# Conserved key players required for *Physcomitrella patens* male fertility are affected by accumulation of (epi-) mutations

**Dissertation**

zur Erlangung der Doktorwürde der

Naturwissenschaften (Dr. rer. nat.)

Philipps Universität Marburg

Der Fakultät für Biologie der Philipps-Universität Marburg

vorgelegt von

**Rabea Meyberg**

aus Bielefeld, Nordrhein-Westfalen, Deutschland

Marburg an der Lahn

September 2019

Von der Philipps-Universität Marburg als Dissertation angenommen am

**Erstgutachter:** Prof. Dr. Stefan A. Rensing

**Zweitgutachterin:** Prof. Dr. Annette Becker

**Drittgutachter:** Prof. Dr. Alfred Batschauer

**Viertgutachter:** Prof. Dr. Michael Bölker

Tag der Disputation am

*„The measure of intelligence is the ability to change.“*

Albert Einstein

# 1 Contents

# 2 Abbreviations

| | |
|---|---|
| µmol | micromol |
| *A. thaliana* | *Arabidopsis thaliana* |
| ABA | abscisic acid |
| ASPB | American Society of Plant Biologists |
| ATG7 | autophagy related protein 7 |
| BB | basal body |
| BELL1 | bellringer 1 |
| bp | base pair |
| BS-seq | bisulfite sequencing |
| *C. reinhardtii* | *Chlamydomonas reinhardtii* |
| C° | celsius |
| ccdc39 | coiled-coil domain containing protein 39 |
| DNA | complementary DNA |
| CLF | curly leaf |
| cm | centimeter |
| CP | central pair |
| cryoSEM | cryo scanning electron microscopy |
| DEG | differentially expressed gene |
| DEK1 | defective kernel 1 |
| DMP | differentially methylated position |
| DMR | differentially methylated region |
| DNA | deoxyribonucleic acid |
| DNMT | DNA methyltransferases |
| e. g. | exempli gratia (for example) |
| EMBO | European Molecular Biology Organization |
| EST | expressed sequence tags |
| F1 | first filial generation |
| FECA | first eukaryoic common ancestor |
| FIE | fertilization-independent endosperm |
| Fig. | figure |
| Gd | Gransden |
| GdJ | Gransden Japan |
| gfp | green fluorescent protein |
| GLR | glutamate-like receptor |
| GO | gene ontology |
| GWAS | genome wide association study |
| h | hour |
| IDA | inner dynein arm |
| iMOSS | international molecular moss science society |
| JGI | Joint Genome Institute |
| Ka | non-synonymous substitutions per non-synonymous site |
| kDA | kilo Dalton |
| KNOX1/2 | knotted homeobox 1/2 |
| KO | knock out |

| | |
|---|---|
| Ks | synonymous substitutions per synonymous site |
| LECA | last eukaryotic common ancestor |
| *M. Polymorpha* | *Marchantia polymorpha* |
| $m^2$ | square meter |
| MARA | Marburg Research Academy |
| MID | Marburg International Doctorate |
| miRNA | micro-RNA |
| MLS | multi-layered structure |
| MRCA | most recent common ancestor |
| mRNA | messenger-RNA |
| MT | microtubule |
| MTOC | microtubule organizing center |
| mya | million years ago |
| ODA | outer dynein arm |
| *P. patens* | *Physcomitrella patens* |
| PCD | Primary Ciliary Dyskinesia |
| PCR | polymerase chain reaction |
| PCR2 | polycomb repressive complex 2 |
| PEAT | Physcomitrella expression atlas tool |
| qPCR | quantitative real-time PCR |
| Re | Reute |
| RE | relative expression |
| resp. | respectively |
| RNA | ribonucleic acid |
| RPKM | reads per kilo base per million mapped reads |
| RSP | radial spoke protein |
| RTPCR | reverse transcriptase PCR |
| rU | Relative units |
| s | second |
| S1-S3 | sporophyte 1-3 |
| SAR | Stramenopiles, Alveolates, Rhizaria |
| SM | sporophyte mature |
| SM | spline microtubule |
| SMC | spermatid mother cell/spore mother cell |
| SNP | single nucleotide polymorphism |
| TAP | transcription associated protein |
| TE | transposable element |
| TEM | transmission electron microscopy |
| TF | transcription factor |
| TR | transcriptional regulator |
| UK | United Kingdom |
| UV | Ultra violet |
| Vx | Villersexel |
| ZCC | Zygnema, Choleochaete, Charales |

# 3 Publications and contributions

## 3.1 Publications contributing to this thesis

Most of the results yielded during my time as a PhD student were published or are submitted for publication to peer-reviewed journals. My contributions to each publication are described below each publication.

**Characterization of evolutionary conserved key players affecting eukaryotic flagellar motility and fertility using a moss model.**

**Rabea Meyberg**, Pierre-François Perroud, Fabian B. Haas, Lucas Schneider, Thomas Heimerl, Karen S. Renzaglia, Stefan A. Rensing

*bioRxiv, 2019: DOI: 10.1101/728691*

**My contributions:** BS-seq and RNA-seq sample preparation and data analysis, quantification of sporophytes in selfing and crossing analysis, spermatozoid analysis, sample preparation for TEM and SEM, analysis of all microscopic data, blast and alignment for ccdc39 phylogeny, RTPCR and qPCR, GO bias analysis and word clouds, KO generation, genotyping, phenotyping, illustrations, wrote the manuscript with the help of Stefan A. Rensing and Karen S. Renzaglia.

***Physcomitrella patens*** **Reute mCherry as a tool for efficient crossing within and between ecotypes.**

Pierre-François Perroud, **Rabea Meyberg**, Stefan A. Rensing

*Plant Biology, 2019: 21(S1): 143-149*

**My contributions:** setup and analysis of part of the crossings.

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

Daniel Lang, Kristian K. Ullrich, Florent Murat, Jörg Fuchs, Jerry Jenkins, Fabian B. Haas, Mathieu Piednoel, Heidrun Gundlach, Michiel Van Bel, **Rabea Meyberg**, Cristina Vives, Jordi Morata, Aikaterini Symeonidi, Manuel Hiss, Wellington Muchero, Yasuko Kamisugi, Omar Saleh, Guillaume Blanc, Eva L. Decker, Nico van Gessel, Jane Grimwood, Richard D. Hayes, Sean W. Graham, Lee E. Gunter, Stuart McDaniel, Sebastian N.W. Hoernstein, Anders Larsson, Fay-Wei Li, Pierre- François Perroud, Jeremy Phillips, Priya Ranjan, Daniel S. Rokshar, Carl J. Rothfels, Lucas Schneider, Shengqiang Shu, Dennis W. Stevenson, Fritz Thümmler, Michael Tillich, Juan Carlos Villarreal A., Thomas Widiez, Gane Ka-Shu Wong, Ann Wymore, Yong Zhang, Andreas D. Zimmer, Ralph S. Quatrano, Klaus F.X. Mayer, David Goodstein, Josep M. Casacuberta, Klaas Vandepoele, Ralf Reski, Andrew C. Cuming, Jerry Tuskan, Florian Maumus, Jérome Salse, Jeremy Schmutz, Stefan A. Rensing

*The Plant Journal, 2018; 93 (3): 515-533*

**My contributions**: sample preparation for BS-seq, BS-seq analysis, contributed to the manuscript, life cycle illustrations.

**Sexual reproduction, sporophyte development and molecular variation in the model moss *Physcomitrella patens*: introducing the ecotype Reute.**

Manuel Hiss, **Rabea Meyberg**, Jens Westermann, Fabian B. Haas, Lucas Schneider, Mareike Schallenberg-Rüdinger, Kristian K. Ullrich, Stefan A Rensing

*The Plant Journal, 2017; 90(3): 606-620*

**My contributions:** analysis of the Reute life cycle, in-depth gametangia analysis of all ecotypes including microscopic analysis, sample preparation for cryoSEM, quantification of sporophytes developed per ecotype, RNA extraction for qPCR, setup of figure 1 & 2, helped writing the manuscript, illustrations.

## 3.2 Additional publications not contributing to the thesis

**A Blind and Independent Benchmark Study for Detecting Differentially Methylated Regions in Plants.**

Clemens Kreutz, S. Nilay Can, Ralph Schulze Bruening, **Rabea Meyberg**, Zusanna Mérai, Noe Fernandez-Pozo, Stefan A. Rensing

*Under review 09/2019, Bioinformatics*

**My contributions:** Initial manual DMR selection, BS-seq data.

**PEATmoss (Physcomitrella Expression Atlas Tool): a unified gene expression atlas for the model plant *Physcomitrella patens.***

Noe Fernandez-Pozo, Fabian B. Haas, **Rabea Meyberg**, Kristian K. Ullrich, Manuel Hiss, Pierre-François Perroud, Sebastian T. Hanke, Viktor Kratz, Adrian Powell, Eleanor F. Vesty, Christopher G. Daum, Matthew Zane, Anna Lipzen, Avinash Sreedasyam, Jane Grimwood, Juliet C. Coates, Kerrie Barry, Jeremy Schmutz, Lukas A. Mueller, Stefan A. Rensing

*Under review 09/2019, The Plant Journal*

**My contributions:** Organisation of sporophyte development metadata, illustrations.

**ABA-Induced Vegetative Diaspore Formation in *Physcomitrella patens*.**

M. Asif Arif, Manuel Hiss, Martha Tomek, Hauke Busch, **Rabea Meyberg**, Stefanie Tintelnot, Ralf Reski, Stefan A. Rensing, Wolfgang Frank

*Frontiers in Plant Science 2019: 10:315, DOI: 10.3389/fpls.2019.00315*

**My contributions:** Micro array data analysis, GO bias analysis, image corrections and false colouring.

# 4 Abstract

*Physcomitrella patens* belongs to the bryophytes and is an extant species of the first land plants. This phylogenetic informative position allows the analysis of key evolutionary steps in (land) plant evolution employing *P. patens* as a model organism. Decades of research mainly focused on the gametophytic generation, probably also due to the lack of sexual reproductive events in the primarily used ecotype Gransden. Long term *in vitro* vegetative reproduction probably led to the accumulation of somatic (epi-) mutations which eventually led to a nearly male sterile phenotype. So far, only few *P. patens* ecotypes are used for scientific work. Thus, to overcome the fertility issues and to apply comparative analyses to study sexual reproduction and sporophyte development as well as species and population divergence in *P. patens*, the establishment of more ecotypes is highly needed. In comparison to other plant model organisms as e.g. *Arabidopsis thaliana*, genome wide epigenetic modifications especially with regard to sexual reproduction are still barely studied in *P. patens* ecotypes.

Here I present the characterization of the sexual reproduction of the recently introduced fertile ecotype Reute which was collected 2006 in Germany. Reute is the most closely related ecotype to Gransden reported so far. In a comparative analysis between the ecotypes Gransden, Reute and Villersexel, I could show no differences in timing and morphology of the sexual reproductive tissues. However, while Reute was as fertile as the more distant ecotype Villersexel, Gransden was nearly self-sterile. Also, I present the fluorescent marker strain Reute-mCherry which can be used in crossing analyses e.g. to determine if female or male sexual reproductive organs are impaired. By employing this method, a clear male defect could be shown in Gransden. Further, I present a comparative multi-omics analysis between Gransden and Reute using different tissues during sexual reproduction. Single nucleotide polymorphisms (SNPs), DNA-methylation and RNA-expression pinpoint a flagellar defect, which presumably leads to the observed male fertility impairment in Gransden. Finally, I present the characterization of key-players which are highly conserved within eukaryotes and are required for flagellar motility in humans as well as in the moss *P. patens*.

# 5 Zusammenfassung

Das Moos *Physcomitrella patens* ist ein rezenter Nachkomme der ersten Landpflanzen. Diese phylogenetisch interessante Position ermöglicht die Analyse von Schlüsselmomenten in der Evolution der (Land-) Pflanzen. *P. patens* wird seit vielen Jahrzehnten als Modellorganismus verwendet. Dennoch konzentrieren sich die meisten Arbeiten bisher auf die gametophytische Generation, welches sich durch das seltene Vorkommen der sexuellen Reproduktion im vorwiegend verwendeten Ökotyp Gransden erklären lässt. Seit vielen Jahren wird das Gewebe in der *in vitro* Kultur ausschließlich vegetativ vermehrt. Die vermutlich daraus resultierende Anhäufung von somatischen (Epi-) Mutationen hat die männliche Sterilität von Gransden zur Folge. Bisher werden nur wenige der bekannten *P. patens* Ökotypen für wissenschaftliche Arbeiten verwendet. Daher ist die Etablierung von weiteren Ökotypen relevant. Zum einen können so Analysen der sexuellen Reproduktion und des daraus resultierendem Sporophyten durchgeführt werden, zum anderen ermöglichen Ökotypen vergleichende Analysen der Spezies- und Populationsvielfalt in *P. patens*. Im Vergleich zu anderen Pflanzenmodellorganismen wie z.B. *Arabidopsis thaliana*, gibt es in *P. patens* bisher nur wenige genomweite Analysen von epigenetischen Modifikationen, vor allem im Bezug auf Gewebe, die in der sexuellen Reproduktion involviert sind.

In dieser Arbeit präsentiere ich die Charakterisierung der sexuellen Reproduktion des vor kurzem vorgestellten fertilen *P. patens* Ökotyps Reute, welcher 2006 in Deutschland gesammelt wurde. Reute ist im Vergleich mit allen sequenzierten Ökotypen genetisch gesehen am nächsten mit dem bisher verwendeten Ökotyp Gransden verwandt. In einer vergleichenden Studie zwischen den drei Ökotypen Gransden, Reute und Villersexel konnte ich zeigen, dass die zeitliche und morphologische Entwicklung der Geschlechtsorgane sich nicht unterscheidet und Reute so fertil wie der genetisch etwas entferntere Ökotyp Villersexel ist. Im Gegensatz dazu konnte ich zeigen, dass Gransden nahezu steril ist. Außerdem präsentiere ich den fluoreszenten Stamm Reute-mCherry, welcher für Kreuzungsanalysen verwendet werden kann um z.B. zu bestimmen, ob die männlichen und/oder weiblichen Geschlechtsorgane funktional sind. Mithilfe dieser Methode konnte eindeutig gezeigt werden, dass Gransden einen Defekt im männlichen Reproduktionsapparat zeigt. Des Weiteren beinhaltet diese Arbeit zwischen Gransden und Reute vergleichende Datensätze von Nukleotidpolymorphismen, der Genexpression und des DNA-Methylierungsstatus von Geweben die in die sexuelle Reproduktion involviert sind. Genetische und epigenetische Unterschiede zwischen den Ökotypen weisen auf Defekte in den Flagellen der männlichen Spermatozoide hin, welche vermutlich zur männlichen Sterilität des Ökotyps Gransden beitragen. Abschließend präsentiere ich die Charakterisierung von Schlüsselgenen die in Eukaryoten konserviert sind und für die flagellare

Motilität der menschlichen Spermien sowie der Spermatozoide des Mooses *P. patens* benötigt werden.

# 6 Introduction

## *6.1* Advantages of using *Physcomitrella patens* as a model organism

The moss *Physcomitrella patens* belongs to the *Funariaceae* and the main laboratory strain Gransden (Gd) has initially been collected in Gransden Wood (UK) in 1962. *In vitro*, *P. patens* completes its lifecycle within two to three months and predominantly self-fertilizes (Nakosteen and Hughes, 1978; Perroud *et al.*, 2011). For the well-established model organism, *in vitro* culture and a large set of methods are established, and it can be propagated via vegetative and sexual reproduction (Hohe *et al.*, 2002; Cove, 2005; Cove *et al.*, 2009; Strotbek *et al.*, 2013). Also, a single cell of any tissue is sufficient to regenerate a whole plant (Cove, 2000), which is employed for transient and stable transfections via protoplasts. Combined with the presence of a highly efficient homologous recombination, which allows precise gene targeting, *P. patens* is a well-established model organism to analyze gene function via reverse genetic approaches in stable mutants (Schaefer and Zrÿd, 1997; Schaefer, 2001; Kamisugi *et al.*, 2006).

*P. patens* is, as all land plants, a haplo-diplont with the haploid gametophyte representing the dominant phase. The prevalence of the dominant haploid generation has the advantage, that genetic modifications are directly visible in the first (haploid) generation and no crossing is necessary to generate a homozygous F2 to analyze the effect of introduced mutations. Also, embryo-lethal mutations as observed in dominant diploid model organisms like e.g. *Arabidopsis thaliana*, are not necessarily lethal in the gametophytic phase, which is highly reduced in angiosperms, but easily accessible in bryophytes like *P. patens* (Cove, 2005). In comparison to gymnosperms and angiosperms, bryophytes develop easily accessible multicellular sporophytes and gametophytes as well as separated male and female reproductive organs. Easily accessible tissues of most developmental stages in both generations makes it applicable for research with regard to the alternation of generations, which is defined as the alternation between a multicellular gametophyte and a multicellular sporophyte (Hofmeister, 1851; O'Donoghue *et al.*, 2013). In 2008, the *P. patens* genome has been published (Rensing *et al.*, 2008) as the first non-seed plant genome. Additionally, DNA methylation (Zemach *et al.*, 2010) and histone modification (Widiez *et al.*, 2014) data are available. Several transcriptomic analyses are available using both array- and RNA-seq approaches covering diverse treatments and developmental stages (Ortiz-Ramírez *et al.*, 2016; Perroud *et al.*, 2018). Since its introduction, *P. patens* has been used for several morphological and genetic (Engel, Paulinus P., 1968; Cove, 2005; Cove *et al.*, 2006), evolutionary-developmental (Sakakibara *et al.*, 2008; Aya *et al.*, 2011; Sakakibara *et al.*, 2013; Kofuji and Hasebe, 2014) as well as functional studies, analyzing mammalian homologs in plants (Horst *et al.*, 2016; Ortiz-Ramírez *et al.*, 2017).

## 6.2 Phylogenetic background

All eukaryotes evolved from the last eukaryotic common ancestor (LECA). The LECA already held many key features of eukaryotes, namely the nucleus, mitochondria, an actin/microtubule skeleton, an endomembrane system, flagella and machineries for splicing and meiosis (Stewart and Mattox, 1975; Mast *et al.*, 2014). It is proposed that mitochondria, and later during the evolution of plants also chloroplasts, derived from bacteria, which were taken up by eukaryotic ancestors and gave rise to the first unicellular algae (Fig. 1, (Keeling, 2004; Keeling, 2010; Zimorski *et al.*, 2014)).

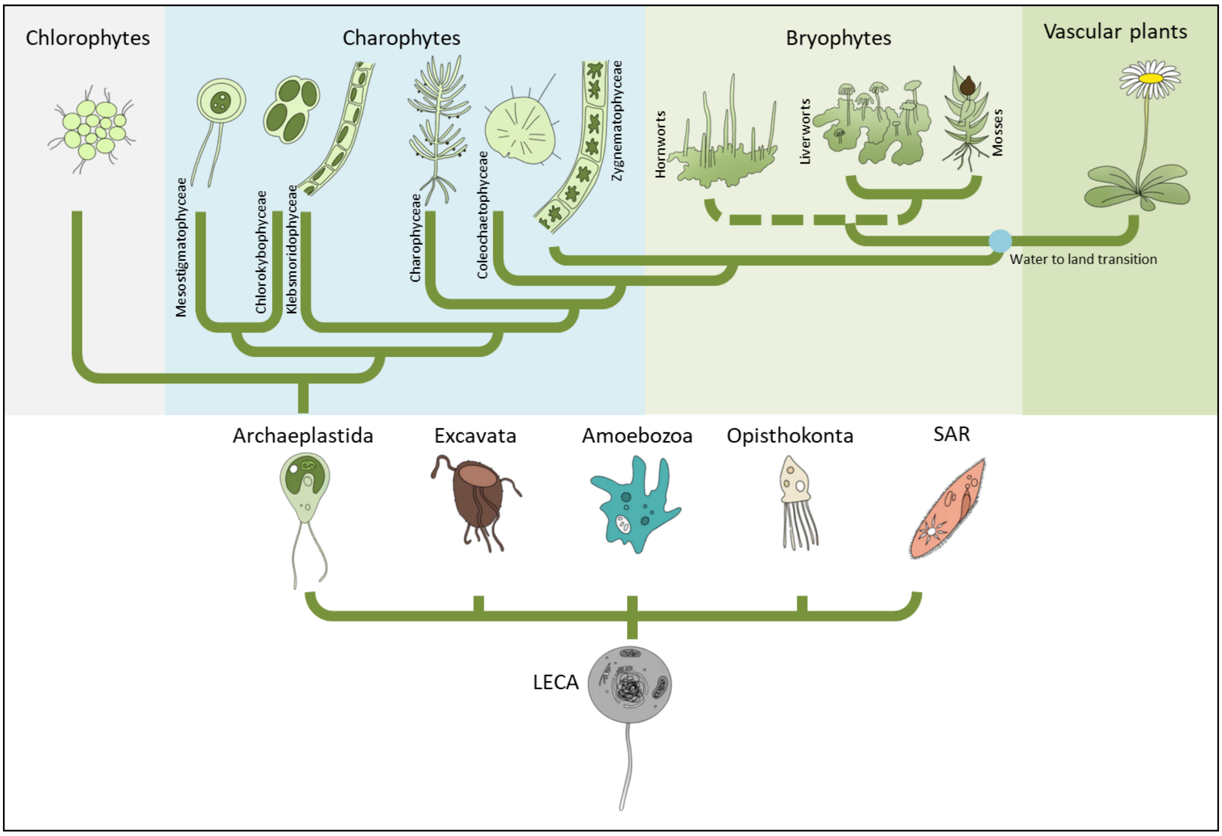


**Figure 1:** All eukaryotic life originates in the last eukaryotic common ancestor (LECA). The LECA gave rise to all five kingdoms of eukaryotic life whereas the Archaeplastida include the green lineage, the plants. Chlorophytes and streptophytes, which comprise the charophytes and all embryophytes (land plants), diverged approx. 1 billion years ago. Within the streptophytes the Zygnematophyceae, Choleochaetophyceae and Charophyceae (ZCC) form a clade with the embryophytes. The water to land transition is indicated by a cyan dot and first land plants were probably akin to bryophytes, which comprise hornworts, liverworts and mosses of which the last two form the Setaphyta clade. Wether hornworts and setaphytes form bryophytes or bryophyta and if hornworts or setaphytes are sister to all land plants still has to be solved. Unsolved branches in the bryophytes indicated with a dashed line. Redrawn from (Mast *et al.*, 2014; Puttick *et al.*, 2018; de Vries and Archibald, 2018), images are not to scale, SAR: Stramenopiles, Alveolates, Rhizaria.

Most green algae are haplontic plants showing a diverse growth pattern ranging from unicellular algae like *Chlamydomonas reinhardtii* to the highly complex *Chara braunii*. The MRCA of all algae, gave rise to the chlorophytic algae, which have a rather simple body plan, and the streptophytic algae, which show complex two- and three-dimensional body plans (Umen, 2014). Interestingly, the haplo-diplontic lifecycle was invented independently in the chloro- and streptophytes (Niklas and Kutschera, 2010). In comparison to the chlorophytic algae, all streptophytic algae are haplonts and haplo-diplonts appeared within the first land plants, the bryophytes. The streptophytic algae comprise the paraphyletic KCM (Klebsormidiophyceae, Chlorokybophyceae, Mestostigmatophyceae) grade and the ZCC (Zygnematophyceae, Choleochaetophyceae, Charophyceae) grade whereas the Zygnematophyceae were shown to be the sister group to all land plants (Wickett *et al.*, 2014; de Vries and Archibald, 2018). Streptophyte algae and all land plants together comprise the monophyletic Phragmoplastophyta (Nishiyama *et al.*, 2018). The bryophytes which comprise hornworts, liverworts and mosses as *P. patens*, are probably the sister group to all vascular plants and positioned at a phylogenetically interesting position, since their ancestors are thought to akin the first land plants after the water to land transition which occurred approx. 430-500 mya (million years ago) (Kenrick and Crane, 1997; Morris *et al.*, 2018). Here, new emerging model organisms slowly start to fill up the gap between streptophytic algae and angiosperms, to get deeper evolutionary insights into (land) plant evolution (Rensing, 2017). This informative phylogenetic position makes them ideal model organisms for the analysis of key evolutionary steps in the (land) plant evolution (Michael J. Prigge and Magdalena Bezanilla, 2010). To gain deeper knowledge about evolutionary forces, acting on the ancestral embryophytes, a comparative analysis between members of this group is needed as well as the solution of the phylogenetic relationship. In the last two years, several laboratories have worked on this topic and could show that bryophytes might be monophyletic and that, based on a parsimonious model, a setaphyta clade is formed, which is comprising mosses and liverworts (Renzaglia and Garbary, 2010). Employing morphological features of the motile male gametes of mosses and liverworts, this relationship recently was reinforced (Renzaglia *et al.*, 2018). Still, it has to be determined if hornworts, or moss and liverwort ancestors emerged first after the water to land transition and if bryophytes are truly monophyletic (Rensing, 2018).

## 6.3 *Physcomitrella patens* ecotypes and strains

*P. patens* has been reported to grow in North America, Europe, Africa, China, Japan, Australia and the land masses of the Holarctic. Thus, several accessions/ecotypes are available (Von Stackelberg *et al.*, 2006; Frey *et al.*, 2009; McDaniel *et al.*, 2010; Beike *et al.*, 2014; Medina *et al.*, 2015; Medina *et al.*, 2019). Per definition, ecotypes are geographically and genetically distinct populations of the same species which adapted to their specific environment (Ferrero-Serrano and Assmann, 2019), whereas a species is defined by the ability to still produce fertile offspring when crossed and adaptation as the development of an inheritable trait (Bock, 1980). Even if *P. patens* has a worldwide distribution, only few accessions were used for scientific work so far. Ecotypes can for example be used for genome wide association studies (GWAS) to identify alleles correlating with a specific trait, which was gained through adaption to a specific environment or to analyze the development of a specific population. In Brassicaceae, ecotypes show between one per 10 base pairs (bp) to one per 285 bp single nucleotide polymorphism (SNPs (Cao *et al.*, 2011; Wei *et al.*, 2017)). Until 2000, the single spore isolate ecotype Gd, which was collected by Whitehouse in the UK, was worldwide used as the main laboratory strain (Beike *et al.*, 2014). In 2003 the Villersexel (Vx) accession has been collected by M. Lüth in Haute Saône (France) and was used to generate a genetic map through crossing with Gd (Kamisugi *et al.*, 2008). The genetic distance between Gd and Vx was determined to be the highest reported so far in *P. patens* ecotypes, showing one SNP in 829 bp (Kasahara *et al.*, 2011) which is still rather low when compared to SNP frequency between Brassicaceae ecotypes. This lower number of SNPs between *P. patens* ecotypes could be explained with the rather low rate of synonymous substitutions per site per year in bryophytes compared to angiosperms: *P. patens*: 1.9 in comparison to Brassicaceae: 7.71 (Rensing *et al.*, 2007; De La Torre *et al.*, 2017), as well as by the lack of ecotype sequencing data. Additionally, the genetic divergence is probably also affected by the mode of sexual reproduction, whereas primary self-fertilizing mosses as *P. patens* probably show less genetic variations in comparison to e.g. obligate crossing species such as the moss *Ceratodon*, in which the F1 generation has been shown to have interspecies haplotypes (Nieto-Lugilde *et al.*, 2018). Decades of research employing *P. patens*, focused on the gametophytic generation, except of very early studies carried out in 1924 by von Wettstein (Wettstein, 1924). Thus, vegetative reproduction was used to keep laboratory strains in culture. When scientists started to analyze the impact of mutations during sexual reproduction and gained interest in the process of sexual reproduction itself in *P. patens*, several laboratories reported fertility issues with their Gd cultures (Ashton and Raju, 2000; Landberg *et al.*, 2013) which are suggested to be based on somatic mutations due to the long term vegetative reproduction *in vitro* (Ashton and Raju, 2000; Perroud *et al.*, 2011).

## 6.4 *Physcomitrella patens* lifecycle and sexual reproduction

The *P. patens* lifecycle starts with the germination of a haploid spore which gives rise to gametophytic protonema cells which grow two-dimensional via tip growth and branching (Menand *et al.*, 2007; Harrison *et al.*, 2009). This filamentous growth state of *P. patens* consists of two cell types, the chloroplast-rich chloronema cells, displaying a perpendicular cell wall and the long and thin caulonema cells which display far less and smaller chloroplasts and diagonal cell walls. Chloronemata mainly grow in the inner part of the filamentous cell cluster and are the photosynthetically active part of this developmental step, whereas caulonema cells grow at the periphery of the cluster, extend the boundaries of the surface covered and acquire nutrients. Eventually, caulonema cell branches give rise to a three-faced bud with an apical stem cell which will now grow three-dimensional instead of two-dimensional (Harrison *et al.*, 2009).

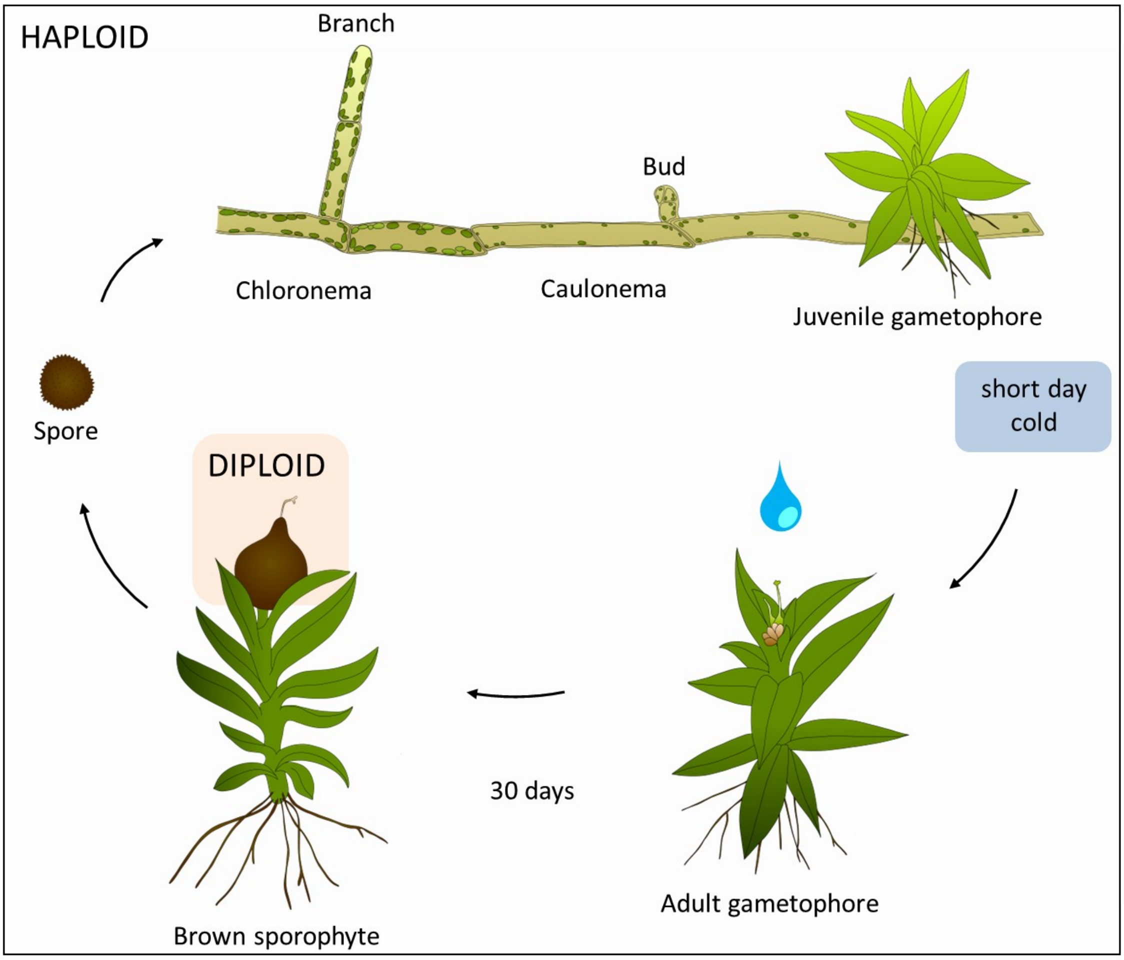


**Figure 2:** Life cycle of *P. patens*. Haploid spores germinate and filamentous protonema develops, consisting of chloronema and caulonema cells. The protonemal tissue gives rise to buds which eventually develop into juvenile gametophores. Upon short and cold day conditions, sexual reproductive organs (gametangia) are formed at the apex of the now adult gametophore. When moistened by water, the egg cell is fertilized by the motile spermatozoids. Subsequently the diploid zygote is formed and embryo-/sporophyte development takes place. After mitotic division and maturation, spores of the new generation can be released. Figure modified from Lang et al., 2018.

The bud subsequently develops into the juvenile gametophore which does not possess sexual reproductive tissue. Gametophores are leafy stems, which consist of a multicellular shoot-like stem bearing several one cell-layer thick non-vascular leaf-like structures, the so called leaflets or phyllids. They show a phenol-enriched cuticula which is ancestral to the lignin evolution in land-plants and probably contributes to the erect growth of the gametophore and the rigidity of the leaflets (Renault *et al.*, 2017). The juvenile gametophyte is anchored on the substrate via multicellular filamentous rhizoids, analogous to roots in angiosperms which deliver water and nutrients to the growing plant (Jones and Dolan, 2012). Vegetative in vitro culture of *P. patens* is usually performed under long day conditions (16 h light, 8 h dark, 21 °C, 70 µmol/m$^2$/s) ensuring efficient and fast cell divisions (Hohe *et al.*, 2002).

To induce the sexual reproduction, which usually takes place in autumn to spring, the plants require colder temperatures and short day conditions (8 h light, 16 h dark, 15 °C, 20 µmol/m$^2$/s) (Engel, Paulinus P., 1968; Nakosteen and Hughes, 1978; Hohe *et al.*, 2002). Since *P. patens* is monecious, both, the female (archegonia) and male (antheridia) sexual reproductive organs (gametangia) develop at the same apex of the gametophore out from an antheridium/archegonium apical stem cell (Kofuji *et al.*, 2009; Landberg *et al.*, 2013). Under reproductive conditions, the gametophore apical stem cell usually producing leaflet apical stem cell precursors develops into the antheridium initial stem cell which gives rise to several antheridium apical stem cells. The antheridium apical stem cells subsequently develop into bundles of antheridia (Kofuji *et al.*, 2018). The antheridium initial cell undergoes several anti-clinal divisions and later, an antheridia jacket is developed which covers the spermatid mother cells (Landberg *et al.*, 2013). Similarly to the liverwort *Marchantia polymorpha*, each spermatid mother cell divides by mitosis to form two spermatids, which undergo nuclear

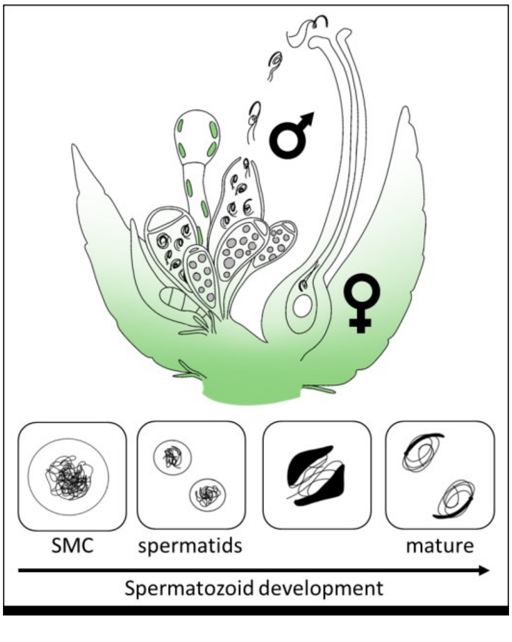


**Figure 3:** Spermatozoid development and fertilization. Each apex of adult gametophores bears several antheridia (male) and at least one archegonium (female). In the antheridia, spermatid mother cells develop, which give rise to two spermatids. They undergo nuclear condensation and morphogenesis to form the mature spermatozoid. The antheridial tip cell (t) swells upon maturity and releases spermatozoids, when moistened by water. Biflagellated spermatozoids subsequently swim through the neck canal cells (nc) of the archegonium to reach the egg cell (e) in the archegonial venter. Fertilization takes place and the diploid zygote is formed.

condensation and morphogenesis to release the mature spermatozoids (Fig.3, (Hisanaga *et al.*, 2019)). During differentiation and maturation processes, autophagy was shown to be required for proper sperm development (Sanchez-Vera *et al.*, 2017). Upon spermatozoid maturity, the antheridial tip cell swells and bursts (when moistened by water) to release the mature biflagellate spermatozoids. A few days after antheridia initiation (Kofuji *et al.*, 2009), the archegonium apical stem cell undergoes, similarly to the antheridium apical stem cell, anti-clinal divisions to form inner and outer cells. The inner cell subsequently divides anti-clinally and gives rise to the egg-initial and canal cells. The lower part develops into the egg-containing venter, whereas the canal cells divide and elongate to become the archegonial neck. Upon maturity, the previously closed archegonial tip opens and the canal cells dissolve. The mature spermatozoids can now reach and fertilize the egg cell to induce sporophyte development. During archegonium maturation, autophagous processes can be detected in the canal cells as well as in the egg and basal cell. Defects in autophagy correlate with more dense material in the canal cells which does not inhibit fertilization (Landberg *et al.*, 2013; Sanchez-Vera *et al.*, 2017).

After successful fertilization, the diploid sporophyte develops from the zygote and undergoes embryogenesis. The first division is asymmetric, forming an apical and a basal cell. The apical cell will develop into the spore capsule, whereas the lower part will penetrate into the gametophore and forms in comparison to other Funariaceae a short seta (Engel, Paulinus P., 1968; Medina *et al.*, 2019). The sporophyte cells contain chloroplasts, but also receive metabolites and nutrients from the gametophyte throughout the sporophyte development (Haig, 2013; Regmi *et al.*, 2017). The sporophyte develops through differentially described and grouped stages. Groups are defined by age (S1 – SM, Ortiz-Ramirez *et al.*, 2016) or by developmental stage (initial to brown sporophyte (O'Donoghue *et al.*, 2013; Daku *et al.*, 2016). First steps of embryo development are performed within the archegonium (S1-S2, initial development). These first developmental steps are mainly defined by cell amplification leading to a growing embryo. During these steps, the archegonium ruptures and becomes the calyptra, which is at least protecting the young sporophyte from dehydration in mosses (Budke and Goffinet, 2016). Afterwards, the capsule enlarges (S3) and sporogenesis is performed. After meiosis of the spore mother cells (SMC) in early green sporophytes, tetrads give rise to each four haploid spores. During spore maturation, the sporophyte turns from mid green (mid sporophyte) to yellow. After maturation, the brown mature sporophyte ruptures and releases up to 3000 spores of the new generation (SM (Nakosteen and Hughes, 1978; Hohe *et al.*, 2002; Sakakibara *et al.*, 2008)). As in all other land plants, bryophytes possess stomata on the sporophytic generation which is, in the case of bryophytes, the three-dimensional sporophyte. In bryophytes, stomata are suggested to play a role during sporophyte maturation, but being a highly discussed topic, additional work will have to be performed to elucidate stomata function (Chater *et al.*, 2017; Merced and Renzaglia, 2017).

## 6.5 Bryophytes possess evolutionary conserved flagella

Bryophytes are flagellated plants, which possess bi-flagellated spermatozoids, which are a defining and name-giving feature of streptophytes (Renzaglia and Garbary, 2001). Flagellated gametes are ancestral to all eukaryotes (Stewart and Mattox, 1975; Mitchell, 2007) and have been secondarily lost several times as e.g. in the algae Zygnematales (Transeau, 1951) and, probably in the MRCA of conifers, Gnetales and flowering plants (Renzaglia *et al.*, 2000; Renzaglia and Garbary, 2001). (Motile) flagella show a common architecture throughout the tree of life (Carvalho-Santos *et al.*, 2011) and are the only motile cells of archegoniates comprising bryophytes, lycophytes, pteridophytes and seed plants (Renzaglia and Garbary, 2001). Still, flagellated male gametes are highly adapted and show myriads of variations between species (Alvarez, 2017). Spermatozoids of the setaphytes show a sinistral coiled architecture which consists of a cylindrical condensed nucleus, two mitochondria, one plastid and two flagella at its distal end (Fig. 4A, (Renzaglia *et al.*, 2000)). Spermatozoids of the charophyceaen lineage and land plants possess a unique multi-layered structure (MLS), that comprises the spline microtubules (SM) and the lamellar strip (LS). The SM form the structural framework of the cell and the organelles are attached to it (Renzaglia *et al.*, 2018). The MLS is located within a microtubule organizing center (MTOC) which is derived from centrosomes. In the MTOC two end to end attached centrioles give rise to the dimorphic basal bodies (BB) which anchors the flagella and serves as the nucleation site of axonemes. Centrioles and BBs show a pattern of nine microtubule (MT) triplets, arranged around a central core with a stellate pattern. This pattern can vary in the BBs with the anterior basal body showing a pattern of up to nine MT triplets whereas the posterior basal body displays only three MT triplets on the bottom (Bernhard and Renzaglia, 1995; Renzaglia and Garbary, 2001; Renzaglia *et al.*, 2017). Between the motile axonemes and the basal body a transition zone with nine peripheral MTs doublets is present. Of the three microtubules originating in the centrioles, the a and b microtubules elongate in the transition zone, whereas the c microtubules are terminated (Fig. 4B,(Hodges *et al.*, 2012)). Motile axonemes show a highly conserved 200nm ultrastructure composed of nine peripheral MT doublets enclosing a central MT pair (CP) (Renzaglia *et al.*, 2000; Renzaglia and Garbary, 2001; Satir *et al.*, 2008; Carvalho-Santos *et al.*, 2011). The outer MT doublets are interconnected via nexin and associated with several motor and radial spoke proteins (RSP) important for stability and motility of the flagella. RSPs are composed of a stalk and a head and extend from the a MT of each outer MT doublet towards the CP MTs and serve as transient links (Fig. 4C). In *C. reinhardtii* RSPs are required for flagellar motility (Piperno *et al.*, 1981; Huang, 1986). Flagellar motility is derived from dynein motor proteins which drive interdoublet sliding (Porter and Sale, 2000; Heuser *et al.*, 2009). Two types of dyneins are shown to be present in the flagellar model organism *C. reinhardtii*: inner dynein arms (IDA) and outer dynein arms (ODA) (Dutcher, 2000), which work together in a highly coordinated complementary way to

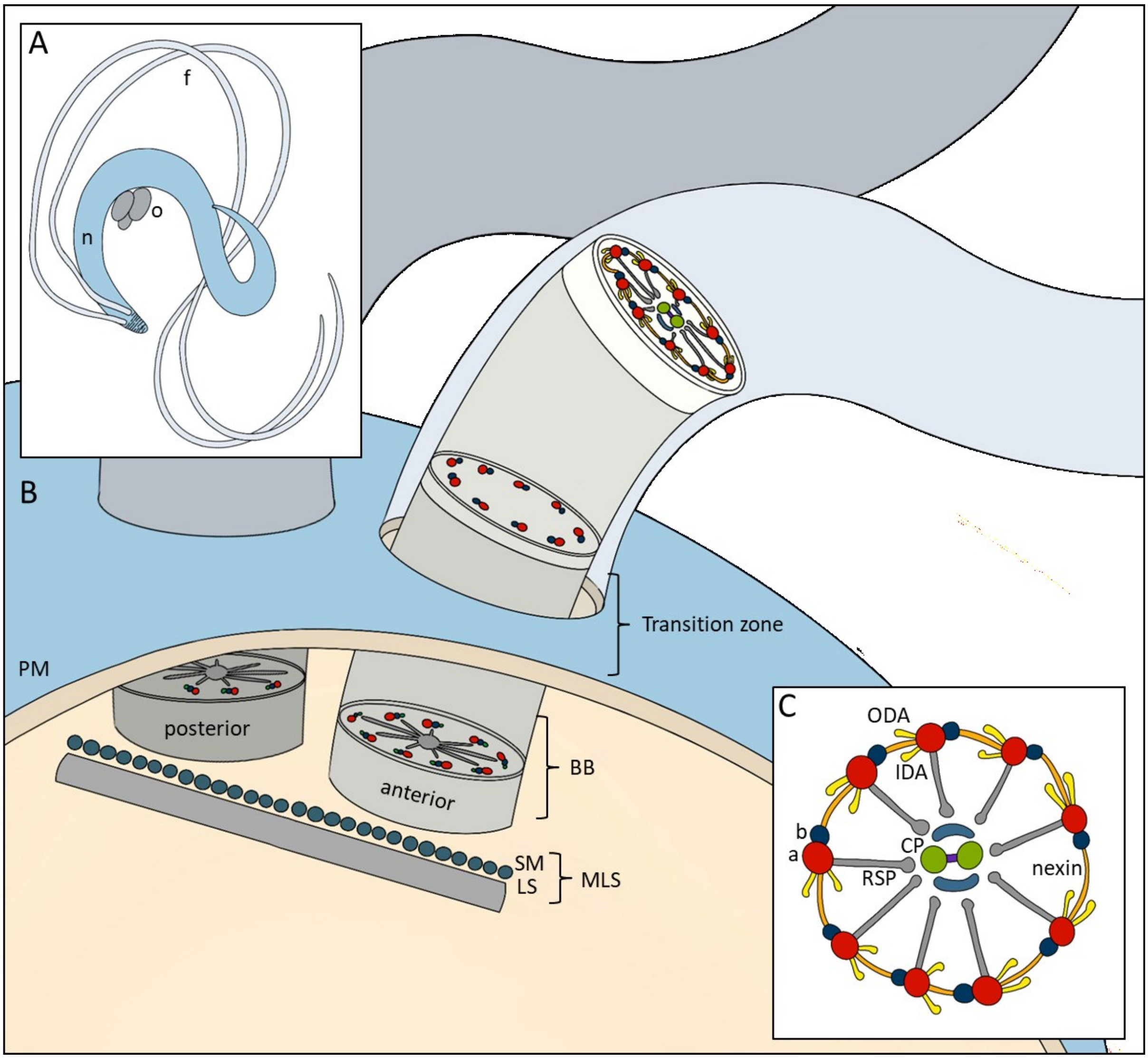


**Figure 4:** Moss spermatozoid morphology and ultrastructure. A: Sinistral coiled spermatozoid with elongated nucleus (n, blue) bearing three organelles, one plastid and two mitochondria (o). A locomotory apparatus is located at its distal end of which two flagella emerge (f). B: The locomotory apparatus consists of a multi-layered structure (MLS) and two dimorphic basal bodies (BB). The MLS comprises a lamellar strip (LS) and overlying spline microtubules (SM). The centriole (not shown) shows nine microtubule (MT) triplets (a: red, b: blue, c: green) arranged around a central core with a stellate pattern. The BBs show dimporphic MT patterns, whereas the posterior BB shows usually three bottom MT triplets and the anterior BB up to nine MT triplets. In the transition zone, a and b MTs are elongated into the transition zone. C: The axoneme shows nine outer MT doublets and a pair of single MTs in the center (CP) which are connected by a bridge (purple). They are also surrounded by protein projections (cyan). The outer MTs are interconnected by nexin (orange) and connect to the CP via radial spoke proteins (RSP, grey). The a MT possess inner arm dyneins (IDA, yellow) and outer arm dyneins (ODA, yellow). Redrawn from (Carvalho-Santos et al., 2011; Renzaglia et al., 2018; Renzaglia and Garbary, 2001; Meyberg et al., 2019).

generate flagellar waveforms (Brokaw and Kamiya, 1987; Kamiya, 1995). With regard to their function it could be shown, that IDAs are linked to the beat pattern whereas ODAs are linked to the beat frequency (Silflow, 2001). In comparison to unicellular flagellated algae, bryophyte spermatozoids seem to be limited in motility time and distance. For *M. polymorpha* it could be shown that spermatozoid motility lasts for ca. 60 minutes, but that spermatozoids where limited in distance and required movement of the surrounding water to reach distances above 2-3 cm (Furuichi and Matsuura, 2016). The dynein setup, namely the occurrence of IDAs and ODAs, is contradictory reported in bryophytes. (Carvalho-Santos *et al.*, 2011) reported only ODAs to be present in *M. polymorpha*, whereas Rensing *et al*., Hodges *et al*., and Wickstead and Gull suggested, that only IDAs and no ODAs are present in bryophytes (Rensing *et al.*, 2008; Hodges *et al.*, 2012; Wickstead and Gull, 2012). Thus, the dynein status of the model moss *P. patens* flagellum, and with it the mechanisms behind motility, will have to be determined in future.

