

# Lawrence Berkeley National Laboratory

## Recent Work

### Title

The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution.

### Permalink

<https://escholarship.org/uc/item/1k13n9zp>

### Journal

The Plant journal : for cell and molecular biology, 93(3)

### ISSN

0960-7412

### Authors

Lang, Daniel  
Ullrich, Kristian K  
Murat, Florent  
[et al.](#)

### Publication Date










2018-02-01

### DOI

10.1111/tpj.13801

Peer reviewed

# The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution

Daniel Lang<sup>1,2,#</sup> , Kristian K. Ullrich<sup>3,#,†</sup> , Florent Murat<sup>4</sup>, Jörg Fuchs<sup>5</sup>, Jerry Jenkins<sup>6</sup>, Fabian B. Haas<sup>3</sup> , Mathieu Piednoel<sup>7</sup>, Heidrun Gundlach<sup>2</sup>, Michiel Van Bel<sup>8,9</sup>, Rabea Meyberg<sup>3</sup>, Cristina Vives<sup>10</sup>, Jordi Morata<sup>10</sup>, Aikaterini Symeonidi<sup>3,‡</sup>, Manuel Hiss<sup>3</sup>, Wellington Muchero<sup>11</sup>, Yasuko Kamisugi<sup>12</sup> , Omar Saleh<sup>1,§</sup>, Guillaume Blanc<sup>13</sup>, Eva L. Decker<sup>1</sup>, Nico van Gessel<sup>1</sup>, Jane Grimwood<sup>6,14</sup>, Richard D. Hayes<sup>14</sup>, Sean W. Graham<sup>15</sup>, Lee E. Gunter<sup>11</sup>, Stuart F. McDaniel<sup>16</sup>, Sebastian N.W. Hoernstein<sup>1</sup>, Anders Larsson<sup>17</sup>, Fay-Wei Li<sup>18</sup>, Pierre-François Perroud<sup>3</sup> , Jeremy Phillips<sup>14</sup>, Priya Ranjan<sup>11</sup>, Daniel S. Rokhsar<sup>14,19</sup>, Carl J. Rothfels<sup>20</sup>, Lucas Schneider<sup>3,¶</sup>, Shengqiang Shu<sup>14</sup>, Dennis W. Stevenson<sup>21</sup>, Fritz Thümmeler<sup>22</sup>, Michael Tillich<sup>23</sup>, Juan C. Villarreal Aguilar<sup>24</sup>, Thomas Widiez<sup>25,26,\*\*</sup>, Gane Ka-Shu Wong<sup>27,28,29</sup>, Ann Wymore<sup>11</sup>, Yong Zhang<sup>30</sup>, Andreas D. Zimmer<sup>1,††</sup>, Ralph S. Quatrano<sup>31</sup>, Klaus F.X. Mayer<sup>2,32</sup>, David Goodstein<sup>14</sup>, Josep M. Casacuberta<sup>10</sup>, Klaas Vandepoele<sup>8,9</sup> , Ralf Reski<sup>1,33</sup> , Andrew C. Cuming<sup>12</sup> , Gerald A. Tuskan<sup>11</sup>, Florian Maumus<sup>34</sup>, Jérôme Salse<sup>4</sup>, Jeremy Schmutz<sup>6,14</sup> and Stefan A. Rensing<sup>3,33,\*</sup> 

<sup>1</sup>Plant Biotechnology, Faculty of Biology, University of Freiburg, Schaezlestr. 1, 79104, Freiburg, Germany,

<sup>2</sup>Plant Genome and Systems Biology, Helmholtz Center Munich, 85764, Neuherberg, Germany,

<sup>3</sup>Plant Cell Biology, Faculty of Biology, University of Marburg, Marburg, Germany,

<sup>4</sup>INRA, UMR 1095 Genetics, Diversity and Ecophysiology of Cereals (GDEC), 5 Chemin de Beaulieu, 63100, Clermont-Ferrand, France,

<sup>5</sup>Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstrasse 3, OT Gatersleben, D-06466, Stadt Seeland, Germany,

<sup>6</sup>HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA,

<sup>7</sup>Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linné Weg 10, D-50829, Cologne, Germany,

<sup>8</sup>VIB Center for Plant Systems Biology, Technologiepark 927, 9052 Ghent, Belgium,

<sup>9</sup>Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, B-9052, Gent, Belgium,

<sup>10</sup>Center for Research in Agricultural Genomics, CRAG (CSIC-IRTA-UAB-UB), Campus UAB, Bellaterra, Cerdanyola del Vallès, 08193, Barcelona, Spain,

<sup>11</sup>Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA,

<sup>12</sup>Centre for Plant Sciences, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, UK,

<sup>13</sup>Structural and Genomic Information Laboratory (IGS), Aix-Marseille Université, CNRS, UMR 7256 (IMM FR 3479), Marseille, France,

<sup>14</sup>DOE Joint Genome Institute, Walnut Creek, CA 94598, USA,

<sup>15</sup>Department of Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada,

<sup>16</sup>Department of Biology, University of Florida, Gainesville, FL 32611, USA,

<sup>17</sup>Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden,

<sup>18</sup>Boyce Thompson Institute, Ithaca, NY 14853, USA,

<sup>19</sup>Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA,

<sup>20</sup>University Herbarium and Department of Integrative Biology, University of California, Berkeley, CA 94720-2465, USA,

<sup>21</sup>New York Botanical Garden, Bronx, NY 10458, USA,

<sup>22</sup>Vertis Biotechnologie AG, Lise-Meitner-Str. 30, 85354, Freising, Germany,

<sup>23</sup>Max Planck Institute of Molecular Plant Physiology, Am Muehlenberg 1, 14476, Potsdam-Golm, Germany,

<sup>24</sup>Department of Biology, Université Laval, Québec G1V 0A6, Canada,

<sup>25</sup>Department of Plant Biology, University of Geneva, Sciences III, Geneva 4 CH-1211, Switzerland,

<sup>26</sup>Department of Plant Biology & Pathology Rutgers, The State University of New Jersey, New Brunswick, NJ 08901, USA,

<sup>27</sup>Department of Biological Sciences, University of Alberta, Edmonton, AB, T6G 2E9, Canada,

<sup>28</sup>Department of Medicine, University of Alberta, Edmonton, AB T6G 2E1, Canada,

<sup>29</sup>BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China,

<sup>30</sup>Shenzhen Huahan Gene Life Technology Co. Ltd, Shenzhen, China,

<sup>31</sup>Department of Biology, Washington University, St. Louis, MO, USA,

<sup>32</sup>WZV, Technical University Munich, Munich, Germany,

<sup>33</sup>BIOSS Centre for Biological Signalling Studies, University of Freiburg, Schaezlestr. 18, 79104, Freiburg, Germany,

<sup>34</sup>URGI, INRA, Université Paris-Saclay, 78026, Versailles, France,

Received 11 October 2017; revised 20 November 2017; accepted 24 November 2017; published online 13 December 2017.

\*For correspondence (e-mail stefan.rensing@biologie.uni-marburg.de).

#These authors contributed equally to this work.

† Present address: Max Planck Institute for Evolutionary Biology, August-Thienemann-Straße 2, 24306, Ploen, Germany.

‡ Present address: Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain.

§ Present address: Plant Molecular Cell Biology, Humboldt-University of Berlin, 10115, Berlin, Germany.

¶ Present address: Institute for Transfusion Medicine and Immunohematology, Johann-Wolfgang-Goethe University and German Red Cross Blood Service, Sandhofstraße 1, 60528, Frankfurt am Main, Germany.

\*\* Present address: Laboratoire Reproduction et Développement des Plantes, Univ Lyon, ENS de Lyon, UCB Lyon 1, CNRS, INRA, F-69342, Lyon, France.

†† Present address: Faculty of Medicine, Institute for Human Genetics, Medical Center – University of Freiburg, Freiburg, Germany.

## SUMMARY

The draft genome of the moss model, *Physcomitrella patens*, comprised approximately 2000 unordered scaffolds. In order to enable analyses of genome structure and evolution we generated a chromosome-scale genome assembly using genetic linkage as well as (end) sequencing of long DNA fragments. We find that 57% of the genome comprises transposable elements (TEs), some of which may be actively transposing during the life cycle. Unlike in flowering plant genomes, gene- and TE-rich regions show an overall even distribution along the chromosomes. However, the chromosomes are mono-centric with peaks of a class of Copia elements potentially coinciding with centromeres. Gene body methylation is evident in 5.7% of the protein-coding genes, typically coinciding with low GC and low expression. Some giant virus insertions are transcriptionally active and might protect gametes from viral infection via siRNA mediated silencing. Structure-based detection methods show that the genome evolved via two rounds of whole genome duplications (WGDs), apparently common in mosses but not in liverworts and hornworts. Several hundred genes are present in colinear regions conserved since the last common ancestor of plants. These syntenic regions are enriched for functions related to plant-specific cell growth and tissue organization. The *P. patens* genome lacks the TE-rich pericentromeric and gene-rich distal regions typical for most flowering plant genomes. More non-seed plant genomes are needed to unravel how plant genomes evolve, and to understand whether the *P. patens* genome structure is typical for mosses or bryophytes.

**Keywords:** evolution, genome, chromosome, plant, moss, methylation, duplication, synteny, *Physcomitrella patens*.

## INTRODUCTION

The original genome sequencing of the model moss *Physcomitrella patens* (Hedw.) Bruch & Schimp. (Funariaceae) reflected its informative phylogenetic position: a very early divergence from the evolutionary path that eventually led to the flowering plants soon after the first plants conquered land ca. 500 Ma ago (Lang *et al.*, 2010). Previous comparisons of the moss genome with those of flowering plants and green algae provided many insights into land plant evolution (Rensing *et al.*, 2008), detailing for example the evolution of abiotic stress responses and phytohormone signaling. Subsequent comparative functional genomic analyses, making use of the ability of *P. patens* for ‘reverse genetics’ by gene targeting, addressed questions of how gene functions evolved to enable the increasing developmental and anatomical complexity that characterizes the dominant forms of plant life on the planet (e.g. Horst *et al.*, 2016; Sakakibara *et al.*, 2013). The initial draft sequence encompassed close to 2000 unordered scaffolds, significantly limiting analyses of chromosomal structure and

evolution, or of the conservation of gene order during land plant evolution. We now present a new assembly accurately representing the chromosomal architecture (pseudochromosomes). Much-increased acquisition of transcriptomic evidence has substantially improved the quality of gene annotation, and acquisition of high-density DNA methylation and histone mark data combined with a detailed analysis of transposable elements (TEs) explain the size and architecture of the moss genome. This study provides unprecedented insights into the genome of a haploid-dominant land plant, such as the peculiar structure and evolution of moss chromosomes, and demonstrates syntenic conservation of important plant genes throughout 500 Ma of evolution.

## RESULTS AND DISCUSSION

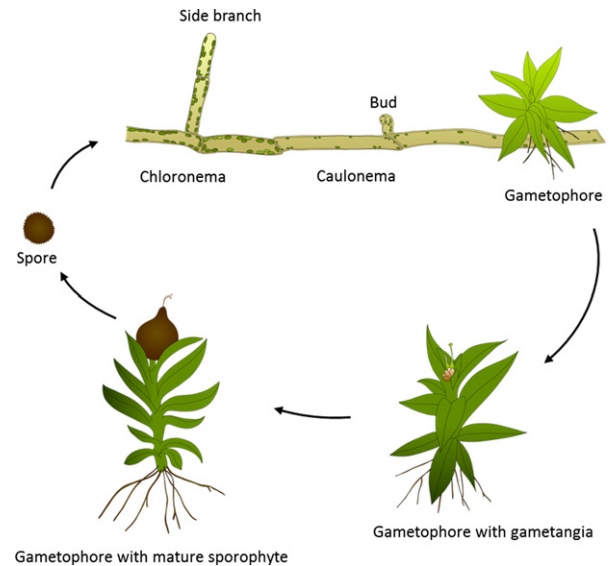
### The moss V3 genome: assembly and annotation

The original genome sequence (V1.2) of *Physcomitrella patens* (strain Gransden 2004) comprised 1995 sequence

scaffolds (Rensing *et al.*, 2008; Zimmer *et al.*, 2013). Here, we integrated the previous sequence data with a high-density genetic linkage map based on 3712 SNP segregating loci in a cross between the 'Gransden 2004' (Gransden) laboratory strain and the genetically divergent 'Villersexel K3' (Villersexel) accession (Kamisugi *et al.*, 2008). The resulting assembly was further improved using novel BAC/fosmid paired end sequence data (cf. Appendix S1, Supplementary Material I for details; see section Availability of gene models and additional data for novel data associated with this study). We screened the subsequent integrated assembly for sequence contamination, producing a pseudomolecule release covering 27 nuclear chromosomes with a total genetic linkage distance of 5502.6–5503.1 centiMorgans (cM). The 27 chromosomal pseudomolecules include 462.3 Mbp of sequence, supplemented by 351 unplaced scaffolds representing 4.9 Mbp (1%) of unintegrated sequence, totaling 90% of the 518 Mbp estimated by flow cytometry (Schween *et al.*, 2003). The reads partitioned as mitochondrial and plastidial were assembled *de novo*, yielding an improved assembly and annotation of both organellar genomes (correcting e.g. the N-terminal sequence of the plastidial RuBisCO). Structural annotation used substantial new transcript evidence (File S3). For parameter optimization it relied on a manually curated reference gene set (Zimmer *et al.*, 2013), yielding gene annotation version 3.1. Of 35 307 predicted protein-coding genes, 27 511 (78%) could be functionally annotated (cf. Appendix S1, Supplementary Material II and File S1), i.e. encode known domains and/or encode homologs of proteins in other species. In total, 20 274 (57%) genes are expressed based on RNA-seq evidence of typical developmental stages covered by the JGI gene atlas project (<http://jgi.doe.gov/our-science/science-programs/plant-genomics/plant-flagship-genomes/>); the remaining genes might be expressed in as yet unrepresented stages such as mature spores or male gametes. We found 13 160 genes to be expressed in the juvenile gametophyte (Figure 1), the filamentous protonemata, 12 714 in the adult gametophyte, the leafy gametophores, and 14 309 in the diploid sporophytes developing from the zygote (overlap: 10 388 genes expressed in all three developmental stages).

