Title: Genome-wide analysis illustrates the effect of polyploidization in *Phragmites australis*

Authors: Cui Wang, Dr.
1 Institute of Ecology and Biodiversity, School of Life Sciences, Shandong University, 72 Binhai Road, Qingdao 266237, China
2 Shandong Provincial Engineering and Technology Research Center for Vegetation Ecology, Shandong University, 72 Binhai Road, Qingdao 266237, China
3 Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki, Viikinkaari 1, Biocentre 3, 00790 Helsinki, Finland
Lele Liu, Dr.
1 Institute of Ecology and Biodiversity, School of Life Sciences, Shandong University, 72 Binhai Road, Qingdao 266237, China
2 Shandong Provincial Engineering and Technology Research Center for Vegetation Ecology, Shandong University, 72 Binhai Road, Qingdao 266237, China
Meiqi Yin, Ms.
1 Institute of Ecology and Biodiversity, School of Life Sciences, Shandong University, 72 Binhai Road, Qingdao 266237, China
2 Shandong Provincial Engineering and Technology Research Center for Vegetation Ecology, Shandong University, 72 Binhai Road, Qingdao 266237, China
Franziska Eller, Assistant Professor,
4 Department of Biology, Aarhus University, Aarhus, Denmark
Hans Brix, Professor,
4 Department of Biology, Aarhus University, Aarhus, Denmark
Tong Wang, Lecturer,
5 College of Landscape Architecture and Forestry, Qingdao Agricultural University, Qingdao, China
Jarkko Salojärvi*, Assistant Professor,
3 Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki, Viikinkaari 1, Biocentre 3, 00790 Helsinki, Finland
6 School of Biological Sciences, Nanyang Technological University
Weihua Guo*, Professor

1 Institute of Ecology and Biodiversity, School of Life Sciences, Shandong University, 72 Binhai Road, Qingdao 266237, China

2 Shandong Provincial Engineering and Technology Research Center for Vegetation Ecology, Shandong University, 72 Binhai Road, Qingdao 266237, China

Correspondence to:

Professor Weihua Guo, whguo@sdu.edu.cn

Assistant Professor Jarkko Salojärvi, jarkko@ntu.edu.sg

Genome-wide analysis illustrates the effect of polyploidization in *Phragmites australis*

Abstract

Polyploidization is a common event in plant evolution, and it plays an important role in plant speciation and adaptation. To address the role of polyploidization in grass diversification, we studied *Phragmites australis*, a species with intraspecific variation of chromosome numbers ranging from 2n=36 to 144. A combined analysis of genome structure, phylogeny and population genetics were used to study the evolution of *P. australis*. Whole-genome sequencing of three representative lineages revealed the allopolyploid origin of the species, with subgenome divergence dating back to approximately 29 million years ago, and the genomes showed hallmarks of relaxed selection associated with asexual propagation. Genome-wide analysis of 88 individuals from different populations around the world using restriction site associated DNA sequencing (RAD-seq) identified seven main intraspecific lineages with extensive genetic admixture. Each lineage was dominated by a distinct ploidy level, mostly tetraploid or octoploid, suggesting several polyploid events. Furthermore, we observed octoploid and hexaploid lineages at contact zones in Romania, Hungary and South Africa, suggestively due to genomic conflicts of allotetraploid parental lineages. Polyploidy may have evolved as a strategy to escape from the evolutionary dead-end of asexual propagation and the resulting decrease in genomic plasticity.

Introduction

The rapidly increasing number of high quality plant genome assemblies have uncovered numerous polyploidization or whole genome duplication (WGD) events immediately prior to or co-occurring with the time of species divergence [1]. These events, originating from autopolyploid or allopolyploid hybrids, are particularly common among grasses [2], for example, all cereals have undergone rho, sigma and tau WGD events. The grass family (Poaceae), which includes more than 10,000 species today, originated from a common ancestor with five chromosomes [3]. In cereals, a series of genome duplication and consecutive genome fractionation events, involving chromosomal structural rearrangements and losses of redundancy, led to a common ancestral genome with 12 chromosomes, shared across the early diverging grass subfamilies Anomochlooideae, Pharoideae, and Puelioideae [3,4]. Hence, all extant Poaceae species are paleopolyploids, in which polyploidization occurred millions of years ago [4]. Depending on their lineages, the current diploids have experienced further WGDs in the form of allo- or autopolyploidization, giving rise to the current grass species.

A recent polyploidization event can be directly inferred when closely related species show different chromosome counts or genome sizes, which differ by an integer-valued factor. This can often be observed in species within the same genus, but in some cases different ploidy levels have been observed also within the same species, such as *Dupontia fisheri*, *Betula pendula*, *Lygeum spartum*, *Cynodon dactylon* and *Phragmites australis* [5-9]. These species are thus ideal models to understand the factors that drive polyploidization. One critical step to characterize the extent of polyploidization is to assess the ploidy level directly by counting chromosome numbers of individual metaphases, or quantifying the genome size in the germplasm using, for example, flow cytometry and comparing it to the size in a known diploid representative. Nevertheless, the tedious procedures and the need of appropriate facilities can be challenging. With the advance of next generation sequencing (NGS) technology, large quantities of NGS reads now provide a chance to infer the ploidy levels through alleles of the genetic markers. Tools such as ploidyNGS [10] and ConPADE [11] and nQuire [12] have been proposed for inferring ploidy levels by fitting a model to the allele frequencies from mapping against the reference genome or calculating the frequency of haplotypic contigs. However, the methods are

still under development and demonstrate relatively high error rates. Therefore, a robust method to determine the ploidy level is still needed to understand the evolution of polyploidy in plants.

Common reed, *Phragmites australis* (Cav.) Trin. ex Steud., is a tenacious cosmopolitan wetland species which provides habitats for small animals and is considered an ecosystem engineer. Similar to many other grass species, *Phragmites* has a basal chromosome count of x=12. However, the euploid chromosome count has proved to be versatile, ranging from 3x to 12x (3x, 4x, 6x, 7x, 8x, 10x, 11x and 12x), or even mixoploid [13,14]. Tetraploid and octoploid individuals are most common in nature [13,14]. Substantial efforts to investigate the evolution of speciation in *Phragmites* have been made, but the evolution of ploidy levels within this species is still unclear [15-19]. To date, seven species, *P. australis*, *P. mauritianus* Kunth, *P. frutescens* H. Scholz, *P. dioica* Hackel ex Conert, *P. berlandieri* E. Fourn., *P. japonicus* Steudel and *P. vallatoria* (Plunk. ex L.) Veldk, have been described in *Phragmites* [17]. Most of them are distributed regionally, with only *P. australis* being cosmopolitan [13,17,20]. Ploidy variation was found in several species, e.g., *P. vallatoria* has 2x, 3x, 4x with aneuploid representatives at all three ploidy levels, *P. japonicus* preserves 4x and aneuploid 8x, and *P. mauritianus* is mostly 4x [17]. Large intraspecific variation was documented in *P. australis* by 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* [17,21,22].

Based on non-coding chloroplast regions *trnT-trnL* and *rbcL-psaI*, researchers have classified lineages to 57 haplotypes [23], roughly corresponding to geographic regions of origin. According to the standard naming system [15], several phylogenetic studies have found haplotype M to be commonly distributed in North America, Europe, and Asia, and later confirmed it to be an invasive lineage in North America [15]. Haplotype I is distributed across South America, southern Pacific islands and Asia; haplotypes U and Q occur in Asia and Australia, whereas haplotype O is mainly distributed in northern China and haplotype P in Korea and middle to southern China. A number of unique haplotypes have emerged endemically in Mexico, or North America, and in southwestern China it is possible to find three main haplotypes I, Q, U [15,16,24,25]. Because the chloroplast haplotypes are not continuously distributed within the same geographic region, it is difficult to conclude how they spread originally, especially when considering the variation observed in intraspecific ploidy levels. Studies based on amplified fragment length polymorphisms (AFLP) and microsatellites have provided further support to the geographic lineage classifications. However, the AFLP produced only 107 variable sites among the species, which may not be sufficient for analyzing allopolyploid species with up to 144 chromosomes [17].

In United States, the range of exotic *Phragmites australis* has expanded from a very limited area to span almost the entire country within just 50 years, forcing the native lineages to the very north; thus it is considered an invasive plant [15]. One hypothesis for the development of invasiveness in *P. australis* is related to polyploidizations [13,26], because polyploids, due to their higher level of genomic plasticity, can outperform their diploid progenitors in environments with high abiotic stress, and thus have access to a wider range of habitats [27]. Successful allopolyploids may thus develop into invasive species and drive diploid ancestors to extinction by recurrent hybridizations with them.

In this study, we investigated the intraspecific genetic divergence, population genetic structure, historical demographics, and hybridization of *P. australis* using RAD-seq approaches that incorporated data from the nuclear genome and chloroplast from 88 individuals sampled throughout the geographic range of the species. To understand the geographic distribution of ploidy variation in the species, we developed a method based on alternative allele frequency of reads mapped to the reference genome to predict the ploidy level of each individual. We obtained high coverage Illumina sequencing data for the whole genomes of three individuals representing three lineages and assembled their draft genomes to investigate the genome evolution associated with polyploidization, and to give insights into the rise of whole genome duplications in grass species.

Results and discussion

Genome assemblies confirm allopolyploid structure of P. australis

We carried out RAD-sequencing for a collection of 88 *P. australis* individuals from Eurasian, North American, Oceanian and African continents (Fig. 1, Table 1). A subset of these samples was

