Nanopore sequence-based genome assembly of the basmati rice

Jae Young Choi[1], Simon C. Groen[1], Sophie Zaaijer[2], and Michael D. Purugganan[1,3*]

[1]Center for Genomics and Systems Biology, Department of Biology, New York University, New York, New York, USA

[2]New York Genome Center, New York, New York, USA

[3]Center for Genomics and Systems Biology, NYU Abu Dhabi Research Institute, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

* Corresponding author, Email: mp132@nyu.edu (MDP)

Keywords: *Oryza sativa*, Asian rice, aromatic rice group, domestication, crop evolution, nanopore sequencing, aus, basmati, indica, japonica, admixture, awnless

ABSTRACT

Basmati rice is an iconic cultivated Asian rice (*Oryza sativa*) variety group that is widespread in the Indian subcontinent. Despite its economic and cultural importance, a high-quality reference genome is currently lacking, and the group's origins have not been fully resolved. To address these gaps we obtained 48x read coverage of a basmati genome using Oxford Nanopore Technologies' long-read sequencing platform. We generated a 373.8-Mbp assembly, consisting of 2,245 contigs with an N50 contig length of ~444.4 kbp. We polished this assembly with an additional 29× read coverage of Illumina short-reads, resulting in a high-quality genome with 98.7% gene completion. Evolutionary genomic analysis with other Asian rice genome assemblies indicated basmati rice was most closely related to the japonica rice group. But the basmati rice origins were complicated as conflicting phylogenetic signals were observed among basmati rice, japonica rice, and the progenitor wild rice *O. rufipogon*, caused by admixture between japonica rice and *O. rufipogon.* Admixture was also detected between the basmati and aus rice groups, consistent with previous a hypothesis suggesting a hybrid origin of the basmati rice group from japonica and aus rices. In addition, we detected basmati rice-specific domestication mutations, one of which was the deletion of a gene that regulates awn length, a deletion found only in the basmati rice group. Single molecule-based long-read sequencing is quickly becoming essential for the *de novo* generation of high-quality plant genome assemblies, and we show that this technology will be particularly useful for studies of crop evolution and genetics.

SIGNIFICANCE STATEMENT

Single molecule-based long-read sequencing is a powerful method for assembling complex plant genomes. Here, using nanopore sequencing we have assembled the genome of the iconic basmati rice (*Oryza sativa*), resulting in an assembly of 373.8 Mbp in 2,245 contigs, with an N50 contig size of 444.4 kbp. Evolutionary genomic analysis of the genome suggests basmati rice has a hybrid origin from the japonica and aus rice groups.

INTRODUCTION

Oryza sativa or Asian rice is an agriculturally important crop that feeds one-third of the world's population (Gnanamanickam, 2009). Based on morphometric differences, *O. sativa* has historically been classified into two major variety groups/subspecies named japonica and indica (Matsuo *et al.*, 1997; Gross and Zhao, 2014). Archaeobotanical remains suggest japonica rice was domesticated ~9000 years ago in the Yangtze Basin of China, while indica rice originated ~4,000 years ago when domestication alleles were introduced from japonica into either *O. nivara* or a proto-indica in the Indian subcontinent (Fuller *et al.*, 2010). With molecular and genome-wide polymorphism data, the japonica and indica rices have since been subdivided into multiple subgroups, and two additional variety groups/subspecies have been identified that are genetically distinct from japonica and indica cultivars: the aus/*circum*-Aus and aromatic/*circum*-Basmati rices (Garris *et al.*, 2005; Wang *et al.*, 2018). Here, we refer to the latter two groups as aus and basmati rices, respectively.

The rich genetic diversity of Asian rice is likely a result from a complex domestication process involving multiple wild progenitor populations and the exchange of important domestication alleles between *O. sativa* subpopulations through gene flow (Meyer and Purugganan, 2013; Huang and Han, 2015; Castillo *et al.*, 2016; Choi *et al.*, 2017; Fuller, 2011; Fuller *et al.*, 2010; He *et al.*, 2010; Choi and Purugganan, 2018). Moreover, many agricultural traits within rice are subpopulation-specific (Sweeney *et al.*, 2006; Wang *et al.*, 1995; Kovach *et al.*, 2009; Xu *et al.*, 2006; Konishi *et al.*, 2006; Bin Rahman and Zhang, 2018), suggesting local adaptation to environments or cultural preferences have partially driven the diversification of rice varieties.

Arguably, the basmati rice group has been the least studied among the four major variety groups/subspecies. Among its members the group boasts the iconic basmati rices (*sensu stricto*) from southern Asia and the sadri rices from Iran (Garris *et al.*, 2005). Many basmati (*sensu stricto*) varieties are characterized by their distinct and highly desirable fragrance and texture (Singh *et al.*, 2000). Nearly all fragrant basmati varieties possess a loss-of-function mutation in the *BADH2* gene that has its origins in ancestral japonica haplotypes, suggesting an introgression between basmati and japonica rices may have led to fragrant basmati rice (Kovach *et al.*, 2009). Genome-wide polymorphism analysis of a smaller array of basmati rice cultivars shows a close association with japonica varieties (Huang *et al.*, 2012; Wang *et al.*, 2018) providing more evidence that the origins of basmati rice may indeed lay at least partially in the japonica variety group/subspecies.

Whole-genome sequences are important for evolutionary biologists studying plant domestication, and breeders aiming to improve crop production. Single-molecule sequencing regularly produces sequencing reads in the range of kilobases (kbp) (Heather and Chain, 2016). This is particularly helpful for assembling plant genomes, which are often highly repetitive, highly heterozygous, and which have often undergone at least one round of polyploidization in the past (Michael and VanBuren, 2015; Jiao and Schneeberger, 2017; Li *et al.*, 2017).

Currently for plants, members of the genus *Oryza* have seen the most progress in terms of generating *de novo* genome assemblies. There are genome assemblies for nine wild species (*Leersia perrieri*, *O. barthii*, *O. brachyantha*, *O. glumaepatula*, *O. longistaminata*, *O. meridionalis*, *O. nivara*, *O. punctata*, and *O. rufipogon*) and two domesticated species (*O. glaberrima* and *O. sativa*) (Zhang *et al.*, 2015; Chen *et al.*, 2013; International Rice Genome Sequencing Project, 2005; Wang *et al.*, 2014; Stein *et al.*, 2018; Yu *et al.*, 2002).

Within the domesticated Asian rice *O. sativa*, genome assemblies for several variety groups/subspecies and cultivars within these are also available (Zhao *et al.*, 2018; International Rice Genome Sequencing Project, 2005; Yu *et al.*, 2002; Zhang *et al.*, 2016; Du *et al.*, 2017; Sakai *et al.*, 2014; Schatz *et al.*, 2014), but for many of these *O. sativa* genomes, the assembly has been generated through short-read sequencing and the assembly status is often incomplete compared to those generated from long-read sequences (Zhang *et al.*, 2016; Du *et al.*, 2017). Nevertheless, these *de novo* genome assemblies have been critical in revealing genomic variations (*e.g.* structural variations, *de novo* species- or population-specific genes, and repetitive DNA) that were otherwise missed from analyzing a single reference genome. Recently, a genome assembly for basmati rice was reported, which was also generated using short-read sequencing data (Zhao *et al.*, 2018). The sequenced basmati cultivar was an elite breeding line, and such modern cultivars are not the best foundations for domestication-related analyses due to higher levels of introgression from other rice populations.

Here, we report the *de novo* sequencing and assembly of the landrace variety Basmati 334 (Vikram *et al.*, 2012; Jain *et al.*, 2004; Kovach *et al.*, 2009) using Oxford Nanopore Technologies' long-read sequencing platform (Jain *et al.*, 2016). Basmati 334 evolved in a rainfed lowland agro-environment in southern Asia and is known to be drought tolerant at the seedling and reproductive stages (GeneSys passport: https://www.genesys-pgr.org/10.18730/2G6Y2) (Vikram *et al.*, 2012). It also possesses several broad-spectrum bacterial blight resistance alleles (Ullah *et al.*, 2012; Chen *et al.*, 2008). These attributes make Basmati 334 desirable for breeding resilience into modern basmati cultivars (Sandhu *et al.*, 2017; Ullah *et al.*, 2012). Using our basmati rice genome assembly we conducted an evolutionary genomic analysis to investigate the origins of basmati rice. In addition, we were able to identify that the domestication trait of reduced awn length may have evolved at least partially independently in basmati rice as evidenced by the basmati-specific deletion of a key gene that regulates this trait. Our basmati rice genome assembly will be a valuable complement to the available rice cultivar genome assemblies, unlocking important genomic variation for rice crop improvement. It also demonstrates the utility of nanopore-based sequencing in a relatively rapid, low-cost approach to generating better genome assemblies using long-read sequence data.

RESULTS

Nanopore sequencing of a basmati rice reference genome. High-molecular-weight genomic DNA was extracted from young seedlings of the variety Basmati 334, and was sequenced on eleven nanopore flow-cells. Base calling was accomplished using the program Albacore ver. 2.1.10 from Oxford Nanopore Technologies. Albacore called 5,389,267 reads constituting a total of 33.2 Gbp of which 2,452,750 reads comprising 18.1 Gbp passed the default Albacore filters, selecting for high quality reads (Table 1). For each flow-cell, the reads that passed the filter had median lengths of ~3.5 kbp to ~11.9 kbp and read length N50s of ~4.6 kbp to ~22.3 kbp. The median quality Q-score for the sequence reads was ~6.8 before and ~8 after filtering. The longest read from each library ranged from ~28 kbp up to ~82 kbp and the quality scores for each long sequence read were comparable to the average quality score seen across all sequenced reads.

De novo assembly of a basmati rice reference genome. Only the reads that passed the Albacore base caller filter were used for downstream genome assembly. For the Basmati 334 genome assembly the 2,452,750 reads that passed the Albacore filter were expected to result in ~48× genome sequence coverage, given that *O. sativa* has average genome sizes in the range of 370-380 Mbp (Stein *et al.*, 2018). We used the genome assembly program Canu ver. 1.6 (Koren *et al.*, 2017) as it was the state-of-the-art genome assembler that supports nanopore sequencing data, and was used in multiple nanopore sequencing-based *de novo* genome assembly projects (Jain *et al.*, 2018; Schmidt *et al.*, 2017; Tyson *et al.*, 2018). Canu assembled the basmati rice genome with an assembly size of 373.8 Mbp on 2,245 contigs, and a contig N50 of 444,364 bp (Supplemental Table 1). We note this is a large improvement from the currently available basmati rice genome assembly of GP295-1 (Zhao *et al.*, 2018), which was generated from 173× sequence coverage Illumina short-read sequencing data and has a contig N50 of 44,407 bp with 50,786 assembled contigs.

The quality of the initial pre-draft genome assembly from Canu was assessed using BUSCO (Simão *et al.*, 2015) to identify and count the number of highly conserved orthologous plant genes. The Canu assembly had a BUSCO estimated genic completeness of only 42.1% (Supplemental Table 1). Raw nanopore sequence-based assembly usually requires a polishing step necessary for fixing the base errors of the genome assembly (Schmidt *et al.*, 2017; Michael *et al.*, 2018), and we implemented two approaches for error correction: (i) using the nanopore sequencing reads itself for polishing, and (ii) using highly accurate short-read sequences from Illumina sequencing.