**Unusual genome structure**

*Transposon content and activity.* *De novo* analyses of repeated sequences revealed that the genome is highly repetitive, with 57% of the assembly comprising TEs, tandem repeats, unclassified repeats, and segments of host genes (cf. Appendix S1, Supplementary Material III and Table S13). The vast majority of TEs are long terminal repeat (LTR) retrotransposons (RT), strongly dominated by Gypsy-type elements that contribute almost 48%, with Copia-type elements much less abundant (3.5%). The estimated relative insertion times of LTR-RTs confirm the



**Figure 1.** The *P. patens* life cycle. Germination of haploid spores yields the juvenile gametophytic generation, the protonema. Protonema grows two-dimensional by apical (tip) growth and side branching. Protonemata consist of chloroplast-rich chloronema cells, and longer, thinner caulonema cells featuring less chloroplasts and oblique cross walls. Three-faced buds featuring single apical stem cells emerge from side branches (Harrison *et al.*, 2009) to form the adult gametophytic phase, the leafy gametophores. Gametophores comprise basal, multicellular rhizoids for nutrient supply, as well as non-vascular leaves (phyllids). Gametangia (female archegonia and male antheridia) develop on the gametophores. Upon fertilization of the egg cell by motile spermatozooids the diploid zygote forms and subsequently performs embryogenesis. Spore mother cells in the diploid sporophyte undergo meiosis to form spores.

limited accumulation of Copia-type elements over a prolonged evolutionary time. By contrast, two peaks of Gypsy-type elements testify to both ancient and recent periods of significant TE activity (Figure S7). Phylogenetic inference revealed the presence of five main LTR-RT groups including three Gypsy-type (RLG1-3) and two Copia-type elements (RLC4-5; Figure S8). Applying a molecular clock based on sequence divergence to the full length, intact LTR-RTs indicates that the latest (<1 Ma) activity of Gypsy-type elements was mostly contributed by RLG1-3 elements, preceded by the amassing of RLG2 and RLC5 copies (around 4–6 Ma, Figures S7 and S36). RLG1 thus comprises the youngest and most abundant group among intact LTR-RTs. In line with these results, analysis of TE insertion polymorphisms between Gransden and Villersexel showed that RLG1 elements are highly polymorphic, accounting for most of the detected insertion variants (Figure S9). Since we detect such insertions in both accessions, the decades long *in vitro* culture of Gransden is not likely to be the major source of transposon activity. RLG1 elements are expressed in non-stressed protonemata (Figure S6), which is uncommon as transposon expression is usually strongly silenced in

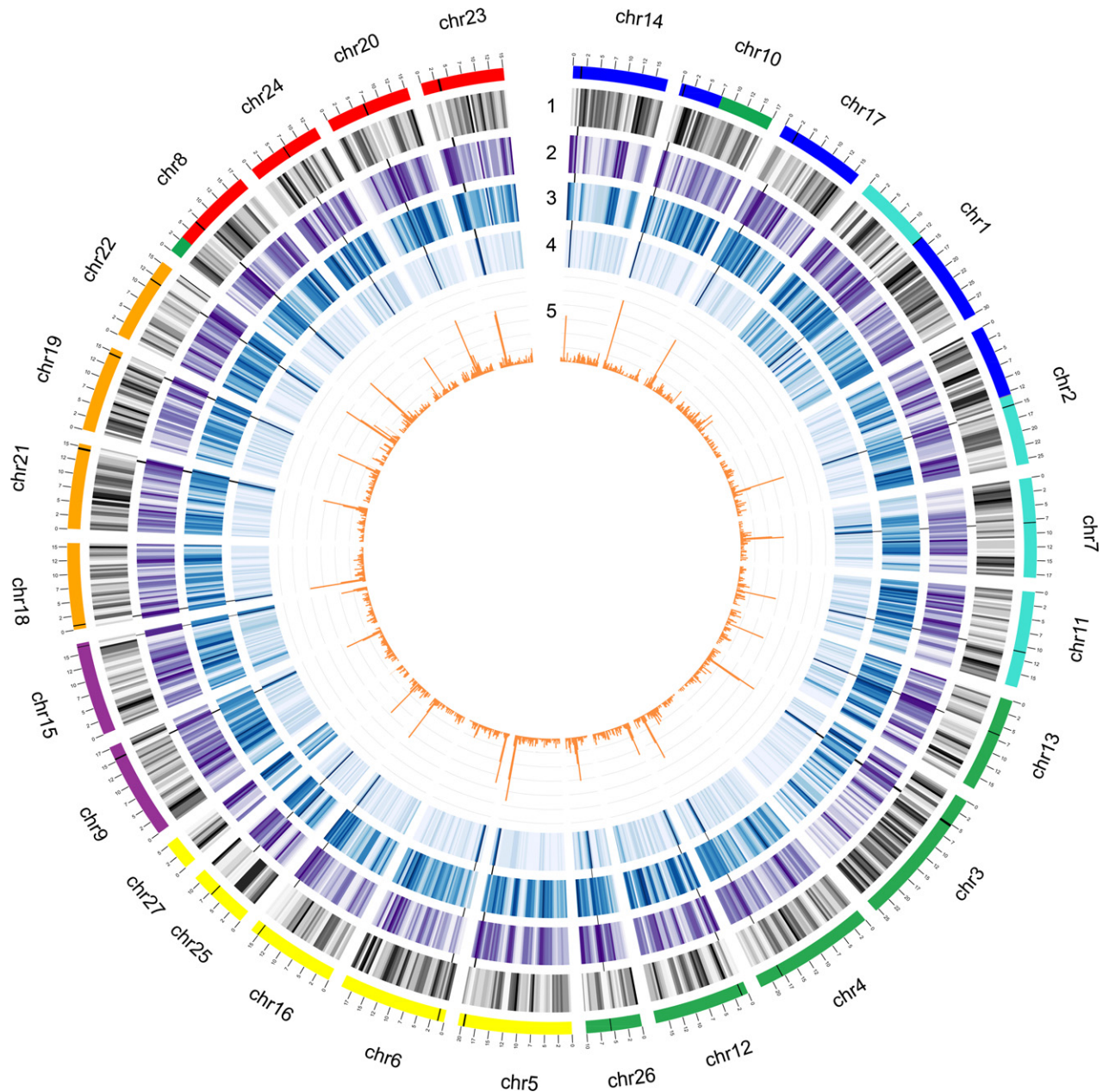
plants and is only detected in very specific tissues such as pollen, in silencing mutants or under stress situations (Martinez and Slotkin, 2012). Moreover, recent data suggest that some stresses that typically induce plant retrotransposons, such as protoplastation, inhibit RLG1 expression (Vives *et al.*, 2016), suggesting that RLG1 may transpose during the *P. patens* life cycle and might play a role in its genome dynamics. The moss germinates from spores that develop into filamentous, tip-growing protonemata (comprising chloroplast-rich chloronemal and fast-growing caulonemal cells; Figure 1). Buds develop from caulonemal cells and grow into gametophores that bear sexual organs (gametangia). Mosses are prone to endopolyploidy (Bainard and Newmaster, 2010) and older *P. patens* caulonema cells endoreduplicate (Schween *et al.*, 2005). Interestingly, endoreduplicated caulonemal cells give rise to somatic sporophytes if PpBELL1 is overexpressed, thus circumventing sexual reproduction (Horst *et al.*, 2016). *De facto* 2n caulonemal cells might constitute a staging ground for (potentially transmitted) somatic changes caused *via* transposon activity.

**Unusual chromatin structure.** The genomes of most flowering plants are typically composed of monocentric chromosomes, whose unique centromeres are surrounded by heterochromatic pericentromeric regions, that are repeat-rich and gene-poor relative to distal (sub-telomeric), euchromatic regions (Lamb *et al.*, 2007; Figure S34). By contrast, the landscape of gene and repeat density along *P. patens* chromosomes is rather homogeneous, we do not detect large repeat-rich regions with relatively low gene density (Figures 2 and 3). At a finer scale, we do detect an alternation of gene-rich and repeat-rich regions all along the chromosomes (Figure S10). Typical plant pericentromeres are more prone to structural variation (e.g. TE insertions and deletions) compared with the remainder of chromosome arms (Li *et al.*, 2014). Yet, analysis of *P. patens* chromosomes failed to identify hotspots of structural variation that could coincide with pericentromeres (Figure S11). It should be noted, however, that the centromeres could be present at least partially in the unassembled parts of the genome. In any case, immuno-labeling of mitotic metaphase chromosomes using a pericentromere-specific antibody demonstrates that they are mono-centric (Figure S5). Unlike in many flowering plant genomes, the *P. patens* chromosomes are characterized by a more uniform distribution of eu- and heterochromatin (Figures 3, S5 and S35), raising questions about the nature and location of centromeres.

**Physcomitrella centromeres seem to coincide with a particular subset of Copia elements.** Plant centromeres typically comprise large arrays of satellite repeats that can be punctuated by some TEs (Wang *et al.*, 2009). However,

plotting the density of tandem repeats along the *P. patens* chromosomes did not reveal peaks likely to reflect the position of centromeres (Figure S11). Computational analysis of tandem repeats in a variety of genomes identified candidate centromeric repeats in *P. patens*, although green algae, mosses, and liverworts contain low abundances of these (Melters *et al.*, 2013). Positioning them on the *P. patens* V3 assembly revealed a patchy distribution, not single peaks that could coincide with centromeres as expected for monocentric chromosomes (Figures S5 and S11). By contrast, the low abundance Copia-type elements exhibited unusually discrete density peaks, typically one per assembled chromosome, spanning hundreds of kbp (Figures 2 and S11). Each Copia density peak principally contains RLC5 elements. A similar situation has been described in the green alga *Coccomyxa subellipsoidea* in which a single peak of a LINE-type retrotransposon, the Zepp element, was proposed to be involved in centromeric function (Blanc *et al.*, 2012). The RLC5 density peak regions are generally punctuated by unresolved gaps in the assembly and by fragments of other TEs (Figure S12). Closer examination revealed that they comprise full length LTR-RTs (FL\_RLC5) as well as highly similar truncated non-autonomous variants (Tr\_RLC5) that lack the integrase (INT) and reverse transcriptase domains (RVT) (Figure S13). Remarkably, all RLC5 clusters appear to be mosaics containing nested insertions of both FL\_RLC5 and Tr\_RLC5 elements, of which additional copies are rare in the genome. A neutral explanation for the distribution of RLC5 clusters is that their target sequences are present at a single location per chromosome, perhaps caused by a preference for self-insertion. Alternatively, a single cluster combining FL\_RLC5 and Tr\_RLC5 copies may be necessary for normal chromosome function. In either case, it is possible that RLC5 clusters might be specific components of centromeres in *P. patens*. The dominant RLC5 peak per chromosome, highlighting the putative centromere, is marked by a radius in Figures 2 and 4.

**Alternation of activating and repressing epigenetic marks.** For the V1.2 scaffolds that harbor histone 3 (H3) ChIP-seq evidence (Widiez *et al.*, 2014), 96% can be mapped to the 27 V3 pseudochromosomes (Figure 4); the remaining 4% map to the unassigned V3 scaffolds, underscoring the quality of the assembly. The alternating structure of genes and TE/DNA methylation (purple in Figure 4) over the full length of the chromosomes is mirrored by activating H3 marks (K4me3, K27Ac, K9Ac; green in Figure 4) corresponding to transcribed genic areas, and repressive H3 marks (K27me3, K9me2; red in Figure 4) coinciding with TEs/intergenic areas. This result contrasts sharply with many flowering plant genomes (Figure S34) in which gene-rich chromosome arms display less heterochromatin than pericentromeres. Similar

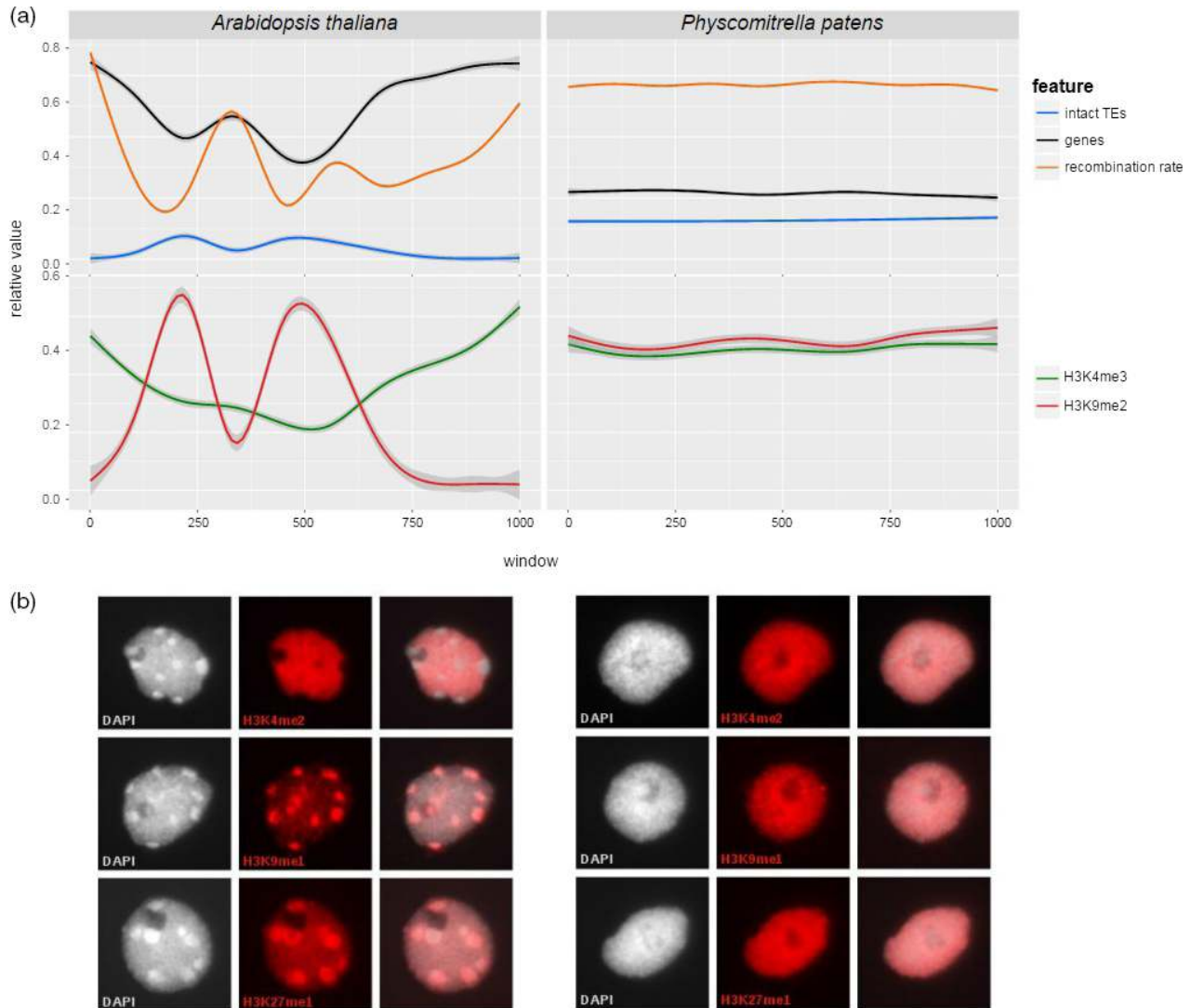


**Figure 2.** Chromosome structure, focus on TEs.

From outer to inner: karyotype bands colored according to ancestral genome blocks as in Figure 5 (scale = Mbp), followed by: (1) gene density (grey, normalized 0,1); (2) repeat density (violet, normalized 0,1); (3) gypsy-type elements (blue, normalized 0,1); (4) Copia-type elements (blue, normalized 0,1); and (5) RLC5 elements (orange, histogram). For each chromosome, a radius marks the dominant RLC5 peak, potentially coinciding with the centromere (see text). All plots are based on a 500 kbp sliding window (400 kbp jump). Chromosomes are arranged according to the ancestral (pre-WGD) seven chromosome karyotype (Figure 5).

