Anthoceros genomes illuminate the origin of land plants and the unique biology of hornworts

Fay-Wei Li, Tomoaki Nishiyama, Manuel Waller, Eftychios Frangedakis, Jean Keller, Zheng Li, Noe Fernandez-Pozo, Michael S. Barker, Tom Bennett, Miguel A. Blázquez, Shifeng Cheng, Andrew C. Cuming, Jan de Vries, Sophie de Vries, Pierre-Marc Delaux, Issa S. Diop, C. Jill Harrison, Duncan Hauser, Jorge Hernández-García, Alexander Kirbis, John C. Meeks, Isabel Monte, Sumanth K. Mutte, Anna Neubauer, Dietmar Quandt, Gane K.-S. Wong, Sophie de Vries, Manuel Waller, Eftychios Frangedakis, Jean Keller, Péter Szövényi, John C. Meeks, Isabel Monte, Shifeng Cheng, Andrew C. Cuming, Dolf Weijers, Susann Wicke, Pierre-Marc Delaux, Gane K.-S. Wong, Sophie de Vries, Manuel Waller, Eftychios Frangedakis, Jean Keller, Péter Szövényi, John C. Meeks, Isabel Monte, Shifeng Cheng, Andrew C. Cuming, Dolf Weijers, Susann Wicke, Pierre-Marc Delaux, Gane K.-S. Wong, Sophie de Vries, Manuel Waller, Eftychios Frangedakis, Jean Keller, Péter Szövényi, John C. Meeks, Isabel Monte, Shifeng Cheng, Andrew C. Cuming, Dolf Weijers, Susann Wicke, Pierre-Marc Delaux, Gane K.-S. Wong, Sophie de Vries, Manuel Waller, Eftychios Frangedakis, Jean Keller, Péter Szövényi

Hornworts comprise a bryophyte lineage that diverged from other extant land plants >400 million years ago and bears unique biological features, including a distinct sporophyte architecture, cyanobacterial symbiosis and a pyrenoid-based carbon-concentrating mechanism (CCM). Here, we provide three high-quality genomes of Anthoceros hornworts. Phylogenomic analyses place hornworts as a sister clade to liverworts plus mosses with high support. The Anthoceros genomes lack repeat-dense centromeres as well as whole-genome duplication, and contain a limited transcription factor repertoire. Several genes involved in angiosperm meristem and stomatal function are conserved in Anthoceros and upregulated during sporophyte development, suggesting possible homologies at the genetic level. We identified candidate genes involved in cyanobacterial symbiosis and found that LCIB, a Chlamydomonas CCM gene, is present in hornworts but absent in other plant lineages, implying a possible conserved role in CCM function. We anticipate that these hornwort genomes will serve as essential references for future hornwort research and comparative studies across land plants.

Land plants evolved from a charophycean algal ancestor 470–515 million years ago and contributed to the greening of the terrestrial environment. The extant land plants consist of vascular plants and three bryophyte lineages—mosses, liverworts and hornworts. While the phylogeny of land plants has been debated, recent evidence indicates that bryophytes are monophyletic with hornworts a sister clade to Setaphyta (liverworts and mosses). The evolution of land plants is underlined by the rise of morphological, molecular and physiological innovations. Tracking the evolutionary origins of these key innovations is prone to errors due to uncertainty in reconstructing the most recent common ancestor (MRCA) of land plants. More than 400 million years of independent evolution of the three bryophyte lineages have provided ample time for evolutionary changes to happen and the availability of model systems for only two bryophyte lineages—mosses (Physcomitrella patens) and liverworts (Marchantia polymorpha)—makes inferences even more difficult. Hornworts, as the earliest diverging lineage in bryophytes, are crucial to infer character evolution and reveal the nature of the MRCA of bryophytes and that of land plants.

Hornworts uniquely possess a combination of traits that connect them with both green algae and other land plant lineages. For instance, most hornworts have a single chloroplast per cell with a pyrenoid capable of carrying out a carbon-concentrating mechanism (CCM). Such pyrenoid-based CCMs cannot be found in...
any other land plants but frequently occur in algae\textsuperscript{41}. Conversely, hornwort sporophytes are long-lived and moderately independent of gametophytes, which have been assumed to be features linking them to vascular plants\textsuperscript{42}. Furthermore, hornwort sporophytes bear stomata that may be homologous with those of vascular plants\textsuperscript{13}.

In addition to having characteristics exclusively shared with algae or with other land plants, hornworts also have a wide range of distinctive biological features. For example, the presence of a basal sporophytic meristem and asynchronous meiosis are unique to hornworts\textsuperscript{13}. Moreover, hornworts are among the very few plants that have a symbiotic relationship with nitrogen-fixing cyanobacteria\textsuperscript{21} and one particular hornwort species, \textit{Anthoceros punctatus}, has been used as a model system to study plant–cyanobacteria interactions\textsuperscript{16}.

Detailed genomic information on hornworts is essential not only to understand the evolutionary assembly of land plant-specific traits, but also to substantiate the full potential of hornworts as a model for studying the genetic basis of cyanobacterial symbiosis and pyrenoid-based CCMs. Here, we provide three high-quality genome assemblies and their annotations for the genus \textit{Anthoceros}. We use these data to refine our inferences on the nature of the land plant MRCA and to gain new insights into hornwort biology.

**Genome assembly and annotation**

We assembled three hornwort genomes from \textit{Anthoceros agrestis} (Bonn and Oxford strains) and \textit{A. punctatus}. For \textit{A. agrestis} Bonn, a combination of short- and long-read data with Chicago and Hi-C libraries resulted in a chromosomal-scale assembly with the six largest scaffolds containing 95% of the assembled genome (\textit{A. agrestis} has six chromosome pairs; Fig. 1 and Supplementary Fig. 1). For \textit{A. agrestis} Oxford strain and \textit{A. punctatus}, we used Oxford Nanopore sequencing to obtain high-quality assemblies composed of roughly 200 contigs with N50 over 1.7 megabase pairs (Mb) (Table 1). The three genomes are highly collinear with a greater collinearity found between the two \textit{A. agrestis} strains (Supplementary Fig. 2 and Supplementary Table 1). The collinearity, BUSCO (Benchmarking Universal Single-Copy Orthologs) (Supplementary Fig. 3) and read mapping statistics (Supplementary Tables 2 and 3), show that the three genomes are of high quality and accuracy.

The total assembly length varied between 117 and 133 Mb, which is consistent with the size estimates based on k-mer analysis (Table 1) but slightly larger than those from flow cytometry\textsuperscript{12,13}. Although these genomes are among the smallest of land plants, their repetitive and transposable element contents are considerable (36–38%). Similar to other plant genomes, the most abundant repeats are long terminal repeat elements (>20%) followed by a large number of unclassified repeats and DNA elements. The genome size variation among the three strains can be largely attributed to the differences in repeat content (Supplementary Fig. 4 and Supplementary Table 4). A combination of ab initio, evidence-based and comparative gene prediction approaches resulted in 24,700–25,800 predicted protein-coding genes (Supplementary Table 5). For \textit{A. agrestis} we also created a pan genome combining genome assemblies and gene annotations of the two strains (Bonn and Oxford) in a non-redundant way (see Methods, Supplementary Table 5 and Supplementary notes). The three hornwort genomes show a high gene density compared to other land plants (Supplementary Table 6). All three genomes and their annotations can be accessed, browsed, searched and downloaded from ref. \textsuperscript{19}.

**\textit{Anthoceros} displays unusual centromere structure**

The chromosomal-level assembly of \textit{A. agrestis} Bonn revealed some peculiarities in the hornwort genome structures. In particular, we could not locate the typical vascular plant centromeric regions, which are usually composed of highly duplicated tandem repeats of 100–1,000 base pairs (bp)\textsuperscript{25}. In \textit{A. agrestis} Bonn, tandem repeats with a unit size over 30 bp gave rise to only very short arrays, and these repeats do not show a clear spatial clustering (Supplementary Fig. 5). While gene density does fluctuate along the scaffolds, extensive regions with low gene density typical for centromeric regions of vascular plants were missing. Similarly, we could not identify stretches of scaffolds having an elevated repeat content (Fig. 1 and Supplementary Fig. 4), other than the putative telomeric regions. In other words, hornwort centromeres may not be characterized by a higher repeat density compared to other parts of the genome (see Supplementary Notes). Similar genome organizations were also discovered in the \textit{P. patens} genome where genes and repeats are evenly distributed along the chromosomes\textsuperscript{41}. While it is tempting to suggest that this genomic organization may be a shared feature of bryophyte genomes, we nevertheless cannot rule out the possibility that the bona fide centromeres were not sequenced or assembled properly despite the long-read and Hi-C data. Future work using immunolabelling is necessary to confirm this suggestion.

