communications biology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-025-07749-x

Comparative analysis using a chromosome-scale genome assembly for *Funaria hygrometrica* suggests greater collinearity in mosses than in seed plants

Alexander Kirbis[1,2], Nasim Rahmatpour[3], Shanshan Dong[4], Jin Yu[5], Lucas Waser[1,2], Huaxing Huang[1,2], Nico van Gessel[6], Manuel Waller[1,2], Ralf Reski[7], Daniel Lang[8], Stefan A. Rensing[9,10], Eva M. Temsch[11], Jill L. Wegrzyn[3], Bernard Goffinet[3], Yang Liu[4,5] & Péter Szövényi[1,2]

Mosses, the largest lineage of seed-free plants, have smaller and less variable genome sizes than flowering plants. Nevertheless, whether this difference results from divergent genome dynamics is poorly known. Here, we use newly generated chromosome-scale genome assemblies for *Funaria hygrometrica* and comparative analysis with other moss and seed plant genomes to investigate moss genome dynamics. Although some aspects of moss genome dynamics are seed plant-like, such as the mechanism of genome size change and de novo gain/loss of genes, moss genomes retain higher synteny, and collinearity over evolutionary time than seed plant genomes. Furthermore, transposable elements and genes are more evenly distributed along chromosomes in mosses than in seed plants, a feature shared with other sequenced seed-free plant genomes. Overall, our findings support the hypothesis that large-scale genome structure and dynamics of mosses and seed plants differ. In particular, our data suggest a lower rate of gene order reshuffling along chromosomes in mosses compared to seed plants. We speculate that such lower rate of structural genomic variation and unique chromosome structure in mosses may contribute to their relatively smaller and less variable genome sizes.

The number of pseudomolecule-scale genome assemblies of seed plants has rapidly increased in the last 20 years revealing their conserved and divergent architectural features[1–5]. In addition, comparative analyses of deep and shallow divergent seed plant genomes provided detailed insights into genome evolution and dynamics over longer and shorter timescales[6–20]. By contrast, the structure and dynamics of seed-free plant genomes are little understood[2,5,21]. Comparison of the few available high-quality genomes suggests that overall genomic architectures of seed-free and seed-plant genomes differ[5,22–26]. In particular, genomes of seed-free plants are smaller, less variable in size (except in the Monilophytes), and the spatial distribution of genes and transposable elements (TEs) along the chromosomes appears to be more even than in seed plants[5,27]. Such pattern may be a consequence of an overall slower rate of genome evolution in seed-free versus seed plants, potentially affecting the spatial distribution of transposable elements (TE) as well as genome size evolution[5]. Nevertheless, relevant aspects of genome evolution in seed-free plants, such as the mechanism of genome size change, the contribution of gene gain/loss and duplication to gene content variation, and their overall effect on genomic collinearity are poorly understood.

[1]Department of Systematic and Evoutionary Botany, University of Zurich, Zurich, Switzerland. [2]Zurich-Basel Plant Science Center, LFW, Universitätsstrasse 2, Zürich, Switzerland. [3]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, USA. [4]Key Laboratory of Southern Subtropical Plant Diversity, Fairy Lake Botanical Garden, Shenzhen & Chinese Academy of Sciences, Shenzhen, Guangdong, China. [5]Key Laboratory of Genomics, Ministry of Agriculture, BGI Research, Shenzhen, Shenzhen, China. [6]Plant Biotechnology, Faculty of Biology, University of Freiburg, Freiburg, Germany. [7]Plant Biotechnology, Faculty of Biology, and Signalling Research Centres BIOSS and CIBSS, University of Freiburg, Freiburg, Germany. [8]Bundeswehr Institute of Microbiology, Microbial Genomics and Bioforensics, Munich, Germany. [9]Plant Cell Biology, Department of Biology, University of Marburg, Marburg, Germany. [10]Faculty of Chemistry and Pharmacy and BIOSS Centre for Biological Signalling Studies, University of Freiburg, Freiburg, Germany. [11]Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria. e-mail: peter.szoevenyi@uzh.ch

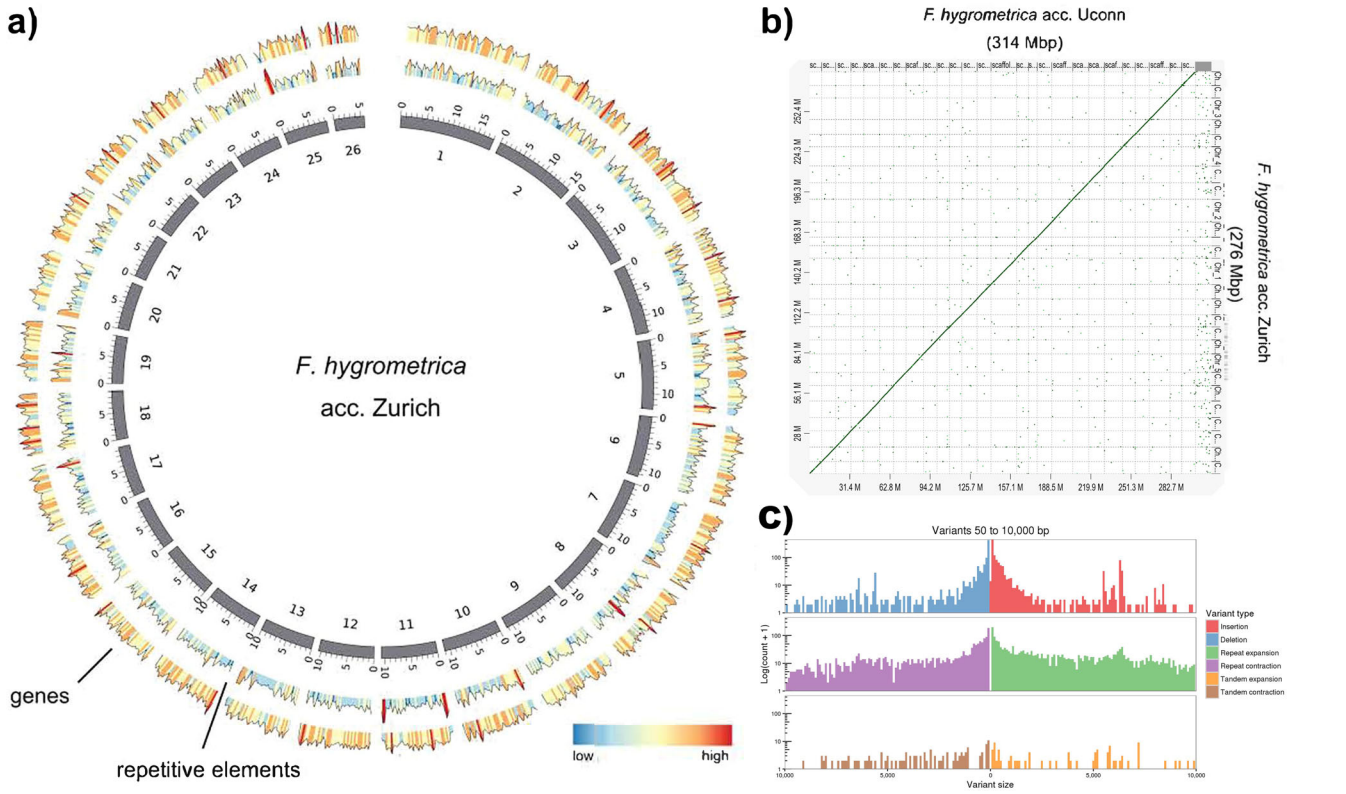

Fig. 1 | Organization and structural variation of the two *F. hygrometrica* genomes. a Circos representation of chromosome-scale pseudomolecules of the Zurich assembly. From outer to inner circle: Gene density in 250-kb windows along the putative chromosomes; Density of repetitive elements in 250-kb windows; representation of the 26 putative chromosomes (tick spacing is one Mbp). **b** Dot plot of alignable regions between the UConn and Zurich assembly of the *F. hygrometrica* genome. The plot was generated with D-GENIES using a similarity threshold of 70%[108]. **c** Length and frequency distribution of structural variants between the Zurich (reference) and UConn assembly. Variants were classified and plotted using assemblytics[111].

Mosses compose the most species-rich lineage of bryophytes (mosses, liverworts, and hornworts) and the group of seed-free plants with the largest number of pseudomolecule-scale genome assemblies[23,24,28–32]. Among mosses, the Funariaceae provide an appropriate model system to investigate genome evolution and dynamics, as they hold the most widely used model organism of bryophytes, the moss *Physcomitrium* (*Physcomitrella*) *patens*, for which a high-quality genome sequence and extensive associated genomic resources are available[28,32–37]. The family also exhibits broad diversity in genome size, ploidy level, and chromosome numbers following 90 million years of diversification[33,37–41] which has been extensively investigated[34,35,37,42,43]. Collectively, these genomic and phylogenetic resources make the Funariaceae a prime model to explore key questions of genome evolution in the largest group of seed-free plants, i.e., the mosses.

Based on two new pseudomolecule-scale genome assemblies for two accessions of the moss *Funaria hygrometrica* (Funariaceae), we investigate through comparisons with that of the related *P. patens* and other moss genomes, three main aspects of genome dynamics: (i) the mechanism of genome size change, (ii) the contribution of gene gain/loss and duplication to gene content variation, (iii) and their overall effect on genomic synteny (occurrence of a set of homologous loci on a set of homologous chromosomes) and collinearity (similar order of a set of homologous loci in two genomes).

We observe that the two investigated moss genomes, *F. hygrometrica* and *P. patens*, show remarkably higher levels of chromosomal synteny and collinearity compared to seed plant genomes. We further show that the enhanced collinearity is not unique to the two investigated species but extends to other moss genomes. The genome structure is conserved despite ancient whole-genome duplications and considerable recent TE activity. We also provide further evidence to the assertion that the spatial distribution of genes and transposable elements along the chromosomes is more even in mosses than in seed plants. Overall, our findings support the hypothesis that large-scale genome structure and dynamics of mosses and seed plants differ. In particular, moss genomes exhibit a lower rate of gene order reshuffling along chromosomes compared to seed plants. We speculate that the unique large-scale genome structure of mosses can potentially contribute to their overall smaller and less variable genome sizes compared to seed plants.

Results

F. hygrometrica accessions have 26 pseudomolecules

We assembled the genome of two *F. hygrometrica* accessions (one collected in St. Gallen, Switzerland [hereafter referred to as "Zurich"]; the other in Willimantic, Connecticut, USA ["UConn"]) using long-reads, Chicago and Hi-C libraries at the level of pseudomolecules (Fig. 1a, Supplementary Data 1, Supplementary Notes 1 and 2, and Supplementary Figs. 1 and 2). Both assemblies were of high-quality resulting in 26 large scaffolds containing 99.10/96.11% (first 26 scaffolds Zurich accession=277,486,149 bp; first 26 scaffolds UConn accession = 301,785,107 bp) of the approx. 300 Mbp (full length of the assemblies: 280 Mbp Zurich, 314 Mbp UConn) genome estimated using k-mer analyses with a minimum proportion of gaps for both accessions (Supplementary Data 2).

The assembled genomes were smaller than their estimated genome sizes using k-mer analysis (Supplementary Note 1, Supplementary Fig. 1, and Table 1), whereas flow cytometry-based estimates suggested a 50–70 Mbp larger genome size. We believe that this difference between assembly lengths and flow cytometry-based estimates is mainly of technical origin, and it is unlikely that a significant portion of the genome is missing from the assemblies. It is well-known that flow cytometry often overestimates genome sizes which is well exemplified by our estimate obtained for the related *P. patens* genome that is 40–50 Mbp larger than the actual length of its telomere-to-telomere genome assembly[32,44]. Whole-genome alignment and

Table 1 | Proportional and absolute abundance of Copia- and Gypsy-like LTR elements in the *F. hygrometrica* and *P. patens* v6 genomes as estimated by the EDTA pipeline[95]

	F. hygrometrica acc. Zurich	*F. hygrometrica* acc. Uconn	*P. patens*
# Copia elements	49,215	64,501	25,512
# intact Copia elements	27,990 (57%)	40,481 (63%)	11,151 (44%)
# Gypsy elements	24,801	69,216	209,285
# intact Gypsy elements	16,204 (65%)	30,697 (44%)	116,257 (55%)

dot plot analysis of the genomes of the two accessions obtained by different long-read sequencing technologies (PacBio and Oxford Nanopore) and assembly algorithms revealed highly collinear scaffolds (Fig. 1b, c and Supplementary Notes 2 and 3, and Supplementary Figs. 2 and 3), suggesting the absence of large-scale misassemblies in either of the assemblies and thus correspondence between the 26 largest scaffolds and the 26 putative chromosomes. These observations are in line with previous chromosome counts reported for *F. hygrometrica*[45–48]. Contigs unanchored to the 26 pseudomolecules were short and contained few genes (Supplementary Data 3, Supplementary Data 4, and Supplementary Data 5; for a more comprehensive comparison of the two genomes see also Supplementary Note 3). Both assemblies represent a significant improvement compared to a previously published *F. hygrometrica* draft genome assembly[49] which was highly fragmented (over 10,000 scaffolds) and showed a considerably lower BUSCO score (See comparison in Supplementary Data 2 and 8).

