies: whole sequence and protein-coding genes.

	Whole sequence			Gene		
	Length	Align. cov. (%)		Length	Align. cov. (%)	
Species	(bp)	Total	Off chr3	(bp)	Total	Off chr3
Oryza rufipogon	4,097,599	95.7	2.5	2,009,158	96.1	1.6
Oryza nivara	4,211,386	98.8	12.5	2,122,186	99.7	12.9
Oryza barthii	4,074,511	97.5	1.8	1,910,161	98.0	1.1
Oryza glumaepatula	4,501,571	94.1	0.6	2,065,291	98.0	0.3
Oryza meridionalis	4,776,129	94.7	4.3	2,343,476	97.0	2.4
Oryza punctata	5,429,790	94.8	1.5	2,376,699	97.4	0.4
Oryza brachyantha	3,883,177	99.1	0.3	1,880,040	99.3	0.2
Leersia perrieri	4,671,510	96.8	0.3	2,324,058	98.8	0.1

Supplementary Table 7. Coverage of chromosome 3 short-arm assemblies of eight species aligned to respective whole-genome assemblies: intergenic regions and transposons.

	Intergenic			Transposons		
	Length	Align. cov. (%)		Length	Align. cov. (%)	
Species	(bp)	Total	Off chr3	(bp)	Total	Off chr3
Oryza rufipogon	2,089,589	95.2	3.3	919,976	93.2	9.8
Oryza nivara	2,090,354	97.8	12.0	920,883	95.3	12.8
Oryza barthii	2,165,434	97.1	2.4	958,051	92.2	3.2
Oryza glumaepatula	2,437,094	90.9	0.9	1,246,175	80.0	1.9
Oryza meridionalis	2,433,374	92.4	6.2	1,288,572	85.3	11.9
Oryza punctata	3,054,283	92.7	2.4	1,676,744	87.1	4.7
Oryza brachyantha	2,004,071	99.0	0.3	792,157	97.7	0.8
Leersia perrieri	2,348,280	94.7	0.4	880,072	86.0	1.0

Supplementary Table 8. Coverage of chromosome 3 short-arm assemblies of eight specie aligned to respective whole-genome assemblies: gene coding exons and introns.

	Coding Exons			Introns		
	Length	Align. cov. (%)		Length	Align. cov. (%)	
Species	(bp)	Total	Off chr3	(bp)	Total	Off chr3
Oryza rufipogon	730,311	96.7	0.5	1,143,670	95.1	2.5
Oryza nivara	776,155	99.2	12.4	1,171,941	99.4	13.3
Oryza barthii	720,501	97.7	1.2	1,041,562	97.5	1.1
Oryza glumaepatula	806,640	98.7	0.2	1,118,439	96.7	0.3
Oryza meridionalis	817,383	98.3	1.3	1,325,495	95.5	3.3
Oryza punctata	831,697	98.2	0.1	1,385,366	96.1	0.6
Oryza brachyantha	652,964	98.5	0.1	1,078,480	99.2	0.3
Leersia perrieri	835,265	98.7	0.0	1,271,591	98.3	0.3

Supplementary Table 9. Coverage of chromosome 3 short-arm assemblies of eight specie aligned to respective whole-genome assemblies: exon untranslated regions.

	5'-UTR			3'-UTR		
	Length	Align. cov. (%)		Length	Align. cov. (%)	
Species	(bp)	Total	Off chr3	(bp)	Total	Off chr3
Oryza rufipogon	26,516	95.8	0.0	78,672	98.1	1.0
Oryza nivara	42,583	99.5	8.9	95,682	99.7	10.3
Oryza barthii	52,588	97.9	0.1	112,635	99.2	0.3
Oryza glumaepatula	51,129	98.5	0.6	103,093	99.7	0.2
Oryza meridionalis	89,456	98.0	0.8	145,603	98.0	1.2
Oryza punctata	35,207	99.3	0.2	82,095	99.7	0.0
Oryza brachyantha	38,924	98.0	0.0	82,524	99.7	0.0
Leersia perrieri	46,939	95.4	0.0	104,117	99.4	0.0

Supplementary Table 10. Alignment coverage of 44 finished BAC sequences versus whole genome assemblies in four species: fraction of BAC aligned by chromosome.

Species	BAC accession	Length (bp)	Alignment coverage (%) total and by chromosome												
			Total	1	2	3	4	5	6	7	8	9	10	11	12
Oryza barthii	KF284071.1	176,785	93.9	-	0.2	93.4	0.1	-	0.3	-	-	-	-	-	0.0
Oryza nivara	FJ032625.1	218,473	95.8	-	-	0.1	-	-	95.6	-	0.1	-	-	-	-
Oryza nivara	FJ266025.1	204,633	94.5	0.4	0.0	-	0.5	18.2	0.6	-	2.7	0.1	-	72.1	-
Oryza nivara	FJ581047.1	324,233	98.0	0.1	0.5	-	-	0.3	96.7	-	0.2	-	-	-	0.4
Oryza nivara	GQ280265.1	119,975	75.4	-	31.1	0.6	0.9	0.5	41.2	0.2	-	0.1	0.4	-	0.5
Oryza nivara	HM999008.1	213,510	72.3	30.5	1.5	0.4	1.4	1.7	-	2.1	32.4	0.3	0.6	0.4	1.0
Oryza punctata	AC213149.2	139,052	96.0	-	2.7	91.6	-	-	0.3	-	1.4	-	-	-	0.0
Oryza punctata	AC214408.3	153,085	97.9	-	56.7	39.8	-	0.0	-	-	-	0.6	0.7	-	-
Oryza punctata	AC215214.2	113,346	84.0	-	-	83.7	-	0.3	-	-	-	-	-	-	-
Oryza punctata	AC215662.3	166,376	99.4	-	-	99.4	-	-	-	-	-	-	-	-	-
Oryza punctata	AC215729.1	134,378	97.3	-	0.2	95.6	-	-	0.3	-	-	-	-	0.3	0.9
Oryza punctata	AC215820.3	148,331	100.0	-	-	99.3	-	-	-	0.7	-	-	-	-	-
Oryza punctata	AC215826.3	162,680	95.5	-	-	95.5	-	-	-	-	-	-	-	-	-
Oryza punctata	AC216669.3	164,554	98.2	-	-	98.2	-	-	-	-	-	-	-	-	-
Oryza punctata	AC217198.1	154,395	99.2	-	-	99.1	-	-	-	-	-	-	-	-	0.1
Oryza punctata	AC217202.1	129,108	99.0	-	-	99.0	-	-	-	-	-	-	-	-	-
Oryza punctata	AC217208.1	141,569	95.9	-	2.6	91.6	-	-	0.3	-	1.4	-	-	-	-
Oryza punctata	AC217578.2	146,405	87.3	-	-	85.2	-	0.9	0.4	-	-	-	-	-	0.8
Oryza punctata	AC217579.1	126,882	94.9	-	-	94.9	-	-	-	-	-	-	-	-	-
Oryza punctata	AC217580.1	151,520	99.6	-	-	99.6	-	-	-	-	-	-	-	-	-
Oryza punctata	AC217581.2	119,156	99.6	-	-	99.6	-	-	-	-	-	-	-	-	-
Oryza punctata	AC217582.2	167,318	99.6	-	-	99.4	-	-	-	-	0.3	-	-	-	-
Oryza punctata†	AC225787.1	153,813	91.3	-	0.5	-	-	0.2	-	0.4	89.4	-	0.8	0.1	-
Oryza punctata	AC232819.1	139,969	93.7	-	1.2	91.9	0.2	0.2	-	-	-	-	-	0.3	-
Oryza punctata†	AC237098.1	158,567	96.9	-	-	-	-	-	-	0.1	96.4	0.4	-	-	0.1
Oryza punctata†	AC237099.1	158,101	81.1	-	1.2	-	0.8	-	0.2	0.3	77.2	0.2	0.6	0.5	-
Oryza punctata†	AC237100.1	123,963	98.9								98.8		0.0		

Supplementary Table 10 continued from previous page

Oryza punctata†	AC240797.1	111,942	85.0	1.9	1.2	0.5	0.5	0.3	-	0.7	77.9	0.5	0.9	0.2	0.5
Oryza punctata†	AC240798.1	181,444	94.0	-	-	-	-	-	-	-	94.0	-	-	-	-
Oryza punctata†	AC240799.1	149,350	97.8	-	-	-	-	-	-	-	97.8	-	-	-	-
Oryza punctata†	AC240800.1	160,973	97.6	-	0.3	-	-	-	-	-	97.1	-	-	-	0.2
Oryza punctata	FJ032628.1	163,589	97.5	-	-	-	-	-	97.0	-	0.5	-	-	-	-
Oryza punctata	FJ266027.1	128,102	96.5	-	-	-	0.1	-	0.3	-	-	-	0.7	95.4	-
Oryza punctata	FJ581043.1	167,491	97.2	0.4	-	-	0.0	-	96.8	-	-	-	-	-	-
Oryza punctata	GQ280266.1	105,467	93.9	0.1	-	-	-	-	93.8	-	-	-	-	-	-
Oryza punctata	HM999007.1	192,133	96.6	-	-	-	0.2	-	1.7	-	93.8	-	0.9	0.0	-
Oryza punctata	HQ827834.1	367,371	98.9	-	0.1	-	98.3	-	0.5	-	-	-	-	-	-
Oryza rufipogon	AY756174.4	102,522	95.3	0.3	88.8	-	0.8	3.2	-	-	0.4	0.9	-	0.4	0.6
Oryza rufipogon	FJ032626.1	134,693	95.0	0.3	0.4	0.1	0.1	1.2	91.2	0.5	0.4	0.3	0.1	0.1	0.4
Oryza rufipogon	FJ266028.1	155,787	92.5	0.8	0.9	0.8	2.0	-	0.2	0.2	0.7	0.8	0.8	84.6	0.6
Oryza rufipogon	FJ581045.1	225,388	90.2	0.5	0.9	2.3	0.4	2.0	76.9	1.0	1.1	0.7	0.8	2.2	1.5
Oryza rufipogon	FO681399.1	108,805	95.5	1.4	1.1	0.3	81.2	1.5	1.5	0.6	2.2	0.9	1.6	0.7	2.7
Oryza rufipogon	FQ377585.1	98,385	99.9	-	-	-	99.2	-	-	-	0.4	0.2	-	-	0.1

† BACs cloned from chromosome 8 centromere.