There are two methods that use the nanopore sequence data for polishing. Nanopolish ver. 0.8.5 (Loman *et al.*, 2015) uses pre-base-call signal-intensity data of each sequencing read for improving the pre-draft genome assembly. A single round of polishing through Nanopolish greatly improved the BUSCO estimated gene completeness to 79.8% (Supplemental Table 1). However, the execution time for Nanopolish was very slow, taking ~5 weeks on a 16-core machine with 128 GB memory and 2.0 GHz CPU. Racon ver. 1.2.1 (Vaser *et al.*, 2017) also uses the nanopore sequence data for polishing but only uses the base-called reads. Racon was on the order of a thousand-fold faster than Nanopolish, but a round of polishing resulted in a gene completion of only 67.1%. However, iteratively increasing the polishing with Racon did increase the BUSCO estimated gene completion to 71.5% (Supplemental Table 1).

Among the massively parallel sequencing technologies, Illumina reads have the highest base-calling accuracy, and Pilon ver. 1.22 (Walker *et al.*, 2014) uses short-read sequencing data for error-correcting pre-draft genome assemblies. The Basmati 334 sample was sequenced on an Illumina platform to a depth of ~29× sequence read coverage, and a single round of Pilon analysis greatly enhanced the BUSCO estimated gene completion to 93%. Iteratively increasing the polishing with Pilon further increased the BUSCO estimated gene completion to 98.2% (Supplemental Table 1). In the end, we chose a strategy that maximizes the speed and accuracy of polishing the pre-draft genome by implementing seven rounds of Racon polishing followed by seven rounds of Pilon polishing (Supplemental Table 1).

Comparing the basmati rice genome assembly to other rice genomes assembled de novo. The final nanopore sequence-based draft assembly for the basmati rice genome had a high quality (Table 2): more then 90% of the contigs were greater than 50 kbp in length with a contig N50 was 455 kbp. The BUSCO gene completion level was 98.7%, comparable to other rice genomes assembled *de novo* (Supplemental Table 2). When the ~29× sequence read coverage Illumina reads from Basmati 334 were aligned to the draft genome, the vast majority (99.6%) of the reads were alignable to the draft genome. The draft genome was also highly syntenic to the Nipponbare japonica and R498 indica genomes, where an ancient chromosomal duplications and translocations (Wang *et al.*, 2005) were visible in synteny dot plots (Supplemental Figure 1).

We then aligned available *de novo*-assembled genomes within the Asian rice complex (see Materials and Method for a complete list of genomes) to the Nipponbare genome, creating a multi-genome alignment to compare and contrast the assembly quality of the draft basmati genome. The Nipponbare cultivar was chosen as the reference genome to be aligned against, as its assembly and gene annotation is a product of years of community-based efforts (International Rice Genome Sequencing

Project, 2005; Kawahara *et al.*, 2013; Sakai *et al.*, 2013). The Basmati 334 genome could be aligned to 80% of the repeat soft-masked Nipponbare genome, which among the Asian rice complex was the second highest value (Supplemental Table 3). Compared to the GP295-1 basmati rice reference genome, over 14.9 Mbp more sequence from our Basmati 334 assembly aligned to the Nipponbare genome, suggesting our Basmati 334 reference genome assembly was more complete with fewer missing regions.

To infer the quality of the genic regions in each of the genome assemblies, we used the multi-genome alignment to extract the coding DNA sequence of each Nipponbare gene and its orthologous regions from each non-Nipponbare genome. The orthologous genes were counted for missing DNA sequences (“N” sequences) or gaps to estimate the percent of Nipponbare gene covered. For all genomes the majority of Nipponbare genes had a near-zero proportion of sites that were missing in the orthologous non-Nipponbare genes (Supplemental Figure 2). There were also Nipponbare genes with close to 100 percent of sites missing in the non-Nipponbare genome. These are likely to be the dispensable genes where its presence/absence status is polymorphic between subspecies (Zhao *et al.*, 2018). The missing proportion of Nipponbare-orthologous genes for the Basmati 334 genome was comparable to those genomes that had higher assembly contiguity (Du *et al.*, 2017; Zhang *et al.*, 2016; Stein *et al.*, 2018). Furthermore, compared to the GP295-1 genome, the Basmati 334 genome had a noticeably lower proportion of sequences that were missing. Thus, we argue that the draft Basmati 334 genome assembly is of high quality and contiguity, and is complete for downstream analysis.

Since Basmati 334 had previously been identified to be a non-fragrant rice and to harbor the wild type allele of the *BADH2* fragrance gene (Kovach *et al.*, 2009), we examined the Basmati 334 genome assembly to check if this was indeed the case. There are multiple different loss-of-function mutations in *BADH2* that cause rice varieties to be fragrant, and the majority of fragrant rices carry a deletion of 8 nucleotides at position chr8: 20,382,861-20,382,868 of the japonica (Nipponbare IRGSP1.0) genome assembly. Using the genome alignment, the *BADH2* sequence region was extracted to compare the gene sequence of the non-fragrant Nipponbare to Basmati 334. Our results showed that the Basmati 334 genome assembly did not carry the deletion, and the non-deleted, 8-bp sequence was identical to that found in the japonica sequence. The *BADH2* protein-coding sequence region was in fact identical between Basmati 334 and Nipponbare, indicating Basmati 334 harbored the expected wild type *BADH2* allele (Kovach *et al.*, 2009). This further suggests the Basmati 334 genome is highly accurate for gene-level analysis.

***Gene annotation of the basmati rice genome assembly*.** Augustus ver. 3.2.3 (Stanke and Waack, 2003) was used to predict 49,943 genes, of which 30,786 genes were not a possible transposable element and had a Pfam-, PANTHER-, or SUPERFAMILY-annotated protein domain (Table 2). The median protein length of the gene annotation was 303 amino acids (aa), which was slightly larger than the median length of the *ab initio* and evidence-based gene annotations of the Nipponbare (267 aa) and R498 (254 aa) genomes (Du *et al.*, 2017), which are arguably the two highest quality genome assemblies for the japonica and indica variety groups/subspecies, respectively. The Basmati 334 gene models were classified into orthogroups, which are sets of genes that are orthologs and recent paralogs to each other, using the gene models from aus (N22), basmati (GP295-1), japonica (Nipponbare), and indica (R498) genomes. The majority of the annotated genes from Basmati 334 were assigned as a member of an orthogroup with a gene annotated from another published rice genome (Figure 1), suggesting the genes predicted *ab initio* for Basmati 334 are likely to be biologically functional.

The number of orthogroups, however, were smallest for Basmati 334. This could be due to the *ab initio* gene prediction algorithm missing several gene annotations, or to an incomplete genome assembly for Basmati 334 that could miss some genes annotated in other genomes. To determine the cause of this anomaly, we focused on the 1,737 orthogroups that had a gene member in every rice variety group/subspecies except Basmati 334 (Figure 1). From those orthogroups, the region orthologous to the japonica (Nipponbare) gene was extracted from the Basmati 334 genome assembly in our multi-genome alignment. For all of these japonica (Nipponbare) genes, the majority of the orthologous Basmati 334 regions had a missing or gapped proportion that was close to zero (Supplemental Figure 3). This suggested the genic sequences had not been completely annotated by Augustus and the genome assembly of Basmati 334 contained those coding DNA sequences.

***Repetitive DNA and LTR transposable element content in the basmati reference genome*.** Repetitive DNA comprises 44.2% of the Basmati 334 genome assembly (Table 2), and consistent with many other plant species (Kumar and Bennetzen, 1999) the repetitive DNA was largely composed of Class I retrotransposons followed by Class II DNA transposons (Figure 2A). Within the retrotransposon class, long terminal repeat (LTR) retrotransposons are most commonly found in plants (Kumar and Bennetzen, 1999). Due to its mechanism of proliferation, a newly inserted LTR retrotransposon will have identical LTR sequences, and in time the accumulation of mutations causes the LTRs to diverge in sequence. Thus, the DNA divergence of LTR sequences can be used to approximate the insertion time for an LTR retrotransposon, and infer the evolutionary history of LTR retrotransposon dynamics within the host genome. For Basmati 334 the majority of retrotransposons were LTR retrotransposons from the *gypsy* and *copia* families, and retrotransposons of both families had a median insertion time of ~3.4 million years ago (MYA) (Figure 2B).

Since the LTR retrotransposon insertion times will be affected by sequencing and assembly errors in the LTR sequences, we examined if the polishing steps were able to fix potential errors in the repetitive

DNA regions. The insertion time estimates for the LTR retrotransposons decreased with each polishing step added (Supplemental Figure 4), suggesting genome polishing not only fixes errors in the coding sequence regions but also in repetitive DNA regions in general. Comparing the LTR retrotransposons and their insertion times in other Asian rice genome assemblies, the aromatic rice group had the oldest insertion times (Figure 2C). This is concordant with results from the basmati rice genome assembled from Illumina sequencing data (GP295-1), suggesting the older insertion time estimates for basmati rice is not likely to be an artifact of assembly errors. Noticeably, reference genomes assembled from short-read sequencing data (GP295-1, DJ123, Kasalath, and IR64) had fewer annotated LTR retrotransposons, suggesting these genome assemblies may be missing certain repetitive DNA regions.

Phylogenomic analysis on the origins of basmati rice. The multi-genome alignment was used to estimate the phylogenetic relationships within and between variety groups/subspecies of domesticated Asian rice. Four-fold degenerate sites from the genes of the Nipponbare genome were extracted and used to build a maximum-likelihood phylogenetic tree (Figure 3A). The tree showed that each cultivar was monophyletic with respect to its variety group/subspecies of origin. In addition, the basmati rice group was sister group to japonica rice, while the aus rice group was sister to indica rice. Consistent with previous observations, the wild rices *O. nivara* and *O. rufipogon* were sister groups to the aus and japonica rices, respectively (Choi *et al.*, 2017), suggesting each domesticated rice variety group/subspecies may have had independent wild progenitors of origin. On the other hand, it should be noted that recent hybridization between wild and domesticated rice can lead to a similar phylogenetic relationship (Wang *et al.*, 2017).

To further investigate the phylogenetic relationships between the basmati (represented by varieties Basmati 334 and GP295-1) and japonica rices, we examined the phylogenetic topologies of each gene involving the trio basmati, japonica, and wild rice *O. rufipogon*. For each gene we tested which of the three possible topologies for a rooted three-species tree [i.e. “((P1,P2),P3),O” where O is outgroup *O. barthii* and P1, P2, and P3 can be any of the combination involving the trio basmati, japonica, and *O. rufipogon*] were found in highest proportion. There were 7,495 genes that significantly rejected one topology over the other two, and the majority of those genes (51.3%) supported a topology grouping basmati and japonica rice as sister to each other (Figure 3B). A smaller proportion of genes (39.7%) supported the topology generated from the genome-wide four-fold degenerate sites, which grouped japonica and *O. rufipogon* as the sister group. The two conflicting topologies can occur if one of the topologies represents the true species topology [i.e. “((P1,P2),P3),O”], while the other is a product of recent admixture that leads to the phylogenetic grouping of the distant species P3 with the species P1 or P2.