Plant and mammalian male gametes are highly similar and share not only structural but also genetic similarities. In plants as well as in mammals, the chromatin is highly condensed, which is shown to protect the genome against mutations (Rathke *et al.*, 2010). Also, autophagous cytoplasmic reduction mediated through the autophagy related protein 7 (ATG7) is required for male fertility in mice and moss (Sanchez-Vera *et al.*, 2017). Finally, sperm attraction of the female in moss and mice is mediated via glutamate-receptor-like ion-channels (Ortiz-Ramírez *et al.*, 2017). Concluding, several aspects of male fertility in flagellated land plants and mammals probably have their roots in the most recent common ancestor (MRCA) and evolved independently afterwards.

## 6.6 Alternation of generation

The life cycle of land plants is diphasic and alternates between two multicellular generations, the haploid gametophyte and the diploid sporophyte, with fertilization and meiosis constituting the switches in between (Hofmeister, 1851). Streptophytic algae as e.g. the model organism *Chara braunii* have a haplontic life cycle with a dominant and multicellular haploid generation whereas the diploid generation is unicellular and encompasses one cell, the zygote (Nishiyama *et al.*, 2018). Some algae also show a haplo-diplontic lifestyle whereas *Ulva* represents the isomorphic and e.g. *Laminaria* the heteromorphic type (Potter *et al.*, 2016; Liu *et al.*, 2017). In angiosperms, the haploid phase is highly reduced to the male multicellular gametophyte, the pollen grain and the multicellular female gametophyte (Drews and Yadegari, 2002).

The alternation of generation is regulated by several genetic and epigenetic factors e.g. FERTILIZATION-INDEPENDENT ENDOSPERM (FIE) and CURLY LEAF (CLF), genes of the POLYCOMB REPRESSIVE COMPLEX 2 (PRC2) group, which was shown to act via the histone modification H3K27me3 on gene expression of downstream genes in *Drosophila melanogaster* (Nekrasov *et al.*, 2005). In *P. patens* knock-out mutants of FIE and CLF, the development is arrested in the gametophytic phase, inhibiting the alternation of generations. FIE acts on the regulation of stem cell maintenance and affects the expression of the HOMEOBOX KNOX1 transcription factors (TF) mkn2 and mkn5 which are involved in sporophyte development (Katz *et al.*, 2004; Sakakibara *et al.*, 2008; Mosquna *et al.*, 2009). KNOX2 genes could be shown to repress apospory (Sakakibara *et al.*, 2013), leading to the conclusion that KNOX HOMEOBOX TFs act on the repression of the diploid body plan of *P. patens*. In *C. reinhardtii*, it could be shown that zygote development, the first stage of the sporophytic generation, depends on a heterodimer of KNOX and another HOMEOBOX TF BELL (Lee *et al.*, 2008) similar to mammals (Bellaoui *et al.*, 2007; Rensing, 2016). In *P. patens*, expression of knox and bell homeobox genes in the sporophytic stage could be shown, suggesting a putative conserved regulation within mammals and algae (Frank and Scanlon, 2015). Contrary to the knox genes, it could be shown that the overexpression of bell1 leads to apogamy, developing a sporophytic body plan in haploid tissue, bypassing the alternation of generation (Horst *et al.*, 2016). Thus, BELL/KNOX heterodimerization is required for the alternation of generation and for the subsequent body plan formation of plants and is probably conserved between green algae, land plants and possibly mammals (Lee *et al.*, 2008; Rensing, 2016). So far, no involvement of HOMEOBOX TFs in the male germline in plants could be shown yet, but in human, RHOX homeobox genes are correlated with male fertility and controlled via DNA methylation (Richardson *et al.*, 2014).

# 7 Research objectives

The aim of this thesis was to identify and analyze factors which are involved in the sexual reproduction of the model moss *Physcomitrella patens*. In early land plant models, little is known about genes required for the sexual reproduction, which includes gametangia initiation, gametangia development, fertilization, embryo and sporophyte development. To achieve this goal, two approaches were chosen. The first approach made use of the worldwide used laboratory strain Gd, which was reported to show defects in its sexual reproduction, which is thought to be based on the accumulation of somatic (epi-) mutations due to long term vegetative reproduction *in vitro*. Thus, characterization and establishment of the recently collected *P. patens* ecotype Re and its establishment in research was planned as well as comparative analyzes between ecotypes to localize and characterize the Gd impairment. The fertility of different Gd strains, Re and Vx was quantified employing the number of sporophytes developed per gametophore. In addition, gametangia were morphologically analyzed. Genetic and epigenetic differences were determined using SNP and differentially methylated positions (DMP) data. Differences in the gene expression during sexual reproduction were quantified via qPCR and RNA-seq analysis.

The second approach was planned to identify candidate genes required for successful sexual reproduction of *P. patens*. The goal was to use publicly available expression data to identify genes exclusively expressed in different developmental stages and tissues during sexual reproduction. The function during the *P. patens* life cycle resp. sexual reproduction was planned to be determined via a reverse-genetic approach. Knock-out mutants were generated and characterized to eventually get new insights into genes important in early land plant sexual reproduction.

# 8 Publications

## 8.1 Sexual reproduction, sporophyte development and molecular variation in the model moss *Physcomitrella patens*: introducing the ecotype Reute

As previously introduced, several laboratories started to realize that their **long term vegetatively propagated cultures were impaired in sexual reproduction**. In this publication, the more recently collected ecotype Re was introduced and characterized, comparing sexual reproduction, genetic variability and gene expression to the frequently used ecotypes Vx and/or Gd. Expression analysis of different developmental stages involved in the sexual reproduction was performed in order to identify genes important for specific developmental time points and processes. No developmental and obvious differences in morphology or timing could be identified during gametangia development. Quantification of the number of sporophytes developed per gametophore, employed as indirect fertility measurements, revealed **Gd to be nearly self-sterile**. In comparison, **Re and Vx were shown to be highly self-fertile** and developed high numbers of sporophytes per gametophore. Genetic variations, namely the number of base pairs per SNP, of the analyzed ecotypes were determined and showed **Re being genetically 10-fold closer to Gd compared to Vx**, who displays a similar genetic distance to Gd as seen in *A. thaliana* ecotypes, updating previous analyses (Kasahara *et al.*, 2011). Nevertheless, micro array data generated from adult gametophores of Re and Gd showed several TAPs to be DEGs, which might be responsible for the nearly sterile Gd phenotype. Comparative analysis of different sporophyte developmental stages allowed to identify several DEGs representing the morphological changes during sporophyte development, e.g. cell wall modifying proteins and proteins involved in the UV-B response of the sporophyte. This publication now allows the community to use the closely related fertile ecotype Re as a proper replacement for studies, previously performed in Gd. **Re enables the analysis of the sexual reproduction and the resulting sporophyte of the model moss *P. patens*** as well as further work on the characterization of the Gd nearly self-sterile phenotype.

the plant journal

SEB Society for Experimental Biology

*The Plant Journal* (2017) **90**, 606–620 doi: 10.1111/tpj.13501

RESOURCE

# Sexual reproduction, sporophyte development and molecular variation in the model moss *Physcomitrella patens*: introducing the ecotype Reute

**Manuel Hiss[1], Rabea Meyberg[1], Jens Westermann[1,†], Fabian B. Haas[1], Lucas Schneider[1,‡], Mareike Schallenberg-Rüdinger[1,§], Kristian K. Ullrich[1,¶] and Stefan A. Rensing[1,2,*]**

[1] *Plant Cell Biology, Faculty of Biology, University of Marburg, Karl-von-Frisch-Str. 8, 35043, Marburg, Germany,*
[2] *BIOSS Centre for Biological Signaling Studies, University of Freiburg, Freiburg, Germany, and*

Received 1 August 2016; revised 20 January 2017; accepted 24 January 2017; published online 5 February 2017.

*For correspondence (e-mail stefan.rensing@biologie.uni-marburg.de).
†Present address: Biocenter, Botanical Institute, University of Cologne, Zülpicherstr. 47b, 50674, Cologne, 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: IZMB–Institut für Zelluläre und Molekulare Botanik, Abteilung Molekulare Evolution, Universität Bonn, Kirschallee 1, 53115, Bonn, Germany.
¶Present address: Max Planck Institute for Evolutionary Biology, August-Thienemann-Straße 2, 24306, Ploen, Germany.

## SUMMARY

**Rich ecotype collections are used for several plant models to unravel the molecular causes of phenotypic differences, and to investigate the effects of environmental adaption and acclimation. For the model moss *Physcomitrella patens* collections of accessions are available, and have been used for phylogenetic and taxonomic studies, for example, but few have been investigated further for phenotypic differences. Here, we focus on the Reute accession and provide expression profiling and comparative developmental data for several stages of sporophyte development, as well as information on genetic variation via genomic sequencing. We analysed cross-technology and cross-laboratory data to define a confident set of 15 mature sporophyte-specific genes. We find that the standard laboratory strain Gransden produces fewer sporophytes than Reute or Villersexel, although gametangia develop with the same time course and do not show evident morphological differences. Reute exhibits less genetic variation relative to Gransden than Villersexel, yet we found variation between Gransden and Reute in the expression profiles of several genes, as well as variation hot spots and genes that appear to evolve under positive Darwinian selection. We analyzed expression differences between the ecotypes for selected candidate genes in the GRAS transcription factor family, the chalcone synthase family and in genes involved in cell wall modification that are potentially related to phenotypic differences. We confirm that Reute is a *P. patens* ecotype, and suggest its use for reverse-genetics studies that involve progression through the life cycle and multiple generations.**

**Keywords: *Physcomitrella patens*, ecotype, Reute, sporophyte, microarray, single-nucleotide polymorphism, spore.**

## INTRODUCTION

### The model moss

The moss *Physcomitrella patens* belongs to the Funariaceae with type species *Funaria hygrometrica*, which has been used for physiological studies for more than half a century (Bryan, 1957; Krupa, 1967). Whereas *P. patens* has been used for similar studies starting nearly as long ago (Engel, 1968), the last decade has seen the completion of the nuclear genome sequence (Rensing *et al.*, 2008) and the development of a plethora of tools for this organism (Reski and Cove, 2004; Frank *et al.*, 2005; Quatrano *et al.*, 2007; Kamisugi *et al.*, 2008; Lang *et al.*, 2008; Prigge and Bezanilla, 2010). Today, *P. patens* is one of the primary plant models for evolutionary developmental and comparative studies (e.g. Mosquna *et al.*, 2009; Okano *et al.*, 2009; Khandelwal *et al.*, 2010; Sakakibara *et al.*, 2013; Horst *et al.*, 2016), and is also employed to study physiology, genome evolution and homologous recombination (e.g. Rensing *et al.*, 2012; Beike *et al.*, 2014, 2015; Charlot *et al.*, 2014).

### Worldwide accessions

*Physcomitrella* has been described to occur in North America, Europe, Africa, China, Japan and Australia (Frey *et al.*, 2009), and is distributed in the land masses of the Holarctic (Medina *et al.*, 2015). In total, 20 *P. patens* accessions, four *Physcomitrella magdalenae* accessions and 15 *Physcomitrella readeri* accessions have been described, and accessions from all these locations have been cultured axenically *in vitro* (von Stackelberg *et al.*, 2006; Beike *et al.*, 2010, 2014; McDaniel *et al.*, 2010; Medina *et al.*, 2015). The single spore isolated near Gransden Wood (Cambridge, UK) by Whitehouse in 1962 was used initially for *in vitro* culture (Engel, 1968), and became the worldwide laboratory strain *P. patens* Gransden, the genome of which was sequenced by Rensing *et al.* (2008). In addition, the genetically divergent (von Stackelberg *et al.*, 2006) isolate Villersexel K3 (Haute Saône, France; collected by Lüth 2003) was used to generate a genetic map through crossing with Gransden (Kamisugi *et al.*, 2008). The accession Reute was collected by Lüth in 2006 close to Freiburg im Breisgau, Germany, from a moist, disturbed field. Its marker-based genetic distance to Gransden is less than that of Villersexel, and all three accessions can be crossed with each other (von Stackelberg *et al.*, 2006; McDaniel *et al.*, 2010; Perroud *et al.*, 2011; Beike *et al.*, 2014). Such crosses produce viable offspring and have been successfully used to generate a genetic map (Kamisugi *et al.*, 2008) and in forward genetics (Stevenson *et al.*, 2016).

### Sexual reproduction and life cycle

The induction of *P. patens* gametangia development by low temperature is well established, with incubation at 17°C leading to gametangia development within 7–14 days (Engel, 1968; Nakosteen and Hughes, 1978). The additional shortening of day length and reduction in light intensity further increases the frequency of gametangia (female archegonia, male antheridia) formation in Gransden, with optimal laboratory induction conditions being 15°C, an 8-h photoperiod and 20 $\mu$mol m$^{-2}$ s$^{-1}$ (Hohe *et al.*, 2002), mimicking a spring or autumn day. Buds were formed 5–7 days after spore germination, gametophores were formed after 11–13 days, and mature spore capsules were formed 21–28 days after induction on agar (Nakosteen and Hughes, 1978; Ortiz-Ramirez *et al.*, 2015). As antheridia mature earlier than archegonia (Nakosteen and Hughes, 1978), selfing will occur if fertilization by sperm from other plants has failed. *Physcomitrella patens* is known to show a low rate of out-crossing and is predominantly self-fertilising (Perroud *et al.*, 2011), but the haploid moss is able to efficiently purge deleterious mutations (Szovenyi *et al.*, 2014).

Bryophyte spores can survive decades in mud or herbaria (Glime, 2007). For example, spore viability in *F. hygrometrica* was demonstrated after 11 years (Hoffmann, 1970). Moss spore germination can be induced by light, but the quality of the light and the quantity necessary for spore germination is likely to depend on the habitat, e.g. under a canopy versus out in the open, whereas the optimum germination temperature can vary in populations of the same species (Glime, 2007). Spore germination in *P. patens* appears to be suppressed by ultraviolet B (UV-B) irradiation in a dose-dependent manner (Wolf *et al.*, 2010), is completely inhibited by a pulse of far-red light (Possart and Hiltbrunner, 2013) or elevated temperature (Vesty *et al.*, 2016), and depends on phytohormone regulation (Vesty *et al.*, 2016). On agar, spores usually germinate within 3 days (Nakosteen and Hughes, 1978). *Physcomitrella patens* spores are 30 $\mu$m in diameter, and around 8000–16 000 are contained per capsule (Nakosteen and Hughes, 1978).

### Transcriptome analyses

Numerous transcriptomic analyses have been performed, using microarrays based on annotation v1.1 or earlier to analyse ABA, drought stress responses and sporophyte development (Cuming *et al.*, 2007; Komatsu *et al.*, 2009; O'Donoghue *et al.*, 2013), followed by a design based on v1.2 analysing different abiotic stresses and developmental stages (Wolf *et al.*, 2010; Busch *et al.*, 2013; Hiss *et al.*, 2014; Beike *et al.*, 2015). More recently, based on v1.6 (Zimmer *et al.*, 2013), array analyses looked into developmental progression and mutants (Ortiz-Ramirez *et al.*, 2015; Yaari *et al.*, 2015). Although the array designs were based on Gransden, for which the complete genome sequence is available (Rensing *et al.*, 2008), the Villersexel ecotype was successfully hybridized to microarrays (O'Donoghue *et al.*, 2013). With the advances in next-generation sequencing technologies, high-throughput cDNA sequencing (RNA-seq) is now frequently used to measure expression strength and to detect differentially expressed genes (DEGs). In *P. patens*, RNA-seq studies (Table S1) have described the development from protoplasts to protonema, and further on to gametophores (Xiao *et al.*, 2011, 2012), but not yet the development of sporophytes. Recently, the flagship Gene Atlas initiative by the U.S. Department of Energy (DoE), has undertaken deep RNA-seq sequencing to cover the most common developmental stages and perturbations. The moss *P. patens* is one of seven plant 'flagship' organisms (http://jgi.doe.gov/our-science/science-programs/plant-genomics/plant-flagship-genomes/) tackled in the Gene Atlas project.

In this study we have compared the sexual reproduction of the Reute ecotype with that of Gransden and

Villersexel. We provide and compare transcriptome data of sporophyte development, and have performed comparative molecular variant analyses on the three ecotypes. Thus we introduce the Reute ecotype for reverse-genetics studies that involve progression through the life cycle, overcoming problems that many labs face in using Gransden for such studies.

## RESULTS AND DISCUSSION

### Reute natural conditions

The accession Reute was collected close to Freiburg, Germany, by Lüth in 2006 (Figure S1a). In November 2006 and in October 2008 mature sporophytes were found on a field that is ploughed in autumn, typically in September or October, probably thus exposing spore capsules resting in the soil (Beike *et al.*, 2014). Wet conditions (residual water in furrows) combined with shortened day length induce spore germination and subsequently sporophyte development. Average temperatures 5 cm above the soil were found to be 1 ± 3°C from October to December in the years 2006–2015, and in some cases temperatures below 0°C were recorded causing frost coverage (Figure S1b). Daily rainfall was between 0 and 6 mm (Figure S1c), with an average of 2.3 mm per day or 71.9 mm per month. As the field is not shaded, direct sunlight reaches the site (Figure S1b): the light fluence rate in November was measured at $>100$ μmol $m^{-2}$ $s^{-1}$ (with cloud cover) and $>700$ μmol $m^{-2}$ $s^{-1}$ (without cloud cover). The natural conditions of growth and sporophyte/spore development of Reute are thus cold temperatures slightly above freezing, wet environment and short days, albeit with direct sunlight, and thus higher light fluence, including UV-B, than a forest floor moss would experience, for example. Noteably, Gransden (although collected at a similar site from furrows of a ploughed field) would experience different weather conditions than Reute, with both less rain and less sun in October at Gransden Wood than at Reute, for example (Figure S2; Table S2). Although the high light conditions found at the Reute site differ from those used to promote sexual reproduction in the laboratory, it should be noted that day length and lower temperature have a larger impact than light fluence rate (Hohe *et al.*, 2002), which is in line with the weather conditions as found at the Reute site in autumn.

### Gametangia and sporophyte development

A number of laboratories working with the sequenced ‘Gransden 2004’ strain (Rensing *et al.*, 2008) observed a low rate of sporophyte production (Ashton and Raju, 2000; Landberg *et al.*, 2013) and instead used Villersexel to study genome-wide expression patterns during sporophyte development (Landberg *et al.*, 2013; O’Donoghue *et al.*, 2013); however, gametangia and sporophytes do not show any evident morphological differences among Reute, Gransden and Villersexel in axenic *in vitro* culture (gametophore/sporophyte $n = 4908/3700$ for Reute, 2965/191 for Gransden and 1529/1137 for Villersexel).

Archegonia of all three ecotypes (Figure 1) comprise a flask shaped egg cell-containing venter (bracket) and tubular neck canal cells (nc) with no visible aberration. Also, antheridia of these three ecotypes show no visible distinctions, growing in septate bundles at the apex of the gametophore (Figure 1a–c, lower panels). Sporophyte development occurs as described for Gransden (Sakakibara *et al.*, 2008).

The time point at which mature gametangia can be found at the gametophore apex in Reute does not differ from Gransden and Villersexel. Under the growth conditions applied here, most apices of all ecotypes carry mature gametangia at day 19–21 after transfer to short-day conditions. If fully developed gametangia are present, watering leads to synchronized fertilization, as the flagellated antherozoids (spermatozoids/sperm cells) need liquid water in order to swim to the archegonia. At 7–10 days after watering (7–10 daw) a pre-meiotic, elongated and inflated early sporophyte (Figure 1e, ES) is observed. At 9–15 daw spore mother cells (SMCs) are formed in a now spherical, translucent green sporophyte (Figure 1e, PM). SMCs undergo meiosis at around 14–15 daw, and the sporophyte turns from translucent to opaque green (Figure 1e, M). During the ripening process the sporophyte turns from yellow (Figure 1e, Y) to light and dark brown (Figure 1e, LB/B), until mature spores are present around 30 daw. Spores start to develop at an early stage when the cells are still surrounded by a cytoplasmic membrane (Figure 1e, Y), culminating in spore coat-covered mature spores (Figure 1e, B).

Reute demonstrates a sporophyte development frequency that is comparable with Villersexel (Figure 1d; median sporophytes per gametophore in Villersexel, 77%, and Reute, 77%), and is therefore well suited for studies focusing on sexual reproduction, fertilization, embryogenesis and sporophyte development. The sporophyte development frequency of Reute versus Gransden and Villersexel versus Gransden differs significantly, however (Wilcoxon rank tests, $P < 0.01$; median/mean sporophytes per gametophore Reute, 77/80%, Villersexel, 77/80%, and Gransden, 5/6%). While gametangia and sporophytes show no morphological differences and develop at the same rate in all ecotypes, Gransden clearly differs significantly in the number of developed sporophytes compared with Reute and Villersexel. Such phenotypic differences, along with their distinct geographic location and their ability to interbreed are hallmarks of distinct ecotypes.

There are several possible explanations for the observed differences. There may be genetic or epigenetic differences between Gransden and Villersexel that account for this. As

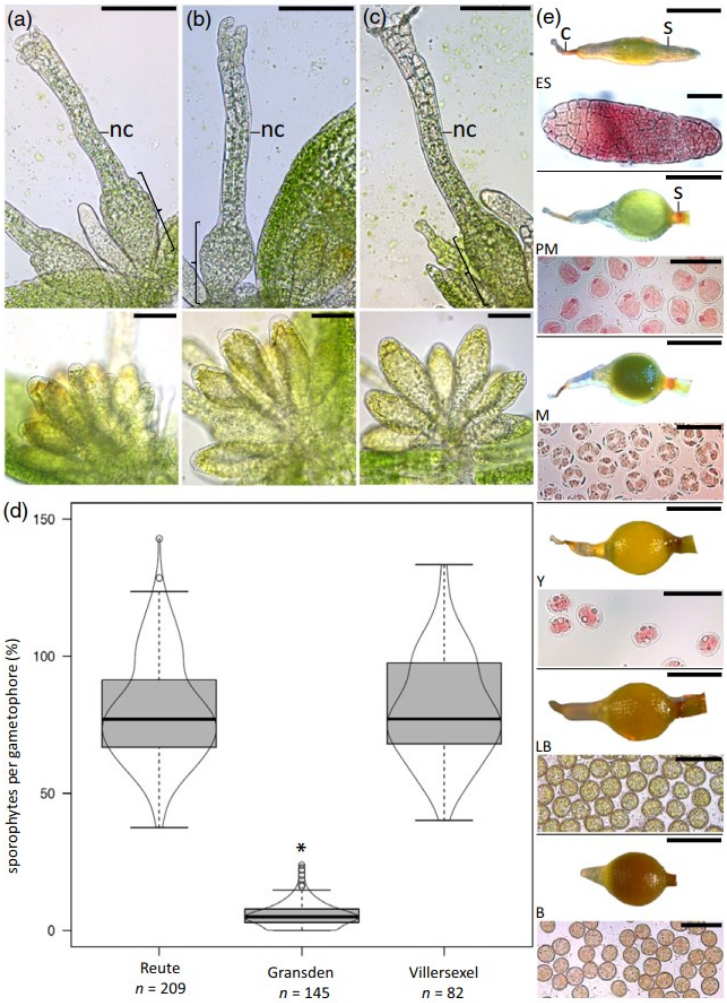


**Figure 1.** Gametangia and sporophyte development. Gametangia (female archegonia and male antheridia) of *Physcomitrella patens* ecotypes Reute (a), Gransden (b) and Villersexel (c). Upper panel shows mature archegonia, consisting of flask-shaped egg-containing venter (bracket) and tubular neck canal cells (nc). Scale bars: 100 μm. Lower panel shows antheridia, occurring in septate bundles at the gametophore apex. Scale bars: 50 μm. (d) Box plot of average number of sporophytes per gametophore (*n* = number of plants) as a percentage. The plot is median-centred, with the grey box representing 50% of the measurements. The whiskers end with the last value in the 1.5 interquartile range (IQR). The edged area shows the distribution of measurements. Sporophyte development of Reute median: 77% versus Gransden 5%. Sporophyte development of Villersexel median: 77% versus Gransden 5%. Differences are significant (Wilcoxon rank test, $P < 0.01$; marked by asterisk), whereas Reute and Villersexel show a comparable proportion of sporophytes (Wilcoxon rank test, $P = 0.84$). (e) Reute sporophyte developmental stages (scale bars: 500 μm), with corresponding spore stages, stained with acetocarmine (scale bars: 50 μm). ES (early sporophyte): elongated premeiotic sporophyte with calyptra (C) and developing seta (S). PM (premeiotic): spherically shaped premeiotic translucent green sporophyte containing spore mother cells, with developed seta (S, brown area). M (meiotic): postmeiotic opaque green sporophyte, cellular content shows spore mother cells with tetrads after metaphase II of meiosis. Y: yellow sporophyte, including ripening spore mother cells. LB: light-brown sporophyte with spores surrounded by a visible spore coat. B: mature brown sporophyte without calyptra, containing mature spores.

only the frequency of sporophyte development appears different there could be a failure of fertilization: either spermatozoids might not be released from Gransden antheridia or they might be less motile. Archegonial development might be affected: during ripening, the archegonial tip cell and inner canal cells degrade (Landberg *et al.*, 2013) to free the way for the entering spermatozoids, a developmental step that might be disrupted. Fertilization or early development of the zygote could be aberrant. A defect during later embryogenesis can be excluded, as no late-stage aborted embryos or sporophytes could be observed. Although the nature of the difference is not the focus of this study, future research might point out why the frequency of sporophyte development differs.

It was suggested that prolonged vegetative cultivation may be the cause for a loss in fertility (Ashton and Raju, 2000), but conditions for sporophyte induction (vessels, substrates, light, temperature) do vary between labs. The conditions used here are adopted from those originally established for Gransden (Hohe *et al.*, 2002).

### Analysis of Reute gametangia and sporophyte development

To address the possibilities described above, we conducted a more detailed developmental analysis. For the Reute ecotype, as is generally known for *P. patens* (Landberg *et al.*, 2013), immature and mature archegonia can be distinguished via the opening of the archegonial tip (Figure 2a,i). This enables spermatozoids to enter and reach the egg cell in the archegonial venter (Figure 2d, arrowhead). During growth and ripening, antheridia undergo an increase of cell size and change color from green (immature; Figure 2b,e) via yellow (mature) to brown (post release; Figure 2c,h). Water is required for fertilization to occur, not only as the transport medium for the flagellated spermatozoids, but also for their release, as the water is taken up by the antheridial tip cells, causing them to swell and finally burst to release spermatozoids (Figure 2c,h, arrowheads). Figure 2(e) shows an early antheridial stage, at which anticlinal cell division can be observed. In the mature archegonium the elongated outer neck canal cells can be seen clearly (Figure 2f, light green, arrowhead), as well as a paraphysis (a sterile organ consisting of elongated cells with a swollen apical cell; Figure 2f,i, blue). In Figure 2(g) a sporophyte with detached calyptra is shown from the top, and the distribution of the outer sporophytic cells can be observed with several cells having recently divided. In opened sporophytes the cellular content can be observed (Figure 2j, ochre). In summary, the detailed analysis of Reute gametangia and sporophyte development confirms its similarity with that of Gransden, including the presence of paraphyses and details of archegonial/antheridial growth (Landberg *et al*., 2013).

### Expression differences between Gransden and Reute gametophores

To determine whether – despite similar development – there are differences in gene expression, we analysed expression profiles. To find differences in gene expression between Gransden and Reute, Hiss *et al*. (2014) performed a whole transcriptome microarray analysis of gametophores with developed gametangia (adult gametophores) for both ecotypes. Here, we analyze these data and find 262 DEGs (Appendix S2), 250 of which were found to show lower and 12 of which were found to show higher expression in the Reute ecotype, when compared with Gransden. Of particular interest are transcription factors (TFs) and transcriptional regulators (TRs) that are differentially expressed between the two ecotypes, as these may underlie phenotypic differences such as the sporophyte development frequency. We find 10 such proteins (Table 1), with PPM3 (Pp3c14_22180V3.1), a member of the moss-specific MADS-box containing subfamily MIKC* (Barker and Ashton, 2013), among them. Members of this subfamily regulate pollen development in *Arabidopsis thaliana* (Gramzow and Theissen, 2010), and therefore could be involved in the development of moss spores, which represent a developmentally analogous structure (Brown and Lemmon, 2011; Daku *et al*., 2016; Vesty *et al*., 2016).

We also find a member of the GRAS TF family among the DEGs. The proteins of this family share the GRAS DNA-binding domain (Li *et al*., 2016), whereas the N-

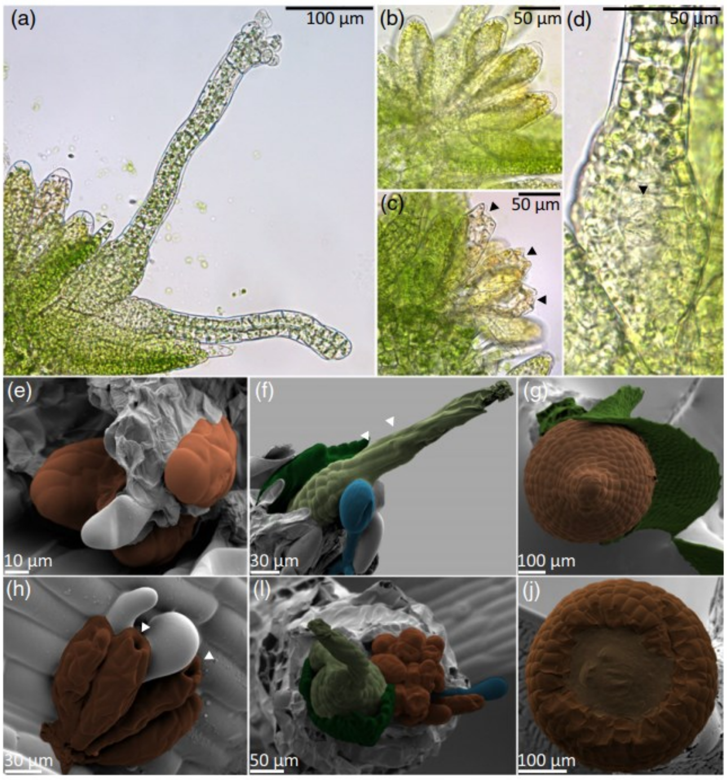


**Figure 2.** Gametangia and sporophytes of *Physcomitrella patens* ecotype Reute. (a) Immature (closed tip) and mature (open tip) archegonia. Immature antheridia (b) and mature antheridia (c), which release spermatozoids through the burst tip cells (arrowheads). (d) Unfertilized egg cell in archegonial venter (arrowhead). (e, f) False-colored cryo-SEM images: (e) early antheridia showing anticlinal cell division; (f) mature archegonium (light green) with elongated outer neck cells (arrowhead), paraphyse (blue) and young phyllid (green). (g) Mature sporophyte, top view (sporophyte brown, phyllid green). (h) Mature antheridia after spermatozoid release through burst tip cell (arrowheads). (i) Gametophore apex, top view with young phyllid (green), mature archegonia (light green), antheridia (brown) and paraphysis (blue). (j) Sporophyte opened at the tip, showing cellular content (ochre).

**Table 1** ID numbers and annotation of 10 transcription factors or transcriptional regulators expressed at a lower level in Reute, as compared with Gransden

| CGI v3 | TF/TR family | Fold change |
|---|---|---|
| Pp3c14_22180V3.1 | MADS | 4.6 |
| Pp3c13_3830V3.1 | AP2/EREBP | 6.7 |
| Pp3c15_3180V3.1 | C2C2_Dof | 4.7 |
| Pp3c1_35770V3.1 | Argonaute | 5.2 |
| Pp3c8_230V3.1 | Zinc finger, AN1 and A20 type | 5.0 |
| Pp3c4_12970V3.1 | ARF | 6.3 |
| Pp3c11_21140V3.1 | GARP_G2-like | 4.4 |
| Pp3c2_20930V3.1 | GRAS | 4.2 |
| Pp3c2_8880V3.1 | bHLH | 4.6 |
| Pp3c11_20170V3.1 | bHLH | 8.5 |

Based on Combimatrix microarray data comparison between adult gametophores (bearing gametangia) of the Gransden and Reute ecotypes.

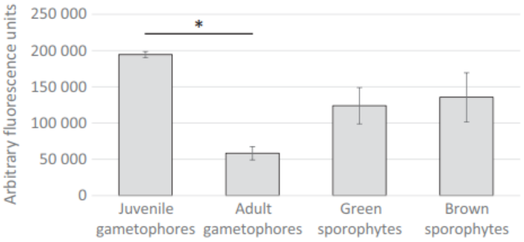


**Figure 3.** Bar chart of GRAS Pp3c2_20930V3.1 microarray expression values. Four different *Physcomitrella patens* Reute developmental stages for the gene *Pp3c2_20930V3.1* are shown based on novel v1.6 NimbleGen microarray data. Error bars show standard deviations for two or three biological replicates. *Significant difference.

terminal parts are predicted to contain molecular recognition features (Morfs) that are important for protein–protein interactions (Sun *et al.*, 2011). GRAS family members are responsible for regulating different plant growth and development steps (Bolle, 2004), and some activate meiosis-specific genes (Morohashi *et al.*, 2003). The intron-less GRAS gene Pp3c2_20930V3.1, encoding a 770 amino acid protein, is expressed at a fourfold lower level in Reute adult gametophores than in Gransden (Figure S3; Table 1). In the Gransden ecotype, Pp3c2_20930V3.1 is more strongly expressed in protoplasts and under UV-B treatment (Figure S4), but does not show reduced expression during the development of sexual organs (i.e. in adult gametophores). Variation in expression can also be seen in the novel Reute array data presented here (Figure 3). Interestingly, the pronounced decrease in expression of this gene in Reute during the development of sexual organs is not visible to this extent in Gransden. Differences between the ecotypes with regards to the frequency of sporophyte development might thus be associated with this GRAS TFs.

In summary, 10 TFs/TRs are differentially regulated between the ecotypes and are good candidates for investigating the differences in frequency of sporophyte development.

### Expression profiling of Reute developmental stages

Upstream regulators such as TFs/TRs often control downstream effector genes that execute the actual phenotypic alterations: here, we focus on such output genes. As we were interested in the sporophyte development of Reute, we generated NimbleGen microarray data for different stages of development: gametophores without gametangia (juvenile), with gametangia (adult), and green sporophytes (developing, pre-meiotic; PM in Figure 2) as well as brown sporophytes (mature, post-meiotic; B in Figure 2). DEGs were computed in pairwise fashion along the developmental progression (Figure 4). We find a high number of DEGs between juvenile and adult gametophores (6021), and also between adult gametophores and green sporophytes (2492). Between green and brown sporophytes we find 313 DEGs.

We focused on genes showing differential expression between adult gametophores and green sporophytes in the Reute ecotype, as this developmental step seems to be affected in the Gransden ecotype. We specifically examined only genes showing differences between the Gransden and Reute ecotypes. We identified 41 genes, 36 of which are expressed at a lower level in the Reute ecotype relative to Gransden (Appendix S2). For selected genes we confirmed the expression profile during sporophyte development by qPCR (Figure 5).

Two of the 10 differentially expressed TFs/TRs between Gransden and Reute (MADS, Pp3c14_22180V3.1; AP2/EREBP, Pp3c13_3830V3.1) also show differential expression during sporophyte development. Aside from that, the list contains genes that are predicted to code for cell wall-modifying enzymes like pectin methylesterase (Pp3c5_23400V3.1) and xylosyltransferase (Pp3c23_380V3.1). Both genes show a stronger expression in adult gametophores than in green and brown sporophytes, suggesting that their products are more active during the late gametophytic c stage. As the development of the sporophyte involves cell wall restructuring (O'Donoghue *et al.*, 2013), the observed expression differences may contribute to the phenotypic differences.

We further find three DEGs that belong to the chalcone synthase (CHS) gene family (Figure S4), namely Pp3c2_32400V3.1 (CHS1a), Pp3c2_32960V3.2 (CHS10 PpASCL; Figure 6) and Pp3c11_2990V1.1 (CHS4; see Table S5 for ID). Chalcone synthases (CHS) catalyze one of the first steps of flavonoid biosynthesis, and are encoded by an expanded gene family in *P. patens* (Koduri *et al.*, 2010; Wolf *et al.*, 2010). The CHS genes Pp3c2_30620V3.1 (CHS01), Pp3c2_32400V3.1 (CHS1a) and Pp3c2_32320V3.1

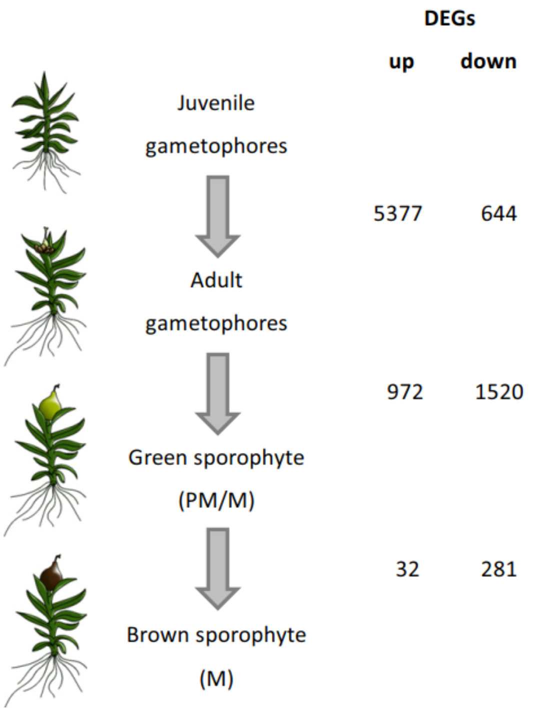


**Figure 4.** Scheme of *Physcomitrella patens* developmental stages and differential gene expression. Left, developmental stages for which NimbleGen microarray data were generated (Reute ecotype). Gametophores were harvested without rhizoids, PM/M and M sporophytes (cf. Figure 2) were separated from the gametophores at harvest. Right, differentially expressed gene (DEG) summary, showing the number of significant DEGs (CyberT test, Benjamini–Hochberg corrected $P < 0.05$; see Experimental procedures for details) that are expressed at higher or lower level between the developmental stages shown, based on the NimbleGen microarray data.

(CHS1c; cf. Figure S6) were found to be induced after UV-B treatment, and are suggested to function as a molecular sunscreen (Wolf *et al.*, 2010); all three genes are less expressed under drought (Stevenson *et al.*, 2016). Here, we find CHS1a to be repressed in brown sporophytes, with a higher expression in all other stages analyzed, namely juvenile gametophores, adult gametophores and green sporophytes (Figure 6). In contrast to CHS1a, CHS10 is induced in green sporophytes as compared with the other three developmental stages, which is also supported by the qPCR analysis (Figure 5). This gene is a functional ortholog of the *A. thaliana* type-III polyketide synthase A (PKSA) belonging to the anther-specific chalcone synthase-like (ASCL) genes, and has been shown to be part of the sporopollenin biosynthesis pathway in *P. patens* (Colpitts *et al.*, 2011; Daku *et al.*, 2016). The repression of the UV-B induced CHS1a/Pp3c2_32400V3.1 accompanied by the induction of the spore coat formation gene Pp3c2_32960V3.2 suggests that biosynthesis of UV-B absorbing quercetin and related flavonoids (Wolf *et al.*, 2010) is no longer needed once sporopollenin is formed. At the same time, the lower expression of CHS1a might be associated with dehydration of the maturing sporophyte.

### Genetic variation among Villersexel, Reute and Gransden

To determine the potential genetic basis for the observed expression and phenotypic differences, we analysed genetic variation of the ecotypes using novel Reute genomic DNA data. Reute and Villersexel can be crossed with each other and with Gransden; however, based on selected markers the genetic distance between Villersexel and Gransden is much greater than between Reute and Gransden (McDaniel *et al.*, 2010). Among different European accessions Villersexel appears most genetically divergent from Gransden (Kamisugi *et al.*, 2008). With the Reute ecotype we present an alternative ecotype with a genetically closer Gransden, yet suitable for both ‘forward’ (map-based) and ‘reverse’ genetics approaches. The lower number of polymorphisms makes reverse-genetics approaches based on the Gransden reference genome easier. We sequenced genomic DNA from Reute gametophores to assess the precise genetic distance by evaluating all single-nucleotide polymorphisms (SNPs) and indels. Comparison of Reute genomic DNA (gDNA) sequence data with the Gransden reference genome identified 264 782 SNPs and 16 292 indels (7874 insertions; 8418 deletions), resulting in a polymorphism density of one SNP every 1783 bases and one indel every 28 857 bases. For Villersexel we find 2 497 294 SNPs and 172 833 indels (77 522 insertions; 95 311 deletions), resulting in a density of one SNP every 188 bases and one indel every 2724 bases. SNP densities between *A. thaliana* ecotypes have been shown to occur between one SNP per 149 bp and one SNP per 285 bp (Cao *et al.*, 2011), similar to the densities found in Villersexel, which is surprising given that the rate of mutation fixation is lower in *P. patens* (Rensing *et al.*, 2007). Reute exhibits an almost 10-fold lower SNP density than Villersexel, although the two collection sites are only 100 km apart geographically, with a negligible difference in latitude, but separated by a mountain range. Reute is found on a field that is regularly plowed in autumn, and Villersexel is found at a dried fish pond, a location that is also regularly flooded. Environmental/microclimatic differences at the two sites might differ, however. The discrepancy between genetic and geographical distance may be explained by the distribution of *P. patens* spores via migrating birds that has been proposed recently (Beike *et al.*, 2014).

We find that most SNPs and indels fall into intergenic regions (about 80%, see Table 2), and into the 2000 bp upstream and downstream of the transcript, consisting of untranslated regions (UTRs) and potential promoter areas (about 15%). Out of the 35 302 genes predicted, about half

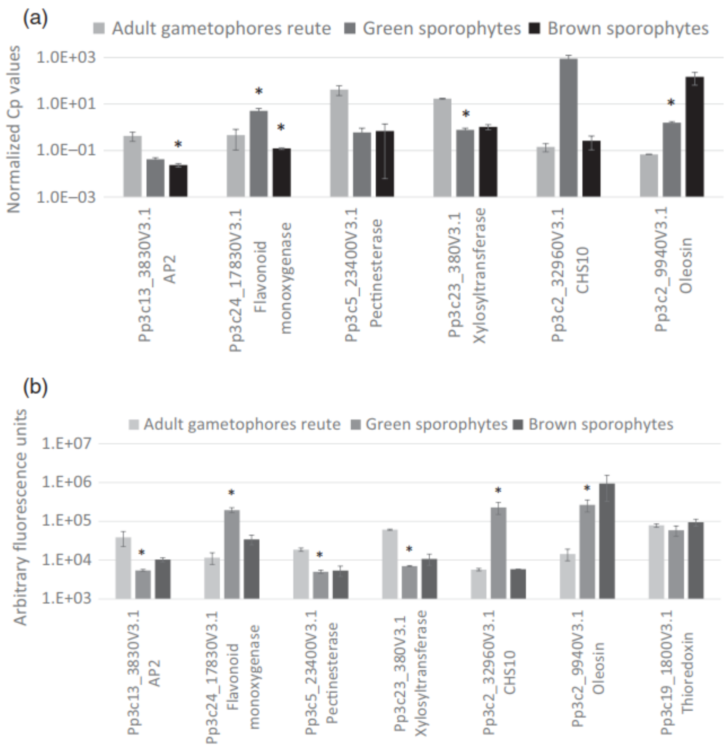


**Figure 5.** Bar chart of expression values derived from qPCR (a) and microarray (b) analysis. qPCR expression values are normalized to the reference gene *Pp3c19_1800V3.1*, which shows a steady expression in the microarray data across the measured tissues. *Significant changes, compared with the preceding developmental stage.

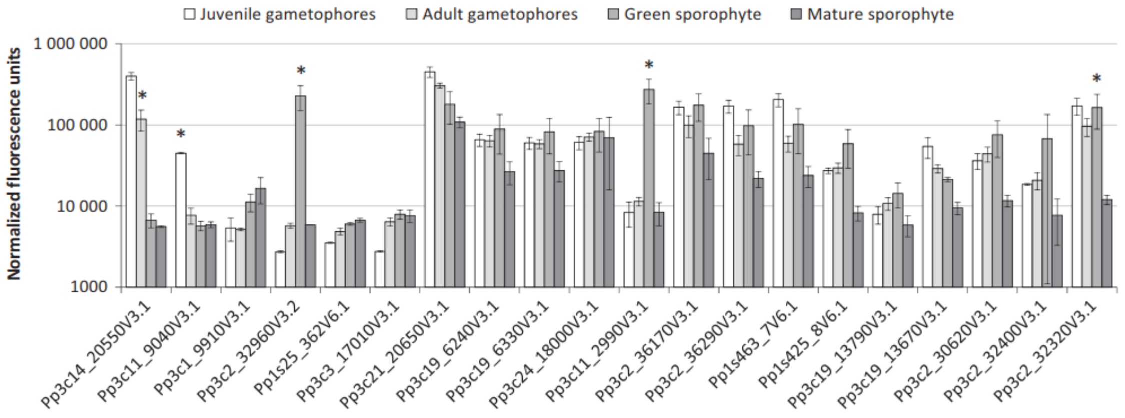


**Figure 6.** Bar chart of microarray expression values (arbitrary units) for 20 chalcone synthase (CHS) genes. Four *Physcomitrella patens* Reute developmental stages are shown, with novel data based on the v1.6 NimbleGen array (corresponding v3.3 gene IDs are shown, except for cases where no v3.3 model was available). Error bars indicate standard deviations of two or three biological replicates. *Significant changes, compared with the following developmental stage (FDR; corrected CyberT test $P < 0.05$).

(15 178) had an SNP within their gene body or promoter in Reute, and 5842 had an indel. In Villersexel almost all (32 473) genes contained an SNP and 27 722 contained an indel. Most SNPs in Reute as well as in Villersexel cause a mis-sense (63%) or a silent (35%) mutation in the coding sequence, and only a few cause a non-sense mutation (2%). Premature stop codons were introduced into 74 genes in Reute (Table S5a) versus 602 genes in Villersexel

by SNPs, and into six genes in Reute (Table S5b) versus 42 genes in Villersexel by an indel. We find 24 genes in Reute and 190 genes in Villersexel that are longer than in Gransden, and therefore may contain a premature stop codon in Gransden.

The Reute SNPs and indels are unevenly distributed on the chromosomes, with several areas of higher SNP density being evident (Figure 7). We selected areas that show significantly higher SNP density (false discovery rate (FDR) corrected $P < 0.01$, see Experimental procedures for details; Table S6), and chose the longest two regions, located on chromosomes 8 and 19, for closer inspection. Within the 1-Mbp peak region on chromosome 8 we find 21 gene models, three of which contain SNPs in their coding sequence; for the 1.7-Mbp peak on chromosome 19 we find 72 gene models, 28 with an SNP in their coding sequence (CDS). For six gene models of the chromosome-8 peak and 11 gene models of the chromosome-19 peak we find close paralogs (BLAST hits with ≥90% identity and length ≥90 aa), often on the same chromosome. Such paralogs, some of them tandemly arrayed genes, can help to provide higher gene-product dosage and might be subject to concerted evolution by gene conversion, in which one copy is 'overwritten' by homologous recombination using the other copy as a template (Wang and Paterson, 2011; Wang *et al.*, 2011). Interestingly, in Reute as well as in Gransden we find three CHS pairs that show identical protein sequences and are located close to their respective partner on chromosomes 2 and 19 (Pp3c2_32400V3.3/CHS1a and Pp3c2_32320V3.3/CHS1c, Pp3c2_36170V3.3/CHS2b.1 and Pp3c2_36290V3.3/CHS2c, Pp3c19_6320V3.3/CHS3.1 and Pp3c19_6330V3.3/CHS3.3), potentially providing higher gene dosage via concerted evolution.

Gene ontology (GO) bias analysis of the genes in the peak regions versus all genes finds diverse GO terms over-represented in the 21 gene models from chromosome 8, e.g. thioredoxin biosynthesis, mRNA processing and superoxide responses. The 72 gene models on chromosome 19 show an over-representation of genes, e.g. involved in cyanate biosynthesis and mitochondrial electron transport (for a full list of GO terms and gene IDs see Tables S7 and S8). Hence, many of the genes in the two SNP hot spots are potentially involved in radical scavenging. This could be an adaptation to environmental conditions that are characterized by higher levels of photonic radiation, as a result of the more southerly latitude and less average cloud cover.