Supplementary Table 11. Alignment coverage of 44 finished BAC sequences versus whole genome assemblies in four species: fraction of whole BAC, annotated genes, repeats, and intergenic regions aligned.

Species	GenBank Accession	Expected Chr.	Aligned Chr.	Summed feature length (bp) and percent aligned (%)			
				Whole BAC	Gene	Transposons	Intergenic
Oryza barthii	KF284071.1	3	3	176,785 (93.9)	49,970 (96.4)	60,720 (86.3)	126,843 (92.9)
Oryza nivara	FJ032625.1	6	6	218,473 (95.8)	116,771 (99.1)	61,727 (87.1)	101,754 (92.0)
Oryza nivara	FJ266025.1	11	5, 11	204,633 (94.5)	106,300 (99.4)	85,629 (86.5)	98,349 (89.2)
Oryza nivara	FJ581047.1	6	6	324,233 (98.0)	100,133 (98.8)	132,959 (94.9)	224,146 (97.7)
Oryza nivara	GQ280265.1	6	2, 6	119,975 (75.4)	40,225 (100.0)	68,747 (56.8)	79,764 (63.0)
Oryza nivara	HM999008.1	8	1, 8	213,510 (72.2)	55,816 (92.7)	144,219 (59.6)	157,726 (65.0)
Oryza punctata	AC213149.2	3	3	139,052 (96.0)	24,693 (100.0)	64,976 (91.7)	114,381 (95.1)
Oryza punctata	AC214408.3	3	2, 3	153,085 (97.9)	60,599 (100.0)	41,479 (91.9)	92,526 (96.5)
Oryza punctata	AC215214.2	3	3	113,346 (83.9)	28,636 (100.0)	65,016 (73.0)	84,726 (78.5)
Oryza punctata	AC215662.3	3	3	166,376 (99.3)	86,453 (99.2)	38,882 (97.5)	79,965 (99.5)
Oryza punctata	AC215729.1	3	3	134,378 (97.2)	63,725 (99.2)	41,152 (92.3)	70,688 (95.4)
Oryza punctata	AC215820.3	3	3	148,331 (100.0)	61,773 (100.0)	44,492 (99.6)	86,586 (100.0)
Oryza punctata	AC215826.3	3	3	162,680 (95.5)	59,686 (100.0)	64,774 (89.4)	103,026 (92.9)
Oryza punctata	AC216669.3	3	3	164,554 (98.2)	77,525 (99.6)	51,799 (94.0)	87,069 (96.9)
Oryza punctata	AC217198.1	3	3	154,395 (99.2)	73,775 (100.0)	25,432 (95.7)	80,664 (98.5)
Oryza punctata	AC217202.1	3	3	129,108 (99.0)	69,262 (99.1)	19,248 (95.4)	59,884 (98.9)
Oryza punctata	AC217208.1	3	3	141,569 (95.9)	64,322 (100.0)	55,446 (90.0)	77,271 (92.5)
Oryza punctata	AC217578.2	3	3	146,405 (87.3)	45,437 (98.0)	65,245 (79.8)	100,986 (82.4)
Oryza punctata	AC217579.1	3	3	126,882 (94.9)	45,940 (100.0)	39,537 (83.5)	80,968 (92.0)
Oryza punctata	AC217580.1	3	3	151,520 (99.6)	75,862 (99.8)	30,625 (98.4)	75,698 (99.3)
Oryza punctata	AC217581.2	3	3	119,156 (99.6)	45,190 (100.0)	33,451 (98.6)	73,996 (99.3)
Oryza punctata	AC217582.2	3	3	167,318 (99.6)	45,218 (100.0)	32,362 (98.1)	122,126 (99.5)
Oryza punctata	AC225787.1†	8	8	153,813 (91.3)	18,955 (100.0)	115,745 (88.6)	134,866 (90.1)
Oryza punctata	AC232819.1	3	3	139,969 (93.7)	63,080 (99.8)	31,881 (73.6)	76,939 (88.7)
Oryza punctata	AC237098.1†	8	8	158,567 (96.9)	21,781 (100.0)	115,445 (95.8)	136,798 (96.4)
Oryza punctata	AC237099.1†	8	8	158,101 (81.0)	ND*	148,512 (79.8)	158,101 (81.0)
Oryza punctata	AC237100.1†	8	8	123,963 (98.9)	ND	109,662 (98.6)	123,963 (98.9)
Oryza punctata	AC240796.1†	8	8	184,672 (89.4)	13,926 (100.0)	169,584 (88.4)	170,754 (88.5)

Table continued next page

Supplementary Table 11 continued from previous page

Oryza punctata	AC240797.1†	8	8	111,942 (85.0)	393 (99.7)	101,971 (84.0)	111,551 (84.9)
Oryza punctata	AC240798.1†	8	8	181,444 (94.0)	14,804 (100.0)	138,398 (92.3)	166,644 (93.5)
Oryza punctata	AC240799.1†	8	8	149,350 (97.8)	11,190 (100.0)	111,587 (97.0)	138,164 (97.6)
Oryza punctata	AC240800.1†	8	8	160,973 (97.6)	20,941 (100.0)	122,690 (96.9)	140,042 (97.3)
Oryza punctata	FJ032628.1	6	6	163,589 (97.5)	50,669 (99.9)	61,997 (93.3)	112,948 (96.4)
Oryza punctata	FJ266027.1	11	11	128,102 (96.5)	41,554 (100.0)	53,172 (92.0)	86,566 (94.8)
Oryza punctata	FJ581043.1	6	6	167,491 (97.2)	60,357 (100.0)	57,698 (92.2)	107,162 (95.6)
Oryza punctata	GQ280266.1	6	6	105,467 (93.9)	30,597 (95.4)	57,315 (92.8)	74,886 (93.2)
Oryza punctata	HM999007.1	8	8	192,133 (96.6)	83,767 (100.0)	50,839 (87.8)	108,412 (94.0)
Oryza punctata	HQ827834.1	4	4	367,371 (98.9)	181,666 (99.3)	90,989 (96.7)	185,815 (98.5)
Oryza rufipogon	AY756174.4	2	2	102,522 (95.3)	52,112 (100.0)	25,791 (81.3)	50,424 (90.3)
Oryza rufipogon	FJ032626.1	6	6	134,693 (94.9)	74,916 (97.1)	39,057 (88.6)	59,811 (92.1)
Oryza rufipogon	FJ266028.1	11	11	155,787 (92.4)	85,173 (96.0)	71,363 (89.3)	70,634 (88.1)
Oryza rufipogon	FJ581045.1	6	6	225,388 (90.2)	69,703 (98.9)	94,028 (87.9)	155,731 (86.2)
Oryza rufipogon	FO681399.1	4	4	108,805 (95.5)	28,486 (100.0)	59,809 (91.6)	80,327 (93.9)
Oryza rufipogon	FQ377585.1	4	4	98,385 (99.9)	56,723 (99.9)	20,703 (99.0)	41,686 (99.8)

* ND, not detected. † BACs cloned from chromosome 8 centromere.

Supplementary Table 12. Alignment coverage of 44 finished BAC sequences versus whole ge assemblies in four species: fraction of annotated BAC coding exons, UTR, and introns aligne

Species	Accession	Summed feature length (bp) and percent aligned (%)			
		Coding exons	3'-UTR	5'-UTR	Intro
Oryza barthii	KF284071.1	18,054 (90.5)	1,786 (99.7)	1,172 (99.6)	29,056 (
Oryza nivara	FJ032625.1	35,472 (98.1)	7,559 (99.7)	3,794 (99.5)	70,666 (
Oryza nivara	FJ266025.1	25,862 (99.7)	3,828 (99.8)	1,689 (99.9)	75,480 (
Oryza nivara	FJ581047.1	33,359 (98.6)	5,346 (99.3)	1,403 (99.1)	60,520 (
Oryza nivara	GQ280265.1	15,861 (99.8)	1,381 (99.7)	1,595 (99.7)	22,782 (
Oryza nivara	HM999008.1	14,186 (98.5)	2,605 (99.7)	739 (98.9)	38,570 (
Oryza punctata	AC213149.2	10,461 (99.6)	2,519 (99.8)	303 (98.7)	11,780 (
Oryza punctata	AC214408.3	23,180 (99.6)	5,235 (99.8)	1,466 (99.1)	31,235 (
Oryza punctata	AC215214.2	11,415 (99.5)	1,710 (99.8)	1,186 (99.6)	14,473 (
Oryza punctata	AC215662.3	26,659 (99.6)	6,639 (99.8)	1,030 (99.0)	52,700 (
Oryza punctata	AC215729.1	23,667 (97.7)	3,481 (99.7)	1,618 (99.3)	35,348 (
Oryza punctata	AC215820.3	22,011 (99.5)	4,636 (99.8)	1,391 (99.4)	34,553 (
Oryza punctata	AC215826.3	16,791 (99.4)	4,295 (99.7)	687 (99.3)	38,210 (
Oryza punctata	AC216669.3	24,288 (99.4)	7,963 (99.8)	2,891 (99.5)	42,629 (
Oryza punctata	AC217198.1	31,757 (99.5)	5,004 (99.8)	1,780 (99.4)	35,569 (
Oryza punctata	AC217202.1	27,453 (97.3)	7,086 (99.8)	2,838 (99.6)	35,008 (
Oryza punctata	AC217208.1	21,465 (99.5)	4,817 (99.8)	862 (99.1)	37,968 (
Oryza punctata	AC217578.2	12,948 (96.1)	2,921 (99.8)	261 (75.1)	29,590 (
Oryza punctata	AC217579.1	16,514 (99.6)	3,181 (99.7)	1,009 (99.4)	26,000 (
Oryza punctata	AC217580.1	27,468 (99.1)	7,945 (99.8)	858 (99.2)	39,919 (
Oryza punctata	AC217581.2	17,442 (99.6)	4,741 (99.8)	813 (99.3)	22,397 (
Oryza punctata	AC217582.2	18,093 (99.6)	3,434 (99.8)	1,920 (99.6)	21,887 (
Oryza punctata†	AC225787.1	7,008 (99.7)	1,306 (99.5)	179 (98.9)	10,508 (
Oryza punctata	AC232819.1	27,917 (99.1)	4,374 (99.7)	711 (98.7)	30,539 (
Oryza punctata†	AC237098.1	8,274 (99.6)	733 (99.6)	ND*	12,828 (
Oryza punctata†	AC237099.1	ND	ND	ND	NI
Oryza punctata†	AC237100.1	ND	ND	ND	NI
Oryza punctata†	AC240796.1	2,859 (99.5)	528 (99.8)	ND	10,561 (
Oryza punctata†	AC240797.1	393 (99.7)	ND	ND	NI
Oryza punctata†	AC240798.1	2,007 (99.2)	3,898 (99.9)	ND	8,927 (
Oryza punctata†	AC240799.1	1,872 (99.2)	758 (99.7)	100 (99.0)	8,486 (
Oryza punctata†	AC240800.1	4,200 (99.3)	729 (99.7)	45 (97.8)	16,015 (
Oryza punctata	FJ032628.1	20,664 (99.4)	2,331 (99.8)	930 (99.0)	26,906 (
Oryza punctata	FJ266027.1	14,289 (99.6)	1,649 (99.8)	217 (99.1)	25,483 (
Oryza punctata	FJ581043.1	19,791 (99.4)	5,449 (99.7)	1,353 (99.3)	34,503 (
Oryza punctata	GQ280266.1	11,736 (88.0)	615 (99.8)	1,716 (99.9)	16,562 (
Oryza punctata	HM999007.1	28,161 (99.6)	4,329 (99.5)	2,035 (99.5)	49,569 (
Oryza punctata	HQ827834.1	70,165 (98.7)	15,858 (99.7)	3,152 (99.1)	94,326 (

