Article:
Li, F-W, Nishiyama, T, Waller, M et al. (31 more authors) (2020) Anthoceros genomes illuminate the origin of land plants and the unique biology of hornworts. Nature Plants, 6 (3). pp. 259-272. ISSN 2055-026X

https://doi.org/10.1038/s41477-020-0618-2

nature plants

ARTICLES

https://doi.org/10.1038/s41477-020-0618-2

Check for updates

OPEN

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

Fay-Wei Li[1,2]✉, Tomoaki Nishiyama[3], Manuel Waller[4], Eftychios Frangedakis[5], Jean Keller[6], Zheng Li[7], Noe Fernandez-Pozo[8], Michael S. Barker[7], Tom Bennett[9], Miguel A. Blázquez[10], Shifeng Cheng[11], Andrew C. Cuming[9], Jan de Vries[12], Sophie de Vries[13], Pierre-Marc Delaux[6], Issa S. Diop[4], C. Jill Harrison[14], Duncan Hauser[1], Jorge Hernández-García[10], Alexander Kirbis[4], John C. Meeks[15], Isabel Monte[16], Sumanth K. Mutte[17], Anna Neubauer[4], Dietmar Quandt[18], Tanner Robison[1,2], Masaki Shimamura[19], Stefan A. Rensing[8,20,21], Juan Carlos Villarreal[22,23], Dolf Weijers[17], Susann Wicke[24], Gane K.-S. Wong[25,26], Keiko Sakakibara[27] and Péter Szövényi[4,28]✉

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[1] 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)[2–6].

The evolution of land plants is underlined by the rise of morphological, molecular and physiological innovations. Tracing 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*)[7] and liverworts (*Marchantia polymorpha*)[8]—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[9]. For instance, most hornworts have a single chloroplast per cell with a pyrenoid capable of carrying out a carbon-concentrating mechanism (CCM)[10]. Such pyrenoid-based CCMs cannot be found in

[1]Boyce Thompson Institute, Ithaca, NY, USA. [2]Plant Biology Section, Cornell University, Ithaca, NY, USA. [3]Advanced Science Research Center, Kanazawa University, Ishikawa, Japan. [4]Department of Systematic and Evolutionary Botany, University of Zurich, Zurich, Switzerland. [5]Department of Plant Sciences, University of Cambridge, Cambridge, UK. [6]LRSV, Université de Toulouse, CNRS, UPS Castanet-Tolosan, Toulouse, France. [7]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA. [8]Faculty of Biology, Philipps University of Marburg, Marburg, Germany. [9]Centre for Plant Sciences, Faculty of Biological Sciences, University of Leeds, Leeds, UK. [10]Instituto de Biología Molecular y Celular de Plantas, CSIC-Universidad Politécnica de Valencia, Valencia, Spain. [11]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. [12]Institute for Microbiology and Genetics, Department of Applied Bioinformatics, Georg-August University Göttingen, Göttingen, Germany. [13]Institute of Population Genetics, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. [14]School of Biological Sciences, University of Bristol, Bristol, UK. [15]Department of Microbiology and Molecular Genetics, University of California, Davis, CA, USA. [16]Department of Plant and Microbial Biology, University of Zurich, Zurich, Switzerland. [17]Laboratory of Biochemistry, Wageningen University & Research, Wageningen, the Netherlands. [18]Nees Institute for Biodiversity of Plants, University of Bonn, Bonn, Germany. [19]Graduate School of Integrated Sciences for Life, Hiroshima University, Hiroshima, Japan. [20]BIOSS Centre for Biological Signalling Studies, University of Freiburg, Freiburg, Germany. [21]LOEWE Center for Synthetic Microbiology (SYNMIKRO), University of Marburg, Marburg, Germany. [22]Department of Biology, Laval University, Quebec City, Quebec, Canada. [23]Smithsonian Tropical Research Institute, Balboa, Panamá. [24]Institute for Evolution and Biodiversity, University of Muenster, Münster, Germany. [25]Department of Biological Sciences, Department of Medicine, University of Alberta, Edmonton, Alberta, Canada. [26]BGI-Shenzhen, Shenzhen, China. [27]Department of Life Science, Rikkyo University, Tokyo, Japan. [28]Zurich-Basel Plant Science Center, Zurich, Switzerland. ✉e-mail: fl329@cornell.edu; peter.szoevenyi@systbot.uzh.ch

any other land plants but frequently occur in algae[11]. Conversely, hornwort sporophytes are long-lived and moderately independent of gametophytes, which have been assumed to be features linking them to vascular plants[12]. Furthermore, hornwort sporophytes bear stomata that may be homologous with those of vascular plants[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[14]. Moreover, hornworts are among the very few plants that have a symbiotic relationship with nitrogen-fixing cyanobacteria[15] and one particular hornwort species, *Anthoceros punctatus*, has been used as a model system to study plant–cyanobacteria interactions[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 *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 *Anthoceros agrestis* (Bonn and Oxford strains) and *A. punctatus*. For *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 (*A. agrestis* has six chromosome pairs; Fig. 1 and Supplementary Fig. 1). For *A. agrestis* Oxford strain and *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 *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[17,18]. 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 *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. [19].

Anthoceros displays unusual centromere structure

The chromosomal-level assembly of *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)[20]. In *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 *P. patens* genome where genes and repeats are evenly distributed along the chromosomes[21]. 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.

Phylogenomic 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[2–6] 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 *P. patens* genome implied that regions showing collinearity between moss and some angiosperms may represent conserved collinear blocks since the MRCA of land plants[21]. However, comparing bryophytes to vascular plants, shared ancestral gene blocks could not be identified, rather that the collinear regions with vascular plants were unique to each of the bryophyte genomes (Supplementary Fig. 6 and Supplementary Table 7). The most genomic blocks collinear with at least one other land plant were found in the moss, followed by the liverwort and hornwort genomes (Supplementary Fig. 6). Within bryophytes, no collinear segment conserved across all three lineages was found, although there were genomic regions exclusively collinear 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 collinear 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 *Anthoceros*

Whole-genome duplications (WGD) have played an important role in shaping plant evolution and possibly underlie several adaptive radiations[22]. A previous study, based on the number of synonymous substitutions per synonymous site (K_S) divergence in transcriptomic datasets, suggested that hornworts may not have experienced any WGD event[21], similar to *M. polymorpha*[8] and *Selaginella moellendorffii*[23]. Our Ks plots on the annotated *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 *M. polymorpha* and *P. patens* genomes for comparison. We found very little