**Phylogenetic evidence for the monophyly of bryophytes**

To investigate the phylogenetic position of hornworts, we compiled 742 mostly single-copy genes from 21 genomes spanning major lineages of land plants and streptophyte algae. Monophyly of bryophytes is maximally supported in all our analyses, regardless of the data types (nucleotide or amino acid), tree inference methods (concatenation- or coalescent-based) and support measures (bootstrap or SH-aLRT or local posterior probability) (Fig. 2). In addition, over 50% of the gene-tree quartets are consistent with hornworts being a sister clade to liverworts and mosses (Fig. 2). Our results add to the growing evidence\textsuperscript{20} supporting two monophyletic groups of land plants: bryophytes and tracheophytes (vascular plants).

**Limited collinearity across bryophyte and vascular plant genomes**

A previous study on the moss \textit{P. patens} genome implied that regions showing collinearity between moss and some angiosperms may represent conserved collarinear blocks since the MRCA of land plants\textsuperscript{12}. However, comparing bryophytes to vascular plants, shared ancestral gene blocks could not be identified, rather that the collarinear regions with vascular plants were unique to each of the bryophyte genomes (Supplementary Fig. 6 and Supplementary Table 7). The most genomic blocks collagen with at least one other land plante plant were found in the moss, followed by the liverwort and hornwort genomes (Supplementary Fig. 6). Within bryophytes, no collarinear segment conserved across all three lineages was found, although there were genomic regions exclusively collarinear between each of two bryophyte genomes (Supplementary Fig. 6). In general, there was more collinearity between the liverwort and the moss than between the hornwort and the liverwort/moss genomes. The numbers of such collarinear regions, however, were small compared to those detected across vascular plants (Supplementary Fig. 6). Altogether, these findings imply that the deep divergence of the moss, hornwort and liverwort genomes may have led to limited collinearity among bryophytes, as well as between bryophytes and vascular plants.

**Absence of large-scale genome duplication in \textit{Anthoceros}**

Whole-genome duplications (WGD) have played an important role in shaping plant evolution and possibly underlie several adaptive radiations\textsuperscript{26}. A previous study, based on the number of synonymous substitutions per synonymous site (K\textsubscript{s}) divergence in transcriptomic datasets, suggested that hornworts may not have experienced any WGD event\textsuperscript{8}, similar to \textit{M. polymorpha} and \textit{Selaginella moellendorffii}\textsuperscript{8}. Our Ks plots on the annotated \textit{Anthoceros} genes similarly show no sign of WGD (Supplementary Fig. 7). To further corroborate this, we investigated patterns of intragenomic synteny in the three hornwort genomes, as well as the published \textit{M. polymorpha} and \textit{P. patens} genomes for comparison. We found very little
Fig. 1 | Genome organizations in the Anthoceros genomes. a, Circos plot of A. agrestis Bonn showing the densities of repeats, genes and single nucleotide polymorphisms (SNPs) with the A. agrestis Oxford genomes. b–e, Anthoceros genomes lack whole genome duplication. No self-synteny can be found in the three Anthoceros genomes (A. agrestis Bonn (b), A. agrestis Oxford (c) and A. punctatus (d)) nor in M. polymorpha (e). f, P. patens, on the other hand, shows a clear 1:1 and some 1:2 syntenic relationship, suggesting paleopolyploidy. In b–f, the bar graphs show the proportion of the genome at different self syntenic levels, with the dot-plots on the right.

Table 1 | Assembly statistics of the three hornwort genomes

|                | Estimated genome size (Mb) | Assembled genome size (Mb) | Contig/scaffold number | Contig/scaffold N50 length | Assembly approach         |
|----------------|---------------------------|----------------------------|------------------------|---------------------------|---------------------------|
| A. agrestis Bonn | 122–132                   | 116.9                      | 1577/322               | 155.5 kb/17.3 Mb         | Illumina + Nanopore + Hi-C |
| A. agrestis Oxford | 123–135                   | 122.9                      | 153/-                  | 1.8 Mb/-                 | Nanopore + Illumina       |
| A. punctatus     | 128–150                    | 132.8                      | 202/-                  | 1.7 Mb/-                 | Nanopore + Illumina       |
self-synteny in the hornwort genomes (Fig. 1b–d), providing strong evidence for the lack of WGD in Anthoceros. The high proportion of the genomes that are not syntenic is comparable to that in M. polymorpha (Fig. 1e). On the other hand, P. patens shows a clear 1:1 (and some 1:2) self-syntenic relationship (Fig. 1f), which is consistent with the earlier report and indicative of two rounds of WGD21.

Small repertoire of TAPs
We found that 2.4–2.6% of the proteomes of the three Anthoceros genomes were annotated as transcription-associated proteins (TAPs) (Fig. 3a and Supplementary Table 8). Compared to other land plants24, this is on the very low end of the spectrum both in terms of proportion and absolute number. Furthermore, about two-thirds (56) of the hornwort TAP families were smaller in size than in M. polymorpha. Given such a minimal TAP repertoire, hornworts can serve as an excellent baseline model to study the evolution and diversification of transcriptional networks. Despite its streamlined nature, some TAPs were only found in hornworts and vascular plants27, P. patens, providing strong sporophytic meristem activity in both land plants but not in the other two bryophyte genomes, with the most notable example being YABBY (Supplementary Table 8 and Supplementary Note). Such TAPs probably evolved in the MRCA of land plants but were lost in the mosses and liverworts. We also detected TAP families that were present in all streptophytes but lost either in the hornwort genomes (for example, SRS transcription factor, TF) or in M. polymorpha (for example type I MADS-box TF). Altogether, our findings suggest a dynamic TAP family turnover in the early evolution of land plants with multiple independent losses in different bryophyte lineages.

Genes related to sporophyte development
While hornworts have a gametophyte-dominant life cycle like other bryophytes, their sporophyte generation (Fig. 3b) shows several unique features25. First, after fertilization, the zygote division in hornworts is longitudinal, whereas zygotes in all other land plants undergo transverse division. Second, the hornwort sporophyte maintains a basal sporophytic meristem producing cells that continuously differentiate into mature tissues towards the tip. A common origin of indeterminate sporophyte development in hornworts and vascular plant shoot apical meristem (SAM) has been hypothesized26. Lastly, hornwort sporophytes have stomata (Fig. 3c) similar to mosses and vascular plants, and the basic regulation may be shared across all stomatous lineages of land plants26. Nevertheless, firm evidence supporting the homology of meristems as well as stomata is scarce. Here, we found that multiple genes critical for flowering plant SAM and stomata function have homologues in the hornwort genomes and are preferentially expressed in the sporophyte phase.

Class 1 Knotted1-like homeobox (KNOX1) genes regulate sporophytic meristem activity in both P. patens and vascular plants27, while Class 2 Knotted1-like homeobox (KNOX2) genes maintain sporophyte cell fate in P. patens28. Interestingly, the KNOX1 orthologue is lost in the Anthoceros genomes and only KNOX2 genes were found (Supplementary Fig. 8 and Supplementary Tables 8 and 9). The KNOX2 orthologues showed strong sporophyte-specific expression (Fig. 3d), which implies that the involvement of KNOX2 in maintaining sporophytic cell fate may be conserved in all land plants. Heterodimerization of KNOX1/KNOX2 and BELL-LIKE HOMEOBOX proteins is a deeply conserved molecular mechanism that is required for the KNOX functions28. We found that in A. agrestis Bonn, a single BELL and a single KNOX2 gene were specifically expressed in the sporophyte phase. Nevertheless, contrary to our expectations, the BELL gene was more strongly expressed in the early stages while the KNOX2 gene in the later stages of sporophyte development (Fig. 3d and Supplementary Tables 8 and 9). This suggests that hornwort sporophyte identity may not be determined by KNOX2 through interaction with BELL. Nevertheless, this hypothesis...
needs functional verification because partially overlapping expression of the KNOX2 and BELL genes does not exclude the possibility of heterodimerization.

