nature genetics

Article

https://doi.org/10.1038/s41588-024-02071-4

The *Marchantia polymorpha* pangenome reveals ancient mechanisms of plant adaptation to the environment

Received: 27 October 2023

Accepted: 16 December 2024

Published online: 17 February 2025

Check for updates

Chloé Beaulieu[1,19], Cyril Libourel[1,17,19], Duchesse Lacourt Mbadinga Zamar[1], Karima El Mahboubi[1], David J. Hoey[2], George R. L. Greiff[2,18], Jean Keller[1], Camille Girou[1], Helene San Clemente[1], Issa Diop[3,4], Emilie Amblard[1], Baptiste Castel[1], Anthony Théron[5], Stéphane Cauet[5], Nathalie Rodde[5], Sabine Zachgo[6], Wiebke Halpape[7,8], Anja Meierhenrich[7,8], Bianca Laker[7,8], Andrea Bräutigam[7,8], The SLCU Outreach Consortium*, Peter Szovenyi[3,4], Shifeng Cheng[9], Yasuhiro Tanizawa[10], Simon Aziz[11], James H. Leebens-Mack[12], Jeremy Schmutz[13,14], Jenell Webber[13], Jane Grimwood[13], Christophe Jacquet[1], Christophe Dunand[1], Jessica M. Nelson[15], Fabrice Roux[16], Hervé Philippe[11], Sebastian Schornack[2], Maxime Bonhomme[1] ✉ & Pierre-Marc Delaux[1] ✉

Plant adaptation to terrestrial life started 450 million years ago and has played a major role in the evolution of life on Earth. The genetic mechanisms allowing this adaptation to a diversity of terrestrial constraints have been mostly studied by focusing on flowering plants. Here, we gathered a collection of 133 accessions of the model bryophyte *Marchantia polymorpha* and studied its intraspecific diversity using selection signature analyses, a genome–environment association study and a pangenome. We identified adaptive features, such as peroxidases or nucleotide-binding and leucine-rich repeats (NLRs), also observed in flowering plants, likely inherited from the first land plants. The *M. polymorpha* pangenome also harbors lineage-specific accessory genes absent from seed plants. We conclude that different land plant lineages still share many elements from the genetic toolkit evolved by their most recent common ancestor to adapt to the terrestrial habitat, refined by lineage-specific polymorphisms and gene family evolution.

Following their transition from an aquatic to a terrestrial habitat, plants had already diversified in two main lineages 400 million years ago[1–6]: the vascular plants (tracheophytes), which include all flowering plants, and the nonvascular plants (bryophytes), which encompass hornworts, mosses and liverworts. The long-lasting colonization of terrestrial habitats by bryophytes and tracheophytes is the result of an initial burst of innovations[3–6] followed by continuous adaptations to new environments, leading to the spread of plants in most ecosystems on Earth.

The genetics of population adaptation to environmental constraints has been explored by genome–environment association (GEA) studies in crops and in nondomesticated model angiosperms[7]. These led to the discovery of genetic variants ranging from SNPs[8] to structural and gene presence–absence variations when species pangenomes were used[9,10], demonstrating the diversity of the genetic bases of adaptation in angiosperms. These approaches facilitated the identification of loci to be used as breeding or genome-editing targets for crop improvement even between species[11–13].

A full list of affiliations appears at the end of the paper. *A list of authors and their affiliations appears at the end of the paper.
✉e-mail: maxime.bonhomme@univ-tlse3.fr; pierre-marc.delaux@cnrs.fr

https://doi.org/10.1038/s41588-024-02071-4

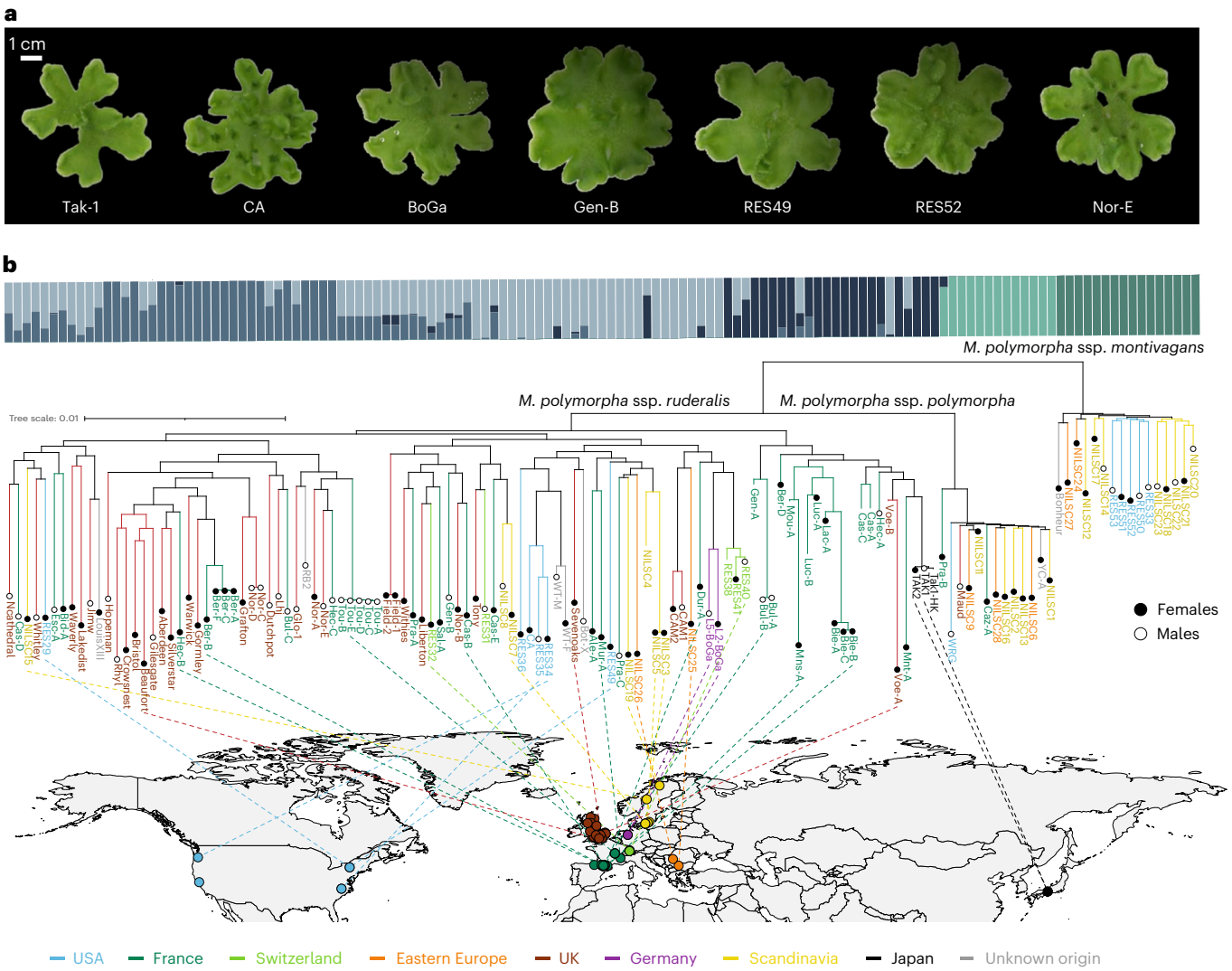


Fig. 1 | *M. polymorpha* collection, phylogenetic tree and population structure. a, Pictures of representative *M. polymorpha* accessions. Tak-1, CA, BoGa, Gen-B, RES49 and Nor-E belong to *M. polymorpha* ssp. *ruderalis*. RES52 belongs to *M. polymorpha* ssp. *montivagans*. **b**, Phylogenetic tree of the *M. polymorpha* subspecies complex, comprising 104 *M. polymorpha* ssp. *ruderalis*, 16 *M. polymorpha* ssp. *montivagans* and 13 *M. polymorpha* ssp. *polymorpha* accessions. The phylogenetic tree was computed using a dataset of 107,744 pruned SNPs. Dots on the world map indicate the sampling location, and colors represent country or geographic regions of origin. Some sampling locations are linked to their tree branch to give examples of the discrepancy between geographical origin and phylogenetic position. The colored branches on the tree also correspond to the geographic origin. Bar plots on the top of the tree indicate the assignment probability of each *M. polymorpha* accession to each of the five main genetic groups (the three subspecies and three populations in the *ruderalis* subspecies, colored in shades of blue and green). These clusters were identified by the Bayesian clustering algorithm implemented in the fastSTRUCTURE software. Map data from CIA World DataBank II (https://www.evl.uic.edu/pape/data/WDB/).

The response of individual plants to environmental factors has been studied in tracheophytes and in bryophytes, leading to the discovery of transcriptomic and physiological responses to abiotic stresses[14,15] and to pathogenic or mutualistic biotic interactions[4,16–18]. Combined with phylogenomics and evo–devo, these approaches revealed that part of the molecular mechanisms regulating these responses to environmental factors are highly conserved across land plants. Although lineage-specific mechanisms occur, this suggests that studying bryophytes can inform work in angiosperms, including for applied purposes.

We hypothesize that, beyond the characterized molecular mechanisms conserved in tracheophytes and bryophytes studied by evo–devo approaches, gene families might have been involved in the adaptation to environmental factors in the two groups. However, adaptive traits evolved by bryophytes over the last 450 million years[19] as well as their potential conservation in other lineages have not been explored.

Among the bryophytes, the liverwort *M. polymorpha* has become a model[20] thanks to its fast life cycle, vegetative reproduction via gemmae and available genetic tools such as efficient transformation[20], CRISPR–Cas9 (ref. 21) and collections of DNA modules[22]. Comparing various aspects of its biology with model tracheophytes helped to infer the nature of the most recent common ancestor of land plants using evo–devo approaches[23]. Beyond evo–devo, genetics conducted in *M. polymorpha* revealed[24] mechanisms defining cell polarity in eukaryotes[25], pathways controlling rhizoid development[26] or a conserved pathway for rapid auxin responses[27]. *M. polymorpha* thus represents the ideal model to study the genetics of adaptation in bryophytes.

https://doi.org/10.1038/s41588-024-02071-4

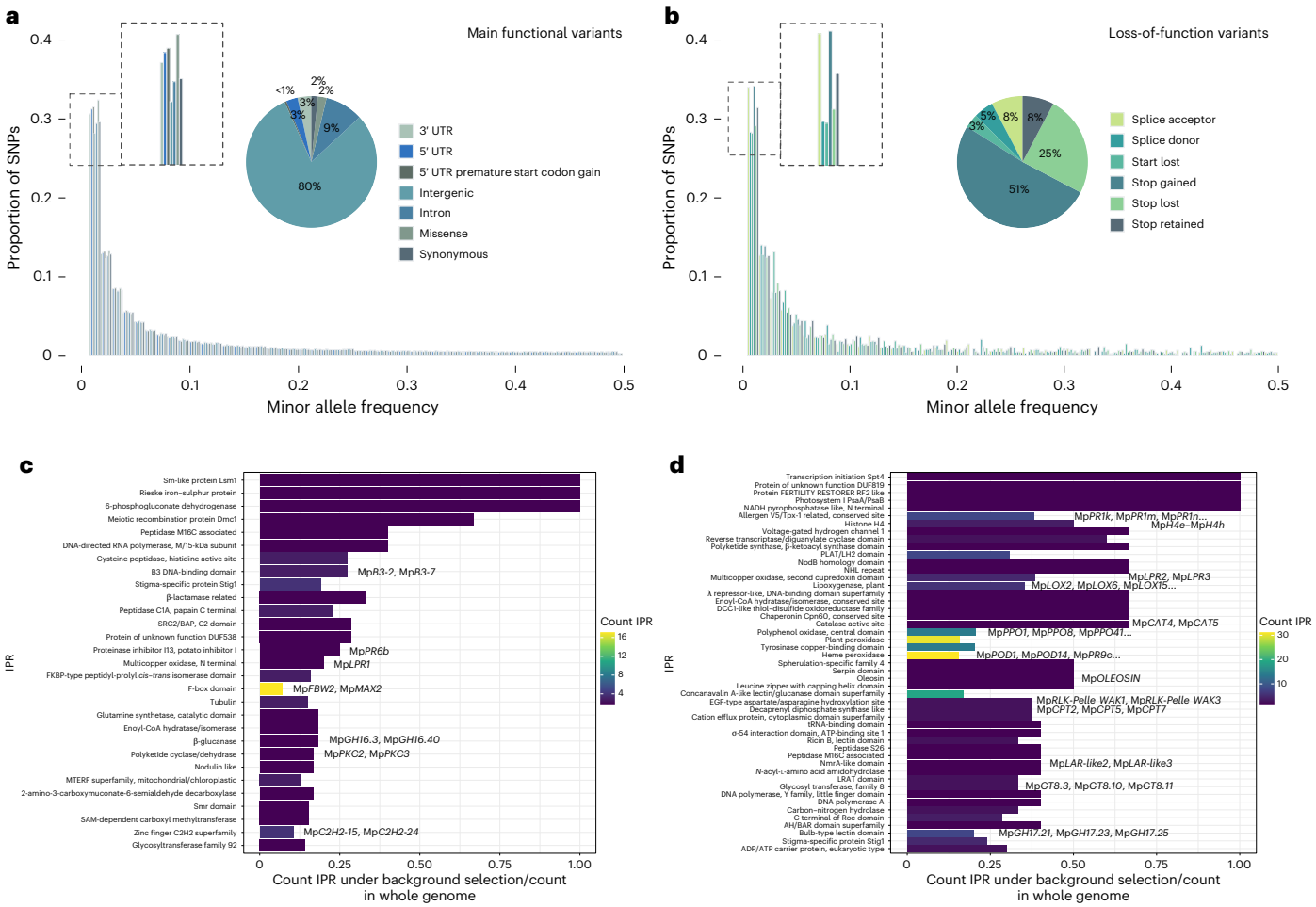


Fig. 2 | Genome-wide distribution and allele frequency spectrum of the main functional variant categories and functional enrichment analysis of genes putatively under background selection or balancing selection in a collection of 104 accessions of *M. polymorpha* ssp. *ruderalis*. In **a**, a total of 5,439,674 SNPs (4,363,099 intergenic, 509,964 in introns, 115,218 missense, 88,163 synonymous, 183,858 3′ UTR, 154,821 5′ UTR and 24,551 5′ UTR premature start codon gain, that is, a variant in a 5′ UTR region producing a three-base sequence that can be a start codon) were used to construct the allele frequency spectrum. In **b**, a total of 7,823 SNPs (4,010 stop gained, 1,948 stop lost, 608 stop retained, 606 splice acceptor, 385 splice donor and 266 start lost) were used to construct the allele frequency spectrum. All SNPs in **a**,**b** are distributed on the eight autosomes and sexual chromosome U or V. In **c**, functional enrichment analysis of a list of 777 genes putatively under background selection (top 10% of genes by Zheng's *E* negative value). In **d**, functional enrichment analysis of a list of 1,814 genes putatively under balancing selection (top 10% of genes by Tajima's *D* positive values). In **c**,**d**, IPR terms are ordered from top to bottom according to decreasing significance (false discovery rate *q* value < 0.05), together with some associated known gene names.

