

<https://doi.org/10.1038/s42003-024-06660-1>

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



Cui Wang^{1,2}, Lele Liu¹, Meiqi Yin¹, Bingbing Liu³, Yiming Wu¹, Franziska Eller⁴, Yingqi Gao³, Hans Brix⁴, Tong Wang⁵, Weihua Guo^{1,7}✉ & Jarkko Salojärvi^{2,6,7}✉

Biological invasions pose a significant threat to ecosystems, disrupting local biodiversity and ecosystem functions. The genomic underpinnings of invasiveness, however, are still largely unknown, making it difficult to predict and manage invasive species effectively. The common reed (*Phragmites australis*) is a dominant grass species in wetland ecosystems and has become particularly invasive when transferred from Europe to North America. Here, we present a high-quality gap-free, telomere-to-telomere genome assembly of *Phragmites australis* consisting of 24 pseudochromosomes and a B chromosome. Fully phased subgenomes demonstrated considerable subgenome dominance and revealed the divergence of diploid progenitors approximately 30.9 million years ago. Comparative genomics using chromosome-level scaffolds for three other lineages and a previously published draft genome assembly of an invasive lineage revealed that gene family expansions in the form of tandem duplications may have contributed to the invasiveness of the lineage. This study sheds light on the genome evolution of Arundinoideae grasses and suggests that genetic drivers, such as gene family expansions and tandem duplications, may underly the processes of biological invasion in plants. These findings provide a crucial step toward understanding and managing the genetic basis of invasiveness in plant species.

Invasive species introduced from exotic regions pose a threat to indigenous ecosystems, and as one of the major drivers of biodiversity loss they can lead to ecological disasters and significant economic losses^{1–3}. The Anthropocene has seen an accelerated spread of invasive species, mediated by the ever-increasing global trade over the last centuries⁴. According to a recent estimate, there are currently more than 3500 harmful invasive alien species and the global economic cost due to alien introductions exceeded 423 billion USD in 2019; considerably more than, for example, the global rice market size, estimated at 287 billion USD in 2021⁵. How and why some species become invasive while others undergoing similar introductions do not is still a conundrum. Certain predisposing phenotypic and physiological traits can be identified, such as short generation time, high growth rate, and high fitness^{6–8}, but from the evolutionary perspective the genome restructurings that lead to invasiveness should be further investigated. A recent genomics study on *Ambrosia artemisiifolia* (common ragweed) suggested the escape of

microbial pathogens, together with introgression from a closely related species to have played a role in its invasiveness⁹. Besides introgression, other hypotheses include the accumulation of transposable elements that may bring in extra genome structural variations and genetic diversity, thus contributing to the fast adaptation to new environments observed in invasive species¹⁰. Another factor commonly suggested to contribute to the invasiveness is polyploidy¹¹, since a higher level of genomic plasticity in polyploids could allow a wider range of habitats and more resilience^{12–14}.

Common reed (*Phragmites australis*; Fig. 1a) is a dominant grass species in wetland ecosystems which occupies large coastal areas all around the globe. Apart from its well-known role in water purification and carbon fixation, *P. australis* is also known for its uses in traditional Chinese medicine¹⁵. As a species with a wide global distribution, the common reed exhibits a high level of intraspecific variation in its functional traits, such as shoot height, shoot density, and flowering occurrence. Genetically, it has

¹Key Laboratory of Ecological Prewarning, Protection and Restoration of Bohai Sea, Ministry of Natural Resources, School of Life Sciences, Shandong University, Qingdao, PR China. ²Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland. ³Institute of Loess Plateau, Shanxi University, Taiyuan, China. ⁴Department of Biology, Aarhus University, Aarhus, Denmark. ⁵College of Landscape Architecture and Forestry, Qingdao Agricultural University, Qingdao, China. ⁶School of Biological Sciences, Nanyang Technological University, Singapore, Singapore. ⁷These authors jointly supervised this work: Weihua Guo, Jarkko Salojärvi. ✉ e-mail: whguo@sdu.edu.cn; jarkko@ntu.edu.sg

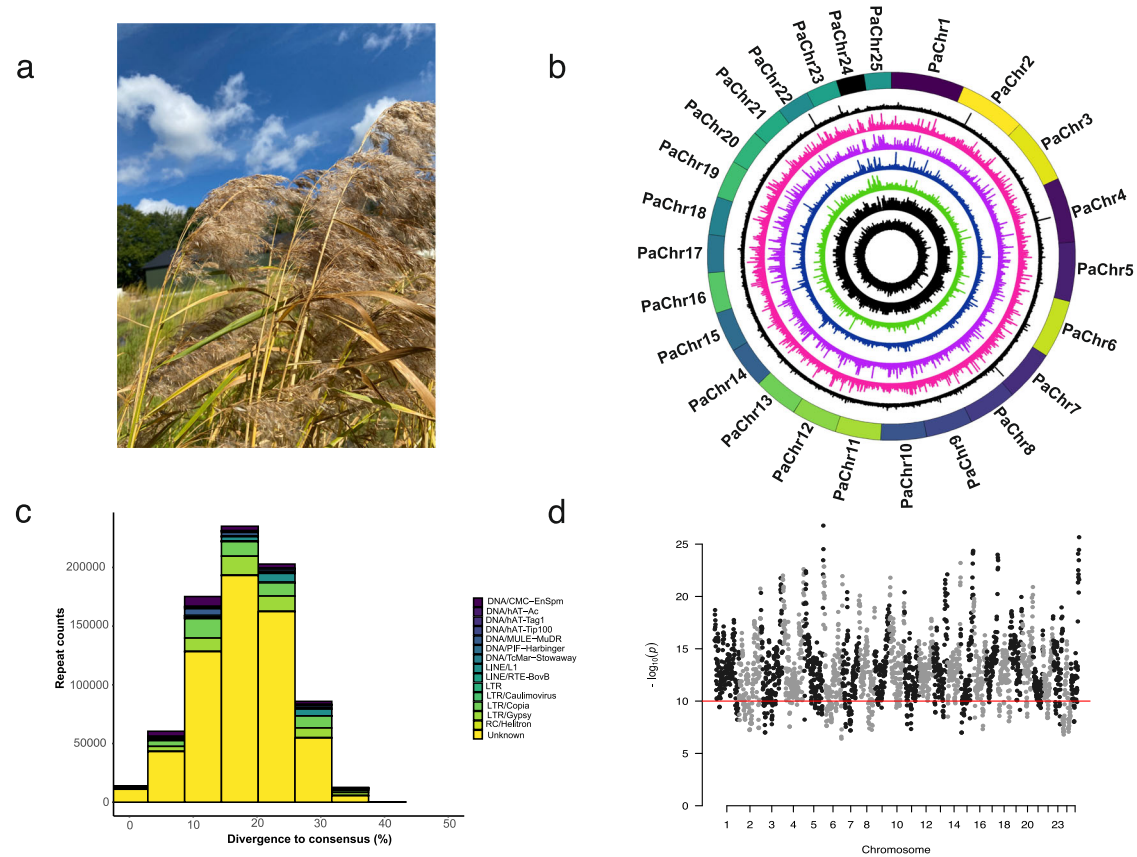


Fig. 1 | Genome features of the chromosome-level genome assembly of *Phragmites australis*. Read coverage distribution, repeat content, and LTR assembly index. **a** Photograph of *Phragmites australis* at a common garden in Denmark (Photo: C. Wang). **b** Circos plot of the 25 pseudo-chromosomes in *P. australis*. The outermost circle indicates the length of each chromosome. The histogram tracks, arranged from the outer to the inner circle represent the read coverages for CN, USnat (Y17), USland (Y7), Med (Y21), and EU invasive lineages estimated from mapping illumina short reads against the reference genome. The last two tracks

illustrate the gene and repeat contents, respectively. Visualization was conducted from an analysis using a sliding window of 10,000 bp. **c** A stacked bar plot of the different repeat types, split according to their divergence from the consensus repeat sequence (x-axis). The major peak shows that a substantial proportion of unknown repeats have similar divergence times, potentially originating from one single expansion event. Expansion events of LTR and LINE repeats may have occurred at different time points. **d** LTR Assembly Index (LAI, y-axis) for the 25 chromosomes of the reference genome.

been classified into seven phylogeographic lineages based on chloroplast and nuclear genes. These lineages encompass the indigenous populations in North America, the Mediterranean region, Europe, Northern China, Southern China and Australia, South Africa, as well as the US land type located at the Gulf of Mexico¹⁶. Common reed is also a super colonizer that shows high adaptability to various environmental conditions^{17–19}. Although there are distinct native populations in Europe and North America, the European lineage has become highly invasive in North America after its introduction to the continent approximately 200 years ago, having expanded its range from an initially limited area to span almost the entire United States. During this process, the invasive European lineage has been slowly replacing the native North American populations. Interestingly, while the European lineage is considered invasive in the United States²⁰, it does not show these characteristics in Europe. In comparison to the native populations in North America, the invasive populations appear to have greener stems, more hairy ligules, and denser seed distribution on the inflorescence²¹. Several factors have been suggested to contribute towards the evolution of invasiveness in *P. australis*, including extensive migration, favorable new environments, asexual propagation, and specific genome evolution²². Also, hybridization has been suggested, as introgression between different *Phragmites* species and lineages are often observed, contributing significantly to species diversity^{23,24}. From the genomic perspective, the invasive population has been found to possess a larger genome size²⁵ than its ancestral populations. Increased genome size could result from (segmental) genome duplications, introgression, hybridization, or

accumulation of transposable elements, which may have led to genomic changes resulting in the evolution of an invasive lineage. A draft genome of the invasive lineage was recently published²⁶, but the lack of a chromosome-level assembly and the absence of an assembly from a non-invasive lineage hindered the identification of the specific genome adaptations.

Similar to rice (*Oryza sativa*), the basal chromosome count in genus *Phragmites* is $n = 12$. In *P. australis*, the euploid chromosome count varies, ranging from 3x to 12x²⁷. Tetraploids and octoploids are most common in nature, but mixoploid and aneuploid lineages are also encountered^{27,28}. Possibly because of this, also a great extent of intraspecific variation has been documented in terms of morphological traits and molecular markers, and at least three subspecies have been suggested: *P. australis* ssp. *australis*, *P. australis* ssp. *altissimus*, and *P. australis* ssp. *americanus*^{29,30}. The versatile ploidy levels in this species allows an opportunity to investigate the potential role of polyploidization in invasiveness.

