

Supplementary Materials for

**Chromosome-level genome assemblies reveal genome evolution of an
invasive plant *Phragmites australis***

Cui Wang *et al.*

*Corresponding author. Email: whguo@sdu.edu.cn, jarkko@ntu.edu.sg

**This PDF file includes:
Figure S1-S11**

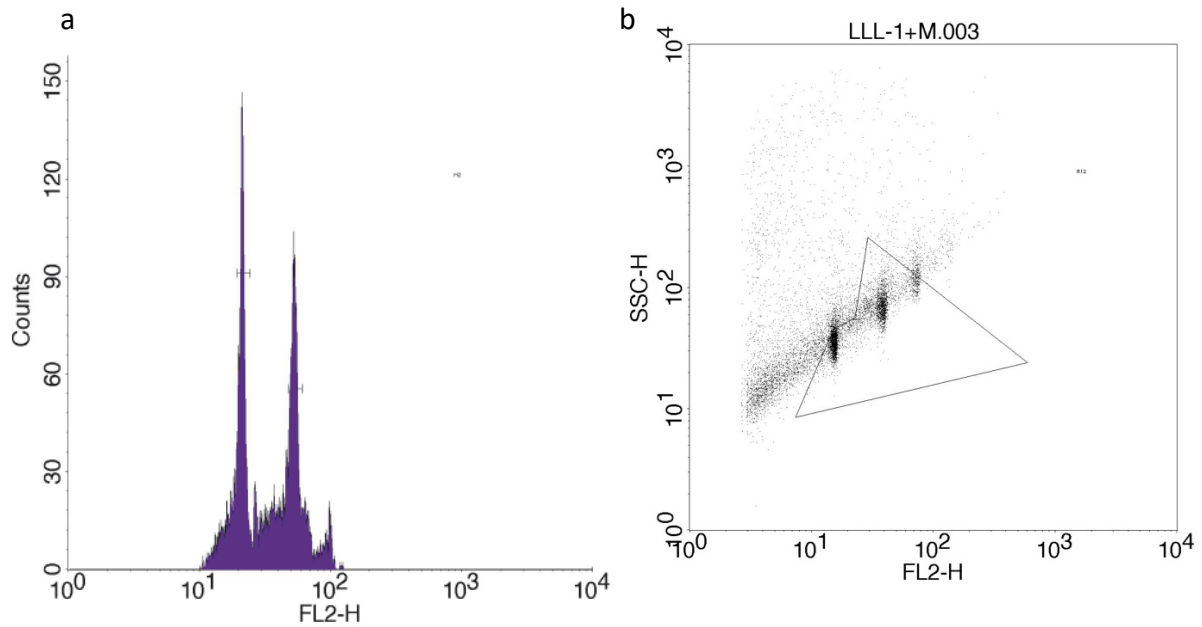


Figure S1. a. The DNA content per cell estimated using flow cytometry suggests a genome size of 920Mb for *P. australis* (left peak). Maize (*Zea mays*) was used as a control (right peak). b. Forward and side scatter gating strategy used for the flow cytometry measurement.

