P. patens
genomic and transcriptomic analyses

Dissertation

„kumulativ“

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

des Fachbereichs Biologie der Philipps-Universität Marburg

Vorgelegt von

Fabian Benjamin Haas

aus Lahr/Schwarzwald, Baden-Württemberg, Deutschland

Marburg an der Lahn

Mai 2020

Die vorliegende Dissertation wurde von März 2015 bis Mai 2020 am Fachbereich Biologie, Pflanzenzellbiologie unter Leitung von Prof. Dr. Stefan Rensing angefertigt.

Vom Fachbereich Biologie der Philipps-Universität Marburg (Hochschulkennziffer 1180) als Dissertation angenommen am .

Erstgutachter: Prof. Dr. Stefan Rensing

Zweitgutachter: Prof. Dr. Gerhard Leubner-Metzger

Prof. Dr. Dominik Heider

Prof. Dr. Lars Voll

Tag der Disputation:

Für meinen Honig

und den ganzen Honigtopf.

„Man bereut nie, was man getan,
sondern nur, was man nicht getan hat.“

Marcus Aurelius

I. Publications and contributions

All research work achieved during my time as a PhD student and presented in this thesis is either published in peer reviewed scientific journals or at the time of submitting this thesis, are under review. The list of publications and my contribution to each publication are listed below.

I.I Publications contributing to this thesis

The *Physcomitrella 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 F. 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 C. Villarreal Aguilar, 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, Gerald A. Tuskan, Florian Maumus, Jérome Salse, Jeremy Schmutz, Stefan A. Rensing

***The Plant Journal** (2018), Vol. 93, Issue 3, 515-533, DOI: 10.1111/tpj.13801*

Awarded silver prize for the best Original Article first published in TPJ in 2018.
SEB-Wiley-TPJ award for outstanding papers published in TPJ in 2018.

My contribution: Performed a genome sequence contamination analysis together with Stefan A. Rensing, called SNPs on the new chloroplast and mitochondrial genome. Set up a new GO database and annotated *P. patens* v3.1 genes using different sources (NCBI nr, TAIR, GO-terms, UniProt and Cosmoss), together with Faezeh Donges. Additionally, calculated the RNA-seq evidence for all gene models and uploaded the genome and associated tracks to the CyVerse Comparative Genomics platform CoGe.

Thesis chapter 5.1

The *Physcomitrella patens* gene atlas project: large-scale RNA-seq based expression data

Pierre-François Perroud, **Fabian B. Haas**, Manuel Hiss, Kristian K. Ullrich, Alessandro Alboresi, Mojgan Amirebrahimi, Kerrie Barry, Roberto Bassi, Sandrine Bonhomme, Haodong Chen, Juliet C. Coates, Tomomichi Fujita, Anouchka Guyon-Debast, Daniel Lang, Junyan Lin, Anna Lipzen, Fabien Nogué, Melvin J. Oliver, Inés Ponce de León, Ralph S. Quatrano, Catherine Rameau, Bernd Reiss, Ralf Reski, Mariana Ricca, Younousse Saidi, Ning Sun, Péter Szövényi, Avinash Sreedasyam, Jane Grimwood, Gary Stacey, Jeremy Schmutz, Stefan A. Rensing

The Plant Journal *(2018), Vol. 95, Issue 1, 168-182, DOI: 10.1111/tpj.13940*

My contribution: Designed and implemented the RNA-seq DEG pipeline based on a draft by Kristian K. Ullrich. This contains RNA-seq sample pre-processing (quality control and trimming, reference mapping) as well as counts and normalized RPKM/FPKM value calculation, DEG calling, GO-bias analysis and DEG clustering. Generated *P. patens* v3.3 gene annotation using different sources as described in (Lang *et al.*, 2018). Setting up a gene model version lookup table and summarizing the results in the supplementary table data S1 are also my contributions.

Thesis chapter 5.2

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 Hanke, Viktor Kratz, Adrian F. 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

The Plant Journal *(2019), Vol. 102, Issue 1, 165-177, DOI: 10.1111/tpj.14607*

My contribution: Calculated the gene expression values for the RNA-seq samples and generated the gene version lookup table as the data input for the PpGML database. Together with Noe Fernandez-Pozo, set up the server infrastructure.

Thesis chapter 5.3

Single nucleotide polymorphism charting of *P. patens* reveals accumulation of somatic mutations during *in vitro* culture on the scale of natural variation by selfing

Fabian B. Haas, Noe Fernandez-Pozo, Rabea Meyberg, Pierre-François Perroud, Marco Göttig, Nora Stingl, Denis Saint-Marcoux, Jane Langdale, Stefan A. Rensing

Frontiers in Plant Science, Plant Systematics and Evolution *(2020), special issue: "Highlights of IAB IMOSS SEB 2019 Joint Conference", DOI: 10.3389/fpls.2020.00813*

My contribution: Established the SNP calling pipeline. The pipeline includes RNA-seq and gDNA samples pre-processing (quality control and trimming, reference mapping), variant calling and validation of each sample and SNP post-processing (filtering, accession clustering). Performed a ploidy test for all samples, classified the different *P. patens* Gransden pedigrees, detected the natural variation at *P. patens* Wisconsin accession and estimated the annual number of mutations per base pair. Extracted restriction enzyme sites, located on SNPs and performed the SNP effect calculation. Designed the graphics (excluding the RFLP figures).Wrote the manuscript together with Stefan A. Rensing, Noe Fernandez-Pozo, Pierre-François Perroud and Rabea Meyberg.

Thesis chapter 5.4

I.II Additional publications not contributing to this thesis

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), Vol. 90, Issue 3, 606-620, DOI: 10.1111/tpj.13501*

My contribution: Supported the validation of mapping tools and SNP calling programs and filtered and clustered the gDNA SNPs.

The Biotrophic Development of *Ustilago maydis* Studied by RNA-Seq Analysis

Daniel Lanver, André N. Müller, Petra Happel, Gabriel Schweizer, **Fabian B. Haas**, Marek Franitza, Clément Pellegrin, Stefanie Reissmann, Janine Altmüller, Stefan A. Rensing, Regine Kahmann

***The Plant Cell** (2018), Vol. 30, Issue 2, 300-323, DOI: 10.1105/tpc.17.00764*

My contribution: Validated the CLC genomic workbench results with our RNA-seq DEG pipeline, together with Kristian K. Ullrich and Stefan A. Rensing.

The *Chara* Genome: Secondary Complexity and Implications for Plant Terrestrialization

Tomoaki Nishiyama, Hidetoshi Sakayama, Jan de Vries, Henrik Buschmann, Denis Saint-Marcoux, Kristian K. Ullrich, **Fabian B. Haas**, Lisa Vanderstraeten, Dirk Becker, Daniel Lang, Stanislav Vosolsobě, Stephane Rombauts, Per K.I. Wilhelmsson, Philipp Janitza, Ramona Kern, Alexander Heyl, Florian Rümpler, Luz Irina A. Calderón Villalobos, John M. Clay, Roman Skokan, Atsushi Toyoda, Yutaka Suzuki, Hiroshi Kagoshima, Elio Schijlen, Navindra Tajeshwar, Bruno Catarino, Alexander J. Hetherington, Assia Saltykova, Clemence Bonnot, Holger Breuninger, Aikaterini Symeonidi, Guru V. Radhakrishnan, Filip Van Nieuwerburgh, Dieter Deforce, Caren Chang, Kenneth

G. Karol, Rainer Hedrich, Peter Ulvskov, Gernot Glöckner, Charles F. Delwiche, Jan Petrášek, Yves Van de Peer, Jiri Friml, Mary Beilby, Liam Dolan, Yuji Kohara, Sumio Sugano, Asao Fujiyama, Pierre-Marc Delaux, Marcel Quint, Günter Theißen, Martin Hagemann, Jesper Harholt, Christophe Dunand, Sabine Zachgo, Jane Langdale, Florian Maumus, Dominique Van Der Straeten, Sven B. Gould, Stefan A. Rensing

***Cell** (2018), Vol. 174, Issue2, 448-464, DOI: 10.1016/j.cell.2018.06.033*

My contribution: Performed the genome sequence contamination analysis and did the functional gene model annotation as described in my contribution to (Lang *et al.*, 2018). I designed figure S5 B.

Genome Improvement and Genetic Map Construction for *Aethionema arabicum*, the First Divergent Branch in the Brassicaceae Family

Thu-Phuong Nguyen, Cornelia Mühlich, Setareh Mohammadin, Erik van den Bergh, Adrian E. Platts, **Fabian B. Haas**, Stefan A. Rensing and M. Eric Schranz

***G3** (2019), Vol. 9, Issue 11, 3521-3530*

My contribution: Set up and did the *in silico* sequencing process of the Oxford Nanopore MinION sequencing, performed the genome sequence contamination analysis and uploaded tracks to the CyVerse Comparative Genomics platform CoGe.

Characterization of evolutionarily 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

***New Phytologist** (2020), DOI: 10.1111/nph.16486*

My contribution: Processed the RNA-seq samples and calculated gene expression and DEGs.

Rocket Science: The Effect of Spaceflight on Germination Physiology, Ageing, and Transcriptome of *Eruca sativa* Seeds

Jake O. Chandler, **Fabian B. Haas**, Safina Khan, Laura Bowden, Michael Ignatz, Eugenia M. A. Enfissi, Frances Gawthrop, Alistair Griffiths, Paul D. Fraser, Stefan A. Rensing and Gerhard Leubner-Metzger

Life, Astrobiology *(2020), special issues: "Frontiers of Astrobiology", Vol. 10, Issue 4, DOI: 10.3390/life10040049*

My contribution: Evaluated the RNA-seq data quality and performed the pre-processing. Implemented the *de novo* transcriptome assembly, validated it, annotated the new transcripts, and generated expression data for each transcript, and called DEGs.

