A genome for gnetophytes and early evolution of seed plants

Tao Wan[1,2‡], Zhi-Ming Liu[3‡], Ling-Fei Li[1‡], Andrew R. Leitch[4‡], Ilia J. Leitch[5‡], Rolf Lohaus[6,7], Zhong-Jian Liu[8‡], Hai-Ping Xin[2,9], Yan-Bing Gong[10], Yang Liu[1], Wen-Cai Wang[4], Ling-Yun Chen[2,11], Yong Yang[12], Laura J. Kelly[4], Ji Yang[13], Jin-Ling Huang[14,15], Zhen Li[6,7], Ping Liu[1], Li Zhang[1], Hong-Mei Liu[1], Hui Wang[1], Shu-Han Deng[3], Meng Liu[3], Ji Li[3], Lu Ma[4], Yan Liu[3], Yang Lei[3], Wei Xu[3], Ling-Qing Wu[3], Fan Liu[2], Qian Ma[10], Xin-Ran Yu[3], Zhi Jiang[3], Guo-Qiang Zhang[8], Shao-Hua Li[16], Rui-Qiang Li[3], Shou-Zhou Zhang[1], Qing-Feng Wang[2,11*], Yves Van de Peer[6,7,17*], Jin-Bo Zhang[3*] & Xiao-Ming Wang[1*]

[1]Key Laboratory of Southern Subtropical Plant Diversity, Fairy Lake Botanical Garden, Shenzhen & Chinese Academy of Science 518004, PR China. [2]Sino-Africa Joint Research Centre, Chinese Academy of Science, Wuhan 430074, PR China. [3]Novogene Bioinformatics Institute, Beijing 100083, PR China. [4]School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, UK. [5]Jodrell Laboratory, Royal Botanic Gardens, Kew, Surrey TW9 3AB, UK. [6]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent 9052, Belgium. [7]Centre for Plant Systems Biology, VIB, Ghent 9052, Belgium. [8]Shenzhen Key Laboratory for Orchid Conservation and Utilization, National Orchid Conservation Centre of China and Orchid Conservation and Research Centre, Shenzhen 518114, PR China. [9]Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan 430074, PR China. [10]State Key Laboratory of Hybrid Rice, College of Life Sciences, Wuhan University, Wuhan 430072, PR China. [11]Key Laboratory of Aquatic Botany and Watershed Ecology, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan 430074, PR China. [12]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, PR China. [13]Education Key Laboratory for Biodiversity Science and Ecological Engineering, Fudan University, Shanghai 200438, PR China. [14]Institute of Plant Stress Biology, State Key Laboratory of Cotton Biology, Henan University, Kaifeng, 475001, PR China. [15]Department of Biology, East Carolina

University, Greenville, NC27858, USA. [16]Beijing Key Laboratory of Grape Sciences and Enology, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, PR China. [17]Genomics Research Institute, University of Pretoria, Private bag X20, Pretoria 0028, South Africa.

[‡]These authors contributed equally to this work. Correspondence should be addressed to Q.F.W. (qfwang@wbgcas.cn), Y.V.d.P. (yves.vandepeer@psb.vib-ugent.be), J.B.Z. (zhangjinbo@novogene.com) or X.M.W. (719921868@qq.com).

Abstract

Gnetophytes are an enigmatic gymnosperm lineage comprising three genera, *Gnetum*, *Welwitschia* and *Ephedra*, which are morphologically distinct from all other seed plants. Their distinctiveness has triggered much debate as to their origin, evolution, and phylogenetic placement amongst seed plants. To increase our understanding of the evolution of gnetophytes, and their relation to other gymnosperms and seed plants, we report here a high-quality draft genome sequence for *Gnetum montanum* - the first for any gnetophyte. By using a novel genome assembly strategy to deal with high levels of heterozygosity, we assembled > 4 Gb of sequence encoding 27,491 protein-coding genes. Comparative analysis of the *G. montanum* genome with other gymnosperm genomes unveiled some remarkable and distinctive genomic features, such as a diverse assemblage of retrotransposons with evidence for elevated frequencies of elimination rather than accumulation, considerable differences in intron architecture, including both length distribution and proportions of (retro) transposon elements, and distinctive patterns of proliferation of functional protein domains. Furthermore, a few gene families showed *Gnetum*-specific copy number expansions (e.g. CesA) or contractions (e.g. LEA), which could be connected with *Gnetum*'s distinctive morphological innovations associated with their adaptation to warm, mesic environments. Overall, the *G. montanum* genome enables a better resolution of ancestral genomic features within seed plants, and the identification of genomic characters that distinguish *Gnetum* from other gymnosperms.

Introduction

The seed plants today are represented by five distinct lineages: the species-rich angiosperms (flowering plants, *c*. 352,000 species) and four gymnosperm lineages (which together comprise *c*. 1,000 species and encompass cycads, *Ginkgo biloba*, conifers and gnetophytes). It is apparent from their long fossil record (dating back to the Late Devonian *c*. 360 million years ago (Mya)) that considerably greater seed plant diversity existed in the past[1]. Nevertheless, widespread extinctions among many gymnosperm lineages means that today's gymnosperms are only a relic of their former diversity, and this has presented a major challenge for reconstructing evolutionary relationships between the extant lineages[2]. Probably the most controversial outstanding question in plant evolution is the phylogenetic position of gnetophytes[3] (comprising the genera *Gnetum*, *Welwitschia* and *Ephedra*, Fig. 1) in relation to the other seed plant lineages. Apparent morphological similarities with angiosperms, such as vessel-like water conducting cells, double fertilization, and leaf morphologies with reticulate venation, have historically led to the proposition that gnetophytes form a group that is sister to angiosperms (termed the 'Anthophyte hypothesis')[4,5]. That hypothesis has, however, largely been rejected by molecular phylogenetic data and a deeper understanding of the developmental pathways that lead to similar morphological features. Nevertheless, the use of molecular data has also been problematic in inferring the exact phylogenetic position of gnetophytes, with topologies differing depending on the type of sequence data (e.g. plastid versus nuclear genes, nucleotide versus amino acid data) and analytical approach used (e.g. maximum parsimony, maximum likelihood, Bayesian, multispecies coalescent based methods)[6-8]. Consequently, several possible hypotheses have been put forward that place gnetophytes as sister to: (i) Pinaceae ('Gnepine' hypothesis); (ii) cupressophytes ('Gnecup' hypothesis); (iii) all conifers ('Gnetifer' hypothesis); (iv) all other gymnosperms; or (v) all seed plants[9]. Currently, the emerging consensus, based on both older and more recent studies, and recently released data from the 1KP initiative (see https://sites.google.com/a/ualberta.ca/onekp/, and Wickett et al. (8)), indicates that gnetophytes are sister to, or within, the conifers.

So far, the availability of whole genome sequences for gymnosperms has been limited to conifers (specifically to Pinaceae)[10-13] and *G. biloba*[14], with no whole genome assemblies available for the two remaining major seed plant lineages - cycads and gnetophytes. This deficiency, together with the conflicting phylogenetic evidence for relationships among these groups, is impeding our understanding of genome evolution across all seed plants. Here, we present a high-quality draft genome of *G. montanum*, the first for gnetophytes. The availability of this genome, as well as survey sequence data and transcriptome data from other vascular plants (including novel data from gnetophytes *Ephedra* and *Welwitschia*), enables us to compare genomic characters with *G. biloba*, conifers, angiosperms and non-seed plants. Comparisons within gymnosperms, and between gymnosperms and angiosperms, highlight the unique nature of the *Gnetum* genome, providing new insights into patterns of genome divergence across seed plants.