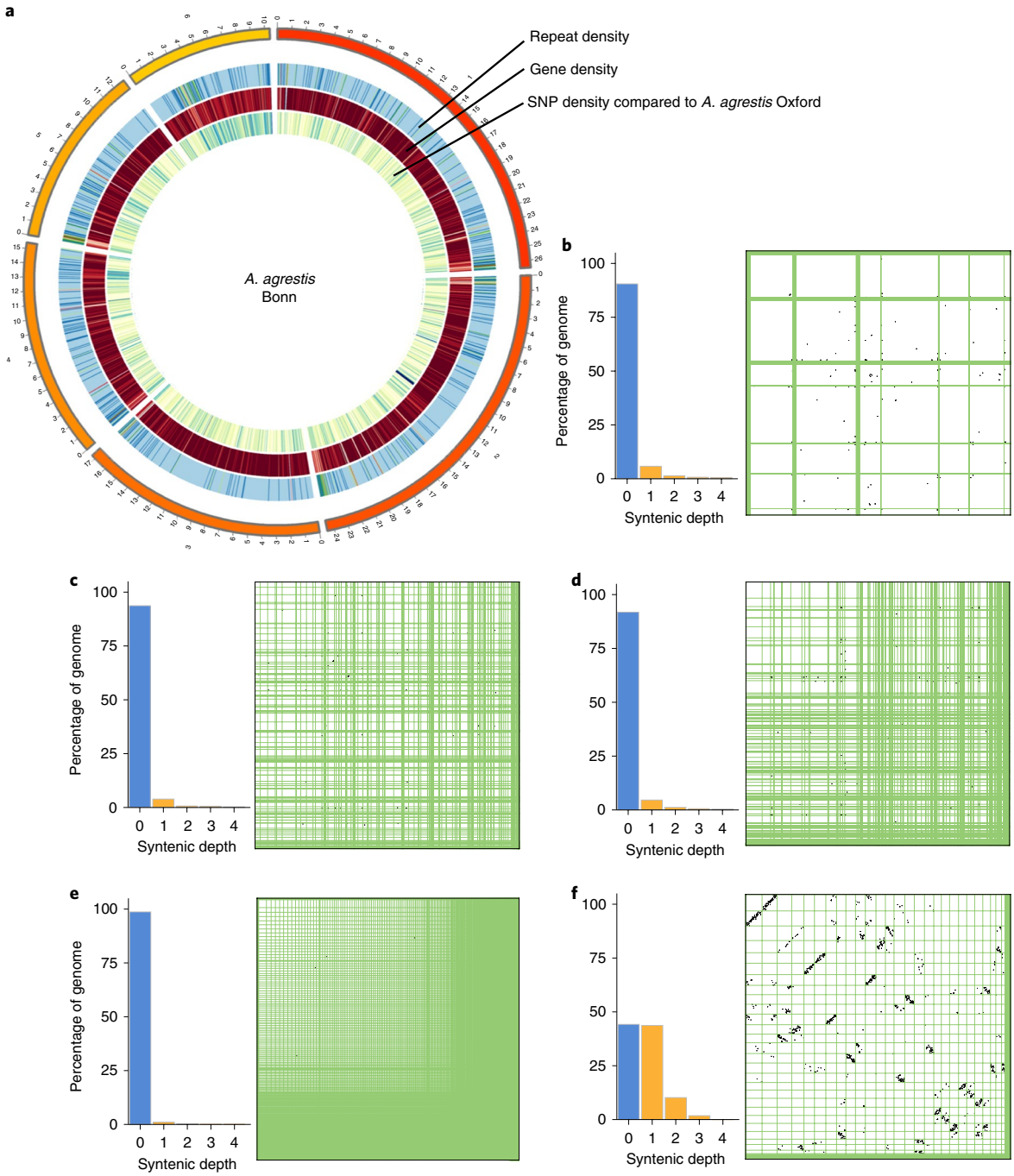

Fig. 1 | Genome organizations in the *Anthoceros* genomes. a, Circos plot of *A. agrestis* Bonn showing the densities of repeats, genes and single nucleotide polymorphisms (SNPs) with the *A. agrestis* Oxford genomes. **b**–**e**, *Anthoceros* genomes lack whole genome duplication. No self-synteny can 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

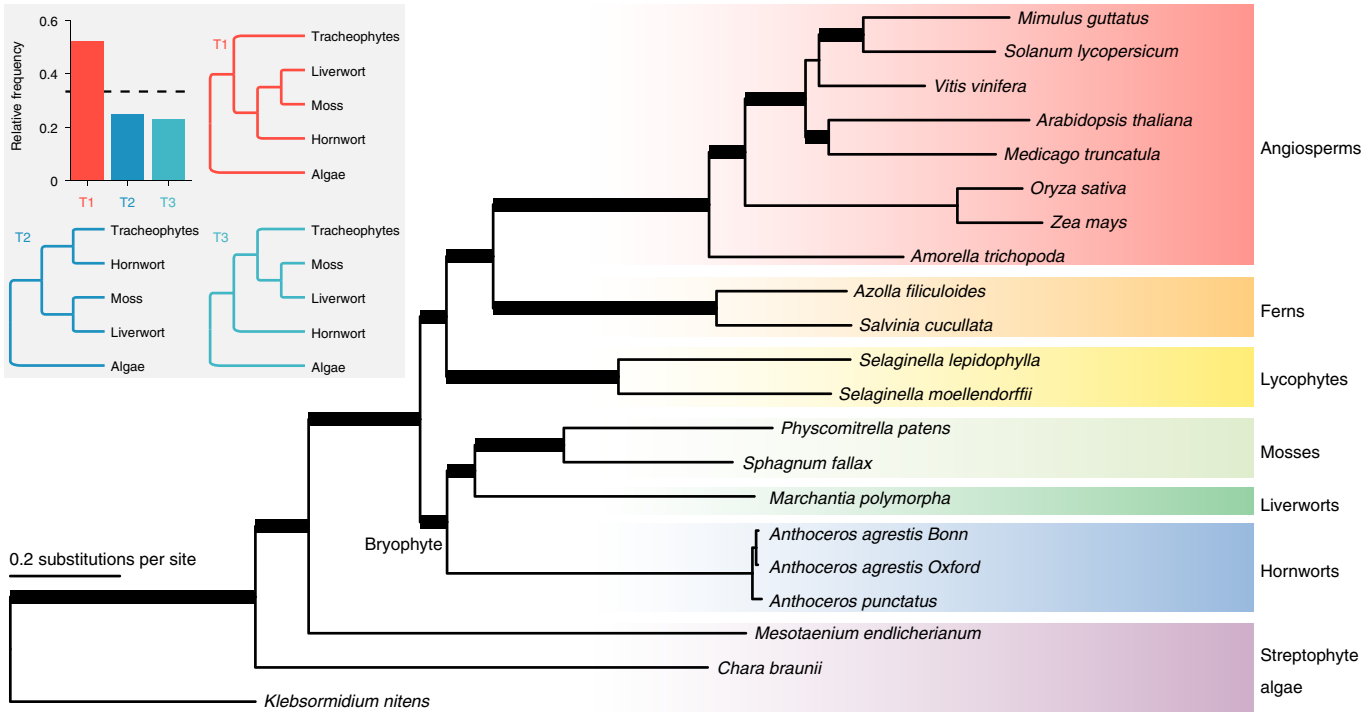

Fig. 2 | Land plant phylogeny inferred from 742 mostly single-copy genes. The monophyly of bryophytes is supported. The topology shown here is based on the maximum likelihood tree from the concatenated amino acid dataset. Thickened branches received maximal (100) bootstrap and SH-aLRT supports from both the concatenated nucleotide and amino acid datasets, as well as maximal posterior probabilities (1.0) from the Astral species-tree analysis (based on both nucleotide and amino acid gene trees). The inset shows the quartet frequencies among the 742 gene trees supporting monophyletic bryophytes (T1) versus two alternative placements of hornworts (T2 and T3). The dotted line shows the one-third threshold.

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 WGD[21].

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 plants[24], 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 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 features[25]. 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 hypothesized[25]. 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 plants[26]. 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 plants[27], while Class 2 *Knotted1*-like homeobox (KNOX2) genes maintain sporophyte cell fate in *P. patens*[28]. 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 functions[29]. 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

