



# Genomic Insights into Genetic Diploidization in the Homosporous Fern *Adiantum nelumboides*

Yan Zhong <sup>1,†</sup>, Yongbo Liu <sup>2,†</sup>, Wei Wu<sup>3</sup>, Jingfang Chen<sup>1</sup>, Chenyu Sun<sup>1</sup>, Hongmei Liu<sup>4</sup>, Jiangping Shu<sup>5</sup>, Atsushi Ebihara<sup>6</sup>, Yuehong Yan<sup>5,\*</sup>, Renchao Zhou<sup>1,\*</sup>, and Harald Schneider<sup>4,\*</sup>

<sup>1</sup>State Key Laboratory of Biocontrol and Guangdong Provincial Key Laboratory of Plant Resources, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China

<sup>2</sup>State Environmental Protection Key Laboratory of Regional Eco-process and Function Assessment, Chinese Research Academy of Environmental Sciences, 8 Dayangfang, Beijing 100012, China

<sup>3</sup>College of Horticulture and Landscape Architecture, Zhongkai University of Agriculture and Engineering, Guangzhou, China

<sup>4</sup>Center for Integrative Conservation, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Yunnan, China

<sup>5</sup>Key Laboratory of National Forestry and Grassland Administration for Orchid Conservation and Utilization, and the Orchid Conservation and Research Centre of Shenzhen, Shenzhen, China

<sup>6</sup>Department of Botany, National Museum of Nature and Science, Tsukuba, Japan

<sup>†</sup>These authors contributed equally to this work.

\*Corresponding authors: E-mails: yhyan@sibs.ac.cn (Y.Y.), zhrench@mail.sysu.edu.cn (R.Z.), harald@xtbg.ac.cn (H.S.).

Accepted: 28 July 2022

## Abstract

Whole-genome duplication has been recognized as a major process in speciation of land plants, especially in ferns. Whereas genome downsizing contributes greatly to the post-genome shock responses of polyploid flowering plants, diploidization of polyploid ferns diverges by maintaining most of the duplicated DNA and is thus expected to be dominated by genic processes. As a consequence, fern genomes provide excellent opportunities to study ecological speciation enforced by expansion of protein families via polyploidy. To test the key predictions of this hypothesis, we reported the de novo genome sequence of *Adiantum nelumboides*, a tetraploid homosporous fern. The obtained draft genome had a size of 6.27 Gb assembled into 11,767 scaffolds with the contig N50 of 1.37 Mb. Repetitive DNA sequences contributed with about 81.7%, a remarkably high proportion of the genome. With 69,568, the number of predicted protein-coding genes exceeded those reported in most other land plant genomes. Intragenomic synteny analyses recovered 443 blocks with the average block size of 1.29 Mb and the average gene content of 16 genes. The results are consistent with the hypothesis of high ancestral chromosome number, lack of substantial genome downsizing, and dominance of genic diploidization. As expected in the calciphilous plants, a notable number of detected genes were involved in calcium uptake and transport. In summary, the genome sequence of a tetraploid homosporous fern not only provides access to a genomic resource of a derived fern, but also supports the hypothesis of maintenance of high chromosome numbers and duplicated DNA in young polyploid ferns.

**Key words:** genome assembly, ecological adaptation, whole-genome duplication, homeologous chromosome pairing, diploidization.

## Significance

Unlike flowering plants, polyploid ferns maintain high chromosome numbers and duplicated DNA. Thus, their diploidization is expected to be dominated by genic processes instead by processes enabling rapid reducing of the duplicated genome components. By sequencing and characterizing the whole genome of a tetraploid homosporous fern, *Adiantum nelumboides*, we provided genomic evidence to support the key predictions of this hypothesis.

## Introduction

Consisting of approximately 10,578 species, ferns are not only the sister lineage to seed plants but also the second-largest group of vascular plants (Smith et al. 2006; PPGI 2016). As their origin in the Paleozoic, they are major components in most terrestrial ecosystems (Mehltreter et al. 2010). Existing evidence supports the hypothesis of rather distinct trends in the genome evolution of ferns compared with other land plants especially to angiosperms as illustrated by recorded characteristics of fern genomes (Klekowski and Baker 1966; Wagner and Wagner 1980; Haufler 1987; Barker 2009; Barker and Wolf 2010; Clark et al. 2016; Sessa and Der 2016; Hidalgo et al. 2017a; Huang et al. 2020; Fujiwara et al. 2021). (1) Ferns accumulate large chromosome numbers including the largest known chromosome number. (2) They possess medium-to-large-sized genomes including the second-largest genome recorded. (3) These two characters, chromosome number ( $n$ ) and genome size ( $1C$ ), are positively correlated. (4) Among land plants, they show the highest frequency of polyploidy enforced speciation. (5) Some evidence supports ancient polyploidy events in the evolutionary history of ferns but arguably they were less common than proposed in the past.

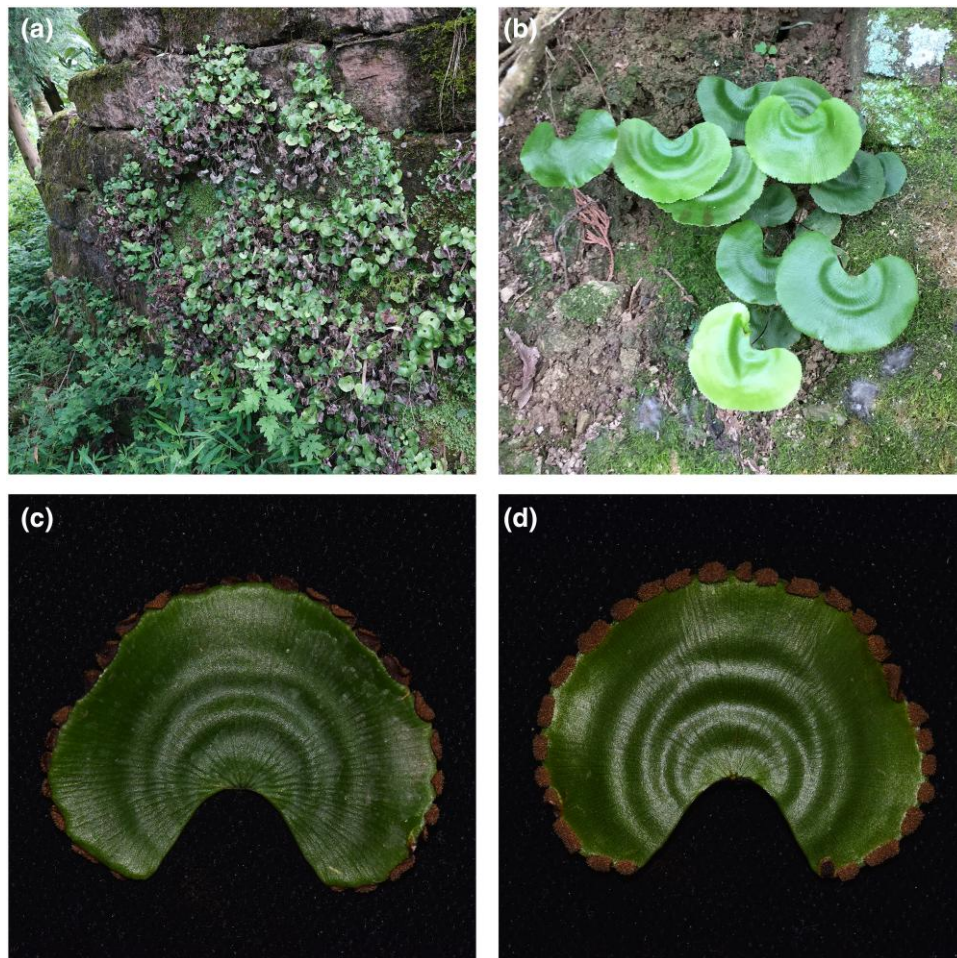
These characteristics together with various lines of evidence ranging from cytology, population genetics, genomics, and proteomics support the hypothesis that the post-whole-genome duplication (WGD) shock response of ferns is highly distinct from angiosperms. Diploidization is mainly achieved by genic process instead of cytological processes enabling rapid genome downsizing as reported from many derived angiosperms (Leitch and Bennett 2004; Wang et al. 2021). Instead, polyploid fern genomes arguably maintain most of the duplicated DNA and thus accumulate large chromosome numbers, a large number of pseudogenes, and a high proportion of noncoding DNA (Wolf et al. 2015; Clark et al. 2016; Marchant et al. 2019).