The late embryogenesis abundant 1 (*LEA1*) gene Pp3c22_8970V3.2 contains a premature stop codon in the first exon. It is expressed, but the gene product does not seem to be necessary for normal growth and development (Kamisugi and Cuming, 2005). The SNP causing the premature stop codon is present in Reute, but not in Villersexel, where a CAG is present, demonstrating again that Reute is genetically closer to Gransden than Villersexel, and that genetic variation of this particular gene varies among ecotypes.

### Reute pseudoreference genome and genes under selection

In order to determine genes under selection, and to make Reute more useful for the community, Reute SNP and deletion data were incorporated into the Gransden reference genome sequence to create a 'pseudogenomic sequence' that can be used for Basic Local Alignment Search Tool (BLAST) searches, phylogenetic analysis or planning of reverse-genetics experiments (available at http://plantco.de/ReutePseudogenome.fa). From the pseudogenome we calculated the ratio of non-synonymous/synonymous ($K_a/K_s$) substitutions between Gransden and Reute, which was possible for 320 genes, and focused on the top and bottom 5% (16 genes each). In the top 5% we found 15 genes showing a $K_a/K_s$ ratio larger than 2 (Table S9), and

**Table 2** Summary table of SNP effect analysis for *Physcomitrella patens* ecotypes Reute and Villersexel, as compared with Gransden

| | Reute – SNPs | | Reute – Indels | | Villersexel – SNPs | | Villersexel – Indels | |
|---|---|---|---|---|---|---|---|---|
| Type | Count | Percent | Count | Percent | Count | Percent | Count | Percent |
| Intergenic | 249 609 | 82.6 | 13 560 | 59.0 | 2 355 783 | 81.1 | 155 247 | 63.0 |
| Upstream | 18 628 | 6.17 | 3729 | 16.2 | 200 512 | 6.91 | 39 690 | 16.1 |
| Downstream | 17 753 | 5.88 | 2835 | 12.3 | 196 593 | 6.77 | 32 902 | 13.4 |
| Intron | 5132 | 1.70 | 1139 | 4.96 | 50 725 | 1.75 | 7868 | 3.19 |
| Exon | 4542 | 1.50 | 381 | 1.66 | 38 405 | 1.32 | 1673 | 0.68 |
| 5′-UTR | 2508 | 0.83 | 507 | 2.21 | 22 546 | 0.78 | 3292 | 1.34 |
| 3′-UTR | 2272 | 0.75 | 487 | 2.12 | 22 994 | 0.79 | 3280 | 1.33 |
| Other | 1683 | 0.56 | 337 | 1.47 | 16 022 | 0.55 | 2356 | 0.96 |
| Sum | 302 127 | | 22 975 | | 2 903 580 | | 246 308 | |

The type column lists the possible locations of SNPs, both upstream and downstream, constituting the 2000-bp regions in front and behind of the transcript sequence and intergenic the region between the genes (excluding up- and downstream regions). Transcript regions were assigned according to the annotation version 3.1 from http://www.cosmoss.org. For each ecotype the count and percentage among all the locations is listed. The last row shows the sum of all effects. The analysis was performed with SnpEff and the type 'Other' summarizes 'none', 'splice site acceptor', 'splice site donor', 'splice site region' and 'transcript'.

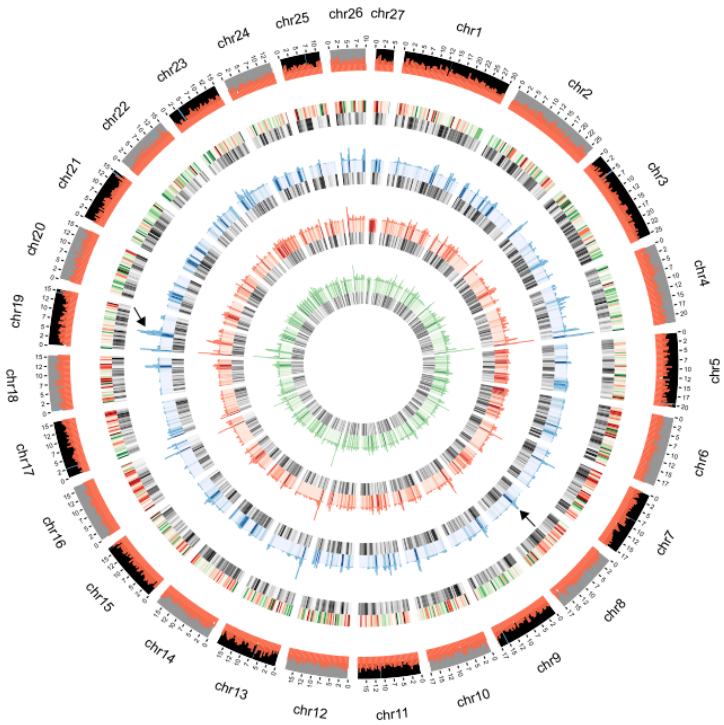


**Figure 7.** Circos plot showing the 27 *P. patens* chromosomes. From outer to inner ring: (i) average gene expression (log2) based on the novel NimbleGen microarray data of four Reute developmental stages (red histogram); (ii) 0–1, normalized gene expression (log2) based on the NimbleGen data, average of four developmental stages (bars in red showing higher expression with values closer to 1 and bars in green showing lower expression with values closer to 0); in the same ring 0–1 normalized gene density is shown (also shown for comparison in the four inner rings, with gray bars of darker color depicting higher density); (iii) SNP density histogram (blue); (iv) deletion density (orange bars); (v) insertion density (green bars). All values were summarized by a sliding window approach with a window size of 500 kbp. Black arrows indicate the two longest high-density SNP peaks on chromosomes 8 and 19.

therefore these genes could be candidates to evolve under positive Darwinian selection.

In conclusion, Reute shows several SNP hot spots that also contain genes with changed coding sequences. Based on the two ecotypes, a number of Reute genes might be subject to positive (Darwinian) selection, which in turn could reflect adaptation to a slightly different niche.

### Determination of a robust set of genes differentially expressed in the *P. patens* sporophyte

Comparison of cross-platform and cross-ecotype data is notoriously difficult. In order to learn whether there is a robust set of genes that are differentially expressed in sporophytes, we analysed all available data sets. Gene expression for green and brown sporophytes from Reute was measured via microarray in this study and via RNA-seq by the JGI Gene Atlas project (http://jgi.doe.gov/our-science/science-programs/plant-genomics/plant-flagship-genomes/). We compared the DEGs derived from both technologies between green and mature sporophytes (Figure 1e, PM/M and B; Appendix S2). For genes with higher expression in mature sporophytes we found 32 genes via the array and 526 genes via RNA-seq data, 15 of which overlap. For genes with lower expression, the array detects 281 and RNA-seq detects 1030 genes, with an overlap of 129. GO annotation for the 15 more highly expressed genes in the overlap shows over-represented categories including 'reproductive structure development' and 'post-embryonic development' (Figure S5), in accordance with expectations. The fraction of DEGs that overlap between the two platforms is comparable with other studies; in general some technology-specific differences identified by microarray and RNA-seq analyses are common (Marioni *et al.*, 2008; Zhao *et al.*, 2014). Nevertheless, combining results obtained with different technologies results in a high-confidence set of genes involved in the examined developmental stage or perturbation. In summary, we have defined a robust set of DEGs during Reute sporophyte maturation.

### Conclusions

In comparing the *P. patens* ecotypes Gransden, Reute and Villersexel we observe a ~15-fold difference in the number of sporophytes that develop under standardized conditions, with Gransden showing just a few, compared with Reute and Villersexel. The reduced Gransden rate may, however, not represent the natural trait, as Gransden has been cultivated under lab conditions for longer time periods than Reute and Villersexel. We do not observe any gross morphological differences during gametangia and sporophyte development; however, several hundred genes are differentially expressed between Gransden and Reute, including transcription factors that are known to control developmental processes, members of the CHS family potentially involved in spore (coat) development, and genes encoding cell wall modification enzymes.

We provide a detailed description of gametangia and sporophyte development, and show Reute to be in accordance with previous descriptions of *P. patens*; however, the Reute genome contains hot spots of genetic variation as well as genes under positive selection.

The Reute ecotype thus combines the advantages of high fertility for forward genetics with the capacity for gene targeting of *P. patens*, enabling studies of the whole life cycle. The Reute ecotype is already used by several labs, and has successfully been used for transient and stable transfections. To facilitate these applications we provide a pseudogenomic sequence as well as a set of expression profiling data. Using cross-platform data we define a confident set of 15 genes expressed in mature sporophytes that can be used, for example as expression markers.

## EXPERIMENTAL PROCEDURES

### Plant material

*Physcomitrella patens* ecotypes Gransden (Rensing *et al.*, 2008), Reute and Villersexel (von Stackelberg *et al.*, 2006) were cultivated on solidified [1% (w/v) agar] mineral medium (Knop's medium; Knop, 1868), on 9-cm Petri dishes enclosed by laboratory film, and maintained at 22°C with a 16-h light/8-h dark regime under 70 μmol m$^{-2}$ s$^{-1}$ white light (long-day conditions), as previously described (Hiss *et al.*, 2014). All ecotypes are available at the international moss stock center (IMSC, http://www.moss-stock-center.org) or from the authors upon request. For sporophyte induction, Petri dishes were transferred to 16°C with an 8-h dark/16-h light regime under 20 μmol m$^{-2}$ s$^{-1}$ white light (short-day conditions), as described by Hohe *et al.* (2002), but using medium without supplements. For sporophyte production, 10 gametophores per Petri dish were evenly distributed into the agar and grown under long-day conditions. After moving the plates to short-day conditions, plants were assessed for gametangia appearance and subsequently watered with sterile tap water (Hohe *et al.*, 2002). As fertilization requires water, this procedure ensures a high rate of synchronization of sporophyte development. Fertilization does not regularly occur under the conditions applied here (including eightfold air exchange per hour and the use of laboratory film to wrap the dishes, allowing gas exchange), as not much condensing water drops onto the gametophores.

### Counting sporophytes per gametophore and statistical analysis

To determine sporophyte development rates, a minimum of five replicate plates were set up as described above for each ecotype. After a minimum of 30 days after watering (30 daw), sporophytes per gametophore were counted (all developmental stages that could be clearly determined as a sporophyte were taken into account) and summarized per plant. Wilcoxon rank tests were performed in R (R Development Core Team, 2008) using the function 'Wilcox.test'. Box plots with added distribution of measurements were generated in R using the function 'boxplot' and the additional package 'caroline', with its function 'violins' (Schruth, 2012).

### Acetocarmine staining

Staining was performed using the method described by Belling (1921). Briefly, acetocarmine staining solution was prepared with 350 ml acetic acid, 650 ml water and 20 g acetocarmine, boiled until completely dissolved, filtered and stored in the dark at room temperature 20-24°C. Sporophytes were fixed for a minimum of 24 h in ethanol/acetic acid (3:1). To stain, fixed tissue was transferred to a microscope slide and squashed to release sporophyte contents; embryonic content was prepared manually using a binocular microscope and forceps. A few drops of staining solution were added and stained for 10 min before image analysis.

### Microscopic imaging of gametangia and sporophytes

The preparation of gametangia was performed using a binocular microscope (S8Apo with MC170HD camera; Leica, http://www.leica.com). Microscopic images were taken with an upright DM6000 microscope (Leica; camera DFC295). Microscopy pictures were processed using Photoshop CC (Adobe Systems Software Ireland Ltd). The brightness and contrast of light microscopy pictures was adjusted, and cryo-SEM images were false colored for enhanced visibility.

### Cryo-SEM analysis of gametangia and sporophyte

For analysis a Philips XL30 ESEM with Cryo Preparation Unit Gatan Alto 2500 was used. Prepared plant material was applied to the specimen holder with freeze-hardening glue and biological samples were preserved by fast-freezing in liquid nitrogen. Afterwards the specimen holder was inserted into the sputter chamber and coated with gold.

### DNA isolation

Genomic DNA was isolated from plant material according to a modified Dellaporta protocol (Dellaporta *et al.*, 1983). After the isopropanol precipitation the dry pellet was dissolved in 700 μl TE buffer (pH 8), 1–3 μl RNaseA (10 mg ml$^{-1}$) was added and incubated for 10 min at 37°C. To purify the DNA, 600 μl phenol/chloroform 1:1 was added, mixed, centrifuged at 10 000 *g* for 1 min and the aqueous phase extracted. To this phase 600 μl chloroform/isoamylalcohol 24:1 was added, mixed, centrifuged at 10 000 *g* for 1 min and the aqueous phase extracted. To precipitate the DNA, 70 μl 3 M Na-acetate and 500 μl isopropanol were added, mixed and centrifuged at 10 000 *g* for 10 min. The pellet was washed with 1 ml 70% ethanol, dried and subsequently dissolved in deionized water. Concentration and quality was tested with the Nanodrop 1000 (ThermoFisher Scientific, http://www.thermofisher.com) and by agarose gel electrophoresis.

### RNA isolation

RNA was isolated from plant material using the RNeasy Plant Micro Kit (Qiagen, http://www.qiagen.com), following the manufacturer's instructions. RNA concentration and size distribution was tested on the 2100 Bioanalyzer (Agilent Technologies, http://www.agilent.com) with the Agilent RNA 6000 Nano Kit to determine quantity and quality.

### NGS analysis

Sequencing data from genomic DNA were retrieved from NCBI SRA in the case of *P. patens* accession Villersexel (SRX030894). For the *P. patens* accession Reute genomic DNA was extracted from gametophores and sequenced on one lane of the Illumina

HiSeq 2500 (100-nt paired end) at the Max Planck-Genome-centre Cologne (http://mpgc.mpipz.mpg.de). The library was prepared according to the Illumina TruSeq protocol with an insert size of 300–400 bp. For Villersexel we started with 201 288 783 paired end reads (SRA: SRX030894), and for Reute we started with 150 711 864 paired end reads (SRA: SRX1528135). Read quality and trimming efficiency was evaluated with FASTQC 0.11.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). All sequence data were trimmed with TRIMMOMATIC 0.32 (Bolger *et al.*, 2014) using the following parameters: -phred33 ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:8:5 SLIDINGWINDOW:4:15 TRAILING:15 MINLEN:35.

After trimming we used 144 872 394 paired end reads for Villersexel and 145 604 667 paired end reads for Reute, for mapping.

All mapping steps were performed with GSNAP 2014-10-22 (Wu and Nacu, 2010), with default parameters. Trimmed reads were mapped against the chloroplast genome (NC_005087.1). Mapped reads were removed from the read pool and the same procedure was repeated for the mitochondrial genome (NC_007945.1) and the ribosomal rRNA genes (HM751653.1, X80986.1, X98013.1). After these steps the reads were mapped against the *P. patens* genome assembly V3 (DoE-JGI, http://phytozome.jgi.doe.gov). Duplicate reads were removed with SAMTOOLS RMDUP (Li *et al.*, 2009) to account for potential PCR artifacts. SNP calling was performed with GATK 3.3.0 (McKenna *et al.*, 2010) and SAMTOOLS 0.1.19. Reference and bam files were indexed with SAMTOOLS FAIDX/INDEX. Adding read-group information and sorting the bam file was achieved with PICCARD-TOOLS 1.115 (http://broadinstitute.github.io/picard). GATK was used as recommended by the Broad Institute for species without reference SNP databases, but the last quality recalibration step was omitted, as no large set of confirmed SNPs is available for *P. patens*. The second SNP calling pipeline used MPILEUP, BCFTOOLS and VARFILTER, with default parameters.

Intersections between.vcf files were extracted with BCFTOOLS 1.2 (http://samtools.github.io/bcftools/bcftools.html). Indel events and SNP events were separated with GATK SELECTVARIANTS. Only homozygous SNPs were used for further analysis because *P. patens* is a haploid organism, and only homozygous SNPs are expected.

Variants were annotated with SNPEFF 4.1 g (Cingolani *et al.*, 2012) using the COSMOSS 3.1 (https://www.cosmoss.org/physcome_project/wiki/Genome_Annotation/V3.1) gene annotation.

Depending on the SNP calling tool employed, we found 3–4 million SNPs between Gransden and Villersexel in the unfiltered data. The overlap of the GATK and SAMTOOLS SNP callers contains almost 1.9 million SNPs and 8900 insertions/deletions (indels). To test the sensitivity of our approach we used a set of 4650 SNPs from the Villersexel accession that were confirmed by an SNP bead array (Appendix S1). Of the SNP array probes, 4628 can be mapped to the V3 genome assembly, and each of the SNP callers (GATK, SAMTOOLS) calls >90% of these test sets on the Villersexel data, showing that the approach used is highly sensitive. Throughout the manuscript we use the data set from the GATK SNP calling because GATK has a quality recalibration step and performs realignment around insertions and deletions.

For the ecotype Reute at each SNP position the corresponding reference allele was replaced by the alternative allele, producing pseudo-chromosome sequences. Using the COSMOSS 3.1 gene annotations, coding sequences were extracted for each gene model and codon alignments were generated for Gransden and Reute. Subsequently, synonymous and non-synonymous nucleotide diversity was calculated using the Yang and Nielson method, as implemented in KAKS_CALCULATOR (Wang *et al.*, 2010). For chromosome-wide plots a sliding window approach (window size, 500 kbp; jump size, 400 kbp) was used and nucleotide diversity values were calculated with VARISCAN (Hutter *et al.*, 2006).

## SNP peak detection

Window-wise (50 kbp with 10-kbp overlap), SNP numbers were extracted from the 'pseudogenome' FASTA file by a custom R script. The R functions fisher.test and p.adjust (method = 'hochberg') were used to select fragments that show a significantly (adjusted $P$ value $< 0.01$) higher SNP number than the chromosome average. An SNP hot spot was called if at least five adjacent fragments showed a significantly higher SNP number.

## Microarray analyses

The NimbleGen 12 × 135k DNA microarray covers 32 851 transcripts, with a total of 130 221 probes in 32 741 probe sets (for 99.66% of the transcripts), of which 353 (1.07%) map redundantly and 87 (0.26%) contain fewer than four probes. Besides the v1.6 transcripts, spikes and negative controls were included, as were 580 v1.2 gene models that lack a v1.6 equivalent but were shown to be differentially expressed based on existing Combimatrix microarray data.

About 200 ng of total RNA was reverse transcribed and amplified using the WTA Kit (Sigma-Aldrich, https://www.sigmaaldrich.com). One microgram of cDNA was labeled with Cy3 according to the NimbleGen One-Color DNA Labeling Kit (Roche, http://www.roche.com); 4 μg of labeled cDNA was used for hybridization on the NimbleGen 12 × 135k DNA microarray, probe design OID33087 (Roche), according to the manufacturers' protocol using the NimbleGen Hybridization Kit (Roche). The NimbleGen Wash Buffer Kit (Roche) was used to prepare the slide for scanning.

The arrays were imaged using a laser scanner Agilent G2565CA Microarray Scanner System (Agilent Technologies). The image of the arrays was cut into single array images using NIMBLESCAN 2.5 (Roche), and the pixel intensities were extracted with the same software.

Microarray expression data were analyzed with ANALYST 7.5 (Genedata, https://www.genedata.com). Median condensed probe set expression values were quantile-normalized and analyzed further, as previously described (Wolf *et al.*, 2010). Box plots, hierarchical clustering (Figures S6, S7) and CyberT test analyses were performed with R (R Development Core Team, 2008).

## Quantitative real-time PCR (qPCR)

For validation of genes found to be differentially expressed in the microarray data, the treatment of *P. patens*, harvesting and RNA extraction was carried out as described above. cDNA was synthesized using the Superscript III kit (Life Technologies, now Thermo-Fisher Scientific, http://www.thermofisher.com) following the manufacturers' protocol. Real-time qPCR was carried out using the SensiMix SYBR green No-ROX kit (Bioline, http://www.bioline.com) with a 10-μL reaction volume. Primers were designed with PRIMER 3, aiming at an annealing temperature of around 60°C for all primers and 3′ clamp and intron-spanning, where applicable (Koressaar and Remm, 2007; Untergasser *et al.*, 2012), and checked for a single genomic locus via BLAST (cosmoss.org; see Table S3 for primer sequences). Melting curve analysis and non-template controls (NTCs) were carried out routinely to ensure the product specificity of individual reactions. After qPCR, reactions with multiple products were not taken into account for further analysis. Data were analyzed with Microsoft Excel 2010, applying the $\Delta\Delta C_t$ method. Expression rates were normalized for variation against the reference gene, a thioredoxin (Pp3c19_1800V3.1), which showed the smallest deviation among a broad range of microarray experiments, including juvenile gametophores, adult gametophores, and green and brown sporophytes (Hiss *et al.*, 2014). For the selection of candidate genes and primer sequences, see Table S2.

### ID conversion

Throughout the text the most recent COSMOSS 3.3 gene identifiers (CGIs) are used. Table S4 shows the corresponding CGIs for the v1.6 annotation and Phypa-IDs (v1.2 annotation) for all genes discussed here.

### Gene ontology (GO) analyses and visualization

The GO bias analyses used Fisher's exact test to calculate *P* values, as described previously (Widiez *et al.*, 2014). Multiple testing-corrected (Benjamini and Hochberg, 1995) *q* values were calculated in R with the function p.adjust (R Development Core Team, 2008). Word-cloud visualizations were created using the online tool WORDLE (http://www.wordle.net). The size of the word is proportional to the −log10(*q* value), and over-represented GO terms were colored dark green if $q \leq 0.0001$ and light green if $q > 0.0001$. Under-represented GO terms were colored dark red if $q \leq 0.0001$ and light red if $q > 0.0001$.

### CIRCOS plots

For the integrative visualization of the individual genomic features one karyotype ideogram was created and tracks were plotted with CIRCOS 0.67-6 (Krzywinski *et al.*, 2009). Chromosomes were split into smaller windows (window size, 500 kbp; window overlaps/jumps, 400 kbp) using values window averages (VWAs), normalized by scaling between a range of 0 and 1 per chromosome using the following equation:

$$\mathit{normalized\ window\ averagechr\ vwaichr} = \mathit{vwaichr} - \mathit{vwachrminvwachrmax} - \mathit{vwachrmin}.$$

## ACCESSION NUMBERS

NimbleGen microarray data for juvenile and adult gametophores, and green and brown sporophytes, are available at ArrayExpress: E-MTAB-4630 (http://www.ebi.ac.uk/arrayexpress). Published data for brown sporophytes (Becker lab): E-MTAB-3069. Combimatrix microarray data for brown sporophytes were deposited at ArrayExpress: E-MTAB-916.

Reute gDNA raw data have been made available via the Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra; SRP068341).

## ACKNOWLEDGEMENTS

The JGI Plant Gene Atlas project conducted by the US Department of Energy Joint Genome Institute was supported by the Office of Science of the US Department of Energy under contract no. DE-AC02-05CH11231. Full Gene Atlas data sets are available at http://phytozome.jgi.doe.gov. This work was supported by the German Federal Ministry of Education and Research (Freiburg Initiative for Systems Biology, 0313921 to SAR). We are grateful to Andrew C. Cuming for language editing. We thank Jörg Becker for access to the sporophyte microarray data prior to publication, and Eva Bieler, Swiss Nanoscience Institute (SNI), Nano Imaging, University of Basel, for help with taking the Cryo-SEM pictures, as well as Faezeh Donges and Marco Göttig for technical assistance, the Deutscher Wetterdienst (DWD) for providing climate data, and Juliet C. Coates and Jörg Becker for comments on the manuscript. The authors declare no conflicts of interest.

## SUPPORTING INFORMATION

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

**Figure S1.** Site and weather data for the Reute collection site.
**Figure S2.** Weather data of the Gransden collection site.
**Figure S3.** Bar chart of microarray expression values for the GRAS family protein Pp3c2_20930V3.1.
**Figure S4.** Phylogenetic tree of CHS genes modified after Wolf *et al.* (2010).
**Figure S5.** Word cloud of gene ontology terms (biological process) of the 15 genes confidently expressed at a higher level in mature sporophytes.
**Figure S6.** Box plot of microarray experiments from Reute developmental stages.
**Figure S7.** Hierarchical clustering of microarray experiments from Reute developmental stages.
**Table S1.** Published microarray and RNA-seq data sets for *Physcomitrella patens*.
**Table S2.** Average weather data from the Gransden and Reute sites.
**Table S3.** Primer sequences used for quantitative real-time PCR.
**Table S4.** Gene IDs for the genes discussed here.
**Table S5.** Reute genes that contain an SNP or an Indel leading to a premature stop codon.
**Table S6.** SNP peaks from Reute (as compared with Gransden).
**Table S7.** Over-represented gene ontology terms of genes found in SNP peaks.
**Table S8.** Gene IDs and their annotation for genes found in the chromosome-19 SNP peak.
**Table S9.** List of genes that show a $K_a/K_s$ ratio >2 between Reute and Gransden.

**Appendix S1.** List of SNPs detected by the *P. patens* bead array.

**Appendix S2.** Differentially expressed genes between Gransden and Reute adult gametophores, between Reute adult gametophores and green sporophytes, and between Reute green sporophytes and brown sporophytes.

## REFERENCES

**Ashton, N.W. and Raju, M.V.S.** (2000) The distribution of gametangia on gametophores of Physcomitrella (Aphanoregma) patens in culture. *J. Bryol.* **22**, 9–12.

**Barker, E.I. and Ashton, N.W.** (2013) A parsimonious model of lineage-specific expansion of MADS-box genes in Physcomitrella patens. *Plant Cell Rep.* **32**, 1161–1177.

**Beike, A.K., Horst, N.A. and Rensing, S.A.** (2010) Axenic bryophyte *in vitro* cultivation. *Endocyt Cell Res.* **20**, 102–108.

**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.

**Beike, A.K., Lang, D., Zimmer, A.D., Wust, F., Trautmann, D., Wiedemann, G., Beyer, P., Decker, E.L. and Reski, R.** (2015) Insights from the cold transcriptome of Physcomitrella patens: global specialization pattern of conserved transcriptional regulators and identification of orphan genes involved in cold acclimation. *New Phytol.* **205**, 869–881.

**Belling, J.** (1921) On counting chromosomes in pollen-mother cells. *Am. Nat.* **55**, 573–574.

**Benjamini, Y. and Hochberg, Y.** (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.* **57**, 289–300.

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

**Bolle, C.** (2004) The role of GRAS proteins in plant signal transduction and development. *Planta*, **218**, 683–692.

**Brown, R.C. and Lemmon, B.E.** (2011) Spores before sporophytes: hypothesizing the origin of sporogenesis at the algal-plant transition. *New Phytol.* **2011**, 1469–8137.

**Bryan, V.S.** (1957) Cytotaxonomic studies in the Ephemeraceae and Funariaceae. *The Bryologist*, **60**, 103–126.

**Busch, H., Boerries, M., Bao, J., Hanke, S.T., Hiss, M., Tiko, T. and Rensing, S.A.** (2013) Network theory inspired analysis of time-resolved expression data reveals key players guiding *P. patens* stem cell development. *PLoS ONE*, **8**, e60494.

**Cao, J., Schneeberger, K., Ossowski, S.** ***et al.*** (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.* **43**, 956–963.

**Charlot, F., Chelysheva, L., Kamisugi, Y.** ***et al.*** (2014) RAD51B plays an essential role during somatic and meiotic recombination in Physcomitrella. *Nucleic Acids Res.* **42**, 11965–11978.

**Cingolani, P., Platts, A., le Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M.** (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92.

**Colpitts, C.C., Kim, S.S., Posehn, S.E., Jepson, C., Kim, S.Y., Wiedemann, G., Reski, R., Wee, A.G., Douglas, C.J. and Suh, D.Y.** (2011) PpASCL, a moss ortholog of anther-specific chalcone synthase-like enzymes, is a hydroxyalkylpyrone synthase involved in an evolutionarily conserved sporopollenin biosynthesis pathway. *New Phytol.* **192**, 855–868.

**Cuming, A.C., Cho, S.H., Kamisugi, Y., Graham, H. and Quatrano, R.S.** (2007) Microarray analysis of transcriptional responses to abscisic acid and osmotic, salt, and drought stress in the moss, Physcomitrella patens. *New Phytol.* **176**, 275–287.

**Daku, R.M., Rabbi, F., Buttigieg, J., Coulson, I.M., Horne, D., Martens, G., Ashton, N.W. and Suh, D.Y.** (2016) PpASCL, the physcomitrella patens anther-specific chalcone synthase-like enzyme implicated in sporopollenin biosynthesis, is needed for integrity of the moss spore wall and spore viability. *PLoS ONE*, **11**, e0146817.

**Dellaporta, S., Wood, J. and Hicks, J.** (1983) A plant DNA minipreparation: version II. *Plant Mol. Biol. Rep.* **1**, 19–21.

**Engel, P.P.** (1968) The induction of biochemical and morphological mutants in the moss *Physcomitrella patens*. *Am. J. Bot.* **55**, 438–446.

**Frank, W., Decker, E.L. and Reski, R.** (2005) Molecular tools to study *Physcomitrella patens. Plant Biol. (Stuttg)* **7**, 220–227.

**Frey, W., Stech, M. and Fischer, E.** (2009) *Syllabus of Plant Families – Part 3 Bryophytes and Seedless Vascular Plants Berlin*. Stuttgart: Borntraeger.

**Glime, J.M.** (2007) Bryophyte ecology. In *Volume 1. Physiological Ecology* (Glime, J.M. ed) Ebook sponsored by Michigan Technological University and the International Association of Bryologists. Available at http://www.bryoecol.mtu.edu/.

**Gramzow, L. and Theissen, G.** (2010) A hitchhiker's guide to the MADS world of plants. *Genome Biol.* **11**, 214.

**Hiss, M., Laule, O., Meskauskiene, R.M.** ***et al.*** (2014) Large-scale gene expression profiling data for the model moss Physcomitrella patens aid understanding of developmental progression, culture and stress conditions. *Plant J.* **79**, 530–539.

**Hoffmann, G.R.** (1970) Spore viability in *Funaria hygrometrica. Bryologist*, **73**, 634–635.

**Hohe, A., Rensing, S.A., Mildner, M., Lang, D. and Reski, R.** (2002) Day length and temperature strongly influence sexual reproduction and expression of a novel MADS-Box gene in the moss *Physcomitrella patens. Plant Biol.* **4**, 762–762.

**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.

**Hutter, S., Vilella, A.J. and Rozas, J.** (2006) Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics*, **7**, 409.

**Kamisugi, Y. and Cuming, A.C.** (2005) The evolution of the abscisic acid-response in land plants: comparative analysis of group 1 LEA gene expression in moss and cereals. *Plant Mol. Biol.* **59**, 723–737.

**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.

**Khandelwal, A., Cho, S.H., Marella, H., Sakata, Y., Perroud, P.F., Pan, A. and Quatrano, R.S.** (2010) Role of ABA and ABI3 in desiccation tolerance. *Science*, **327**, 546.

**Knop, W.** (1868) *Der Kreislauf des Stoffs: Lehrbuch der Agricultur-Chemie*. Leipzig: H. Haessel.

**Koduri, P.K.H., Gordon, G., Barker, E., Colpitts, C., Ashton, N. and Suh, D.-Y.** (2010) Genome-wide analysis of the chalcone synthase superfamily genes of Physcomitrella patens. *Plant Mol. Biol.* **72**, 247–263.

**Komatsu, K., Nishikawa, Y., Ohtsuka, T., Taji, T., Quatrano, R.S., Tanaka, S. and Sakata, Y.** (2009) Functional analyses of the ABI1-related protein phosphatase type 2C reveal evolutionarily conserved regulation of abscisic acid signaling between Arabidopsis and the moss Physcomitrella patens. *Plant Mol. Biol.* **6**, 6.

**Koressaar, T. and Remm, M.** (2007) Enhancements and modifications of primer design program Primer3. *Bioinformatics*, **23**, 1289–1291.

**Krupa, J.** (1967) Studies on the physiology of germination of spores of *Funaria hygrometrica*. III. The influence of monochromatic light on the germination of the spores. *Acta Soc. Bot. Polon.* **36**, 57–65.

**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.

**Landberg, K., Pederson, E.R., Viaene, T., Bozorg, B., Friml, J., Jonsson, H., Thelander, M. and Sundberg, E.** (2013) The MOSS Physcomitrella patens reproductive organ development is highly organized, affected by the two SHI/STY genes and by the level of active auxin in the SHI/STY expression domain. *Plant Physiol.* **162**, 1406–1419.

**Lang, D., Zimmer, A.D., Rensing, S.A. and Reski, R.** (2008) Exploring plant biodiversity: the Physcomitrella genome and beyond. *Trends Plant Sci.* **13**, 542–549.

**Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.** and **1000 Genome Project Data Processing Subgroup.** (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

**Li, S., Zhao, Y., Zhao, Z., Wu, X., Sun, L., Liu, Q. and Wu, Y.** (2016) Crystal structure of the GRAS domain of SCARECROW-LIKE 7 in Oryza sativa. *Plant Cell*, **28**, 1025–1034.

**Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. and Gilad, Y.** (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517.

**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.

**Medina, R., Liu, Y., Li-Song, W., Shuiliang, G., Hylander, K. and Goffinet, B.** (2015) DNA based revised geographic circumscription of species of Physcomitrella sl (Funariaceae): *P. patens* new to East Asia and *P. magdalenae* new to East Africa. *The Bryologist*, **1**, 22–31.

**Morohashi, K., Minami, M., Takase, H., Hotta, Y. and Hiratsuka, K.** (2003) Isolation and characterization of a novel GRAS gene that regulates meiosis-associated gene expression. *J. Biol. Chem.* **278**, 20865–20873.

**Mosquna, A., Katz, A., Decker, E.L., Rensing, S.A., Reski, R. and Ohad, N.** (2009) Regulation of stem cell maintenance by the Polycomb protein FIE has been conserved during land plant evolution. *Development*, **136**, 2433–2444.

**Nakosteen, P.C. and Hughes, K.W.** (1978) Sexual Life cycle of three species of Funariaceae in culture. *Bryologist*, **81**, 307–314.

**O'Donoghue, M.T., Chater, C., Wallace, S., Gray, J.E., Beerling, D.J. and Fleming, A.J.** (2013) Genome-wide transcriptomic analysis of the sporophyte of the moss Physcomitrella patens. *J. Exp. Bot.* **64**, 3567–3581.

**Okano, Y., Aono, N., Hiwatashi, Y., Murata, T., Nishiyama, T., Ishikawa, T., Kubo, M. and Hasebe, M.** (2009) A polycomb repressive complex 2 gene regulates apogamy and gives evolutionary insights into early land plant evolution. *Proc. Natl Acad. Sci. USA*, **106**, 16321–16326.

**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.

**Possart, A. and Hiltbrunner, A.** (2013) An evolutionarily conserved signaling mechanism mediates far-red light responses in land plants. *Plant Cell,* **25**, 102–114.

**Prigge, M.J. and Bezanilla, M.** (2010) Evolutionary crossroads in developmental biology: *Physcomitrella patens. Development,* **137**, 3535–3543.

**Quatrano, R.S., McDaniel, S.F., Khandelwal, A., Perroud, P.F. and Cove, D.J.** (2007) *Physcomitrella patens*: mosses enter the genomic age. *Curr. Opin. Plant Biol.* **10**, 182–189.

**R Development Core Team** (2008) *R: A Language and Environment for Statistical Computing.* Available at http://www.R-project.org/.

**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

**Reski, R. and Cove, D.J.** (2004) Quick guide: *Physcomitrella patens. Curr. Biol.* **14**, R261–R262.

**Sakakibara, K., Nishiyama, T., Deguchi, H. and Hasebe, M.** (2008) Class 1 KNOX genes are not involved in shoot development in the moss Physcomitrella patens but do function in sporophyte development. *Evol. Dev.* **10**, 555–566.

**Sakakibara, K., Ando, S., Yip, H.K., Tamada, Y., Hiwatashi, 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.

**Schruth, D.** (2012) *Caroline: A Collection of Database, Data Structure, Visualization, and Utility Functions for R, R package version 0.7. 4.* Available at https://rdrr.io/cran/caroline/.

**von Stackelberg, M., Rensing, S.A. and Reski, R.** (2006) Identification of genic moss SSR markers and a comparative analysis of twenty-four algal and plant gene indices reveal species-specific rather than group-specific characteristics of microsatellites. *BMC Plant Biol.* **6**, 9.

**Stevenson, S.R., Kamisugi, Y., Trinh, C.H. *et al.*** (2016) Genetic analysis of Physcomitrella patens identifies ABSCISIC ACID NON-RESPONSIVE (ANR), a regulator of ABA responses unique to basal land plants and required for desiccation tolerance. *Plant Cell.* **28**, 1310–1327.

**Sun, X., Xue, B., Jones, W.T., Rikkerink, E., Dunker, A.K. and Uversky, V.N.** (2011) A functionally required unfoldome from the plant kingdom: intrinsically disordered N-terminal domains of GRAS proteins are involved in molecular recognition during plant development. *Plant Mol. Biol.* **77**, 205–223.

**Szovenyi, P., Devos, N., Weston, D.J., Yang, X., Hock, Z., Shaw, J.A., Shimizu, K.K., McDaniel, S.F. and Wagner, A.** (2014) Efficient purging of deleterious mutations in plants with haploid selfing. *Genome Biol. Evol.* **6**, 1238–1252.

**Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M. and Rozen, S.G.** (2012) Primer3–new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115.

**Vesty, E.F., Saidi, Y., Moody, L.A. *et al.*** (2016) The decision to germinate is regulated by divergent molecular networks in spores and seeds. *New Phytol.* **211**, 952–966.

**Wang, X.Y. and Paterson, A.H.** (2011) Gene conversion in angiosperm genomes with an emphasis on genes duplicated by polyploidization. *Genes (Basel),* **2**, 1–20.

**Wang, D., Zhang, Y., Zhang, Z., Zhu, J. and Yu, J.** (2010) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics,* **8**, 77–80.

**Wang, X., Tang, H. and Paterson, A.H.** (2011) Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major Poaceae lineages. *Plant Cell,* **23**, 27–37.

**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.

**Wolf, L., Rizzini, L., Stracke, R., Ulm, R. and Rensing, S.A.** (2010) The Molecular and Physiological Responses of Physcomitrella patens to Ultraviolet-B Radiation. *Plant Physiol.* **153**, 1123–1134.

**Wu, T.D. and Nacu, S.** (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics,* **26**, 873–881.

**Xiao, L., Wang, H., Wan, P., Kuang, T. and He, Y.** (2011) Genome-wide transcriptome analysis of gametophyte development in Physcomitrella patens. *BMC Plant Biol.* **11**, 177.

**Xiao, L., Zhang, L., Yang, G., Zhu, H. and He, Y.** (2012) Transcriptome of protoplasts reprogrammed into stem cells in Physcomitrella patens. *PLoS ONE,* **7**, e35961.

**Yaari, R., Noy-Malka, C., Wiedemann, G., Auerbach Gershovitz, N., Reski, R., Katz, A. and Ohad, N.** (2015) DNA METHYLTRANSFERASE 1 is involved in (m)CG and (m)CCG DNA methylation and is essential for sporophyte development in Physcomitrella patens. *Plant Mol. Biol.* **88**, 387–400.

**Zhao, S., Fung-Leung, W.P., Bittner, A., Ngo, K. and Liu, X.** (2014) Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE,* **9**, e78644.

**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 Genomics,* **14**, 498.

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

Today, the quality and efficiency of molecular and bioinformatic work is highly dependent on the resources available, e.g. the genome, transcriptome, methylome. This second genome paper, introducing the ***P. patens* genome at chromosomal scale**, was a huge milestone for the bryophyte and plant evo-devo communities. The improved genome assembly and annotation eased the bioinformatic and molecular work and lead to new findings. E.g. the distribution of genes and TEs over the chromosome could be shown to be homogenous in *P. patens* whereas for flowering plants, gene- and TE-rich regions are known. It also could be shown that the chromosomes, with a uniform distribution of eu- and heterochromatin, display a peak of copia-type TEs on each chromosome, coinciding with the centromeric region as well as histone marks co-localizing with genic areas (activating marks) and intergenic/TE regions (repressive marks). In addition, a **high-quality dataset of DNA methylation** was published, showing **gene body methylation to be present in some genes**, **coinciding with gene silencing**, contrary to flowering plants. Interestingly, it was found that the *P. patens* genome probably underwent **two rounds of whole genome duplications** which could not be shown for hornworts or liverworts. Thus, this publication demonstrated the ***P. patens* genome** to be **different from seed plant and other bryophyte genomes** and provides a strong basis for future analysis of genome evolution.

the plant journal

SEB Society for Experimental Biology

The Plant Journal (2018) 93, 515–533

doi: 10.1111/tpj.13801

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

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

[1]*Plant Biotechnology, Faculty of Biology, University of Freiburg, Schaenzlestr. 1, 79104, Freiburg, Germany,*
[2]*Plant Genome and Systems Biology, Helmholtz Center Munich, 85764, Neuherberg, Germany,*
[3]*Plant Cell Biology, Faculty of Biology, University of Marburg, Marburg, Germany,*
[4]*INRA, UMR 1095 Genetics, Diversity and Ecophysiology of Cereals (GDEC), 5 Chemin de Beaulieu, 63100, Clermont-Ferrand, France,*
[5]*Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstrasse 3, OT Gatersleben, D-06466, Stadt Seeland, Germany,*
[6]*HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA,*
[7]*Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linné Weg 10, D-50829, Cologne, Germany,*
[8]*VIB Center for Plant Systems Biology, Technologiepark 927, 9052 Ghent, Belgium,*
[9]*Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, B-9052, Gent, Belgium,*
[10]*Center for Research in Agricultural Genomics, CRAG (CSIC-IRTA-UAB-UB), Campus UAB, Bellaterra, Cerdanyola del Vallès, 08193, Barcelona, Spain,*
[11]*Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA,*
[12]*Centre for Plant Sciences, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, UK,*
[13]*Structural and Genomic Information Laboratory (IGS), Aix-Marseille Université, CNRS, UMR 7256 (IMM FR 3479), Marseille, France,*
[14]*DOE Joint Genome Institute, Walnut Creek, CA 94598, USA,*
[15]*Department of Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada,*
[16]*Department of Biology, University of Florida, Gainesville, FL 32611, USA,*
[17]*Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden,*
[18]*Boyce Thompson Institute, Ithaca, NY 14853, USA,*
[19]*Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA,*
[20]*University Herbarium and Department of Integrative Biology, University of California, Berkeley, CA 94720-2465, USA,*
[21]*New York Botanical Garden, Bronx, NY 10458, USA,*
[22]*Vertis Biotechnologie AG, Lise-Meitner-Str. 30, 85354, Freising, Germany,*
[23]*Max Planck Institute of Molecular Plant Physiology, Am Muehlenberg 1, 14476, Potsdam-Golm, Germany,*
[24]*Department of Biology, Université* Laval, Québec G1V 0A6, Canada,
[25]*Department of Plant Biology, University of Geneva, Sciences III, Geneva 4 CH-1211, Switzerland,*
[26]*Department of Plant Biology &* Pathology Rutgers, The State University of New Jersey, New Brunswick, NJ 08901, USA,
[27]*Department of Biological Sciences, University of Alberta, Edmonton, AB, T6G 2E9, Canada,*
[28]*Department of Medicine, University of Alberta, Edmonton, AB T6G 2E1, Canada,*
[29]*BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China,*
[30]*Shenzhen Huahan Gene Life Technology Co. Ltd, Shenzhen, China,*
[31]*Department of Biology, Washington University, St. Louis, MO, USA,*
[32]*WZW, Technical University Munich, Munich, Germany,*
[33]*BIOSS Centre for Biological Signalling Studies, University of Freiburg, Schaenzlestr. 18, 79104, Freiburg, Germany,*
[34]*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 D_eveloppement 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 plastidal were assembled *de novo*, yielding an improved assembly and annotation of both organellar genomes (correcting e.g. the N-terminal sequence of the plastidal 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).

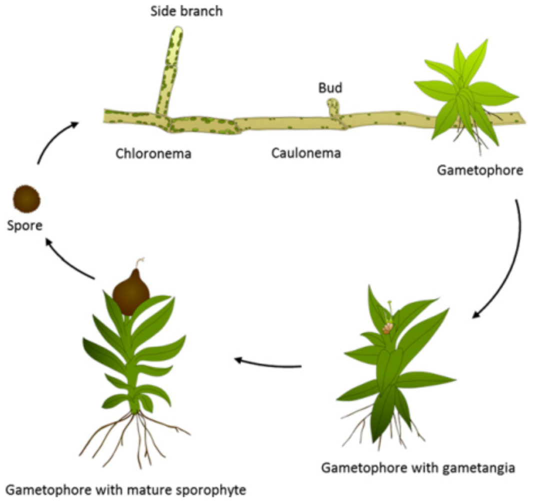

Side branch
Bud
Chloronema
Caulonema
Gametophore
Spore
Gametophore with gametangia
Gametophore with mature sporophyte

**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 spermatozoids the diploid zygote forms and subsequently performs embryogenesis. Spore mother cells in the diploid sporophyte undergo meiosis to form spores.