Table continued next

Supplementary Table 12 continued from previous page

Oryza rufipogon	AY756174.4	21,147 (99.8)	1,868 (99.7)	545 (99.3)	28,988 (99.8)
Oryza rufipogon	FJ032626.1	19,518 (97.4)	4,024 (99.7)	2,642 (98.5)	49,960 (96.4)
Oryza rufipogon	FJ266028.1	19,120 (94.1)	1,902 (99.8)	319 (71.8)	63,940 (96.4)
Oryza rufipogon	FJ581045.1	26,718 (98.7)	2,314 (99.7)	336 (97.9)	40,521 (98.5)
Oryza rufipogon	FO681399.1	6,828 (99.6)	1,452 (99.8)	ND	20,288 (99.9)
Oryza rufipogon	FQ377585.1	20,136 (99.6)	2,956 (99.8)	395 (99.0)	33,362 (99.7)

* ND, not detected; † BACs cloned from chromosome 8 centromere.

Supplementary Table 13. Evaluation of seven new wild assemblies using paired BAC end sequences (P-BES).

Species	P-BESs that map to assemblies	% of mapped P-BESs that support assemblies*	Fold coverage of supporting P-BESs	Physical coverage (bp)	Fraction of assembly covered (%)
Oryza nivara	19,374	98	6.86	287,917,102	85
Oryza rufipogon	15,151	98	4.39	302,375,796	89
Oryza barthii	19,521	97	6.27	298,335,181	97
Oryza glumaepatula	11,622	99	3.44	339,522,898	91
Oryza meridionalis	1,808	96	0.59	175,547,045	52
Oryza punctata	23,160	99	7.65	388,043,058	99
Leersia perrieri	21,859	99	8.85	264,327,971	99

* Mapped pairs in expected orientation within assembly.

Supplementary Table 14. Assembly alignments to independently sequenced BAC clones and chromosome 3 short-arm assemblies.

Species	matched location (chr:start-end)	GenBank Accession #	Clone Length (bp)	% aligned	% identical
L. perrieri	3:35016-520052	ALNV00000000.1	500,000	95.48	99.94
O. barthii	3:23061562-23227436	KF284071	176,785	92.22	99.55
O. barthii	3:45704-557469	ABRL00000000.1	500,040	91.79	99.85
O. glumaepatula	3:139931-742755	ALNU00000000.1	500,000	89.74	99.84
O. meridionalis	3:1138-793402	ALNW00000000.1	500,000	88.87	99.87
O. nivara	1:35199800-35249039	HM999008	213,510	30.70	99.62
O. nivara	6:10330448-10532913	GQ280265	119,975	41.22	99.85
O. nivara	6:23124318-23336361	FJ032625	218,473	91.80	99.91
O. nivara	6:9514993-9839416	FJ581047	324,233	97.00	99.95
O. nivara	11:5343063-5494291	FJ266025	204,633	82.30	99.19
O. punctata	3:49310-181279	AC217208	141,569	92.87	99.97
O. punctata	3:114113-243767	AC213149	139,052	92.81	99.95
O. punctata	3:224829-344782	AC217579	126,882	94.25	99.98
O. punctata	3:338226-489942	AC232819	139,969	99.62	100.00
O. punctata	3:461824-618847	AC217198	154,395	98.41	99.89
O. punctata	3:613221-741846	AC217202	129,108	98.72	99.88
O. punctata	3:1195849-1318139	AC217581	119,156	99.59	99.20
O. punctata	3:1245550-1400078	AC217580	151,520	99.64	99.20
O. punctata	3:1525537-1686799	AC215820	148,331	99.87	99.51
O. punctata	3:1718895-1898191	AC215662	166,376	98.41	99.92
O. punctata	3:3494614-3624499	AC215729	134,378	95.62	99.96
O. punctata	3:9604814-9604246	AC216669	164,554	98.21	99.64
O. punctata	3:14588737-14652569	AC214408	153,085	41.58	100.00
O. punctata	3:14669580-14773728	AC215214	113,346	83.98	99.26
O. punctata	4:31941831-32304147	HQ827834	367,371	95.73	99.93
O. punctata	6:9384616-9547664	FJ581043	167,491	96.99	99.98
O. punctata	6:10446807-10567376	GQ280266	105,467	92.36	99.91
O. punctata	6:27819305-27978678	FJ032628	163,589	97.19	99.71
O. punctata	6:27819305-27978678	FJ032628	163,589	97.43	99.12
O. punctata	8:1081079-1205329	AC217578	146,405	98.70	99.65
O. punctata	8:1690249-1846464	HM999007	192,133	87.03	99.90
O. punctata	8:11294295-11450681	AC240800	160,973	97.60	99.52
O. punctata	8:11425461-11578430	AC240799	149,350	98.10	99.82
O. punctata	8:11601499-11761903	AC237098	158,567	97.34	99.77
O. punctata	8:11761898-11884595	AC237100	123,963	98.82	99.96
O. punctata	8:11936836-12060704	AC237099	158,101	78.98	99.78
O. punctata	8:12048599-12213780	AC240796	184,672	90.05	99.66
O. punctata	8:12199455-12288947	AC240797	111,942	82.67	99.47
O. punctata	8:12262235-12399971	AC225787	153,813	89.62	99.86
O. punctata	8:12396649-12567530	AC240798	181,444	93.75	99.97
O. punctata	11:6110313-6234861	FJ266027	128,102	95.55	99.84
O. rufipogon	2:2650689-2742167	AY756174	102,522	88.88	98.62
O. rufipogon	4:13174249-13261924	FO681399	108,805	77.53	99.93
O. rufipogon	4:18944993-19043000	FQ377585	98,385	99.50	99.71
O. rufipogon	6:8406322-8673998	FJ581045	225,388	55.30	97.25
O. rufipogon	6:21448289-21572951	FJ032626	134,693	91.06	97.86
O. rufipogon	11:5096184-5247528	FJ266028	155,787	86.41	97.67

Supplementary Table 15. Evaluation of assembly completeness with respect to gene-space using CEGMA[8] and BUSCO[9].

Species	CEGMA		BUSCO	
	Raw	Normalized*	Raw	Normalized*
Oryza sativa vg. japonica	92.3	100.0	97.9	100.0
Oryza sativa vg. indica [93-11]	92.7	100.4	97.6	99.7
Oryza sativa vg. indica [IR 8]	93.6	101.4	97.9	100.0
Oryza sativa vg. aus [N 22]	91.9	99.6	97.2	99.3
Oryza rufipogon	91.9	99.6	97.8	99.9
Oryza nivara	92.3	100.0	98.0	100.1
Oryza glaberrima	85.5	92.6	90.6	92.5
Oryza barthii	92.3	100.0	97.1	99.1
Oryza glumaepatula	92.7	100.4	98.0	100.1
Oryza meridionalis	89.9	97.4	91.8	93.8
Oryza punctata	92.3	100.0	97.2	99.3
Oryza brachyantha	93.2	100.9	95.9	98.0
Leersia perrieri	94.0	101.7	97.6	99.7

*Normalized values were obtained by dividing each raw result by the raw value obtained using the Nipponbare RefSeq.

Supplementary Table 16. Paired-end read counts by RNA-seq in nine species by three tissues. Units are millions of reads.

Species	Leaf	Root	Panicle
Oryza sativa vg. japonica	122	162	207
Oryza rufipogon	209	98	192
Oryza nivara	200	168	250
Oryza glaberrima	216	187	127
Oryza barthii	40	37	243
Oryza glumaepatula	233	166	215
Oryza meridionalis	258	247	252
Oryza punctata	36	40	138
Oryza brachyantha	NA	NA	231
Leersia perrieri	201	178	209

NA = Not applicable

Supplementary Table 17. Transcript counts after *de novo* assembly of RNA-seq reads in ten species by three tissues.