Here, we collected 133 accessions of the liverwort *M. polymorpha*, covering the three subspecies (ssp. *ruderalis*, ssp. *montivagans* and ssp. *polymorpha*), and resequenced their nuclear genomes. This dataset was augmented with two long-read genomes for one European and one North American accession. This sequencing effort provides the genomic and resource needed to explore intraspecific diversity in a bryophyte. In addition, a pangenome for *M. polymorpha* was built. Performing GEA on climate variables, identifying the signatures of selection across the genome and investigating gene presence–absence variation revealed the strategies deployed by *M. polymorpha* to adapt to diverse environments. Our findings suggest that environmental adaptation in *M. polymorpha* is partly achieved by mechanisms similar to those described in angiosperms, thus likely ancestral in land plants, and by lineage-specific polymorphisms, some originating from horizontal gene transfer.

Results

M. polymorpha intraspecific diversity resources

M. polymorpha encompasses three subspecies, ssp. *ruderalis*, ssp. *montivagans* and ssp. *polymorpha*. Separation of the three subspecies is supported by phylogenetic analysis[28] and infertility between some intersubspecific crosses. *M. polymorpha* is distributed mostly in the northern hemisphere, excluding the arctic regions, with ecological specificities for each subspecies. *M. polymorpha* ssp. *polymorpha* grows in natural riparian sites. *M. polymorpha* ssp. *montivagans* is usually found at higher elevation. *M. polymorpha* ssp. *ruderalis* is a rapid colonizer adapted to human-disturbed habitats and is the most common of the three subspecies[29]. Two accessions from Japan, Tak-1 and Tak-2, have been widely used by the community. The Tak-1 nuclear genome, composed of eight autosomes and completed by the two sexual chromosomes from Tak-1 and Tak-2, was sequenced[24]. To maximize the intraspecific diversity, we collected 133 additional accessions from four main geographic areas: south of France, the UK, a transect between Switzerland and Sweden and the USA (Fig. 1a and Supplementary Table 1). The collection sites ranged from sidewalks to soil near ponds and from sea level to an elevation of 1,124 m. Most of the UK accessions were collected in the framework of a citizen science project[30]. Gemmae were collected from a single individual for each accession and propagated in vitro following sterilization. This collection includes 103 accessions from *M. polymorpha* ssp. *ruderalis*, 16 from *M. polymorpha* ssp. *montivagans* and 14 from *M. polymorpha* ssp. *polymorpha*.

https://doi.org/10.1038/s41588-024-02071-4

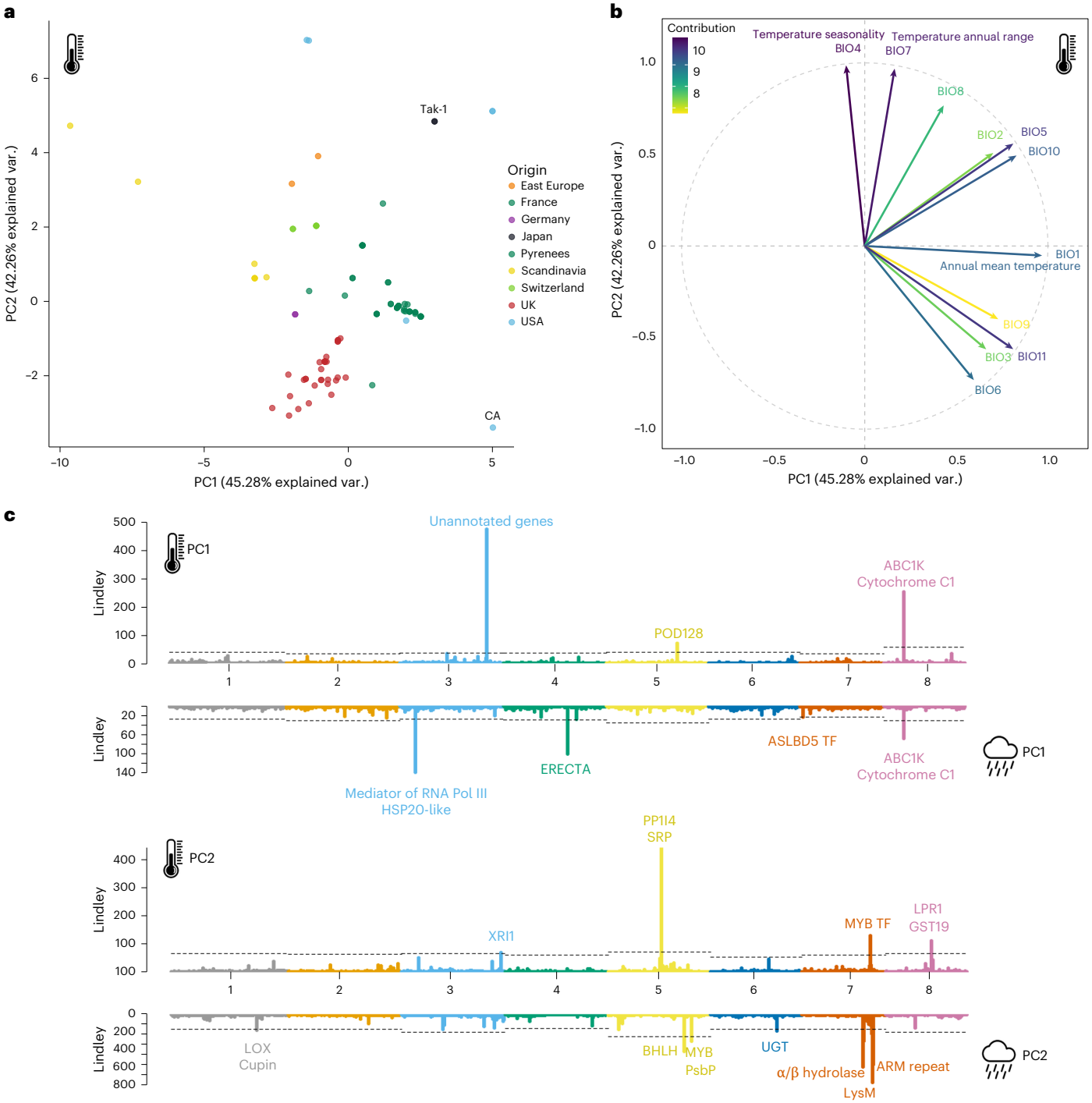


Fig. 3 | Identification of loci associated with climatic conditions by gene–environment association analyses in *M. polymorpha* ssp. *ruderalis*. **a**, Principal-component (PC) plot of 96 *M. polymorpha* ssp. *ruderalis* individuals based on temperature-linked bioclimatic variables (BIO1 to BIO11). Var., variance. **b**, Contributions of temperature-linked bioclimatic variables (BIO1 to BIO11) to the two first principal components. **c**, Miami plots (mirror comparison of two Manhattan plots) of the GEA results on the first components of distinct PCA for two categories of bioclimatic variables from the WorldClim 2 data: temperature-linked (BIO1 to BIO11) and precipitation-linked (BIO12 to BIO19) variables. These Miami plots result from a classical genome-wide association analysis performed with GEMMA, followed by use of the local score technique on SNP *P* values (bilateral Wald test) to amplify the signal between SNPs in LD. The result of this process is a Lindley value that has been used instead of a *P* value to plot the *y* values of the Manhattan plots. The dashed lines represent significance thresholds for each chromosome (resampling thresholds from the local score method). TF, transcription factor.

All accessions were sequenced by Illumina with a mean genome coverage of 110× (Supplementary Table 1). Scaffolds were de novo assembled ($N_{50average}$ = 49.8 kbp; Supplementary Table 2) and annotated, leading to an average of 19,705 genes predicted per accession. The average benchmarking universal single-copy orthologs (BUSCO) completeness reached 88.5% (Supplementary Table 2). In addition, the *M. polymorpha* ssp. *ruderalis* accessions CA (USA) and BoGa crossline 5 (BoGa-L5) (Osnabrück, Germany) were sequenced using long-read technologies (N_{50} = 5.991 Mb and L_{50} = 12 for CA and N_{50} = 26.28 Mbp and L_{50} = 4 for BoGa-L5). The CA genome assembly and gene annotations for the

133 accessions are available on MarpolBase (https://marchantia.info). The BoGa genome assembly is available at https://doi.org/10.4119/unibi/2991996. As *M. polymorpha* is a dioicous species, the sex of the collected accessions was determined using PCR markers on individuals in the in vitro collection and bioinformatic analyses of the sequenced genomes. Among the 133 accessions, 80 appeared to be females and 44 were males. For 11 accessions, the results of the sexing analyses are uncertain. Two accessions (NILSC14 and Mns-A) displayed inconsistent sexing between the in vivo and in silico collections.

We provide here a genomic and physical collection covering the intraspecific diversity of a bryophyte.

M. polymorpha population structure

To characterize *M. polymorpha* intraspecific diversity, reads from all accessions were mapped on the reference Tak-1 genome (V6.1r2, which includes the female U chromosome from Tak-2), and SNPs were extracted. We identified 12,281,497 autosomal SNPs in the collection including all three *M. polymorpha* subspecies. Among these SNPs, 5,307,425 (~1 SNP every 40 bp) segregated in *M. polymorpha* ssp. *ruderalis*. In addition, 51,607 SNPs (~1 SNP every 88 bp) and 12,314 SNPs (~1 SNP every 613 bp) segregated on the U and V chromosomes of 61 females and 34 males from this subspecies, respectively. The SNP data mapped on the Tak-1 assembly can be explored on MarpolBase (https://marchantia.info/).

We reconstructed the phylogeny of the accessions using a dataset of 107,744 selected SNPs and performed Bayesian clustering population structure analyses (using a dataset of 20,146 SNPs; Methods). Although the exact relationship within each subspecies is likely not precisely captured, these analyses resolved three main clades corresponding to the three *M. polymorpha* subspecies. In addition, we observed three main genetic groups within the subspecies *ruderalis* (Fig. 1b). The accession Pra-B, from the subspecies *polymorpha*, displayed mixed subspecies ancestry (Fig. 1b). Although a denser sampling effort in the UK and France increased the phylogenetic clustering in these regions, we observed a weak correlation between genetic and geographical distances among accessions within the *M. polymorpha* ssp. *ruderalis* clade (Mantel statistic $r = 0.087$, $P = 0.074$), suggesting substantial gene flow and recombination events across its broad geographical range and extending previous findings at a regional scale (Southern Ontario, Canada) in *M. polymorpha*[31]. Additional identity by descent (IBD) analyses revealed contrasted patterns of genetic structure across autosomes and U and V chromosomes, suggesting differential dispersal ability between sexes in ssp. *ruderalis*, and distance-based analyses confirmed the accession Pra-B as a recent intersubspecific hybrid and identified multiple historical introgressions in the genome of *M. polymorpha* subspecies (Supplementary Note 1).

These results support *M. polymorpha* as a subspecies complex[28] undergoing gene flow within subspecies throughout a broad geographical range, as shown in *M. polymorpha* ssp. *ruderalis*, but also between subspecies as the result of historical introgression events.

Selection in the genome of *M. polymorpha ruderalis*

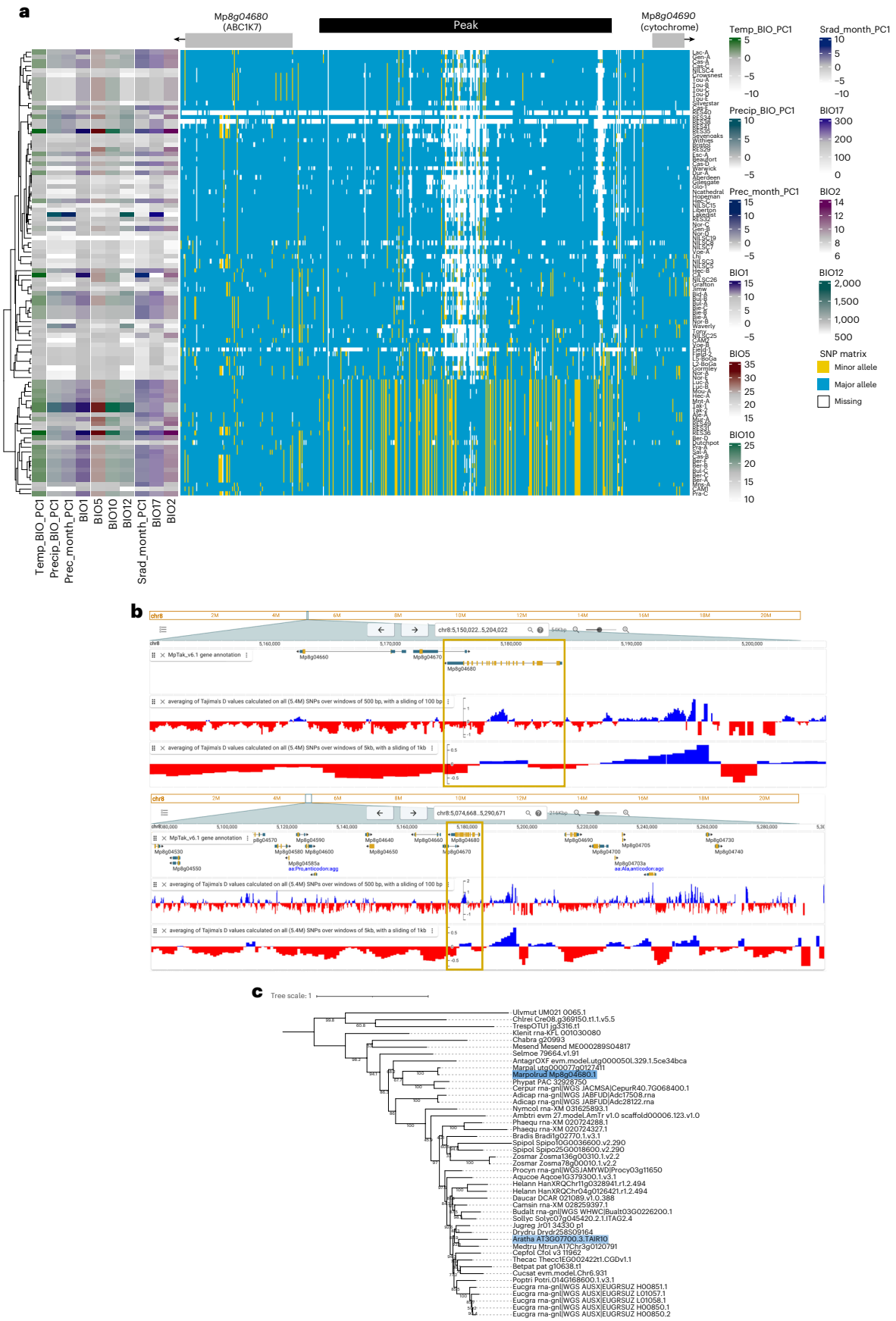
Using the SNP data for *M. polymorpha* ssp. *ruderalis*, we determined the genome-wide population-scaled mutation rate (Watterson's θ estimator, θ_w[32]) to be 0.00539, which is similar to that of angiosperm species with comparable sample sizes[33,34]. A sliding window scan of Tajima's *D*, Wu's *H* and Zheng's *E* was carried out on the whole genome (available at https://marchantia.info/) as well as on gene sequences. The average values of gene-based Tajima's *D* and Zheng's *E* statistics were negative ($D = -0.356$, $E = -0.317$), indicating a genome-wide excess of low-frequency variants (Supplementary Fig. 3). This signature suggests demographic expansion, with low population substructure in *M. polymorpha* ssp. *ruderalis*, in agreement with phylogenetic and population structure analyses.