Our knowledge on the genomic diversity of seed plants and most other plant lineages have been rapidly improved since the publication of the first plant genome of *Arabidopsis thaliana* in 2000 as a consequence of hundreds of sequenced whole genomes that have been made available in recent years (Sun et al. 2022). In contrast, only two heterosporous ferns, *Azolla filiculoides* and *Salvinia cucullata*, and one homosporous fern, *Alsophila spinulosa*,

have been fully sequenced to date (Li et al. 2018; Huang et al. 2022). Two sequenced heterosporous ferns diverge from the majority of ferns by having relative small genomes (750 Mb for *Az. filiculoides* and 260 Mb for *S. cucullata*), whereas most homosporous fern lineages accumulate medium-to-large genomes (Clark et al. 2016; Fujiwara et al. 2021), including the second-largest genomes ever recorded, the genome of *Tmesipteris obliqua* (Hidalgo et al. 2017a, 2017b). In a whole-genome shotgun sequencing project, nuclear genome contigs were reported for a homosporous fern, *Ceratopteris richardii* (Marchant et al. 2019). However, this incompletely assembled genome with a coverage of only 38% genome assembled is insufficient to comprehensively address questions regarding genome structure, content, and evolution in homosporous ferns. Despite the above-mentioned efforts to obtain a whole genome of *C. richardii*, we still lack a completely assembled genome for polyploid ferns that contribute >80% of the extant diversity (PPGI 2016). Therefore, understanding the evolution of ferns in particular and vascular plants in general is incomplete due to the lack of a whole-genome sequence of one or several derived ferns.

Genome sequencing of ferns with typical large genome size and high chromosome number is essential for a comprehensive understanding of fern biology. Instead of hunting for the smallest fern genome, this study focused on a rather typical polyploid homosporous fern. Here we fully sequenced, assembled, and annotated the genome of a homosporous fern, *Adiantum nelumboides* X. *C. Zhang* that was previously recorded as *Adiantum reniforme* var. *sinense* Y. X. Lin. The genus *Adiantum* belongs to Pteridaceae, which comprises ca. 1211 species over 10% of extant ferns (PPGI 2016). Species in this family not only occupy a wide range of niches, including terrestrial, epiphytic, xeric-adapted rupestral, and even aquatic habitats, but also comprise many highly specialized species (Schuettpelez et al. 2007).

The species studied here was firstly discovered in Chongqing, China in 1978 (Lin 1980). It has a narrow distribution along the Yangtze River from Shizhu County to Wanzhou District of Chongqing (Tianquan et al. 1987) and usually occurs in exposed karst rocks and rocky crevices (Kang et al. 2008; fig. 1). Because of its sparse distribution, limited population number, and small population size, *Ad. nelumboides* was listed as a critically endangered species in Threatened Species List of China's Higher Plants (Qin et al.



**Fig. 1.**—*Adiantum nelumboides*. (a) habitat of a natural population in Shizhu County, Chongqing, China. (b) An individual in this population. (c) Adaxial frond surface with sporangia on the margin. (d) Abaxial frond surface with sporangia on the margin.

2017). In addition to the loss of habitats, over-collection by local people as a consequence of its medicinal value has further reduced the population size of *Ad. nelumboides* (Kang et al. 2008).

This ecologically highly specialized species has a chromosome number of  $2n = 120$  (Lin 1989) and a holoploid genome size of 7.39 Gb based on flow cytometry analysis (Fujiwara et al. 2021). Considering the high conserved basic chromosome number of *Adiantum* of  $x = 30$ , *Ad. nelumboides* is interpreted as a tetraploid, whereas the sister taxon *Ad. reniforme* occurring in Canary Islands and Madeira has been recorded as hexaploid ( $6x$ ) with  $2n = 180$  (Wang et al. 2015) or a decaploid ( $10x$ ) with  $2n = 300$  (Manton and Vida 1968).

1. The study was designed to confirm several predictions about the genomes of polyploid fern species that achieved to re-establish homeosis and fully functional meiosis after the WGD event. Firstly, the characteristics of the generated genome are expected to be consistent

with the post-WGD diploidization without genome downsizing including maintenance of high chromosome number enabling homeologous chromosome pairing during meiosis (Grusz et al. 2017). Thus, the genome is expected to possess a large number of genes and noncoding DNA especially transposable elements, as observed in other ferns (Wolf et al. 2015). Components relevant to gene regulation such as microRNA (miRNA) and small nuclear RNA (snRNA) are also expected to be present with a high copy number. Consistent with reports on the low frequency of large syntenic gene blocks in *Ceratopteris* (Nakazato et al. 2006; Marchant et al. 2019), the genome of *Ad. nelumboides* is expected to lack conserved large syntenic blocks of genes. This expectation is in conflict to the assumed conservation of large syntenic blocks as the consequence of chromosome duplication in newly formed polyploids (MacKintosh and Ferrier 2017). Finally, we predict the occurrence of a large number of protein-coding genes regulating calcium

transport and uptake, which enable the adaptation of *Ad. nelumboides* to grow on limestone rocks—arguably contributed a lot by the WGD

## Results

### Genome Assembly and Annotation

We generated 834 Gb PacBio reads and 389 Gb of Illumina paired-end 150-bp reads from an *Ad. nelumboides* individual for genome assembly (supplementary table 1, Supplementary Material online). The genome size was estimated to be 5.94 Gb, the heterozygosity was estimated to be 0.26% and the repeat content was estimated to be 59.0% using K-mer statistics of a subset (235 Gb) of the Illumina reads (supplementary fig. 1, Supplementary Material online). The genome assembled with the PacBio reads had a total length of 6.27 Gb sorted into 11,767 contigs with an N50 contig size of 1.37 Mb (table 1). The Benchmarking Universal Single-Copy Orthologs (BUSCO) assessment showed that 412 of the 425 BUSCO genes (96.9%) in the viridiplantae\_odb10 data set were recovered in the assembled genome, including 35.3% single-copy BUSCO genes and 61.6% duplicated BUSCO genes (supplementary table 2, Supplementary Material online). The high fidelity of the assembly was supported by the high genome coverage rate of 97.3% and high mapping rate of 94.8% of Illumina reads. The overall read-mapping rates for frond and rhizome transcriptomes were 92.4% and 86.0%, respectively. These observations suggest high quality of the genome assembly. The overall Guanine+Cytosine content of *Ad. nelumboides* is 40.2%, which is comparable with those of other ferns, ranging from 37.9% to 42.9%.

Noncoding repetitive DNA contributed about 81.7% of the assembled genome (5.12 Gb; table 1), which is a much higher proportion than those reported for two heterosporous ferns, containing 53.6% in *Azolla* and 44.5%

in *Salvinia* and even exceeding the 74.6% reported for the homosporous tree fern *Als. spinulosa*. Long terminal repeat (LTR) retrotransposons contributed 54.1% of the genome, with Ty3-gypsy (46.6% of the genome) and Ty1-copia (7.4% of the genome) being the most abundant types, and DNA transposons contributed 3.1% of the genome (supplementary table 3, Supplementary Material online). A total of 59,841 intact LTRs were obtained and most of the intact LTRs (99.9%) were inserted in the relative recent phylogenetic past (0–15 Ma), with the highest proportion of LTRs inserted ca. 0.46 Ma (supplementary fig. 2, Supplementary Material online).

The *Ad. nelumboides* genome contained 69,568 protein-coding genes based on de novo prediction, transcriptome sequences, and homology with other known plant proteins (table 1). The majority (94.6%) of the predicted genes were functionally annotated in at least one public database (supplementary table 4, Supplementary Material online). The BUSCO assessment for the predicted proteome showed that 87.1% complete BUSCO genes was recovered in the viridiplantae\_odb10 data set (supplementary table 2, Supplementary Material online), higher than 72.1% detected in *Als. spinulosa* from eukaryote\_odb database. The mapping rates of frond and rhizome RNA-seq to annotated coding sequences were 85.8% and 84.5%, respectively. The average exon and intron sizes were 328 and 3,380 bp, respectively (table 1). The average intron size of *Ad. nelumboides* is among the largest reported in plants, compared with those observed in the genomes of conifers (Nystedt et al. 2013; Niu et al. 2022). Nearly all (99.0%) long introns (>3000 bp in size) contained repeats, suggesting repeat insertion and expansion contributed to intron size expansion.