To determine which of the two topologies represented the tree from the initial lineage splitting event versus the tree from an introgression event, we estimated the divergence times of the two internal nodes (T_1 and T_2) for each topology (Figure 3C). Because introgression reduces the time to the most recent common ancestor, the divergence time estimates from the introgression topology would be lower than the divergence time estimates from the lineage-splitting topology (Fontaine *et al.*, 2015). Results showed that the divergence time estimates from the topology grouping basmati and japonica as sister to each other were older than the topology grouping japonica and *O. rufipogon* (Figure 3C). Thus, basmati has origins that are closely related to japonica, whereas the admixture between japonica and *O. rufipogon* created topologies grouping the two together.

Detecting admixture between basmati, and aus or indica rices. Due to the extensive admixture occurring between rice variety groups/subspecies (Choi *et al.*, 2017) we examined whether the basmati genome was influenced by gene flow with aus or indica rices. For each gene, the trio aus, indica, and basmati were tested to see which of the three possible topologies fit significantly for that gene. There were 7,752 genes that significantly rejected one topology over the other two, and the majority (53.4%) supported the genome-wide topology, grouping aus and indica as sister groups (Figure 3D). There was a larger proportion of genes that supported the topology that groups aus and basmati as sisters (26.5%), than the topology that groups indica and basmati as sisters (20.1%) suggesting the aus variety group may have contributed a larger proportion of genes into basmati than the indica variety group through gene flow.

The pattern observed could have arisen from gene flow, or may just be the product of incomplete lineage sorting. To test for introgression, we employed the D-statistics from the ABBA-BABA test (Green *et al.*, 2010; Durand *et al.*, 2011). Initially, we employed the D-statistics to detect evidence of introgression between basmati rice, and the aus or indica rice groups. The D-statistics were calculated from groups involving basmati and aus/indica Asian rices, as well as wild rice *O. nivara*/*O. rufipogon* (Table 3, row 1 to row 4). D-statistics were significant for introgression between aus and basmati rice (Table 3 row 1 and row 3), but for indica rice the D-statistics were only significant when calculated from groups involving *O. rufipogon*, basmati, and indica (Table 3 row 2).

We then calculated the D-statistics for groups involving japonica, basmati, and aus/indica rices. Results showed all D-statistics were significant and detected evidence of introgression between basmati and both aus and indica rices (Table 3 row 5 and row 6). Given the significant evidence of introgression between japonica and both aus and indica rices (Choi *et al.*, 2017), our results suggest the gene flow between the basmati and aus or indica variety groups was greater than the gene flow between the japonica and aus or variety groups. To determine whether aus or indica rice shared more alleles with basmati rice, we calculated the D-statistic from a group involving indica, aus, and basmati rices. The D-statistic was not significant (Table 3 row 7). However, consistent with the topology test result (Figure 3D), the D-statistic was positive, suggesting a higher proportion of aus alleles than indica alleles segregated within the basmati genome.

Domestication of basmati rice involved variety group-specific gene deletions**.** Gene deletions can have drastic phenotypic consequences and several domestication phenotypes in rice are caused by gene deletions (Wang *et al.*, 2014; Shomura *et al.*, 2008; Zhou *et al.*, 2009). We identified basmati rice-specific gene deletions by comparing the gene annotations of the Basmati 334 reference genome to the gene models from the aus (N22), indica (R498), and japonica (Nipponbare) references. In addition, we confirmed the gene deletions by comparing the genome alignment of the Basmati 334 genome and the japonica (Nipponbare) rice genome (see Materials and Methods for detail). There were 186 orthogroups where the aus, indica, and japonica rices all had at least one gene member to the orthogroup, but where no gene was annotated in Basmati 334 rice. This was not due to a mis-annotation of the basmati genome, since the genome alignment of the japonica and basmati rice genomes was used to verify that the gene was indeed missing in the basmati genome sequence.

Out of the 186 orthogroups with basmati-specific gene deletions, 91 orthogroups had single-copy gene members in each of the aus, indica, and japonica rice genomes (see Supplemental Table 4 for the japonica RAP-DB gene ID numbers and functions), suggesting the deletion of these genes may have strong phenotypic consequences in basmati rice. One of these involved the deletion of a gene with japonica RAP-DB (Sakai *et al.*, 2013) identifier Os03g0418600 and gene symbol *Awn3-1*. *Awn3-1* was identified in a previous study where a change in expression level was associated with altered awn length in japonica rice (Li *et al.*, 2016). Reduced awn length was an important domestication trait that was selected for ease of harvesting and storing rice seeds (Hua *et al.*, 2015).

The japonica (Nipponbare) and aus (N22) genomes were used to investigate the synteny around the deleted *Awn3-1* region of the basmati rice genome. The two genes that were up- and downstream of *Awn3-1* were syntenic between our basmati rice genome and both the japonica and aus rice genome assemblies (Figure 4). However, two orthologous basmati genes were found at the edges of two different contigs (tig00005191 and tig00002596), while the *Awn3-1* gene was deleted in the Basmati 334 genome.

Surrounding the *Awn3-1* gene deletion region, both contigs from the Basmati 334 genome assembly showed an increase in repetitive DNA sequences. These increased repetitive sequences may have caused difficulty in assembling the region and it is possible the *Awn3-1* gene may not have been properly assembled. To address this possibility we aligned the Basmati 334 nanopore sequencing reads to both the japonica and aus genomes, and investigated if there were any reads that could be aligned to the *Awn3-1* gene. Using the long-read-specific read alignment software ngmlr ver. 0.2.6 (Sedlazeck, Rescheneder, *et al.*, 2018) and long-read-specific structural variation detection program sniffles ver. 1.0.7 (Sedlazeck, Rescheneder, *et al.*, 2018), the *Awn3-1* gene was found to be deleted in Basmati 334 and no significant number of reads aligned to the *Awn3-1* gene region in either the japonica or aus genomes (Supplemental Figure 5).

We then determined the frequency of this deletion by examining the polymorphisms surrounding the *Awn3-1* region, using the 3,010 rice genomes sequencing project (Wang *et al.*, 2018). The majority of the basmati rice group did not have any variant calls for the *Awn3-1* gene, suggesting that no sequencing reads existed for making genotype calls and that hence the gene was deleted in most of the basmati rice group (Supplemental Figure 6). Thus, the *Awn3-1* deletion found in Basmati 334 is a basmati rice-specific gene deletion, and is found in high frequency specifically within the basmati rice group.

DISCUSSION

In this study, using long-read sequences generated with Oxford Nanopore Technologies' sequencing platform we assembled the draft genome for a basmati rice cultivar of *O. sativa*. This is the third plant genome, after *Arabidopsis thaliana* (Michael *et al.*, 2018) and *Solanum pennellii* (Schmidt *et al.*, 2017), and the first monocot plant that was assembled using nanopore sequencing technology. Our results suggest nanopore sequencing is a highly effective platform for generating a genome assembly. With modest genome coverage we were able to assemble a genome that was a significant improvement over the currently available basmati genome, which was generated from over three times our sequencing coverage but with short-read sequencing data (Zhao *et al.*, 2018). With additional short-read sequencing data we were able to correct errors in the draft genome assembly, resulting in a basmati genome that was highly complete and accurate. Although the assembly is fragmented into multiple contigs, plant genomes still require additional technologies such as optical mapping or Hi-C sequencing for improving the contiguity of the assembly (Goodwin *et al.*, 2016; Howe and Wood, 2015; Sedlazeck, Lee, *et al.*, 2018; Udall and Dawe, 2018). Regardless, with this genome assembly we were able to make unique inferences on the origin and evolution of the basmati rice group.

Due to the sequencing methodology, single molecule-based long-read sequencing requires high-quality DNA in abundant amounts, which can be an obstacle for plant materials. We found that commercially available DNA extraction kits coupled with a downstream DNA fragment size selection protocol, can be a rapid and reliable method for generating long-read sequencing libraries. Nanopore sequencing using this DNA library preparation resulted in an average read length N50 of ~11 kbp, and we were able to sequence high-quality, long reads of up to ~82 kbp. In addition, compared to using only short-read sequencing data, ~48× sequencing read coverage obtained with nanopore sequencing was enough for generating a more contiguous genome assembly. However, the noisy error-prone sequencing reads were problematic for the genome assembly, as assemblies based solely on nanopore sequencing are often full of base errors that require subsequent error-correcting steps to be resolved (Schmidt *et al.*, 2017; Michael *et al.*, 2018). Several software packages are currently under development to improve the accuracy of base calling from raw nanopore sequencing signal-intensity data (Teng *et al.*, 2018; Stoiber and Brown, 2017; David *et al.*, 2017; Boža *et al.*, 2016). These algorithms will be important for generating higher quality sequencing reads for genome assembly. For now, high-quality short-read sequencing data, such as can be generated with Illumina's sequencing methodology, is necessary for decreasing the error rates in the genome assembly (Michael *et al.*, 2018; Schmidt *et al.*, 2017). However, our study shows that the short-read sequencing data does not need to be obtained at high coverage for the error correction to be effective, and with only ~29× Illumina sequencing read coverage we were able to obtain a polished genome with gene accuracies that were comparable to genomes generated from multiple sequencing platforms (International Rice Genome Sequencing Project, 2005; Zhang *et al.*, 2016; Du *et al.*, 2017; Stein *et al.*, 2018).

Repetitive DNA constitutes large proportions of the plant genome. Transposable elements dominate the repetitive DNA landscape and are the main determinants of genome size variation in plants (Wendel *et al.*, 2016). Effects of transposable elements can be dependent on the genomic context, as their presence can affect nearby gene expression, prompting selection to favor or remove the insertion (Lisch, 2012). Hence, it is important that genome assemblies should represent the repetitive DNA of the sequenced organism well. Our long-read sequencing-based assembly of basmati rice had repeat content similar to that of other rice genome assemblies. Further, compared to the genomes assembled from short-read sequencing data, the long-read sequencing-based genome assemblies of this study and others enabled annotation of a larger number of transposable elements. Thus, there is an advantage of using long-read sequences for genome assembly: since short-read sequences are often shorter than the repeat unit of a repetitive or transposable element, using longer reads will lead to a better assembly of those regions (Chaisson *et al.*, 2015).