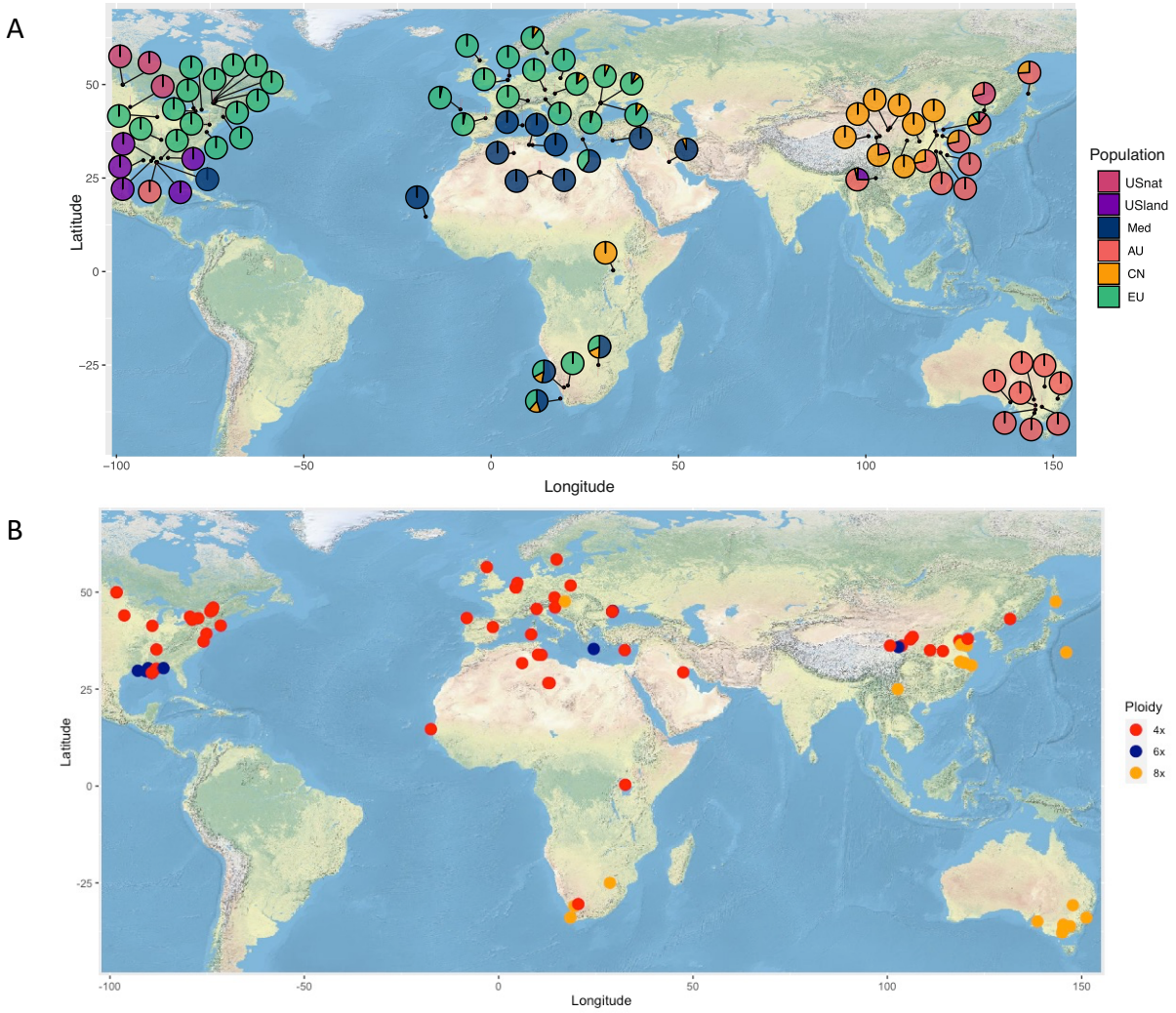

Figure 1 Sampling locations of *Phragmites australis* individuals in the study. A) Pie charts indicate the genetic admixture proportions. B) Ploidy level estimates from flow cytometry. The ploidy levels of the individuals with no flow cytometry data were predicted using sequencing data, see Methods for a detailed description. Different colors represent distinct ploidy level.

quantified with flow cytometry to obtain genome size estimate (Table 1; flow cytometry info provided by H. B., the samples have been used in previous studies [17]). For initial exploration of the RAD-seq data we carried out a *de novo* analysis using Stacks v2.4[28]. Even though the method assumes the organisms to be diploid and only considers biallelic sites, we were able to obtain enough qualified variant sites to describe the population grouping of the species.

Based on this initial analysis, we identified three non-admixed representatives of distinct genetic lineages; the so-called US land type occurring in Gulf Coast (USland; Y7)[18,29], native US lineage limited to the North[17] (USnat; Y17) and Mediterranean lineage (Med; Y21)[17]. The three accessions were sequenced to 90x coverage using Novaseq short read sequencing and assembled using MaSurCA genome assembler after which contigs representing haplotype variants of the same locus were removed. The final assembly sizes were 922 Mb for US native lineage (Y17), 815 Mb for Mediterranean lineage (Y21), which is nearly twice the monoploid Cx value (0.490, 479.22 Mbp) [30].

Finally, the assembly size for USland (Y7) was 1.2 Gb. Quality assessment using universally conserved single-copy genes (BUSCO v5.2.1) yielded high values (92.3-98.0% completeness, Table 2), suggesting that the draft assemblies were of comparable quality and that they had captured the gene coding space with high accuracy. We further noticed equal proportions of duplicated genes across all assemblies (34-48.9%). The duplicated genes showed significant overlap among the three assemblies (Fig. 2A, Table 2), suggesting that the duplication is biological and likely associated with the polyploid nature of *P. australis*. Genome annotation was carried out next, by combining homology information from two Poales species, *ab initio* gene predictors and publicly available RNAseq data, into consensus predictions using Evidence Modeler [9,31]. The highest number of protein coding genes were predicted in USland (Table 2).

Phragmites australis has been suggested to be an allopolyploid species [7] but this hypothesis has been largely ignored due to the limited evidential support available at the time. To study this hypothesis, we carried out syntenic alignments against *Oropetium thomaeum* genome assembly. For all assemblies two peaks in the Ks spectrum were detected, suggesting that the two subgenomes have different divergence times from *Oropetium*, thus confirming an allopolyploid origin for *P. australis* (Fig 2 B). All assemblies had roughly equal numbers of syntenic blocks, providing yet no explanation to the larger assembly size in Y7 (Table S1).

We next used the Ks (synonymous substitutions rate) spectrum to phase syntenic blocks with Ks values between 0 - 0.63 to subgenome A, and blocks with Ks value between 0.63 - 2 to subgenome B (Table S1). Divergence time estimation for the phased duplicates of universally conserved single-copy genes (BUSCOs), using divergence time with *O. sativa* as calibration point, showed that the subgenomes diverged at ca. 29.07 Mya (95% Highest Posterior Density, HPD 26.74 – 31.70 MYA) (Fig. 2C).

Both subgenomes gave largely concordant evidence about the later divergence between the three lineages. In subgenome A, USnat lineage diverged from others at 5.58 MYA (95% HPD 4.83 - 6.40MYA), followed by divergence between Med and USland lineages at 2.44 MYA (95% HPD 2.02 - 2.87MYA). In subgenome B, USnat diverged from others at 6.84 MYA (95% HPD 6.02 - 7.72MYA), followed by divergence between USland and Med at 2.66 MYA (95% HPD 2.28 – 3.03Mya) (Fig.2 C), indicating a similar evolutionary rate of the two subgenomes of *P. australis.* It is not known when the allopolyploidization event occurred, but the clear separation of the two subgenomes suggests that the two ancestral species have evolved independently for a considerable time to produce barriers for gene flow between the two subgenomes. It is worth noticing that the divergence time between *Arundo* and *Phragmites* species was also estimated to be 29 Mya [32], therefore one possible scenario is that the current genome is generated by ancestor of *Phragmites* hybridizing with an ancestral species of the *Arundo* genus.

Alternative allele percentage spectrum predicts differing ploidy levels

The percent of reads supporting reference vs. alternative allele in a locus can be used as a tool to assess the ploidy levels of the individuals [33]. For all individuals with whole genome sequencing data the allele percentage histograms showed a main peak at 0.5, but the overall shape differed. In USnat and Med, two side peaks were observed at 0.09 and 0.91, whereas in USland the major peak was wider and less steep (Fig. 2D-F). We hypothesized that the inconclusive shape was due to repeat elements and other ambiguously mapping regions in the genomes, and therefore focused the analysis on the gene models with assignations to subgenome A or B based on the Ks values. Both USnat and Med displayed a clear main peak at 0.5 and retained two side peaks at 0.09 and 0.91 (Fig. 2D-E), indicating the subgenomes were diploid, and these two lineages should be regarded as allotetraploid. However, for USland two major peaks at 0.337 and 0.669 were observed (Fig. 2F), suggesting a hexaploid genome organization; higher ploidy level also explains the 1.5x larger assembly size of the USland. Moreover, the number of BUSCOs with three duplicates is much higher in Y7 (106 copies)

than in Y17 (51 copies) and Y21 (58 copies), which further supports the hypothesis (Table 2). Therefore, USland lineage most likely resulted from a tetraploid – octoploid hybridization event.

Although both USnat and Med lineages were allotetraploid, the genome size and gene count of USnat was 13.12 – 14.15% higher than Med lineage. The two lineages diverged at 5.58 – 6.84 Mya in late Miocene, suggesting that different lineages may have been going through differential gene loss in diploidization [34]. There were 1916 duplicated BUSCOs shared by USnat and Med lineage, but only 1203 (62.79%) of these were also shared with USland with 1665 duplicated BUSCOs. One possible reason for the increased amount of copies in USland could be that the remaining duplicates were inherited from another closely related species or an allopatric diverged population.

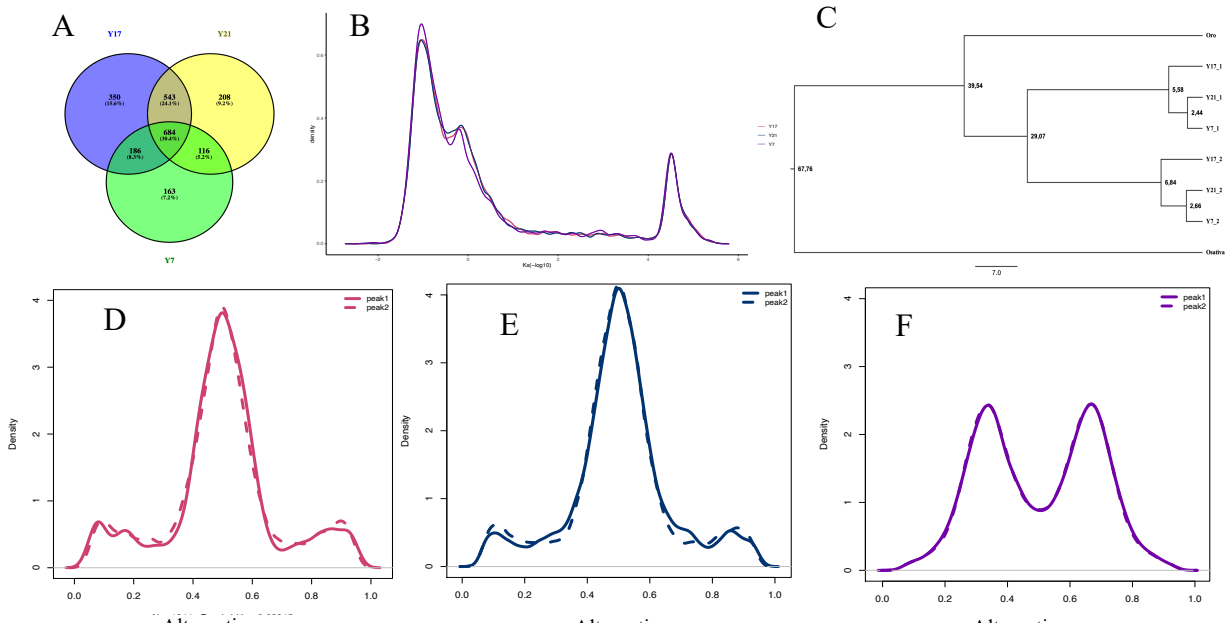

Figure 2 A) Overlap of the multicopy BUSCOs among genomes Y17, Y21, and Y7. B) Histograms of Ks values of syntenic genes of Y17, Y21 and Y7 genomes when aligned against *Oropetium thomaeum* genome. C) Divergence time of *Phragmites australis* subgenomes estimated from single copy genes in the ancestors and duplicated genes in current *P. australis* lineages. D-F) Allele frequency of reads mapped to peak1 and peak 2 in Y17, Y21 and Y7.

Population genetics analysis reveals links between polyploidization and lineage divergence

The initial *de novo* assembly of RAD-tags generated 4,565,536 loci with the average length of 521.3 bp and mean effective coverage of 39.0x per sample (with minimum 22.8x and maximum at 65.7x). Altogether the regions contained 40,507,058 single nucleotide polymorphism (SNP) sites. In contrast, a reference-guided alignment against USnat assembly resulted in 7,045,580 loci and 19,535,449 SNP positions, suggesting that many of the *de novo* loci were collapsed assemblies of the two sub-genomes. In the reference-based RAD analysis the region length was 305.9, and the mean coverage of each sample was 39.6x (minimum coverage 23.9x, maximum coverage 65.8x). From this set, 16,500 SNPs were mapped in chloroplast genome.