to flowering plant genomes, TE bodies are generally depleted for histone marks, excepting the silencing mark H3K9me2 that is above background levels in the filamentous protonemata, and at background level in unstressed and stressed leafy gametophores (File S2). The previously described (Widiez *et al.*, 2014) deposition of H3K27me3 at developmental genes that takes place with the switch from protonema to gametophore (Figure 1)

can be observed genome-wide (File S2). All TE bodies are methylated in similar fashion, with CG and CHG more abundant than CHH (>80% CG and CHG, >40% CHH; Figures S15 and S25–S28), whereas gene bodies remain barely methylated (Figures S15 and S25–S29). RLC4 has the sharpest boundary pattern (File S2), with almost no methylation outside the TE, followed by RLC5 with more outside-TE methylation, especially CHH. RLG1



**Figure 3.** Comparative analysis of genome structures.

Comparative data of *Arabidopsis thaliana* (left) and *Physcomitrella patens* (right) reveals the lack of large heterochromatic blocks (b) that is mirrored by even distribution of recombination rate, gene and LTR-RT distribution (a) in the moss.

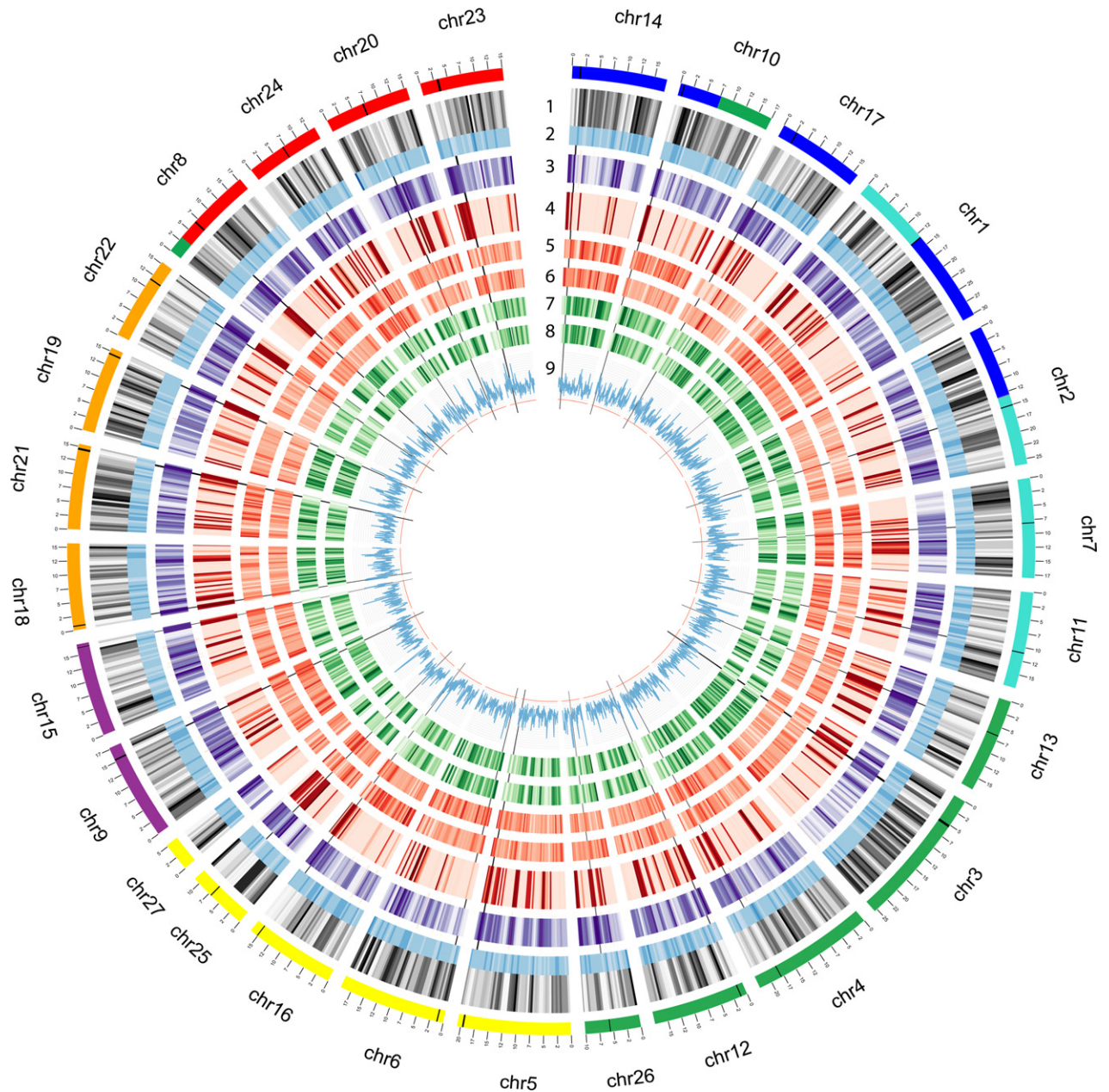
(a) Averaged topology of genomic features based on 1000 non-overlapping windows per chromosome (averaged over all chromosomes); arbitrary units, 1000 representing the full length of the averaged chromosomes. Upper track: Smoothed chromosomal densities of intact LTRs, protein-coding genes and the normalized mean recombination rate. Lower track: Smoothed density curves of H3K4me3 and H3K9me2 histone modification peak regions.

(b) Immunostaining of typical eu- and heterochromatin-associated histone methylation marks (H3K4me2, H3K9me1 and H3K27me1) on flow-sorted interphase nuclei.

follows in a similar fashion, although the relatively sharp pattern of RLG1 and RLC5 can in part be attributed to the fact that in case of nested insertions no 'outside' TE region is present next to the TE boundary. RLG2 shows a broad pattern of all three contexts, RLG3 shows the broadest pattern with no discernible body peak. As the methylation pattern of the main TE categories differs in how sharply they define the TE proper, TE families might have different impacts on the proximal epigenome.

*Gene body methylation marks low GC genes.* Interestingly, intron-containing genes (Figure S25) show a much

sharper methylation contrast between gene body and surrounding DNA, and a more pronounced difference between CHH and the other contexts, than intron-less genes (Figure S26). As the latter genes might in part be retrocopies (Kaessmann, 2010), they might be more prone to silencing and be embedded in more homogeneously methylated areas. Gene-body methylation (GBM) is found in many eukaryotic lineages and is thought to have been present in the last common eukaryotic ancestor (Feng *et al.*, 2010). GBM in flowering plants is characterized by CG methylation of the coding sequence, not extending to transcriptional start and stop (Niederhuth *et al.*, 2016).



**Figure 4.** Chromosome structure, focus on epigenetic marks.

From outer to inner: karyotype bands colored according to ancestral genome blocks as in Figure 5, followed by: (1) gene density (grey) normalized 0,1; (2) GC content 0.25–0.45 (blue); (3) all TEs density (violet) normalized 0,1, NCLDV evidence is shown as radial orange lines; (4) methylation (red): CHH+CHG+CG, each median per window normalized 0,1, 0.0–3.0 (individual tracks see Figure S32); (5) gametophore H3 repression marks (red, K27me3, K9me2) percent per window normalized, 0.0–2.0 (for more detailed plots see File S1); (6) protonema H3 repression marks (red, K27me3, K9me2) normalized as in (5); (7) gametophore H3 activation marks (green, K4me3, K27Ac, K9Ac) normalized as in (5); (8) protonema H3 activation marks (green, K4me3, K27Ac, K9Ac) normalized as in (5); (9) Nucleotide diversity (blue histogram) 0.0–0.01. Dominant RLC5 peak radius as in Figure 2. (9) 100 kbp sliding window and 100 kbp jump, all other plots as in Figure 2. Chromosomes are arranged according to the ancestral (pre-WGD) seven chromosome karyotype (Figure 5).

Such genes are typically constitutively expressed and evolutionarily conserved; however, the functional relevance of GBM in flowering plants remains unclear (Zilberman, 2017). The low incidence of genic methylation in *P. patens*, although all DNA methyltransferase classes are present (Dangwal *et al.*, 2014), probably reflects secondary

reduction. Despite the generally low genic methylation, 2012 (5.7%) protein-coding genes contain at least one methylated position in gametophores (Figure S29), and 1155 (3.3%) of the genes show more than 50% of methylatable positions to be methylated (Figure S30), making them GBM candidates. Most methylated genes are not



expressed in gametophores (1608 genes, 79.9%), suggesting that, contrary to flowering plants, GBM might silence them. They are also significantly less often annotated (21.7% of methylated genes carry GO terms, versus 48.7% of all genes;  $P < 0.01$ , chi-squared test). CHH-type methylation is most abundant (1409 genes), followed by CHG (1306) and CG (1162); one-third of the genes share methylation in all three contexts. The presence of CG methylation in *P. patens* gene bodies is in contrast with a previous report (Bewick *et al.*, 2017), potentially due to different coverage or filtering applied. Surprisingly, given that cytosines are methylated, the average GC content of GBM genes (36.5%) is significantly ( $P < 0.01$ , T-test) lower than the genome-wide GC (45.9%). Genes without expression evidence in gametophores have lower GC content and GBM than those that are weakly expressed (Table S18, RPKM 0–2), while confidently expressed genes (RPKM >2) are more GC-rich and less methylated. In summary, in contrast with flowering plants low GC genes with no conserved function are principally more often found to be targeted (silenced) by DNA methylation, suggesting their potential conditional activation. GO bias analysis of the methylated genes expressed in gametophores shows enrichment of genes involved in protein phosphorylation (Figure S30(b)). Most (290, 59%) of the expressed methylated genes are expressed in protonema, gametophores and green sporophytes (Figure S30(c)), but 12.5% are expressed in two tissues each, while 17 (3.5%) are exclusively expressed in protonemata, 28 (5.7%) in gametophores and 93 (19%) in green sporophytes.

*Do giant virus remnants guard gametes?* We mapped the genomic segments that were likely acquired horizontally from nucleocytoplasmic large DNA virus relatives [NCLDV, (Maumus *et al.*, 2014); Table S16, and Figures 4 and S14–S22] and found that 87 integrations (NCLDVI) harbor 257 regions homologous to NCLDV protein-coding genes and 163 sRNA clusters. Colinearity and molecular dating analyses of NCLDVI (Figures S19 and S20) suggest four groups of regions that have been either amplified by recombination events or represent simultaneous integrations. The timing of these integrations (comprising both relatively young and older insertions/duplications) appears independent from the periods of LTR-RT activity. NCLDVI regions are the most variable annotated loci in terms of nucleotide diversity (Figure S18). Previous evidence suggested that NCLDVI represent non-functional, decaying remnants of ancestral infections that are transcriptionally inactivated by methylation (Maumus *et al.*, 2014). By screening available sRNA-seq libraries we could record repetitive, but specific sRNA clusters for these loci. Strikingly, we identified two NCLDV genes harboring sRNA loci that exhibit high transcriptional activity, coinciding with lower levels of DNA methylation as compared with other

NCLDVI (Figures S14 and S15). Consistent with the predicted potential to form hairpin structures, sRNA northern blots (Figure S22) of wild type and Dicer-like (DCL) deletion mutants (Khraiwesh *et al.*, 2010; Arif *et al.*, 2012) suggest that RNA transcribed from these loci might be processed by distinct DCL proteins to generate siRNAs. These siRNAs in turn might act to target viral mRNA during a potential NCLDV infection, or to guide DNA methylation to silence these regions (Kawashima and Berger, 2014). Regions harboring corresponding antisense sRNA loci are enriched for stop-codon-free (i.e. non-degrading) NCLDV genes and deviate from the remainder of NCLDVI in terms of cytosine versus histone modifications (Figures S15 and S16). Based on the similarity with intact LTR-RTs in terms of methylation and low GC (Figure S17), and the absence of H3K9me2, we hypothesize that (like intact TEs) these ancient, retained NCLDVI are euchromatic. We propose that they are demethylated during gametogenesis by DEMETER (which in *Arabidopsis* preferentially targets small, AT-rich, and nucleosome-depleted euchromatic TEs (Ibarra *et al.*, 2012)). Given the proposed time point of activation of these regions during gametangiogenesis, NCLDVI might provide a means to provide large numbers of siRNAs which, besides ensuring the transgenerational persistence of silencing, could also provide protection against cytoplasmically replicating viruses *via* RNAi and methylation of the viral genome. This would provide efficient protection for moss gametes which, due to their dependency on water, might be the most exposed to NCLDV infections. This hypothesis provides a plausible answer to the question why endogenous NCLDV relatives have only been found in embryophytes with motile sperm cells (Maumus *et al.*, 2014).