Genome assembly and annotation

The genome of *G. montanum* (2n = 44) is small compared with other gymnosperms (flow cytometry: 4.2 Gb / 1C; k-mer analysis: 4.11 Gb), and is highly heterozygous and rich in repeats (Supplementary Fig. 1a-c, and Supplementary Note 1). To overcome problems caused by repeats and heterozygosity, we generated deep coverage (~302 ×, Supplementary Table 1) Illumina sequence data and applied a novel genome assembly strategy (Supplementary Note 2, Supplementary Fig. 2) to assemble 4.07 Gb of sequence (contig N50 size = 25.02 kb, scaffold N50 size = 475.17 kb, Supplementary Table 2), to which > 99% of genome reads, > 90% ESTs and > 99% of BACs were mapped (Supplementary Fig. 1d, e, Supplementary Table 3 and Note 3).

A total of 27,491 protein-coding genes were predicted from this assembly (Supplementary Table 4 and Note 4), 97% of which were supported by orthology (> 50% coverage of high-scoring segment pair, Supplementary Fig. 3a) with existing protein sequences and/or RNA-seq data from multiple tissues (Supplementary Table 5). A BUSCO analysis to assess the quality of the genome and annotation completeness suggested that 81% of the genes have been recovered (Supplementary Table 6). Unlike conifer genomes, which contain numerous pseudogenes[15] (e.g. 8,328 in *Picea abies*, 13,550 in *Pinus taeda*), many fewer were found in the *G. montanum* genome (3,122, Supplementary Note 5). The read depth distribution across genic regions (Supplementary Fig. 3b) suggested little sequence redundancy caused by heterozygosity (see Supplementary Fig. 3c for further confirmation of gene assembly quality).

Repetitive sequence dynamics

Repetitive sequences have been shown to account for the major component of all gymnosperm genomes that have been sequenced to date[11-14], with diverse and ancient transposable elements (TEs), especially LTR retrotransposons (LTR-RTs), being particularly prevalent. Overall, the repetitive element content of *G. montanum* was also high (85.9%) and dominated by LTR-RTs (especially *gypsy*-like elements), which comprised 77.4% of the genome (Supplementary Table 8 and Supplementary Note 6). The genome assembly of *G. montanum* is likely to be sufficient to represent most of the LTR-RTs, since their length is typically around 25 kb[16], whilst 90% of the scaffolds are larger than 34 kb. Phylogenetic reconstructions of the reverse transcriptase domains of LTR-RTs in *G. montanum* and *P. taeda* revealed that most of the *gypsy*- and *copia*-like elements in *G. montanum* were restricted to just a few clades, representing only a small minority of the diversity encountered in *P. taeda* (Supplementary Fig. 4, Supplementary Note 6).

Comparative analyses of repeats identified by RepeatExplorer using survey sequence data from multiple gnetophytes (*G. montanum*, *G. gnemon*, *W. mirabilis* and *E. altissima*) and *P. taeda* revealed substantial differences in the abundance of the major repeat classes (Supplementary Fig. 5a, Supplementary Table 9 and Supplementary Notes 1, 7). Further, the majority of individual repeat types (repeat clusters in RepeatExplorer) were shown to be species-specific (i.e. containing Illumina reads from just one species, data not shown). The species-specific nature of the repeat profiles probably reflects the long estimated divergence times between species (e.g. the two *Gnetum* species likely diverged between *c.* 25 Mya and 75 Mya)[17,18].

Previously, it was reported from conifers and *G. biloba* that LTR-RTs have accumulated steadily over the last *c*. 25 Mya, especially between 16-24 Mya, a process contributing to their large genome sizes[11,12,14]. This interpretation is consistent with the data here (Supplementary Table 10), which shows that most LTR-RTs in conifers are intact (solo LTR / intact LTR ratio ranged from 0.16:1 to 0.72:1, Supplementary Table 10). It is notable that the solo LTR / intact LTR ratio was substantially higher in *G. montanum* (~1.94:1), which together with its small genome and similar profile of accumulation (Supplementary Fig. 5b), suggests higher frequencies of LTR-RT elimination than amplification compared with *G. biloba* and conifers.

Most angiosperm genomes analysed to date have far fewer ancient repeats and less divergent LTR-RT subsets than conifers and *G. biloba*, presumably due to more efficient elimination and replacement processes operating within these angiosperm genomes[19] (e.g. in *Oryza sativa* the half-life of LTR-RTs is estimated to be less than five million years[20], leading to "genome turnover"[21]). However, an exception to this pattern has been observed in *Amborella trichopoda*. The genome of this species is considered to have retained many features that were likely present in the ancestral angiosperm genome[22]. It is notable that its repeat content[13] and lower abundance of intact LTR-RTs (i.e. solo LTR / intact LTR ratio = 2.43/1.0; Supplementary Table 10) is similar to that observed in *G. montanum*. These observations suggest that neither *A. trichopoda* nor *G. montanum* genomes have experienced recent, extensive (retro) transposon activity, although they continue to eliminate repetitive sequences. Both these species seem to differ from conifers and *G. biloba* with respect to the dynamics of repeat accumulation[11,12,14], and from other angiosperms in terms of the levels of repeat amplification/removal.

Intron morphologies

Although intron size has been positively correlated with genome size across eukaryotes as a whole[23], this trend does not translate well across broad and some narrow taxonomic distances in seed plants (Fig. 2a). Previous studies of *G. biloba*[14] and conifers[11,12] have reported larger introns than angiosperms, probably arising from the long-term, steady amplification of LTR-RTs (Fig. 2b), as also observed here, where LTR-RTs account for 51% and 59% of the large intron sequences in *P. taeda* and *G. biloba*, respectively (Fig. 2a, Supplementary Table 12). The evolution of these large introns may have arisen from similar repeat accumulation processes that are operating across the genome as a whole.

When comparing these observations with introns of *G. montanum*, it is apparent that their introns are substantially smaller (minimum, mean and maximum intron lengths) than those of *P. taeda* and *G. biloba* (Fig. 2a, see also statistics test in Supplementary Table 11). In addition, the repeat composition of *G*. *montanum*'s introns is dominated by both long interspersed nuclear elements (LINEs) as well as LTR-RTs, rather than predominantly LTR-RTs, as in conifers and *G. biloba* (Fig. 2b, Supplementary Table 12). The correlation between smaller intron sizes and smaller genome size in *G. montanum* compared with conifers and *G. biloba* may reflect the repeat dynamic processes operating across its genome as a whole. In contrast, the variable length distributions of introns in angiosperms suggest that the evolution of repeats in their introns do not necessarily reflect the repeat dynamics observed across the rest of their genomes[24]. In the highly dynamic repetitive genome of *Z. mays*, the profile of repeats across the genome[25] and within the whole intron set (Supplementary Fig. 6a) both suggests many recent insertions. However, in *A. trichopoda*, the intron sizes are overall larger, and the genome size smaller than in *Z. mays* (Fig. 2a, b). In addition, an analysis of introns in *A. trichopoda* and *G. montanum* highlighted a closer similarity to each other (in terms of length distributions, repeat composition and divergence) than either species has to conifers and *G. biloba*, despite a 4.8-fold difference in their genome sizes (Fig. 2a, 2b, Supplementary Table 12).