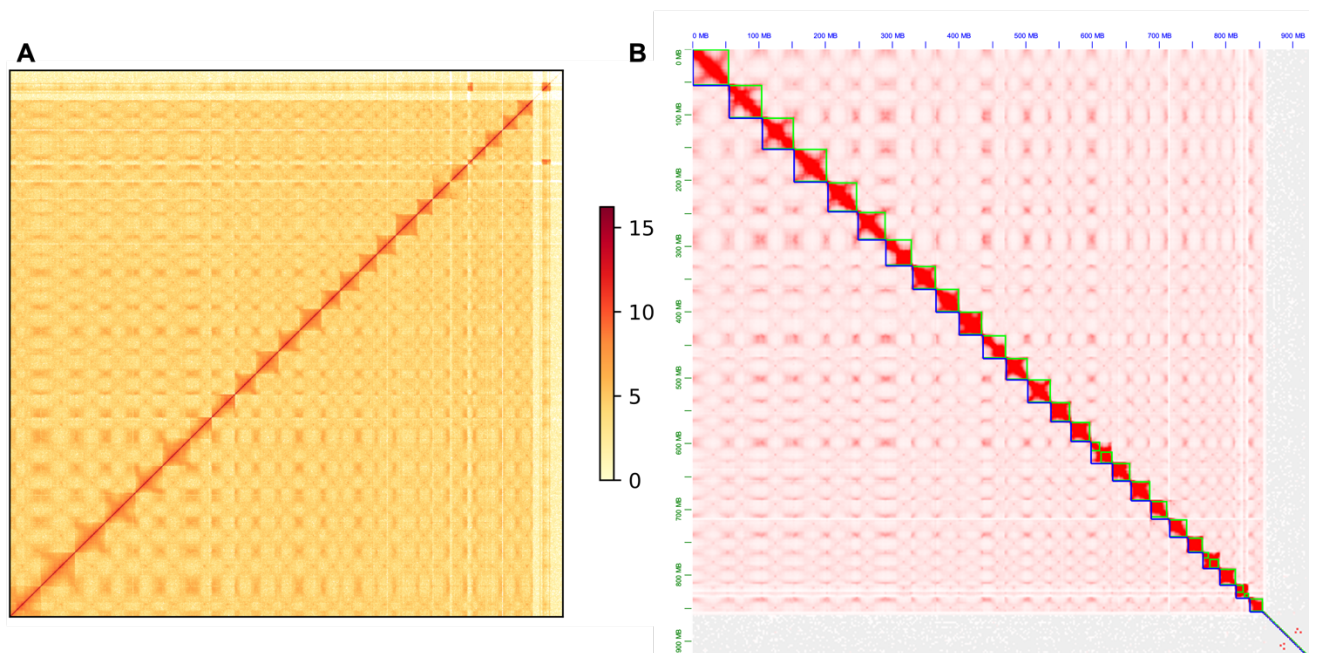


Figure S2. Hi-C contact heat maps generated by aligning Hi-C short read sequences against the draft assembly. **A.** The initial draft was assembled from with Pacbio HiFi reads and then scaffolded into 25 pseudomolecules using Hi-C contacts using 100x coverage data and the ALLHiC pipeline. The heatmap displays the contiguous blocks (pseudochromosomes) after scaffolding the draft assembly into 25 pseudochromosomes. One small block in the top right corner, showing substantial contacts with the fourth block from the top was manually corrected. **B.** After manual correction using Juicebox and the 3d-dna pipeline, the assembly errors were resolved.

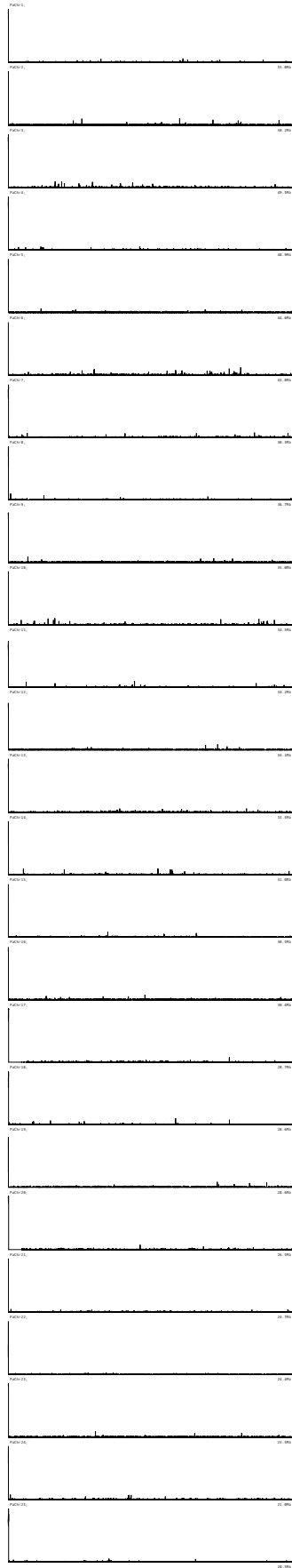


Figure S3. Chromosomal regions with telomere repeat counts detected using a sliding window size of 10,000 bp. Telomeres were detected at both ends of all of the 25 pseudochromosomes.