Multiple ploidy levels present in P. australis

For RAD-seq data, a histogram of the proportion of reads supporting alternative allele calls showed modes at around 0.5 in tetraploids, at 0.35, 0.65 in hexaploids, and between 0.35 – 0.65 in octoploids (Fig. S1). Of the total of 88 *Phragmites* individuals, the ploidy levels of 55 individuals were quantified with flow cytometry, and only five samples showed different patterns between flow cytometry and

our prediction method (Y12, Y24, E3, Y37, K2); this could be an error in flow cytometry or aneuploidy of the same individual; therefore, the prediction accuracy is at least 90.9% (Table 1), indicating that the prediction of ploidy levels from RAD-seq data can be done accurately. In addition, our approach is intuitive to visualize and is suited for large plant genomes with high ploidy levels. However, it should not be treated as a replacement of traditional methods for determining ploidy level, but can be used as a complementary method in addition to the flow cytometry results.

Using our approach we predicted the ploidy levels for the remaining 33 individuals. Most individuals from North America, Northern China, Mediterranean regions, and Europe (representing USnat, CN, Med, EU lineages; see Fig. 4) were predicted to be tetraploid, showing unimodal distribution peaking at 0.5 (Table 1; Fig 3A-D). All representatives from Gulf Coast

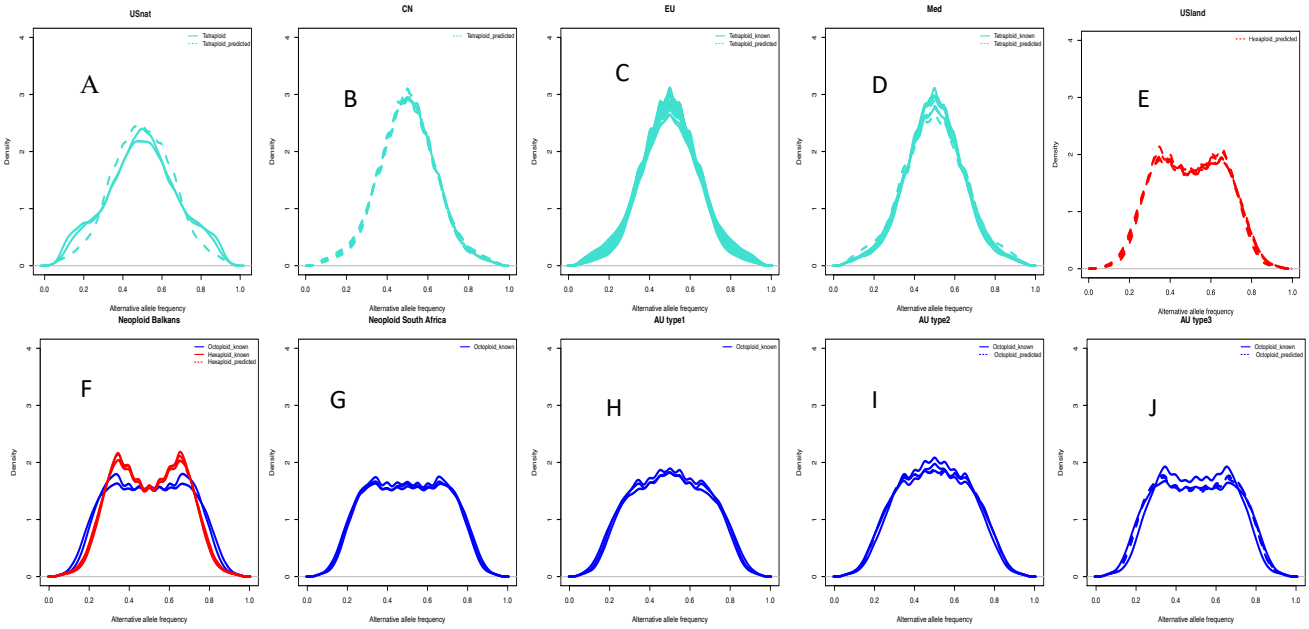

Figure 3 The alternative allele frequency histograms from aligning the RAD-seq reads to the Y17 reference genome. Solid line shows frequency distribution of alternative alleles in individuals where the result was confirmed with flow cytometry. Dashed line shows frequency distribution of alternative alleles in individuals with missing or differing flow cytometry measurement. Turquoise, red and blue color represented tetraploid, hexaploidy and octoploid respectively. Pattern of alternative allele frequency in population USnat (A), CN (B), EU (C), Med (D), USland (E), Neoploid (F-G), and AU (H-J).

(USland lineage, see below Fig. 4), and a few admixed individuals (Fig. 4), were predicted to be hexaploids (Fig. 3 E-F), whereas Australian (AU lineage, Fig. 4) and a few admixed individuals from South Africa and Mediterranean region (Fig. 4) were predicted to be octoploids (Table 1, Table 3; Fig3 F-J). Out of 20 admixed individuals, 13 showed increased ploidy levels and two individuals showed decreased ploidy levels compared to parental populations. Five individuals, including samples from South Africa, Romania and Hungary, were admixed offspring of tetraploid lineages CN, EU and Med (see below Fig. 4), but they were found to be octoploid, while another three individuals of the same background were predicted to be hexaploids, and we refer to these individuals as Neoploid (Balkans); see Fig 3F-G , samples Y3, Y36, A5, A7, A8, B2, F2, H2 and K2 (Table 3). The octoploids in the Danube delta were classified into EU lineage previously [35].

Population structure shows several independent polyploidization events within P. australis

After selecting one random SNP from each Stacks locus, we obtained 538,761 variant sites in total, and 13,536 variants remained after selecting loci with no more than 10% missing values and minor allele frequency greater than 5%. Maximum likelihood phylogenetic trees revealed seven distinct

genetic clusters, namely Native American lineage (USnat), Australian lineage (AU), Chinese lineage (CN), European lineage (EU), South African lineage, Mediterranean lineage (Med), and US land type (USland) (Fig. 4B). USnat and AU lineages diverged from other lineages at an early stage. Subsequently, CN lineage split from EU, South Africa, Med and USland. Population structure analyses using fastSTRUCTURE supported K=6 as the best number of clusters, in terms of highest marginal likelihood. These six genetic groups coincided with the geographical area, corresponding to USnat, USland, Med, AU, CN, and EU in the phylogenetic tree (Fig. 4B). Admixed individuals were mainly detected between AU vs CN, EU vs CN, and Med vs EU vs CN (Table 3).

The phylogeny based on RADseq loci mapping on chloroplast showed a similar split into six genetic groups corresponding to USnat, AU, USland, Med, CN and EU (Fig. 5), with varying bootstrap support values. Around 14 individuals showed discordant placement between chloroplast and nuclear trees, which is very likely a result from different origins of maternal lineages, again pointing towards several independent hybridization events and extensive gene flow. For example, five individuals of Med lineage carry chloroplast from EU lineage, whereas two individuals from EU lineage and three individuals from Med lineage carry CN chloroplast; another four genetically admixed individuals carry the chloroplast of the group intermediate between main genetic groups (Fig. 5).

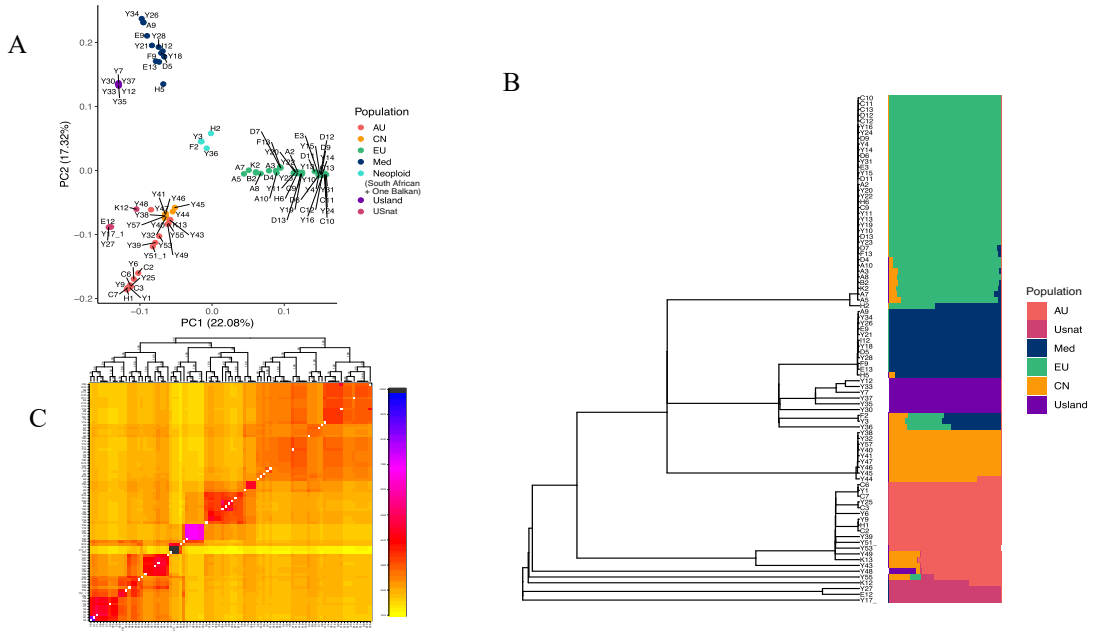

Figure 4 Nuclear phylogeny and population structure of *Phragmites australis* lineages. A) Population structure of *P. australis* samples from principal component analysis based on variants called against the reference genome Y17. B) Left: Maximum likelihood phylogenetic tree estimated from 3,023,618 variants aligned to Y17, with 10% of missing data allowed; Right: fastStructure population assignments based on variants. C) Genetic clusters revealed by fineRADstructure based on genetic similarities among individuals from *de novo* analysis. Darker colors indicate higher level of similarity.

Our data showed a rather clear geographic distribution of each genetic group: USnat lineage is distributed at the border of United States and Canada, whereas USland type is distributed at the Gulf of Mexico; Med lineage are found in Mediterranean regions, North Africa and Gulf of Mexico; CN lineage are mainly located along the Yellow River in China, but one individual was also found in Africa; AU lineage is colonizing Australia, Southern China, Pacific peninsula and islands, Korean peninsula; EU lineage is dominant across Europe, and also the invasive lineage in United States; finally, the Neoploid (South Africa) lineage is distributed in South Africa (Fig. 1). Although we could not obtain the sequences of *trn*T-*trn*L and *rbc*L–*psa*I regions to directly compare the results with

previous studies based on chloroplast haplotypes [15,25,36], we could infer the corresponding haplotypes by the corresponding geographic locations (Table S2). For example, the chloroplast haplotype P which is distributed in Eastern China may represent AU lineage, as they have overlapping geographic occurrences [19,37]. Each lineage was characterized by its own ploidy level: The AU is octoploid, the USland is hexaploid, and the Neoploid (South Africa) is octoploid (Fig. 3). The other lineages such as CN, Med, and most EU individuals are tetraploid.