Are fungi-derived genomic regions related to antagonism toward fungi in mosses?

Guiling Sun, Shenglong Bai, Yanlong Guan, Shuanghua Wang, Qia Wang, Yang Liu, Huan Liu, Bernard Goffinet, Yun Zhou, Mathieu Paoletti, Xiangyang Hu, **Fabian B. Haas**, Noe Fernandez-Pozo, Alia Czyrt, Hang Sun, Stefan A. Rensing, Jinling Huang

Under revision: ***New Phytologist***

My contribution: Analysed the gene expression, gene body methylation and TE evidence at the two HET gene regions in *P. patens*.

II. Zusammenfassung

Der in dieser Arbeit verwendete Modelorganismus *Physcomitrium patens*, früher *Physcomitrella patens* (zu Deutsch Kleines Blasenmützenmoos), ist ein Laubmoos (Bryopsida) der Familie Funariaceae. Aufgrund einer sehr effizienten homologen Rekombination wurden Wissenschaftler bereits früh auf das Moos aufmerksam. *P. patens* war die erste Pflanze, außerhalb der Samenpflanzen, deren Genom vollständig sequenziert wurde (V1). Durch kontinuierliches Voranbringen und Verbessern der Genomassemblierung konnte diese in eine pseudo-chromosomale Struktur gegliedert werden (V3). Diese V3-Genomversion ist die Basis aller in dieser Arbeit durchgeführt Analysen.

Mit was für einer Genantwort reagieren Organismen infolge eines induzierten Stresses? Diese und weitere Fragen versucht das *U.S. Department of Energy Joint Genome Institute* (DOE JGI) im Rahmen des *Gene Atlas* Projekts[1] zu beantworten. In dessen Folge hunderte RNA-seq-Experimente durchgeführt wurden. Diese *P. patens* JGI *Gene Atlas* Daten sowie duzende weiterer RNA-seq-Experimente, unterschiedlichster Projekte, durfte ich in meiner Zeit als Doktorand analysieren. Die gleichbleibend hohe Qualität und Effizienz der Datenanalyse konnte mittels einer neu entwickelten RNA-seq-Pipeline gewährleistet werden, welche unter anderem zuverlässig differenziell exprimierte Gene (DEG) detektiert. Die Leistungsfähigkeit der RNA-seq-Pipeline wurde in unterschiedlichsten Projekten getestet und konnte sich auch gegen kommerzielle Software behaupten.

Die auf Basis meiner RNA-seq-Pipeline berechneten Expressionswerte wurden zusammen mit Microarray-Expressionsdaten, aus bereits veröffentlichten Projekten, auf der für diesen Zweck neu entwickelten Onlineplattform PEATmoss zur Verfügung gestellt. Benutzer können Expressionsdaten unterschiedlichster Experimente interaktiv miteinander vergleichen und Resultate in übersichtlicher Form herunterladen. Des Weiteren ermöglicht PEATmoss das Konvertieren aller bisher verwendeten *P. patens* Genmodellversionen untereinander.

Die zuvor beschriebenen Sequenzdaten beinhalten fünf unterschiedliche *P. patens* Ökotypen (Gransden, Kaskaskia, Reute, Villersexel und Wisconsin). Sequenzvariation unter den Ökotypen wurde bereits in früheren Studien aufgezeigt. In dieser Arbeit wurden erstmals Sequenzvariationen für fünf unterschiedliche Ökotypen auf Basis von RNA-seq-Daten untersucht. Hierfür wurde die RNA-seq-Pipeline modifiziert und die Funktionalität auf Variationsdetektion ausgeweitet. Eine klare Gruppierung der einzelnen Ökotypen und Gransden-Stämme kann beobachtet werden. Zusätzlich stellen wir mittels Restriktionsfragmentlängenpolymorphismus (RFLP) eine Methode bereit, die eine klare Identifikation einzelner *P. patens* Pflanzen in Laboren ermöglicht.

[1] https://jgi.doe.gov/doe-jgi-plant-flagship-gene-atlas/

III. Abstract

The model organism *Physcomitrium patens*, formerly *Physcomitrella patens* is a moss in the Funariaceae family. Due to *P. patens* ability to generate easily transgenic plants via homologous recombination, the interest of scientists worldwide was attracted. *P. patens* was the world's first completely sequenced non-seed plant genome (V1). Constant improvements of the genome assembly and the associated gene annotations resulted in the current *P. patens* pseudo-chromosomal genome version (V3). This genome version is the basis of all analyses performed in this thesis.

Since *P. patens* became a U.S. Department of Energy Joint Genome Institute (DOE JGI) plant flagship genome[1] and a member of the JGI Gene Atlas project[2], hundreds of *P. patens* RNA-seq samples were generated. During my time as a PhD student, I analysed the JGI Gene Atlas RNA-seq samples and several dozen other RNA-seq samples from different projects. These RNA-seq samples contained data from five different *P. patens* ecotypes/accessions (Gransden, Kaskaskia, Reute, Villersexel, and Wisconsin).To efficiently analyse this data, I developed a powerful RNA-seq pipeline to perform differentially expressed gene (DEG) calling. The performance of the RNA-seq pipeline was tested by comparing its results to commercial software solutions and multiple RNA-seq samples from different species.

My newly generated gene expression results, together with previous published expression data from a variety of other projects, were stored at our novel online tool PEATmoss. Furthermore, my gene version lookup tables were implemented in a database structure. This, allows PEATmoss users to find gene models of different gene annotation versions and to use them in PEATmoss.

With an updated version of my RNA-seq pipeline, I identified and analysed sequence variations in *P. patens* accessions. A clear clustering by individual accessions could be shown. I could demonstrate, that due to decades of vegetative propagation in laboratories, somatic mutations have accumulated in Gransden laboratory plants. In addition, we used restriction fragment length polymorphism (RFLP) to offer a simple method for quick identification of unknown *P. patens* plants.

[1] https://jgi.doe.gov/our-science/science-programs/plant-genomics/plant-flagship-genomes/

[2] https://jgi.doe.gov/doe-jgi-plant-flagship-gene-atlas/

1 Contents

2 Abbreviations

chr	(pseudo-)Chromosome
bp	Base pair
BS-seq	Bisulfite sequencing
RNA-seq	cDNA sequencing
CoV	Coefficient of variation
CoV	Coefficient of variation
CoGe	Comparative Genomics suite of web-tools
cDNA	Complementary DNA
DB	database
DNA	Deoxyribonucleic acid
DEG	Differentially expressed gene
FPKM	Fragments per kilobase and million fragments
GO	Gene Ontology
v1.2	Genome annotation version 1.2
v1.6	Genome annotation version 1.6
v3.3	Genome annotation version 3.3
V1	Genome assembly version 1
V3	Genome assembly version 3
gDNA	Genomic DNA
Gd	Gransden
Gd_DE	Gransden Germany
Gd_JP	Gransden Japan
Gd_CH	Gransden Switzerland
Gd_UK	Gransden United Kingdom
InDel	Insertion or deletion
JGI	Joint Genome Institute
Ka	Kaskaskia
kbp	Kilo base pair
LTR	Long terminal repeats
Mbp	Mega base pair
mRNA	Messenger RNA
MA	Mutation accumulation
NGS	Next-generation sequencing

nt	Nucleotide
PpGML	*P. patens* Gene Model Lookup database
P. patens	*Physcomitrium patens*
Pp	*Physcomitrium patens*
PCR	Polymerase chain reaction
PCA	Principal component analysis
qPCR	Quantitative (real-time) polymerase chain reaction
RPKM	Reads per kilobase and million reads
RFLP	Restriction fragment length polymorphism
Re	Reute
RNA	Ribonucleic acid
scaf	Scaffold
SSR	Simple sequence repeats
SNP	Single nucleotide polymorphism
TSS	Transcription start sites
TE	Transposable element
DOE	U.S. Department of Energy
Vx	Villersexel
WT	Wild type
Wi	Wisconsin

3 Introduction

3.1 Physcomitrium patens

Physcomitrium patens (formerly known as *Physcomitrella patens*) belongs to the Funariaceae, a family of the taxonomic division Bryopsida. The moss *P. patens* was first characterized by (Hedwig and Schwägrichen, 1801). Today, after several phylogenetic rearrangements, *P. patens* is placed in the genus Physcomitrium as propagated by (Mitten, 1851, Rensing *et al.*, 2020). The current full name is *Physcomitrium patens* (Hedwig) Mitten (Medina *et al.*, 2019).

Several requirements need to be fulfilled for an organism to become a model organism (Hedges, 2002). These are, among others, a relatively short lifecycle, uncomplicated in genetic and cultivation purposes. The results should be widely transferable to other organisms. In the early 1960', H.L.K. Whitehouse collected the moss *Physcomitrium* (*Physcomitrella*) *patens* at Gransden Wood (Huntingdonshire, UK). Since then, the descendants of this plant became one of the most important non-seed plant model organisms (Rensing *et al.*, 2020). Plant evolutionary questions like the water to land transition (de Vries and Rensing, 2020) as well as the use of modern omics technologies (Reski *et al.*, 2018) are parts of the wide field of studies *P. patens* is involved.

