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], Jill Harrison[14], Duncan Hauser[1], Jorge Hernández-García[10], Alexander Kirbis[4], John C. Meeks[15], Isabel Monte, Sumanth K. Mutte[16], Anna Neubauer[4], Dietmar Quandt[17], Tanner Robison[1,2], Masaki Shimamura[18], Stefan A. Rensing[8,19], Juan Carlos Villarreal[20,21], Dolf Weijers[22], Susann Wicke[23], Gane K.-S. Wong[24,25], Keiko Sakakibara[26], Peter Szövényi[4,27*]

[1]Boyce Thompson Institute, Ithaca, New York, USA
[2]Plant Biology Section, Cornell University, Ithaca, New York, USA
[3]Advanced Science Research Center, Kanazawa University, Ishikawa, Japan
[4]Department of Systematic and Evolutionary Botany, University of Zurich, Switzerland
[5]Department of Plant Sciences, University of Cambridge, UK
[6]Laboratoire de Recherche en Sciences Végétales, Université de Toulouse, Castanet-Tolosan, France
[7]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona, USA
[8]Faculty of Biology, Philipps University of Marburg, Germany
[9]Center for Plant Sciences, Faculty of Biological Sciences, University of Leeds, UK
[10]Instituto de Biología Molecular y Celular de Plantas, CSIC-Universidad Politécnica de Valencia, Valencia, Spain
[11]Lingnan Guangdong Laboratory of Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong, China
[12]Institute for Microbiology and Genetics, Department of Bioinformatics, Georg-August University Göttingen, Germany
[13]Institute of Population Genetics, Heinrich Heine University Düsseldorf, Germany
[14]School of Biological Sciences, University of Bristol, UK
[15]Department of Microbiology, University of California, Davis, California, USA
[16]Department of Plant and Microbial Biology, University of Zurich, Switzerland
[17]Nees Institute for Biodiversity of Plants, University of Bonn, Germany
[18]Graduate School of Integrated Sciences for Life, Hiroshima University, Japan
[19]BIOSS Centre for Biological Signalling Studies, University of Freiburg, Freiburg, Germany
[20]Department of Biology, Laval University, Quebec City, Quebec, Canada
[21]Smithsonian Tropical Research Institute, Balboa, Ancon, Panamá
[22]Laboratory of Biochemistry, Wageningen University & Research, Wageningen, the Netherlands
[23]Institute for Evolution and Biodiversity, University of Muenster, Germany
[24]Department of Biological Sciences, Department of Medicine, University of Alberta, Edmonton, Alberta, Canada
[25]BGI-Shenzhen, Shenzhen, China
[26]Department of Life Science, Rikkyo University, Tokyo, Japan
[27]Zurich-Basel Plant Science Center, Zurich, Switzerland

*Authors of correspondence: F.-W.L. (fl329@cornell.edu); P.S. (peter.szoevenyi@systbot.uzh.ch)

Abstract

Hornworts are a bryophyte lineage that diverged from other extant land plants >400 million years ago and bear unique biological features, including a distinct sporophyte architecture, cyanobacterial symbiosis, and carbon-concentrating mechanism (CCM). Here we provide three high-quality genomes of *Anthoceros* hornworts. 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 stomata function are conserved in *Anthoceros* and up-regulated 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 these hornwort genomes will serve as essential references for future hornwort research and comparative studies across land plants.

Introduction

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 being sister to Setaphyta (liverworts plus mosses)[2–5].

The evolution of land plants is underlined by the rise of morphological, molecular and physiological innovations. Tracing the evolutionary origins of these key novelties is prone to errors due to the 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 limited availability of model systems for only two bryophyte lineages—mosses (*Physcomitrella patens*)[6] and liverworts (*Marchantia polymorpha*)[7]—makes the 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[8]. For instance, most hornworts have a single chloroplast per cell with a pyrenoid capable of carrying out a carbon-concentrating mechanism (CCM)[9]. Such pyrenoid-based CCMs cannot be found in any other land plants but frequently occur in algae[10]. Conversely, hornwort sporophytes are long-lived and moderately independent of gametophytes, which have been assumed to be a feature linking them to vascular plants[11]. Furthermore, hornwort sporophytes bear stomata which may be homologous with those of vascular plants[12].

In addition to having characters exclusively shared with algae or with other land plants, hornworts also boast a wide range of distinctive biological features. For example, the presence of a basal sporophytic meristem and the asynchronous meiosis are unique to hornworts[13]. Moreover, hornworts are among the very few plants that have a symbiotic relationship with nitrogen-fixing cyanobacteria[14], and one particular hornwort species, *Anthoceros punctatus*, has been used as a model system to study plant-cyanobacteria interaction[15].

Detailed genomic information on hornworts is essential to not only 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 CCM. 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.

Table 1. Assembly statistics of the three hornwort genomes.

	Estimated genome size	Assembled genome size	Contig/Scaffold number	Contig/Scaffold N50 length	Assembly approach
Anthoceros agrestis Bonn	124.5 Mb	116.9 Mb	1577/322	155.5Kb/17.3Mb/	Illumina + Nanopore + Hi-C
Anthoceros agrestis Oxford	126.1 Mb	122.9 Mb	153/-	1.8Mb/-	Nanopore + Illumina
Anthoceros punctatus	129.4 Mb	132.8 Mb	202/-	1.7Mb/-	Nanopore + Illumina

Results

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 six largest scaffolds containing 95% of the assembled genome (*A. agrestis* has 5-6 chromosome pairs; Fig. 1 and Supplementary Figure 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 1.7Mb (Table 1). The three genomes are highly collinear with a greater collinearity found between the two *A. agrestis* strains (Supplementary Figure 2; Supplementary Table 1). The collinearity, BUSCO scores (Supplementary Figure 3), and read mapping statistics (Supplementary Tables 2-3), show that the three genomes are of high quality and accuracy. The genome assemblies and annotations are available in CoGe (https://genomevolution.org; reviewer username: “hornwortgenome” and password: “lrm36ns89.DGcm”), and also on Figshare (private link: https://figshare.com/s/e3ebfc9104663c5d08de).

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[16,17]. Despite 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 LTR 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 Figure 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). The three hornwort genomes exhibit a high gene density compared to other land plants (Supplementary Table 6).