A total of 2,586 copies of tRNAs, 621 copies of ribosomal RNAs (rRNAs), 736 copies of snRNAs, and 9,453 copies of miRNAs (in 22 miRNA families) were identified (supplementary tables 5 and 6, Supplementary Material online). Among the 22 miRNA families, 4,139 copies of miRNA408 accounted for 43.8% of all miRNA copies, which is the largest copy number for a miRNA family. Other two miRNA families also showed abundant copies, with 2,463 for miRNA287 and 1,063 for miRNA672, in the genome of *Ad. nelumboides*, whereas miR156/157, miR170/171, miR396, miR165/166, miR159, miR160, miR168, and miR169 belong to the conserved miRNA families previously identified in plants (Berrueto et al. 2017; You et al. 2017).

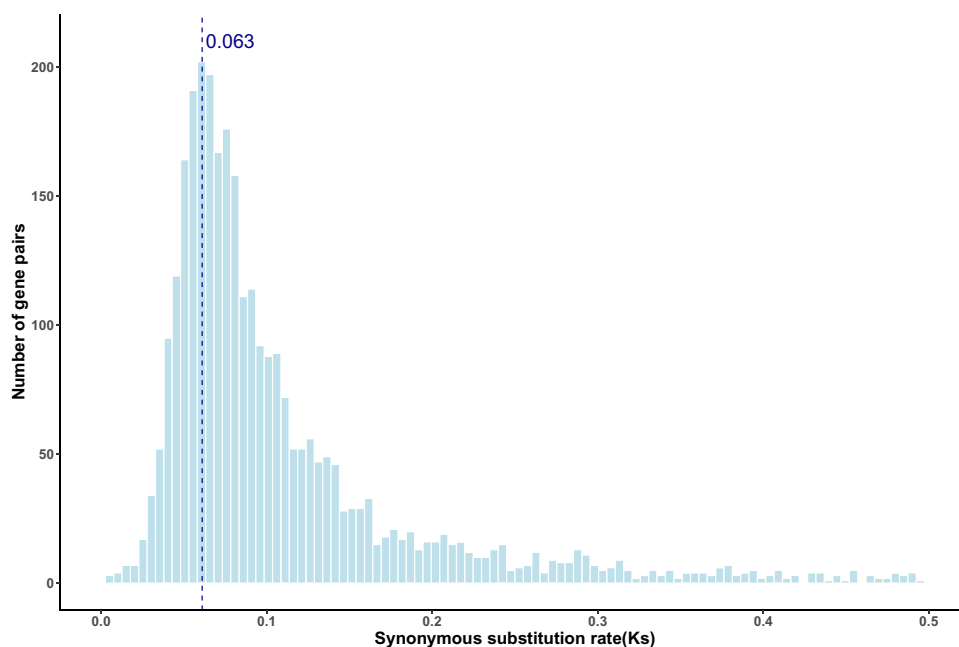
### Intragenomic Synteny and WGD

A total of 443 intragenomic syntenic blocks in the genome were identified, and they contain 14,996 genes and 3,485 gene pairs. On average, each syntenic block contains eight homologous gene pairs, with the longest block containing

**Table 1**

Statistics of the Genome Assembly for *Adiantum nelumboides*

Assembly Features	
Contig length (bp)	6,272,116,485
Contig number	11,767
N50 (bp)	1,373,929
L50	1,324
N90 (bp)	271,030
L90	5,099
Guanine+Cytosine content	40.2%
Repeat content (% of the genome assembly)	81.7
Number of predicted gene models	69,568
Average coding sequence length (bp)	1,309.9
Average exon number per gene	4.1
Average exon length (bp)	328
Average intron length (bp)	3,380



**FIG. 2.**—Frequency distribution of synonymous substitution rate (Ks) between gene pairs on syntenic blocks in the genome of *Adiantum nelumboides*. The x axis shows the Ks with a peak corresponding to 0.063, and the y axis represents the number of gene pairs. Only the distribution of Ks values of 0–0.5 was shown here.

23 gene pairs. With a small fraction of genes having more than one counterpart, these 3,485 gene pairs include 6,928 genes, indicating that approximately 10% of the annotated *Ad. nelumboides* genes exhibit synteny-based signals.

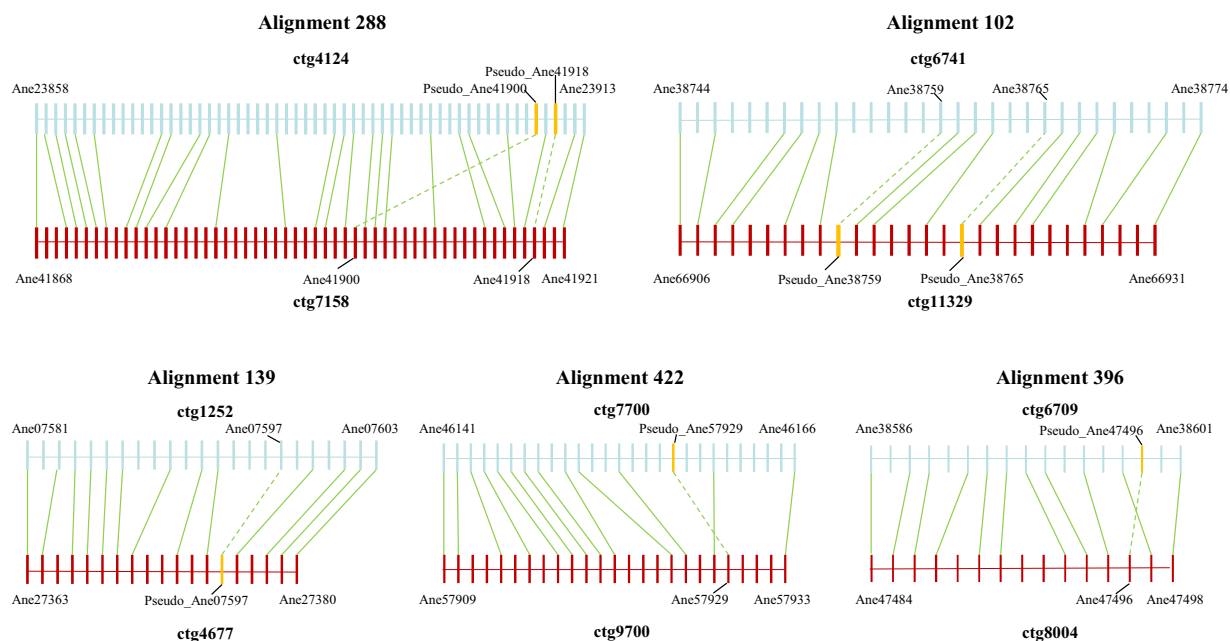
The peak value of Ks (the number of substitutions per synonymous site) distribution for all paralogous gene pairs within the 443 blocks is 0.063 (fig. 2 and supplementary fig. 3, Supplementary Material online), suggesting a very recent WGD event shaping the genome of this fern. Applying a synonymous substitution rate of  $4.79 \times 10^{-9}$  in polypodiaceous nuclear genomes, this WGD event was dated to approximately 6.6 Ma, whereas the absolute dating approach based on phylogenetic analysis showed that the WGD event occurred at 10.7 Ma (7.5–12.6 Ma).

With 69,568 genes in the 6.27 Gb genome, its average gene density is about 1 gene per 90 kb. On average, the segment length for five genes is [the default minimum number of gene pairs for syntenic blocks ( $m$ ) is 5 in MCScanX] about 450 kb. Because there are a lot of genomic contigs <450 kb, they may escape the detection of syntenic block identification. We then set  $m$  to 4, 3, and 2 separately, and a very limited increase for the number of syntenic blocks as well as the number of genes on these blocks was observed (supplementary table 7, Supplementary Material online).

In the whole genome, a large number of pseudogenes (89,510) were identified in the intergenic spacers, and they correspond to 13,373 genes. After excluding the internal stop codons from the pseudogenes, the Ks

distribution between pseudogenes and their corresponding genes showed a Ks peak at 0.068 (supplementary fig. 4, Supplementary Material online), which is very close to the Ks peak related to the WGD (0.063). Because pseudogenes usually exhibit an accelerated mutation rate due to the relaxation of selective constraints, we considered pseudogenes with a Ks value <1.5-fold of 0.063 from their corresponding genes as those produced after the WGD event. According to this criterion, 8,604 pseudogenes (9.6%), which correspond to 2,174 genes (16.3%), arguably originated after the detected WGD. Among the 8,604 pseudogenes, 1,548 were present in the syntenic blocks and corresponded to 661 genes. The 1,548 pseudogenes were most promising candidates originated after the detected WGD. This implies that there have been quite a few gene silencing events in the genome since the WGD.