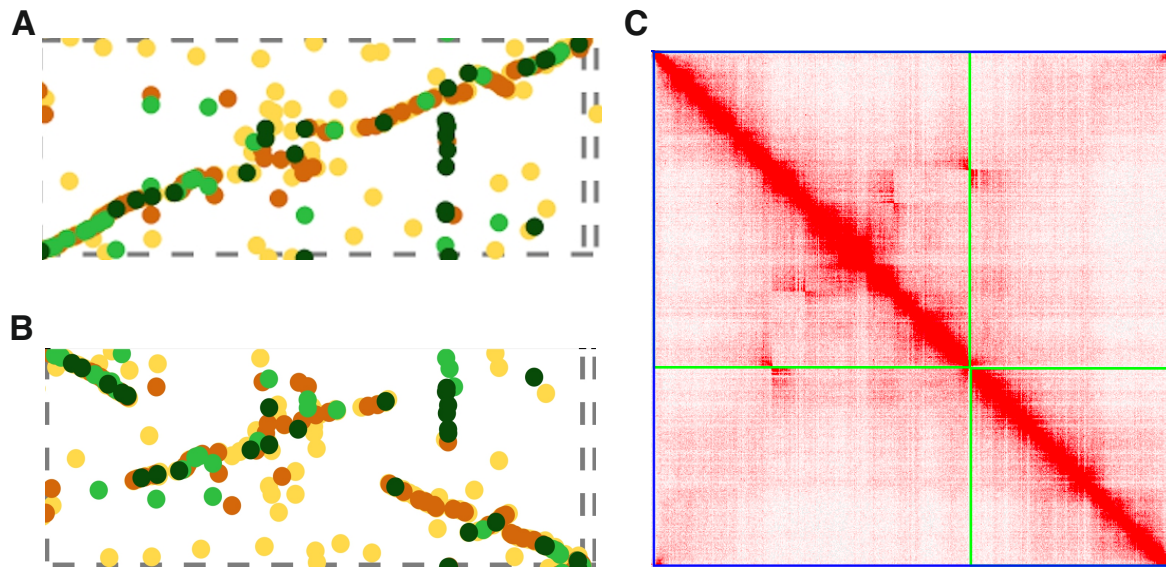


Figure S4. A large haplotypic inversion was detected on Chromosome 22. The two haplotig assemblies from Hifiasm synteny analysis between the two haplotypes and rice (**A** and **B**) indicated an inversion in one haplotype (**B**). This inversion is also evident from the Hi-C contact heatmap (**C**).

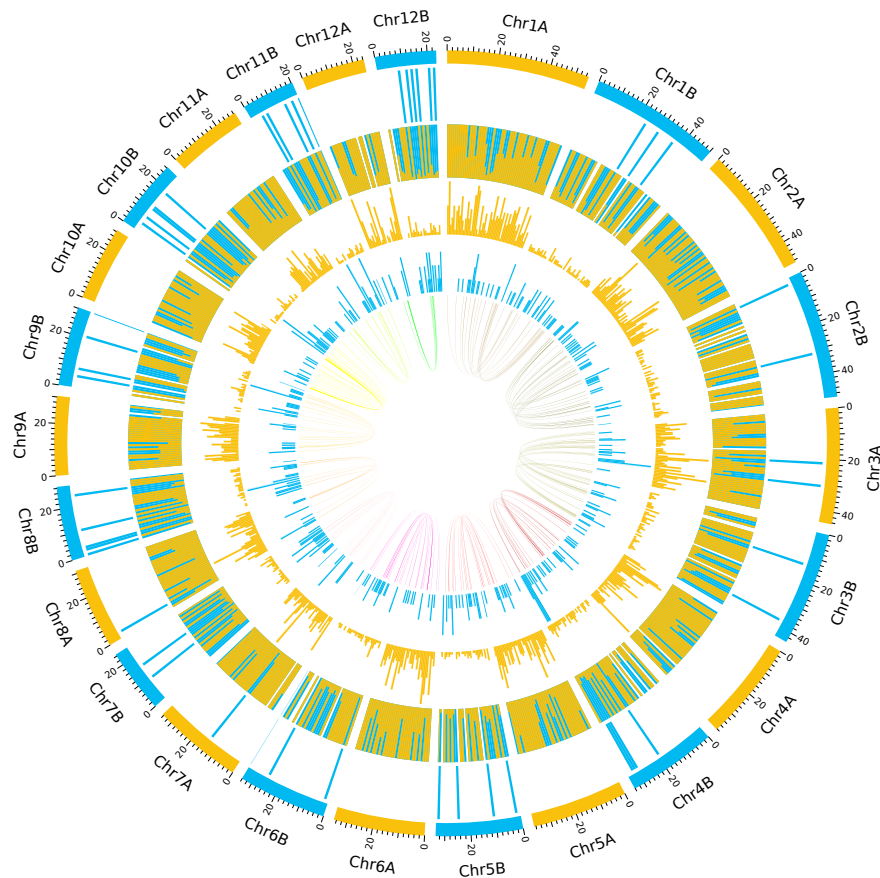


Figure S5. The homoeologous chromosome pairs for the two subgenomes revealed by Subphaser. Chromosomal characteristics from outer to inner circles indicates (1-6): (1) subgenome assignments based on k-means algorithm; (2) significant enrichment of subgenome-specific k-mers – the same color as the subgenome indicates significant enrichment for those subgenome-specific k-mers; white areas are not significantly enriched; (3) normalized proportion (relative) of subgenome-specific k-mers; (4 – 5) count (absolute) of each subgenome-specific k-mer set; (6) homoeologous blocks.

