The Plant Cell, Vol. 27: 1567–1578, June 2015, www.plantcell.org

Horsetails Are Ancient Polyploids: Evidence from *Equisetum giganteum*[OPEN]

Kevin Vanneste,[a,b] Lieven Sterck,[a,b] Alexander Andrew Myburg,[c,d] Yves Van de Peer,[a,b,d,1] and Eshchar Mizrachi[c,d,1]

[a] Department of Plant Systems Biology, VIB, Ghent B-9052, Belgium
[b] Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent B-9052, Belgium
[c] Department of Genetics, Forestry, and Agricultural Biotechnology Institute, University of Pretoria, Pretoria 0028, South Africa
[d] Department of Genetics, Genomics Research Institute, University of Pretoria, Pretoria 0028, South Africa

ORCID ID: 0000-0001-7116-4000 (L.S.)

Horsetails represent an enigmatic clade within the land plants. Despite consisting only of one genus (*Equisetum*) that contains 15 species, they are thought to represent the oldest extant genus within the vascular plants dating back possibly as far as the Triassic. Horsetails have retained several ancient features and are also characterized by a particularly high chromosome count (n = 108). Whole-genome duplications (WGDs) have been uncovered in many angiosperm clades and have been associated with the success of angiosperms, both in terms of species richness and biomass dominance, but remain understudied in nonangiosperm clades. Here, we report unambiguous evidence of an ancient WGD in the fern linage, based on sequencing and de novo assembly of an expressed gene catalog (transcriptome) from the giant horsetail (*Equisetum giganteum*). We demonstrate that horsetails underwent an independent paleopolyploidy during the Late Cretaceous prior to the diversification of the genus but did not experience any recent polyploidizations that could account for their high chromosome number. We also discuss the specific retention of genes following the WGD and how this may be linked to their long-term survival.

INTRODUCTION

Although ferns (monilophytes) represent the most diverse plant lineage on Earth second only to angiosperms, almost all of these species are contained within the monophyletic (leptosporangiate) polypod fern lineage that contains 267 genera and ~9000 species. The remaining 20% of species lie outside this derived group and represent earlier diverging lineages (Schuettpelz and Pryer, 2007). Ferns were the dominant plant clade on Earth both in terms of species richness and biomass during the Paleozoic but had to compete with the more derived angiosperms during the Mesozoic. Originally it was thought that ferns experienced a drastic decline (Crane, 1987), although latter work readjusted this view by demonstrating that polypod ferns, in particular, diversified during the Cretaceous in the wake of novel ecological opportunities that opened up by the rise to dominance of angiosperms (Schneider et al., 2004).

Over half a century ago, the high chromosome counts of homosporous fern genomes caught the attention of researchers (Manton, 1950). Ferns have especially high chromosome numbers, with *Ophioglossum reticulatum* possessing the highest known chromosome count among extant eukaryotes (n > 600; Khandelwal, 1990). However, a distinction should be made between heterosporous ferns that have chromosome counts similar to angiosperms (n = 15.99 on average) and homosporous ferns where the chromosome count is more than 3 times higher (n = 57.05 on average; Klekowski and Baker, 1966). This early work investigated a limited number of homosporous ferns and concluded that selfing was a common mechanism of reproduction (Klekowski and Baker, 1966). Combined with later observations of unusually high numbers of chromosomes in homosporous ferns (and the resultant assumption that this was due to polyploidy), the hypothesis emerged that polyploidy would present a plausible mechanism allowing the generation of novel genetic material to avoid homozygous inbreeding depression by pairing of homoeologous rather than homologous chromosomes during meiosis (Klekowski, 1973). This view was later challenged and revised through findings that species having the lowest chromosome number within each genus exhibit diploid genetic expression (Haufler and Soltis, 1986) and that these genetically diploid species experience selfing rates near zero (Soltis and Soltis, 1990) through breeding systems such as temporal separation between sperm and egg release (Soltis and Soltis, 1992). It was therefore proposed that a few ancient cycles of polyploidy followed by gene silencing and maintenance of chromosomes could explain the high chromosome counts of homosporous ferns (Haufler, 1987) or alternatively that the ancestor of ferns and seed plants had a high chromosome number that was retained only in ferns (Soltis and Soltis, 1987). Hence, the paradox between the high chromosome number of ferns and their diploid state remains an intriguing open question (Haufler, 2002, 2014).

Among the monilophytes, horsetails in particular have received much attention. They consist of a single extant genus (*Equisetum*), which is the only remaining representative from the more ancient Sphenophyta that once were a very abundant and diverse clade (Husby, 2013). The Equisetopsida are estimated to have diverged from other monilophytes somewhere around ~354 million

[1] Address correspondence to yves.vandepeer@psb.ugent.be or eshchar.mizrachi@fabi.up.ac.za.
The authors responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) are: Yves Van de Peer (yves.vandepeer@psb.ugent.be) and Eshchar Mizrachi (eshchar.mizrachi@fabi.up.ac.za).
[OPEN]Articles can be viewed online without a subscription.
www.plantcell.org/cgi/doi/10.1105/tpc.15.00157

years ago (mya) (Schneider et al., 2004). This extreme isolation of *Equisetum*, combined with fossil findings that show that horsetails have retained multiple ancient features dating back to the Jurassic (Channing et al., 2011) and possibly even the Triassic (Hauke, 1963), indicate that *Equisetum* could possibly be the oldest extant genus within the vascular plants (Husby, 2013).

Equisetum contains only 15 species, all of which (with the exception of a single reported triploidy event) have a chromosome number of $n = 108$ and are diploid (Soltis, 1986; Leitch and Leitch, 2013). Nevertheless, the genus has a subcosmopolitan distribution with most species found between 40° and 60° northern latitude. They typically thrive under a very wide range of conditions due to diverse adaptations that enable efficient nutrient uptake and nitrogen fixation mechanisms and make them tolerant to disturbance, soil anoxia, high metals, and salinity (Husby, 2013). Rhizomatous clonal growth is a universal feature of the genus and is extremely important for its ecology in moist conditions such as woods, ditches, and wetlands (Hauke, 1963). Despite its limited number of species, its plasticity and survival through multiple geological time scales and extinction events render *Equisetum* a very successful lineage (Rothwell, 1996). The relationships between the different horsetail species is now well resolved with *Equisetum bogotense* basal to both the subgenera *Hippochaete* and *Equisetum* that each contain seven species (Guillon, 2004, 2007), while the Equisetopsida are most likely sister to both the whisk ferns (Psilotales) and ophioglossoid ferns (Ophioglossales; Grewe et al., 2013). Both molecular dating studies (Des Marais et al., 2003; Pryer et al., 2004) and fossil evidence (Stewart and Rothwell, 1993) indicate that extant horsetails diverged in the Early Cenozoic, not long after the Cretaceous-Paleogene (K-Pg) boundary ~66 mya.

Recently, we demonstrated that multiple whole-genome duplication (WGD) events in the angiosperms occurred in association with the K-Pg boundary, especially within the more derived eudicots and monocots (Vanneste et al., 2014a). Several WGDs have also been found in earlier diverging angiosperm lineages such as the magnolids, although dating of these events remains uncertain due to their more isolated position and lack of sequence data (Cui et al., 2006; Soltis et al., 2009). Moreover, the ancestor of all angiosperms most likely underwent a WGD entailing that all extant angiosperms are in fact paleopolyploids (Jiao et al., 2011). Although the prevalence of WGDs in the angiosperms has become firmly established, their attributed importance remains controversial, ranging between a road toward evolutionary success and an evolutionary dead end (Van de Peer et al., 2009, 2010; Abbasi, 2010; Mayrose et al., 2011, 2015; Arrigo and Barker, 2012; Soltis et al., 2014). Recent advances suggest that the success of WGDs should be viewed in the context of environmental and ecological conditions at the time of their establishment and their subsequent evolutionary routes taken and that this might perhaps help to reconcile both fates (Schranz et al., 2012; Vanneste et al., 2014b). Contrary to angiosperms, much less is known about the prevalence of WGDs in nonangiosperm clades, mostly due to a lack of genomic data that allows exploring such events (Soltis and Soltis, 2013). The ancestor of gymnosperms was most likely also a paleopolyploid species, although no recent polyploids have been described yet (Jiao et al., 2011; Nystedt et al., 2013). The bryophyte *Physcomitrella patens* appears to have undergone a WGD (Rensing et al., 2008), coinciding again with the K-Pg boundary.

In this study, we performed transcriptome sequencing of the giant horsetail (*Equisetum giganteum*) to gain more insight into the evolution of this enigmatic genus. We found that horsetails underwent a paleopolyploidy during their evolutionary history that occurred somewhere in the Late Cretaceous. We interpret these results in the light of current knowledge concerning the importance of WGDs in plant species and argue that the WGD might have contributed to the evolutionary success of *Equisetum*.