Many instances of gene rearrangements in the syntenic blocks of *Ad. nelumboides* were found (see fig. 3 for examples), besides a great number of pseudogenes identified in the intergenic spacers. Given this data, pseudogenization of one homeolog was common in the identified syntenic blocks. For example, no corresponding annotated duplicates for the two genes, Ane38759 and Ane38765, were found in the collinear block alignment #102, but their pseudogenes were detected (fig. 3). Ks between the two genes and their pseudogenes in the syntenic blocks (only for the translatable regions) were 0.058 and 0.059, respectively, resembling the Ks peak value of 0.063 identified for the WGD event.



**FIG. 3.**—Examples for homeolog loss due to pseudogenization after the recent whole-genome duplication. Five syntenic blocks are shown. Each color bar represents a gene and the gene names are coded consecutively. The gene names at the beginning and end are labeled. Blue and red bars represent genes in the syntenic blocks and only gene pairs are connected with solid green lines. Yellow bars represent pseudogenes formed after the recent whole-genome duplication, which are connected with corresponding genes with dotted green lines.

### Gene Supporting the Adaptation to Limestone Habitats

Given the preference of this species as well as several other species of *Adiantum* to grow on limestone rocks, the study focused specifically to identify gene families considered to play a role in the adaptation to the specific needs of these habitats. These observations will enhance the usage of this species as a model to study the role of genes potentially supporting the adaptation of these plants to calcium-rich habitats.

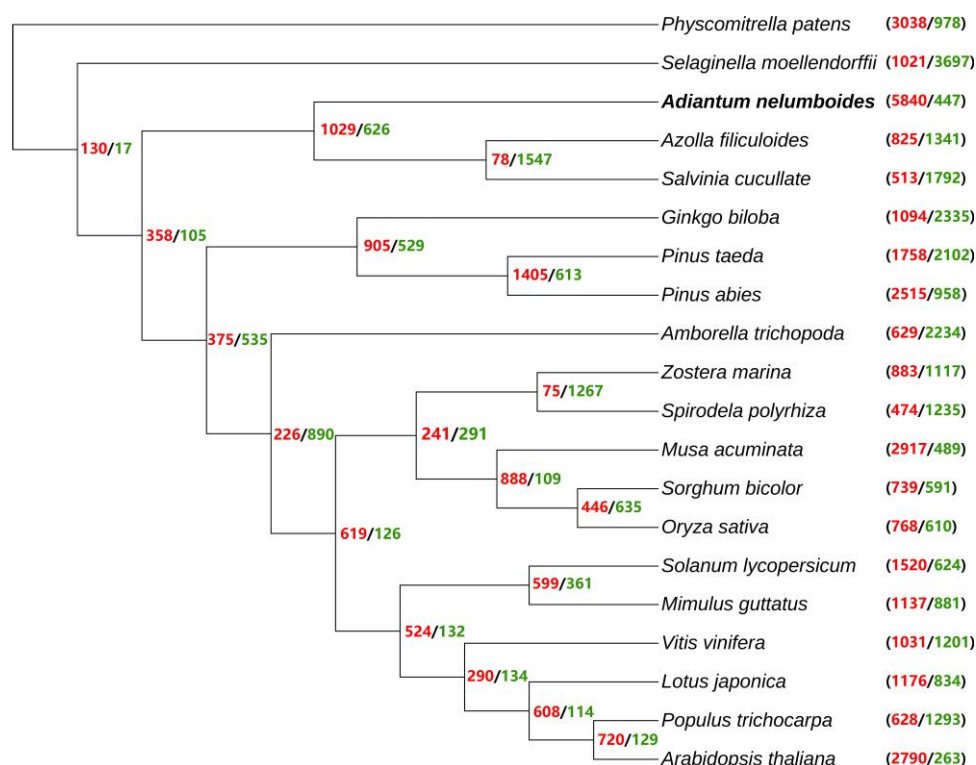
Genes from 20 species were clustered into 46,961 gene families with two or more members. We detected 5,840 gene families that were significantly expanded ( $P < 0.05$ ) and 447 gene families that were significantly contracted ( $P < 0.05$ ) in *Ad. nelumboides* (fig. 4). Although there are a large number of expanded gene families in *Ad. nelumboides*, functional enrichment analysis showed that neither the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways nor the GO terms are significantly enriched for these expanded gene families. There was one KEGG pathway, ascorbate and aldarate metabolism, and 169 GO terms are significantly underrepresented in *Ad. nelumboides* (Fisher's exact test,  $P < 0.05$ ; supplementary Excel file 1, Supplementary Material online). The underrepresented GO terms include L-ascorbic acid biosynthetic process, L-ascorbic acid metabolic process, response to reactive oxygen species, vitamin biosynthetic process, vitamin metabolic process, etc.

Five protein families related to calcium uptake and transport were detected in the genome of *Ad. nelumboides* (table 2 and supplementary Excel file 2, Supplementary Material online), including  $\text{Ca}^{2+}$ -ATPase proteins, calmodulins, IQD gene family of calmodulin binding, calmodulin-binding receptor-like kinase proteins, and calcineurin B-like calcium sensor. There were also three potassium transport-related protein families expanded, namely potassium proton antiporter family, potassium transporter family, and potassium channel family (table 2). In addition, we found three salt-related protein families in *Ad. nelumboides*, including  $\text{Na}^+/\text{H}^+$  exchanger subfamily, another  $\text{Na}^+/\text{H}^+$  exchanger subfamily, and vacuolar sodium/proton exchanger family (table 2). In 8 of these 11 gene families, 13–58% members likely originated in the WGD based on pairwise  $K_s$  analysis.

## Discussion

### Rapid Genome Evolution Following a Recent WGD

Given the conserved basic chromosome number of  $n = 30$  or less frequent  $n = 29$  (Rice et al. 2015), *Ad. nelumboides* is with  $2n = 120$  a tetraploid (supplementary fig. 5, Supplementary Material online). Consistently, the genome size of 7.39 Gb corresponds to  $4n$  given the reported genome size of diploids such as the genome size of 3.78 Gb of *Adiantum caudatum* (Kuo and Li 2019), a closest relative



**Fig. 4.**—Gene family expansion/contraction analysis for *Adiantum nelumboides* and 19 other species. Phylogeny of these species is reconstructed based on single- or low-copy orthologous genes. The number of expanded/contracted gene families is indicated at each node and after species.

with available genome size information. Furthermore, according to the results of the synteny analysis, *Ad. nelumboides* has undergone a WGD. The separation of *Ad. nelumboides* from its sister species *Ad. reniforme* has been dated back to about 4.94 Ma (Wang et al. 2015), so the establishment of the WGD event (6.6/10.7 Ma) occurred before their divergence. We did not find the signal of two earlier WGDs predating the origins of core leptosporangiate and Polypodiales, respectively. These events have been suggested before but these claims are controversial (Li et al. 2018; Huang et al. 2020). The very recent origin of this WGD event appears to be in contrast with relatively short syntenic blocks and a small fraction of genes on these blocks. Low gene density is the most plausible explanation (1 gene per 90 kb) in the *Ad. nelumboides* genome. Changes in the minimum number of gene pairs used for identifying syntenic blocks have little effect on the results of syntenic blocks, which implied that syntenic blocks in the *Ad. nelumboides* genome are indeed short and the proportion of genes on these syntenic blocks is indeed small. These observations are consistent with the hypothesis that the *Ad. nelumboides* genome has experienced rapid genome evolution following a rather recent WGD event. This argument is inconsistent with a previous suggestion that fern genomes are thought to evolve slowly and be less dynamic in general (Leitch and Leitch 2013; Clark et al. 2016).