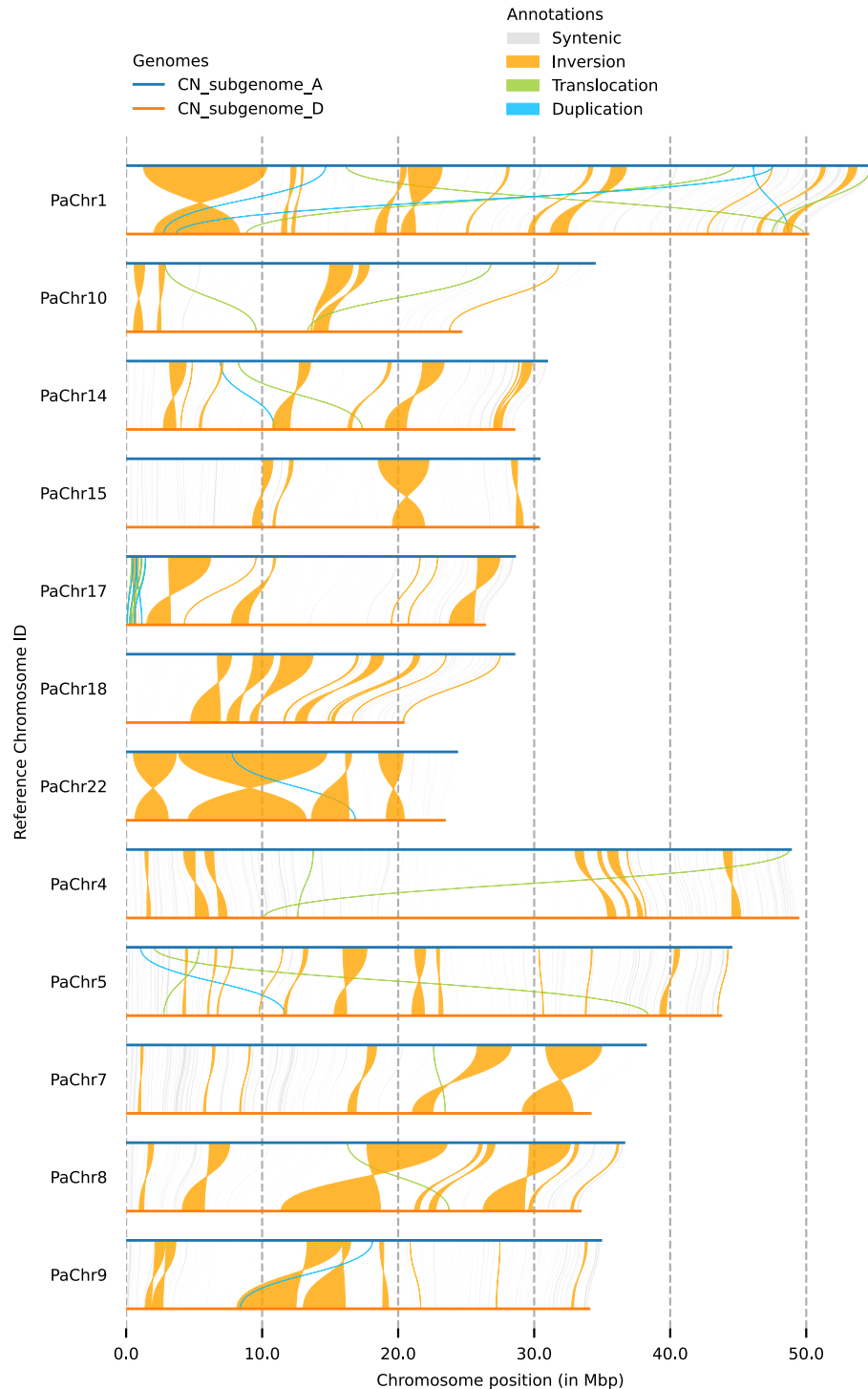


Figure S6. Genome rearrangements between the two subgenomes of *P. australis*. Large inversions were detected in several homoeologous chromosome pairs.