## 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 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

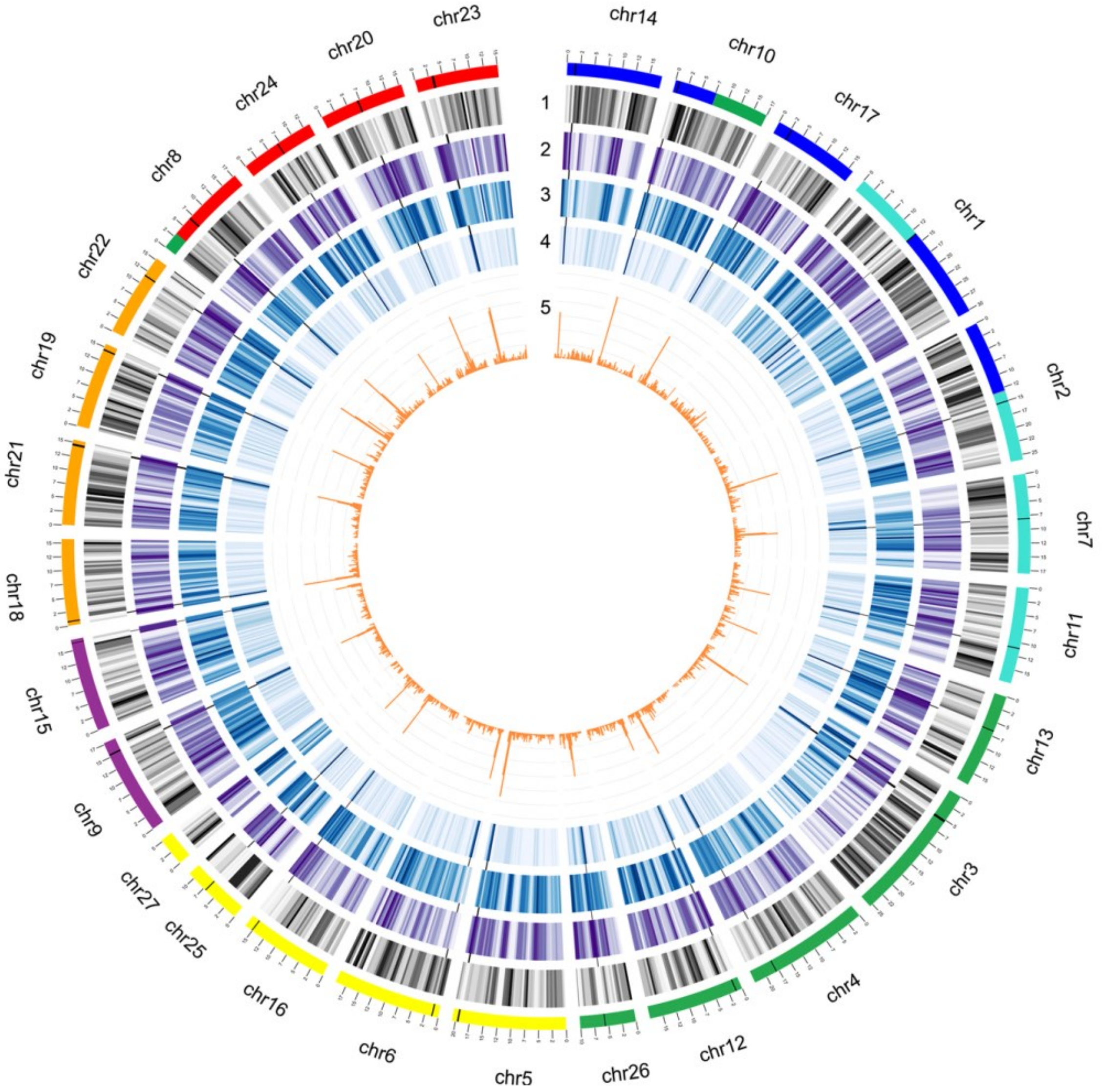


**Figure 2.** Chromosome structure, focus on TEs.

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 karytope (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

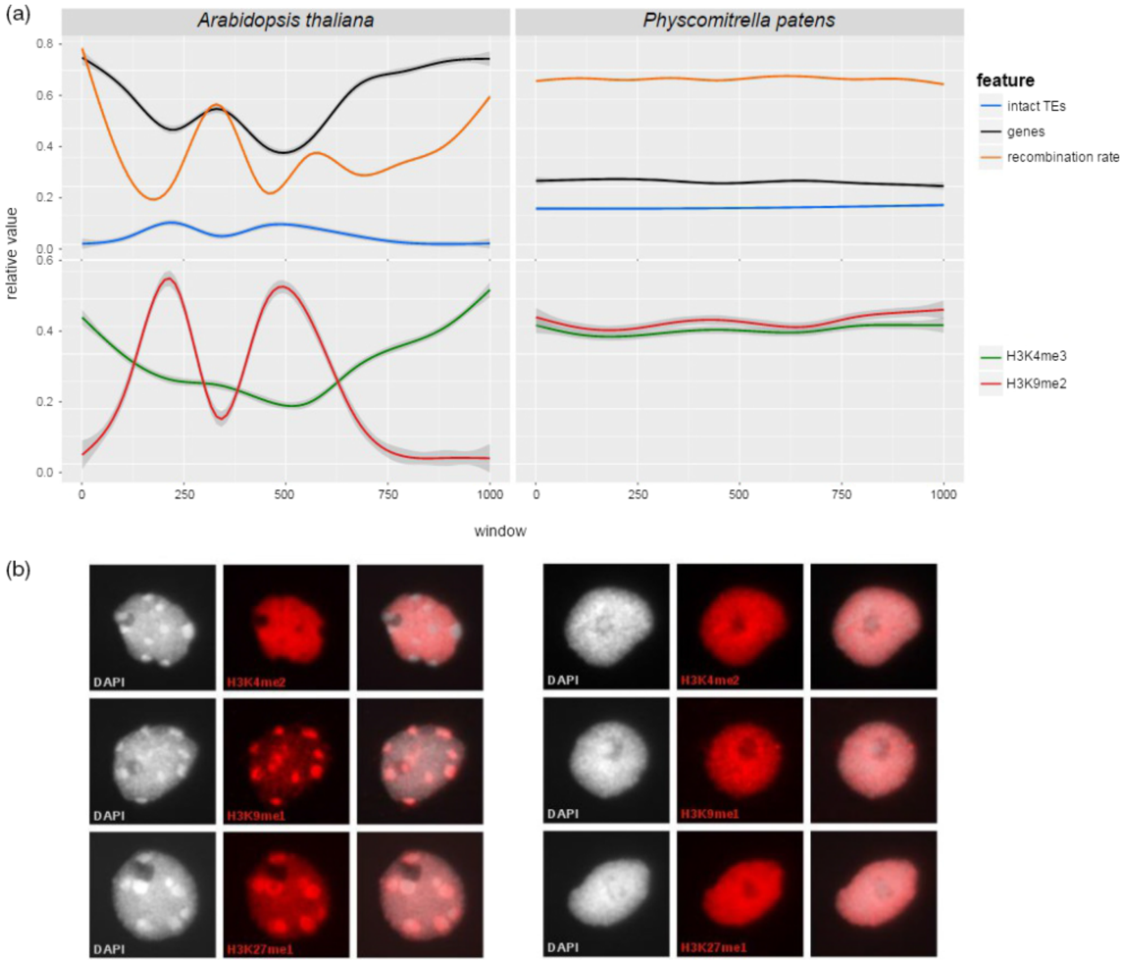


**Figure 3.** Comparative analysis of genome structures.

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).

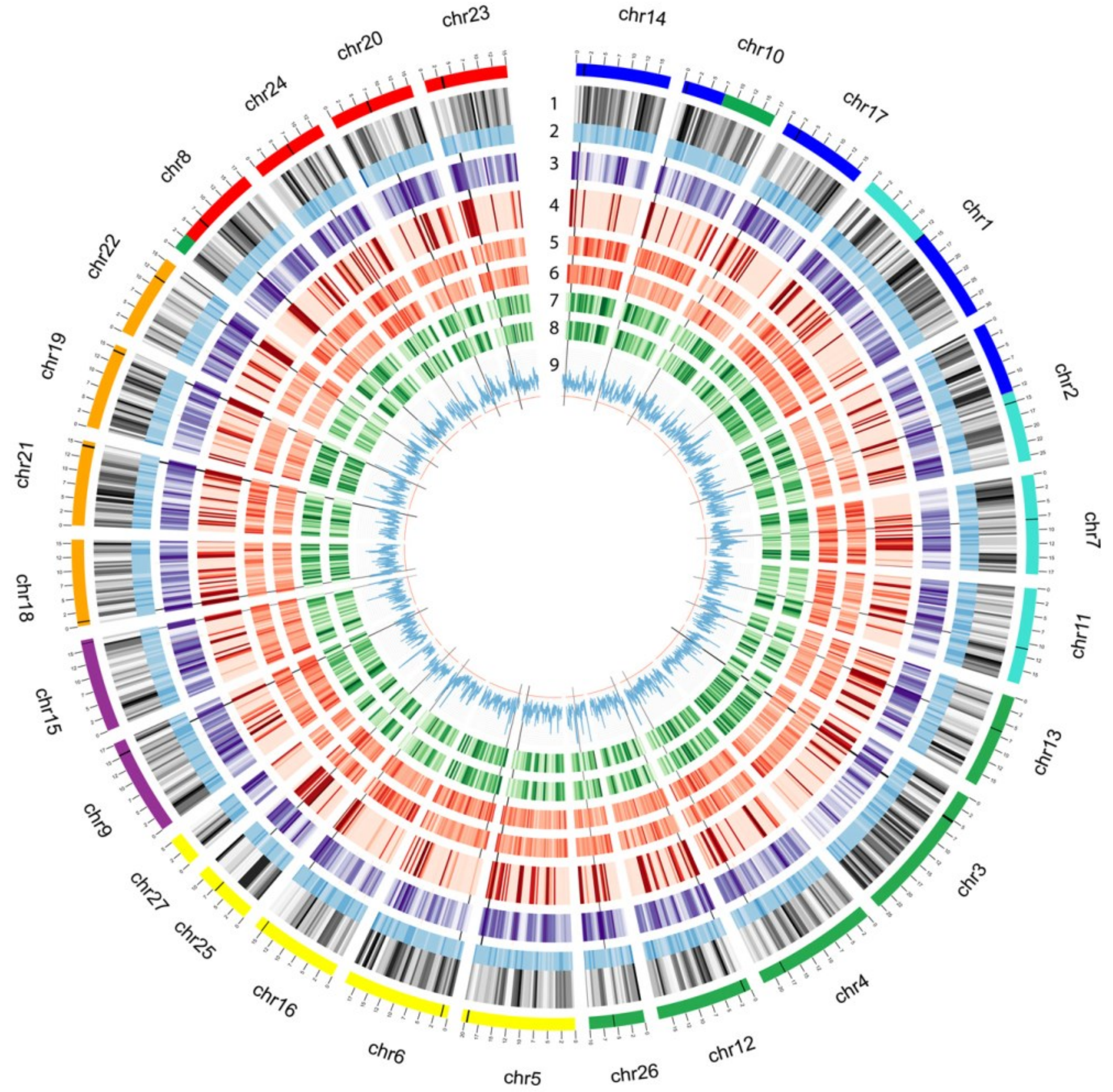


**Figure 4.** Chromosome structure, focus on epigenetic marks.
From outer to inner: karyotype bands colored according to ancestral genome blocks as in Figure 5, followed by: (1) gene density (grey) normalized 0,1; (2) GC content 0.25–0.45 (blue); (3) all TEs density (violet) normalized 0,1, NCLDV evidence 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 karyotope (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 gametohores 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 exlusively 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 NCLDVIs (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, NCLDVIs might provide a means to provide large numbers of siRNAs which, besides ensuring the transgenerational persistence of silencing, could also provide protection against cytoplasmatically 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 644p) 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) 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

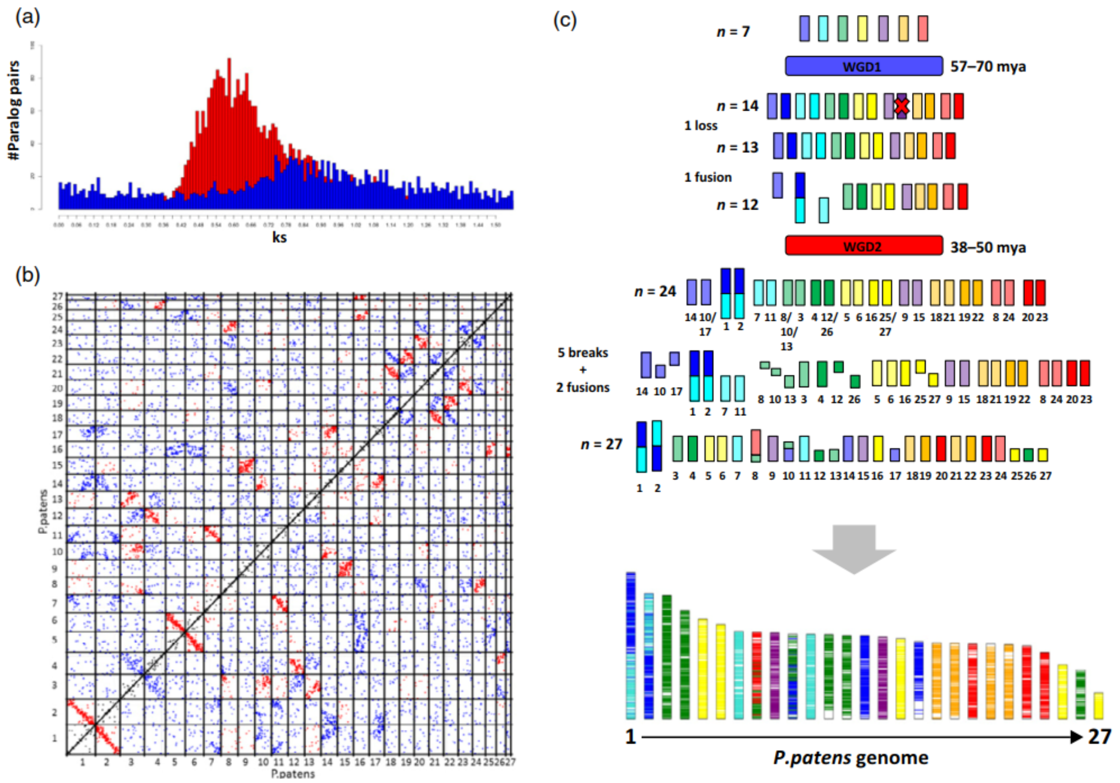


**Figure 5.** Evolutionary scenario leading to the modern *P. patens* genome.
(a) Ks distribution (*y*-axis) of paralogous pairs (*x*-axis) inherited from two (blue for older and red for more recent) WGD events.
(b) Dotplot representation of the paralogous pairs belonging to two WGD events.
(c) Karyotype 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.

moss 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/prots_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 GAII) 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) = Σ [(linkage group length) + 2 (linkage group length/no. markers)] (Fishman *et al.*, 2001) and one in which L = Σ[(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 discontiguity 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 ≥90% bp identity and ≥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 ABEU00000000. 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 ≥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 TEdenovo 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 × 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-genomes/) 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 μmol $sec^{-1}$ $m^{-2}$, for 6 weeks at 22°C/19°C day/night temperature following 8 h/16 h light/dark cycle, 20 μmol $sec^{-1}$ $m^{-2}$, for 7 weeks at 16°C/16°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-Shenzen, Shenzen, 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 GAII 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 ≥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;≥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 = ' 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%. Ks (rate of synonymous substitutions) distribution of the identified pairs unveiled two peaks illuminating two WGDs, one older and one more recent, included between Ks 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 Ks ranges between 0.75–0.9 and 0.5–0.65 and a mean substitution rate (r) of 9.4 × $10^{-9}$ 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 ≤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) 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_correection 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 karyotpyes 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 −log10(*q*-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:

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

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) 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), 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.

## 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 GKSW, 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 (CNGB). 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

**Akalin, 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 minipreparation: 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., Coggill, 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., Hiwatashi, 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., Egener, 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., Cuming, 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., Mizrachi, 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 Mizrachi, 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 paleo-hexaploidy 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.

## 8.3 *Physcomitrella patens* Reute-mCherry as a tool for efficient crossing within and between ecotypes

The monecious model moss ***P. patens*** **predominantly selfs** when cultivated *in vitro*. In most laboratories, Gd is the predominantly used ecotype which could be shown to be nearly self-sterile (Hiss et al., 2017). In order to analyze e.g. fertility, segregation of genetic markers and to determine species distance, crossing is the method of choice. For fast and efficient analysis, lines with fluorescent tags have already been established using the Gd and Vx ecotypes (Perroud et al., 2011). In this publication, the closely to Gd related ecotype Re (Hiss et al., 2017) has been established as an efficient crossing partner and is now available with the fluorescent tag mCherry. Crossing analysis with a less fertile Gd mutant showed the newly generated **Re-mCherry line can cross fertilize** with a high efficiency. The Re background shows similar growth behavior to Gd. Thus, this publication now allows the community to **separate already established mutants from the Gd background** and to transfer them in the fertile and highly similar Re background for analysis of processes like sporophyte development, requiring sexual reproduction.

plant biology

Plant Biology ISSN 1435-8603

RESEARCH PAPER

# *Physcomitrella patens* Reute mCherry as a tool for efficient crossing within and between ecotypes

P.-F. Perroud[1], R. Meyberg[1] & S. A. Rensing[1,2]

1 Plant Cell Biology, Faculty of Biology, University of Marburg, Marburg, Germany

2 BIOSS Centre for Biological Signalling Studies, University of Freiburg, Freiburg, Germany

**Keywords**

Crossing; ecotype; *Physcomitrella patens* Reute; sporophyte; transgenic.

**Correspondence**

P.-F. Perroud & S. A. Rensing, Plant Cell Biology, Faculty of Biology, University of Marburg, Marburg, Germany.
E-mails: pierre-francois.perroud@biologie.uni-marburg.de; stefan.rensing@biologie.uni-marburg.de

**Editor**

D. Staiger

Received: 7 March 2018; Accepted: 9 May 2018

doi:10.1111/plb.12840

## ABSTRACT

- *Physcomitrella patens* is a monoecious moss that is predominantly selfing in the wild. Laboratory crossing techniques have been established and crosses between the sequenced Gransden ecotype and the genetically divergent Villersexel ecotype were used for genetic mapping. The recently introduced ecotype Reute has a high fertility rate and is genetically more closely related to the Gransden ecotype than the Villersexel ecotype. Reute sexual reproduction phenology is similar to Gransden, which should allow successful crossing.
- Using the Reute ecotype and an existing Gransden mutant as a test case, we applied a normalised crossing approach to demonstrate crossing potential between these ecotypes. Also, using a standard transformation approach, we generated Reute fluorescent strains expressing mCherry that allow an easy detection of crossed offspring (sporophyte).
- We show that Reute can be successfully crossed with a self-infertile DR5:DsRed2 mutant generated in the Gransden background. Using newly established Reute fluorescent strains, we show that they can efficiently fertilise Reute as well as Gransden wild type. The resulting progeny display Mendelian 1:1 segregation of the fluorescent marker(s), demonstrating the suitability of such strains for genetic crossing.
- Overall our results demonstrate that Reute is highly suitable for genetic crossing. The Reute mCherry strain can be used as a suitable background for offspring selection after crossing.

## INTRODUCTION

During the last 15 years, *Physcomitrella patens* has become a leading model organism representing early divergence of land plant lineages. The sequencing of its genome (Rensing *et al.* 2008), followed by its constant improvement through re-annotation (Zimmer *et al.* 2013) and reassembly (Lang *et al.* 2018) had led to the establishment of a solid genome platform for evolutionary, functional and biotechnological approaches. Most of the sequence data and publications are based on the ecotype Gransden (Gd), based on a plant isolated in Gransden Wood, Huntingdonshire, UK, in 1962, by H.L.K. Whitehouse. The laboratory strain was subsequently established from a single spore and used for the first time in 1968 in a genetic study (Engel 1968). Since then this strain has been widely distributed worldwide and used for most of the published genomic work. Within Gd, crossing has provided a powerful tool to analyse mutants, but the low frequency of crossing events has historically restrained most results to strains displaying strong auxotrophic (Engel 1968; Ashton & Cove 1977; Ulfstedt *et al.* 2017) or self-fertility (e.g. Sakakibara *et al.* 2014) defects, facilitating crossing event detection. The first published *P. patens* genetic map was based on crosses between the Gd ecotype tagged with a neomycin phosphotransferase II (*nptII*) resistance marker or the male sterile *nicB5/ylo6* mutant established in the Gd background (Ashton & Cove 1977) with the Villersexel K3 (Vx) ecotype (Kamisugi *et al.* 2008). Vx was chosen specifically among *P. patens* isolates for being the most divergent, coupled with the capacity to produce sporophytes upon crossing (von Stackelberg *et al.* 2006). This divergence was subsequently confirmed by comparing genomic sequences between Gd, Vx and two other accessions: Reute (Re; Hiss *et al.* 2017a) and Kaskaskia (Ka; Lang *et al.* 2018). Vx displays a single nucleotide polymorphism (SNP) rate five to ten times higher than either Ka or Re, as compared to Gd (Table S1; Lang *et al.* 2018). Vx is also the only other ecotype besides Gd to have been used so far in the context of forward genetics studies. The genetic analysis of progeny of a cross between a mutant insensitive to ABA (ANR for ABA non-responsive) generated in the Gd background with Vx allowed identification of the causal gene for the detected phenotype (Stevenson *et al.* 2016). In order to facilitate rapid detection of crossed sporophytes, fluorescent marker strains where established in both the Gd and Vx background (Perroud *et al.* 2011). The presence of a fluorescent marker in one parental background allows efficient tracking of successful fertilisation events by non-invasive visual observation of the sporophyte in culture, even if this event is rare (a typical crossing jar contains up to 500 sporophytes). Additionally, segregation analysis can also be performed visually on the resulting offspring, displaying the characteristic haploid 1:1 Mendelian segregation of the marker. Hence, the crossing of such fluorescent mutants with wild types allows rapid detection of crossed sporophytes and permits assessment of male fertility through outcrossing rate evaluation, a method used successfully with mutants displaying fertility impairment (Hackenberg *et al.* 2016; Ortiz-Ramírez *et al.* 2017). The experimental crossing rate between and within the tested ecotypes (Vx, Gd and Ka) was low in all crosses, but one (Perroud *et al.* 2011). When Gd was used as female mate with Vx as a male mate the outcrossing rate was more variable and could reach 71%. Although it was not possible to conclude definitively the origin of this phenomenon, the best explanation was that during its long laboratory propagation history, the Gd strain may have accumulated deleterious (epi-)mutations that reduced its male fertility (Perroud *et al.* 2011). Most recently, the traceable Gd and Vx strains were used successfully to identify the causal sterile mutation in a novel somatic hybridisation coupled SNP approach (Moody *et al.* 2018a,b).

The *P. patens* ecotype Reute (Re) is genetically closer to Gd than to Vx, but it is clearly distinguishable from Gd by its high self-fertility rate (Hiss *et al.* 2017a). Re has been used successfully in the framework of genomic studies (Hiss *et al.* 2014, 2017a) for reverse genetic studies (Sanchez-Vera *et al.* 2017) and biotechnological approaches (Hiss *et al.* 2017b). Here we report the successful cross of the Re ecotype with a double transgenic strain (Dr5:DsRed2 and NLS-4) established in the Gd background (Bezanilla *et al.* 2003; Lavy *et al.* 2016). Albeit this mutant was not able to self, multiple crossed sporophytes were detected when paired with Re. The two marker transgenes, DsRed2 and NLS-GFP, segregated in the progeny in Mendelian fashion (F1 1:1:1:1) allowing us to recover a strain without a marker, with both markers and individual Dr5:DsRed2 and NLS-4 marker strains. In parallel, we established two fluorescent Re transgenic strains expressing mCherry and successfully used them in crosses with both Re wild type and Gd. Similar to the cross with Re and Dr5:DsRed2 and NLS-4, the red fluorescent marker segregates in Mendelian fashion (F1 1:1) in the progeny. We present here three fertile single marker strains, Dr5:DsRed2, Re-mCherry#43 and Re-mCherry#63, that can be used for efficient intra- and inter-*P. patens* ecotype crossing, expanding the palette of previously established fluorescent marker lines (Gd-green and Vx-red; Perroud *et al.* 2011).

## MATERIAL AND METHODS

### Plant material and culture conditions

*Physcomitrella patens* Reute (Hiss *et al.* 2017a), *P. patens* Gransden (controlled-descendant of the Gransden lab strain used in previous crossing experiments; Perroud *et al.* 2011) and the double mutant Dr5:DsRed2 (Lavy *et al.* 2016) were routinely cultivated as previously described (Hiss *et al.* 2017a) on solidified (0.7% [w/v] agar) mineral medium, also known as modified (Reski & Abel 1985) Knop's medium (Knop 1868), on 9-cm Petri dishes sealed using 3M Micropore tape or Parafilm. Incubator temperature was set at 22 °C with a long-day light cycle of 16-h light/8-h dark (70 $\mu$mol·m$^{-2}$·s$^{-1}$ white light). For protoplast generation, gametophytic tissue was blended weekly in sterile water and plated on cellophane overlaying Knop medium enriched with 5 mM di-ammonium tartrate to obtain a homogenous pure protonemal culture. In order to facilitate crossing events, sporulation and crossing experiments were performed as modification of different protocols (Engel 1968; Hohe *et al.* 2002; Cove *et al.* 2009). Either growing protonema or gametophore tissue was inoculated in 220 ml bulbous Weck jars (Weck, Wehr-Öflingen**,** Germany) containing 100 ml solidified Knop medium, sealed with 3M Micropore tape, and grown for 5 weeks in standard incubator conditions (see above). Subsequently, the jar was transferred to 16 °C with a short-day light cycle of 8-h light/16-h dark and a reduced fluence rate of 20 $\mu$mol·m$^{-2}$·s$^{-1}$ white light for the rest of the experiment (Hohe *et al.* 2002). After 2 weeks at 16 °C (upon gametangia formation), the culture was submerged in sterile Type I/MilliQ water for 24 h, after which the excess water was drained off, leaving a moist culture. The same operation was repeated 1 week later to assure complete fertilisation. The first observable young green sporophytes were observed after one more week, *i.e.* 70 days after inoculation or 7–14 days after fertilisation. Sporophyte harvesting was performed 2–4 weeks later, once the sporophytes were uniformly dark brown (mature brown sporophyte; Hiss *et al.* 2017a). Sporophytes were subsequently stored for a least 1 week in the dark at 4 °C before performing germination assays. Germination was routinely performed on Knop medium enriched with 5 mM di-ammonium tartrate, and germination rate was scored 7 days after inoculation. This procedure is similar to that recently described (Vesty *et al.* 2016), but does not use cellophane to plate the spores and is scored only at a fixed time point after inoculation.

### Molecular procedures

*Vector construction and preparation*

If not mentioned otherwise, all restriction enzymes were purchased from New England Biolabs (Frankfurt, Germany) and the reactions performed as per the manufacturer's instructions. PCR reactions are performed with OneTaq® DNA polymerase with a final the reaction volume of 15 $\mu$l. The primer sequences used in this study are listed in Table S2. The expression vector was constructed from three functional sections: two targeting sequences flanking the construct to permit efficient transformation, a resistance cassette Lox-35S:Zeocin-Noster-Lox (Perroud & Quatrano 2008) for selection, and the expression cassette Ubiquitin promoter-mCherry-Nos terminator for red fluorescent protein expression. In order to ensure efficient genome transformation, the locus Pp3c10_12220V3.1 was chosen, as it is expressed but codes for one of two copies of ARPC3 present in the *P. patens* genome. Pp3c10_12220V3.1 displays much lower expression values than Pp3c3_23320V3.1, the other ARPC3 gene, in all tissues tested (Perroud *et al.* 2018; Figure S4). The ARPC3 protein is part of a seven subunit functional complex (Arp2/3 complex) and the two copies are expressed at a higher level than the other genes of the complex, hence its targeting was not anticipated to generate a severe phenotype. First, 5′ and 3′ targeting sequences were PCR amplified using the primer pairs P1/P2 and P3/P4, respectively, and each cloned into pTOPO2.1 to generate p5′T and p3′T vectors. The 5′ targeting fragment was cloned from p5′T using the restriction enzymes *Sph*I and *Sac*I into pLox-35S:Zeocin-Noster-Lox to generate a p5′T-pLox-35S:Zeocin-Noster-Lox vector. Then, the 3′ targeting fragment was cloned from p3′T using *Sac*II restriction enzyme into p5′T-sequence-pLox-35S:Zeocin-

Noster-Lox to create a p5′T-sequence-pLox-35S:Zeocin-Noster-Lox-p3′T vector. Orientation of the targeting sequences was confirmed by sequencing from the vector backbone. The expression cassette was generated in two steps. First, using the primer pair P5, P6 the cloning unit *Zm*Ubiquitin promoter-Gateway cassette-Nos terminator was amplified out of the pTHUBI-Gateway vector (Perroud *et al.* 2011) and cloned into pGEM-Teasy to generate a pUbi-Gateway vector. mCherry cDNA (Perroud *et al.* 2011) was cloned into the pUbi-Gateway using the Clonase LR reaction (Life Technologies, Darmstadt, Germany) to create pUbi-mCherry. Finally, the red fluorescent protein expression cassette was cut out of pUbi-mCherry using the *Hpa*I restriction enzyme and ligated into the p5′T-sequence-pLox-35S:Zeocin-Noster-Lox-p3′T in the blunted (using Klenow fragment) *Not*I restriction site to create the final pZTUbi:mCherry vector (for map see Figure S1).

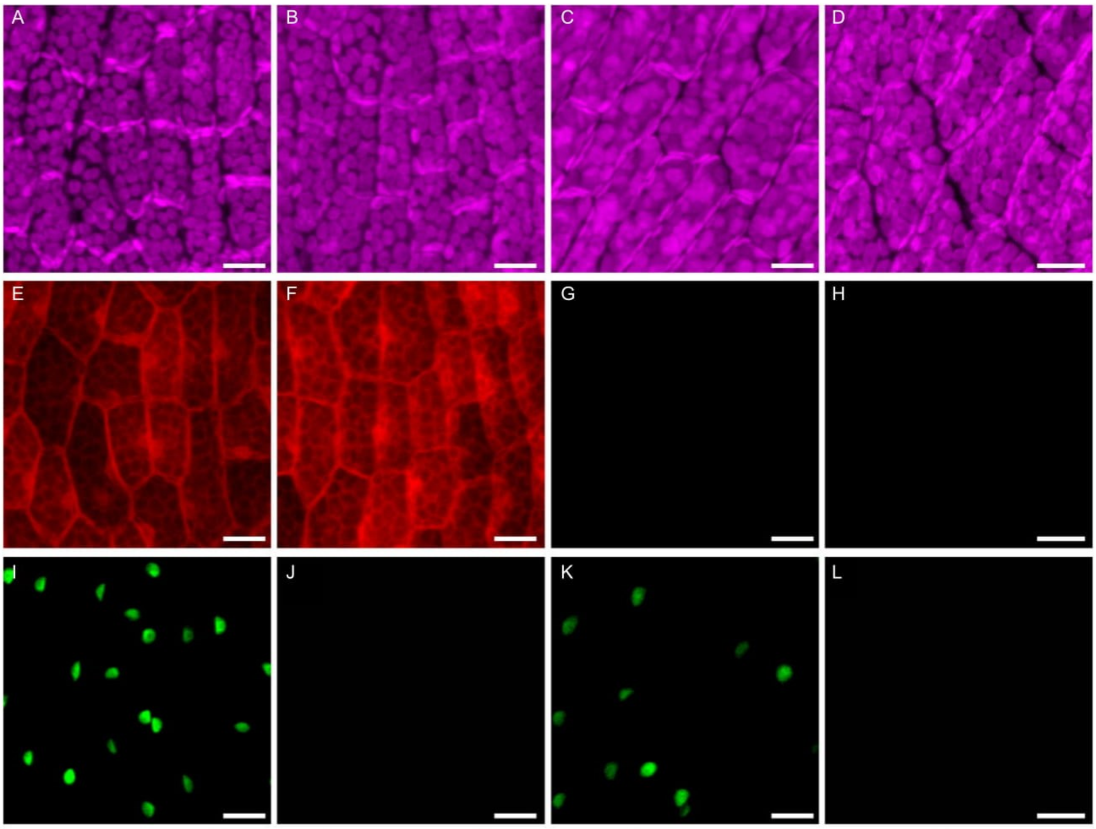
Plasmid DNA generation for transformation was performed with the NucleoBond Xtra Midi kit (Macherey-Nagel, Düren, Germany). The plasmid for stable transformation was cut using the restriction enzyme *Swa*I to separate the targeting vector from the plasmid backbone. After restriction evaluation, the cut DNA was ethanol-precipitated and resuspended in sterile TE.

### *Genomic DNA extraction*

Rapid genomic DNA (gDNA) extraction for genotyping purposes using less than 50 mg of moss tissue was performed as previously described (Cove *et al.* 2009).

### *Moss transformation*

Standard transformation protocols developed with the Gd ecotype are transferable to the Re ecotype. Protoplast



| M | Spores number | No fluorescence | Green + red Fluorescence | Green Fluorescence | Red Fluorescence | $X^2$ (1:1:1:1) |
|---|---|---|---|---|---|---|
| Re16 x Dr5:Dsred2 | 184 | 56 | 45 | 42 | 41 | 0.423723025 |

**Fig. 1.** Segregant analysis of the Re × Dr5:DsRed2 cross. A–L: confocal images of phyllid cells of the four segregating phenotypes: with both fluorescent markers (A, E, I), Dr5:DsRed2 fluorescent marker (B, F, J), NLS-4 fluorescent maker (C, G, K) and no fluorescent marker (D, H, L). Chloroplast auto-fluorescence signal is shown in magenta false colour (A–D), Dr5:DsRed2 specific signal in red false colour (E–H), GFP specific signal in green false colour (I–L). Bar: 10 µm. M: Distribution of fluorescent signal in germinating segregants, the *P*-value of the Chi-squared test to reject an expected 1:1:1:1 ratio (three degrees of freedom) is not significant.

transformation and selection were performed as described previously (Cove *et al.* 2009) with slight modifications. Namely, after protoplast regeneration all media used for selection and selection release were standard Knop medium supplemented with 5 mM di-ammonium tartrate; due to the high light sensitivity of the antibiotic Zeocin (ThermoFischer Scientific, Schwerte, Germany; # R25001), the selection plates were renewed every 3 days (Perroud & Quatrano 2008). Selection with Zeocin was performed at a final concentration of 100 μg·ml$^{-1}$.

### Microscopy

Microscopy was performed with a Leica DM6000 bright field microscope equipped with epi-fluorescence capacity (Leica Microsystems, Wetzlar, Germany). mCherry was excited with a mercury HBO100 lamp using a BP562/40 nm 593 dichroic mirror. Images were acquired with a Leica DFC295 camera using Leica Application suite version 4.4 software. Routine macroscopic observations were performed with a fluorescence stereomicroscope SteREO Lumar.V12 (Carl Zeiss, Oberkochen, Germany). Confocal microscopy observation and image acquisition were performed with a TCS-SP8 confocal microscope (Leica Microsystems) equipped with a tuneable white light laser. GFP was excited at 488 nm and emitted fluorescence was acquired between 495 and 530 nm. mCherry was excited at 561 nm and emission was acquired between 590 and 600 nm. DsRed2 was excited at 558 nm, emission was acquired between 575 and 595 nm. The chloroplast auto-fluorescence signal was

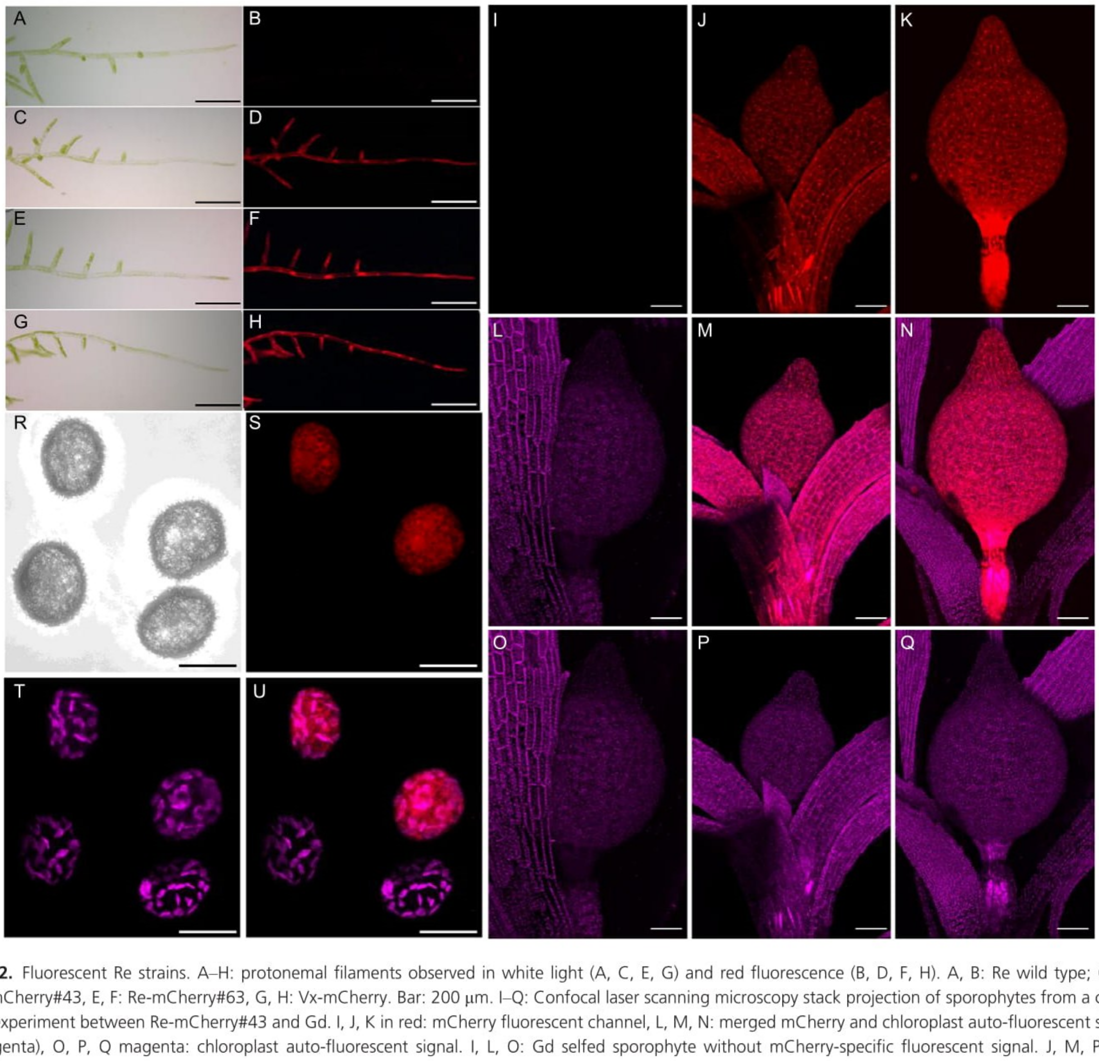


**Fig. 2.** Fluorescent Re strains. A–H: protonemal filaments observed in white light (A, C, E, G) and red fluorescence (B, D, F, H). A, B: Re wild type; C, D: Re-mCherry#43, E, F: Re-mCherry#63, G, H: Vx-mCherry. Bar: 200 μm. I–Q: Confocal laser scanning microscopy stack projection of sporophytes from a crossing experiment between Re-mCherry#43 and Gd. I, J, K in red: mCherry fluorescent channel, L, M, N: merged mCherry and chloroplast auto-fluorescent signal (magenta), O, P, Q magenta: chloroplast auto-fluorescent signal. I, L, O: Gd selfed sporophyte without mCherry-specific fluorescent signal. J, M, P: Re-mCherry#43 strain selfed, both gametophore and sporophyte display mCherry fluorescent signal. K, N, Q: crossed sporophyte between Re-mCherry#43 and Gd; only the sporophyte displays mCherry fluorescent signal. Bar: 20 μm. R, S, T, U: Confocal laser scanning microscopy stack projection of imbibed segregating spores of a crossed sporophyte between Re-mCherry#43 and Gd. R: bright field. S: in red: mCherry specific fluorescent signal. T: in magenta, auto-fluorescent chloroplast signal. U: Merged signal from S and T. Top spores display mCherry fluorescent signal, bottom spores do not show any signal. Bar: 20 μm.

acquired between 660 and 710 nm, with all three aforementioned excitation wavelengths. Images were processed using ImageJ version 1.51t (Schneider *et al.* 2012).

## RESULTS AND DISCUSSION

### *Physcomitrella patens* Re can efficiently cross with Gd

In order to evaluate the crossing potential of *P. patens* ecotype Re we established a crossing experiment between Re and the Gd Dr5:DsRed2 transgenic strain. The Dr5:DsRed2 strain contains two different transgenes, Dr5:DsRed2 leading to the nucleo-cytoplasmic accumulation of red fluorescent protein as a reporter of auxin accumulation level, and NLS-4 generating a nuclear green fluorescence (Lavy *et al.* 2016). Additionally, both parental strains were assayed in independent jars as a selfing control. At the end of the crossing/selfing procedure, all Re gametophores with more than ten phyllids in the control jar displayed a mature sporophyte, as previously reported (Hiss *et al.* 2017a). On the other hand, the Dr5:DsRed2 transgenic did not generate a single sporophyte in three independent trials. This self-infertility is potentially attributable to the intermediary transgenic strain NLS-4 (Bezanilla *et al.* 2003), which has to our knowledge not been regularly selfed. Also, its retransformation to generate the Dr5:DsRed2 strain may have allowed (epi-)mutations to occur and thus render the final Dr5:dsRed2 self-sterile. Under crossing conditions (in co-culture), the Dr5:dsRed2 gametophores yielded sporophytes in a similar time to Re, demonstrating that Re can cross efficiently with other ecotypes. Moreover, this indicates that the fertility impairment of the Gd Dr5:DsRed2 mutant is of male origin. Weak Gd self-fertility has also been reported in other studies (Perroud *et al.* 2011; Hiss *et al.* 2017a). Spores from the crossed sporophytes germinated successfully, and the two markers segregated in the population with the four predicted visual phenotypes assumed by single insertion loci of the transgenes: double fluorescent signal (Fig. 1A, E and I), no fluorescent signal (Fig. 1D, H and L) as well as single DsRed2 (Fig. 1B, F and J) and NLS-4 (Fig. 1C, G and K). Segregation frequency distribution was not statistically different (Chi-squared test, $P = 0.42$) to the predicted 1:1:1:1 ratio (Fig. 1M), indicating that both transgenes act like two independent single loci. This segregation pattern is no different from that reported previously for the cross using Vx and Gd as parental strains (Perroud *et al.* 2011). Finally, the segregants of this cross allowed us to isolate a self-fertile single marker strain carrying the Dr5:DsRed2 transgene, which can be used without interference from the NLS4 transgene and permit its use in sporophytic tissue. 18 F1 segregants were submitted to the self-sporulation procedure, and only one displayed sporophytes similar to the Re rate. This strain allows analysis of auxin accumulation with a fluorescent reporter in the diploid stage of *P. patens*. The small population size, lack of information about the cause of Dr5:DsRed2 self-sterility and fertility segregation 1 (Re fertile):1 (low fertility):18 (unfertile) in this specific Gd and Re ecotype does not allow us to conclude the source of this severe fertility segregation distortion, but the fertile single marker Re Dr5:DsRed2 S19 strain is now available for further study.

**Table 1.** Crossing efficiency between transgenic Reute strains with Reute and Gransden wild type.

| parent B<br>parent A | Re-mCherry-43<br>% crossed | n | Re-mCherry-63<br>% crossed | n |
|---|---|---|---|---|
| Re | 1.82 | 56 | 1.86 | 55 |
| Gd | 99.2 | 115 | 100 | 117 |

### *Physcomitrella patens* Re mCherry mutant permits rapid crossed sporophyte detection

In order to facilitate the easy detection of crossed sporophytes using the Re ecotype, we used the same approach already established in the Gd and Vx background (Perroud *et al.* 2011). We constructed the vector pZTUbi:mCherry to generate a moss strain accumulating easily visually detectable red fluorescence. The vector contains a Zeocin resistance cassette for plant selection, targeting sequence to ensure integration into the genome *via* homologous recombination and the expression cassette with the ubiquitin promoter driving the mCherry cDNA (see Material and Methods for vector construction and Figure S1 for map of pZTUbi:Cherry). This vector was transfected into protoplasts of *P. patens* ecotype Reute. Mutants generated by targeted insertion with a linear construct featuring homologous flanks can result in correct single locus integration through homologous recombination (SHR), as well as in concatemeric single locus integration (CHR) and integration into (additional) non-targeted loci (NTL) through non-homologous end joining (Kamisugi *et al.* 2006). Upon selection, 86 antibiotic-resistant transformants were isolated, 60 of which were initially selected for uniformly detectable fluorescent signal. Yet, most of these transformants displayed a relatively weak signal compared to existing strains. Hence, only five lines with strong fluorescence (potentially expressing more than one mCherry copy) were selected for the growth test and sporophyte formation evaluation. They all produced successfully selfed sporophytes and viable spores on all gametophores with more than ten phyllids. However, only two strains displayed a wild-type phenotype in the course of the life cycle. For these two strains, protonemal growth (Figure S2A–D), gametophore morphological development and sporophyte development, as well as number of sporophytes per gametophore with more than ten phyllids are not distinguishable to the wild type grown in parallel (Figure S2E), suggesting that they are SHR/CHR insertions, while the other three probably harbour NTL. It also confirms that removal of the locus used for targeting does not impact *P. patens* life cycle morphologically or phenologically. These two will be referred to as Re-mCherry-43 and Re-mCherry-63 hereafter. Indeed, genotyping of the transformed locus confirmed

**Table 2.** Segregation analysis of the fluorescent marker in the cross progeny between transgenic Reute and Gransden wild type. The *P*-values of the Chi-squared test for rejection of an expected 1:1 ratio with one degree of freedom are not significant.

| | number of spores | wild type phenotype | fluorescent | X2 (1:1) |
|---|---|---|---|---|
| Re16-mCherry#43 × Gd | 1202 | 584 | 618 | 0.326751277 |
| Re16-mCherry#63 × Gd | 1524 | 760 | 764 | 0.918389097 |

the targeting events in these two strains. Re-mCherry-43 displayed an SHR pattern with the wild type signal absent (Figure S3B) and 5′ and 3′ insertion signal present (Figure S3C and D); Re-mCherry-63 still carries the wild type locus (Figure S3B) but displays the 5′ insertion signal (Figure S3C), suggesting CHR. The intensity of the red fluorescent signal is in the same range as the transgenic Vx-red previously published (Fig. 2C–-H).

In order to evaluate the crossing capacity of the fluorescent lines, and to confirm single site (SHR/CHR) insertion, both strains were crossed independently with their parental strain Re as well as with Gd. Crossing was successful and the fluorescent sporophytes could be easily detected on wild-type gametophores (Fig. 2J–Q). In the sporophyte, the mCherry fluorescent signal was detectable in both homozygote (Fig. 2J, M and P) and heterozygote (Fig. 2K, N and Q) context. Similarly to observations in the Vx transgenic (Perroud *et al.* 2011), the selfing rate was very low and the crossing rate between Re and Gd was very high (Table 1). This very high crossing rate between Re and Gd confirms the weakness of Gd as a male partner, a trait potentially acquired during its long laboratory use.

The segregation analysis performed on germinated spores sampled from 15 different crossed sporophytes between both Re-mCherry-43 × Gd and Re-mCherry-63 × Gd crosses showed a clear 1:1 segregation of the fluorescent signal. This pattern is consistent for a single locus segregating in a haploid organism (Table 2), indicating that the transgene is confined to a single locus (SHR or CHR) in both Re-mCherry-43 and Re-mCherry-63. Interestingly, dry spores did not show a detectable fluorescent signal, but imbibed spores displayed a detectable fluorescent signal (Fig. 2R–U).

In conclusion, we show that the *P. patens* Re ecotype can be used as an alternative to the Gd ecotype for both reverse genetics and crossing within and between ecotypes. Not only its higher fertility as compared to Gd makes Re a good candidate to cross these two ecotypes, but it also allows efficient crossing within the Re ecotype. With an outcrossing rate of almost 2%, screening of spores from sporophytes derived from a cross between two strains with different selectable markers is feasible, since a standard crossing jar typically contains more than 200 sporophytes. In addition, established specific genetic approaches, such as the use of auxotroph mutants to facilitate selection (Ulfstedt *et al.* 2017), can be transferred into the Re background to perform reverse genetic screens.

## ACKNOWLEDGEMENTS

The authors are grateful to Michael Prigge and Mark Estelle for the generous gift of *P. patens* strain Dr5:DsRed2+NLS-4. The authors thank Rebecca Hinrichs for excellent technical assistance. We also thank Stefanie Reißmann (Kahmann laboratory) at the MPI for terrestrial microbiology in Marburg for use of the laboratory confocal microscopy equipment.

## SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article:

**Figure S1.** Schematic representation of the transformation vector pT3ZUbi:mCherry.

**Figure S2.** Growth of Re-mCherry#43 and Re-mCherry#63 is not distinguishable from Re wild-type.

**Figure S3.** Genotyping of Re-mCherry strains.

**Figure S4.** Expression profile of *PpARPc3a* (Pp3c3_23320V3.1) *PpARPC3b* (Pp3c10_12220V3.1).

**Table S1.** Summary table of SNPs and indels between Vx, Ka and Re as compared with Gransden (data from Lang *et al.* 2018).

**Table S2.** primers used in this study.

## REFERENCES

Ashton N.W., Cove D.J. (1977) Isolation and preliminary characterization of auxotrophic and analog resistant mutants of moss, physcomitrella-patens. *Molecular and General Genetics*, **154**, 87–95.

Bezanilla M., Pan A., Quatrano R.S. (2003) RNA interference in the moss *Physcomitrella patens*. *Plant Physiology*, **133**, 470–474.

Cove D.J., Perroud P.F., Charron A.J., McDaniel S.F., Khandelwal A., Quatrano R.S. (2009) The moss *Physcomitrella patens*: a novel model system for plant development and genomic studies. *Cold Spring Harbor Protocols*, **2009**, https://doi.org/10.1101/pdb.emo115.

Engel P.P. (1968) The induction of biochemical and morphological mutants in the moss *Physcomitrella patens*. *American Journal of Botany*, **55**, 438–446.

Hackenberg D., Perroud P.-F., Quatrano R., Pandey S. (2016) Sporophyte formation and life cycle completion in moss requires heterotrimeric G-proteins. *Plant Physiology*, **172**, 1154–1166.

Hiss M., Laule O., Meskauskiene R.M., Arif M.A., Decker E.L., Erxleben A., Frank W., Hanke S.T., Lang D., Martin A., Neu C., Reski R., Richardt S., Schallenberg-Rudinger M., Szovenyi P., Tiko T., Wiedemann G., Wolf L., Zimmermann P., Rensing S.A. (2014) Large-scale gene expression profiling data for the model moss *Physcomitrella patens* aid understanding of developmental progression, culture and stress conditions. *The Plant Journal*, **79**, 530–539.

Hiss M., Meyberg R., Westermann J., Haas F.B., Schneider L., Schallenberg-Rudinger M., Ullrich K.K., Rensing S.A. (2017a) Sexual reproduction, sporophyte development and molecular variation in the model moss *Physcomitrella patens*: introducing the ecotype Reute. *The Plant Journal*, **90**, 606–620.