The *P. patens* lifecycle (Figure 1) starts with a germinating spore that grows long branching protonema tissue. First, it consists of chloronema with a high density of chloroplasts while later caulonemal cells with less chloroplasts emerge (Figure 1 upper part). Small buds grow at caulonema cells and form the gametophore with a stem and phyllids (leaflets) (Rensing *et al.*, 2020). An adult gametophore reaches five millimeters in size. Small rhizoids, as analogous to roots, help to anchor the plant to the substrate and supply water and nutrients. *P. patens* is a monoecious moss and thus develops gametangia, antheridia (male) and archegonia (female) structures, apical of the gametophore. A single adult gametophore accommodates several archegonia and antheridia formed in a single cluster (Hiss *et al.*, 2017) (Figure 1 lower left part). Antheridia produce biflagellated gametes (spermatozoids). In the presence of water, hundreds of spermatozoids are released by the antheridium. They swim to the archegonium to reach the mature egg cell at the bottom of the archegonium (Rensing *et al.*, 2020). The self-fertilization rate of *P. patens* is between 92 % and 97 % (Perroud *et al.*, 2011). A fertilized egg cell forms the diploid (2n) zygote and turns into an embryo which develops into the sporophyte (Figure 1 lower right part). This is the only diploid stage in the whole lifecycle of *P. patens*. A small leftover of the archegonium can be found at the top of the sporophyte, the Calyptra, which remains haploid (1n). Usually, only a single sporophyte develops on the apex of a gametophore. During the process of maturation, the sporophyte turns from green to brown. The brown, mature sporophyte releases haploid spores. A full lifecycle is completed within approximately three months (Schaefer and Zrÿd, 2001).

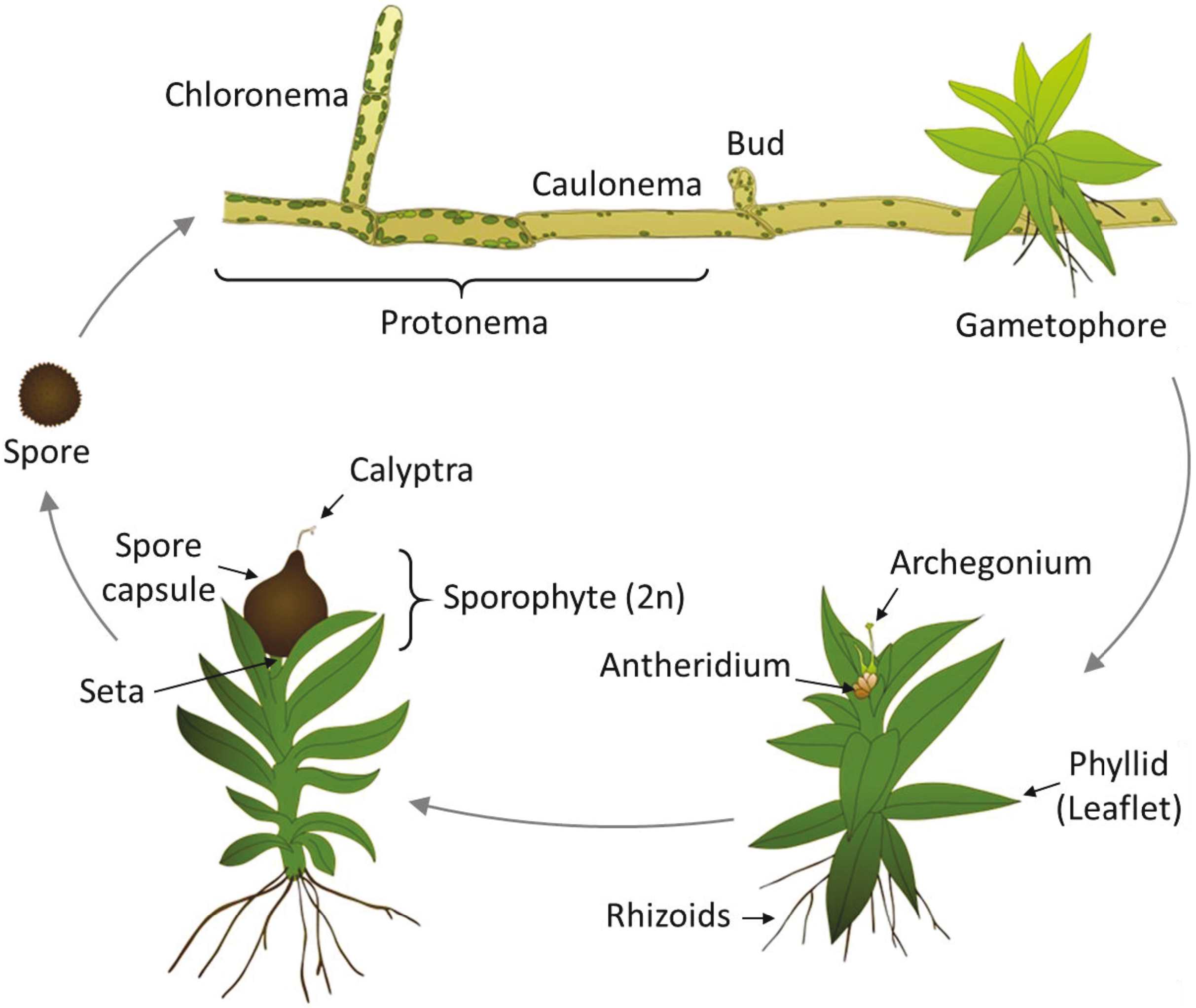


Figure 1: P. patens lifecycle. *A germinating spore grows protonema. Buds grown on protonema develop to the gametophore. Rhizoids anchor the plant. The sexual organs, archegonium and antheridium, are build apical of the gametophore. A fertilized egg cell develops the sporophyte. After maturation, the sporophyte releases new spores. All tissues, except the sporophyte, are haploid (1n). Tissues are not drawn at a similar scale. Modified after (Lang et al., 2018, Rensing et al., 2020).*

3.2 Accessions and Gransden pedigrees

During the last decades, several samplings of different *P. patens* ecotypes (in this thesis called accessions) were reported (von Stackelberg *et al.*, 2006, Kamisugi *et al.*, 2008, McDaniel *et al.*, 2010, Perroud *et al.*, 2011, Beike *et al.*, 2014, Medina *et al.*, 2019) and (https://www.moss-stock-center.org). In this work, five different *P. patens* accessions were used (Table 1 and Figure 2). Two strains were collected in the Northern American region: The isolate Kaskaskia (Perroud *et al.*, 2011) was collected near St. Louis, Illinois, USA and the isolate Wisconsin came from the AUGIE herbarium, Rock Island, Illinois, USA (Haas *et al.*, 2020). Two accessions are from central Europe: The isolate Villersexel

(Kamisugi *et al.*, 2008) was sampled from a site close to Villersexel, France and the isolate Reute (Hiss *et al.*, 2017) from Reute near Freiburg, Germany. The remaining isolate is Gransden (Engel, 1968, Rensing *et al.*, 2008, Lang *et al.*, 2018), collected in Gransden Wood, Huntingdonshire, UK. *P. patens* can be found at temporarily flooded creeks or lake banks and moist open fields (Cove, 2005, Rensing *et al.*, 2020) (Figure 4).

Table 1: Origin of the five P. patens accessions used in this work.

Isolate	**Where**	**Who**	**When**	**Publication**
Gransden	Collected at Gransden Wood, Cambridgeshire, England, UK	H.L.K. Whitehouse	1962	(Engel, 1968)
Villersexel	Collected at Villersexel, Villers la Ville, Haute Saône, France	M. Lüth	2003	(Kamisugi *et al.*, 2008)
Kaskaskia	Collected near the Kaskaskia River, Illinois, USA.	D. Vitt and M. Sargeant	2003	(Perroud *et al.*, 2011)
Reute	Collected near Reute, Freiburg i.B., Germany	M. Lüth	2006	(Hiss *et al.*, 2017)
Wisconsin	Collected in Wisconsin, original specimen in AUGIE herbarium, Rock Island, Illinois, USA	R. Medina	2017	(Haas *et al.*, 2020)

As previously mentioned, H.L.K. Whitehouse collected the most commonly used *P. patens* accession Gransden in 1962. Engel started to cultivate Whitehouse's *P. patens* Gransden sample. He used a single spore to initiate the culture (Engel, 1968). This accession was named Gransden. In 1974 Ashton and Cove (Ashton and Cove, 1977, Cove, 2005) started to distribute Gransden plants globally. It is important to be aware of the fact that all Gransden plants are derived from a single spore isolate. *P. patens* is mainly cultivated and propagated vegetatively. Therefore, the offspring can be considered to be clonal and should be genetically identical. However, genetic variation between the different Gransden laboratory strains (Gd pedigrees) can be observed (Haas *et al.*, 2020). All currently existing Gransden plants are derived from laboratories. New Gransden material cannot be collected, because its habitat at Gransden Wood was destroyed. In this thesis, different Gransden pedigrees around the world are grouped by their origin. Four main Gransden pedigree clusters were defined. These pedigrees are Gransden United Kingdom (Gd_UK), Gransden Germany (Gd_DE), Gransden Switzerland (Gd_CH), and Gransden Japan (Gd_JP) (Figure 2).