Species	Leaf	Root	Panicle
Oryza sativa vg. japonica	54,439	125,762	210,604
Oryza rufipogon	80,692	260,585	154,051
Oryza nivara	187,016	268,705	147,922
Oryza glaberrima	94,762	252,628	147,351
Oryza barthii	83,513	170,712	132,479
Oryza glumaepatula	156,220	238,106	163,959
Oryza meridionalis	160,066	355,433	203,905
Oryza punctata	79,543	163,496	124,004
Oryza brachyantha	NA	NA	163,547
Leersia perrieri	149,085	157,795	199,332

NA = Not applicable

Supplementary Table 18. Contig N50 of *de novo* assembled transcripts using RNA-seq of ten species by three tissues.

Species	Leaf	Root	Panicle
Oryza sativa vg. japonica	1,189	1,512	1,636
Oryza rufipogon	1,320	809	1,406
Oryza nivara	1,289	1,316	1,311
Oryza glaberrima	1,433	560	1,690
Oryza barthii	1,252	941	1,475
Oryza glumaepatula	1,674	546	1,599
Oryza meridionalis	1,475	970	1,437
Oryza punctata	1,330	1,388	913
Oryza brachyantha	NA	NA	1,401
Leersia perrieri	1,381	1,313	1,524

NA = Not applicable

Supplementary Table 19. Percentage of high confidence Trinity-assembled transcript clusters that mapped to the reference genome assembly*.

Species	Leaf	Root	Panicle	Ave. % ± S.D.
Oryza sativa vg. japonica	31264/31761 (98.4%)	54026/55406 (97.5%)	76676/80758 (94.9%)	96.5 ± 1.48
Oryza rufipogon	36558/39378 (92.8%)	43660/46669 (93.6%)	52548/56338 (93.3%)	93.2 ± 0.33
Oryza nivara	76642/81853 (93.6%)	49808/51491 (96.7%)	66138/69684 (94.9%)	94.9 ± 1.27
Oryza glaberrima	38846/45665 (85.1%)	11056/13081 (84.5%)	48530/57329 (84.7%)	84.8 ± 0.25
Oryza barthii	38199/40423 (94.5%)	35731/38743 (92.2%)	51003/56320 (90.6%)	92.2 ± 1.60
Oryza glumaepatula	51073/52755 (96.8%)	27922/29156 (95.8%)	58799/61556 (95.5%)	96.0 ± 0.56
Oryza meridionalis	58834/69462 (84.7%)	57989/68896 (84.2%)	63311/75550 (83.8%)	84.2 ± 0.37
Oryza punctata	34245/35540 (96.4%)	37882/40154 (94.3%)	50166/53370 (94.0%)	94.8 ± 1.07
Oryza brachyantha	NA	NA	51485/54928 (93.7%)	90.9 ± 2.16
Leersia perrieri	50907/52166 (97.6%)	47375/49161 (96.4%)	56624/58687 (96.5%)	96.8 ± 0.54

Denominator is number of transcript clusters that competitively align to *Oryza* taxon sequences in the NCBI RefSeq database. Numerator is the subset that aligns the species' reference genome. NA = Not applicable

Supplementary Table 20. Percentage of unigenes (Trinity-assembled transcripts clustered across three tissues) that mapped to reference genome assemblies.

Species	Unigene mapping
Oryza sativa vg. japonica	71286/81751 (96.8)
Oryza rufipogon	55954/61373 (94.2)
Oryza nivara	76291/88436 (95.9)
Oryza glaberrima	52485/51845 (86.1)
Oryza barthii	52925/57631 (92.7)
Oryza glumaepatula	57698/65026 (96.4)
Oryza meridionalis	78381/81503 (85.0)
Oryza punctata	53737/60292 (95.6)
Oryza brachyantha	63314/65535 (92.4)
Leersia perrieri	58799/74367 (97.4)

Supplementary Table 21. Repeat abundance and composition in 13 assembled *Oryzeae* genomes. The genome fraction occupied by each different repeat type is reported.

	Species:	*O. sativa* vg. **japonica**	*O. sativa* vg. **indica [93-11]**	*O. sativa* vg. **indica [IR 8]**	*O. sativa vg.* **aus [N 22]**	*O. rufipogon*	*O. nivara*	*O. glaberrima*	*O. barthii*	*O. glumaepatula*	*O. meridionalis*	*O. punctata*	*O. brachyantha*	*Leersia perrieri* Outgroup
	Genome Type:	**AA**	**AA**	**AA**	**AA**	**AA**	**AA**	**AA**	**AA**	**AA**	**AA**	**BB**	**FF**	
Class I (Retrotransposons)														
LTR	*Copia*	4.08	3.56	3.13	3.66	3.62	3.07	3.8	3.32	2.57		5.17	3.14	3.91
	Gypsy	19.44	15.87	21.89	18.77	13.78	10.57	12.55	11.36	9.12	7.51	24.13	5.98	8.32
	Retrovirus	0.14	0.07	0.1	0.09	0.07	0.03	0.04	0.01	0.02	0.01	0.07	0.01	0.02
LINE	*L1*	1.48	1.51	1.15	1.37	1.57	1.54	1.6	1.72	1.35	1.26	1.26	0.68	2.04
SINE		0.49	0.51	0.09	0.50	0.55	0.53	0.53	0.59	0.47	0.46	0.27	0.16	0.27
Other Class I		0.36	0.33	0.28	0.32	0.32	0.28	0.26	0.26	0.23	0.14	0.27	0.26	0.46
Class II (DNAt) subclass 1														
TIR	*Tc1–Mariner*	2.83	2.95	0.34	2.56	3.13	3.04	2.96	2.84	2.73	2.06	1.46	2.5	1.11
	hAT	1.29	1.25	0.46	1.23	1.3	1.27	1.22	1.31	1.06	0.99	1.22	1.03	0.92
	Mutator	4.69	4.36	1.95	4.41	4.43	4.2	4.25	3.99	3.57	3.11	3.64	2.41	1.76
	PIF–Harbinger	3.63	3.77	0.7	3.29	4.01	3.89	3.77	3.61	3.39	2.84	2.05	5.42	1.9
	CACTA	4.06	2.84	2.76	3.70	3.01	2.26	2.78	2.18	1.83	1.51	5.45	1.13	1.34
Class II (DNAt) subclass 2														
Helitron		2.16	2.12	1.13	2.00	2.31	2.13	1.89	2.06	1.79	1.05	1.59	0.54	0.33
Other Class II		2.14	2.14	0.54	2.79	2.28	2.2	2.18	3.53	1.94	2.6	1.9	3.63	2.71
Total TEs		46.8	**41.27**	**34.52**	**44.70**	**40.37**	**34.99**	**37.83**	**36.77**	**30.09**	**26.02**	**48.48**	**26.91**	**25.08**
Ribosomal DNA		0.11	0.09	0.04	0.07	0.06	0.05	0.06	0.03	0.04	0.05	0.05	0.05	0.04
Structural Repeats		1.28	0.97	2.29	1.00	1.22	0.95	0.92	0.86	0.83	0.8	0.66	1.16	1.02
norgDNA		0.40	0.29	NA	NA	0.41	0.30	0.19	0.17	0.15	0.27	0.30	0.08	0.06
Unclassified		0.45	0.44	0.14	0.55	0.47	0.44	0.45	0.65	0.40	0.35	0.45	0.56	0.63
Total Repeats		**49.04**	**43.06**	**36.99**	**46.32**	**42.53**	**36.73**	**39.45**	**38.48**	**31.51**	**27.49**	**49.94**	**28.76**	**26.83**

Supplementary Table 22. Transcript counts after reference-guided assembly of RNA-seq reads in nine species by three tissues.

Species	Leaf	Root	Panicle
Oryza sativa vg. japonica	29,012	45,781	53,228
Oryza rufipogon	39,565	38,778	57,471
Oryza nivara	45,739	47,530	56,738
Oryza glaberrima	59,699	42,651	65,069
Oryza barthii	30,758	30,765	48,580
Oryza glumaepatula	44,408	20,792	47,628
Oryza meridionalis	38,176	39,032	44,621
Oryza punctata	29,090	26,775	50,167
Oryza brachyantha	NA	NA	51,194
Leersia perrieri	66,701	67,803	74,786

NA = Not applicable

Supplementary Table 23. Detection and intersection of 13,397 highly conserved orthologous genes within gene annotations and transcriptome data in 11 *Oryzeae* species.*

Species	ort_ann+ (%)	ort_txp+ (%)	ort_ann–	% ort_ann+ in ort_txp+	% ort_ann– in ort_txp+	chi-square
Oryza sativa vg. japonica†	13397 (100)	13131 (98.0)	0	98.01	n/a	n/a
Oryza sativa vg. indica [93-11]‡	13300 (99.3)	n/a	97	n/a	0.00	n/a
Oryza nivara	13256 (98.9)	13082 (97.6)	141	97.70	92.91	1.88E-04
Oryza rufipogon	13308 (99.3)	12941 (96.6)	89	96.63	92.13	1.99E-02
Oryza glaberrima	12576 (93.9)	12839 (95.8)	821	95.94	94.28	2.10E-02
Oryza barthii	13210 (98.6)	12925 (96.5)	187	96.62	86.63	1.92E-13
Oryza glumaepatula	13230 (98.8)	12981 (96.9)	167	96.94	93.41	9.05E-03
Oryza meridionalis	12575 (93.9)	12742 (95.1)	822	95.17	94.28	2.55E-01
Oryza punctata	13201 (98.5)	12885 (96.2)	196	96.51	73.98	6.04E-60
Oryza brachyantha	12781 (95.4)	13066 (97.5)	616	98.61	75.16	1.75E-293
Leersia perrieri†	13397 (100)	13015 (97.1)	0	97.15	n/a	n/a

* A collection of 13,397 sets of orthologous genes was defined with characteristics of conservation in *Arabidopsis thaliana* and having annotated members in both *O. sativa* vg. japonica and *L. perrieri*, and thus expected to exist in all *Oryza*. Presence (ort_ann+) or absence (ort_ann-) of each ortholog among annotated loci was scored for each species. *De novo* assembled transcripts from each species were assigned to orthologous sets after alignment against the entire set of annotated proteins across all eleven species. This enabled the discovery of transcripts corresponding to each species' ortholog even if not annotated in that species, and the number of expected orthologs having transcriptome evidence was scored (ort_txp+). Analysis compared discovery rates between annotated vs. non-annotated (e.g. putatively missing) orthologs within the transcriptome data, with a null-hypothesis predicting no difference. †Selected ortholog sets always included representatives from these species and therefore do not possess non-annotated orthologs. ‡ RNA-seq data was not collected for this species.