*WUSCHEL*-related homeobox 13 like (WOX13L) genes are involved in zygotephyt development and stem cell formation in the moss *P. patens*. *A. thaliana* WOX13 promotes replum formation in the fruit (Fig. 3d) and WOX14 promotes vascular cell differentiation. The *Anthoceros* genomes have four WOX13L members (Supplementary Fig. 8 and Supplementary Tables 8 and 9) and WOX13La is specifically expressed in sporophytes while WOX13Lbc have expression at both gametophyte and sporophyte generations (Fig. 3d) and may have diverse roles in stem cell maintenance and sporophyte development. The *Anthoceros* genomes also have a single *FLORICAULA*/*LEAFY* (FLO/LEF) gene (Supplementary Fig. 8 and Supplementary Tables 8 and 9), which in *P. patens* and *A. thaliana* controls zygotephyt development and SAM maintenance, respectively. In hornworts, *LFY* is predominantly expressed in the gametophyte stages (Fig. 3d) while in *P. patens* it is expressed both in the gametophyte and the sporophyte. It is possible that such differences may contribute to the unique developmental pattern of hornwort sporophytes.

Stomatal development in *A. thaliana* and *P. patens* is regulated by a conserved genetic toolbox, including the basic helix–loop–helix (bHLH) transcription factors SMF (SPCH, MUTE and FAMA), ICE/SCREAMs (SCRM)s, EPIDERMAL PATTERNING FACTOR (EPF), ERECTA and TOO MANY MOUTHS (TMM) genes. FAMA in particular is involved in the final guard cell differentiation step and serves as the key switch. Orthologues of SMF, TMM and EPF were absent in *M. polymorpha*, consistent with the fact that liverworts do not have stomata. We found orthologues of FAMA (SMF), SCRM, ERECTA and EPF in TMM in the *Anthoceros* genomes (in line with a previous study based on our earlier genome draft; Supplementary Table 10 and Supplementary Fig. 9). SMF, SCRM, TMM and EPF showed sporophyte-specific expression patterns (Fig. 3d), suggesting that they may have similar roles in stomatic patterning in hornworts. While ERECTA was also expressed during early sporophyte development, its expression fluctuated between replicates and results were inconclusive. EPF expression showed similar inconsistency among replicates but did not influence our conclusion about its sporophyte-specific expression. In addition to EPF, an EPF-like gene in the EPFL4-6 clade, was found in hornworts (Supplementary Fig. 9), and is specifically expressed in gametophytes with a higher expression toward maturity and thus perhaps involved in a different cell–cell signalling other than stomatal regulation. EPF4 and EPF6 in *A. thaliana* are involved in coordination of the central and peripheral zone in SAM. Taken together, our data are consistent with a single origin of stomatal differentiation mechanism among all stomatous land plants, though positional determination may have evolved differently (Supplementary Notes).

**Genes related to phytohormone synthesis and signalling**

The *Anthoceros* genomes contain the genetic chassis for the biosynthesis and signalling of abscisic acid, auxin, cytokinin, ethylene and jasmonate (see Supplementary Notes, Supplementary Figs. 10–12 and Supplementary Table 10), reaffirming the origins of these pathways in the MRCA of land plants. Similar to *M. polymorpha* and *P. patens*, salicylic acid signalling components, but not the receptor-related genes, are found in hornworts. While DELLA is present, orthologues of gibberellin (GA) receptor GID1 and GA oxidases are missing from the *Anthoceros* genomes. This is consistent with the recent suggestion that DELLA was recruited to the GA signalling pathway later in plant evolution. Hornworts also possess enzymes to synthesize strigolactones but genes involved in strigolactone signalling are absent. This supports the idea that strigolactones are an ancient non-hormonal signal for rhizospheric communication with mycorrhizal fungi.

**Genetic network for arbuscular mycorrhizal symbiosis was present in the MRCA of land plants**

The symbiotic relationship with arbuscular mycorrhizal fungi (AMF) is one of the key innovations underlying the successful colonization and diversification on land of plants. Evidence of AMF can be found inside plant megafossils 407 million years ago and in almost all extant plant lineages (hornworts, liverworts and vascular plants). Recent genetic studies have identified a suite of genes in the angiosperms that regulate the establishment and maintenance of AMF symbiosis. Some of these genes are also required for...
legume–rhizobial interaction and are often referred to as the common symbiosis genes (CSGs) [41].

While a few components can be traced back to as far as chlorella alga (CSGs), the question of when exactly did the entire AMF symbiosis genetic network originate remains open. This is partly because both the bryophytes that have published genomes to date (P. patens and M. polymorpha) are incapable of AMF symbiosis and may have secondarily lost the symbiosis genes, as exemplified in some angiosperms [41]. Here, we show that all the key angiosperm AMF symbiosis genes have orthologues in the three hornwort genomes (Fig. 4, Supplementary Table 11 and Supplementary Fig. 13). Although their roles in hornwort AMF symbiosis remain to be tested, this result provides strong evidence that the genetic infrastructure required for AMF symbiosis was already present in the MRCA of land plants. Importantly, the presence of these genes in liverworts [42] and hornworts makes this conclusion insensitive to any uncertainty of the land plant phylogeny. We have not succeeded in reconstituting hornwort–AMF symbiosis in vitro and hence are unable to test expression of these orthologues in the context of AMF.

Nevertheless, we found that in both A. agrestis (Oxford strain) and A. punctatus, one of the AMF symbiosis genes, RAM1, was upregulated when plants were nitrogen-starved (Fig. 4). Nitrogen limitation is a major trigger for cyanobacteria symbiosis in hornworts, which might implicate the involvement of RAM1 in symbiosis, but further genetic studies are needed.

Genes related to cyanobacterial symbiosis
Symbiosis with nitrogen-fixing cyanobacteria is a rare trait, with limited appearances in a few plant lineages: bryophytes, Azolla (ferns), cycads (gymnosperms) and Gunnera (angiosperms) [31,42]. In bryophytes, although mosses frequently harbour epiphytic cyanobacteria [31], only hornworts and two liverwort species host cyanobacteria endophytically within specialized slime-filled cavities [42,43]. Amongst all the plant associations with cyanobacteria, most of the research has been done on hornworts, using A. punctatus (sequenced here) and the cyanobacterium Nostoc punctiforme as the study system. Although several cyanobacterial genes from N. punctiforme have been identified that are key to initiation of symbiotic association [45], nothing is known about the hornwort genetics. Here we generated RNA-seq data to compare the gene expression of symbiont-free (either nitrogen-starved or nitrogen-fed) and symbiosis-reconstituted hornworts (Fig. 4). This experiment was conducted with both A. punctatus and the A. agrestis Oxford isolate. We identified 40 genes that, when the cyanobionts are present, are highly induced (>16-fold) in both hornwort species (Fig. 4 and Supplementary Table 12). These include a number of receptor kinases, transcription factors and transporters. Of particular interest is a SWEET sugar transporter in the SWEET16/17 clade (Fig. 4 and Supplementary Fig. 14), which is minimally transcribed under the symbiont-free states but is among the highest expressed genes in symbiosis (>105 fold-change). The upregulation of SWEET in symbiosis is interesting because it implies that this sugar transporter is dedicated to supplying carbon rewards to the cyanobionts. This implication is supported by the fact that only exogenous glucose, fructose or sucrose sustained dark nitrogen fixation in the A. punctatus–N. punctiforme association [44] and by the observation that inactivation of a carbohydrate permease in N. punctiforme resulted in a defective symbiotic phenotype [45]. In parallel, SWEET is involved in mycorrhizal symbiosis as well, but a different orthologue, in the SWEET1 clade, was recruited [46].