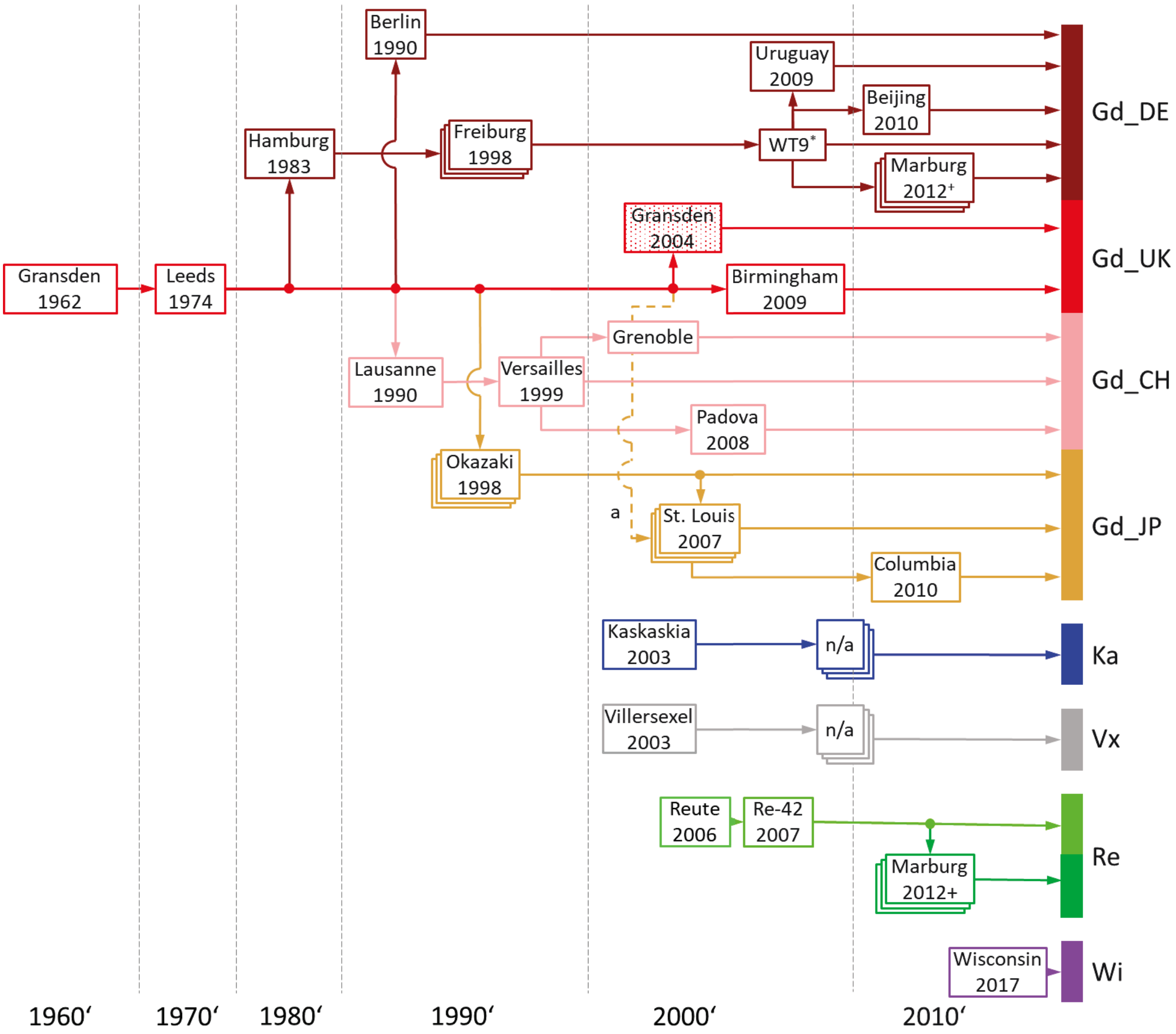


Figure 2: P. patens pedigree. *The pedigree shows five different P. patens accessions sub-clustered by 13 different Gransden and two different Reute laboratory strains (Gd and Re pedigrees). Gransden (reddish) was collected in 1962 and is the oldest accession in laboratories. Kaskaskia (blue) and Villersexel (grey) were collected in 2003. The accession Reute (green) joined the collection in 2006. The youngest accession in this pedigree is Wisconsin (violet), collected in 2017. This pedigree shows only those P. patens accessions that were used in this thesis. More accessions can be found at https://www.moss-stock-center.org. Trackable sexual propagation history is shown by stacked boxes. + Since 2011 yearly selfing, expect 2013. * Since 1999 Gransden Freiburg went through nine generations leading to WT9.*

3.3 *P. Patens* genome and gene model annotation

The US DOE JGI started to sequence the *P. patens* genome after a genome consortium was founded in 2004 (Rensing *et al.*, 2020). Back in 2004, just a few plant genomes were fully sequenced. The phylogenetic position of *P. patens* promoted a high interest in sequencing the full genome, as the first non-seed plant. In 2008 (Rensing *et al.*, 2008) published the first draft of the *P. patens* genome. It contained 480 Mbp, spread over ~ 2,000 scaffold sequences. The first published genome annotation version consisted of 35,938 gene models (Rensing *et al.*, 2008). Subsequently, the genome annotation was continuously improved over the years. Gene model annotation version 1.2 and 1.6 were published on the recently discontinued cosmoss.org webpage. Nevertheless, data documentation is still available (https://www.cosmoss.org/physcome_project/wiki/Main_Page).

Exactly 10 years after the publication of the first genome version, the most recent genome version 3 (V3) was published by (Lang *et al.*, 2018). This genome version represents the assembled 472 Mbp in 27 pseudo-chromosomes and 330 unassigned scaffolds. The associated genome annotation version v3.3 contains 32,458 gene models. This resource is available at the CyVerse comparative genomics platform CoGe (https://genomevolution.org/coge/OrganismView.pl?gid=33928) (Figure 3).

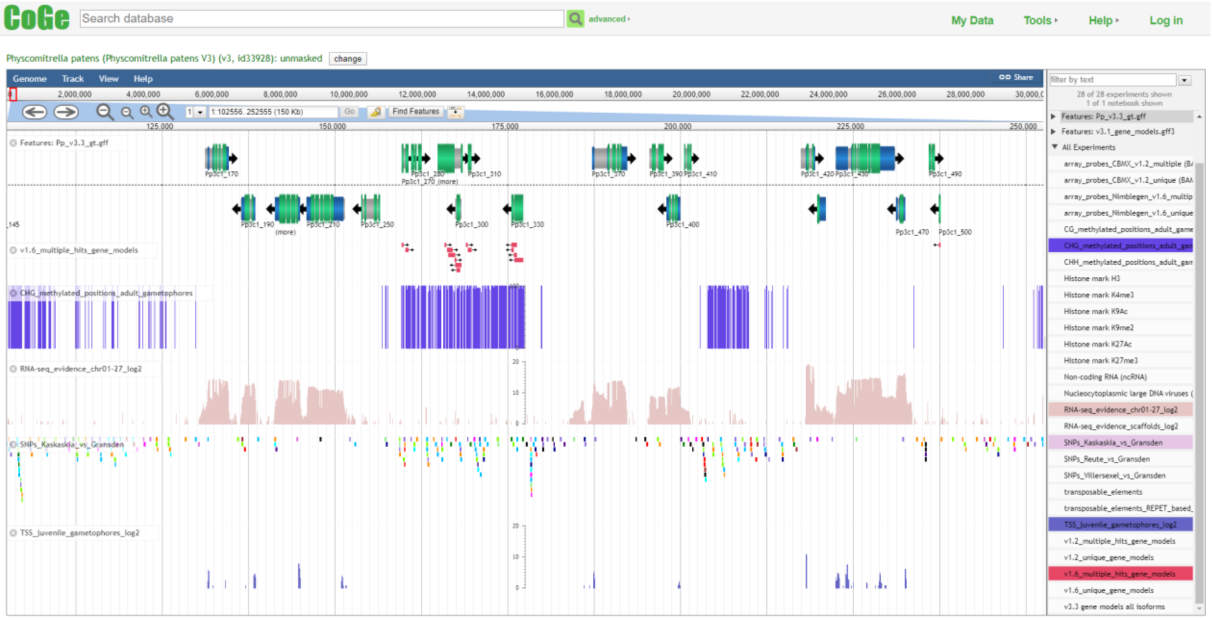

CoGe
Search database
advanced
My Data
Tools
Help
Log in
Physcomitrella patens (Physcomitrella patens V3) (v3, id33928): unmasked
change
Genome
Track
View
Help
Share
2,000,000
4,000,000
6,000,000
8,000,000
10,000,000
12,000,000
14,000,000
16,000,000
18,000,000
20,000,000
22,000,000
24,000,000
26,000,000
28,000,000
30,000,0
1:102556..252555 (150 Kb)
Go
Find Features
125,000
150,000
175,000
200,000
225,000
250,000
Features: Pp_v3.3_gt.gff
Pp3c1_170
Pp3c1_280
Pp3c1_270 (more)
Pp3c1_310
Pp3c1_370
Pp3c1_390
Pp3c1_410
Pp3c1_420
Pp3c1_450
Pp3c1_490
_145
Pp3c1_190
(more)
Pp3c1_210
Pp3c1_250
Pp3c1_300
Pp3c1_330
Pp3c1_400
Pp3c1_470
Pp3c1_500
v1.6_multiple_hits_gene_models
CHG_methylated_positions_adult_gametophores
RNA-seq_evidence_chr01-27_log2
20
10
0
SNPs_Kaskaskia_vs_Gransden
TSS_juvenile_gametophores_log2
20
10
0
filter by text
28 of 28 experiments shown
1 of 1 notebook shown
Features: Pp_v3.3_gt.gff
Features: v3.1_gene_models.gff3
All Experiments
array_probes_CBMX_v1.2_multiple (BA
array_probes_CBMX_v1.2_unique (BAM
array_probes_Nimblegen_v1.6_multip
array_probes_Nimblegen_v1.6_unique
CG_methylated_positions_adult_game
CHG_methylated_positions_adult_gan
CHH_methylated_positions_adult_gan
Histone mark H3
Histone mark K4me3
Histone mark K9Ac
Histone mark K9me2
Histone mark K27Ac
Histone mark K27me3
Non-coding RNA (ncRNA)
Nucleocytoplasmic large DNA viruses (
RNA-seq_evidence_chr01-27_log2
RNA-seq_evidence_scaffolds_log2
SNPs_Kaskaskia_vs_Gransden
SNPs_Reute_vs_Gransden
SNPs_Villersexel_vs_Gransden
transposable_elements
transposable_elements_REPET_based_
TSS_juvenile_gametophores_log2
v1.2_multiple_hits_gene_models
v1.2_unique_gene_models
v1.6_multiple_hits_gene_models
v1.6_unique_gene_models
v3.3 gene models all isoforms