Pseudomolecules of *F. hygrometrica* and its relative *P. patens* are highly collinear and syntenic

The two *F. hygrometrica* assemblies suggest the presence of 26 chromosomes (Fig. 2a) which aligns well with the 26 chromosomes recently reported for the related *P. patens* v6 genome[32] (see Supplementary Note 2 and Supplementary Fig. 2).

Dot plots of the two genomes showed a 1 to 1 correspondence between all 26 chromosomes with virtually no exchange of genomic segments among chromosomes (Fig. 2a–c). That is, homologous chromosomes of *F. hygrometrica* and *P. patens* housed homologous set of genes hence are highly syntenic. The *F. hygrometrica* and *P. patens* genomes were also highly collinear, showing a highly conserved linear order of homologous genes (Fig. 2a–c). This implies that in contrast to previous assumptions[28], both species have 26 chromosomes, and this karyotype, including structure and gene content of the chromosomes, remained stable over at least 60–80 million years (Myr) of evolution[37,50].

The *F. hygrometrica* genome has higher gene density but is smaller than that of *P. patens*

Evidence from flow cytometry measurements (*P. patens*: 511 Mbp [*C* value 0.53 pg]; *F. hygrometrica* Zurich: 380 Mbp, [*C* value 0.4 pg], for further information see Supplementary Note 1) and genome assemblies (assembly length: *P. patens* v6: 482 Mbp; *F. hygrometrica* Zurich accession: 280 Mbp; UConn accession: 314 Mbp) reveal that the *F. hygrometrica* genome is at least 160–200 Mbp smaller than the *P. patens* genome (Supplementary Data 2, see also in Supplementary Note 1–4 and Supplementary Figs. 1–4)[28,32,51]. In contrast, the genome of each *F. hygrometrica* accession harbored somewhat more genes than that of *P. patens*, with 36,301 (UConn) and 36,804 (Zurich) high-quality filtered gene models (Supplementary Data 7), which is about 250–800 genes more than the 36,044 genes reported for the *P. patens* v6 genome[32]. Importantly, about 80% of the predicted gene models were supported by expression evidence (TPM ≥5), and 85% of Zurich and 74% of UConn gene models have RNA-seq coverage higher than 80% (Supplementary Data 7). Furthermore, BUSCO scores of both the annotated gene set and that of the genome sequence were among the top of currently published chromosome-scale genomes (including *P. patens*) (Supplementary Data 8). Due to the smaller genome size and slightly larger gene set, gene density of the *F. hygrometrica* genome is nearly twice that of *P. patens* (13.08/12.20 genes/100 kbp [*F. hygrometrica* Zurich/UConn] vs 7.48 genes/100 kbp [*P. patens* v6]).

Genome size difference is mainly due to TE expansion/contraction

The size difference between the *F. hygrometrica* and *P. patens* v6 genomes (i.e., +150–200 Mbp) may mainly be explained by the higher proportion of repeat elements in the latter. Indeed, nonalignable parts of the *F. hygrometrica* and *P. patens* v6 genomes were enriched in their respective dominant LTR elements (see below), whereas the alignable segments were enriched in exonic and intronic regions (Fig. 3a and Supplementary Data 6). Nonalignable regions were also enriched in segments of the genomes containing not annotated features (regions outside of exons, introns, and repeat elements). Furthermore, the length and proportional TE content increase of homologous *P. patens* v6 and *F. hygrometrica* chromosomes is significantly and positively correlated (Spearman's Rho=0.517265, *P* value = 0.00753), strengthening the hypothesis that the larger genome size in *P. patens* resulted, at least in larger part, from an increased representation of repeat elements and intergenic regions.

The repeat element content of the *F. hygrometrica* and *P. patens* genomes differ both quantitatively and qualitatively (Fig. 3b and Table 1). About a third of the *F. hygrometrica* genomes (32% and 37% of the Zurich and UConn accessions' genome, respectively) is composed of repeats (Fig. 3b and Supplementary Data 6), compared to 60% in the *P. patens* v6 genome, reflecting a 23-28% difference in the fraction of repetitive elements between the genomes of these species. In other words, almost 80% of the genome size difference between *P. patens* and *F. hygrometrica* can be attributed to differences in repetitive element content alone. Furthermore, the class of LTR dominating the repeat content differed between the two species: Gypsy elements in *P. patens* v6 (3% [Copia] vs. 27% [Gypsy]) versus Copia elements (16–17% [Copia] vs. 9–12% [Gypsy]) in *F. hygrometrica* (Fig. 3b and Supplementary Data 6).

The overall difference in repeat content and abundance of Copia and Gypsy elements between the *F. hygrometrica* and *P. patens* genomes could have arisen by lineage-specific expansion of LTRs. Intriguingly, our reanalysis of the temporal activity of Copia and Gypsy element insertions in the *P. patens* v6 and *F. hygrometrica* genomes revealed shared histories (Fig. 4). Although Copia elements exhibited rather continuous activity through time (albeit with a recent and an older peak of activity in the genome of both species) and Gypsy elements a mostly recent activity (Fig. 4), the temporal dynamics of dominant LTR elements did not differ between the two genomes. Further, the absolute number of intact Gypsy elements was more than three-fold higher in the repeat-rich and Gypsy-dominated *P. patens* v6 than in *F. hygrometrica* (Table 1). The significantly larger number of all and intact Gypsy elements in *P. patens* and the similar proportion of intact elements in *P. patens* and *F. hygrometrica* suggest that the difference between the two genomes likely arose via a more massive activation of Gypsy elements in *P. patens*. By contrast, the activity of Copia and Gypsy LTRs was more balanced in *F. hygrometrica*, but overall, at a lower level compared to *P. patens*.

Proportion of species-specific genes is similar to seed plants

To compare homology of the gene sets of the two species using a phylogenetic approach, we created orthogroups using proteomes of 38 plant species including 12 bryophytes, the two *F. hygrometrica* accessions, various

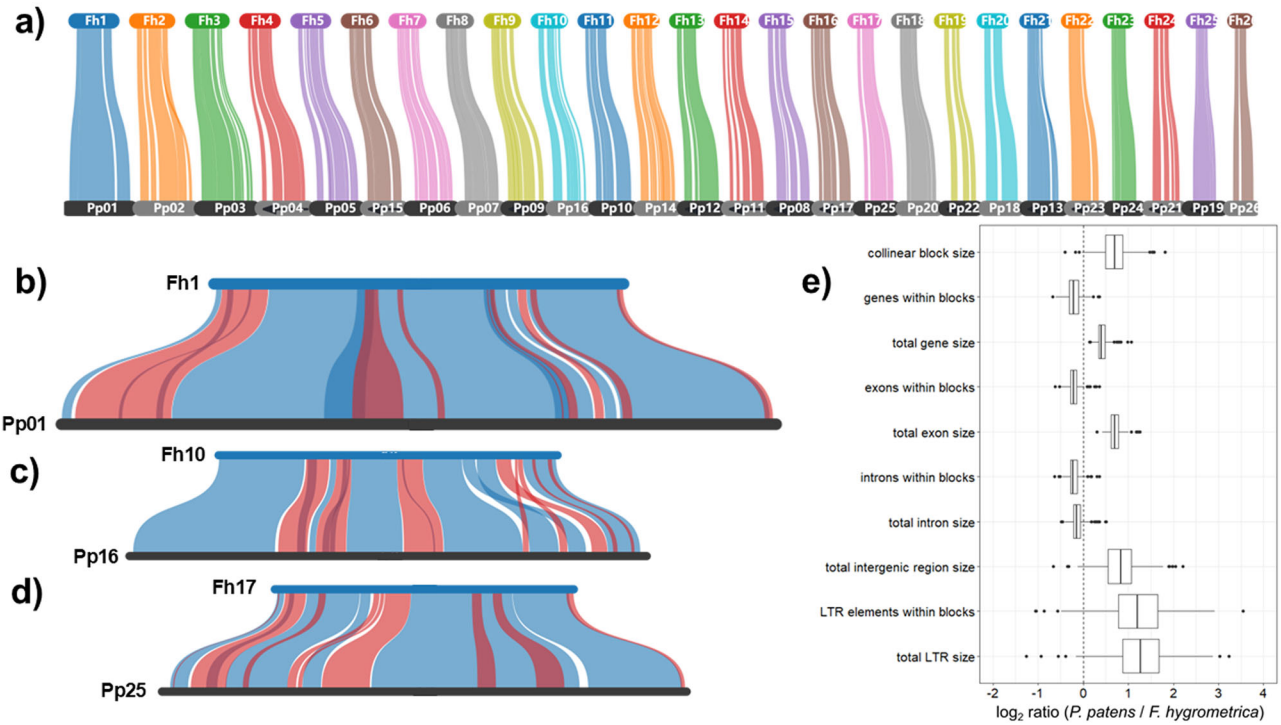

Fig. 2 | Intergenomic collinearity between the *P. patens* v6 and *F. hygrometrica* acc. Zurich genome. a Collinearity between the *P. patens* v6 and *F. hygrometrica* pseudomolecules. Collinear blocks of the two genomes are connected with colored ribbons, and pseudomolecules are drawn to scale. **b**, **c** Collinearity between syntenic pseudomolecules of *F. hygrometrica* and *P. patens* v6 (fh1 vs. pp01, and fh10 vs. pp16). Blue ribbons connect collinear blocks with the same directionality, while red ribbons depict blocks with inverted positions. **d** Collinearity and synteny between fh17 and pp25. **e** Comparison of genomic features in the corresponding collinear blocks of *P. patens* v3 and *F. hygrometrica*. The box plot shows the median and interquartile ranges, whiskers represent values up to 1.5 times the interquartile range, values outside this range are represented as individual data points.

vascular plants, chlorophyte and streptophyte algae (Supplementary Data 9). We recovered 60,901 orthogroups including 83.5% of the genes with only 0.5% of the orthogroups being species-specific (Supplementary Data 10).

A considerable proportion (41.11%, 15,026 genes) of the *F. hygrometrica* gene set occurs in species-specific gene families (containing *F. hygrometrica* but no *P. patens* gene models), compared to 30.5% (i.e., 10,044 genes) of the *P. patens* v6 gene set. Therefore, shared gene families housed about 60% (58.89% 21,527/36,553) and 70% (69.50% 22,882/32,926) of the *F. hygrometrica* and *P. patens* gene sets, respectively (Supplementary Data 10). Presence/absence polymorphism of genes was not an artifact of gene prediction. Virtually all predicted *F. hygrometrica* gene models had RNA-seq coverage (Supplementary Data 7) and gene families with species-specific genes had genes predicted for both accessions (Supplementary Data 10). Importantly, *F. hygrometrica*-specific genes showed a considerable level of expression as well as expression variation across developmental stages. Furthermore, only 4.92% (739 gene models) of the 15,026 *Funaria*-specific genes could be partially (50% coverage threshold) mapped to the *P. patens* genomic sequence, and of these 42.63% (315) produced truncated gene models with one or more frameshifts. *P. patens* holds a similar set of species-specific genes. Only 8.69% (873) of the 10,044 *P. patens*-specific genes had partial matches in the *F. hygrometrica* genome sequence, and of these 20.96% (183) were further interrupted by frameshifts. Thus, more than 90% of the species-specific genes cannot be detected in the alternate genome, and therefore likely represent de novo gene gains/losses following the divergence of the two species. The remaining and considerably smaller proportion of species-specific genes likely underwent gene degeneration/pseudogenization in one of the two genomes. The proportion of species-specific genes in the *F. hygrometrica* (41%) and *P. patens* (30%) genomes, falls within the range that we found between seed plants with similar depth of divergence (i.e., 14–49%, median of 28% and mean of 29.17%, Supplementary Note 6) and therefore does not seem to be exceptional, or indicative of distinct modes of genome evolution in mosses.