Eighty percent of SNPs were located in intergenic regions. In genic regions, SNPs were similarly frequent in intronic (9%) and in non-intronic (11%) regions (Fig. 2a). The genome-wide allele frequency spectrum was L shaped as expected under neutral evolution. However, we observed an excess of rare missense, 5′ untranslated region (UTR), 3′ UTR and gain of premature start codon in 5′ UTR SNPs, relative to intergenic, intronic or synonymous SNPs, most probably reflecting stronger purifying selection on the former, compared, for instance, to intergenic variants (Fig. 2a). Loss-of-function SNPs predominantly led to stop codon gain or loss (51% or 25%) (Fig. 2b). The allele frequency spectrum of such SNPs was also L shaped, with an excess of rare stop gains and splice acceptor SNPs, reflecting strong purifying selection against variants leading to molecular phenotypes associated with truncated proteins and retained introns (Fig. 2b).

To identify genes under selective pressures in *M. polymorpha* ssp. *ruderalis*, we computed Tajima's *D, H* and *E* statistics (Methods). Of 18,920 genes, 3,965 genes exhibited the most pronounced selection signatures (777 under background selection, 1,374 under soft or hard selective sweep and 1,814 under balancing selection; Methods and Supplementary Table 5). Using functional enrichment analysis, we found that genes with pronounced signature of background selection in *M. polymorpha* ssp. *ruderalis* were enriched in various molecular functions (Fig. 2c and Supplementary Table 6), among them, the ones encoding the DNA repair proteins DNA meiotic recombinase 1 and RADiation sensitive protein 51 that are essential for meiotic recombination and highly conserved in eukaryotes, or cytoskeleton components. Functional enrichment analysis of genes under balancing selection revealed a wider array of functions involving highly polymorphic genes,

Fig. 4 | An atypical kinase linked to the response of *M. polymorpha* to temperature and precipitation variation. a, Haplotype block illustration of the genomic region on chromosome 8 associated with both temperature-linked and precipitation-linked variables in *M. polymorpha* ssp. *ruderalis* and containing two genes: one coding for an ABC1 atypical kinase (Mp8g04680) and one for a cytochrome C1 protein (Mp8g04690). The block on the left side of the figure represents bioclimatic variable values at the sampling site of each accession that led to the identification of the ABC1 cytochrome C1 region via GEA. The main matrix represents the allelic status of the SNPs in this region for each accession. The clustered matrix was generated with the ComplexHeatmap R package. Precip, precipitation; temp, temperature; srad, solar radiation; prec, precipitation. **b**, Close-up view of Tajima's *D* sliding window analysis in the genomic region surrounding the *ABC1K* gene of *M. polymorpha* ssp. *ruderalis* (*Mp8g04680*) on the genome browser (https://marchantia.info/variants/). This view highlights the specificity of balancing selection (high Tajima's *D* values) observed in the coding sequence of *Mp8g04680* and in its promoter region compared to the local genomic region (50- and 200-kb windows at the top and bottom, respectively). The gold boxes represent the genomic location of *Mp8g04680*. M, million. **c**, Phylogenetic tree of the orthologs of the *ABC1K* gene of *M. polymorpha* ssp. *ruderalis* illustrating direct orthology with the *ABC1K7* gene in *A. thaliana*. The names of the different species from which the *ABC1K* genes originate are specified with a six-letter code (Ulvmut, *Ulva mutabilis*; Chlrei, *Chlamydomonas reinhardtii*; TrespOTU1, *Trebouxia* spOTU1; Klenit, *Klebsormidium nitens*; Chabra, *Chara braunii*; Mesend, *Mesotaenium endlicherianum*; Selmoe, *Selaginella moellendorffii*; AntagrOXF, *Anthoceros agrestis* cv. *oxford*; Marpal, *M. paleacea*; Marpolrud, *M. polymorpha* ssp. *ruderalis*; Phypat, *Physcomitrium patens*; Cerpur, *Ceratodon purpureus*; Adicap, *Adiantum capillus-veneris*; Nymcol, *Nymphaea colorata*; Ambtri, *Amborella trichopoda*; Phaequ, *Phalaenopsis equestris*; Bradis, *Brachypodium distachyon*; Spipol, *Spirodela polyrhiza*; Zosmar, *Zostera marina*; Procyn, *Protea cynaroides*; Aqucoe, *Aquilegia coerulea*; Helann, *Helianthus annuus*; Daucar, *Daucus carota*; Camsin, *Camellia sinensis*; Budalt, *Buddleja alternifolia*; Sollyc, *Solanum lycopersicum*; Jugreg, *Juglans regia*; Drydru, *Dryas drummondii*; Aratha, *A. thaliana*; Medtru, *M. truncatula*; Cepfol, *Cephalotus follicularis*; Thecac, *Theobroma cacao*; Betpat, *Beta patula*; Cucsat, *Cucumis sativus*; Poptri, *Populus trichocarpa*; Eucgra, *Eucalyptus grandis*).

https://doi.org/10.1038/s41588-024-02071-4



https://doi.org/10.1038/s41588-024-02071-4

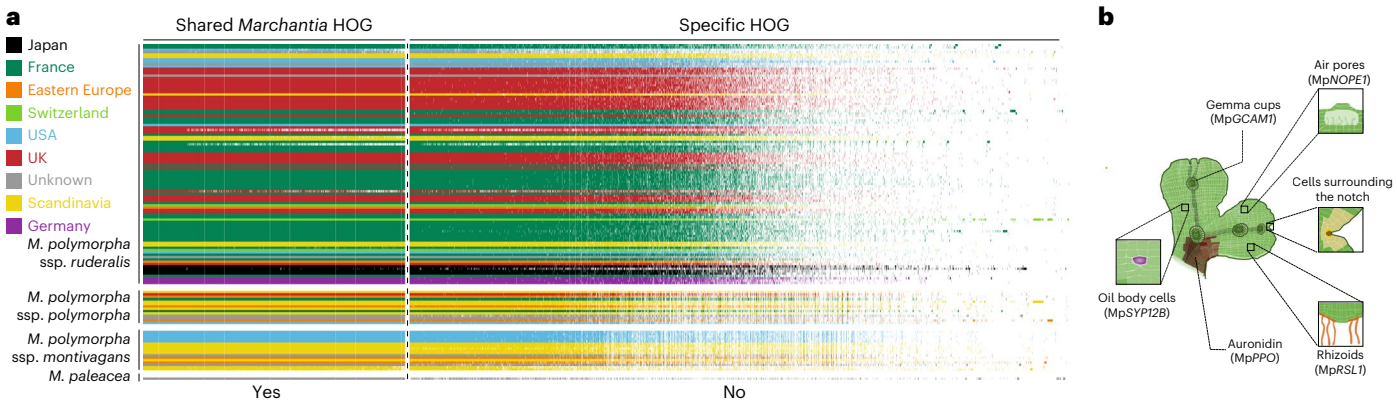


Fig. 5 | Landscape of gene presence–absence variation in the *Marchantia* genus pangenome. a, HOGs are sorted by their occurrence, with the genes shared by all the *Marchantia* species on the left (separated from the rest by a dashed line) and the very rare genes on the right part of the matrix. The accessions (in rows) are sorted according to their position in the phylogenetic tree of the accessions generated by the software OrthoFinder to infer these HOGs. Colors represent the geographic origin of each accession. **b**, Genetic features and traits shared by all species in the *Marchantia* genus.

including those encoding receptor-like kinase–Pelle and lectins, known for their function in mediating plant–environment interactions. Some genes involved in fundamental molecular functions also seemed to be subjected to balancing selection, such as the four histone H4 genes *H4e*, *H4f*, *H4g* and *H4h* (Fig. 2d and Supplementary Table 6). Finally, genes putatively under selective sweep showed enrichment in a wide variety of functions, among them genes encoding for MADS box transcription factors or glycoside hydrolases from family 16 (Supplementary Fig. 4 and Supplementary Table 6).

To identify gene families that might be under a similar selection regime in both angiosperms and bryophytes, we computed the same statistics for two model angiosperms with available population genomic SNP datasets, *Arabidopsis thaliana* and *Medicago truncatula* (Supplementary Table 7). We compared the statistics of *M. polymorpha* ssp. *ruderalis* to those of *A. thaliana* and *M. truncatula* using orthogroups, a proxy for gene families, which were filtered to identify those containing genes under a particular selection regime (background, balancing or sweep). The number of overlapping orthogroups for each selection regime was significantly higher than random expectation. We detected 40 orthogroups, encompassing genes under strong background selection in all species, including genes encoding the cellulose synthase MpCSLD10, members of the histone fold Hap3–NF-YB family[35] or the receptor-like kinase of the CrRLKL1L-1 family MpFERONIA or MpTHESEUS crucial for *Marchantia* and angiosperm development (Supplementary Table 8)[36–38]. In addition, we detected 86 orthogroups containing genes under balancing selection in all species (Supplementary Table 8), such as those encoding peroxidases, polyphenol oxidases and PR5 (pathogenesis-related) proteins as well as lipoxygenases or the terpenoid synthases MpTPS1 and MpTPS4 (Fig. 2d and Supplementary Table 8). Two evolutionary scenarios may explain the shared selection patterns between gene families in bryophytes and in tracheophytes. First, these genes families may have maintained genes under balancing selection, irrespective of lineage-specific expansions, since their most recent common ancestor. Alternatively, the selection signatures have evolved independently in *M. polymorpha* and the two angiosperms in a convergent manner. In both cases, these gene families can be considered tunable nodes repeatedly recruited for adaptation across the land plants. Exploring intraspecific diversity in other plant lineages will allow us to test these two hypotheses.

To conclude, analyzing the selection signatures in *M. polymorpha* ssp. *ruderalis* (1) revealed that purifying selection is pervasive on functional variants at the genome level, as probably the consequence of the predominant haploid phase and (2) identified highly conserved and highly polymorphic genes and gene families either specific to *M. polymorpha* or showing similar intraspecific diversity in two angiosperms.

GEA in *M. polymorpha* ssp. *ruderalis*

The autosomal genome-wide average recombination rate was estimated at 1.5×10^{-9} crossing-over per base pair per meiosis ($\pm 0.2 \times 10^{-9}$). We identified a genome-wide average one-half linkage disequilibrium (LD) decay of 3.6 kbp. Considering the low genetic structure and the shape of LD observed in our sampling, a genetic mapping approach can be thus implemented through genome-wide association approaches (Supplementary Fig. 5). We performed GEA on 96 accessions using 19 bioclimatic variables related to precipitation and temperature as well as data on monthly average precipitation, solar radiation, water vapor pressure and altitude (see the list of all detected genomic regions in Supplementary Table 9). To mitigate the correlations among bioclimatic variables, we first ran two principal-component analyses (PCA) on the bioclimatic variables related to (1) temperature and (2) precipitation. The first three principal components of temperature- and precipitation-related variables, explaining 95% and 98.4% of the total variance, respectively, were retained to describe the main results of the GEA. A clear geographical structuring of temperature variation is illustrated by principal-component plots of *M. polymorpha* ssp. *ruderalis* individuals (Fig. 3a) and contributions of bioclimatic variables to these principal components (Fig. 3b).

We identified genomic regions and candidate genes robustly associated with temperature and precipitation variation across the range of *M. polymorpha* ssp. *ruderalis* (Fig. 3c). Among the identified loci was a region on chromosome 8 coding for a cytochrome C1 protein (Mp8g04690) and an ABC1 atypical kinase (Mp8g04680). In this locus, the minor haplotype, present in 25% of the accessions (24 of 96), is associated with high values of temperature and precipitation (Fig. 4a). The *ABC1K* gene in this locus is in the top 16% of genes under balancing selection, and Tajima's *D* sliding window analysis indicates the specificity of this selection signature to the coding sequence but also to the promoter region of the gene, compared to the local genomic region (Fig. 4b). ABC1K is orthologous to the ABC1K7 protein in *A. thaliana* (Fig. 4c), which acts in concert with ABC1K8 to regulate the response to the stress hormone abscisic acid moderating the oxidative stress response[39]. Abscisic acid has been proven to have a role in desiccation tolerance in *M. polymorpha*, which could explain the link between the *ABC1K* gene and fluctuations in temperature and precipitation. Other loci detected are detailed in Supplementary Note 2, such as the one encoding the peroxidase MpPOD128 (Mp5g17500).

Comparisons of the 198 candidate genes with RNA-sequencing (RNA-seq) studies on the response of *M. polymorpha* ssp. *ruderalis* to the pathogen *Phytophthora palmivora*[40] and to various abiotic stresses[14] allowed the identification of 97 genes differentially regulated in at least one tested condition (Supplementary Table 10). Among them, Mp*NBS-LRR11* (Mp4g08790) was found differentially expressed in response to abiotic stresses and pathogen infection, and the associated locus was detected in GEA for bioclimatic variables linked to hot and humid conditions, often linked to higher pathogen pressure[41,42]. This gene encoding a coiled-coil NOD-like-receptor (NLR) may thus play a role in the adaptation of *M. polymorpha* to pathogens.

Many *M. polymorpha* ssp. *ruderalis* genes identified via GEA are from families important for abiotic and biotic stress responses in angiosperms (for example, peroxidases, leucine-rich repeat receptor-like kinases, NLRs). This suggests that plant adaptation relies on ancestral mechanisms, enhanced by lineage-specific duplications that diversify paralogs across plant lineages.

Pangenomic variations in *M. polymorpha*

Intraspecific diversity can be characterized using SNPs and by focusing on gene content variation between accessions. To identify presence–absence variation, 128 Illumina-based and long-read genomes of *M. polymorpha* ssp. *ruderalis* and reference genomes of *M. polymorpha* ssp. *montivagans* and ssp. *polymorpha*[28] and *Marchantia paleacea*[43] were used to generate *a Marchantia* genus pangenome, although limited to only two *Marchantia* species. This gene-centered pangenome represents gene repertoires across genomes, excluding structural variations not detectable with the available data (Methods; for a detailed version of this section, see Supplementary Note 3).

To explore the distribution of the gene space annotated for each accession, we clustered the predicted proteomes in hierarchical orthogroups (HOGs) (Supplementary Table 11). These HOGs represent groups of orthologs, which are either specific to a few accessions or shared by most of them (Fig. 5a). Of the 25,648 total HOGs, 3,542 were specific to one of the subspecies (Supplementary Table 12), while 7,292 HOGs were shared by accessions from all three subspecies and *M. paleacea* (Supplementary Table 13). These 'shared *Marchantia* HOGs' include genes present across land plants for conserved functions, such as the formation of epidermal structures (that is, *RSL1*). Cross-referencing the list of 'shared *Marchantia* HOGs' with single-cell RNA-seq data[44] revealed significant enrichment in genes expressed in the 'cells surrounding the notch' cluster (Supplementary Table 13), aligning with the role of these cell populations in structuring *Marchantia*'s development. In addition, genes thus far only functionally characterized in *M. polymorpha* Tak-1 clustered in the 'shared *Marchantia* HOGs', indicating that the molecular mechanisms controlling traits such as gemma cup formation (*MpGCAM1*) are likely conserved in the *Marchantia* genus beyond Tak-1 (Supplementary Table 13 and Fig. 5b).