Supplementary Table 24. Success rates mapping transcripts of annotated or non-annotated (putative missing) genes to reference assemblies*.

	Annotated (ort_ann+)		Non-annotated (ort_ann–)		
Species	# ort_txp+	% ref–	# ort_txp+	% ref–	chi-square
Oryza sativa vg. japonica†	13131	0.07	n/a	n/a	n/a
Oryza sativa vg. indica [93-11]‡	n/a	n/a	n/a	n/a	n/a
Oryza nivara	12951	0.12	131	4.58	5.96E-37
Oryza rufipogon	12859	0.19	82	2.44	5.61E-06
Oryza glaberrima	12065	0.65	774	61.37	0
Oryza barthii	12763	0.52	162	15.43	1.03E-112
Oryza glumaepatula	12825	0.09	156	0.0	0.71
Oryza meridionalis	11967	0.49	775	58.06	0
Oryza punctata	12740	0.25	145	19.31	2.40E-246
Oryza brachyantha	12603	0.21	463	15.77	0
Leersia perrieri†	13015	0.08	n/a	n/a	n/a

*13,397 highly conserved sets of orthologous genes showing homology to *Arabidopsis thaliana* and having annotated members in both *O. sativa* vg. japonica and *L. perrieri*, thus expected to have existed in the common ancestor of all *Oryza*. Presence (ort_ann+) or absence (ort_ann-) of each ortholog among annotated loci was scored for each species. *De novo* assembled transcripts from each species were assigned to orthologous sets after aligning to the entire set of annotated proteins across all eleven species and the number of expected orthologs having transcriptome evidence was scored (ort_txp+). Transcripts were also scored with respect to their ability to align to the genome reference assembly, and the percent of ort_txp+ orthologs whose transcripts failed to align to its reference assembly is given as %ref-. Testing the relationship between these categories, the null hypothesis predicts no difference in mappability of transcripts to the reference assembly between those having annotation (ort_ann+) and those missing annotation (ort_ann-). †Selected ortholog sets always included representatives from these species and therefore did not possess non-annotated orthologs. ‡RNA-seq data was not collected for this species.

Supplementary Table 25. Putative "split model" annotation artifacts.

Species	Count	Pct. loci (%)
Oryza sativa vg. indica [93-11]	650	1.69
Oryza glumaepatula	608	1.70
Oryza barthii	538	1.56
Oryza glaberrima	488	1.56
Oryza sativa vg. aus [N 22]	487	1.35
Oryza sativa vg. japonica	478	1.24
Oryza rufipogon	434	1.17
Oryza meridionalis	432	1.45
Oryza punctata	422	1.33
Oryza nivara	392	1.08
Oryza sativa vg. indica [I R8]	368	1.04
Leersia perrieri	330	1.13
Oryza brachyantha	230	0.95

Supplementary Table 26. **Per-chromosome summary of data used in phylogenomic analyses.**

Chromosome	11-taxon clusters	Supermatrix length (bp)
1	1042	2,007,918
2	888	1,737,433
3	956	1,828,388
4	577	1,118,748
5	585	1,105,228
6	530	1,074,990
7	393	799,690
8	292	541,352
9	277	555,435
10	158	320,466
11	142	277,934
12	175	376,961
total	6015	11,744,543

Supplementary Table 27. Divergence time estimates (MYR) within *Oryza*, by chromosome.

	Node name (node #)							
Chr.	AA-BB (1)	AA (2)	glum.-Asian-African (3)	Asian AA-African AA (4)	African AA (5)	Asian AA (6)	indica-nivara (7)	japonica-rufipogon (8)
1	6.67	2.39	0.97	0.75	0.17	0.54	0.31	0.27
2	6.68	2.56	0.9	0.78	0.19	0.5	0.3	0.23
3	6.71	2.48	0.89	0.72	0.17	0.48	0.31	0.27
4	6.74	2.47	0.92	0.75	0.12	0.46	0.34	0.26
5	6.63	2.49	0.93	0.82	0.2	0.53	0.33	0.24
6	6.93	2.68	1.06	NA	0.18	NA	0.79	0.28
7	6.98	2.39	0.9	0.78	0.2	0.52	0.36	0.3
8	6.83	2.43	0.93	0.74	0.16	0.48	0.34	0.24
9	6.7	2.32	0.92	0.79	0.13	0.53	0.29	0.28
10	6.57	2.17	0.95	0.82	0.19	0.52	0.34	0.34
11	6.9	2.2	1.26	1.14	0.37	0.81	0.64	0.32
12	6.83	2.3	1.02	NA	0.21	0.64	0.38	0.25
mean (stderr)	6.76 (0.13)	2.41 (0.15)	0.97 (0.11)	0.81 (0.12)	0.19 (0.06)	0.55 (0.10)	0.39 (0.16)	0.27 (0.03)

"NA" indicates a node not appearing in the ML supermatrix tree inferred for a given chromosome.

Supplementary Table 28. ABBA-BABA analysis of introgression by chromosome.

Chromosome	Alignment length[1]	D-score	Block jackknife std. error	Z score[2]	(+) sites[3]	(-) sites[3]	(*) sites[3]
1	8722728	+0.3848	0.0964	+3.9896*	12586	5592	12849
2	6847560	+0.2124	0.0852	+2.4926	8167	5306	9538
3	7769376	+0.3801	0.0511	+7.4432*	7997	3592	10094
4	4573584	+0.4486	0.0790	+5.6800*	7820	2977	8486
5	4516200	+0.4592	0.0641	+7.1611*	5963	2210	6882
6	4654368	+0.2997	0.0829	+3.6143*	7621	4106	8318
7	3814056	+0.3988	0.0776	+5.1365*	7524	3234	5044
8	3146328	+0.3836	0.0566	+6.7744*	4711	2099	5649
9	2913768	+0.3740	0.0775	+4.8290*	3896	1775	4518
10	2368008	+0.4262	0.1300	+3.2789*	4280	1722	4434
11	2317536	+0.1625	0.1519	+1.0696	4678	3370	5599
12	1817784	+0.4069	0.1222	+3.3294*	3762	1586	3623

[1]Total length of aligned k-mer blocks used (nt).
[2]Significance level above 3.0 indicated by asterisk.
[3] (+) sites imply *O. glumaepatula* is sister to *O. barthii*; (-) sites imply *O. glumaepatula* is sister to *O. rufipogon*; (*) sites agree with species tree: *O. barthii* and *O. rufipogon* are sister groups.

Supplementary Table 29. Apparent divergence times within and between *Oryza* species and *Leersia perrieri* for the initial 2.2Mb (start) and the remaining region (end) of chromosomes 11 and 12. Dates were calculated based on the median values for Bayesian distances (Online Methods). Values (in millions of years) are given for the intra-species distance (chr11 *vs*. chr12, first 7 lines) and for orthologous chromosomes inter-species *(*chr11 *vs*. chr11 and chr11 *vs*. chr12, last 12 lines). osa = *O. sativa*, oru=*O. rufipogon*, oba = *O. barthii*, ogl = *O. glaberrima*, opu = *O. punctata*, obr = *O. brachyantha*, lpe = *Leersia perrieri*.

	start		end	
species	**distance**	**date (MYR)**	**distance**	**date (MYR)**
O. sativa vg. japonica	0.07	3.4	0.16	8.55
O. rufipogon	0.07	3.77	0.19	9.99
O. barthii	0.07	3.68	0.16	8.34
O. glaberrima	0.08	3.94	0.17	8.97
O. punctata	0.06	3.3	0.18	9.52
O. brachyantha	0.05	2.65	0.24	12.71
L. perrieri	0.09	4.9	0.23	11.96
osa11/oru11	0.01	0.56	0.01	0.78
osa11/oru12	0.01	0.56	0.01	0.78
osa11/oba11	0.02	1.18	0.03	1.58
osa11/oba12	0.02	1.13	0.03	1.71
osa11/ogl11	0.02	1.17	0.03	1.63
osa11/ogl12	0.02	1.21	0.03	1.71
osa11/opu11	0.12	6.46	0.13	6.68
osa11/opu12	0.12	6.45	0.14	7.08
osa11/obr11	0.23	11.74	0.19	9.76
osa11/obr12	0.22	11.72	0.19	9.72
osa11/lpe11	0.23	12.17	0.21	10.81
osa11/lpe12	0.23	12.12	0.21	10.78

Supplementary Table 30. Apparent divergence times within and between *Oryza* AA and BB genome species for the initial 2.2Mb of chromosomes 11 and 12. Dates were calculated based on the median values for Bayesian distances (see Online Methods). Values (in millions of years) are given for the intra-species distance (chr11 *vs.* chr12, first 8 lines) and for orthologous chromosomes inter-species *(average of* chr11 *vs*. chr11 and chr12 *vs*. chr12, last 7 lines). osa = *O. sativa*, oru=*O. rufipogon*, oba = *O. barthii*, ogl = *O. glaberrima*, oni = *O. nivara*, ogu = *O. glumaepatula*, ome = *O. meridionalis*, opu = *O. punctata*.

Species	date (MYR)
O. sativa vg. japonica	3.76
O. rufipogon	3.92
O. barthii	4.1
O. glaberrima	3.97
O. nivara	4.53
O. glumaepatula	3.74
O. meridionalis	3.69
O. punctata	3.68
osa/oru	0.41
osa/oba	1.04
osa/ogl	1.09
osa/oni	1.12
osa/ogu	1.13
osa/ome	3.62
osa/opu	6.8

Supplementary Table 31. Twelve chromosomal inversions of 5 or more genes within internal branches of the *Oryza* genus.

Branch*	Length (kb)	Gene count	Chromosome	Start position
R-A	100	8	2	8,106,666
A-B	300	19	2	18,778,609
A-B	241	15	5	16,465,742
A-B	170	14	2	34,564,579
A-B	138	14	2	23,653,591
A-B	105	9	3	12,542,586
A-B	67	10	12	3,176,080
B-C	67	9	5	25,295,457
D-E	60	6	11	18,728,880

*Branches between nodes as labeled in Fig. 3.