Hiss M., Schneider L., Grosche C., Barth M.A., Neu C., Symeonidi A., Ullrich K.K., Perroud P.F., Schallenberg-Rudinger M., Rensing S.A. (2017b) Combination of the endogenous lhcsr1 promoter and codon usage optimization boosts protein expression in the moss *Physcomitrella patens*. *Frontiers in Plant Science*, **8**, 1842.

Hohe A., Rensing S.A., Mildner M., Lang D., Reski R. (2002) Day length and temperature strongly influence sexual reproduction and expression of a novel MADS-box gene in the moss *Physcomitrella patens*. *Plant Biology*, **4**, 595–602.

Kamisugi Y., Schlink K., Rensing S.A., Schween G., von Stackelberg M., Cuming A.C., Reski R., Cove D.J. (2006) The mechanism of gene targeting in *Physcomitrella patens*: homologous recombination, concatenation and multiple integration. *Nucleic Acids Research*, **34**, 6205–6214.

Kamisugi Y., von Stackelberg M., Lang D., Care M., Reski R., Rensing S.A., Cuming A.C. (2008) A sequence-anchored genetic linkage map for the moss, *Physcomitrella patens*. *The Plant Journal*, **56**, 855–866.

Knop W. (1868) *Der Kreislauf des Stoffs: Lehrbuch der Agricultur-Chemie*. Haessel, H., Leipzig, Germany.

Lang D., Ullrich K.K., Murat F., Fuchs J., Jenkins J., Haas F.B., Piednoel M., Gundlach H., Van Bel M., Meyberg R., Vives C., Morata J., Symeonidi A., Hiss M., Muchero W., Kamisugi Y., Saleh O., Blanc G., Decker E.L., van Gessel N., Grimwood J., Hayes R.D., Graham S.W., Gunter L.E., McDaniel S., Hoernstein S.N.W., Larsson A., Li F.W., Perroud P.F., Phillips J., Ranjan P., Rokshar D.S., Rothfels C.J., Schneider L., Shu S., Stevenson D.W., Thummler F., Tillich M., Villarreal A.J., Widiez T., Wong G.K., Wymore A., Zhang Y., Zimmer A.D., Quatrano R.S., Mayer K.F.X., Goodstein D., Casacuberta J.M., Vandepoele K., Reski R., Cuming A.C., Tuskan J., Maumus F., Salse J., Schmutz J., Rensing S.A. (2018) The *P. patens* chromosome-scale assembly reveals moss genome structure and evolution. *The Plant Journal*, **93**, 515–533.

Lavy M., Prigge M.J., Tao S., Shain S., Kuo A., Kirchsteiger K., Estelle M. (2016) Constitutive auxin response in *Physcomitrella* reveals complex interactions between Aux/IAA and ARF proteins. *eLife*, **5**, e13325.

Moody L.A., Kelly S., Coudert Y., Nimchuk Z.L., Harrison C.J., Langdale J.A. (2018a) Somatic hybridization provides segregating populations for the identification of causative mutations in sterile mutants of the moss *Physcomitrella patens*. *New Phytologist*, **218**, 1270–1277. https://doi.org/10.1111/nph.15069.

Moody L.A., Kelly S., Rabbinowitsch E., Langdale J.A. (2018b) Genetic regulation of the 2D to 3D growth transition in the moss *Physcomitrella patens*. *Current Biology*, **28**, 473–478.

Ortiz-Ramírez C., Michard E., Simon A.A., Damineli D.S.C., Hernández-Coronado M., Becker J.D., Feijó J.A. (2017) Glutamate receptor-like channels are essential for chemotaxis and reproduction in mosses. *Nature*, **549**, 91–95.

Perroud P.F., Quatrano R.S. (2008) BRICK1 is required for apical cell growth in filaments of the moss *Physcomitrella patens* but not for gametophore morphology. *The Plant Cell*, **20**, 411–422.

Perroud P.F., Cove D.J., Quatrano R.S., McDaniel S.F. (2011) An experimental method to facilitate the identification of hybrid sporophytes in the moss *Physcomitrella patens* using fluorescent tagged lines. *New Phytologist*, **191**, 301–306.

Perroud P.-F., Haas F.B., Hiss M., Ullrich K.K., Alboresi A., Amirebrahimi M., Barry K., Bassi R., Bonhomme S., Chen H., Coates J., Fujita T., Guyon-Debast A., Lang D., Lin J., Lipzen A., Nogué F., Oliver M.J., León I.P.d., Quatrano R.S., Rameau C., Reiss B., Reski R., Ricca M., Saidi Y., Sun N., Szövényi P., Sreedasyam A., Grimwood J., Stacey G., Schmutz J., Rensing S.A. (2018) The *Physcomitrella patens* gene atlas project: large-scale RNA-seq based expression data. *The Plant Journal*, In press. https://doi.org/10.1111/tpj.13940.

Rensing S.A., Lang D., Zimmer A.D., Terry A., Salamov A., Shapiro H., Nishiyama T., Perroud P.F., Lindquist E.A., Kamisugi Y., Tanahashi T., Sakakibara K., Fujita T., Oishi K., Shin I.T., Kuroki Y., Toyoda A., Suzuki Y., Hashimoto S., Yamaguchi K., Sugano S., Kohara Y., Fujiyama A., Anterola A., Aoki S., Ashton N., Barbazuk W.B., Barker E., Bennetzen J.L., Blankenship R., Cho S.H., Dutcher S.K., Estelle M., Fawcett J.A., Gundlach H., Hanada K., Heyl A., Hicks K.A., Hughes J., Lohr M., Mayer K., Melkozernov A., Murata T., Nelson D.R., Pils B., Prigge M., Reiss B., Renner T., Rombauts S., Rushton P.J., Sanderfoot A., Schween G., Shiu S.H., Stueber K., Theodoulou F.L., Tu H., Van de Peer Y., Verrier P.J., Waters E., Wood A., Yang L., Cove D., Cuming A.C., Hasebe M., Lucas S., Mishler B.D., Reski R., Grigoriev I.V., Quatrano R.S., Boore J.L. (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64–69.

Reski R., Abel W.O. (1985) Induction of budding on chloronemata and caulonemata of the moss, *Physcomitrella patens*, using isopentenyladenine. *Planta*, **165**, 354–358.

Sakakibara K., Reisewitz P., Aoyama T., Friedrich T., Ando S., Sato Y., Tamada Y., Nishiyama T., Hiwatashi Y., Kurata T., Ishikawa M., Deguchi H., Rensing S.A., Werr W., Murata T., Hasebe M., Laux T. (2014) WOX13-like genes are required for reprogramming of leaf and protoplast cells into stem cells in the moss *Physcomitrella patens*. *Development*, **141**, 1660–1670.

Sanchez-Vera V., Kenchappa C.S., Landberg K., Bressendorff S., Schwarzbach S., Martin T., Mundy J., Petersen M., Thelander M., Sundberg E. (2017) Autophagy is required for gamete differentiation in the moss *Physcomitrella patens*. *Autophagy*, **13**, 1–13.

Schneider C.A., Rasband W.S., Eliceiri K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, **9**, 671–675.

von Stackelberg M., Rensing S.A., Reski R. (2006) Identification of genic moss SSR markers and a comparative analysis of twenty-four algal and plant gene indices reveal species-specific rather than group-specific characteristics of microsatellites. *BMC Plant Biology*, **6**, 9.

Stevenson S.R., Kamisugi Y., Trinh C.H., Schmutz J., Jenkins J.W., Grimwood J., Muchero W., Tuskan G.A., Rensing S.A., Lang D., Reski R., Melkonian M., Rothfels C.J., Li F.-W., Larsson A., Wong G.K.S., Edwards T.A., Cuming A.C. (2016) Genetic analysis of *Physcomitrella patens* identifies ABSCISIC ACID NON-RESPONSIVE, a regulator of ABA responses unique to basal land plants and required for desiccation tolerance. *The Plant Cell*, **28**, 1310–1327.

Ulfstedt M., Hu G.-Z., Johansson M., Ronne H. (2017) Testing of auxotrophic selection markers for use in the moss *Physcomitrella* provides new insights into the mechanisms of targeted recombination. *Frontiers in Plant Science*, **8**, 1850.

Vesty E.F., Saidi Y., Moody L.A., Holloway D., Whitbread A., Needs S., Choudhary A., Burns B., McLeod D., Bradshaw S.J., Bae H., King B.C., Bassel G.W., Simonsen H.T., Coates J.C. (2016) The decision to germinate is regulated by divergent molecular networks in spores and seeds. *New Phytologist*, **211**, 952–966.

Zimmer A.D., Lang D., Buchta K., Rombauts S., Nishiyama T., Hasebe M., Van de Peer Y., Rensing S.A., Reski R. (2013) Reannotation and extended community resources for the genome of the nonseed plant *Physcomitrella patens* provide insights into the evolution of plant gene structures and functions. *BMC Genomics*, **14**, 498.

## 8.4 Identification of genes involved in *Physcomitrella patens* sexual reproduction

To identify genes which are involved in the process of sexual reproduction in *P. patens*, publicly available array data have been used. The archegonial data set published by (Ortiz-Ramírez *et al.*, 2016) was used to select genes expressed during sexual reproduction. An expression level cut-off was applied, based on the expression level of the low abundance KNOX TF mkn2 (relative expression (RE) >= 750 (Sakakibara *et al.*, 2008)). Second, only genes expressed exclusively in the archegonial data set were kept by selection against expression in all other tissues from (RE < 750, (Ortiz-Ramírez *et al.*, 2016)) and the RNA-seq data by (reads per kilo base per million mapped reads (RPKM) < 2, (Perroud *et al.*, 2018)). Additionally, all genes showing no homolog in *M. polymorpha* were discarded, to get rid of putative artefacts and to positively select conserved genes.

Finally, 15 candidate genes, showing the highest expression in the archegonial sample were selected. Nine additional candidate genes found via literature research were added, to analyse their expression pattern. cDNA, obtained from juvenile and adult gametophore apices of the ecotypes Gd and Re, was used to confirm *in vivo* expression in *P. patens* reproductive tissues via RTPCR, and for five selected candidate genes via qPCR. Sample preparation, RTPCR, qPCR and qPCR expression normalization employing act5 (Pp3c10_17070V3.1) as reference gene was carried out as previously described ((Hiss *et al.*, 2017; Meyberg *et al.*, 2019) Fig. 5, chapter 12). The qPCR results confirmed the functionality of the chosen candidate gene approach. The five selected candidate genes showed nearly no expression in juvenile, but expression in adult gametophore apices whereas Re showed higher expression levels in general when compared Gd. Thus, the chosen candidate genes might be involved in the described fertility reduction of Gd.

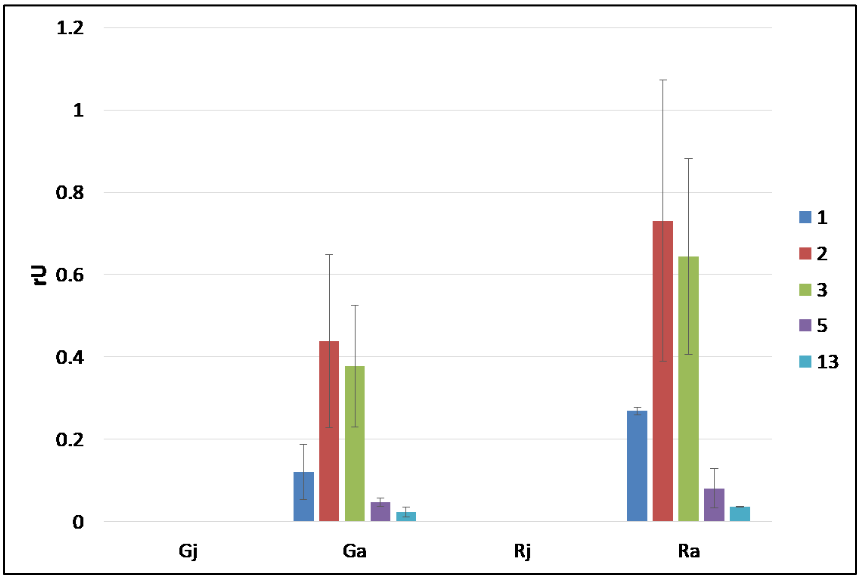


**Figure 5:** *Gene expression of selected candidate genes in Re (R) and Gd (G) juvenile (j) and adult (a) gametophore apices. Average expression of two independent biological replicates was used and normalized to the reference gene act5. Expression of chosen candidates show nearly no expression in the juvenile but expression in the adult samples, whereas Re shows overall a slightly higher gene expression compared to Gd. Error bars: standard deviation. rU: relative units.*

To learn more about the putative gene function and differences between both ecotypes, the five selected candidate genes were analyzed *in silico.* SNPs (Hiss *et al.*, 2017), epigenetic marks (Widiez *et al.*, 2014; Lang *et al.*, 2018) and the functional annotation (phytozome.org) were compared. The selected candidate gene function was chosen to be determined via a reverse genetic approach and for all candidates, KO mutants were generated as recently described (Meyberg *et al.*, 2019, chapter 12). For the further work, only candidate gene 13, namely the coiled-coil domain containing protein 39 (ccdc39) was used (chapter 6.5).

**Table 1:** Candidate gene overview. qPCR expression marked in green for Gd and Re juvenile (j) and adult (a) apices. SNPs (Gd vs. Re), no expression marked in red. Presence of DNA methylation in Gd and differences in histone marks between protonema and juvenile gametophores marked in green, red marks absence. X: no data available.

| No. | qPCR | | | | SNPs Gd vs. Re | DNA methylation | Different histone marks | Description |
|---|---|---|---|---|---|---|---|---|
| | Gd j | Re j | Gd a | Re a | | | | |
| 1 | 🟥 | 🟥 | 🟩 | 🟩 | intergenic | 🟩 | 🟥 | Leucine-rich repeat,<br>N-terminal domain |
| 2 | 🟥 | 🟥 | 🟩 | 🟩 | intergenic | 🟥 | X | 17,6 kDa heat shock protein |
| 3 | 🟥 | 🟥 | 🟩 | 🟩 | intergenic | 🟥 | 🟩 | Spiral1-like 2-related |
| 5 | 🟥 | 🟥 | 🟩 | 🟩 | 🟥 | 🟥 | 🟩 | Late embryogenesis abundant protein |
| 13 | 🟥 | 🟥 | 🟩 | 🟩 | synonymous | 🟥 | 🟥 | Coiled-coil domain-containing protein |

## 8.5 Characterization of evolutionary conserved key players affecting eukaryotic flagellar motility and fertility using a moss model

To finally understand the infertile Gd phenotype, work from the three presented publications was combined: Hiss *et al.,* 2017 set a basis by the quantification of the lack of developed sporophytes in Gd compared to Re and Vx and supplied this publication with the genetic divergence between the analyzed ecotypes. The Gd DNA methylation data which was published in Lang *et al.,* 2018 is now comparatively analyzed with the corresponding methylome of Re. Finally, a crossing analysis was employed to pinpoint the impairment of Gd using the Re-mCherry strain which was introduced in Perroud *et al.*, 2019.

As previously introduced, **flagella are an ancestral feature of eukaryotes** and were already present in the LECA. Independent of the organisms motility, flagella are required for sexual reproduction in most eukaryotic species and **defects within flagella are often associated with infertility**. Bryophytes belong to the clade of flagellated plants and possess bi-flagellated male gametes. As previously reported, several laboratories including our laboratory, recognized fertility issues with the most frequently used ecotype Gd and it was shown that the number of sporophytes developed per gametophore is significantly reduced in Gd (Hiss et al., 2017). Additionally, only little was known about genes important during *P. patens* sexual reproduction. Thus, morphological and molecular phenotyping of Gd, as well as a candidate gene approach (chapter 6.4) were performed, in order to identify genes involved in *P. patens* sexual reproduction and to characterize the Gd impairment. Crossing experiments with the previously introduced marker strain Re:mCherry was employed to show **Gd archegonia are fully functional** and subsequently the male defective phenotype was analyzed. Comparative genetic, epigenetic as well as expression analysis identified a set of genes required for male fertility in algae and mammals, showing a **conservation of flagellar genes across kingdoms**. A conserved coiled-coil domain 39 containing protein was identified to be important for flagellar assembly in *P. patens* and also to contribute to the Gd phenotype. This study does not only show Gds impairment within several flagella associated genes, probably due to **accumulation of somatic (epi-) mutations** during long vegetative reproductive periods, but also demonstrates, how ***P. patens* can be used as an easily accessible model system to study genes involved in male defects**.

# Characterization of evolutionarily conserved key players affecting eukaryotic flagellar motility and fertility using a moss model

Rabea Meyberg[1], Pierre-François Perroud[1], Fabian B. Haas[1], Lucas Schneider[1,5], Thomas Heimerl[4], Karen Renzaglia[3], Stefan A. Rensing[1,2,4]

[1] Plant Cell Biology, Faculty of Biology, University of Marburg, Karl-von-Frisch Str. 8, 35043 Marburg, Germany

[2] BIOSS Centre for Biological Signalling Studies, University of Freiburg, Freiburg, Germany

[3] Plant Cell Biology, Southern Illinois University, Carbondale, USA

[4] LOEWE Center for Synthetic Microbiology (SYNMIKRO), Philipps University of Marburg, Germany

[5] Current address: Institute for Transfusion Medicine and Immunohematology, Goethe-University and German Red Cross Blood Service, Frankfurt am Main, Germany

Keywords

*Physcomitrella patens*, moss, male infertility, flagella, spermatozoid, sperm, cilia

## Abstract

Defects in flagella/cilia are often associated with infertility and disease. Motile male gametes (sperm cells) with flagella are an ancestral eukaryotic trait that has been lost in several lineages, for example in flowering plants. Here, we made use of a phenotypic male fertility difference between two moss (*Physcomitrella patens*) strains to explore spermatozoid function. We compare genetic and epigenetic variation as well as expression profiles between the Gransden and Reute strain to identify a set of genes associated with moss male infertility. Defects in mammal and algal homologs of these genes coincide with a loss of fertility, demonstrating the evolutionary conservation of flagellar function related to male fertility across kingdoms. As a proof of principle, we generated a loss-of-function mutant of a coiled-coil domain containing 39 (ccdc39) gene that is part of the flagellar hydin network. Indeed, the *Ppccdc39* mutant resembles the male infertile Gransden strain phenotype. Potentially, several somatic (epi-)mutations occurred during prolonged vegetative propagation of *P. patens* Gransden, causing regulatory differences of e.g. the homeodomain transcription factor BELL1. Probably these somatic changes are causative for the observed male fertility. We propose that *P. patens* spermatozoids might be employed as an easily accessible system to study male infertility of human and animals.

## Introduction

### *Motile plant gametes allow studying sperm cell fertility*

Motile (flagellated) gametes are an ancestral character of eukaryotes (Mitchell 2007, Stewart *et al.* 1975) that has been secondarily lost in lineages such as the Zygnematales (Transeau 1951), which reproduce via conjugation of aplanogametes, and flowering plants, in which pollen transport non-motile gametes to the female (Renzaglia *et al.* 2001, Southworth *et al.* 1997). Flagella play an important role in sexual reproduction. In the unicellular algal model organism *Chlamydomonas reinhardtii* flagella are responsible for motility of the organism and serve during sexual reproduction to mediate a species-specific adhesion between two cells of different mating types (van den Ende *et al.* 1990). In streptophyte algae (Streptophyta comprise charophycean green algae as well as land plants), male gametes (spermatozoids) are the only motile cells of sessile multicellular algae, swimming to the oogonia to fertilize the egg cell (Hackenberg *et al.* 2019, McCourt *et al.* 2004). After the water-to-land-transition of plants this system was retained, i.e. motile (flagellated) spermatozoids require water to swim to the egg cell. Flagella of spermatozoids from flagellated organisms show a common architecture (Carvalho-Santos *et al.* 2011) and were already present in the most recent common ancestor (MRCA) of land plants (Stewart *et al.* 1975). During land plant evolution, loss of motile sperm occurred in seed plants, after the evolution of cycads and *Ginkgo* (both of which have pollen and motile male gametes), probably in the MRCA of conifers, Gnetales and flowering plants (Renzaglia *et al.* 2000, Renzaglia *et al.* 2001). Flagellated plants, that have retained motile spermatozoids, are an easily accessible system to study the fertility and other characteristics of male gametes. The moss *Physcomitrella patens* is particularly attractive in that regard because it develops spermatozoids in superficial male sex organs (antheridia) on an independent generation (gametophyte) that is readily cultured and manipulated (Cove 2005, Landberg *et al.* 2013).

### *The moss model* Physcomitrella patens

Bryophytes comprise the mosses, liverworts and hornworts and probably represent the monophyletic sister clade to vascular plants (Puttick *et al.* 2018). Due to this informative phylogenetic position they are increasingly being used to address evolutionary-developmental questions, e.g. (Aya *et al.* 2011, Horst *et al.* 2016, Sakakibara *et al.* 2013, Sakakibara *et al.* 2008), but also for functional studies of mammalian homologs in easily accessible model organisms (Ortiz-Ramirez *et al.* 2017, Sanchez-Vera *et al.* 2017). *P. patens* is a well-developed functional genomics moss model system. The genome was published in 2008 (Rensing *et al.* 2008) and recently updated to chromosome scale (Lang *et al.* 2018). Collections of worldwide accessions are available (Beike *et al.* 2014), of which some are likely to constitute ecotypes (Lang *et al.* 2018). The predominantly used ecotype Gransden (Gd) was derived from a single spore isolate collected 1962 in Gransden Wood (UK). The Gd cultures used in most labs worldwide were typically propagated vegetatively and several labs reported fertility issues over the past decades (Ashton *et al.* 2000, Hiss *et al.* 2017, Landberg *et al.* 2013, Perroud *et al.* 2011), potentially the result of somatic (epi-)mutations through decades of vegetative propagation (Ashton *et al.* 2000). The ecotype Reute (Re) collected 2006 in Reute (Germany) has a low genetic divergence to Gd, is highly self-fertile and thus has recently been introduced as a Gd alternative to enable studies involving sexual reproduction (Hiss *et al.* 2017).

### *P. patens sexual reproduction*

*P. patens'* sexual reproduction is initiated (under natural conditions in autumn) when day length shifts towards short days and temperature drops (Engel 1968, Hohe *et al.* 2002, Nakosteen *et al.* 1978). On the tip (apex) of the leafy gametophores the apical stem cell produces the antheridium initial stem cell (Kofuji *et al.* 2018), which subsequently gives rise to a bundle of antheridia, the male reproductive organs (Fig. 1). Each antheridium consists of a single outer cell layer or jacket surrounding a mass of spermatogenous tissue that produces up to 160 motile spermatozoids. Upon maturity the tip cell of the antheridia swells and bursts (when moistened by water) to release mature, swimming male gametes (spermatozoids). A few days after initiation of antheridia development, the female gametangia (archegonia) start to develop (Kofuji *et al.* 2009). They comprise a flask shaped egg-containing venter and elongated neck that opens following

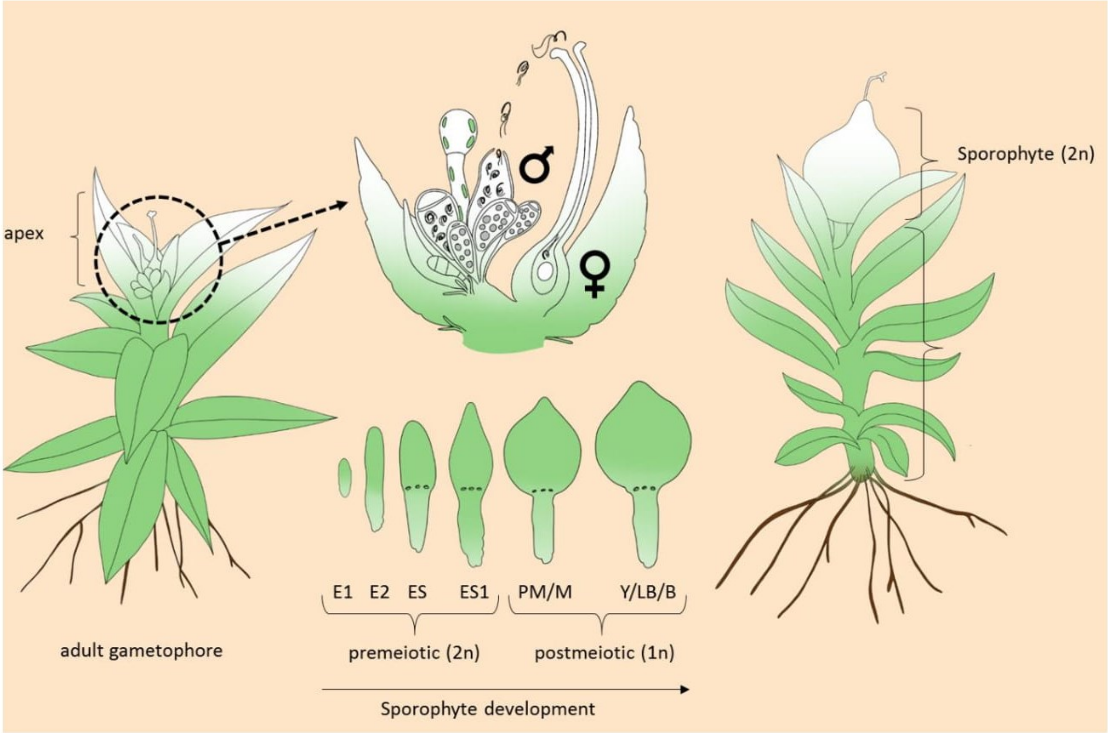


**Figure 1: Illustration of *P. patens* sexual reproduction.**
Upon environmental stimulus, reproductive organs (gametangia) develop on the apex of each gametophore, archegonia (female) and antheridia (male). Mature antheridia release the motile (flagellated) spermatozoids upon watering. The sperm cells swim through the archegonial venter to fertilize the egg cell. After fertilization, embryo development (E1/E2) and sporophyte development (ES-B) occurs. The mature sporophyte is located on the apex of the gametophore and releases haploid spores of the next generation. *P. patens* is predominantly selfing. Embryo/sporophyte developmental stages according to Hiss et al. (2017).

degeneration of the neck canal cells upon maturity, allowing swimming spermatozoids to reach the egg cell (Hiss *et al.* 2017, Landberg *et al.* 2013, Sanchez-Vera *et al.* 2017). After fertilization, the diploid zygote undergoes embryogenesis and develops into the sporophyte, which eventually generates haploid spores via meiosis (Fig. 1). Considering the difference in fecundity of Gd and Re (Hiss *et al.* 2017) we aim here to determine the genes underlying this difference, employing *P. patens* as a model for studying sperm fertility. For this purpose, we analysed the nearly self-sterile ecotype Gransden (Gd) in comparison with the highly fertile ecotype Reute (Re), using genetic and epigenetic variation data, fertility phenotyping as well as transcript profiling. Based on publicly available array data of developmental stages during sexual reproduction and network analyses, we generated a deletion mutant of a gene exclusively expressed in sexual reproductive organs. The mutant displays a loss of male fertility similar to that observed in Gd, and to that in human/mammal mutants of the ortholog. Our study demonstrates that the moss male germ line can serve as a model for flagellar dysfunction disease.

## Results & Discussion

### *The Gransden ecotype displays a reduced male fertility and a defect in spermatozoid motility*

The ecotype Reute (Re) was shown to develop a significantly higher number of sporophytes per gametophore in comparison to the broadly used ecotype Gransden (Gd, (Hiss *et al.* 2017)). To determine whether the reduced fertility of Gd is based on the female or male reproductive apparatus, crossing analyses between Gd and Re were performed, using the fluorescent marker strain Re-mCherry (Perroud *et al.* 2011, Perroud *et al.* 2019). Re developed on average 100% sporophytes per gametophore, of which 0.79% were crosses (Fig. 2A).

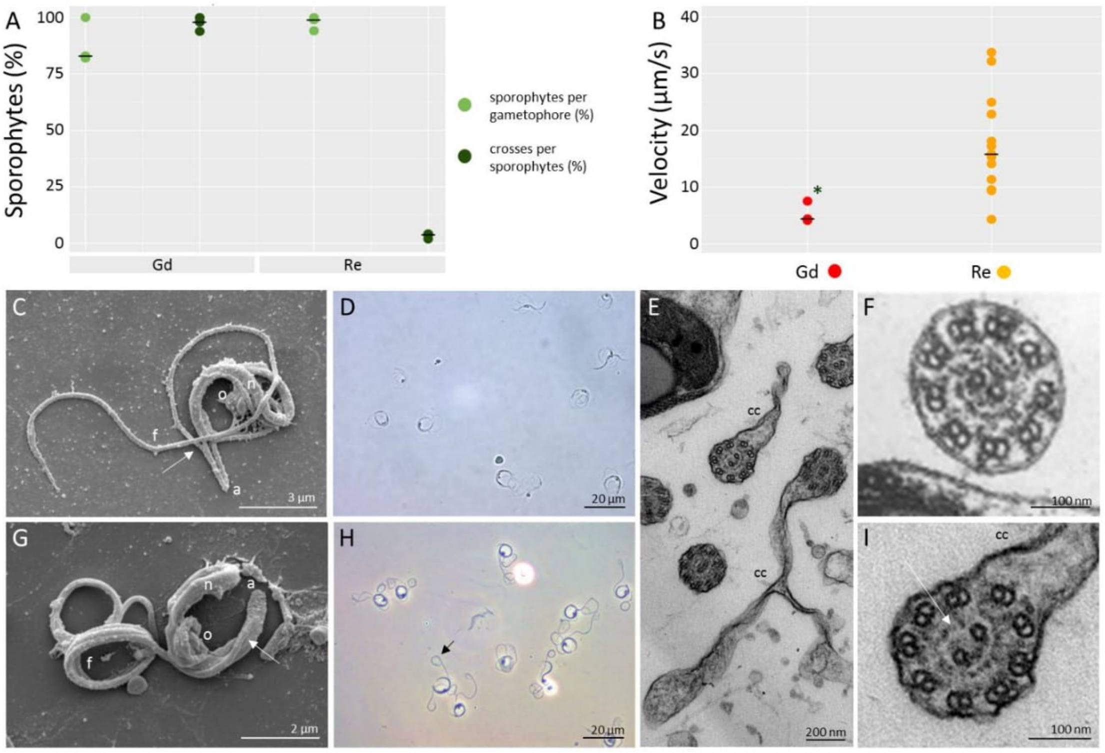


**Figure 2: Comparative analysis of Gd and Re male reproductive apparatus.**

A: Crossing analysis between Re (n = 430), Gd (n = 300) and fluorescent marker strain Re-mcherry. Re develops 100% of sporophytes per gametophore of which 0.79% are crosses. Gd develops 88.13% sporophytes per gametophore of which 99.69% are crosses. Median represented by black line. Significance shown by asterisk (Chi-square test, $p < 0.01$).
B: Re (n = 10) and Gd (n = 3) spermatozoids differ significantly in velocity (two-sided t-test, $p < 0.01$,*). Median represented by black line.
C: SEM of Re sperm cell showing cell architecture. From the cell anterior (a), the two flagella emerge (arrow) at staggered locations from the coiled nucleus (n). Organelles (o) include one mitochondrion and one plastid that attach to the mid-section of the nucleus. One of two long flagella (f) is visible and extends beyond the cell proper.
D: Phase contrast image of swimming Re spermatozoids.
E: TEM of mature Gd axonemes. Cytoplasmic connections (cc) appear on irregular flagella.
F: TEM of a mature spermatozoid of Re. The axoneme exhibits clear outer dynein arms and the plasmalemma is closely associated with the nine doublets.
G: SEM of Gd spermatozoid with architecture as in Re spermatozoids in C, except the flagella remain coiled.
H: Phase contrast image of swimming Gd spermatozoids showing loops at the posterior end of the flagella (arrow).
I: TEM of mature axonemes of Gd spermatid showing missing parts of the central pair microtubuli projections (white arrow) and cytoplasmic connections (cc).

This rate is expected, since *P. patens* is known to predominantly self-fertilize (Perroud *et al.* 2011). In comparison, Gd developed 88.13% sporophytes per gametophore of which 99.69% were crosses (significantly more than Re: $p < 0.01$, t-test, Fig. 2A). Thus, Gd archegonia are fully functional and can develop comparable numbers of sporophytes when fertilized by a male fertile partner. Hence, the Gd male reproductive apparatus is impaired. Analysis of the spermatozoid number per antheridium showed no significant differences between Re (median 123) and Gd (median 125, Figure S1A). This is less than determined by an approximation method (Horst *et al.* 2017), counting DAPI stained spermatozoids not released from the antheridium. Motility measurements showed that only a small number of Gd spermatozoids are motile (Fig. S1B), and show a significantly reduced velocity (Re median: 15,75 µm/s; Gd median: 4,37µm/s; $p < 0.01$, t-test; Fig. 2B). After release of the spermatozoids through the bursting tip cells of the antheridia, Re spermatozoids started moving a few seconds after release, whereas Gd spermatozoids showed motility only rarely. With regard to their structure, Gd spermatozoids display significantly more spermatozoids with the ends of flagella remaining in a coil (90.2%) than Re (6.6%, t-test, $p < 0,01$, Fig. 2C, D, G, H, Fig. S1C). Interestingly, similar coiled flagella are also known from mouse and human infertile flagella (Dong *et al.* 2018, He *et al.* 2018, Tang *et al.* 2017). There are ultrastructural differences between Re and Gd sperm cells. Two cylindrical flagella develop around the outside of the spermatozoid as the nucleus condenses and elongates (Fig. 2C, D, G, H). The axonemes of both ecotypes exhibit the typical nine doublets and two central pair microtubules that characterize flagella of most eukaryotes (Fig. 2E, F, I). Axonemes in Re mature spermatozoids demonstrate visible outer dynein arms and protein projections at the central pair of microtubuli (Fig. 2F). In contrast to Re (Fig. 2C, D), gametes of Gd appear to arrest in the final stages of flagellar elongation and thus the posterior coils of the flagella fail to individualize, resulting in the coiled posterior loop (Fig. 2G, H). Mature Gd axonemes developed cytoplasmic connections, which probably result in the coiled posterior loop of the flagella. Additionally the protein projections around the central pair of microtubules seems to lack some proteins (Fig. 2 E,I).

*Genes related to flagellar assembly and motility harbour (epi-)mutations between Gd and Re*

Similar to motile sperm, DNA methylation is an ancestral eukaryotic feature (Feng *et al.* 2010) and is supposed to regulate gene expression on the DNA level (Zemach *et al.* 2010). In *P. patens*, methylated gene bodies usually are associated with lower gene expression (Lang *et al.* 2018) and loss of the DNA methyltransferase PpMET1, which is involved in CG DNA methylation of gene bodies, inhibits sporophyte development (Yaari *et al.* 2015). Recently, it has been shown, that the male reproductive organs undergo severe changes in DNA methylation in the liverwort *M. polymorpha* during sexual reproduction (Schmid *et al.* 2018). Hence, we expected DNA methylation to play a role during regulation of sexual reproduction in moss. Whole genome bisulfite sequencing (bs-seq) of Re and Gd adult gametophores (bearing gametangia; Fig. 1) was performed. Differentially methylated positions (DMPs) in all three methylation contexts (CHG, CHH, CG) were determined. In total, 671 genes harboring DMPs were found (Table S1). GO bias analysis of the genes containing DMPs showed enriched terms related to cilium movement and motile cilium assembly, as well as protein and macromolecule modification, which includes post-translational modifications like ubiquitination (Fig. S2A). Using the intersect of genes that contain DMPs and single nucleotide polymorphisms (SNPs, Table S2, (Hiss *et al.* 2017)) the number of GO terms associated with cilia and microtubule based movement was found to be increased (from 3 to 6, Fig. 3A), suggesting genetic and epigenetic effects on the Gd phenotype. In the intersection, additional GO terms related to protein phosphorylation were found to be over-represented (Fig. 3A). In *C. reinhardtii,* it was shown that protein phosphorylation is a key event of flagellar disassembly (Pan *et al.* 2011) and that the phosphorylation state of an aurora-like protein kinase coincides with reduced flagellar length (Luo *et al.* 2011).

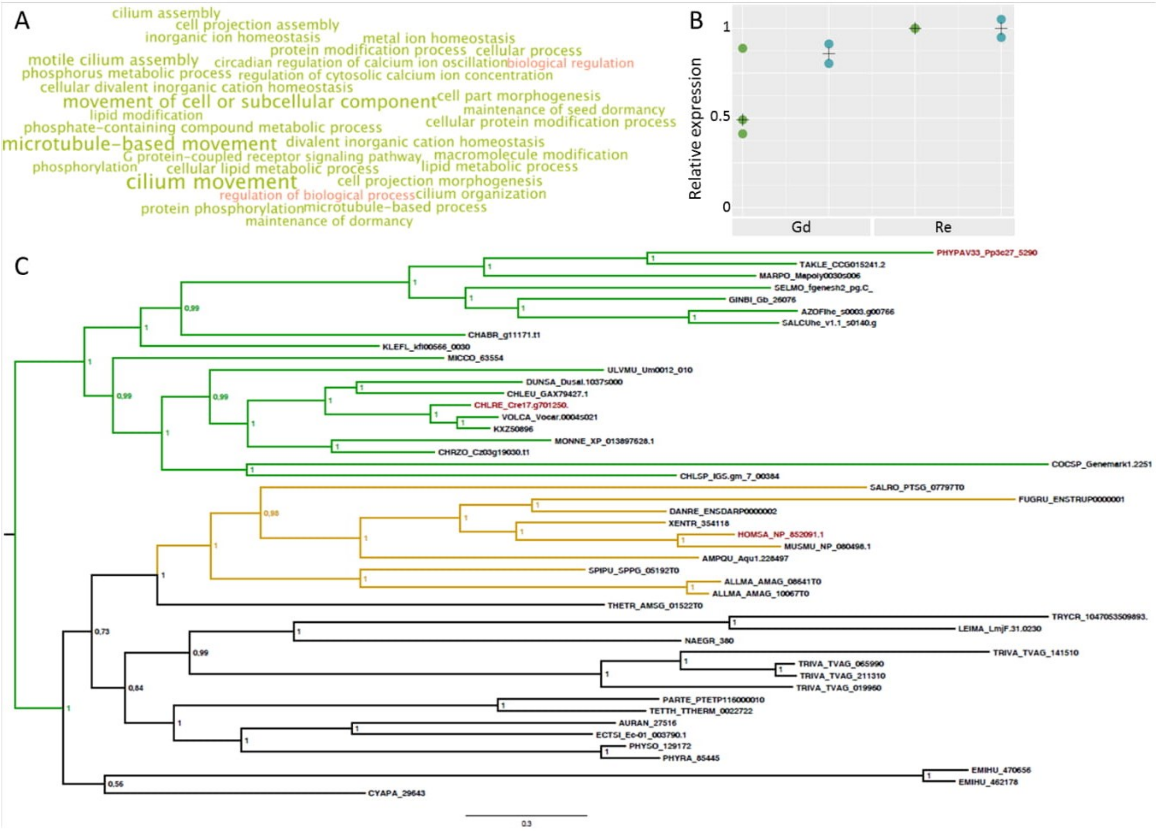


**Figure 3: DMP/SNP overlap and CCDC39.**

A: GO bias analysis of genes differentially methylated and SNP- containing in comparison of Gd vs. Re. The intersect of DMP and SNP affected genes shows over-representation of GO terms associated with cilia motility. Over-represented terms are shown in green, whereas under-represented terms are shown in red. Larger font size correlates with a higher significance level.

B: Relative expression of ccdc39 in qPCR (green, n = 3) and RNA-seq (blue, n = 2) data shows higher expression in Re in comparison to Gd. Median marked by black cross.

C: Phylogenetic analysis of CCDC39 shows clear correlation with the presence of flagella and generally reflects species relationships. Planta are shown in green, opisthokonts in yellow, protozoa, SAR and fungi in black. *Homo sapiens*, *Chlamydomonas reinhardtii* and *Physcomitrella patens* are marked in red.

### *RNA-seq of male reproductive organs suggests differential transcriptional regulation to be connected to infertility*

To gain deeper insight into gene expression differences, cDNA sequencing (RNA-seq) of Gd and Re antheridial bundles was performed. In total 19,218 genes were found to be expressed (RPKM

>= 2) in Gd and 19,254 in Re (Table S3/S4). 865 genes were uniquely expressed in Gd and 901 genes in Re (Table S5/S6). GO bias analysis of all genes expressed in antheridia of Gd (Fig. S2B) and Re (Fig. S2C) showed an over-representation of terms like amide biosynthetic process and peptide biosynthetic process. Spermine and spermidine belong to the amides and play a major role in male fertility in mammals (Lefevre *et al.* 2011) which, based on the RNA-seq GO analysis, also appears to be the case in *P. patens*. The few under-represented terms found are all connected to mRNA capping, potentially indicating that during gamete development mRNA capping could be performed by less characterized capping enzymes, or that translation occurs independently of mRNA capping. It was postulated that the first eukaryotic mRNA translation was probably driven by internal ribosomal entry sites (IRES) and that the cap-dependent (CD) mechanism evolved later. The former mechanism, according to this hypothesis, was retained as an additional regulation of the translation in specific situations like stress responses (Godet *et al.* 2019, Hernandez 2008). Also, cap-independent (CI) translation was recently suggested to be a robust partner of CD translation and to appear prevalently in germ cells (Keiper 2019), which matches with the presented results.

In human, transcriptional regulation is important to control spermatogenesis (Bettegowda *et al.* 2010). Within the expressed genes, in total 1,259 transcription associated proteins (TAPs; comprising transcription factors, TFs, and transcriptional regulators, TRs) could be identified using TAPscan (Wilhelmsson *et al.* 2017). Within the Re uniquely expressed genes , in total 74 TAPs of 27 different TAP families could be identified, including HD_BEL and MADS_MIKC$^c$ type proteins (Table S7). Interestingly, bell1 (HD_BEL) previously was described not to be expressed in antheridia (Horst *et al.* 2016), which indeed is true for Gd (used in the study by Horst *et al.*), but not for Re (Fig. S3A). Therefore, BELL1 might play a role in male fertility in *P. patens*. MADS-box genes are important for flower, pollen and fruit development (Theißen *et al.* 2016) and have previously been associated with sexual reproduction in *P. patens* (Hohe *et al.* 2002, Quodt *et al.* 2007, Singer *et al.* 2007). *P. patens* MIKC$^c$-type genes are important for motile flagella and external water conduction (Koshimizu *et al.* 2018). In the latter study, PPM1, PPM2 and PPMC6 are the key players for flagellar motility, but since the triple KO did not completely phenocopy the sextuple KO (*ppm1, ppm2, ppmads1, ppmc5, ppmc6, ppmadss*), the other genes apparently also play a role. While ppm1, ppm2 and ppmc6 are expressed in both, Re and Gd antheridia, mads1 (Hohe *et al.* 2002) and ppmc5 are exclusively expressed in Re, implying a putative involvement of these two genes in flagellar motility. Among the Gd uniquely expressed genes (Table S5), 40 TAPs belonging to 23 different TAP families were found, including MADS and HD proteins. TIM1 (Pp3c17_24040V3.1) is a type I MADS-box protein, basal to all other type I MADS-box proteins (Gramzow *et al.* 2010), whereas the HD gene duxa-like (Pp3c5_6470V3.1) is mainly expressed in lung and testis tissue in human, which supports a potential function in cilia/flagella (Booth *et al.* 2007).

*Analysis of differentially expressed genes between Gd and Re*

From all expressed genes in Gd and Re, a set of 110 differentially expressed genes (DEGs) could be identified using a stringent intersect of three tools. Detailed analysis of the DEGs showed many spermatozoid/sperm and pollen related genes (Table S8). Of the genes expressed significantly higher in Re, many showed no expression in Gd, but moderate to high expression in Re, e.g. the arl13b homolog (Pp3c1_40600V3.1), which plays a role in flagella and cilia stability and signaling via axonemal poly-glutamylation, and the protein phosphatase hydin homolog Pp3c3_14230V3.1 (Fig. S3B). The central microtubule pair protein hydin is a well analysed gene in human and algae, and connected to the human disease Primary Ciliary Dyskinesia (PCD). Mutations in this gene lead to the loss of flagellar motility and occasionally to loss of one or both central pair microtubules due to missing C2b projections (Lechtreck *et al.* 2007, Olbrich *et al.* 2012). Interestingly, the Gd ultrastructure showed missing protein projections (Fig. 2E,I) which could might result from the missing hydin expression.

Genes that are expressed significantly higher in Gd often lacked annotation (28/37 showed no annotation in the latest gene annotation v3.3 (Lang *et al.* 2018)) and hence might be lineage specific. Among the annotated genes more highly expressed in Gd were e.g. the membrane associated ring finger march1 homolog (Pp3c18_16700V3.1), which is known to negatively affect the sperm quality and quantity in Chinese Holstein bulls (Liu *et al.* 2017), and an arabinogalactan 31 homolog (Pp3c5_9210V3.1). Arabinogalactan proteins are known to be part of mucilage (Lord *et al.* 1992). Analysis of the mucilage content during antheridia maturation in the charophytic alga *Chara vulgaris* showed a reduction of mucilage upon maturity (Gosek *et al.* 1991). Arabinogalactan proteins are abundant in the matrix around fern spermatozoids during development and are virtually negligible when antheridia release spermatozoids (Lopez *et al.* 2018). Thus, an overexpression of an arabinogalactan gene could represent an impairment of the spermatozoid maturation process in the Gd background.

RNA-seq expression data of hydin and march1 have been confirmed by RT-PCR (Fig. S3B), and SNP and DNA methylation data have been analysed. The hydin gene body and potential promoter region (2kbp upstream of the coding sequence) displays six SNPs and one insertion or deletion (InDel) between Gd and Re. One SNP is located in the putative promoter region, three are intron variants and two are located within an exon, causing moderate amino acid changes (Table S9). In Gd, CHG and CHH context DMPs are present in the promoter and gene body, while a single CG mark is present in Re (Fig. S4). The lack of expression in Gd matches the presence of gene body methylation, shown to coincide with lack of expression (Lang *et al.* 2018). March1 does show very low levels of DNA methylation but two SNPs between Gd and Re in the 5'-UTR and five SNPs and one InDel in the putative promoter region, which could affect a potential regulatory function of the UTR (Fig. S5, Table S9). Network analysis was performed to gain more insights into putative protein interaction partners. The March1 network revealed proteins known to be involved in degradation of mis-folded proteins via ubiquitination, and modification of proteins via phosphorylation (Fig. S6A). Since protein phosphorylation and ubiquitination pathways are known to act together (Hunter 2007), march1 disregulation in Gd might be involved in the determined phenotype. The hydin network revealed many proteins involved in ciliary function and motility e.g. radial spoke head protein 9 (RSPH9, (Castleman *et al.* 2009) and CCDC39 (Merveille *et al.* 2011, Oda *et al.* 2014), Fig. S6B.

CCDC39 (Pp3c27_5290V3.1) was detected via a candidate gene approach because it shows expression only in adult gametophores (bearing gametangia), and a difference between Re and Gd (Fig. 3B). It is part of the hydin network, is a coiled-coil domain containing protein which, as well as hydin and rsph9, belongs to the genes that cause PCD when mutated (Antony *et al.* 2013, Horani *et al.* 2018). In human it is required for the assembly of inner dynein arms and the dynein regulatory complex (Merveille *et al.* 2011) and in *Chlamydomonas reinhardtii* it acts as a molecular ruler for the determination of the flagellar length (Oda *et al.* 2014). Phylogenetic analysis showed the presence of ccdc39 orthologs in all major kingdoms, providing evidence that it was already present in the MRCA of all eukaryotes (Fig. 3C). Interestingly, all species with ccdc39 orthologs also possess flagella, implicating a gene function unique to these motile organelles. Ccdc39 does not show amino acid sequence-affecting SNPs between Gd and Re in the gene body, and no SNPs in the promoter region (Table S9). The promoter region and gene body of ccdc39 display low levels of DNA methylation (Fig. S6). Interestingly, CCDC39 and MARCH1 network analysis showed connections to the same two protein phosphatases (Pp3c16_18360V3.1, Pp3c25_7050V3.1, Fig. S6A, C), implicating involvement of phosphatases in spermatogenesis of *P. patens*.