Although the proportion of genes unique to either of the two moss species was considerable, over 60% of the gene set occurred in shared gene families (Supplementary Data 10). Our gene family expansion/contraction analysis indicated that very few gene families (about 1%, see Supplementary Note 7 and Supplementary Data 11) showed significant expansion or contraction between *F. hygrometrica* and *P. patens*. This supports that the gene set difference between *P. patens* and *F. hygrometrica* primarily achieved by de novo gain/loss of genes and not by gene family expansion/contraction.

The genomes of *F. hygrometrica*, *P. patens*, and other mosses show higher level of collinearity than those of seed plants

The *P. patens* genome appears to have significantly expanded via the activation of LTRs, which could have triggered extensive spatial reshuffling of the gene set, resulting in lower collinearity (similar order of homologous genes in two or more genomes) with the *F. hygrometrica* genome. Despite considerable genome expansion and more than 60 million years of independent evolution, the genomes of the two species remain remarkably collinear (Fig. 2a–d and Supplementary Data 12). More specifically, about half of the *F. hygrometrica* and *P. patens* v6 gene set (50.67% or 18,263 genes and 49.15% or 18,088 genes in *F. hygrometrica* [Zurich accession] and *P. patens* v6, respectively) occurred in 759 collinear blocks containing at least five collinear genes. About 70% and 60% of the *P. patens* v6 and *F. hygrometrica* genes, respectively, have homologs in the alternate genome, and almost all shared genes (80–90%) occur in collinear blocks, of which about half (52.17%, 396) were inverted in one of the two genomes. Inverted regions did not serve as hotspots of genome evolution, as inverted and noninverted collinear blocks in the *F. hygrometrica* [Zurich accession] and *P. patens* v6 genomes, were similar in terms of length, overall number of genes and number of collinear gene pairs (Fig. 2a–d).

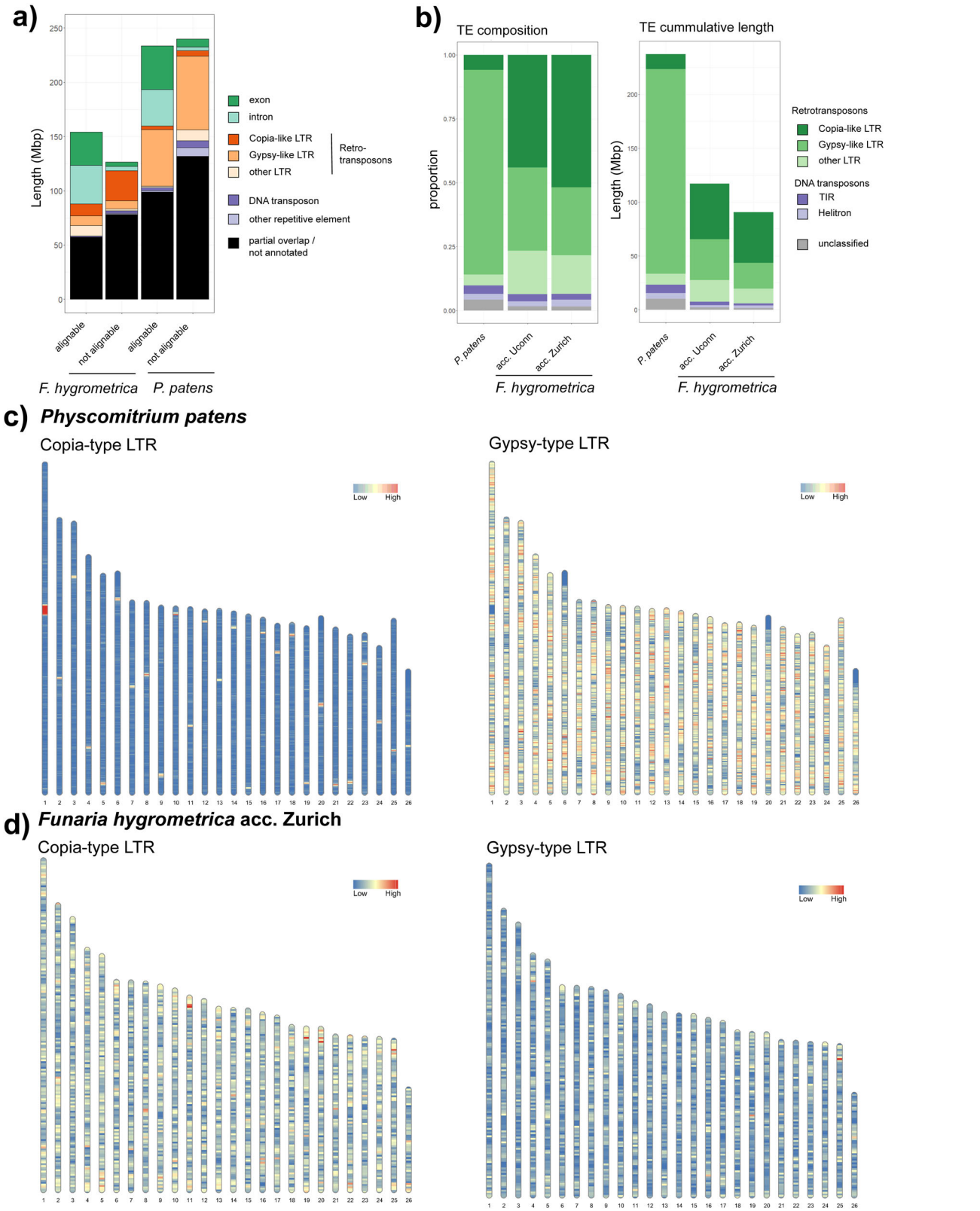

Fig. 3 | Transposable element annotation of Funariaceae genomes. a Genome features in alignable and not alignable regions between the *F. hygrometrica* acc. Zurich and *P. patens* v6 genome. Only features fully overlapping with the respective genomic regions are shown as annotated, partially overlapping features are shown in black alongside regions without annotation. The whole-genome alignment was computed using the minimap2 algorithm with default parameters[109]. **b** Transposable element annotation summary of the three studied genomes. The bar plot on the left shows the composition of transposable element families as a fraction based on their total sequence length in the genome. The plot on the right shows the absolute length the different families occupy in the respective genome. **c**, **d** Density of Long terminal repeat (LTR) retrotransposons on the chromosomal scaffolds of *P. patens* v6 (c) and *F. hygrometrica* (d) in 100 kb windows.

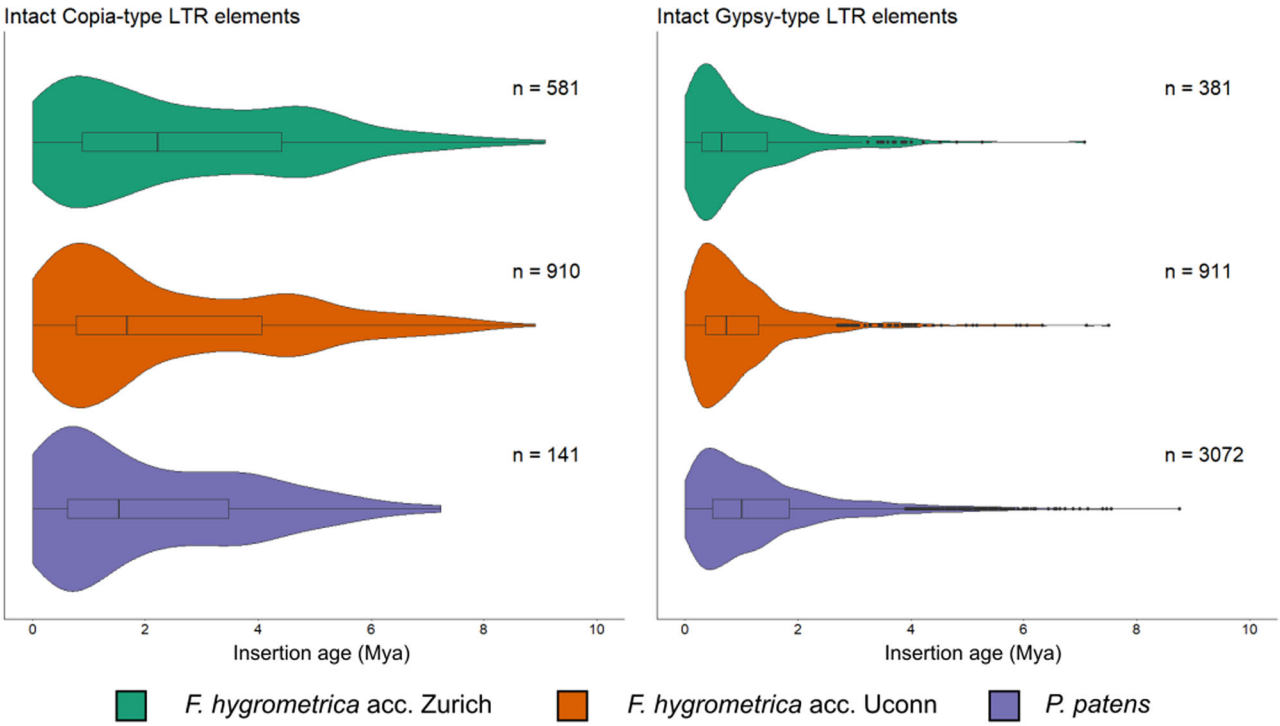

Fig. 4 | Insertion age distribution of intact full-length LTR elements in the *Funaria* genomes and in *P. patens* (*P. patens* v6). Insertion time is estimated by calculating the sequence divergence between left and right terminal repeats and provided in millions of years ago (Mya). "*n*" refers to the number of full-length elements included in the analysis.

We further compared the extent of collinearity between the two moss genomes with that of seed plants with similar divergence times. We first used pairs of grass genome with chromosome-scale assemblies known to show exceptional collinearity among seed plants with a splitting time of 46–62 million years ago (Mya), which is similar to the estimated divergence time of *F. hygrometrica* and *P. patens* (see methods). Only about 49–65% of the homologous genes were in collinear blocks in these pairwise comparisons (Fig. 5a–c), a proportion roughly 20% smaller than that observed between the two current moss genomes. Altogether, our observations indicate that despite new LTR insertions and 60 million years of divergence, exceptional collinearity was retained over most of the genome (Fig. 5a–c).

To test whether exceptional collinearity is a specific feature of the *F. hygrometrica*–*P. patens* genome pair, we extended our collinearity analysis to other seed plant and moss genomes. To do so, we calculated the proportion of homologous genes located in collinear blocks between pairs of moss and seed plant genomes with comparable divergence times. We only included moss and seed plant genomes with chromosome-scale assemblies. We compared species pairs with approx. 30 and 175 million years (Myrs) of divergence time because high-quality moss genomes were only available for these two splitting time categories (see species list in methods). Our results confirmed that pairs of moss genomes always showed greater collinearity than seed plants with comparable divergence depth (Fig. 6). Altogether, our observations indicated that despite new LTR insertions and 60 million years of divergence, the two Funariaceae genomes show exceptionally high collinearity compared to seed plant genomes of similar age of divergence and suggest that such pattern may characterize moss genomes with highly contiguous genomes available so far (Figs. 5 and 6).