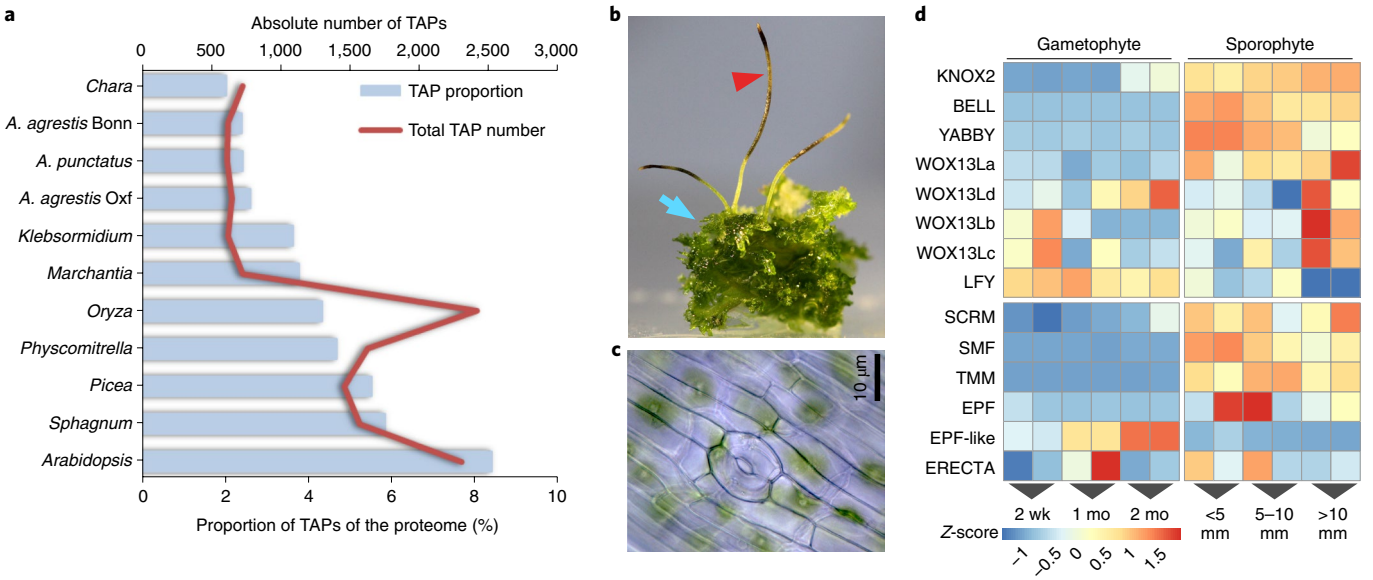

Fig. 3 | TAPs and sporophyte development. a, The *Anthoceros* genomes have the smallest TAP repertoire among land plants. **b**, Sporophytes (red arrowhead) and gametophytes (blue arrow) of *A. agrestis* Bonn. **c**, Stomata of *A. agrestis* Bonn. **d**, Gene expression profiles across different developmental stages in *A. agrestis* Bonn ($n=12$ biologically independent samples; two-sided test for differential expression, false-discovery rate ≤0.05 and $\log_2$-fold-change ≥2). wk, week; mo, month.

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 zygote development and stem cell formation in the moss *P. patens*[30]. *A. thaliana WOX13* promotes replum formation in the fruit[31] and *WOX14* promotes vascular cell differentiation[32]. The *Anthoceros* genomes have four WOX13L members (Supplementary Fig. 8 and Supplementary Tables 8 and 9) and *WOX13La* is specifically expressed in sporophytes while *WOX13Lbcd* 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/LFY*) gene (Supplementary Fig. 8 and Supplementary Tables 8 and 9), which in *P. patens* and *A. thaliana* controls zygote development and SAM maintenance, respectively[33]. 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 (*SCRMs*), *EPIDERMAL PATTERNING FACTOR* (*EPF*), *ERECTA* and *TOO MANY MOUTHS* (*TMM*) genes[34,35]. *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[8]. We found orthologues of *FAMA (SMF)*, *SCRM*, *ERECTA*, *EPF* and *TMM* in the *Anthoceros* genomes (in line with a previous study based on our earlier genome draft[26]; 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 stomatal 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[36]. 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[7,8,37]. 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[38]. 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[39].

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[40,41] 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[42]. Some of these genes are also required for

legume–rhizobial interaction and are often referred to as the common symbiosis genes[43].

While a few components can be traced back to as far as charophyte algae[44], 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[45]. 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[44] 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)[15,46]. In bryophytes, although mosses frequently harbour epiphytic cyanobacteria[47], only hornworts and two liverwort species host cyanobacteria endophytically within specialized slime-filled cavities[15,46]. 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[16], 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 ($>10^3$ 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[48] and by the observation that inactivation of a carbohydrate permease in *N. punctiforme* resulted in a defective symbiotic phenotype[49]. In parallel, SWEET is involved in mycorrhizal symbiosis as well, but a different orthologue, in the *SWEET1* clade, was recruited[50].

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[51], mycorrhizal[52] and actinorhizal[53–55] interactions. RNA interference knockdown of a subtilase (*SBTM1*) in the legume *Lotus japonicus* resulted in a decreased arbuscule formation[52]. 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[56]. Algal and hornwort chloroplasts use inorganic carbon transporters and carbonic anhydrases to locally concentrate CO_2 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[57].

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[58], 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[57] do not have clear homologues in hornworts (nor in any other land plants), we did find *LCIB* (low CO_2 inducible B) to be present in the hornwort genomes and hornwort transcriptomes of the 1,000 plant transcriptomes project (1KP)[6] (Fig. 5). 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 *LCIB*s 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 CO_2 limitation and the encoded proteins localize around pyrenoids to presumably block CO_2 leakage[59,60]. All the hornwort LCIB sequences have the conserved amino acid residues at the active sites that are shared with other algal LCIBs[61] (Fig. 5b). However, unlike *C. reinhardtii*, we did not find *LCIB* to be differentially expressed when plants are grown at different CO_2 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 CO_2 level[62]. 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 KNOX1, 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 *WOX13L* 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

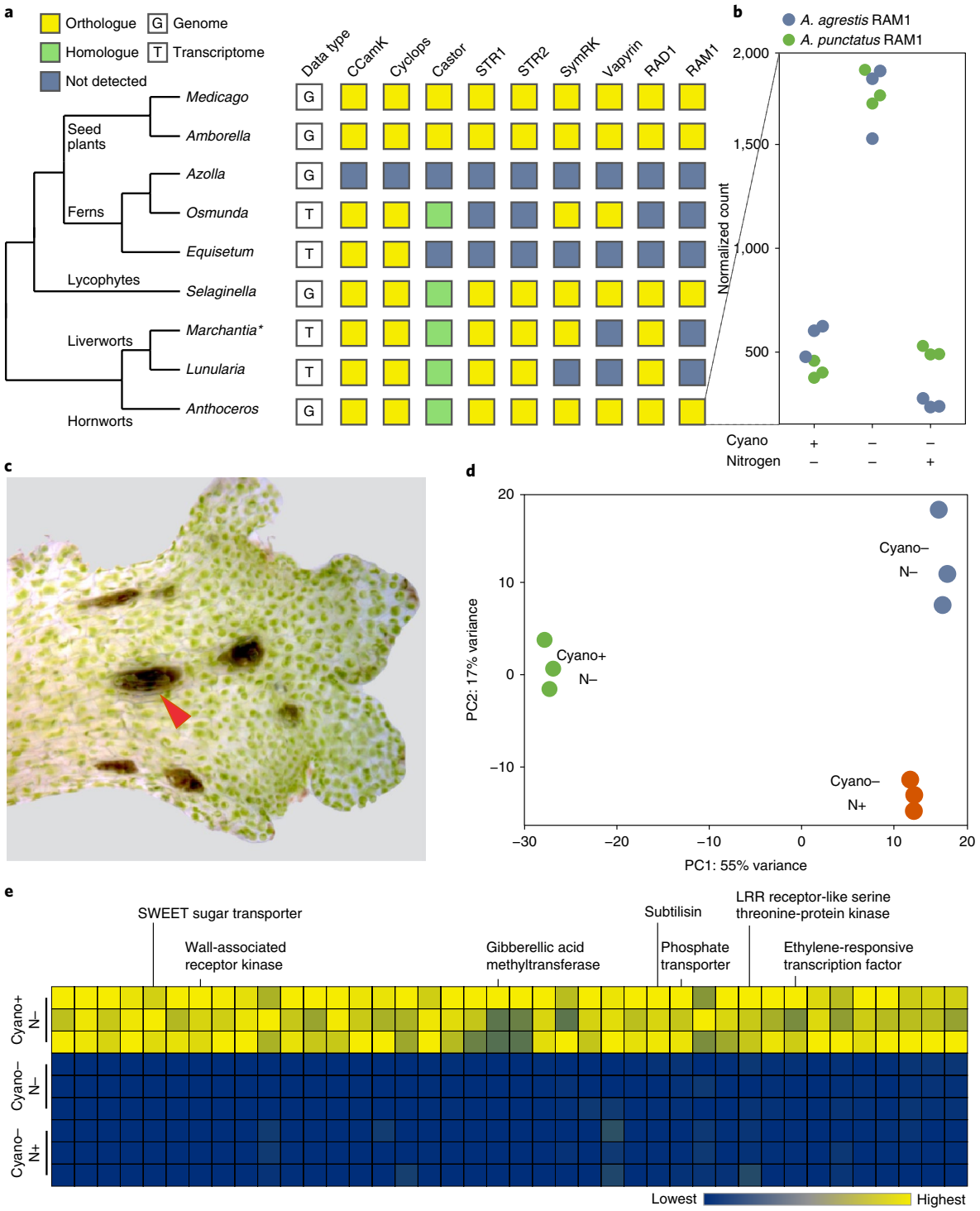

Fig. 4 | Evolution and genetics of symbiosis in hornworts. a, Orthologues of AMF symbiosis pathway genes can be found in hornworts, indicating their presence in the common ancestor of land plants. The asterisk indicates that the *M. paleacea* transcriptome was searched instead of *M. polymorpha* genome 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 $\log_2$-fold-change >4).