The geographic distribution of *P. australis* lineages was not precisely concordant with their origins, most likely due to extensive gene flow between the populations or frequent artificial transfer. For example, the nuclear phylogeny showed a subset of North American samples clustering together with a subclade of EU, indicating an artificial introduction from EU to North America [15]. In addition, some introduced individuals in US (sample information obtained from H. B.) were found in several subclusters of the EU lineage suggesting multiple introductions to North America. These individuals spread quickly and are known to be invasive in North America [36]. The US invasive individuals were again introduced to South Africa, which may have initiated another cycle of cryptic invasion in Southern Africa [38]. South Africa is a common shipping route, so it is not surprising that seeds or plants of *Phragmites* have been transported, for example as ballast, from their original locations. Confusing results such as one pure Chinese individual observed in central Africa and characterized as *P. mauritianus* proved to be *P. australis* in this study – this could be an artificial error made during laboratory handling or the transfer happened in reality. One individual from US delta was also clustered with individuals from Australia. Coincidentally haplotype P has been found in southwestern US since the beginning of 20th century [39], hinting that the AU octoploids were also introduced to North America.

We next assessed the population structure in *P. australis* with principal component analysis. The first three principal components explained 52.1% of the data and showed segregation pattern according to geography and ploidy level. The hexaploids largely fell between the tetraploid and octoploid population groups (Fig. 4A). The EU, USland and South African octoploids were separated from the others by the first PCA axis, and the second axis further distinguished Med and AU/CN lineages (Fig. 4A). Although AU and CN were of different ploidy levels, they were not clearly distinguishable from the PCA plot. One rare individual, Y48, with admixture from AU, USland and CN, was located between the AU and USland populations, in agreement with the admixture result (Fig. 4A).

For more accurate quantification of the covariates contributing to genetic variation we carried out redundancy analysis (RDA; Fig. 6)[40,41]. Altogether 56.1% of the total genetic variation were explained by the population groups, whereas ploidy levels accounted for 15.8% of the variation. Geographic coordinates accounted for a small but significant proportion of 1.04% of variance (latitude, p= 0.01; longitude, p= 0.01, respectively) after modeling the effect of population group as covariate. The ploidy levels were confounded within the population groups, and thus the ploidy did not explain any additional variance after including the groups as covariate.

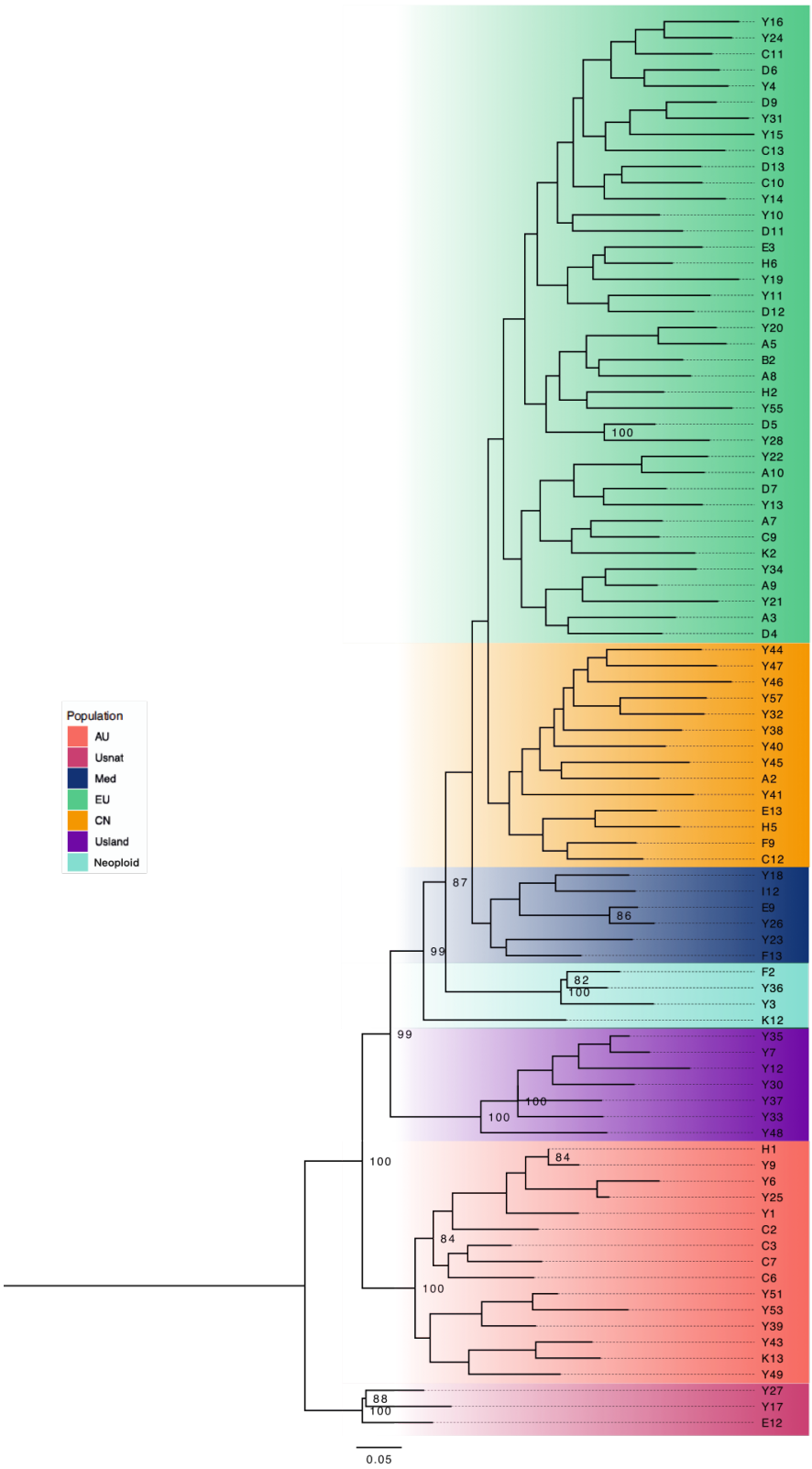

Figure 5 Phylogenetic tree constructed from chloroplast reads subtracted from RAD-seq dataset and mapped to chloroplast genome. Branch colors indicate the genetic groups corresponding to nuclear trees. Number on the node shows bootstrap value from 100 replicates, and only bootstrap value higher than 80 are showed.

The genetic coancestry estimation with fineRADstructure also revealed seven clusters (Fig. 4C), with most of the groups corresponding to the geographic locations, showing high genetic coancestry of AU, USland, Med, South African, USnat, CN and EU lineages. Because the South African individuals formed a group of neopolyploids with similar admixture from the tetraploid CN, EU and Med lineages, we refer the lineage as Neoploid (South Africa) in further analyses.

We next carried out modeling of population divergence times using Approximate Bayesian Computations (ABC) modeling in Fastsimcoal2 by fitting the joint Site Frequency Spectrum (SFS) two-population models. Based on the two-population models, the divergence time of the pairwise populations yield rather concordant order following the phylogenetic tree. To facilitate the visualization of divergence times, we constructed a rough UPGMA tree based on the pairwise divergences between populations. The topology of the UPGMA tree slightly differs from the phylogeny (Fig. 7C) but still provides a viable alternative hypothesis, showing a split between Asian and European populations, with the hexaploid hybrid USland in between. MRCA of USnat and the remaining lineages fell in the range 3.42 - 9.16 Mya, using generation time range of 2 - 4 years (Table 9). Moreover, AU diverged from CN at 1.74 - 3.48 Mya, and EU diverged from Med at 1.64 - 3.28 Mya, and Med diverged from USland at 1.34 - 2.68Mya (Fig. 7 C; Table 9). These populations diverged within the last 1.34 - 3.48 million years in Pliocene and Pleistocene, indicating that the climate change during Pleistocene glaciation cycles may have enhanced lineage divergence and polyploidization.

According to the leading-edge hypothesis, source populations would preserve more genetic diversity than populations in the new territories during the spread of the species [42], and studies on human populations have displayed the so-called serial founder effect, in which the diversity of the populations decrease in proportion to the distance to Africa, due to a series of founding events from diminishing populations [43]. To test whether the spread routes of *Phragmites* populations could be identified in a similar manner, we compared geographically isolated individuals from the same genetic group. For pure EU lineage, nucleotide diversity of the North American populations originating from Europe (0.12726, n = 14) is lower than within European populations (0.17707, n = 16), in support of the known founding event. For AU lineage, the nucleotide diversity of Southeastern China population (0.16872, n = 6) was higher than Australian populations (0.13283, n = 8). However, if only pure individuals of AU lineage were considered, the nucleotide diversity of Chinese population (0.13337, n = 3) was lower than the Australian population (0.16872, n = 6), illustrating the effect of admixture on nucleotide diversity. Tetraploid lineages of *P. australis* showed a decrease of genetic diversity from CN (0.096761, n=8) to EU (0.095245, n=26), Med (0.074062, n=11) and USnat (0.0207, n=3), suggesting the core lineage of *P. australis* originated in temperate grasslands of the former Laurasia (see eg [44] for a reconstruction of a late Miocene vegetation map), split into Asian and European populations and further into Mediterranean population. Based on our phylogenies USnat lineage diverged early from the main Laurasian population, most likely already during the Tortonian period.

Extensive gene flow in Phragmites australis

Genetic divergence between USnat and any other lineage was very high ($F_{st} > 0.44$, Table 8), especially with "Land-type", see [18,45], confirming the distant genetic relationship with the rest of the lineages (Table 8). USland showed exceptionally high divergence with Neoploid ($F_{st} = 0.44689$), but moderate divergence with AU, EU, CN and Med ($0.29 < F_{st} < 0.35$, Table 8). The lowest F_{st} value was observed between Neoploid (South Africa) and EU lineages (0.17126), suggesting recent gene flow (Table 5, 8). We therefore analysed the geneflow using formal tests of introgression.

Population migrations estimated with Treemix yielded a concordant topology with the maximum likelihood phylogeny tree, suggesting one early migration event among the *P. australis*

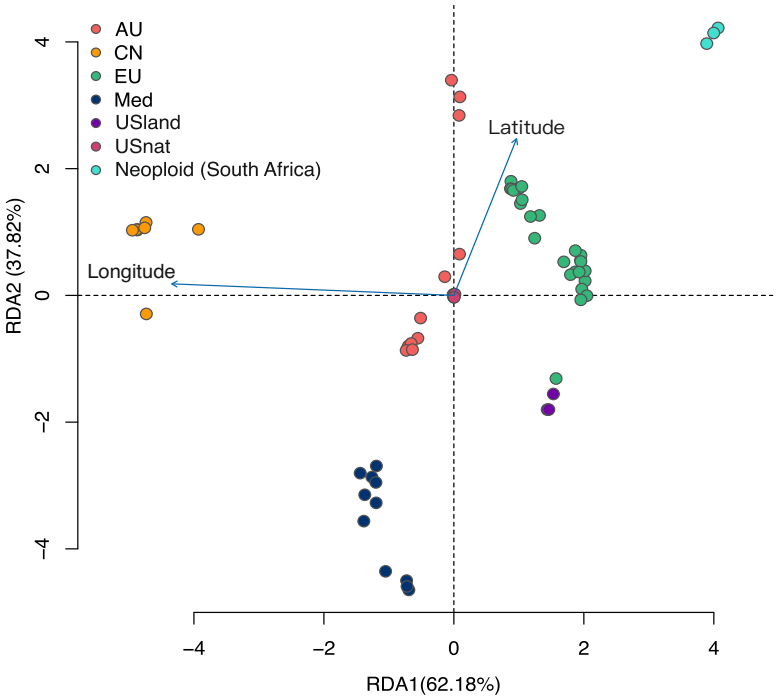

Figure 6 RDA analysis under constraint model using genetic groups as covariates. RDA1 expains 62.18% of the data, and RDA2 explains 37.82% of the data. The arrows lat and long indicated the latitude and longititude of sampling sites of each individual. Genetic group explains 43.31% of the variance and latitude and longititude explains 1.04% of variance.

species (Fig. 7A) from an ancestral lineage not related to USnat to USland lineage (Fig. 7A) as the best model. However, F3 tests of admixture yielded positive values for all seven groups, providing no evidence of hybridization (Table 4). Therefore, gene flow suggested by Treemix is probably from a ghost lineage not sampled in our study (Fig. 7A). In fact, USland lineage, habituated in the Gulf Coast of North America, has been regarded as a hybrid of Mediterranean *P. australis* and *P. mauritianus* [18]. Similar to Mediterranean *P. australis*, the USland lineage shows significantly higher photosynthetic efficacy than EU lineage, which is probably an adaptive mechanism to the climate in its origins at tropical Africa and the Mediterranean area [46]. On the other hand, previous studies have showed that USland lineage holds a 200bp band in its *waxy* gene which specifically exists in *P. australis*, and also a 100bp DNA which is specific in *P. mauritianus* [18], confirming its hybrid status. Hybridization between *P. australis* and *P. mauritianus* has also been observed in Southern Africa in recent decades [38]. Due to the lack of proper outgroup for USland, we were not able to test the gene flow between USland and Med using D statistics. However, USland groups together with Med lineage in nuclear phylogenetic topologies (Fig. 4B; Fig. 7A), and the absolute genetic distance d_{xy} between USland and Med is lower than for any other lineage (Table 8), in support of the hypothesis that Med is probably the second parental lineage of USland.