The *M. polymorpha* ssp. *ruderalis* pangenome was obtained using gene predictions from 100 *ruderalis* accessions, three *ruderalis* long-read genomes and two outgroup genomes (ssp. *polymorpha* and ssp. *montivagans)*. The specific OrthoFinder analysis carried out on this sampling identified 28,143 HOGs representing *M. polymorpha* ssp. *ruderalis* gene content (Supplementary Table 14 and Fig. 6a). These HOGs were categorized as core (containing genes in nearly all accessions, tolerating six lower-quality ones), accessory (HOG containing genes from more than four accessions but not core) and cloud genome (HOG containing genes from four accessions or less) (Supplementary Table 15 and Methods). Core HOGs represent 35.2% of the total pangenome HOG content, while accessory HOGs account for 49.2% (Fig. 6b). The cloud HOGs should be taken with caution because they are enriched in genes from long-read genomes and therefore may be absent from other accessions for technical reasons (Supplementary Fig. 6). The sex ratio in all HOGs was shown to be 1.8 female for one male (similar to the sex ratio of the whole collection, $r^2 = 0.98$) (Supplementary Fig. 7 and Supplementary Table 15); therefore, the sex of the accessions should not impact the orthogroup categories.

Cross-referencing core and accessory genes with scRNA-seq data revealed that the expression cluster corresponding to 'cells surrounding the notch' was the most significantly enriched in core *M. polymorpha* ssp. *ruderalis* genes (Supplementary Table 15), similar to the *Marchantia* pangenome. The expression clusters with a significant proportion of accessory genes were the ones linked to gemma and oil body cells. Among the 283 genes of the gemma cell cluster, 83 were accessory and mainly linked to specialized metabolism (for example, chalcone synthases, cytochrome P450). For the oil body cell cluster, 52 of the 177 assigned genes were accessory, many of which were also involved in specialized metabolism (for example, microbial terpene synthase-like (MTPSL) and polyphenol oxidase genes). Copy number variation had already been observed in *M. polymorpha*'s MTPSL between two accessions[45], corroborating enrichment in the accessory genome.

Core and accessory HOGs in the pangenome were cross-referenced with Tak-1 RNA-seq data. Conditions with significantly more core or accessory genes deregulated were scrutinized (Supplementary Table 16). Most stress conditions showed downregulation of significantly more core genes than accessory genes (Supplementary Fig. 9). On the contrary, an equal number of stress conditions significantly upregulate more accessory genes or more core genes. (Supplementary Fig. 9). This suggests that core genes are more often downregulated during stress and that specific environmental constraints can trigger either core or accessory genome upregulation.

We performed an InterPro (IPR) domain enrichment on core and accessory compartments of the *M. polymorpha* ssp. *ruderalis* pangenome. As expected from other pangenomic studies[46,47], the core genome was enriched in domains involved in housekeeping functions (Supplementary Table 17). The enrichment on the accessory genome included many functions linked to multiple stress responses (peroxidases and isoprenoid synthases) or more specifically to biotic (NLRs, chitinases) or abiotic (aquaporins) stresses (Fig. 6c and Supplementary Table 17).

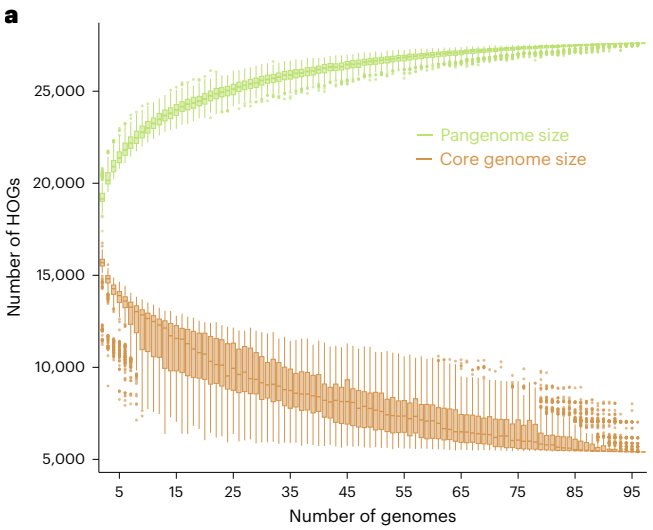
The *M. polymorpha* ssp. *ruderalis* pangenome highlights the accessory genome's role in the stress response, extending findings from flowering plants to a distant lineage and possibly to the entire diversity of land plants.

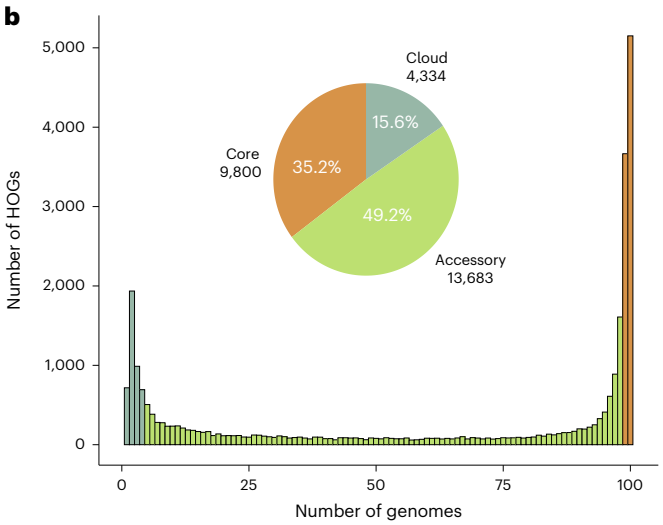
Horizontal gene transfer aided plant terrestrial adaptation

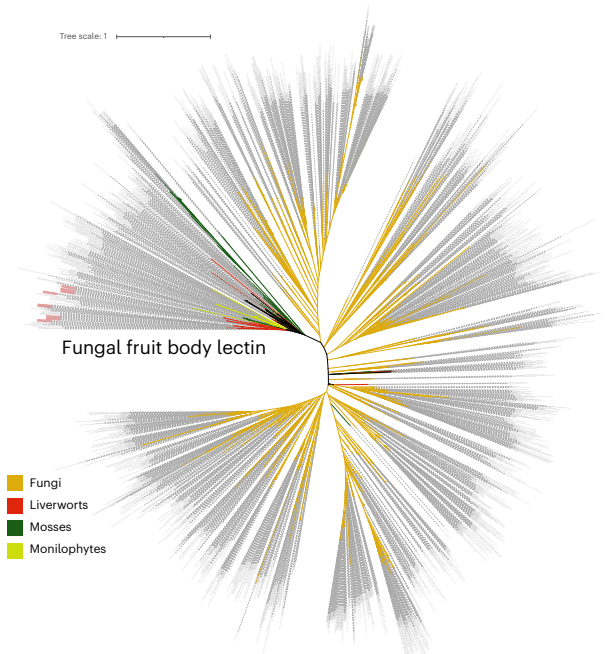
The IPR domain analysis identified one atypical domain enriched in the accessory genome compartment, the fungal fruit body lectin. Fungal fruit body lectins have been reported in *M. polymorpha* ssp. *ruderalis* Tak-1 and some other bryophytes and described as fungal horizontal gene transfer (HGT)[48]. To verify their distribution in land plants, we searched for orthologs of the nine fungal fruit body lectin genes present in the reference genome of *M. polymorpha* ssp. *ruderalis* Tak-1 in a database containing genomes from all the main clades of land

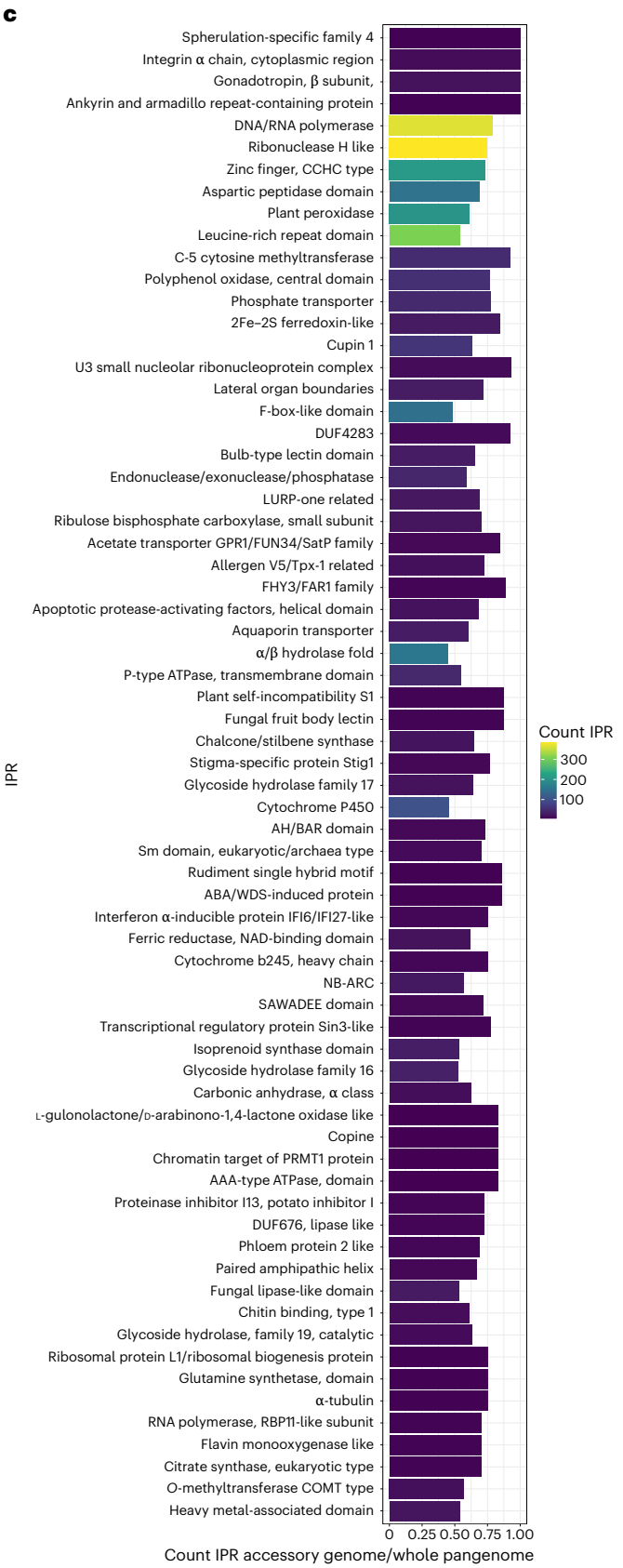
Fig. 6 | The pangenome of *M. polymorpha* ssp. *ruderalis*. a, Increase in pangenome size and decrease in core genome size when adding accessions to the *M. polymorpha* ssp. *ruderalis* pangenome (500 random samplings for the box plot, for which the center is the median of values, bounds of the boxes are the first and third quantiles, that is, the 25th and 75th percentiles, and the whiskers are distant from the quantiles by 1.5 times the interquantile range). This saturation plot indicates a closed pangenome in *M. polymorpha* ssp. *ruderalis*. **b**, Composition of the pangenome regarding the HOG categories (core, accessory and rare). The histogram shows the number of HOGs with different frequencies of presence in accessions, and the pie chart shows the proportion of HOGs in each category. **c**, Bar plot representing IPR terms significantly enriched in the accessory genome. Redundant IPR terms and IPR terms linked with transposable elements and viral proteins were discarded to improve readability. **d**, Phylogeny of the orthologs of fungal lectin genes present in *M. polymorpha* ssp. *ruderalis*.

https://doi.org/10.1038/s41588-024-02071-4









plants (Supplementary Table 18) and transcriptomes[49]. We detected orthologs in liverworts and mosses but also in ferns (two genomes and 21 transcriptomes) (Fig. 6d). These lectins had not been identified in ferns or any tracheophytes. Indeed, these genes are missing from the two previously screened genomes of aquatic ferns[50] and may represent a loss following the reversion to an aquatic environment[51]. This presence of fungal fruit body lectins in terrestrial ferns gives a whole other perspective on the gain of these domains in plants: they might have been transferred from a fungus to the ancestor of land plants and then lost in three clades (hornworts, lycophytes and seed plants) (Fig. 6d). Seven of the nine fungal fruit body lectin genes present in Tak-1 are differentially expressed in stress conditions, with most being upregulated in response to drought (Supplementary Table 16).

Our results suggest that a gene family originating from fungal HGT before the diversification of land plants contributes to drought adaptation in *M. polymorpha* ssp. *ruderalis*, while it was independently lost in other lineages, illustrating the different adaptive landscape in land plants after 450 million years of evolution on land.

Discussion

By gathering a large collection of wild collected accessions of *M. polymorpha*, we aimed to provide a resource for the community, to identify genetic polymorphism correlating with environmental variables and to discover gene families sharing patterns of selection across land plants. Three approaches were combined: population genomic-based selection analyses, genotype–environment association analyses and a gene-centered pangenome. Comparing the genes identified through these approaches with knowledge from angiosperms, we identified commonalities in the genetic strategy for adaptation to the environment but also *Marchantia*-specific innovations. Two gene families stood out when comparing adaptation in *M. polymorpha* and angiosperms: those encoding class III peroxidases and NLRs. Class III peroxidases are extremely diversified in angiosperms (75 in *A. thaliana*) as well as in *M. polymorpha* (190 genes), which is surprising considering that liverworts did not experience whole-genome duplication like angiosperms[52]. In *M. polymorpha*, this diversity results from clade-specific tandem duplications, as illustrated by *POD128* identified in the GEA (Fig. 3c). In addition to these independent gene family expansions, peroxidase-encoding genes are among those with the strongest balancing selection signature in *M. polymorpha*, *A. thaliana* and *M. truncatula*, correlating with their known roles in development and adaptation in angiosperms[53]. NLRs were the second class of proteins with diversity and polymorphism patterns in *M. polymorpha* reminiscent of the one found in angiosperms, clearly associating NLRs with adaptation. In the future, conducting genome-wide association studies in response to pathogens would reveal the genetic basis of resistance in *M. polymorpha* and could confirm the link between NLRs and immunity, well known in angiosperms. Peroxidases and NLRs associated with adaptation in both bryophytes and angiosperms may reflect ancestral traits retained since the most recent common ancestor of land plants.