Another gene of interest is subtilase. Members of this gene family have been shown to be highly upregulated in a wide variety of microbial symbioses, including rhizobial [41], mycorrhizal [42] and actinorhizal [31–33] interactions. RNA interference knockdown of a subtilase (SBT1M1) in the legume Lotus japonicus resulted in a decreased arbuscle formation [42]. Here, we found that in both A. punctatus and A. agrestis, a subtilase homologue was similarly induced by cyanobacteria symbiosis. Phylogenetic reconstruction showed that this hornwort subtilase is not orthologous to those involved in other plant symbioses (Supplementary Fig. 15). Taken together, our results imply that hornworts might have convergently recruited SWEET and subtilase for cyanobacterial symbiosis, although in both cases not the same orthologues were used as in other plant–microbe symbioses.

Pyrenoid-based CCM
To enable a more efficient photosynthesis, hornworts, cyanobacteria and many eukaryotic algae have evolved biophysical CCM inside their cells (cyanobacteria) or individual chloroplasts [34]. Algal and hornwort chloroplasts use inorganic carbon transporters and carbonic anhydrases to locally concentrate CO2 in the pyrenoids, a specialized chloroplast compartment where RuBisCOs aggregates. Pyrenoids can thus boost photosynthetic efficiency and reduce photorespiration. Such pyrenoid-based CCM has been extensively studied in the model green alga Chlamydomonas reinhardtii with the hope of installing a CCM in crop plants [36].

Hornworts are the only land plants with a pyrenoid-based CCM. Interestingly, for the past 100 million years, pyrenoids in hornworts are inferred to have been repeatedly lost and gained [37,38], suggesting that pyrenoid development and function is controlled by a few master switches. The genetics behind hornwort pyrenoids, however, has remained completely unknown. We explored whether hornwort genomes have genes that are known to be required for pyrenoid-based CCM in C. reinhardtii. While many of the C. reinhardtii CCM genes do not have clear homologues in hornworts (nor in any other land plants), we did find LCIB (low CO2 inducible B) to be present in the hornwort genomes and hornwort transcriptomes of the 1,000 plant transcriptomes project (1KP) [39]. Apart from hornworts, no LCIB homologue could be found in other plant genomes sequenced to date. The uniquely shared presence of LCIB in pyrenoid-bearing algae and hornworts implies that LCIB might have a role in the hornwort CCM. The phylogenetic tree indicates that the hornwort LCIBs form a sister clade to the Klebsormidium nitens homologue (Fig. 5) and thus is consistent with the organisms tree with many losses in various lineages. In this scenario, the MRCA of land plants had LCIB.

In C. reinhardtii, LCIB gene expression is highly induced by CO2 limitation and the encoded proteins localize around pyrenoids to presumably block CO2 leakage [39,40]. All the hornwort LCIB sequences have the conserved amino acid residues at the active sites that are shared with other algal LCIBs [41] (Fig. 5b). However, unlike C. reinhardtii, we did not find LCIB to be differentially expressed when plants are grown at different CO2 levels (Supplementary Fig. 12). This, nevertheless, cannot rule out the involvement of LCIB in CCM because hornwort CCM was reported to be constitutively expressed and not regulated by CO2 level [42]. Whether LCIB homologues have a similar function and localization in hornworts remains to be experimentally tested.

Discussion
The hornwort genomes presented here offer a unique window into the biology of land plant MRCA. For example, the Anthoceros genomes lack KNOXI, while P. patens and M. polymorpha lack YABBY genes. This suggests that the MRCA of land plants had both of these key developmental genes and independent gene losses occurred in different bryophyte lineages. While LEAFY expression is predominantly in the gametophyte stage, YABBY, KNOX2, BELL and some WOX3L genes are up-regulated in the hornwort sporophytes (Fig. 3). In addition, several stomata-related genes are present in the Anthoceros genomes and expressed in early sporophyte development (Fig. 3), implying a homology of stomata at the genetic level. Finally, we found that the genes required for AMF symbiosis
are conserved in Anthoceros (Fig. 4), providing evidence that the MRCA of land plants was already equipped with the genetic network for AMF symbiosis. In-depth analysis on the evolution of the plant hormones (abscisic acid, auxin, gibberellin, jasmonate, salicylic acid and strigolactone), light signalling, peptidoglycan synthesis and chloroplast development can be found in Supplementary Notes.

**Fig. 4 | Evolution and genetics of symbiosis in hornworts.** a, Orthologues of AMF symbiosis pathway genes can be found in hornworts, indicating their presence in the common ancestor of land plants. The asterisk indicates that the M. paleacea transcriptome was searched instead of M. polymorpha genome because the latter secondarily lost AMF. b, RAM1 is upregulated during nitrogen starvation in both A. agrestis and A. punctatus. c, Reconstituted Anthoceros-cyanobacteria symbiosis. Arrowhead points to a cyanobacteria colony. d, Transcriptomic responses to nitrogen starvation and cyanobacterial symbiosis in A. agrestis (n = 9 biologically independent samples). PC1 and PC2 refer to the first and second axes of principal component analysis on gene expression values. e, A suite of genes were highly upregulated under symbiosis in both A. agrestis and A. punctatus (two-sided test for differential expression, false-discovery rate ≤ 0.05 and log₂-fold-change > 4).
The Anthoceros genomes shared several features with the two other published bryophyte genomes. Most notable is the absence of tandem repeats that make up the typical centromeric regions. Further studies are needed to identify the centromeric regions and understand their structure. While P. patens has experienced two rounds of WGD\(^{15}\), none can be found in Anthoceros. This might explain the minimal representation of transcription factors in the last two genomes.

Furthermore, our functional genomic data shed light on the genetic framework that underpins features that are unique to hornworts. We identified a suite of candidate genes underlying hornwort–cyanobacteria symbiosis (Fig. 4). This includes a SWEET transporter that might be involved in nutrient transfer with the cyanobionts. A well-characterized C. reinhardtii CCM gene, LCIB, was conserved in hornworts but apparently lost in all other plant lineages (Fig. 5). Whether LCIB also participates in hornwort CCM awaits future functional characterization.

The recent advances of ‘seed-free genomics’ have greatly improved our understanding of streptophyte evolution\(^1,2,11,13,17,63-66\). Here, our hornwort genomes fill in yet another critical gap and are beginning to illuminate the dawn of land plants as well as the unique biology of hornworts.

**Methods**

**Plant materials.** Cultures of A. agrestis (Oxford and Bonn strains) and A. punctatus were all derived from a single spore and axenically propagated and maintained on either RCD\(^{16}\) or Hatcher’s media. Supplementary Table 13 shows the origin and specimen voucher for each of the three strains.

**Chromosome count.** The tip of an A. agrestis Oxford gametophyte thallus was cut into small pieces and fixed with 4% glutaraldehyde in 0.05 M phosphate buffer (pH 7.0) for 12 h at 4 °C. After washing with the buffer for 10 min, cell walls were digested for 2 h with a solution containing 1% Driselase (Sigma-Aldrich), 1% Cellulase Onozuka RS (Yakult), 1% Pectolyase (Kikkoman), 0.5% IGEPA-L CA-630 and 1% bovine serum albumin (BSA) at 30 °C. After several washes with the buffer, the samples were incubated in 0.05 M phosphate buffer containing 0.1% Triton-X-100 for 12 h at 4 °C. After several washes with the buffer, the samples were transferred onto MAS coated slide glasses (Matsunami Glass) and coverslipped. The slides were then pressed with a thumb directly over the coverslip. After removal of the coverslip, the slides were air-dried for 10 min at room temperature and then extracted with methanol at −20 °C for 10 min. After the staining with the buffer containing 1 μg 4,6-diamidino-2-phenylindole (DAPI) for 5 min, the slides were mounted with Vectashield mounting medium (Vector Laboratory Burlingame) and observed with a fluorescence microscope under ultraviolet-light excitation.

**DNA sequencing.** Hornwort DNA was extracted using a CTAB-precipitation method modified from ref. \(^70\). Nanopore libraries were prepared by SQK-LSK108 and sequenced on MinION R9 flow cells for 48 h. Basecalling was done by Albacore.