Rapid genome reorganization has also been reported in the polyploid fern, *C. richardii*. This distantly related species, which also has a relative large genome (11.25 Gb) and has experienced a WGD event (Barker 2009). Its genome does not show evidence of large-scale synteny based on a high-resolution genetic linkage map, although 76% loci were duplicated (Nakazato et al. 2006). While these results may be less unexpected in the genome of *C. richardii* (Barker 2009; Barker and Wolf 2010), our findings for rapid genome evolution following such a recent WGD event in *Ad. nelumboides* are most striking. This may explain why genomic evidence for polyploidy is not as evident in homosporous ferns as expected in the past. Compared with most angiosperms showing evidence for rapid genome downsizing after WGDs, the *Ad. nelumboides* genome does not show evident genome downsizing. In contrast, the characteristics of the *Ad. nelumboides* genome suggest rapid evolution at segmental and genic levels.

The lack of evidence for extensive polyploidy in homosporous ferns with high chromosome numbers has been referred to as the “polyploidy paradox” (Soltis and Soltis 2000). For example, many homosporous ferns with high chromosome numbers exhibit diploid gene expression at isozyme loci (Haufler and Soltis 1986; Soltis 1986; Soltis and Soltis 1988a, 1988b; Gastony 1991). Soltis and Soltis (2000) considered two explanations for this paradox. In the

**Table 2**Expanded Gene Families Related to Calcium, Potassium, and Sodium Uptake and Transport in the *Adiantum nelumboides* Genome

Category	OrthoID	Gene Family Description	Number of Members in <i>Adiantum nelumboides</i>	Number of Members likely Produced by the WGD (%)	Average Number of Members in Other Species <sup>a</sup>	Reference for Gene Family Description
Calcium uptake and transport	OG0000331	Ca <sup>2+</sup> -ATPases proteins: Ca <sup>2+</sup> -efflux transporters responsible for maintaining homeostasis of cytosolic Ca <sup>2+</sup> concentration	23	13 (57%)	7	Huda et al. (2013)
	OG0000341	Calmodulins: predominant calcium receptors and small Ca <sup>2+</sup> binding protein that acts to transduce second messenger signals into a wide array of cellular responses	11	0	7	Chin and Means (2000)
	OG0000187	IQD gene family of calmodulin binding: linking calcium-signaling pathways to the regulation of gene expression	19	8 (42%)	10	Abel et al. (2005)
	OG0000729	Calmodulin-binding receptor-like kinase proteins	15	2 (13%)	5	-
	OG0000245	Calcineurin B-like calcium sensor: Ca sensors that interact with CIPK for stress responses, including mediating salt tolerance by regulating ion homeostasis in Arabidopsis	14	6 (43%)	9	Mao et al. (2016), Deng et al. (2013), Yu et al. (2014), Jin et al. (2016)
Potassium transport	OG0001821	Potassium proton antiporter family	6	2 (33%)	3	-
	OG0002459	Potassium transporter family	6	2 (33%)	2	-
	OG0002487	Potassium channel family	6	0	2	-
Salt tolerance	OG0000437	Na <sup>+</sup> /H <sup>+</sup> exchanger subfamily	12	6 (50%)	6	Qiu et al. (2003)
	OG0002550	Na <sup>+</sup> /H <sup>+</sup> exchanger subfamily	12	7 (58%)	2	Qiu et al. (2003)
	OG0000946	Vacuolar sodium/proton exchanger family	7	0	4	Glenn et al. (1999)

<sup>a</sup>Number of members in other species can be found in [Supplementary Excel file 2, Supplementary Material](#) online.

first explanation, ancient polyploidy in ferns coincides with extensive gene silencing to produce genetic diploids, whereas the second explanation assumes frequent chromosomal fission resulting in high chromosome numbers.

The second explanation does not hold for *Ad. nelumboides* because its genome size is roughly two-fold as large as its diploid relatives. Although *Ad. nelumboides* is a tetraploid species, a previous microsatellite study suggested that this species behaves a genetic diploid (Kang et al. 2008). Rapid chromosomal and genomic changes following WGD (Song et al. 1995; Leitch and Bennett 1997; Chester et al. 2012; Soltis et al. 2012; Chester et al. 2015), including karyotypic variation (intrachromosomal and intragenomic rearrangements) and gene silencing (and even loss), may explain genetic diploidization in *Ad. nelumboides*. In fact, pseudogenization and gene deletion by recombination have been considered to be major mechanisms for fractionation, as a particularly important component of diploidization (Li et al. 2021). A great number of pseudogenes, including pseudogenization of one homeolog, are detected in the *Ad. nelumboides* genome, which

is consistent with the genetic diploidization prediction. Moreover, with a long interval between neighboring genes in the *Ad. nelumboides* genome, the chance for recombination between genes is higher than other species with compact genomes. Correspondingly, more genomic rearrangements are expected between neighboring genes and very shorter syntenic blocks and even the loss of syntenic blocks in the *Ad. nelumboides* genome supports this. Meanwhile, a high level of recombination should be expected for *Ad. nelumboides* with such a high chromosome number of  $2n = 120$ , because increased chromosome numbers in nuclear genomes tend to cause increased genetic recombination (Qumsiyeh 1994). Overall, these data suggest rapid and substantial gene rearrangements and gene silencing following polyploidy in *Ad. nelumboides*. Thus, our genome fits well with the first explanation proposed by Soltis and Soltis (2000).

### Genetic Adaptation to High-Calcium Habitats

*Adiantum nelumboides* usually occurs in thin soils on limestone or purplish shale rocks and rocky crevices (Tianquan



et al. 1987; Kang et al. 2008). Limestone is a sedimentary rock composed primarily of calcium carbonate ( $\text{CaCO}_3$ ). Soils developed from purplish shale contain relatively high contents of calcium, potassium, and sodium other than silicon, aluminum, and iron (Du et al. 2013). Expansion and contraction of gene families have been suggested as major mechanisms underlying phenotypic differences as well as the key contributor to adaptive evolution (Purugganan et al. 1995; Lashbrook et al. 1998). Five protein families related to calcium uptake and transport in *Ad. nelumboides* play various roles related to maintaining homeostasis of cytosolic  $\text{Ca}^{2+}$  concentration and regulating gene expression and cellular responses via calcium-signaling pathway. Potassium ( $\text{K}^+$ ) transport is critical for enzyme activation, osmotic adjustment, turgor generation, cell expansion, regulation of membrane electric potential, and pH homeostasis (Hawkesford et al. 2012). In plants,  $\text{Na}^+/\text{H}^+$  exchangers in the plasma membrane are critical for growth in high levels of salt, removing toxic  $\text{Na}^+$  from the cytoplasm by transport out of the cell (Qiu et al. 2003). Vacuolar sodium/proton exchanger, on the other hand, plays essential role in transporting  $\text{Na}^+$  from the cytoplasm to vacuoles (Glenn et al. 1999). These protein families related to calcium uptake and transport, potassium, and sodium transport are expected to contribute to the adaptation of *Ad. nelumboides* to the harsh habitats.

Ascorbate has a role in protecting against reactive oxygen species produced during photosynthesis (Awad et al. 2015). The KEGG pathway of ascorbate is significantly underrepresented in *Ad. nelumboides*. The underrepresented GO terms, including L-ascorbic acid biosynthetic process and L-ascorbic acid metabolic process, are related to reactive oxygen species, vitamin biosynthetic process, vitamin metabolic process, etc. The underrepresented KEGG pathway and GO terms in *Ad. nelumboides* are consistent with its ecological characteristics, given that it inhabits shady habitats with relatively weak sunlight intensity and that less reactive oxygen species are expected to be produced.

Furthermore, *Ad. nelumboides* has a large copy number of miRNA408, which might be associated with its adaptation in the barren rocky habitats with various abiotic stresses. The miRNA408 family has been considered to be major contributors to abiotic stress tolerance through regulation of redox status during heat, cold, salt, and oxidative stress conditions in *Arabidopsis* (Ma et al. 2015).

## Conclusions

The genome of the tetraploid *Ad. nelumboides* provides important insights into the genomic structure of polyploid ferns that are geologically relatively young. The recovered structure of the genome is highly consistent with the first explanation of the fern genome paradox proposed by Soltis and Soltis (2000). The WGD has resulted in a highly

expanded genome by conserving the vast majority of the duplicated DNA and its packaging in chromosomes. However, several genic processes including gene silencing and expansion of regulation mechanisms enabled these genomes to establish homeostasis. In turn, the genome of this rare fern also supports the prediction that some of the expanded gene families contribute to the adaptation to stressful environments such as calcium-rich substrates. Besides providing the key to resolve the fern genome paradox, the complete genome of a derived homosporous fern also provides crucial information about the evolution of ferns and their relatives.