Here we present a high quality chromosome-level assembly of *P. australis* from PacBio HiFi sequencing. The high-quality reference was used to scaffold four individuals, representing global variation of the species, into chromosome level. The set of genome assemblies was then used to study the genome evolution and demographic histories in *P. australis* lineages to understand the genetic drivers underlying its invasiveness. More specifically, we tested two hypotheses on invasiveness in common reed. First, we evaluated whether the increased genome content in the invasive lineage was involved in proliferation. Second, we tested whether asexual reproduction has facilitated the population growth of common reed and increased

Table 1 | Genome assembly statistics for the *P. australis* accessions analyzed in this study

<i>P. australis</i> lineage	CN	Y17 (USnat)	Y21 (Med)	Y7 (USland)	US invasive (Oh et al.) rev*	US invasive (Oh et al.)
Assembled genome size (bp)	920,401,352	922,056,673	814,995,062	1,217,674,372	866,025,734	1,139,927,050
Largest contig size (bp)	55,032,922	608,995	506,639	421,189	3,219,705	3,219,705
Smallest contig size (bp)	4080	150	150	150	3741	1771
N50 contig length (bp)	34,219,654	78,980	52,722	35,489	279,365	194,574
Total number of contigs	1313	19,648	27,190	58,654	5335	13,411
Sequencing platform	PacBio HiFi + HiC	Illumina WGS	Illumina WGS	Illumina WGS	PacBio RSII	PacBio RSII
Percentage of Ns (%)	0	4.98	5.34	3.7	2.09	-
GC (%)	43.94	44.15	43.76	43.62	44.03	-
Number of protein-coding gene loci	42,498	40,688	40,969	40,124	46,625	64,857
BUSCO genome (poales_odb10)						
Complete gene models	98.9	98.1	97.4	92.5	95.9	97.2
Single-copy	47.2	48.8	51.5	58.0	55.6	41.9
Duplicated	51.7	49.3	45.9	34.5	40.3	55.3
Fragmented	0.1	0.6	0.9	1.9	1.3	0.9
Missing	1.0	1.3	1.7	5.6	2.8	1.9
Busco annotation (poales_odb10)						
Complete gene models	98.2	96.6	95.9	90.3	88.7	92.0
Single-copy	46.9	51.8	53.0	61.0	59.7	49.2
Duplicated	51.3	44.8	42.9	29.3	29.0	42.8
Fragmented	0.4	1.0	1.6	3.2	3.9	2.9
Missing	1.4	2.4	2.5	6.5	7.4	5.1

The seventh column provides the statistics for the published draft genome of an invasive lineage, and the sixth column shows the statistics for the published draft genome of the invasive lineage when haplotigs were removed.

deleterious load in the species. This comprehensive study gives deep insights to genome evolution in grass species, and particularly it suggests the specific changes underlying invasiveness, not only in *P. australis*, but possibly also other grasses in general.

Results and discussion

Chromosome-level genome assemblies of *P. australis* lineages

Flow cytometry-based estimate of the haploid genome size of the reference *P. australis* (Chinese lineage, denoted as CN) selected for genome assembly was 0.92 Gb (Supplementary Fig. S1), with cytogenetic analysis on cell metaphase revealing $2n = 50$ chromosomes. PacBio HiFi sequencing yielded a 61-fold coverage of the predicted genome size with an average read length of 16 kb. Contig-level assembly from Hifiasm was 920 Mb (Table 1), with an N50 of 34.22 Mb. Scaffolding with chromosome conformation capture (Hi-C) data allowed us to map 94.5% of the assembly into 25 pseudomolecules (Fig. 1b; Supplementary Fig. S2; Table 1), and subsequent gapfilling resulted in gap-free pseudochromosome assemblies. Telomeres were detected at both ends of all pseudomolecules, suggestive of complete telomere-to-telomere assembly for all chromosomes (Supplementary Fig. S3). Genome completeness assessed by quantifying the presence of universally conserved single copy genes (BUSCO v5 with Poales odb10 database) suggested a high-quality assembly with 98.9% completeness (Table 1). An intrachromosomal inversion on Chr 22 was detected by phased haplotypes and Hi-C contact maps (Supplementary Fig. S4).

Annotation using a hybrid pipeline³¹ combining ab initio predictions, conserved proteins in model species, as well as RNA-seq data yielded 42,498 gene model predictions, with 98.2% BUSCO completeness (Table 1). Altogether 42,369 genes (99.7%) were anchored to the 25 chromosomes (Fig. 1b). Repeat modeling annotated 63.95% of the genome as repeats, with *Gypsy* (10.73%) and *Copia* (7.18%) being the most abundant types (Fig. 1c, Source data <https://doi.org/10.6084/m9.figshare.26103652>, Supplementary Data 1). We additionally identified 851 tRNA on the 25 chromosomes, with

the highest number detected on Chr 24 (74 genes). Additionally, we predicted 3749 rDNA genes, with most of them located on Chr 24 (776 genes), Chr 17 (579 genes), Chr 20 (403 genes) and Chr 16 (317 genes).

To assess the genomic diversity in *P. australis* we conducted further sequencing on three representative genetically distinct lineages using Illumina short reads: the so-called US land type occurring in Gulf Coast (USland; Y7)²³, native US lineage limited to the north of United States (USnat; Y17)²⁰ and Mediterranean lineage (Med; Y21)³². The draft assemblies exhibited sizes comparable to the chromosome-level reference; 922 Mb for the USnat lineage (Y17), 815 Mb for the Med lineage (Y21), and 1.2 Gb for the USland lineage (Y7). The USland lineage has been hypothesized to be a hybrid between *P. australis* and *P. mauritius*, which would explain its larger genome size²³. Quality assessment using universally conserved single-copy genes (BUSCO v5.2.1) yielded values similar to the high-quality reference (92.5–98.1% completeness; Table 1), with equal proportions of duplicated genes across all assemblies (34.5–51.7%). These figures were again similar to the number of duplicated genes in the reference and their original contig-level draft assemblies (Supplementary Data 2). Altogether, this suggests that the draft assemblies were of similar quality and accurately captured the gene coding space. Additionally, we included the previously published draft genome of the invasive lineage from North America (EUinvasive)²⁶ in our analyses. Using the CN genome as a reference, we proceeded to scaffold these four draft assemblies to chromosome-level pseudomolecules. Based on homology, we successfully mapped 94.4–97.1% of the contigs to pseudochromosomes for the draft assemblies of the three main lineages, and 79% for the USland lineage (Supplementary Data 3, Supplementary Data 4). The scaffolded assemblies displayed significantly higher levels of repeat content compared to the contig-level assemblies. As LTR retrotransposons are good indicators of assembly continuity and completeness³³, we assessed the LTR Assembly Index (LAI). The reference genome yielded a relatively high value of 12.71 (Fig. 1d, Source data <https://doi.org/10.6084/m9.figshare.26104750>) while the LAI for the chromosome-

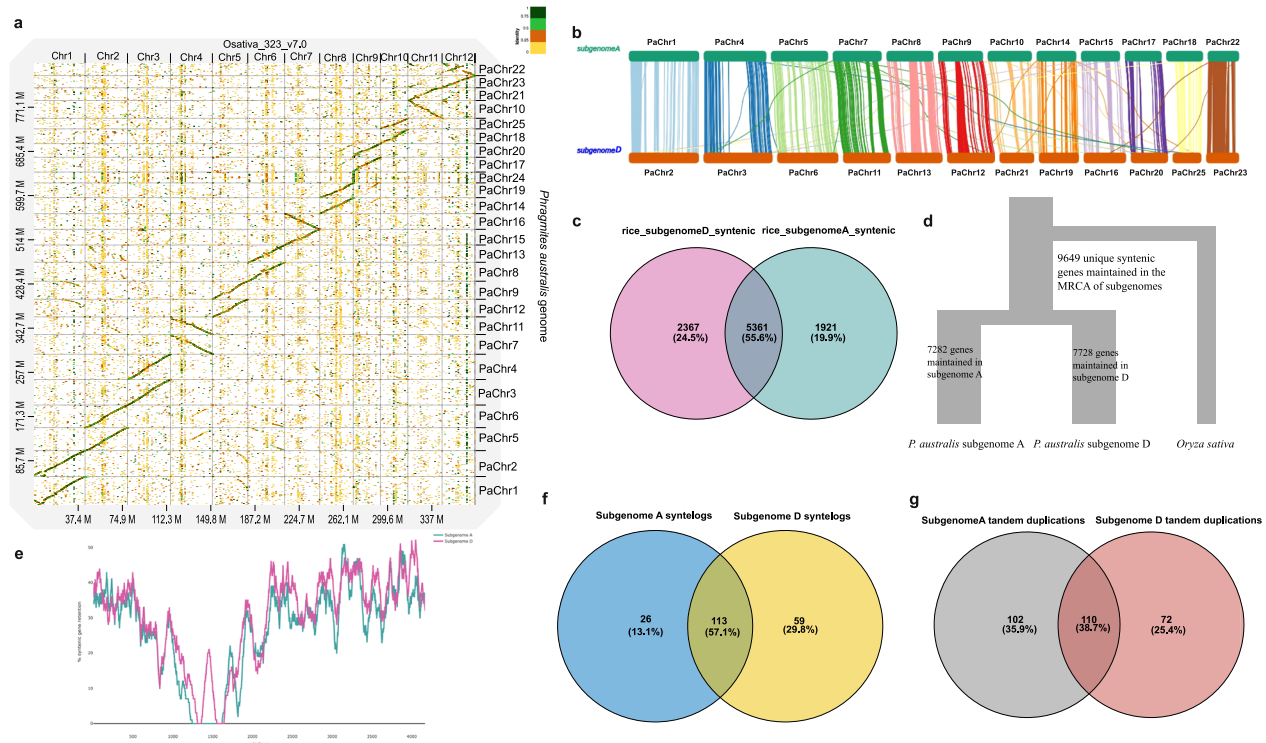


Fig. 2 | Subgenome synteny, structural rearrangements, gene retention, fractionation bias and tandem duplications in *Phragmites australis*. **a** Dotplot of the syntenic regions between *P. australis* and rice (*O. sativa*). The plot shows that two homeologous chromosomes of *P. australis* align to one rice chromosome, representing the two subgenomes of *P. australis*. **b** Large-scale synteny and structural rearrangements between subgenomes A and D of *P. australis* reference genome, with a minimum alignment length of 10,000 bp. **c** Venn diagram showing the number of shared syntenic genes with rice in subgenomes D (left) and A (right). **d** A comparison of the numbers of syntenic gene pairs (syntelogs) with rice illustrates

differential loss rates between the two subgenomes. **e** Fractionation bias of subgenome D over subgenome A in *P. australis* chromosomes homeologous to rice Chromosome 5. The line chart shows the number of retained syntenic genes computed using a sliding genomic window of 100 genes for homeologous chromosome pairs in *P. australis* and *O. sativa japonica* genome assemblies. **f** A Venn diagram illustrating the number of significantly enriched GO terms and their overlap among subgenome A and D syntelogs resulting from whole genome duplications. **g** A Venn diagram illustrating the number of significantly enriched GO terms and their overlap among subgenome A and D tandemly duplicated genes.

level USnat, Med, USland and EU invasive lineage assemblies was 7.07, 6.97, 6.72, and 9.47, respectively. Since the EU invasive lineage was initially assembled using PacBio long reads, the scaffolding resulted in a higher-quality assembly compared to the contig-level draft assemblies.