RESULTS

Assessment of the Transcriptome Assembly

The Oases assembly based on the Velvet k-mer 41 output resulted in 34,282 transcripts with an average sequence length of 1251.19 bases and an average GC and GC3 content of 47.12 and 43.81%, respectively. The completeness of our transcriptome assembly was assessed by estimating the coverage of the gene space based on the core eukaryotic gene mapping approach (Parra et al., 2007) that assesses how many genes out of a set of 248 genes shared by all eukaryotic species are present in our assembly. In total, 239 partial or complete genes (96.37%), of which 224 were complete genes (90.32%) (based on a minimum protein alignment length of 70% against the Hidden Markov Model), could be identified, which is comparable to similar transcriptome studies in other plants (Nakasugi et al., 2013) and suggests a relatively high level of gene completeness. Additionally, we found on average 2.37 and 2.66 orthologs for the complete and partial set, respectively, while in total 140 and 165 of detected core proteins had more than one ortholog in the complete and partial set, respectively. These levels are comparable with *Arabidopsis thaliana*, which has on average two orthologs per core protein (Parra et al., 2007).

The *E. giganteum* Transcriptome Contains the Remnants of a Paleopolyploidy Event

The K_S-based age distribution was constructed and is presented in Figure 1 for all duplicated gene pairs in the *E. giganteum* transcriptome. Synonymous substitutions do not change the amino acid and are therefore considered to behave putatively neutral (Kimura, 1977), so that they accumulate at an approximately constant rate and hence serve as a proxy for the time since duplication of paralogous genes (Li, 1997). However, genes are most frequently lost after duplication, resulting in an L-shaped pattern corresponding to many recently duplicated genes that exponentially decay over time with only few genes surviving across large evolutionary time spans (Lynch and Conery, 2000, 2003). Since a large-scale duplication event such as a WGD results in the birth of a large set of genes that are created contemporarily, they are recognizable as peaks superimposed upon this L-shaped small-scale duplication (SSD) distribution (Blanc and Wolfe, 2004). To provide evidence that the large peak present in the *E. giganteum* age distribution corresponds to a bona fide WGD feature rather than stochastic variation on the background SSD distribution, we employed mixture modeling to separate the contribution of

log-transformed SSD background exponential and WGD Gaussian functions to the overall age distribution (Schlueter et al., 2004; Cui et al., 2006). We limited this approach to a maximum K_S of 2, as after this boundary evolutionary models used in K_S estimation are prone to K_S saturation and stochasticity effects that misguide mixture modeling (Vanneste et al., 2013). Results are presented in Figure 1 and Table 1. A Gaussian function is fitted to a putative WGD feature with its peak located at a K_S of 0.81, but several other smaller Gaussian functions are also fitted to the age distribution.

Stop criteria such as BIC (Schwarz, 1978) used in identifying the optimal number of components in the mixture are prone to overfitting (Naik et al., 2007), so that small stochastic deviations in the background SSD distribution are also often fitted (Vekemans et al., 2012). We therefore used the SiZer package (Chaudhuri and Marron, 1999) that searches for significant changes in the first derivative of a distribution in conjunction with our mixture modeling approach, as this technique allows a more robust evaluation of whether a fitted Gaussian component truly corresponds to a WGD signature (Barker et al., 2008). SiZer identified an unambiguous change in the age distribution at the location of the Gaussian function that was fitted to the large peak located at a K_S of 0.81 and therefore confirmed this to be a bona fide WGD signature.

Regulatory and Developmental Genes Are Overrepresented in the Set of Homoeologs

We performed functional annotation of the *E. giganteum* transcriptome by employing Blast2GO to assign Gene Ontology (GO) terms to genes (Conesa et al., 2005) and then employed BiNGO (Maere et al., 2005b) to investigate whether particular GO terms were significantly overrepresented in homoeologs compared with the remainder of the gene space. Results for GO-slim terms (that provide a higher level overview of individual GO terms and therefore allow a broad overview of all of the top GO categories) are presented on Figure 2, while full results are presented in Supplemental Data Set 1. In total, 15 GO-slim labels showed enrichment. These included many genes in categories such as regulation of biological process, growth, and ribosome and signal transduction, typical for regulatory and developmental genes that are often highly connected and present in macromolecular complexes and/or stoichiometric pathways, which have frequently been observed before to be overrepresented and retained long after the WGD event itself in several angiosperm and metazoan genomes (Hakes et al., 2007; Freeling, 2009; Bekaert et al., 2011; Rodgers-Melnick et al., 2012). This is confirmed by the more extensive list of full GO terms that are found to be overrepresented in the homoeologs, where again many genes with roles that can be associated with plant regulation and development are found (Supplemental Data Set 1). This pattern of biased gene retention for certain functional classes of genes following WGD in horsetails therefore appears to be similar to what is found in angiosperms.

However, we also observed that a large proportion of the enriched categories included those related to abiotic stress response and nutrient metabolism, including detection of (endogenous and abiotic) stimulus, response to abiotic stimulus, response to (osmotic) stress, response to (organic and inorganic) substance, hyperosmotic response, response to metal ions, chemicals, and nutrients, regulation of phosphorus and phosphate metabolism,

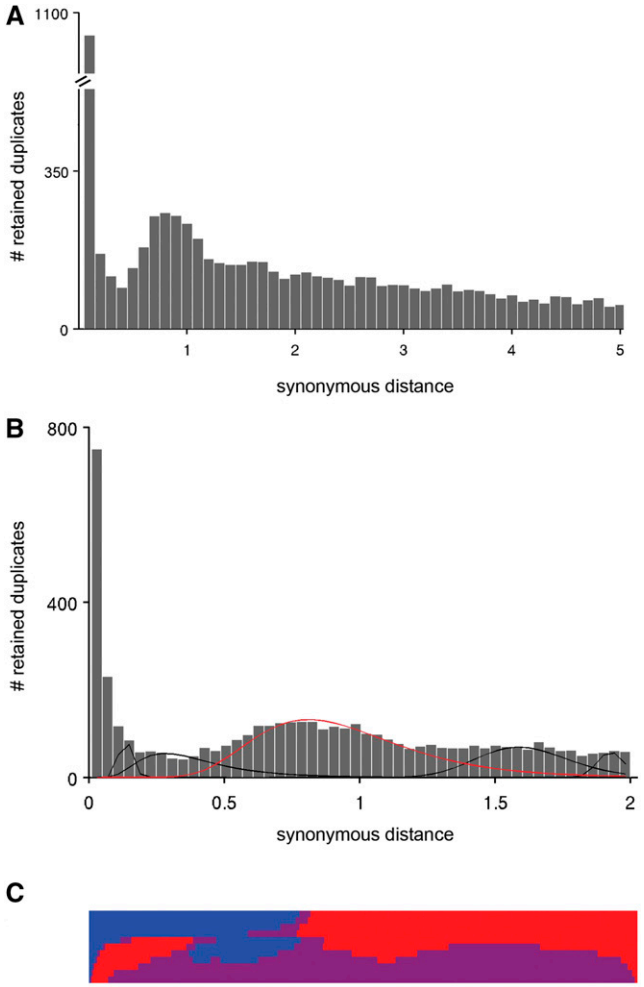

Figure 1. The K_S-Based Age Distribution of the *E. giganteum* Transcriptome Provides Support for a Paleopolyploidy Event.

(A) The full K_S-based age distribution of the paranome is presented. The *x* axis shows the synonymous distance until a K_S cutoff of 5 in bins of 0.1, while the *y* axis shows the number of retained duplicated paralogous gene pairs. Note that the *y* axis is broken because the first bin contains a very high number of recent duplicates.

(B) A subset of **(A)** is shown with K_S values < 2 in bins of 0.04, containing the K_S values that were used for mixture modeling (excluding those with a K_S < 0.1). Note that the *y* axis has a different scale compared with **(A)** because of the different bin size. The component of the Gaussian mixture model as identified by EMMIX (McLachlan et al., 1999), which corresponds to a significant WGD feature based on SiZer analysis (Chaudhuri and Marron, 1999), is plotted on the age distribution in red, while other components are colored in black. An ancient WGD is identified with its peak centered around a K_S of 0.81 (see Table 1).

(C) SiZer output. The transition from the blue to the red color at a K_S of 0.81 indicates a significant bump in the distribution.

and cellular cation homeostasis. Characteristic traits of horsetails include physiological adaptations to nutrient- and oxygen-poor waterlogged soils, where they accumulate disproportionately high amounts of calcium, phosphorus, and potassium, thus playing a keystone role by contributing to the productivity of the total

Table 1. Mixture Modeling and SiZer Analysis of the *E. giganteum* Age Distribution Presented in Figure 1

No. of Duplicates	No. of Components	BIC	Mixture Means (K_S)	Mixture Peaks (K_S)	Variance (K_S)	Proportion	WGD Signature[a]
4834	5	6010.58494	0.137	0.132	0.0005	0.036	No
4834	5	6010.58494	0.395	0.280	0.0401	0.137	No
4834	5	6010.58494	0.944	0.811	0.0949	0.592	Yes
4834	5	6010.58494	1.613	1.586	0.0299	0.192	No
4834	5	6010.58494	1.927	1.925	0.0023	0.044	No

[a]Fitted components corresponding to WGD features are selected based on SiZer analysis.

plant biomass in their ecosystems (Marsh et al., 2000; Husby, 2013).