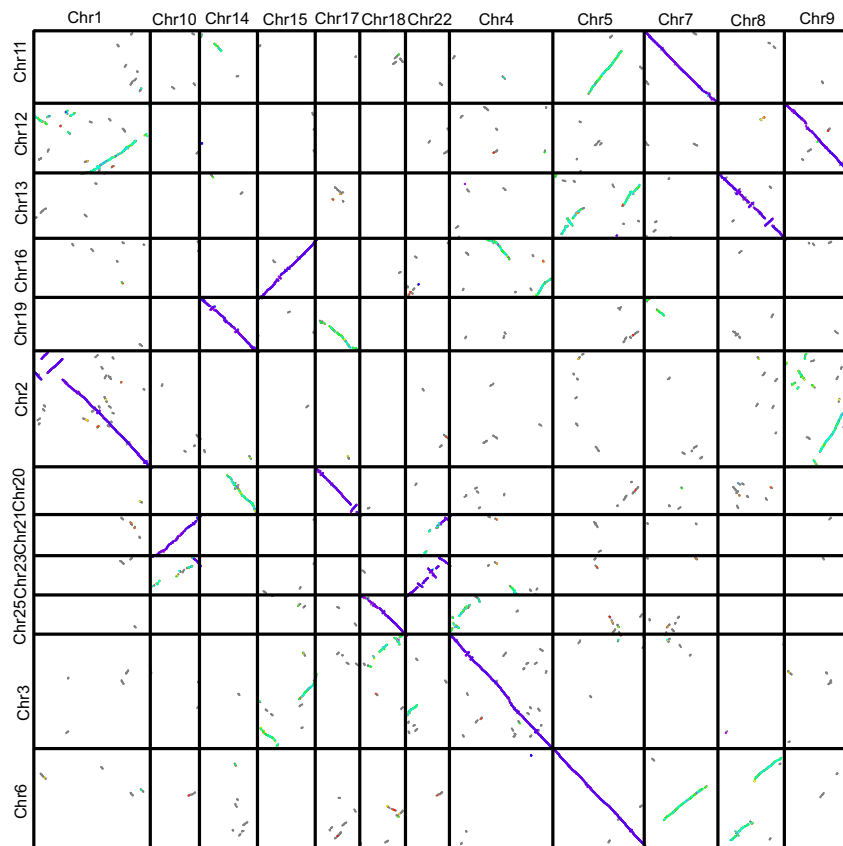


Figure S7. Syntenic dotplot of the two subgenomes of *P. australis* aligned against each other. Colors correspond to the number of synonymous mutations (Ks) between syntelogs, with purple syntenic regions originating from the common reed allopolyploid event (more recent) and green, yellow and red correspond to syntenic blocks originating from the earlier whole genome duplication events.