For A. agrestis Bonn, the TrueSeq DNA Nano Kit (Illumina) was used to prepare paired-end (PE) sequencing libraries which were sequenced (PE 150 bp) on HiSeq4000 at the Functional Genomic Center Zurich (FGCZ). For A. agrestis Oxford 251 PE reads, a PCR-Free library was prepared using a KAPA Hyper Prep Kit according to the protocol published by Broad Institute\(^{70}\). The library was mixed (3%) with other barcoded libraries and sequenced on Illumina HiSeq1500 (two lanes with Rapid mode; OnBoardClustering) at the National Institute of Basic
Biologically, for *A. punctatus*, Illumina genomic libraries were prepared by BGI and sequenced on HiSeq4000. Read quality and adapter trimming was done by Iasp [1] with the default setting. For *A. agrestis* Bonn, additional Chicago and Hi-C libraries were prepared by Oxford Nanopore Technologies. A total of two *Chicago* libraries and one Hi-C library were prepared with a physical coverage of 300× and 200×.

To calculate the read mapping rates, trimmed reads were mapped to the final assemblies using bwa mem -M [2] and sorted with sortools [3]. The mean insert size and #READPAIRS were calculated using picard CollectInsertSizeMetrics. Unmapped reads were counted with samtools view -c [4] (ref. 5) and divided with the total number of reads to calculate percentage mapped. High-quality mapped reads were counted with -q 20. Reads mapped to chloroplast and mitochondrial genomes were counted with samtools [6]. The bam files were assessed with qualimap v2.2.1 [7] (ref. 8) bamqc and observed error rates (total, mismatch, insertions and deletions), as well as the genome coverage were recovered.

**Genome assembly.** Genomes sizes for the three *Anthoceros* were estimated based on k-mer distribution by Jellyfish [9] in conjunction with GenomeScope [10]. Draft assembly for *A. agrestis* Bonn strain was first generated using a hybrid approach including Oxford nanopore (60×) and Illumina paired-end reads (150×) using Miniasm [11] and SSPACE [12] (ref. 5) and divided with the two rounds of Pilon polishing [13]. We mapped Chicago and Hi-C reads back to the draft assembly and used DoveTail’s HiRise assembler v.2.1.2 (ref. 14) for scaffolding. Contigs of the draft assembly were first scaffolded using the Chicago library to correct smaller scale errors and improve contiguity. Finally, the output assembly was further scaffolded using the Hi-C libraries and DoveTail’s HiRise assembler v.2.1.2 (ref. 14) to derive the final assembly.

Genome assemblies of *A. agrestis* Oxford strain and *A. punctatus* were generated with the minimap2-miniasm assembly [15] using only the nanopore reads. We then used four iterations of minimap2-rac0 [16] to derive the consensus sequence, followed by six rounds of Pilon polishing [17].

**Contamination removal.** While our cultures were grown in a putative axenic condition, low level of contamination cannot be completely ruled out. We therefore used bloopools [18] to identify scaffolds/contigs primarily consisting of contaminating sequences. The Hi-C library theoretically should sort DNA sequences originating from different organisms because cross-linking occurred within the nuclei. Therefore, we hypothesized that dropping scaffolds mainly with non-sterotrophyle affiliation will effectively remove contaminants from our assembly. For *A. agrestis* Bonn, we used both the full uniprot and the National Center for Biotechnology Information (NCBI) nucleotide database and bloopools to assign the taxonomic affiliation of scaffolded reads. For *A. punctatus*, some of the small scaffolds were classified as of non-sterotrophyle origin with high confidence; these scaffolds were then removed from the assembly. For the *A. agrestis* Oxford and *A. punctatus* genomes, assemblies were contamination-filtered in a similar way.

The detailed summary can be found in Supplementary Table 14.

**RNA-seq dataset and analysis on developmental stages.** To study the expression pattern of transcription factor genes across developmental stages, we generated RNA-seq libraries for the following stages of the *A. agrestis* Bonn strain in two biological replicates: (1) spores after 2 weeks of germination, (2) 4-week-old gametophytes, (3) 2-month-old gametophytes, (4) sporophytes shorter than 5 mm, (5) sporophytes 5–10 mm (6) sporophytes 10–20 mm (7) fronds with brown or black tips. Plants were grown on agar plates containing BCD medium [26, 27] at 22 °C. RNA was extracted with the Spectrum Total RNA Plant Kit (Sigma-Aldrich). The Illumina libraries were prepared by BGI and sequenced on HiSeq4000. One of the low CO2 samples failed to produce high-quality library, and as a result the low CO2 condition has only two replicates. RNA-seq dataset and analysis on CO2 response.

**Gene prediction.** We used RNA-seq, transcript and protein evidence to train Augustus (ref. 28) within Braker2 [29]. Because the resulting gene models were heavily dependent on the training data, we decided to generate multiple gene predictions and build consensus gene models using EvidenceModeler (EVM) [30]. We used both individual approaches (prediction of genes for each genome separately, see (1)–(5) below) and comparative (simultaneous prediction of gene models for the genomes, see (5) below) approaches to increase the accuracy and compatibility of gene annotations. Comparative genome annotation approaches use whole-genome alignment and external evidence (RNA-seq, protein and expressed sequence tag) to simultaneously predict genes in multiple genomes and are able to correct errors may arise during individual-based predictions. The following gene prediction approaches were used. (1) We trained Augustus with only the RNA-seq evidence and predicted gene models by taking into account RNA-seq, protein, Mikado and PASA assembled transcripts. (2) We used the previously trained model (in (1)) species model but with a modified weighting file (extrinsic.cfg) to give more weight to the protein evidence. (3) We trained Augustus using both protein and RNA-seq evidence within Braker2 (EPT mode of Braker2). (4) We used the RNA-seq evidence to automatically train genemark and obtain gene predictions. (5) Finally, we ran Augustus in the comparative mode with RNA-seq, transcript and protein evidence and genome alignments inferred by maps [31]. Generating this series of genome predictions was necessary as our preliminary analyses suggested that none of the predictions was superior but rather complementary. The proteomes used can be found in Supplementary Table 15.

**Generating consensus gene models.** We used EVM to derive consensus gene models best supported by the various evidence. We used all the previously generated gene predictions (gff files) and selected the best consensus gene models using protein (exonerate-mapped proteomes of species and the uniprot_sport plant dataset) and transcript evidence (Mikado and PASA assembled transcripts). We gave equal weights to each ab initio predictions, transcript evidence (weight 1), but increased the weight for Mikado loci (2) and PASA assembled transcripts (10). After deriving the consensus gene models, we used PASA and the PASA assembled transcripts to correct erroneous gene models, add UTRs (untranslated regions), and predict alternative splice variants in two rounds. Finally, we extensively manually curated these three annotations (revised and corrected various gene models) and generated genes for all further downstream analyses.

Our annotation pipeline resulted about 1,000 more predicted gene models for the *A. agrestis* Bonn compared to the Oxford strain. This suggested that despite high collinearity, gene content of the two strains may differ. To aid future comparative analyses we created an *A. agrestis* pan genome containing a non-redundant set of genomic sequences and annotations of the two strains.
Furthermore, we carefully analysed the predicted gene set of the two strains to show that gene number difference is not due to annotation issues. Methods and results of the pan-genome construction as well as gene set comparison can be found in the Supplementary notes.

Genome completeness assessment. We used BUSCO v.3 (ref. 100) with the Viridiplantae set to assess the completeness of our genomes and annotations. We did not use the Embryophyta set because it was constructed based almost exclusively on flowering plant genomes (29 out of 30)103, which does not offer an appropriate benchmark for non-flowering plant genomes. Supplementary Fig. 3 shows that our genomes have similar (if not better) BUSCO scores compared to many published non-flowering plant genomes. It should be noted that while Physcomitrella, Sphenophyllum and Marchantia all have much higher BUSCO scores, this is probably reflecting the fact that these genomes were used to compile the Viridiplantae set.