Supplementary Table 32. LTR-RT families in the 13 assembled *Oryzeae* genomes. The 50 most abundant *Gypsy*, *Copia*, and the unclassified LTR-RT families are listed. The three most abundant families in each genome (highlighted) reveal high uniformity among AA genomes, in contrast with high heterogeneity outside of AA genomes. Underlined numbers represent families of elements found only in one genome.

Genome type	**AA**	**AA**	**AA**	**AA**	**AA**	**AA**	**AA**	**AA**	**AA**	**AA**	**BB**	**FF**	**-**	
Family #	***O. s.* vg. japonica**	***O. s.* vg. indica – IR 8**	***O. s.* vg. indica- N 22**	***O. s.* vg. indica-93-11**	***O. rufipogon***	***O. nivara***	***O. glaberrima***	***O. barthii***	***O. glumaepatula***	***O. meridionalis***	***O. punctata***	***O. brachyantha***	***L. perrieri***	**Family total**
1	185	304	185	80	58	43	82	49	33	13	367	-	-	1399
2	403	319	262	95	-	-	-	-	-	-	-	-	-	1079
3	182	196	179	91	83	60	90	71	32	14	-	-	-	998
4	119	106	114	75	105	64	103	96	52	26	-	-	-	860
5	89	164	150	60	34	58	43	33	24	3	127	7	23	815
6	106	103	90	81	98	82	111	77	23	-	-	-	-	771
7	162	261	202	27	27	17	32	11	9	1	1	-	-	750
8	29	396	181	30	5	7	6	3	1	-	3	-	-	661
9	158	311	122	21	7	-	20	1	-	-	2	-	-	642
10	149	130	115	59	27	19	55	33	15	3	-	-	-	605
11	-	1	1	-	-	-	-	-	-	-	598	-	-	600
12	147	109	128	38	38	30	55	22	3	-	9	7	6	592
13	90	66	68	46	69	51	85	62	28	18	-	-	-	583
14	77	83	80	37	44	42	60	29	8	2	30	7	-	499
15	81	100	85	30	26	16	32	18	5	-	19	2	1	415
16	118	180	81	8	-	-	-	-	-	-	-	-	-	387
17	45	41	44	19	26	28	30	20	3	4	25	-	-	285
18	9	126	96	38	3	3	2	1	-	-	-	-	-	278
19	37	38	29	27	27	30	25	20	11	2	3	-	-	249
20	44	42	36	18	13	9	15	6	2	2	29	10		226
21	38	45	44	35	8	5	17	5	-	-	14	6	4	221
22	31	28	31	19	27	21	25	18	16	4	-	-	-	220
23	57	52	49	14	3	2	10	1	-	3	5	-	-	196
24	31	26	27	18	26	8	18	5	7	11	-	-	-	177
25	26	33	36	14	8	8	22	2	1		27	-	-	177
26	-	-	-	-	-	-	-	-	-	-	175	-	-	175
27	-	-	-	-	-	-	-	-	-	-	149	-	-	149
28	16	17	19	12	13	13	15	12	10	-	-	3		130
29	12	19	26	6	14	15	19	7	-	-	12	-	-	130
30	10	16	18	6	7	10	7	3	1	1	26	16	4	125

Table continued next page

Supplementary Table 32 continued from previous page

31	13	6	9	3	9	7	7	6	2	2	50	1	4	119
32	20	10	14	5	15	3	19	3	2	-	25	-	-	116
33	18	35	25	11	3	2	2	2	1	-	14	-	-	113
34	18	20	20	6	14	11	15	5	1	-	-	-	-	110
35	10	33	20	7	5	9	11	3	1	-	4	-	7	110
36	-	-	-	-	-	-	-	-	-	-	109	-	-	109
37	14	10	12	11	12	10	16	13	6	3	-	-	-	107
38	15	20	28	1	2	4	8	-	-	-	24	-	-	102
39	19	15	14	10	5	6	16	4	5	-	-	-	-	94
40	13	4	4	8	9	11	12	10	11	3	2	-	-	87
41	-	-	-	-	-	-	-	-	-	-	-	-	78	78
42	8	2	7	2	7	9	5	7	5		21	-	-	73
43	9	5	4	6	8	5	5	3	2	1	21	-	-	69
44	8	11	9	7	3	5	9	4	3	2	8	-	-	69
45	4	6	6	4	6	8	4	3	4	1	3	-	1	50
46	2	3	2	1	3	2	2	2	2	1	30	-	-	49
47	-	-	-	-	-	-	-	-	-	-	-	-	40	40
48	-	-	-	-	-	-	-	-	-	-	39	-	-	39
49	5	-	-	4	4	3	8	8	4	-	-	-	-	36
50	-	-	-	-	-	-	-	-	-	-	36	-	-	36
Total	3,224	3,491	2,672	1,885	1,491	1,275	1,517	1,038	632	337	2,882	395	859	16,000

Supplementary Table 33. Summary of TRIMs identified in 13 genomes. The numbers in parenthesis indicate elements located in exons, intron, and 1.5 kb upstream of a gene, respectively.

	Copy	Genome coverage		Gene-related TRIMs	
Species	**number**	**kb**	**%**	**Counts**	**Percentage (%)**
O. sativa vg. japonica	2,911	689.0	0.18	1,157 (160/855/292)	39.7
O. sativa vg.indica [93-11]	3,252	761.0	0.19	997 (106/665/226)	30.7
O. sativa vg.indica [IR 8]	2,954	682.5	0.18	1,113(181/639/293)	37.7
O. sativa vg. aus [N 22]	2,747	681.5	0.18	1,101 (154/629/318)	40.0
O. rufipogon	2,724	664.9	0.20	1,101 (157/666/278)	40.4
O. nivara	2,248	574.4	0.19	855 (108/506/241)	38.0
O. glaberrima	2,297	545.1	0.18	736 (94/444/198)	32.0
O. barthii	2,384	539.7	0.18	826 (71/498/257)	34.6
O. glumipatula	2,470	577.8	0.19	1,008 (110/660/238)	40.8
O. meridionalis	2,351	564.2	0.22	829 (53/558/218)	35.3
O. punctata	1,659	466.3	0.12	470 (65/323/82)	28.3
O. brachyantha	1,699	402.4	0.17	701 (50/536/115)	41.3
Leersia perrieri	1,624	413.9	0.16	485 (29/323/133)	29.9
Average	**2,409**	**581.8**	**0.18**	**875 (103/562/210**	**36.1**

Supplementary Table 34. Indels inferred from 6 comparisons of *Oryza* genomes.*

Pairwise comparisons	**Total (Putative)**	**Inferable Indels**	**Insertions / Deletions (Target species)**	**Insertions / Deletions (Query species)**	**Outgroup species**
japonica *vs*. indica [93-11]	495,777[H]	253,386[A] (51%)[F]	41,986[B] / 58,795[C]	67,471[D] / 85,134[E]	*O. barthii*
japonica *vs*. O. *rufipogon*	352,510	161,934 (46%)	28,079 / 38,506	46,258 / 49,091	*O. barthii*
japonica *vs*. *O. nivara*	431,420	228,184 (53%)	45,022 / 61,555	49,162 / 72,445	*O. barthii*
japonica *vs*. *O. glaberrima*	593,950	308,955 (52%)	71,249 / 90,700	59,450 / 87,556	*O. glumaepatula*
japonica *vs*. *O. barthii*	699,587	360,625 (52%)	82,599 / 107,123	67,783 / 103,120	*O. glumaepatula*
O. glaberrima vs. *O. barthii*	216,059	91,434 (42%) (46,307 >1bp)	19,783 / 24,80 (8,249 / 11,656)	20,154 / 26,692 (10,326 / 16,076)	*O. sativa* vg. japonica

*Note: A = B+C+D+E; F=A/H

Supplementary Table 35. Polymorphism and fixation of derived indels in populations of *O. glaberrima* and *O. barthii*.

a. Polymorphic distribution of the derived indels (>1 bp) in *O. glaberrima* that inferred from the comparison with *O. barthii* by using *O. sativa* vg. japonica as the outgroup species

Types	**Polymorphic in *O. glaberrima* populations**		**Fixed in *O. glaberrima* populations**		**Total**
	Polymorphic in *O. barthii* group	**Absent in *O. barthii* group**	**Polymorphic in *O. barthii* group**	**Absent in *O. barthii* group**	
Insertions	**3,192**	**129**	**984**	**0**	**4,305**
Deletions	**5,242**	**537**	**2,110**	**2**	**7,891**

b. Polymorphic distribution of the derived indels (>1 bp) in *O. barthii* that inferred from the comparison with *O. glaberrima* by using *O. sativa* vg. japonica as the outgroup species.

Types	**Polymorphic in *O. barthii* populations**		**Fixed in *O. barthii* populations**		**Total**
	Polymorphic in *O. glaberrima* group	**Absent in *O. glaberrima* group**	**Polymorphic in *O. glaberrima* group**	**Absent in *O. glaberrima* group**	
Insertions	**2,683**	**1,429**	**1**	**0**	**4,113**
Deletions	**3,482**	**3,579**	**2**	**0**	**7,063**

Supplementary Table 36. Gene Ontology enrichment in *Poaceae*-derived families in *O. sativa* vg. japonica.