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.

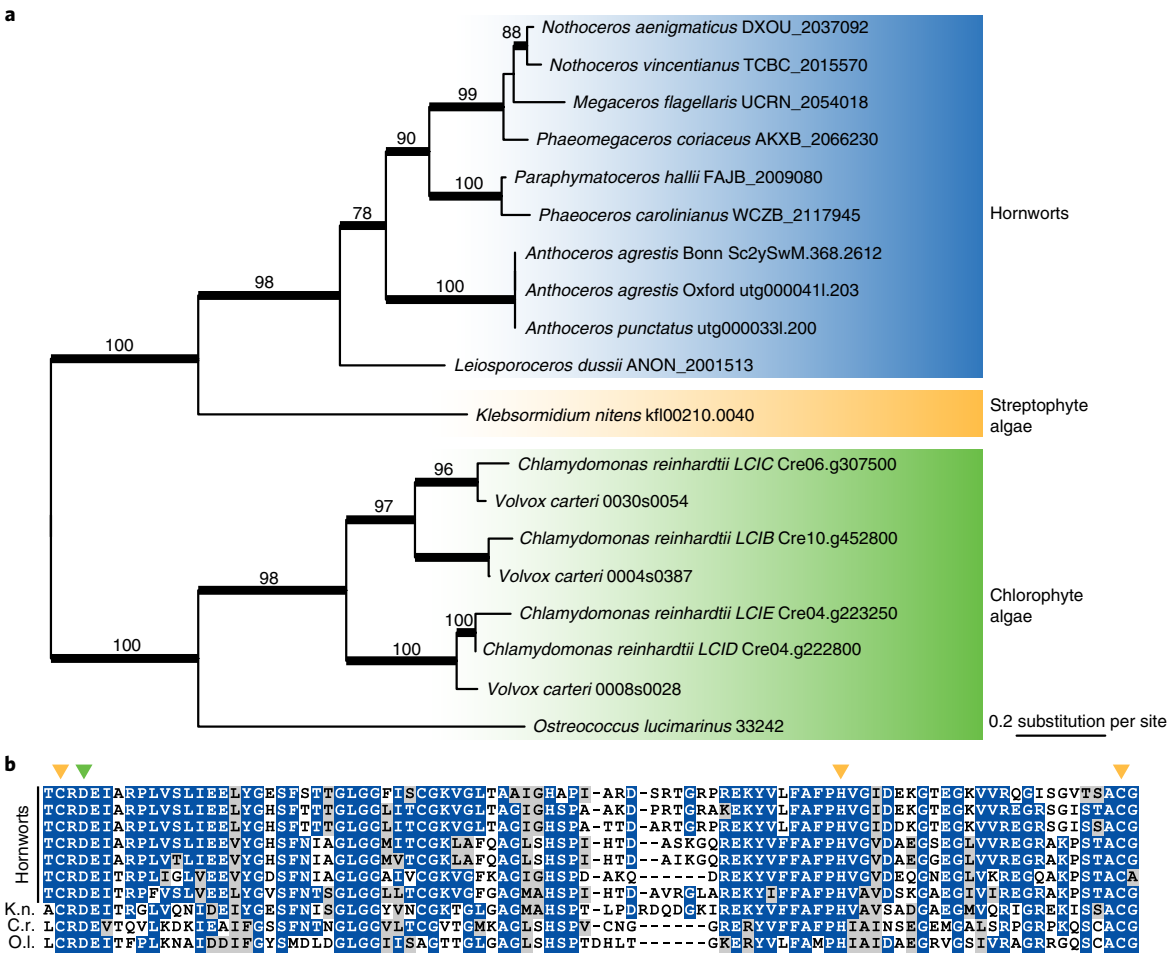

Fig. 5 | Relationship between *LCIB* and pyrenoid-based CCM. a, Phylogeny of *LCIB*. Numbers above branches are bootstrap support values (branches thickened when bootstrap >70). **b**, Hornwort *LCIB*s have conserved amino acid residues at the active site. Yellow and green arrowheads point to the zinc-binding and catalytic residues, respectively. K.n., *K. nitens*; C.r., *C. reinhardtii*; O.l., *Ostreococcus lucimarinus*.

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[21], none can be found in *Anthoceros* and *M. polymorpha* (Fig. 1). 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[8,21,23,37,63–66]. 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 BCD[67] or Hatcher's[68] 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% IGEPAL 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% TritonX-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 l 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. [69]. 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 (5%) with other barcoded libraries and sequenced on Illumina HiSeq1500 (two lanes with Rapid mode; OnBoardClustering) at the National Institute of Basic

Biology. For *A. punctatus*, Illumina genomic libraries were prepared by BGI and sequenced on HiSeq4000. Read quality and adaptor trimming was done by fastp[71] with the default setting. For *A. agrestis* Bonn, additional Chicago and Hi-C libraries were prepared by DoveTail Genomics. 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[72] and sorted with samtools[73]. The mean insert size and #READ_PAIRS were calculated using picard CollectInsertSizeMetrics. Unmapped reads were counted with samtools view -c -f 4 (ref. [73]) 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. The bam files were assessed with qualimap v.2.2.1 (ref. [74]) bamqc and observed error rates (total, mismatch, insertions and deletions), as well as the genome coverage were recovered.

Genome assembly. Genome sizes for the three *Anthoceros* were estimated based on *k*-mer distribution by Jellyfish[75] in conjunction with GenomeScope[76]. 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 MaSuRCA v.3.2.8 (ref. [77]). After assembly, base call quality was improved by two rounds of Pilon polishing[78]. We mapped Chicago and Hi-C reads back to the draft assembly and used DoveTail's HiRise assembler v.2.1.2 (ref. [79]) 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. [79]) to derive the final assembly.

Genome assemblies of *A. agrestis* Oxford strain and *A. punctatus* were generated with the minimap2-miniasm assembler[80] using only the nanopore reads. We then used four iterations of minimap2-racon[81] to derive the consensus sequence, followed by six rounds of Pilon polishing[78].

Contamination removal. While our cultures were grown in a putative axenic condition, low level of contamination cannot be completely ruled out. We therefore used blobtools[82] to identify scaffolds/contigs primarily consisting of contaminant 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-streptophyte 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 blobtools to assign the taxonomic affiliation to each scaffold with an *e*-value of 10^{-4}. We found that some of the small scaffolds were classified as of non-streptophyte 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 of 5–10 mm, (6) sporophytes longer than 1 cm with brown or black tips. Plants were grown on agar plates containing BCD medium[67] at 22 °C. RNA was extracted with the Spectrum Total RNA Plant Kit (Sigma-Aldrich) and stranded RNA-seq libraries were prepared using the TrueSeq Stranded mRNA Library Prep Kit (Illumina). Libraries were sequenced at the FGCZ on a HiSeq4000 machine. We used trimmomatic[83] to quality filter and trim the raw reads. Gene expression was estimated using Salmon[84] and differential expression done by DESeq2 ($\log_2$-fold ≥2, false-discovery rate ≤0.05 and normalized reads counts)[85].

We also generated separate thallus RNA-seq data for the Oxford strain (for annotation purpose). The plants were cultured on solid BCD plates and total RNA was extracted using RNeasy Plant Mini Kit (QIAGEN). The library was prepared using the TrueSeq stranded mRNA Library Prep Kit (Illumina) and sequenced on Hiseq1500.

RNA-seq dataset and analysis on cyanobacterial symbiosis. Liquid cultures of *A. agrestis* Oxford and *A. punctatus* were used in this experiment. To establish liquid cultures, plants were transferred from solid BCD plates to flasks with 100 ml of BCD media solution and placed on an orbital shaker at 130 r.p.m. for 2 weeks. For the cyano–/N+ and cyano–/N– conditions, plants were transferred to fresh new BCD solution with and without KNO_3, respectively and grown for 10 d before harvest. To reconstitute cyanobacterial symbiosis (with *N. punctiforme* ATCC 29133), we followed the method of Enderlin and Meeks[86] but using BCD as the growth medium. Three biological replicates were done for each condition. RNA was extracted by the Spectrum Total RNA Plant Kit (Sigma-Aldrich). The Illumina libraries were prepared by BGI and sequenced on HiSeq4000. Sequencing reads were mapped to the respective genomes by HiSat2 (ref. [87]) and transcript abundance quantified by Stringtie[88]. We used DESeq2 (ref. [85]) to carry out differential gene expression analysis, with false-discovery rate set to 0.005 and $\log_2$-fold-change threshold set to 1. To identify genes that are differentially expressed in both *A. agrestis* Oxford and *A. punctatus*, we used the Orthofinder gene family classification results (see below) coupled with phylogenetic analysis if needed.