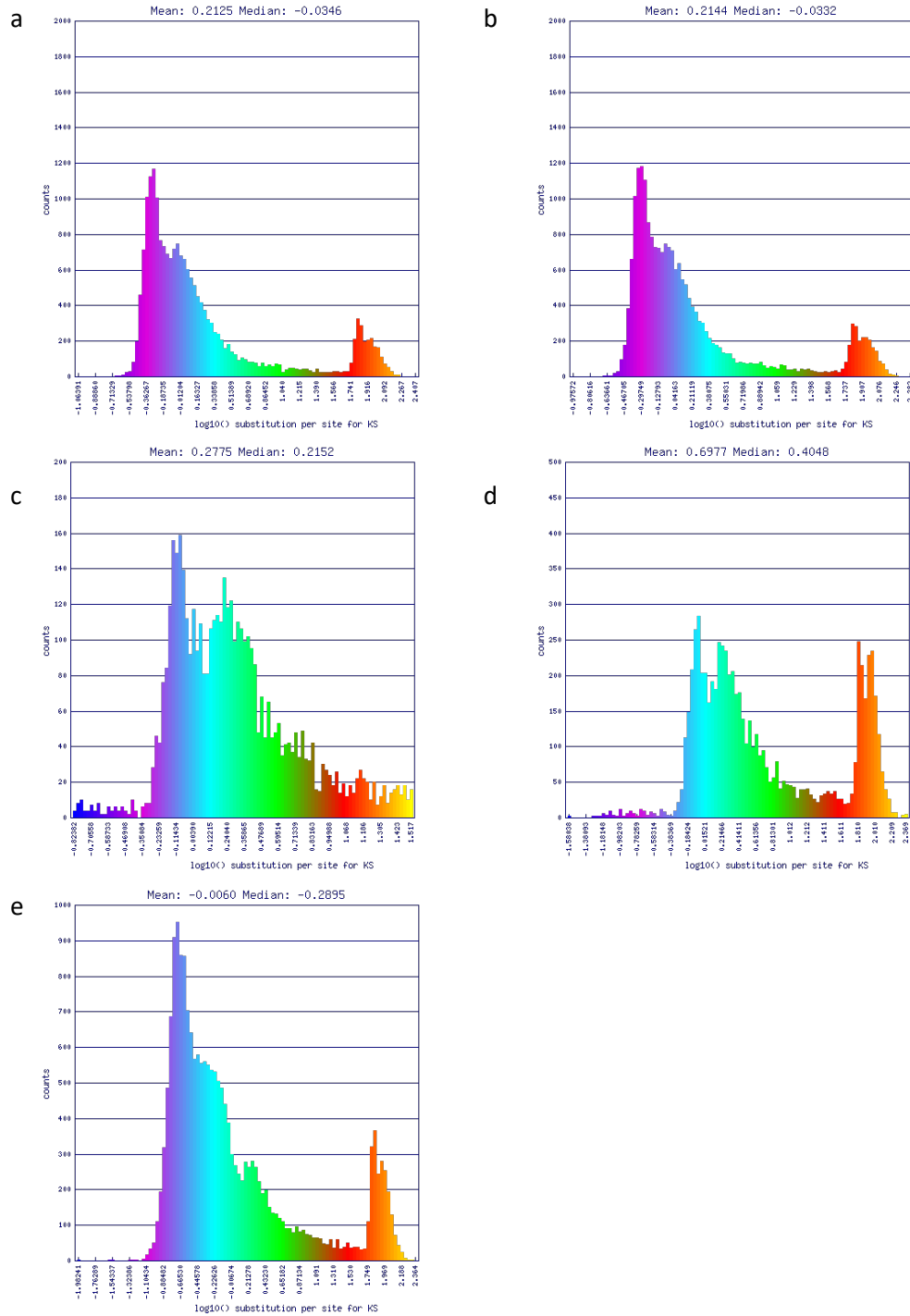


Figure S8. Synonymous substitution (Ks) histograms for the syntenic genes between subgenomes of the reference *P. australis* and rice (*Oryza sativa*), as well as self-self alignments of subgenomes. The panels show histograms of Ks values transformed with log10 when aligning (a) subgenome A to rice genome, (b) subgenome D to rice genome, (c) subgenome A to itself, (d) subgenome D to itself, (e) the two subgenomes against each other. Ks value on the x-axis was. For all the figures, the orange peaks on the right with the transformed Ks values higher than 1.6 are noise due to mis-called syntenic pairs or wrong prediction of gene models. The time of whole genome duplication (WGD) events in subgenome A and D can be inferred from the self-self alignments from each subgenome (c, d) with two peak values higher than -0.13, suggests all the WGD events are predating the

divergence date with rice which shows a most recent peak at around -0.29. Divergence time between the subgenomes can be inferred from the most recent peak of E, at around -0.63.

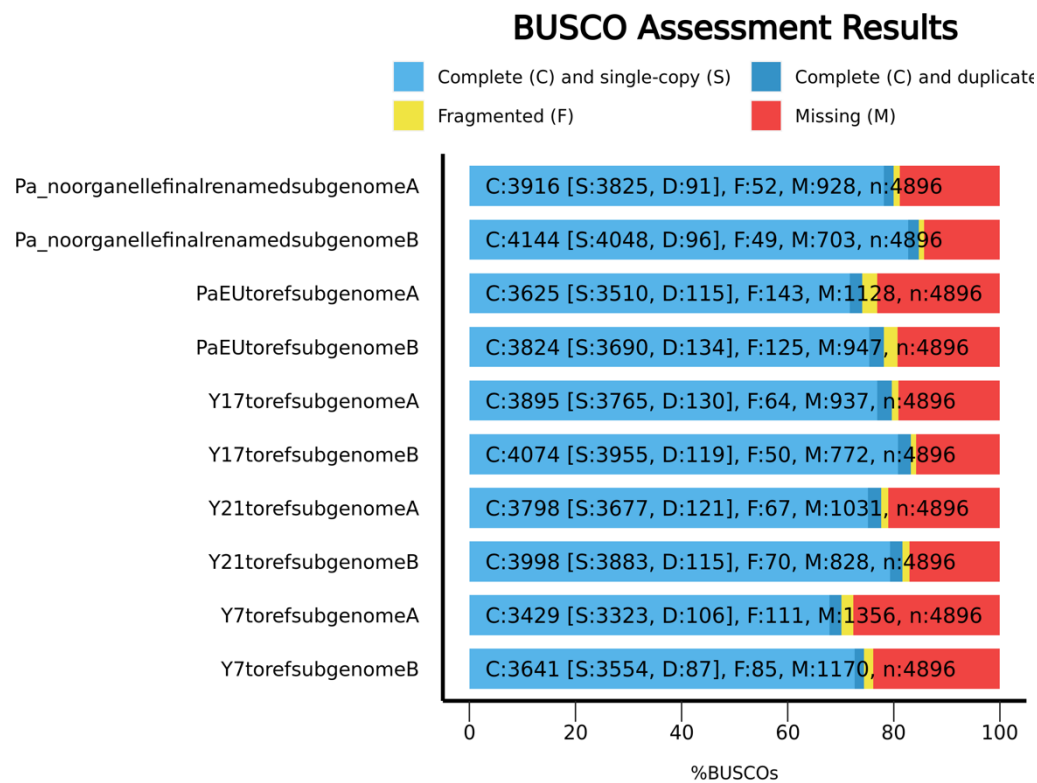


Figure S9. BUSCO assessment for all the subgenomes of the five-representative lineages. From the top to bottom, the first two bar plots show the BUSCO qualities for the subgenome A and D for Chinese reference individual genome assembly. Draft genomes of the other four genomes were scaffolded into chromosome-level using their homology to the reference genome, subgenomes were phased in this process. From the third bar plot to the bottom, the chart shows BUSCO scores for the EU invasive lineage subgenomes A, and D, North American native lineage (Y17, USnat) subgenomes A and D, Mediterranean lineage (Y21, Med) subgenomes A and D, as well as Gulf Coast Land type (Y7, USland) subgenomes A and D.