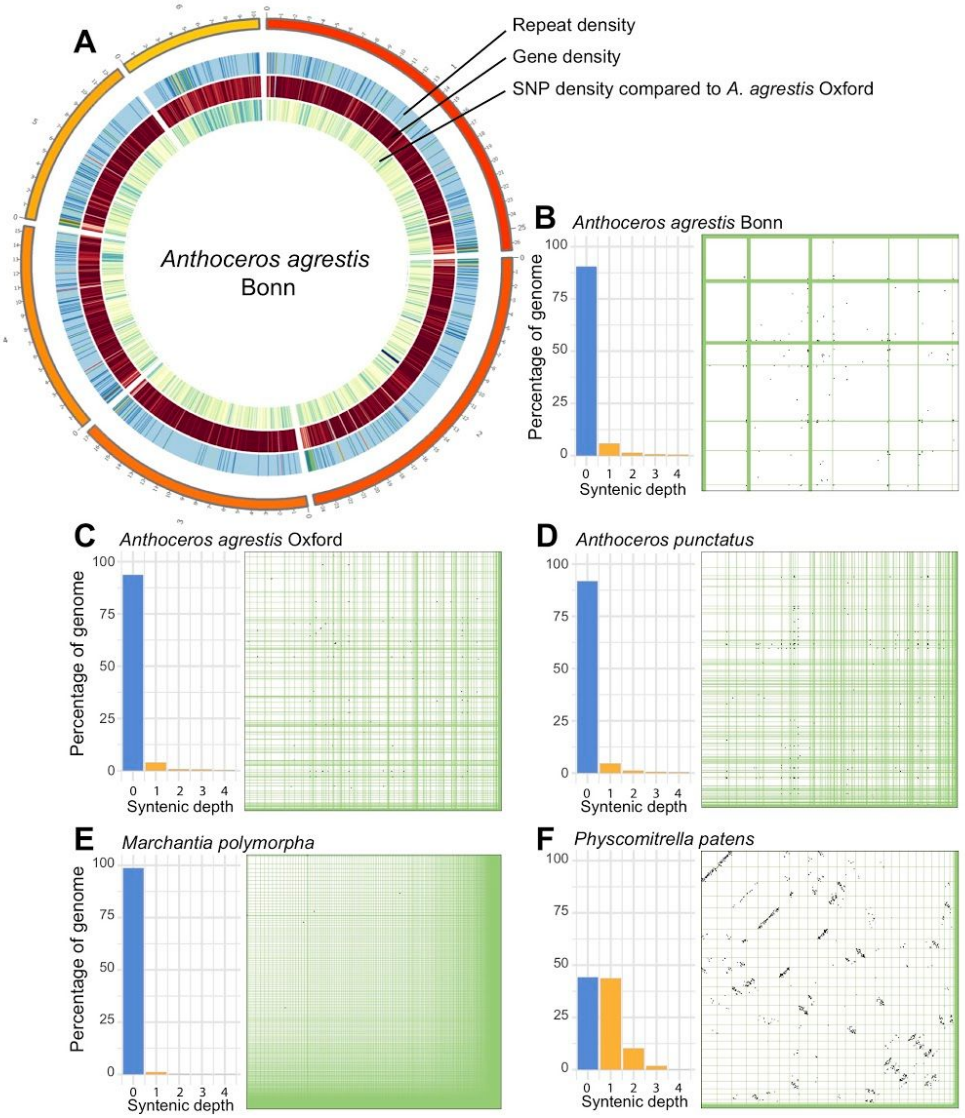

Figure 1. **Genome organizations in the *Anthoceros* genomes**. **a**, Circos plot of *A. agrestis* Bonn showing the densities of repeats, genes, and SNP with the *A. agrestis* Oxford genomes. No self-synteny can be found in the three *Anthoceros* genomes (**b-d**) nor in *Marchantia polymorpha* (**e**). *Physcomitrella patens* (**f**), on the other hand, shows a clear 1:1 and some 1:2 syntenic relationship, suggesting paleopolyploidy. In each panel the bar graph shows the proportion of the genome at different self-self syntenic levels, with the dotplot on the right.

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 100-1000bp[18]. 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 Figure 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; Supplementary Figure 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[19]. While it is tempting to hypothesize 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 assembled properly despite the long-read and Hi-C data. Future work using immuno-labeling is necessary to confirm this hypothesis.

Limited collinearity across bryophyte and vascular plant genomes

A previous study on the moss *P. patens* genome implied that regions showing collinearity between the moss and some angiosperms may represent conserved collinear blocks since the MRCA of land plants[19]. 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 Figure 6; Supplementary Table 7). The largest number of genomic blocks collinear with at least one other land plant was found in the moss, followed by the liverwort and hornwort genomes (Supplementary Figure 6). Within bryophytes, no collinear segment conserved across all the three lineages was found, although there were genomic regions exclusively collinear between each of two bryophyte genomes (Supplementary Figure 6). In general, there was more collinearity between the liverwort and the moss than between the hornwort and the liverwort/moss genomes. The number of such collinear regions, however, were small compared to those detected across vascular plants (Supplementary Figure 6). Altogether, these findings imply that the deep divergence of the moss, hornwort, and liverwort genomes may have led to limited collinearity both among bryophytes, as well as between bryophytes and vascular plants. Our results also suggest that each bryophyte genome potentially retained a unique, non-overlapping set of collinear regions from the MRCA of land 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[20]. A previous study, based on Ks divergence in transcriptomic datasets, suggested that hornworts may not have experienced any WGD event[19], similar to *M. polymorpha*[7] and *Selaginella moellendorffii*[21]. Our Ks plots on the annotated *Anthoceros* genes similarly show no sign of WGD (Supplementary Figure 7). To further corroborate this, we investigated patterns of intra-genomic synteny in the three hornwort genomes, as well as the published *M. polymorpha* and *P. patens* genomes for comparison. We found very little self-synteny in the hornwort genomes (Fig. 1b-d), thus providing a 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[19].

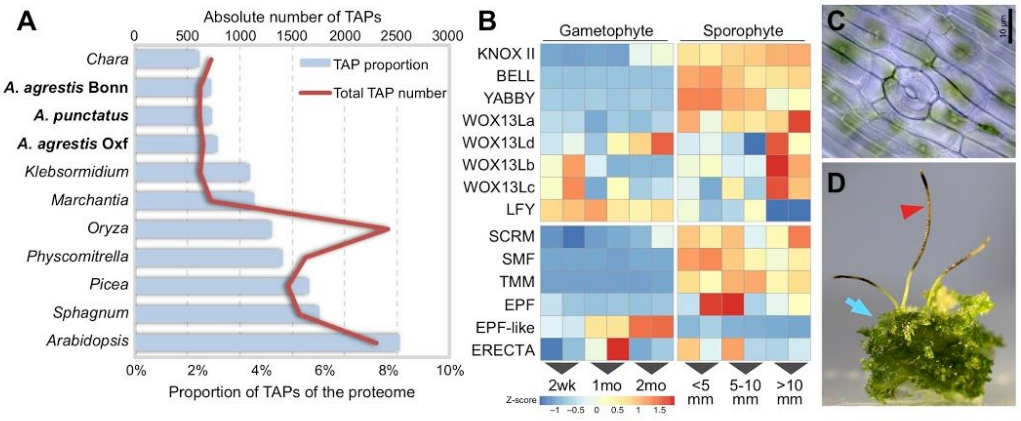

Figure 2. **Transcription associated proteins (TAPs) and sporophyte development**. **a,** The *Anthoceros* genomes have the smallest TAP repertoire among land plants. **b,** Gene expression profiles across different developmental stages in *A. agrestis* Bonn. Only selected genes known to regulate sporophyte and stomata development are shown here. **c,** Stomata of *A. agrestis* Bonn. **d,** Sporophytes (red arrowhead) and gametophytes (blue arrow) of *A. agrestis* Bonn.