Reconstructing the land plant phylogeny. We used Orthofinder2 (ref. 104) to identify mostly single-copy genes, with 21 genomes (Fig. 2) included in the run to represent angiosperms, ferns, lycophytes, mosses, liverworts, hornworts and the grade of streptophyte algae. A total of 742 mostly single-copy orthogroups were identified. Protein alignments for individual orthougroup were done by MAFFT v.7.427 (ref. 105) and back translated to nucleotides by TranslatorX106. We translated DNA sequences and identified reading frames by comparing the proteomes of 19 species using Orthofinder2 (ref. 104). The dataset included all the 742 loci and was used as a human dataset to visually assess collinearity of the three genome assemblies. We also aligned the genomic sequences using the nucler module of mummer111 and assessed their differences using Assemblytics112.

To study the collinearity across all plants, we first created orthogroups with proteomes of 19 species using Orthofinder2 (ref. 104). The dataset included representatives from each major groups of land plants (Supplementary Table 15), and species experienced different numbers of large-scale duplication events113. Gff files and proteomes were retrieved from Phytozome v.12 (ref. 118). We used BUSCO v.3 (ref. 103) with the Viridiplantae set to assess the completeness of our genomes and annotations. We did not use the Embryophyta set because it was constructed based almost exclusively on flowering plant genomes (29 out of 30)103, which does not offer an appropriate benchmark for non-flowering plant genomes. We used BUSCO v.3 (ref. 103) with the Viridiplantae set to assess the completeness of our genomes and annotations. We did not use the Embryophyta set because it was constructed based almost exclusively on flowering plant genomes (29 out of 30)103, which does not offer an appropriate benchmark for non-flowering plant genomes. We used BUSCO v.3 (ref. 103) with the Viridiplantae set to assess the completeness of our genomes and annotations. We did not use the Embryophyta set because it was constructed based almost exclusively on flowering plant genomes (29 out of 30)103, which does not offer an appropriate benchmark for non-flowering plant genomes. We used BUSCO v.3 (ref. 103) with the Viridiplantae set to assess the completeness of our genomes and annotations. We did not use the Embryophyta set because it was constructed based almost exclusively on flowering plant genomes (29 out of 30)103, which does not offer an appropriate benchmark for non-flowering plant genomes. We used BUSCO v.3 (ref. 103) with the Viridiplantae set to assess the completeness of our genomes and annotations. We did not use the Embryophyta set because it was constructed based almost exclusively on flowering plant genomes (29 out of 30)103, which does not offer an appropriate benchmark for non-flowering plant genomes. We used BUSCO v.3 (ref. 103) with the Viridiplantae set to assess the completeness of our genomes and annotations. We did not use the Embryophyta set because it was constructed based almost exclusively on flowering plant genomes (29 out of 30)103, which does not offer an appropriate benchmark for non-flowering plant genomes. We used BUSCO v.3 (ref. 103) with the Viridiplantae set to assess the completeness of our genomes and annotations. We did not use the Embryophyta set because it was constructed based almost exclusively on flowering plant genomes (29 out of 30)103, which does not offer an appropriate benchmark for non-flowering plant genomes. We used BUSCO v.3 (ref. 103) with the Viridiplantae set to assess the completeness of our genomes and annotations. We did not use the Embryophyta set because it was constructed based almost exclusively on flowering plant genomes (29 out of 30)103, which does not offer an appropriate benchmark for non-flowering plant genomes. We used BUSCO v.3 (ref. 103) with the Viridiplantae set to assess the completeness of our genomes and annotations. We did not use the Embryophyta set because it was constructed based almost exclusively on flowering plant genomes (29 out of 30)103, which does not offer an appropriate benchmark for non-flowering plant genomes. We used BUSCO v.3 (ref. 103) with the Viridiplantae set to assess the completeness of our genomes and annotations. We did not use the Embryophyta set because it was constructed based almost exclusively on flowering plant genomes (29 out of 30)103, which does not offer an appropriate benchmark for non-flowering plant genomes. We used BUSCO v.3 (ref. 103) with the Viridiplantae set to assess the completeness of our genomes and annotations. We did not use the Embryophyta set because it was constructed based almost exclusively on flowering plant genomes (29 out of 30)103, which does not offer an appropriate benchmark for non-flowering plant genomes. We used BUSCO v.3 (ref. 103) with the Viridiplantae set to assess the completeness of our genomes and annotations. We did not use the Embryophyta set because it was constructed based almost exclusively on flowering plant genomes (29 out of 30)103, which does not offer an appropriate benchmark for non-flowering plant genomes. W
WOX sequences, then aligned with einsi —maxiterate 1,000 in MAFFT v.7.429 (ref. 18). The alignment was manually inspected with Mesquite v.3.6 (ref. 18) and a matrix consisting of 38 included sites of 142 sequences was constructed. Sequences identical in the included region were treated as a single operational taxonomic unit (OTU) during the phylogenetic analysis. The best-fitting model was chosen as JJT with ProteinModelSelection.pl. The maximum likelihood (ML) tree was inferred by RAxML10 with -f a -r 100 -m PROTGAMMAJTT including and -x from random number generator. Bootstrap samples were generated with seqboot from PHYLIP package v.3.697 (ref. 18) and RAxML was run for each of them.

For YABBY, the 107 OTU dataset from Finet et al.135 was downloaded from treebase and combined with YABBY genes from Huperzia and Anthoceros. The sequences were aligned using einsi of MAFFT v.7.450 (ref. 105). The alignment was manually inspected with MacClade 4.08 and 123 well-conserved sites were included to result in alignment of 66 sequences. The sequences identical in the included region were treated as a single OTU during the phylogenetic analysis. The best-fitting model was chosen as JJTTDCMUTF with ProteinModelSelection8.pl. The ML tree was inferred by RAxML with -r 100 -m PROTGAMMAJTTDCMUTF including and -x from random number generator. Bootstrap samples (1,000 replicates) were generated with seqboot from PHYLIP package v.3.697 (ref. 18) and RAxML was run for each of them.

Identification of orthologues to AMF symbiosis genes. Homologues to symbiotic genes were retrieved in 31 species covering the different plant lineages (Supplementary Table 1) using protein from the model plant Medicago truncatula and the iBLASTn v.2.9.0 + (ref. 30) with a threshold e-value of 1e-05. Sequences were aligned using MAFFT v.7.470 (ref. 18) with default parameters and alignments were cleaned using TrimAl v.1.4 (ref. 4) to remove positions with more than 20% of gaps. Resulting alignments were subjected to ML tree inference using IQ-TREE (v.1.6.1.6) with 1,000 replicates of UltraFast Bootstraps12.

Phylogenetic reconstruction of LCIB. The orthogroup OG0009668 was identified as the LCIB gene family containing C. reinhardtii LCIB-E genes. Additional hornwort LCIB homologues were retrieved from the 1,000 plant transcriptome database. To find other LCIB homologues, we ran BLASTp against the Physcomitrella database using both the Anthoceros and C. reinhardtii sequences as the query and no hit could be obtained. Gene phylogeny was reconstructed on the basis of the amino acid alignment done by MUSCLE13. IQ-TREE v.1.6.1.6 (ref. 18) was used to obtain the ML tree as outlined above.

Data availability
All three genomes and their annotations can be accessed, browsed, searched and downloaded at https://www.hornworts.uzh.ch/en.html. All the raw sequences are deposited in the NCBI Sequence Read Archive under the BioProject PRJNA574124 and PRJNA574453, and to European Nucleotide Archive (ENA) under the study accession PRJEB34763 and PRJEB34743 (Supplementary Tables 2 and 3). The genome assemblies, annotations (Submitted.zip) as well as alignment matrices and tree files (phylogeny_dataset.zip) can be found at Figshare: https://doi.org/10.6084/m9.figshare.9974999.