*Genetic variability.* Sequencing three different accessions we find 264 782 SNPs (1 per 1783 bp) for Reute (collected close to Freiburg, Germany), 2 497 294 (1 per 188 bp) for Villersexel (Haute-Saône, France) and 732 288 (1 per 644bp) for Kaskaskia (IL, USA) as compared with Gransden. There are 42 490 polymorphisms shared among all three accessions relative to Gransden, with other SNPs present in only one or two of the accessions (Figure S31). SNP densities of *Arabidopsis thaliana* ecotypes occur at one SNP per 149–285 bp (Cao *et al.*, 2011), similar to that in Villersexel, which is surprising given that the rate of neutral mutation fixation is lower in *P. patens* (Rensing *et al.*, 2007). However, Villersexel has an extraordinarily high divergence compared with other *P. patens* accessions (McDaniel *et al.*, 2010). Due to the fact that all accessions are inter-fertile, yet genetically divergent (Beike *et al.*, 2014), and exhibit phenotypic differences (File S2; Hiss *et al.*, 2017), we consider them potential ecotypes. For all accessions, most SNPs (>80%) are found in intergenic and adjacent (potential regulatory) regions of genes (Table S19). Less than 5%

of all SNPs are found in genic regions, of those 34–36% are silent (synonymous), 62–64% missense (non-synonymous) and 1.6% cause a nonsense mutation. Overall, Reute showed 72 regions of SNP accumulation, whereas Villersexel and Kaskaskia showed 30 and 32, respectively (Table S20–S22). The SNP accumulation regions in Reute are more gene-rich with 18 genes/region compared with 8 and 10 in Villersexel and Kaskaskia. One peak on chromosome 16 is found in all accessions and contains genes involved in sterol catabolism and chloroplast light sensing/movement (Figure S33). Sterols have been implicated in cell proliferation, in regulating membrane fluidity and permeability, and in modulating the activity of membrane-bound enzymes (Hartmann, 1998). The over-represented terms detected in the genes commonly harboring SNPs might be the signature of evolutionary modification of dehydration tolerance, for which membrane stability has been shown to be an important factor in mosses (Oliver *et al.*, 2004; Hu *et al.*, 2016).

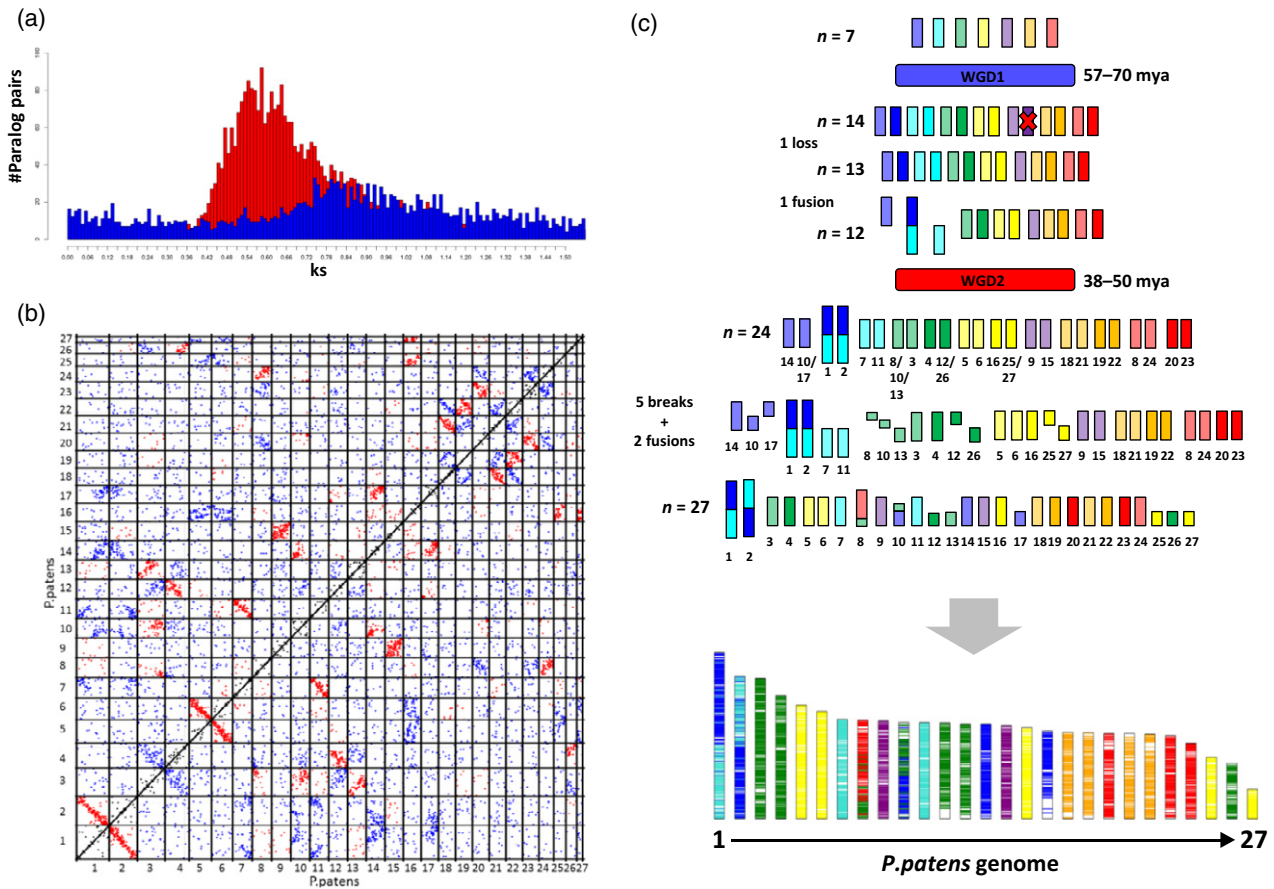
*Recombination might be needed for purging TEs.* Many genomes have higher densities of TEs in centromeres, sub-telomeres (Figure S34), and sex chromosomes, i.e. regions of low recombination (Dolgin and Charlesworth, 2008). One potential explanation for this biased distribution is that TEs insert with more or less equal frequencies across the genome, but are heterogeneously distributed because purifying selection is weaker in regions of low recombination. This hypothesis can be put to test using the *Physcomitrella* genome: the species is mostly selfing (it practises *de facto* asexual reproduction using sexual gametes; Perroud *et al.*, 2011), and thus the effective rate of recombination is low (since genetic variants are seldom mixed as heterozygotes), and purifying selection is correspondingly weak (Szovenyi *et al.*, 2013). If recombination (in outcrossed offspring) is indeed critical for making purifying selection effective at purging weakly deleterious TEs, we would predict that selection against TE disruption of gene expression may be playing an important role in the chromosomal distribution of TEs (Wright *et al.*, 2003). Hence, the unusual chromosomal structure might be a function of predominant inbreeding. We expect that the genomes of bryophytes that are outcrossers, like *Marchantia polymorpha*, *Ceratodon purpureus*, *Funaria hygrometrica* or *Sphagnum magellanicum*, might show a more biased distribution of TEs along their chromosomes.

### Genome evolution

*Two whole genome duplication events.* Based on synonymous substitution rates (Ks) of paralogs, at least one WGD event was evident in *P. patens* (Rensing *et al.*, 2007, 2008). However, gene family trees often show nested paralog pairs, and the ancestral moss karyotype is hypothesized to be seven (Rensing *et al.*, 2012), while the extant

chromosome number of *P. patens* is  $n = 27$  (Reski *et al.*, 1994), suggesting two ancestral WGD events (Rensing *et al.*, 2007, 2012). Using the novel pseudochromosome structure, Ks-based analyses support two WGDs dating back to 27–35 and 40–48 Ma (Figure 5), respectively (cf. supplementary material IV.). Given the detected synteny, the most parsimonious explanation for the extant chromosome number is the duplication of seven ancestral chromosomes in WGD1, followed by one chromosomal loss and one fusion event during the subsequent haploidization. In WGD2 the 12 chromosomes would have duplicated again, followed by five breaks and two fusions, leading to 27 modern chromosomes. The Ks values of the above-mentioned structure-based peaks (Figure 5) fall approximately between 0.5–0.65 (younger WGD2) and 0.75–0.9 (older WGD1). The structural and Ks information can be used to trace those genes that were present in the ancestral (pre-WGD) karyotype and have since been retained (Figure S37 and File S3). In total, 484 genes can be traced to the pre-WGD1 karyotype (denoted ancestor 7), and 3112 genes to the pre-WGD2 karyotype (ancestor 12). GO bias analysis of the ancestor 7 genes shows over-representation of many genes involved in regulation of transcription and metabolism (Figure S38). This accords with previous evidence that metabolic genes were preferentially retained after the *P. patens* WGD (Rensing *et al.*, 2007), and with the trend that genes involved in transcriptional regulation are preferentially retained after plant WGDs (De Bodt *et al.*, 2005).

*WGDs are common in mosses, but not in other bryophytes.* Detecting WGD events using paranome-based Ks distributions is notoriously difficult (Vekemans *et al.*, 2012; Vanneste *et al.*, 2014). Here we compared several methods for deconvolution of such distributions and found that a mixture model based on log-transformed values was able to detect four potential WGDs (Figure S39), including the two that we observed based on the pseudochromosomal structure (Figure 5). By excluding very young/low and very old/high Ks ranges, we restricted the data to the two structure-based events. Using low bandwidth (smoothing) we find that such methodology is able to detect relatively young WGDs with a clear signature (Figure S39(e, f)), whereas overlapping distributions (here the older WGD1) are hinted at via significant changes in the distribution curve at higher bandwidth settings (Figure S39(i, j); cf. Experimental Procedures and Appendix S1 Supplementary Material IV/2 for details). We applied this paranome-based WGD prediction to transcriptome data obtained from the onekp project ([www.onekp.com](http://www.onekp.com)) on 41 moss, 7 hornwort and 28 liverwort datasets and overlaid them with a molecular clock tree (Figures S40–S42) (Newton *et al.*, 2006). For 24 of the moss samples at least one WGD signature was supported. For four out of these 24



**Figure 5.** Evolutionary scenario leading to the modern *P. patens* genome.

(a) Ks distribution (y-axis) of paralogous pairs (x-axis) inherited from two (blue for older and red for more recent) WGD events.

(b) Dotplot representation of the paralogous pairs belonging to two WGD events.

(c) Karyotype evolution of the *P. patens* genome from an  $n = 7$  ancestor through two WGDs. The modern *P. patens* genome is illustrated as a mosaic of coloured chromosomal blocks highlighting chromosome ancestry.

mass datasets, mixture model components were merged into one WGD signature with the possibility of additional hidden WGD signatures. Among these species is *Physcomitrium* sp. which is a close relative of *P. patens*; shared WGD events are in accordance with previous studies (Beike *et al.*, 2014). The three *Sphagnum* species show overlap and significant gradient change support for a young WGD event and in *Sphagnum lescurii* also significant support for an older WGD event, supporting a recent report (Devos *et al.*, 2016). While only a chromosome-scale assembly would be able to detect WGD events with high confidence, we note that evidence of WGDs is not detected in any of the liverwort and hornwort datasets, while the majority of moss lineages appears to have been subject to ancient WGDs. In contrast with mosses (Rensing *et al.*, 2012; Szovenyi *et al.*, 2014), most liverworts and are known for low levels of neopolyploidy and endopolyploidy with rather constant chromosome numbers within each lineage (Bainard *et al.*, 2013). The three-fold fluctuations in genome size in nested hornwort lineages without a chromosomal

change (Bainard and Villarreal, 2013) is thus most likely due to variable TE content. The karyotype evolution of *P. patens* can thus be considered as typical for moss genomes, but probably different from the genomes of hornworts and liverworts. While we do not know why mosses might be more prone to fixation of genome duplications than other bryophytes, the associated paralog acquisition and retention might be a foundation for the relative species richness of mosses (Rensing, 2014; Rensing *et al.*, 2016; Van de Peer *et al.*, 2017).

*Ancient colinearity reveals conserved plant-specific functions.* Have gene orders been conserved since the last common ancestor of land plants (LAP)? Colinearity analyses with 30 other plant genomes (cf. Experimental Procedures and Appendix S1 Supplementary Material IV/3) revealed 180 colinear regions, harbouring around 1700 genes. *P. patens* chromosomes contain 0.5–10 of these genes per Mbp (Figure S43), most chromosomes hence containing a number of syntenic genes that follows

random expectation. Chromosomes 1, 8, 11, 14, 16 and 27, however, contain significantly more ancient colinear genes than expected ( $q < 0.05$ , Fisher's exact test; File S3). GO bias analyses revealed that chromosome 8 is enriched for genes encoding functions for plant cell and tissue growth and development (Figure S44). Surprisingly, several hundred genes are present in colinear regions that involve 5–21 other species. Moreover, 17 of these regions showed elevated levels of gene co-expression ( $P < 0.05$ , permutation statistics; File S3), indicating potential co-regulation of neighboring genes, thus corroborating the existence of conserved plant regulons (Van de Velde *et al.*, 2016) or genomic regions exposed similarly to the transcriptional machinery. GO bias analyses of these ancient syntenic genes demonstrate that they are involved in land plant-specific cell growth and tissue organization (Figure S45), akin to chromosome 8. Apparently, genes encoded in the LAP genome that enabled the distinct cell and tissue organization of land plants have been retained as colinear blocks throughout land plant evolution. In total, 10 genes on chromosome 7 can be traced back to chromosome 4 of ancestor 12 (pre-WGD2), and to chromosome 2 of ancestor 7 (pre-WGD1). GO bias of chromosome 7 (Figure S46) further supports the notion that genes enabling plant-specific development have been conserved since the LAP.