Subgenome phasing reveals subgenome dominance patterns in *P. australis*

Syntenic alignment of the HiFi reference assembly against *O. sativa* revealed a structure of duplicated syntenic blocks, implying the presence of two subgenomes within *P. australis*, characterized by remarkably conserved synteny (Fig. 2a). Subphaser³⁴ partitioned the genome into subgenomes denoted as A and D, each consisting of 12 pseudochromosomes (Supplementary Fig. S5). Although synteny between the two subgenomes was highly conserved, some large chromosome inversions and minor rearrangements were still identified (Fig. 2b; Supplementary Fig. S6). Out of a total of 42,396 genes on pseudomolecules, 21,152 were assigned to subgenome A, and 20,931 to subgenome D, respectively. The histogram of synonymous mutations (Ks values) between syntenic gene pairs (syntelogs) from self-self alignments for each of the subgenomes indicated three whole genome duplication events in their ancestry, with highly conserved, large syntenic blocks from the most recent event (Supplementary Fig. S7). Compared against the Ks spectrum from syntenic alignments against rice, the modes of the Ks peaks suggested that the events predate the divergence with rice, indicating them to be the pan-Poales *rho*, *sigma* and *tau* WGD events³⁵. Assuming that *rho* event occurred at ~70Mya³⁶ and using this event to estimate a mutation rate, the Ks peak resulting from the syntenic alignment of the two subgenomes suggests their divergence at ~30.9 Mya (recent Ks peaks at 0.2; Supplementary Fig. S8).

Following the allopolyploid event, during the subsequent diploidization process, the genome undergoes fractionation, loss of redundant genes. This loss can occur either in an unbiased random manner, where neither of the genomes dominates the process³⁷, or it can be biased towards one of the subgenomes, resulting in subgenome dominance^{38,39}. In the case of *P. australis*, we observed a significant subgenome bias among the syntelogs counts between rice and the two subgenomes (Chi-square test: $p = 4.471 \times 10^{-07}$). In total, 2367 rice syntelogs from subgenome D were lost in subgenome A, while, in contrast, 1921 rice syntelogs present in subgenome A were lost from subgenome D. Altogether, we identified 5361 syntelogs retained in both subgenomes (Fig. 2c, d). The same asymmetry was reflected in the BUSCO genes, as a total of 79.6% of single copy genes from the Poales database were identified in subgenome A, while 84.7% were identified in subgenome D, respectively. This suggests that subgenome D dominated over A during genome fractionation subsequent to the allopolyploid event (Supplementary Fig. S9; Table 1). The bias towards subgenome D was more emphasized in nine out of 12 chromosomes in terms of fractionation patterns as well as the numbers of retained syntelogs with rice (Fig. 2e, Supplementary Fig. S10). We next analyzed Gene Ontology (GO) term enrichments for the genes originating from whole genome duplications versus tandem duplications to find out whether these genes are involved in biological processes having adaptive potential. Following the dosage-balance hypothesis⁴⁰, the WGD syntelogs were mostly enriched for regulatory activities such as regulation of biological process, various metabolic processes, regulation of gene expression, developmental process, and biosynthesis in both subgenomes (Fig. 2f, Supplementary Fig. S11, Supplementary Data 5). The accumulation of tandem duplications is an evolutionarily more rapid mechanism and, accordingly, these genes were

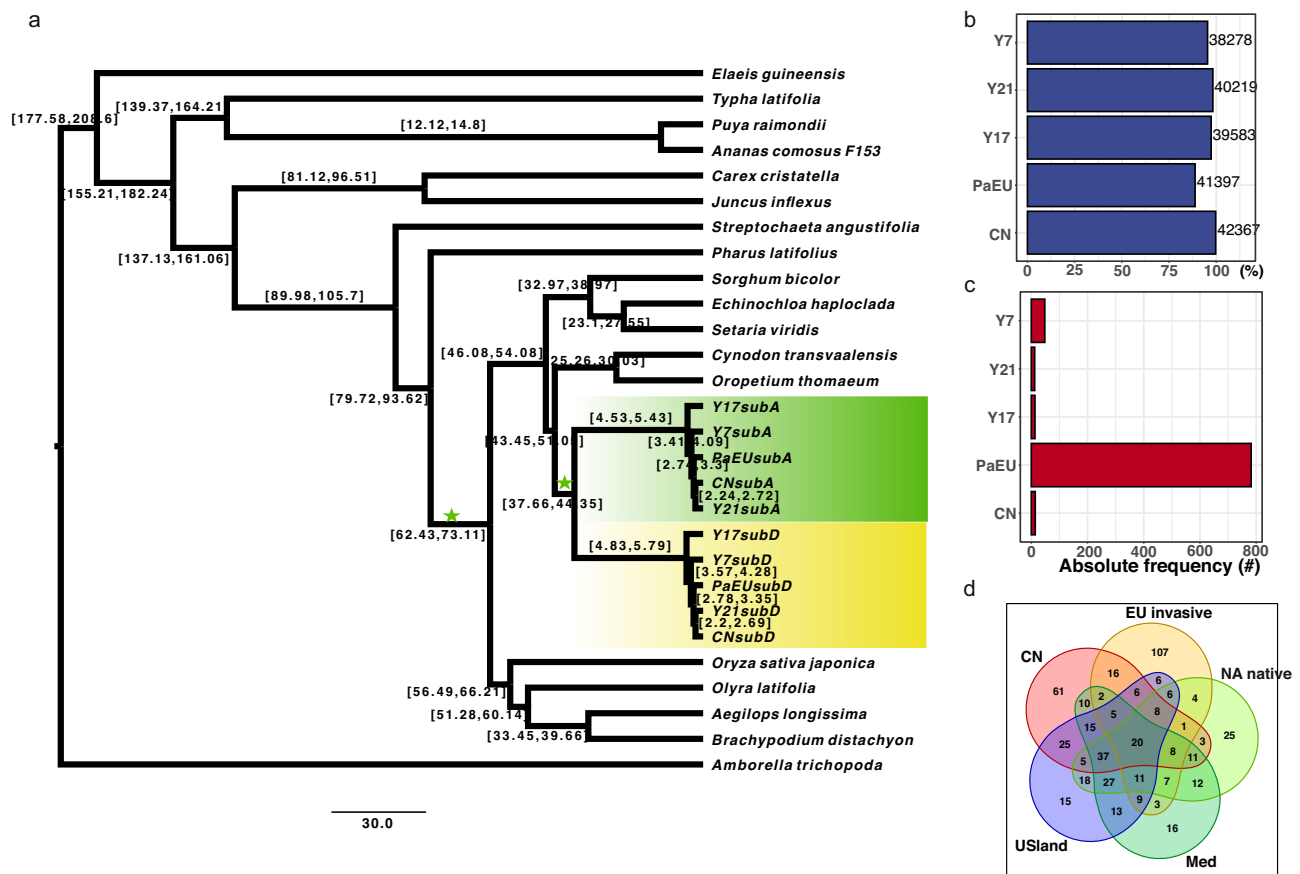


Fig. 3 | Comparative genomics of *Phragmites australis*. Analysis of subgenome divergence, phylogenetic relationships, distribution of orthologous genes, and patterns of tandem duplications. **a** A phylogenetic tree estimated from 189 single copy BUSCOs of 23 species, with *Amborella trichopoda* as an outgroup. The subgenomes of each *P. australis* lineage were fully phased and used separately in estimation. The 95% confidence interval (CI) is shown in the square brackets along each of the branches, with Million Years ago (Mya) as unit. **b** The proportions (x-axis) and the

numbers of genes (next to bar plots) clustered into orthogroups for the five *P. australis* lineages. **c** The number of genes in lineage-specific orthogroups for the five *P. australis* assemblies. **d** A Venn diagram illustrating the number of shared and lineage-specific Biological Process gene ontology (GO) terms among tandem duplications in each of the lineages. Unlike other lineages, tandem duplications in the EU lineage were significantly enriched for 14 GO categories, primarily associated with DNA structure maintenance and replication.

mostly enriched for response to stimulus, response to stress and other organisms, as well as defense responses (Fig. 2g, Supplementary Fig. S11; Supplementary Data 6).

Common reed lineages originate from an ancestral hybridization

Genome assembly and population analysis of Arundiacae species has historically been challenging due to high ploidy level variance. To assess phylogenetic placement and divergence times between the subgenomes, we identified 189 orthologous genes present in 19 species as a single copy but duplicated in all *P. australis* assemblies, with one copy per subgenome (Supplementary Data 7). Using *Amborella trichopoda* as an outgroup, the phylogeny estimated from this set grouped the *P. australis* lineages as a monophyletic clade, with the North American lineage sister to all other lineages (Fig. 3a). According to our calibrated trees, with two calibration points including the common ancestor of BOP clade (69.6–81.9 Million years ago, Mya) as well as the divergence time between monocots and dicots (179.9–205 Mya), the two subgenomes of *P. australis* appeared to diverge at an early age (37.7–44.4 Mya). This estimate appears to be much earlier than ~17 Mya reported by Huang et al.⁴¹ who used several additional secondary calibration points and relatively similar to the ~30.9 Mya obtained with Ks peak comparisons above. The differences may stem from the conservative selection of calibration points in our data and a fixed molecular clock, varied species diversification rate among branches that could affect phylogenetic estimates, or purifying selection acting on the highly conserved BUSCO genes while being more relaxed with the newly duplicated genes⁴². Both

subgenomes yielded concordant phylogenies for the *P. australis* lineages, following the divergent order of USnat, USland, EUinvasive, CN and Med lineages (Fig. 3a), with the lineage splits occurring during the last ~5 Mya.

Gene families involved in DNA replication are expanded in the invasive lineage

To investigate gene family evolution specific to the invasive lineage, we clustered the proteomes of the five representative *Phragmites* assemblies using Orthofinder. The assemblies yielded similar numbers of genes in the orthogroups (Fig. 3b), with the EUinvasive lineage showing the highest number of lineage-specific orthogroups and duplications (Fig. 3c; Source data: <https://doi.org/10.6084/m9.figshare.26356390>). The genes encoding the proteins in the EUinvasive-specific orthogroups were enriched for multiple biological processes including telomere organization, telomere maintenance, DNA geometric change, DNA replication, chromosome organization, recognition of pollen, DNA metabolic process and DNA biosynthetic process (Supplementary Data 8). These processes are closely related to proliferative activities such as the growth rate, the accumulation of biomass, and the balanced allocation of energy to sexual and asexual reproduction of common reed.