For the other lineages, from the 40 quartets of lineages compared with the ABBABABA analysis using USnat as outgroup, altogether 30 quartets tested significant (Table 5), suggesting gene flow among most of the lineages (Table 5). The gene flow between the lineages was further confirmed by D statistics for randomly sampled four representative individuals from each lineage (except South Africa lineage); the individual tests confirmed the gene flow results from population-level data but also revealed possible gene flow between Med and AU, and Med and USland (Table S3). The complex pattern of admixture between nearly every population tested suggests extensive global gene

flow between populations; similar conclusion can be drawn from the admixture analysis where 25% of all individuals were found to be admixed, as well as from the incongruence between nuclear and organellar phylogeny. The admixture analyses and ABBABABA tests (Fig. 2B, Table 5) supported extensive hybridization between AU and CN lineages, suggesting the AU lineage may have been widely spread throughout the Asian continent. The CN clade was mainly confined within the Yellow River watershed, so this lineage may have occurred later from the west of China and spread along the

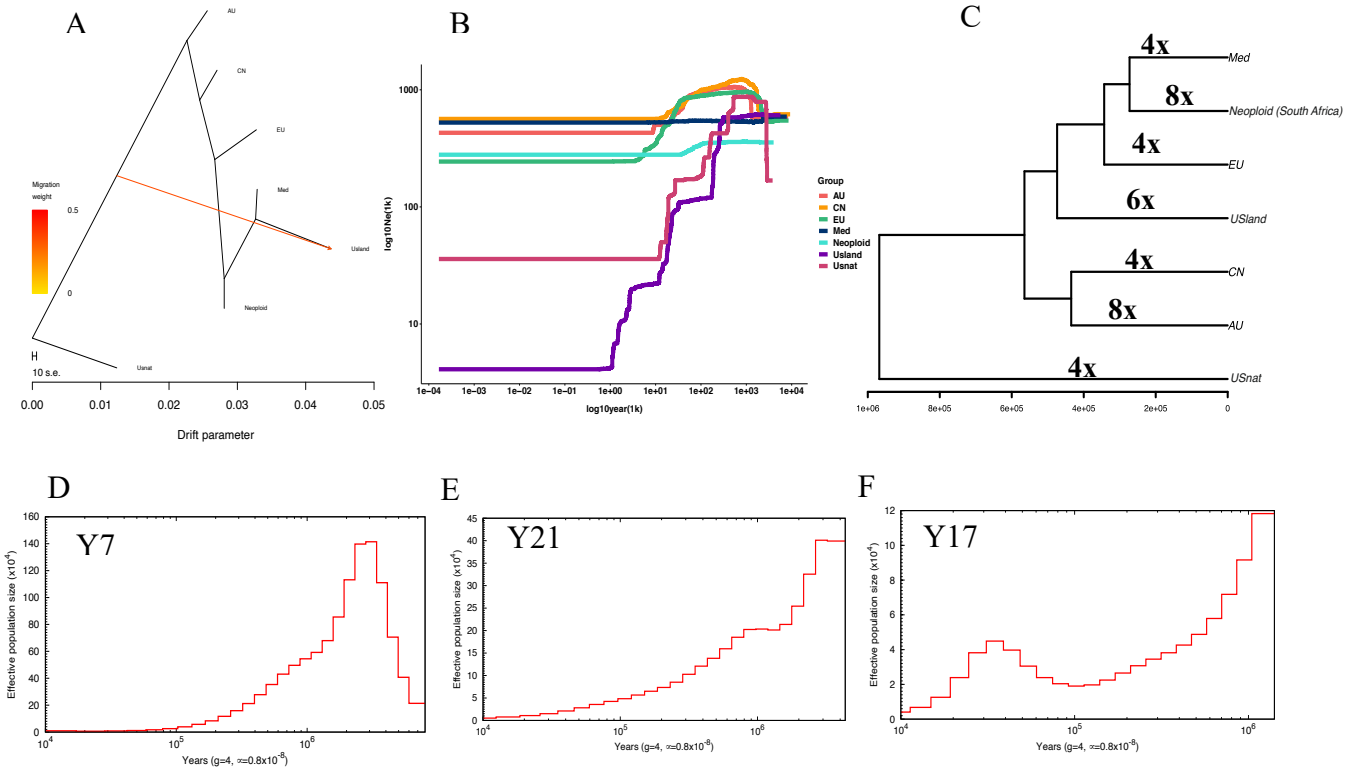

Figure 7 A) Migration events inferred with Treemix. B) Historical effective population sizes of seven *P. australis* lineages estimated with Stairway plot (AU, CN, EU, Med, Neoploid, USland, and USnat). Admixed individuals were excluded from the population analyses, except Neoploid. C) An UPGMA tree estimated from pairwise divergence times inferred with fastsimcoal2. The divergence times are shown as generations, and the number on each brach indicates the ploidy levels . D-F) Population histories for Y7, Y21 and Y17 inferred using PSMC with mutation rate of 6.0277e–09 substitutions per site per year and generation time of 4 years.

river to eastern China.

Hybridization and polyploid diversity

For RADseq data, nucleotide diversity (π) was estimated to be highest in AU lineage (0.099908, n=12) and lowest in USnat lineage (0.0207, n = 3) (Table 6), while other lineages showed intermediate levels of polymorphism (π ranging between 0.067431-0.096761, Table 6). On the other hand, the genome assemblies yielded low nucleotide diversities of neutral sites (π_s) for USnat (Y17; 0.001142), Med (Y21; 0.003121), and higher for USland (Y7; 0.00947) (Table 7). Similar results were obtained for RADseq data when looking at individuals separately (Table 1). The inbreeding F coefficient was positive for all accessions, being highest among USnat lineage and lowest in USland lineage, suggesting a higher level of inbreeding in USnat and relatively more outbreeding in USland, correspondingly (Table 6). Tajima's D value was positive in all populations, being moderately high in AU (0.75 ± 0.85), EU (0.66 ± 1.03), CN (0.575 ± 0.91), Med (0.63 ± 0.98), Neoploid (South Africa) (0.81 ± 0.77) and USnat (0.39 ± 0.94) lineages, and exceptionally high in USland lineage (1.54 ± 1) (Table 6).

Low nucleotide diversity and high Tajima's D are in general hallmarks of a population bottleneck, and we therefore continued by assessing the population history of the species. Using Stairway plot with the site frequency spectrum from RADseq data and four years as generation time, an early population increase was detected between 1-5 Mya for USnat, AU, CN and EU lineages (Fig. 7), followed by a recent decline of effective population size in nearly all the lineages except Med (Fig. 7B). The decline of AU, CN and EU populations started from 50Kya. Population of USland dropped dramatically three times with first one at 500Kya, followed by a drop at 40Kya, and 5Kya. USnat started to decrease monotonously from 700Kya.

Using the same generation time, PSMC analysis on the genome assemblies showed population decreases in Med lineage (Y21) starting from around 3 Mya. In USland the decrease started from 1500Kya, and in USnat from 1000Kya, the latter demonstrating a slight increase at 35Kya followed by another decline (Fig. 7 D-F). Divergence time estimation had showed USland to diverge from Med lineage at 1.34 – 2.68 Mya (Fig. 2C, Table 9), and the PSMC showed the sharp increase of the population size of USland at around 2.5 Mya (Fig. 7E), indicating the hybridization event may happened at a suture zone when the population went through contraction and expansion during glaciation cycles, or then the increase in effective population size may reflect the hybridization event and resulting increase in heterozygosity. USnat on the other hand was also characterized with a population increase at around 5 Mya (Fig. 7B), coincidentally in line with expansion of North American Prairie in late Miocene, suggesting that the climate cooling and drying may have facilitated the lineage divergence[47].

Towards holocene, the USland and Med populations went through an acute decline as seen in PSMC and Stairway plots (Fig. 7D, Fig. 7B) and also from the high Tajima's D value (Table 6).

Genetic diversity suggests transitions in reproductive mode

The decline seen in population histories was unexpected, since the population-level diversities were relatively high, and especially when considering the invasiveness of the expanding EU lineage. We suspected that the results might reflect a break of self-incompatibility and transition into clonal propagation, since the transit of reproductive mode strongly affects genetic diversity regarding to the mutation accumulation, allele frequency alteration, and population size changes as time passes. Population genetic simulations predict that compared to outcrossing and selfing lineages, a clonal lineage experiences relaxed selection and thus accumulates more deleterious mutations results in an overall high level of heterozygosity. However, due to lack of recombination, genotypic diversity in asexual population would be low since genotypes which have accumulated a high number of deleterious alleles, and therefore have reduced fitness, would be purged from the population (the so-called Muller's ratchet). Additionally, clonal populations may undergo frequent bottlenecks, since they can colonize habitats with just a few individuals. Altogether, these effects result in effective population decline and possibly an evolutionary dead-end in the long run [48]. The predictions have been verified empirically in both animals and plants, such as *Primula* species [49], stick insects [50], duckweed [51], bdelloid rotifers[52], and wasps [53], to name a few. Relaxed selection would result in an increased ratio of non-synonymous to synonymous mutations in asexual and selfing proliferation[54,55], but a low rate in outcrossing species because purifying selection would limit the amount of deleterious alleles in outcrossers [48].

P. australis uses both sexual and asexual reproduction mode. Genome evidence pointed towards frequent selfing among *Phragmites* populations, since the genome assemblies demonstrated low levels of heterozygosity and the RAD-seq data yielded high values of inbreeding coefficient. However, seed set rate of *P. australis* is very low (mean 9.7 -12%) due to its partial self-incompatibility [56,57], indicating that asexual reproduction may also be playing an important role. In fact, clonal propagation has been shown to work efficiently in *Phragmites* [58-60]. Nucleotide diversities of deleterious sites (π_n) were Y17 (0.001417), Y21 (0.002784), Y7 (0.008596) (Table 7), and the

corresponding ratios of non-synonymous to synonymous diversity, π_n/π_s, were 1.24 for USnat, 0.89 and 0.91 for Med and USland, respectively. The ratio is very high compared to other plants [61] and suggests relaxed to no selection on deleterious sites[62]. The USnat lineage had the highest ratio of nonsynonymous π_n to synonymous π_s (Table 7). The USnat population has long been isolated from other *Phragmites* lineages, which was also reflected in the high F*st* statistic (Table 8). The site-wise genetic diversity in USnat population is rather high 0.0207 (Table 6) but the heterozygosity of each individual quite low, suggesting the individuals have accumulated mainly private mutations, complying with the predictions of clonal lineage, and these alleles may have fixed into homozygosity due to selfing. The overall diversity is possibly a result of 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 *Phragmites* populations (increased F_{st}). Therefore, we infer that the secluded USnat lineage may rely even more on clonal propagation than the other lineages. In fact, 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 propagation[63]. The combination of selfing and clonal propagation possibly also explains the sharp decline of effective population size of Usnat since 700Kya (Fig. 7B).