Phylogenomic Dating of the Paleopolyploidy in *E. giganteum*

We used a phylogenomics dating pipeline to obtain an absolute estimate for the age of the WGD in *E. giganteum*. Our approach is based on a broad taxonomic sampling that includes sequence information from in total 36 angiosperm genomes (Supplemental Figure 1), uses a relaxed clock model that assumes a lognormal distribution on evolutionary rates that should better deal with rate shifts between different branches when taxon sampling is limited (Smith et al., 2010), and allows the implementation of calibrations as lognormal distributions (Supplemental Methods) that represent the error associated with fossil calibration in a more intuitive manner (Forest, 2009). We obtained an age estimate of 92.42 mya for the paleopolyploidy with a lower and higher 90% confidence interval of 75.16 and 112.53 mya, respectively (Figure 3). This is a particularly large confidence interval due to the diffuseness of the resulting absolute age distribution, especially compared with similar WGD age estimates within the angiosperms where such confidence intervals are typically much smaller (Vanneste et al., 2014a). This large error on the WGD age estimate is most likely due to the isolated long branch leading to the node joining the two *E. giganteum* paralogs representing the WGD in each orthogroup.

Because we were not able to obtain a very reliable WGD age estimate for the *E. giganteum* paleopolyploidy through phylogenomic dating, we revisited the use of its K_S distribution. There are a number of caveats in using K_S values to obtain age estimates for WGDs (see Discussion). Nevertheless, careful absolute dating of WGDs has revealed that unambiguous WGD signature peaks in K_S-based age distributions between K_S values of 0.6 and 1.1 usually correspond with a WGD that has occurred somewhere between 50 and 70 mya (Vanneste et al., 2014a). Figure 4 illustrates K_S-based age distributions for a variety of angiosperm species, such as the eudicots Arabidopsis, *Lactuca sativa*, *Solanum tuberosum*, and *Cicer arietinum* as well as the monocots *Oryza sativa*, *Sorghum bicolor*, *Brachypodium distachyon*, *Musa acuminata*, and *Phalaenopsis equestris*. It also includes the K_S-based age distribution for the bryophyte *P. patens*. Although there are some exceptions to this correlation in species that underwent drastic rate shifts (see Discussion), there is a striking correlation between the location of the peaks within the K_S-based age distribution and the inferred absolute age of the WGD event (Vanneste et al., 2014a). Therefore, assuming no drastic rate shifts for *E. giganteum*, based on previous absolute dating of WGDs in different plant lineages, the K_S-based age distribution of *E. giganteum* would suggest that the WGD in horsetails could be as young as 50 to 70 mya. Therefore, given the difficulties in obtaining an absolute date for the WGD based on protein distances, we think we cannot rule out the possibility that the WGDs might be considerably much younger and even coinciding with the K/PG boundary, as shown for many other plants (Fawcett et al., 2009; Vanneste et al., 2014a; see Discussion).

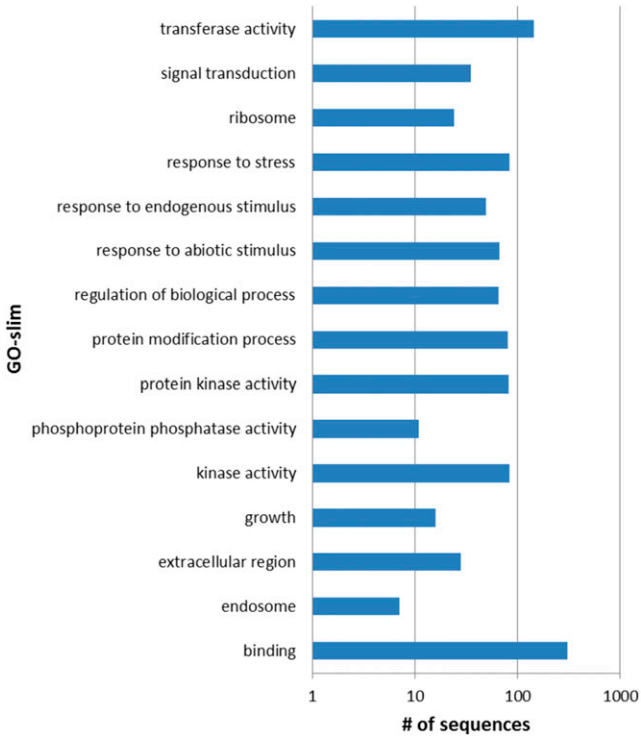

Figure 2. Genes Generated by the Paleopolyploidy Are Enriched for GO Terms with Roles in Plant Regulation and Development.

The bar plot shows the number of genes on the *x* axis (note the logarithmic scale) and the GO-slim categories that were found to be overrepresented in homoeologs on the *y* axis. Many GO-slim categories are related to plant regulation and development, indicating that such genes have preferentially been retained after the WGD. See also Supplemental Data Set 1.

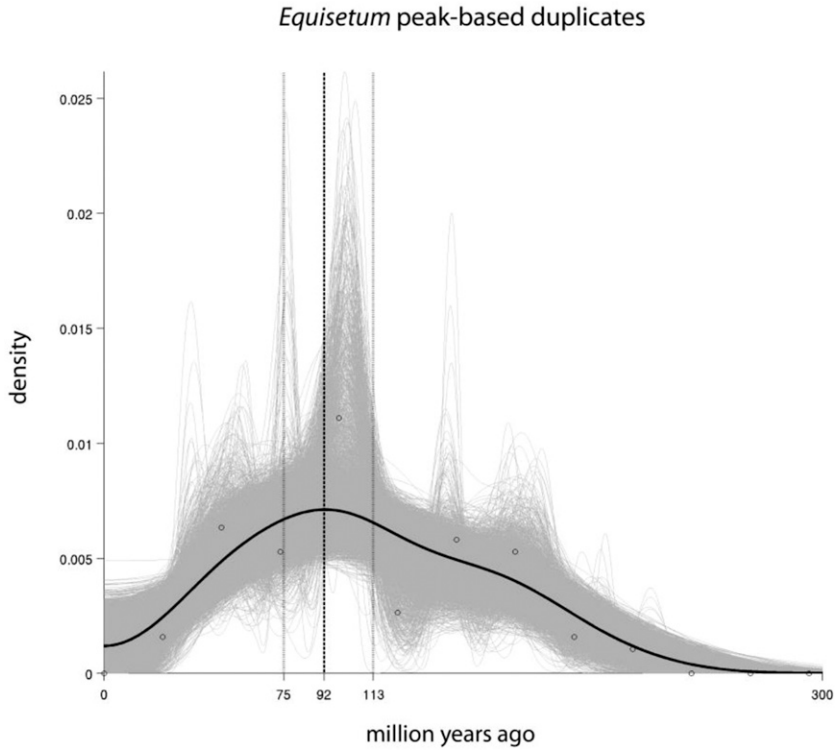

Figure 3. Absolute Age Distribution for the Dated Peak-Based Duplicates Representing the Paleopolyploidy in *E. giganteum*.

The solid black line represents the kernel density estimate of the dated peak-based duplicates, while the vertical dashed black line represents its peak, used as the WGD age estimate. Gray lines represent the density estimates for the 1000 bootstrap replicates, and the vertical black dotted lines represent the corresponding 90% confidence intervals for the WGD age estimate. The original raw distribution of dated peak-based duplicates is also indicated by open circles. The mode used as an estimate for the consensus WGD age is found at 92.42 mya with lower and upper 90% confidence interval boundaries at 75.16 and 112.53 mya, respectively.

DISCUSSION

The *E. giganteum* Transcriptome Confirms That High Chromosome Counts Did Not Evolve through Repeated Rounds of (Recent) Polyploidy

Here, we present unambiguous evidence for a paleopolyploidy event within the horsetails based on a large-scale transcriptome assembly generated by next-generation sequencing of *E. giganteum*. Although transcriptome data cover only genes that are actively being transcribed, these are ideally suited for exploratory analysis of the gene space (Matasci et al., 2014), and (partial) EST-based gene catalogs have been used extensively in the past for the discovery of WGD features in K_S-based age distributions (Blanc and Wolfe, 2004; Schlueter et al., 2004; Sterck et al., 2005; Cui et al., 2006). Our assembly contained more than 90% of core eukaryotic genes (Parra et al., 2007), which suggests it is complete enough to rely on the use of mixture modeling techniques to identify WGD signature peaks before a K_S threshold of 2.0 where K_S saturation and stochasticity can become problematic (Vanneste et al., 2013).