Lineage-specific genes and adaptation signatures were also detected. These may have coevolved with morphological adaptation. In that respect, the importance of the gemma and oil body cell clusters in adaptation processes in *M. polymorpha* has been underscored by cross-referencing scRNA-seq data with GEA candidates and the accessory genome. Oil bodies are the biosynthesis and storage sites of many specialized metabolites and contribute to resistance to abiotic stress in *M. polymorpha*[54,55]. By examining oil body clusters, which mainly express genes related to specialized metabolite synthesis, a clear pattern between the core and accessory genome arises. Core genes (such as *IDS1*, encoding an upstream enzyme in terpenoid biosynthesis) are more likely to be upstream genes, whereas accessory genes, such as those encoding terpene synthases, likely contribute to the final steps of the pathways. The transfer of such terminal genes might represent an option for the production of antimicrobial compounds in crops. A fern-specific insecticidal protein-coding gene originating from HGT, Tma12, was successfully transferred to cotton, conferring resistance to whiteflies[56,57]. Such a case of HGT expanding the abilities to interact with the microbiota is reminiscent of fungal fruit body lectin proteins, identified in the *M. polymorpha* accessory genome and originating from HGT (Fig. 6d), previously linked with resistance to insects[48,58]. We traced their acquisition to the last common ancestor of land plants. It is therefore possible that other horizontally transferred genes present in bryophytes and lost in angiosperms were acquired by their common ancestor and that the role of HGT in the terrestrialization process is even more important than suspected[58].

Our study provides a genomic and physical resource to explore intraspecific diversity in the model liverwort *M. polymorpha*, providing a reservoir of genetic novelties with potential for use in crop improvement. We identified genetic polymorphisms correlating with environmental variables in this species, revealing similarities with angiosperms together with lineage-specific novelties. *M. polymorpha* also retains adaptation-related genes inherited from the most recent common ancestor of land plants but lost in angiosperms.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-02071-4.

References

1. Delaux, P.-M. & Schornack, S. Plant evolution driven by interactions with symbiotic and pathogenic microbes. *Science* **371**, eaba6605 (2021).
2. Morris, J. L. et al. The timescale of early land plant evolution. *Proc. Natl Acad. Sci. USA* **115**, E2274–E2283 (2018).
3. Beerling, D. *The Emerald Planet: How Plants Changed Earth's History* https://doi.org/10.1093/oso/9780192806024.001.0001 (Oxford University Press, 2007).
4. Rich, M. K. et al. Lipid exchanges drove the evolution of mutualism during plant terrestrialization. *Science* **372**, 864–868 (2021).
5. Clark, J. W. et al. The origin and evolution of stomata. *Curr. Biol.* **32**, R539–R553 (2022).
6. González-Valenzuela, L., Renard, J., Depège-Fargeix, N. & Ingram, G. The plant cuticle. *Curr. Biol.* **33**, R210–R214 (2023).
7. Lasky, J. R., Josephs, E. B. & Morris, G. P. Genotype–environment associations to reveal the molecular basis of environmental adaptation. *Plant Cell* **35**, 125–138 (2023).
8. Alseekh, S., Kostova, D., Bulut, M. & Fernie, A. R. Genome-wide association studies: assessing trait characteristics in model and crop plants. *Cell. Mol. Life Sci.* **78**, 5743–5754 (2021).
9. Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J. & Edwards, D. Plant pan-genomes are the new reference. *Nat. Plants* **6**, 914–920 (2020).
10. He, Q. et al. A graph-based genome and pan-genome variation of the model plant *Setaria*. *Nat. Genet.* **55**, 1232–1242 (2023).
11. Lin, X. et al. *Solanum americanum* genome-assisted discovery of immune receptors that detect potato late blight pathogen effectors. *Nat. Genet.* **55**, 1579–1588 (2023).
12. Lemmon, Z. H. et al. Rapid improvement of domestication traits in an orphan crop by genome editing. *Nat. Plants* **4**, 766–770 (2018).
13. Khan, A. W. et al. *Cicer* super-pangenome provides insights into species evolution and agronomic trait loci for crop improvement in chickpea. *Nat. Genet.* **56**, 1225–1234 (2024).
14. Tan, Q. W. et al. Cross-stress gene expression atlas of *Marchantia polymorpha* reveals the hierarchy and regulatory principles of abiotic stress responses. *Nat. Commun.* **14**, 986 (2023).

15. Wu, T.-Y. et al. Evolutionarily conserved hierarchical gene regulatory networks for plant salt stress response. *Nat. Plants* **7**, 787–799 (2021).
16. Adachi, H. et al. Jurassic NLR: conserved and dynamic evolutionary features of the atypically ancient immune receptor ZAR1. *Plant Cell* **35**, 3662–3685 (2023).
17. Ngou, B. P. M., Heal, R., Wyler, M., Schmid, M. W. & Jones, J. D. G. Concerted expansion and contraction of immune receptor gene repertoires in plant genomes. *Nat. Plants* **8**, 1146–1152 (2022).
18. Sucher, J. et al. Phylotranscriptomics of the Pentapetalae reveals frequent regulatory variation in plant local responses to the fungal pathogen *Sclerotinia sclerotiorum*. *Plant Cell* **32**, 1820–1844 (2020).
19. Song, B. et al. Plant genome resequencing and population genomics: current status and future prospects. *Mol. Plant* **16**, 1252–1268 (2023).
20. Bowman, J. L. et al. The renaissance and enlightenment of *Marchantia* as a model system. *Plant Cell* **34**, 3512–3542 (2022).
21. Sugano, S. S. et al. Efficient CRISPR/Cas9-based genome editing and its application to conditional genetic analysis in *Marchantia polymorpha*. *PLoS ONE* **13**, e0205117 (2018).
22. Romani, F. et al. The landscape of transcription factor promoter activity during vegetative development in *Marchantia*. *Plant Cell* **36**, 2140–2159 (2024).
23. Delaux, P.-M. et al. Reconstructing trait evolution in plant evo–devo studies. *Curr. Biol.* **29**, R1110–R1118 (2019).
24. Bowman, J. L. et al. Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell* **171**, 287–304 (2017).
25. Van Dop, M. et al. DIX domain polymerization drives assembly of plant cell polarity complexes. *Cell* **180**, 427–439 (2020).
26. Honkanen, S. et al. The mechanism forming the cell surface of tip-growing rooting cells is conserved among land plants. *Curr. Biol.* **27**, 3238–3224 (2017).
27. Kuhn, A. et al. RAF-like protein kinases mediate a deeply conserved, rapid auxin response. *Cell* **187**, 130–148 (2024).
28. Linde, A.-M., Sawangproh, W., Cronberg, N., Szövényi, P. & Lagercrantz, U. Evolutionary history of the *Marchantia polymorpha* complex. *Front. Plant Sci.* **11**, 829 (2020).
29. Shimamura, M. *Marchantia polymorpha*: taxonomy, phylogeny and morphology of a model system. *Plant Cell Physiol.* **57**, 230–256 (2016).
30. Hoey, D. J., Greiff, G. R. L., SLCU Outreach Consortium & Schornack, S. The Great British Liverwort Hunt — collecting wild accessions for molecular biology research while engaging the public. *Zenodo* https://doi.org/10.5281/ZENODO.10040685 (2023).
31. Sandler, G., Agrawal, A. F. & Wright, S. I. Population genomics of the facultatively sexual liverwort *Marchantia polymorpha*. *Genome Biol. Evol.* **15**, evad196 (2023).
32. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
33. Branca, A. et al. Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. *Proc. Natl Acad. Sci. USA* **108**, E864–E870 (2011).
34. Fischer, M. C. et al. Estimating genomic diversity and population differentiation — an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. *BMC Genomics* **18**, 69 (2017).
35. Bhattacharjee, B. & Hallan, V. NF-YB family transcription factors in *Arabidopsis*: structure, phylogeny, and expression analysis in biotic and abiotic stresses. *Front. Microbiol.* **13**, 1067427 (2023).
36. Galindo-Trigo, S., Gray, J. E. & Smith, L. M. Conserved roles of CrRLK1L receptor-like kinases in cell expansion and reproduction from algae to angiosperms. *Front. Plant Sci.* **7**, 1269 (2016).
37. Malivert, A. & Hamant, O. Why is FERONIA pleiotropic? *Nat. Plants* **9**, 1018–1025 (2023).
38. Mecchia, M. A. et al. The single *Marchantia polymorpha FERONIA* homolog reveals an ancestral role in regulating cellular expansion and integrity. *Development* **149**, dev200580 (2022).
39. Manara, A., DalCorso, G. & Furini, A. The role of the atypical kinases ABC1K7 and ABC1K8 in abscisic acid responses. *Front. Plant Sci.* **7**, 366 (2016).
40. Carella, P. et al. Conserved biochemical defenses underpin host responses to oomycete infection in an early-divergent land plant lineage. *Curr. Biol.* **29**, 2282–2294 (2019).
41. Singh, B. K. et al. Climate change impacts on plant pathogens, food security and paths forward. *Nat. Rev. Microbiol.* **21**, 640–656 (2023).
42. MacQueen, A. & Bergelson, J. Modulation of *R*-gene expression across environments. *J. Exp. Bot.* **67**, 2093–2105 (2016).
43. Radhakrishnan, G. V. et al. An ancestral signalling pathway is conserved in intracellular symbioses-forming plant lineages. *Nat. Plants* **6**, 280–289 (2020).
44. Wang, L. et al. The maturation and aging trajectory of *Marchantia polymorpha* at single-cell resolution. *Dev. Cell* **58**, 1429–1444 (2023).
45. Takizawa, R. et al. Fungal-type terpene synthases in *Marchantia polymorpha* are involved in sesquiterpene biosynthesis in oil body cells. *Plant Cell Physiol.* **62**, 528–537 (2021).
46. 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).
47. Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176 (2020).
48. Peumans, W. J. et al. The liverwort *Marchantia polymorpha* expresses orthologs of the fungal *Agaricus bisporus* agglutinin family. *Plant Physiol.* **144**, 637–647 (2007).
49. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
50. Van Holle, S. & Van Damme, E. J. M. Messages from the past: new insights in plant lectin evolution. *Front. Plant Sci.* **10**, 36 (2019).
51. Delaux, P.-M. et al. Comparative phylogenomics uncovers the impact of symbiotic associations on host genome evolution. *PLoS Genet.* **10**, e1004487 (2014).
52. Clark, J. W. & Donoghue, P. C. J. Whole-genome duplication and plant macroevolution. *Trends Plant Sci.* **23**, 933–945 (2018).
53. Kidwai, M., Ahmad, I. Z. & Chakrabarty, D. Class III peroxidase: an indispensable enzyme for biotic/abiotic stress tolerance and a potent candidate for crop improvement. *Plant Cell Rep.* **39**, 1381–1393 (2020).
54. Kanazawa, T. et al. The liverwort oil body is formed by redirection of the secretory pathway. *Nat. Commun.* **11**, 6152 (2020).
55. Romani, F. et al. Oil body formation in *Marchantia polymorpha* is controlled by MpC1HDZ and serves as a defense against arthropod herbivores. *Curr. Biol.* **30**, 2815–2828 (2020).
56. Li, F.-W. et al. Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nat. Plants* **4**, 460–472 (2018).
57. Shukla, A. K. et al. Expression of an insecticidal fern protein in cotton protects against whitefly. *Nat. Biotechnol.* **34**, 1046–1051 (2016).
58. Ma, J. et al. Major episodes of horizontal gene transfer drove the evolution of land plants. *Mol. Plant* **15**, 857–871 (2022).

[1]Laboratoire de Recherche en Sciences Végétales, Université de Toulouse, CNRS, UPS, Toulouse INP, Castanet-Tolosan, France. [2]Sainsbury Laboratory, University of Cambridge, Cambridge, UK. [3]Department of Systematic and Evolutionary Botany, University of Zurich, Zurich, Switzerland. [4]Zurich–Basel Plant Science Center, Zurich, Switzerland. [5]CNRGV French Plant Genomic Resource Center, INRAE, Castanet-Tolosan, France. [6]Division of Botany, School of Biology, Osnabrueck University, Osnabrueck, Germany. [7]Computational Biology, Faculty of Biology, Bielefeld University, Bielefeld, Germany. [8]CeBiTec, Bielefeld University, Bielefeld, Germany. [9]Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China. [10]Department of Informatics, National Institute of Genetics, Mishima, Japan. [11]Station d'Ecologie Théorique et Expérimentale de Moulis, UMR CNRS 5321, Moulis, France. [12]University of Georgia, Athens, GA, USA. [13]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA. [14]Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [15]Maastricht Science Programme, Maastricht University, Maastricht, the Netherlands. [16]Laboratoire des Interactions Plantes–Microbes–Environnement, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement, CNRS, Université de Toulouse, Castanet-Tolosan, France. [17]Present address: Unité de Recherche Physiologie, Pathologie et Génétique Végétales, INP PURPAN, Université de Toulouse, Toulouse, France. [18]Present address: University of Bristol, Bristol, UK. [19]These authors contributed equally: Chloé Beaulieu, Cyril Libourel. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: maxime.bonhomme@univ-tlse3.fr; pierre-marc.delaux@cnrs.fr

The SLCU Outreach Consortium

David J. Hoey[2], Edwige Moyroud[2], Alan Wanke[2], Alessandra Bonfanti[2], Stefano Gatti[2], Alexander Summers[2], Elisabeth Burmeister[2], Kathy Grube[2], Andreea Alexa[2], Nataliia Kuksa[2], Lauren Gardiner[2], Martin Balcerowicz[2], Jemma Salmon[2], Bryony Yates[2], Lucie Riglet[2] & Elena Salvi[2]

Methods

Plant material

Plants were collected directly in sampling pots and brought to the laboratory. After growing individual thalli from each accession in potting soil, gemmae were collected and sterilized as described by Delaux et al.[59]. Cultures were initiated from single gemma, and a single plant was conserved for each accession and grown on 0.5× B5 medium (G0209-0025, Duchefa Biochemie) as described by Rich et al.[4] and supplemented with 1% sucrose (200-301-B, Euromedex), with 16:8-h day:night cycles at 22 °C under fluorescent illumination with a light intensity of 100 µmol μm^{-2} s^{-1}. Note that, due to the COVID-19 pandemic, a total of 38 accessions, at the time being propagated for the first round of gemmae, were lost.

BoGa-L5 accession long-read sequencing

DNA extraction, library preparation and long-read sequencing of BoGa-L5. *M. polymorpha* ssp. *ruderalis* ecotype BoGa-L5 plants were raised on Gamborg B5 medium (Duchefa Biochemie) at room temperature with a day–night cycle of 16 h of light to 8 h of darkness. Male plants were collected at 44 d after germination, and high-molecular-weight DNA was extracted from 1 g of material according to an adapted CTAB DNA extraction protocol in which chloroform–isoamyl alcohol was replaced with dichloromethane (https://www.protocols.io/view/plant-dna-extraction-and-preparation-for-ont-seque-kxygxenmkv8j/v1). DNA quality was checked with the Invitrogen Qubit 4 fluorometer (Thermo Fisher Scientific). DNA was size selected with the short-read eliminator kit (PacBio), and sequencing libraries were prepared with the ligation sequencing gDNA kit (SQK-LSK109-XL, Oxford Nanopore Technologies (ONT)). The libraries were sequenced on a GridION platform using one R9.4.1 and one R10.0 flow cell, and bases were called with MinKNOW v.21.02.5 software (high-accuracy basecalling with Guppy 4.3.4) (ONT).