Previous comparisons of orthologous introns have led to the suggestion that the expansion of introns occurred early in the evolutionary history of conifers[12]. Comparisons of orthologous introns (with identical adjacent exons) between *P. taeda* and *G. biloba* showed that introns identified as being long (> 6 kb) in *P. taeda* were also typically long in their orthologues in *G. biloba*, containing, in both cases, abundant LTR-RTs (both *gypsy*- and *copia*-like elements, Fig. 2c). These features were likely to have been present in their most recent common ancestor (MRCA). Using similar approaches to analyse the length and repeat content of 4,348 orthologous introns of *G. montanum* shared with *P. taeda* (Supplementary Note 8) highlighted notable differences. Whilst the length of exons remained similar, a substantial fraction of orthologous genes had longer introns in *P. taeda* (Supplementary Fig. 6b). The introns identified as 'short' in *P. taeda* comprised *c*. 4% repeats, rising to *c*. 56% in 'long' introns, largely through the accumulation of LTR-RTs (especially *copia* elements) (Fig. 2d, Supplementary Table 13). In contrast, introns in *G. montanum* that are orthologous to the 'long' introns of *P. taeda* (36% of introns analysed) showed high proportions of LINEs. As with comparisons of all introns, pairwise comparisons of orthologous introns in *G. montanum* and *A. trichopoda* again showed some similarities in their introns, with both species having abundant LINEs (Fig. 2e). Collectively, these data reveal a different repeat dynamic within introns of *G. montanum* compared with the other gymnosperms.

('Lack of') Whole genome duplication (WGD)

All angiosperms are reported to have undergone at least one round of ancient WGD, and in many lineages WGDs are recurrent and ongoing[26]. In addition, a WGD event has been proposed at the base of all seed plants *c*. 341 Mya (= *zeta* WGD[27]), although the underlying evidence for these two ancient WGD events has been recently questioned[28]. In gymnosperms, WGDs have been reported for conifers, *G. biloba* and cycad (a likely shared WGD)[14,29,30]. Although recent polyploidy seems common in extant *Ephedra*[31], evidence for ancient WGDs in gnetophytes is missing (Supplementary Note 9 and Supplementary Fig. 7), except for a WGD in *Welwitchia* which likely occurred after the divergence of its lineage from that leading to *Ephedra* (Supplementary Fig. 7)[29]. If indeed the ancient zeta WGD is shared by all seed plants, the absence of evidence for this event in gnetophytes is best explained by their faster rates of gene evolution compared with other gymnosperms[32,33], erasing all evidence of this more than 300 million year old event (Supplementary Note 9 and Supplementary Fig. 7).

Organization of functional protein domains

To characterize the patterns of functional diversification in gene domains across land plants, we used principal component analysis (PCA) to analyse the number of pfam domains (conserved protein domains) in multiple species (Supplementary Note 10, Supplementary Table 13). Our approach showed that angiosperms formed a discrete cluster that was separate from the gymnosperms (Fig. 3a), with *G. montanum* being an outlier. Indeed, heatmaps compiled from the pfam data that contributed most (top 10%) to PCA1 and PCA2 showed that *G. montanum* formed a clade with the lycophyte *S. moellendorffii* and the moss *Physcomitrella patens* (Fig. 3b), whilst the non-gnetophyte gymnosperms formed a separate clade (Fig. 3b).

Given the distinct distributions of *G. montanum*, non-gnetophyte gymnosperms and angiosperms in the PCA analysis, the data suggest that significant functional diversification of the conserved protein domains has occurred since these major lineages split. It may be surprising given the long divergence times (*c*. 300 Mya)[2], that *G. biloba* and conifers retain similar conserved domain organizations (with similar eigenvector values). This could reflect their relatively low substitution rates (on average 7 × lower) compared with angiosperms[33].

An analysis of the pfam domain expansions that contributed most to the PCA1 and PCA2 distributions amongst angiosperms (except *A. trichopoda*). included genes associated with flower and organ development (Supplementary Table 15). In contrast, non-gnetophyte gymnosperms showed large-scale specific expansions of pfam domains in genes associated with defence and secondary metabolism, as previously suggested (Supplementary Table 16)[10,11]. The clustering of *G. montanum* with non-seed plants in the heatmap (Fig. 3b) was a surprise, and may indicate the approach has identified proteins that have diverged very little since the MRCA of seed plants. Nevertheless, such an explanation is at odds with the hypothesis that the genes of gnetophytes have diverged rapidly, given their comparatively high substitution rate compared with other gymnosperms[33].

Growth form (shrubs and lianas) and leaf morphology

Gnetophytes differ from other extant gymnosperms in growth form, with the unusual and distinct form of *Welwitschia*, the shrub habit of *Ephedra* and the shrub and liana habit and specialized leaf morphologies of *Gnetum*[34]. Cellulose synthase (*Ces*A) and cellulose synthase-like (*Csl*) genes are considered to play a role in influencing the biomechanical properties of the cell[35], hence potentially the distinctive growth forms of gnetophytes are associated with the divergence of these genes. To explore this hypothesis, *Ces*A and *Csl* family members were examined in *G. montanum* and compared with those in other seed plants. The total number of *Ces*A and *Csl* family members ranged about 3-fold amongst the seed plants analysed (*P. abies*, *P. taeda*, *A. trichopoda*, *A. thaliana* and *O. sativa*). However, only *G. montanum* showed a large expansion of the *Csl*B/H gene subfamily (to 20 genes, Supplementary Table 17), involving tandem duplications (Supplementary Fig. 9), and accounting for two-thirds of its total *Csl* gene repertoire. Furthermore, transcriptome analysis showed that these *Csl*B/H genes were differentially expressed in leaves, stems and roots of *G. montanum*, supporting an association with distinct growth forms and leaf morphologies (Supplementary Fig. 9). In contrast, all other species analysed, including *Welwitschia* and *Ephedra*, were seen to have only 1-6 *Csl*B/H genes (at least based on transcriptome analysis) (Supplementary Note 11, Supplementary Table 16, Supplementary Fig. 8).

Another gene family associated with leaf morphology and development is the *WOX* (*WUSCHEL-related homeobox*) family[36]. Recent studies have shown that the conserved family members WOX3 and WOX4, which play a role in leaf development, show diffuse *WOX3* expression at the leaf bases of *Arabidopsis* and

Gnetum, with such patterns being associated with the distinctive reticulate venation observed in their leaves[37]. Two unusual paralogues, GgWOXX and GgWOXY, were previously reported to occur only in gnetophytes[37], and this is confirmed here in phylogenetic reconstructions of gene family members (Supplementary Note 12, Supplementary Fig. 10). These paralogues are unlikely to have arisen by *Gnetum*-specific gene amplifications, as this would group them with other *Gnetum* paralogues. Alternatively, these genes may correspond to ancestral seed plant sequences that have been lost in other plant lineages. Potentially the different patterns of gene loss, retention and amplification compared with other gymnosperms may be associated with their distinctive growth forms.