The gene family expansions as such provide no direct evolutionary interpretation of the adaptive potential in the invasive lineage, and furthermore, the results could be susceptible to the particular clustering approach used. Dosage increases in the form of tandem duplications have been proposed as the general rapid evolutionary mode of adaptation^{43,44}, and

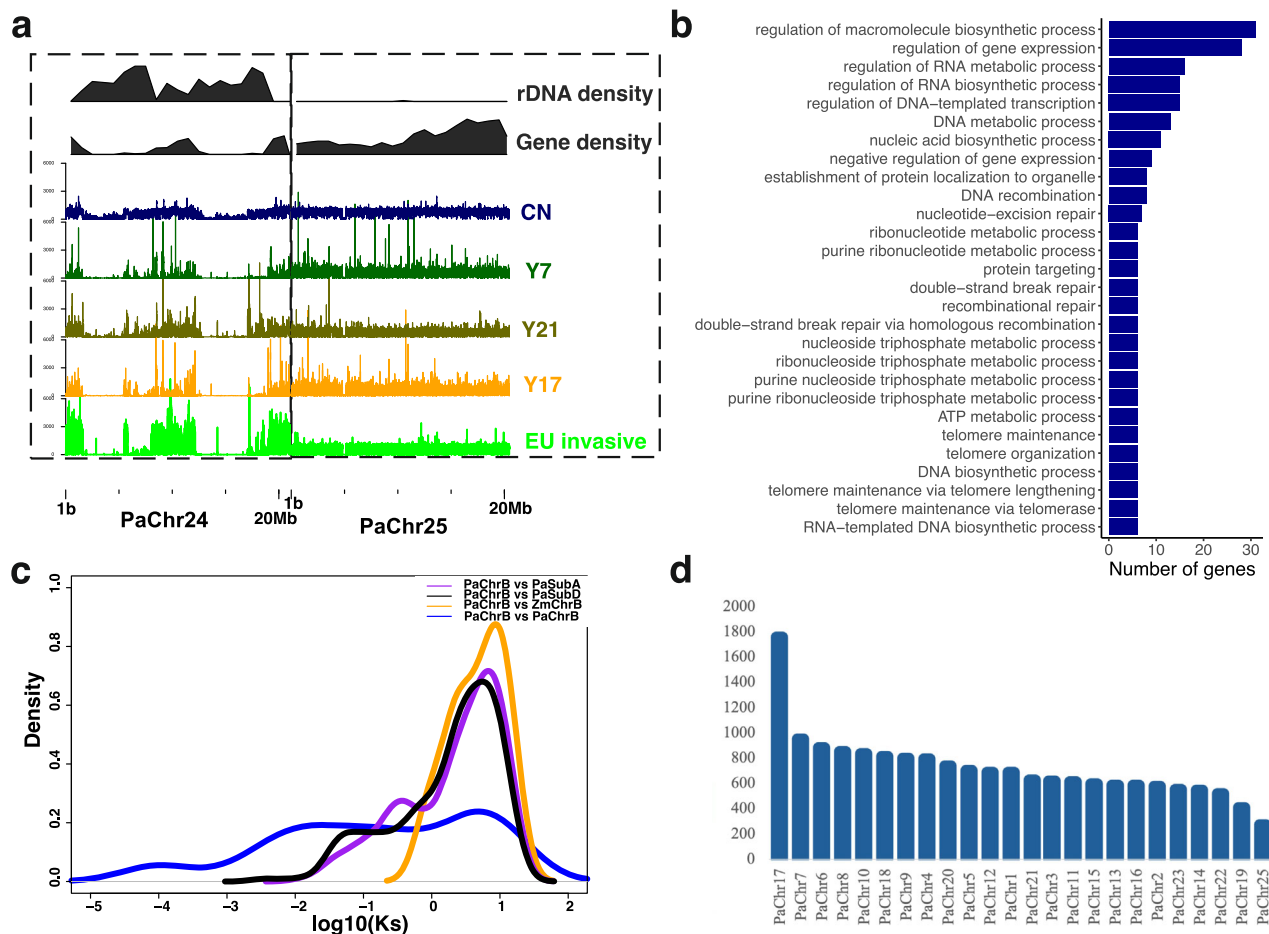


Fig. 4 | Genomic insights into the B chromosome of *Phragmites australis*.

a Sequencing coverage over B chromosomes. The tracks (from top to the bottom) illustrate the rDNA density, gene density, followed by average read coverage on CN, USland (Y7), Med (Y21), USnat (Y17), and EU invasive individuals, when mapped on the B chromosome reference (Chr 24; left) or a non-B chromosome Chr 25 (right window). Dashed line indicates the borders of the chromosomes, while the y-axis illustrates read coverage. The invasive individual exhibits notably higher read coverage on the B chromosome compared to Chr 25. **b** The Gene Ontology (GO) terms significantly enriched among the genes on the B chromosome. The enrichment

analysis was carried out for GO Biological Processes categories, and the significances were adjusted for multiple comparisons using Bonferroni correction. **c** Density plot of the synonymous substitution rate (Ks) distributions of the syntenic regions between the B Chromosome and subgenomes A (purple) and D (black) of the main genome. Additionally, the panel shows Ks plots of *P. australis* B chromosome against the maize B chromosome (gold) and B chromosome against itself (blue). **d** A bar plot of the numbers of homologous fragments (300 bp size) between the B chromosome and *P. australis* main genome.

therefore we carried out a complementary analysis for the tandemly duplicated genes to test the hypothesis. Overall, the highest amount of tandem duplicates were identified in the EUinvasive lineage (13,114 genes), followed by the reference CN assembly (7572 genes), while the short-read assemblies demonstrated somewhat lower numbers (5657, 5950, 6162 duplicates), likely due to their lower contiguity. However, these sets were clearly large enough and of comparable sizes to allow comparison of process-specific expansions between the lineages. A majority of the enriched gene ontology categories demonstrated overlap across the lineages, but we additionally detected 14 GO processes comprising 255 genes that were specific to the invasive lineage and, intriguingly, most were related to DNA replication and telomere maintenance (Fig. 3d, Supplementary Data 9). Tandem duplications in plant species have been generally found to play a role in environmental responses⁴³, whereas in invasive common reed there seems to be a greater emphasis on maintaining DNA integrity.

To test whether the gene family expansions detected in the clustering approach were due to hybridization or adaptation in the form of tandem duplications we tested the overlap between the two approaches. The overlap was significant ($p < 3.7 \times 10^{-62}$; Fisher Exact test) with 42.5% of the expanded genes being present in tandem duplications. Altogether, the results suggest that lineage-specific gene family expansions are associated with the well-

documented increased proliferation in the invasive lineage⁴⁵ and that they at least partly result from rapid adaptation in the form of increased dosage through tandem duplications. Considering the young evolutionary age of the invasive lineage this expansion has been relatively rapid.

Phragmites australis has a large B chromosome

Interestingly, karyotyping of the reference individual identified 25 diploid chromosomes instead of the expected 24, leading us to suspect that the reference individual possessed a B chromosome. While the assembly size of Chr 24 was relatively large in the reference CN lineage (21 M), its size was much smaller in the other assemblies (Supplementary Data 4). We speculated that this discrepancy might be due to high repeat content that gets assembled accurately with the PacBio HiFi, while the other accessions would either contain collapsed assemblies (due to platform limitations) or no assembly at all due to the absence of B chromosome in those individuals. When the short reads were mapped against the scaffolded chromosome-level assembly from the same individual, the read coverage on Chr 24 was consistently higher than on other chromosomes, suggesting collapsed contigs in these draft assemblies. Mapping against the reference assembly largely resolved these issues, as the coverages became more similar to the overall genome coverages (Fig. 4a). Elevated coverage was still observed in

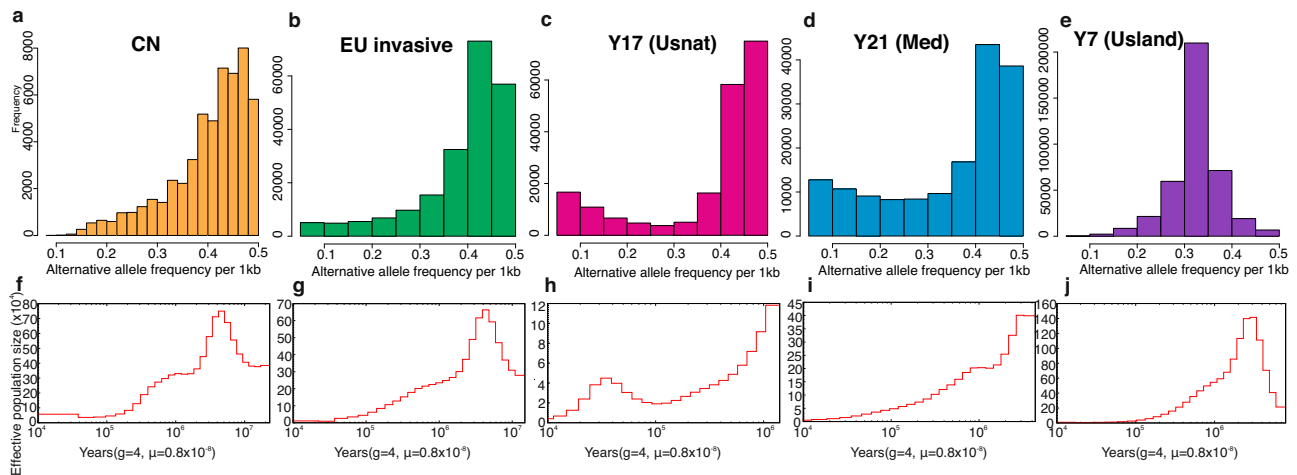


Fig. 5 | Ploidy levels and historical effective population sizes of *Phragmites australis* accessions. The figures in the upper panel show the inferred ploidy levels for each of the accessions. The estimates were obtained by aligning short-read sequencing data against the reference genome. Geographic lineages CN (a), EU invasive (b), USnat lineage (Y17; c) and Med lineage (Y21; d) were inferred to be allotetraploid with a minor allele frequency (MAF) distribution peak at 0.4–0.5. The

MAF for the USland type of Gulf Coast (Y7; e) was 0.3–0.35, indicating that this individual is allohexaploid. The lower panel shows historical effective population sizes for these accessions (f: CN, g: EU invasive, h: Y17, i: Y21, j: Y7). The demography was estimated from short-read sequences mapped to their respective assemblies and using Pairwise sequential Markovian coalescent (PSMC) model.

the North American invasive lineage, suggesting aneuploidy for this chromosome, while USnat and USland demonstrated lower coverages.

A closer analysis of Chr 24 revealed high repeat contents (89.8%), along with a redundant accumulation of non-coding RNA genes such as rDNA (776 genes) and tRNA (74 genes), and a lack of large syntenic blocks with any of the other chromosomes in *P. australis*, hence we may conclude it to be a B chromosome⁴⁶. In *Arabidopsis* the length of rDNA clusters can reach up to millions of base pairs^{37,47}, and copy number variation in multiple species of hybrid origin such as *Arabidopsis*³⁷, strawberry⁴⁸, and synthetic wheat⁴⁹ for these loci have been identified. Therefore, the apparent copy number variation of B chromosome among *P. australis* lineages may also be confounded by variation in rDNA amounts. However, an in-depth analysis of the read alignments showed mappings on the B chromosomes mainly on gene rich regions rather than on rDNA regions, suggesting that the variations in coverage were not due to rDNA (Fig. 4a). To estimate the B chromosome copy number in the different accessions we calculated the ratio of averaged read coverage on B chromosome versus on all other chromosomes, identifying a much higher value for the invasive lineage (2.29), and medium value for CN (0.93) and Med lineages (1.01), but only 0.36 for USnat (Y17) and 0.29 for USland (Y7). This suggests that the sequenced invasive accession likely contains a higher number of B chromosomes (2x) than the reference, while the low ratio in Y17 and Y7 suggests that B chromosome may not exist in these individuals.