GO term	GO definition	Foreground	Background	FDR-corrected p-value	Associated InterPro structures
GO:0006952	defense response	304/2117	451/18273	5.00E-161	IPR002182:NB-ARC IPR001611:Leucine-rich repeat IPR000916:Bet v I domain IPR002411:Cereal allergen/alpha-amylase inhibitor, rice-type
GO:0005515	protein binding	904/2117	4628/18273	1.20E-40	IPR001810:F-box domain IPR000210:BTB/POZ-like
GO:0004857	enzyme inhibitor activity	78/2117	133/18273	4.30E-36	IPR000864:Proteinase inhibitor I13, potato inhibitor I IPR003465:Proteinase inhibitor I20, Pin2 IPR000877:Proteinase inhibitor I12, Bowman-Birk IPR006501:Pectinesterase inhibitor domain IPR002411:Cereal allergen/alpha-amylase inhibitor, rice-type
GO:0009611	response to wounding	19/2117	22/18273	3.70E-13	IPR000864:Proteinase inhibitor I13, potato inhibitor I IPR002411:Cereal allergen/alpha-amylase inhibitor, rice-type
GO:0051248	negative regulation of protein metabolic process	22/2117	40/18273	3.70E-09	IPR001574:Ribosome-inactivating protein

Supplementary Table 37. Gene Ontology enrichment in *Oryzeae*-derived families in *O. sativa* vg. japonica.

GO term	GO definition	Foreground	Background	FDR-corrected p-value	Associated InterPro structures
GO:0003676	nucleic acid binding	121/533	2753/18273	2.30E-03	IPR002119:Histone H2A IPR001356:Homeobox domain IPR003441:NAC domain IPR003902:Transcription regulator, GCM-like IPR017930:Myb domain IPR001471:AP2/ERF domain IPR003340:B3 DNA binding domain IPR001739:Methyl-CpG DNA binding IPR001606:ARID/BRIGHT DNA-binding domain IPR003957:Transcription factor, NFYB/HAP3 subunit IPR003657:DNA-binding WRKY IPR027725:Heat shock transcription factor family

Supplementary Table 38. Fraction of loci at conserved syntenic positions among 13 *Oryzeae* species, fractionated by taxon bin.

Species	*Magnoliophyta*		*Poaceae*		*Oryzeae*	
Oryza sativa vg. japonica	21787/21997	(99.0)	4328/4439	(97.5)	9516/10448	(91.1)
Oryza sativa vg. indica [93-11]	22142/22860	(96.9)	4302/4677	(92.0)	7855/8942	(87.8)
O. sativa vg. indica [IR 8]	19846/20096	(98.8)	3634/3721	(97.7)	8414/9414	(89.4)
O. sativa vg. aus [N 22]	19049/19584	(97.3)	3496/3692	(94.7)	7734/9047	(85.5)
Oryza rufipogon	21505/21714	(99.0)	4201/4376	(96.0)	8998/9872	(91.1)
Oryza nivara	21048/21670	(97.1)	4031/4259	(94.6)	8128/9225	(88.1)
Oryza glaberrima	19000/19242	(98.7)	3578/3694	(96.9)	6841/7486	(91.4)
Oryza barthii	20932/21278	(98.4)	3946/4114	(95.9)	7472/8247	(90.6)
Oryza glumaepatula	21267/21590	(98.5)	3982/4143	(96.1)	7051/8191	(86.1)
Oryza meridionalis	17428/18200	(95.8)	3037/3359	(90.4)	3710/5537	(67.0)
Oryza punctata	20536/21240	(96.7)	3305/3861	(85.6)	1744/3291	(53.0)
Oryza brachyantha	17532/18315	(95.7)	1939/2255	(86.0)	692/980	(70.6)
Leersia perrieri	19186/20567	(93.3)	2939/3676	(80.0)	917/1333	(68.8)
Totals:	261258/268353	(97.4)	46718/50266	(92.9)	79072/92013	(85.9)

Supplementary Table 39. Taxon origin of putative MULE-derived loci in protein-coding gene annotations.

Species	Count	Fraction of MULE-derived loci				
		Magnoliophyta	*Poaceae*	*Oryzeae*	Species-specific	Syntenic
Oryza sativa vg. japonica	1507	0.18	0.05	0.64	0.13	0.75
Oryza sativa vg. indica [93-11]	1058	0.18	0.06	0.71	0.05	0.82
Oryza nivara	1312	0.19	0.06	0.66	0.09	0.78
Oryza rufipogon	1191	0.18	0.05	0.67	0.10	0.83
Oryza glaberrima	827	0.18	0.05	0.70	0.07	0.82
Oryza barthii	985	0.19	0.05	0.69	0.07	0.84
Oryza meridionalis	922	0.60	0.12	0.19	0.09	0.81
Oryza glumaepatula	1057	0.19	0.05	0.63	0.13	0.74
Oryza brachyantha	85	0.42	0.04	0.08	0.46	0.19
Oryza punctata	208	0.22	0.10	0.25	0.43	0.11
Leersia perrieri	54	0.33	0.04	0.00	0.63	0.15

Supplementary Table 40. NB-ARC domain genes (n = 5,408) identified in 13 *Oryzeae* species and distribution in clusters.

Species	Gene count	Gene fraction in clusters (%)*	Cluster count	Cluster size (mean, max)	Cluster fraction heterogeneous (%)†
O. sativa vg. japonica	464	59.9	82	3.4, 10	28.0
O. sativa vg. indica [93-11]	535	59.6	109	2.9, 12	33.9
O. sativa vg. indica [IR 8]	439	58.8	77	3.4, 22	26
O. sativa vg. aus [N 22]	407	57.0	77	3.0, 12	41.6
O. nivara	453	55.4	89	2.8, 10	34.8
O. rufipogon	474	58.2	87	3.2, 16	32.2
O. glaberrima	362	62.4	73	3.1, 11	38.4
O. barthii	468	64.7	94	3.2, 16	34.0
O. glumaepatula	447	60.4	85	3.2, 10	34.1
O. meridionalis	398	57.3	70	3.3, 13	34.3
O. punctata	331	55.9	62	3.0, 10	33.9
O. brachyantha	237	53.6	47	2.7, 8	34.0
L. perrieri	393	65.1	85	3.0, 12	38.8

*Genes are positionally clustered if no more than five genes distant from a second NB-ARC gene.
†A cluster is heterogeneous if composed of genes from different gene families

Supplementary Table 41. Species counts and root taxon of NLR gene families in 13 *Oryzeae*.

Family	Root taxon*	Family size (genes/species)†	Gene counts‡
A	Mag	28.2 ± 7.5	27/21/25/18/33/29/25/24/40/25/26/47/27
B	Poa	22.8 ±3.4	27/16/18/25/19/28/23/23/23/27/23/22/22
D	Poa	22.5 ± 5.2	25/14/11/21/25/24/16/26/28/27/26/24/26
C	Poa	20.7 ± 5.3	18/9/19/16/28/29/20/21/25/26/21/21/16
H	Mag	20.2 ± 4.4	19/13/11/21/22/24/17/27/26/22/20/21/20
F	Poa	19.3 ± 4.0	26/11/18/22/22/20/15/18/25/18/22/19/15
G	Mag	19.7 ± 5.0	18/8/14/19/20/20/18/23/25/30/20/19/22
I	Poa	18.6 ± 5.1	19/13/15/19/20/19/7/21/26/20/27/21/15
J	Poa	18.7 ± 3.5	16/10/19/21/21/19/14/23/23/19/20/19/19
K	Poa	17.9 ± 4.0	16/9/16/23/23/15/20/21/22/16/21/14/17
M	Poa	16.9 ± 3.3	11/13/12/20/17/19/15/19/21/21/20/16/16
O	Poa	16.5 ± 5.0	17/5/10/15/14/20/20/19/27/16/18/18/15
E	Poa	16.1 ± 2.9	18/10/17/12/17/20/13/16/20/18/18/15/15
L	Mag	16.4 ± 2.5	14/12/15/17/20/20/16/18/17/14/16/20/14
N	Poa	16.4 ± 4.3	15/4/16/15/19/21/13/16/20/21/19/18/16
P	Poa	15.1 ± 4.4	22/10/10/10/19/19/10/15/17/21/18/11/14
R	Poa	13.2 ± 4.3	8/7/7/14/15/14/16/18/20/18/18/18/12
Q	Poa	12.9 ± 3.4	7/5/12/15/14/15/11/16/18/14/15/12/13
U	Mag	12.9 ± 3.8	7/7/7/13/14/14/11/14/19/15/18/16/13
T	Poa	12.3 ± 3.7	10/6/8/8/12/17/14/12/18/17/16/11/11
S	Poa	13.6 ± 4.6	11/7/8/11/9/15/15/14/18/17/12/15/25
V	Poa	9.8 ± 2.6	11/5/10/9/11/13/6/13/12/13/9/8/7
W	Poa	9.8 ± 3.1	6/7/5/8/8/12/9/9/11/10/14/13/16
X	Poa	6.4 ± 0.9	5/4/7/7/7/7/6/7/7/6/6/7/7
Y	Poa	5.6 ± 1.6	7/3/6/7/4/5/3/5/9/7/6/6/5
Z	Mag	3.7 ± 0.5	4/3/4/4/4/3/4/4/4/4/4/3/3
AA	Poa	2.8 ± 1.1	2/3/5/4/3/2/1/4/3/3/3/1/3
CC	Mag	2.0 ± 1.0	2/1/2/1/3/2/2/2/3/4/3/1/0
DD	Poa	0.9 ± 0.3	1/1/1/0/1/1/1/1/1/1/1/1/1
BB	Mag	0.8 ± 0.4	1/0/1/1/1/1/0/1/1/1/1/1/0
FF	Poa	0.8 ± 0.4	1/0/1/1/0/1/1/1/1/1/1/0/1
HH	Poa	0.6 ± 0.6	0/0/0/1/0/0/0/1/2/1/1/1/1
GG	Poa	0.3 ± 0.6	0/0/0/0/1/0/0/0/2/0/1/0/0
II	Ory	0.2 ± 0.4	0/0/1/0/0/0/0/1/0/1/0/0/0
JJ	Poa	0.2 ± 0.4	1/0/0/0/1/0/0/0/0/0/0/0/0
KK	Poa	0.2 ± 0.4	1/0/0/0/0/0/0/0/1/0/0/0/0
			Species totals: 393/237/331/398/447/468/362/453/535/474/464/439/407

*Root taxon node of family abbr. Poa = Poaceae, Mag = Magnoliophyta, Ory = Oryzeae
†Mean ± SD
‡Species counts: *L. perrieri* / *O. brachyantha* / *O. punctata* / *O. meridionalis* / *O. glumaepatula* / *O. barthii* / *O. glaberrima* / *O. nivara* / *O. sativa* vg. *indica* [93-11]/ *O. rufipogon* / *O. sativa* vg. japonica / *O. sativa* vg. indica [IR 8] / *O. sativa* vg. aus [N 22]

Supplementary Table 42. Domain structures and named disease-resistance genes associated with NLR gene families in the *Oryzeae*.