Figure 3: CoGe JBrowse screen shot of P. patens V3 with multiple tracks loaded. *Among other functions, CoGe has a integrated JBrowse (Skinner et al., 2009, Buels et al., 2016) instance. On the most right side, all gene annotations (top) and experiment tracks (below) are shown. Each track can be loaded and will be visible on the main screen. The current shown main screen displays (top to bottom): v3.3 and v1.6 gene models, methylation and expressions data, SNPs and TSS evidence. Each single track can be downloaded. https://genomevolution.org/coge/GenomeView.pl?embed=&gid=33928*

3.4 Sequencing data

A breakthrough in the DNA sequencing technology was the first fully sequenced genome of the bacteriophage phi-X with 5,386 bp (Sanger *et al.*, 1977). Subsequently, the technology for RNA sequencing, more precisely cDNA sequencing (RNA-seq), was developed (St. John and Davis, 1979, Weinstock *et al.*, 1994) that paved the way for wide gene expression studies. The next boost in sequencing technologies occurred when next-generation sequencing (NGS) methods appeared (Bennett, 2004, Margulies *et al.*, 2005, Shendure *et al.*, 2005). Supported by the huge impact of NGS high throughput RNA-seq emerged (Mortazavi *et al.*, 2008, Tang *et al.*, 2009). While several NGS techniques, all with advantages and disadvantages exist, the technology used by Illumina (Bennett, 2004) is the most common sequencing technology. This technology is based on a large number of short DNA fragments that bind to adapters fixed on a surface. Both, the forward and the reverse complementary strands are amplified before sequencing. The sequencing itself is done by imaging a fluorescent signal. This signal is created by fluorescently labelled reversible terminators each time a deoxyribonucleoside triphosphate (dNTP) is added to the target sequence (Bennett, 2004). An output sequence (read) can be up to 300 nt long and a single sequencing run can generate up to one billion reads (https://www.illumina.com/systems/sequencing-platforms.html, visited 2020-04-06). Illumina provides a good reading accuracy and high sequence output. The latest, third-generation sequencing technologies, Oxford Nanopores MinION (Jain *et al.*, 2016) and Pacific Biosciences SMRT (Eid *et al.*, 2009) do not need to amplify DNA and can produce multiple kbp long reads. However, these technologies need to improve their reading accuracy.

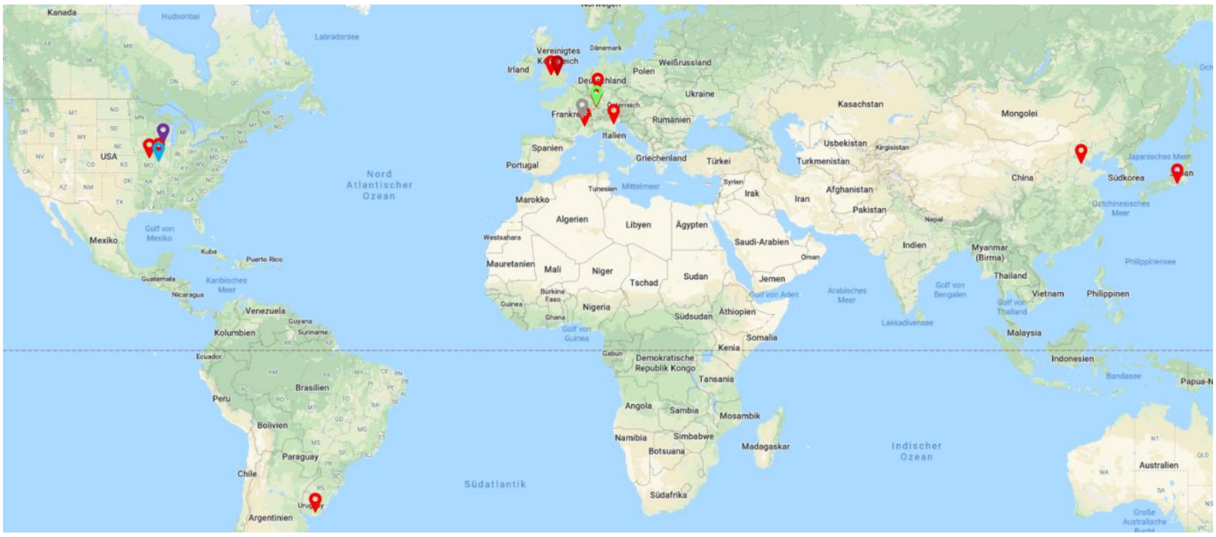


Figure 4: Origin of the P. patens accessions and Gransden pedigrees. The sampling sites of the five different accessions are shown in dark red (Gransden, UK), blue (Kaskaskia, USA), violet (Wisconsin, USA), green (Reute, DE) and grey (Villersexel, FR). Gransden pedigrees originating from laboratories worldwide resulted in RNA-seq samples for this thesis (red indicators) Map based on https://www.google.com/maps.

3.4.1 The DOE JGI Gene Atlas project

In 2010, *P. patens* became a US DOE JGI flagship genome (https://jgi.doe.gov/our-science/science-programs/plant-genomics/plant-flagship-genomes). Subsequently, the JGI funded the DOE JGI Plant Flagship Gene Atlas project (https://jgi.doe.gov/doe-jgi-plant-flagship-gene-atlas/). This project aimed to advance the expression catalog of the JGI Plant Flagship genomes. A special focus on nitrogen metabolism was put forward. The first five species, alga *Chlamydomonas reinhardtii*, soybean *Glycine max*, moss *Physcomitrium patens*, poplar tree *Populus trichocarpa* and foxtail millet *Setaria italica*, were complemented by *Arabidopsis thaliana*, *Medicago truncatula*, *Brachypodium distachyon, Panicum virgatum*, *Panicum halli*, *Setaria viridis* and *Sorghum bicolor*. Across all 12 species, 887 RNA-seq samples were sequenced by the DOE JGI using Illumina technology (https://genome.jgi.doe.gov/portal/JGIFlaPGeneAtlas/JGIFlaPGeneAtlas.info.html). These RNA-seq datasets enable broad inferences of shared gene function across phyla.

3.4.2 *P. patens* expression data

As part of the ambitious JGI Gene Atlas project, 61 *P. patens* experiments were sequenced with RNA-seq, separated by two sequencing rounds (Figure 5). In the first round (1st) in 2014, 13 different laboratories produced 34 tissue/treatment conditions (Chapter 5.2 and Supporting information 9.2, Table S6 and S7, Figure S2) (Perroud *et al.*, 2018). The second round (2nd) in 2016 included three laboratories and 27 experiments (Supporting information 9.3, Table S8) (Fernandez-Pozo *et al.*, 2019). In both sequencing rounds two *P. patens* accessions, Gransden (Figure 4, red marks) and Reute were used. Multiple tissues were covered, *inter alia*, protonemata with young gametophore, green and brown sporophytes, adult gametophores, phyllids (leaflets) and germinating spores (Figure 1). The 1st round had its focus on protonemata tissue and different nitrogen sources. Studies in the second round used more Reute plants and the experimental focus changed to phosphate time courses as well as mutant studies. Due to library preparation issues, 54 samples of the 2nd round were sequenced twice (Supporting information 9.5).

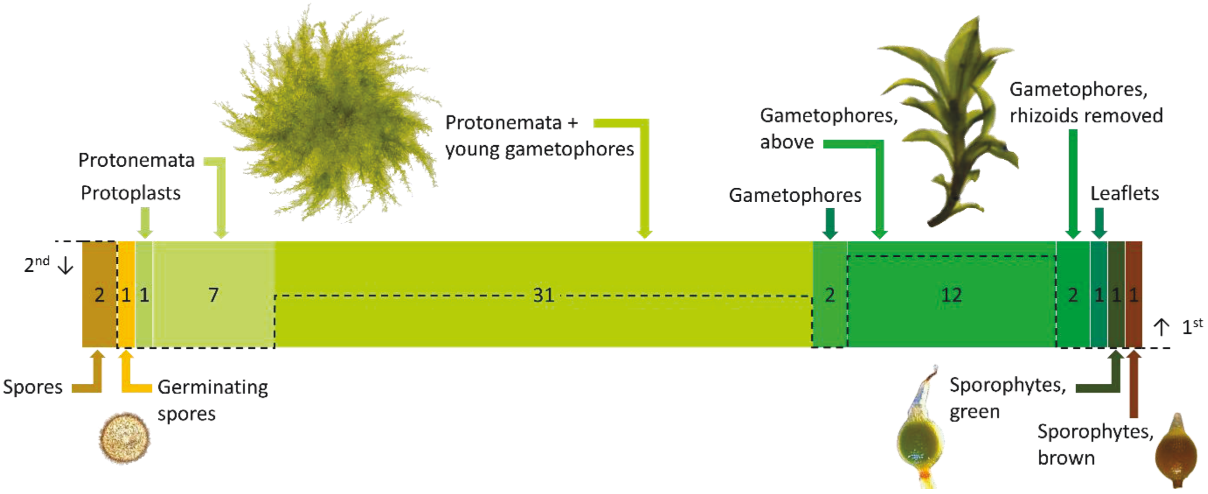


Figure 5: Number of P. patens JGI Gene Atlas experiments and used tissues. *39 protonema tissue experiments, 17 gametophore tissue experiments, three spore tissue experiments, and two sporophyte experiments were done. In the 1st sequencing round, 34 experiments were sequenced, in the 2nd round, 27 experiments (dashed line). Tissues are not shown at a similar scale. P. patens tissue pictures were taken from (Prigge and Bezanilla, 2010, Hiss et al., 2017).*

3.5 Genetic variation

Dissimilarity in the genome sequence of individuals is called genetic variation. Substantial sources of variation is sequence mutation. The emergence of mutations has multiple causes, e.g. environmental influences, mutagenic chemicals or DNA synthesis errors. Mitotic and meiotic processes can force sequence errors, like genetic recombination. Examples for genetic recombination are chromosomal cross overs and shifting transposable elements. For inheritance questions, the affected cell type is important. Only variations in germ cells will influence the offspring, thus genetic variation is the driving force of evolution. However, vegetatively propagated organisms can obtain variations from somatic cells as well.