## Materials and Methods

### Sampling and Sequencing

An adult individual of *Ad. nelumboides* from a wild population in Shizhu County, Chongqing, China was collected and cultivated at the Shanghai Chenshan Botanical Garden, Shanghai, China. The voucher specimen (YYH15116) has been deposited in the herbarium of Shanghai Chenshan Botanical Garden (CSH). Several mature leaves were used for DNA isolation and library construction. Genomic DNA was isolated using the HiPure Plant DNA Mini Kit (Magen, Guangzhou, China). DNA integrity was tested on 1% agarose gel, whereas DNA purity was tested and quantified using Qubit Fluorometer (Invitrogen, Carlsbad, CA, USA). The genome libraries with an insert size of 30 kb were constructed and then sequenced on the PacBio Sequel II System based on the single-molecule real-time (SMRT) sequencing technology. To estimate the genome size and polish the genome assembly, a short genome fragment library with an insert size of 350 bp was also constructed and sequenced on the Illumina Novaseq platform with 150 bp paired-end reads obtained. All Illumina reads were filtered using Trimmomatic v0.39 (Bolger et al. 2014) with default parameters.

Fron and rhizome tissues used for RNA isolation and transcriptome sequencing were sampled from the specimen mentioned above. RNA-sequencing libraries were constructed using the Illumina TruSeq RNA Sample Prep Kit. The paired-end libraries with an insert size of 300 bp were sequenced on an Illumina Novaseq platform. For each tissue, about 8 Gb sequence data were generated (supplementary table 1, Supplementary Material online).

### Genome Size Estimation

We used Jellyfish v.2.2.10 (Marçais and Kingsford 2011) to obtain k-mers and calculate k-mers frequency for a subset (235 Gb) of the Illumina data. Genome size, heterozygosity, and repeat content were estimated based on k-mer ( $k = 25$ ) frequency distributions by GenomeScope (<http://qb.cshl.edu/genomescope>).

## Genome Assembly

The PacBio reads obtained were corrected, trimmed, and assembled into contigs using Canu v1.5 (Koren et al. 2017) with default parameters. The primary assembly was polished by referring to the Illumina reads with Pilon (Walker et al. 2014) with default parameters.

## Identification of Repetitive Sequences

Known repeat sequences were identified by RepeatMasker version 4.0.3 ([www.repeatmasker.org](http://www.repeatmasker.org)) with the Repbase library (Jurka et al. 2005). A de novo repeat library was constructed using RepeatModeler version 1.0.7 (RepeatModeler Open-1.0. 2008–2015 <<http://www.repeatmasker.org>>). The results from different methods were merged to generate a final nonredundant set of repetitive sequences. The repeat-masked genome sequences were used for subsequent gene prediction. LTRs were predicted by combining LTR\_finder\_parallel (Ou and Jiang 2019) and LTRharvest (Ellinghaus et al. 2008). The final identification was achieved using by LTR\_retriever (Ou and Jiang 2018). For intact LTR transposons, their insertion times were estimated based on the genetic divergence between the two LTRs, assuming their sequences were identical upon integration. Using the formula  $T = K/(2 * r)$ , where  $K$  and  $r$  denote the divergence and mutation rate. A synonymous substitution rate of  $4.79 \times 10^{-9}$  per site per year for polypodiaceous nuclear genomes was applied (Barker 2009).

## Gene Prediction and Annotation

We predicted protein-coding genes using a combination of homologous sequence search, ab initio gene prediction, and transcriptome-data comparison with the genome sequence implemented in an automatic genome annotation tool GETA v2.4.5 (<https://github.com/chenlianfu/geta>). Illumina RNA-seq reads from different tissues were mapped to the genome assembly using HISAT2 (Kim et al. 2019). Protein sequences of two heterosporous ferns (*Az. filiculoides* v1.1 and *S. cucullata* v1.2 available at <https://www.fernbase.org/>), and three flowering plants (*Populus trichocarpa* v4.1, *Ar. thaliana* TAIR10 and *Oryza sativa* v7.0 accessed via Phytozome version 13; Goodstein et al. 2012) were employed for homology-based prediction with GeneWise (<https://www.ebi.ac.uk/~birney/wise2/>). Ab initio prediction was performed in Augustus v3.3.3 (Stanke and Morgenstern 2005), trained with intron and exon information generated above. These prediction results were integrated and then were searched against the Pfam database for screening to get the final gene prediction result. The rRNAs, miRNAs, small nucleolar RNAs, and snRNAs were predicted using INFERNAL version 1.1rc4 (Nawrocki and Eddy 2013) against the Rfam database version 11.0 (Burge et al. 2012). The transfer RNAs (tRNAs) were

identified using tRNAscan-SE (Lowe and Eddy 1997) with the parameters: -X 20 -z 8.

Gene functions were assigned based on the best matches in the alignments using Blastp (Altschul et al. 1997) with the SwissProt database. Functional annotation of genes was also performed by using InterproScan (Jones et al. 2014) and eggNOG-mapper (<http://eggNOG-mapper.embl.de/>). The SwissProt, eggNOG, and InterPro annotation results were then integrated.

## Genome Assembly and Annotation Quality Assessment

The quality of the genome assembly and annotation was assessed by QUILT v5.1.0 (Gurevich et al. 2013), BUSCO (Simão et al. 2015), and the mapping rates of Illumina DNA-seq and RNA-seq reads. QUILT was employed to calculate N50, L50, N90, and L90, whereas Illumina reads were mapped to the genome assembly using BWA-mem v0.7.17 (Li and Durbin 2009) with default parameters. The mapping rates were summarized using the flagstat module in SAMtools (Li et al. 2009). RNA-seq reads was aligned to the genome and the predicted protein-coding sequences, and the mapping rate was recorded using the flagstat module in SAMtools and HISAT2 (Kim et al. 2019). The completeness of the genome assembly and predicted protein-coding sequences were further evaluated using the BUSCO viridiplantae\_odb10 database with default parameters.

## Identification of Collinear Blocks and WGD Analysis

All-versus-all alignment of the protein sequences of *Ad. nelumboides* was constructed using the Blastp algorithm (Altschul et al. 1997). To detect the signature of WGD, the program MCScanX (Wang et al. 2012) with default parameters was used to define duplicated blocks. For the default setting, at least five genes ( $m = 5$ ) in a block were required to call synteny. For each gene pair in the duplicated blocks, the Ks values were calculated using the YN00 method in PAML4 (Yang 2007) and the distribution of Ks values was plotted. To assess the influence of  $m$  on the number of syntenic blocks, we also set  $m$  to be 4, 3, and 2. Homeolog loss after the WGD caused by pseudogenization was checked for the selected collinear block pairs.

To estimate the absolute time of the WGD, gene pairs in the syntenic blocks with Ks ranging from 0.043 to 0.083 ( $0.063 \pm 0.02$ ) were selected. Protein sequences of three related species (*Pteris vittata*, *Aleuritopteris chrysophylla*, and *Ad. caudatum*), which were obtained from transcriptome sequencing, were downloaded from GigaDB (Shen et al. 2018). Orthologs were searched for *Ad. nelumboides* (using only one of the two homeologs produced by the WGD) and the three related species with OrthoFinder (Emms and Kelly 2019). Single-copy orthologs of the three

species and the homeologs of *Ad. nelumboides* produced by the WGD were aligned using MAFFT v7.0 (Kato and Standley 2013). RAXML-NG (Kozlov et al. 2019) was used to construct the maximum likelihood tree for each protein. Trees inconsistent with the species phylogeny were discarded. Using the divergence time of *Pt. vittata* and *Ale. chrysophylla* (Huang et al. 2020) as a calibration point, the divergence time between the homeologs in *Ad. nelumboides* was estimated using the mcmctree program implemented in PAML.

### Pseudogene Identification and Analysis

To enable the identification of putative pseudogenes, protein sequences of *Ad. nelumboides* were first aligned with the repeat-masked genome using TBLASTN (Altschul et al. 1997). Pseudogenes were then identified following the PseudogenePipeline (<https://github.com/ShiuLab/PseudogenePipeline>). Only sequences >300 bp in length were kept as pseudogenes.