Discussion

Intraspecific hybrids give rise to new polyploids at contact zones

Altogether, genetic admixture was commonly seen between the borders of lineages, suggesting low reproductive barriers among lineages. Neoploid lineages occurring in South Africa and Balkans are derived from admixture of the tetraploid Med, EU and CN lineages (Fig. 3, Table 3). The neopolyploids are special in the sense that all the parental lineages are tetraploid, but the resulting offspring are either hexaploids or octoploids. Three out of four individuals from South Africa are admixed by Med, EU and CN tetraploid lineages. All the three samples, located far from each other, are octoploids (Fig. 1), concordant with earlier records [13].

Admixed octoploid individuals were found also in Hungary and Romania, with a slightly different allele frequency pattern from South African populations (Fig. 3 F-G), suggesting this may be an independent polyploidization event caused by intraspecies genomic conflicts. Not all admixed individuals had their chromosomes doubled (Table 3). In fact, only the individuals simultaneously carrying alleles of Med, EU and CN ancestry had higher ploidy levels. These individuals group within EU lineage in both chloroplast and RAD-seq genomic phylogeny, implying recursive hybridization with EU lineage after polyploidization. Beyond this study, other researchers have found octoploids in countries around Mediterranean Sea, including Romania, Greece, Afganistan, Iran, Algeria, Morocco, Tunisia, France, Spain and previous Yugoslavian area [14,64]. These regions overlap with the borders of Med and EU lineage, possibly also the CN lineage, and thus hybridization and polyploidization events are very likely to happen frequently there. Concordantly, intensive investigation has shown the existence of a mixture of octoploids, tetraploids and hexaploids in the Danube Delta, likewise inferred to originate from recent polyploidization and hybridization [7,35]. The hexaploids in Greece and Romania were found to be admixed individuals with tetraploid ancestors. This is supported from previous studies showing seeds from the same inflorescence producing offspring of different ploidy levels, suggesting the hexaploidy may be generated by hybridization of tetraploid and octoploid [65]. The same circumstance occurred in *Spartina,* where the dodecaploid *S. anglica* was resulted from whole genome duplication after the hybridization of two hexaploid species *S. alterniflora* and *S. maritima*, and then hybridized with coexisting parental species *S. alterniflora* again, giving rise to nonaploid individuals [66].

Asexual propagation may contribute to genomic conflict

Autopolyploidization and allopolyploidization have been regarded as the main ways of polyploidization, with the underlying mechanism of how whole genome duplication happens still being largely unknown (Stuessy and Weiss-Schneeweiss 2019). If the parental species are well diverged, the genomic incompatibility after hybridization may give more chance of mistakes in meiosis, thus leading to new polyploids. In this study, neopolyploids were inferred to result as admixture from three lineages of the same species, implying the regulation of genomic networks may be broken down. One possibility is the frequent misregulation due to stressful environmental conditions. Extreme environment has been observed as one way to induce polyploidization, especially autopolyploids. One example is the *Brachypodium distachyon* in Iberian Penisula, where the distribution of diploids and allotetraploids was significantly associated with the dryness of living environment; arid environment may cause problems with cytotype segregation and thus enable the polyploidization in Poaceae species [67]. In *P. australis*, the occurrence of octoploids is mainly distributed in Australia, South Africa, or between Middle Europe and Black Sea, where temperature is high throughout the year. Thus, the highly stressing, dry and hot climate may have caused errors in meiosis resulting in unreduced gametes, and the resulting plasticity may have enhanced their ability to adapt to the new ecological niche [67,68]. However, this conclusion is controversial since an experimental study sampling *Brachypodium* species along the aridity gradient in Israel denied the hypothesis that dryness increases the chance of allopolyploidization [69].

However, since these polyploids were observed mainly in the contact zones, a possible genomic component contributing to the breakup follows from the extensive clonal propagation in *Phragmites,* due to relaxed selection and accumulation of deleterious alleles. Long term accumulation in allopatric populations may have led into increased genomic divergence such that intraspecific genomic conflicts arise among progeny from these diverged populations. This could be due to misregulation during meiosis; since sexual propagation is of lower importance in a predominantly clonally propagating species, the genes associated with it are not under strict purifying selection. Altogether, we suggest that decreasing population sizes in *Phragmites* first led into adopting selfing as propagation strategy. This resulted in further decline of population sizes and perhaps widespread inbreeding depression, and therefore clonal propagation was adopted as a solution. The accumulation of deleterious alleles then eventually results in genome incompatibility and polyploid hybrids are formed when long diverged populations encounter at contact zones.

Methods

DNA extraction, RAD sequencing and whole genome sequencing

A total of 88 *P. australis* individuals were obtained from Eurasian, North American, Oceanian and African continents (Fig. 1, Table 1) and planted in Aarhus University and Shandong University. Some of the individuals have already been used in previous phylogeographic studies, and thus we were able to link them to the defined lineages in other studies (see more in Table 1). One *Arundo donax* (giant reed) individual was included as outgroup. For RAD sequencing, DNA was extracted from fresh leaves using CTAB method. The leaves were ground into powder and mixed with 500 µl extraction buffer (containing 100mM Tris pH8, 1.4 M NaCl, 20mM EDTA, 2% CTAB) and 2% β-mercaptoethanol. After incubation at 65 °C for 10 minutes, the suspension was mixed with 500 µl of chloroform and spun at 10000 rpm for 5min. Then, 500 µl of chloroform was added to the supernatant and spun at 10000 rpm for 10 minutes, and finally 800 µl ice cold ethanol was added. The tube was left in the freezer overnight and subsequently spun at 13000 rpm at 4°C for 20 minutes, followed by washing with 500 µl 70% ethanol. The pellet was precipitated at 6000rpm at 4 °C, and let dry and dissolve in 200 µl H_2O. Paired-end RAD-seq library preparation and sequencing was completed by

Shanghai Honsunbio Limited company (http://www.honsunbio.com/), using Illumina HiSeqX10. Whole genome sequences of *Oropetium thomaeum* (genome accession: Phytozome [70].), *Oryza sativa* (genome accession: Phytozome), and *Miscanthus sacchariflorus* (genome assembly accession: GCA_002993905.1, NCBI) were used as outgroups. Illumine reads of *Oropetium thomaeum* (SRR2083762), *Miscanthus sinensis* (SRR486617), *Arundo donax* (SRR4319201), *Arundo plinii* (SRR4319202) and *Sorghum bicolor* (SRR12628364) were used to construct ancestral alleles of *Phragmites* lineages.

Draft genome assembly

Based on initial analyses with RAD-seq data, 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 90x high coverage (assuming 1 Gb genome size) by NovogeAIT at Singapore using Novaseq platform. After receiving the sequencing data, the genome size was first estimated using Kmergenie [71]. Then, after quality assessment of the reads using FastQC, the three genomes were assembled using MaSuRCA assembler [72] with default settings. The assembly was passed through purge haplotigs v1.0.4 [73] to remove the redundant haplotigs resulted from the highly heterozygous regions. The completeness of the genome assemblies was assessed using BUSCO v5 [74]. and odb_database version 10, and the quality of the assembly was estimated using Quast [75]. The draft genome assemblies were next filtered using Purge Haplotigs pipeline to remove heterozygous genome regions which may have been accidentally assembled into two haplotigs [76].

Gene prediction

Gene prediction for the three draft genome assemblies was performed by combining homology - based method and *ab initio* methods. The draft genomes were then searched against the Poaceae grasses *O. thomaeum* and *M. sinensis* to obtain the annotation information based on homology prediction. *Ab initio* gene prediction was performed using GeneMark-ES [77], BRAKER2 [78], PASA [79] and AUGUSTUS [80]. RNAseq data of *P. australis* was obtained from NCBI (accession number: SAMN04544298), aligned to each draft assemblies by STAR aligner [81], and used as evidence to define intron borders in BRAKER2 prediction. In addition, the RNAseq data was also assembled *de novo* using TRINITY [82] and used in PASA for generating a high quality dataset for downstream *ab initio* gene predictions. Finally, all evidence of gene prediction was combined in EvidenceModeler [31] to get consensus gene model predictions. The completeness of predicted gene models was assessed using BUSCO by searching against poales_odb10 protein database. Significance of overlap between duplicated BUSCOs in different genome assemblies were tested with Fisher's exact test using R package "GeneOverlap"[83].

Syntenic blocks and divergence of subgenomes

The three draft genomes were each aligned against *O. thomaeum* genome (Genome ID 51527 in CoGe database, https://genomeevolution.org/coge) using SynMap tool in CoGe [84] that identifies syntenic blocks. The Ks (synonymous substitutions rate) rates for the syntenic genes were then downloaded from CoGe and syntenic blocks were phased into two subgenomes based on the average Ks values throughout the block.

To date the time of WGD in relation to speciation, we inspected BUSCO genes that were present in rice (*O. sativa*), *O. thomaeum*, Y7, Y17 and Y21. *Oryza sativa* and *O. thomaeum* were used as outgroups. For polyploids, multi-copy genes sourced from several polyploidizations often result in incongruent phylogeny. Single copy orthologs are better choice to reflect the species evolution due

to less complicated history, and allopolyploidization events could be traced by analyzing the two copies of paralogous gene sets belonging to their own parental species [85]. Following this idea, we used the BUSCOs which were present as a single copy in the outgroups and duplicated in Y7, Y17 and Y21 to find out the divergence time of the subgenomes. These BUSCOs were considered to be the orthologs in all species, but paralogs in *Phragmites* species, representing the two ancestral subgenomes. In total 580 such BUSCOs among *O. sativa*, O. *thomaeum* and *Phragmites* lineages were obtained. From this set, we randomly selected 25 BUSCOs which showed coherent topology with species tree and estimated divergence time of *Phragmites* subgenomes using BEAST2 with calibrated Yule model. A site model of HKY+Gamma was used, with four categories of Gamma distribution, and the molecular clock was set to be random local clock which allows certain variation of clock rate among lineages. Based on fossil record in Oryzeae and calibration point of BOP and PACMAD lineages [86,87], we set the 95% confidence interval of total most recent common ancestor (tmrca) to be 65.6 -73.1Mya, by assigning a lognormal prior, with an offset 65.0, a mean of 3.0, and a standard deviation of 0.8 (Mean In Real Space). The total MCMC sampling chain length was 10 million iterations, and the states were stored every 5000 samples. The divergence estimation was run twice with the same parameter to guarantee the convergence, with each run achieving ESS value of all parameters to be higher than 200.