Vessels

The presence of vessel-like water-conducting cells, morphologically distinct from tracheids, is another feature that sets gnetophytes apart from other gymnosperms. However, there has been long-standing debate as to whether gnetophyte "vessels" are homologous to the "vessels" of angiosperms. In angiosperms, VASCULAR-RELATED NAC-DOMAIN (VND) proteins *VND1*-7 are members of the *NAC* domain class of transcription factors, *VND7* being a master regulator of vessel formation in *Arabidopsis thaliana*[38], and *VND1-6* being upstream regulators of *VND7*[39]. Although five *NAC* domain genes were identified in the genome of *G. montanum*, no orthologues of *VND*7 or *VND1-3* in the sister clade were identified, consistent with previous analyses of other gymnosperms[12], and suggesting that these proteins are restricted to angiosperms (Supplementary Fig. 11). Nevertheless, *Gnetum* does share the *VND4-6* clade with angiosperms and other gymnosperms. Furthermore, *A. trichopoda,* which lacks angiosperm vessels, also lacks orthologues of *VND1-3*, but it does have *VND7* (Supplementary Fig. 11), indicating that the ability to form vessels may have occurred after angiosperms diverged. Taken together, these data suggest a greater dependency of vessel development on *VND1-3* than is apparent from experiments on *A. thaliana*. The most parsimonious explanation of our data is that angiosperm vessel formation requires genes from the *VND7* clade (and potentially its sister clade *VND1-3*), and that gymnosperms, including gnetophytes, which lack sequences from both these clades cannot form structures that are homologous to angiosperm vessels. Such an interpretation supports Carlquist's[40] morphological interpretations of vessels. It is therefore most likely that different molecular mechanisms underpin the origin and development of vessels in *Gnetum* and angiosperms. Indeed, these new molecular data support the hypothesis based on morphological studies that *Gnetum* vessels are actually more closely related to conifer tracheids than angiosperm vessels and that vessels in the two groups are convergent characters[40].

Water stress

Extant species of *Gnetum* are unusual amongst gymnosperms in being restricted to warm, mesic habitats[41], this contrasts to conifers that are adapted to cold and water-stressed environments. An analysis of genes involved in water and cold stress revealed some substantial differences between conifers and *Gnetum*. The Late Embryogenesis Abundant protein (LEA) gene family encodes crucial proteins that are involved in protecting plants from desiccation or osmotic stresses associated with low temperature[42,43]. An analysis of LEA family members suggests that some members have been reduced in number in *Gnetum* or expanded in conifers (e.g. LEA-3), or lost completely in *Gnetum* (i.e. LEA-4, 5, 6). In addition, dehydrins, which play a role in the response to cold/drought[44], had only two members in *G. montanum*, compared with 38 in *P. abies*, 28 in *P. taeda* and 3-15 in angiosperms (Supplementary Table 19). Further analysis of the *G. montanum* genome also revealed relatively few gene family members of the AP2 domain containing protein families, which are involved in the cold stress response[45,46], and GPX and GST families, involved in the oxidant stress response[47,48]. Taken together, these data appear consistent with the hypothesis that the ecological shift to a warm, wet forest habitat is associated with a relaxation of selection pressure on genes associated with water stress and low temperature.

Conclusion

Here, we have described the assembly, annotation, and comparative analysis of the first gnetophyte genome, namely that of *G. montanum*. Its genome is particularly enigmatic given a phylogenetic position within or sister to conifers. It also carries genomic peculiarities that may reflect its morphological and ecological uniqueness amongst gymnosperms. Comparisons of these genome features with the genomes of conifers and *G. biloba* provide opportunities to predict the nature and direction of genomic change accompanying the evolution of the lineage leading to *Gnetum* (Fig. 4). Assuming that gnetophytes do indeed form a clade that is sister to, or within, the conifers, the following genomic features can be predicted to have been present in the MRCA of the gymnosperms, as observed in *G. biloba*[14] and conifers[11,12]: (1) A large genome size (1C > 10 Gb) comprised predominantly of a heterogeneous set of large numbers of LTR-RTs associated with low levels of repeat deletion[14]; (2) Long introns predominantly shaped by insertions of LTR-RTs (*gypsy* and *copia* elements); (3) Pfam domains that show a profile distinct from angiosperms; If this is so, and assuming a common ancestry of gnetophytes and conifers, these genomic characters, or their signatures, have subsequently been lost or diverged considerably in the lineage leading to *Gnetum*. This most likely involved the following genomic processes: (1) Genome downsizing, leading to the relatively (for a gymnosperm) small genomes of *Gnetum* species (1C= 2.25-4.11 Gb). This is supported by the high ratio of solo LTR / intact LTR-RTs observed in the genome of *Gnetum* compared with conifers, and is indicative of the activity of recombination-based processes, which can eliminate DNA from the genome. Similar processes leading to genome downsizing have also been reported in many angiosperms, resulting in small genomes despite the occurrence of multiple rounds of polyploidy detected in many lineages[49]; (2) Reduction in the size of introns in *G. montanum* and a replacement of many of the LTR-RTs repeats with LINEs to give rise to introns that are more similar to those of, for instance, *A. trichopoda* than to other gymnosperms; (3) Elevated rates of sequence divergence causing the erosion of a hypothesised shared seed-plant WGD event and leading to a pattern of Pfam domains, which is distinct from the remaining gymnosperms; (4) Expansion and contraction of specific gene families associated with adaptation to new ecologies.

Methods summary

The sequenced *G. montanum* is a single mature female individual growing naturally in Fairy Lake Botanical Garden, Shenzhen, China. Genome sequences were generated using an Illumina platform and assembled with a novel hierarchical assembly strategy. Gene annotations were determined by integrating results from both *de novo* prediction approaches and alignment-based methods based on orthology and transcriptomic data. RNA-seq was performed using an Illumina platform. All methods and bioinformatic analyses are detailed in the Supplementary Information.

Data availability

The *G. montanum* genome project has been deposited at the NCBI under the BioProject number PRJNA339497. The whole genome sequencing data were deposited in the Sequence Read Archive (SRA) database under the accession number SRX2052734, SRX2098865, SRX2099144, SRX2114825, SRX2114827, SRX2134147, SRX2134160, SRX2134177, SRX2134180, SRX2134596, and SRX2134624. And the *G. montanum* assemblies, gene sequences, and annotation data are also available at the DRYAD website. The data or related program scripts that support the findings of this study are available from the corresponding author upon request.

Acknowledgements

Genome sequencing, assembly and annotation were conducted by the Novogene Bioinformatics Institute, Beijing, China, mutual contracts were No. NHT140016 and NVT140016004. This work was supported by funding from the Scientific Project of Shenzhen Urban Administration (201519) and a Major Technical Research Project of the Innovation of Science and Technology Commission of Shenzhen (JSGG20140515164852417) Additional funding was provided in particular by the Scientific Research Program of Sino-Africa Joint Research Center (SAJL201607). We thank X.Q. Wang, G.W. Hu, Z.D. Chen and Y.H. Guo for comments on gnetophyte phylogenetic relationships and ecological issues; H. Wu and X.P. Ning for discussion of related organ development; K.K. Wan and S. Sun for additional help on the analysis of repeats. We also thank X.Y for support of funding coordination. YVdP acknowledges the Multidisciplinary Research Partnership 'Bioinformatics: from nucleotides to networks' Project (no. 01MR0310W) of Ghent University, and funding from the European Union Seventh Framework Programme (FP7/2007-2013) under European Research Council Advanced Grant Agreement 322739-DOUBLEUP.