Small repertoire of transcription associated proteins

We found that 2.4-2.6% of the gene set of the three *Anthoceros* genomes was annotated as transcription associated proteins (TAPs) (Fig. 2; Supplementary Table 8). Compared to other land plants[22], this is on the very low end of the spectrum. Furthermore, approximately 2/3rd (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; Supplementary Note). Such TAPs likely 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 (e.g. SRS transcription factor [TF]) or in *M. polymorpha* (e.g. 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 exhibits several unique features[23]. 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 that continuously differentiates into mature tissues towards the tip. A common origin of the indeterminate sporophyte development in hornworts and the vascular plant shoot apical meristem (SAM) has been hypothesized[23]. Lastly, hornwort sporophytes have stomata (Fig. 2c) similar to mosses and vascular plants, and the basic regulation may be shared across all stomatous lineages of land plants[24]. 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 homologs 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[25], while Class 2 *Knotted1*-like homeobox (KNOX2) genes maintain the sporophyte cell fate in *P. patens*[26]. Interestingly, the KNOX1 ortholog is lost in the *Anthoceros* genomes, and only KNOX2 genes were found (Supplementary Figure 8; Supplementary Tables 8-9). The KNOX2 orthologs showed a strong sporophyte-specific expression (Fig. 2), which implies that KNOX2's involvement 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[27]. We found that in hornworts, a single BELL and a single KNOX2 gene were expressed in the early stages of sporophyte development (Fig. 2; Supplementary Tables 8-9). This suggests that hornwort sporophyte identity may be determined by KNOX2 through interaction with BELL.

WUSCHEL-related homeobox 13 like (WOX13L) genes are involved in the zygote development and stem cell formation in the moss *P. patens*[28]*. A. thaliana WOX13* promotes replum formation in the fruit[29] and *WOX14* promotes vascular cell differentiation[30]*.* The *Anthoceros* genomes have four WOX13L members (Supplementary Figure 8; Supplementary Tables 8-9), and *WOX13La* is specifically expressed in sporophytes while *WOX13Lbcd* have expression at both gametophyte and sporophyte generations (Fig. 2b) 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 Figure 8; Supplementary Tables 8-9), which in *P. patens* and *A. thaliana* controls zygote development and SAM maintenance, respectively[31]. In hornworts, *LFY* is predominantly expressed in the gametophyte stages (Fig. 2) while in *P. patens* it is expressed both in the gametophyte and the sporophyte. It is possible that such differences may contribute to the unique developmental pattern of hornwort sporophytes.

Stomatal development in *A. thaliana* and *P. patens* is regulated by a conserved genetic toolbox, including the basic helix-loop-helix (bHLH) transcription factors SMF (*SPCH*, *MUTE*, and *FAMA*), ICE/SCREAMs (*SCRM*s), *EPIDERMAL PATTERNING FACTOR* (*EPF*), *ERECTA*, and *TOO MANY MOUTHS* (*TMM*) genes[32,33]. *FAMA* in particular is involved in the final guard cell differentiation and serves as the key switch. Orthologs of *SMF*, *TMM,* and *EPF* were absent in *M. polymorpha*, consistent with the fact that liverworts do not have stomata[7]. We found orthologs of *FAMA (SMF)*, *SCRM*, *ERECTA*, *EPF*, and *TMM* in the *Anthoceros* genomes (in line with a previous study based on our earlier genome draft[24]; Supplementary Table 10, Supplementary Figure 9). *SMF*, *SCRM*, *TMM* and *EPF* showed sporophyte-specific expression pattern (Fig. 2), suggesting that they may have similar roles in stomata patterning in hornworts. While *ERECTA* was also expressed during early sporophyte development, its expression fluctuated between replicates and 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 Figure 9), which is specifically expressed in gametophytes with a higher expression toward maturity and thus perhaps involved in a different cell-cell signaling other than stomatal regulation. *EPF4* and *EPF6* in *A. thaliana* are involved in coordination of central and peripheral zone in shoot apical meristem[34]. Taken together, our data are consistent with a single origin of

stomatal differentiation mechanism among all stomatous land plants, though the positional determination may have evolved differently (Supplementary Notes).

Genes related to phytohormone synthesis and signaling

The *Anthoceros* genomes contain the genetic chassis for the biosynthesis and signaling of abscisic acid, auxin, cytokinin, ethylene, and jasmonate (see Supplementary Notes; Supplementary Figures 10-12; Supplementary Table 10), reaffirming the origins of these pathways in the MRCA of land plants[6,7,35]. Similar to *M. polymorpha* and *P. patens*, salicylic acid signaling components are found in hornworts, but not the receptor-related genes. While *DELLA* is present, orthologs of gibberellin (GA) receptor GID1 and GA oxidases are missing from the *Anthoceros* genomes. This is consistent with the recent hypothesis that DELLA was recruited to GA signaling pathway later in plant evolution[36]. Hornworts also possess enzymes to synthesize strigolactones but genes involved in strigolactone signaling are absent. This supports the idea that strigolactones are an ancient non-hormonal signal for rhizospheric communication with mycorrhizal fungi[37]

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 plants' successful colonization and diversification on land. Evidence of AMF can be found inside plant megafossils 407 million years ago[38,39], and in almost all extant plant lineages (hornworts, liverworts, and vascular plants). Recent genetic studies have identified a suite of genes in the angiosperms that regulate the establishment and maintenance of AMF symbiosis. Some of these genes are also required for legume-rhizobial interaction, and are often referred to as the common symbiosis genes[40].

While a few components can be traced back to as far as charophyte algae[41], the question of when exactly did the entire AM symbiosis genetic network originate still remains open. This is partly because both of the bryophytes that have published genomes to date (i.e. *P. patens* and *M. polymorpha*) are incapable of AMF symbiosis and may have secondarily lost the symbiosis genes, as exemplified in some angiosperms[42]. Here we show that all the key angiosperm AMF symbiosis genes have orthologs in the three hornwort genomes (Fig. 3, Supplementary Table 11, Supplementary Figure 13). Although their roles in hornwort AM symbiosis remains to be tested, this result provides strong evidence that the genetic infrastructure required for AM symbiosis was already present in the MRCA of land plants. Importantly, the presence of these genes in liverworts[41] 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 these ortholog expressions 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 up-regulated when plants were nitrogen-starved (Fig. 3). Nitrogen limitation is a major trigger for cyanobacteria symbiosis in hornworts, which might implicate RAM1's involvement in symbiosis but further genetic studies are needed.