## CONCLUSIONS

Our analyses show that the genome of the model moss is organized differently from seed plant genomes. In particular, no central TE-rich and distal gene-rich chromosomal areas are detected, and centromeres are potentially marked by a subclass of Copia elements. There is evidence for activation of TE and viral elements during the life cycle of *P. patens* that might be related to its haploid-dominant life style and motile gametes. Surprisingly, syntenic blocks harboring genes involved in plant-specific cell organization were conserved for *ca.* 500 Ma of land plant evolution. Chromosome-scale assemblies of other non-seed plants will be needed in order to understand how plant genomes from diverse lineages evolve, and to determine whether the genomes of haploid-dominant plants are generally different from those of seed plants.

## EXPERIMENTAL PROCEDURES

### Sequencing and assembly

We sequenced *Physcomitrella patens* Gransden 2004 using a whole genome shotgun sequencing strategy. Most sequencing reads were collected with standard Sanger sequencing protocols on ABI 3730XL capillary sequencing machines at the Department of Energy Joint Genome Institute in Walnut Creek, California, USA ([http://www.jgi.doe.gov/sequencing/protocols/protos\\_production.html](http://www.jgi.doe.gov/sequencing/protocols/protos_production.html)) as previously reported (Rensing *et al.*, 2008). BAC end sequences were collected using standard protocols at the

HudsonAlpha Institute in Huntsville, Alabama, USA. The sequencing (see Table S1) consisted of two libraries of 3 kbp (4.01x), 3 libraries of 8 kbp (4.58x), four fosmid libraries (0.43x), and two BAC libraries (0.22x) on the Sanger platform for a total of 9.25x Sanger based coverage. In total, 7 572 652 sequence reads (9.25x assembled sequence coverage, see Table S1 for library size summary) were assembled using our modified version of Arachne v.20071016 (Jaffe *et al.*, 2003) with parameters correct1\_passes=0 maxcliq1 = 140 BINGE\_AND\_PURGE=True max\_bad\_look=2000 (see Table S2 for overall scaffold and contigs statistics). This produced a raw assembly consisting of 1469 scaffolds (4485 contigs) totaling 475.8 Mb of sequence, with a scaffold N50 of 2.8 Mb, 271 scaffolds larger than 100 kbp (464.3 Mb). Scaffolds were screened against bacterial proteins, organellar sequences and the GenBank 'nr' database, and removed if found to be a contaminant. Additional scaffolds were removed if they were: (i) scaffolds smaller than 50 kbp consisting of >95% 24-mers that occurred four other times in scaffolds larger than 50 kbp; (ii) contained only unanchored RNA sequences; (iii) were less than 1 kbp in length; or (iv) contaminated. Post-screening, we integrated the resulting sequence with the genetic map reported here (3712 markers), and BAC/fosmid paired end link support. An additional map (9080 markers) was developed for chromosome 16 that resolved ordering problems present in the original map, and was used for the integration of chromosome 16. The integrated assembly was screened for contamination to produce a pseudomolecule reference covering 27 nuclear chromosomes. The pseudomolecules include 462.3 Mb of base pairs, an additional 351 unplaced scaffolds consist of 4.9 Mb of unanchored sequence. The total release includes 467.1 Mb of sequence assembled into 3077 contigs with a contig N50 of 464.9 kbp and an N content of 1.5%. Chromosome numbers were assigned according to the physical length of each linkage group (1 = largest and 27 = smallest).

### Genetic mapping

In order to assign the sequenced scaffolds representing the release version V1.2 *Physcomitrella* genome sequence to chromosomes, we used a genetic mapping approach based on high-density SNP markers. SNP loci between the Gransden 2004 ('Gd') and genetically divergent Villersexel K3 ('Vx') genotype were identified by Illumina sequencing (100 bp end reads; Illumina GAI) of the Vx accession. The sequence data have been deposited in the NCBI Sequence Read Archive as accessions SRX037761 (two Illumina Genome Analyzer II runs: 176.1 M spots, 26.8 G bases, 93.4 Gb downloads) and SRX030894 (three Illumina Genome Analyzer II runs: 277.9 M spots, 42.2 G bases, 56 Gb downloads). SNPs for linkage mapping were selected for the construction of an Illumina Infinium bead array for the GoldenGate genotyping platform, based on their distribution across the 1921 scaffolds representing the V1.2 genome sequence assembly, with an average physical distance between SNP loci of *ca.* 110 kbp. Segregants of a mapping population [539 progeny from Gd×Vx crosses: (Kamisugi *et al.*, 2008)] were genotyped at 5542 loci to construct a linkage map using JoinMap 4.0 (Van Ooijen JW, 2006, Kyazma B.V., Wageningen, The Netherlands), with a minimum independence LOD threshold of 22, a recombination threshold of 0.4, a ripple value of 1, a jump threshold of 5 and Haldane's mapping function. Of the 5542 SNPs, 4220 loci were represented in the final map. The map contained 27 linkage groups, covering 5432.9 cM. Map lengths were calculated using two methods: one in which  $L$  (total map length) =  $\Sigma$  [(linkage group length) + 2 (linkage group length/no. markers)] (Fishman *et al.*, 2001) and one in which  $L$  =  $\Sigma$  [(linkage group length (no. markers + 1)/(no. markers - 1)] (Chakravarti *et al.*, 1991). The map corresponded to 467 985 895 bp distributed

across the previously predicted 27 *P. patens* chromosome (Table S3). Chromosome numbers were assigned according to the overall physical length of each linkage group (1 = largest and 27 = smallest).

### Pseudochromosome construction

The combination of the existing genetic map (4220 markers), and BAC/fosmid paired end link support was used to identify 12 misjoins in the overall assembly. Misjoins were identified as linkage group discontinuity coincident with an area of low BAC/fosmid coverage. In total, 12 breaks were executed, and 295 scaffolds were oriented, ordered and joined using 268 joins to form the final assembly containing 27 pseudomolecule chromosomes, capturing 462.3 Mb (98.97%) of the assembled sequence. Each chromosome join is padded with 10 000 Ns. The final assembly contains 378 scaffolds (3077 contigs) that cover 467.1 Mb of the genome with a contig L50 of 464.9 kbp and a scaffold L50 of 17.4 Mb.

Completeness of the euchromatic portion of the genome assembly was assessed using 35 940 full-length cDNAs. The aim of this analysis was to obtain a measure of completeness of the assembly, rather than a comprehensive examination of gene space. The cDNAs were aligned to the assembly using BLAT (Kent, 2002); Parameters: `-t=dna -q=rna -extendThroughN`, and alignments  $\geq 90\%$  bp identity and  $\geq 85\%$  coverage were retained. The screened alignments indicate that 34 984 (97.3%) of the FLcDNAs aligned to the assembly. The ESTs that failed to align were checked against the NCBI nucleotide repository (nr), and a large fraction was found to be prokaryotic in origin. Significant telomeric sequence was identified using the TTTAGGG repeat, and care was taken to make sure that it was properly oriented in the production assembly. Plots of the marker placements for the 27 chromosomes are shown in File S2. For contamination screening, further assessment of assembly accuracy and organellar genomes please refer to Appendix 1, Supplementary Material, Section I.

### Mapping of the v1.6 genome annotation

Gene models of the v1.6 annotation (Zimmer *et al.*, 2013) were mapped against the V3 assembly using GenomeThreader (Gremme *et al.*, 2005) and resulting spliced alignments were filtered and classified for consistency with the original gene structures. 93.9% of the 38 357 v1.6 transcripts could be mapped with unaltered gene structure. This comprised 29 371 loci (91.4% of the v1.6 loci). The majority of the unmappable v1.6 models represented previously unidentified bacterial or human contaminations in the V1 assembly (492 loci). Nevertheless, 49 loci with expression evidences remained unmappable in the current assembly. The mapped annotation is made available via the *cosmoss.org* genome browser and under the download section.

### Generation of the v3.1 genome annotation

All available RNA-seq libraries (File S3 and Table S10) were mapped to the V3 assembly using TopHat (Trapnell *et al.*, 2009). Based on a manually curated set of *cosmoss.org* reference genes (Zimmer *et al.*, 2013), libraries and resulting splice junctions were filtered to enrich evidence from mature mRNAs. Sanger and 454 EST evidence used in the generation of the v1.6 annotation was mapped using GenomeThreader. The resulting splice junctions and exonic features were used as extrinsic evidences to train several gene finders, which were evaluated using the *cosmoss.org* reference gene set. Based on this evaluation, five predictive models derived with EuGene (Foissac *et al.*, 2003) resulting from

different parameter combinations, including the original model used to predict v1.6, were retained for genome-wide predictions. RNA-seq libraries were assembled into virtual transcripts using Trinity (Grabherr *et al.*, 2011). The resulting 1 702 106 assembled transcripts with a mean length of 1219 bp were polyA trimmed using seqclean (part of the PASA software), of which 96% could be mapped against the V3 genome using GenomeThreader. Together with the 454 and Sanger ESTs 2 755 148 transcript sequences were used as partial cDNA evidence in the PASA software to derive 266 051 assemblies falling in 68 382 subclusters. For these, transdecoder was trained and employed to call open reading frames based on PFAM (Finn *et al.*, 2016) domain evidence. Gene models from transdecoder, EuGene and the JGI V3.0 predictions were combined and evaluated using the eval software (Keibler and Brent, 2003) on the reference gene set. Based on the resulting gene and exon sensitivity and specificity scores a rank-based weight was inferred (Table S9), which was used to infer combined CDS models using EvidenceModeler, resulting in a gene sensitivity/specificity of 0.76/0.76 and an exon sensitivity/specificity of 0.93/0.98. For these combined CDS features, UTR regions were annotated using PASA in six iterations. All transcript evidence and alternative gene models are available via tracks in the *cosmoss.org* genome browser. From the resulting set of gene models, protein-coding gene loci and representative isoforms were inferred using a custom R script implementing a multiple feature weighting scheme that employed information about CDS orientation, proteomic, sequence similarity and expression evidence support, feature overlaps, contained repeats, UTR-introns and UTR lengths of the gene models in a Machine Learning-guided approach. This approach was optimized and trained based on a manually curated training set in order to ideally select the functional, evolutionary conserved 'major' isoform for each protein-coding gene locus. The v3.1 annotation comprises only the 'major' (indicated by the isoform index 1 in the CGI), while v3.3 also includes other splice variants with isoform indices >1.

### Availability of gene models and additional data

The analyses in this publication rely on the structural annotation v3.1. Subsequently, this release was merged with the phytozome-generated release v3.2, leading to the current release v3.3 which is available from <http://cosmoss.org> and <https://phytozome.jgi.doe.gov/>. Both v3.1 and v3.3 are available in CoGe (<https://genomevolution.org/coge/GenomeView.pl?gid=33928>), and v1.6 and v1.2 can be loaded as tracks for backward compatibility. Available experiment tracks can be downloaded and are listed in Table S12. Organellar genomes are also available at CoGe under the id 35274 (chloroplast) and 35275 (mitochondrion). For gene annotation version 3.2/3.3, locus naming, non-protein coding genes and functional annotation refer to Appendix S1, Supplementary Material, Section II. Annotations v3.1 and v3.3 are available in File S1, including a lookup of gene names for versions 3.3, 3.1, 1.6, 1.2 and 1.1. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession ABEU000000000. The version described in this paper is version ABEU02000000.

**Cytological analyses.** The chromosome arrangement during mitotic metaphase as well as the punctate labelling at pericentromeric regions after immunolabelling with a pericentromere-specific antibody against H3S28ph (Gernand *et al.*, 2003) indicate a monocentric chromosome structure in *P. patens* (Figure S5). Furthermore, many plant genomes, as for example *A. thaliana* (Fuchs *et al.*, 2006), are organized in well defined heterochromatic pericentromeric regions, decorated with typical heterochromatic marks

(H3K9me1, H3K27me1) and gene-rich regions presenting the typical euchromatic marks (H3K4me2). By contrast, immunostaining experiments with antibodies against these marks label the entire chromatin of flow-sorted interphase *P. patens* nuclei homogeneously (Figure 3(b)). Obviously, *P. patens* nuclei are thus characterized by a uniform distribution of euchromatin and heterochromatin.

### Transposon and repeat detection and annotation

TRharvest (Ellinghaus *et al.*, 2008) which scans the genome for LTR-RT specific structural hallmarks (like long terminal repeats, tRNA cognate primer binding sites and target site duplications) was used to identify full length LTR-RTs. The input sequences comprised the 27 pseudochromosomes plus all genomic scaffolds with a length of  $\geq 10$  kbp together with a non-redundant set of 183 *P. patens* tRNAs, identified beforehand via tRNA scan (Lowe and Eddy, 1997). The used parameter settings of LTRharvest were: 'overlaps best -seed 30 -minlenltr 100 -maxlenltr 2000 -mindistltr 3000 -maxdistltr 25000 -similar 85 -mintsd 4 -maxtsd 20 -motif tgca -motifmis 1 -vic 60 -xdrop 5 -mat 2 -mis -2 -ins -3 -del -3'. All of the resulting 9290 candidate sequences were annotated for PfamA domains with hmmer3 (<http://hmmer.org/>) and stringently filtered for false positives by several criteria, the main ones being the presence of at least one typical retrotransposon domain (e.g. RT, RH, INT, GAG) and a tandem repeat content below 25%. The filtering steps led to a final set of 2785 high confident full-length LTR RTs. Transposons were annotated by RepeatMasker (Smit *et al.*, 1996) against a custom-built repeat library (Spannagl *et al.*, 2016) which included *P. patens* specific full length LTR-retrotransposons.