RNA-seq dataset and analysis on CO_2 response. For the CO_2 experiment, we grew hornworts in magenta boxes with vented lids to allow air circulation while maintaining sterility. *A. agrestis* Oxford strain was used in this experiment and kept on solid BCD medium. We subjected the plant cultures to one of the three CO_2 environments at 150 ppm (low), 400 ppm (ambient) and 800 ppm (high) in a CO_2-controlled growth chamber for 10 d (12 h/12 h day/night cycle). Three biological replicates were done for each treatment. RNA was extracted by the Spectrum Total RNA Plant Kit (Sigma-Aldrich). The Illumina libraries were prepared by BGI and sequenced on HiSeq4000. One of the low CO_2 samples failed to produce high-quality library, and as a result the low CO_2 condition has only two replicates. RNA-seq data analysis was done following the same procedure as described above. We used BiNGO[89] for gene ontology enrichment analysis and REVIGO[90] to summarize and visualize the results.

Repeat annotation. For repeat annotation, we first built custom repeat libraries for each genome using RepeatModeler[91] and LTR_retriever[92]. The libraries were filtered to remove protein-coding genes by blasting against the UniProt plant database. We then used RepeatMasker[93] to annotate and mask the repetitive regions for each genome.

RNA-seq, transcript and protein evidence. We pooled *A. agrestis* Bonn, Oxford and *A. punctatus* RNA-seq reads together and mapped them onto each of the genome assemblies using HiSat2 (ref. [87]). We used all RNA-seq evidence available owing to the low nucleotide divergence among the three genomes. Transcriptomes were assembled for each species/strain separately. We used Portcullis[94] to filter out bad splice junctions and Stringtie[88] to assemble the transcripts. We additionally used Trinity[95] to generate both de novo and genome-guided transcriptome assemblies. We combined Trinity transcripts using the Program to Assemble Spliced Alignments (PASA) pipeline[96] and derived high-quality transcripts with Mikado[97]. To obtain protein homology information, we retrieved the 19 proteomes (only primary transcripts; Supplementary Table 15) and aligned them to the genome assemblies using exonerate[98]. We kept only hits with at least 60% coverage and a similarity above 60%.

Gene prediction. We used RNA-seq, transcript and protein evidence to train Augustus (ref. [99]) within Braker2 (ref. [100]). 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)[101]. 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 (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 mugsy[102]. 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 used them 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. [103]) with the Viridiplantae set to assess the completeness of our genomes and annotations. We did not use the Embryophya 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*, *Sphagnum* 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 orthogroup were done by MAFFT v.7.427 (ref. [105]) and back translated to nucleotides by TranslatorX[106]. The alignments were processed to remove sites with over 50% gaps or Ns and remove sequences shorter than 50% of the alignment length. When a species had more than one copy in an orthogroup, none from that species was included. To infer gene trees, we used both the amino acid and nucleotide matrices, and employed the maximum likelihood method implemented in IQ-Tree v.1.6.12 (ref. [107]). The best-fitting substitution models were selected by ModelFinder[108]. To reduce saturation in nucleotide substitution at this large time scale, the third codon position was excluded.

To infer the species tree, we used both concatenation and multispecies coalescent approach. The concatenated dataset included all the 742 loci and was analysed using IQ-Tree with ModelFinder model selection. To assess branch supports, we carried out ultrafast bootstrap[109] and SH-aLRT[110] analyses (both with 1,000 replicates). For the multispecies coalescent approach, we used ASTRAL-III (ref. [111]) to summarize all the 742 gene trees and measured branch supports as local posterior probabilities[112]. Gene-tree/species-tree discordance in terms of quartet frequencies was calculated by DiscoVista[113].

Collinearity of the three hornwort genomes and collinearity across Viridiplantae. We used the D-GENIES dot-plot tool[114] with the default options to visually assess collinearity of the three genome assemblies. We also aligned the genomic sequences using the nucmer module of mummer[115] and assessed their differences using Assemblytics[116].

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 events[117]. Gff files and proteomes were retrieved from Phytozome v.12 (ref. [118]). We used I-ADHore3 (ref. [119]) to detect highly degenerate collinear blocks among bryophytes and vascular plants requiring a minimum of three, four and five anchor points within each collinear region (gap_size=30, cluster_gap=35, q_value=0.75, prob_cutoff=0.01, anchor_points=5, alignment_method=gg2, level_2_only=false).

Identification of tandem repeats and centromeres. We run Tandem Repeats Finder[120] to identify tandem repeats with a minimum alignment score of 50 and a maximum period size of 2,000 bp. We then plotted repeat unit size against tandem array size to look for bimodal distribution. To localize centromeric regions in the *A. agrestis* Bonn genome, we generated dot-plots between a short-read-only assembly and the final chromosome-scale assembly. Because centromeric repeats are difficult to assemble using short-reads we expected that they will be missing from the Illumina assembly but will be present in the chromosomal-scale assembly. We also generated a self dot-plot of the *A. agrestis* Bonn genome to search for regions that are highly similar across scaffolds and are repetitive. Finally, we used the output of Tandem Repeats Finder[120] to search for tandem arrays with a period length of minimum 10 bp and with a minimum tandem array length of 30 repeat units. We plotted the location of these tandem arrays along the chromosomes to visually assessed their distribution.

Screening for whole-genome duplication. We used a combination of synonymous divergence (Ks) and synteny analyses to look for evidence of whole-genome duplication in the *Anthoceros* genomes. For each genome, we used the DupPipe pipeline to construct gene families and estimate the age of gene duplications[121]. We translated DNA sequences and identified reading frames by comparing the Genewise[122] alignment to the best-hit protein from a collection of proteins from 25 plant genomes from Phytozome[118]. For each analysis, we used protein-guided DNA alignments to align our nucleic acid sequences while maintaining reading frame. We then used single-linkage clustering to construct gene families and estimate K_S divergence using phylogenetic analysis by maximum likelihood (PAML)[123] with the F3X4 model for each node in the gene family phylogenies. Because the *Anthoceros* genomes contain large numbers of pentatricopeptide repeat genes (PPR), we also repeated the analysis with all the PPR genes removed. PPR genes were identified based on the Orthofinder results (see later).

For synteny analysis, we used MCscan's 'jcvi.compara.catalog ortholog'[124] function to search for and visualize intragenomic syntenic regions. The default *C*-score of 0.7 is used to filter low-quality hits. To calculate syntenic depths, the 'jcvi.compara.synteny depth' function was used. For comparison, we also carried out the same analysis for *P. patens* v.3.3 and *M. polymorpha* v.3.0 genomes; the former is known to have two rounds of WGD while the latter has none[8,21].

Transcription factor annotation. TAPs were annotated using TAPscan, according to Wilhelmsson et al.[24] and compared with selected other organisms using the major protein of each gene model ('.1' splice variant). TF annotations were further manually checked and adjusted for annotation errors or missing annotations.