Smaller genome size of *F. hygrometrica* is mirrored by shorter collinear regions

The majority of the genome of *F. hygrometrica* was composed of gene blocks collinear with portions of the *P. patens* v6 genome. Given the significantly smaller genome of *F. hygrometrica* (i.e., by 150–200 Mbp), we expected that its collinear blocks are shorter (Fig. 2e and Supplementary Data 12). Indeed, the overall size of the genomic segments containing the collinear blocks was about half as large in the *F. hygrometrica* compared to the *P. patens* genome (Fh_{median}=271,894 bp, interquartile range [IQR] = 166,567-455,659 bp; Pp_{median} = 444,906 bp, IQR = 260,162-779,820, Wilcoxon rank sum test W = 248333, $P < 2.2e-16$). This difference was largely due to the smaller intergenic regions in *F. hygrometrica* compared to *P. patens* (Fh_{median}=190,100 bp, IQR = 109400-328200; Pp_{median} = 325,400 bp, IQR = 182,000-617,800), while genomic segments of the collinear blocks contained somewhat more genes in *F. hygrometrica* than in *P. patens* (Fh_{median}=39.00, IQR = 25-64; W = 110432; Pp_{median} = 34.00, IQR = 22.00-56.00, $P < 2.2e-16$). This is in line with our above observation that genome size change was primarily achieved by repeat expansion leading to an overall increased gene density in collinear blocks in *F. hygrometrica* versus in *P. patens* (Fig. 2e).

The most recent whole-genome duplication is shared by *F. hygrometrica* and *P. patens*

Previous analyses suggested that the ancestor of mosses may have had seven chromosomes, and that two whole genome duplications (WGD) subsequently occurred in the line leading to *P. patens*[23,24,28]. Signatures of the older WGD dated to about 200 Mya are shared by *Ceratodon purpureus*[23,52] and *Syntrichia caninervis*[24], whereas the more recent one likely predated the origin of the Funariaceae[28,38,42,53,54]. We found abundant collinearity and synteny between the *F. hygrometrica* and *P. patens* v6 chromosomes (Fig. 2a–d), and our Ks analysis resulted in very similar Ks distributions in *F. hygrometrica* and *P. patens* v6 (Fig. 7a, b). Furthermore, both species' Ks distribution showed two major peaks at Ks = ~0.8 and Ks = ~1.2 representing two potential WGDs. Self-synteny maps of both genomes were also very similar, further confirming the presence of two shared WGDs (Fig 7a, b). Therefore, our Ks and self-synteny-based analyses suggest that both the old and the more recent WGDs are shared by the two species (Fig. 8a–c). This confirms that both WGDs preceded the split of *F.*

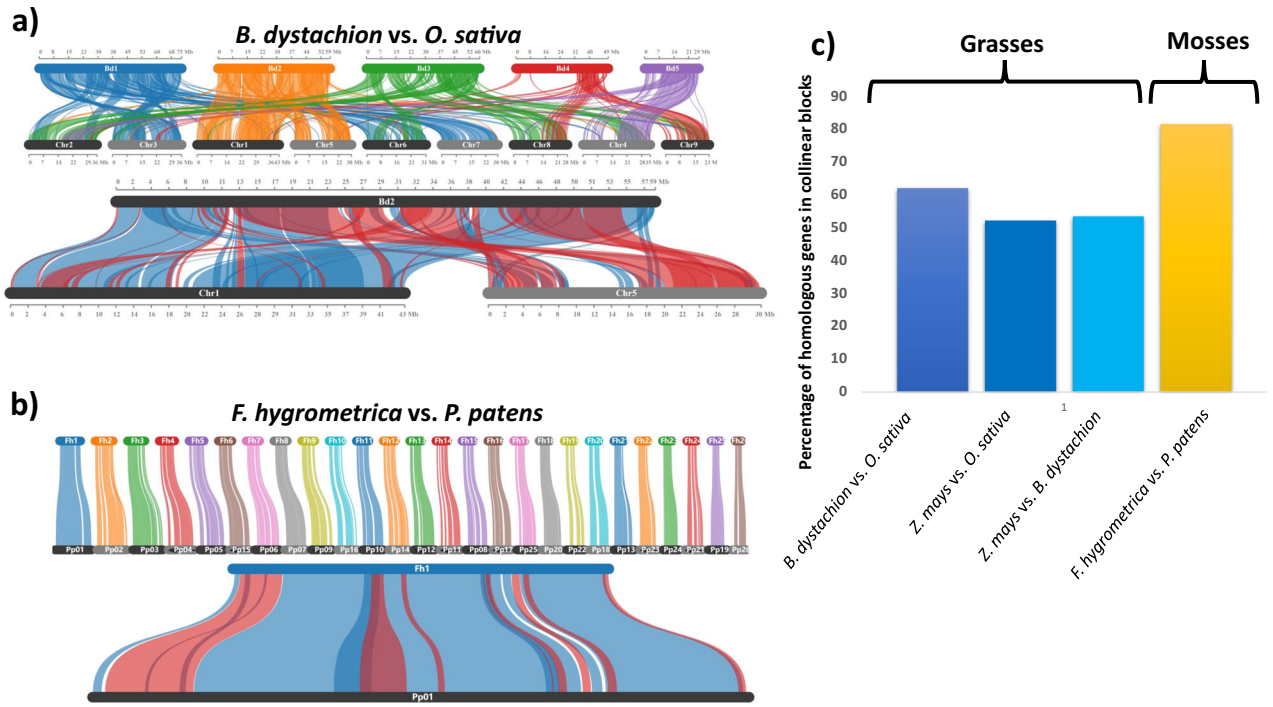

Fig. 5 | Comparison of collinearity between pairs of grass and moss genomes. **a** Collinearity between the complete *Brachypodium dystachion* and *Oryza sativa* genomes (top) and between the second chromosome of *B. dystachion* and chromosome one and five of *O. sativa* (bottom). **b** Collinearity between the complete genomes of *F. hygrometrica* and *P. patens* v6 (top) and between their fully syntenic first chromosomes (bottom). **a**, **b** Ribbons connect collinear blocks of the two genomes, and the first genome in the comparison is always shown on the top of the figures. In the close-up figures inverted and noninverted collinear blocks are shown in red and blue colors, respectively. **c** Percentage of homologous genes found in collinear blocks in the grass species pairs and in the moss species investigated. Abbreviations: *Z. mays Zea mays*, *O sativa*, *Oryza sativa*, *B. dystachion*, *Brachypodium dystachion*, *F. hygrometrica*, *Funaria hygrometrica*, *P. patens Physcomitrium patens*.

hygrometrica and *P. patens* and may be specific to the Funarioideae or any prior node (e.g., Funariaceae) in the evolution of the Funariidae[37].

The *F. hygrometrica* genome lacks the specific telomeric repeat reported for *P. patens* v6

The newly released *P. patens* v6 genome reported that 13 out of the 52 telomeres show an unusual sequence composition[32]. In particular, the 13 telomere sequences consist of the unusual (GGCCCA)n instead of the frequently observed (TTTAGGG)n repeat. It was argued that this unusual telomere repeat may be the result of chromosomal fission and healing. Using this assumption a new scenario of chromosomal evolution was proposed including 14 breaks, three fusions and nine small chromosome fragment losses following the most recent WGD. Because we observed 1 to 1 homology between the *P. patens* and the *F. hygrometrica* chromosomes and both WGDs are shared by the two species, similar telomeric repeats would be expected in *F. hygrometrica* and *P. patens*. While not all of the *F. hygrometrica* chromosome ends contained properly assembled telomeres, the (GGCCCA)n repeat pattern did not occur in either of the assemblies. Nevertheless, besides the often observed (TTTAGGG)n repeat we found that multiple chromosomes contained a telomeric repeat with the unusual (CCCAGGG)n pattern (Supplementary Note 9: Supplementary Table 2). This discordance between the telomeric repeats of the two species may either suggest that healing is unrelated to the unusual telomere structure of *P. patens* or it produces different telomeric repeats in the two species.

Overall chromosome structure of *F. hygrometrica* resembles that of *P. patens* and other bryophytes

The overall chromosome structure of the *F. hygrometrica* genomes is very similar to that of the other published bryophyte genomes. Pericentromeric regions enriched for TEs could not be identified and gene and repeat features were rather uniformly distributed along the chromosomes (Fig. 1a, Fig. 3c, d, Supplementary Note 8, and Supplementary Figs. 5–7). Although pericentromeric regions enriched in TEs are not present in the *P. patens* genome, the distribution of RLC type 5 Copia elements shows a peak at the centromeric regions[28,32]. In the *F. hygrometrica* genome, we could not identify a single Copia or Gypsy subfamily that showed a clear and single peak along any pseudomolecule (Fig. 1a and Supplementary Note 8). Therefore, we conclude that the association of RLC elements with the putative centromeres is specific to *P. patens* and does not occur in *F. hygrometrica*.

Discussion

Comparison of sequenced seed-free and seed-plant genomes suggests that their overall genome structure may differ in multiple aspects[5]. Yet given the paucity of information on the genome evolution and dynamics of seed-free plants, the ultimate causes of their differential genome structure is poorly understood[21]. By conducting comparative analyses of the *F. hygrometrica*, *P. patens*, and some additional moss genomes, we found that structural variation, which rearranges the physical position of genes along chromosomes, is likely to be slower in mosses than in seed plants. Moreover, we provide further evidence to the assertion that the spatial distribution of genes and transposable elements (TE) along the chromosomes is more even in mosses than in seed plants. We speculate that the slower pace of structural genomic variation and the unusual spatial distribution of TEs and genes may be a general characteristic of moss genomes and contribute to their overall smaller and less variable genome sizes compared to seed plants. This hypothesis needs further testing once an extensive data set of highly contiguous moss genome assemblies are available.

Moss genomes are smaller and more conserved in size than those of seed plants[27]. This could potentially be explained by a genome expansion/contraction mechanism that is different from that of seed plants. In most

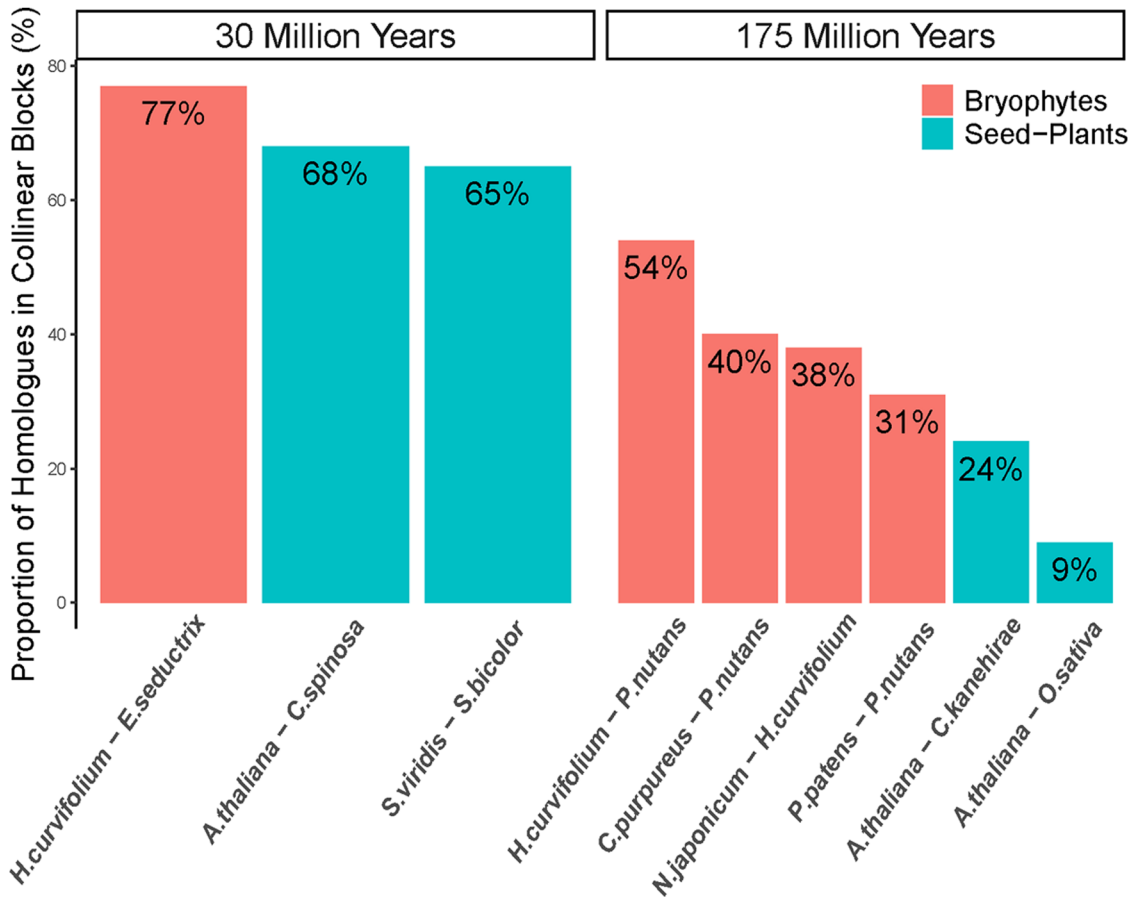

Fig. 6 | Comparison of collinearity between pairs of seed plant and moss genomes. The percentage of homologous genes found in collinear blocks are shown between pairs of moss and seed plant genomes with 30 Myr and 175 Myr of divergence depth. Abbreviations: *H. curvifolium Hypnum curvifolium*, *E. seductrix Entodon seductrix*, *A. thaliana Arabidopsis thaliana*, *C. spinosa Capparis spinosa*, *S. viridis Setaria viridis*, *S. bicolor Sorghum bicolor*, *C. purpureus Ceratodon purpureus* GG1, *P. nutans Pohlia nutans*, *N. japonicum Niphotrichum japonicum*, *P. patens Physcomitrium patens* v6, *C. kanehirae Cinnamomum kanehirae*, *O. sativa Oryza sativa*.