Due to the lack of archaeobotanical data the origins of the basmati rice group remain elusive. Studies of the basmati rice group's origins have primarily focused on the genetic differences that exist between the basmati and other Asian rice groups (Garris *et al.*, 2005; Wang *et al.*, 2018). Recently, a study suggested basmati rice (called 'aromatic' in that study) was a product of hybridization between the aus and japonica rice variety groups (Civáň *et al.*, 2015). This was based on phylogenetic relationships across genomic regions with evidence of domestication-related selective sweeps, where these regions mostly grouped the basmati rice group with the japonica or aus rice groups. Our evolutionary analysis of the basmati rice genome assembly suggests the basmati rice group has origins that are closely related to the japonica rice group. In addition, with our basmati rice genome assembly we were able to infer that the genome assembly of the wild rice *O. rufipogon* had evidence of admixture with the japonica rice group. This admixture between wild and domesticated rice has been reported from a reanalysis of published polymorphism data (Wang *et al.*, 2017), and here we suggest it is likely to have also influenced the reference *O. rufipogon* genome sequence (Stein *et al.*, 2018). Researchers involved with rice domestication and evolution studies should be aware of the admixed history of the *O. rufipogon* reference genome, and be cautious of using it as the wild progenitor genome sequence.

Because basmati rice has its origins in both the japonica and aus rice groups, there might be basmati rice-specific genomic features that were specifically selected only within the basmati rice group. Detecting and studying the evolution of these subpopulation-specific features is important for understanding the domestication process of each rice variety group and how Asian rice in general was domesticated. The *Awn3-1* gene deletion was an example of such a feature where the deletion had occurred only within the basmati rice group. Compared to awned rice plants, awnless rice with the japonica allele had enhanced expression of this gene (Li *et al.*, 2016), suggesting a gene deletion of *Awn3-1* would necessitate another mechanism of shortening awns in the basmati rice group. The awnless phenotype is a quantitative trait where many genes are involved in controlling the development of the awn structure (Luo *et al.*, 2013; Gu *et al.*, 2015; Hua *et al.*, 2015; Xiong *et al.*, 1999). For one of these genes, *LABA1*, the haplotype carrying the domestication mutation is found in both the japonica and indica rice groups (Hua *et al.*, 2015; Choi and Purugganan, 2018) suggesting a single domestication origin for the awnless trait. However, here we suggest that since the deletion of the *Awn3-1* gene in basmati rice has the opposite effect of the enhanced expression of the gene in japonica rice, the evolutionary paths of domestication mutations for awn shortening cannot be the same in the basmati and japonica rice groups. Thus, there is an at least partially independent domestication history inferred from the genes that control the short-awn phenotype. Similar observations have been made for the non-shattering domestication trait, where the genes that control seed shattering each have a different selection history (Lin *et al.*, 2007; Li *et al.*, 2006; Konishi *et al.*, 2006). Increasing evidence suggests the domestication process of Asian rice is more complex (Wang *et al.*, 2018) than the single- *versus* multiple-origin dichotomy that has dominated the debate of its domestication history (Gross and Zhao, 2014). Here, analyzing genome assemblies to detect genome-level variations that were largely missed by previous studies would be crucial for our understanding of the complex domestication history of Asian rice.

In conclusion, our study shows that generating a high-quality genome assembly is feasible with modest amounts of resources and data. Using nanopore sequencing we were able to generate a contiguous genome assembly for a rice cultivar. After our polishing for fixing assembly errors, our reference genome sequence has the potential to be an important genomic resource for identifying single nucleotide variations and larger structural variations that are unique to the basmati rice group. Analyzing *de novo* genome assemblies for a larger sample of the Asian rice population will be important for uncovering and studying the hidden population genomic variations that were too complex to study with only short-read sequencing technology.

MATERIALS AND METHODS

Plant material. Basmati 334 (IRGC 27819) is a basmati landrace from Pakistan and was originally donated to the International Rice Research Institute (IRRI) by the Agricultural Research Council (ARC) in Karachi (donor accession ID: PAK. SR. NO. 39). Seeds from accession IRGC 27819 were obtained from the IRRI seed bank, surface-sterilized with bleach, and germinated in the dark on a wet paper towel for four days. Seedlings were transplanted individually in pots containing continuously wet soil in a greenhouse at New York University's Center for Genomics and Systems Biology and cultivated under a 12h day-12h night photoperiod at 30°C. Plants were kept in the dark in a growth cabinet under the same climatic conditions for four days prior to tissue harvesting. Continuous darkness induced chloroplast degradation, which diminishes the amount of chloroplast DNA that would otherwise end up in the DNA extracted from the leaves.

DNA extractions. Thirty-six 100-mg samples (3.6 g total) of leaf tissue from a total of 10 one-month-old plants were flash-frozen at harvest and stored at -80ºC. For extraction, samples were ground using mortar and pestle in liquid nitrogen. DNA was extracted using the Qiagen DNeasy Plant Mini Kit following the manufacturer's protocol. Yields ranged between 60ng/ul and 150ng/ul.

Library preparation and nanopore sequencing. Genomic DNA was visualized on an agarose gel to determine shearing by Covaris g-tube. DNA fragments were size-selected by BluePippin (Sage Science) or directly used for library preparation using Oxford Nanopore Technologies' standard ligation sequencing kits SQK-LSK108. FLO-MIN106 (R9.4) and FLO-MIN107 (R9.5) flowcells were used for sequencing on either GridION X5 or MinION platform.

Library preparation and Illumina sequencing. Extracted genomic DNA was prepared for short-read sequencing using the Illumina Nextera DNA Library Preparation Kit. Sequencing was done on the Illumina HiSeq 2500 – HighOutput Mode v3 with 2×100 bp read configuration, at the New York

University Genomics Core Facility. A total of 116,689,522 reads with 11,225,650,361 bps were generated.

Nanopore sequencing read analysis. After completion of sequencing, the raw signal intensity data was used for base calling using the program Albacore ver. 2.1.10 from Oxford Nanopore Technologies. Default parameter settings were used for base calling and for flagging reads as "pass" to filter out low-quality sequencing reads. The NanoPack package (De Coster *et al.*, 2018) was used to calculate the statistical summary of the base-called sequencing reads.

Genome assembly. Base-called reads that passed the Albacore filter were used by the program Porechop (https://github.com/rrwick/Porechop) to trim adapters at the end of reads. Reads with internal adapters were discarded. Genome assembly was conducted using the program Canu ver. 1.6 (Koren *et al.*, 2017) assuming a genome size of 380 Mbp for basmati rice. There were 48 cores and 400 GB of memory available for Canu during the assembly. We applied the parameters "overlapper=mhap utgReAlign=true" to speed up the assembly, as assembly without those parameters took over 6 weeks. The size of the resulting assembly was not so different from that of the genome assembly with those options (results not shown).

Genome assembly statistics were calculated using the bbmap stats.sh script from the BBTools suite (https://jgi.doe.gov/data-and-tools/bbtools/). Completeness of the genome assembly was evaluated using BUSCO ver. 2.0 (Simão *et al.*, 2015). Synteny between the Basmati 334 genome and the japonica (Nipponbare) and indica (R498) genomes were visualized using SynMap (Haug-Baltzell *et al.*, 2017) from the CoGe web platform (https://genomevolution.org).

Genome polishing. Three different methods were used to polish the genome assembly. Adapter-trimmed FASTQ reads from our nanopore sequencing were aligned to the unpolished draft genome assembly using bwa-mem ver. 0.7.15 (Li, 2013) with the "–x ont2d" option enabled to align nanopore reads. Nanopolish ver. 0.8.5 (Loman *et al.*, 2015) was used for polishing the genome in 50 kbp intervals.

The program minimap ver. 0.2-r124-dirty (Li, 2016) was also used to align the same adapter-trimmed FASTQ reads to the draft genome assembly, which generated a pairwise overlap file. The overlaps were used by Racon ver. 1.2.1 (Vaser *et al.*, 2017) to polish the draft genome assembly.

Reads from Illunima sequencing were used by bwa-mem to align to the draft genome assembly. The alignment file was then used by Pilon ver. 1.22 (Walker *et al.*, 2014) for polishing.

For racon and pilon we implemented an iterative approach for polishing the genome where the polished genome assembly was used again for further polishing steps. We implemented a total of 7 rounds of polishing.

Gene annotation and analysis. Genes were predicted with augustus ver. 3.2.3 (Stanke and Waack, 2003; Stanke *et al.*, 2006) using maize gene models as a training set. To differentiate genes from transposable elements we used InterProScan (Jones *et al.*, 2014) and annotated potential protein domains for each gene model. Gene models that had protein domains with the following terms: "transposable, transposase, transposon, retroelement, retroid, retrotransposon, retroviral, retrovirus, LTR" (Zdobnov *et al.*, 2005) were assumed to be a transposable element and removed. In addition to select a core set of biologically relevant gene models, only genes that had a protein domain annotated in the databases of Pfam (Finn *et al.*, 2016), PANTHER (Mi *et al.*, 2017), or SUPERFAMILY (Wilson *et al.*, 2009) were selected.

Orthology between the genes from different rice genomes were determined with Orthofinder ver. 1.1.9 (Emms and Kelly, 2015). Ortholog status were visualized with the UpSetR package (Conway *et al.*, 2017).

Repetitive DNA annotation. The repeat content of a genome assembly was determined using Repeatmasker ver. 4.0.5 (http://www.repeatmasker.org/RMDownload.html). We used the *Oryza*-specific repeat sequences that were identified from Choi *et al.* (2017) (DOI: 10.5061/dryad.7cr0q), who used Repeatmodeler ver. 1.0.8 (http://www.repeatmasker.org/RepeatModeler.html) to annotate repetitive elements across wild and domesticated *Oryza* genomes *de novo* (Stein *et al.*, 2018).

LTR retrotransposons were annotated using the program LTRharvest (Ellinghaus *et al.*, 2008) with parameters adapted from Copetti *et al.* (2015). LTR retrotransposons were classified into superfamilies (Wicker *et al.*, 2007) using the program RepeatClassifier from the RepeatModeler suite. Insertion times for the LTR retrotransposons were estimated using the DNA divergence between the pair of LTR sequences (SanMiguel *et al.*, 1998). The L-INS-I algorithm in the alignment program MAFFT ver. 7.154b (Katoh and Standley, 2013) was used. PAML ver. 4.8 (Yang, 2007) was used to estimate the DNA divergence between the LTR sequences with the Kimura-2-parameter base substitution Model (Kimura, 1980). DNA divergence was converted to divergence time (i.e. time since the insertion of a LTR retrotransposon) approximating a base substitution rate of 1.3×10^{-8} (Ma and Bennetzen, 2004), which is two times higher than the synonymous site substitution rate.

***Whole genome alignment of* Oryza *genomes assembled* de novo.** Several genomes from published studies that were assembled *de novo* were analyzed. These include domesticated Asian rice genomes from the japonica variety group cv. Nipponbare (International Rice Genome Sequencing Project, 2005); the indica variety group cvs. 93-11 (Yu *et al.*, 2002), IR8 (Stein *et al.*, 2018), IR64 (Schatz *et al.*, 2014), MH63 (Zhang *et al.*, 2016), R498 (Du *et al.*, 2017), and ZS97 (Zhang *et al.*, 2016); the aus variety group cvs. DJ123 (Schatz *et al.*, 2014), Kasalath (Sakai *et al.*, 2014), and N22 (Stein *et al.*, 2018); and the basmati variety group cv. GP295-1 (Zhao *et al.*, 2018). Three genomes from wild rice species were also analyzed; these were *O. barthii* (Wang *et al.*, 2014), *O. nivara* (Stein *et al.*, 2018), and *O. rufipogon* (Stein *et al.*, 2018).