Gene family classification and curation. We used Orthofinder2 (ref. [104]) to classify gene families of 25 plant and algal complete genomes, including the three hornworts reported here (Supplementary Table 16) into orthogroups. Orthofinder was run using the default setting, except that the 'msa' option was used. A total of 31,001 orthogroups were circumscribed. The detailed gene count and classification results can be found in Supplementary Table 16. While Orthofinder2 (ref. [104]) provides an automatic circumscription of gene families, they rarely correspond to their expert-based circumscriptions and can contain a substantial number of misclassified gene models due to the inherent limitations of the automatic classification algorithms. Therefore, all Orthofinder2 generated gene families selected for detailed evolutionary analyses were manually curated by their experts. In particular, members of all extensively investigated gene families were checked for the presence of their domain structure either using InterPro[125], Pfam[126] or CCD[127] to remove false positives and/or correct improperly predicted gene models. Furthermore, to ensure that a gene is truly absent (and not just unannotated), we carried out additional searches on the genome assemblies. For each extensively analysed gene family, we directly searched the raw genomic sequence using bryophyte or vascular plant homologues as query sequences to find additional gene models that might have been missed by our gene prediction pipeline. These searches were done using tBLASTn[128] and, in case no hit was found, were repeated with the hmmsearch module of HMMER[129] using the corresponding hmmer profiles from Pfam. Indeed, the manual curation helped us to add, revise and correct a substantial number of existing and/or missing gene models. Therefore, we believe that our careful manual curation ensures that the number of false positives and negatives are kept low and allows us to make statements about the presence/absence of particular genes.

Phylogenetic reconstruction of KNOX, LEAFY, WOX and YABBY. For KNOX, AagrBONN.evm.model.Sc2ySwM.368.1986.6 was used as a query to BLASTp search at NCBI on 13 Sept 2019. The search database was NCBI non-redundant (nr) database limited to records that include: *A. thaliana*, *Oryza sativa* (japonica cultivar-group), *Phalaenopsis equestris*, *Amborella trichopoda*, *Ceratopteris richardii*, *Selaginella moellendorffii*, *M. polymorpha*, *P. patens*, *K. nitens*, *Ostreococcus tauri* and *C. reinhardtii*. The search parameters were otherwise as default. The hit sequences were downloaded and combined with the *Anthoceros* KNOX sequences, then aligned with FFT-NS-2 in MAFFT v.7.427 (ref. [105]). The alignment was manually inspected in Mesquite v.3.6 (ref. [130]) and 149 well-conserved sites of 51 sequences were included. Phylogenetic analysis based on maximum likelihood (ML) was conducted in MEGA X[131]. The best-fitting model was chosen as LG+G+I using the FindBestProteinModel function. A total of 100 bootstrap replicates were performed to evaluate branch support. 'ML Heuristic Method' was set to 'Subtree-Pruning-Regrafting – Extensive (SPR level 5)' and 'No. of Discrete Gamma Categories' set to 5.

For LEAFY, AagrOXF evm.model.utg000049l.76.4 was used as a query to BLASTp search at NCBI on 30 August 2019. The search database was nr limited to records that include: *A. thaliana*, *O. sativa* (japonica cultivar-group), *P. equestris*, *A. trichopoda*, *P. radiata*, *P. armandii*, *P. abies*, *C. richardii*, *S. moellendorffii*, *M. polymorpha* and *P. patens*. The search parameters were otherwise as default. The hit sequences were downloaded and combined with the *Anthoceros* LEAFY sequence and AHJ90704.1, AHJ90706.1, AHJ90707.1 from Sayou et al.[132], then aligned with FFT-NS-2 in MAFFT v.7.427 (ref. [105]). The alignment was manually inspected and processed as described above to include 194 conserved sites of 20 sequences. Phylogenetic inference was done similarly as above but with LG selected as the best-fitting model.

For WOX, WOX genes in *Anthoceros* genomes were searched using the corresponding *A. thaliana*, *P. patens* and *M. polymorpha* proteins. Based on comparison among the three genomes, three gene models with excess intron predictions were manually revised and one model was added. AagrOXF_evm.model.utg000018l.552.1 was used as a query to BLASTp search at NCBI on 9 October 2019. The search database was nr limited to records that include: *A. thaliana*, *O. sativa* (japonica cultivar-group), *P. equestris*, *A. trichopoda*, *C. richardii*, *S. moellendorffii*, *M. polymorpha*, *P. patens*, *K. nitens* and *C. braunii*. Maximum target was set to 250 and the word size as 2. The search parameters were otherwise as default. The hit sequences were downloaded and combined with the *Anthoceros*

WOX sequences, then aligned with einsi —maxiterate 1,000 in MAFFT v.7.429 (ref. [105]). The alignment was manually inspected with Mesquite v.3.6 (ref. [130]) and a matrix consisting of 58 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 JTT with ProteinModelSelection8.pl. The maximum likelihood (ML) tree was inferred by RAxM[133] with -f a -\# 100 -m PROTGAMMAJTT and supplying -p and -x from random number generator. Bootstrap samples were generated with seqboot from PHYLIP package v.3.697 (ref. [134]) 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 aligned sequences were manually inspected with Mesquite and short sequences were removed and ambiguously aligned or gap containing sites were excluded. The best-fitting model was chosen as HIVB by ProteinModelSelection8.pl and ML tree search followed what was described for WOX.

Phylogenetic reconstruction of stomata-related genes. An *Anthoceros* ICE/SCRM homologue sequence AagrBONN_evm.model.Sc2ySwM_368.1570.1 was used as a query to BLASTp search at NCBI on 7 October 2019. The search database was nr limited to records that include: *A. thaliana*, *O. sativa* (japonica cultivar-group), *P. equestris*, *A. trichopoda*, *S. moellendorffii*, *P. patens*, *M. polymorpha*, *C. braunii* and *K. nitens*. The word size was set to 2 and maximum target sequences as 250. The search parameters were otherwise set as the default. The hit sequences (100) were downloaded and combined with the *Anthoceros* ICE/SCRM sequences, then aligned with einsi —maxiterate 1,000 in MAFFT v.7.429 (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 JTTDCMUTF with ProteinModelSelection8.pl. The ML tree was inferred by RAxML with -f a -# 100 -m PROTGAMMAJTTDCMUTF and supplying -p and -x from random number generator. Bootstrap samples (1,000 replicates) were generated with seqboot from PHYLIP package v.3.697 (ref. [134]) and RAxML[133] was run for each of them. For ERECTA and TMM, the sequences of AagrOXF_evm.model.utg000083l.351.1 and AagrOXF_evm.model.utg000012l.100.1 were respectively used as the query and processed as in ICE/SCRM. Phylogenetic analyses were performed as for the ICE/SCRM case, but with LG selected as the best-fitting model. For the EPF and EPF-like gene family, we used the matrix compiled by Takata et al.[136] and added the *Anthoceros* and *M. polymorpha* homologues. ML tree inference was done by IQ-TREE v.1.6.1 with 1,000 replicates of UltraFast Bootstraps[109].

Identification of orthologues to AMF symbiosis genes. Homologues to symbiotic genes were retrieved in 31 species covering the different plant lineages (Supplementary Table 11) using protein from the model plant *Medicago truncatula* and the tBLASTn v.2.9.0+ (ref. [128]) with a threshold e-value of $1e^{-10}$. Sequences were aligned using MAFFT v.7.407 (ref. [105]) with default parameters and alignments were cleaned using TrimAl v.1.4 (ref. [137]) to remove positions with more than 20% of gaps. Resulting alignments were subjected to ML tree inference using IQ-TREE v.1.6.1 (ref. [107]). Before ML analysis, the best-fitting evolutionary model was tested using ModelFinder[108] and according to the Bayesian Information Criteria Branch support was tested using 10,000 replicates of UltraFast Bootstraps[109]. Trees were visualized with the iTOL platform v.4.4.2 (ref. [138]).

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[6]. To find other *LCIB* homologues, we ran BLASTp against the Phytozome 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 MUSCLE[139]. IQ-TREE v.1.6.1 (ref. [107]) was used to obtain the ML tree as outlined above.

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

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 PRJNA574424 and PRJNA574453, and to European Nucleotide Archive (ENA) under the study accessions 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.