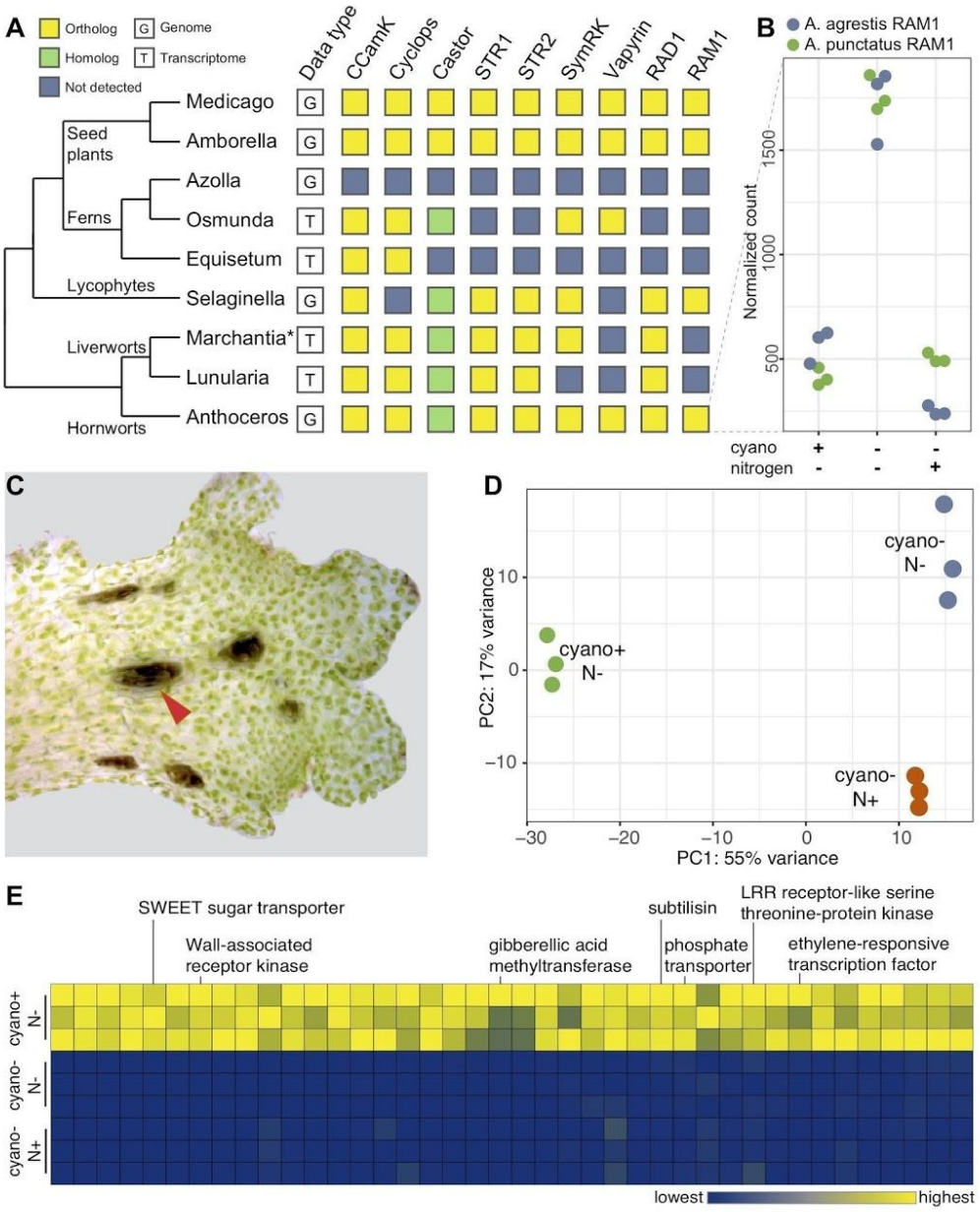

Figure 3. **Evolution and genetics of symbiosis in hornworts**. **a**, Orthologs of AMF symbiosis pathway genes can be found in hornworts, indicating their presence in the common ancestor of land plants. **Marchantia paleacea* transcriptome was searched instead of *M. polymorpha* genome because the latter secondarily lost AMF. **b**, *RAM1* is up-regulared 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*. **e**, A suite of genes were highly up-regulated (>16 folds) under symbiosis in both *A. agrestis* and *A. punctatus*.

Genes related to cyanobacterial symbiosis

Symbiosis with nitrogen-fixing cyanobacteria is a rare trait, with limited appearances in a few plant lineages[14]. In bryophytes, although mosses frequently harbor epiphytic cyanobacteria[43], only hornworts and two liverwort species host cyanobacteria endophytically within specialized slime-filled cavities[14]. 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[15], 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. 3). 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. 3; 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. 3; Supplementary Figure 14), which is minimally transcribed under the symbiont-free states but is among the highest expressed genes in symbiosis (>10^3 fold-change). The up-regulation of *SWEET* in symbiosis is interesting because it implies that this sugar transporter is dedicated to supplying carbon rewards to the cyanobionts. This implication is supported by the fact that only exogenous glucose, fructose or sucrose sustained dark nitrogen fixation in the *A. punctatus*-*N. punctiforme* association[44], and the observation that inactivation of a carbohydrate permease in *N. punctiforme* resulted in a defective symbiotic phenotype[45]. Parallelly, SWEET is involved in mycorrhizal symbiosis as well, but a different ortholog, in the *SWEET1* clade, was recruited[46].

Another gene of interest is subtilase. Members of this gene family have been shown to be highly up-regulated in a wide variety of microbial symbioses, including rhizobial[47], mycorrhizal[48], and actinorhizal[49–51] interactions. RNAi knockdown of a subtilase (*SBTM1*) in the legume *Lotus japonicus* also resulted in a decreased arbuscule formation[48]. Here we found that in both *A. punctatus* and *A. agrestis*, a subtilase homolog was similarly induced by cyanobacteria symbiosis. Phylogenetic reconstruction showed that this hornwort subtilase is not orthologous to those involved in other plant symbioses (Supplementary Figure 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 orthologs were used as in other plant-microbe symbioses.

Pyrenoid-based carbon concentrating mechanism

To enable a more efficient photosynthesis, hornworts, cyanobacteria, and many eukaryotic algae have evolved biophysical carbon-concentrating mechanism (CCM) inside individual chloroplasts[52]. In this case, 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 to install a CCM in crop plants[53].

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[54], suggesting that pyrenoid development and function is controlled by a few master switches. The genetics behind hornwort pyrenoids, however, has remained completely unknown. We hence 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[53] do not have clear homologs 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 1KP transcriptomes[5] (Fig. 4). Apart from hornworts, no *LCIB* homolog could be found in other plant genomes sequenced to date. The uniquely shared presence of LCIB in pyrenoid-bearing algae and hornworts implies that *LCIB* might have a role in the hornwort CCM. The phylogenetic tree indicates that the hornwort *LCIBs* form a clade sister to the *Klebsormidium nitens* homolog (Fig. 4) 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[55,56]. All the hornwort LCIB sequences have the conserved amino acid residues at the active sites that are shared with other algal LCIBs[57] (Fig. 4b). However, unlike *C. reinhardtii*, we did not find *LCIB* to be differentially expressed when plants are grown at different CO_2 levels (Supplementary Figure 12). This, nevertheless, cannot rule out *LCIB*'s involvement in CCM because hornwort CCM was reported to be constitutively expressed and not regulated by CO_2 level[58]. Whether *LCIB* homologs have a similar function and localization in hornworts remains to be experimentally tested.