References
1. Morris, J. L. et al. The timescale of early land plant evolution. Proc. Natl Acad. Sci. USA 115, E2274–E2283 (2018).
2. Nishiyama, T. et al. Chloroplast phylogeny indicates that bryophytes are monophyletic. Mol. Biol. Evol. 21, 1813–1819 (2004).
3. Wickett, N. J. et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl Acad. Sci. USA 111, E4859–E4868 (2014).
4. Putton, M. N. et al. The interrelationships of land plants and the nature of the ancestral bryophyte. Curr. Biol. 28, 733–745 (2018).
5. de Sousa, F., Foster, P. G., Donoghue, P. C. J., Schneider, H. & Cox, C. J. Nuclear protein phylogenies support the monophyly of the three bryophyte groups (Bryophyta Schimp). New Phytol. 222, 565–575 (2019).
6. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. Nature 574, 679–688 (2019).
7. Rensing, S. A. et al. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science 319, 64–69 (2008).
8. Bowman, J. L. et al. Insights into land plant evolution garnered from the Marchantia polymorpha genome. Cell 171, 287–299 (2017).
9. Renzagla, K. S. Comparative morphology and developmental anatomy of the Anthocerotophyta. J. Hattori Bot. Lab. 44, 31–90 (1978).
10. Smith, E. C. & Griffiths, H. A pyrenoid-based carbon-concentrating mechanism is present in terrestrial bryophytes of the class Anthoceraceae. Planta 200, 203–212 (1996).
11. F.-W. Villarreal Aguilar, J. C. & Szövényi, P. Horwnorts: an overlooked window into carbon-concentrating mechanisms. Trends Plant Sci. 22, 275–277 (2017).
12. Qu, Y.-L. et al. The deepest divergences in land plants inferred from phylogenomic evidence. Proc. Natl Acad. Sci. USA 103, 15511–15516 (2006).
13. Renzagla, K. S., Villarreal Aguilar, J. C., Piatkowski, B. T., Lucas, J. R. & Merced, A. Hornwort stoma: architecture and fate shared with 400-Million-year-old fossil plants without leaves. Plant Physiol. 174, 788–797 (2017).
14. Renzagla, K. S., Villarreal, J. C. & Duff, R. J. in Bryophyte Biology Vol. 2 (eds Goffinet, B. & Shaw, J.) 139–171 (Cambridge Univ. Press, 2009).
15. Meeks, J. C. Symbiosis between nitrogen-fixing cyanobacteria and plants. Bioscience 48, 266–276 (1998).
16. Meeks, J. C. Physiological adaptations in nitrogen-fixing Nostoc–plant symbiotic associations. Microbiol. Monogr. 8, 181–205 (2009).
17. Szövényi, P. et al. Establishment of Anthoceros agrestis as a model species for studying the biology of hornworts. BMC Plant Biol. 15, 98 (2015).
18. Bannard, J. D. & Villarreal Aguilar, J. C. Genome size increases in recently diverged hornwort clades. Genome 56, 431–435 (2013).
19. Hornworts (Anthocerotophyta). University of Zurich https://www.hornworts.uzh.ch/en.html (2020).
20. Jiang, J., Birchler, J. A., Parrott, W. A. & Dawe, R. K. A molecular view of plant centromeres. Trends Plant Sci. 8, 570–575 (2003).
21. Lang, D. et al. The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. Plant J. 93, 515–533 (2018).
22. Landis, J. B. et al. Impact of whole-genome duplication events on diversification rates in angiosperms. Am. J. Bot. 105, 348–363 (2018).
23. Banks, J. A. et al. The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. Science 332, 960–963 (2011).
24. Wilhelmsson, P. K. I., Mühlich, C., Ullrich, K. K. & Rensing, S. A. Comprehensive genome-wide classification reveals that many plant-specific transcription factors evolved in streptophyte algae. Genome Biol. Evol. 9, 3384–3397 (2017).
25. Ligrone, R., Duckett, J. G. & Renzagla, K. S. The origin of the sporophyte shoot in land plants: a bifurational perspective. Am. J. Bot. 110, 935–941 (2012).
26. Chater, C. C. C., Caine, R. S., Fleming, A. J. & Gray, J. E. Origins and evolution of stomatal development. Plant Physiol. 174, 624–638 (2017).
27. Courdert, Y., Novák, O., Dossat, C. J. & Novák, O. A KNIX-cytokinin regulatory module preads the origin of indeterminate vascular plants. Current Biology 29, 2743–2750 (2019).
28. Sakakibara, K. et al. KNOX2 genes regulate the haploid-to-diploid morphological transition in land plants. Science 339, 1067–1070 (2013).
29. Arun, A. et al. Convergent recruitment of TALE homeodomain life cycle regulators to direct sporophyte development in land plants and brown algae. Elife 8, e43101 (2019).
30. Sakakibara, K. et al. WOX13-like genes are required for reprogramming of leaf and protoxylem cells into stem cells in the moss Physcomitrella patens. Development 141, 1660–1670 (2014).
31. Romanzo-Branchat, M., Ripoll, J. I., Yanovsky, M. F. & Pelaz, S. The WOX 13 homeobox gene promotes repeated formation in the Arabidopsis thaliana fruit. Plant J. 73, 37–49 (2013).
32. Denis, E. et al. WOX14 promotes bioactive gibberellin synthesis and vascular cell differentiation in Arabidopsis. Plant J. 90, 560–572 (2017).
57. Rae, B. D. et al. Progress and challenges of engineering a biophysical
56. Meyer, M. T., Whittaker, C. & Griffiths, H. The algal pyrenoid: key
54. Ribeiro, A., Akkermans, A. D., van Kammen, A., Bisseling, T. & Pawlowski,
52. Takeda, N., Sato, S., Asamizu, E., Tabata, S. & Parniske, M. Apoplastic plant
48. Steinberg, N. A. & Meeks, J. C. Physiological sources of reductant for
43. Parniske, M. Arbuscular mycorrhiza: the mother of plant root
40. Remy, W., Taylor, T. N., Hass, H. & Kerp, H. Four hundred-million-year-old
37. Nishiyama, T. et al. The
36. Kosentka, P. Z., Overholt, A., Maradiaga, R., Mitoubsi, O. & Shpak, E. D.
59. Wang, Y. & Spalding, M. H. LCIB in the
60. Atkinson, N. et al. Introducing an algal carbon‐concentrating mechanism
270
3717–3737 (2017).
50. Puchta, H. Fungi filamentous microorganisms associated with plants
365–374 (2006).
89. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess
278, 2114–2120 (2014).
408, 764–770 (2011).
67. Weissenfeld, N. I. et al. Comprehensive variation discovery in single
452, 344–350 (2015).
517–518 (2014).
174, 122–128 (2018).
62. Sasaki, H., Ishii, T., Kikawa, K., Hara, H. & Kikuchi, Y. A novel subtilase
550 (2014).
576–585 (2003).
60. Atkinson, N. et al. Introducing an algal carbon‐concentrating mechanism
140, 1302–1315 (2016).
61. Jin, S. et al. Structural insights into the LCIB protein family reveals a new
37–47 (2017).
1416–1420 (2018).
1047–1058 (2008).
638–642 (2018).
546–553 (2018).
133, 722–727 (2014).
542–556 (2018).
122–128 (2018).
219, 1018–1030 (2018).
17, 1085–1100 (2018).
13, 7324–7329 (1991).
4. Nistor, P. et al. Cellular dedifferentiation and reprogramming of the root
16179 (2016).
124. Tang, H. et al. Unraveling ancient hexaploidy through multiply-aligned transcriptome assemblies. *Nucleic Acids Res.* 31, 5663–5666 (2003).

125. Venturini, L., Cajus, S., Thakalottal, G., Maplestone, D. L. & Swarbrick, D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *Gigascience* 7, giy093 (2018).

126. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinf.* 6, 31 (2005).

127. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 33, W465–W467 (2005).

128. Hoef, K. J., Lange, S., Lamosadze, A., Borodovsky, M. & Stanke, M. BRAKER1: unsupervised RNA-seq-based genome annotation with GeneMark-ES. *Bioinformatics* 32, 767–769 (2016).

129. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVIDENCEModeler and the program to assemble spliced alignments. *Genome Biol.* 9, R7 (2008).

130. Angiuoli, S. V. & Salzberg, S. L. M. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* 27, 334–342 (2011).

131. 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).

132. Gence, M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. * Genome Biol.* 20, 238 (2019).

133. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. * Mol. Biol. Evol.* 30, 772–780 (2013).

134. Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* 38, W7–W13 (2010).

135. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. * Mol. Biol. Evol.* 32, 268–274 (2015).

136. 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).

137. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35, 518–522 (2018).

138. 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).

139. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Biol.* 19, 153 (2015).

140. Sayyari, E. & Mirarab, S. Fast coalcecent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* 33, 1654–1668 (2016).

141. Sayyari, E., Whitfield, J. B. & Mirarab, S. DiscoVista: interpretable visualizations of gene tree discordance. * Mol. Phylogenet. Evol.* 122, 110–118 (2018).

142. Cabanetès, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* 6, e4958 (2018).

143. Marçais, G. et al. MUMmer: a fast and versatile genome alignment system. *PLoS Comput. Biol.* 14, e1005944 (2018).

144. Nattestad, M. & Schatz, M. C. Assemblies: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* 32, 3021–3023 (2016).

145. Qiao, X. et al. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. * Genome Biol.* 20, 38 (2019).

146. Goodstein, D. M. et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186 (2012).

147. Pratola, S. et al. ANViz 3.0—fast and sensitive detection of genomic polymorphism in extremely large data sets. * Nucleic Acids Res.* 40, e112 (2012).

148. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573 (1999).

149. Barker, M. S. et al. EvoPipes: bioinformatic tools for ecological and evolutionary genomics. *Evol. Bioinform.* 6, 143–149 (2010).

150. Birney, E., Clamp, M. & Durbin, R. Genevestigator and genowise. * Genome Res.* 14, 988–995 (2004).

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

152. Tang, H. et al. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene models. *Genome Res.* 18, 1944–1954 (2008).

153. Mitchell, A. L. et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* 47, D351–D360 (2019).

154. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* 47, D427–D432 (2019).
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41477-020-0618-2.

Correspondence and requests for materials should be addressed to F.-W.L. or P.S.

Peer review information Nature Plants thanks Burkhard Becker and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Content of the page:

**Reporting Summary**

Nature Research 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 Research policies, see Authors & Referees 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](https://natureresearch.com/guidelines/statistics) contains articles on many of the points above.

**Software and code**

Policy information about availability of computer code

**Data collection**

No software was used. We constructed paired-end DNA sequencing libraries with approx. 400 bp insert sizes for standard WGS sequencing using Illumina HiSeq and Novaseq machines. We also prepared RNA-seq libraries and sequenced them on Hiseq and Novaseq machines. We generated long-reads using high-molecular weight DNA on the Oxford Nanopore Minion machine using R9 flow cells.

**Data analysis**

Trimmomatic v0.39, samtools v1.9, bwa v0.7.17, Albacore, v2.3.3, Jellyfish v1.1.12, Genomescop, MaSuRCAv3.2.8, pilon v1.22, HiRise assembler v2.2.12, minimap2 v2.14-r883, miniasm v0.3-r179, racon v1.3.1, blobtools v1.1.1, Salmon v0.13.1, Deseq2 v3.9, HiSat2 v2.1.0, stringtie v2.0.3, BINGO v3.0.3, REVIKO, RepeatModeler v1.0.11, LTR_retriever v2.0, RepeatMasker v4.0.8, portcullis v1.1.2, Trinity v2.8.4, PASA v2.3.3, Mikado v1.5, exonerate v2.2.0, BRAKER v2.1.2, augustus v3.3.2, EvidenceModeler v1.1.1, Mugsy v1.2.3, D-GENIES dot plot v1.2.0, MUMmer v3.0 and v4.0, nucmer v3.0, Assemblytics, OrthoFinder v2.3.3, i-AdHoRe v3.0, Tandem Repeats Finder v4.09, DupPipe, Genewise v2.4.1, PAML v3.1, MScancX v1.0, MAFFT v7.427, Mesquite v3.6, MEGA-X, RAxML v8.2.12, PHYLYP v3.697, MAFFT v7.450, BLAST v2.9.0+, TrimAl v1.4, IQ-TREE v1.6.1 and v1.6.1.2, ModelFinder v1.6.1, iTOI, platform v4.4.2, MUSCLE v3.8.31, fastp v0.19.10, qualimap v2.2.1, blat v35, picard v2.21.4, BUSCO v3, TranslatorX, ASTRAL-III v5.6.3, DiscoVista v1.0, TAPscan, InterPro 77.0, Pfam 32.0, CDD v3.17, HMMER v3.3

For manuscripts utilising custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

All the raw sequences are deposited to NCBI SRA under the BioProject PRINAS74424, PRINAS74453, and to ENA under the study accessions PRJEB34763.
The genome assemblies, annotations ("Submitted.zip") as well as alignment matrices and tree files ("phylogeny_dataset.zip") can be found on Figshare (private link: https://figshare.com/s/e3ebfc9104663c5d08de). A future genome browser will be available for the public upon publication.

### 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
- [x] Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](http://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
Here we provide three high-quality genome assemblies and their annotations for the genus Anthoceros. We use these data to refine our inferences on the nature of land plant MRCA and to gain new insights into hornwort biology.

#### Research sample
Cultures of Anthoceros agrestis (Oxford and Bonn strains) and A. punctatus were all derived from a single spore, (haploid gametophyte tissue) and axenically propagated and maintained on either BCD or Hatcher’s medium. Supplementary Table 13 shows the origin and specimen voucher for each of the three strains. We have been developing the three Anthoceros isolates as model systems for multiple years. Our selection was tailored by the potential of these strains to become model species for hornworts.

#### Sampling strategy
No statistical test was used to determine sample size. In gene expression studies three or two biological replicates were used to estimate differential gene expression (significance and fold change). The number of biological replicates used was tailored by the difficulty in obtaining tissue samples and extracting high-quality RNA from Anthoceros tissues.

#### Data collection
DNA was derived from axenic isolates of the three Anthoceros accessions. For gene expression studies RNA was extracted from tissues of the very same isolates after vegetative propagation. Data was recorded and analyzed as described in the Authors Contribution section of the main text.

#### Timing and spatial scale
Samples for DNA-sequencing were collected when available. Samples for RNA-seq experiments followed well-defined developmental stages described in the manuscript. For the CO2 response experiment, we subjected the plant cultures to one of the three CO2 environments at 150 (low), 400 (ambient), and 800 (high) ppm in a CO2-controlled growth chamber for 10 days (12/12hr day/night cycle). These CO2 concentrations match up with those used in previous experiments investigating hornwort pyrenoid function. Therefore, our results are directly comparable with observations of previous investigations. Sampling intervals also followed previous experiments to ensure comparability.

For the cyanobacterial symbiosis experiment, plants were transferred from solid BCD plates to flasks with 100 ml BCD media solution, and placed on an orbital shaker with 130 rpm for two weeks. For the cyano-/N+ and cyano-/N- conditions, plants were transferred to fresh new BCD solution with and without KNO3, respectively and grown for 10 days before harvest. These conditions and time intervals correspond to those that were previously applied in studies investigating hornwort-cyanobacteria symbiosis.

#### Data exclusions
Raw sequence data was quality filtered and trimmed using either fastp or trimmomatic (default parameters). Genome assemblies were filtered for contaminant scaffolds with blobtools and were excluded. Our data exclusion strategy was not pre-established. We used well accepted thresholds to filter out low-quality sequence data.

#### Reproducibility
RNA-seq experiments were carried out using two or three biological replicates. Genome assemblies were done with and without using long-reads and their collinearity compared. Bootstrap analyses were estimated for all nodes in gene trees.

#### Randomization
Bootstrap support for nodes in gene trees were estimated in a standard fashion through random resampling of columns in sequence alignments.

#### Blinding
No blinding was done for any of our analyses.

#### Did the study involve field work?
- [ ] Yes
- [x] No

### 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 | Methods |
|----------------------------------|---------|
| n/a | Involved in the study |
| ❌ | Antibodies |
| ❌ | Eukaryotic cell lines |
| ❌ | Palaeontology |
| ❌ | Animals and other organisms |
| ❌ | Human research participants |
| ❌ | Clinical data |
| n/a | Involved in the study |
| ❌ | ChIP-seq |
| ❌ | Flow cytometry |
| ❌ | MRI-based neuroimaging |