Received: 14 October 2019; Accepted: 11 February 2020;
Published online: 13 March 2020

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. Puttick, M. N. et al. The interrelationships of land plants and the nature of the ancestral embryophyte. *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**, 679–685 (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. Renzaglia, 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 Anthocerotae. *Planta* **200**, 203–212 (1996).
11. Li, F.-W., Villarreal Aguilar, J. C. & Szövényi, P. Hornworts: an overlooked window into carbon-concentrating mechanisms. *Trends Plant Sci.* **22**, 275–277 (2017).
12. Qiu, Y.-L. et al. The deepest divergences in land plants inferred from phylogenomic evidence. *Proc. Natl Acad. Sci. USA* **103**, 15511–15516 (2006).
13. Renzaglia, K. S., Villarreal Aguilar, J. C., Piatkowski, B. T., Lucas, J. R. & Merced, A. Hornwort stomata: architecture and fate shared with 400-Million-year-old fossil plants without leaves. *Plant Physiol.* **174**, 788–797 (2017).
14. Renzaglia, 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. Bainard, 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. & Renzaglia, K. S. The origin of the sporophyte shoot in land plants: a bryological perspective. *Ann. 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. Coudert, Y., Novák, O. & Harrison, C. J. A KNOX-cytokinin regulatory module predates 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 protoplast cells into stem cells in the moss *Physcomitrella patens*. *Development* **141**, 1660–1670 (2014).
31. Romera-Branchat, M., Ripoll, J. J., Yanofsky, M. F. & Pelaz, S. The WOX 13 homeobox gene promotes replum 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).

33. Tanahashi, T., Sumikawa, N., Kato, M. & Hasebe, M. Diversification of gene function: homologs of the floral regulator FLO/LFY control the first zygotic cell division in the moss *Physcomitrella patens*. *Development* **132**, 1727–1736 (2005).
34. Lee, L. R. & Bergmann, D. C. The plant stomatal lineage at a glance. *J. Cell Sci.* **132**, jcs228551 (2019).
35. Chater, C. C. et al. Origin and function of stomata in the moss *Physcomitrella patens*. *Nat. Plants* **2**, 16179 (2016).
36. Kosentka, P. Z., Overholt, A., Maradiaga, R., Mitoubsi, O. & Shpak, E. D. EPFL signals in the boundary region of the SAM restrict its size and promote leaf initiation. *Plant Physiol.* **179**, 265–279 (2019).
37. Nishiyama, T. et al. The *Chara* genome: secondary complexity and implications for plant terrestrialization. *Cell* **174**, 448–464 (2018).
38. Hernandez-Garcia, J. & Briones-Moreno, A. Origin of gibberellin-dependent transcriptional regulation by molecular exploitation of a transactivation domain in DELLA proteins. *Mol. Biol. Evol.* **36**, 908–918 (2019).
39. Walker, C. H., Siu-Ting, K., Taylor, A., O'Connell, M. J. & Bennett, T. Strigolactone synthesis is ancestral in land plants, but canonical strigolactone signalling is a flowering plant innovation. *BMC Biol.* **17**, 70 (2019).
40. Remy, W., Taylor, T. N., Hass, H. & Kerp, H. Four hundred-million-year-old vesicular arbuscular mycorrhizae. *Proc. Natl Acad. Sci. USA* **91**, 11841–11843 (1994).
41. Strullu-Derrien, C. Fossil filamentous microorganisms associated with plants in early terrestrial environments. *Curr. Opin. Plant Biol.* **44**, 122–128 (2018).
42. MacLean, A. M., Bravo, A. & Harrison, M. J. Plant signaling and metabolic pathways enabling arbuscular mycorrhizal symbiosis. *Plant Cell* **29**, 2319–2335 (2017).
43. Parniske, M. Arbuscular mycorrhiza: the mother of plant root endosymbioses. *Nat. Rev. Microbiol.* **6**, 763–775 (2008).
44. Delaux, P.-M. et al. Algal ancestor of land plants was preadapted for symbiosis. *Proc. Natl Acad. Sci. USA* **112**, 13390–13395 (2015).
45. Delaux, P.-M. et al. Comparative phylogenomics uncovers the impact of symbiotic associations on host genome evolution. *PLoS Genet.* **10**, e1004487 (2014).
46. Adams, D. G. & Duggan, P. S. Cyanobacteria–bryophyte symbioses. *J. Exp. Bot.* **59**, 1047–1058 (2008).
47. Rousk, K., Jones, D. L. & DeLuca, T. H. Moss–cyanobacteria associations as biogenic sources of nitrogen in boreal forest ecosystems. *Front. Microbiol.* **4**, 150 (2013).
48. Steinberg, N. A. & Meeks, J. C. Physiological sources of reductant for nitrogen-fixation activity in *Nostoc* sp. strain UCD 7801 in symbiotic association with *Anthoceros punctatus*. *J. Bacteriol.* **173**, 7324–7329 (1991).
49. Ekman, M., Picossi, S., Campbell, E. L., Meeks, J. C. & Flores, E. A *Nostoc punctiforme* sugar transporter necessary to establish a cyanobacterium–plant symbiosis. *Plant Physiol.* **161**, 1984–1992 (2013).
50. An, J. et al. A *Medicago truncatula* SWEET transporter implicated in arbuscule maintenance during arbuscular mycorrhizal symbiosis. *New Phytol.* **224**, 396–408 (2019).
51. Kistner, C. et al. Seven *Lotus japonicus* genes required for transcriptional reprogramming of the root during fungal and bacterial symbiosis. *Plant Cell* **17**, 2217–2229 (2005).
52. Takeda, N., Sato, S., Asamizu, E., Tabata, S. & Parniske, M. Apoplastic plant subtilases support arbuscular mycorrhiza development in *Lotus japonicus*. *Plant J.* **58**, 766–777 (2009).
53. Fournier, J. et al. Cell remodeling and subtilase gene expression in the actinorhizal plant *Discaria trinervis* highlight host orchestration of intercellular *Frankia* colonization. *New Phytol.* **219**, 1018–1030 (2018).
54. Ribeiro, A., Akkermans, A. D., van Kammen, A., Bisseling, T. & Pawlowski, K. A nodule-specific gene encoding a subtilisin-like protease is expressed in early stages of actinorhizal nodule development. *Plant Cell* **7**, 785–794 (1995).
55. Svistoonoff, S. et al. cg12 expression is specifically linked to infection of root hairs and cortical cells during *Casuarina glauca* and *Allocasuarina verticillata* actinorhizal nodule development. *Mol. Plant Microbe Interact.* **16**, 600–607 (2003).
56. Meyer, M. T., Whittaker, C. & Griffiths, H. The algal pyrenoid: key unanswered questions. *J. Exp. Bot.* **68**, 3739–3749 (2017).
57. Rae, B. D. et al. Progress and challenges of engineering a biophysical CO_2-concentrating mechanism into higher plants. *J. Exp. Bot.* **68**, 3717–3737 (2017).
58. Villarreal Aguilar, J. C. & Renner, S. S. Hornwort pyrenoids, carbon-concentrating structures, evolved and were lost at least five times during the last 100 million years. *Proc. Natl Acad. Sci. USA* **109**, 18873–18878 (2012).
59. Wang, Y. & Spalding, M. H. LCIB in the *Chlamydomonas* CO_2-concentrating mechanism. *Photosyn. Res.* **121**, 185–192 (2014).
60. Atkinson, N. et al. Introducing an algal carbon-concentrating mechanism into higher plants: location and incorporation of key components. *Plant Biotechnol. J.* **14**, 1302–1315 (2016).
61. Jin, S. et al. Structural insights into the LCIB protein family reveals a new group of β-carbonic anhydrases. *Proc. Natl Acad. Sci. USA* **113**, 14716–14721 (2016).
62. Hanson, D. T., Renzaglia, K. & Villarreal, J. C. in *Photosynthesis in Bryophytes and Early Land Plants* (eds Hanson, D. T. & Rice, S. K.) 95–111 (Springer, 2014).
63. Li, F.-W. et al. Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nat. Plants* **4**, 460–472 (2018).
64. Hori, K. et al. *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nat. Commun.* **5**, 3978 (2014).
65. Cheng, S. et al. Genomes of subaerial Zygnematophyceae provide insights into land plant evolution. *Cell* **179**, 1057–1067 (2019).
66. VanBuren, R. et al. Extreme haplotype variation in the desiccation-tolerant clubmoss *Selaginella lepidophylla*. *Nat. Commun.* **9**, 13 (2018).
67. Cove, D. J. et al. Culturing the moss *Physcomitrella patens*. *Cold Spring Harb. Protoc.* **2009**, db.prot5136 (2009).
68. Hatcher, R. E. Towards the establishment of a pure culture collection of Hepaticae. *Bryologist* **68**, 227–231 (1965).
69. Nagar, R. & Schwessinger, B. High purity, high molecular weight DNA extraction from rust spores via CTAB based DNA precipitation for long read sequencing v1. *protocols.io* https://doi.org/10.17504/protocols.io.n5ydg7w (2018).
70. Weisenfeld, N. I. et al. Comprehensive variation discovery in single human genomes. *Nat. Genet.* **46**, 1350–1355 (2014).
71. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
72. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
73. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
74. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).
75. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
76. Vurture, G. W. et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
77. Zimin, A. V. et al. The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
78. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
79. Putnam, N. H. et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
80. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
81. 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).
82. Laetsch, D. R. & Blaxter, M. L. BlobTools: interrogation of genome assemblies. *F1000Res.* **6**, 1287 (2017).
83. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
84. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
85. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
86. Enderlin, C. S. & Meeks, J. C. Pure culture and reconstitution of the *Anthoceros–Nostoc* symbiotic association. *Planta* **158**, 157–165 (1983).
87. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
88. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
89. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**, 3448–3449 (2005).
90. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**, e21800 (2011).
91. Smit, A. F. A. & Hubley, R. *RepeatModeler Open-1.0* (Institute for Systems Biology, accessed February 2019); www.repeatmasker.org
92. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
93. Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open-4.0* (Institute for Systems Biology, accessed February 2019); www.repeatmasker.org
94. Mapleson, D., Venturini, L., Kaithakottil, G. & Swarbreck, D. Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. *Gigascience* **7**, giy131 (2018).

95. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
96. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
97. Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L. & Swarbreck, D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *Gigascience* **7**, giy093 (2018).
98. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinf.* **6**, 31 (2005).
99. 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).
100. 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).
101. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
102. Angiuoli, S. V. & Salzberg, S. L. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* **27**, 334–342 (2011).
103. 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).
104. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
105. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
106. 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).
107. 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).
108. 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).
109. 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).
110. 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).
111. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinf.* **19**, 153 (2018).
112. Sayyari, E. & Mirarab, S. Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* **33**, 1654–1668 (2016).
113. Sayyari, E., Whitfield, J. B. & Mirarab, S. DiscoVista: interpretable visualizations of gene tree discordance. *Mol. Phylogenet. Evol.* **122**, 110–115 (2018).
114. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ.* **6**, e4958 (2018).
115. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
116. Nattestad, M. & Schatz, M. C. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021–3023 (2016).
117. Qiao, X. et al. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol.* **20**, 38 (2019).
118. Goodstein, D. M. et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186 (2012).
119. Proost, S. et al. i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.* **40**, e11 (2012).
120. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573 (1999).
121. Barker, M. S. et al. EvoPipes.net: bioinformatic tools for ecological and evolutionary genomics. *Evol. Bioinform.* **6**, 143–149 (2010).
122. Birney, E., Clamp, M. & Durbin, R. Genewise and genomewise. *Genome Res.* **14**, 988–995 (2004).
123. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
124. Tang, H. et al. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18**, 1944–1954 (2008).
125. Mitchell, A. L. et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).
126. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
127. Marchler-Bauer, A. et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45**, D200–D203 (2017).
128. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinf.* **10**, 421 (2009).
129. Wheeler, T. J. & Eddy, S. R. Nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**, 2487–2489 (2013).
130. Maddison, W. P. & Maddison, D. R. *Mesquite: A Modular System for Evolutionary Analysis* v.3.04 (Mesquite, accessed 5 July 2016); http://mesquiteproject.org
131. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
132. Sayou, C. et al. A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. *Science* **343**, 645–648 (2014).
133. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
134. Felsenstein, J. *PHYLIP (Phylogeny Inference Package)* Version 3.697 (University of Washington, 2015); http://evolution.genetics.washington.edu/phylip/oldversions.html
135. Finet, C. et al. Evolution of the YABBY gene family in seed plants. *Evol. Dev.* **18**, 116–126 (2016).
136. Takata, N. et al. Evolutionary relationship and structural characterization of the EPF/EPFL gene family. *PLoS ONE* **8**, e65183 (2013).
137. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
138. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
139. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

Acknowledgements

This project was supported by: National Science Foundation (NSF) grant no. DEB1831428 to F.-W.L.; Swiss National Science Foundation grant nos 160004 and 131726 to P.S.; funding from the Georges and Antoine Claraz Foundation to P.S.; funding from The Forschungskredit and the University Research Priority Program 'Evolution in Action' of the University of Zurich to M.W. and P.S.; NSF grant nos IOS-1339156 and EF-1550838 to M.S.B.; National Institute for Basic Biology (NIBB) Collaborative Research Program grant no. 13-710 to T.N.; Japan Society for the Promotion of Science (JSPS) grant no. KAKENHI 15H04413 to T.N.; JSPS grant nos KAKENHI 25113001, 26650143, 18H04843 and 18K06367 to K.S.; JSPS Short Term Postdoctoral Fellowship grant no. PE14780 to E.F.; Bill & Melinda Gates Foundation grant no. OPP11772165 to P.-M.D.; Spanish Ministry of Science, Innovation and Universities grant no. BFU2016-80621-P to J.H.-G. and M.A.B.; European Research Council starting grant 'TerreStriAL' to J.d.V.; funding from Foundation of German Business (sdw), Georges and Antoine Claraz Foundation and URPP Evolution in Action to A.N.; German Science Foundation grant no. WI4507/3-1 to S.W.; Special Grant for Innovation in Research Program of the Technical University of Dresden (Germany) to D.Q. and S.W.; Funding from the Earl S. Tupper Fellowship, STRI to J.C.V.; and Netherlands Organization for Scientific Research VICI grant no. 865.14.001 to D.W. and S.K.M. We thank K. Yamaguchi and S. Shigenobu of Functional Genomics Facility at NIBB and L. Poveda, C. Aquino and A. Patrignani of FGCZ for sequencing support. Computational resources were partly provided by the Data Integration and Analysis Facility, NIBB and the NIG supercomputer at ROIS National Institute of Genetics.

Author contributions

F.-W.L., P.S., K.S. and T.N. coordinated the project. M.S. carried out chromosome work. F.-W.L., P.S., K.S., T.N., D.H., S.C. and G.K.-S.W. sequenced the genomes. F.-W.L. and P.S. assembled the genomes. P.S. and T.N. annotated the genomes. T.R. and P.S. assembled and annotated organellar genomes P.S. and F.-W.L. performed synteny analyses. Z.L. and M.S.B. performed Ks analyses. A.N. and P.S. conducted RNA-seq experiment on developmental stages. F.-W.L. and J.C.M. conducted RNA-seq experiment on cyanobacterial symbiosis. F.-W.L. conducted RNA-seq experiment on CO_2 response. F.-W.L., M.W., A.K., I.D. and P.S. analysed RNA-seq data. N.P., S.R., M.W., K.S., P.S. and E.F. characterized transcription factors. F.-W.L. and P.S. performed gene family classification. J.K. and P.-M.D. conducted analysis on AMF symbiosis genes. I.M. conducted analysis on jasmonates. S.M. and D.W. conducted analysis on auxin signalling. A.C. conducted analysis on abscisic acid signalling. T.B. conducted analysis on strigolactone signalling. J.H.-G. and M.A.B. conducted analysis on gibberellin signalling. S.d.V. conducted analysis on salicylic acid signalling. J.d.V. and E.F. conducted analysis on genes associated with polyplastidy. C.J.H. conducted analysis on PIN proteins. S.W. and D.Q. conducted analysis on plastid-targeted genes and established the cultures of the *A. agrestis* Bonn strain. E.F., T.N. and F.-W.L. conducted analysis on stomatal development genes. F.-W.L., P.S., K.S., T.N., J.C.V., E.F. and M.W. synthesized and wrote the manuscript.

Competing interests

The authors declare no competing interests.

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.

nature research

Corresponding author(s): Peter Szovenyi

Last updated by author(s): 01/31/2020

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 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, Genomescope, MaSuRCAv3.2.8, pilon v1.22, HiRise assembler v2.1.2, 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, REVIGO, 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 v1.3.1, MCscanX v1.0, MAFFT v7.427, Mesquite v3.6, MEGA X, RAxML v8.2.12, PHYLIP 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, iTOL 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 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/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 PRJNA574424, PRJNA574453, and to ENA under the study accessions PRJEB34763,

PRJEB34743 (Supplementary Tables 2-3) and will become public upon publication.

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 ☒ 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	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 ☒ 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

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

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