Alignment of the genomes assembled *de novo* was conducted using the approach outlined in Choi *et al.*, (2017). Briefly, this involved using the japonica (Nipponbare) genome as the reference for aligning all other genome assemblies. Alignment between japonica and a query genome was conducted using LASTZ ver. 1.03.73 (Harris, 2007), and the alignment blocks were chained together using the UCSC Kent utilities (Kent *et al.*, 2003). For japonica genomic regions with multiple chains, the chain with the highest alignment score was chosen as the single-most orthologous region. This analyses only one of the multiple possibly paralogous regions between japonica and the query genome, but was not expected to affect the downstream phylogenomic analysis of determining the origin and evolution of basmati rice. All pairwise genome alignments between the japonica and query genomes were combined into a multi-genome alignment using MULTIZ (Blanchette *et al.*, 2004).

Phylogenomic analysis. The multi-genome alignment was used to reconstruct the phylogenetic relationships between the domesticated and wild rices. Four-fold degenerate sites based on the gene model of the reference japonica genome were extracted using the msa_view program from the phast package ver. 1.4 (Hubisz *et al.*, 2011). The four-fold degenerate sites were used by RAxML ver. 8.2.5 (Stamatakis, 2014) to build a maximum likelihood-based tree, using a general time-reversible DNA substitution model with gamma-distributed rate variation.

To investigate the genome-wide landscape of introgression and incomplete lineage sorting we examined the phylogenetic topologies of each genes (Martin and Jiggins, 2017). For a three-species phylogeny using *O. barthii* as an outgroup there are three possible topologies. For each gene, topology-testing methods (Goldman *et al.*, 2000) can be used to determine which topology significantly fits the gene of interest (see Choi *et al.*, (2017) for detail). RAxML-estimated site-likelihood values were calculated for each gene and the significant topology was determined using the Approximately Unbiased (AU) test (Shimodaira, 2002) from the program CONSEL ver 0.20 (Shimodaira and Hasegawa, 2001). Genes with AU test results with a likelihood difference of zero were omitted and the topology with an AU test p-value of greater then 0.95 was selected.

Testing for evidence of admixture. Evidence of admixture between subpopulations was detected using the ABBA-BABA test D-statistics (Durand *et al.*, 2011; Green *et al.*, 2010). In a rooted three-taxon phylogeny [i.e. "((P1,P2),P3),O" where P1, P2, and P3 are the variety groups of interest and O is outgroup *O. barthii*], admixture can be inferred from the combination of ancestral ("A") and derived ("B") allelic states of each individual. The ABBA conformation arises when subpopulations P2 and P3 share derived alleles, while the BABA conformation is found when P1 and P3 share derived alleles. The difference in the frequency of the ABBA and BABA conformations is measured by the D-statistics, where significantly positive D-statistic indicate admixture between the P2 and P3 variety groups, while significantly negative D-statistics indicate admixture between the P1 and P3 variety groups. The genome was divided into 1-Mbp bins for jackknife resampling and calculating the standard errors. The significance of the D-statistic was calculated using the Z-test, and D-statistics with Z-scores greater than |3.9| ($p < 0.0001$) were considered significant.

Admixture can create phylogenetic signals that conflict with the underlying species phylogenetic tree. To differentiate the species topology created by lineage splitting events (i.e. speciation events) versus relatedness through admixture, we estimated the divergence times of each internal node to infer the correct topology that represented the species phylogeny. In a rooted three-taxon phylogeny ("((P1,P2),P3),O"), T_2 represents the divergence time of the two closely related sister species pair P1 and P2; and T_1 represents the divergence time of the distantly related species P3. T_1 and T_2 can be estimated from the allele combinations of the biallelic sites, where $T_1 = \frac{1}{Total_{Sites}}\left(\frac{Total_{ABAA}+Total_{BAAA}}{2} + Total_{BBAA}\right)$ and $T_2 = \frac{1}{Total_{Sites}}\left(\frac{Total_{ABAA}+Total_{BAAA}}{2}\right)$ (Fontaine *et al.*, 2015). Calculating T_1 and T_2 for the two conflicting topologies, if P3 is the source of introgression to P1 or P2, then T_2 will be lower in the topology consistent with the introgression; while if P1 or P2 is the source of introgression to P3, then both T_1 and T_2 will be lower in the topology consistent with the introgression (Wu *et al.*, 2018).

Detecting basmati-specific gene deletions. Basmati 334-specific gene deletions were detected and verified using several different methods. Initially, we searched the orthofinder results to find orthogroups for which there was at least one gene member from each of the genomes of the japonica (Nipponbare), indica (R498), and aus (N22) variety groups, but for which there was a gene missing for Basmati 334. We made sure that the Basmati 334 gene was indeed missing, and was not falsely inferred as missing due to a possible mis-annotation by the gene prediction software, by examining the genome alignments between the japonica and Basmati 334 genomes. If the orthologous region in the Basmati 334 genome lacked more than 90% of the japonica gene, then the deletion of the gene in Basmati 334 was determined to be "true", and not "false" due to a mis-annotation.

It is also possible a gene was deemed deleted because the region containing the deletion was not assembled correctly, and the gene was omitted from the genome alignment and gene annotation steps. To check for this possibility we aligned the raw nanopore sequencing reads to the japonica and aus reference genomes using ngmlr ver. 0.2.6 (Sedlazeck, Rescheneder, *et al.*, 2018). We then used the long read-specific structural variation detection program sniffles ver. 1.0.7 (Sedlazeck, Rescheneder, *et al.*, 2018) for calling deletions that have occurred in Basmati 334. IGV ver. 2.4 (Robinson *et al.*, 2011) was also used to visualize the alignment regions and verify the candidate deletions. Polymorphisms of the *Awn3-1* region was visualized using Rice SNP-Seek Database (Mansueto *et al.*, 2017).

DATA AVAILABILITY

Sequencing data generated from this study are available at the European Nucleotide Archive under bioproject ID PRJEB28274.

ACKNOWLEDGEMENTS

We thank Katherine Dorph for assistance with the DNA extractions, and Adrian Platts for computational support. We also thank Sissel Juul, David Dai, Eoghan Harrington, and Priyesh Rughani from Oxford Nanopore Technologies and the New York University Genomics Core Facility for sequencing support and the New York University High Performance Computing for supplying the computational resources. This work was supported by grants from the Gordon and Betty Moore Foundation through Grant GBMF2550.06 to S.C.G., and from the National Science Foundation Plant Genome Research Program (IOS-1546218), the Zegar Family Foundation (A16-0051) and the NYU Abu Dhabi Research Institute (G1205) to M.D.P.

Figure Legend

Figure 1. Groups of orthologous genes identified in the domesticated Asian rice genomes of basmati cv. Basmati 334, aus cv. N22, indica cv. R498, japonica cv. Nipponbare, and basmati cv. GP295-1.

Figure 2. Repetitive DNA landscape within the Basmati 334 genome. (A) Proportion of repeat family in Basmati 334 genome. (B) Distribution of insert time for the Gypsy and Copia LTR retrotransposons. (C) Distribution of insertion times for LTR retrotransposons across *Oryza* species with a *de novo* genome assembly.

Figure 3. Phylogenomics of the Asian rice complex. (A) Maximum-likelihood tree based on four-fold degenerate sites. Nodes with bootstrap support less then 100% are indicated. (B) Percentage of genes supporting the topology involving japonica cv. Nipponbare (J), basmati cv. Basmati 334 (B), and *O. rufipogon* (R). (C) Estimating the divergence times from two competing topologies involving the genomes from japonica (J), basmati (B), and *O. rufipogon* (R). (D) Percentage of genes supporting the topology involving aus cv. N22 (A), basmati cv. Basmati 334 (B), and indica cv. R498 (I).

Figure 4. Visualization of the genome region orthologous to the japonica cv. Nipponbare gene Os03g0418600 (*Awn3-1*) in the basmati cv. Basmati 334 and aus cv. N22 genome region.

Supplemental Figures

Supplemental Figure 1. Genome alignment dotplot comparing the genomes of basmati cv. Basmati 334 and japonica cv. Nipponbare and indica cv. R498.

Supplemental Figure 2. Distribution of the proportion of missing nucleotides for japonica cv. Nipponbare gene models across non-japonica genome gene models.

Supplemental Figure 3. Distribution of the proportion of missing nucleotides for japonica cv. Nipponbare gene models in basmati cv. Basmati 334 genome gene models, which were missed by the *de novo* gene annotation program but were detected in genome alignments.

Supplemental Figure 4. Distribution of LTR retrotransposon insertion times for each genome polishing steps.

Supplemental Figure 5. IGV view of the *Awn3-1* region from the genome alignment BAM file generated from aligning Basmati 334 reads to japonica cv. Nipponbare and aus cv. N22 reference genomes.

Supplemental Figure 6. Visualization of the SNPs from the 3,010 rice genomes across the *Awn3-1* region.

REFERENCES

Blanchette, M., Kent, W.J., Riemer, C., et al. (2004) Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Res.*, **14**, 708–715.

Boža, V., Brejová, B. and Vinař, T. (2016) DeepNano: Deep Recurrent Neural Networks for Base Calling in MinION Nanopore Reads D. Zhi, ed. *PLoS One*, **12**, e0178751.

Castillo, C.C., Tanaka, K., Sato, Y.-I., et al. (2016) Archaeogenetic study of prehistoric rice remains from Thailand and India: evidence of early japonica in South and Southeast Asia. *Archaeol. Anthropol. Sci.*, **8**, 523–543.

Chaisson, M.J.P., Huddleston, J., Dennis, M.Y., et al. (2015) Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, **517**, 608–611.

Chen, J., Huang, Q., Gao, D., et al. (2013) Whole-genome sequencing of Oryza brachyantha reveals mechanisms underlying Oryza genome evolution. *Nat. Commun.*, **4**, 1595.

Chen, S., Huang, Z., Zeng, L., Yang, J., Liu, Q. and Zhu, X. (2008) High-resolution mapping and gene prediction of Xanthomonas Oryzae pv. Oryzae resistance gene Xa7. *Mol. Breed.*, **22**, 433–441.

Choi, J.Y., Platts, A.E., Fuller, D.Q., Hsing, Y.-I., Wing, R.A. and Purugganan, M.D. (2017) The rice paradox: Multiple origins but single domestication in Asian rice. *Mol. Biol. Evol.*, **34**, 969–979.

Choi, J.Y. and Purugganan, M.D. (2018) Multiple Origin but Single Domestication Led to Oryza sativa. *G3 (Bethesda).*, **8**, 797–803.

Civáň, P., Craig, H., Cox, C.J. and Brown, T.A. (2015) Three geographically separate domestications of Asian rice. *Nat. plants*, **1**, 15164.