Family	LRR?	N-terminus*	C-terminus*	Named genes†
A	+	CC, RPW8, UPF0261	.	Pit, NBSt1
B	+	CC, UPF0261, B3	Jacalin, Thioredoxin	NLS1
D	+	CC, RPW8, UPF0261	Pkinase, Thioredoxin, WRKY, PP2C, AvrRpt	RGA5, Pi-ta
C	+	CC, zf-BED	Pkinase	Xa1
H	+	CC, RPW8, UPF0261	.	Pid3
F	+	CC	.	Pi9, Pib
G	+	CC, RPW8, B3	WRKY	.
I	+	CC, UPF0261	.	.
J	+	CC, UPF0261	.	Pikm2-TS
K	-	CC, RPW8	.	.
M	+	CC, RPW8	.	.
O	+	CC, UPF0261	Jacalin	.
E	+	CC	.	Pb1
L	+	CC, RPW8	TAXI, AvrRpt	Pi5-1, Pi5-2
N	+	CC, RPW8, B3, ALIX	Pkinase	.
P	+	CC	Pkinase	.
R	+	CC, RPW8, UPF0261	.	Pi36
Q	+	CC, RPW8	Pkinase	.
U	+	CC	.	.
T	+	CC, Myb	.	.
S	+	CC, UPF0261	.	Pi37
V	-	CC	.	.
W	+	CC, UPF0261	.	RGA4
X	+	CC	.	.
Y	+	CC	.	Pikm1-TS
Z	-	TIR	.	.
AA	+	CC	.	pi54
CC	-	CC	.	.
DD	-	CC	.	RLS1
BB	+	CC	.	.
FF	-	.	.	.
HH	+	CC	.	.
GG	+	CC	.	.
II	-	CC	.	.
JJ	-	.	.	.
KK	-	.	.	.

* Pfam domain names where abbreviated: AvrRpt = AvrRpt-cleavage; Myb = Myb_DNA_bind_4; ALIX = ALIX_LYPXL_bnd;
† Citations in Online Methods

Supplementary Table 43. NLR disease resistance gene counts by chromosome in13 *Oryzeae* species.

Chromosome	*L. perrieri*	*O. brachyantha*	*O. punctata*	*O. meridionalis*	*O. glumaepatula*	*O. barthii*	*O. glaberrima*	*O. nivara*	*O. sativa* vg. indica [93-11]	*O. rufipogon*	*O. sativa* vg. *japonica*	*O. sativa* vg. indica [IR 8]	*O. sativa* vg. aus [N 22]
1	33	22	26	40	44	41	25	50	58	42	44	38	42
2	20	11	21	23	38	30	26	33	40	30	32	27	28
3	15	12	10	14	20	17	14	24	27	17	18	16	24
4	23	11	20	28	25	28	26	22	39	30	25	28	30
5	17	10	10	20	20	18	10	26	39	16	16	16	18
6	27	13	23	33	28	31	29	34	43	36	31	24	24
7	30	16	26	29	29	22	23	28	33	29	26	35	30
8	36	21	26	39	47	38	28	45	50	49	45	46	42
9	17	14	19	14	21	26	19	19	24	20	24	27	25
10	28	16	24	24	25	35	26	26	29	35	28	25	22
11	117	71	86	99	98	123	101	104	103	121	127	114	83
12	30	20	40	35	52	59	35	42	50	49	48	43	39
Total:	393	237	331	398	447	468	362	453	535	474	464	439	407

Supplementary Table 44. Paired arrangements of NB-ARC containing genes in 13 *Oryzeae* species.

Species	Tandem pairs (count)	Heterogeneous pairs (%)*	Gene counts in pairs (%)†
O. sativa vg. japonica	105	25.7	180 (38.8)
O. sativa vg. indica [93-11]	130	23.1	221 (41.3)
O. sativa vg. *indica* [IR 8]	87	18.4	150 (34.2)
O. sativa vg. *aus* [N 22]	93	30.1	161 (39.6)
O. nivara	98	23.5	173 (38.2)
O. rufipogon	112	27.7	191 (40.3)
O. glaberrima	114	26.3	194 (53.6)
O. barthii	137	27.0	224 (47.9)
O. glumaepatula	117	22.2	199 (44.6)
O. meridionalis	105	23.8	175 (44.0)
O. punctata	92	26.1	156 (47.1)
O. brachyantha	61	29.5	112 (47.3)
L. perrieri	135	29.6	223 (56.7)

*Percent of adjacent pairs composed of genes from different gene families.
†Percent of total NB-ARC domain genes in a species that form adjacent pairs. Note that a single gene was counted in two pairs if flanked on both sides by an NB-ARC gene, which explains why the number of genes is not equal to double the number of gene pairs.

Supplementary Table 45. Adjacent pairs of NB-ARC genes in 13 *Oryzeae* species.

Family origin	Type	Counts (fraction)*	Pairs w/conserved synteny (%)	Pairs w/ unusual C-terminal domains count (%)†	Pairs w/ unusual N-terminal domains count (%)††
homogeneous	h2h	177 (0.17)	151 (85.3)	11 (6.2)	28 (15.8)
homogeneous	h2t	762 (0.74)	506 (66.4)	32 (4.2)	49 (6.4)
homogeneous	t2t	92 (0.09)	56 (60.9)	0 (0)	4 (4.3)
heterogeneous	h2h	165 (0.47)	120 (72.7)	50 (30.3)	8 (4.8)
heterogeneous	h2t	126 (0.35)	72 (57.1)	14 (11.1)	7 (5.6)
heterogeneous	t2t	64 (0.18)	40 (62.5)	9 (14.1)	5 (7.8)

*Fraction of total homogeneous (n=1031) or heterogeneous (n=355) pairs with indicated configuration type.
†Data includes 108 genes with unusual C-terminal domains, which participate in 116 adjacent pair formations.
††Data includes 96 genes with unusual N-terminal domains, which participate in 101 adjacent pair formations.

Supplementary Table 46. Family composition of head-to-head heterogeneous NB-ARC gene pairs in 11 *Oryzeae* species.

Species	D:W	J:Y	B:W	A:J	B:I	Other	Totals
O. sativa vg. japonica	7	3	1	1	1	2	15
O. sativa vg. indica [93-11]	5	5	2	1	1	0	14
O. sativa vg. indica [IR 8]	8	2	2	0	1	1	14
O. sativa vg. aus [N 22]	11	3	1	0	1	2	18
O. nivara	2	2	2	0	0	3	9
O. rufipogon	4	4	1	1	1	2	13
O. glaberrima	4	1	2	2	1	1	11
O. barthii	7	2	2	1	0	4	16
O. glumaepatula	4	3	1	1	0	5	14
O. meridionalis	5	3	1	1	1	0	11
O. punctata	3	3	3	1	1	3	14
O. brachyantha	5	3	0	0	1	3	12
L. perrieri	2	3	2	1	0	2	10

Supplementary Note Table 1. BioProject and cultivar information.

Species	BioProject	Cultivar/Accession
O. sativa vg. japonica	N/A	Nipponbare
O. sativa vg. indica	PRJNA353946	IR 8
O. sativa vg. aus	PRJNA315689	N 22
O. rufipogon	PRJEB4137	W1943
O. nivara	PRJNA48107	IRGC100897
O. barthii	PRJNA30379	IRGC105608
O. glaberrima	PRJNA13765	IRGC96717
O. glumaepatula	PRJNA48429	GEN1233_2
O. meridionalis	PRJNA48433	W2112
O. punctata	PRJNA13770	IRGC105690
O. brachyantha	PRJNA70533	IRGC101232
Leersia perrieri	PRJNA163065	IRGC105164

Supplementary Note Table 2. NCBI SRA accessions for whole genome shotgun sequence reads in six species.

Species	Samples	Accession
O. nivara	5	SRX663049-SRX663053
O. rufipogon	1	ERX096841
O. barthii	9	SRX662937-SRX662945
O. glumaepatula	4	SRX663040-SRX663043
O. meridionalis	5	SRX663044-SRX663048
O. punctata	27	SRX662909-SRX662935
Leersia perrieri	1	SRX663039

Supplementary Note Table 3. GenBank and INSDC numbers of reference genomes used.

Species	INSDC	WGS accession
O. sativa vg. indica [IR 8]	MPPV00000000.1	GCA_001889745.1
O. sativa vg. aus [N 22]	LWDA00000000.1	GCA_001952365.1
O. nivara	AWHD00000000.1	GCA_000576065.1
O. rufipogon	CBQP000000000.1	GCA_000817225.1
O. barthii	ABRL00000000.2	GCA_000182155.2
O. glaberrima	ADWL00000000.1	GCA_000147395.2
O. glumaepatula	ALNU00000000.2	GCA_000576495.1
O. meridionalis	ALNW00000000.2	GCA_000338895.2
O. punctata	AVCL00000000.1	GCA_000573905.1
O. brachyantha	AGAT00000000.1	GCA_000231095.2
L. perrieri	ALNV00000000.2	GCA_000325765.3

Supplementary Note Table 4. NCBI SRA accessions for RNA-seq reads collected from three tissues in 10 species.

Species	Leaf	Panicle	Root
O. sativa vg. japonica	SRX477950	SRX477951	SRX477952
O. rufipogon	SRX512340	SRX512341	SRX512342
O. nivara	SRX472708	SRX472710	SRX472709
O. barthii	SRX471823	SRX472434	SRX472435
O. glaberrima	SRX474528	SRX474529	SRX474530
O. glumaepatula	SRX475002	SRX475003	SRX475004
O. meridionalis	SRX475006	SRX475007	SRX475008
O. punctata	SRX472098	SRX472099	SRX472100
O. brachyantha	ND	SRX475011	ND
Leersia perrieri	SRX472913	SRX472914	SRX472915

NA = Not determined