The number of genetic modifications between *P. patens* accessions varies. A high degree of polymorphism in the genome of Villersexel was observed by using simple sequence repeats (SSRs) (von Stackelberg *et al.*, 2006). The same trend was shown while generating the *P. patens* genetic linkage map (Kamisugi *et al.*, 2008). Genetic distances in Villersexel found by (Beike *et al.*, 2014) were proposed to be evidence for high intrapopulation diversity or long-range dispersal. Based on recognized polymorphisms, (McDaniel *et al.*, 2010) argued that *P. patens* emerged at least three times from various ancestors from the genus Physcomitrium. NGS data was used to detect single nucleotide polymorphisms (SNPs). These SNP studies showed a higher variation of Villersexel, where Gransden was used as reference, compared to other analysed accessions (Ding *et al.*, 2018, Lang *et al.*, 2018, Haas *et al.*, 2020). SNP calling based on RNA-seq data could strengthen the circumstance, that somatic mutation occurs and accumulate in *P. patens* laboratory strains.

3.6 Bioinformatic procedure

Next-generation sequencing, can generate hundreds of millions reads per sequencing run. The sheer size of this amount of data requires predictable and standardized analysis methods. Many of the high-performance software tools for analytical processing of RNA-seq data are designed for one specific task only, e.g. sequence quality control or read mapping. It is challenging to run these programs fast and efficient. Issues that users face can be incompatible software versions, different input formats or unknown calculation parameters. Predefined process structures can help to overcome such issues. As an example: A pipeline is a chain of predefined process elements. The output of a finished process is the input of the follow-up process. Rulesets define the workflow. The same methodology is used for developing computational pipelines. For this thesis, several hundred RNA-seq samples were processed. To manage the analysis, a functional and effective RNA-seq pipeline, adapted on the concept of scientific workflows, was developed. The scheme of the pipeline is shown in Figure 6. The design fulfils specific and important requirements such as flexibility with respect to data input, modular structure, user-friendly interface, easy maintenance, and fast computing, to mention only some advantages.

While focusing on the computing part and the informatics behind it, it is important not to neglect the biological interpretation of the results. The choice between biological or statistical normalization methods and the interpretation of the corresponding values has a major impact on the results (Qin *et al.*, 2013, Evans *et al.*, 2018). Likewise, the origin and composition of the biological samples need to be considered.

The interaction of the newly released V3 genome with the associated v3.3 gene annotation together with the novel designed RNA-seq pipeline for the *P. patens* accessions' variant calling and the JGI Gene Atlas RNA-seq data analyses, in association with expression data published on PEATmoss, is the key of my successful analyses.

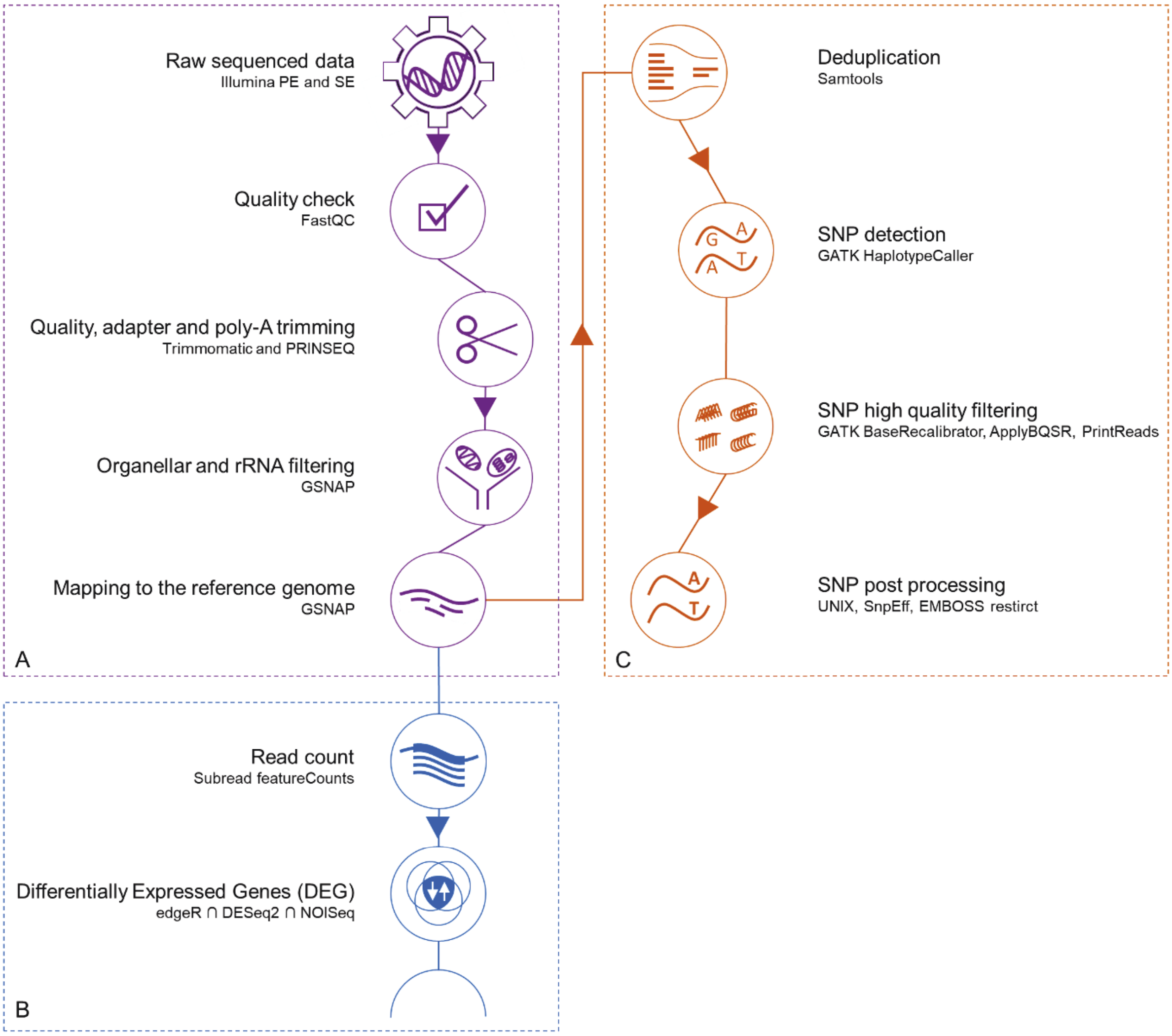

Raw sequenced data
Illumina PE and SE
Quality check
FastQC
Quality, adapter and poly-A trimming
Trimmomatic and PRINSEQ
Organellar and rRNA filtering
GSNAP
Mapping to the reference genome
GSNAP
A
Read count
Subread featureCounts
Differentially Expressed Genes (DEG)
edgeR ∩ DESeq2 ∩ NOISeq
B
Deduplication
Samtools
SNP detection
GATK HaplotypeCaller
G
A
A
T
SNP high quality filtering
GATK BaseRecalibrator, ApplyBQSR, PrintReads
SNP post processing
UNIX, SnpEff, EMBOSS restirct
A
T
C

Figure 6: The RNA-seq pipeline design. *The pipeline is structured in three functional units: A) RNA-seq pre-processing, includes sequence quality control, filtering and reference mapping, B) Gene expression calculation, normalization and DEG calling, and C) Variant detection/SNP calling.*

4 Research objectives

The main focus of this work was to develop resources for the important non-seed plant model *P. patens* to improve genome annotation, expression data, and methods to analyse sequence variation within accessions and pedigrees.

The first step was to develop methods to analyse large numbers of RNA-seq samples. The JGI Gene Atlas project contributed the majority of the analysed data. This project includes the first extensive high throughput RNA-seq dataset of *P. patens*. Our novel universal RNA-seq pipeline uses RNA-seq and gDNA data form different species. The pipeline results were tested by comparison to the commercial CLC workbench tool (Lanver *et al.*, 2018) and by the usage of data from different organisms. Pipeline output data like gene expression evidence and DEGs were used to accomplish gene expression profiles.

The *P. patens* genome V3 and its annotation were upgraded by using transcriptomic data and publicly available sequence databases. Further genome version upgrades are in progress.

Our approach to share results with the community and establishing a good data availability was to upload genome and annotation data to CoGe and expression data to the new developed web tool PEATmoss.

Subsequently, pipeline extensions were developed. The functionality was expended by adding SNP calling and transcriptome assembly branches. Dozens of already publicly obtainable RNA-seq samples were collected for variant detection. This data was combied with new RNA-seq samples of *P. patens* accessions Gransden, Kaskaskia, Reute and Villersexel plus the JGI Gene Atlas RNA-seq samples. Somatic mutation, as well as natural variation, was observed by analysing *P. patens* Wisconsin gDNA samples. Single nucleotide polymorphisms based on transcriptomic data were detected and relations between five different *P. patens* accessions and 13 different Gransden pedigrees were demonstrated.