B chromosomes have been reported to evolve by accumulating genes from A chromosomes (with A denoting the standard chromosome set), as gene fragments or DNA homologous to the A chromosomes have often been detected⁵⁰. In total, 286 protein coding genes were predicted on the B chromosome. These genes were enriched for biological process GO terms such as fusion of sperm to egg plasma membrane involved in double fertilization forming a zygote and endosperm, chromosomal segregation during cell cycles, telomere maintenance, and associated biochemical activities such as ATP synthesis and metabolism, protein localization, DNA synthesis and repair, attachment of spindle microtubules to kinetochore in mitotic and meiotic cell cycles (Fig. 4b; Source data: <https://doi.org/10.6084/m9.figshare.26356621>), potentially contributing to increased proliferation.

To study whether the genes in B chromosome were expressed in the different accessions, we reanalyzed the transcriptomic data from Oh et al.²⁶. A mapping of RNAseq reads onto the reference genome suggested that genes on the B chromosomes were expressed in all tissues of all individuals. However, since the number of B chromosome copies in the individuals used

for RNAseq was not known, the presence could also be attributed to recent duplications of homologous genes on other chromosomes. We found the highest number of unique transcripts expressed in the shoot tissue of the invasive lineage (Supplementary Data 10), supporting the possible higher copy number at least in the invasive lineage.

To determine the duplication modes of B chromosome genes, we identified regions that were syntenic with subgenomes A and D and determined their synonymous substitution rates (Ks). When filtering for genes with high Ks values (Ks>5) to exclude artefacts from synteny alignment or errors from gene prediction, we detected altogether 254 and 284 genes in subgenomes A and D, respectively. A histogram of Ks values suggested the B chromosome duplicates to be of ancient origin, with the peak of Ks values comparable, or possibly even older in time, than the *tau* whole genome duplication event common to monocots. The Ks values displayed a similar peak when aligning this B chromosome against the B chromosome in *Zea mays* (Fig. 4c) with default parameters and then reducing the syntenic block length to one gene. According to *P. australis* self-self alignments the B chromosome had undergone two subsequent, possibly segmental, duplication events (Fig. 4c). Together, this suggests that even though rapidly evolving, the B chromosomes across different species may have a common origin, and the core set of genes has been shaped by repeated transfer of genetic material from the main chromosomes.

To further inspect the origins of the B chromosome outside of the gene coding regions, we divided the chromosome into 300 bp sequence fragments and mapped them against the main chromosomes. Since the unanchored contigs may be part of B chromosome or other chromosomes, we only targeted chromosomal level scaffolds to ensure the accuracy of tracing the origins. Out of the 70,085 fragments, 17,013 mapped primarily, suggesting that at least 24.27% of the B chromosome has been sourced from the main chromosomes. Of these, 1758 sequences mapped to pseudochromosome Chr 17, while 953 sequences mapped to Chr 7 (Fig. 4d; Source data: <https://doi.org/10.6084/m9.figshare.26356660>).

Complicated life history involved in common reed evolution

To assess the impact of polyploidization events on intraspecific species divergence, we estimated the ploidy level of each individual using the ratio of reads supporting the minor allele versus all reads aligned to a locus⁵¹. We followed the logic that in a diploid (or, in this case, an allotetraploid with clearly separating subgenomes) genome, the number of reads supporting the minor allele in a heterozygous locus should be half of the total read

Table 2 | Nucleotide diversities at neutral and deleterious sites in five different *P. australis* accessions

	Nonsynonymous π	Synonymous π	$\pi_{N/S}$
Y17	0.001417	0.001142	1.24
Y21	0.002784	0.003121	0.89
Y7	0.008596	0.009470	0.91
PaEU invasive	0.004491	0.007634	0.59
CN	0.004636	0.013686	0.34

The diversities were estimated separately for each assembly using gene models transferred from the reference assembly.

count. Likewise, in a triploid genome, the proportion of reads mapping to minor allele should be one third. The scaffolded reference assemblies for CN, North American invasive, Med and USnat lineages all displayed minor allele frequency peaks at 0.4–0.5 (Fig. 5a–e; Source data: <https://doi.org/10.6084/m9.figshare.26356684>), suggesting that reads are uniquely mapping to either of the two subgenomes. This observation aligns with the notion of common reed being an allotetraploid, exhibiting two highly differentiated subgenomes. Additionally, a peak at 0.33 was observed for USland, suggesting a hexaploid genome organization. The higher ploidy level is also supported by the 1.5x larger assembly size compared to others and a large proportion of contigs that were not mapped to the pseudomolecules (Table 1).

Next, we modeled demography using the pairwise sequentially Markovian coalescent (PSMC) model, using four years as generation time. PSMC analysis showed consistent decreases in CN, EU invasive and Med lineages (Y21), all starting from around 4 Mya. In USland, the decrease started from around 3Mya, and in USnat from 1–2Mya, with the latter demonstrating a slight recovery at 30–40Kya followed by another decline (Fig. 5f–j).

The decreasing population sizes may be related to biogeographic events in Pliocene, such as glaciation cycles. Another possibility is a transition into asexual reproductive mode. Population genetic simulations predict that compared to outcrossing and selfing lineages, a clonal lineage experiences relaxed selection and thus accumulates more deleterious mutations⁵². Contemporary common reed uses both sexual and asexual reproductive modes. The seed set rate of *P. australis* is very low (mean 9.7–12%) due to its partial self-incompatibility^{53,54} while, in contrast, clonal propagation has been shown to be highly efficient^{55–57}. Due to the lack of recombination, genotypic diversity would be expected to be low in clonal propagation since genotypes that have accumulated a high number of deleterious alleles, and therefore have reduced fitness, will be purged from the population (the so-called Muller's ratchet). Additionally, clonal populations may undergo frequent bottlenecks, as they can colonize habitats with just a few individuals. Altogether, these effects result in effective population decline in the long run⁵⁸. These predictions have been empirically verified in both animals and plants, such as *Primula* species⁵⁹, stick insects⁶⁰, duckweed⁶¹, and wasps⁶². Relaxed selection would result in an increased ratio of non-synonymous to synonymous mutations in asexual and selfing proliferative strategies, but a low ratio in outcrossing species because purifying selection would limit the amount of deleterious alleles in outcrossers⁵⁸.

In support of this hypothesis, nucleotide diversities of deleterious sites (π_n) were low for USnat (Y17; $\pi_n = 0.001417$) and Med Y21 ($\pi_n = 0.002784$), while being higher for USland Y7 ($\pi_n = 0.008596$), EU invasive ($\pi_n = 0.004491$) and CN ($\pi_n = 0.004636$) (Table 2). To assess the proportion of deleterious load we compared their ratios to synonymous diversity and observed high values for USnat ($\pi_n/\pi_s = 1.24$), followed by 0.89 for Med and 0.91 for USland. For EU invasive lineage the proportion was considerably lower, 0.59, and 0.34 for CN. Altogether, the ratios are high compared to other plants⁶³, and can be interpreted to reflect predominantly clonal propagation, since empirical studies and theory have shown that clonal propagation accumulates more deleterious mutations due to relaxed purifying selection, resulting in higher π_n/π_s ratio than among outcrossers^{64,65}. The proportion was highest in USnat population, which has long been isolated

from other lineages. The population statistics for the USnat thus possibly result from a combination of clonal propagation (high π_n , high population level π_s), inbreeding depression (low individual level π_s), and little gene flow from other *P. australis* populations (early divergence in species tree). In addition, it has been found that in North America, the long-distance dispersal of USnat is facilitated by seedlings, but local growth is mainly proceeding through vegetative propagation⁶⁶. Interestingly, the ratios in invasive lineage show a decrease in the proportion of deleterious alleles, suggesting that selection is more active in this population. Therefore, asexual propagation through cloning may accumulate the mutational load, reduce the fitness for the lineages and result in decreased population sizes⁶⁷.

Conclusion

We present a chromosome-level genome assembly for an invasive grass species *P. australis*, and chromosome-level scaffoldings of four draft assemblies representing distinct lineages. This genome assembly marks a major advancement in our understanding of the Arundinoideae species renowned for their complex genomic history⁶⁸. These species may have undergone multiple whole genome duplication events and hybridizations, rendering the tracing of their origins challenging⁶⁸. Utilizing the unique repeat patterns of each parental genome, we phased 24 chromosomes of common reed to two subgenomes. Our analysis revealed biased subgenome fractionation following the allopolyploid event, with subgenome D playing a more dominant role. Phylogeny estimated from phased subgenomes of the five representative lineages and the divergence time estimate from Ks values indicate that the allotetraploid *P. australis* has formed as a hybrid of two ancestral closely related species at around 30.9 million years ago. Hybridizations between *Phragmites* species may still occur and can lead to higher ploidy levels within the species, as evidenced by the hexaploid USland lineage⁶⁹. Therefore, hybridization and polyploidization are essential drivers of genome evolution in grass species, but they are not direct causes of invasiveness, as both North American native and invasive lineages are allotetraploids. Decreasing effective population size and high deleterious load suggests that *P. australis* is going through a transition of reproduction mode from sexual to asexual propagation which is one of the major features in invasive plants.

North American invasive accessions consistently outperform North American native and European native lineages in terms of shoot biomass⁴⁵. While hybridization with a yet unknown *Phragmites* species cannot be ruled out, this difference could also be linked to higher B chromosome copy numbers in the invasive lineage, as observed for the single representative in this study. B chromosomes potentially promote the adaptability of some plant species; for example, individuals with B chromosomes have demonstrated a higher survival rate under stress in *Allium schoenoprasum*⁷⁰ and *Lolium perenne*⁷¹. B chromosomes also affect phenotypic traits and gene expression of A chromosome genes, with a more emphasized effect in the presence of more copies^{72–75}. In maize, thousands of genes on A chromosomes were upregulated in the presence of B chromosomes, while the B chromosome gene expression was positively correlated with its copy number in a dosage-specific manner⁷⁶. However, exceptionally high copy number of B chromosomes may also reduce the fertility and viability of offspring, thus restricting sexual reproduction⁷⁷.

Our study also demonstrated lineage-specific gene family expansions, particularly tandem duplications, as potential contributors to invasiveness; similar effects have been observed in a malignant invasive plant *Heracleum sosnowskyi*⁷⁸. The expanded gene families were mostly enriched for DNA replication and telomere maintenance, suggesting their essential role in maintaining genome integrity as well as DNA synthesis, thereby enabling fast regeneration and proliferation.

In conclusion, we suggest that common reed may have gradually acquired traits facilitating invasiveness, although the observed invasion has occurred within a short time frame. Future controlled experiments are needed to validate the function of these genetic elements in common reed, and to study their associations with physiological traits at the population level. Ultimately, this will contribute to a greater understanding of the decisive factors in grass invasiveness.

Methods

DNA extraction, library construction and whole genome sequencing of the reference individual

For the reference genome sequencing, young leaves were sampled from an individual originated from Yuncheng, Shanxi province. DNA extraction was done using Qiagen DNA extraction kit. One paired-end library was constructed for short-read sequencing on BGI T7 sequencing platform in Hongshu Shanghai limited company, which generated around 230 G (255x) of high-quality data. High Weight Molecular DNA was extracted by Biozeron Biotech Ltd. High fidelity long reads were obtained from Pacific Biosciences Sequel IIe platform, with about 61-fold coverage. Long DNA of the same individual were used for Hi-C library preparation. After crosslink, cut with DpnII enzyme, biotin labeling, ligation, shearing and biotin capture, the DNA fragments were sequenced using Illumina Novaseq platform, with an average depth of 110-fold across the genome. Transcriptomic RNA was extracted from the same individual to help with the prediction of gene models.