Author contributions

T.W. and X.M.W. conceived and initiated the study, managing the gnetophytes (*Gnetum*, *Welwitschia*, *Ephedra*) genome sequencing project. T.W. designed the major scientific objectives and led the manuscript preparation together with A.R.L., I.J.L., J.B.Z, L.J.K and Y.V.d.P. The collaboration between groups was close in all aspects of the project. T.W., Z.M.L., L.F.L., A.R.L., I.J.L. and Z.J.L. are joint first authors, H.P.X., Y.B.G., Y.L., L.Y.C. and W.C.W. are joint second authors. Z.M.L., J.B.Z., J.L., Y.L. performed the genome assembly and annotation; H.P.X., L.F.L., L.Y.C., L.M., X.R.Y. contributed to the RNA-seq and corresponding analysis. A.R.L., I.J.L., and W.C.W. coordinated the *RepeatExplorer* analysis in gnetophytes and contributed to the design of the analysis for investigating the dynamics of genome evolution. Z.M.L., J.B.Z., L.F.L., F.L., H.M.L., T.W., A.R.L., I.J.L., W.X. and Y.L. participated in the analyses of LTR-RTs and comparisons of introns. R.L., T.W., Y.V.d.P., Z.L., Z.J.L. and Z.M.L. were involved in the WGD determination; M.L., L.F.L., J.B.Z., J.Y., T.W., L.Z., Y.B.G. and Y.H.D. conducted PCA analysis of pfam domains. J.B.Z., T.W., J.L., L.F.L., L.J.K., Y.L. and Z.M.L. performed the analysis investigating the divergence of gene families. J.L.H., P.L., Q.M., Y.L. and G.Q.Z. contributed to the analysis of pseudogenes; Q.F.W., S.H.L. and S.Z.Z. helped with the collecting of *Welwitschia* and *Ephedra*. Y.Y. provided experimental information on the taxonomic identity of the species used for genome sequencing and collated the distribution records of gnetophytes.

Additional Information

Information on reprints and permissions is available at http://www.nature.com/reprints. Correspondence and requests for materials should be addressed to Q.F.W. (qfwang@wbgcas.cn) or Y.V.d.P. (yves.vandepeer@psb.vib-ugent.be) or J.B.Z. (zhangjinbo@novogene.com) or X.M.W. (719921868@qq.com).

Competing interests

The authors declare no competing financial interests.

References

1 Rothwell, G. W. & Scheckler, S. E. *Biology of Ancestral Gymnosperms. Origin and Evolution of Gymnosperms* (Columbia Uni. Press, 1988).

2 Lu, Y., Ran, J. H., Guo, D. M., Yang, Z. Y. & Wang, X. Q. Phylogeny and divergence times of gymnosperms inferred from single-copy nuclear genes. *PLoS One* **9**, e107679 (2014).

3 Doyle, J. A. Molecular and Fossil Evidence on the Origin of Angiosperms. *Annu. Rev. Earth Planet. Sci.* **40**, 301-326 (2012).

4 Doyle, J. A. & Donoghue, M. J. Seed plant phylogeny and the origin of angiosperms: An experimental cladistic approach. *Bot. Rev.* **52**, 321-431 (1986).

5 Crane, P. R. Phylogenetic Analysis of Seed Plants and the Origin of Angiosperms. *Annals of the Missouri Botanical Garden* **72**, 716-793 (1985).

6 Mathews, S. Phylogenetic relationships among seed plants: Persistent questions and the limits of molecular data. *Am. J. Bot.* **96**, 228-236 (2009).

7 Wang, X. Q. & Ran, J. H. Evolution and biogeography of gymnosperms. *Mol. Phylogenet. Evol.* **75**, 24-40 (2014).

8 Wickett, N. J. *et al.* Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl Acad. Sci. USA* **111**, E4859-4868 (2014).

9 Li, Z. *et al.* Single-Copy genes as molecular markers for phylogenomic studies in seed plants. *Genome Biol. Evol.* **9**, 1130-1147 (2017).

10 Warren, R. L. *et al.* Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. *Plant J.* **83**, 189-212 (2015).

11 Neale, D. B. *et al.* Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol.* **15**, R59 (2014).

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

13 Stevens, K. A. *et al.* Sequence of the sugar pine megagenome. *Genetics* **204**, 1613-1626 (2016).

14 Guan, R. *et al.* Draft genome of the living fossil *Ginkgo biloba*. *GigaScience* **5**, 49 (2016).

15 Garcia-Gil, M. R. Evolutionary aspects of functional and pseudogene members of the phytochrome gene family in Scots pine. *J. Mol. Evol.* **67**, 222-232 (2008).

16 Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973-982 (2007).

17 Won, H. & Renner, S. S. Dating dispersal and radiation in the gymnosperm Gnetum (Gnetales)--clock calibration when outgroup relationships are uncertain. *Syst. Biol.* **55**, 610-622 (2006).

18 Hou, C., Humphreys, A. M., Thureborn, O. & Rydin, C. New insights into the evolutionary history of *Gnetum* (Gnetales). *Taxon* **64**, 239-253 (2015).

19 Kejnovsky, E., Leitch, I. J. & Leitch, A. R. Contrasting evolutionary dynamics between angiosperm and mammalian genomes. *Trends Ecol. Evol.* **24**, 572-582 (2009).

20 Ma, J., Devos, K. M. & Bennetzen, J. L. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.* **14**, 860-869 (2004).

21 Lim, K. Y. *et al.* Sequence of events leading to near-complete genome turnover in allopolyploid *Nicotiana* within five million years. *New Phytol.* **175**, 756-763 (2007).

22 Amborella Genome Project. The Amborella genome and the evolution of flowering plants. *Science* **342**, 1241089 (2013).

23 Vinogradov, A. E. Intron-genome size relationship on a large evolutionary scale. *J. Mol. Evol.* **49**, 376-384 (1999).

24 Wendel, J. F. *et al.* Intron size and genome size in plants. *Mol. Biol. Evol.* **19**, 2346-2352 (2002).

25 Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112-1115 (2009).

26 Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* (2017).

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

28 Ruprecht, C. *et al.* Revisiting ancestral polyploidy in plants. *Sci. Adv.* **3**, e1603195 (2017).

29 Li, Z. *et al.* Early genome duplications in conifers and other seed plants. *Sci. Adv.* **1**, e1501084 (2015).

30 Roodt, D. *et al.* Evidence for an ancient whole genome duplication in the cycad lineage. *PLoS One* **12**, e0184454 (2017).

31 Wu, H. *et al.* A high frequency of allopolyploid speciation in the gymnospermous genus *Ephedra* and its possible association with some biological and ecological features. *Mol. Ecol.* **25**, 1192-1210 (2016).