Despite the high chromosome count of *E. giganteum* (n = 108; Leitch and Leitch, 2013), recent polyploidizations were not detected and therefore could not have contributed to this high number. This affirms what had been observed already before based on the analysis of gel electrophoresis banding patterns (Haufler and Soltis, 1986; Haufler, 1987), namely, that high chromosome counts in homosporous ferns did not evolve by repeated rounds of recent polyploidy, at least not within the horsetails. However, this does leave open the question whether the observed chromosome count is the result of a high chromosome number in the ancestor of both ferns and seed plants (Soltis and Soltis, 1987), or rather the result of the paleopolyploidy in *E. giganteum* coupled with a slow loss of genetic material (Haufler, 1987).

It is important to note that a putative paleopolyploidy has also been reported in the ancestor of the polypod fern lineage, estimated to be ~180 million years old, based on a WGD signature peak found in a K_S-based age distribution of duplicate pairs in partial next-generation sequencing data sets of *Ceratopteris richardii* and *Adiantum capillus-veneris* (Barker, 2009; Barker and Wolf, 2010). The latter was interpreted in favor of ferns having undergone a limited set of paleopolyploidies that resulted in high chromosome counts because fern genomes have retained more genetic material from these rare events compared with angiosperms (Barker, 2013). The low gene density of active genes in some fern genomes (Rabinowicz et al., 2005) and their strong correlation between genome size and chromosome number

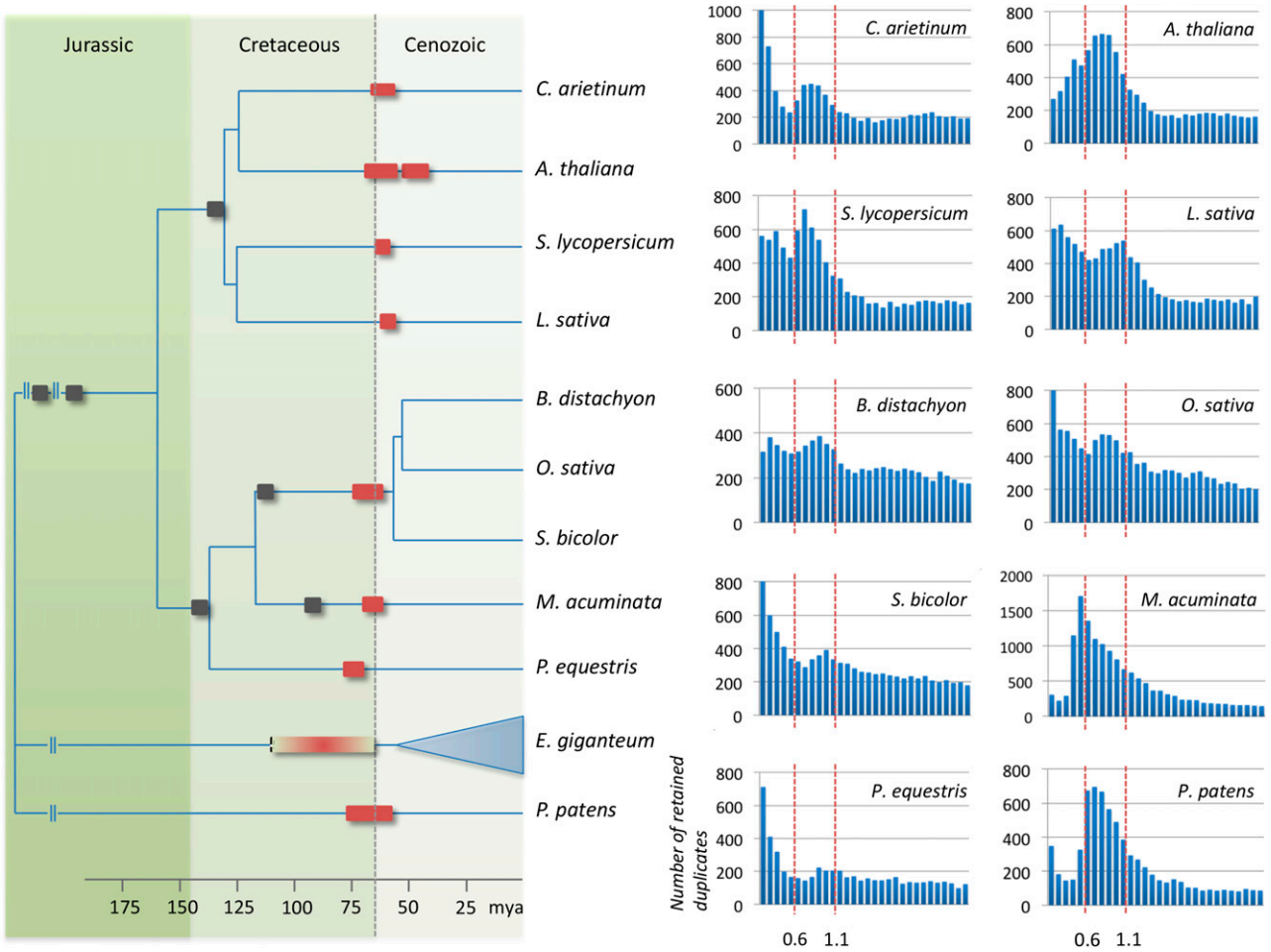

Figure 4. Pruned Tree Topology and K_S Age Distributions for the Green Land Plants.

The left panel presents the topology of a few representative key plant species: *C. arietinum* (chickpea), Arabidopsis (thale cress), *Solanum lycopersicum* (tomato), *L. sativa* (lettuce), *B. distachyon* (purple false brome), *O. sativa* (rice), *S. bicolor* (sorghum), *M. acuminata* (banana), *P. equestris* (orchid), *E. giganteum* (giant horsetail), and *P. patens* (moss). Genome duplications are indicated by colored boxes. Ages for the WGDs are taken from Vanneste et al. (2014a). Red boxes indicate ages of WGDs obtained through absolute dating. Black boxes indicate ages of WGDs obtained from literature. The orange box represents the WGD for *E. giganteum*, as inferred in this study, while the triangle denotes the divergence of the different horsetail crown group species in the Early Cenozoic (Des Marais et al., 2003). The right panel presents K_S-based age distributions for all species included in the tree, with the exception of the distribution for *E. giganteum*, which is presented in more detail on Figure 1. Age distributions were constructed as described in Methods, while sequence data for these species were taken from Vanneste et al. (2014a). The *x* axis always shows the synonymous distance until a K_S cutoff of 3 in bins of 0.1, while the *y* axis shows the number of retained duplicated paralogous gene pairs. The red dotted lines indicate K_S boundaries of 0.6 and 1.1 on all individual age distributions.

(Nakazato et al., 2008) support that gene silencing is not accompanied with gene loss in ferns (Haufler, 2014). Our results for horsetails mirror these from the polypods in so far that both fern clades underwent a separate paleopolyploidy and are characterized by high chromosome counts.

However, it remains to be determined whether the high chromosome count in horsetails is the result of slow loss of genetic material generated by the paleopolyploidy or whether the ancestor of horsetails already possessed a high chromosome number. It appears unlikely that a single paleopolyploidy could have resulted in a chromosome count of $n = 108$ if the ancestor of ferns and seed plants had a low chromosome number, requiring alternative explanations to resolve this phenomenon. Older paleopolyploidies could be hidden in the tail of the age distribution, masked by K_S saturation and stochasticity effects coupled with ongoing gene loss (Vanneste et al., 2013). The high chromosome count of horsetails could then indeed be caused by a few paleopolyploidies of which a large fraction of genetic material has been retained (Haufler, 1987). Alternatively, the ancestor of all vascular plants could have exhibited a relatively high chromosome number (Soltis and Soltis, 1987) so that a single paleopolyploidy could have resulted in the very high chromosome number in horsetails. We emphasize that both theories are not mutually exclusive.

Untangling these events will most likely remain very challenging, even as full genome sequence information becomes available in the future for other horsetails and their nearest relatives. A powerful integrated syntenic and phylogenomic approach has recently been used to demonstrate the occurrence of a shared WGD among the monocots that was too old to be detected within any single monocot genome (Jiao et al., 2014). However, because all extant horsetails diverged relatively recent in the Early Cenozoic and their nearest extant neighbors diverged ~354 mya ago, close to the origin of the vascular plants (Schneider et al., 2004), almost the entire history of the horsetails is contained within one long branch spanning ~300 million years for which no other species are available that could facilitate either syntenic or phylogenomic approaches.

Phylogenomic Dating Places the Paleopolyploidy Somewhere in the Late Cretaceous

Using a powerful phylogenomic approach (Vanneste et al., 2014a), we dated the paleopolyploidy event at ~92 mya, in the Late Cretaceous. However, this date should be handled with caution for several reasons. Dating of long isolated branches has generally proven difficult for other species such as eucalyptus (*Eucalyptus grandis*; Myburg et al., 2014) and orchid (*Phalaenopsis equestris*; Cai et al., 2015), where their more isolated position typically results in a larger uncertainty of the eventual estimate. This effect appears particularly pronounced here as evidenced by the almost 50% of orthogroups that failed to converge (see Methods) and the relatively large 90% confidence interval between 75.16 and 112.53 mya (Figure 3). As mentioned before, this is most likely due to the fact that the entire branch carrying the node joining the homoeologs is extremely long so that it most likely encompasses a vast number of combinations of rate and time that are notoriously difficult to disentangle (Magallón, 2010).