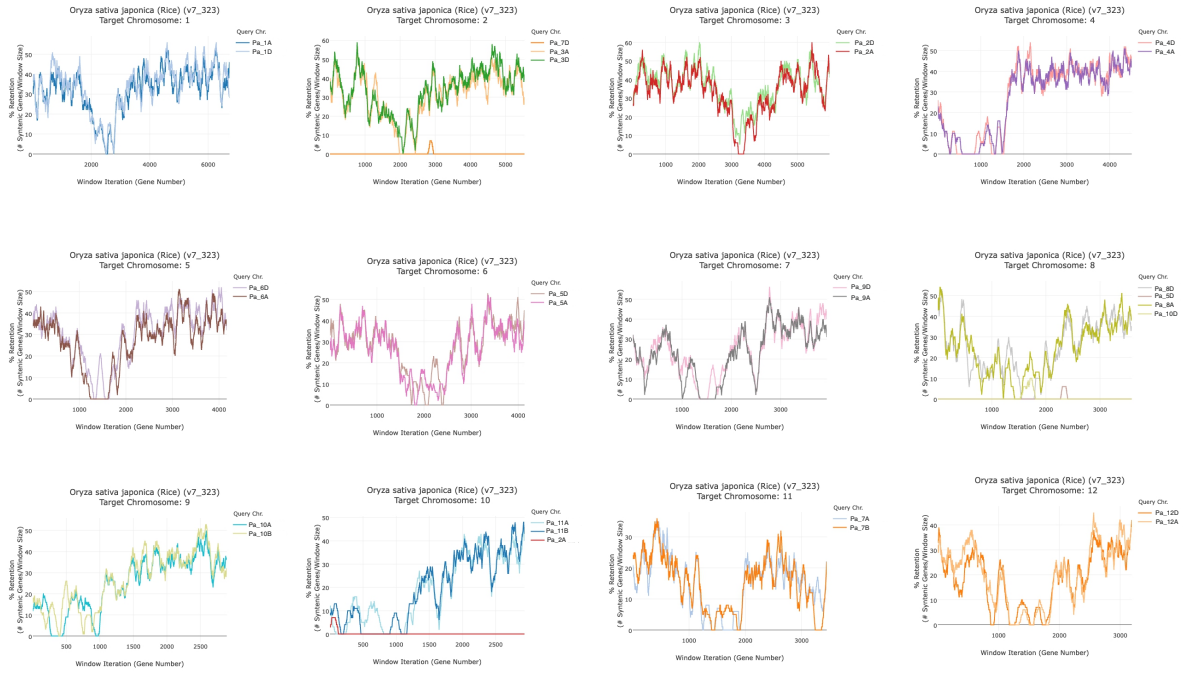
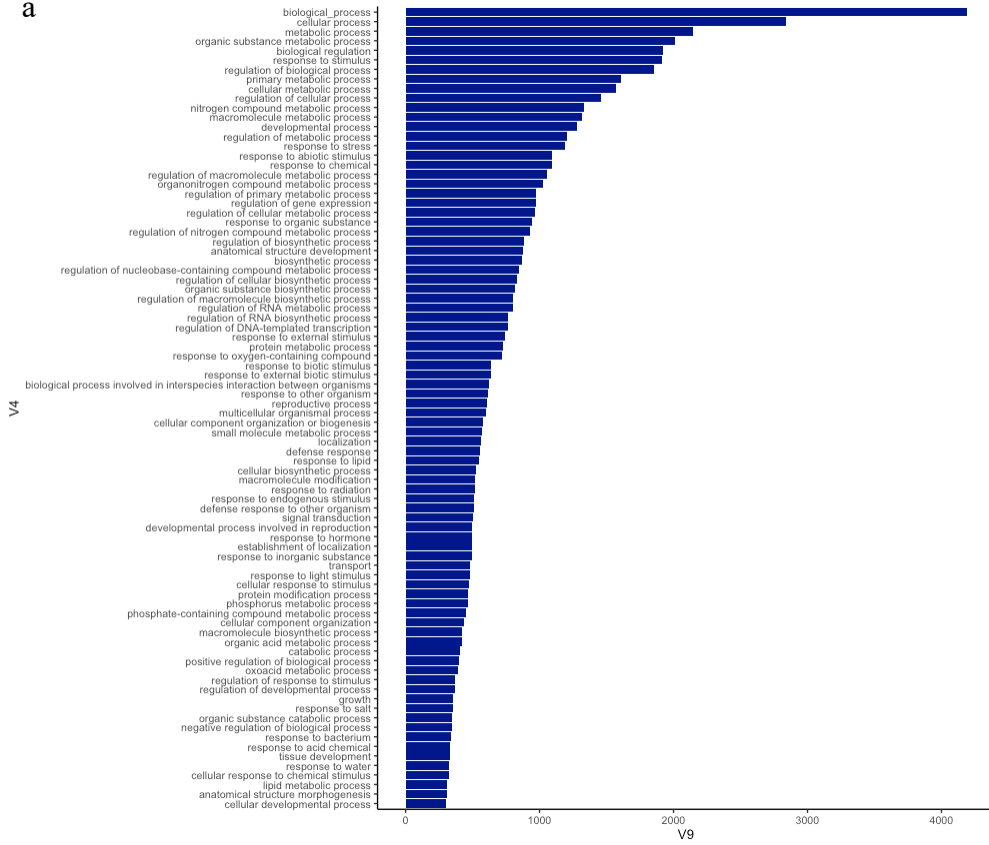