32 Hajibabaei, M., Xia, J. & Drouin, G. Seed plant phylogeny: gnetophytes are derived conifers and a sister group to Pinaceae. *Mol. Phylogenet. Evol.* **40**, 208-217 (2006).

33 De La Torre, A. R., Li, Z., Van de Peer, Y. & Ingvarsson, P. K. Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. *Mol. Biol. Evol.* **34**, 1363-1377 (2017).

34 Frohlich, M. W. & Chase, M. W. After a dozen years of progress the origin of angiosperms is still a great mystery. *Nature* **450**, 1184-1189 (2007).

35 Popper, Z. A. *et al.* Evolution and diversity of plant cell walls: from algae to flowering plants. *Annu. Rev. Plant Biol.* **62**, 567-590 (2011).

36 Nakata, M. *et al.* Roles of the middle domain-specific *WUSCHEL-RELATED HOMEOBOX* genes in early development of leaves in *Arabidopsis*. *Plant Cell* **24**, 519-535 (2012).

37 Nardmann, J. & Werr, W. Symplesiomorphies in the *WUSCHEL* clade suggest

that the last common ancestor of seed plants contained at least four independent stem cell niches. *New Phytol.* **199**, 1081-1092 (2013).

38 Yamaguchi, M. *et al.* VASCULAR-RELATED NAC-DOMAIN7 directly regulates the expression of a broad range of genes for xylem vessel formation. *Plant J.* **66**, 579-590 (2011).

39 Endo, H. *et al.* Multiple classes of transcription factors regulate the expression of *VASCULAR-RELATED NAC-DOMAIN7*, a master switch of xylem vessel differentiation. *Plant Cell Physiol.* **56**, 242-254 (2015).

40 Carlquist, S. Wood, bark and stem anatomy of New World species of *Gnetum*. *Bot. J. Linn. Soc.* **120**, 1-19 (1996).

41 Ickert-Bond, S. M. & Renner, S. S. The Gnetales: Recent insights on their morphology, reproductive biology, chromosome numbers, biogeography, and divergence times. *J. Syst. Evol.* **54**, 1-16 (2016).

42 Hundertmark, M. & Hincha, D. K. LEA (Late Embryogenesis Abundant) proteins and their encoding genes in *Arabidopsis thaliana*. *BMC Genomics* **9**, 1-22 (2008).

43 Gao, J. & Lan, T. Functional characterization of the late embryogenesis abundant (LEA) protein gene family from Pinus tabuliformis (Pinaceae) in Escherichia coli. *Sci. Rep.* **6**, 19467 (2016).

44 Richard, S., Morency, M. J., Drevet, C., Jouanin, L. & Seguin, A. Isolation and characterization of a dehydrin gene from white spruce induced upon wounding, drought and cold stresses. *Plant Mol. Biol.* **43**, 1-10 (2000).

45 Chinnusamy, V., Zhu, J. & Zhu, J. K. Cold stress regulation of gene expression in plants. *Trends Plant Sci.* **12**, 444-451 (2007).

46 Du, C. *et al.* Dynamic transcriptome analysis reveals AP2/ERF transcription factors responsible for cold stress in rapeseed (*Brassica napus* L.). *Mol. Genet. Genomics* **291**, 1053-1067 (2016).

47 Roxas, V. P., Smith Jr, R. K., Allen, E. R. & Allen, R. D. Overexpression of glutathione S-transferase/glutathioneperoxidase enhances the growth of transgenic tobacco seedlings during stress. *Nat. Biotechnol.* **15**, 988 (1997).

48 Zhao, J. *et al.* Global transcriptional profiling of a cold-tolerant rice variety under moderate cold stress reveals different cold stress response mechanisms. *Physiol. Plant* **154**, 381-394 (2015).