Then, we extract-coding sequences of the identified pseudogenes and their corresponding genes and translated them into amino acid sequences. To calculate Ks between each pseudogene and its corresponding gene, the internal stop codons of pseudogenes were removed. For each pseudogene and its corresponding gene, their protein sequences were aligned, and then converted into nucleotide sequence alignments by ParaAT (Zhang et al. 2012). Ks was calculated for each pseudogene and its corresponding gene using KaKs\_Calculator 2.0 (Wang et al. 2010), and then the Ks distribution was plotted.

### Gene Family Analysis

Protein-coding genes from 19 other species (1 moss: *Physcomitrella patens*; one lycopodiales: *Selaginella moellendorffii*; 2 ferns: *Az. filiculoides* and *S. cucullata*; 3 gymnosperms: *Pinus taeda*, *Pinus abies*, and *Ginkgo biloba*; and 12 angiosperms: *Ar. thaliana*, *Po. trichocarpa*, *Lotus japonica*, *Vitis vinifera*, *Solanum lycopersicum*, *Mimulus guttatus*, *O. sativa*, *Sorghum bicolor*, *Musa acuminata*, *Zostera marina*, and *Spirodela polyrhiza*) were downloaded from Phytozome version 13 (Goodstein et al. 2012). For each gene with alternative splicing variants, the longest transcript was selected to represent that gene. Proteins from *Ad. nelumboides* and the 19 species were then combined to perform an all-against-all comparison using Blastp in OrthoFinder with default parameters. Orthologous groups were then established. Based on these orthologous groups, orthologous single- or two-copy nuclear genes were identified among *Ad. nelumboides* and 19 other species. For each orthologous gene, protein sequences were aligned using MAFFT v7.0, and all the alignments were then concatenated into a super matrix, and subject to substitution model test using jModelTest2 with the Akaike information

criterion (Darriba et al. 2012). Following the mode and using *Ph. patens* as an outgroup, the phylogenetic tree was constructed using RAXML v8.2.10 (Stamatakis 2014) with 1,000 bootstrap replicates.

Gene family expansions and contractions were inferred using the program CAFE v 4.1 (De Bie et al. 2006). Given equal birth ( $\lambda$ ) and death ( $\mu$ ) rate, global (one  $\lambda$ ) and multiple local rates ( $\lambda_1, \lambda_2, \dots$ ) were determined using a likelihood ratio test. Then the maximum likelihood value of turnover rates across the phylogenetic tree was estimated, and inferences on ancestral states of gene family sizes for each node and changes along each branch in the phylogenetic tree were drawn. Gene families with an accelerated rate of expansion and contraction were determined with a threshold conditional *P*-value ( $P < 0.05$ ). For the expanded and contracted gene families, KEGG pathway and GO category enrichment analysis was conducted using KOBAS 3.0 (Xie et al. 2011) and statistical significance was tested by Fisher's exact test in combination with the False Discovery Rate correction. For those expanded gene families related to calcium, potassium and sodium uptake and transport in the *Ad. nelumboides* genome, we calculated pairwise Ks between members within each gene family and counted the number of genes likely produced by the WGD with Ks values ranging from <25% to >25% 0.063.

### Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online (<http://www.gbe.oxfordjournals.org/>).

### Acknowledgment

This work was financially supported by the Biodiversity Survey, Observation and Assessment of the Ministry of Ecology and Environment, China (2019HJ2096001006).

### Data Availability

The high-quality genome assembly and annotation of *Ad. nelumboides* have been deposited in NCBI under the accession number: JAKNSL000000000 (BioProject accession: PRJNA802344). The annotation details of the genome are available on the website <https://doi.org/10.6084/m9.figshare.19736395.v1>.

### Literature Cited

- PPGI. 2016. A community-derived classification for extant lycophytes and ferns. *J Syst Evol.* 54:563–603.
- Abel S, Savchenko T, Levy M. 2005. Genome-wide comparative analysis of the IQD gene families in *Arabidopsis thaliana* and *Oryza sativa*. *BMC Evol Biol.* 5(1):7.
- Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.

- Awad J, et al. 2015. 2-Cysteine peroxidoredoxins and thylakoid ascorbate peroxidase create a water-water cycle that is essential to protect the photosynthetic apparatus under high light stress conditions. *Plant Physiol.* 167:1592–1603.
- Barker MS. 2009. Evolutionary genomic analyses of ferns reveal that high chromosome numbers are a product of high retention and fewer rounds of polyploidy relative to angiosperms. *Am Fern J.* 99:136–141.
- Barker MS, Wolf PG. 2010. Unfurling fern biology in the genomics age. *BioScience* 60:177–185.
- Berrueto F, et al. 2017. Sequencing of small RNAs of the fern *Pleopeltis minima* (Polypodiaceae) offers insight into the evolution of the microRNA repertoire in land plants. *PLoS One* 12:e0177573.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120.
- Burge SW, et al. 2012. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.* 41:D226–32.
- Chester M, et al. 2012. Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). *Proc Natl Acad Sci U S A.* 109:1176–1181.
- Chester M, Riley RK, Soltis PS, Soltis DE. 2015. Patterns of chromosomal variation in natural populations of the neoallotetraploid *Tragopogon mirus* (Asteraceae). *Heredity* 114:309–317.
- Chin D, Means AR. 2000. Calmodulin: a prototypical calcium sensor. *Trends Cell Biol.* 10(8):322–328.
- Clark J, et al. 2016. Genome evolution of ferns: evidence for relative stasis of genome size across the fern phylogeny. *New Phytol.* 210:1072–1082.
- Darriba D, Taboada GL, Doallo R, Posada D. 2012. Jmodeltest 2: more models, new heuristics and parallel computing. *Nat Methods* 9:772–772.
- De Bie T, Cristianini N, Demuth JP, Hahn MW. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22:1269–1271.
- Deng X, et al. 2013. TaCIPK29, a CBL-interacting protein kinase gene from wheat, confers salt stress tolerance in transgenic tobacco. *PLoS ONE* 8(7):e69881.
- Du J, Luo Y, Zhang W, Xu C, Wei C. 2013. Major element geochemistry of purple soils/rocks in the red Sichuan Basin, China: implications of their diagenesis and pedogenesis. *Environ Earth Sci.* 69:1831–1844.
- Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* 9:18.
- Emms DM, Kelly S. 2019. Orthofinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20:238.
- Fujiwara T, et al. 2021. Evolution of genome space occupation in ferns: linking genome diversity and species richness. *Ann Bot.* mcab094:1–12.
- Gastony GJ. 1991. Gene silencing in a polyploid homosporous fern: paleopolyploidy revisited. *Proc Natl Acad Sci U S A.* 88:1602–1605.
- Glenn EP, Brown JJ, Blumwald E. 1999. Salt tolerance and crop potential of halophytes. *Crit Rev Plant Sci.* 18:227–255.
- Goodstein DM, et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40:D1178–D1186.
- Grusz AL, Sigel EM, Witherup C. 2017. Homoeologous chromosome pairing across the eukaryote phylogeny. *Mol Phylogenet Evol.* 117:83–94.
- Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29:1072–1075.
- Haufler CH. 1987. Electrophoresis is modifying our concepts of evolution in homosporous pteridophytes. *Am J Bot.* 74:953–966.
- Haufler CH, Soltis DE. 1986. Genetic evidence suggests that homosporous ferns with high chromosome numbers are diploid. *Proc Natl Acad Sci U S A.* 83:4389–4393.
- Hawkesford M, et al. 2012. Chapter 6—functions of macronutrients. In: Marschner P, editor. *Marschner's mineral nutrition of higher plants* (3rd ed). San Diego: Academic Press. p. 135–189.
- Hidalgo O, et al. 2017a. Is there an upper limit to genome size? *Trends Plant Sci.* 22:567–573.
- Hidalgo O, Pellicer J, Christenhusz MJ, Schneider H, Leitch IJ. 2017b. Genomic gigantism in the whisk-fern family (Psilotaceae): *Tmesipteris obliqua* challenges record holder *Paris japonica*. *Bot J Linn Soc.* 183:509–514.
- Huang X, et al. 2022. The flying spider-monkey tree fern genome provides insights into fern evolution and arborescence. *Nat Plants.* 8(5):500–512.
- Huang CH, Qi X, Chen D, Qi J, Ma H. 2020. Recurrent genome duplication events likely contributed to both the ancient and recent rise of ferns. *J Integr Plant Biol.* 62:433–455.
- Huda K, et al. 2013. Genome-wide analysis of plant-type II Ca<sup>2+</sup>-ATPases gene family from rice and Arabidopsis: Potential role in abiotic stresses. *Plant Physiol Biochem.* 65:32–47.
- Jin X, et al. 2016. Wheat CBL-interacting protein kinase 25 negatively regulates salt tolerance in transgenic wheat. *Sci Rep.* 6(1):2457.
- Jones P, et al. 2014. Interproscan 5: genome-scale protein function classification. *Bioinformatics* 30:1236–1240.
- Jurka J, et al. 2005. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res.* 110:462–467.
- Kang M, Huang H, Jiang M, Lowe AJ. 2008. Understanding population structure and historical demography in a conservation context: population genetics of an endangered fern. *Divers Distrib.* 14:799–807.
- Katoh K, Standley DM. 2013. MAFFT Multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30:772–780.
- Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol.* 37:907–915.
- Klekowski EJ, Baker HG. 1966. Evolutionary significance of polyploidy in the Pteridophyta. *Science* 153:305–307.
- Koren S, et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27:722–736.
- Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAXML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35:4453–4455.
- Kuo L-Y, Li F-W. 2019. A roadmap for fern genome sequencing. *Am Fern J.* 109:212–223.
- Lashbrook CC, Tieman DM, Klee HJ. 1998. Differential regulation of the tomato ETR gene family throughout plant development. *Plant J.* 15:243–252.
- Leitch IJ, Bennett MD. 1997. Polyploidy in angiosperms. *Trends Plant Sci.* 2:470–476.
- Leitch I, Bennett M. 2004. Genome downsizing in polyploid plants. *Biol J Linn Soc.* 82:651–663.
- Leitch IJ, Leitch AR. 2013. Genome size diversity and evolution in land plants. In: Greilhuber J, Dolezel J, and Wendel JF, editors. *Plant genome diversity volume 2: physical structure, behaviour and evolution of plant genomes*. Vienna: Springer Vienna. p. 307–322.
- Li H, et al. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079.
- Li F-W, et al. 2018. Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nat Plants* 4:460–472.
- Li Z, et al. 2021. Patterns and processes of diploidization in land plants. *Annu Rev Plant Biol.* 72(1):387–410.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25:1754–1760.
- Lin Y-X. 1980. New taxa of *Adiantum* L. in China. *J Syst Evol.* 18:101–105.