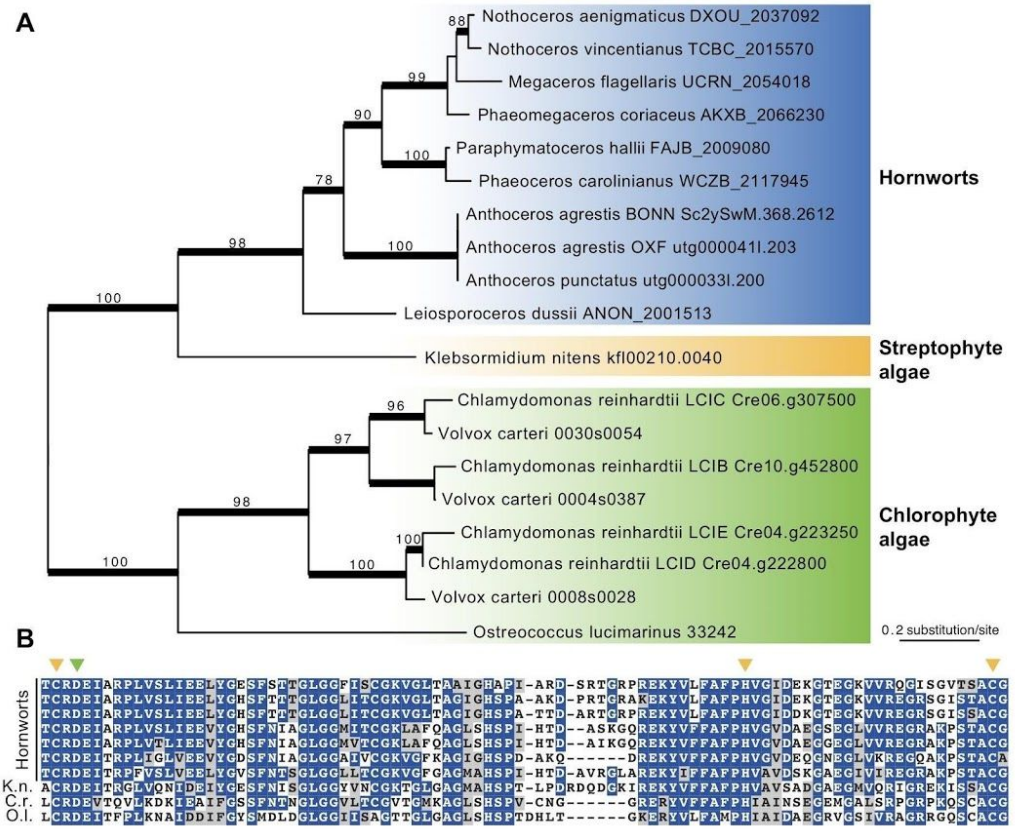

Figure 4. **Relationship between *LCIB* and pyrenoid-based carbon-concentrating mechanism**. **a**, Phylogeny of *LCIB*. Numbers above branches are bootstrap support values (branches thickened when bootstrap > 70). **b**, Hornwort *LCIBs* have the conserved amino acid residues at the active site. Yellow and green arrowheads point to the zinc-binding and catalytic residues, respectively.

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. 2). In addition, several stomata related genes are present in the *Anthoceros* genomes and expressed in early sporophyte development (Fig. 2), implying a homology of stomata at the genetic level. Finally, we found that the genes required for AM symbiosis are conserved in *Anthoceros* (Fig. 3), providing the first solid evidence that the MRCA of land plants was already equipped with the genetic network for AM symbiosis. In-depth analysis on the evolution of the plant hormones (abscisic acid, auxin, gibberellin, jasmonate, salicylic acid, and strigolactone), light signaling, peptidoglycan synthesis, and chloroplast development can be found in Supplementary Notes.

The *Anthoceros* genomes shared several features with the two other published bryophyte genomes. Most notably is the absence of tandem repeats that make up the typical centromeric regions. Further studies are needed to identify the centromeric regions and understand its structure. While *P. patens* has experienced two rounds of WGD[19], none can be found in *Anthoceros* and *M. polymorpha* (Fig. 1). This might explain the minimal representation of transcription factors in the latter 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. 3). 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. 4). Whether or not *LCIB* also participates in hornwort CCM awaits future functional characterization.

The recent advances of “seed-free genomics” have significantly improved our understanding of streptophyte evolution[7,19,21,35,59,60]. 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 *Anthoceros agrestis* (Oxford and Bonn strains) and *A. punctatus* were all derived from a single spore, and axenically propagated and maintained on either BCD[61] or Hatcher's[62] medium. Supplementary Table 13 shows the origin and specimen voucher for each of the three strains.

Chromosome count

The tip of an *A. agrestis* Oxford thallus was cut into small pieces and fixed with 4% glutaraldehyde in 0.05M phosphate buffer (pH 7.0) for 12 hours at 4°C. After washing with the buffer for 10 minutes, cell walls were digested for 2 hours with a solution containing 1% Driselase (Sigma-Aldrich, St. Louis, MO, USA), 1% Cellulase Onozuka RS (Yalult, Tokyo, Japan), 1% Pectolyase (Kikkoman, Tokyo, Japan), 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 0.05M phosphate buffer containing 0.1% TritonX-100 for 12 hours at 4°C. After several washes with the buffer, the samples were transferred onto MAS coated slide glasses (Matsunami Glass, Osaka, Japan) 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 minutes 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 minutes, the slides were mounted with Vectashield mounting medium (Vector Laboratory Burlingame, CA, USA) and observed with a fluorescence microscope under UV-light excitation.

DNA sequencing

Hornwort DNA was extracted using a CTAB-precipitation method modified from[63]. Nanopore libraries were prepared by SQK-LSK108, and sequenced on MinION R9 flow cells for 48 hours. Basecalling was done by Albacore.

For *A. agrestis* Bonn, the TrueSeq DNA Nano kit (Illumina) was used to prepare 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[64]). The library was mixed (5%) with other barcoded libraries, and sequenced on Illumina HiSeq1500 (2 lanes with Rapid mode; OnBoardClustering) at the National Institute of Basic Biology (NIBB). For *A. punctatus*, Illumina genomic libraries were prepared by BGI and sequenced on HiSeq4000. Read quality and adaptor trimming was done by fastp[65] with the default setting. For *A. agrestis* Bonn, additional Chicago and HiC libraries were prepared by DoveTail Genomics. A total of two Chicago libraries and one HiC library were prepared with a physical coverage of 300 and 200x.