Conway, J.R., Lex, A., Gehlenborg, N. and Hancock, J. (2017) UpSetR: an R package for the visualization of intersecting sets and their properties J. Hancock, ed. *Bioinformatics*, **33**, 2938–2940.

Copetti, D., Zhang, J., Baidouri, M. El, et al. (2015) RiTE database: a resource database for genus-wide rice genomics and evolutionary biology. *BMC Genomics*, **16**, 538.

Coster, W. De, D'Hert, S., Schultz, D.T., Cruts, M. and Broeckhoven, C. Van (2018) NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*.

David, M., Dursi, L.J., Yao, D., Boutros, P.C., Simpson, J.T. and Birol, I. (2017) Nanocall: an open source basecaller for Oxford Nanopore sequencing data. *Bioinformatics*, **33**, 49–55.

Du, H., Yu, Y., Ma, Y., et al. (2017) Sequencing and de novo assembly of a near complete indica rice genome. *Nat. Commun.*, **8**, 15324.

Durand, E.Y., Patterson, N., Reich, D. and Slatkin, M. (2011) Testing for Ancient Admixture between Closely Related Populations. *Mol. Biol. Evol.*, **28**, 2239–2252.

Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, **9**, 18.

Emms, D.M. and Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.*, **16**, 157.

Finn, R.D., Coggill, P., Eberhardt, R.Y., et al. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.

Fontaine, M.C., Pease, J.B., Steele, A., et al. (2015) Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science (80-.).*, **347**, 1258524.

Fuller, D.Q. (2011) Finding Plant Domestication in the Indian Subcontinent. *Curr. Anthropol.*, **52**, S347–S362.

Fuller, D.Q., Sato, Y.-I., Castillo, C., Qin, L., Weisskopf, A.R., Kingwell-Banham, E.J., Song, J., Ahn, S.-M. and Etten, J. van (2010) Consilience of genetics and archaeobotany in the entangled history of rice. *Archaeol. Anthropol. Sci.*, **2**, 115–131.

Garris, A.J., Tai, T.H., Coburn, J., Kresovich, S. and McCouch, S. (2005) Genetic structure and diversity in Oryza sativa L. *Genetics*, **169**, 1631–8.

Gnanamanickam, S.S. (2009) Rice and Its Importance to Human Life. In *Biological Control of Rice Diseases*. Dordrecht: Springer Netherlands, pp. 1–11.

Goldman, N., Anderson, J.P., Rodrigo, A.G. and Olmstead, R. (2000) Likelihood-Based Tests of Topologies in Phylogenetics R. Olmstead, ed. *Syst. Biol.*, **49**, 652–670.

Goodwin, S., McPherson, J.D. and McCombie, W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**, 333–351.

Green, R.E., Krause, J., Briggs, A.W., et al. (2010) A draft sequence of the Neandertal genome. *Science (80-.).*, **328**, 710–22.

Gross, B.L. and Zhao, Z. (2014) Archaeological and genetic insights into the origins of domesticated rice. *Proc. Natl. Acad. Sci.*, **111**, 6190–6197.

Gu, B., Zhou, T., Luo, J., et al. (2015) An-2 Encodes a Cytokinin Synthesis Enzyme that Regulates Awn Length and Grain Production in Rice. *Mol. Plant*, **8**, 1635–1650.

Harris, R.S. (2007) Improved pairwise alignment of genomic dna. *Ph.D. Thesis, Pennsylvania State Univ.*

Haug-Baltzell, A., Stephens, S.A., Davey, S., Scheidegger, C.E., Lyons, E. and Hancock, J. (2017) SynMap2 and SynMap3D: web-based whole-genome synteny browsers J. Hancock, ed. *Bioinformatics*, **33**, 2197–2198.

He, G., Zhu, X., Elling, A.A., et al. (2010) Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. *Plant Cell*, **22**, 17–33.

Heather, J.M. and Chain, B. (2016) The sequence of sequencers: The history of sequencing DNA. *Genomics*, **107**, 1–8.

Howe, K. and Wood, J.M. (2015) Using optical mapping data for the improvement of vertebrate genome assemblies. *Gigascience*, **4**, 10.

Hua, L., Wang, D.R., Tan, L., et al. (2015) LABA1, a Domestication Gene Associated with Long, Barbed Awns in Wild Rice. *Plant Cell*, **27**, 1875–1888.

Huang, X. and Han, B. (2015) Rice domestication occurred through single origin and multiple introgressions. *Nat. Plants*, **2**, 15207.

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

Hubisz, M.J., Pollard, K.S. and Siepel, A. (2011) PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinform.*, **12**, 41–51.

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

Jain, M., Koren, S., Miga, K.H., et al. (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*, **36**, 338–345.

Jain, M., Olsen, H.E., Paten, B. and Akeson, M. (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.*, **17**, 239.

Jain, S., Jain, R.K. and McCouch, S.R. (2004) Genetic analysis of Indian aromatic and quality rice (Oryza sativa L.) germplasm using panels of fluorescently-labeled microsatellite markers. *Theor. Appl. Genet.*, **109**, 965–977.

Jiao, W.-B. and Schneeberger, K. (2017) The impact of third generation genomic technologies on plant genome assembly. *Curr. Opin. Plant Biol.*, **36**, 64–70.

Jones, P., Binns, D., Chang, H.-Y., et al. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

Katoh, K. and Standley, D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.*, **30**, 772–780.

Kawahara, Y., la Bastide, M. de, Hamilton, J.P., et al. (2013) Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice*, **6**, 4.

Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. and Haussler, D. (2003) Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci.*, **100**, 11484–11489.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16**, 111–20.

Konishi, S., Izawa, T., Lin, S.Y., Ebana, K., Fukuta, Y., Sasaki, T. and Yano, M. (2006) An SNP Caused Loss of Seed Shattering During Rice Domestication. *Science (80-.).*, **312**, 1392–1396.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, **27**, 722–736.

Kovach, M.J., Calingacion, M.N., Fitzgerald, M.A. and McCouch, S.R. (2009) The origin and evolution of fragrance in rice (Oryza sativa L.). *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 14444–9.

Kumar, A. and Bennetzen, J.L. (1999) Plant Retrotransposons. *Annu. Rev. Genet.*, **33**, 479–532.

Li, B., Zhang, Y., Li, J., Yao, G., Pan, H., Hu, G., Chen, C., Zhang, H. and Li, Z. (2016) Fine Mapping of Two Additive Effect Genes for Awn Development in Rice (Oryza sativa L.). *PLoS One*, **11**, e0160792.

Li, C., Lin, F., An, D., Wang, W. and Huang, R. (2017) Genome Sequencing and Assembly by Long Reads in Plants. *Genes (Basel).*, **9**.

Li, C., Zhou, A. and Sang, T. (2006) Rice domestication by reducing shattering. *Science (80-.).*, **311**, 1936–9.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 1303.3997v2.

Li, H. (2016) Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, **32**, 2103–2110.

Lin, Z., Griffith, M.E., Li, X., et al. (2007) Origin of seed shattering in rice (Oryza sativa L.). *Planta*, **226**, 11–20.

Lisch, D. (2012) How important are transposons for plant evolution? *Nat. Rev. Genet.*, **14**, 49–61.

Loman, N.J., Quick, J. and Simpson, J.T. (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods*, **12**, 733–735.

Luo, J., Liu, H., Zhou, T., et al. (2013) An-1 encodes a basic helix-loop-helix protein that regulates awn development, grain size, and grain number in rice. *Plant Cell*, **25**, 3360–76.

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

Mansueto, L., Fuentes, R.R., Borja, F.N., et al. (2017) Rice SNP-seek database update: new SNPs, indels, and queries. *Nucleic Acids Res.*, **45**, D1075–D1081.

Martin, S.H. and Jiggins, C.D. (2017) Interpreting the genomic landscape of introgression. *Curr. Opin. Genet. Dev.*, **47**, 69–74.

Matsuo, T., Futsuhara, Y., Kikuchi, F. and Yamaguchi, H. (1997) *Science of the Rice Plant*, Tokyo: Food and Agriculture Policy Research Center.

Meyer, R.S. and Purugganan, M.D. (2013) Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.*, **14**, 840–852.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D. and Thomas, P.D. (2017) PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.*, **45**, D183–D189.

Michael, T.P., Jupe, F., Bemm, F., Motley, S.T., Sandoval, J.P., Lanz, C., Loudet, O., Weigel, D. and Ecker, J.R. (2018) High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nat. Commun.*, **9**, 541.

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

Rahman, A.N.M.R. Bin and Zhang, J. (2018) Preferential Geographic Distribution Pattern of Abiotic Stress Tolerant Rice. *Rice*, **11**, 10.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.

Sakai, H., Kanamori, H., Arai-Kichise, Y., et al. (2014) Construction of Pseudomolecule Sequences of the aus Rice Cultivar Kasalath for Comparative Genomics of Asian Cultivated Rice. *DNA Res.*, **21**, 397–405.

Sakai, H., Lee, S.S., Tanaka, T., et al. (2013) Rice Annotation Project Database (RAP-DB): An Integrative and Interactive Database for Rice Genomics. *Plant Cell Physiol.*, **54**, e6–e6.

Sandhu, N., Kumar, A., Sandhu, N. and Kumar, A. (2017) Bridging the Rice Yield Gaps under Drought: QTLs, Genes, and their Use in Breeding Programs. *Agronomy*, **7**, 27.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. and Bennetzen, J.L. (1998) The paleontology of intergene retrotransposons of maize. *Nat. Genet.*, **20**, 43–45.

Schatz, M.C., Maron, L.G., Stein, J.C., et al. (2014) 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.

Schmidt, M.H., Vogel, A., Denton, A.K., et al. (2017) De novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. *Plant Cell*, tpc.00521.2017.

Sedlazeck, F.J., Lee, H., Darby, C.A. and Schatz, M.C. (2018) Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.*, **19**, 329–346.

Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Haeseler, A. von and Schatz, M.C. (2018) Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods*, 1.

Shimodaira, H. (2002) An Approximately Unbiased Test of Phylogenetic Tree Selection. *Syst. Biol.*, **51**, 492–508.

Shimodaira, H. and Hasegawa, M. (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, **17**, 1246–7.

Shomura, A., Izawa, T., Ebana, K., Ebitani, T., Kanegae, H., Konishi, S. and Yano, M. (2008) Deletion in a gene associated with grain size increased yields during rice domestication. *Nat. Genet.*, **40**, 1023–1028.

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

Singh, R., Singh, U. and Khush, G. eds. (2000) *Aromatic rices*, Oxford & IBH Publishing Co Pvt Ltd.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Stanke, M., Schöffmann, O., Morgenstern, B. and Waack, S. (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62.

Stanke, M. and Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19**, ii215-ii225.

Stein, J.C., Yu, Y., Copetti, D., et al. (2018) Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. *Nat. Genet.*, **50**, 285–296.

Stoiber, M. and Brown, J. (2017) BasecRAWller: Streaming Nanopore Basecalling Directly from Raw Signal. *bioRxiv*, 133058.