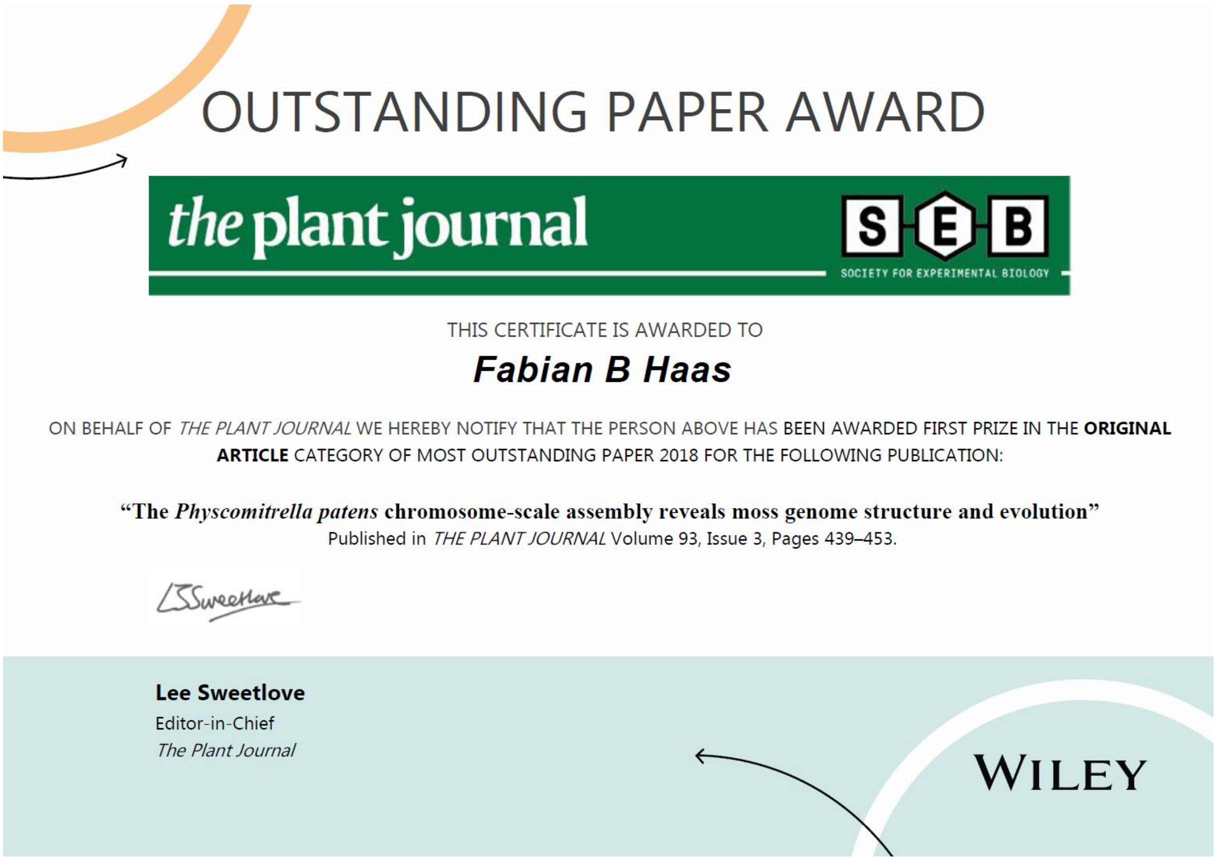
5 Publications

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

Exactly 10 years after the first *P. patens* genome version (V1) was published (Rensing *et al.*, 2008), the new version 3 (V3) was released (Lang *et al.*, 2018). The new genome assembly has reached a pseudo-chromosomal status. Its 472 Mbp are split into 27 pseudo-chromosomes and 330 unassigned scaffolds (Paper 5.1, page 515f). Besides the genome assembly, new chloroplast and mitochondrial genomes were presented (Supporting information, 9.1.1). Before the final genome assembly version was released, the full assembly went into a contamination check (Supporting information, 9.1.2). Since V1, it was known, that *Bacillus subsp.* occurred in the *P. patens* sequencing libraries and later assigned in the genome assembly. The first contamination removal of these bacterial fragments was done for the V1.1 release (https://www.cosmoss.org/physcome_project/wiki/Contaminations). Additionally, the genome annotation and the corresponding gene models were improved. Expression datasets and different gene prediction methods were evaluated and merged into final gene annotation versions 3.1 (v3.1) and 3.3 (v3.3) (Supporting information, 9.1.3) (Paper 5.1, page 526).

Several fundamental analyses, like TE studies, that detected an unusual distribution of LTRs and protein-coding genes (Paper 5.1, page 517f), DNA methylation analyses show gene body methylation (Paper 5.1, page 519-522), and SNP detection to identify variation between three different accessions (Paper 5.1, page 522f), were done. In total, 21 experimental tracks were generated and uploaded to CoGe (https://genomevolution.org/coge/GenomeInfo.pl?gid=33928) (Supporting information, 9.1.4).

This publication introduces the new *P. patens* V3 genome. This is the reference for all my work in this thesis.



OUTSTANDING PAPER AWARD

the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

THIS CERTIFICATE IS AWARDED TO

Fabian B Haas

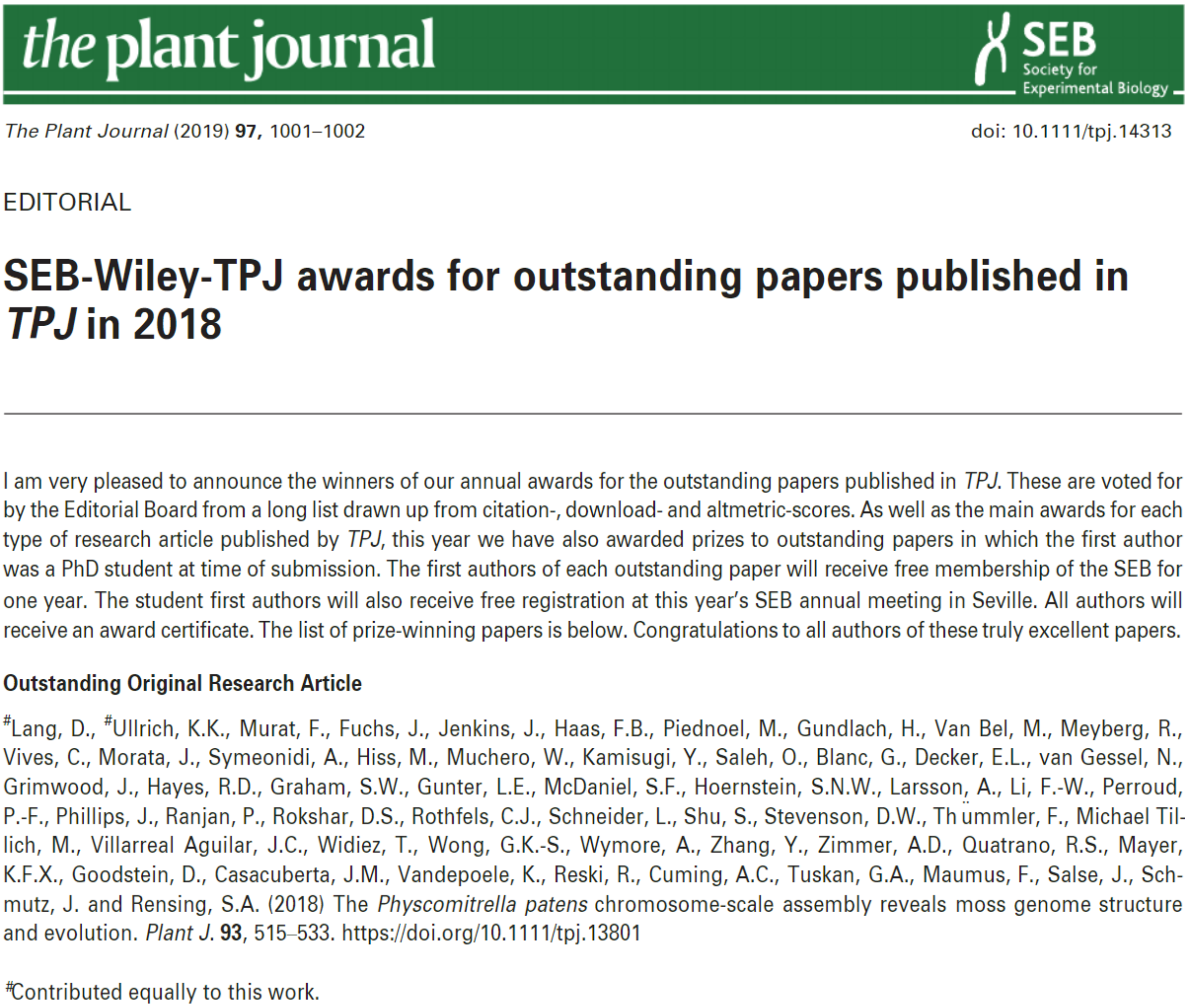
ON BEHALF OF *THE PLANT JOURNAL* WE HEREBY NOTIFY THAT THE PERSON ABOVE HAS BEEN AWARDED FIRST PRIZE IN THE **ORIGINAL ARTICLE** CATEGORY OF MOST OUTSTANDING PAPER 2018 FOR THE FOLLOWING PUBLICATION:

"The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution"
Published in *THE PLANT JOURNAL* Volume 93, Issue 3, Pages 439–453.

LJSweetlove

Lee Sweetlove
Editor-in-Chief
The Plant Journal

WILEY



the plant journal SEB Society for Experimental Biology

The Plant Journal (2019) **97,** 1001–1002 doi: 10.1111/tpj.14313

EDITORIAL

SEB-Wiley-TPJ awards for outstanding papers published in *TPJ* in 2018

I am very pleased to announce the winners of our annual awards for the outstanding papers published in *TPJ*. These are voted for by the Editorial Board from a long list drawn up from citation-, download- and altmetric-scores. As well as the main awards for each type of research article published by *TPJ*, this year we have also awarded prizes to outstanding papers in which the first author was a PhD student at time of submission. The first authors of each outstanding paper will receive free membership of the SEB for one year. The student first authors will also receive free registration at this year's SEB annual meeting in Seville. All authors will receive an award certificate. The list of prize-winning papers is below. Congratulations to all authors of these truly excellent papers.