To calculate the read mapping rates, trimmed reads were mapped to the final assemblies using bwa mem[66]. The mean and median insert sizes were calculated using picard collectInsertSizemetrics. Unmapped reads were counted with samtools view -f 4[67] and

divided with the total number of reads to calculate percent mapped. Reads mapped to chloroplast (Cp) and mitochondrial (Mt) genomes were counted with samtools and divided by the number of records in the bam file that is not flagged as unmapped to obtain percentage of Cp and Mt reads. 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).

Genome assembly

Genome sizes for the three *Anthoceros* were first estimated by Jellyfish[68] in conjunction with GenomeScope[69]. Draft assembly for *A. agrestis* Bonn strain was first generated using a hybrid approach including Oxford nanopore (approx. 60x) and Illumina paired-end reads (approx. 150x) using MaSuRCA version 3.2.8[70]. After assembly, base call quality was improved by two rounds of pilon polishing[71]. We mapped Chicago and Hi-C reads back to the draft assembly and used DoveTail's HiRise assembler v2.1.2[72] 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 HiC libraries and DoveTail's HiRise assembler v2.1.2[72] to derive the final assembly.

Genome assemblies of *A. agrestis* Oxford strain and *A. punctatus* were generated with the minimap2-miniasm assembler[73] using only the nanopore reads. We then used four iterations of minimap2-racon[74] to derive the consensus sequence, followed by six rounds pilon polishing[71].

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[75] 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 NCBI nt 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 Ascomycota and Cyanobacteria; 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.

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 two weeks of germination, (2) four-week-old gametophytes, (3) two-month-old gametophytes, (4) sporophytes shorter than 5 mm, (5) sporophytes 5-10 mm, (6) sporophytes longer than 1 cm with brown or black tips. Plants were grown on agar plates containing BCD medium[61] 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[76] to quality filter and trim the raw reads. Gene expression was estimated using Salmon[77] and differential expression done by DESeq2 (log2fold ≥ 2, false-discovery rate ≤ 0.05, and normalized reads counts)[78].

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 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 KNO_3, respectively and grown for 10 days before harvest. To reconstitute cyanobacterial symbiosis (with *Nostoc punctiforme* ATCC 29133), we followed the method of Enderlin and Meeks[79], 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[80], and transcript abundance quantified by Stringtie[81]. We used DESeq2[78] to carry out differential gene expression analysis, with false discovery rate set to 0.005 and log2fold change threshold set to 1. To identify genes that are differentially expressed in both *A. agrestis* Oxford and *A. punctatus*, we used on 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 in order 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 (low), 400 (ambient), and 800 (high) ppm in a CO_2-controlled growth chamber for 10 days (12/12hr 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[82] for gene ontology enrichment analysis, and REVIGO[83] to summarize and visualize the results.

Repeat annotation

For repeat annotation, we first built custom repeat libraries for each genome using RepeatModeler[84] and LTR_retriever[85]. The libraries were filtered to remove protein-coding

genes by blasting against the UniProt plant database. We then used RepeatMasker[86] 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[80]. 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[87] to filter out bad splice junctions, and Stringtie[81] to assemble the transcripts. We additionally used Trinity[88] to generate both *de novo* and genome-guided transcriptome assemblies. We combined Trinity transcripts using the PASA pipeline[89] and derived high-quality transcripts with Mikado[90]. To obtain protein homology information, we retrieved the 19 proteomes (only primary transcripts; Supplementary Table 14) and aligned them to the genome assemblies using exonerate[91]. 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[92] within Braker2[93]. 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)[94]. The following gene prediction approaches were used: A) 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. B) We used the previously trained (in A) species model but with a modified weighting file (extrinsic.cfg) to give more weight to the protein evidence. C) We trained Augustus using both protein and RNA-seq evidence within Braker2 (EPT mode of Braker2). D) We used the RNA-seq evidence to automatically train genemark and obtain gene predictions. E) Finally, we run Augustus in the comparative mode with RNA-seq, transcript, and protein evidence and genome alignments inferred by mugsy[95]. 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 14.

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, and predict alternative splice variants in two rounds.

Collinearity of the three hornwort genomes and collinearity across viridiplantae

We used the D-GENIES dot-plot tool[96] 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[97] and assessed their differences using Assemblytics[98].

To study the collinearity across all plants, we first created orthogroups with 19 species' proteomes using Orthofinder2[99]. The dataset included representatives from each major groups of land plants (Supplementary Table 14), and species experienced different numbers of large-scale duplication events[100]. Gff files and proteomes were retrieved from Phytozome v12[101]. We used I-ADHore3[102] to detect highly degenerate collinear blocks among bryophytes and vascular plants requiring a minimum of 3, 4 and 5 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[103] 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[103] to search for tandem arrays with a period length of minimum 10bp 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[104]. We translated DNA sequences and identified reading frames by comparing the Genewise[105] alignment to the best-hit protein from a collection of proteins from 25 plant genomes from Phytozome[101]. 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 constructed gene families and estimated Ks divergence using PAML[106] 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 below).

For synteny analysis, we used MCscan's "jcvi.compara.catalog ortholog"[107] function to search for and visualize intra-genomic syntenic regions, and used the "jcvi.compara.synteny depth" function for calculating syntenic depths. For comparison, we also carried out the

same analysis for *P. patens* v3.3 and *M. polymorpha* v3.0 genomes; the former is known to have two rounds of whole genome duplications while the latter has none[7,19].

Transcription factor annotation

TAPs were annotated using TAPscan, according to Wilhelmsson et al[22] 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

We used Orthofinder2[99] to classify gene families of 25 plant and algal complete genomes, including the three hornworts reported here (Supplementary Table 15) into orthogroups. Orthofinder was ran using the default setting, except 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 15.

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 13th Sept 2019. The search database was nr 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*, *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 v7.427[108]. The alignment was manually inspected in Mesquite v3.6[109] and well conserved 149 sites of 51 sequences were included. Phylogenetic analysis based on Maximum Likelihood (ML) was conducted in MEGA X[110]. 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 30th Aug 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.[111], then aligned with FFT-NS-2 in MAFFT v7.427[108]. 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*. Max 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 1000 in MAFFT v7.429[108]. The alignment was manually inspected with Mesquite v3.6[109] and a matrix consisting of 58 included sites of 142 sequences were constructed. Sequences identical in the included region were treated as a single OTU during the phylogenetic analysis. The best-fitting model was chosen as JTT with ProteinModelSelection8.pl. The ML tree was inferred by RAxM[112] with -f a -\# 100 -m PROTGAMMAJTT and supplying -p and -x from random number generator. Bootstrap samples were generated with seqboot from PHYLIP package v3.697[113] and RAxML was run for each of them.