Although some future advances still can be attained by the development of novel uncorrelated relaxed clock models that are better equipped to deal with rate shifts and long isolated branches (Li and Drummond, 2012; Baele et al., 2013), due to the lack of any near outgroup species, dating the paleopolyploidy will also remain very difficult to resolve even when full genome sequence information for other horsetail species and their nearest neighbors becomes available. We therefore revisited the use of K_S-based age distributions for inferring absolute dates. Converting K_S values to absolute ages can be problematic because it is heavily dependent on the assumption of a strict molecular clock and the employed rate of synonymous substitutions per give time period, which can lead to drastically different estimates (Fawcett et al., 2009). However, Figure 4 demonstrates that there is now sufficient evidence through absolute dating that, in the absence of any severe rate accelerations or decelerations, a WGD signature peak K_S value between 0.6 and 1.1 corresponds to a WGD that took place somewhere between 50 and 70 mya (Vanneste et al., 2014a). Whether this also applies to *Equisetum* depends on whether this genus experienced any strong rate shifts in the last ~100 million years that could misplace its WGD signature peak at lower or higher K_S values. Unfortunately, to the best of our knowledge, no such information is available. However, evolutionary rates seem to be correlated strongly with some life history traits such as generation time. Faster and slower generating species generally exhibit faster and slower evolutionary rates, both in flowering plants (Smith and Donoghue, 2008) and in leptosporangiate ferns (Korall et al., 2010; Zhong et al., 2014). Since *E. giganteum* is neither a small, fast-growing, and reproducing weed nor possesses any arborescence, which typically is associated with rate accelerations and decelerations, respectively, its WGD signature peak location at a K_S of 0.81 would suggest an alternative estimate for the paleopolyploidy somewhere between 50 and 70 mya. Because fossil evidence indicates that horsetails have an extended history dating back to the Late Jurassic, in which they experienced little evolutionary change, their associated evolutionary rate in the last ~100 million years most likely also has been relatively stable. The alternative date of 50 to 70 mya for the WGD in *E. giganteum* is critically dependent on the assumption that horsetails experienced average substitution rates similar to angiosperms, to which it is worth mentioning that at least some clades of polypod ferns do not seem to conform, as they experienced an elevated rate of mutation compared with angiosperms (Rothfels and Schuettpelz, 2014).

In conclusion, the paleopolyploidy that horsetails experienced most likely took place somewhere in the Late Cretaceous between 100 and 66 mya. Although we acknowledge that this presents a broad distribution, we stress that resolving this will remain a particularly challenging task even if additional sequence data is obtained, due to the long and isolated branch of this lineage. Absolute molecular dating places the paleopolyploidy at ~92 mya, but has great difficulty in separating the contribution of rate and time in this single branch than spans almost ~300 million years. Interpolating K_S-based age distributions places the paleopolyploidy in association with the K-Pg boundary, as for many other angiosperms and the moss *P. patens*, but hinges heavily on the assumption that horsetails share similar evolutionary rates with angiosperms.

The Paleopolyploidy May Have Contributed to the Long-Term Survival and Evolutionary Success of Horsetails

The discovery of a WGD within the horsetails helps to shed some light on the debate concerning the evolutionary importance of polyploidy. Although the prevalence of WGD within the angiosperms is well established and the discussion has now shifted toward which degree of importance can be attributed to WGDs in the light of evolutionary success (Van de Peer et al., 2009; Bekaert et al., 2011; Mayrose et al., 2011, 2015; Arrigo and Barker, 2012; Soltis et al., 2014), both the prevalence and evolutionary consequences of WGDs remain largely unexplored in nonangiosperm plant clades (Soltis and Soltis, 2013). Here, we demonstrated that a paleopolyploidy event has also occurred within the horsetails. In combination with suggested evidence for a WGD in the polypod ferns (Barker, 2009; Barker and Wolf, 2010) and the moss *P. patens* (Rensing et al., 2008), this would indicate that (paleo)polyploidy, despite not being responsible for the high chromosome counts in some homosporous fern lineages, might be a widespread phenomenon within the land plants that is not limited to just the angiosperms.

However, elucidating the importance of WGD for evolutionary success in the horsetails and ferns in general remains problematic. Although estimates of polyploid speciation within ferns range from 7% (Otto and Whitton, 2000) to 31.37% (Wood et al., 2009), no link

between polyploidy and species richness has been found (Wood et al., 2009). Consisting of only 15 species, the horsetails are not a particularly successful example in terms of species richness. However, there are more ways to measure evolutionary success than just species numbers. The evolutionary history of horsetails as perhaps the oldest extant vascular plant lineage that has retained multiple ancestral features despite having survived past multiple geological times scales and extinction events (Husby, 2013) is of particular interest in light of the paleopolyploidy they experienced. Although the exact timing of this event remains to be determined (see previous section), its position toward the end of the Mesozoic in the Late Cretaceous coincides with a very turbulent period for ferns as a transition phase between the Paleozoic and Cenozoic, during which they had to surrender their title of most dominant plant life form to the angiosperms (DiMichele and Phillips, 2002). In response, many fern lineages underwent radiations during the Cretaceous as an ecological opportunistic reaction to the rise of angiosperms by exploiting and diversifying in novel generated niches such as forest floors and canopies (Schneider et al., 2004; Schuettpelz and Pryer, 2009), which indicates that the survival and success of ferns at that time most likely was determined by their capacity to adapt quickly to changing ecological and environmental conditions. Similarly, the K-Pg mass extinction event at the end of the Mesozoic represented a very drastic challenge for ferns and angiosperms alike (Rohde and Muller, 2005; Vajda and McLoughlin, 2007).

The ancestors of horsetails, the Sphenophyta, which included both herbaceous and arborescent forms, were successful and dominant components of fern vegetation during the Paleozoic, but they became less diverse and increasingly limited to their herbaceous forms during the Mesozoic by specializing in colonizing disturbed and moist habitats (Husby, 2013). The placement of the paleopolyploidy during the Late Cretaceous, where a quick response to both the rise of the angiosperms and the K-Pg extinction event was required to survive, therefore opens up the intriguing possibility that WGD contributed toward the survival of the horsetails during this period. Although fossil data indicate that the morphology of horsetails has remained remarkably stable since the Jurassic (Channing et al., 2011) and possibly even the Triassic (Hauke, 1963), essential changes in plant physiology that allowed survival during this turbulent period are not necessarily captured in the fossil record. Functional enrichment analysis of homoeologs generated by the WGD unveiled that in particular many genes with roles in plant regulation and development, as well as abiotic stress response and nutrient uptake, have been retained (Figure 2; Supplemental Data Set 1). We speculate that duplicated genes with functions that aid in adaptation to moist and disturbed habitats may have been generated by this event, allowing horsetails to specialize their physiology despite their overall remarkably stable morphology.

METHODS

Next-Generation Sequencing and de Novo Assembly of the *Equisetum giganteum* Transcriptome

Tissues were collected from the stem, leaf, and strobili (the sporangium-bearing organs) of *E. giganteum*. Biological material was obtained with permission from the botanical collections of the University of Pretoria. RNA was extracted using a modified CTAB protocol as described previously (Chang et al., 1993). Illumina library preparation and RNA-sequencing were performed as described by Mizrachi et al. (2010). Approximately 220 million reads (paired-end, 40-40 bases) were assembled using Velvet (Zerbino and Birney, 2008), followed by Oases using a k-mer of 41 (Schulz et al., 2012), which resulted in 34,282 transcripts. TransDecoder was subsequently employed (Haas et al., 2013), using the Pfam search option and a minimal length of 99 amino acids, resulting in a set of 25,763 open reading frames that were used as our gene catalog.

We used CEGMA (v2.5) (Parra et al., 2007) to assess the completeness of our transcriptome assembly. This software identifies the presence of a set of core genes that are highly conserved and expected to be present in all eukaryotic species by comparing the assembly with a set of 248 precomputed core eukaryotic genes. Since the latter are housekeeping genes expected to be basally expressed, they can be used to assess the completeness of the gene space in our transcriptome assembly.