Genome size estimation and cytogenetic confirmation

To estimate the genome size of the reference individual, we used flow cytometry method. Briefly, plant tissue was initially digested with pre-cooled MG solution (45 mM MgCl₂·6H₂O, 20 mM MOPS, 30 mM Sodium citrate, 1% (W/V) PVP40, 0.2% (v/v) Tritonx-100, 10 mM Na₂EDTA, 20 µL/mL β-mercaptoethanol, pH 7.5). The tissue was then broken apart thoroughly using sharp blade and placed still on ice for 10 min, followed by filtration using 40 µm diameter mesh to obtain the cell suspension. Add 50 µg/mL propidium iodide and RNAase, then place the cell suspension on ice in dark for staining. We selected *Zea mays* of which the haploid genome content is 2.3 Gbp as an internal reference and used BD FACS calibur flow cytometer to detect the fluoresce intensity of the stained cell nucleus suspension. To count the number of chromosomes of the reference individual, we collected roots and treated it with HCl: ethanol (1:1) under 65 °C for 10 min, wash the treated tissue in deionized water for 5 min, stained the tissue with Carbol fuchsin, squeeze it on a glass slide, and observe the metaphase under microscope.

Assembly of organelle genomes and nuclear reference genome

The Pacbio HiFi reads were first used to perform a de novo assembly of chloroplast genome for *P. australis* using PMAT⁷⁹. The mitochondrial genome was assembled with Flye⁸⁰ and extracted using GetOrganelle⁸¹ based on the embplant_mt database. The resulting circular organellar genomes were visualized and corrected using Bandage software⁸². After removing the Pacbio HiFi reads that were primarily mapping to the organelle genome, the remaining Pacbio HiFi reads were assembled using Hifiasm⁸³, resulting in a primary phased haplotig assembly with genome size 920 Mb in 1350 contigs. The draft genome was then scaffolded to pseudomolecules using HiC contact reads and ALLHiC pipeline⁸⁴. To verify the accuracy of scaffolding, the HiC reads were also mapped to the draft genome using juicer⁸⁵, and the ALLHiC result was then converted to the format that could be visualized in juicebox⁸⁶ for manual correction to improve the contig orders and chromosome boundaries. The final assembly was obtained from the postreview procedure in 3D-DNA pipeline⁸⁷, and finalized with gap closing using TGS-GapCloser⁸⁸.

Genome assemblies of USland, USnat and Med lineages

Based on previous biogeographic lineages identified using chloroplast markers, we selected three individuals (Y7, Y17, Y21) representing USland, USnat and Med lineages, respectively, to be sequenced for whole genome assemblies. Following DNA extraction with CTAB protocol, the genomes were sequenced to 90-fold high coverage (assuming 1 Gb genome size) by NovogeneAIT at Singapore using illumina Novaseq platform. The genome size was first estimated using KmerGenie⁸⁹. Then, after quality assessment of the reads using FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), three draft genomes were assembled using MaSuRCA assembler⁹⁰ with default settings. Assemblies based on short reads may

capture gene coding space to a lower accuracy, for example due to collapsed assemblies or separate haplotig assemblies. To avoid these biases, we purged the draft assemblies for haplotigs, thus obtaining similar coverages across the contigs. The assembly was passed through purge_haplotigs v1.0.4⁹¹ to remove the redundant haplotigs resulting from the highly heterozygous regions. The completeness of the genome assemblies was assessed using BUSCO v5⁹² using odb_database version 10. The quality of the assembly was estimated using Quast⁹³. Next, we scaffolded to pseudomolecules with RagTag⁹⁴, which aligns the contigs against the chromosome-level reference assembly. LTR Assembly Index (LAI) was calculated using LTR_retriever⁹⁵.

Reference genome annotation

Repeat libraries were built with repeatmodeler⁹⁶, and the repeats were masked using RepeatMasker (Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013–2015 <http://www.repeatmasker.org>). Telomeric repeats were detected using telomeric-identifier (<https://github.com/tolkmit/telomeric-identifier>). The tRNA was predicted using tRNAscan-SE⁹⁷. We identified rRNA and non-coding RNA sequences using Infernal⁹⁸.

Gene prediction was performed by combining ab initio prediction, genome homology to close related species, and gene structure prediction from RNAseq evidence. Ab initio gene prediction was performed using GeneMark-ES⁹⁹, BRAKER¹⁰⁰, PASApipeline¹⁰¹ and AUGUSTUS¹⁰². Annotation information of *Sorghum bicolor*, *Setaria viridis* and *O. sativa* were obtained from Phytozome¹⁰³ and used for homology-based annotation using GeMoMa¹⁰⁴. To improve the gene model predictions, we incorporated 20 RNAseq libraries for genome annotation, including 13 RNAseq samples downloaded from NCBI (Supplementary Data 10). Some of these data were originally used to annotate the genome assembly of the EU invasive lineage²⁶. Additionally, we sequenced seven RNAseq libraries representing three different genetic lineages (CN, Med, and South Africa)^{16,69}. These reads were aligned against the reference genome with STAR aligner v2.7.11b¹⁰⁵, and used as evidence in BRAKER¹⁰⁶ prediction. In addition, RNAseq data was also assembled into genome-guided assembly using Trinity¹⁰⁷ and stringtie¹⁰⁸, which was further processed using PASApipeline¹⁰⁹, and TransDecoder (Haas, BJ. <https://github.com/TransDecoder/TransDecoder>) for generating a high-quality dataset for downstream ab initio gene predictions. All the evidence created with different methods were combined using EVidenceModeler v2.1.0¹¹⁰. Functional annotation was performed by aligning the proteins against Uniprot and NCBI nr databases using DIAMOND v2.1.8¹¹¹, with parameters -evalue 1e-5 -max_target_seqs 5. The completeness of predicted gene models was assessed using Benchmarking Universal Single Copy Orthologs v5.1.2¹¹² by searching against poales_odb10 protein database with 4650 conserved single-copy genes. Significance of overlap between duplicated BUSCOs in different genome assemblies were tested with Fisher's exact test using R package "GeneOverlap"¹¹³. Repeat counts and their divergence to the consensus repeat sequences were plotted using repeatR package (<https://github.com/dwinter/repeatR>).

Five genome assemblies and subgenome phasing

With the Chinese assembly (CN) as a reference, we scaffolded the draft assemblies of Y17, Y21, Y7 and the published EU lineage to chromosome level using RagTag⁹⁴. For the reference genome, we detected syntenic blocks among each pair of homoeologous chromosomes by aligning them against the rice genome assembly (*O. sativa japonica* v7.0, downloaded from Phytozome) using minimap2 and visualizing the dotplot created using D-GENIES¹¹⁴. Based on this, we initially phased the largest 25 chromosomes (except Chr 24) to two subgenomes (A, D). The chromosome assignment to each subgenome was then refined using SubPhaser³⁴, which calculates subgenome specific repetitive kmer, with the parameter set -k 13 -q 100 -f 2. The visualization of collinearity was performed using NGenomeSyn¹¹⁵ and Syri¹¹⁶. Accordingly, subgenomes of all the four extra geographic lineages were phased to chromosome level, for each subgenome the completeness was assessed using BUSCO v5.1.2. In addition, the five *P. australis* draft genome assemblies were aligned against *O. sativa* to get the synonymous

substitutions for further investigation of multiple whole genome duplication events. Gene models for the three synteny-based chromosome level assemblies were predicted using GeMoMa pipeline, with the annotation of the reference assembly as a reference. Gene models for the EU invasive lineage was obtained from Oh *et al.* 2022, with haplotigs purged. To evaluate the gene loss along each of the homologous chromosome pairs in the two subgenomes, we performed genome fractionation analysis in CoGe platform using FractBias tool by aligning the *P. australis* reference genome against rice (*O. sativa japonica*). For each rice chromosome, the percentage of gene retention within each window iteration (100 genes) from *P. australis* homologous chromosome pairs was displayed in a graph. We evaluated the bias in syntelog losses between the two subgenomes by selecting the syntelogs with Ks value smaller than 0.7. Significance of subgenome fractionation bias was assessed using R package GeneOverlap¹¹³. The chromosome rearrangements in subgenomes were detected using syri¹¹⁶, and plotted using plotsr¹¹⁷. Tandem duplications and syntenic genes of each subgenomes were analyzed for GO term enrichment using goatools¹¹⁸. To find out lineage specific gene families, we performed all-versus-all blast search for the reference proteome and the predicted proteomes from the four contig-level genome assemblies using OrthoFinder v2.5.5¹¹⁹, and visualized the results using R package cogeqc¹²⁰.

B chromosome copy detection

To analyse the copy number of B chromosomes in *P. australis* lineages, we calculated the read depth within 1 kb window sizes for each chromosome with repeats masked. Several regions were found to have high depth of coverage (Chr14 and Chr 25), peaking in all the five lineages, which may indicate chloroplastic or mitochondrial DNA, or assembly errors. We also masked these regions to avoid noise. We then calculated the ratio of read depth mapped to B chromosome and the average depth of other chromosomes in all the five individuals, assuming the copy number of B chromosome in an individual should be proportional to the ratio of the read coverages. The read depth across chromosome 24 and Chromosome 25 were visualized using R package karyoploteR¹²¹.

Grass genome evolution

To investigate the parental species and allopolyploidization of *P. australis*, we used 192 single copy orthologous genes from BUSCO evaluation to estimate a phylogenetic tree for 23 plant species including the subgenomes of five *P. australis* assemblies. Species *Amborella trichopoda* var. *Santa Cruz* was used as outgroup. The gene trees were inferred by maximum likelihood method using RAxML¹²² with 100 bootstraps, and combined into species tree using Astral-III¹²³. To help with dating the time of allopolyploidization, we estimated divergence time using mcmctree¹²⁴ in PAML¹²⁵, with most common ancestor of BOP clade (69.6–81.9 Mya) as well as the divergence time between monocots and dicots (179.9–205 Mya) as calibration point^{126,127}. Two runs of mcmctree runs were compared to check whether convergence was reached.

Ploidy level prediction

Since the number of alternative alleles other than the reference allele in the genome can reflect the copies of homologous chromosome number¹²⁸, we predicted the ploidy level information using minor allele frequency (MAF) after mapping the reads to the reference genome. By setting the total read coverage to range from 20 to 200, only alleles with coverage higher than 7 were considered as variants. For example, in diploids or allotetraploid in current study, each genomic region has two copies of alleles, the proportion of reads supporting alternative alleles would be around 0.5 in heterozygous sites, and thus the genomic distribution of the proportion of reads supporting alternative alleles should demonstrate a single peak at 0.5. Similarly, for an allohexaploid individual, the minor allele frequency should peak at around 0.33 etc. Based on this pattern, we predicted the ploidy levels by plotting the density of MAF and compare the predictions using the flow cytometry data.