*Ccdc39 is essential for proper flagellar development in* P. patens

Expression analysis of ccdc39 via RNA-seq and qPCR in *P. patens* showed no expression in any tissue except for the antheridia, and a reduced expression level in Gd as compared to Re (Fig. 3C), suggesting a functional connection to the observed fertility phenotype. To elucidate the function of ccdc39 and to separate the phenotype from hydin function, a loss-of-function mutant was generated, hereafter referred to as *ccdc39*. Protoplast regeneration, protonemal and gametophytic growth and gametangia development did not reveal an obvious aberrant phenotype (Fig. S8, S9). However, no sporophytes were formed under selfing conditions (Fig. 4A, B yellow). 30 days after watering Re had developed mature brown sporophytes, whereas *ccdc39* apices showed several unfertilized archegonia and bundles of antheridia in a cauliflower-like structure (Fig. 4C). When no fertilization takes place, *P. patens* is known to continue gametangiogenesis, leading to the accumulation of multiple archegonia per gametophore (Landberg *et al.* 2013, Sanchez-Vera *et al.* 2017). Under crossing conditions using the male fertile Re-mCherry line (Perroud *et al.* 2019) a close to normal amount of sporophytes (94%) could be observed, with 100% of them being crosses (Fig. 4A, light and dark green), suggesting a defect in the male reproductive apparatus. Variable phenotypes of flagella could be detected in *ccdc39*. Some spermatozoids developed very short flagella, others normal length flagella, but with ends that remained in coils, comparable to the Gd phenotype (Fig. 4D-F). Pp*ccdc39* hence showed a somewhat different phenotype compared to the alga *C. reinhardtii*, in which a severe reduction

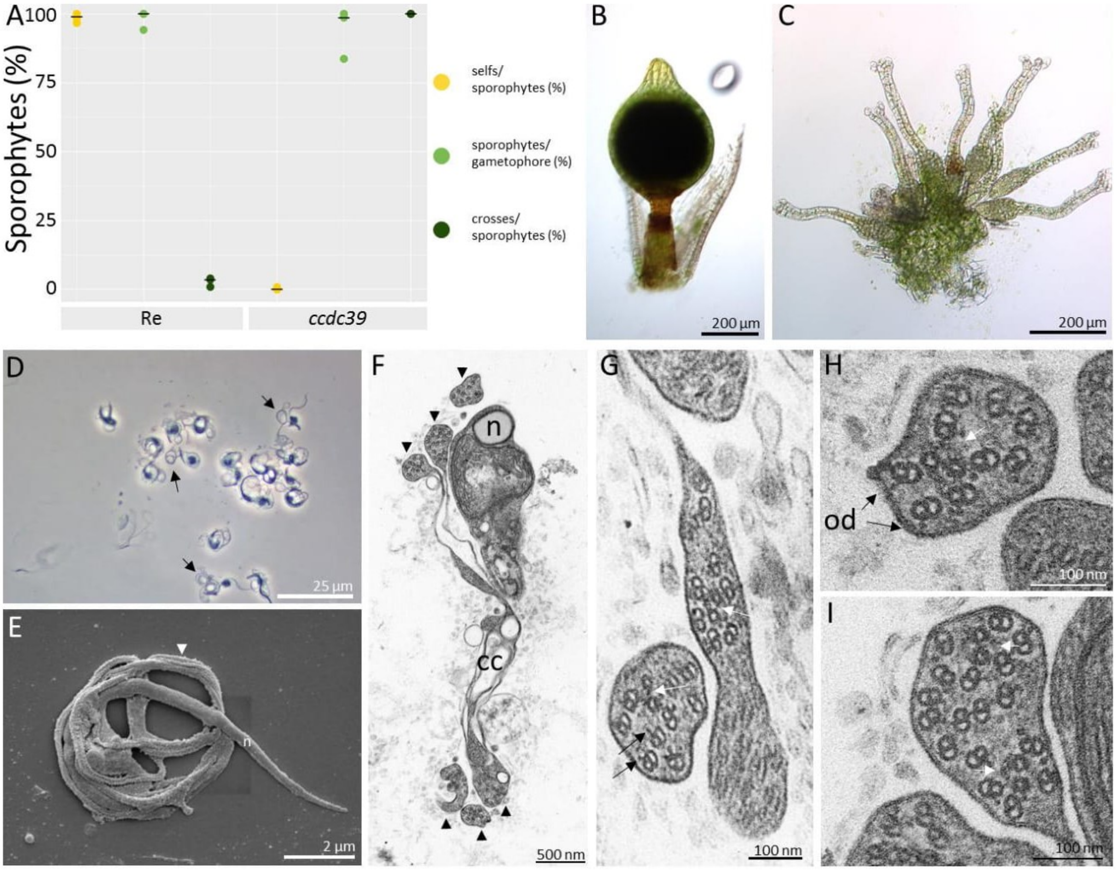


**Figure 4: Analysis of *P. patens ccdc39*.**
A: Selfing and crossing analysis between *ccdc39* (n = 630), Re (n = 331) and fluorescent marker strain Re-mcherry. *ccdc39* develops 0% and Re develops 99% sporophytes per gametophore under selfing conditions. Under crossing conditions, Re develops 98% of sporophytes per gametophore of which 0.03% are crosses. *ccdc39* develops 94% sporophytes per gametophore of which 100% are crosses (asterisk shows significant deviation, Chi-square test, p < 0.01). Median marked by black line.
Re (B) and *ccdc39* (C) apices 30 days after watering show properly developed sporophytes on Re apices and bundles of gametangia on *ccdc39* apices.
D: Phase contrast image of fixed mature *ccdc39* spermatozoids displaying posterior flagellar loops (arrows).
E: SEM of *ccdc39* spermatozoid that remains in coils. The nucleus (n) is a long coiled cylinder that narrows posteriorly. Irregular flagella coil around the cell (arrowhead).
F: TEM of whole mature spermatozoid in cross section with compacted nucleus (n) in upper coil. Arrowheads indicate irregular flagella that encircle the cell for over 2 revolutions. Cytoplasmic connections (cc) adjoin axonemes across their length. Bar= 500 nm.
G-I: TEM cross sections of irregular flagella showing intact acentric central pair complexes (white arrowheads), outer dynein arms (od, black arrows), and double axonemes in single flagellar shafts in G and I.

of the flagellar length occurred (Oda *et al.* 2014), an observation only occasionally seen in *P. patens ccdc39*. No *ccdc39* null mutant has been reported for *H. sapiens* yet, but cilia of epithelial cells of persons bearing mutations in ccdc39 showed a reduction of length in 2/4 cases (Merveille *et al.* 2011), implying flagellar length variations to be part of the phenotype, comparative to *P. patens* ccdc39. Interestingly, all three species share the loss of the inner arm dyneins upon mutation of ccdc39 ((Merveille *et al.* 2011, Oda *et al.* 2014); Fig.4 G-I), indicating a conserved function. Ccdc39 gametes undergo normal cellular morphogenesis into a streamlined coiled architecture, including nuclear compaction/elongation, similar to Re and Gd gametes (Fig. 4F). The locomotory apparatus develops normally and the two flagella elongate around the cell perimeter. However, axonemal organization is completely disrupted in the KO spermatozoids. The nine doublet microtubules are randomly arranged within an abnormally-shaped flagellar shaft that is rarely cylindrical (Fig. 4E-I). The central pair complex remains intact but it is not anchored in the center of the axoneme (Fig. 4F-I). Floating doublet microtubules show evidence of outer dynein but inner dynein is unclear (Fig. 4G, H). Typically, only a single complement of nine doublets and one central pair occurs within an axoneme, but the occasional two complements are visible within a single flagellar shaft (Fig. 4F, I). Cytoplasmic connections across coils provide evidence of incomplete separation of cellular coils that is likely responsible for the posterior loop in flagella evident in the light microscope (Fig. 4D). The cytoplasmatic connections observed in the *ccdc39* background are similar to those displayed by Gd (Fig. 2E), which also showed a reduced expression of ccdc39 in comparison to Re (Fig. S3B). Together these observations suggest a functional connection between ccdc39 expression and the mutant phenotype.

### ***Conclusions***

Our analyses suggest that the long term vegetative propagation of the Gd lab strain led to accumulation of several (epi-)mutations, affecting the male reproductive apparatus. Unfortunately, the original Gd collection site is not available anymore due to changes in land use. However, the Gd isolate was shipped to Japan (GdJp) in the early 90s and has been propagated involving regular sexual reproduction. To proof the assumption that the original Gd isolate had not been infertile, we analyzed the fertility of GdJp and found that the sporophyte development was significantly higher in comparison with the predominantly vegetatively propagated Gd (Fig. S10). Hence, the genetic and epigenetic differences observed between Gd and Re are probably due to somatic changes acquired during vegetative culture as previously proposed (Ashton *et al.* 2000). To avoid such accumulation of deleterious (epi-)mutations it appears advisable to regularly include sexual reproduction into the lab strain propagation, and to test the offspring for fertility.

Interestingly, known key players affecting cilia motility in mammal model organisms and human could be identified via their *P. patens* expression profile as well as through (epi-)mutations between Gd and Re. Some of these were characterized via transcript profiling, demonstrating biased expression between the ecotypes. Loss-of-function analysis of ccdc39 showed involvement in proper flagella formation in *P. patens* as well. Hence, flagellated plants (like mosses) apparently harbor male gametes with a flagellar architecture that is conserved with mammals, and thus could serve as easily accessible models for human disease or animal foodstock fertility associated with flagella. Whether only flagella or the process of male motile gamete development is evolutionarily conserved as well remains to be determined.

Our network analyses show that the components of the protein-protein interaction network of the eukaryotic flagellar structure are well conserved. This network is also supported by the loss-of-function mutant: ccdc39 was found via the network and expression analysis as well as through a candidate gene approach, and resembles the Gd phenotype in terms of male infertility. Most probably, given the lack of SNPs and DMPs in the underlying gene, Ppccdc39 is not the causative master mutation. Rather, it appears probable that one or several upstream regulators are affected, causing (among others) the mis-regulation of Ppccdc39. In this regard it is interesting to note that the homeodomain (HD) TALE TF BELL1, known to be involved in the alternation of generations (haploid to diploid phase transition) in algae and plants (Horst *et al.* 2016, Lee *et al.* 2008), shows expression in Re but not in Gd. Intriguingly, the heterodimeric interaction of HD TALE TFs observed in animals and plants has recently been suggested to be an ancestral eukaryotic feature to control the haploid-to-diploid transition only when gametes were correctly fused (Joo *et al.* 2018). Potentially, master regulators like BELL1 not only control zygote formation, but even earlier processes of sexual reproduction.

Primary Ciliary Dyskinesia (PCD) is a heterogeneous genetic defect that results in abnormal structure and function of cilia (Noone *et al.* 2004). In humans, the condition has a range of biological and clinical phenotypes that include respiratory infections, reduced fertility and situs inversus or developmental left-right inversion of organs. Ultrastructural observations of cilia in patients with PCD reveal perturbations of outer and inner dynein arms, radial spokes and the central pair complex, with 10% exhibiting normal axonemal structure (Papon *et al.* 2010). The conserved nature of the 9+2 core structure of cilia and flagella is particularly underlined in the present study as a universal phenotype of ccdc39 loss-of-function mutants, since it occurs in moss, dogs and humans (Merveille *et al.* 2011). All show axonemal disorganization and inner dynein arm aberrations. Functional analyses indicate that the ccdc39 gene is necessary to assemble the dynein regulatory complex and inner dynein arms. The striking defects evident in *ccdc39* moss spermatozoids are even more exaggerated than those in humans and dogs, perhaps due to the extreme length of flagellar axonemes compared with cilia. Nevers (Nevers *et al.* 2017) classified motility-associated genes responsible for ciliopathies such as PCD, including ccdc39, in a relatively small subset of ciliary genes necessary to construct a motile flagellum. These genes are found in all ciliated organisms, including eukaryotes such as land plants in which cilia are restricted to gamete-producing cell lineages. As an easily manipulated experimental system, *P. patens* provides unparalleled opportunities to systematically and individually examine gene structure, function and evolution in the core genes responsible for the development of the eukaryotic cilium/flagellum.

## Methods

### *Plant material*

*P. patens* ecotypes Gd (Rensing *et al.* 2008) and Reute (Hiss *et al.* 2017) were cultivated under the same conditions as described in (Hiss *et al.* 2017) for sporophyte development, and for crossing analysis as described in (Perroud *et al.* 2019). For gametophytic stage phenotyping, protonemal cells were placed on 9cm Petri dishes with solid minimal medium (1% w/v agar, Knop’s medium, (Knop 1868)) enclosed with 3M micropore tape and inoculated 7 days for protonema and in total 10 days for gametophore analysis in long-day (LD) conditions (70 µmol $m^{-2}$ $s^{-1}$, 16h light, 8h dark, 22°C). Gametangia harvest and analysis was performed at 21 days after short-day (SD, 20 µmol $m^{-2}$ $s^{-1}$, 8h light, 16h dark, 15°C) transfer.

*Counting and statistical analysis of sporophytes per gametophore under selfing and crossing conditions, data visualization*

To determine the number of sporophytes developed by ecotypes and knock-out mutant, counting of the selfed F1 of three independent mutant lines (*ccdc39#8*, *ccdc39#41*, *ccdc39#115*, Fig. S11) was performed at 30 days after watering as described in (Hiss *et al.* 2017) using a Leica S8Apo binocular (Leica, Wetzlar, Germany). Counting of the crosses was performed according to (Perroud *et al.* 2019) using a fluorescence stereomicroscope SteREO LumarV12 (Carl Zeiss, Oberkochen, Germany). For counting of the selfed F1, at least five independent biological replicates with in total 537 to 742 gametophores per sample were analysed. For the crossings, three biological replicates with in total 300 to 630 gametophores per sample were analysed. For the ecotype comparison between Re, Gd and GdJp three replicates with in total 523 to 730 gametophores were analysed. Statistical analysis was carried out using Microsoft Excel 2016 (Microsoft) and R. Plots in R were done using ggplot2 (Wickham 2016).

*Microscopic analysis of gametangia and sporophytes*

Harvest and preparation of sporophytes, gametangia and spermatozoids was performed with a Leica S8Apo binocular (Leica, Wetzlar, Germany). Microscopic images were taken with an upright DM6000 equipped with a DFC295 camera (Leica). For both devices the Leica Application suite version 4.4 was used as the performing software Images were processed (brightness and contrast adjustment) using Microsoft PowerPoint.

*Spermatozoid microscopy, DAPI staining and counting*

Preparation for 4',6-diamidino-2-phenylindole (DAPI) staining/flagellar analysis and counting of spermatozoids per antheridium was performed after 21 days of SD inoculation. Per sample, a single antheridium was harvested in 4 µl sterile tap water applied to an objective slide. The spermatozoids were released using two ultra-fine forceps (Dumont, Germany) and the sample was dried at room temperature (RT). Samples were fixed with 3:1 ethanol/acetic acid and after drying spermatozoid nuclei were stained applying 0,7 ng/µl DAPI (Roth, Germany) in tap water. The samples were sealed with nail polish (www.nivea.de). Microscopic images were taken with an upright DM6000 equipped with a DFC295 camera (Leica). Brightness and contrast of microscopy images was adjusted using Microsoft PowerPoint.

### *Electron microscopy*

For SEM images, mature spermatozoids (21 d after SD transfer) were harvested on an objective slide covered with polilysine. Chemical fixation was performed with 2.5% glutaraldehyde and an ascending alcohol preparation using 30, 50, 70, 90, and two times 100 % of EtOH was performed. Subsequently, the samples were critical point dryed (Tousimis, Rockville, USA) and coated with 20nm gold using a Leica Sputter (Leica, Wetzlar, Germany). The samples were then analysed with a SEM (Philips XL30, Hamburg, Germany). For TEM, adult apices (21 d after SD transfer) were used. Sample preparation was done as previously described with slightly modified infiltration (25% Epon (Fluka® Analytical, USA) in propylene oxide for 2h; 50% Epon in propylene oxide overnight; next day pure Epon and polymerization).

### *Nucleic acid Isolation*

Genomic DNA for genotyping was isolated with a fast extraction protocol, using one to two gametophores as described in (Cove *et al.* 2009). To determine candidate gene expression levels in juvenile (5 weeks LD growth) and adult (21 days after SD transfer) apices, each at least 25 apices were harvested in 20µl RNA later (Qiagen, Hilden, Germany). Tissue was stored at -20°C, until RNA was extracted using the RNeasy plant mini Kit (Qiagen). For RNA-seq RNA was isolated from 40 antheridia bundles, comprising 4-6 antheridia each, harvested at 21 day after SD transfer and directly stored in 20µl RNAlater. RNA was extracted with a combination of RNeasy plant mini and RNeasy micro kit (Qiagen). The RNeasy plant mini kit was used until step 4. Flowthrough of the QIAshredder spin column was diluted in 0,5 volume 100% EtOH and then transferred to the RNEasy spin columns of the micro kit for further treatment with the micro kit. RNA concentration and quality was analysed with Agilent RNA 6000 Nano Kit at a Bioanalyzer 2100 (Agilent Technologies).

*Real-time PCR (RT-PCR) & quantitative real-time PCR (qPCR)*

To determine gene expression in juvenile vs. adult, respectively Gd vs. Re adult apices, RNA was extracted as described above. cDNA synthesis was performed using the Superscript III (SSIII) kit (ThermoFisher Scientific) according to the manufacturers` protocol but using ½ of the suggested SSIII amount. Primers were designed manually with an annealing temperature +/- 60°C and a product length of ca. 300 bp. Single genomic locus binding properties were tested by using BLAST against the V3.3 genome of *P. patens*. Real-time PCR was performed using 5 ng of cDNA as input for PCR reaction with OneTaq from New England Biolabs). PCR products were visualized via gel electrophoresis using peqGREEN from (VWR, Germany). As size standard, the 100 bp ladder (NEB) was used. Real-time qPCR was carried out with two (for juvenile vs. adult comparison) or three (expression determination of ccdc39) biological replicates as published in Hiss et al., 2017. As reference gene, act5 (Pp3c10_17070V3.1, (Le Bail *et al.* 2013) was chosen, due to homogenous expression in juvenile and adult apices in Gd and Re (Fig. S12). For primer sequences see Table S10.

*Antheridia bundle RNA-seq & analysis*

RNA-library preparation and RNA-seq was performed at the MPI genome center Cologne (mpgc.mpipz.mpg.de). For each ecotype, two libraries were prepared with the ultra low input RNA-seq protocol followed by sequencing with Illumina HiSeq3000 (150 nt, single ended). For gene expression and DEG analysis for Gd 53,6 Mio. and for Re 56,4 Mio. uniquely mapped reads were used. The analysis was performed as previously described for the *Physcomitrella patens* gene atlas project (Perroud *et al.* 2018), using the strict consensus of three DEG callers, defining transcripts as expressed if their RPKM was greater than or equal to 2. For RPKM data see Table S11 and for DEGs Table S8.

*Methylated and differentially methylated positions (DMP) and overlap with single nucleotide polymorphisms (SNP)*

The Re data was generated and treated as described for Gd in (Lang *et al.* 2018) and only positions with a coverage equal to 9 or above were used for further analysis. DMPs between Gd and Re were called using methylKit using a q-value < 0.01 and a methylation difference of at least 25% (Akalin *et al.* 2012). Differentially methylated positions were associated with gene models by using bedtools intersect (Quinlan *et al.* 2010). For GO bias calculation all hyper- and hypo-methylated genes within all contexts (CHG, CHH, CG) were used. This gene list was used to identify genes also affected by a SNP between Gd and Re (Hiss *et al.* 2017) to generate GO bias data of the intersection.

*Vector construction and preparation*

Deletion constructs were built using pBHRF (Schaefer *et al.* 2010) as backbone. PCR reactions were performed with OneTaq (NEB), restriction enzymes were purchased from NEB as well as the T4 ligase and 100bp respectively 1kb ladder. For primer sequences see Table S1. 5'- and 3'-homologous regions of Ppccdc39 were amplified from gDNA using primer pairs ccdc39_HR1_for/rev and ccdc39HR2_for/rev. Flanks were subsequently cloned into pMETA via TA cloning to generate 5'- and 3' flank containing vectors. The 5' flank was cut and cloned into pBHRF using the restriction enzymes *Pme*I and *Xho*I. The 3' flank was cut and cloned using the restriction enzymes *Nar*I and *Asc*I. The final deletion cassette consists out of the two homologous regions flanking the 35S:Hygromycin:CamVter selection marker (Fig. S13). Plasmid amplification was performed in *E. coli* Top10 cells, plasmid extraction was performed using the NucleoBond Xtra midi kit (Machery-Nagel, Düren, Germany). For stable transfection, the plasmid was cut using *Pme*I and *Asc*I, and after ethanol-precipitation solved in sterile TE.

*Moss transfection*

Moss transfection was carried out in the Re wild type background (Hiss *et al.* 2017) with small modifications as published in (Cove *et al.* 2009): after protoplast regeneration all subsequently used media was Knop medium supplemented with 5mM di-ammonium tartrate (Sigma-Aldrich, Germany). For selection, the antibiotic hygromycin B (Roth, Germany) was used as an additive, with a concentration of 20mg/l.

*Genotyping*

To identify correctly integrated mutants a set of PCR controls was performed (Fig. S14). WT locus absence was tested by using the primer pair ccdc39_ingDNA_for/rev located in the coding sequence of ccdc39. Proper insertion of the 5’ flank was tested using ccdc39_HR1in (located upstream of the 5’ flank) and p35S_rev (located in the resistance cassette promotor). The 3’ flank was tested using ccdc39_HR2in (located downstream of the 3’ flank) and tCMV_for (located in the terminator of the resistance cassette). Full length PCR was performed using ccdc39_HR1in and ccdc39_HR2in.

*Gene Ontology analysis and visualization*

Gene ontology analysis was performed as described in (Widiez *et al.* 2014) based on the *P. patens* V3.3 annotation. Visualization was done using Wordle (http://www.wordle.net/). Green font colors mark over-represented, red font colors mark under-represented GO terms. Size and color correlate with significance. Darker colors and larger font size are correlated with higher significance (q < 0.0001) whereas lighter colors and smaller font size indicate a lower *q*-value (q > 0.0001 but < 0.05).

*Network analysis and gene identifier conversion*

For network analysis www.string-db.org (Szklarczyk *et al.* 2015) was used, using V3.3 protein models. To convert gene IDs from the most recent V3.3 annotation to V1.6 used by string-db.org, the conversion table published by (Perroud *et al.* 2018) was used.

*Authors' contributions*

RM and SAR wrote the manuscript with contributions by KSR. RM carried out the experiments and analysed the data. SAR conceived of and supervised the project. PFP helped with the crossing experiments. TH supervised and did the sample preparation for TEM analyses, KSR performed TEM and analysed the images together with RM. FH and RM analysed the RNA-seq data, LS and RM the bs-seq data.

*Acknowledgements*

We thank Eva Bieler of the Swiss Nanoscience Institute (SNI), Nano Imaging, University of Basel, for the help taking the SEM pictures, as well as Marion Schön for TEM sample preparation and Rebecca Hinrichs and Jason Backes for assistance. We also thank Karl Ferdinand Lechtreck (University of Georgia) for his helpful comments with regard to the hydin phenotype. The authors declare no conflicts of interest.

*Data availability*

All sequencing data has been uploaded to the NCBI SRA. Bs-seq Gd data: SRR4454535 (Lang *et al.* 2018); bs-seq Re data: SRR9901085. RNA-seq data BioProject PRJNA559055.

## References

**Akalin, 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.

**Antony, D., Becker-Heck, A., Zariwala, M.A., Schmidts, M., Onoufriadis, A., Forouhan, M., . . . Mitchison, H.M.** (2013) Mutations in CCDC39 and CCDC40 are the major cause of primary ciliary dyskinesia with axonemal disorganization and absent inner dynein arms. *Hum Mutat*, **34**, 462-472.

**Ashton, N. and V. S. Raju, M.** (2000) *The distribution of gametangia on gametophores of Physcomitrella (Aphanoregma) patens in culture*.

**Aya, K., Hiwatashi, Y., Kojima, M., Sakakibara, H., Ueguchi-Tanaka, M., Hasebe, M. and Matsuoka, M.** (2011) The Gibberellin perception system evolved to regulate a pre-existing GAMYB-mediated system during land plant evolution. *Nat Commun*, **2**, 544.

**Beike, A.K., von Stackelberg, M., Schallenberg-Rudinger, M., Hanke, S.T., Follo, M., Quandt, D., . . . Rensing, S.A.** (2014) Molecular evidence for convergent evolution and allopolyploid speciation within the Physcomitrium-Physcomitrella species complex. *BMC Evol Biol*, **14**, 158.

**Bettegowda, A. and Wilkinson, M.F.** (2010) Transcription and post-transcriptional regulation of spermatogenesis. *Philos Trans R Soc Lond B Biol Sci*, **365**, 1637-1651.

**Booth, H.A. and Holland, P.W.** (2007) Annotation, nomenclature and evolution of four novel homeobox genes expressed in the human germ line. *Gene*, **387**, 7-14.

**Carvalho-Santos, Z., Azimzadeh, J., Pereira-Leal, J.B. and Bettencourt-Dias, M.** (2011) Evolution: Tracing the origins of centrioles, cilia, and flagella. *J Cell Biol*, **194**, 165-175.

**Castleman, V.H., Romio, L., Chodhari, R., Hirst, R.A., de Castro, S.C., Parker, K.A., . . . Mitchison, H.M.** (2009) Mutations in radial spoke head protein genes RSPH9 and RSPH4A cause primary ciliary dyskinesia with central-microtubular-pair abnormalities. *Am J Hum Genet*, **84**, 197-209.

**Cove, D.** (2005) The moss Physcomitrella patens. *Annu Rev Genet*, **39**, 339-358.

**Cove, D.J., Perroud, P.F., Charron, A.J., McDaniel, S.F., Khandelwal, A. and Quatrano, R.S.** (2009) The moss Physcomitrella patens: a novel model system for plant development and genomic studies. *Cold Spring Harb Protoc*, **2009**, pdb emo115.

**Dong, F.N., Amiri-Yekta, A., Martinez, G., Saut, A., Tek, J., Stouvenel, L., . . . Coutton, C.** (2018) Absence of CFAP69 Causes Male Infertility due to Multiple Morphological Abnormalities of the Flagella in Human and Mouse. *Am J Hum Genet*, **102**, 636-648.

**Engel, P.P.** (1968) The induction of biochemical and morphological mutants in the moss *Physcomitrella patens*. *American Journal of Botany*, **55**, 438-446.

**Feng, S., Cokus, S.J., Zhang, X., Chen, P.Y., Bostick, M., Goll, M.G., . . . Jacobsen, S.E.** (2010) Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A*, **107**, 8689-8694.

**Godet, A.C., David, F., Hantelys, F., Tatin, F., Lacazette, E., Garmy-Susini, B. and Prats, A.C.** (2019) IRES Trans-Acting Factors, Key Actors of the Stress Response. *Int J Mol Sci*, **20**.

**Gosek, A. and Kwiatkowska, M.** (1991) Cytochemical studies on the antheridial mucilage and changes in its concentration and amount during the spermatogenesis in Chara vulgaris L. *Folia Histochem Cytobiol*, **29**, 91-99.

**Gramzow, L. and Theissen, G.** (2010) A hitchhiker's guide to the MADS world of plants. *Genome Biol*, **11**, 214.

**Hackenberg, D. and Twell, D.** (2019) The evolution and patterning of male gametophyte development. *Curr Top Dev Biol*, **131**, 257-298.

**He, K., Ma, X., Xu, T., Li, Y., Hodge, A., Zhang, Q., . . . Hu, J.** (2018) Axoneme polyglutamylation regulated by Joubert syndrome protein ARL13B controls ciliary targeting of signaling molecules. *Nature Communications*, **9**, 3310.

**Hernandez, G.** (2008) Was the initiation of translation in early eukaryotes IRES-driven? *Trends Biochem Sci*, **33**, 58-64.

**Hiss, M., Meyberg, R., Westermann, J., Haas, F.B., Schneider, L., Schallenberg-Rudinger, M., . . . 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.

**Hohe, A., Rensing, S., Mildner, M., Lang, D. and Reski, R.** (2002) *Day Length and Temperature Strongly Influence Sexual Reproduction and Expression of a Novel MADS-Box Gene in the Moss Physcomitrella patens*.

**Horani, A. and Ferkol, T.W.** (2018) Advances in the Genetics of Primary Ciliary Dyskinesia: Clinical Implications. *Chest*, **154**, 645-652.

**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.

**Horst, N.A. and Reski, R.** (2017) Microscopy of Physcomitrella patens sperm cells. *Plant Methods*, **13**, 33.

**Hunter, T.** (2007) The age of crosstalk: phosphorylation, ubiquitination, and beyond. *Mol Cell*, **28**, 730-738.

**Joo, S., Wang, M.H., Lui, G., Lee, J., Barnas, A., Kim, E., . . . Lee, J.H.** (2018) Common ancestry of heterodimerizing TALE homeobox transcription factors across Metazoa and Archaeplastida. *BMC Biol*, **16**, 136.

**Keiper, B.D.** (2019) Cap-Independent mRNA Translation in Germ Cells. *Int J Mol Sci*, **20**.

**Knop, W.** (1868) *Der Kreislauf des Stoffs: Lehrbuch der Agriculture-Chemie. Leipzig: H. Haessel.*

**Kofuji, R., Yagita, Y., Murata, T. and Hasebe, M.** (2018) Antheridial development in the moss Physcomitrella patens: implications for understanding stem cells in mosses. *Philos Trans R Soc Lond B Biol Sci*, **373**.

**Kofuji, R., Yoshimura, T., Inoue, H., Sakakibara, K., Hiwatashi, Y., Kurata, T., . . . Hasebe, M.** (2009) Gametangia Development in the Moss Physcomitrella patens, pp. 167-181.

**Koshimizu, S., Kofuji, R., Sasaki-Sekimoto, Y., Kikkawa, M., Shimojima, M., Ohta, H., . . . Hasebe, M.** (2018) Physcomitrella MADS-box genes regulate water supply and sperm movement for fertilization. *Nat Plants*, **4**, 36-45.

**Landberg, K., Pederson, E.R., Viaene, T., Bozorg, B., Friml, J., Jonsson, H., . . . Sundberg, E.** (2013) The MOSS *Physcomitrella patens* reproductive organ development is highly organized, affected by the two SHI/STY genes and by the level of active auxin in the SHI/STY expression domain. *Plant Physiol*, **162**, 1406-1419.

**Lang, D., Ullrich, K.K., Murat, F., Fuchs, J., Jenkins, J., Haas, F.B., . . . Rensing, S.A.** (2018) The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. *Plant J*, **93**, 515-533.

**Le Bail, A., Scholz, S. and Kost, B.** (2013) Evaluation of reference genes for RT qPCR analyses of structure-specific and hormone regulated gene expression in Physcomitrella patens gametophytes. *PLoS ONE*, **8**, e70998.

**Lechtreck, K.F. and Witman, G.B.** (2007) Chlamydomonas reinhardtii hydin is a central pair protein required for flagellar motility. *J Cell Biol*, **176**, 473-482.

**Lee, J.H., Lin, H., Joo, S. and Goodenough, U.** (2008) Early sexual origins of homeoprotein heterodimerization and evolution of the plant KNOX/BELL family. *Cell*, **133**, 829-840.

**Lefevre, P.L., Palin, M.F. and Murphy, B.D.** (2011) Polyamines on the reproductive landscape. *Endocr Rev*, **32**, 694-712.

**Liu, S., Yin, H., Li, C., Qin, C., Cai, W., Cao, M. and Zhang, S.** (2017) Genetic effects of PDGFRB and MARCH1 identified in GWAS revealing strong associations with semen production traits in Chinese Holstein bulls. *BMC Genet*, **18**, 63.

**Lopez, R.A. and Renzaglia, K.S.** (2018) The Ceratopteris (fern) developing motile gamete walls contain diverse polysaccharides, but not pectin. *Planta*, **247**, 393-404.

**Lord, E.M. and Sanders, L.C.** (1992) Roles for the extracellular matrix in plant development and pollination: a special case of cell movement in plants. *Dev Biol*, **153**, 16-28.

**Luo, M., Cao, M., Kan, Y., Li, G., Snell, W. and Pan, J.** (2011) The phosphorylation state of an aurora-like kinase marks the length of growing flagella in Chlamydomonas. *Curr Biol*, **21**, 586-591.

**McCourt, R.M., Delwiche, C.F. and Karol, K.G.** (2004) Charophyte algae and land plant origins. *Trends Ecol Evol*, **19**, 661-666.

**Merveille, A.C., Davis, E.E., Becker-Heck, A., Legendre, M., Amirav, I., Bataille, G., . . . Amselem, S.** (2011) CCDC39 is required for assembly of inner dynein arms and the dynein regulatory complex and for normal ciliary motility in humans and dogs. *Nat Genet*, **43**, 72-78.

**Mitchell, D.R.** (2007) The evolution of eukaryotic cilia and flagella as motile and sensory organelles. *Adv Exp Med Biol*, **607**, 130-140.

**Nakosteen, P.C. and Hughes, K.W.** (1978) Sexual Life Cycle of Three Species of Funariaceae in Culture. *The Bryologist*, **81**, 307-314.

**Nevers, Y., Prasad, M.K., Poidevin, L., Chennen, K., Allot, A., Kress, A., . . . Lecompte, O.** (2017) Insights into Ciliary Genes and Evolution from Multi-Level Phylogenetic Profiling. *Mol Biol Evol*, **34**, 2016-2034.

**Noone, P.G., Leigh, M.W., Sannuti, A., Minnix, S.L., Carson, J.L., Hazucha, M., . . . Knowles, M.R.** (2004) Primary ciliary dyskinesia: diagnostic and phenotypic features. *Am J Respir Crit Care Med*, **169**, 459-467.

**Oda, T., Yanagisawa, H., Kamiya, R. and Kikkawa, M.** (2014) A molecular ruler determines the repeat length in eukaryotic cilia and flagella. *Science*, **346**, 857-860.

**Olbrich, H., Schmidts, M., Werner, C., Onoufriadis, A., Loges, N.T., Raidt, J., . . . Omran, H.** (2012) Recessive HYDIN mutations cause primary ciliary dyskinesia without randomization of left-right body asymmetry. *Am J Hum Genet*, **91**, 672-684.

**Ortiz-Ramirez, C., Michard, E., Simon, A.A., Damineli, D.S.C., Hernandez-Coronado, M., Becker, J.D. and Feijo, J.A.** (2017) GLUTAMATE RECEPTOR-LIKE channels are essential for chemotaxis and reproduction in mosses. *Nature*, **549**, 91-95.

**Pan, J., Naumann-Busch, B., Wang, L., Specht, M., Scholz, M., Trompelt, K. and Hippler, M.** (2011) Protein phosphorylation is a key event of flagellar disassembly revealed by analysis of flagellar phosphoproteins during flagellar shortening in Chlamydomonas. *J Proteome Res*, **10**, 3830-3839.

**Papon, J.F., Coste, A., Roudot-Thoraval, F., Boucherat, M., Roger, G., Tamalet, A., . . . Escudier, E.** (2010) A 20-year experience of electron microscopy in the diagnosis of primary ciliary dyskinesia. *Eur Respir J*, **35**, 1057-1063.

**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.

**Perroud, P.F., Haas, F.B., Hiss, M., Ullrich, K.K., Alboresi, A., Amirebrahimi, M., . . . Rensing, S.A.** (2018) The Physcomitrella patens gene atlas project: large-scale RNA-seq based expression data. *Plant J*, **95**, 168-182.

**Perroud, P.F., Meyberg, R. and Rensing, S.A.** (2019) Physcomitrella patens Reute mCherry as a tool for efficient crossing within and between ecotypes. *Plant Biol (Stuttg)*, **21 Suppl 1**, 143-149.

**Puttick, M.N., Morris, J.L., Williams, T.A., Cox, C.J., Edwards, D., Kenrick, P., . . . Donoghue, P.C.J.** (2018) The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte. *Curr Biol*, **28**, 733-745 e732.

**Quinlan, A.R. and Hall, I.M.** (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.

**Quodt, V., Faigl, W., Saedler, H. and Munster, T.** (2007) The MADS-domain protein PPM2 preferentially occurs in gametangia and sporophytes of the moss Physcomitrella patens. *Gene*, **400**, 25-34.

**Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., . . . Boore, J.L.** (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64-69.

**Renzaglia, K.S., Duff, R.J.T., Nickrent, D.L. and Garbary, D.J.** (2000) Vegetative and reproductive innovations of early land plants: implications for a unified phylogeny. *Philos Trans R Soc Lond B Biol Sci*, **355**, 769-793.

**Renzaglia, K.S. and Garbary, D.J.** (2001) Motile Gametes of Land Plants: Diversity, Development, and Evolution. *Critical Reviews in Plant Sciences*, **20**, 107-213.

**Renzaglia, K.S., Lopez, R.A., Henry, J.S., Flowers, N.D. and Vaughn, K.C.** (2017) Transmission Electron Microscopy of Centrioles, Basal Bodies and Flagella in Motile Male Gametes of Land Plants. *Bio-protocol*, **7**, e2448.

**Sakakibara, K., Ando, S., Yip, H.K., Tamada, Y., Hiwatashi, Y., Murata, T., . . . Bowman, J.L.** (2013) KNOX2 genes regulate the haploid-to-diploid morphological transition in land plants. *Science*, **339**, 1067-1070.

**Sakakibara, K., Nishiyama, T., Deguchi, H. and Hasebe, M.** (2008) Class 1 KNOX genes are not involved in shoot development in the moss Physcomitrella patens but do function in sporophyte development. *Evol Dev*, **10**, 555-566.

**Sanchez-Vera, V., Kenchappa, C.S., Landberg, K., Bressendorff, S., Schwarzbach, S., Martin, T., . . . Sundberg, E.** (2017) Autophagy is required for gamete differentiation in the moss Physcomitrella patens. *Autophagy*, **13**, 1939-1951.

**Schaefer, D.G., Delacote, F., Charlot, F., Vrielynck, N., Guyon-Debast, A., Le Guin, S., . . . Nogue, F.** (2010) RAD51 loss of function abolishes gene targeting and de-represses illegitimate integration in the moss Physcomitrella patens. *DNA Repair (Amst)*, **9**, 526-533.

**Schmid, M.W., Giraldo-Fonseca, A., Rovekamp, M., Smetanin, D., Bowman, J.L. and Grossniklaus, U.** (2018) Extensive epigenetic reprogramming during the life cycle of Marchantia polymorpha. *Genome Biol*, **19**, 9.

**Singer, S.D., Krogan, N.T. and Ashton, N.W.** (2007) Clues about the ancestral roles of plant MADS-box genes from a functional analysis of moss homologues. *Plant Cell Rep*, **26**, 1155-1169.

**Southworth, D. and Cresti, M.** (1997) Comparison of Flagellated and Nonflagellated Sperm in Plants. *American Journal of Botany*, **84**, 1301-1311.

**Stewart, K.D. and Mattox, K.R.** (1975) Comparative cytology, evolution and classification of the green algae with some consideration of the origin of other organisms with chlorophylls A and B. *The Botanical Review*, **41**, 104-135.

**Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., . . . von Mering, C.** (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*, **43**, D447-452.

**Tang, S., Wang, X., Li, W., Yang, X., Li, Z., Liu, W., . . . Zhang, F.** (2017) Biallelic Mutations in CFAP43 and CFAP44 Cause Male Infertility with Multiple Morphological Abnormalities of the Sperm Flagella. *Am J Hum Genet*, **100**, 854-864.

**Theißen, G. and Gramzow, L.** (2016) Structure and Evolution of Plant MADS Domain Transcription Factors, pp. 127-138.

**Transeau, E.N.** (1951) *The Zygnemataceae (fresh-water Conjugate Algae) with Keys for the Identification of Genera and Species: And Seven Hundred Eighty-nine Illustrations*: Ohio State University Press.

**van den Ende, H., Musgrave, A. and Klis, F.M.** (1990) The Role of Flagella in the Sexual Reproduction of Chlamydomonas Gametes. *Springer, Boston, MA.*

**Wickham, H.** (2016) ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York.*

**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.

**Wilhelmsson, P.K.I., Mühlich, C., Ullrich, K.K. and Rensing, S.A.** (2017) Comprehensive Genome-Wide Classification Reveals That Many Plant-Specific Transcription Factors Evolved in Streptophyte Algae. *Genome Biology and Evolution*, **9**, 3384-3397.

**Yaari, R., Noy-Malka, C., Wiedemann, G., Auerbach Gershovitz, N., Reski, R., Katz, A. and Ohad, N.** (2015) DNA METHYLTRANSFERASE 1 is involved in (m)CG and (m)CCG DNA methylation and is essential for sporophyte development in Physcomitrella patens. *Plant Mol Biol*, **88**, 387-400.

**Zemach, A., McDaniel, I.E., Silva, P. and Zilberman, D.** (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, **328**, 916-919.

# 9 Concluding remarks and outlook

In summary, the work presented here did not only help to understand the loss of fertility in the Gd ecotype, but also introduced and characterized the fertile ecotype Re and several newly generated resources for *P. patens.* Namely, a comparative DNA methylation data set of sexual reproductive tissue was generated using Gd and Re; the fluorescent marker strain Re-mCherry was introduced which can be employed in crossing analyses; comparative RNA-seq analysis of antheridia bundles was generated for Gd and Re which allows to get deeper insights into spermatozoid development and differences between both ecotypes.

Finally, this work also could show that key players required in *P. patens* and mammal sexual reproduction are evolutionary conserved.

## 9.1 Gransden infertile phenotype probably due to somatic (epi-) mutations

The finally determined locus of impairment in the world wide used ecotype Gransden which indeed, should be defined as a lab strain, with Gd ecotype background, was shown to be the male reproductive apparatus. Interestingly, SNP, KaKs, DNA methylation and RNA-seq analysis from the presented publications and tissues showed differences between Gd and Re, in which GO bias analysis and/or selected marker genes pinpointed towards impairments in the sexual reproduction. My assumption here is that Gd and Re are even more closely related than thought before and that the shown differences are due to the long term vegetative reproduction *in vitro,* which led to the accumulation of (epi-) mutations visible in the comparative analysis (Ashton and Raju, 2000; Perroud *et al.*, 2011; Hiss *et al.*, 2017; Meyberg *et al.*, 2019). The distance between both collection points could easily be bridged by migrating birds which was proposed earlier (Beike *et al.*, 2014; Hiss *et al.*, 2017). This theory is supported by the low number of SNPs found between Gd and Re in comparison to the ecotypes Vx and Kaskaskia (Ka) and in comparison to ecotypes in e.g. *A. thaliana* (Cao *et al.*, 2011; Hiss *et al.*, 2017; Lang *et al.*, 2018). To prove this theory, it might be worthwhile to analyze the SNPs between the japanese Gransden (GdJ) laboratory strain which still can develop a solid number of sporophytes (~60%, Meyberg *et al.*, 2019) and Gd resp. Re. With this approach, also the effect of natural variations as a putative trigger of infertility can be reduced: genes, affected by the negative selection pressure due to the missing sexual reproduction of Gd, should show less divergence in comparison of Re and GdJ, but several SNPs when Gd and GdJ are compared.

## 9.2 *Physcomitrella patens* can be employed for analysis of defects in the mammal male germline

So far, several *P. patens* genes were shown to have homologs in mammals, but in particular genes involved in the male sexual reproduction, show highly similar functionalities of the proteins as well, e.g. GLR (Ortiz-Ramírez *et al.*, 2017), ATG7 (Sanchez-Vera *et al.*, 2017) and CCDC39 (Meyberg *et al.*, 2019). This leads to the assumption that the development of flagellated male gametes and sessile female gametes is not convergent but homologous between plants and mammals. Usually, *C. reinhardtii* is employed for the analysis of flagellar related genes which are responsible for male infertility in human (Lechtreck and Witman, 2007; Oda *et al.*, 2014). In comparison to the used algal model so far, bryophytes possess not only mating types, but spermatozoids and egg cells which fuse to yield the zygote after successful fertilization. The flagella are not present during most of the life cycle and only develop for sexual reproduction as in mammals. Also several resources are available, as e.g. the genome, transcriptomes and methylomes and efficient transformation and gene targeting protocols are available. The two suggested model organisms are the liverwort *M. polymorpha* and the moss *P. patens.* Both have different advantages and can be used for different approaches: *M. polymorpha* is diocious and thus has the advantage of a clear separation of the sexes for e.g. DNA/RNA extraction for DNA methylation or expression analysis. In comparison, *P. patens* is easily cultured and transformed, the induction of sexual reproduction is fast and the life cycle *in vitro* is shorter. Additionally, antheridia, archegonia and the sporophytes are not hidden inside the antheridio-/archegoniophores and thus, easily accessible for harvest or microscopical analysis. Efficient crossing analysis can be performed without effort and also multiple mutants can be generated this way. Thus, organisms should be chosen depending on the scientific topic which should be addressed. Based on the presented work, I propose bryophytes as a model system for easy and quick analysis of genes affecting female or male reproductive organs as e.g. genes associated with flagellar dysfunctionalities in mammal and particulary human diseases.

## 9.3 Bryophytes possess inner and outer dynein arms

As described in the introduction, contrary information about the presence of IDAs and ODAs in bryophytes is published. The motility of eukaryotic flagella is dependent on these dynein motor proteins which drive interdoublet sliding (Porter and Sale, 2000; Heuser *et al.*, 2009) whereas most eukaryotic flagella probably show inner and outer dynein arms (Silflow, 2001). The data underlying the previous publications which were published in 2011 and 2012 was, with regard to bioinformatic analysis quite fragmented. Also the development of the tools used for analysis was not as far as it is now. Thus, probably the lack of high quality and in depth sequencing data of the genomes as well as transcriptomes was one part leading to the contrary results. With regard to the morphological analysis, several studies were probably performed on laboratory strains. As this work could show, the accumulation of mutations and epi-mutations due to *in vitro* culture leading to the loss of fertility through the loss of IDAs is possible. Thus, the possibility to gather mutations in ODAs or IDAs or in proteins affecting their function could also have led to the contrary results. Additionally, the process needed to finally get a TEM image of a flagellar ultrastructure is complicated and e.g. images with little contrast could easily hide the presence of IDAs or ODAs. Finally, the here presented ultrastructural analysis of mature Gd and Re axonemes showed the presence of probably both, inner and outer dynein arms in *P. patens* and solved at least for mosses, the contrary information published so far.

## 9.4 Outlook

### 9.4.1 Future usage of data and methods established in this thesis

Data sets and findings from the presented work will be and have already been used for further research. The DNA methylation data for the Gd ecotype, published in Lang *et al.*, 2018, is the first part of a comparative DNA methylation analysis between Gd and the previously introduced ecotype Re which was employed in Meyberg *et al.*, 2019. This data set was also used for the development of a pipeline identifying differentially methylated regions (DMR) in plants (Kreutz *et al.*, under revision).

The fluorescent marker strain Re-mCherry is employed for all crossing analysis performed in Meyberg *et al.*, 2019 and additionally in the latest analysis of DEK1 localization and fertility analysis (Perroud *et al.*, in preparation). The RNA-seq analysis of antheridia bundles presented in Meyberg *et al.*, 2019 will be made available in the Physcomitrella expression analysis tool (PEATmoss, Fernandez-Pozo *et al.*, under revision), which then can be used by the community to assess expression in sexual reproductive tissues as well as comparative analyses between Gd and Re. The technique developed for microscopical antheridia and spermatozoid analysis will also be employed for further publications as e.g. in Perroud *et al.*, (in preparation) characterizing DEK1 localization in *P. patens* antheridia and spermatozoids as well as for mutant spermatozoid characterization in research performed by Anne C. Genau.

### 9.4.2 Project specific outlook

Based on the results of this work, I am highly interested in the complementation analysis of the *ccdc39* mutant. To determine the level of evolutionary conservation, I would like to employ homologs from alga, liverworts and human. To analyze the temporal and spatial protein presence, a *P. patens* ccdc39:gfp would serve well. Since the *C. reinhardtii* mutant did always show flagella with a reduced length (Oda *et al.*, 2014) which is not the case in *P. patens*, I would also like to analyze ccdc40 which is a direct interaction partner of ccdc39 in human and algae. Defects in ccdc40 also lead to PCD in human displaying a comparative phenotype to mutations in ccdc39 (Becker-Heck *et al.*, 2011; Antony *et al.*, 2013). In this case, the analysis of single as well as double deletion mutant phenotypes with regard to the flagella and the flagellar ultrastructure would be interesting in particular, if the deletion of both interaction partners led to a reduction of the flagellar length. This would give new insights into the evolution of ccdc39 and ccdc40 function in eukaryotic cilia. Additionally, in depth analysis based on the newly released expression data sets and high quality TEM images stained with specific antibodies of IDAs and ODAs would be highly interesting to get deeper knowledge about IDAs and ODAs in early diverging land plants.