seed plants, genome size variation is caused by the expansion/contraction of non-coding DNA, especially TE-elements[55–58]. By contrast, direct genomic evidence for the predominant effect of TEs in the genome size variation of mosses and bryophytes was lacking. We provide evidence that the 150–200 Mbp larger size of the *P. patens* (versus the *F. hygrometrica*) genome is primarily due to its increased TE content, suggesting that genome size variation in the Funariaceae, and hence perhaps in mosses and bryophytes in general, is driven, like in seed plants, by expansion/contraction in TE-elements. Therefore, our observations provide little support to the idea that genome size change would follow radically different mechanisms in mosses and seed plants. Nevertheless, the possibility that genome size change via TE expansions/contractions is more constrained or genome downsizing is more effective via TE deletion in mosses than in seed plants cannot be ruled out.

Although the mechanism underlying change in genome size may be similar in seed plants and mosses, the genomes of the latter may be structurally less dynamic (e.g., fewer chromosomal translocations, inversions, insertions, deletions than in seed plants). This could potentially explain why moss genomes are on average smaller, less variable in size, and partitioned among a more conserved number of chromosomes than those of seed plants. Indeed, previous comparative analyses among moss genomes that diverged over 170–200 Mya revealed detectable synteny among chromosomes (so-called "ancestral elements")[23,24]. Our analysis recovered unexpectedly strong collinearity between the *F. hygrometrica* and *P. patens* genomes, despite 60–80 million years of divergence[37], which is in line with recent findings[29,59] and overriding observations made between highly collinear grass genomes with similar depth of divergence[60–66]. Furthermore, we also showed that greater collinearity of moss genomes versus seed plant genomes is not a specific characteristic of *F. hygrometrica* and *P. patens* but appears to be a general feature of the moss genomes available so far. Therefore, moss genomes are characterized by a lower rate of chromosomal translocations and rigidity of collinear gene blocks compared to seed plants, a feature is probably shared with another group of bryophytes, the liverworts[59]. The mechanisms and/or constraints leading to lower structural dynamics of chromosomes are unclear but could be related to the haploidy of the moss genome directly exposing mutations to natural selection. This is expected to increase the efficacy of purifying selection potentially leading to extended synteny and collinearity[67,68]. It is also possible that high efficiency of homologous recombination facilitates homology-mediated repair, which could increase genomic synteny and collinearity[28,33,69].

Our study provides further evidence to the hypothesis that bryophytes and potentially all seed-free plant genomes are characterized by an unusual chromosome structure with repeat and gene features relatively evenly spread along the chromosomes[22,26,70–75] lacking a broad TE enriched pericentromeric region (*Marchantia polymorpha*[76], *Ceratodon purpureus*[23], *Syntrichia caninervis*[24], *Anthoceros agrestis*[77], *Pohlia nutans*[78]). This is in stark contrast to the usual chromosome structure of flowering plant genomes wherein gene density is highest in the middle of the chromosome arms and repeats more dominant in the pericentromeric regions. The processes potentially contributing to this unique arrangement of genes and repeats are obscure and remain to be investigated[5].

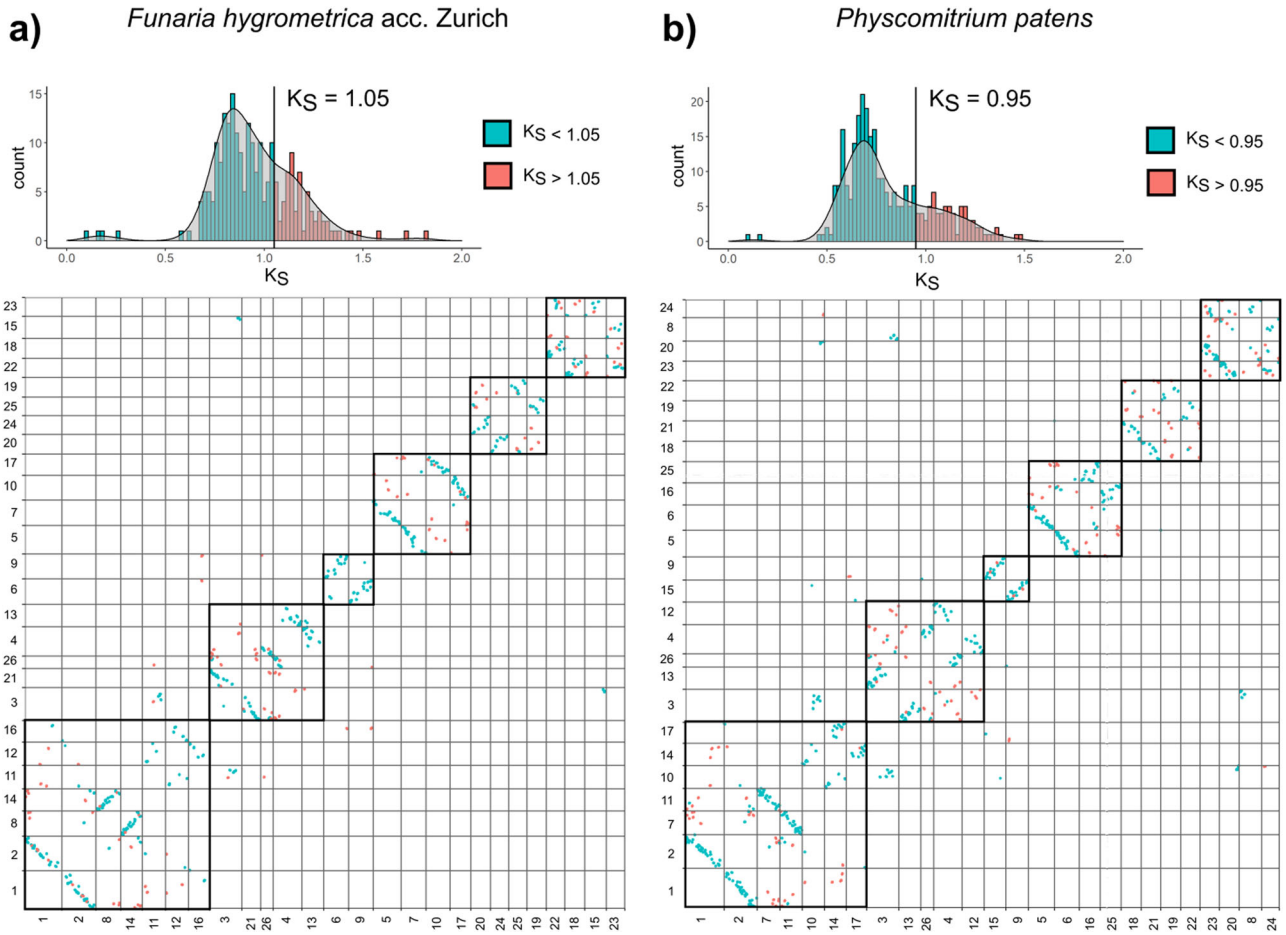

Fig. 7 | Dot plot of self-synteny among pseudomolecules of the *F. hygrometrica* Zurich accession and *P. patens* v6. Pseudomolecule blocks corresponding to the putative ancestral chromosomes are framed. Histograms above the dot plots show the distribution of average K_s values per collinear block. Histograms are colored according to the two whole-genome duplication events. Pseudomolecules in the dot plots are ordered according to intergenomic synteny between **a** *F. hygrometrica* and **b** *P. patens* v6.

We can only speculate about if, and to what extent, the unique distribution of TEs and genes as well as the overall low rate of structural genomic variation may contribute to the smaller and less variable genome sizes in mosses and potentially in most bryophytes versus seed plants. It is possible that increased purifying selection and homology-mediated repair facilitate TE deletions thereby constraining genome size change. This hypothesis assumes that TE insertions are deleterious and genome size change is primarily driven by natural selection. We can also hypothesize that the even distribution of gene-rich islands embedded in the sea of TEs may contribute to smaller and less variable genome sizes in bryophytes. This is because tight intermingling of TEs and genes may facilitate TE deletion rates owing to the greater recombination rate of genic regions[5,79,80].

Conclusions

In comparison to the rapidly growing understanding of genome evolution and dynamics in flowering plants, very little is known about patterns and processes pertaining to changes in the genomes of seed-free plants[2,5]. Our analyses and the integration of previous observations suggest that moss genomes show more extensive synteny/collinearity than those of flowering plants. Therefore, our results provide further support to the hypothesis that genome dynamics of mosses and potentially other bryophytes differ from those of seed plants[5,29]. Our study provides a solid basis for a more extensive exploration of genome dynamics within the Funariaceae, to test for the generality of our observations. Moreover, the availability of a high-quality genome sequence for two species representing endpoints of the morphological and ecological diversity within the Funariaceae will open the way for detailed investigations on the genetic basis of phenotypic diversity within the family[42,49,53].

Methods

DNA sequencing

For both the Zurich and UConn accessions (Supplementary Data 1) high molecular weight DNA was extracted using a modified CTAB protocol[81]. The Genome of the Zurich accession was sequenced with Illumina and PacBio technology. Illumina libraries were generated with insert sizes of 250 bp, 350 bp, 2 kbp, and 5kbp and sequenced on Hiseq 2000, Hiseq 2500, and Hiseq 4000 systems (paired-end, 150 bp read length). PacBio data was generated on the RS II platform using C1 chemistry (3 cells) and P6-C4 chemistry (10 cells). Illumina sequencing yielded over 62 Gbp of raw sequencing data in total, while PacBio sequencing resulted in 13 Gbp of sequence data.

For the UConn accession we generated two Illumina libraries with an insert size of 400 bp and sequenced them using the HiSeq Xten platform (paired-end, 150 bp read length). Using the very same DNA, we also prepared a single DNA library for Oxford Nanopore sequencing using the ligation kit and sequenced it on the Nanopore X5 platform. Illumina sequencing resulted in a total of 62 Gbp raw data. Nanopore sequencing resulted in about 1.8 million reads longer than 10,000 bp after clean-up.