49 Wendel, J. F. The wondrous cycles of polyploidy in plants. *Am. J. Bot.* **102**, 1753-1756 (2015).

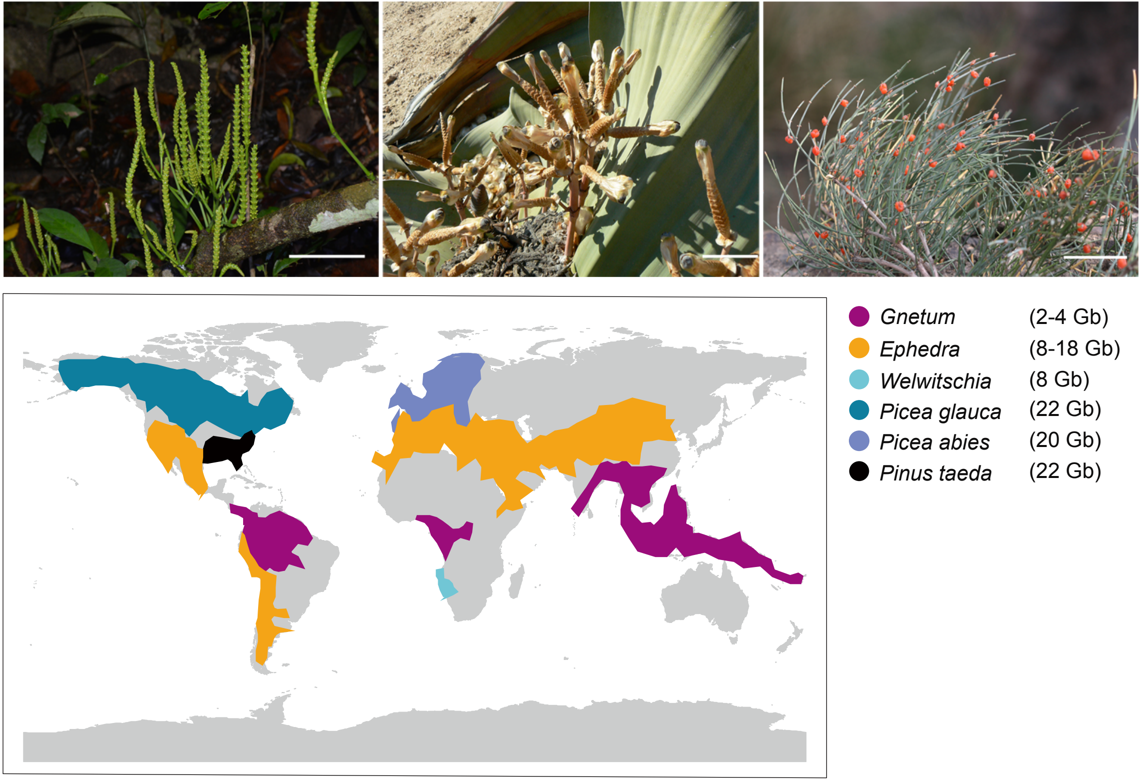
Figure Legends

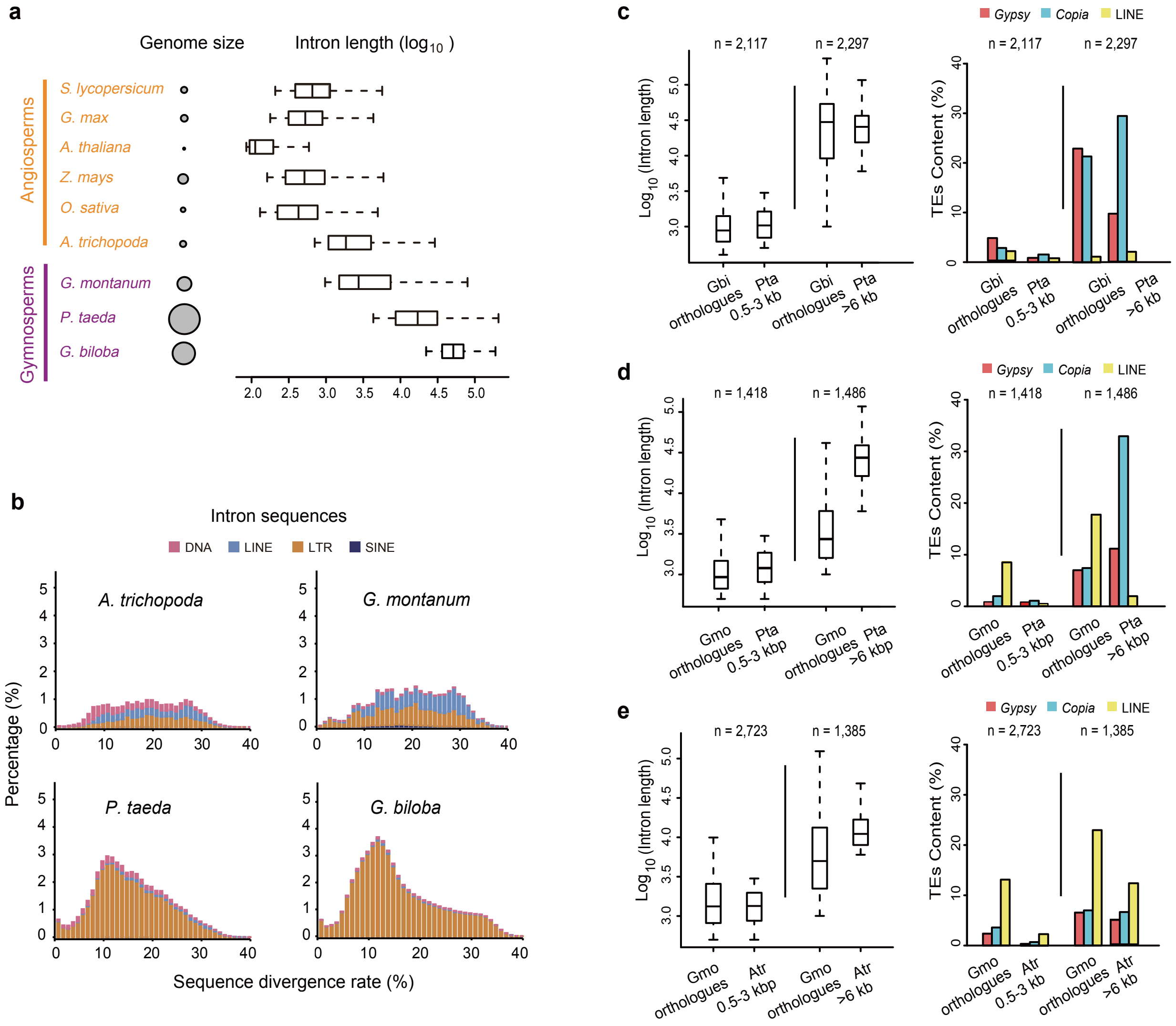
Fig. 1 | Morphological variation and geographical distribution of gnetophytes and some other gymnosperms. Top row from left to right, female cones of *Gnetum montanum*, male cones of *Welwitschia mirabilis* and female cones of *Ephedra equisetina* (Bar = 5 cm). Below, pantropical distribution of the three gnetophyte genera, compared with three conifer species that are most abundant at higher latitudes and altitudes. The range of genomes sizes (1C-values) found in the three genera comprising gnetophytes and the three conifer species are also shown (data taken from http://data.kew.org/cvalues/ and unpublished data).

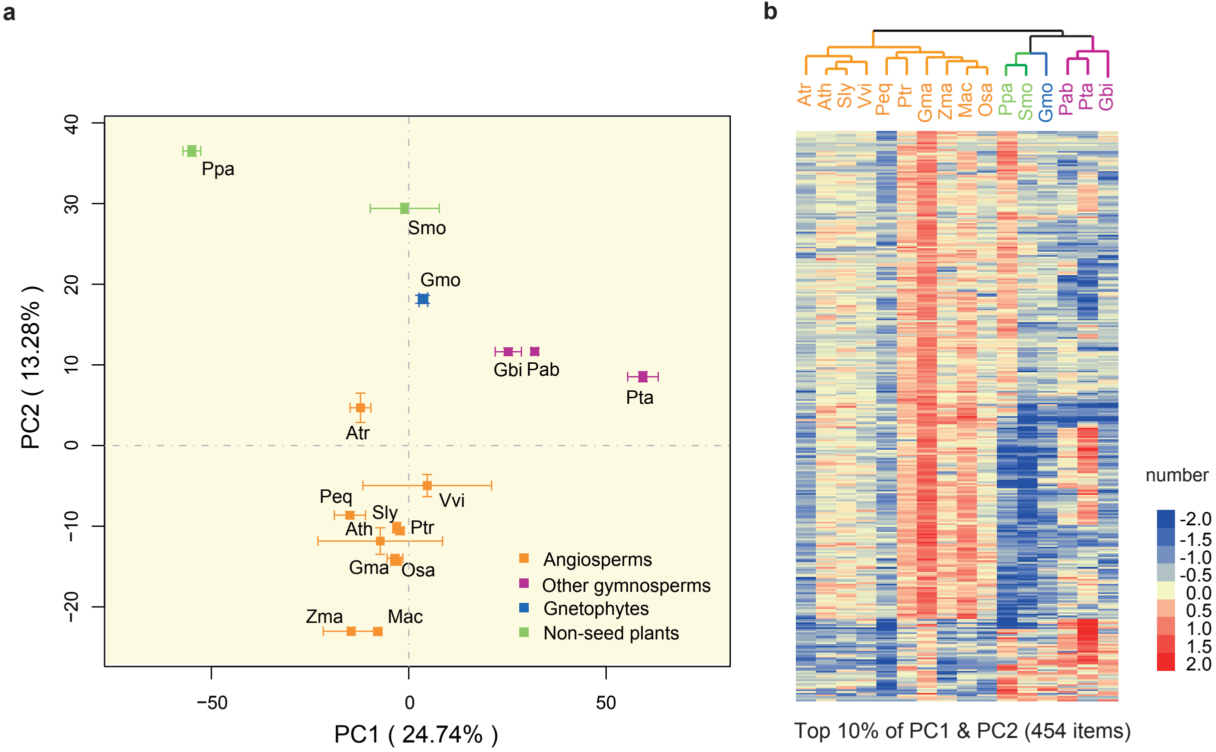
Fig. 2 | Comparative analysis of seed plant intron morphologies. (**a**) Intron length distributions and genome sizes (1C-values, depicted by the relative circle size) are shown for nine representative seed plants. (**b**) Distribution of sequence divergence for four types of transposable elements (TEs) in introns of *A. trichopoda*, *G. montanum*, *P. taeda*, and *G. biloba*. The data show that TEs in *G. montanum* and *A. trichopoda* are more diverse than in *P. taeda* and *G. biloba.* The latter two species also show a peak at around 10% sequence divergence probably reflecting a pulse of LTR-RT expansions. (**c**), (**d**) and **(e)**, Comparison of orthologous introns between *P. taeda* (Pta) vs. *G. biloba* (Gbi) **(c)**, *P. taeda* (Pta) vs. *G. montanum* (Gmo) **(d)** *G. montanum* (Gmo) vs. *A. trichopoda* (Atr) **(e)**. Two orthologous intron sets that differed more than two-fold in length were examined, i.e. 'short' introns = 0.5-3 kb and 'long' introns ≥ 6 kb. Orthologous introns that were 'long' in one species were also found to be 'long' in the other species of the pair. Analysis of the TEs in orthologous introns showed the 'long' introns of *G. montanum* and *A. trichopoda* carried a high proportion of LINEs, contributing to intron expansion. In contrast, *gypsy* and *copia* LTR-RT elements contributed most to intron expansion in *P. taeda* and *G. biloba*.