Identification of putative deleterious mutations

The nucleotide diversites 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 mem to obtain the variants along the genome, sorted with Samtools [88], and genome coverage variants were called using bcftools [88]. Site-wise diversities were estimated using ANGSD [89] and used as input in the R package PiNSiR. The ratio of nonsynonymous/synonymous mutations of each subgenome was performed using the same method and esimtating the diversity using the coordinates of the syntenic regions with *O. thomaeum.*

Variant calling

RAD-tag sequences were processed following Stacks pipeline refmap methods by aligning the reads to the reference genome Y17 [28]. We first demultiplexed the reads and removed the barcode using process_radtags. In the refmap pipeline, the paired reads were aligned to Y17 assembly representing USnat lineage using Bowtie2 [90]. The RAD-seq data was also aligned to the already published chloroplast assembly (accession number: KJ825856), so as to draw the genetic information carried by organelles, and to compare the evolutionary history between nuclear and chloroplast genome.

Ploidy level prediction

Although the ploidy level of most individuals was known from flow cytometry, the knowledge of genome size of all individuals would give a clearer picture about the evolutionary path of polyploidization. To find the link between genomic alleles and ploidy levels, we developed a method that predicts ploidy level by counting the proportions of alternative alleles based on the reference guided aligned reads. After mapping the short reads to the reference genome, similarly but differently from ploidyNGS [10], we count the number of alleles that are different from 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. The number of chromosome copies were predicted by counting the number of peaks of alternative alleles. For example, in diploids harboring 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 a tetraploid individual, the alternative allele proportion should show peaks at 0.25,0.5,0.75, and at 0.167, 0.33, 0.5, 0.67 and, 0.83 in hexaploids etc. Based on this pattern, we predicted the ploidy levels by plotting the density of alternative allele frequency, and verified the predictions using the flow cytometry data. We performed the analysis by counting the read depth of alternative alleles in bam files after aligning the filtered reads to the reference genome Y17.

Phylogeny

For further analyses a filtered set of high quality SNPs was obtained by selecting SNPs that are present in >90% of the samples using vcftools [91]. Phylogeny was estimated for the set of filtered SNPs using RAxML [92] under the GTRGAMMA substitution model with 100 bootstrap replicates. The final tree was annotated and viewed in Figtree (http://tree.bio.ed.ac.uk/software/figtree/). Variants from chloroplast were also used for estimating the phylogenetic tree using same methods.

Population structure and reticulate evolution

The population structure was estimated on the whole dataset using fastSTRUCTURE with K ranging between 1…15 [93], best K was selected by the highest marginal likelihood. To further evaluate the genetic ancestry of each independent allele, we reduce the effect of linkage disequilibrium by selecting a random SNP from each stacks loci, and removing ninety percent of the missing data were removed using vcftools v 0.1.16 [91], and filtered out the data with minor allele frequency less than 0.05 using Plink 1.9 [94]. The processed data was used to performed genetic clustering using fineRADstructure [95].

Principal component analysis (PCA) was calculated using Plink 1.9 [94] and plotted in R to reveal the ordination of the genetic data. In addition, we performed Redundancy Analysis (RDA) using R package "vegan" [96] to quantify the proportion of the genetic variation explained by covariates such as different lineages, geographical location and ploidy level, with 100 permutations to assess the importance of fitted models.

To understand the genomic structure of each lineages, we grouped individuals to populations based on fastSTRUCTURE and phylogeny results (Table 4). Inbreeding *F* coefficients of each individual were calculated using Plink v1.07 [94], and group mean and standard deviation group were measured. Nucleotide diversity (pi), private alleles (N_A) and genetic divergence between populations (F_{st} and d_{xy}) were calculated using populations program in Stacks [97]. F coefficient for each individual was calculated using Plink 1.9 [94], and the nucleotide diversity (pi) of each individual was measured using vcftools v 0.1.16 with 1000bp window size [91]. To test for admixture, we performed f3 tests using Admixtools 2 [98]. ABBABABA were considered to be significant only when the absolute value of Z score is higher than 3. To test the potential gene flow between lineages, we selected four pure individuals from each lineage (except for USnat lineage where one individual was used as an outgroup) and performed ABBABABA statistics among all combinations using AdmixTools v7.0.1 [98]. Tajima's D of each population was measured using vcfkit [99] using non-overlapping sliding windows of size 10,000 bp to infer whether population demographics has been heterogenous along the genome.

Population demographics in history

Site frequency spectrum (SFS) of derived alleles were used in Stairway plot [100] to infer the historical population size changes. To further validate our inference about demographic changes, we aligned the reads of the three whole-genome-sequenced individuals to their own reference genome using bwa [101], and performed demographic analysis using PSMC [102]. Although it usually takes 1 year for *P. australis* to sprout and mature (Phragmites resource, 1989:15), we set the generation time to 2-4 years because it is perennial. We set the mutation rate to be the Poaceae silent-site mutation rate, 6.0277e–

09 substitutions per site per year [103]. Number of migration events between *P. australis* lineages were tested using Treemix [104]. All pure individuals belonging to the lineages were included in the analysis, and SNPs were filtered to only include the ones that appear in all populations. We analyzed 1- 5 migration events, and each scenario was tested with 5 replicates. The optimal number of migration were determined using R package "optM" based on the second-order rate of change in likelihood [105].

To determine the derived alleles in *Phragmites* species, we aligned the genome reads of *O. thomaeum*, *A. donax*, *A. plinii*, *S. bicolor* and *M. sinensis* to Y17 using bwa mem, and then obtained consensus ancestral genome using ANGSD [89]. Ancestral alleles of the variants called by refmap pipeline were called and annotated using samtools and bcftools. The multidimensional unfolded Site Frequency Spectrum (SFS) for all seven lineages was calculated using easySFS program (https://github.com/isaacovercast/easySFS). We selected the optimal projection by balancing the number of segregating sites and adequate number of individuals, resulting between 933 and 2895 segregating sites per population group. The parameters with regard to population splitting were estimated with fastsimcoal version 2.6 [106], and mainly focused on the divergence time between populations. In each cycle, the program run for 200,000 iterations to estimate the expected SFS and conduct 40 optimization (ECM) cycles to estimate the parameters. To find the best fitting parameter, we performed 100 runs for each scenario, and selected the best run by the highest likelihood. We used a mutation rate of $6.0277E^{-09}$ site/year [103] as a standard, assuming two to four years of generation time for *Phragmites* species.

Acknowledgements

This work was supported by International Postdoctoral Exchange Fellowship Program of China Postdoctoral Science Foundation (C.W.), the National Natural Science Foundation of China (Grant No. 31770361 to WH. G, Grant No. 31800299 to T. W., and Nanyang Technological University startup grant and Academy of Finland (decisions 318288 and 319947) 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. Finally, we want to acknowledge CSC–IT Center for Science, Finland, and NTU HPCC, Singapore, for computational resources.

Data Availability Statements

The data underlying this article are available in NCBI SRA database and can be accessed with the BioProject ID PRJNA753984.

Author Contributions

C.W., WH.G. and J.S. conceived the study; H.B., LL.L., F.E. and T.W. maintained and collected the samples as well as provided sampling information; C.W., LL.L. and MQ.Y. performed DNA extraction; C.W. and J.S. lead the data analysis; C.W., J.S., WH.G., LL.L., F.E., H.B., T.W., and MQ.Y. contributed to the writing.

Competing Interests statement

The authors declared no competing interests.

References

1. Alix, K., Gérard, P.R., Schwarzacher, T. & Heslop-Harrison, J. Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. *Annals of Botany* **120**, 183-194 (2017).
2. Estep, M.C. *et al.* Allopolyploidy, diversification, and the Miocene grassland expansion. *Proceedings of the National Academy of Sciences* **111**, 15149-15154 (2014).

3. Salse, J. *et al.* Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *The Plant Cell* **20**, 11-24 (2008).
4. Levy, A.A. & Feldman, M. The impact of polyploidy on grass genome evolution. *Plant physiology* **130**, 1587-1593 (2002).
5. Brysting, A.K., Fay, M.F., Leitch, I.J. & Aiken, S.G. One or more species in the arctic grass genus Dupontia?–a contribution to the Panarctic Flora project. *Taxon* **53**, 365-382 (2004).
6. Abdeddaim-Boughanmi, K. *et al.* A single species, two basic chromosomal numbers: case of Lygeum spartum (Poaceae). *Plant Biosystems-An International Journal Dealing with all Aspects of Plant Biology* **153**, 775-783 (2019).
7. Raicu, P., Staicu, S., Stoian, V. & Roman, T. The Phragmites communis Trin. chromosome complement in the Danube delta. *Hydrobiologia* **39**, 83-89 (1972).
8. Zhang, J. *et al.* Variation in ploidy level and genome size of Cynodon dactylon (L.) Pers. along a latitudinal gradient. *Folia Geobotanica*, 1-12 (2020).
9. Salojärvi, J. *et al.* Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nature genetics* **49**, 904-912 (2017).
10. dos Santos, R.A.C., Goldman, G.H. & Riaño-Pachón, D.M. ploidyNGS: Visually exploring ploidy with next generation sequencing data. *bioRxiv*, 086488 (2016).
11. Margarido, G.R. & Heckerman, D. ConPADE: genome assembly ploidy estimation from next-generation sequencing data. *PLoS Comput Biol* **11**, e1004229 (2015).
12. Weiß, C.L., Pais, M., Cano, L.M., Kamoun, S. & Burbano, H.A. nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC bioinformatics* **19**, 1-8 (2018).
13. Clevering, O.A. & Lissner, J. Taxonomy, chromosome numbers, clonal diversity and population dynamics of Phragmites australis. *Aquatic Botany* **64**, 185-208 (1999).
14. Connor, H., Dawson, M., Keating, R. & Gill, L. Chromosome numbers of Phragmites australis (Arundineae: Gramineae) in New Zealand. *New Zealand Journal of Botany* **36**, 465-469 (1998).
15. Saltonstall, K. Cryptic invasion by a non-native genotype of the common reed, Phragmites australis, into North America. *Proceedings of the National Academy of Sciences* **99**, 2445-2449 (2002).
16. Tanaka, T.S., Irbis, C. & Inamura, T. Phylogenetic analyses of Phragmites spp. in southwest China identified two lineages and their hybrids. *Plant Systematics and Evolution* **303**, 699-707 (2017).
17. Lambertini, C. *et al.* A phylogeographic study of the cosmopolitan genus Phragmites (Poaceae) based on AFLPs. *Plant Systematics and Evolution* **258**, 161-182 (2006).
18. Lambertini, C. *et al.* Tracing the origin of Gulf Coast Phragmites (Poaceae): A story of long-distance dispersal and hybridization. *American Journal of Botany* **99**, 538-551 (2012).
19. Liu, L.L. *et al.* Cryptic lineages and potential introgression in a mixed-ploidy species (Phragmites australis) across temperate China. *Journal of Systematics and Evolution* (2020).
20. Greuter, W. & Scholz, H. Phragmites in Crete, Cenchrus frutescens, and the nomenclature of the common reed (Gramineae). *Taxon* **45**, 521-523 (1996).
21. Peterson, K.S.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).
22. Saltonstall, K., Peterson, P.M. & Soreng, R.J. RECOGNITION OF PHRAGMITES AU STRAUS SUBSP. AMERICANUS (POACEAE: ARUNDINOIDEAE) IN NORTH AMERICA: EVIDENCE FROM MORPHOLOGICAL AND GENETIC ANALYSES. *SIDA, Contributions to Botany*, 683-692 (2004).