To improve the continuity of the genome assemblies, we created Hi-C libraries for the Zurich accession (using the Dovetail Hi-C kit for genome assembly) and for the UConn accession using the protocol described in ref. 82. Furthermore, a Chicago library was also prepared by Dovetail Genomics for the Zurich accession aiding scaffolding and assembly

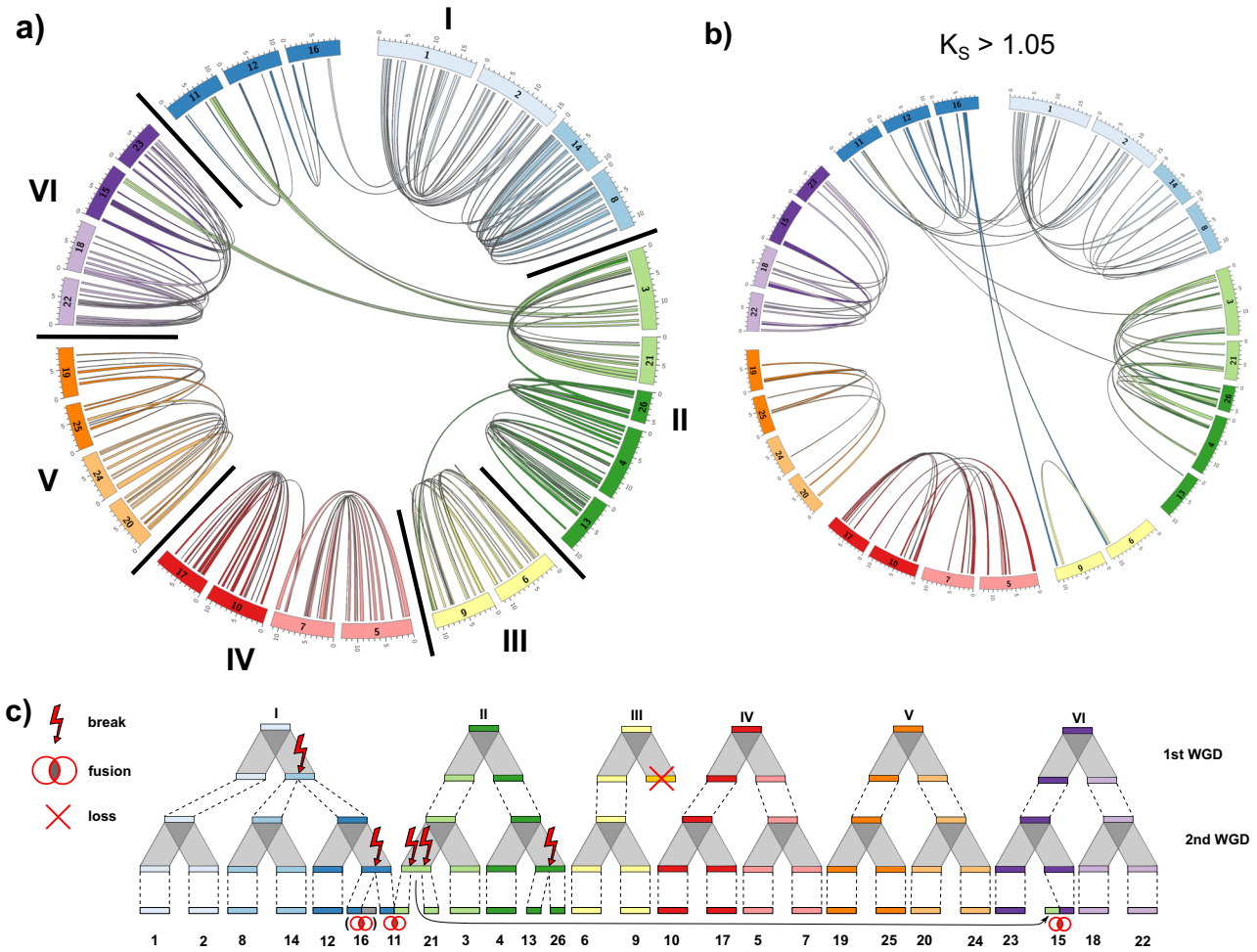

Fig. 8 | Intragenomic collinearity and karyotype evolution model of the *F. hygrometrica* genome. a Circular visualization of the 26 pseudomolecules of the *F. hygrometrica* genome (Zurich accession). Blocks of collinear genes with a mean synonymous substitution rate (Ks) ≤1.05 (corresponding to the most recent whole-genome duplication [WGD]) are connected by colored ribbons. Pseudomolecules are arranged to reflect their putative evolutionary relationship. Pseudomolecules potentially originating from the same ancestral chromosome are grouped together and labeled with I – VI. **b** Visualization of collinear blocks with a mean synonymous substitution rate (Ks) >1.05 (corresponding to the older WGD). **c** Hypothetical model of karyotype evolution in the *F. hygrometrica* lineage. Six ancestral chromosomes underwent two whole genome duplication events accompanied by one chromosome loss, five chromosome breaks, and three chromosome fusions, resulting in 26 recent chromosomes.

correction at shorter distances, in the several 100kbp range. We sequenced the Hi-C and Chicago libraries using Illumina Hiseq 4500 and Novaseq machines in paired-end mode (150 bp read length). Details of the DNA sequencing data used for the genome assembly can be found in Supplementary Data 1.

Genome assembly

For the Zurich accession the initial assembly was generated with the Canu assembler v1.5[83] using all PacBio data. Afterward, we employed HiRise[84] together with the Chicago sequencing data to scaffold the original reads into larger scaffolds and improve assembly contiguity. The resulting assembly was further consolidated by a second HiRise run employing the Hi-C data. For manual curation, Hi-C sequencing data was aligned to the assembly using the juicer pipeline v1.6[85] and files for visual inspection of the Hi-C contact map were created with scripts supplied with the 3D-DNA software package[86]. Visual review with the Juicebox v1.6 software[87] revealed no obvious misassemblies, but we identified a misjoin in the largest pseudo-molecule. We manually corrected the misjoin, and the genome assembly was updated to accommodate for the introduced scaffold split (see Supplementary Note 2).

For the UConn accession Nanopore raw reads were first corrected by Canu v1.9. The corrected reads were then assembled into contigs by NextDenovo v2.3.0 (https://github.com/Nextomics/NextDenovo) with default parameters. After assembly polishing (see below), Hi-C raw reads were processed by Juicebox v1.6 to extract valid reads which contain Hi-C contact information. The 3D-DNA pipeline was then used to cluster, orient, and order the contigs, generating chromosome-scale scaffolds. We also used Juicebox to manually adjust the scaffolding according to the contact map. After manual curation in Juicebox, the post-process module of 3D-DNA pipeline was used to generate the corrected chromosome-level scaffolds (see Supplementary Note 2).

Polishing

The assembly of the Zurich accession was first polished with the quiver tool, which is part of the PacBio SMRT Analysis software package v2.3.0.140936, using all PacBio reads obtained with the P6-C4 chemistry. We used the default thresholds to remove very low coverage scaffolds from the assembly. This polishing step also corrected base calls, filled in Ns, and corrected repeat regions. After that we mapped Illumina reads to the quiver-polished assembly using BWA mem v0.7.18[88] to correct indels and SNPs in the non-repetitive parts of the genome assembly using Pilon v1.23[89] in three rounds. Final polishing was done with PBSuite v.15.8.24[90] to fill up some of the remaining gaps of the assembly using all available PacBio reads.

To fix SNPs, indels and SVs originated from sequencing errors, the UConn genome's assembly was first corrected using all Oxford Nanopore

reads and the algorithm provided in racon v1.4.10 [91]. We further corrected the racon polished contigs using all Illumina reads and Pilon v1.23. We repeated the successive polishing with racon and pilon three times. We used both tools with default parameters.

Contamination detection and filtering

Initial assessment of the Hi-C contact maps of the Zurich accession suggested that some scaffolds of the assembly had very low coverage of Hi-C data and therefore represented potential contaminations. We made similar observation using the UConn accession's genome assembly. To identify contaminant scaffolds, we used BlobTools v1.1.1[92]. More specifically, we used all Illumina reads and the full NCBI nucleotide collection (nt) and the uniprot database to assess sequencing depth along the genomic scaffolds and to assign them to broad taxonomic categories, respectively. After that, we removed all scaffolds from the assembly that were assigned to bacterial or other non-eukaryotic taxonomic classes.

Repetitive element identification and annotation: RepeatModeler2

We used the automated approach implemented in the RepeatModeler2 package v2.0[93] to generate a de novo annotation of repetitive elements in the genomes. The resulting library of TE families was then used for annotation of repetitive elements in the genome sequence using RepeatMasker v4.1.0[94] (Supplementary Data 13–14A and B).

Transposable element annotation

We used the Extensive de novo TE Annotator (EDTA) pipeline v1.9.6[95] to get a comprehensive TE annotation for both *F. hygrometrica* genome assemblies. The contamination-filtered pseudomolecule-scale assemblies were used as an input to the pipeline together with coding sequences of genes annotated by BRAKER v2.1.0[96] and genBlastG v1.39[97] as described in the "Gene prediction" paragraphs. To avoid introducing a bias when comparing TE composition and distribution between the *F. hygrometrica* accessions and *P. patens* caused by differing annotation pipelines, we retrieved the most recent *P. patens* genome assembly v6 (Bi et al. [32], https://phytozome-next.jgi.doe.gov/info/Ppatens_v6_1) and re-annotated transposable elements using the EDTA pipeline as described above (Supplementary Data 15–17). Phylogenetic analysis of LTR Copia and Gypsy super-families are described in Supplementary Note 8.

LTR insertion time estimation

The sequences of 5' and 3' terminal repeats are supposed to be identical to each other when LTR retrotransposons are newly inserted into the genome[98]. Therefore, the degree of sequence conservation between left and right terminal repeats can be used as a proxy for insertion age of individual LTR elements. To assess the recent history of LTR retroelement insertions in *F. hygrometrica* and *P. patens*, we extracted 5' and 3' terminal repeat sequences for each LTR element classified as full-length and intact by the EDTA pipeline[95] and aligned them using MUSCLE v3.8.31 with default parameters[99]. We then calculated the Kimura 2 parameter distance[100] for each aligned pair using a custom Python script and modules from the Biopython library[101]. The divergence time between LTR pairs was estimated by dividing the distance parameter by two times the synonymous substitution rate. We used a substitution rate of 9.4×10^{-9} synonymous substitutions per synonymous site per year for both genomes, which was established for *P. patens* elsewhere[102]. We plotted LTR insertion age distributions using the R package ggplot2 v3.3.3[103].

Genome annotation

Transcriptome assembly. To aid gene prediction we generated RNA-seq data covering three developmental stages of the gametophyte and four developmental stages of the sporophyte generations in three replicates (six developmental stages in total) for the Zurich accession. Gametophyte and sporophyte RNA-seq data was also obtained for sporophyte and gametophyte tissues of the UConn accession (Supplementary Data 1). RNAseq data was first trimmed for low-quality bases and adapter sequences using Trimmomatic v0.36[104]. The strand-specific RNA-seq reads were then mapped to their respective genome using Hisat2 v2.1.0[105] and a genome-guided transcriptome assembly was generated using default options in Trinity v2.13.2[106]. A second transcript assembly was generated using StringTie2 v2.1.6[107]. Here, transcripts were assembled independently for each sample and a final set of unique transcripts was computed using the –merge function.

Gene prediction. Gene models were initially predicted separately for both *F. hygrometrica* accessions, using BRAKER2 v2.1.0[96] and consolidated afterwards as described in the Supplementary Note 5.

Whole-genome alignments and collinearity analyses

We used dot plots generated with D-GENIES v1.2.0[108] to assess collinearity between genomes of the two *F. hygrometrica* accessions as well as between *F. hygrometrica* and *P. patens* v6 (Bi et al., 2023, https://phytozome-next.jgi.doe.gov/info/Ppatens_v6_1). We aligned the genomes using Minimap2 v2.17[109]. We excluded matches with less than 90% sequence identity when aligning genomes of the two *F. hygrometrica* accessions, while a threshold of 70% was used for alignments between *F. hygrometrica* and *P. patens* v6.

To assess structural variation between assemblies of the two *F. hygrometrica* accessions, we aligned them using the NUCmer module of the MUMmer package v4.0.0[110]. To visualize and classify the observed differences, we submitted the resulting .delta file to the Assemblytics web service v1.2.1[111] (for further details, see Supplementary Note 3).

We utilized the MCScan algorithm[112] and MCScanX toolkit[113] to assess collinearity within and between the studied genomes. We used peptide sequences of primary transcripts as input to an all-vs-all homology search with the BLAST v2.9.0 algorithm[114], as recommended in the MCScanX documentation. The resulting tabular output was fed into the MCScan algorithm[112] to establish blocks of collinear genes using default parameters. We calculated synonymous and non-synonymous substitution rates for each syntenic gene pair using the tools supplied with the MCScanX toolkit[113]. We visualized collinearity within genomes using circular plots generated with Circos v0.69-9[115], while SynVisio[116] was employed to visualize collinearity between genomes.