Repetitive elements have also been annotated *de novo* with the REPET package (v2.2). The TEde novo pipeline from REPET (Flutre *et al.*, 2011) was launched on the contigs of size  $>350$  kbp in the v3 assembly (representing approximately 310 Mb, gaps excluded) to build a library of consensus sequences representative of repetitive elements. Consensus sequences were built if at least five similar hits were detected in the sub-genome. Each consensus was classified with PASTEC (Hoede *et al.*, 2014) followed by semi-manual curation. The library was used for a first genome annotation with the TEannot pipeline (Quesneville *et al.*, 2005) from REPET to select the consensus sequences that are present for at least one full length copy ( $n = 349$ ). Each selected consensus was then used to perform final genome annotation with TEannot with default settings (BLASTER sensitivity set to 2). The REPET annotations absent from the mipsREdat annotation were added to the latter to build the final repeat annotation. Tandem repeats Finder (Benson, 1999) was launched with the following suite of parameters: 2 7 7 80 10 50 2000. The putative centromeric repeat previously identified through tandem repeats analysis (Melters *et al.*, 2013) was compared with the whole V3 assembly using RepeatMasker (Smit *et al.*, 1996) with default settings (filter divergence  $<20\%$ ). Besides Copy and Gypsy-type elements (see main text), other types of TEs, including LINEs and Class II (DNA transposon) elements, appear at very low frequency (0.1% each). Simple sequence repeats represent only 2% of the assembly. For TE phylogenetic, age and expression analyses as well as NCLDV analyses refer to Appendix S1, Supplementary Material, Section III.

### ChIP-seq data

Published ChIP-seq data (Widiez *et al.*, 2014) for *P. patens* were re-analysed by mapping read libraries against the *P. patens* V3.0 genome sequence. Briefly, the FASTA and QUAL files were converted into FASTQ data files, which were aligned against the *P. patens* v3.0 genome using BWA v0.5.9 (Li and Durbin, 2010), employing a seed length of 25, allowing a maximum of two

mismatches on the seed and a total maximum of 10 mismatches between the reference and the reads. In order to avoid redundancy problems, all reads that were mapped to more than one genomic locus were omitted as already applied elsewhere (Zemach *et al.*, 2010; Stroud *et al.*, 2012). SAM files were converted into BED files using an in-house Python script.

### Identification of histone-modified enriched regions

For the identification of the histone-modified enriched regions (peaks) the software MACS2 v2.0.10 (Zhang *et al.*, 2008; Feng *et al.*, 2012) with parameters tuned for histone modification data was used. The parameters used were 'no model', shift size set as 'sonication fragment size', 'no lambda', 'broad', bandwidth 300 following the developer's instructions, fold change between 5 and 50 and q-value 0.01. As control for the peak identification the combination of Input-DNA and Mock-IP of the corresponding tissues was used as in Widiez *et al.* (2014). The number of identified peaks per tissue and histone mark is shown in Table S17.

### Extension of unannotated genomic regions

For several gene models in the *P. patens* v3.1 genome annotation the prediction of UTR regions (either 5' or 3') failed. In total there are 9769 genes lacking the 5'-UTR and 11 385 genes lacking the 3'-UTR. Additionally, gene promoters are also unannotated. Using an approach already used in (Widiez *et al.*, 2014), UTRs and promoters were assigned to gene models. In brief, a Python script was implemented that takes as input any valid GFF3 file and: (i) creates UTR regions of 300 bp for genes lacking either one or both of them; and (ii) creates potential promoter regions of 1500 bp upstream and downstream of each gene in the file. In the case that the space between the gene and the next element is not wide enough for the extension of the gene model by 300 bp, the new UTR region is shrunk to the available space. In the case that two consecutive genes have to be extended and the space between them is less than  $2 \times 300$  bp the new UTRs are assigned half the space between the two genes. For the assignment of promoters the same rules apply. In no case is an element created that overlaps with existing elements of the annotation file used as input.

### Filtering for expressed genes

Based on all the available JGI gene atlas (<http://jgi.doe.gov/our-science/science-programs/plant-genomics/plant-flagship-genome-s/>) RNA-seq data downloaded from Phytozome (File S3), we filtered for genes that had a certain minimal RPKM value in at least one condition. At RPKM 2, 20 274 genes are expressed, at RPKM 4 18 281 genes. The RPKM cutoff of four was based on quantitative real-time PCR (qRT-PCR) results of a recent microarray transcriptome atlas study (Ortiz-Ramirez *et al.*, 2015), in which genes with this expression level were reliably detected by qPCR.

### BS-seq data: plant material and culture conditions

*Physcomitrella patens* accession Gransden was grown in 9-cm Petri dishes on 0.9% agar solidified minimal (Knop's) medium. Cultures were grown under the following experimental conditions: 16 h/8 h light/dark cycle,  $70 \mu\text{mol sec}^{-1} \text{m}^{-2}$ , for 6 weeks at  $22^\circ\text{C}/19^\circ\text{C}$  day/night temperature following 8 h/16 h light/dark cycle,  $20 \mu\text{mol sec}^{-1} \text{m}^{-2}$ , for 7 weeks at  $16^\circ\text{C}/16^\circ\text{C}$  day/night temperature. Adult gametophores were harvested after 13 weeks and DNA was isolated according to Dellaporta *et al.* (1983) with minor modifications (Hiss *et al.*, 2017).

### Bisulfite conversion, library preparation and sequencing

Bisulfite conversion and library preparation was conducted by BGI-Shenzhen, Shenzhen, China according to the following procedure: DNA was fragmented to 100–300 bp by sonication, followed by blunt end DNA repair adding 3'-end dA overhang and adapter ligation. The ZYMO EZ DNA Methylation-Gold kit was used for bisulfite conversion and after desalting and size selection a PCR amplification step was conducted. After an additional size selection step the qualified library was sequenced using an Illumina GAI instrument according to manufacturer instructions resulting in 66 108 645 paired end reads of 90 bp length.

### Processing of BS-seq reads

Trimmomatic v0.32 (Bolger *et al.*, 2014) was used to clean adapter sequences, to trim and to quality-filter the reads using the following options: ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 SLIDINGWINDOW:4:5 TRAILING:3 MINLEN:35 resulting in cleaned paired-end and orphan single-end reads. Further, the paired-end and single-end reads were mapped with Bismark v0.14 (Krueger and Andrews, 2011) against *P. patens* chloroplast (NC\_005087.1) and mitochondrion (NC\_007945.1) sequences using the *-non\_directional* option due to the nature of the library. After mapping the remaining single-end and paired-end reads with Bismark v0.14 separately against the genome of *P. patens* both SAM alignment files were sorted and merged with samtools v0.1.19 (Li *et al.*, 2009) and deduplicated with the *deduplicate\_bismark* program of Bismark v0.14. To call methylation levels for the different cytosine contexts (CG, CHG, CHH), deduplicated SAM files and the R package *methylkit* (Akalin *et al.*, 2012) were used, only considering sites with a coverage of at least nine reads and a minimal mapping quality of 20.

### Gene- and TE-body methylation

Gene- and TE-body methylation levels were calculated for individual cytosine contexts (CG, CHG, CHH). For each gene and TE, all annotated feature regions (promoter, 5'-UTR, CDS, intron, 3'-UTR, TE-fragment) were combined and divided into 10 quartiles. For each quartile the mean methylation level (CG, CHG, CHH) was calculated and the average, 5% and 95% distribution per quartile and feature type were plotted. For the TE-body methylation plots TEs were further subdivided into TE-groups. For gene body methylation (GBM) analysis positions were filtered according to  $\geq 90\%$  of the reads showing methylation. Distribution of affected genes over the three different contexts was analysed with Venny (Figure S29; <http://bioinfogp.cnb.csic.es/tools/venny/>) and visualized via a stacked column diagram (Figure S30). Genes were grouped by RPKM value ( $0 > 0 < 2; \geq 2$ ) and compared with regard to GC and methylation content (Table S18).

### Read mapping and variant calling

Genomic DNA sequencing data for *P. patens* accessions Reute (SRP068341), Villersexel (SRX030894) and Kaskaskia (SRP091316) are available from the NCBI Sequence Read Archive (SRA). The libraries were trimmed for adapters and quality filtered using trimmomatic v32 (Bolger *et al.*, 2014) applying the following parameters: *-phred33 ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:8:5 SLIDINGWINDOW:4:15 TRAILING:15 MINLEN:35*. After trimming, the single-end and paired-end reads were initially mapped to the chloroplast genome (NC\_005087.1), the mitochondrial genome (NC\_007945.1) and ribosomal DNAs (HM751653.1, X80986.1, X98013.1) using GSNAP v2014-10-22 (Wu *et al.*, 2016) with default parameters. The remaining unmapped single-end and paired-end

reads were used for reference mapping using GSNAP with default parameters and both resulting SAM alignment files were sorted and merged with samtools v0.1.19 (Li *et al.*, 2009). Duplicated reads were further removed with *rmdup* from samtools to account for potential PCR artifacts. GATK tools v3.3.0 (McKenna *et al.*, 2010) were used for SNP calling as recommended by the Broad institute for species without a reference SNP database including the 'ploidy 1' option for the first and second haplotype calling step.

### SNP validation

Called SNPs of the accession Villersexel were validated by comparing them to the Illumina Infinium bead array dataset (File S3) used for map construction (see Map construction method section). The 4650 bead array probes were mapped to the genome using GSNAP (Wu *et al.*, 2016) and SNPs were called using mpileup and bcftools. In total, 4628 SNPs could be unequivocally mapped, out of those 4466 (96%) were also called as SNPs in the gDNA-seq based Villersexel GSNAP/GATK dataset. Thus, the vast majority of SNPs called based on deep sequence data could be independently confirmed (File S3).

### SNP divergence estimates

To obtain window-wise (100 kbp non-overlapping windows) nucleotide diversity  $\pi$  and Tajima's  $D$  values, a 'pseudogenome' was constructed for each accession using a custom python script. In brief, based on the VCF file output generated by GATK all given variants were reduced to SNPs and InDels and for each accession (Kaskaskia, Reute and Villersexel) the corresponding reference sequence was substituted with the ALT allele at the given positions. These 'pseudogenome' FASTA files were additionally masked for all sites which had a read coverage  $< 5$  which might lead to erroneous SNP calling. The masked 'pseudogenome' FASTA files were further converted into PHYLIP format and used as input for Variscan v2.0 (Hutter *et al.*, 2006), settings 'RunMode = 12', 'Sliding Window = 1; WidthSW = 100 000; JumpSW = 100 000; WindowType = 0' and excluding alignment gaps via 'CompleteDeletion = 1' (Figure S32).

### SNP accumulation detection

Window-wise (50 kbp with 10 kbp overlap) SNP numbers were extracted from the 'pseudogenome' FASTA files by a custom R script. The R functions *fisher.test* and *p.adjust* (*method = 'p'*) were used to select fragments that show a significantly (adjusted  $P$ -value  $< 0.01$ ) higher SNP number than the chromosome average. A region of accumulated SNPs (hotspot) was called if at least five adjacent fragments showed a significantly higher SNP number (Tables S20–S22 and Figure S33).

### Structure-based ancestral genome reconstruction and associated karyotype evolutionary model

The *P. patens* genome was self-aligned to identify duplicated gene pairs following the methodology previously described (Salse *et al.*, 2009). Briefly, gene pairs are identified based on blastp alignment using CIP (cumulative identity percentage) and CALP (cumulative alignment length percentage) filtering parameters with respectively 50% and 50%.  $K_s$  (rate of synonymous substitutions) distribution of the identified pairs unveiled two peaks illuminating two WGDs, one older and one more recent, included between  $K_s$  0.75–0.9 (WGD1) and 0.5–0.65 (WGD2).

We performed a classical dating procedure of the two WGD events based on the observed sequence divergence, taking into account the  $K_s$  ranges between 0.75–0.9 and 0.5–0.65 and a mean

substitution rate ( $r$ ) of  $9.4 \times 10^{-9}$  substitutions per synonymous site per year (Rensing *et al.*, 2007). The time ( $T$ ) since gene insertion is thus estimated using the formula  $T = Ks/2r$ .

Mapping of the identified gene pairs on the *P. patens* chromosomes defines seven independent (non-overlapping) groups (or CARs for Contiguous Ancestral Regions) of four duplicated regions (representing two rounds of WGDs; Figure S37). Based on the seven CARs identified, we determined the most likely evolutionary scenario based on the assumption that the proposed evolutionary history involves the smallest number of shuffling operations (including inversions, deletions, fusions, fissions, translocations) that could account for the transition from the reconstructed ancestral genome to modern karyotype (Salse, 2012). The ancestor 7 and 12 genes were mapped to the extant chromosomes and visualized as circular plots (Figure S37). These two ancestors (7 and 12) correspond respectively to the pre-WGD1 ancestor (quadruplicated by WGD1 and WGD2 in the modern *P. patens* genome), and the pre-WGD2 ancestor that is the result of the duplication of ancestor 7 (leading to ancestor 14) after one fusion and one chromosome loss (duplicated by WGD2 in the modern *P. patens* genome).

### Paranome-based WGD prediction

For species samples and Ks distribution calculation refer to Appendix 1, Supplementary Material, Section IV. We employed mixture modeling to find WGD signatures using the *mclust* v5.1 R package to fit a mixture model of Gaussian distributions to the raw Ks and log-transformed Ks distributions. All Ks values  $\leq 0.1$  were excluded for analysis to avoid the incorporation of allelic and/or splice variants and to prevent the fitting of a component to infinity (Schlueter *et al.*, 2004; Vanneste *et al.*, 2015), while Ks values  $> 5.0$  were removed because of Ks saturation. Further, only WGD signatures were evaluated between the Ks range of 0.235 (12.5 Mya) to account for recently duplicated gene pairs to Ks of 2.0 to account for misleading mixture modeling above this upper limit (Vanneste *et al.*, 2014, 2015). Because model selection criteria used to identify the optimal number of components in the mixture model are prone to overfitting (Vekemans *et al.*, 2012; Olsen *et al.*, 2016) we also used SiZer and SiCon (Chaudhuri and Marron, 1999; Barker *et al.*, 2008) as implemented in the *feature* v1.2.13 R package to distinguish components corresponding to WGD features at a bandwidth of 0.0188, 0.047, 0.094 and 0.188 (corresponding 1, 2.5, 5 and 10 Mya) and a significance level of 0.05.