- Lin Y-X. 1989. The sexual propagation and chromosome number of *Adiantum reniforme* L. var. *sinense* YX Lin. *Cathaya* 1:143–148.
- Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25:955–964.
- Ma C, Burd S, Lers A. 2015. Mir408 is involved in abiotic stress responses in *Arabidopsis*. *Plant J.* 84:169–187.
- MacKintosh C, Ferrier DEK. 2017. Recent advances in understanding the roles of whole genome duplications in evolution. *F1000Research* 6:1623–1623.
- Manton I, Vida G. 1968. Cytology of the fern flora of Tristan da Cunha. *Proc R Soc Lond B Biol Sci.* 170:361–379.
- Mao J, et al. 2016. Mechanisms and physiological roles of the CBL-CIPK networking system in *Arabidopsis thaliana*. *Genes* 7(9):62.
- Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27:764–770.
- Marchant DB, et al. 2019. The C-Fern (*Ceratopteris richardii*) genome: insights into plant genome evolution with the first partial homosporous fern genome assembly. *Sci Rep.* 9:1–14.
- Mehlreter K, Walker LR, Sharpe JM. 2010. *Fern ecology*. New York: Cambridge University Press.
- Nakazato T, Jung M-K, Housworth EA, Rieseberg LH, Gastony GJ. 2006. Genetic map-based analysis of genome structure in the homosporous fern *Ceratopteris richardii*. *Genetics* 173:1585–1597.
- Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29:2933–2935.
- Niu S, et al. 2022. The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell* 185:204–217.e214.
- Nystedt B, et al. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* 497:579–584.
- Ou S, Jiang N. 2018. LTR\_RetrieveR: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* 176:1410–1422.
- Ou S, Jiang N. 2019. LTR\_FINDER\_parallel: parallelization of LTR\_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mobile DNA* 10:48.
- Purugganan MD, Rounsley SD, Schmidt RJ, Yanofsky MF. 1995. Molecular evolution of flower development: diversification of the plant MADS-box regulatory gene family. *Genetics* 140:345–356.
- Qin H, et al. 2017. Threatened species list of China's higher plants. *Biodivers Sci.* 25:696–744.
- Qiu Q-S, Barkla BJ, Vera-Estrella R, Zhu J-K, Schumaker KS. 2003. Na<sup>+</sup>/H<sup>+</sup> exchange activity in the plasma membrane of *Arabidopsis*. *Plant Physiol.* 132:1041–1052.
- Qumsiyeh M. 1994. Evolution of number and morphology of mammalian chromosomes. *J Hered.* 85:455–465.
- Rice A, et al. 2015. The chromosome counts database (CCDB)—a community resource of plant chromosome numbers. *New Phytol.* 206:19–26.
- Schuettpelz E, Schneider H, Huet L, Windham MD, Pryer KM. 2007. A molecular phylogeny of the fern family Pteridaceae: assessing overall relationships and the affinities of previously unsampled genera. *Mol Phylogenet Evol.* 44:1172–1185.
- Sessa E, Der J. 2016. Evolutionary genomics of ferns and lycophytes. *Adv Bot Res.* 78:215–254.
- Shen H, et al. 2018. Large-scale phylogenomic analysis resolves a backbone phylogeny in ferns. *GigaScience* 7(2):1–11.
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–3212.
- Smith AR, et al. 2006. A classification for extant ferns. *Taxon* 55:705–731.
- Soltis DE. 1986. Genetic evidence for diploidy in *Equisetum*. *Am J Bot.* 73:908–913.
- Soltis DE, et al. 2012. The early stages of polyploidy: rapid and repeated evolution in *Tragopogon*. In: Soltis PS, Soltis DE, editors. *Polyploidy and genome evolution*. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 271–292.
- Soltis DE, Soltis PS. 1988a. Are lycopods with high chromosome numbers ancient polyploids? *Am J Bot.* 75:238–247.
- Soltis PS, Soltis DE. 1988b. Electrophoretic evidence for genetic diploidy in *Pilotum nudum*. *Am J Bot.* 75:1667–1671.
- Soltis PS, Soltis DE. 2000. The role of genetic and genomic attributes in the success of polyploids. *Proc Natl Acad Sci U S A.* 97:7051–7057.
- Song K, Lu P, Tang K, Osborn TC. 1995. Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. *Proc Natl Acad Sci U S A.* 92:7719–7723.
- Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.
- Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 33:W465–W467.
- Sun Y, Shang L, Zhu Q-H, Fan L, Guo L. 2022. Twenty years of plant genome sequencing: achievements and challenges. *Trends Plant Sci.* 27:391–401.
- Tianquan X, Zhong Z, Yixing J. 1987. On the distribution characteristic of the variety *Adiantum reniforme* var. *sinense*. *Plant Sci J.* 5:247–252.
- Wagner WH, Wagner FS. 1980. Polyploidy in pteridophytes. In: Lewis WH, editor. *Polyploidy: biological relevance*. Boston, MA: Springer US. p. 199–214.
- Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963.
- Wang Y, et al. 2012. MCScanx: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40:e49.
- Wang AH, et al. 2015. Identification of the relationship between Chinese *Adiantum reniforme* var. *sinense* and Canary *Adiantum reniforme*. *BMC Plant Biol.* 15:36.
- Wang X, Morton JA, Pellicer J, Leitch IJ, Leitch AR. 2021. Genome downsizing after polyploidy: mechanisms, rates and selection pressures. *Plant J.* 107:1003–1015.
- Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. 2010. Kaks\_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics.* 8:77–80.
- Wolf PG, et al. 2015. An exploration into fern genome space. *Genome Biol Evol.* 7:2533–2544.
- Xie C, et al. 2011. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* 39:W316–W322.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.
- You C, et al. 2017. Conservation and divergence of small RNA pathways and microRNAs in land plants. *Genome Biol.* 18:1–19.
- Yu Q, An L, Li W. 2014. The CBL–CIPK network mediates different signaling pathways in plants. *Plant Cell Rep.* 33(2):203–214.
- Zhang Z, et al. 2012. ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochem Biophys Res Commun.* 419:779–781.

Associate editor: Itay Mayrose