To quantitatively assess whether the collinearity between the two moss genomes (*P. patens* vs *F. hygrometrica* acc. Zurich) is greater than what is observed between grass genome pairs, we estimated the extent of collinearity between three pairs of grass genomes having chromosomal-level assemblies and an estimated splitting age (60 Mya, estimates obtained from http://www.timetree.org/) similar to the *P. patens* v6 - *F. hygrometrica* pair. We selected the following grass genomes: *Brachypodium dystachion* vs. *Oryza sativa* (split median 46 Mya, adjusted 53 Mya), *Zea mays* vs. *O. sativa* (split median 47 Mya, adjusted 62 Mya), and *Z. mays* and *B. dystachion* (split median 46 Mya, adjusted time 62 Mya). We retrieved protein files of primary transcripts and gff files from Phytozomev13[117]. We utilized the MCScan algorithm[112] and the MCScanX toolkit[113] to assess collinearity applying the very same parameters used to assess collinearity between the two moss genomes. To quantitatively describe collinearity, we calculated the proportion of homologous genes (Blastx *e* value 10^{-10}) that are in collinear blocks.

We further tested whether our finding that moss genomes show greater collinearity than seed plant genomes can be generalized. To do so, we calculated the proportion of homologous genes located in collinear blocks between pairs of moss and seed plant genomes with comparable divergence time. We used the very same software and parameters described above. We only included moss and seed plant genomes with chromosome-scale assemblies obtained using long-read data. More specifically, we compared species pairs with approx. 30 and 175 Myr of divergence time because high-quality moss genomes were only available for these two splitting time categories. For the 30 million years depth of divergence we compared the moss genomes *Hypnum curvifolium* vs. *Entodon sedutrix* ([29] genomes were retrieved from https://db.cngb.org/search/assembly/CNA0030131/ and

https://db.cngb.org/search/assembly/CNA0030130/) and the seed plant genome pairs *A. thaliana* (genome retrieved from https://phytozome-next.jgi.doe.gov/info/Athaliana_TAIR10) vs. *Capparis spinosa* (genome retrieved from https://doi.org/10.6084/m9.figshare.17702051[118]), *Setaria viridis* (genome retrieved from https://phytozome-next.jgi.doe.gov/info/Sviridis_v2_1) vs. *Sorghum bicolor* (genome retrieved from https://phytozome-next.jgi.doe.gov/info/Sbicolor_v3_1_1). For the 175 Myr of divergence depth we used (he moss genome pairs *P. patens* v6 (genome retrieved from https://phytozome-next.jgi.doe.gov/info/Ppatens_v6.1) vs. *Pohlia nutans* ([78] genome retrieved from https://www.ncbi.nlm.nih.gov/assembly/GCA_022496805.1/), *Hypnum curvifolium* (data source see above) vs. *Niphotrichum japonicum*[119] (data retrieved https://doi.org/10.6084/m9.figshare.23573514.v5), *Hypnum curvifolium* (data source see above) vs. *Pohlia nutans* (data source see above), *Ceratodon purpureus* GG1[23] (data source: https://phytozome-next.jgi.doe.gov/info/CpurpureusGG1_v1_1) vs *Pohlia nutans* (data source see above) and the seed plant genome pairs *A. thaliana* (genome retrieved from https://phytozome-next.jgi.doe.gov/info/Athaliana_TAIR10) vs. *Oryza sativa* (genome retrieved from https://phytozome-next.jgi.doe.gov/info/Osativa_v7_0) and *A. thaliana* vs. *Cinnamomum kanehirae* (genome retrieved https://phytozome-next.jgi.doe.gov/info/Ckanehirae_v3).

Gene set comparison of the *F. hygrometrica* and *P. patens* genomes

We created orthogroups, groups of genes descended from a common ancestor, using 38 plant proteomes including six species of green and streptophyte algae, 12 bryophytes, and 20 vascular plants representing all major lineages of land plants (Supplementary Data 9 and 10). OrthoFinder v.2.5.2[120] analysis was run using default parameters. We obtained the species tree from orthogroup gene trees using the algorithm provided in OrthoFinder v.2.5.2. The species tree was converted into a time tree (ultrametric tree with branch length in time units) using the Environment for Tree Exploration (ETE) toolkit v3[121]. To infer gene family evolution on the branches leading to *F. hygrometrica* and *P. patens* from their common ancestor we used COUNT[122]. COUNT applies a phylogenetic birth-and-death model to reconstruct the evolution of gene numbers in gene families along a phylogenetic tree taking into account the processes of gene loss, gene gain and duplication. All three parameters vary by the edges of the phylogenetic tree and by family, the latter according to a discretized gamma distribution. We used likelihood optimization to obtain numerical estimates for these parameters. To do so we performed model optimization in a model hierarchy starting with the simplest model and changing only one parameter at a time and retained parameters that led to the most significant improvement of the likelihood value. The final model included variable duplication rates, and edge length as well as duplication and loss rates varied according to a discrete gamma distribution with two parameters (length_k = 2_dupl_k = 2_loss_k = 2). The gain was modeled with a simple gamma distribution as its inclusion did not influence the likelihood value significantly. Using the model parameter estimates, we calculated posterior probabilities for gene family expansion/contraction as well as gain/loss for each family running the posteriors module of COUNT (Supplementary Data 11).

Functional annotation of predicted genes

To obtain the functional annotation for the *F. hygrometrica* genes, we used two approaches. In particular, we assigned GO annotations to the gene models of *F. hygrometrica* using the eggNOG-mapper v2[123] and InterProScan v5[124] (Supplementary Data 18–21).

Reporting summary

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

Data availability

Raw DNA and RNA sequencing data used in this publication were submitted to NCBI Short Read Archive (SRA) under the BioProject ID PRJNA816911 (SRA submission SUB11197892) and to the European Nucleotide Archive (ENA) under study accession number PRJEB36328 for the Zurich accession, respectively. For the UConn strain, raw DNA and RNA data were deposited at the CNGB data center (https://db.cngb.org/) under the project number CNP0002793. Genome assembly files, their annotations, all supplementary data files, and figure source data have been deposited on figshare under the following link https://doi.org/10.6084/m9.figshare.19720216.

Code availability

Custom scripts used to estimate the absolute age of insertion of LTRs in the genome are provided at figshare https://doi.org/10.6084/m9.figshare.28369454.v1. No other custom code was used.

Received: 6 June 2024; Accepted: 14 February 2025;
Published online: 28 February 2025

References

1. Marks, R. A., Hotaling, S., Frandsen, P. B. & VanBuren, R. Representation and participation across 20 years of plant genome sequencing. *Nat. Plants* **7**, 1571–1578 (2021).
2. Kress, W. J. et al. Green plant genomes: what we know in an era of rapidly expanding opportunities. *Proc. Natl. Acad. Sci. USA* **119**, 1–9 (2022).
3. Chen, F. et al. The sequenced angiosperm genomes and genome databases. *Front. Plant Sci.* **9**, 1–14 (2018).
4. Wendel, J. F., Jackson, S. A., Meyers, B. C. & Wing, R. A. Evolution of plant genome architecture. *Genome Biol.* **17**, 1–14 (2016).
5. Szövényi, P., Gunadi, A. & Li, F.-W. Charting the genomic landscape of seed-free plants. *Nat. Plants* **7**, 554–565 (2021).
6. Hufford, M. B. et al. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, 655–662 (2021).
7. Li, H. et al. Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. *Nat. Commun.* **13**, 682 (2022).
8. Zhao, Q. et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* **50**, 278–284 (2018).
9. Wang, W. et al. Genomic variation in 3010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
10. Gordon, S. P. et al. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun.* **8**, 2184 (2017).
11. Golicz, A. A. et al. The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat. Commun.* **7**, 13390 (2016).
12. Li, Y. et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045–1052 (2014).
13. Hoopes, G. et al. Phased, chromosome-scale genome assemblies of tetraploid potato reveal a complex genome, transcriptome, and predicted proteome landscape underpinning genetic diversity. *Mol. Plant* **15**, 520–536 (2022).
14. Qiao, Q. et al. Evolutionary history and pan-genome dynamics of strawberry (*Fragaria* spp.). *Proc. Natl. Acad. Sci. USA* **118**, e2105431118 (2021).
15. Lovell, J. T. et al. Four chromosome scale genomes and a pan-genome annotation to accelerate pecan tree breeding. *Nat. Commun.* **12**, 4125 (2021).
16. Qin, P. et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542–3558.e16 (2021).
17. Tao, Y. et al. Extensive variation within the pan-genome of cultivated and wild sorghum. *Nat. Plants* **7**, 766–773 (2021).
18. Li, J. et al. Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. *Genome Biol.* **22**, 119 (2021).