Construction of the *E. giganteum* K_s Age Distribution

The K_S-based age distribution of *E. giganteum* was constructed as described previously (Vanneste et al., 2013). In summary, the paranome was built by performing an all-against-all BLASTP search of all genes with an E-value cutoff of e^{-10}, after which gene families were built with the mclblastline pipeline (v10-201) (micans.org/mcl; Enright et al., 2002). Each gene family was aligned using MUSCLE (v3.8.31) (Edgar, 2004), and K_S estimates for all pairwise comparisons within a gene family were obtained through maximum likelihood estimation using the CODEML program (Goldman and Yang, 1994) of the PAML package (v4.4c) (Yang, 2007). Gene families were then divided into subfamilies for which K_S estimates between members did not exceed a value of 5. To correct for the redundancy of K_S values [a gene family of n members produces $n(n-1)/2$ pairwise K_S estimates for $n-1$ retained duplication events], for each subfamily an amino acid phylogenetic tree was constructed using PhyML (Guindon et al., 2010) under default settings. For each duplication node in the resulting phylogenetic tree, all m K_S estimates between the two child clades were added to the K_S distribution with a weight $1/m$, so that the weights of all K_S estimates for a single duplication event sum up to one.

Mixture Modeling

The *E. giganteum* transcriptome displays a prominent peak in its age distribution between a K_S of 0.6 and 1.2. This feature most likely corresponds to a WGD in the evolutionary past of this species because WGDs result in peaks of contemporarily created genes that are recognizable as a peak superimposed on an exponential L-shaped decay distribution of paralogous genes created by SSDs (Blanc and Wolfe, 2004), so that the compounded distribution consists of both log-transformed SSD exponential functions and WGD Gaussian functions (Schlueter et al., 2004). We employed mixture modeling to confirm this WGD signature using the EMMIX software (McLachlan et al., 1999) to fit a mixture model of Gaussian distributions to the log-transformed K_S distribution of the *E. giganteum* transcriptome. All observations ≤ 0.1 K_S were excluded for analysis to avoid the incorporation of allelic and/or splice variants and to prevent the fitting of a component to infinity (Schlueter et al., 2004), while all observations >2.0 K_S were removed because K_S saturation and stochasticity can mislead mixture modeling above this range (Vanneste et al., 2013). One to five components per mixture model were fitted increasingly to the data, using 1000 random and 100 k-mean starts. The Bayesian Information Criterion (BIC; Schwarz, 1978), which strongly penalizes increases in the number of model parameters to avoid overfitting, was used to select the best number of components. The mean and variance of each component were back-transformed afterwards. BIC selected the maximum number of allowed components, hinting that the usage of the BIC model selection criterion still

results in overfitting of components (Naik et al., 2007), as also observed before (Vekemans et al., 2012). We therefore employed the SiZer software (Chaudhuri and Marron, 1999) to help identify significant features ($\alpha = 0.05$) in the age distribution. This software searches for changes in the first derivative of a range of kernel density estimates with different smoothing bandwidths to distinguish peaks in the distribution that represent true features from those that represent noise. Components of the mixture model corresponding to significant features as identified by SiZer were then considered as bona fide WGD signature peaks.

Functional Enrichment

We performed functional enrichment analysis to gain insight into the functionality of novel genes generated by the WGD (i.e., homoeologs). To assign homoeologs, we extracted duplicate pairs lying under the WGD signature peak in the age distribution between a K_S of 0.6 and 1.2. Although duplicate pairs lying under this peak can also have been generated by SSDs that occurred during the same time frame, mechanistic modeling approaches have demonstrated previously that their majority consists of homoeologs (Maere et al., 2005a). K_S saturation effects are expected to be absent in this K_S range, but K_S variation can already be problematic for individual duplicate pairs (Vanneste et al., 2013). To increase the robustness of our results, rather than accepting all duplicate pairs lying under the WGD peak, we therefore scored each duplication event separately based on the phylogenetic trees built previously for each (sub)family. The median K_S was calculated for each individual node in all trees based on all possible K_S comparisons between its terminals. Nodes with less than three possible K_S comparisons were not considered. Nodes with a median K_S between 0.6 and 1.2 were considered as WGD nodes, and their terminals were assigned as homoeologs. In total, 975 homoeologs were collected.

All genes were then BLASTed against the nonredundant protein database at NCBI with an E-value cutoff of e^{-10}. BLAST results were used as input for the command line version of Blast2GO (v2.3.5), which we used to assign GO terms to our gene catalog based on the functional transfer from a standard set of homologous sequences (Conesa et al., 2005). GO terms are labels that aim to assign attributes concerning functionality to genes based on a universal controlled vocabulary (Ashburner et al., 2000). We then used BiNGO (v2.42) (Maere et al., 2005b) for Cytoscape (v2.8.2) to find GO terms that were overrepresented in the set of homoeologs compared with the total gene catalog of *E. giganteum*.

Absolute Dating

We tried to obtain an absolute date for the WGD using an absolute dating pipeline as described previously (Vanneste et al., 2014a). In summary, homoeologs, also referred to as peak-based duplicates in this context, were used for dating by constructing orthogroups that contain a homoeologous gene pair plus several orthologs from other plant species as identified by InParanoid (v4.1) (Ostlund et al., 2010) using a broad taxonomic sampling within the angiosperms as described in the Supplemental Methods. The required percentage of segment overlap and coverage used by InParanoid were lowered to 0.25 and 0.2%, respectively, to accommodate for the large evolutionary distance between *E. giganteum* and other plant clades. In total, 145 orthogroups based on peak-based duplicates could be collected. The node joining the two *E. giganteum* homoeologs was then dated using the BEAST package (v1.7) (Drummond et al., 2012) under an uncorrelated relaxed clock model and a LG+G (four rate categories) evolutionary model. A starting tree with branch lengths satisfying all fossil prior constraints was created according to the consensus APGIII phylogeny (Bremer et al., 2009). Fossil calibrations are described in the Supplemental Methods. A run without data was performed to ensure proper placement of the marginal calibration prior distributions (Heled and Drummond, 2012). The Markov chain Monte Carlo for each orthogroup was run for 10 million generations, sampling every 1000 generations resulting in a sample size of 10,000. The resulting trace files of all orthogroups were processed automatically with LogAnalyzer (part of the BEAST package) using a burn-in of 1000 samples to ensure proper convergence (minimum effective sample size for all statistics at least 200).

However, only 78 out of these 145 orthogroups (53.8%) passed this criterion, which is substantially lower than WGDs dated within the angiosperms where on average 93.7% of orthogroups based on peak-based duplicates passed (Vanneste et al., 2014a). We therefore manually evaluated the trace files of all orthogroups with Tracer (v1.5) (Drummond et al., 2012) that allows visual exploration of the behavior of the Markov chain Monte Carlo. This inspection indicated that a substantial fraction of orthogroups did not pass our stringent criterion of ESS > 200 for all statistics because the trace for the WGD node did not converge properly and resulted in a very diffuse and skewed distribution with in particular very long upper tails (Supplemental Figure 2), most likely due to the lack of a proper upper bound on this very long and isolated branch. Visual inspection indicated that this was also an issue in orthogroups that passed our stringent criterion, although to a much lesser extent. To cope with this issue and to ensure that the best WGD age estimate was always taken for each accepted orthogroup, rather than taking the LogAnalyser estimate that is based on the mean of the resulting distribution for orthogroups where all statistics had an ESS > 200, we used the peak of this distribution as the best estimate for the age of the WGD node using the lognfit function in Matlab (MATLAB R2011a; The MathWorks), since unbounded distributions of such estimates have been demonstrated to follow a lognormal distribution (Morrison, 2008). The WGD age estimates obtained from all these orthogroups were then grouped into one absolute age distribution, for which we employed the KDE toolbox for Matlab (available at http://www.mathworks.com/matlabcentral/fileexchange/17204-kernel-density-estimation; retrieved March 21, 2013) to find the density estimate that best described the overall absolute age distribution (Botev et al., 2010) by taking its mode as the best WGD age estimate for the *E. giganteum* WGD and using a bootstrapping procedure (Hall and Kang, 2001) to obtain 90% confidence intervals on the WGD age estimate.

Accession Numbers

Raw reads and assembled contigs are available at the European Nucleotide Archive under project PRJEB9341 (http://www.ebi.ac.uk/ena/data/view/PRJEB9341). Assembled transcripts and predicted genes are available at http://bioinformatics.psb.ugent.be/supplementary_data/kenes/Equisetum/.

Supplemental Data

Supplemental Figure 1. Representation of the structure of an orthogroup.

Supplemental Figure 2. Many orthogroups display diffuse and skewed distributions for the age of the node joining the two WGD paralogs.

Supplemental Methods. Fossil calibrations.

Supplemental Data Set 1. GO terms significantly enriched in homoeologs.

ACKNOWLEDGMENTS

We thank the Manie van der Schijff Botanical Gardens of the University of Pretoria, a division of the Plant Science Department, for botanical material and Martin Ranik, Andrew Dos Santos, and Charles Hefer for assistance with RNA isolation, RNA-sequencing library preparation, and transcriptome assembly. This work was supported by Sappi (South Africa) through the Forest Molecular Genetics Programme, the National Research Foundation (UID 5571255 and 86936), and the Department of

Science and Technology of South Africa. Further support was provided by Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks"). Y.V.d.P. acknowledges support from the European Union Seventh Framework Programme (FP7/2007-2013) under ERC Advanced Grant Agreement 322739-DOUBLE-UP. The work was carried out using the Stevin Supercomputer Infrastructure at Ghent University, funded by Ghent University, the Hercules Foundation, and the Flemish Government Department EWI.