Deconvolution of the overlapping distributions that can be derived from paranome-based Ks values without structural information shows that using mixture model estimation based on log-transformed Ks values mimics structure-based WGD predictions better than using raw Ks values, resulting however in the prediction of four WGD signatures (pbSIG1: 0.15–0.32; pbSIG2: 0.48–0.60; pbSIG3: 0.7–1.12; pbSIG4: 1.66–3.45; Figure S39(a, b)). As WGD signature prediction based on paranome-based Ks values can be misleading and is prone to overprediction (Schlueter *et al.*, 2004; Vekemans *et al.*, 2012; Vanneste *et al.*, 2015; Olsen *et al.*, 2016) we only considered Ks distribution peaks in a range of 0.235–2.0 as possible WGD signatures, thus excluding young paralogs potentially derived from tandem or segmental duplication and those for which accurate dating cannot be achieved due to high age. The paranome-based WGD signatures pbSIG2 (25–32 Ma) overlaps with the younger WGD2, and pbSIG3 (37–60 Ma) overlaps with the older WGD1. Further testing for significant gradient changes in the Ks distribution applying different bandwidths showed that only pbSIG2 is detected as a significant WGD signature (significance level 0.05; Figure S39(h)), whereas pbSIG3 overlaps with a significant change of the Ks distribution

curve at a bandwidth of 0.047 but shows no significant gradient change. These results show that even if one paranome-based WGD signature can be found which perfectly overlaps with a structure-based WGD signature (WGD1 and pbSIG3) it is still hard to significantly distinguish it from the younger WGD signatures (WGD2 and pbSIG2) which tend to collapse using higher bandwidths (Figure S39(i, j)). Showing that log-transformed Ks value mixture modeling at least can predict young WGD signatures and can pinpoint older WGD signatures, we applied paranome-based WGD prediction to transcriptome data obtained from the onekp project ([www.onekp.com](http://www.onekp.com)) on 41 moss samples, 7 hornwort samples and 28 liverwort samples and overlaid them with an existing time tree (Figures S40–S42). After evaluating the overlap of significant gradient changes on mixture model components, for 24 out of 41 moss samples at least one WGD signature was supported. For four out of these 24 moss samples mixture model components were merged into one WGD signature with the possibility of additional hidden WGD signatures. Among these samples is *Physcomitrium* sp. which belongs like *P. patens* to the Funariaceae with WGD signatures 3 (0.43–0.66) and 4 (0.80–1.07), overlapping with pbSIG2 and pbSIG3 from *P. patens* and hinting at WGD events in *Physcomitrium* 23–35 Ma and 43–57 Ma ago, respectively. For all liverwort samples and almost all hornwort samples no single predicted WGD signature was supported by three different bandwidth kernel densities. For one hornwort, namely *Megaceros flagellaris*, one WGD signature was supported by a significant gradient change (significance level 0.05), which disappeared using a more stringent significance level of 0.01 and represents more likely a mixture model artifact than a true WGD signature.

### Colinearity analyses

For set of species refer to Appendix S1, Supplementary Material, Section IV. Initially, all chromosomes from all species were compared against each other and significant colinear regions are identified. To detect colinearity within and between species i-ADHoRe 3.0 was used (Proost *et al.*, 2012) with the following settings: alignment\_method gg2, gap\_size 30, cluster\_gap 35, tandem gap 30, q\_value 0.85, prob\_cutoff 0.01, multiple\_hypothesis\_correction FDR, anchor\_points 5 and level\_2\_only false. *P. patens* v3.1 genes were assigned to PLAZA 3.0 gene families based on the family information for the best BLASTP match (27 895 genes were assigned to 10 153 gene families). The profile-based search approach of i-ADHoRe combines the gene content information of multiple homologous genomic regions and therefore allows detection of highly degenerated though significant genomic homology (Simillion *et al.*, 2008). In total, 180 regions were found showing significant colinearity with genomes from flowering plants (colinearity with green algal genomes was not found), comprising 1717 genes involved in syntenic regions, representing 660 unique conserved moss genes. Whereas 94/180 of the ultra-conserved colinear (UCC) regions showed genomic homology with one other species, 45 UCC regions showed colinearity with five or more other plant genomes. One UCC region (multiplicon 1440, File S3) grouped 27 genomic segments from 21 species showing colinearity, while 70% of the UCC regions contained five or more conserved moss genes. Starting from the V1 moss genome assembly, only 11/180 UCC regions were recovered, demonstrating that the superior assembly V3 significantly improves the detection of ancient genomic homology. Mapping of the 660 UCC genes reveals their chromosomal location (Figure S43). Co-expression analysis of neighboring UCC genes was performed using the Pearson Correlation Coefficient (PCC) on the JGI gene atlas data (File S3) and permutation



statistics were used to identify UCC regions showing significant levels of gene co-expression (i.e. based on 1000 iterations, in how many cases was the expected median PCC for  $n$  randomly selected genes larger than the observed median PCC for  $n$  UCC genes).

We tested whether the actual number of genes detected to be present in ancient colinear blocks deviated from the expected number, if all genes were randomly distributed on the chromosomes. Chromosomes significantly deviating (Fisher's exact test and false discovery rate correction) are mentioned in the main text and are shown in File S3 and Figure S43. Genes detected to be derived from ancestor 7 and ancestor 12 karyotypes can be traced to extant chromosomes (File S3).

### GO bias analyses and GO word cloud presentation

Analyses were conducted as described previously (Widiez *et al.*, 2014), using the GOstats R package and Fisher's exact test with *fdr* correction. Visualization of the GO terms was implemented using word clouds via the <http://www.wordle.net> application. The weight of the given terms was defined as the  $-\log_{10}(q\text{-values})$  and the colour scheme used for the visualization was red for under-represented GO terms and green for those over-represented. Terms with stronger representation, i.e. weight  $>4$ , were represented with darker colours.

### Circos plots

For the integrative visualization of the individual genomic features a karyotype ideogram was created and tracks were plotted with CIRCOS v0.67-6 (Krzywinski *et al.*, 2009). For each feature track it is highlighted in the corresponding figure legend whether feature raw counts/values were used for visualization or if chromosomes were split into smaller windows (specifying the window size in kbp and window overlaps/jumps in kbp) using the counts/values window average for visualization. If indicated, feature counts/values window averages (cvwa) were normalized by scaling between a range of 0 and 1 per chromosome using the following equation:

$$\text{normalized window average}_{\text{chr}}(\text{cvwa}_{i_{\text{chr}}}) = \frac{\text{cvwa}_{i_{\text{chr}}} - \text{cvwa}_{\text{chr}_{\text{min}}}}{\text{cvwa}_{\text{chr}_{\text{max}}} - \text{cvwa}_{\text{chr}_{\text{min}}}}$$

For normalized comparison of embryophyte chromosome structure refer to Appendix S1, Supplementary Material, Section III; for phylostratigraphy analyses to Appendix S1, Supplementary Material, Section IV.

### Availability of data and material

The data reported in this paper are tabulated in Experimental Procedures and Supporting Information, are archived at the NCBI SRA and have been made available using the comparative genomics (CoGe) environment of CyVerse ([cyverse.org](http://cyverse.org)) via <https://genomevolution.org/coge/GenomeView.pl?gid=33928>. Novel data presented with this study comprise Villersexel and Kaskaskia genomic DNA (SRX037761, SRX030894, SRP091316), genomic BAC end data (KS521087–KS697761), RNA-seq data (Table S6 and File S3 – available from [phytozome.org](http://phytozome.org)), CAP-capture and BS-seq data (Table S10), and Goldengate SNP bead array data (File S3). See also section Availability of gene models and additional data.

Requests for materials should be addressed to [stefan.rensing@biologie.uni-marburg.de](mailto:stefan.rensing@biologie.uni-marburg.de).

### AUTHORS' CONTRIBUTIONS

AS, ADZ, ACC, AW, CVC, DL, FH, FMu, FMa, GB, HG, JP, JSa, JJ, GAT, JM, JF, JMC, KV, Kku, LEG, LS, MH, MT,

MP, MvB, NvG, OS, PR, RM, RH, SNWH, SS, SAR, SFM, TW, WM, YK, YZ analysed data or performed experiments. AL, CR, DWS, ELD, FT, FWL, GW, JCVA, JG, PFP, SAR, SG, RR, RSQ, YZ contributed samples, materials or data. DSR, DG, JSc, JSa, GAT, JMC, KV, KFXM, RR, SAR supervised part of the research. ACC, DL, FMa, SAR wrote the paper with help by SG, KFXM, DWS and contributions by all authors. JSc and SAR coordinated the project.

### ACKNOWLEDGEMENTS

We thank Richard Haas, Faezeh Donges, Marco Göttig and Katrin Kumke for technical assistance. We thank Walter Sanseverino and Riccardo Aliese (Sequentia Biotech) for assistance in TE RNAseq analyses. The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. Support to RR and SAR by the German Research Foundation (DFG RE 837/10-2), the Excellence Initiative of the German Federal and State Governments (EXC 294), and by the German Federal Ministry of Education and Research (BMBF FRISYS), is highly appreciated. CoGe is supported by the US National Science Foundation under Award Numbers IOS-339156 and IOS-1444490, CyVerse is supported by the U.S. National Science Foundation under Award Numbers DBI-0735191 and DBI-1265383. YK and ACC are grateful for support from the UK Biological Sciences and Biotechnology Research Council (Grant BB/F001797/1). KV acknowledges the Multidisciplinary Research Partnership 'Bioinformatics: from nucleotides to networks' Project (no 01MR0410W) of Ghent University. JC is grateful for support from the Spanish Ministerio de Economía y Competitividad (Grant AGL2013-43244-R). RSQ is grateful to Monsanto (St. Louis, MO, USA) for sequencing genomic DNA of *P. patens* accession Kaskaskia. The 1000 Plants (1 KP) initiative, led by GKSU, is funded by the Alberta Ministry of Innovation and Advanced Education, Alberta Innovates Technology Futures (AITF), Innovates Centres of Research Excellence (iCORE), Musea Ventures, BGI-Shenzhen and China National Genebank (CNCB). TW was supported by EMBO Long-Term Fellowships (ALTF 1166-2011) and by Marie Curie Actions (European Commission EMBOCOFUND2010, GA-2010-267146). The work conducted at PGSB was supported by the German Research Foundation (SFB924) and German Ministry of Education and Research (BMBF, 031A536/de.NBI). The authors declare that they have no competing interests.

### SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

**Appendix S1.** Supplementary Materials I–IV, Experimental Procedures, and Results including Tables S1–S23, Figures S1–S50, and References.

**File S1.** v3.1 + v3.3 annotation.

**File S2.** Plots of markers, TE methylation and histone modification, phenotypic differences of *P. patens* accessions, sRNA northern blots.