Short-read sequencing of BoGa-L5. A paired-end sequencing library was prepared using the extracted genomic DNA of BoGa and the TruSeq DNA Nano Kit according to the TruSeq DNA Sample Preparation v.2 Guide (Illumina). The library was sequenced in 2 × 150-bp paired-end mode on a NovaSeq 500 sequencer (Illumina). Raw reads were quality trimmed using Trimmomatic v.0.39 (ref. 60) with parameters 'LEADING:34 TRAILING:34 SLIDINGWINDOW:4:15 ILLUMINACLIP:2:34:15 MINLEN:100'.

Computation of a chromosome-scale BoGa genome assembly. BoGa long reads were assembled with Canu v.2.2 (ref. 61) (setting 'genomeSize=280m'). The genome assembly was polished with Racon v.1.4.20 (ref. 62) and minimap v.2.22-1101 (ref. 63) using the parameters '-m 8 -x -6 -g -8 -t 40', with medaka v.1.4.3 (https://github.com/nanoporetech/medaka) and with Pilon v.1.24 (ref. 64). Medaka polishing was run one round with the raw ONT reads from the R9.4.1 flow cell and the parameter 'r941_min_high_g360' and one round with reads from the R10.0 flow cell and the parameter 'r103_min_high_g360'. Pilon polishing was carried out three times using BWA-MEM v.0.7.17 (ref. 65) and the trimmed genomic short-read data of BoGa-L5. The polished contig sequences were anchored to pseudochromosomes based on the *Marchantia* reference genome sequence assembly Tak v.6.1 (ref. 24,66) with RagTag v.2.0.1 (ref. 67), allowing 100-bp gaps between sequences. For this, chromosome U, which represents the female sex chromosome, and unplaced scaffolds of the Tak assembly were disregarded and an artificial chromosome 0 was generated from unplaced sequences of the BoGa assembly. The assembly was repeat masked with RepeatModeler v.2.0.2 (ref. 68) with the parameter '-LTRStruct' and RepeatMasker v.4.1.2 (ref. 69) with the softmasking option enabled.

The completeness of the BoGa genome assembly was determined with the BUSCO tool v.5.4.4 (ref. 70) using the database Viridiplantae odb10 to 93.1%.

Chloroplast and mitochondrion contig sequences in the BoGa assembly were identified through a blastn search (package v.2.11.0+) against the *M. polymorpha* chloroplast (National Center for Biotechnology Information (NCBI) accession number NC_037507.1) and the mitochondrion genome sequence assembly (NCBI accession number NC_037508.1), resulting in one full sequence each. Overlaps were trimmed for the chloroplast contig with Berokka v.0.2 (https://github.com/tseemann/berokka) and for the mitochondrion contig through identification of overlapping ends with a blastn search against itself. The contigs were oriented according to the references used with blastn. The gene annotation of the BoGa chloroplast sequence was computed with GeSeq (v.2.03)[71], applying tRNAscan-SE (v.2.0.7), Chloe version 0.1.0, HMMER v.3.3.1 and BLAT[72]. The gene annotation of the BoGa mitochondrion genome sequence was calculated with GeSeq using tRNAscan-SE (v.2.0.7) and BLAT. The results were filtered, discarding annotations with HMMER score < 15, BLAT score < 20 and tRNAscan score < 20. As GeSeq outputs overlapping genes resulted from the different prediction programs, only one gene annotation was kept at a sequence region. If the overlapping genes had the same start and end position, the prediction of BLAT was prioritized, followed by HMMER, followed by Chloe and followed by tRNAscan. If the start or end position or both differed and the overlapping genes were named equally, the longest gene annotation was kept. The assembly and all associated data can be accessed at https://doi.org/10.4119/unibi/2982437.

CA accession long-read sequencing

Preparation of high-molecular-weight DNA for sequencing. DNA was isolated from dark-treated 3-week-old thalli using the QIAGEN Genomic-tips 500/G kit (10262) following the tissue extraction protocol. Briefly, 13.5 g of young leaf material was frozen and ground in liquid nitrogen with a mortar and a pestle. After 3 h of lysis at 50 °C and one centrifugation step (15,000*g*), the DNA was immobilized on the column. After several washing steps, DNA was eluted from the column and then desalted and concentrated by alcohol precipitation. DNA was resuspended in EB buffer. DNA quality and quantity were assessed respectively using the NanoDrop One spectrophotometer (Thermo Scientific) and the Qubit 3 fluorometer using the Qubit dsDNA BR assay (Invitrogen). DNA size was assessed using the Femto Pulse system (Agilent).

Preparation and sequencing of the HiFi PacBio library. A HiFi SMRTbell library was constructed using the SMRTbell Template Prep Kit 2.0 (Pacific Biosciences) according to PacBio recommendations (PN 101-853-100, v.5). Briefly, high-molecular-weight DNA was sheared by using the Megaruptor 3 system (Diagenode) to obtain an average size of 20 kb. Following enzymatic treatment of 5 µg of sheared DNA sample for removing single-stranded overhangs and DNA damage repair, ligation with overhang adaptors to both ends of the targeted double-stranded DNA molecule was performed to create closed, single-stranded circular DNA. Nuclease treatment was performed by using the SMRTbell Enzyme Clean-Up Kit 2.0 (Pacific Biosciences). Due to the limited quantity of libraries, no size selection was performed. Size and concentration of the final library were assessed using the Femto Pulse system (Agilent) and the Qubit fluorometer and the Qubit dsDNA HS reagents assay kit (Thermo Fisher), respectively.

Sequencing primer v2 and Sequel II DNA Polymerase 2.0 were annealed and bound, respectively, to the SMRTbell library. The library was loaded on two SMRT Cells 8M at an on-plate concentration of 70 pM using adaptive loading. Sequencing was performed on the Sequel II system using the Sequel II sequencing kit 2.0, a run movie time of 30 h with a 120-min pre-extension step and software v.9.0 (PacBio) by the Gentyane Genomic Platform (INRAE).

Whole-genome assembly. The data were assembled using hifiasm assembler (v.16.1 (ref. 73)). Hifiasm is able to produce a primary

assembly and an alternative assembly (incomplete alternative assembly consisting of haplotigs under heterozygous regions). We then filtered the primary assembly from organelles and low-quality contigs. Organelles were identified by mapping contigs on NCBI reference assemblies. The closest references (AP025455.1 and AP025456.1) were found with MitoHiFi software[74]. We finally removed 601 contigs from the primary assembly (mitochondrial and chloroplastic contigs identified, high-coverage (>100×), low-coverage (<1×) and high GC percent (>50%) contigs). All the metrics correspond to the filtered primary assembly used for the next analysis. We obtained 527,046 corrected reads (N_{50} = 17.6 kbp, genome coverage = 23×) that were assembled and filtered, resulting in 109 contigs (N_{50} = 5.99 Mbp, L_{50} = 12 contigs, GC content = 43.34%).

To assess the completeness and quality of the genome, we used the BUSCO v.5.4.4 (ref. 75) tool on the Viridiplantae odb10 database (n = 425 (ref. 70)). We obtained a 99.1% complete BUSCO score on the assembly.

Short-read sequencing

DNA extraction, library preparation and DNA sequencing (Illumina). Genomic DNA was extracted for all accessions, either from in vitro-propagated material or from plants grown in pots. For plants sequenced at BGI Shenzhen or at a service company, genomic DNA was extracted using a CTAB modified protocol from Healey et al.[76]. Approximately 0.1 g of cleaned young thalli (2–3 weeks) from *Marchantia* populations was reduced in a fine powder in nitrogen liquid. Because *Marchantia* tissues contain high amounts of phenolic compounds and other molecules that can interfere with DNA extraction, 1 ml STE (sucrose, 0.25 M (200-301-B, Euromedex); Tris, 0.1 M (26-128-3094-B, Euromedex); and EDTA, 0.05 M (EU0007, Euromedex)) wash buffer was previously added to the fine powder in the Eppendorf tube. Next, the mixture was thoroughly vortexed and centrifuged for 5 min at 15,000*g*. After centrifugation, the supernatant was removed, and washing of ground tissues with STE buffer was repeated until the supernatant was clear. Once ground tissues were clearly washed, DNA extraction and NanoDrop quality assessment were performed as described by Healey et al.[76] by adjusting volumes to the 0.1 g of starting powder. Illumina libraries were prepared according to manufacturer instructions. Illumina sequencing was performed using either Illumina sequencing technology (HiSeq 2000 and HiSeq 4000) or the Illumina NovaSeq 6000 at BGI Shenzhen or the service company, respectively, in both cases using a 2 × 150 paired-end configuration. For accessions sequenced at JGI, DNA was extracted using the CTAB–STE method[77]. Libraries were prepared following manufacturer instructions and sequenced on an Illumina NovaSeq 6000 instrument using a 2 × 150 paired-end configuration.

All sequenced libraries are available under BioProject PRJNA931118 at the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra/).

Assembly and cleaning of 129 *M. polymorpha* Illumina sequences. For 129 accessions that were of good enough quality, short reads were processed and adaptors were removed with Trim Galore v.0.6.5 (ref. 78) with the -q 30 option and then assembled with MEGAHIT v.1.1.3 (ref. 79) with default parameters. These assemblies were then cleaned by discarding the contigs not complying with the following characteristics: GC content inferior to 55, at least one hit but no bacterial one in the top ten blast hits of the contig against the RefSeq database from the NCBI (from February 2023), length of the contig superior to 500 bp.

Prediction and functional annotation of protein-coding genes

Genome assemblies were softmasked using Red v.2.0 (ref. 80), and structural annotations were made using the BRAKER2.1.6 pipeline[81–90]. BRAKER2 was run with '--epmode --softmasking --gff3 --cores 1' options. In epmode, the ProtHint pipeline generates hints for AUGUSTUS training and predicts protein-coding genes. The OrthoDB input proteins used by ProtHint are a combination of ones from https://busco-data.ezlab.org/v5/data/lineages/viridiplantae_odb10.2024-01-08.tar.gz and proteins from seven species (*A. agrestis* cv. BONN, *A. agrestis* cv. OXF, *Anthoceros punctatus*[91], *C. purpureus* strain R40 (NCBI GCA_014871385.1), *M. paleacea*[43], *M. polymorpha* ssp. *ruderalis* Tak-1 (https://marchantia.info/download/MpTak_v6.1/), *P. patens*[92] and *Sphagnum fallax*[93]). Only the predicted genes partly supported by this protein database were kept using the script at https://github.com/Gaius-Augustus/BRAKER/blob/report/scripts/predictionAnalysis/selectSupportedSubsets.py.

The long-read genomes of *M. polymorpha* ssp. *ruderalis*, *M. polymorpha* ssp. *polymorpha*, *M. polymorpha* ssp. *montivagans*[28] and *M. paleacea* were reannotated in the same fashion as well as the two new long-read genomes from the accessions CA and BoGa-L5.

The completeness of the predictions in each accession was assessed with BUSCO v.5.4.4 against the Viridiplantae odb10 (n = 425).

The predictions were functionally annotated with InterProScan 5.51-85.0 (refs. 94,95) with options –iprlookup and –goterms.

Mapping and SNP calling of the 135 accessions

Trim Galore v.0.6.5 (ref. 78) was used to remove 3′ bases with a quality score lower than 30, trim Illumina adaptors and only keep reads longer than 20 bp. Bowtie 2 v.2.3.5.1 (ref. 96) was then used to map reads from the 135 accessions to the reference genome of *M. polymorpha* ssp. *ruderalis* (Tak-1 accession, MpTak v.6.1 genome version[66]) without permissiveness for discordant and mixed alignment. Polymorphic variants with a minimum of four supporting reads, a minimum base quality of 30, a minimum variant allele frequency of 0.97 and a P-value threshold of 0.01 were then called with VarScan v.2.4.2 (refs. 97,98). The resulting VCF was then filtered with VCFtools v.0.1.16 to discard indels and two accessions of low quality and keep only biallelic sites with information in at least 50% of the accessions, leading to a total of 12,519,663 SNPs in the 133 remaining accessions and 5,414,844 in the 104 *M. polymorpha* ssp. *ruderalis* accessions. The SnpEff v.5.0e program was used to predict the effect of SNPs in the *ruderalis* subspecies.

Sex assignment of the 135 accessions

The sex of the 135 accessions was determined by evaluating the mapping of reads on the U and V chromosomes of the reference genome with Indexcov[99]. This was complemented by PCR sexing for 104 accessions, based on the primers presented by Hoey et al.[30], as well as an additional bioinformatic check up by performing a blastn search (*e*-value threshold of 1×10^{-15})[100] of these same primers on the accessions' assemblies.

This led to a sex assignment consensus for the 124 remaining ones displaying an ambiguous signal. Eighty of these accessions are females, and 44 are males. PCR and bioinformatic prediction disagreed for two accessions: NILSC14 and Mns-A.

Phylogeny and population structure analysis

A total of 107,744 multiallelic and biallelic SNPs from the autosomes, pruned with the R package SNPRelate v.1.30.1 (ref. 101) for an LD < 0.3 on 1,500-bp windows and with missing data <4%, were kept for phylogenetic analysis on the 133 *M. polymorpha* accessions. A SNP-based phylogenetic tree was inferred with IQ-TREE v.2.1.2 (ref. 102) and ModelFinder[103] with an SH-like approximate likelihood ratio test[104] and ultrafast bootstrap[105] (with 10,000 replicates). The tree was then visualized using the phytools R package[106].

Population structure analysis was performed on 124 nonredundant *M. polymorpha* accessions from the SNPs with the same filters as the ones used in the phylogeny, and an additional filter on 5% MAF (20,146 SNPs) was used in fastSTRUCTURE 1.0 (ref. 107). Population numbers (K) from 2 to 7 were tested with five cross-validation test sets, pointing to an optimal K between 4 and 6. Tracts of IBD among pairs of 116 nonredundant accessions were identified by using the program hmmIBD v.2.1.3 (ref. 108), with the same parameters as those

reported by Sandler et al.[31]: recombination rate of 4.54×10^{-6} cM per bp and number of chromosomes set to eight for the autosomes. The IBD analysis was also performed independently on the 44 male and 80 female sex chromosomes, setting the same recombination rate and the number of chromosomes to one.

Detecting genomic fragments resulting from recent introgressions using a distance-based method

We took advantage of the shape of the phylogenetic tree constructed based on the whole SNP dataset (Fig. I1 from the file Figure-introgressions-Marchantia-v4 at https://doi.org/10.6084/m9.figshare.25574100; very long basal branches and very short terminal branches) to develop a method to detect short (for example, 500 nucleotides) fragments resulting from recent introgressions. For each accession of a given subspecies, we expect that its distance to the other accessions of this subspecies (terminal branches) should be much smaller than the mean distance between subspecies (internal branches). Our method therefore looks for accessions too dissimilar from their own subspecies; this allows us to detect introgression from the ghost lineage, which is not possible with IBD. Simulations allowed us to determine parameters (for example, window size) that produce a very limited number of false positives (window size = 25). Details of the distance-based method and additional results are provided at https://doi.org/10.6084/m9.figshare.25574100.