AUTHOR CONTRIBUTIONS

E.M. is the lead investigator and conceived the WGD study with Y.V.d.P. K.V. wrote the article with Y.V.d.P. and E.M. L.S. identified open reading frames and protein sequences and contributed to WGD analysis. K.V. and E.M. performed most of the data analysis. A.A.M. was a co-investigator for transcriptome sequencing and assembly and edited the manuscript. All authors have read and commented on the article.

Received February 18, 2015; revised April 3, 2015; accepted April 28, 2015; published May 22, 2015.

REFERENCES

Abbasi, A.A. (2010). Piecemeal or big bangs: correlating the vertebrate evolution with proposed models of gene expansion events. Nat. Rev. Genet. **11:** 166.

Arrigo, N., and Barker, M.S. (2012). Rarely successful polyploids and their legacy in plant genomes. Curr. Opin. Plant Biol. **15:** 140–146.

Ashburner, M., et al.; The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. Nat. Genet. **25:** 25–29.

Baele, G., Li, W.L., Drummond, A.J., Suchard, M.A., and Lemey, P. (2013). Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. Mol. Biol. Evol. **30:** 239–243.

Barker, M.S. (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, M.S. (2013). Karyotype and Genome Evolution in Pteridophytes. (Vienna, Austria: Springer).

Barker, M.S., Kane, N.C., Matvienko, M., Kozik, A., Michelmore, R.W., Knapp, S.J., and Rieseberg, L.H. (2008). Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol. Biol. Evol. **25:** 2445–2455.

Barker, M.S., and Wolf, P.G. (2010). Unfurling fern biology in the genomics age. Bioscience **60:** 177–185.

Bekaert, M., Edger, P.P., Pires, J.C., and Conant, G.C. (2011). Two-phase resolution of polyploidy in the Arabidopsis metabolic network gives rise to relative and absolute dosage constraints. Plant Cell **23:** 1719–1728.

Blanc, G., and Wolfe, K.H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell **16:** 1667–1678.

Botev, Z.I., Grotowski, J.F., and Kroese, D.P. (2010). Kernel density estimation via diffusion. Ann. Stat. **38:** 2916–2957.

Bremer, B., et al. (2009). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot. J. Linn. Soc. **161:** 105–121.

Cai, J., et al. (2015). The genome sequence of the orchid *Phalaenopsis equestris*. Nat. Genet. **47:** 65–72.

Chang, S., Puryear, J., and Cairney, J. (1993). A simple and efficient method for isolating RNA from pine trees. Plant Mol. Biol. Rep. **11:** 113–116.

Channing, A., Zamuner, A., Edwards, D., and Guido, D. (2011). Equisetum thermale sp. nov. (Equisetales) from the Jurassic San Agustín hot spring deposit, Patagonia: anatomy, paleoecology, and inferred paleoecophysiology. Am. J. Bot. **98:** 680–697.

Chaudhuri, P., and Marron, J. (1999). SiZer for exploration of structures in curves. J. Am. Stat. Assoc. **94:** 807–823.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics **21:** 3674–3676.

Crane, P.R. (1987). Vegetational consequences of the angiosperm diversification. In The Origins of Angiosperms and their Biological Consequences, E.M. Friis, W.G. Chaloner, and P.R. Crane, eds (Cambridge, UK: Cambridge University Press), pp. 107–144.

Cui, L., et al. (2006). Widespread genome duplications throughout the history of flowering plants. Genome Res. **16:** 738–749.

Des Marais, D.L., Smith, A.R., Britton, D.M., and Pryer, K.M. (2003). Phylogenetic relationships and evolution of extant horsetails, equisetum, based on chloroplast DNA sequence data (rbcL and trnL-F). Int. J. Plant Sci. **164:** 737–751.

DiMichele, W.A., and Phillips, T.L. (2002). The ecology of Paleozoic ferns. Rev. Palaeobot. Palynol. **119:** 143–159.

Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. **29:** 1969–1973.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:** 1792–1797.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. **30:** 1575–1584.

Fawcett, J.A., Maere, S., and Van de Peer, Y. (2009). Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc. Natl. Acad. Sci. USA **106:** 5737–5742.

Forest, F. (2009). Calibrating the Tree of Life: fossils, molecules and evolutionary timescales. Ann. Bot. (Lond.) **104:** 789–794.

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu. Rev. Plant Biol. **60:** 433–453.

Goldman, N., and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11:** 725–736.

Grewe, F., Guo, W., Gubbels, E.A., Hansen, A.K., and Mower, J.P. (2013). Complete plastid genomes from *Ophioglossum californicum*, *Psilotum nudum*, and *Equisetum hyemale* reveal an ancestral land plant genome structure and resolve the position of Equisetales among monilophytes. BMC Evol. Biol. **13:** 8.

Guillon, J.M. (2004). Phylogeny of horsetails (Equisetum) based on the chloroplast rps4 gene and adjacent noncoding sequences. Syst. Bot. **29:** 251–259.

Guillon, J.M. (2007). Molecular phylogeny of horsetails (Equisetum) including chloroplast atpB sequences. J. Plant Res. **120:** 569–574.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. **59:** 307–321.

Haas, B.J., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. **8:** 1494–1512.

Hakes, L., Pinney, J.W., Lovell, S.C., Oliver, S.G., and Robertson, D.L. (2007). All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol. **8:** R209.

Hall, P., and Kang, K.H. (2001). Bootstrapping nonparametric density estimators with empirically chosen bandwidths. Ann. Stat. **29:** 1443–1468.

Haufler, C.H. (1987). Electrophoresis is modifying our concepts of evolution in homosporous pteridophytes. Am. J. Bot. **74:** 953–966.

Haufler, C.H. (2002). Homospory 2002: An odyssey of progress in pteridophyte genetics and evolutionary biology. Bioscience **52:** 1081–1093.

Haufler, C.H. (2014). Ever since Klekowski: testing a set of radical hypotheses revives the genetics of ferns and lycophytes. Am. J. Bot. **101:** 2036–2042.

Haufler, C.H., and Soltis, D.E. (1986). Genetic evidence suggests that homosporous ferns with high chromosome numbers are diploid. Proc. Natl. Acad. Sci. USA **83:** 4389–4393.

Hauke, R. (1963). A taxonomic monograph of the genus Equisetum subgenus Hippochaete. Beih. Nova Hedwigia **8:** 123.

Heled, J., and Drummond, A.J. (2012). Calibrated tree priors for relaxed phylogenetics and divergence time estimation. Syst. Biol. **61:** 138–149.

Husby, C. (2013). Biology and functional ecology of Equisetum with emphasis on the giant horsetails. Bot. Rev. **79:** 147–177.

Jiao, Y., Li, J., Tang, H., and Paterson, A.H. (2014). Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. Plant Cell **26:** 2792–2802.

Jiao, Y., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. Nature **473:** 97–100.

Khandelwal, S. (1990). Chromosome evolution in the genus Ophioglossum L. Bot. J. Linn. Soc. **102:** 205–217.

Kimura, M. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature **267:** 275–276.

Klekowski, E.J. (1973). Sexual and subsexual systems in homosporous pteridophytes: New hypothesis. Am. J. Bot. **60:** 535–544.

Klekowski, E.J., Jr., and Baker, H.G. (1966). Evolutionary significance of polyploidy in the pteridophyta. Science **153:** 305–307.

Korall, P., Schuettpelz, E., and Pryer, K.M. (2010). Abrupt deceleration of molecular evolution linked to the origin of arborescence in ferns. Evolution **64:** 2786–2792.

Leitch, I., and Leitch, A. (2013). Genome size diversity and evolution in land plants. In Plant Genome Diversity, Vol. 2, I.J. Leitch, ed (Vienna, Austria: Springer), pp. 307–322.

Li, W.H. (1997). Molecular Evolution. (Sunderland, MA: Sinauer Associates).

Li, W.L., and Drummond, A.J. (2012). Model averaging and Bayes factor calculation of relaxed molecular clocks in Bayesian phylogenetics. Mol. Biol. Evol. **29:** 751–761.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science **290:** 1151–1155.

Lynch, M., and Conery, J.S. (2003). The evolutionary demography of duplicate genes. J. Struct. Funct. Genomics **3:** 35–44.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. (2005a). Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. USA **102:** 5454–5459.

Maere, S., Heymans, K., and Kuiper, M. (2005b). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics **21:** 3448–3449.

Magallón, S. (2010). Using fossils to break long branches in molecular dating: a comparison of relaxed clocks applied to the origin of angiosperms. Syst. Biol. **59:** 384–399.

Manton, I. (1950). Problems of Cytology and Evolution in the Pteridophyta. (Cambridge, UK: Cambridge University Press).