Sweeney, M.T., Thomson, M.J., Pfeil, B.E. and McCouch, S. (2006) Caught red-handed: Rc encodes a basic helix-loop-helix protein conditioning red pericarp in rice. *Plant Cell*, **18**, 283–94.

Teng, H., Cao, M.D., Hall, M.B., Duarte, T., Wang, S. and Coin, L.J.M. (2018) Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. *Gigascience*, **7**.

Tyson, J.R., O'Neil, N.J., Jain, M., Olsen, H.E., Hieter, P. and Snutch, T.P. (2018) MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome. *Genome Res.*, **28**, 266–274.

Udall, J.A. and Dawe, R.K. (2018) Is It Ordered Correctly? Validating Genome Assemblies by Optical Mapping. *Plant Cell*, **30**, 7–14.

Ullah, I., Jamil, S., Iqbal, M.Z., Shaheen, H.L., Hasni, S.M., Jabeen, S., Mehmood, A. and Akhter, M. (2012) Detection of bacterial blight resistance genes in basmati rice landraces. *Genet. Mol. Res.*, **11**, 1960–1966.

Vaser, R., Sović, I., Nagarajan, N. and Šikić, M. (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.*, **27**, 737–746.

Vikram, P., Swamy, B.P.M., Dixit, S., Ahmed, H., Cruz, M.T.S., Singh, A.K., Ye, G. and Kumar, A. (2012) Bulk segregant analysis: "An effective approach for mapping consistent-effect drought grain yield QTLs in rice." *F. Crop. Res.*, **134**, 185–192.

Walker, B.J., Abeel, T., Shea, T., et al. (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS One*, **9**, e112963.

Wang, H., Vieira, F.G., Crawford, J.E., Chu, C. and Nielsen, R. (2017) Asian wild rice is a hybrid swarm with extensive gene flow and feralization from domesticated rice. *Genome Res.*, **27**, 1029–1038.

Wang, M., Yu, Y., Haberer, G., et al. (2014) The genome sequence of African rice (Oryza glaberrima) and evidence for independent domestication. *Nat. Genet.*, **46**, 982–988.

Wang, W., Mauleon, R., Hu, Z., et al. (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, **557**, 43–49.

Wang, X., Shi, X., Hao, B., Ge, S. and Luo, J. (2005) Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytol.*, **165**, 937–946.

Wang, Z.Y., Zheng, F.Q., Shen, G.Z., Gao, J.P., Snustad, D.P., Li, M.G., Zhang, J.L. and Hong, M.M. (1995) The amylose content in rice endosperm is related to the post-transcriptional regulation of the waxy gene. *Plant J.*, **7**, 613–22.

Wendel, J.F., Jackson, S.A., Meyers, B.C. and Wing, R.A. (2016) Evolution of plant genome architecture. *Genome Biol.*, **17**, 37.

Wicker, T., Sabot, F., Hua-Van, A., et al. (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.*, **8**, 973–982.

Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C. and Gough, J. (2009) SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.*, **37**, D380–D386.

Wu, M., Kostyun, J. and Moyle, L. (2018) Genome sequence of Jaltomata addresses rapid reproductive trait evolution and enhances comparative genomics in the hyper-diverse Solanaceae. *bioRxiv*, 335117.

Xiong, L.Z., Liu, K.D., Dai, X.K., Xu, C.G. and Zhang, Q. (1999) Identification of genetic factors controlling domestication-related traits of rice using an F 2 population of a cross between Oryza sativa and O. rufipogon. *TAG Theor. Appl. Genet.*, **98**, 243–251.

Xu, K., Xu, X., Fukao, T., et al. (2006) Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature*, **442**, 705–708.

Yang, Z. (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.

Yu, J., Hu, S., Wang, J., et al. (2002) A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica). *Science (80-.).*, **296**, 79–92.

Zdobnov, E.M., Campillos, M., Harrington, E.D., Torrents, D. and Bork, P. (2005) Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. *Nucleic Acids Res.*, **33**, 946–54.

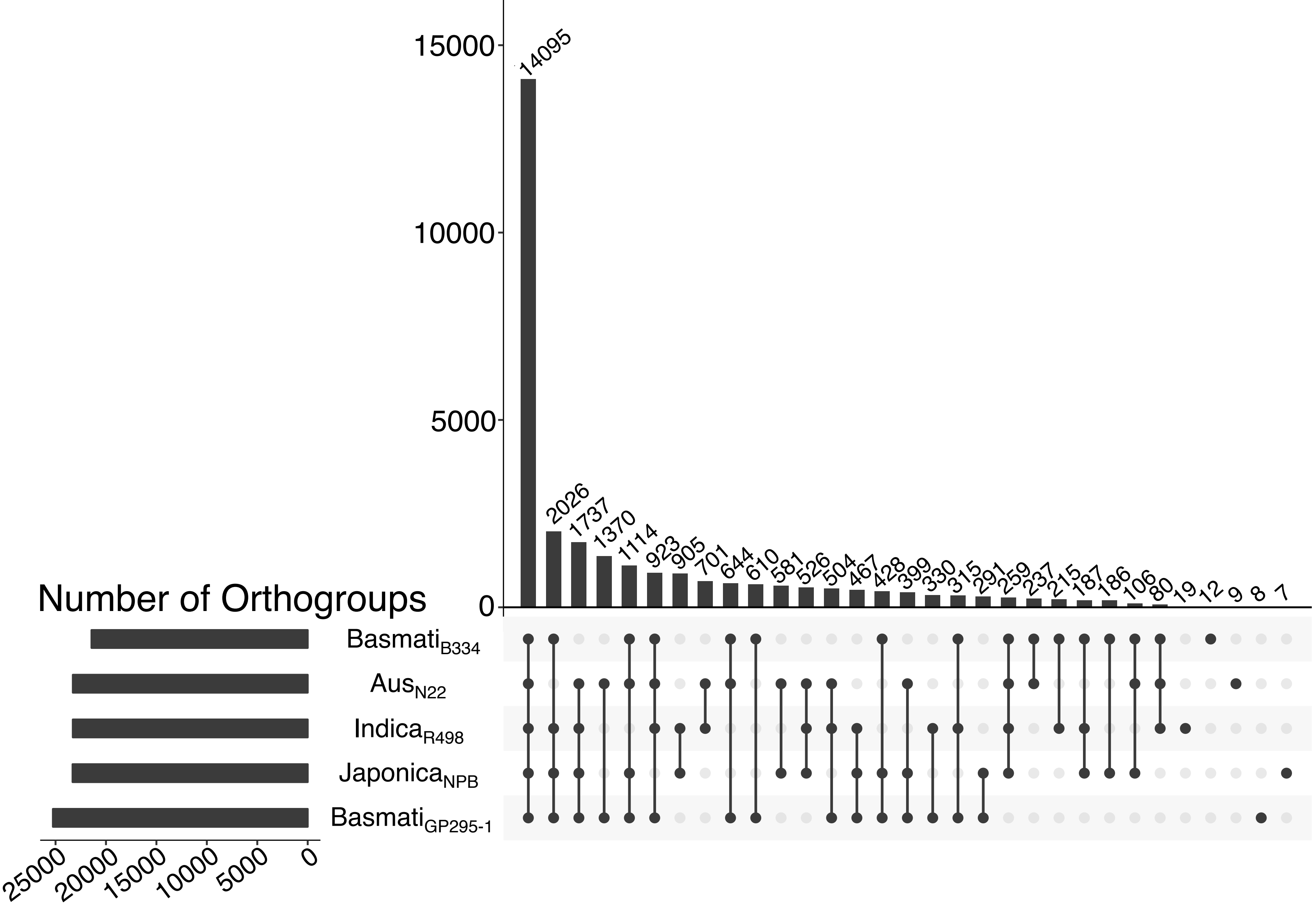
Zhang, J., Chen, L.-L., Xing, F., et al. (2016) Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. *Proc. Natl. Acad. Sci. U. S. A.*, **113**, E5163-71.

Zhang, Y., Zhang, S., Liu, H., et al. (2015) Genome and Comparative Transcriptomics of African Wild Rice Oryza longistaminata Provide Insights into Molecular Mechanism of Rhizomatousness and Self-Incompatibility. *Mol. Plant*, **8**, 1683–1686.

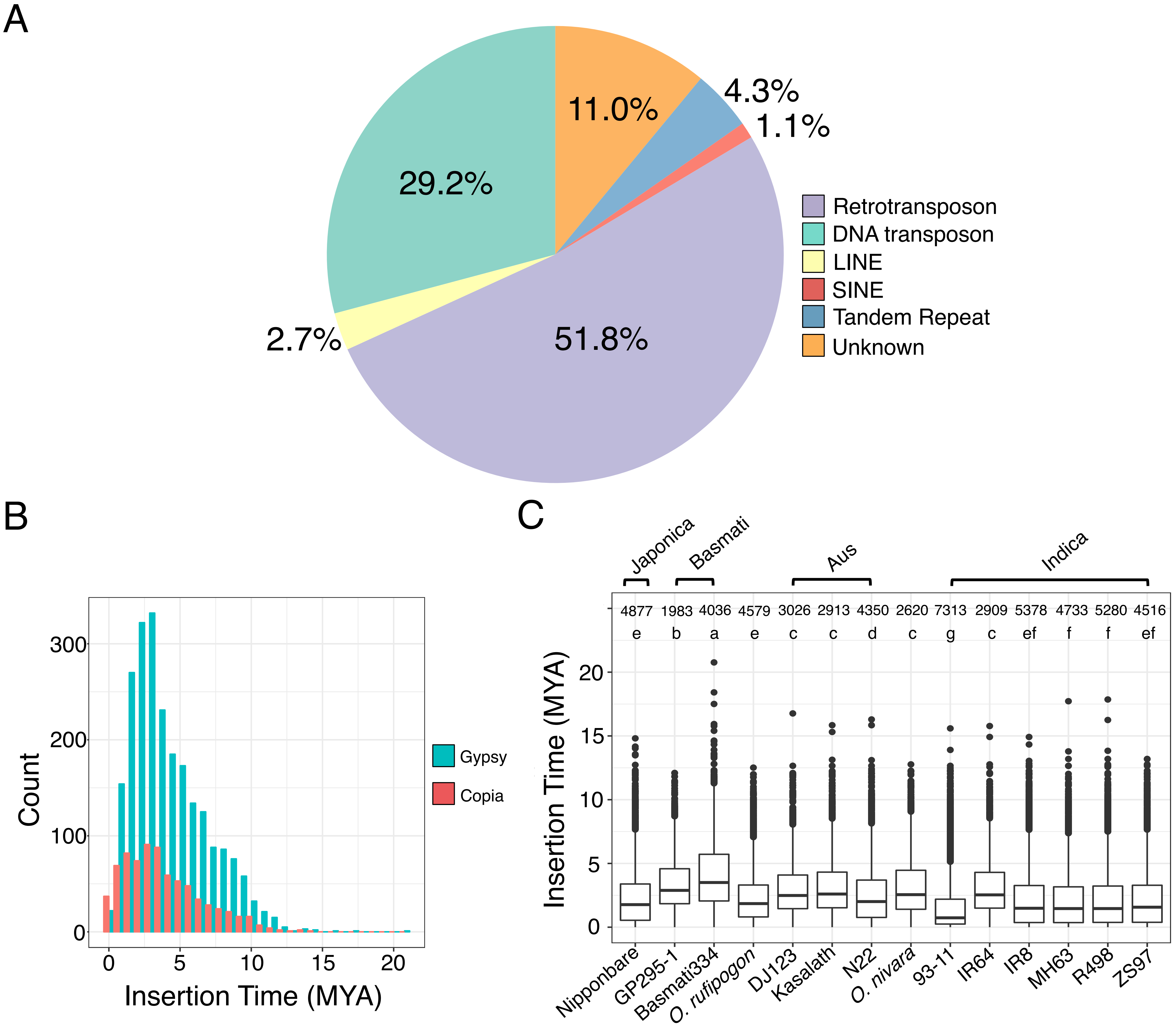
Zhao, Q., Feng, Q., Lu, H., et al. (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.*, **50**, 278–284.