Figure S10. Fractionation biases of subgenome D vs subgenome A in *P. australis* reference, when aligned against rice. The charts show the number of syntenic genes retained in sliding genomic windows of 100 genes along the homoeologous chromosome pairs. The gene order on x axis follows the chromosome ordering of *O. sativa japonica*. The overall fractionation is a result from the divergence between *O. sativa* and the common ancestor of *P. australis* progenitors, while the differences between the subgenome graphs suggest biased gene retention.

a

common GO enrichment for *P. australis* syntelogs



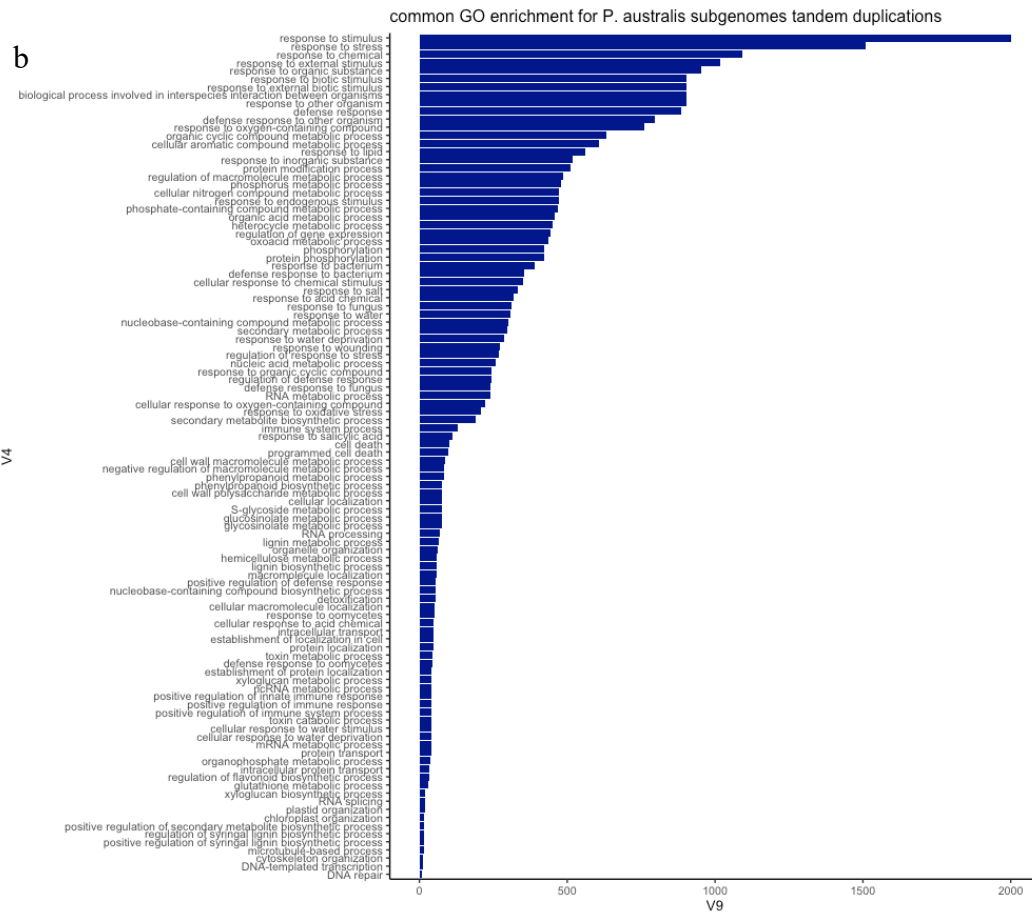


Figure S11. Significantly enriched GO terms shared between subgenome A and subgenome D for the genes originating from **(a)** whole genome duplications and **(b)** tandem duplications. The statistics showed their enrichment in subgenome A.