Population demographics in history

We aligned the reads of the five whole-genome-sequenced individuals to their own reference genome using bwa¹²⁹, and performed demographic analysis using PSMC¹³⁰. Although it usually takes 1 year for *P. australis* to sprout and mature, we set the generation time to 4 years because it is perennial. We set the mutation rate to be the Poaceae silent-site mutation rate, 8e^{-09} substitutions per site per year¹³¹.

Identification of putative deleterious mutations

The nucleotide diversities of neutral and deleterious sites were calculated following the procedure in <https://github.com/jsalojar/PiNSiR>. Briefly, the short reads of Y17, Y21 and Y7 were aligned to their own draft genome using bwa mem¹²⁹ to obtain the variants along the genome, sorted with samtools¹³², and genome coverage variants were called using bcftools¹³³. Site-wise diversities were estimated using ANGSD¹³⁴ and used as input in the R package PiNSiR. The deleterious mutations were determined using SnpEff¹³⁵, and assuming the loci with “HIGH | MODERATE” flags to be deleterious.

Statistics and reproducibility

The gene overlap tests were carried out with either Fisher’s exact test or Chi-square tests. The GO enrichment was tested using Fisher’s exact test followed by *p*-value correction for multiple testing using Bonferroni, Sidak, Holm, and false discovery rate adjustment. The analyses in this study were carried out with genome assemblies from five individuals.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw sequencing data are available in NCBI SRA database under the BioProject ID PRJNA849004. This reference genome assembly has been deposited at DDBJ/ENA/GenBank under the accession JBCHVN000000000. The version described in this paper is version JBCHVN010000000. The assemblies of all the accessions and their annotations are available at CoGe (<https://genomevolution.org/coge/>): the chromosome-level reference CN individual (id 67975), scaffolded North American invasive lineage (id67997), Med lineage (Y21, id68003), USland lineage (Y7, id68004), and USnat lineage (Y17, id68002). In addition, the assemblies, the annotations, chloroplast and mitochondrial genome assemblies as well as the numerical source data underlying the graphs and charts can all be downloaded from the collection in Figshare (<https://doi.org/10.6084/m9.figshare.c.7368682.v1>).

Code availability

The scripts used for the construction of genome assembly and genome annotation, as well as data analysis in this article could be found from github https://github.com/smallfishcui/Phrag_Genome, and zenodo (<https://doi.org/10.5281/zenodo.12567126>).

Received: 22 January 2024; Accepted: 30 July 2024;

Published online: 17 August 2024

References

- Cuthbert, R. N. *et al.* Global economic costs of aquatic invasive alien species. *Sci. Total Environ.* **775**, 145238 (2021).
- Obiri, J. F. Invasive plant species and their disaster-effects in dry tropical forests and rangelands of Kenya and Tanzania. *J. Disaster Risk Stud.* **3**, 417–428 (2011).
- Diagne, C. *et al.* High and rising economic costs of biological invasions worldwide. *Nature* **592**, 571–576 (2021).
- Seebens, H. *et al.* No saturation in the accumulation of alien species worldwide. *Nat. Commun.* **8**, 14435 (2017).

5. Roy, H. E. et al. IPBES Invasive Alien Species Assessment: Summary for Policymakers. <https://zenodo.org/record/8314303> (2023).
6. Montesinos, D. Fast invasives fastly become faster: invasive plants align largely with the fast side of the plant economics spectrum. *J. Ecol.* **110**, 1010–1014 (2022).
7. Zhang, X., Oduor, A. M. & Liu, Y. Invasive plants have greater growth than co-occurring natives in live soil subjected to a drought-rewetting treatment. *Funct. Ecol.* **37**, 513–522 (2023).
8. Van Kleunen, M., Weber, E. & Fischer, M. A meta-analysis of trait differences between invasive and non-invasive plant species. *Ecol. Lett.* **13**, 235–245 (2010).
9. Bieker, V. C. et al. Uncovering the genomic basis of an extraordinary plant invasion. *Sci. Adv.* **8**, eabo5115 (2022).
10. Schrader, L. et al. Transposable element islands facilitate adaptation to novel environments in an invasive species. *Nat. Commun.* **5**, 5495 (2014).
11. Moura, R. F., Queiroga, D., Vilela, E. & Moraes, A. P. Polyploidy and high environmental tolerance increase the invasive success of plants. *J. Plant Res.* **134**, 105–114 (2021).
12. Leitch, A. & Leitch, I. Genomic plasticity and the diversity of polyploid plants. *Science* **320**, 481–483 (2008).
13. Dubcovsky, J. & Dvorak, J. Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* **316**, 1862–1866 (2007).
14. Van de Peer, Y., Mizrahi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).
15. Ren, Y. et al. Traditional uses, phytochemistry, pharmacology and toxicology of rhizoma phragmitis: a narrative review. *Chin. J. Integr. Med.* **28**, 1127–1136 (2022).
16. Wang, C. et al. Transcriptome analysis of tetraploid and octoploid common reed (*Phragmites australis*). *Front. Plant Sci.* **12**, 653183 (2021).
17. Ren, L. et al. Intraspecific variation in *Phragmites australis*: clinal adaption of functional traits and phenotypic plasticity vary with latitude of origin. *J. Ecol.* **108**, 2531–2543 (2020).
18. Song, H. et al. Is there evidence of local adaptation of *Phragmites australis* to water level gradients and fluctuation frequencies? *Sci. Total Environ.* **756**, 144065 (2021).
19. Zhu, X., Chen, G. & Zhang, C. Photosynthetic electron transport, photophosphorylation, and antioxidants in two ecotypes of reed (*Phragmites communis* Trin.) from different habitats. *Photosynthetica* **39**, 183–189 (2001).
20. Saltonstall, K. Cryptic invasion by a non-native genotype of the common reed, *Phragmites australis*, into North America. *Proc. Natl. Acad. Sci. USA* **99**, 2445–2449 (2002).
21. Wani, G. A. et al. Phenotypic variability and genetic diversity of *Phragmites australis* in Quebec and Kashmir reveal contrasting population structure. *Plants* **9**, 1392 (2020).
22. Facon, B. et al. A general eco-evolutionary framework for understanding bioinvasions. *Trends Ecol. Evol.* **21**, 130–135 (2006).
23. Lambertini, C. et al. Tracing the origin of Gulf Coast *Phragmites* (Poaceae): a story of long-distance dispersal and hybridization. *Am. J. Bot.* **99**, 538–551 (2012).
24. Liu, L. L. et al. Cryptic lineages and potential introgression in a mixed-ploidy species (*Phragmites australis*) across temperate China. *J. Syst. Evol.* **60**, 398–410 (2022).
25. Pyšek, P. et al. Small genome separates native and invasive populations in an ecologically important cosmopolitan grass. *Ecology* **99**, 79–90 (2018).
26. Oh, D. H. et al. Novel genome characteristics contribute to the invasiveness of *Phragmites australis* (common reed). *Mol. Ecol.* **31**, 1142–1159 (2022).
27. Connor, H., Dawson, M., Keating, R. & Gill, L. Chromosome numbers of *Phragmites australis* (Arundineae: Gramineae) in New Zealand. *N.Z. J. Bot.* **36**, 465–469 (1998).
28. Clevering, O. A. & Lissner, J. Taxonomy, chromosome numbers, clonal diversity and population dynamics of *Phragmites australis*. *Aquat. Bot.* **64**, 185–208 (1999).
29. Saltonstall, K., Peterson, P. M. & Soreng, R. J. Recognition of *Phragmites australis* subsp. *americanus* (poaceae: arundinoideae) in North America: evidence from morphological and genetic analyses. *Sida* **21**, 683–692 (2004).
30. Lambertini, C. et al. A phylogeographic study of the cosmopolitan genus *Phragmites* (Poaceae) based on AFLPs. *Plant Syst. Evol.* **258**, 161–182 (2006).
31. Salojärvi, J. et al. Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nat. Genet.* **49**, 904–912 (2017).
32. Lambertini, C., Sorrell, B. K., Riis, T., Olesen, B. & Brix, H. Exploring the borders of European *Phragmites* within a cosmopolitan genus. *AoB Plants* **2012**, pls020 (2012).
33. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126–e126 (2018).
34. Jia, K. H. et al. SubPhaser: a robust allopolyploid subgenome phasing method based on subgenome-specific k-mers. *New Phytol.* **235**, 801–809 (2022).
35. McKain, M. R. et al. A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. *Genome Biol. Evol.* **8**, 1150–1164 (2016).
36. Ma, L. et al. Diploid and tetraploid genomes of *Acorus* and the evolution of monocots. *Nat. Commun.* **14**, 3661 (2023).
37. Burns, R. et al. Gradual evolution of allopolyploidy in *Arabidopsis suecica*. *Nat. Ecol. Evol.* **5**, 1367–1381 (2021).
38. Liu, C. & Wang, Y.-G. Does one subgenome become dominant in the formation and evolution of a polyploid? *Ann. Bot.* **131**, 11–16 (2023).
39. Alger, E. I. & Edger, P. P. One subgenome to rule them all: underlying mechanisms of subgenome dominance. *Curr. Opin. Plant Biol.* **54**, 108–113 (2020).
40. Birchler, J. A. & Veitia, R. A. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl. Acad. Sci. USA* **109**, 14746–14753 (2012).
41. Huang, W. et al. A well-supported nuclear phylogeny of Poaceae and implications for the evolution of C4 photosynthesis. *Mol. Plant* **15**, 755–777 (2022).
42. Hardion, L. et al. Does infraspecific taxonomy match species evolutionary history? A phylogeographic study of *Arundo formosana* (Poaceae). *Botanical J. Linn. Soc.* **183**, 236–249 (2017).
43. Huang, Y.-L. et al. The impact of tandem duplication on gene evolution in Solanaceae species. *J. Integr. Agricult.* **21**, 1004–1014 (2022).
44. Hanada, K., Zou, C., Lehti-Shiu, M. D., Shinozaki, K. & Shiu, S.-H. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol.* **148**, 993–1003 (2008).
45. Pyšek, P. et al. Competition among native and invasive *Phragmites australis* populations: An experimental test of the effects of invasion status, genome size, and ploidy level. *Ecol. Evol.* **10**, 1106–1118 (2020).
46. Camacho, J. P. M., Sharbel, T. F. & Beukeboom, L. W. B-chromosome evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **355**, 163–178 (2000).
47. Copenhaver, G. P. & Pikaard, C. S. Two-dimensional RFLP analyses reveal megabase-sized clusters of rRNA gene variants in *Arabidopsis thaliana*, suggesting local spreading of variants as the mode for gene homogenization during concerted evolution. *Plant J.* **9**, 273–282 (1996).
48. Liu, B. & Davis, T. M. Conservation and loss of ribosomal RNA gene sites in diploid and polyploid *Fragaria* (Rosaceae). *BMC Plant Biol.* **11**, 1–13 (2011).