Determination of selection signature on genes

Selection signatures were detected with indicators based on the AFS: Tajima's D[109], Fay and Wu's H[110] and Zeng's E[111]. By exploiting the biallelic SNP dataset in the three subspecies with PLINK v.1.90b6.21 (ref. 112), the ancestral or derived state of each allele in *M. polymorpha* ssp. *ruderalis* was determined. For the polymorphic positions in *M. polymorpha* ssp. *ruderalis* that were fixed the same way in the two other subspecies (allowing missing information in one accession of each subspecies), the *M. polymorpha* ssp. *ruderalis* allele corresponding to the fixed allele in *M. polymorpha* ssp. *montivagans* and *M. polymorpha* ssp. *polymorpha* was considered to be the ancestral allele. This led to a reduced dataset of 1,344,013 unfolded SNPs that were used to calculate the H and E indicators. We used the folded SNP dataset (5,414,844 SNPs) to calculate Tajima's D. Calculation of the D, H and E statistics was performed with a custom R script[113]: θ_W, θ_π and θ_L were calculated for each SNP, with $\theta_W = \left(\sum_{i=1}^{n-1} \frac{1}{i}\right)^{-1}$, $\theta_\pi = \frac{2n_A n_D}{n(n-1)}$ and $\theta_L = \frac{n_D}{n-1}$ (with n_A and n_D being the copy number of the ancestral and derived alleles, respectively, in the case of H and E calculations), which allowed the correction of the values for the missing data present at each site. The $D = \frac{\theta_\pi - \theta_W}{\sqrt{\mathrm{var}(\theta_\pi - \theta_W)}}$, $H = \frac{\theta_\pi - \theta_L}{\sqrt{\mathrm{var}(\theta_\pi - \theta_L)}}$ and $E = \frac{\theta_L - \theta_W}{\sqrt{\mathrm{var}(\theta_L - \theta_W)}}$ values obtained for each site were then averaged over the SNPs present in each of the 18,920 gene models, allowing us to determine the global selective pressure acting on each gene. The same method was used to determine the selection signatures in sliding windows of different sizes (20 kb, 5 kb and 0.5 kb, with a sliding of 20% of the window size) over the whole genome of *M. polymorpha* (available for visualization on the genome browser hosted by MarpolBase at https://marchantia.info/). Genes with pronounced selection signatures were determined based on a dataset of 18,140 genes with at least four folded SNPs to calculate Tajima's D statistics. Based on this, we selected the top 10% of genes (that is, 1,814) with the highest values of Tajima's D as a gene set with the most pronounced signature of balancing selection. Among the remaining 16,326 genes, based on the ancestral–derived SNP dataset (1,344,013 SNPs), we selected 11,932 genes for which the ancestral–derived allele identification could be inferred for at least 50% of the SNPs. This gene set allowed us to select 777 genes with a marked signature of background selection based on Zeng's E value < 10% genome-wide (18,140 genes) quantile value. Finally, the same gene set allowed us to select 1,374 genes with marked signatures of soft or hard selective sweep based on Fay and Wu's H value < 10% or E value > 90% genome-wide (18,140 genes) quantile value. Hence, 3,965 genes were considered as having pronounced selection signatures based on the three neutrality test statistics D, H and E.

The calculation of D, H and E statistics and the selection of genes with pronounced selection signatures were performed in the same way for the angiosperm model species *A. thaliana* and *M. truncatula*. For *A. thaliana*, the SNP dataset from the 1001 Genomes Consortium (https://1001genomes.org/data/GMI-MPI/releases/v3.1/ (ref. 114)) was used, coupled with mapping and calling of *A. thaliana* SNPs on sequencing data from 30 *Arabidopsis halleri* (PRJNA592307) and 29 *Arabidopsis lyrata* (PRJNA357372) accessions on the *Arabidopsis* TAIR10 reference genome. Reads from the two outgroups were processed with Trim Galore similarly to what was done with *M. polymorpha* and then mapped with Bowtie 2 v.2.3.5.1 with permissiveness for discordant and mixed alignment and with an alignment score lower than the minimal score threshold of $-1.5 + -1.5 \times L$ to allow mapping of reads from distant species. The polymorphic variants present in *A. thaliana* were called with VarScan v.2.4.2 in the two other subspecies, with a minimum of four supporting reads, a minimum base quality of 30, a minimum variant allele frequency of 0.97 and a P-value threshold of 0.01, leading to 9,500,949 sites for ssp. *halleri* and 9,206,703 sites for ssp. *lyrata*. The unfolded *A. thaliana* sites were the ones bearing the same allele in 100% of the accessions from ssp. *halleri* and ssp. *lyrata* and with less than 30% of missing values. This led to 3,641,556 SNPs with ancestral allele information, with which H and E were calculated. Tajima's D was calculated on the 11,458,975 SNPs from the 1,135 accessions of the 1001 Genomes Consortium. For *M. truncatula*, the SNP dataset based on mapping of 317 accessions (among which 285 are from the *M. truncatula* species) on v.5 of the A17 genome was used (https://data.legumeinfo.org/Medicago/truncatula/diversity/A17.gnm5.div.Epstein_Burghardt_2022/ (refs. 115,116)). The 50,885,956 biallelic SNPs with minimum base quality of 30, no half calls and a maximum of 50% of missing data were used to calculate Tajima's D. SNP calling of eight accessions from various other species phylogenetically close to the *M. truncatula* species (species *Medicago turbinata*, *Medicago italica*, *Medicago doliata*, *Medicago littoralis*, *M. turbinata*, *Medicago tricycla* and *Medicago soleirolii*), performed by Epstein and collaborators, was taken as the outgroup dataset. The ancestral allele was determined for sites that were fixed the same way in all eight outgroup accessions, with a tolerance for missing data in one accession only, leading to a dataset of 12,129,573 unfolded *M. truncatula* SNPs, on which H and E were calculated. The values of the D, H and E statistics calculated for *A. thaliana* and *M. truncatula* SNPs were then averaged over their gene models. We compared *M. polymorpha* ssp. *ruderalis*, *A. thaliana* and *M. truncatula* D, H and E datasets by using orthology relationships to compare selection signatures across species (OrthoFinder analysis on a representative sample of land plants). To account for the deep and complex evolutionary divergence between bryophytes and angiosperms, notably due to several rounds of whole-genome duplication, we focused on HOGs showing in each of the three species at least one gene that is in their top 20% of genes under background, balancing or sweep selection at the intraspecific level (Supplementary Table 8). To determine whether the number of orthogroups sharing genes under the same selection regime for all three species was significantly different from random expectation, the value of selection signature for each species was shuffled and the overlap was determined. This shuffling was conducted 1,000 times, leading to a mean of ten orthogroups (highest value of 20 orthogroups) with common background selection and 29 orthogroups with common balancing selection (highest value of 46 orthogroups), values well below those found in the analysis on real selection signatures of the three species.

Estimation of recombination rates and of LD decay

The genome-wide landscape of recombination was determined based on the SNP data, using ReLERNN v.1.0.0 (ref. 117). For the simulation

phase, the assumed mutation rate was 10^{-8} and the assumed generation time was 1 year. The other phases (train, predict and BScorrect) were run with default parameters. LD decay in *M. polymorpha* ssp. *ruderalis* was determined on the dataset of 5,414,844 SNPs called on the 104 ssp. *ruderalis* accessions, using PopLDdecay[118] with a MAF of 0.05 and a maximum pairwise SNP distance of 40 kb. The LD decay was also computed on each chromosome separately.

Genome–environment association analyses

GEA analyses were carried out to identify genomic regions in which SNP alleles displayed statistical association with the bioclimatic variables extracted from the WorldClim 2 database (https://www.worldclim.org/ (ref. 119)). We performed analyses on each of the 19 bioclimatic variables (BIO1 to BIO19), on monthly average precipitation, solar radiation (kJ m^{-2} d^{-1}) and water vapor pressure (kPa) and on elevation (meters above sea). To integrate the correlations among bioclimatic variables, we performed PCA separately on temperature (BIO1 to BIO11) and precipitation (BIO12 to BIO19) variables, monthly precipitation, solar radiation (kJ m^{-2} d^{-1}) and water vapor pressure (kPa). The first three principal components were then used for the GEA analyses. GEA analyses were performed using the mixed linear model implemented in GEMMA software (v.0.98.1 (ref. 120)). We used a dataset of 2,155,434 SNPs with minor allele frequencies of 0.05 and a maximum of 10% missing data on a set of 96 accessions of *M. polymorpha* ssp. *ruderalis* for which climatic data were available. To estimate the SNP effects and their significance, the model used a centered kinship matrix as a covariable with random effects and a Wald test. SNP *P* values from GEMMA were processed using a local score approach[121,122] to help detect robust loci across analyses. The local score is a cumulative score that takes advantage of local LD among SNPs. This score, defined as the maximum of the Lindley process over a SNP sequence (that is, a chromosome), was calculated using a tuning parameter value of $\xi = 2$, as suggested by simulation results[121]. Chromosome-specific significance thresholds ($\alpha = 5\%$) were estimated using a resampling approach. The R scripts used to compute the local score and significance thresholds are available at https://forge-dga.jouy.inra.fr/projects/local-score/documents. Visualization of the candidate loci was carried out with the ComplexHeatmap R package[123].

OrthoFinder analysis on a representative sample of land plants

To compare genes across evolutionarily distant species, *A. thaliana*, *M. truncatula* and *M. polymorpha* orthogroups were reconstructed using OrthoFinder v.2.5.2 (ref. 124). To ensure an accurate reconstruction of orthogroups, 36 species covering the main lineages of land plants were added to the above-mentioned species. A first round of OrthoFinder was performed, and a careful inspection of the estimated species was conducted to ensure the absence of errors compared to the phylogeny of land plants[49,125]. A consistent tree was reconstructed, and the tree was used for a second run of OrthoFinder with the option 'msa', enabling us to reconstruct the orthogroups using alignment and phylogenetic reconstruction methods.

Gene-centered pangenome construction

The presence–absence variation of genes in *M. polymorpha* and in the subset of *M. polymorpha* ssp. *ruderalis* accessions was determined based on orthogroup inference. The proteins predicted from the short-read and long-read assemblies were pooled together, and OrthoFinder was run independently on the dataset of 104 genomes for the ssp. *ruderalis* pangenome and on the dataset of 134 genomes for the *Marchantia* complex pangenome. The OrthoFinder analyses were run a first time with default parameters, the resulting species tree was then rerooted (on *M. paleacea* for the *Marchantia* complex pangenome and on the ssp. *montivagans* reference genome for the ssp. *ruderalis* pangenome), and a second OrthoFinder run forced on this tree topology was performed with the multiple-sequence alignment parameter. To remove any remaining contamination, all the proteins used in this orthogroup inference were blasted against the nonredundant protein database from the NCBI (NR version from 20 February 2023) using DIAMOND v.2.0.8 (ref. 126) with an e-value threshold of 10^{-3}, a maximum of ten target sequences and a block size of 4. HOGs with 75% of Viridiplantae DIAMOND hits over all genes were considered reliable; the other ones were discarded. This cleaning led to 25,648 HOGs in the *Marchantia* complex pangenome (initially 33,418 HOGs) and 28,143 HOGs in the ssp. *ruderalis* pangenome (initially 35,004 HOGs). Two accessions were removed from the ensuing analyses because they had less than 10,000 genes assigned to HOGs. Core orthogroups were determined in each subspecies. *M. polymorpha* ssp. *montivagans* orthogroups containing 15 of the total 17 accessions were considered core (because two accessions from this subspecies were absent from many core orthogroups), while the *M. polymorpha* ssp. *polymorpha* core orthogroups had to contain all accessions. For the ssp. *ruderalis*, the core orthogroups were defined as orthogroups containing all accessions, with tolerance for six specific accessions with lower quality, meaning that some or all of them could be absent from the HOG and the HOG would still be considered as core in the subspecies. The list of shared orthogroups among the *Marchantia* complex was determined by crossing those core orthogroup lists in each subspecies with the list of orthogroups present in *M. paleacea*. HOGs were considered specific to a subspecies if they contained at least one accession of the subspecies and were absent from all accessions of other subspecies. For the *M. polymorpha* ssp. *ruderalis* pangenome, the core genome was determined in the same way as the *Marchantia* complex pangenome (shared by all accessions with the tolerance margin for the six lower-quality accessions) but on the orthogroups generated by the OrthoFinder run on the 104 genomes. The accessory genome is constituted of the remaining HOGs that are present in more than four accessions, and the cloud genome comprises HOGs present in four accessions or less.

Pangenome saturation

The ssp. *ruderalis* pangenome and core genome saturation were studied by determining the number of HOGs present (whole pangenome) and the number of HOGs shared between the accessions (core genome) for a given number of accessions. For each given number of accessions, 500 random samplings were conducted. The corresponding numbers of HOGs present in the sample and shared between accessions were plotted separately as box plots.

Pangenome InterPro term enrichment

Enrichment analyses were performed on the compartments determined in the *M. polymorpha* ssp. *ruderalis* pangenome, comparing the IPR domains present in the core and accessory genome to the content of the whole pangenome. The presence of multiple genes for an accession in the same HOG was accounted for by multiplying the IPR count of each HOG by the median number of genes per accession in the HOG. The IPR content of each compartment was then compared to the IPR content of the whole pangenome with a hypergeometric test, IPR terms present less than five times in the whole pangenome were filtered out, and multiple-testing correction was computed on the remaining IPR terms (Benjamini–Hochberg, false discovery rate). IPR terms with a corrected *P* value less than 0.05 were considered significant.

Candidate gene phylogenetic analysis

The orthologs of some GEA candidates (like *ABC1K7*) were determined by BLASTp+ v.2.12.0 (maximum of 4,000 target sequences and *e* value of 10^{-30}) against a database of 37 streptophytes (Supplementary Table 18). The resulting sequences were aligned with Muscle5 v.5.1 (ref. 127) and trimmed with trimAl v.1.4 to discard positions with more than 60% of gaps. The phylogenetic tree was computed with IQ-TREE v.2.1.2, and the best-fitting evolutionary model was selected using ModelFinder according to the Bayesian information criterion. Branch support was

estimated with 10,000 replicates of both SH-like approximate likelihood ratio and ultrafast bootstrap (details of the models chosen and log likelihood for each tree are in Supplementary Table 19).

Phylogenetic analysis of *M. polymorpha* fungal lectins was performed by blasting (BLASTp+ v.2.12.0) their protein sequence against transcriptomes of the 1KP initiative[49], a database of fungal genomes from MycoCosm[128] (last time consulted, February 2019), a database of algal genomes and a database of plant genomes (Supplementary Table 18), each with an *e* value of 10^{-5} and maximum of 2,000 target sequences. Muscle5 was used to align the sequences, trimAl was used to discard positions with more than 60% of gaps, and IQ-TREE 2 was used to compute the tree, in the same way as for the previous phylogenies.

Statistical analyses

Statistics applied in the present study were performed in R (v.4.2.2)[129] and are provided alongside the respective analysis in Methods, the main text and figure legends. The statistical significance tests in GEA and functional enrichment analyses have been described above.