19. Hübner, S. et al. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat. Plants* **5**, 54–62 (2019).
20. Gao, L. et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* **51**, 1044–1051 (2019).
21. Rensing, S. Why we need more non-seed plant models. *New Phytol.* **216**, 355–360 (2017).
22. Wickell, D. et al. Underwater CAM photosynthesis elucidated by *Isoetes* genome. *Nat. Commun.* **12**, 6348 (2021).
23. Carey, S. B. et al. Gene-rich UV sex chromosomes harbor conserved regulators of sexual development. *Sci. Adv*. **7**, eabh2488 (2021).
24. Silva, A. T. et al. To dry perchance to live: Insights from the genome of the desiccation-tolerant biocrust moss *Syntrichia caninervis*. *Plant J.* **105**, 1339–1356 (2021).
25. Li, C. et al. Extraordinary preservation of gene collinearity over three hundred million years revealed in homosporous lycophytes. *Proc. Natl. Acad. Sci*. USA **121**, e2312607121 (2024).
26. Marchant, D. B. et al. Dynamic genome evolution in a model fern. *Nat. Plants* **8**, 1038–1051 (2022).
27. Pellicer, J., Hidalgo, O., Dodsworth, S. & Leitch, I. Genome size diversity and its impact on the evolution of land plants. *Genes* **9**, 88 (2018).
28. Lang, D. et al. The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. *Plant J.* **93**, 515–533 (2018).
29. Yu, J. et al. Chromosome-level genome assemblies of two hypnales (mosses) reveal high intergeneric synteny. *Genome Biol. Evol.* **14**, 1–6 (2022).
30. Buck, W. R., Shaw, A. J. & Goffinet, B. Morphology, anatomy, and classification of the Bryophyta. In *Bryophyte Biology* (ed. Shaw, A. J.) 55–138 (Cambridge University Press, 2008).
31. Li, X., Wu, H. X., Dillon, S. K. & Southerton, S. G. Generation and analysis of expressed sequence tags from six developing xylem libraries in Pinus radiata D. Don. *BMC Genomics* **10**, 41 (2009).
32. Bi, G. et al. Near telomere-to-telomere genome of the model plant *Physcomitrium patens*. *Nat. Plants* **10**, 327–343 (2024).
33. Rensing, S. A., Goffinet, B., Meyberg, R., Wu, S. Z. & Bezanilla, M. The moss *Physcomitrium* (*Physcomitrella*) *patens*: a model organism for non-seed plants. *Plant Cell* **32**, 1361–1376 (2020).
34. Medina, R. et al. Phylogenomic delineation of Physcomitrium (Bryophyta: Funariaceae) based on targeted sequencing of nuclear exons and their flanking regions rejects the retention of Physcomitrella, Physcomitridium and Aphanorrhegma. *J. Syst. Evol.* **57**, 404–417 (2019).
35. Liu, Y., Budke, J. M. & Goffinet, B. Phylogenetic inference rejects sporophyte based classification of the Funariaceae (Bryophyta): rapid radiation suggests rampant homoplasy in sporophyte evolution. *Mol. Phylogenet. Evol.* **62**, 130–145 (2012).
36. Fernandez-Pozo, N., Haas, F. B., Gould, S. B. & Rensing, S. A. An overview of bioinformatics, genomics and transcriptomics resources for bryophytes. *J. Exp. Bot.* **54**, 1–54 (2022).
37. Medina, R. et al. Evolutionary dynamism in bryophytes: Phylogenomic inferences confirm rapid radiation in the moss family Funariaceae. *Mol. Phylogenet. Evol.* **120**, 240–247 (2018).
38. Beike, A. K. et al. Molecular evidence for convergent evolution and allopolyploid speciation within the *Physcomitrium-Physcomitrella* species complex. *BMC Evol. Biol.* **14**, 158 (2014).
39. Fife, A. A generic revision of the Funariaceae (Bryophyta: Musci). Part I. *J. Hattori Bot. Lab.* **58**, 149–196 (1985).
40. Budke, J. M. & Goffinet, B. Comparative cuticle development reveals taller sporophytes are covered by thicker calyptra cuticles in mosses. *Front. Plant Sci.* **7**, 1–11 (2016).
41. Patel, N., Medina, R., Johnson, M. & Goffinet, B. Karyotypic diversity and cryptic speciation: have we vastly underestimated moss species diversity? *Bryophyt. Divers. Evol*. **43**, 150–163 (2021).
42. McDaniel, S. F. et al. The speciation history of the *Physcomitrium–Physcomitrella* species complex. *Evolution* **64**, 217–231 (2010).
43. Ostendorf, A. K. et al. Polyploidization within the Funariaceae—a key principle behind speciation, sporophyte reduction and the high variance of spore diameters? *Bryophyte Diversity Evol.* **43**, 164–179 (2021).
44. Schween, G., Gorr, G., Hohe, A. & Reski, R. Unique tissue-specific cell cycle in *Physcomitrella. Plant Biology* **5**, 50–58 (2003).
45. Fritsch, R. Index to bryophyte chromosome counts. *Bryophytorum Bibliotheca* **40**, J. Cramer Berlin, Stuttgart (1991).
46. Rice, A. et al. The Chromosome Counts Database (CCDB)—a community resource of plant chromosome numbers. *New Phytol.* **206**, 19–26 (2015).
47. Kapila, S. Cytological observations on some west Himalayan mosses. *Proc. Indian Sci. Congr. Assoc.* **79**, 115–116 (1992).
48. Kapila, S. & Kumar, S. S. Cytological observations on some West Himalayan mosses. *Hikobia* **12**, 215–219 (1997).
49. Kirbis, A. et al. Transcriptional landscapes of divergent sporophyte development in two mosses, Physcomitrium (Physcomitrella) patens and *Funaria hygrometrica*. *Front. Plant Sci.* **11**, 747 (2020).
50. Bechteler, J. et al. Comprehensive phylogenomic time tree of bryophytes reveals deep relationships and uncovers gene incongruences in the last 500 million years of diversification. *Am. J. Bot.* **110**, e16249 (2023).
51. Pellicer, J. & Leitch, I. J. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytol.* **226**, 301–305 (2020).
52. Szövényi, P. et al. De novo assembly and comparative analysis of the *Ceratodon purpureus* transcriptome. *Mol. Ecol. Resour.* **15**, 203–215 (2015).
53. Rahmatpour, N., Perera, N. V., Singh, V., Wegrzyn, J. L. & Goffinet, B. High gene space divergence contrasts with frozen vegetative architecture in the moss family Funariaceae. *Mol. Phylogenet. Evol.* **154**, 106965 (2021).
54. Gao, B., Chen, M. X., Li, X. S., Zhang, D. Y. & Wood, A. J. Ancestral gene duplications in mosses characterized by integrated phylogenomic analyses. *J. Syst. Evol.* **60**, 144–159 (2022).
55. Beric, A. et al. Comparative phylogenetics of repetitive elements in a diverse order of flowering plants (Brassicales). *G3 (Bethesda)*. **11**, jkab140 (2021).
56. Elliott, T. A. & Gregory, T. R. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370**, 20140331 (2015).
57. Lisch, D. How important are transposons for plant evolution? *Nat. Rev. Genet.* **14**, 49–61 (2013).
58. Lysak, M. A., Koch, M. A., Beaulieu, J. M., Meister, A. & Leitch, I. J. The dynamic ups and downs of genome size evolution in Brassicaceae. *Mol. Biol. Evol.* **26**, 85–98 (2009).
59. Linde, A.-M. et al. Genome evolution in plants: complex thalloid liverworts (Marchantiopsida). *Genome Biol. Evol.* **15**, 1–10 (2023).
60. Zhang, G. et al. The reference genome of *Miscanthus floridulus* illuminates the evolution of Saccharinae. *Nat. plants* **7**, 608–618 (2021).
61. Shi, J. et al. Chromosome conformation capture resolved near complete genome assembly of broomcorn millet. *Nat. Commun.* **10**, 464 (2019).
62. Dvorak, J. et al. Structural variation and rates of genome evolution in the grass family seen through comparison of sequences of genomes greatly differing in size. *Plant J.* **95**, 487–503 (2018).
63. Massa, A. N. et al. Gene space dynamics during the evolution of *Aegilops tauschii*, *Brachypodium distachyon*, *Oryza sativa*, and *Sorghum bicolor* genomes. *Mol. Biol. Evol.* **28**, 2537–2547 (2011).

64. Huang, K. & Rieseberg, L. H. Frequency, origins, and evolutionary role of chromosomal inversions in plants. *Front. Plant Sci.* **11**, 1–13 (2020).
65. Yang, Z. et al. Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. *Nat. Commun.* **10**, 2989 (2019).
66. Yates, T. B. et al. The ancient Salicoid genome duplication event: a platform for reconstruction of de novo gene evolution in *Populus trichocarpa*. *Genome Biol. Evol.* **13**, 1–14 (2021).
67. Bengtsson, B. O. & Cronberg, N. The effective size of bryophyte populations. *J. Theor. Biol.* **258**, 121–126 (2009).
68. Szövényi, P. et al. Selfing in haploid plants and efficacy of selection: codon usage bias in the model moss *Physcomitrella patens*. *Genome Biol. Evol*. **9**, 1528–1546 (2017).
69. Quatrano, R. S., McDaniel, S. F., Khandelwal, A., Perroud, P. F. & Cove, D. J. *Physcomitrella patens*: mosses enter the genomic age. *Curr. Opin. Plant Biol.* **10**, 182–189 (2007).
70. Yu, J.-G. et al. The first homosporous lycophyte genome revealed the association between the recent dynamic accumulation of LTR-RTs and genome size variation. *Plant Mol. Biol.* **112**, 325–340 (2023).
71. Fang, Y. et al. The genome of homosporous maidenhair fern sheds light on the euphyllophyte evolution and defences. *Nat. Plants* **8**, 1024–1037 (2022).
72. Banks, J. A. et al. The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. *Science* **332**, 960–963 (2011).
73. VanBuren, R. et al. Extreme haplotype variation in the desiccation-tolerant clubmoss *Selaginella lepidophylla*. *Nat. Commun.* **9**, 13 (2018).
74. Huang, X. et al. The flying spider-monkey tree fern genome provides insights into fern evolution and arborescence. *Nat. Plants* **8**, 500–512 (2022).
75. Hisanaga, T. et al. The ancestral chromatin landscape of land plants. *New Phytol.* **240**, 2085–2101 (2023).
76. Bowman, J. L. et al. Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell* **171**, 287–304.e15 (2017).
77. Li, F.-W. et al. *Anthoceros* genomes illuminate the origin of land plants and the unique biology of hornworts. *Nat. Plants* **6**, 259–272 (2020).
78. Liu, S. et al. The Antarctic moss *Pohlia nutans* genome provides insights into the evolution of bryophytes and the adaptation to extreme terrestrial habitats. *Front. Plant Sci.* **13**, 920138 (2022).
79. Diop, S. I. et al. A pseudomolecule-scale genome assembly of the liverwort *Marchantia polymorpha*. *Plant J.* **101**, 1378–1396 (2020).
80. Kent, T. V., Uzunović, J. & Wright, S. I. Coevolution between transposable elements and recombination. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, 20160458 (2017).
81. Porebski, S., Bailey, L. G. & Baum, B. R. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol. Biol. Report.* **15**, 8–15 (1997).
82. Xie, T. et al. De novo plant genome assembly based on chromatin interactions: a case study of *Arabidopsis thaliana*. *Mol. Plant* **8**, 489–492 (2015).
83. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
84. Putnam, N. H. et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
85. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
86. Dudchenko, O. et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
87. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
88. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2023).
89. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
90. English, A. C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **7**, e47768 (2012).
91. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
92. Laetsch, D. R. & Blaxter, M. L. BlobTools: interrogation of genome assemblies. *F1000Research* **6**, 1287 (2017).
93. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
94. Smit, A., Hubley, R. & Green, P. RepeatMasker open-4.0. http://www.repeatmasker.org (2015).
95. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
96. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinforma.* **3**, 1–11 (2021).
97. She, R. et al. genBlastG: using BLAST searches to build homologous gene models. *Bioinformatics* **27**, 2141–2143 (2011).
98. Perlman, P. S. & Boeke, J. D. Ring around the retroelement. *Science* **303**, 182–184 (2004).
99. Edgar, R. C. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinforma.* **5**, 113 (2004).
100. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980).
101. Cock, P. J. A. et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
102. Rensing, S. et al. An ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*. *BMC Evol. Biol.* **7**, 130 (2007).
103. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016).
104. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
105. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
106. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
107. Kovaka, S. et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
108. Cabanettes, F. & Klopp, C. D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **2018**, e4958 (2018).
109. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
110. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
111. Nattestad, M. & Schatz, M. C. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021–3023 (2016).
112. Tang, H. et al. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18**, 1944–1954 (2008).

113. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
114. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
115. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
116. Bandi, V. & Gutwin, C. Interactive exploration of genomic conversation. In *Proceedings of Graphics Interface 2020* 74–83 (Canadian Human-Computer Communications Society, 2020).
117. Goodstein, D. M. et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, 1178–1186 (2012).
118. Wang, L. et al. The *Capparis spinosa var. herbacea* genome provides the first genomic instrument for a diversity and evolution study of the Capparaceae family. *Gigascience* **11**, 1–14 (2022).
119. Zhou, X. et al. Chromosome-level genome assembly of *Niphotrichum japonicum* provides new insights into heat stress responses in mosses. *Front. Plant Sci*. **14**, 1271357 (2023).
120. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
121. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
122. Csurös, M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**, 1910–1912 (2010).
123. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
124. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

Acknowledgements

This study was financially supported by grants of the Swiss National Science Foundation (grant nos. 131726, 160004, 184826 and 212509 to P.S.), a pilot and a PhD grant of the University Research Priority Program "Evolution in Action" of the University of Zurich to P.S., A.K. and L.W., a Georges and Antoine Claraz Foundation grant to A.K., M.W., L.W. and P.S., and a doctoral research grant (UZH Candoc, formerly Forschungskredit) to M.W. This project was also carried out in the framework of MAdLand (https://madland.science, DFG priority programme 2237, PS-1111/1 to P.S., RE 1697/19-1 and 20-1 to S.A.R.). N.R. and B.G. were supported by a grant from the US National Science Foundation (DEB-1146295 & 1753811). The study was also funded by the Shenzhen Urban Management Bureau Fund (202005) to YL. The authors thank Yang Peng at the Shenzhen Fairy Lake Botanical Garden for laboratory assistance. This work was supported by China National GeneBank (CNGB; https://www.cngb.org/). RR received funding from the German Research Foundation (DFG) under CRC-Transregio 141 (project B02) and Germany's Excellence Strategy (CIBSS – EXC-2189 – Project ID 390939984). The authors are grateful to Andrea Patrignani, Lucy Poveda, Weihong Qi, and Catharine Aquino for assistance in sequencing at the Functional Genomic Center Zurich (FGCZ). We are also thankful to the S3IT team and the ScienceCloud infrastructure at the University of Zurich for providing computational resources.

Author contributions

P.S. and Y.L. conceptualized the study. P.S., Y.L., S.D., B.G., J.L.W., R.R., N.R., M.W., and S.A.R. generated primary sequence data. E.M.T. carried out genome size measurements. P.S. and JY assembled the genomes. A.K., Y.L., J.Y., N.v.G., D.L., L.W., and P.S. carried out genome annotation. A.K., L.W., H.H., and P.S. carried out detailed comparative genomic analyses. P.S. and A.K. wrote the manuscript. All co-authors revised and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-07749-x.

Correspondence and requests for materials should be addressed to Péter Szövényi.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Matteo Dell'Acqua and David Favero.