With regard to the methylation and RNA-seq data I would like to dive deeper into the flood of genes that are differentially expressed and methylated between Gd and Re and, if possible, identify the reason for the hydin hypermethylation which, at least partly, led to the Gd infertile phenotype. Also, some of the candidate genes from the candidate gene approach should be analyzed further, since promising first phenotypic data are available. Finally, I propose several putative scientific projects can arise from my presented work and I am looking forward to find out more about the factors affecting sexual reproduction in *P. patens*.

# 10 Literature

**Alvarez, L.** (2017) The tailored sperm cell. *J. Plant Res.*, **130**, 455–464.

**Antony, D., Becker-Heck, A., Zariwala, M.A., et al.** (2013) Mutations in CCDC39 and CCDC40 are the Major Cause of Primary Ciliary Dyskinesia with Axonemal Disorganization and Absent Inner Dynein Arms. *Hum. Mutat.*, **34**, 462–472.

**Ashton, N.W. and Raju, M.V.S.** (2000) The distribution of gametangia on gametophores of Physcomitrella (Aphanoregma) patens in culture. *J. Bryol.*, **22**, 9–12. Available at: http://dx.doi.org/10.1179/jbr.2000.22.1.9.

**Aya, K., Hiwatashi, Y., Kojima, M., Sakakibara, H., Ueguchi-Tanaka, M., Hasebe, M. and Matsuoka, M.** (2011) The Gibberellin perception system evolved to regulate a pre-existing GAMYB-mediated system during land plant evolution. *Nat. Commun.*, **2**, 544–549. Available at: http://dx.doi.org/10.1038/ncomms1552.

**Becker-Heck, A., Zohn, I.E., Okabe, N., et al.** (2011) The coiled-coil domain containing protein CCDC40 is essential for motile cilia function and left-right axis formation. *Nat. Genet.*

**Beike, A.K., Stackelberg, M. Von, Schallenberg-Rüdinger, M., et al.** (2014) Molecular evidence for convergent evolution and allopolyploid speciation within the Physcomitrium-Physcomitrella species complex. *BMC Evol. Biol.*

**Bellaoui, M., Pidkowich, M.S., Samach, A., Kushalappa, K., Kohalmi, S.E., Modrusan, Z., Crosby, W.L. and Haughn, G.W.** (2007) The Arabidopsis BELL1 and KNOX TALE Homeodomain Proteins Interact through a Domain Conserved between Plants and Animals. *Plant Cell*.

**Bernhard, D.L. and Renzaglia, K.S.** (1995) Spermiogenesis in the Moss Aulacomnium palustre. *Bryologist*.

**Bock, W.J.** (1980) The definition and recognition of biological adaptation. *Integr. Comp. Biol.*

**Brokaw, C.J. and Kamiya, R.** (1987) Bending patterns of Chlamydomonas flagella: IV. Mutants with defects in inner and outer dynein arms indicate differences in dynein arm function. *Cell Motil. Cytoskeleton*.

**Budke, J.M. and Goffinet, B.** (2016) Comparative cuticle development reveals taller sporophytes are covered by thicker calyptra cuticles in mosses. *Front. Plant Sci.*

**Cao, J., Schneeberger, K., Ossowski, S., et al.** (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.*

**Carvalho-Santos, Z., Azimzadeh, J., Pereira-Leal, J.B. and Bettencourt-Dias, M.** (2011) Tracing the origins of centrioles, cilia, and flagella. *J. Cell Biol.*, **194**, 165–175.

**Chater, C.C.C., Caine, R.S., Fleming, A.J. and Gray, J.E.** (2017) Origins and Evolution of Stomatal Development. *Plant Physiol.*

**Cove, D.** (2000) The moss, Physcomitrella patens. *J. Plant Growth Regul.*

**Cove, D.** (2005) The Moss *Physcomitrella patens*. *Annu. Rev. Genet.*, **39**, 339–358. Available at: http://www.annualreviews.org/doi/10.1146/annurev.genet.39.073003.110214.

**Cove, D., Bezanilla, M., Harries, P. and Quatrano, R.** (2006) Mosses As Model Systems for the Study of Metabolism and Development. *Annu. Rev. Plant Biol.*, **57**, 497–520.

**Cove, D.J., Perroud, P.F., Charron, A.J., McDaniel, S.F., Khandelwal, A. and Quatrano, R.S.** (2009) The moss Physcomitrella patens: A novel model system for plant development and genomic studies. *Cold Spring Harb. Protoc.*

**Daku, R.M., Rabbi, F., Buttigieg, J., Coulson, I.M., Horne, D., Martens, G., Ashton, N.W. and Suh, D.Y.** (2016) PpASCL, the Physcomitrella patens anther-specific chalcone synthase-like enzyme implicated in sporopollenin biosynthesis, is needed for integrity of the moss spore wall and spore viability. *PLoS One*.

**Drews, G.N. and Yadegari, R.** (2002) Development and Function of the Angiosperm Female Gametophyte. *Annu. Rev. Genet.*

**Dutcher, S.K.** (2000) Chlamydomonas reinhardtii: Biological rationale for genomics. In *Journal of Eukaryotic Microbiology*.

**Engel, Paulinus P.** (1968) The Induction of Biochemical and Morphological Mutants in the Moss Physcomitrella patens. *Am. J. Bot.*

**Ferrero-Serrano, Á. and Assmann, S.M.** (2019) Phenotypic and genome-wide association with the local environment of Arabidopsis. *Nat. Ecol. Evol.*

**Frank, M.H. and Scanlon, M.J.** (2015) Transcriptomic evidence for the evolution of shoot meristem function in sporophyte-dominant land plants through concerted selection of ancestral gametophytic and sporophytic genetic programs. *Mol. Biol. Evol.*

**Frey, W., Stech, M. and Fischer, E.** (2009) Syllabus of Plant Families – Part 3 Bryophytes and Seedless Vascular Plants Berlin. *Stuttgart: Borntraeger*.

**Furuichi, T. and Matsuura, K.** (2016) Kinetic Analysis on the Motility of Liverwort Sperms Using a Microscopic Computer-Assisted Sperm Analyzing System. *Environ. Control Biol.*

**Haig, D.** (2013) Filial mistletoes: The functional morphology of moss sporophytes. *Ann. Bot.*

**Harrison, C.J., Roeder, A.H.K., 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.*

**Heuser, T., Raytchev, M., Krell, J., Porter, M.E. and Nicastro, D.** (2009) The dynein regulatory complex is the nexin link and a major regulatory node in cilia and flagella. *J. Cell Biol.*

**Hisanaga, T., Yamaoka, S., Kawashima, T., Higo, A., Nakajima, K., Araki, T., Kohchi, T. and Berger, F.** (2019) Building new insights in plant gametogenesis from an evolutionary perspective. *Nat. Plants*, **5**, 663–669. Available at: http://www.nature.com/articles/s41477-019-0466-0.

**Hiss, M., Meyberg, R., Westermann, J., Haas, F.B., Schneider, L., Schallenberg-Rüdinger, 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.

**Hodges, M.E., Wickstead, B., Gull, K. and Langdale, J.A.** (2012) The evolution of land plant cilia. *New Phytol.*, **195**, 526–540.

**Hofmeister, W.** (1851) *Vergleichende Untersuchungen der Keimung, Entfaltung und Fruchtbildung h€oherer Kryptogamen (Moose, Farne, Equisetaceen, Rhizokarpeen und Lykopodiaceen) und der Samenbildung der Coniferen*, Leipzig.

**Hohe, A., Rensing, S.A., Mildner, M., Lang, D. and Reski, R.** (2002) Hohe2002. , **4**.

**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**, 1–6. Available at: http://dx.doi.org/10.1038/nplants.2015.209.

**Huang, B.P.H.** (1986) Chlamydomonas reinhardtii: A Model System for the Genetic Analysis of Flagellar Structure and Motility. *Int. Rev. Cytol.*

**Jones, V.A.S. and Dolan, L.** (2012) The evolution of root hairs and rhizoids. *Ann. Bot.*

**Kamisugi, Y., Schlink, K., Rensing, S.A., Schween, G., Stackelberg, M. von, Cuming, A.C., Reski, R. and Cove, D.J.** (2006) The mechanism of gene targeting in Physcomitrella patens: Homologous recombination, concatenation and multiple integration. *Nucleic Acids Res.*, **34**, 6205–6214.

**Kamisugi, Y., Stackelberg, M. Von, 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.*

**Kamiya, R.** (1995) Exploring the function of inner and outer dynein arms with Chlamydomonas mutants. *Cell Motil. Cytoskeleton*.

**Kasahara, M., Hiwatashi, Y., Ishikawa, T., et al.** (2011) Genetic map of Physcomitrella patens based on SNP identification with Illumina sequencing. In *Conference Proceedings of Moss 2011*. Germany: University of Freiburg.

**Katz, A., Oliva, M., Mosquna, A., Hakim, O. and Ohad, N.** (2004) FIE and CURLY LEAF polycomb proteins interact in the regulation of homeobox gene expression during sporophyte development. *Plant J.*

**Keeling, P.J.** (2004) Diversity and evolutionary history of plastids and their hosts. *Am. J. Bot.*

**Keeling, P.J.** (2010) The endosymbiotic origin, diversification and fate of plastids. *Philos. Trans. R. Soc. B Biol. Sci.*

**Kenrick, P. and Crane, P.R.** (1997) The origin and early evolution of plants on land. *Nature*.

**Kofuji, R. and Hasebe, M.** (2014) Eight types of stem cells in the life cycle of the moss Physcomitrella patens. *Curr. Opin. Plant Biol.*

**Kofuji, R., Yagita, Y., Murata, T. and Hasebe, M.** (2018) Antheridial development in the moss physcomitrella patens: Implications for understanding stem cells in mosses. *Philos. Trans. R. Soc. B Biol. Sci.*, **373**, 1–7.

**Kofuji, R., Yoshimura, T., Inoue, H., Sakakibara, K., Hiwatashi, Y., Kurata, T., Aoyama, T., Ueda, K. and Hasebe, M.** (2009) Gametangia development in the moss Physcomitrella patens. *Annu. Plant Rev.*

**La Torre, A.R. De, Li, Z., Peer, Y. Van De and Ingvarsson, P.K.** (2017) Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. *Mol. Biol. Evol.*

**Landberg, K., Pederson, E.R.A., Viaene, T., Bozorg, B., Friml, J., Jonsson, H., Thelander, M. and Sundberg, E.** (2013) The Moss Physcomitrella patens Reproductive Organ Development Is Highly Organized, Affected by the Two SHI/STY Genes and by the Level of Active Auxin in the SHI/STY Expression Domain. *Plant Physiol.*, **162**, 1406–1419.

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

**Lechtreck, K.F. and Witman, G.B.** (2007) Chlamydomonas reinhardtii hydin is a central pair protein required for flagellar motility. *J. Cell Biol.*, **176**, 473–482.

**Lee, J.H., Lin, H., Joo, S. and Goodenough, U.** (2008) Early Sexual Origins of Homeoprotein Heterodimerization and Evolution of the Plant KNOX/BELL Family. *Cell*.

**Liu, X., Bogaert, K., Engelen, A.H., Leliaert, F., Roleda, M.Y. and Clerck, O. De** (2017) Seaweed reproductive biology: Environmental and genetic controls. *Bot. Mar.*

**Mast, F.D., Barlow, L.D., Rachubinski, R.A. and Dacks, J.B.** (2014) Evolutionary mechanisms for establishing eukaryotic cellular complexity. *Trends Cell Biol.*, **24**, 435–442. Available at: http://dx.doi.org/10.1016/j.tcb.2014.02.003.

**McDaniel, S.F., Stackelberg, M. Von, Richardt, S., Quatrano, R.S., Reski, R. and Rensing, S.A.** (2010) The speciation history of the physcomitrium - Physcomitrella species complex. *Evolution (N. Y).*, **64**, 217–231.

**Medina, R., Johnson, M.G., Liu, Y., Wickett, N.J., Shaw, A.J. and Goffinet, B.** (2019) Phylogenomic delineation of Physcomitrium (Bryophyta: Funariaceae) based on targeted sequencing of nuclear exons and their flanking regions rejects the retention of Physcomitrella , Physcomitridium and Aphanorrhegma . *J. Syst. Evol.*

**Medina, R., Liu, Y., Li-Song, W., Shuiliang, G., Hylander, K. and Goffinet, B.** (2015) DNA based revised geographic circumscription of species of Physcomitrella s.l. (Funariaceae): P. patens new to East Asia and P. magdalenae new to East Africa . *Bryologist*.

**Menand, B., Calder, G. and Dolan, L.** (2007) Both chloronemal and caulonemal cells expand by tip growth in the moss Physcomitrella patens. *J. Exp. Bot.*

**Merced, A. and Renzaglia, K.S.** (2017) Structure, function and evolution of stomata from a bryological perspective. *Bryophyt. Divers. Evol.*, **39**, 7.

**Meyberg, R., Perroud, P.-F., Haas, F.B., Schneider, L., Heimerl, T., Renzaglia, K. and Rensing, S.A.** (2019) Characterization of evolutionarily conserved key players affecting eukaryotic flagellar motility and fertility using a moss model. *bioRxiv*, 728691. Available at: http://biorxiv.org/content/early/2019/08/08/728691.abstract.

**Michael J. Prigge and Magdalena Bezanilla** (2010) Evolutionary crossroads in developmental biology: Physcomitrella patens. *Development*.

**Mitchell, D.R.** (2007) The evolution of eukaryotic cilia and flagella as motile and sensory organelles. *Adv. Exp. Med. Biol.*

**Morris, J.L., Puttick, M.N., Clark, J.W., et al.** (2018) The timescale of early land plant evolution. *Proc. Natl. Acad. Sci.*, **115**, E2274–E2283.

**Mosquna, A., Katz, A., Decker, E.L., Rensing, S.A., Reski, R. and Ohad, N.** (2009) Regulation of stem cell maintenance by the Polycomb protein FIE has been conserved during land plant evolution. *Development*, **136**, 2433–2444.

**Nakosteen, P.C. and Hughes, K.W.** (1978) Sexual Life Cycle of Three Species of Funariaceae in Culture. *Bryologist*.

**Nekrasov, M., Wild, B. and Müller, J.** (2005) Nucleosome binding and histone methyltransferase activity of Drosophila PRC2. *EMBO Rep.*

**Nieto-Lugilde, M., Werner, O., McDaniel, S.F. and Ros, R.M.** (2018) Environmental variation obscures species diversity in southern european populations of the moss genus ceratodon. *Taxon*.

**Niklas, K.J. and Kutschera, U.** (2010) The evolution of the land plant life cycle. *New Phytol.*

**Nishiyama, T., Sakayama, H., Vries, J. de, et al.** (2018) The Chara Genome: Secondary Complexity and Implications for Plant Terrestrialization. *Cell*.

**O'Donoghue, M.T., Chater, C., Wallace, S., Gray, J.E., Beerling, D.J. and Fleming, A.J.** (2013) Genome-wide transcriptomic analysis of the sporophyte of the moss Physcomitrella patens. *J. Exp. Bot.*

**Oda, T., Yanagisawa, H., Kamiya, R. and Kikkawa, M.** (2014) A molecular ruler determines the repeat length in eukaryotic cilia and flagella. *Science (80-. ).*

**Ortiz-Ramírez, C., Hernandez-Coronado, M., Thamm, A., Catarino, B., Wang, M., Dolan, L., Feijó, J.A.A. and Becker, J.D.D.** (2016) A Transcriptome Atlas of Physcomitrella patens Provides Insights into the Evolution and Development of Land Plants. *Mol. Plant*.

**Ortiz-Ramírez, C., Michard, E., Simon, A.A., Damineli, D.S.C., Hernández-Coronado, M., Becker, J.D. and Feijó, J.A.** (2017) GLUTAMATE RECEPTOR-LIKE channels are essential for chemotaxis and reproduction in mosses. *Nature*, **549**, 91–95. Available at: http://dx.doi.org/10.1038/nature23478.

**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.*, **191**, 301–306.

**Perroud, P.F., Haas, F.B., Hiss, M., et al.** (2018) The Physcomitrella patens gene atlas project: large-scale RNA-seq based expression data. *Plant J.*, **95**, 168–182.

**Piperno, G., Huang, B., Ramanis, Z. and Luck, D.J.L.** (1981) Radial spokes of Chlamydomonas flagella: Polypeptide composition and phosphorylation of stalk components. *J. Cell Biol.*

**Porter, M.E. and Sale, W.S.** (2000) The 9 + 2 axoneme anchors multiple inner arm dyneins and a network of kinases and phosphatases that control motility. *J. Cell Biol.*

**Potter, E.E., Thornber, C.S., Swanson, J.D. and McFarland, M.** (2016) Ploidy distribution of the harmful bloom forming macroalgae ulva spp. in Narragansett Bay, Rhode Island, USA, using flow cytometry methods. *PLoS One*.

**Puttick, M.N., Morris, J.L., Williams, T.A., et al.** (2018) The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte. *Curr. Biol.*, **28**, 733-745.e2. Available at: https://doi.org/10.1016/j.cub.2018.01.063.

**Rathke, C., Barckmann, B., Burkhard, S., Jayaramaiah-Raja, S., Roote, J. and Renkawitz-Pohl, R.** (2010) Distinct functions of Mst77F and protamines in nuclear shaping and chromatin condensation during Drosophila spermiogenesis. *Eur. J. Cell Biol.*

**Regmi, K.C., Li, L. and Gaxiola, R.A.** (2017) Alternate modes of photosynthate transport in the alternating generations of physcomitrella patens. *Front. Plant Sci.*

**Renault, H., Alber, A., Horst, N.A., et al.** (2017) A phenol-enriched cuticle is ancestral to lignin evolution in land plants. *Nat. Commun.*

**Rensing, S.A.** (2016) (Why) Does Evolution Favour Embryogenesis? *Trends Plant Sci.*

**Rensing, S.A.** (2018) Plant Evolution: Phylogenetic Relationships between the Earliest Land Plants. *Curr. Biol.*

**Rensing, S.A.** (2017) Why we need more non-seed plant models. *New Phytol.*

**Rensing, S.A., Ick, J., Fawcett, J.A., Lang, D., Zimmer, A., Peer, Y. Van De and Reski, R.** (2007) An ancient genome duplication contributed to the abundance of metabolic genes in the moss Physcomitrella patens. *BMC Evol. Biol.*

**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 (80-. ).*, **319**, 64–69.

**Renzaglia, K., Lopez, R., Henry, J., Flowers, N. and Vaughn, K.** (2017) Transmission Electron Microscopy of Centrioles, Basal Bodies and Flagella in Motile Male Gametes of Land Plants. *BIO-PROTOCOL*.

**Renzaglia, K.S., Duff, R.J., Nickrent, D.L. and Garbary, D.J.** (2000) Vegetative and reproductive innovations of early land plants: Implications for a unified phylogeny. In *Philosophical Transactions of the Royal Society B: Biological Sciences*.

**Renzaglia, K.S. and Garbary, D.J.** (2010) Motile Gametes of Land Plants : Diversity , Development , and Evolution Motile Gametes of Land Plants : Diversity, Development, and Evolution. *CRC. Crit. Rev. Plant Sci.*, 37–41.

**Renzaglia, K.S. and Garbary, D.J.** (2001) Motile gametes of land plants: Diversity, development, and evolution. *CRC. Crit. Rev. Plant Sci.*, **20**, 107–213.

**Renzaglia, K.S., Villarreal Aguilar, J.C. and Garbary, D.J.** (2018) Morphology supports -the setaphyte hypothesis: mosses plus liverworts form a natural group. *Bryophyt. Divers. Evol.*

**Richardson, M.E., Bleiziffer, A., Tü ttelmann, F., Gromoll, J. and Wilkinson, M.F.** (2014) Epigenetic regulation of the RHOX homeobox gene cluster and its association withhumanmale infertility. *Hum. Mol. Genet.*

**Sakakibara, K., Ando, S., Yip, H.K., Tamada, Y., Hiwatashi, 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 (80-. ).*, **339**, 1067–1070.

**Sakakibara, K., Nishiyama, T., Deguchi, H. and Hasebe, M.** (2008) Class 1 KNOX genes are not involved in shoot development in the moss Physcomitrella patens but do function in sporophyte development. *Evol. Dev.*, **10**, 555–566.

**Sanchez-Vera, V., Kenchappa, C.S., Landberg, K., et al.** (2017) Autophagy is required for gamete differentiation in the moss Physcomitrella patens. *Autophagy*, **13**, 1939–1951.

**Satir, P., Mitchell, D.R. and Jékely, G.** (2008) *Chapter 3 How Did the Cilium Evolve?* 1st ed., Elsevier Inc. Available at: http://dx.doi.org/10.1016/S0070-2153(08)00803-X.

**Schaefer, D.G.** (2001) Gene targeting in physcomitrella patens. *Curr. Opin. Plant Biol.*

**Schaefer, D.G., Delacote, F., Charlot, F., Vrielynck, N., Guyon-Debast, A., Guin, S. Le, Neuhaus, J.M., Doutriaux, M.P. and Nogué, F.** (2010) RAD51 loss of function abolishes gene targeting and de-represses illegitimate integration in the moss Physcomitrella patens. *DNA Repair (Amst).*

**Schaefer, D.G. and Zrÿd, J.P.** (1997) Efficient gene targeting in the moss Physcomitrella patens. *Plant J.*

**Silflow, C.D.** (2001) Assembly and Motility of Eukaryotic Cilia and Flagella. Lessons from Chlamydomonas reinhardtii. *Plant Physiol.*, **127**, 1500–1507.

**Stackelberg, M. Von, Rensing, S.A. and Reski, R.** (2006) Identification of genic moss SSR markers and a comparative analysis of twenty-four algal and plant gene indices reveal species-specific rather than group-specific characteristics of microsatellites. *BMC Plant Biol.*

**Stewart, K.D. and Mattox, K.R.** (1975) Comparative cytology, evolution and classification of the green algae with some consideration of the origin of other organisms with chlorophylls A and B. *Bot. Rev.*, **41**, 104–135.

**Strotbek, C., Krinninger, S. and Frank, W.** (2013) The moss Physcomitrella patens: Methods and tools from cultivation to targeted analysis of gene function. *Int. J. Dev. Biol.*, **57**, 553–564.

**Transeau, E.N.** (1951) *The Zygnemataceae (fresh-water conjugate algae) with keys for the identification of genera and species, and seven hundred eighty-nine illustrations.*,.

**Umen, J.G.** (2014) Green Algae and the Origins of Multicellularity in the Plant Kingdom. *Cold Spring Harb. Perspect. Biol.*

**Vries, J. de and Archibald, J.M.** (2018) Plant evolution: landmarks on the path to terrestrial life. *New Phytol.*

**Wei, D., Cui, Y., He, Y., et al.** (2017) A genome-wide survey with different rapeseed ecotypes uncovers footprints of domestication and breeding. *J. Exp. Bot.*

**Wettstein, F. v** (1924) Morphologie und physiologie des formwechsels der moose auf genetischer grundlage. I. *Z. Indukt. Abstamm. Vererbungsl.*

**Wickett, N.J., Mirarab, S., Nguyen, N., et al.** (2014) Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci.*

**Wickstead, B. and Gull, K.** (2012) Evolutionary biology of dyneins. In *Dyneins*.

**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.

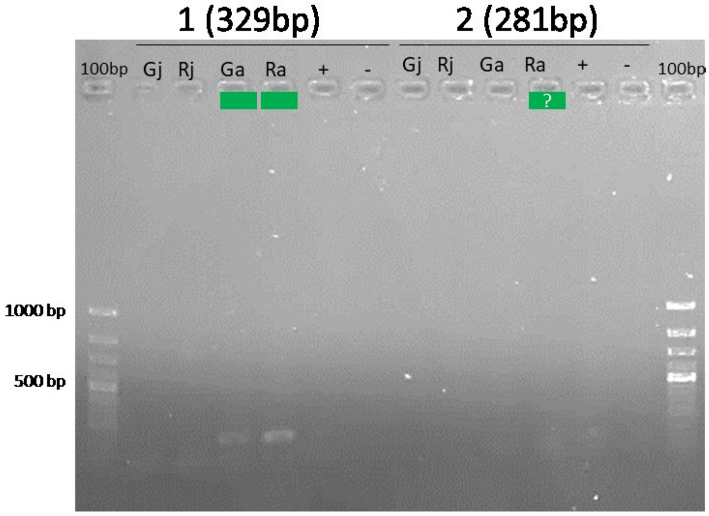
**Zemach, A., McDaniel, I.E., Silva, P. and Zilberman, D.** (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science (80-. ).*, **328**, 916–919.

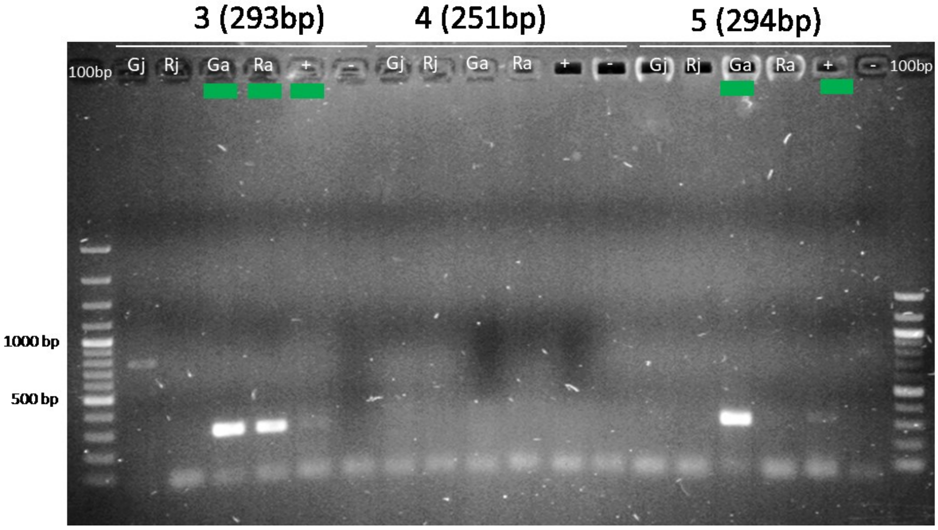
**Zimorski, V., Ku, C., Martin, W.F. and Gould, S.B.** (2014) Endosymbiotic theory for organelle origins. *Curr. Opin. Microbiol.*

# 11 Attachment

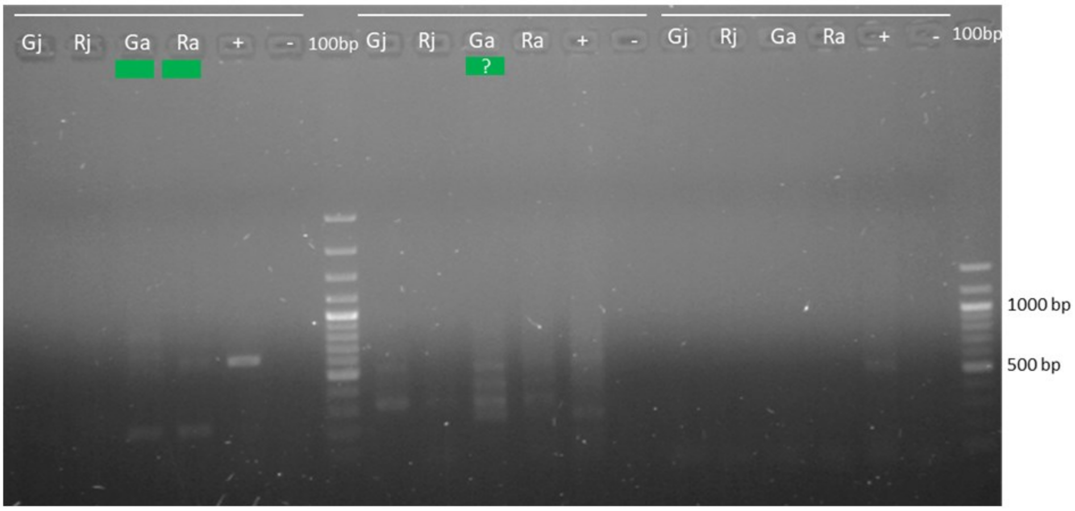
## 11.1 RTPCR analysis

RTPCR and qPCR primer were designed to, if possible, include at least one exon-exon border to restrict DNA amplification. Gel images of RTPCR for all 24 candidate genes in Gransden (G) and Reute (R) juvenile (j) and adult (a) apices were analysed. Green rectangle marks presence, or putative presence (question mark), of a PCR product with the predicted size, which is indicated in brackets next to the candidate gene number.

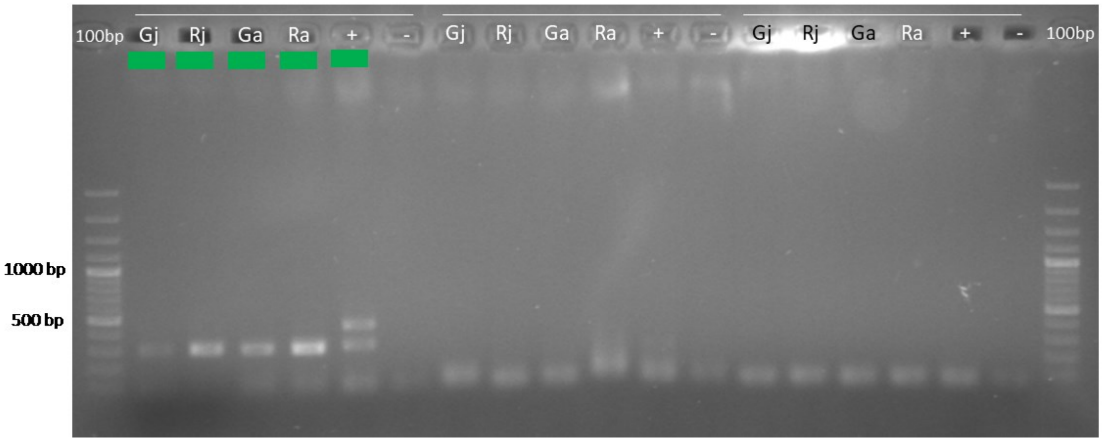




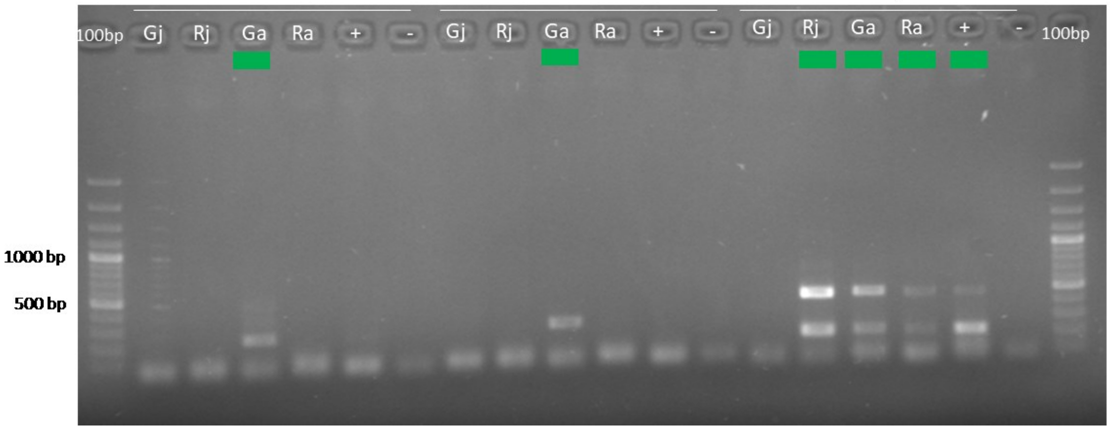
6 (237bp) 7 (256bp) 8 (239bp)

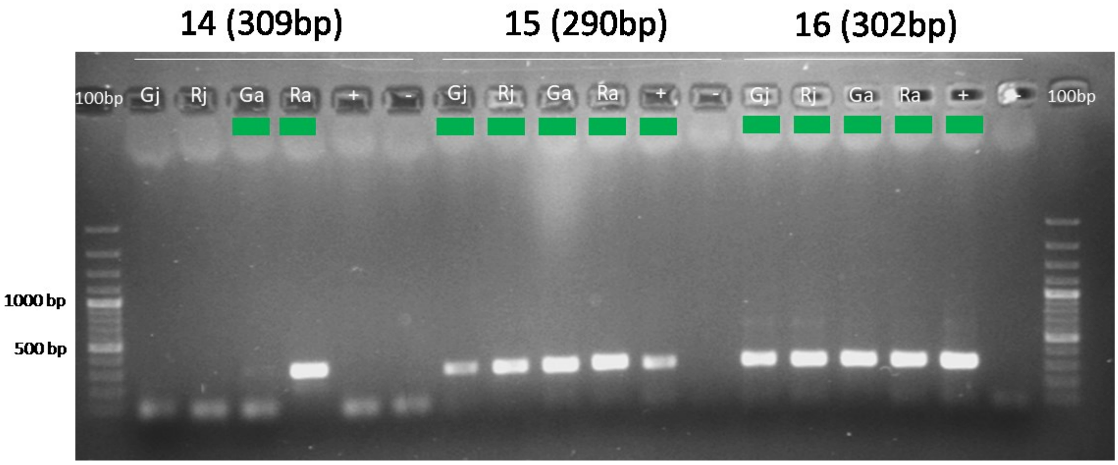


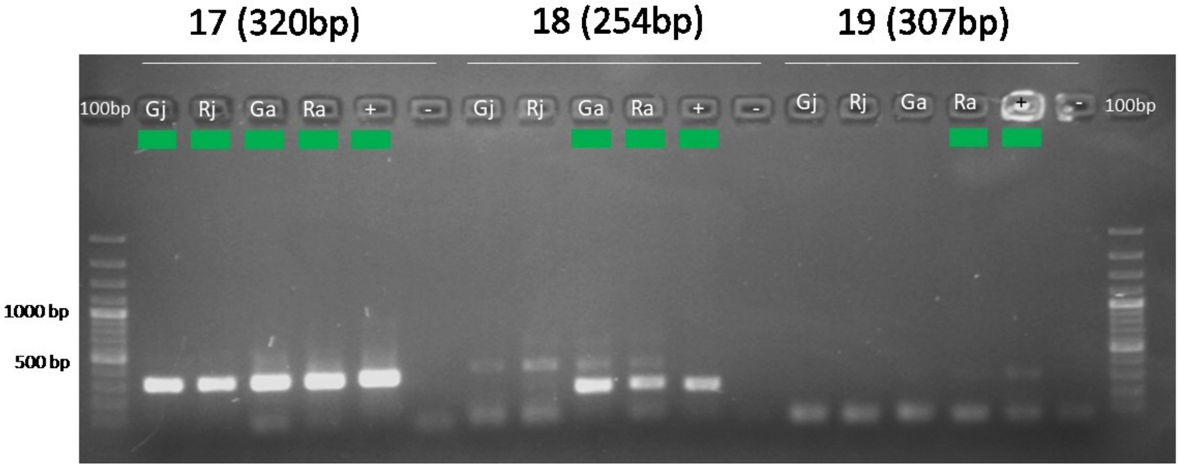
9 (288bp) 10 (259bp) 11 (363bp)

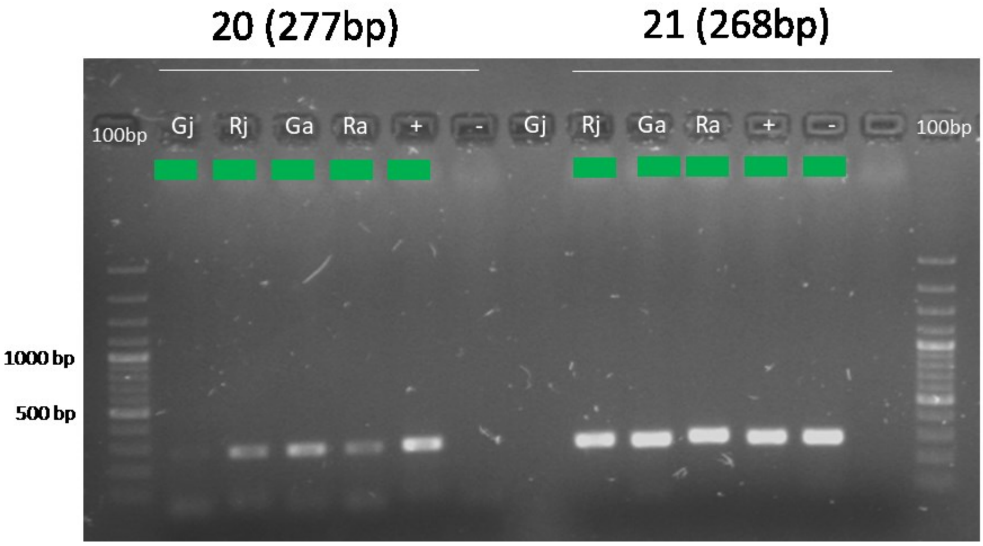


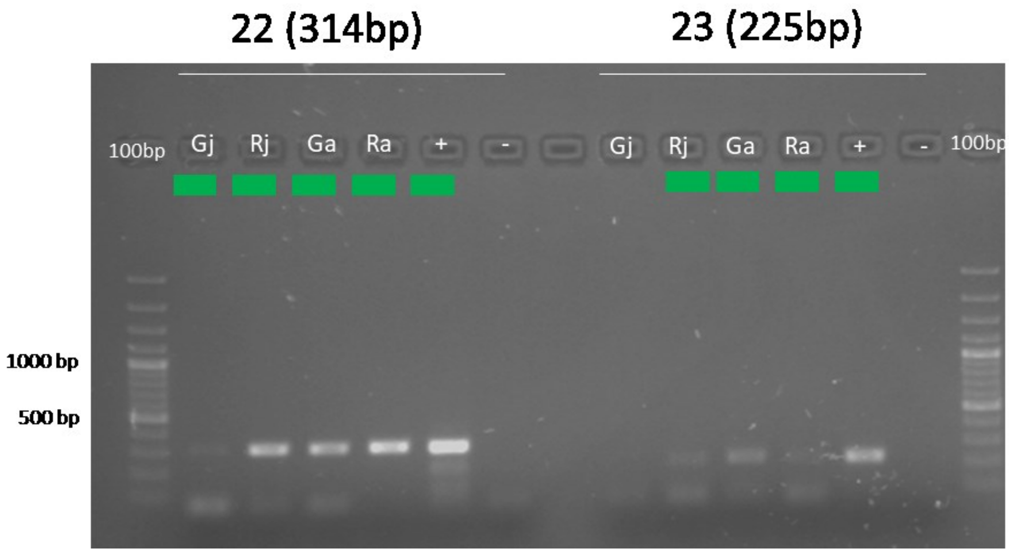
12 (253bp) 13 (305bp) 24 (223bp)




14 (309bp)
15 (290bp)
16 (302bp)
100bp
Gj
Rj
Ga
Ra
+
-
Gj
Rj
Ga
Ra
+
-
Gj
Rj
Ga
Ra
+
-
100bp
1000 bp
500 bp


17 (320bp)
18 (254bp)
19 (307bp)
100bp
Gj
Rj
Ga
Ra
+
-
Gj
Rj
Ga
Ra
+
-
Gj
Rj
Ga
Ra
+
-
100bp
1000 bp
500 bp

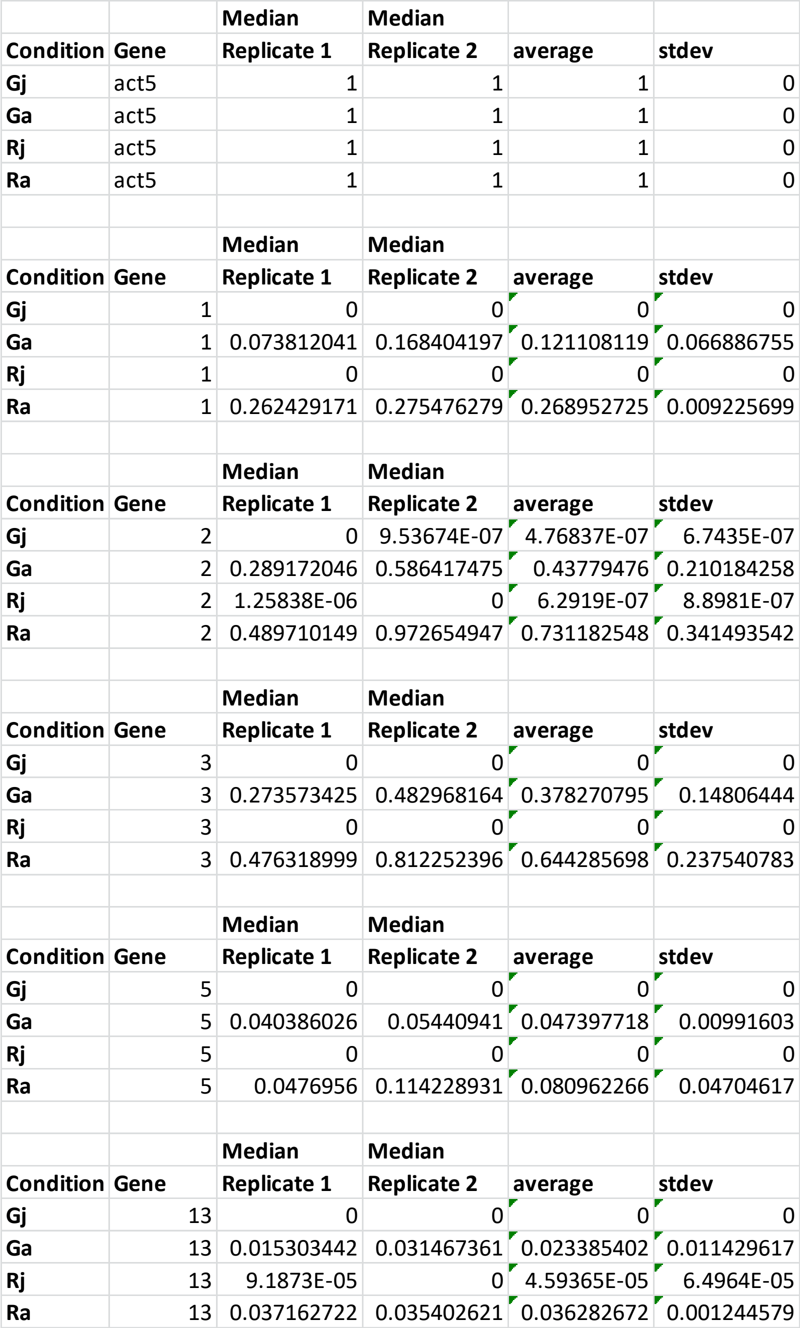

20 (277bp)
21 (268bp)
100bp
Gj
Rj
Ga
Ra
+
-
Gj
Rj
Ga
Ra
+
-
100bp
1000 bp
500 bp


22 (314bp)
23 (225bp)
100bp
Gj
Rj
Ga
Ra
+
-
Gj
Rj
Ga
Ra
+
-
100bp
1000 bp
500 bp

## 11.2 Candidate gene qPCR results

2-cq was calculated for each sample and normalized by reference gene median expression. Technical replicates were normalized by median, biological replicates by average. Average and standart deviation of all biological replicates was determined and used for visualisation (Tab. 2, Fig. 5).

*Table 2: qPR expression of the five selected candidate genes.*



| | | **Median** | **Median** | | |
|---|---|---|---|---|---|
| **Condition** | **Gene** | **Replicate 1** | **Replicate 2** | **average** | **stdev** |
| **Gj** | act5 | 1 | 1 | 1 | 0 |
| **Ga** | act5 | 1 | 1 | 1 | 0 |
| **Rj** | act5 | 1 | 1 | 1 | 0 |
| **Ra** | act5 | 1 | 1 | 1 | 0 |
| | | | | | |
| | | **Median** | **Median** | | |
| **Condition** | **Gene** | **Replicate 1** | **Replicate 2** | **average** | **stdev** |
| **Gj** | 1 | 0 | 0 | 0 | 0 |
| **Ga** | 1 | 0.073812041 | 0.168404197 | 0.121108119 | 0.066886755 |
| **Rj** | 1 | 0 | 0 | 0 | 0 |
| **Ra** | 1 | 0.262429171 | 0.275476279 | 0.268952725 | 0.009225699 |
| | | | | | |
| | | **Median** | **Median** | | |
| **Condition** | **Gene** | **Replicate 1** | **Replicate 2** | **average** | **stdev** |
| **Gj** | 2 | 0 | 9.53674E-07 | 4.76837E-07 | 6.7435E-07 |
| **Ga** | 2 | 0.289172046 | 0.586417475 | 0.43779476 | 0.210184258 |
| **Rj** | 2 | 1.25838E-06 | 0 | 6.2919E-07 | 8.8981E-07 |
| **Ra** | 2 | 0.489710149 | 0.972654947 | 0.731182548 | 0.341493542 |
| | | | | | |
| | | **Median** | **Median** | | |
| **Condition** | **Gene** | **Replicate 1** | **Replicate 2** | **average** | **stdev** |
| **Gj** | 3 | 0 | 0 | 0 | 0 |
| **Ga** | 3 | 0.273573425 | 0.482968164 | 0.378270795 | 0.14806444 |
| **Rj** | 3 | 0 | 0 | 0 | 0 |
| **Ra** | 3 | 0.476318999 | 0.812252396 | 0.644285698 | 0.237540783 |
| | | | | | |
| | | **Median** | **Median** | | |
| **Condition** | **Gene** | **Replicate 1** | **Replicate 2** | **average** | **stdev** |
| **Gj** | 5 | 0 | 0 | 0 | 0 |
| **Ga** | 5 | 0.040386026 | 0.05440941 | 0.047397718 | 0.00991603 |
| **Rj** | 5 | 0 | 0 | 0 | 0 |
| **Ra** | 5 | 0.0476956 | 0.114228931 | 0.080962266 | 0.04704617 |
| | | | | | |
| | | **Median** | **Median** | | |
| **Condition** | **Gene** | **Replicate 1** | **Replicate 2** | **average** | **stdev** |
| **Gj** | 13 | 0 | 0 | 0 | 0 |
| **Ga** | 13 | 0.015303442 | 0.031467361 | 0.023385402 | 0.011429617 |
| **Rj** | 13 | 9.1873E-05 | 0 | 4.59365E-05 | 6.4964E-05 |
| **Ra** | 13 | 0.037162722 | 0.035402621 | 0.036282672 | 0.001244579 |

## 11.3 Knock-out strategy and constructs

To generate the knock-out mutants, homologous recombination was used to exchange the complete gene locus (Fig. 6). This has the advantage of generating mutants with a known genome sequence and the complete loss of gene function.