23. Saltonstall, K. The naming of Phragmites haplotypes. *Biological Invasions* **18**, 2433-2441 (2016).
24. Colin, R. & Eguiarte, L.E. Phylogeographic analyses and genetic structure illustrate the complex evolutionary history of Phragmites australis in Mexico. *American journal of botany* **103**, 876-887 (2016).
25. An, J.X., Wang, Q., Yang, J. & LIU, J.Q. Phylogeographic analyses of Phragmites australis in China: native distribution and habitat preference of the haplotype that invaded North America. *Journal of Systematics and Evolution* **50**, 334-340 (2012).
26. Lambertini, C. Why are tall-statured energy grasses of polyploid species complexes potentially invasive? A review of their genetic variation patterns and evolutionary plasticity. *Biological Invasions*, 1-23 (2019).
27. Godfree, R.C., Marshall, D.J., Young, A.G., Miller, C.H. & Mathews, S. Empirical evidence of fixed and homeostatic patterns of polyploid advantage in a keystone grass exposed to drought and heat stress. *Royal Society open science* **4**, 170934 (2017).
28. Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A. & Cresko, W.A. Stacks: an analysis tool set for population genomics. *Molecular ecology* **22**, 3124-3140 (2013).
29. Meyerson, L.A., Lambertini, C., McCormick, M.K. & Whigham, D.F. Hybridization of common reed in North America? The answer is blowing in the wind. *AoB plants* **2012**(2012).
30. Pyšek, P. *et al.* Small genome separates native and invasive populations in an ecologically important cosmopolitan grass. *Ecology* **99**, 79-90 (2018).
31. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, R7 (2008).
32. Hardion, L. *et al.* Does infraspecific taxonomy match species evolutionary history? A phylogeographic study of Arundo formosana (Poaceae). *Botanical Journal of the Linnean Society* **183**, 236-249 (2017).
33. 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).
34. Jones, N. & Pašakinskienė, I. Genome conflict in the gramineae. *New Phytologist* **165**, 391-410 (2005).
35. Lambertini, C. *et al.* Revisiting Phragmites australis variation in the Danube Delta with DNA molecular techniques. in *International Conference Proceedings: Water Resources and Wetlands* 14-16 (2012).
36. Saltonstall, K. Microsatellite variation within and among North American lineages of Phragmites australis. *Molecular Ecology* **12**, 1689-1702 (2003).
37. Lambertini, C. *et al.* Phylogenetic diversity shapes salt tolerance in Phragmites australis estuarine populations in East China. *Scientific reports* **10**, 1-12 (2020).
38. Canavan, K., Paterson, I.D., Lambertini, C. & Hill, M.P. Expansive reed populations—alien invasion or disturbed wetlands? *AoB Plants* **10**, ply014 (2018).
39. Lambert, A.M., Saltonstall, K., Long, R. & Dudley, T.L. Biogeography of Phragmitesaustralis lineages in the southwestern United States. *Biological invasions* **18**, 2597-2617 (2016).
40. Israels, A.Z. Redundancy analysis for qualitative variables. *Psychometrika* **49**, 331-346 (1984).
41. Capblancq, T., Luu, K., Blum, M.G. & Bazin, E. Evaluation of redundancy analysis to identify signatures of local adaptation. *Molecular Ecology Resources* **18**, 1223-1233 (2018).

42. Hewitt, G.M. Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological journal of the Linnean Society* **58**, 247-276 (1996).
43. DeGiorgio, M., Jakobsson, M. & Rosenberg, N.A. Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proceedings of the National Academy of Sciences* **106**, 16057-16062 (2009).
44. Pound, M.J. *et al.* A Tortonian (late Miocene, 11.61–7.25 Ma) global vegetation reconstruction. *Palaeogeography, Palaeoclimatology, Palaeoecology* **300**, 29-45 (2011).
45. Liu, L. *et al.* Genetic and epigenetic changes during the invasion of a cosmopolitan species (Phragmites australis). *Ecology and evolution* **8**, 6615-6624 (2018).
46. Nguyen, L.X. *et al.* Photosynthesis of co-existing Phragmites haplotypes in their non-native range: are characteristics determined by adaptations derived from their native origin? *AoB Plants* **5**(2013).
47. Retallack, G.J. Neogene expansion of the North American prairie. *Palaios* **12**, 380-390 (1997).
48. Glemin, S. & Galtier, N. Genome evolution in outcrossing versus selfing versus asexual species. *Evolutionary Genomics*, 311-335 (2012).
49. Wang, X.-J. *et al.* The Genomic Selfing Syndrome Accompanies the Evolutionary Breakdown of Heterostyly. *Molecular biology and evolution* **38**, 168-180 (2021).
50. Bast, J. *et al.* Consequences of asexuality in natural populations: insights from stick insects. *Molecular biology and evolution* **35**, 1668-1677 (2018).
51. Ho, E.K., Bartkowska, M., Wright, S.I. & Agrawal, A.F. Population genomics of the facultatively asexual duckweed Spirodela polyrhiza. *New Phytologist* **224**, 1361-1371 (2019).
52. Barraclough, T.G., Fontaneto, D., Ricci, C. & Herniou, E.A. Evidence for inefficient selection against deleterious mutations in cytochrome oxidase I of asexual bdelloid rotifers. *Molecular Biology and Evolution* **24**, 1952-1962 (2007).
53. Tvedte, E.S., Ward, A.C., Trendle, B., Forbes, A.A. & Logsdon, J.M. Genome evolution in a putatively asexual wasp. *bioRxiv* (2020).
54. Henry, L., Schwander, T. & Crespi, B.J. Deleterious mutation accumulation in asexual Timema stick insects. *Molecular Biology and Evolution* **29**, 401-408 (2012).
55. Lovell, J.T., Williamson, R.J., Wright, S.I., McKay, J.K. & Sharbel, T.F. Mutation accumulation in an asexual relative of Arabidopsis. *PLoS genetics* **13**, e1006550 (2017).
56. Ishii, J. & Kadono, Y. Factors influencing seed production of Phragmites australis. *Aquatic Botany* **72**, 129-141 (2002).
57. Tachibana, Y. Phragmites communis Trin. in Lake Biwa. *Bull. Water Plant Soc. Japan* **18**, 2-6 (1984).
58. Amsberry, L., Baker, M.A., Ewanchuk, P.J. & Bertness, M.D. Clonal integration and the expansion of Phragmites australis. *Ecological applications* **10**, 1110-1118 (2000).
59. 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).
60. 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).
61. Chen, J., Glémin, S. & Lascoux, M. Genetic diversity and the efficacy of purifying selection across plant and animal species. *Molecular Biology and Evolution* **34**, 1417-1428 (2017).

62. Johnson, S.G. & Howard, R.S. Contrasting patterns of synonymous and nonsynonymous sequence evolution in asexual and sexual freshwater snail lineages. *Evolution: International Journal of Organic Evolution* **61**, 2728-2735 (2007).
63. 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. *Journal of Ecology* **103**, 1529-1537 (2015).
64. Gorenflot, R. Degres et niveaux de la variation du nombre chromosomique chez Phragmites australis (Cav.) Trin. ex Steud. *Veroff. Geobot. Inst. ETH, Stiftung Rubel, Zurich* **87**, 53-65 (1986).
65. Paucã-Comãnescu, M., Clevering, O.A., Hanganu, J. & Gridin, M. Phenotypic differences among ploidy levels of Phragmites australis growing in Romania. *Aquatic Botany* **64**, 223-234 (1999).
66. Renny-Byfield, S. *et al.* Flow cytometry and GISH reveal mixed ploidy populations and Spartina nonaploids with genomes of S. alterniflora and S. maritima origin. *Annals of Botany* **105**, 527-533 (2010).
67. Manzaneda, A.J. *et al.* Environmental aridity is associated with cytotype segregation and polyploidy occurrence in Brachypodium distachyon (Poaceae). *New phytologist* **193**, 797-805 (2012).
68. Stuessy, T. & Weiss-Schneeweiss, H. What drives polyploidization in plants? *The New phytologist* **223**, 1690 (2019).
69. Achenbach, L., Lambertini, C. & Brix, H. Phenotypic traits of Phragmites australis clones are not related to ploidy level and distribution range. *AoB Plants* **2012**(2012).
70. Goodstein, D.M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic acids research* **40**, D1178-D1186 (2012).
71. Chikhi, R. & Medvedev, P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **30**, 31-37 (2014).
72. Zimin, A.V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669-2677 (2013).
73. Roach, M.J., Schmidt, S.A. & Borneman, A.R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC bioinformatics* **19**, 1-10 (2018).
74. 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).
75. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-1075 (2013).
76. Roach, M.J., Schmidt, S.A. & Borneman, A.R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC bioinformatics* **19**, 460 (2018).
77. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic acids research* **33**, 6494-6506 (2005).
78. Hoff, K.J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: incorporating protein homology information into gene prediction with GeneMark-EP and AUGUSTUS. *Plant and Animal Genomes XXVI* (2018).
79. Haas, B.J., Zeng, Q., Pearson, M.D., Cuomo, C.A. & Wortman, J.R. Approaches to fungal genome annotation. *Mycology* **2**, 118-141 (2011).
80. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637-644 (2008).
81. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

82. Grabherr, M.G. *et al.* Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology* **29**, 644 (2011).
83. Shen, L. GeneOverlap: An R package to test and visualize gene overlaps. *R Package* (2014).
84. Lyons, E. & Freeling, M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal* **53**, 661-673 (2008).
85. Brysting, A.K., Mathiesen, C. & Marcussen, T. Challenges in polyploid phylogenetic reconstruction: A case story from the arctic-alpine Cerastium alpinum complex. *Taxon* **60**, 333-347 (2011).
86. Christin, P.-A. *et al.* Molecular dating, evolutionary rates, and the age of the grasses. *Systematic biology* **63**, 153-165 (2014).
87. Prasad, V. *et al.* Late Cretaceous origin of the rice tribe provides evidence for early diversification in Poaceae. *Nature communications* **2**, 1-9 (2011).
88. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
89. Korneliussen, T.S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC bioinformatics* **15**, 1-13 (2014).
90. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357 (2012).
91. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).
92. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690 (2006).
93. Raj, A., Stephens, M. & Pritchard, J.K. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**, 573-589 (2014).
94. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics* **81**, 559-575 (2007).
95. Malinsky, M., Trucchi, E., Lawson, D.J. & Falush, D. RADpainter and fineRADstructure: population inference from RADseq data. *Molecular biology and evolution* **35**, 1284-1290 (2018).
96. Oksanen, J. *et al.* Vegan: community ecology package. R package version 1.17-4. *http://cran. r-project. org>. Acesso em* **23**, 2010 (2010).
97. Rochette, N.C., Rivera-Colón, A.G. & Catchen, J.M. Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular ecology* **28**, 4737-4754 (2019).
98. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-1093 (2012).
99. Cook, D.E. & Andersen, E.C. VCF-kit: assorted utilities for the variant call format. *Bioinformatics* **33**, 1581-1582 (2017).
100. Liu, X. & Fu, Y.-X. Exploring population size changes using SNP frequency spectra. *Nature genetics* **47**, 555 (2015).
101. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* **25**, 1754-1760 (2009).
102. Li, H. & Durbin, R. Inference of human population history from whole genome sequence of a single individual. *Nature* **475**, 493 (2012).
103. 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. *Molecular Biology and Evolution* **34**, 1363-1377 (2017).

104. Pickrell, J. & Pritchard, J. Inference of population splits and mixtures from genome-wide allele frequency data. *Nature Precedings*, 1-1 (2012).
105. RRs, F. optM: an R package to optimize the number of migration edges using threshold models. *Journal of Heredity* (2019).
106. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet* **9**, e1003905 (2013).