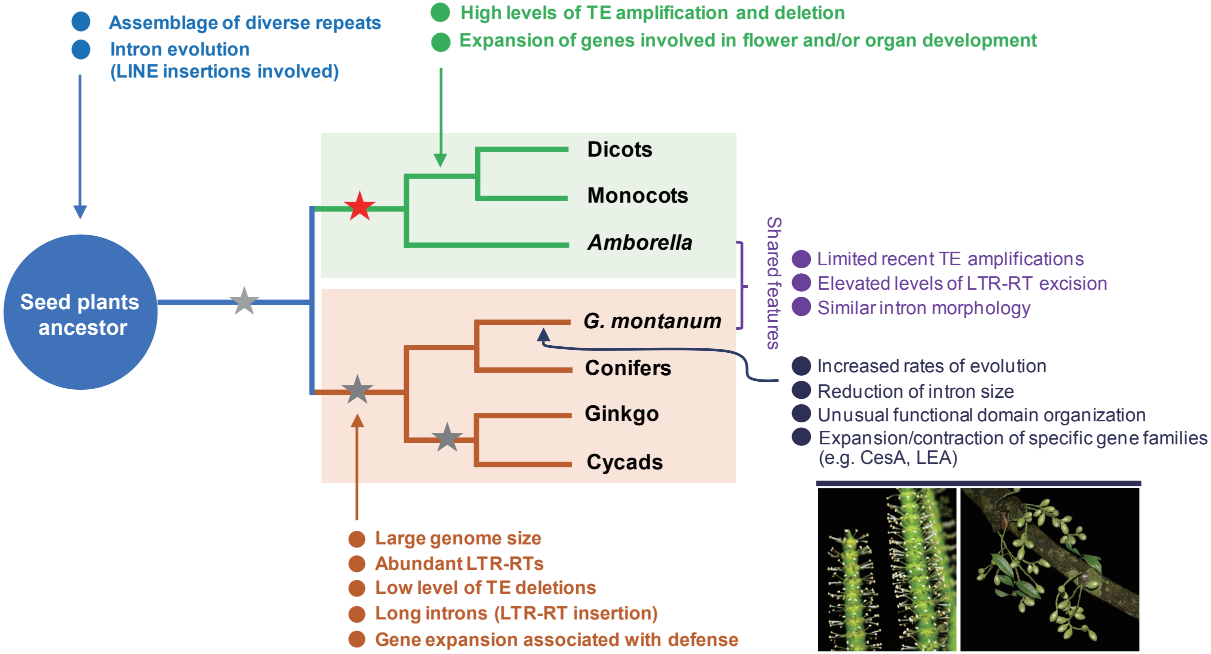
Fig. 3 | Genome-wide analysis to show the contrasting diversification of functional protein domains across land plants. (**a**) PCA analysis of the occurrence and number of pfam domains in multiple orthologous genes across land plants. Plotting PC1 against PC2 reveals that monocots and eudicots cluster together, as do conifers with *G. biloba*, whilst the remaining species are separate from these clusters. (**b**) Heatmaps reveal the ancestral coding repertories shared by *S. moellendorffii* and *G. montanum*. Different patterns of expansion and contraction of the pfam domains are seen for other gymnosperms and angiosperms (see **Supplementary Table 7** for species name list and corresponding abbreviations).

Fig. 4 | Prediction of patterns of genome divergence across seed plants. The origin and evolution of distinctive genomic features observed in *G. montanum* genome are inferred, assuming a phylogenetic placement of gnetophytes as sister to, or within conifers. The predicted features shared by respective lineages are marked by coloured circles. Likely whole genome duplication (WGD) events (red stars) and a putative WGD event (grey star) are shown.


● Gnetum (2-4 Gb)
● Ephedra (8-18 Gb)
● Welwitschia (8 Gb)
● Picea glauca (22 Gb)
● Picea abies (20 Gb)
● Pinus taeda (22 Gb)


a
Genome size
Intron length (log_{10})
Angiosperms
S. lycopersicum
G. max
A. thaliana
Z. mays
O. sativa
A. trichopoda
Gymnosperms
G. montanum
P. taeda
G. biloba
2.0 2.5 3.0 3.5 4.0 4.5 5.0
b
Intron sequences
DNA LINE LTR SINE
A. trichopoda
G. montanum
P. taeda
G. biloba
Percentage (%)
Sequence divergence rate (%)
0 10 20 30 40
0 1 2 3 4 5
c
Gypsy Copia LINE
n = 2,117 n = 2,297
Log_{10} (Intron length)
TEs Content (%)
Gbi orthologues
Pta 0.5-3 kb
Gbi orthologues
Pta >6 kb
d
n = 1,418 n = 1,486
Gmo orthologues
Pta 0.5-3 kbp
Gmo orthologues
Pta >6 kbp
e
n = 2,723 n = 1,385
Gmo orthologues
Atr 0.5-3 kbp
Gmo orthologues
Atr >6 kb
Atr 0.5-3 kb
3.0 3.5 4.0 4.5 5.0
0 10 20 30 40


a
b
PC2 (13.28%)
PC1 (24.74%)
40
30
20
10
0
−10
−20
−50
0
50
Ppa
Smo
Gmo
Gbi
Pab
Pta
Atr
Peq
Vvi
Sly
Ath
Ptr
Gma
Osa
Zma
Mac
Angiosperms
Other gymnosperms
Gnetophytes
Non-seed plants
Atr
Ath
Sly
Vvi
Peq
Ptr
Gma
Zma
Mac
Osa
Ppa
Smo
Gmo
Pab
Pta
Gbi
number
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
Top 10% of PC1 & PC2 (454 items)



- Assemblage of diverse repeats
- Intron evolution (LINE insertions involved)

- High levels of TE amplification and deletion
- Expansion of genes involved in flower and/or organ development

- Limited recent TE amplifications
- Elevated levels of LTR-RT excision
- Similar intron morphology

- Increased rates of evolution
- Reduction of intron size
- Unusual functional domain organization
- Expansion/contraction of specific gene families (e.g. CesA, LEA)

- Large genome size
- Abundant LTR-RTs
- Low level of TE deletions
- Long introns (LTR-RT insertion)
- Gene expansion associated with defense