Marsh, A.S., Arnone, J.A., Bormann, B.T., and Gordon, J.C. (2000). The role of Equisetum in nutrient cycling in an Alaskan shrub wetland. J. Ecol. **88:** 999–1011.

Matasci, N., et al. (2014). Data access for the 1,000 Plants (1KP) project. Gigascience **3:** 17.

Mayrose, I., Zhan, S.H., Rothfels, C.J., Arrigo, N., Barker, M.S., Rieseberg, L.H., and Otto, S.P. (2015). Methods for studying polyploid diversification and the dead end hypothesis: a reply to Soltis et al. (2014). New Phytol. **206:** 27–35.

Mayrose, I., Zhan, S.H., Rothfels, C.J., Magnuson-Ford, K., Barker, M.S., Rieseberg, L.H., and Otto, S.P. (2011). Recently formed polyploid plants diversify at lower rates. Science **333:** 1257.

McLachlan, G., Peel, D., Basford, K., and Adams, P. (1999). The EMMIX software for the fitting of mixtures of normal and t-components. J. Stat. Softw. **4:** 2.

Mizrachi, E., Hefer, C.A., Ranik, M., Joubert, F., and Myburg, A.A. (2010). De novo assembled expressed gene catalog of a fast-growing Eucalyptus tree produced by Illumina mRNA-Seq. BMC Genomics **11:** 681.

Morrison, D.A. (2008). How to summarize estimates of ancestral divergence times. Evol. Bioinform. Online **4:** 75–95.

Myburg, A.A., et al. (2014). The genome of *Eucalyptus grandis*. Nature **510:** 356–362.

Naik, P.A., Shi, P., and Tsai, C.L. (2007). Extending the akaike information criterion to mixture regression models. J. Am. Stat. Assoc. **102:** 244–254.

Nakasugi, K., Crowhurst, R.N., Bally, J., Wood, C.C., Hellens, R.P., and Waterhouse, P.M. (2013). De novo transcriptome sequence assembly and analysis of RNA silencing genes of *Nicotiana benthamiana*. PLoS ONE **8:** e59534.

Nakazato, T., Barker, M., Rieseberg, L., and Gastony, G. (2008). Evolution of the nuclear genome of ferns and lycophytes. In Biology and Evolution of Ferns and Lycophytes, T.A. Ranker and C. H. Haufler, eds (Cambridge, UK: Cambridge University Press), pp. 175–198.

Nystedt, B., et al. (2013). The Norway spruce genome sequence and conifer genome evolution. Nature **497:** 579–584.

Ostlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D.N., Roopra, S., Frings, O., and Sonnhammer, E.L. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. **38:** D196–D203.

Otto, S.P., and Whitton, J. (2000). Polyploid incidence and evolution. Annu. Rev. Genet. **34:** 401–437.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23:** 1061–1067.

Pryer, K.M., Schuettpelz, E., Wolf, P.G., Schneider, H., Smith, A.R., and Cranfill, R. (2004). Phylogeny and evolution of ferns (monilophytes) with a focus on the early leptosporangiate divergences. Am. J. Bot. **91:** 1582–1598.

Rabinowicz, P.D., Citek, R., Budiman, M.A., Nunberg, A., Bedell, J.A., Lakey, N., O'Shaughnessy, A.L., Nascimento, L.U., McCombie, W.R., and Martienssen, R.A. (2005). Differential methylation of genes and repeats in land plants. Genome Res. **15:** 1431–1440.

Rensing, S.A., et al. (2008). The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science **319:** 64–69.

Rodgers-Melnick, E., Mane, S.P., Dharmawardhana, P., Slavov, G.T., Crasta, O.R., Strauss, S.H., Brunner, A.M., and Difazio, S.P. (2012). Contrasting patterns of evolution following whole genome versus tandem duplication events in Populus. Genome Res. **22:** 95–105.

Rohde, R.A., and Muller, R.A. (2005). Cycles in fossil diversity. Nature **434:** 208–210.

Rothfels, C.J., and Schuettpelz, E. (2014). Accelerated rate of molecular evolution for vittarioid ferns is strong and not driven by selection. Syst. Biol. **63:** 31–54.

Rothwell, G.W. (1996). Pteridophytic evolution: An often underappreciated phytological success story. Rev. Palaeobot. Palynol. **90:** 209–222.

Schlueter, J.A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J.J., and Shoemaker, R.C. (2004). Mining EST databases to resolve evolutionary events in major crop species. Genome **47:** 868–876.

Schneider, H., Schuettpelz, E., Pryer, K.M., Cranfill, R., Magallón, S., and Lupia, R. (2004). Ferns diversified in the shadow of angiosperms. Nature **428:** 553–557.

Schranz, M.E., Mohammadin, S., and Edger, P.P. (2012). Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. Curr. Opin. Plant Biol. **15:** 147–153.

Schuettpelz, E., and Pryer, K.M. (2007). Fern phylogeny inferred from 400 leptosporangiate species and three plastid genes. Taxon **56:** 1037–1050.

Schuettpelz, E., and Pryer, K.M. (2009). Evidence for a Cenozoic radiation of ferns in an angiosperm-dominated canopy. Proc. Natl. Acad. Sci. USA **106:** 11200–11205.

Schulz, M.H., Zerbino, D.R., Vingron, M., and Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics **28:** 1086–1092.

Schwarz, G. (1978). Estimating the dimension of a model. Ann. Stat. **6:** 461–464.

Smith, S.A., Beaulieu, J.M., and Donoghue, M.J. (2010). An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. Proc. Natl. Acad. Sci. USA **107:** 5897–5902.

Smith, S.A., and Donoghue, M.J. (2008). Rates of molecular evolution are linked to life history in flowering plants. Science **322:** 86–89.

Soltis, D.E. (1986). Genetic evidence for diploidy in *Equisetum.* Am. J. Bot. **73:** 908–913.

Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D., Depamphilis, C.W., Wall, P.K., and Soltis, P.S. (2009). Polyploidy and angiosperm diversification. Am. J. Bot. **96:** 336–348.

Soltis, D.E., and Soltis, P.S. (1987). Polyploidy and breeding systems in homosporous pteridophyta - a reevaluation. Am. Nat. **130:** 219–232.

Soltis, D.E., and Soltis, P.S. (1992). The distribution of selfing rates in homosporous ferns. Am. J. Bot. **79:** 97–100.

Soltis, D.E., Visger, C.J., and Soltis, P.S. (2014). The polyploidy revolution then…and now: Stebbins revisited. Am. J. Bot. **101:** 1057–1078.

Soltis, P.S., and Soltis, D.E. (1990). Evolution of inbreeding and outcrossing in ferns and fern-allies. Plant Species Biol. **5:** 1–11.

Soltis, P.S., and Soltis, D.E. (2013). A conifer genome spruces up plant phylogenomics. Genome Biol. **14:** 122.

Sterck, L., Rombauts, S., Jansson, S., Sterky, F., Rouzé, P., and Van de Peer, Y. (2005). EST data suggest that poplar is an ancient polyploid. New Phytol. **167:** 165–170.

Stewart, W.N., and Rothwell, G.W. (1993). Paleobotany and the Evolution of Plants. (Cambridge, UK: Cambridge University Press).

Vajda, V., and McLoughlin, S. (2007). Extinction and recovery patterns of the vegetation across the Cretaceous-Palaeogene boundary - a tool for unravelling, the causes of the end-Permian mass-extinction. Rev. Palaeobot. Palynol. **144:** 99–112.

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. Nat. Rev. Genet. **10:** 725–732.

Van de Peer, Y., Maere, S., and Meyer, A. (2010). 2R or not 2R is not the question anymore. Nat. Rev. Genet. **11:** 166.

Vanneste, K., Baele, G., Maere, S., and Van de Peer, Y. (2014a). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. Genome Res. **24:** 1334–1347.

Vanneste, K., Maere, S., and Van de Peer, Y. (2014b). Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. Philos. Trans. R. Soc. Lond. B Biol. Sci. **369:** 20130353.

Vanneste, K., Van de Peer, Y., and Maere, S. (2013). Inference of genome duplications from age distributions revisited. Mol. Biol. Evol. **30:** 177–190.

Vekemans, D., Proost, S., Vanneste, K., Coenen, H., Viaene, T., Ruelens, P., Maere, S., Van de Peer, Y., and Geuten, K. (2012). Gamma paleohexaploidy in the stem lineage of core eudicots: significance for MADS-box gene and species diversification. Mol. Biol. Evol. **29:** 3793–3806.

Wood, T.E., Takebayashi, N., Barker, M.S., Mayrose, I., Greenspoon, P.B., and Rieseberg, L.H. (2009). The frequency of polyploid speciation in vascular plants. Proc. Natl. Acad. Sci. USA **106:** 13875–13879.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24:** 1586–1591.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18:** 821–829.

Zhong, B., Fong, R., Collins, L.J., McLenachan, P.A., and Penny, D. (2014). Two new fern chloroplasts and decelerated evolution linked to the long generation time in tree ferns. Genome Biol. Evol. **6:** 1166–1173.