**File S3.** Synteny analyses, JGI gene atlas samples, NCLDV clusters/genes, JGI bead array SNP QC.

## REFERENCES

- Akalın, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A. and Mason, C.E. (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13**, R87.
- Arif, M.A., Fattash, I., Ma, Z., Cho, S.H., Beike, A.K., Reski, R., Axtell, M.J. and Frank, W. (2012) DICER-LIKE3 activity in *Physcomitrella patens* DICER-LIKE4 mutants causes severe developmental dysfunction and sterility. *Mol. Plant*, **5**, 1281–1294.
- Bainard, J.D. and Newmaster, S.G. (2010) Endopolyploidy in bryophytes: widespread in mosses and absent in liverworts. *J. Bot.* **2010**, 7.
- Bainard, J.D. and Villarreal, J.C. (2013) Genome size increases in recently diverged hornwort clades. *Genome*, **56**, 431–435.
- Bainard, J.D., Forrest, L.L., Goffinet, B. and Newmaster, S.G. (2013) Nuclear DNA content variation and evolution in liverworts. *Mol. Phylogenet. Evol.* **68**, 619–627.
- Barker, M.S., Kane, N.C., Matvienko, M., Kozik, A., Michelmore, R.W., Knapp, S.J. and Rieseberg, L.H. (2008) Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol. Biol. Evol.* **25**, 2445–2455.
- Beike, A.K., von Stackelberg, M., Schallenberg-Rudinger, M., Hanke, S.T., Follo, M., Quandt, D., McDaniel, S.F., Reski, R., Tan, B.C. and Rensing, S.A. (2014) Molecular evidence for convergent evolution and allopolyploid speciation within the *Physcomitrium-Physcomitrella* species complex. *BMC Evol. Biol.* **14**, 158.
- Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580.
- Bewick, A.J., Niederhuth, C.E., Ji, L., Rohr, N.A., Griffin, P.T., Leebens-Mack, J. and Schmitz, R.J. (2017) The evolution of CHROMOMETHYLASES and gene body DNA methylation in plants. *Genome Biol.* **18**, 65.
- Blanc, G., Agarkova, I., Grimwood, J. *et al.* (2012) The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biol.* **13**, R39.
- Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
- Cao, J., Schneeberger, K., Ossowski, S. *et al.* (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43**, 956–963.
- Chakravarti, A., Lasher, L.K. and Reefer, J.E. (1991) A maximum likelihood method for estimating genome length using genetic linkage data. *Genetics*, **128**, 175–182.
- Chaudhuri, P. and Marron, J.S. (1999) SiZer for exploration of structures in curves. *J. Am. Stat. Assoc.* **94**, 807–823.
- Dangwal, M., Kapoor, S. and Kapoor, M. (2014) The PpCMT chromomethylase affects cell growth and interacts with the homolog of LIKE HETEROCHROMATIN PROTEIN 1 in the moss *Physcomitrella patens*. *Plant J.* **77**, 589–603.
- De Bodt, S., Maere, S. and Van de Peer, Y. (2005) Genome duplication and the origin of angiosperms. *Trends Ecol. Evol.* **20**, 591–597.
- Dellaporta, S.L., Wood, J. and Hicks, J.B. (1983) A plant DNA miniprep: version II. *Plant Mol. Biol. Rep.* **1**, 19–21.
- Devos, N., Szovenyi, P., Weston, D.J., Rothfels, C.J., Johnson, M.G. and Shaw, A.J. (2016) Analyses of transcriptome sequences reveal multiple ancient large-scale duplication events in the ancestor of Sphagnopsida (Bryophyta). *New Phytol.* **211**, 300–318.
- Dolgin, E.S. and Charlesworth, B. (2008) The effects of recombination rate on the distribution and abundance of transposable elements. *Genetics*, **178**, 2169–2177.
- Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, **9**, 18.
- Feng, S., Cokus, S.J., Zhang, X. *et al.* (2010) Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl Acad. Sci. USA*, **107**, 8689–8694.
- Feng, J., Liu, T., Qin, B., Zhang, Y. and Liu, X.S. (2012) Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740.
- Finn, R.D., Coghill, P., Eberhardt, R.Y. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285.
- Fishman, L., Kelly, A.J., Morgan, E. and Willis, J.H. (2001) A genetic map in the *Mimulus guttatus* species complex reveals transmission ratio distortion due to heterospecific interactions. *Genetics*, **159**, 1701–1716.
- Flutre, T., Duprat, E., Feuillet, C. and Quesneville, H. (2011) Considering transposable element diversification in de novo annotation approaches. *PLoS ONE*, **6**, e16526.
- Fuchs, J., Demidov, D., Houben, A. and Schubert, I. (2006) Chromosomal histone modification patterns – from conservation to diversity. *Trends Plant Sci.* **11**, 199–208.
- Foissac, S., Bardou, P., Moisan, A., Cros, M.J. and Schiex, T. (2003) EUGENE/HOM: A generic similarity-based gene finder using multiple homologous sequences. *Nucleic Acids Res.* **31**, 3742–3745.
- Gernand, D., Demidov, D. and Houben, A. (2003) The temporal and spatial pattern of histone H3 phosphorylation at serine 28 and serine 10 is similar in plants but differs between mono- and polycentric chromosomes. *Cytogenet. Genome Res.* **101**, 172–176.
- Grabherr, M.G., Haas, B.J., Yassour, M. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652.
- Gremme, G., Brendel, V., Sparks, M.E. and Kurtz, S. (2005) Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.* **47**, 965–978.
- Harrison, C.J., Roeder, A.H., Meyerowitz, E.M. and Langdale, J.A. (2009) Local cues and asymmetric cell divisions underpin body plan transitions in the moss *Physcomitrella patens*. *Curr. Biol.* **18**, 18.
- Hartmann, M.A. (1998) Plant sterols and the membrane environment. *Trends Plant Sci.* **3**, 170–175.
- Hiss, M., Meyberg, R., Westermann, J., Haas, F.B., Schneider, L., Schallenberg-Rudinger, M., Ullrich, K.K. and Rensing, S.A. (2017) Sexual reproduction, sporophyte development and molecular variation in the model moss *Physcomitrella patens*: introducing the ecotype Reute. *Plant J.* **90**, 606–620 <https://doi.org/10.1111/tpj.13501>.
- Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V. and Quesneville, H. (2014) PASTEC: an automatic transposable element classification tool. *PLoS ONE*, **9**, e91929.
- Horst, N.A., Katz, A., Pereman, I., Decker, E.L., Ohad, N. and Reski, R. (2016) A single homeobox gene triggers phase transition, embryogenesis and asexual reproduction. *Nat. Plants*, **2**, 15209.
- Hu, R., Xiao, L., Bao, F., Li, X. and He, Y. (2016) Dehydration-responsive features of *Atrichum undulatum*. *J. Plant Res.* **129**, 945–954.
- Hutter, S., Vilella, A.J. and Rozas, J. (2006) Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics*, **7**, 409.
- Ibarra, C.A., Feng, X., Schoft, V.K. *et al.* (2012) Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. *Science*, **337**, 1360–1364.
- Jaffe, D.B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J.P., Zody, M.C. and Lander, E.S. (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91–96.
- Kaessmann, H. (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20**, 1313–1326.
- Kamisugi, Y., von Stackelberg, M., Lang, D., Care, M., Reski, R., Rensing, S.A. and Cuming, A.C. (2008) A sequence-anchored genetic linkage map for the moss, *Physcomitrella patens*. *Plant J.* **56**, 855–866.
- Kawashima, T. and Berger, F. (2014) Epigenetic reprogramming in plant sexual reproduction. *Nat. Rev. Genet.* **15**, 613–624.
- Keibler, E. and Brent, M.R. (2003) Eval: a software package for analysis of genome annotations. *BMC Bioinformatics*, **4**, 50.
- Kent, W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664.

- Khraiwesh, B., Arif, M.A., Seumel, G.I., Ossowski, S., Weigel, D., Reski, R. and Frank, W. (2010) Transcriptional control of gene expression by microRNAs. *Cell*, **140**, 111–122.
- Krueger, F. and Andrews, S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645.
- Lamb, J.C., Yu, W., Han, F. and Birchler, J.A. (2007) Plant chromosomes from end to end: telomeres, heterochromatin and centromeres. *Curr. Opin. Plant Biol.* **10**, 116–122.
- Lang, D., Weiche, B., Timmerhaus, G., Richardt, S., Riano-Pachon, D.M., Correa, L.G., Reski, R., Mueller-Roeber, B. and Rensing, S.A. (2010) Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. *Genome Biol. Evol.* **2**, 488–503.
- Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R.; Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
- Li, Y.H., Zhou, G., Ma, J. et al. (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045–1052.
- Lowe, T.M. and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964.
- Martinez, G. and Slotkin, R.K. (2012) Developmental relaxation of transposable element silencing in plants: functional or byproduct? *Curr. Opin. Plant Biol.* **15**, 496–502.
- Maumus, F., Epert, A., Nogue, F. and Blanc, G. (2014) Plant genomes enclose footprints of past infections by giant virus relatives. *Nat. Commun.* **5**, 4268.
- McDaniel, S.F., von Stackelberg, M., Richardt, S., Quatrano, R.S., Reski, R. and Rensing, S.A. (2010) The speciation history of the Physcomitrium-Physcomitrella species complex. *Evolution*, **64**, 217–231.
- McKenna, A., Hanna, M., Banks, E. et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303.
- Melters, D.P., Bradnam, K.R., Young, H.A. et al. (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.* **14**, R10.
- Newton, A.E., Wikström, N., Bell, N., Forrest, L.L. and Ignatov, M.S. (2006) Dating the diversification of the pleurocarpous mosses. In *Pleurocarpous mosses: Systematics and Evolution*. (Tangney, N., ed). Boca Raton: CRC Press, Systematics Association, pp. 329–358.
- Niederhuth, C.E., Bewick, A.J., Ji, L. et al. (2016) Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* **17**, 194.
- Oliver, M.J., Dowd, S.E., Zaragoza, J., Mauget, S.A. and Payton, P.R. (2004) The rehydration transcriptome of the desiccation-tolerant bryophyte *Tortula ruralis*: transcript classification and analysis. *BMC Genom.* **5**, 89.
- Olsen, J.L., Rouze, P., Verhelst, B. et al. (2016) The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. *Nature*, **530**, 331–335.
- Ortiz-Ramirez, C., Hernandez-Coronado, M., Thamm, A., Catarino, B., Wang, M., Dolan, L., Feijo, J.A. and Becker, J.D. (2015) A transcriptome atlas of *Physcomitrella patens* provides insights into the evolution and development of land plants. *Mol. Plant*, **9**, 205–220.
- Perroud, P.F., Cove, D.J., Quatrano, R.S. and McDaniel, S.F. (2011) An experimental method to facilitate the identification of hybrid sporophytes in the moss *Physcomitrella patens* using fluorescent tagged lines. *New Phytol.* **2**, 1469–8137.
- Proost, S., Fostier, J., De Witte, D., Dhoedt, B., Demeester, P., Van de Peer, Y. and Vandepoele, K. (2012) i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.* **40**, e11.
- Quesneville, H., Bergman, C.M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M. and Anxolabehere, D. (2005) Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput. Biol.* **1**, 166–175.
- Rensing, S.A. (2014) Gene duplication as a driver of plant morphogenetic evolution. *Curr. Opin. Plant Biol.* **17C**, 43–48.
- Rensing, S.A., Ick, J., Fawcett, J.A., Lang, D., Zimmer, A., Van de Peer, Y. and Reski, R. (2007) An ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*. *BMC Evol. Biol.* **7**, 130.
- Rensing, S.A., Lang, D., Zimmer, A.D. et al. (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64–69.
- Rensing, S.A., Beike, A.K. and Lang, D. (2012) Evolutionary importance of generative polyploidy for genome evolution of haploid-dominant land plants. In *Plant Genome Diversity* (Greilhuber, J., Wendel, J.F., Leitch, I.J. and Doležel, J., eds). Vienna, New York: Springer, pp. 295–305.
- Rensing, S.A., Sheerin, D.J. and Hiltbrunner, A. (2016) Phytochromes: more than meets the eye. *Trends Plant Sci.* **21**, 543–546.
- Reski, R., Faust, M., Wang, X.H., Wehe, M. and Abel, W.O. (1994) Genome analysis of the moss *Physcomitrella patens* (Hedw.) B.S.G. *Mol. Gen. Genet.* **244**, 352–359.
- Sakakibara, K., Ando, S., Yip, H.K., Tamada, Y., Hiwatahi, Y., Murata, T., Deguchi, H., Hasebe, M. and Bowman, J.L. (2013) KNOX2 genes regulate the haploid-to-diploid morphological transition in land plants. *Science*, **339**, 1067–1070.
- Salse, J. (2012) In silico archeogenomics unveils modern plant genome organisation, regulation and evolution. *Curr. Opin. Plant Biol.* **15**, 122–130.
- Salse, J., Abrouk, M., Murat, F., Quraishi, U.M. and Feuillet, C. (2009) Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. *Brief. Bioinform.* **10**, 619–630.
- Schlueter, J.A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J.J. and Shoemaker, R.C. (2004) Mining EST databases to resolve evolutionary events in major crop species. *Genome*, **47**, 868–876.
- Schween, G., Gorr, G., Hohe, A. and Reski, R. (2003) Unique tissue-specific cell cycle in *Physcomitrella*. *Plant Biol.* **5**, 50–58.
- Schween, G., Egner, T., Fritzkowsky, D. et al. (2005) Large-scale analysis of 73,329 gene-disrupted *Physcomitrella* mutants: production parameters and mutant phenotypes. *Plant Biol.* **7**, 238–250.
- Simillion, C., Janssens, K., Sterck, L. and Van de Peer, Y. (2008) i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics*, **24**, 127–128.
- Smit, A.F.A., Hubley, R. and Green, P. (1996) RepeatMasker Open-3.0. URL <http://www.repeatmasker.org> (unpublished), 2004.
- Spannagl, M., Nussbaumer, T., Bader, K.C., Martis, M.M., Seidel, M., Kugler, K.G., Gundlach, H. and Mayer, K.F. (2016) PGSB PlantsDB: updates to the database framework for comparative plant genome research. *Nucleic Acids Res.* **44**, D1141–D1147.
- Stroud, H., Otero, S., Desvoyes, B., Ramirez-Parra, E., Jacobsen, S.E. and Gutierrez, C. (2012) Genome-wide analysis of histone H3.1 and H3.3 variants in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA*, **109**, 5370–5375.
- Szovenyi, P., Ricca, M., Hock, Z., Shaw, J.A., Shimizu, K.K. and Wagner, A. (2013) Selection is no more efficient in haploid than in diploid life stages of an angiosperm and a moss. *Mol. Biol. Evol.* **30**, 1929–1939.
- Szovenyi, P., Perroud, P.F., Symeonidi, A., Stevenson, S., Quatrano, R.S., Rensing, S.A., Cumig, A.C. and McDaniel, S.F. (2014) De novo assembly and comparative analysis of the *Ceratodon purpureus* transcriptome. *Mol. Ecol. Resour.* **15**, 203–215.
- Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
- Van de Peer, Y., Mizrahi, E. and Marchal, K. (2017) The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424.
- Van de Velde, J., Van Bel, M., Van Eechoutte, D. and Vandepoele, K. (2016) A collection of conserved non-coding sequences to study gene regulation in flowering plants. *Plant Physiol.* **171**, 2586–2598.
- Vanneste, K., Baele, G., Maere, S. and Van de Peer, Y. (2014) Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome Res.* **24**, 1334–1347.
- Vanneste, K., Sterck, L., Myburg, A.A., Van de Peer, Y. and Mizrahi, E. (2015) Horsetails are ancient polyploids: evidence from *Equisetum giganteum*. *Plant Cell*, **27**, 1567–1578.

- Vekemans, D., Proost, S., Vanneste, K., Coenen, H., Viaene, T., Ruelens, P., Maere, S., Van de Peer, Y. and Geuten, K. (2012) Gamma paleohexaploidy in the stem lineage of core eudicots: significance for MADS-box gene and species diversification. *Mol. Biol. Evol.* **29**, 3793–3806.
- Vives, C., Charlot, F., Mhiri, C., Contreras, B., Daniel, J., Epert, A., Voytas, D.F., Grandbastien, M.A., Nogue, F. and Casacuberta, J.M. (2016) Highly efficient gene tagging in the bryophyte *Physcomitrella patens* using the tobacco (*Nicotiana tabacum*) Tnt1 retrotransposon. *New Phytol.* **212**, 759–769.
- Wang, G., Zhang, X. and Jin, W. (2009) An overview of plant centromeres. *J. Genet. Genomics* **36**, 529–537.
- Widiez, T., Symeonidi, A., Luo, C., Lam, E., Lawton, M. and Rensing, S.A. (2014) The chromatin landscape of the moss *Physcomitrella patens* and its dynamics during development and drought stress. *Plant J.* **79**, 67–81.
- Wright, S.I., Agrawal, N. and Bureau, T.E. (2003) Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. *Genome Res.* **13**, 1897–1903.
- Wu, T.D., Reeder, J., Lawrence, M., Becker, G. and Brauer, M.J. (2016) GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Methods Mol. Biol.* **1418**, 283–334.
- Zemach, A., McDaniel, I.E., Silva, P. and Zilberman, D. (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, **328**, 916–919.
- Zhang, Y., Liu, T., Meyer, C.A. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137.
- Zilberman, D. (2017) An evolutionary case for functional gene body methylation in plants and animals. *Genome Biol.* **18**, 87.
- Zimmer, A.D., Lang, D., Buchta, K., Rombauts, S., Nishiyama, T., Hasebe, M., Van de Peer, Y., Rensing, S.A. and Reski, R. (2013) Reannotation and extended community resources for the genome of the non-seed plant *Physcomitrella patens* provide insights into the evolution of plant gene structures and functions. *BMC Genom.* **14**, 498.