For *YABBY*, the 107 OTU dataset from Finet et al[114] was downloaded from treebase and combined with *YABBY* genes from *Huperzia* and *Anthoceros*. The sequences were aligned using einsi of MAFFT v7.450[108]. 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 homolog sequence AagrBONN_evm.model.Sc2ySwM_368.1570.1 was used as a query to BLASTp search at NCBI on 7 Oct 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*, *K. nitens*. The word size was set to 2 and max 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 1000 in MAFFT v7.429[108]. The alignment was manually inspected with MacClade 4.08 and well-conserved 123 sites were included to result in alignment of 66 sequences. The sequence 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 (1000 replicates) were generated with seqboot from PHYLIP package v3.697[113] and RAxML[112] 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 same as ICE/SCRM case, but with LG selected as the best-fitting model. For EPF and EPF-like gene family, we used the matrix compiled by Takata et al[115] and added the *Anthoceros* and *M. polymorpha* homologs. ML tree inference was done by IQ-TREE v1.6.1 with 1,000 replicates of UltraFast Bootstraps[116].

Identification of orthologs to AMF symbiosis genes

Homologs 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 v2.9.0+[117] with a threshold e-value of $1e^{-10}$. Sequences were aligned using MAFFT v7.407[108] with default parameters and alignments were cleaned using TrimAl v1.4[118] to remove positions with more than 20% of gaps. Resulting alignments were subjected to ML tree inference using IQ-TREE v1.6.1[119]. Prior to ML analysis, the best-fitting evolutionary model was tested using ModelFinder[120] and according to the Bayesian Information Criteria. Branch support was tested using 10,000 replicates of UltraFast Bootstraps[116]. Trees were visualized with the iTOL platform v4.4.2[121].

Phylogenetic reconstruction of LCIB

The orthogroup OG0009668 was identified as the *LCIB* gene family containing *C. reinhardtii LCIB-E* genes. Additional hornwort *LCIB* homologs were retrieved from the 1KP transcriptome database[5]. To find other *LCIB* homologs, 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 based on the amino acid alignment done by MUSCLE[122]. IQ-TREE v1.6.1[119] was used to obtain the ML tree as outlined above.

Acknowledgements

This project was supported by National Science Foundation DEB1831428 to F.-W.L., Swiss National Science Foundation (grants 160004 and 131726) to P.SZ., the Georges and Antoine Claraz Foundation (Switzerland) to P.SZ., The Forschungskredit and the University Research Priority Program “Evolution in Action” of the University of Zurich to M.W and P.SZ., IOS-1339156 and EF-1550838 to M.S.B., National Institute for Basic Biology (NIBB) Collaborative Research Program (13-710) to T.N., Japan Society for the Promotion of Science KAKENHI 26650143, 18K06367 to K.S., Short Term Postdoctoral Fellowship (PE14780) to E.F., Bill & Melinda Gates Foundation grant (OPP11772165) to P.-M.D., Spanish Ministry of Science, Innovation and Universities (BFU2016-80621-P) to J.H.-G. and M.A.B., European Research Council starting grant (“TerreStriAL”) to J.dV., Foundation of German Business (sdw), Georges and Antoine Claraz Foundation, URPP Evolution in Action to A.N., German Science Foundation (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., Netherlands Organization for Scientific Research VICI grant (865.14.001) to D.W. and S.K.M. We thank Dr. Katsushi Yamaguchi and Prof. Shuji Shigenobu of Functional Genomics Facility at NIBB, Japan and Lucy Poveda, Catharine Aquino, Andrea Patrignani of the Functional Genomics Center Zurich (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., T.N. coordinated the project, M.S. carried out chromosome work, F.-W.L., P.S., T.N., D.H., S.C., G.K.-S.W. sequenced the genomes, F.-W.L., P.S. assembled the genomes, P.S., T.N. annotated the genomes, T.R., P.S. assembled and annotated organellar genomes, P.S., F.-W.L. performed synteny analyses, Z.L., M.S.B. performed Ks analyses, A.N., P.S. conducted RNA-seq experiment on developmental stages, F.-W.L., J.C.M. conducted RNA-seq experiment on cyanobacterial symbiosis, F.-W.L. conducted RNA-seq experiment on CO2 response, F.-W.L., M.W., A.K., I.D., P.S. analyzed RNA-seq data, N.P., S.R., M.W., K.S., P.S., E.F. characterized transcription factors, F.-W.L., P.S. performed gene family classification, J.K., P.-M.D. conducted analysis on AM symbiosis genes, I.M. conducted analysis on jasmonates, S.M., D.W. conducted analysis on auxin signaling, A.C. conducted analysis on ABA signaling, T.B. conducted analysis on strigolactone signaling, J.H.-G., M.A.B. conducted analysis on gibberellin signaling, S.dV. conducted analysis on salicylic acid signaling, J.dV., E.F. conducted analysis on genes associated with polyplastidy, J.H. conducted analysis on PIN proteins, S.W. conducted analysis on plastid targeted genes, E.F., T.N., F.-W.L. conducted analysis on stomatal development genes, F.-W.L., P.S., K.S., T.N., J.C.V., E.F., M.W. synthesized and wrote the manuscript.

References

1. Morris, J. L. *et al.* The timescale of early land plant evolution. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E2274–E2283 (2018).
2. 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).
3. Puttick, M. N. *et al.* The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte. *Curr. Biol.* **28**, 733–745.e2 (2018).
4. Wickett, N. J. *et al.* Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E4859–68 (2014).
5. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and phylogenomics of green plants. *Nature* **in press**, (2019).
6. Rensing, S. A. *et al.* The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science* **319**, 64–69 (2008).
7. Bowman, J. L. *et al.* Insights into land plant evolution garnered from the Marchantia polymorpha genome. *Cell* **171**, 287–299.e15 (2017).
8. Renzaglia, K. S. Comparative morphology and developmental anatomy of the Anthocerotophyta. *J. Hattori Bot. Lab.* **44**, 31–90 (1978).
9. 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).
10. Li, F.-W., Villarreal Aguilar, J. C. & Szövényi, P. Hornworts: An overlooked window into carbon-concentrating mechanisms. *Trends Plant Sci.* **22**, 275–277 (2017).

11. Qiu, Y.-L. *et al.* The deepest divergences in land plants inferred from phylogenomic evidence. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 15511–15516 (2006).

12. 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).

13. Renzaglia, K. S., Villarreal, J. C. & Duff, R. J. New insights into morphology, anatomy, and systematics of hornworts. in *Bryophyte Biology* (eds. Goffinet, B. & Shaw, J.) **2**, 139–171 (Cambridge University Press, 2009).

14. Adams, D. G. & Duggan, P. S. Cyanobacteria-bryophyte symbioses. *J. Exp. Bot.* **59**, 1047–1058 (2008).

15. Meeks, J. C. Physiological adaptations in nitrogen-fixing Nostoc–plant symbiotic associations. *Microbiological Monograph* **8**, 181–205 (2009).

16. Szövényi, P. *et al.* Establishment of Anthoceros agrestis as a model species for studying the biology of hornworts. *BMC Plant Biol.* **15**, 98 (2015).

17. Bainard, J. D. & Villarreal Aguilar, J. C. Genome size increases in recently diverged hornwort clades. *Genome* **56**, 431–435 (2013).