Reporting summary

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

Data availability

Raw illumina reads generated for this study are available at NCBI under BioProject PRJNA931118. The SNP data mapped on the Tak-1 assembly (VCF format) and genome assemblies (FASTA) as well as their annotations (protein FASTA and GFF) for the 133 accessions sequenced in this paper and a graphical representation of the sliding window scan of Tajima's *D*, Wu's *H* and Zheng's *E* are available at MarpolBase (https://marchantia.info/). The *M. polymorpha ruderalis* accession CA annotated genome assembly and raw sequencing data are also available at NCBI under BioProject PRJNA1021402. The BoGa genome assembly is available at https://doi.org/10.4119/unibi/2991996.

Code availability

Custom code used in this study is available at https://doi.org/10.6084/m9.figshare.24428083 (ref. 113).

References

59. Delaux, P. et al. Origin of strigolactones in the green lineage. *New Phytol.* **195**, 857–871 (2012).
60. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
61. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
62. 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).
63. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
64. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
65. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
66. Montgomery, S. A. et al. Chromatin organization in early land plants reveals an ancestral association between H3K27me3, transposons, and constitutive heterochromatin. *Curr. Biol.* **30**, 573–588 (2020).
67. Alonge, M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
68. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).
69. Smit, A., Hubley, R. & Green, R. RepeatMasker Open v.4.0 http://www.repeatmasker.org (Institute for Systems Biology, 2013).
70. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
71. Tillich, M. et al. GeSeq — versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **45**, W6–W11 (2017).
72. Kent, W. J. BLAT — the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
73. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
74. Uliano-Silva, M. et al. MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC Bioinformatics* **24**, 288 (2023).
75. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: assessing genomic data quality and beyond. *Curr. Protoc.* **1**, e323 (2021).
76. Healey, A., Furtado, A., Cooper, T. & Henry, R. J. Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant Methods* **10**, 21 (2014).
77. Shepherd, L. D. & McLay, T. G. B. Two micro-scale protocols for the isolation of DNA from polysaccharide-rich plant tissue. *J. Plant Res.* **124**, 311–314 (2011).
78. Krueger, F., James, F., Ewels, P., Afyounian, E. & Schuster-Boeckler, B. FelixKrueger/TrimGalore: v0.6.7 — DOI via *Zenodo*. *Zenodo* https://doi.org/10.5281/ZENODO.5127899 (2021).
79. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
80. Girgis, H. Z. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics* **16**, 227 (2015).
81. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom. Bioinform.* **2**, lqaa026 (2020).
82. 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 Genom. Bioinform.* **3**, lqaa108 (2021).
83. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
84. Gotoh, O., Morita, M. & Nelson, D. R. Assessment and refinement of eukaryotic gene structure prediction with gene-structure-aware multiple protein sequence alignment. *BMC Bioinformatics* **15**, 189 (2014).
85. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2016).
86. Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-genome annotation with BRAKER. In *Gene Prediction* Vol. 1965 (ed. Kollmar, M.) 65–95 (Springer, 2019).
87. Iwata, H. & Gotoh, O. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Res.* **40**, e161 (2012).
88. Lomsadze, A. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506 (2005).

89. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006).
90. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
91. 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).
92. Lang, D. et al. The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution. *Plant J.* **93**, 515–533 (2018).
93. Healey, A. L. et al. Newly identified sex chromosomes in the *Sphagnum* (peat moss) genome alter carbon sequestration and ecosystem dynamics. *Nat. Plants* **9**, 238–254 (2023).
94. Blum, M. et al. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2021).
95. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
96. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
97. Koboldt, D. C. et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283–2285 (2009).
98. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
99. Pedersen, B. S., Collins, R. L., Talkowski, M. E. & Quinlan, A. R. Indexcov: fast coverage quality control for whole-genome sequencing. *Gigascience* **6**, 1–6 (2017).
100. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
101. Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
102. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
103. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
104. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
105. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
106. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things): phytools: R package. *Methods Ecol. Evol.* **3**, 217–223 (2012).
107. Raj, A., Stephens, M. & Pritchard, J. K. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589 (2014).
108. Schaffner, S. F., Taylor, A. R., Wong, W., Wirth, D. F. & Neafsey, D. E. hmmIBD: software to infer pairwise identity by descent between haploid genotypes. *Malar. J.* **17**, 196 (2018).
109. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
110. Fay, J. C. & Wu, C. I. Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413 (2000).
111. Zeng, K., Fu, Y.-X., Shi, S. & Wu, C.-I. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* **174**, 1431–1439 (2006).
112. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
113. Beaulieu, C. Code repository for 'The *Marchantia polymorpha* pangenome reveals ancient mechanisms of plant adaptation to the environment'. *Figshare* figshare.com/s/715a36bc7585d46d0279 (2024).
114. Alonso-Blanco, C. et al. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**, 481–491 (2016).
115. Epstein, B. et al. Combining GWAS and population genomic analyses to characterize coevolution in a legume–rhizobia symbiosis. *Mol. Ecol.* **32**, 3798–3811 (2022).
116. Pecrix, Y. et al. Whole-genome landscape of *Medicago truncatula* symbiotic genes. *Nat. Plants* **4**, 1017–1025 (2018).
117. Adrion, J. R., Galloway, J. G. & Kern, A. D. Predicting the landscape of recombination using deep learning. *Mol. Biol. Evol.* **37**, 1790–1808 (2020).
118. Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M. & Yang, T.-L. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2019).
119. Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).
120. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
121. Bonhomme, M. et al. A local score approach improves GWAS resolution and detects minor QTL: application to *Medicago truncatula* quantitative disease resistance to multiple *Aphanomyces euteiches* isolates. *Heredity* **123**, 517–531 (2019).
122. Fariello, M. I. et al. Accounting for linkage disequilibrium in genome scans for selection without individual genotypes: the local score approach. *Mol. Ecol.* **26**, 3700–3714 (2017).
123. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
124. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
125. The Angiosperm Phylogeny Group et al. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* **181**, 1–20 (2016).
126. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
127. Edgar, R. C. Muscle5: high-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.* **13**, 6968 (2022).
128. Grigoriev, I. V. et al. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* **42**, D699–D704 (2014).
129. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2023).

Acknowledgements

We thank the GenoToul bioinformatics platform in Toulouse, Occitanie (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12) for providing computing resources and public contributors of material from the Great British Liverwort Hunt (I. Burrow, D.G. Hill, the Roberts family, P. and T. Rooney, J. Showers, A. Wilson, an additional 59 anonymous contributors from the British public and Y. Coudert). Research at the Laboratoire de Recherche en Sciences Végétales and the LIPME is supported by the Laboratoires d'Excellence TULIP (ANR-10-LABX-41). J.K., C.G., C.L. and P.-M.D. were supported by the project Engineering Nitrogen Symbiosis for Africa currently funded through

a grant to the University of Cambridge by the Bill and Melinda Gates Foundation (OPP1172165) and the UK Foreign, Commonwealth and Development Office as Engineering Nitrogen Symbiosis for Africa (OPP1172165). The work (proposal: award at https://doi.org/10.46936/10.25585/60001405) conducted by the US Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the US Department of Energy operated under contract no. DE-AC02-05CH11231. This project received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement no. 101001675-ORIGINS) and from the FRM/FSER (FRM/FSER202302017064) to P.-M.D., from the CNRS to P.-M.D. and H.P. (80|PRIME MicMac), from the Deutsche Forschungsgemeinschaft (ZA, 259/9) to S.Z., from the National Science Foundation to J.M.N. (NSF 1501826), from the Japan Society for the Promotion of Science KAKENHI (JSPS 20K15783), from the Agence Nationale de la Recherche (ANR LEVEL-UP ANR-21-CE20-0010-01) to C.J., from ZhuJiang (2019ZT08N628) and the National Natural Science Foundation of China (32022006) to S. Cheng and from the URPP Evolution in Action of the University of Zurich, grants from the Swiss National Science Foundation (160004, 131726), the EU's Horizon 2020 research and innovation program (PlantHUB 722338), the Georges and Antoine Claraz Foundation and the Forschgungskredit of the University of Zurich (FK-20-089) to P.S.

Author contributions

P.-M.D., J.M.N., F.R., D.J.H., G.R.L.G., S.S., the SLCU Outreach Consortium led by D.J.H. and E.M., P.S. and S.Z. collected and contributed accessions. C.B., M. Bonhomme, J.K., D.L.M.Z., P.-M.D., G.R.L.G., N.R., A.T., C.G., P.S., I.D., A.T., S. Cauet, N.R., K.E.M., E.A., C.G., S.A., J.W., J.G., W.H., A.M., B.L., A. Bräutigam, B.C. and H.S.C. carried out experiments. C.B., C.L., M. Bonhomme, P.-M.D., S. Cauet, J.K., H.P. and S.A. analyzed data. C.L., M. Bonhomme, P.-M.D., C.J., C.D., H.P., J.H.L.-M., J. Schmutz, A. Bräutigam and S. Cheng coordinated experiments. Y.T. developed the website. C.B., M. Bonhomme and P.-M.D. wrote the manuscript. C.L., F.R., P.S., S.Z. and S.S. edited the manuscript. M. Bonhomme and P.-M.D. coordinated the project.

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/s41588-024-02071-4.

Correspondence and requests for materials should be addressed to Maxime Bonhomme or Pierre-Marc Delaux.

Peer review information *Nature Genetics* thanks John Bowman, Fay-Wei Li and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Pierre-Marc Delaux, Maxime Bonhomme

Last updated by author(s): 12/06/2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	no software was used
Data analysis	Trimmomatic v0.39, Canu v2.2; Racon v1.4.20; Medaka v1.4.3; Pilon v1.24; RagTag v2.0.1; RepeatModeler v2.0.2; RepeatMasker v4.1.2; BUSCO v5.4.4; blastn v2.11.0+; Berokka v0.2; GeSeq v2.03; tRNAscan-SE (v2.0.7); Chloe v0.1.0; HMMER v3.3.1 ; BLAT; HiFiasm assembler v16.1; MitoHiFi v2.1; Trim Galore! V 0.6.5; megahit v1.1.3; Red v2.0; BRAKER v2.1.6; interproscan-5.51-85.0; Bowtie2 v2.3.5.1; VarScan.v2.4.2; VCFTools v0.1.16, SnpEff v5.0e, SNPRelate_1.30.1; IQ-TREE v2.1.2; phytools_1.5-1; mapdata_2.3.1; ape_5.7-1; fastStructure 1.0; hmmIBD v2.1.3; PLINK v1.90b6.21; ReLERNN v1.0.0; PopLDdecay v3.42; gemma v0.98.1; OrthoFinder v2.5.2; diamond v2.0.8; BLASTp+ v2.12.0; muscle5 v5.1; trimAl v1.4; ComplexHeatmap v2.22.0; custom code doi: 10.6084/m9.figshar.24428083 (D, H and E statistics calculation) and 10.6084/m9.figshare.25574100 (introgressed fragments detection)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The SNP data mapped on the Tak-1 assembly (VCF format) and genome assemblies (FASTA) as well as their annotations (protein FASTA and GFF) for the 133 accessions sequenced in this paper are available at MarpolBase (https://marchantia.info/). The Marchantia polymorpha ruderalis accession CA annotated genome assembly and raw sequenced data are also available in NCBI under BioProject PRJNA1021402. The BoGa genome assembly is available under https://doi.org/10.4119/unibi/2982437, the WorldClim database is available at the following link: https://www.worldclim.org/, the MycoCosm database is available at the following link: https://mycocosm.jgi.doe.gov/mycocosm/home

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	The research did not involve human participants
Reporting on race, ethnicity, or other socially relevant groupings	The research did not involve human participants
Population characteristics	The research did not involve human participants
Recruitment	The research did not involve human participants
Ethics oversight	The research did not involve human participants

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The aim of this study was to explore the intraspecific diversity in the model species Marchantia polymorpha (135 accessions) in order to understand the genetic mechanisms used by this bryophyte to adapt to its surroundings and compare them to the ones of other land plants.
Research sample	The sample is composed of 135 accessions from the three subspecies of the model species Marchantia polymorpha, with an enhanced sampling effort on the ruderalis subspecies, since the reference genome of M. polymorpha is from this subspecies.
Sampling strategy	Previous GEA and GWAS in plants used a number of accessions ranging from ~50 to 1000+. We aimed at collecting accessions within that range. We targeted a diversity of latitude and geographic areas to maximize diversity. We collected samples in areas where collection was allowed in the frame of the Nagoya protocol. We did not design the sampling to address biogeographical questions, as stated in the manuscript. This would have required a denser and more homogeneous sampling effort, including areas where collection is extremely regulated.
Data collection	One to five thalli for each accessions were collected in sealed plastic bags, location and GPS coordinate collected for accessions from the UK, France and the USA. Collection in eastern europe and scandinavia was done independently in plastic tubes, and GPS coordinates assessed afterward based on the collection site locations. Plants were maintained in control environment before being sterilized (single gemma) and maintained in vitro, in controled environment.
Timing and spatial scale	Accessions were collected between 2013 and 2018 for all accessions, except the UK accessions which were also collected in 2020-2022.
Data exclusions	Data from the accessions NISLC10 and RES37 were excluded from the analysis due to limited sequencing quality

Reproducibility	Sequencing was performed using multiple libraries for most accessions to limit single-sample effect. In addition, for three accessions (CA, Tak-1 and BoGa) both long-read and short-read sequencing was performed and compared.
Randomization	This does not apply here
Blinding	Blinding was not relevant for our study since sample collection, sequencing and statistical anlysis were performed independently and with standardized procedures.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	This does not apply as no data recording was performed on site.
Location	Sampling took place at multiple locations in Europe and USA. The GPS coordinates for each accession are listed in Table_S1
Access & import/export	Collection was done following the Nagoya protocol rules in each country. Collection in the UK performed after 2017 were recorded and authorization obtained from the individuals who are listed, or mentionned, in the manuscript (see 10.5281/ZENODO.10040685). In France, Marchantia polymorpha belongs to the exception for "model species" and where collected on public, not protected, areas before 2017 (when the APA was initiated). The TOU accessions were collected in a private farm from one of the Author's parents. Accessions transfered from the UK or the USA to France were already in in vitro culture, in double confinment, devoid of any microbioal or other biological contaminations.
Disturbance	Only one to five pieces of thalli were collected for each accessions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☐	☒ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes
☒	☐ Public health
☒	☐ National security
☒	☐ Crops and/or livestock
☒	☐ Ecosystems
☒	☐ Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☒	☐	Demonstrate how to render a vaccine ineffective
☒	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐	Increase transmissibility of a pathogen
☒	☐	Alter the host range of a pathogen
☒	☐	Enable evasion of diagnostic/detection modalities
☒	☐	Enable the weaponization of a biological agent or toxin
☒	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks	Marchantia polymorpha is a non-seed plants. We have collected and stored multiple samples of gemmae produced in vitro for most of the accessions. Some (listed in the manuscript) have been lost during the COVID pandemics and no physical record remains.
Novel plant genotypes	no novel genotypes were produced in the frame of this project.
Authentication	no novel genotypes were produced in the frame of this project.