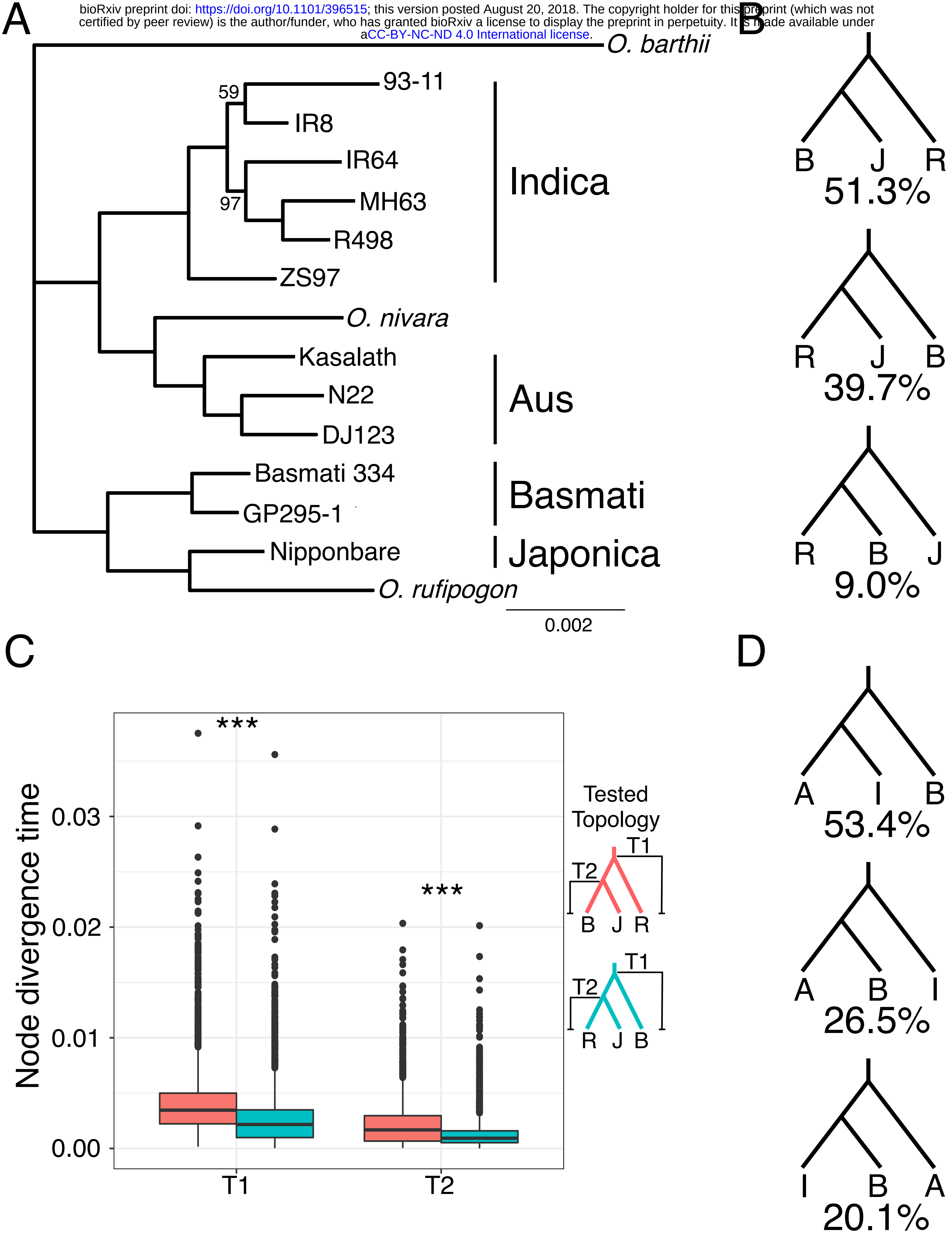
Zhou, Y., Zhu, J., Li, Z., Yi, C., Liu, J., Zhang, H., Tang, S., Gu, M. and Liang, G. (2009) Deletion in a Quantitative Trait Gene qPE9-1 Associated With Panicle Erectness Improves Plant Architecture

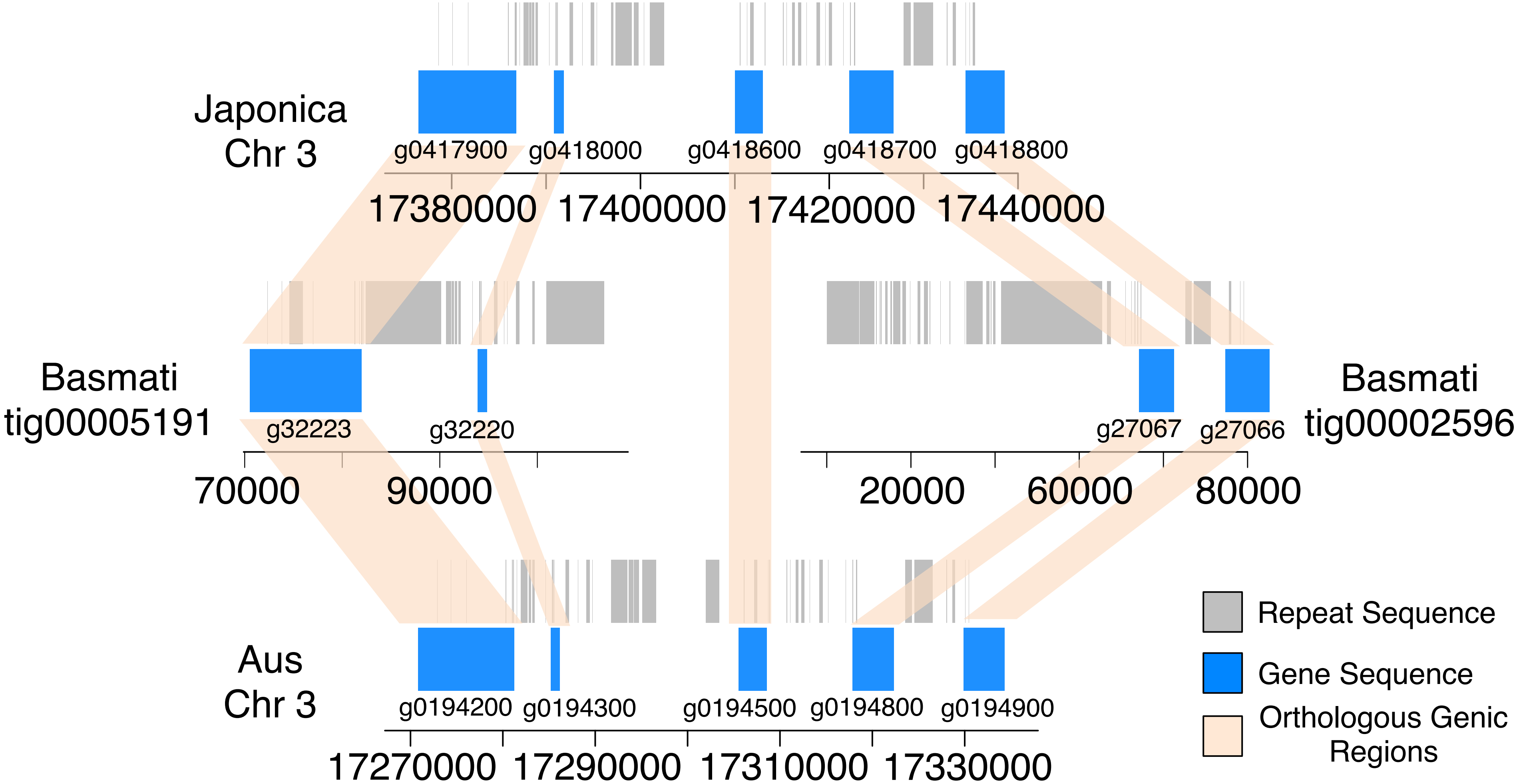
During Rice Domestication. *Genetics*, **183**, 315–324.



15000
10000
5000
0
14095
2026
1737
1370
1114
923
905
701
644
610
581
526
504
467
428
399
330
315
291
259
237
215
187
186
106
80
19
12
9
8
7
Number of Orthogroups
Basmati$_{B334}$
Aus$_{N22}$
Indica$_{R498}$
Japonica$_{NPB}$
Basmati$_{GP295-1}$
25000
20000
15000
10000
5000
0


A
29.2%
11.0%
4.3%
1.1%
2.7%
51.8%
Retrotransposon
DNA transposon
LINE
SINE
Tandem Repeat
Unknown
B
Count
300
200
100
0
0
5
10
15
20
Insertion Time (MYA)
Gypsy
Copia
C
Japonica
Basmati
Aus
Indica
4877 1983 4036 4579 3026 2913 4350 2620 7313 2909 5378 4733 5280 4516
e b a e c c d c g c ef f f ef
Insertion Time (MYA)
20
15
10
5
0
Nipponbare
GP295-1
Basmati334
O. rufipogon
DJ123
Kasalath
N22
O. nivara
93-11
IR64
IR8
MH63
R498
ZS97


A
O. barthii
59
93-11
IR8
IR64
97
MH63
R498
ZS97
Indica
O. nivara
Kasalath
N22
DJ123
Aus
Basmati 334
GP295-1
Basmati
Nipponbare
Japonica
O. rufipogon
0.002
B
B J R
51.3%
R J B
39.7%
R B J
9.0%
C
Node divergence time
0.00
0.01
0.02
0.03
***
***
T1
T2
Tested Topology
T1
T2
B J R
T1
T2
R J B
D
A I B
53.4%
A B I
26.5%
I B A
20.1%


Japonica
Chr 3
g0417900
g0418000
g0418600
g0418700
g0418800
17380000
17400000
17420000
17440000
Basmati
tig00005191
g32223
g32220
70000
90000
Basmati
tig00002596
g27067
g27066
20000
60000
80000
Aus
Chr 3
g0194200
g0194300
g0194500
g0194800
g0194900
17270000
17290000
17310000
17330000
Repeat Sequence
Gene Sequence
Orthologous Genic
Regions