18. Jiang, J., Birchler, J. A., Parrott, W. A. & Dawe, R. K. A molecular view of plant centromeres. *Trends Plant Sci.* **8**, 570–575 (2003).

19. Lang, D. *et al.* The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. *Plant J.* **93**, 515–533 (2018).

20. Landis, J. B. *et al.* Impact of whole-genome duplication events on diversification rates in angiosperms. *Am. J. Bot.* **105**, 348–363 (2018).

21. Banks, J. A. *et al.* The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. *Science* **332**, 960–963 (2011).

22. 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).

23. 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).

24. 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).

25. Hay, A. & Tsiantis, M. KNOX genes: versatile regulators of plant development and diversity. *Development* **137**, 3153–3165 (2010).

26. Sakakibara, K. *et al.* KNOX2 genes regulate the haploid-to-diploid morphological transition in land plants. *Science* **339**, 1067–1070 (2013).

27. Arun, A. *et al.* Convergent recruitment of TALE homeodomain life cycle regulators to direct sporophyte development in land plants and brown algae. *Elife* **8**, (2019).

28. 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).

29. 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).

30. Denis, E. *et al.* WOX14 promotes bioactive gibberellin synthesis and vascular cell differentiation in Arabidopsis. *Plant J.* **90**, 560–572 (2017).

31. 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).

32. Lee, L. R. & Bergmann, D. C. The plant stomatal lineage at a glance. *J. Cell Sci.* **132**, (2019).

33. Chater, C. C. *et al.* Origin and function of stomata in the moss Physcomitrella patens. *Nat Plants* **2**, 16179 (2016).

34. 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).

35. Nishiyama, T. *et al.* The Chara genome: secondary complexity and implications for plant terrestrialization. *Cell* **174**, 448–464.e24 (2018).

36. 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).

37. 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).

38. Remy, W., Taylor, T. N., Hass, H. & Kerp, H. Four hundred-million-year-old vesicular arbuscular mycorrhizae. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 11841–11843 (1994).

39. Strullu-Derrien, C. Fossil filamentous microorganisms associated with plants in early terrestrial environments. *Curr. Opin. Plant Biol.* **44**, 122–128 (2018).

40. Parniske, M. Arbuscular mycorrhiza: the mother of plant root endosymbioses. *Nat. Rev. Microbiol.* **6**, 763–775 (2008).

41. Delaux, P.-M. *et al.* Algal ancestor of land plants was preadapted for symbiosis. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 13390–13395 (2015).

42. Delaux, P.-M. *et al.* Comparative phylogenomics uncovers the impact of symbiotic associations on host genome evolution. *PLoS Genet.* **10**, e1004487 (2014).

43. 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).

44. 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).

45. 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).

46. An, J. *et al.* A Medicago truncatula SWEET transporter implicated in arbuscule maintenance during arbuscular mycorrhizal symbiosis. *New Phytol.* **224**, 396–408 (2019).

47. 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).

48. 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).

49. 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).

50. 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).

51. 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).

52. Meyer, M. T., Whittaker, C. & Griffiths, H. The algal pyrenoid: key unanswered questions. *J. Exp. Bot.* **68**, 3739–3749 (2017).

53. Rae, B. D. *et al.* Progress and challenges of engineering a biophysical CO2-concentrating mechanism into higher plants. *J. Exp. Bot.* **68**, 3717–3737 (2017).

54. 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. U. S. A.* **109**, 18873–18878 (2012).

55. Wang, Y. & Spalding, M. H. LCIB in the Chlamydomonas CO2-concentrating mechanism. *Photosynth. Res.* **121**, 185–192 (2014).

56. 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).

57. Jin, S. *et al.* Structural insights into the LCIB protein family reveals a new group of β-carbonic anhydrases. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 14716–14721 (2016).

58. Hanson, D. T., Renzaglia, K. & Villarreal, J. C. Diffusion Limitation and CO2 Concentrating Mechanisms in Bryophytes. in *Photosynthesis in bryophytes and early land plants* (eds. Hanson, D. T. & Rice, S. K.) 95–111 (Springer Netherlands, 2014).

59. Li, F.-W. *et al.* Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nat. Plants* **4**, 460–472 (2018).

60. Hori, K. *et al.* Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. *Nat. Commun.* **5**, 3978 (2014).

61. Cove, D. J. *et al.* Culturing the moss Physcomitrella patens. *Cold Spring Harb. Protoc.* **2009**, db.prot5136 (2009).

62. Hatcher, R. E. Towards the establishment of a pure culture collection of Hepaticae. *Bryologist* **68**, 227–231 (1965).

63. 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* (2018). doi:10.17504/protocols.io.n5ydg7w

64. Laboratory Methods | DISCOVAR. Available at: https://software.broadinstitute.org/software/discovar/blog/?page_id=375. (Accessed: 2nd October 2019)

65. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

66. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

67. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

68. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).

69. Vurture, G. W. *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).

70. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).

71. Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).

72. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).

73. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).

74. 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).

75. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies. *F1000Res.* **6**, (2017).

76. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

77. 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).

78. 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).

79. Enderlin, C. S. & Meeks, J. C. Pure culture and reconstitution of the Anthoceros-Nostoc symbiotic association. *Planta* **158**, 157–165 (1983).

80. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).

81. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

82. 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).

83. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).

84. Smit, A. F. A. & Hubley, R. *RepeatModeler Open-1.0. Available online at http://www.repeatmasker.org*. (2015).

85. 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).

86. Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open-4.0. Available online at: http://www.repeatmasker.org*. (2015).

87. Mapleson, D., Venturini, L., Kaithakottil, G. & Swarbreck, D. Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. *Gigascience* **7**, (2018).

88. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).

89. Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).

90. 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).

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

92. 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).

93. 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).

94. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using

EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).

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

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

97. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).

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

99. Emms, D. M. & Kelly, S. OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences. *bioRxiv* **13**, 466201 (2018).

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

101. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–86 (2012).

102. 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–e11 (2012).

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

104. Barker, M. S. *et al.* EvoPipes.net: Bioinformatic Tools for Ecological and Evolutionary Genomics. *Evol. Bioinform.* **6**, 143–149 (2010).

105. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).

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

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

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

109. Maddison, W. P. & Maddison, D. R. Mesquite: a modular system for evolutionary analysis. Version 3.04. 2015. *Available from: http://mesquiteproject. org (Accessed 5 Jul. 2016)* (2015).

110. 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).

111. Sayou, C. *et al.* A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. *Science* **343**, 645–648 (2014).

112. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of

large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

113. Felsenstein, J. PHYLIP (phylogeny inference package) version 3.695. *Distributed by the Author* (2013).

114. Finet, C. *et al.* Evolution of the YABBY gene family in seed plants. *Evol. Dev.* **18**, 116–126 (2016).

115. Takata, N. *et al.* Evolutionary relationship and structural characterization of the EPF/EPFL gene family. *PLoS One* **8**, e65183 (2013).

116. 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).

117. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

118. 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).

119. 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).

120. 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).

121. 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**, W242–5 (2016).

122. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).