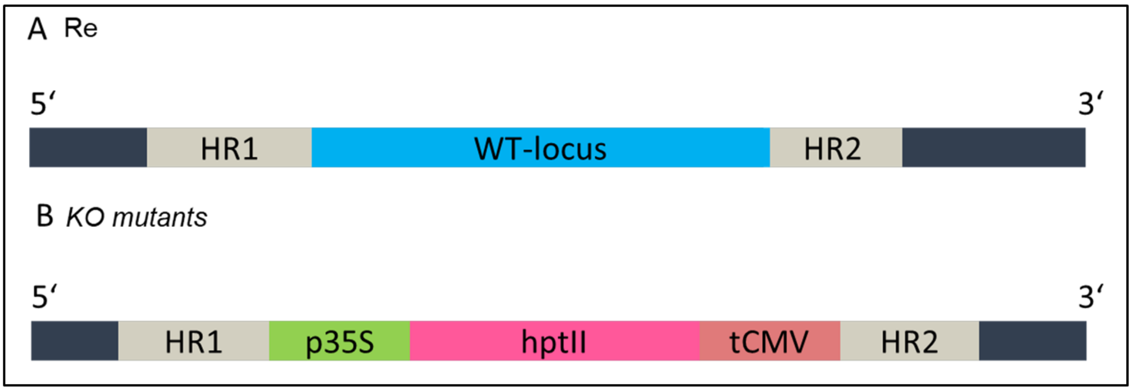


*Figure 6: Sketch of KO generation via homologous recombination. A: Wild type locus (blue) flanked by the homologous regions (HR) one and two. B: Mutant locus with the resistance cassette between both HRs. Resistance cassette consists out of a 35S promotor (p35S), a hygromycin resistance (hptII) and a cauliflower-mosaic-virus terminator (tCMV).*

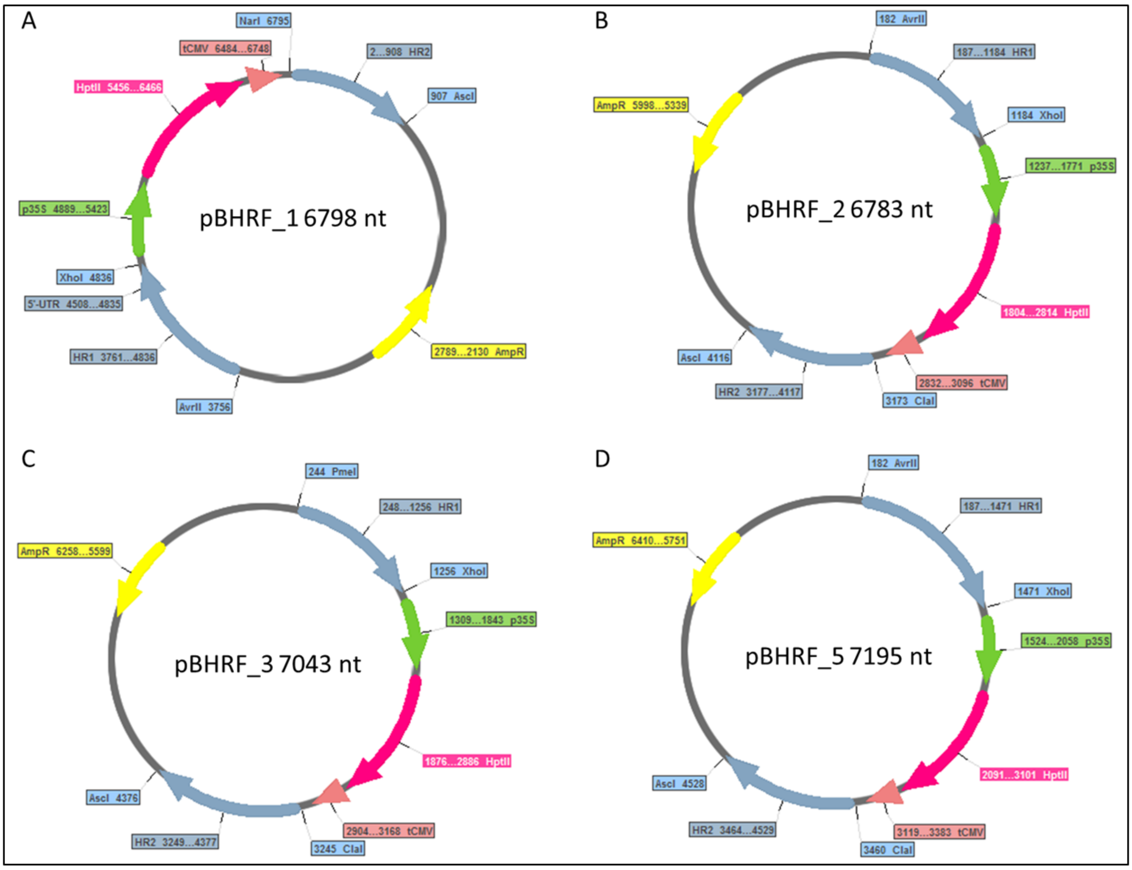


*Figure 7: KO-constructs for candidates 1,2,3 and 5 (A-D) and their corresponding size in nucleotides (nt). Each construct is based on pBHRF* (Schaefer *et al.*, 2010). *All constructs encompass an ampicillin resistance (AmpR) for amplification in* Escherichia coli *and a hyromycin resistance (HptII) for mutant selection in* P. patens. *Used restriction enzymes are shown in light blue and homologous regions (HR) in blue.*

# 12 Supporting information

Published supplements of publications contributing to this thesis. Partial supplements provide the link to the supporting information at the journal homepage.

## 12.1 Meyberg *et al.,* 2019

Supplementary tables are available at BioRxiv:
https://www.biorxiv.org/content/10.1101/728691v1.supplementary-material

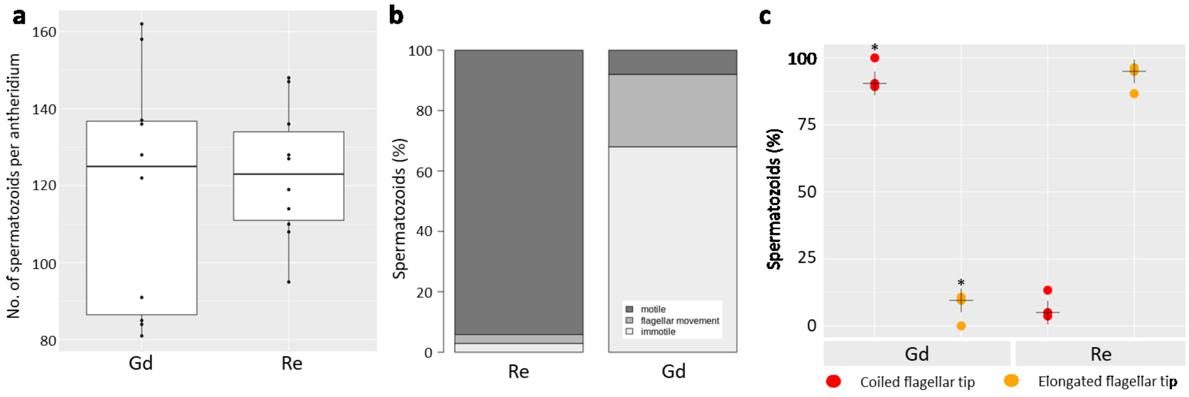


Figure S1: Detailed analysis of Gd and Re spermatozoids.

The number of spermatozoids per antheridium does not vary significantly between Gd and Re (A, n = 10 of each ecotype). Median-cantered box-dot plots representing 50% of the measurements within the white box, whereas the whiskers show the 1.5 interquartile range (IQR). Dots show individual measurements. In Re, 94% of the spermatozoids were motile, 6% were immotile of which 3% showed flagellar movement. In Gd, 8% of the spermatozoids where motile, 92% were immotile, of which 24% showed flagellar movement. The number of motile spermatozoids is significantly different between ecotypes (B, Gd n = 50, Re n = 103, $p < 0.01$ chi-square test). C: Gd spermatozoids (n = 51) show coiled flagellar tips statistically significantly more often compared to Re spermatozoids (n = 62, $p < 0.01$, t-test). Median: black cross.

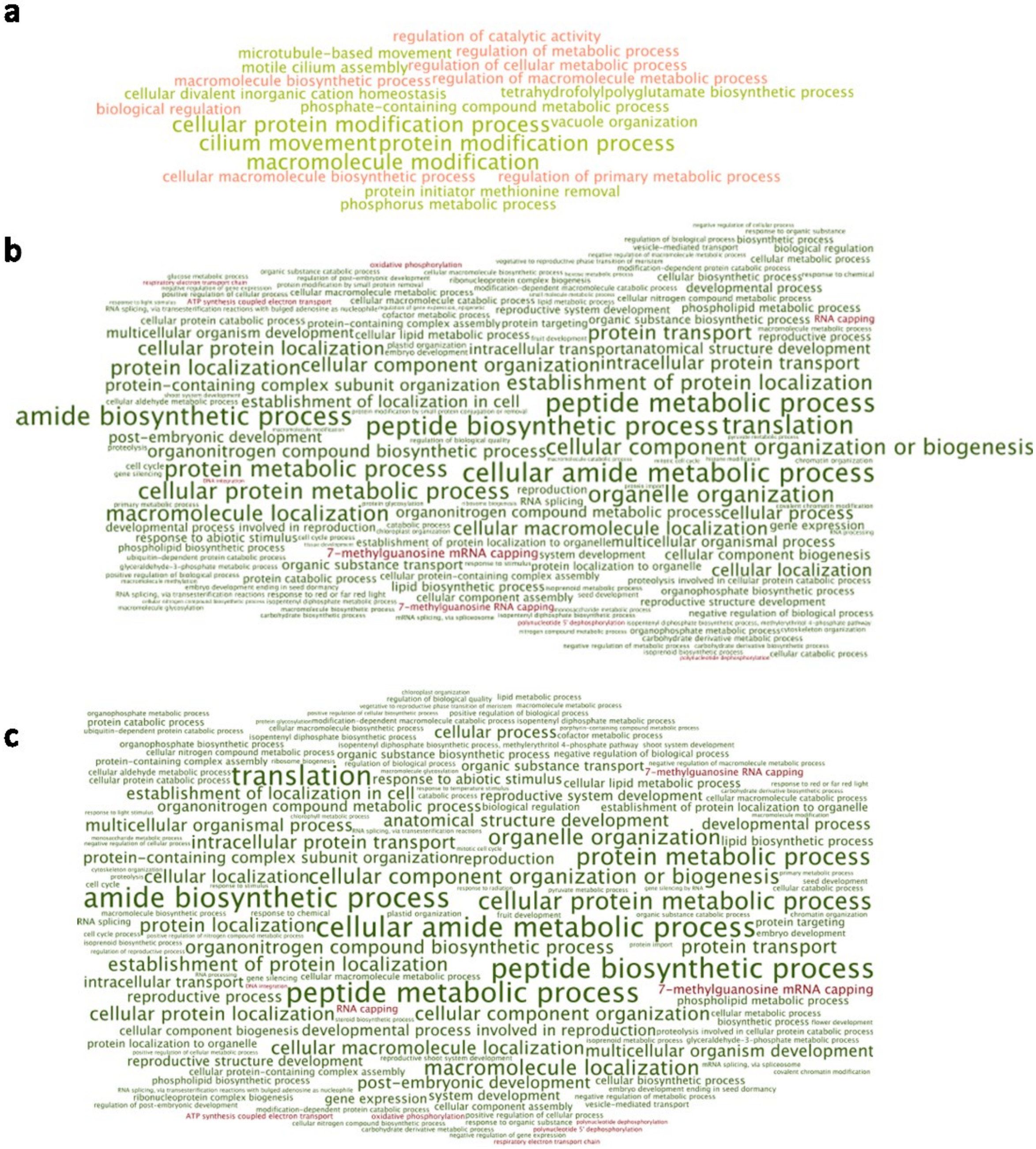


Figure S2: GO bias analysis and word cloud visualization.

Over-represented terms are shown in green, whereas under-represented terms are shown in red. Larger font size correlates with a higher significance level. Genes affected by DMPs between Re and Gd in any of the contexts (CG, CHH, CHG) show over-representation of terms connected to cilia motility as well as terms related to macromolecule modification (A). Genes expressed in Gd (B) and Re (C) antheridia bundles show over-representation of polyamine and peptide biosynthetic processes as well as under-represented terms associated with mRNA capping.

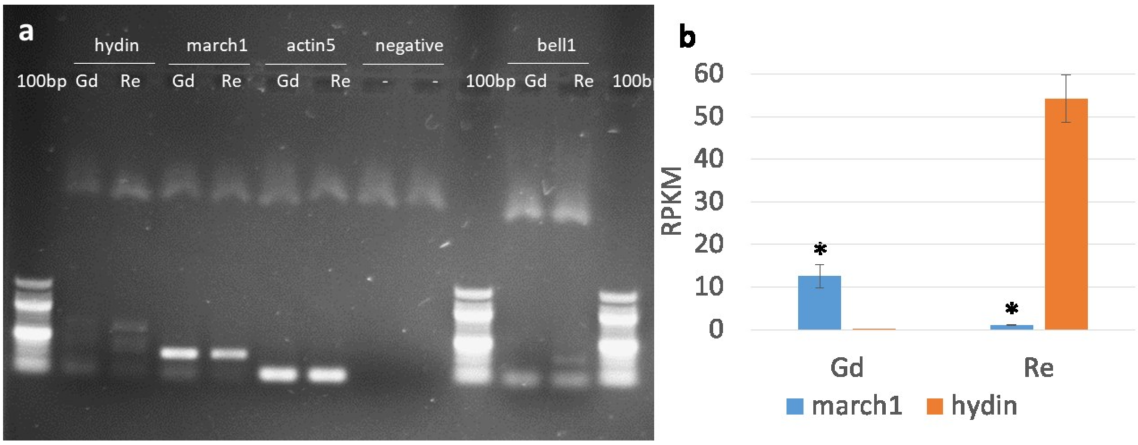


Figure S3: Expression analysis of genes expressed in antheridia bundles.

A: RT-PCR expression in adult gametophore apices of Gd and Re of hydin, march1 and bell1 genes matches RNA-seq data. B: RNA-seq expression of Gd and Re march1 (blue) and hydin (orange) show significant differences between Gd and Re (*, p < 0.01 t-test).

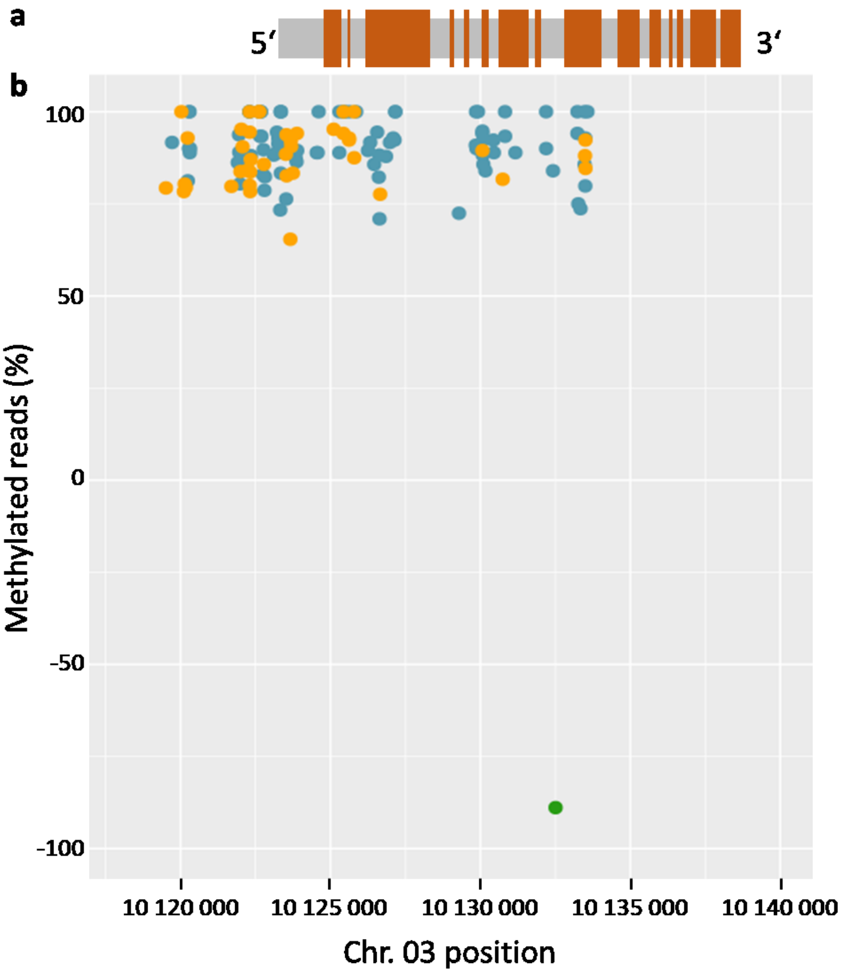


Figure S4: Hydin gene.

A: Exon and intron pattern of hydin. B: Aligned with the differentially methylated positions (DMPs) whereas 0 to 100 represents positions methyated with 0-100% in Gd and 0 to -100 represents positions methylated with 0-100% in Re. Gene body and 5‘-UTR of Gd hydin are highly affected by DNA methylation showing CHG (94, blue) and CHH (41, yellow) methylation, whereas in Re a single CG (green) methylation could be detected in an intron.

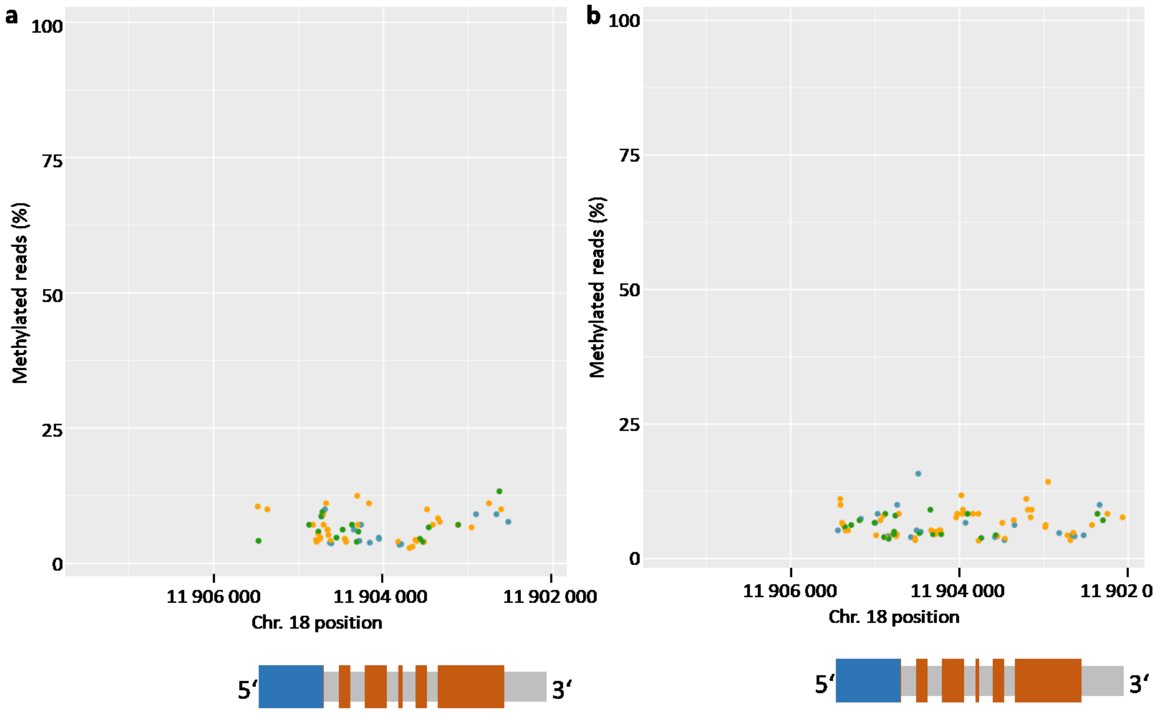


Figure S5: March1 gene body methylation.

Percentage of methylated reads per position per pattern of Gd (A) and Re (B) adult gametophores. The V3.3 gene model of march1 is shown in grey, the UTRs are marked in blue and exons are shown in orange. Very low levels of methylation are detectable in march1 and 2kb upstream in both Gd and Re. CHG (blue), CHH (yellow) and CpG (green) marks are shown.

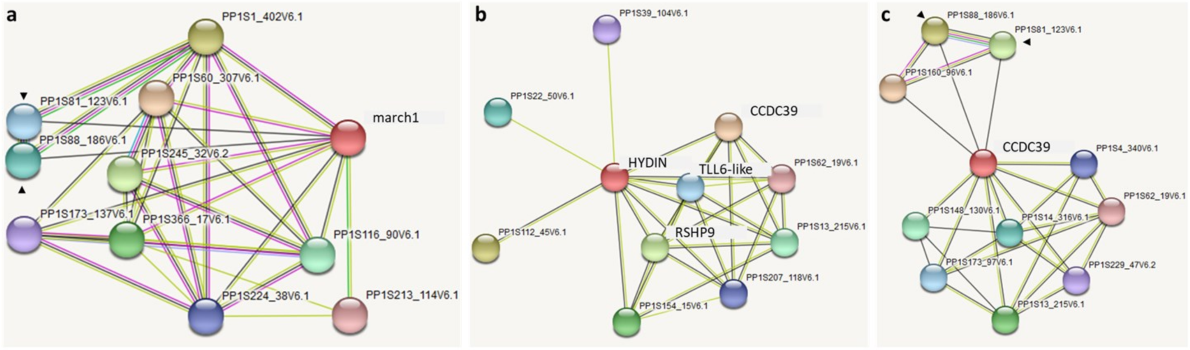


Figure S6: Network analysis of genes associated with proper flagellar functionality.

A: March1 network shows two protein phosphatases being co-expressed in common with CCDC39 (C, black arrows). B: Network analysis of hydin shows connectivity with genes affecting sperm cells (CCDC39, RSPH9, TLL6-like). C: Network analysis of CCDC39. Coexpressed protein phosphatases marked by black arrows. Line color specifies connection type between analysed proteins: dark grey marks co-expression, light green marks literature analysis, sky blue marks protein homology, grass green marks gene neighborhood, cyan marks interaction shown by curated databases, pink marks experimentally determined interactions.

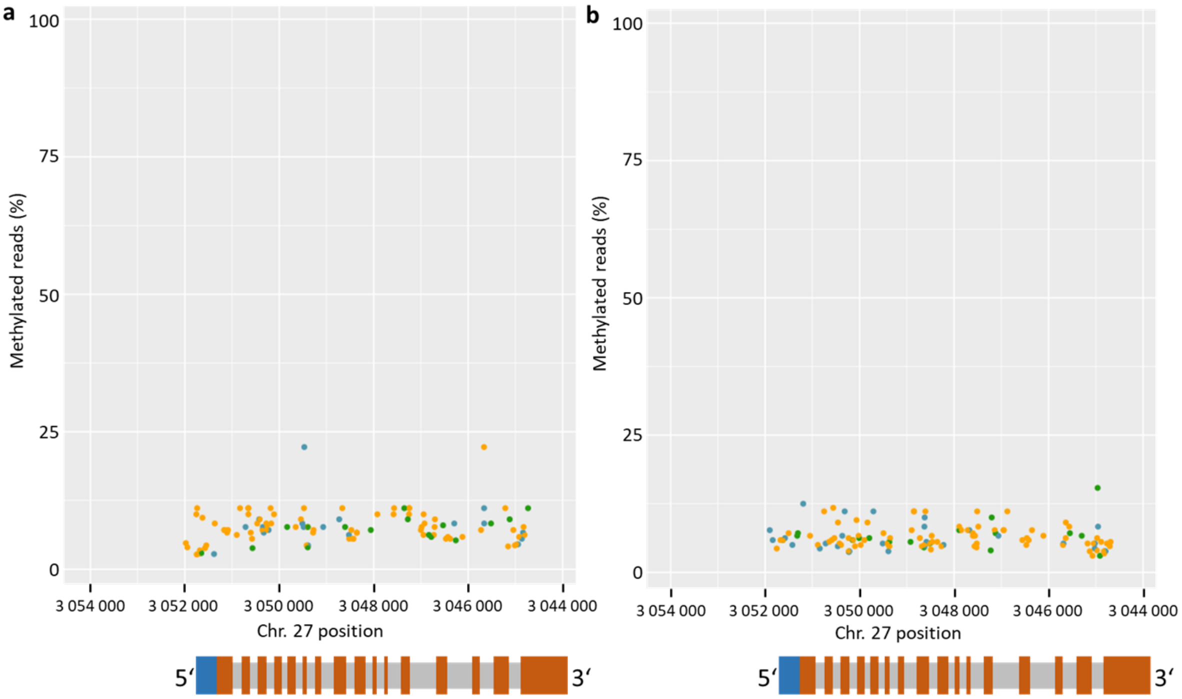


Figure S6: Percentage of ccdc39 methylated reads per position.

Gd (A) and Re (B) adult gametophores. Gene model annotation as described in Fig. S5. Low gene body methylation is detectable in ccdc39 and 2kbp upstream in both Gd and Re.

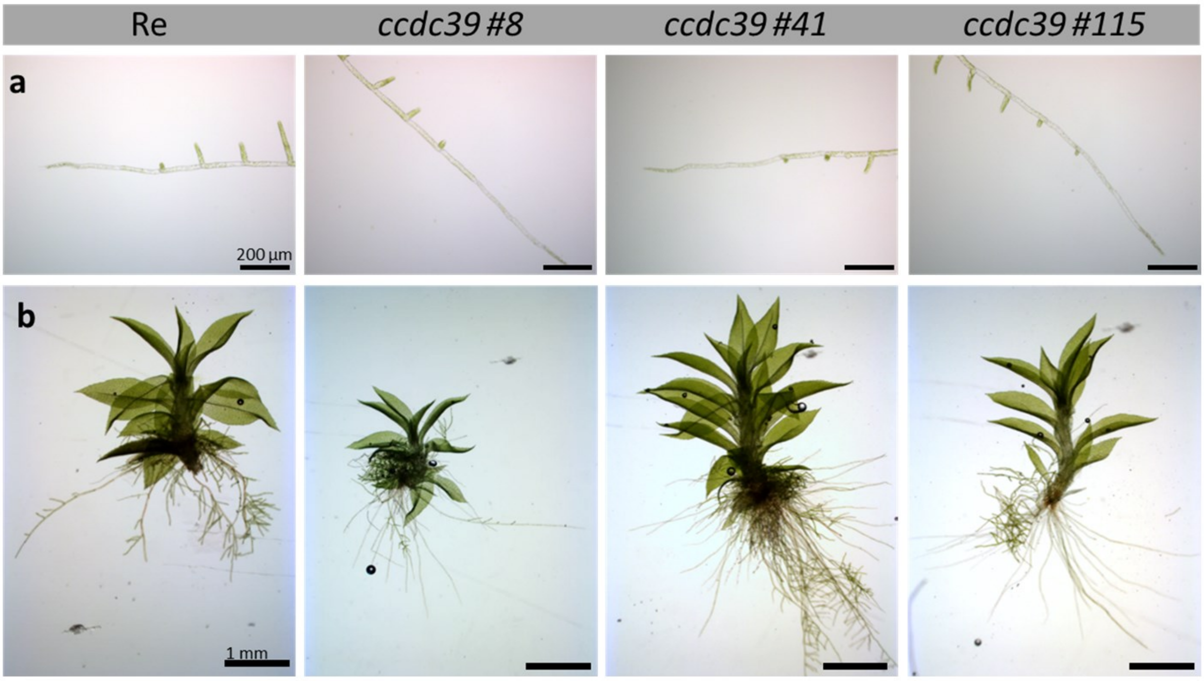


Figure S7: Protonema (A) and gametophore (B) development of Re compared to *ccdc39* mutant strains.

No obvious morphological differences could be observed in the analysed mutant lines.

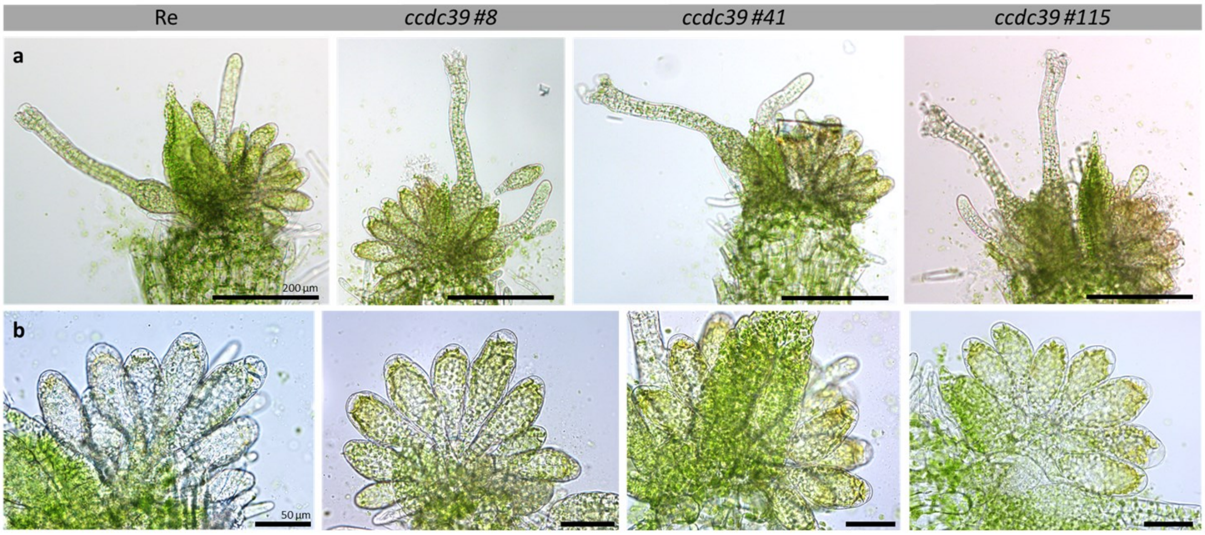


Figure S8: Gametangia development of *ccdc39* mutant strains compared to Re.

No obvious morphological differences can be detected. A: dissected apices with both archegonia and antheridia, B: detailed images of the antheridia bundles.

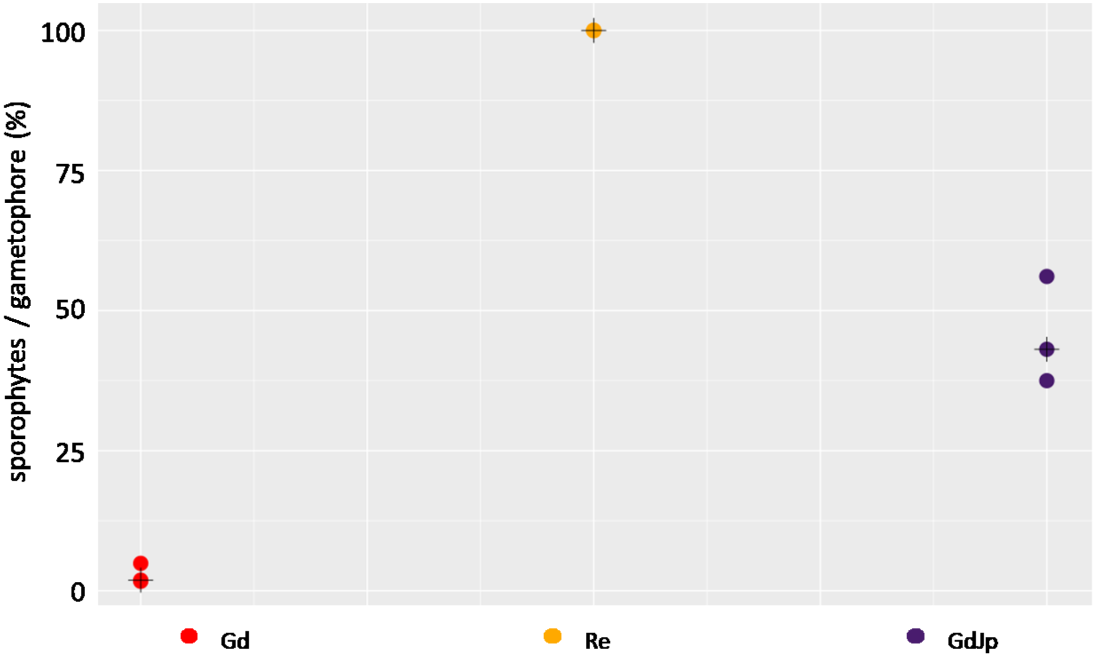


Figure S9: Number of sporophytes per gametophore developed under selfing conditions for *P. patens* ecotypes Re, Gd and GdJp.

Re (n = 3, 730 gametophores) shows median 100% sporophyte per gametophore, Gd (n = 3, 523 gametophores) shows 1.9% sporophytes per gametophore and GdJp (n = 3, 653 gametophores) 43.1%. GdJp develops significantly more sporophytes per gametophore in comparison with Gd ($p < 0.05$, t-test).

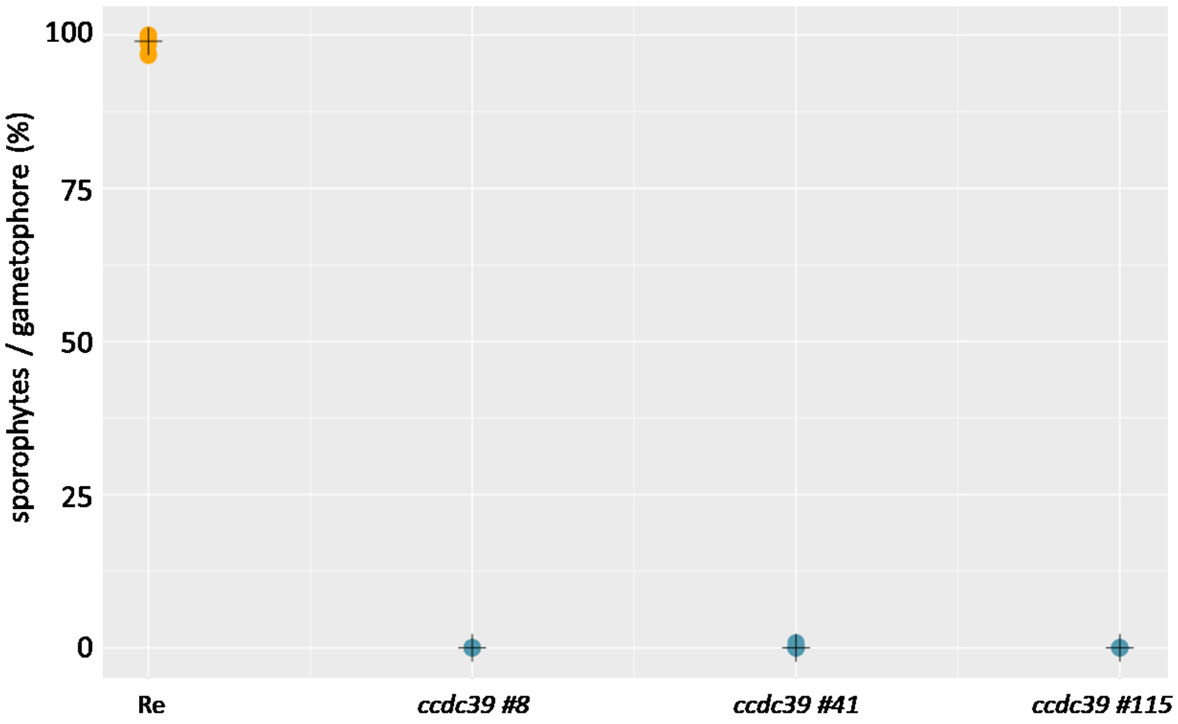


Figure S10: Sporophyte per gametophore development for all three independently analysed *ccdc39* strains.

All *ccdc39#8* (n = 546), *ccdc39#41* (n = 714) and *ccdc39#115* (n = 537) show in median 0% of sporophytes in comparison to median 99% of sporophytes for the corrosponding wildtype background Re.

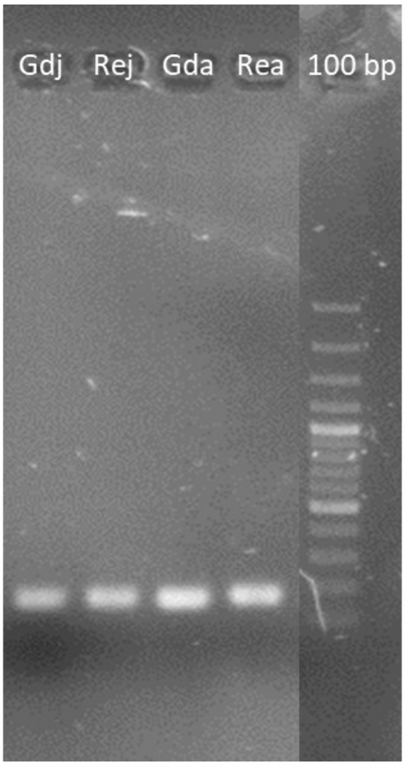


Figure S11: Expression level of act5 in juvenile (j) and adult (a) apices of Gd and Re show similar expression.

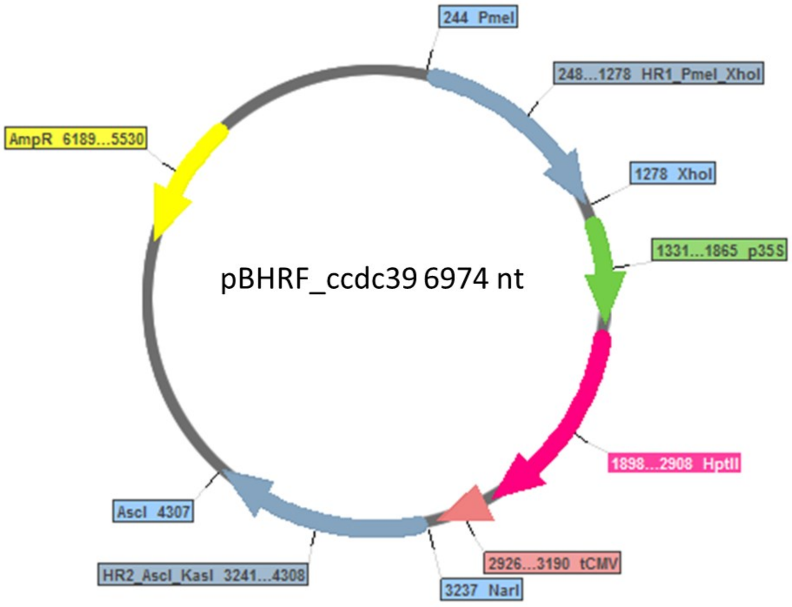


Figure S12: Final vector for amplification via ampicillin selection (yellow) of the knock out cassette flanked by the enzymes *Pme*I and *Asc*I. HR1 and HR2 (blue) flank the resistance cassette consisting out of the 35S promotor (green), the *hptII* resistance gene (pink) and the CMV terminator (red).

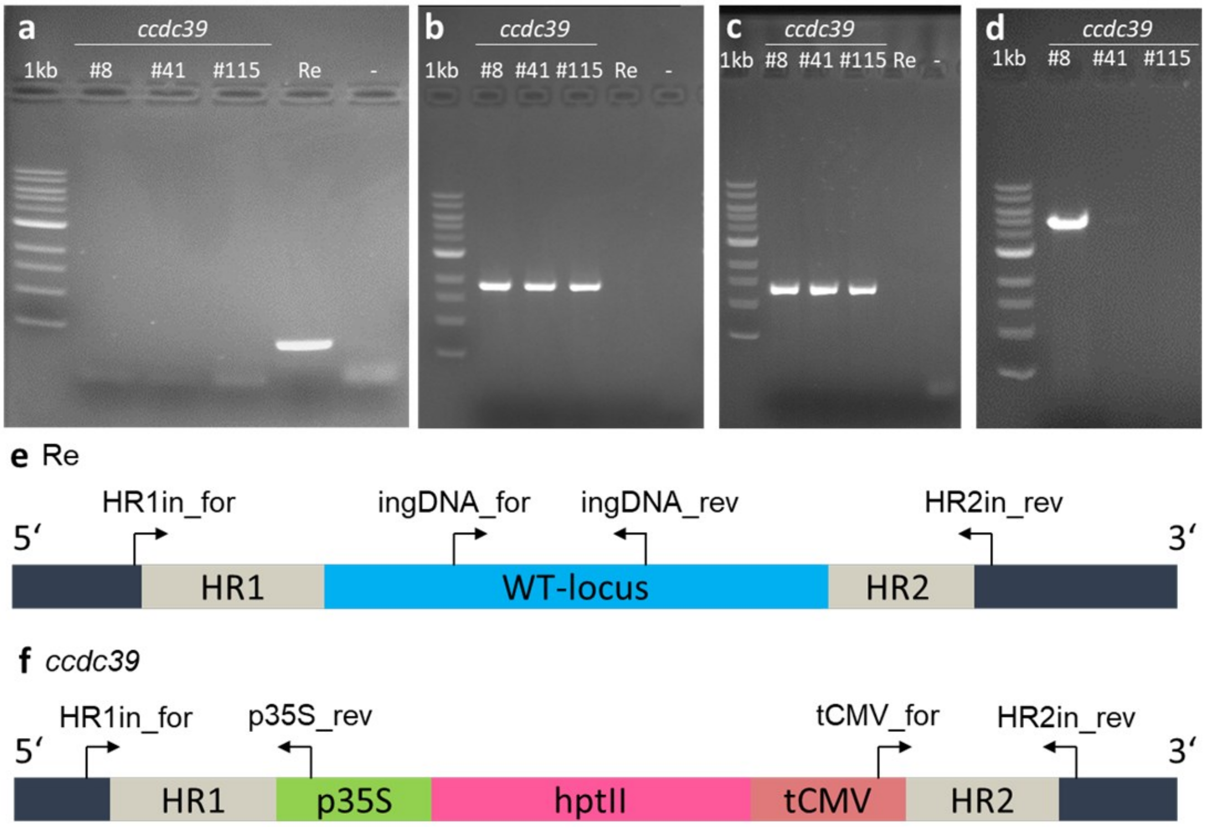


Figure S13: Gel images and sketch (E,F) of performed genotyping on Re and *ccdc39* gDNA.

Presence of the wild type locus was tested using ingDNA_for/rev (A). HR1 (B) and HR2 (C) presence was verified using HR1in_for/p35S_rev and tCMV_for/HR2in_rev. Full length amplification was performed using HR1in_for/HR2in_rev (D).

## 12.2 Perroud et al., 2019

Supporting information is accessible at the journal homepage: https://onlinelibrary.wiley.com/doi/full/10.1111/plb.12840

## 12.3 Lang *et al.*, 2018

Download complete supplemental material: https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Ftpj.13801&file=tpj13801-sup-0001-AppendixS1.docx

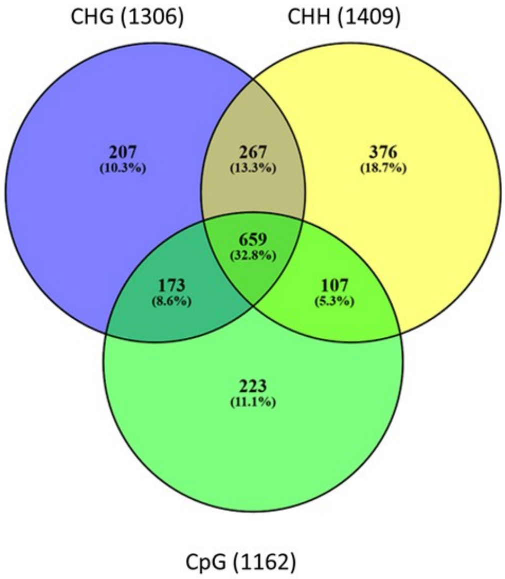
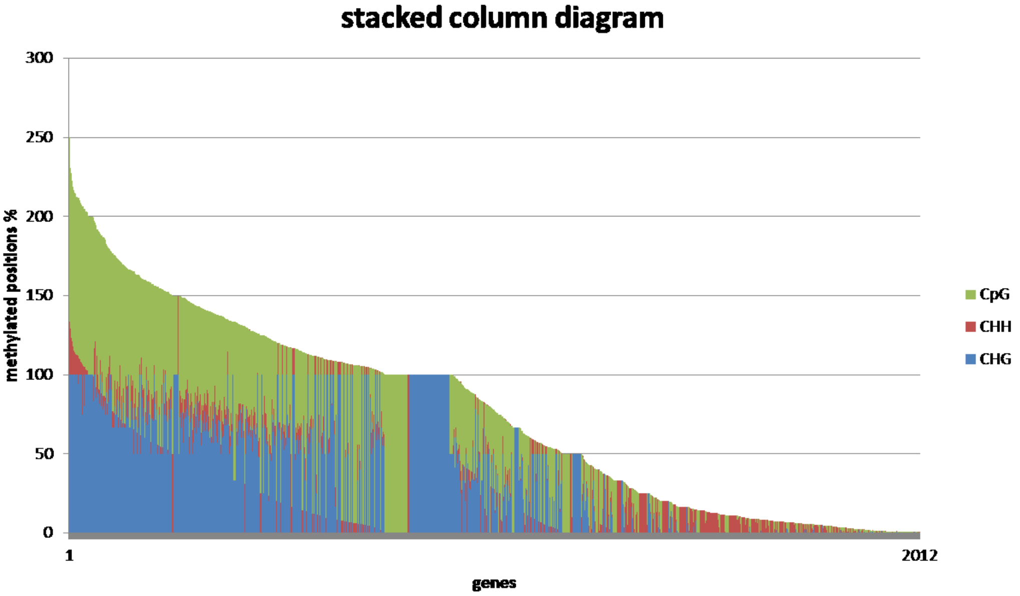
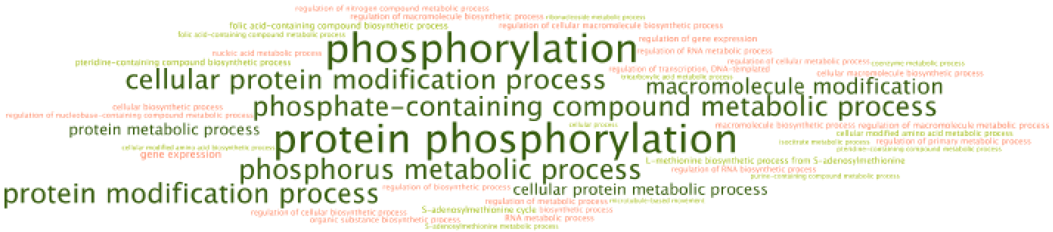


**Figure S29:** Genes affected by body methylation.

A



B



C

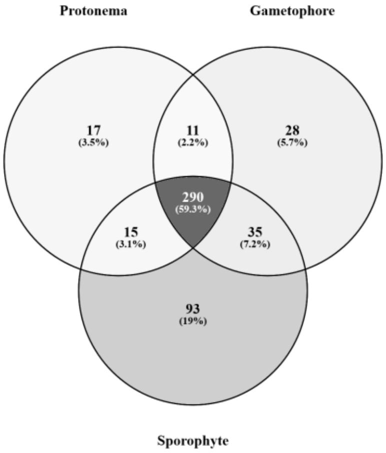


**Figure S30: Methylated genes.**
A) Distribution of methylated positions (%) across all three contexts in genes showing GBM. B) GO bias analysis of methylated genes with RPKM > 0, biological process ontology. C) Venn diagram showing expression of methylated genes in the three principal developmental stages.

| | G_CHG | | | G_CHH | | | G_CG | |
|---|---|---|---|---|---|---|---|---|
| average | methylated % | GC % | average | methylated % | GC % | average | methylated % | GC % |
| ALL | 62,45841279 | 37,9643415 | ALL | 10,72077003 | 35,7739744 | ALL | 48,26382088 | 39,7191394 |
| NO RPKM | 69,65318023 | 37,5602105 | NO RPKM | 11,52950977 | 35,4695635 | NO RPKM | 58,38698921 | 39,2857488 |
| RPKM >0 <2 | 42,10277254 | 38,7865405 | RPKM >0 <2 | 5,734804923 | 36,7334532 | RPKM >0 <2 | 26,75197331 | 39,3490395 |
| RPKM >=2 | 13,08027413 | 41,5197368 | RPKM >=2 | 1,406773056 | 43,1433333 | RPKM >=2 | 5,705589374 | 43,4942735 |

**Table S18**: Comparison of methylation contexts with GC and expression evidence.

Most methylated genes show no expression (no RPKM); increasing RPKM is associated with increasing GC and decreasing methylation content. Percentage methylation varies between contexts whereas GC content is relatively stable.

## 12.4 Hiss *et al.*, 2017

Supporting information is available at the journal homepage: https://onlinelibrary.wiley.com/doi/full/10.1111/tpj.13501

# 13 Acknowledgements

Stay awhile and listen…

At this moment, which I always feared to never happen, I want to show my gratitude to you!

My very special gratitude goes to you Stefan, for your support and encouragement. I want to thank you for all the little nerdy moments, scientific discussions and glasses of awesome gin we had throughout the last years. Your trust in my abilities helped me like hell in my personal and scientific development. Your supervision made me a so much better scientist and I am very grateful for this.

I also want to thank you Manuel Hiss, Fabian Haas and Pierre-François, for the extraordinary time we spent together during the time of my PhD. For all the fruitful discussions, your support and the pleasant breaks we had in between. I am glad we will share co-authorships forever.

A very warm thank you, I want to address to my family. Thank you for all the love, encouragement and support! Without all of you, this work would not have been possible. You were the nutrients I needed to grow!

I also want to thank all of my lab members. You made it every day and without your support, life would have been rough.

Thank you, to all my friends who accompanied me! No matter when our paths met, your friendship matters and I am very grateful to call all of you my friends!

Finally, I want to thank you Anke! Thank you for giving me my beloved and unforgotten Zeiss microscope, which probably flattened the way I am walking on now!

So long, and thanks for all the fish!

# 14 Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe und dies mein erster Promotionsversuch ist. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quellen gekennzeichnet. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- beziehungsweise Beratungsdiensten (Promotionsberater oder anderer Personen) in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt. Die Bestimmungen der Promotionsordnung der Fakultät für Biologie der Universität Marburg sind mir bekannt, insbesondere weiß ich, dass ich vor Vollzug der Promotion zur Führung des Doktortitels nicht berechtigt bin.

Marburg,