49. Guo, X. & Han, F. Asymmetric epigenetic modification and elimination of rDNA sequences by polyploidization in wheat. *Plant Cell* **26**, 4311–4327 (2014).
50. Dhar, M. K., Kour, J. & Kaul, S. Origin, behaviour, and transmission of B chromosome with special reference to *Plantago lagopus*. *Genes* **10**, 152 (2019).
51. Augusto Corrêa dos Santos, R., Goldman, G. H. & Riaño-Pachón, D. M. ploidyNGS: visually exploring ploidy with Next Generation Sequencing data. *Bioinformatics* **33**, 2575–2576 (2017).
52. Ramu, P. et al. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat. Genet.* **49**, 959–963 (2017).
53. Ishii, J. & Kadono, Y. Factors influencing seed production of *Phragmites australis*. *Aquat. Bot.* **72**, 129–141 (2002).
54. Kaneko, Y. Biology of the Common Reed (*Phragmites australis*) surrounding Lake Biwa. In: *Lake Biwa: Interactions between Nature and People*, (eds Kawanabe, H., Nishino, M. & Maehata, M.) 137–138 (Springer, 2012).
55. Amsberry, L., Baker, M. A., Ewanchuk, P. J. & Bertness, M. D. Clonal integration and the expansion of *Phragmites australis*. *Ecol. Appl.* **10**, 1110–1118 (2000).
56. Alvarez, M. G., Tron, F. & Mauchamp, A. Sexual versus asexual colonization by *Phragmites australis*: 25-year reed dynamics in a Mediterranean marsh, southern France. *Wetlands* **25**, 639–647 (2005).
57. Čuda, J., Skálová, H., Meyerson, L. A. & Pyšek, P. Regeneration of *Phragmites australis* from rhizome and culm fragments: an experimental test of environmental effects, population origin and invasion status. *Preslia* **93**, 237–254 (2021).
58. Glemin, S. & Galtier, N. Genome evolution in outcrossing versus selfing versus asexual species. *Evolut. Genomics Stat. Comput. Methods* **1**, 311–335 (2012).
59. Wang, X.-J. et al. The genomic selfing syndrome accompanies the evolutionary breakdown of heterostyly. *Mol. Biol. Evol.* **38**, 168–180 (2021).
60. Bast, J. et al. Consequences of asexuality in natural populations: insights from stick insects. *Mol. Biol. Evol.* **35**, 1668–1677 (2018).
61. Ho, E. K., Bartkowska, M., Wright, S. I. & Agrawal, A. F. Population genomics of the facultatively asexual duckweed *Spirodela polyrrhiza*. *N. Phytol.* **224**, 1361–1371 (2019).
62. Tvedte, E. S., Ward, A. C., Trendle, B., Forbes, A. A. & Logsdon J. R. J. M. Genome evolution in a putatively asexual wasp. *bioRxiv*. (<https://doi.org/10.1101/2020.12.23.424202>) (2020).
63. Chen, J., Glémin, S. & Lascoux, M. Genetic diversity and the efficacy of purifying selection across plant and animal species. *Mol. Biol. Evol.* **34**, 1417–1428 (2017).
64. Henry, L., Schwander, T. & Crespi, B. J. Deleterious mutation accumulation in asexual *Timema* stick insects. *Mol. Biol. Evol.* **29**, 401–408 (2012).
65. Lovell, J. T., Williamson, R. J., Wright, S. I., McKay, J. K. & Sharbel, T. F. Mutation accumulation in an asexual relative of *Arabidopsis*. *PLoS Genet.* **13**, e1006550 (2017).
66. Albert, A., Brisson, J., Belzile, F., Turgeon, J. & Lavoie, C. Strategies for a successful plant invasion: the reproduction of *Phragmites australis* in north-eastern North America. *J. Ecol.* **103**, 1529–1537 (2015).
67. Dwivedi, S. L. et al. Evolutionary dynamics and adaptive benefits of deleterious mutations in crop gene pools. *Trends Plant Sci.* **28**, 685–697 (2023).
68. Jike, W. et al. Phylogenomic proof of recurrent demipolyploidization and evolutionary stalling of the “Triploid Bridge” in *Arundo* (Poaceae). *Int. J. Mol. Sci.* **21**, 5247 (2020).
69. Wang, C. et al. Genome-wide analysis tracks the emergence of intraspecific polyploids in *Phragmites australis*. *bioRxiv*. (<https://doi.org/10.1101/2021.09.05.458733>) (2021).
70. Holmes, D. & Bougourd, S. B-chromosome selection in *Allium schoenoprasum* L. Experimental populations. *Heredity* **67**, 117–122 (1991).
71. Teoh, S., Rees, H. & Hutchinson, J. B chromosome selection in *Lolium*. *Heredity* **37**, 207–213 (1976).
72. Banaei-Moghaddam, A. M., Meier, K., Karimi-Ashtiyani, R. & Houben, A. Formation and expression of pseudogenes on the B chromosome of rye. *Plant Cell* **25**, 2536–2544 (2013).
73. Randolph, L. Genetic characteristics of the B chromosomes in maize. *Genetics* **26**, 608 (1941).
74. Huang, W., Du, Y., Zhao, X. & Jin, W. B chromosome contains active genes and impacts the transcription of A chromosomes in maize (*Zea mays* L.). *BMC Plant Biol.* **16**, 1–14 (2016).
75. Ma, W., Liu, Z., Beier, S., Houben, A. & Carpentier, S. Identification of rye B chromosome-associated peptides by mass spectrometry. *N. Phytol.* **230**, 2179–2185 (2021).
76. Shi, X. et al. Effect of aneuploidy of a non-essential chromosome on gene expression in maize. *Plant J.* **110**, 193–211 (2022).
77. Romera, F., Vega, J., Diez, M. & Puertas, M. B chromosome polymorphism in Korean rye populations. *Heredity* **62**, 117–121 (1989).
78. Schelkunov, M. I. et al. The genome of the toxic invasive species *Heracleum sosnowskyi* carries an increased number of genes despite absence of recent whole-genome duplications. *Plant J.* **117**, 449–463 (2024).
79. Bi, C. et al. PMAT: an efficient plant mitogenome assembly toolkit using low coverage HiFi sequencing data. *Horticult. Res.* **11**, <https://doi.org/10.1093/hr/uhae023> (2024).
80. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
81. Jin, J.-J. et al. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* **21**, 1–31 (2020).
82. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).
83. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
84. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
85. Durand, N. C. et al. Juicebox provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
86. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
87. Dudchenko, O. et al. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under \$1000. *BioRxiv*, 254797 (2018).
88. Xu, M. et al. TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience* **9**, g1aa094 (2020).
89. Chikhi, R. & Medvedev, P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **30**, 31–37 (2014).
90. Zimin, A. V. et al. The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
91. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinforma.* **19**, 1–10 (2018).
92. 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. *Bioinformatics* **31**, 3210–3212 (2015).

93. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
94. Alonge, M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 1–19 (2022).
95. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinforma.* **9**, 1–14 (2008).
96. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
97. 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).
98. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
99. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506 (2005).
100. Hoff, K., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: Incorporating Protein Homology Information into Gene Prediction with GeneMark-EP and AUGUSTUS. In: *Plant and Animal Genomes XXVI*, (San Diego, California, 2018).
101. Haas, B. J., Zeng, Q., Pearson, M. D., Cuomo, C. A. & Wortman, J. R. Approaches to fungal genome annotation. *Mycology* **2**, 118–141 (2011).
102. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
103. Goodstein, D. M. et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186 (2012).
104. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol. Biol.* **1962**, 161–177 (2019).
105. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
106. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinforma.* **3**, lqaa108 (2021).
107. Grabherr, M. G. et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **29**, 644 (2011).
108. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
109. Haas, B. J. et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
110. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, 1–22 (2008).
111. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
112. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
113. Shen, L. GeneOverlap: an R package to test and visualize gene overlaps. <https://bioconductor.org/packages/release/bioc/html/GeneOverlap.html> (2023).
114. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958 (2018).
115. He, W. et al. NGenomeSyn: an easy-to-use and flexible tool for publication-ready visualization of syntenic relationships across multiple genomes. *Bioinformatics* **39**, btad121 (2023).
116. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 1–13 (2019).
117. Goel, M. & Schneeberger, K. plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics* **38**, 2922–2926 (2022).
118. Klopfenstein, D. et al. GOATOOLS: a python library for gene ontology analyses. *Sci. Rep.* **8**, 1–17 (2018).
119. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14 (2019).
120. Almeida-Silva, F. & Van de Peer, Y. Assessing the quality of comparative genomics data and results with the cogeqc R/Bioconductor package. *Methods in Ecology and Evolution* **14**, 2942–2952 (2023).
121. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).
122. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
123. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinforma.* **19**, 15–30 (2018).
124. dos Reis, M. & Yang, Z. Bayesian Molecular Clock Dating Using Genome-Scale Datasets. in *Evolutionary Genomics: Statistical and Computational Methods* (ed. Anisimova, M.) 309–330 (Springer, 2019).
125. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
126. Ramirez-Barahona, S., Sauquet, H. & Magallon, S. The delayed and geographically heterogeneous diversification of flowering plant families. *Nat. Ecol. Evol.* **4**, 1232–1238 (2020).
127. Wolfe, K. H., Gouy, M., Yang, Y.-W., Sharp, P. M. & Li, W.-H. Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. *Proc. Natl. Acad. Sci. USA* **86**, 6201–6205 (1989).
128. Santos, R.A.C.d., Goldman, G. H. & Riaño-Pachón, D. M. ploidyNGS: visually exploring ploidy with next generation sequencing data. *Bioinformatics* **33**, 2575–2576 (2016).
129. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
130. Li, H. & Durbin, R. Inference of human population history from whole genome sequence of a single individual. *Nature* **475**, 493 (2011).
131. De La Torre, A. R., Li, Z., Van de Peer, Y. & Ingvarsson, P. K. Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. *Mol. Biol. Evol.* **34**, 1363–1377 (2017).
132. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
133. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
134. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinforma.* **15**, 1–13 (2014).
135. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).

Acknowledgements

This study was supported by Young Scientists Fund of Shandong Provincial Natural Science Foundation ZR2021QC119 (to C.W.), International

Postdoctoral Exchange Fellowship Program of China Postdoctoral Science Foundation (to C.W.), National Natural Science Foundation of China U22A20558 (to W.H.G.), National Natural Science Foundation of China 32100304 (to L.L.Liu), National Natural Science Foundation of China 31800299 (to T.W.), Academy of Finland 318288 (to J.S.), Academy of Finland 319947 (to J.S.) and Nanyang Technological University startup grant (to J.S.). We would like to thank Pasi Rastas for providing the script to calculate alternative allele count, and Sitaram Rajaraman for sharing the work pipeline of genome annotation. We appreciate the assistance from James Ord in improving the English language throughout this article. Finally, we want to acknowledge CSC–IT Center for Science, Finland, and NTU HPCC, Singapore, for computational resources.

Author contributions

C.W., W.G. and J.S. conceived the study; H.B., L.L., F.E., M.Y., Y.W. and T.W. maintained and collected the samples as well as provided sampling information; C.W., L.L. and M.Y. performed DNA extraction; C.W. and J.S. lead the data analysis; L.L. performed the flow cytometry analysis, B.L. and Y.G. performed karyotype analysis; C.W. and J.S. wrote the paper with input from W.G., L.L., F.E., H.B., T.W., and M.Y.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-024-06660-1>.

Correspondence and requests for materials should be addressed to Weihua Guo or Jarkko Salojärvi.

Peer review information *Communications biology* thanks Wanqiang Qian and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: David Favero. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024