Outstanding Original Research Article

#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.F., 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., Th ̈ummler, F., Michael Tillich, M., Villarreal Aguilar, J.C., Widiez, T., Wong, G.K.-S., 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, G.A., Maumus, F., Salse, J., Schmutz, J. and Rensing, S.A. (2018) The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution. *Plant J.* **93**, 515–533. https://doi.org/10.1111/tpj.13801

#Contributed equally to this work.

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).

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

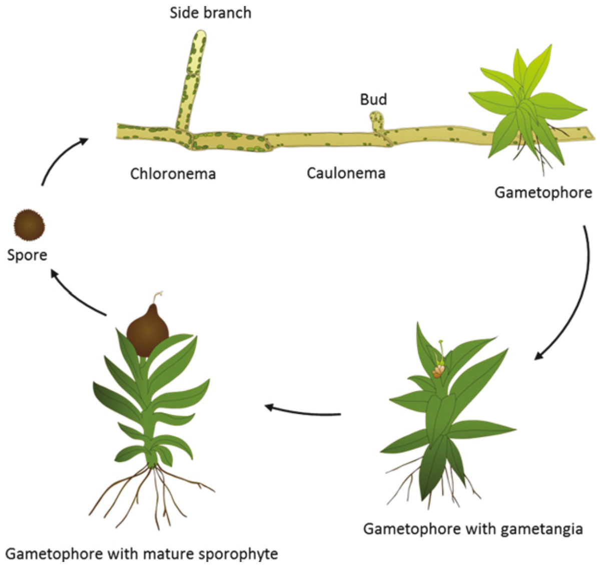

Side branch
Bud
Chloronema
Caulonema
Gametophore
Spore
Gametophore with mature sporophyte
Gametophore with gametangia

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.

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

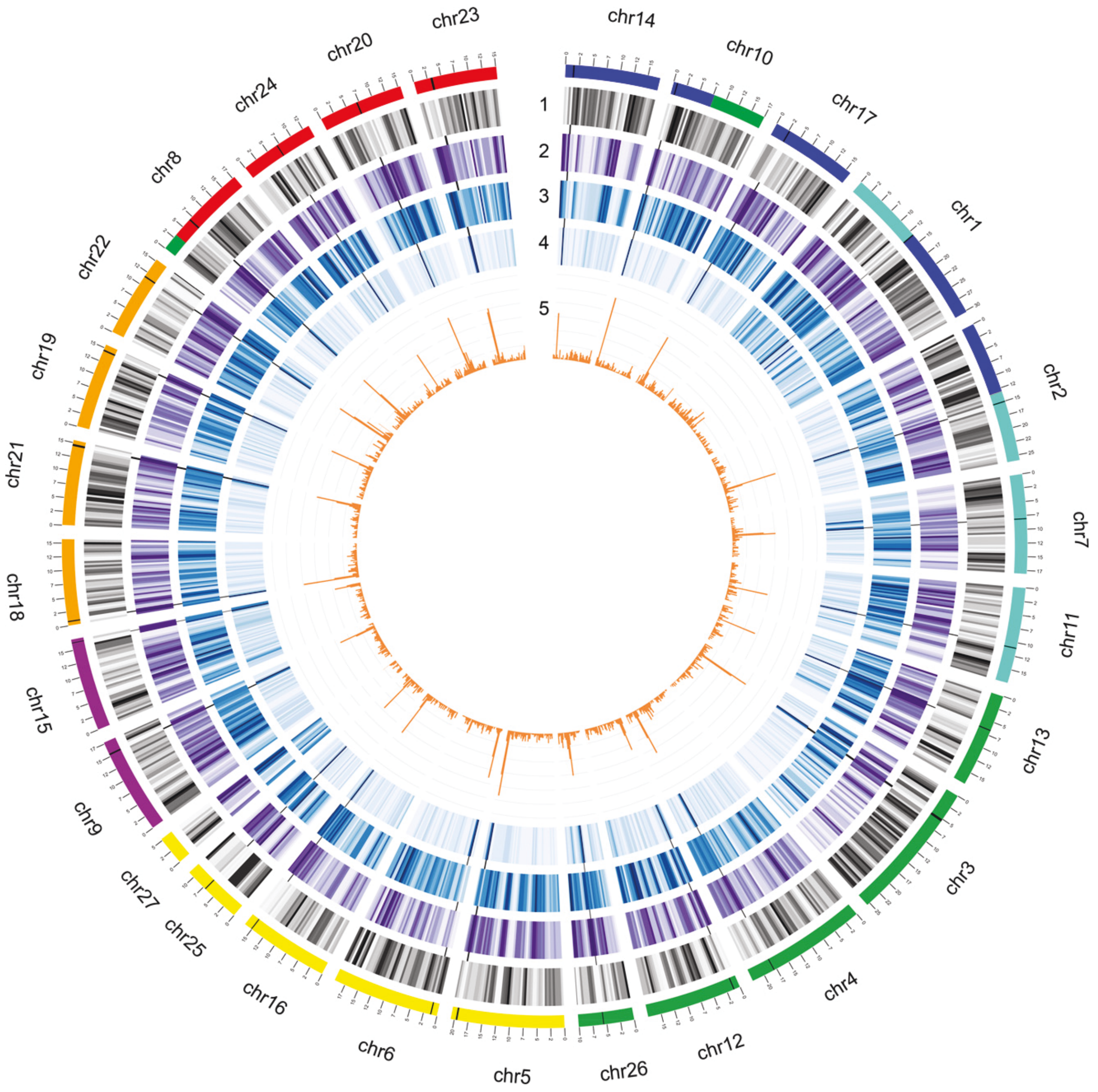


Figure 2. Chromosome structure, focus on TEs.
From outer to inner: karyotype bands colored according to ancestral genome blocks as in Figure 5 (scale = Mbp), followed by: (1) gene density (grey, normalized 0,1); (2) repeat density (violet, normalized 0,1); (3) gypsy-type elements (blue, normalized 0,1); (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

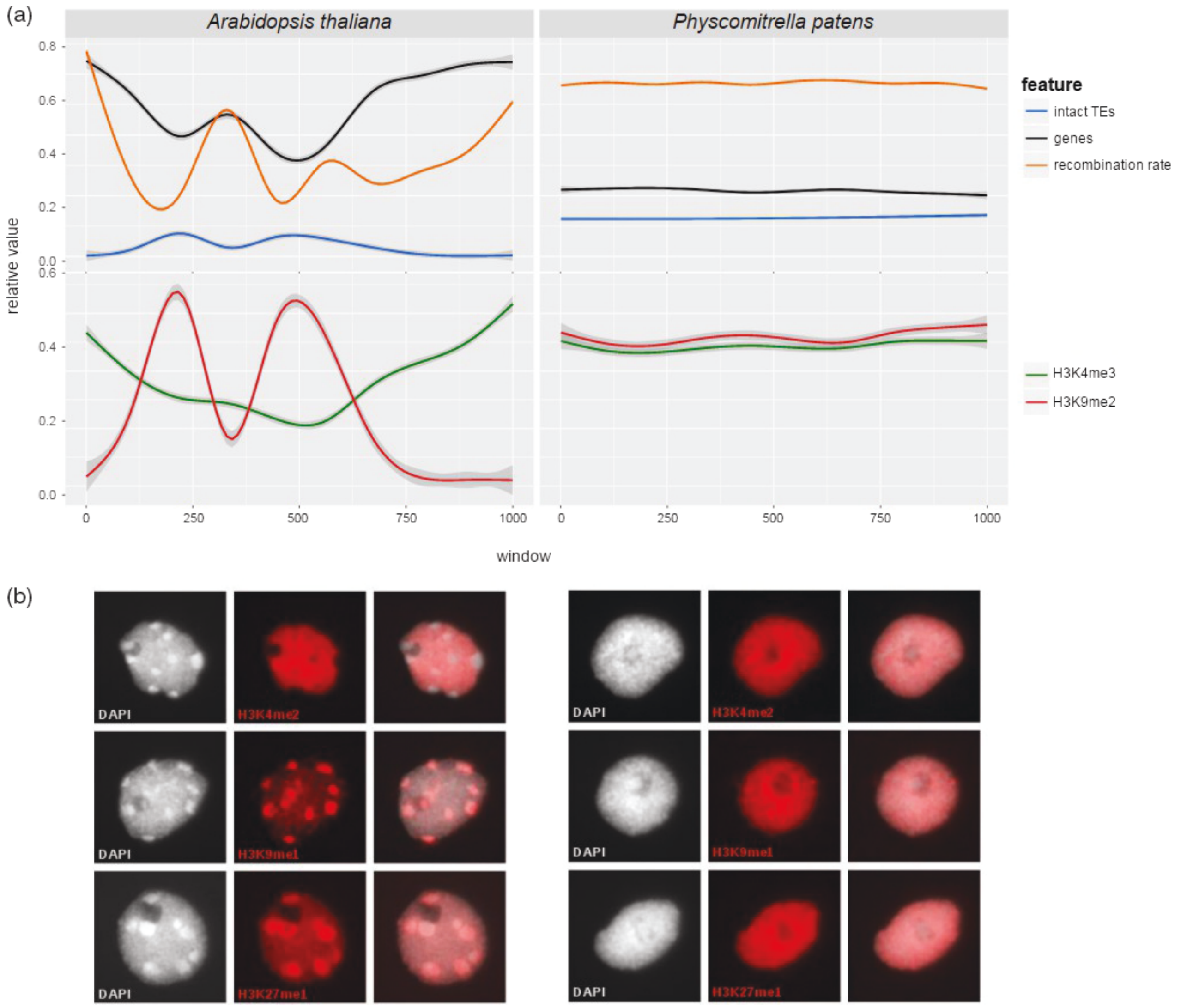


Figure 3. Comparative analysis of genome structures.

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).

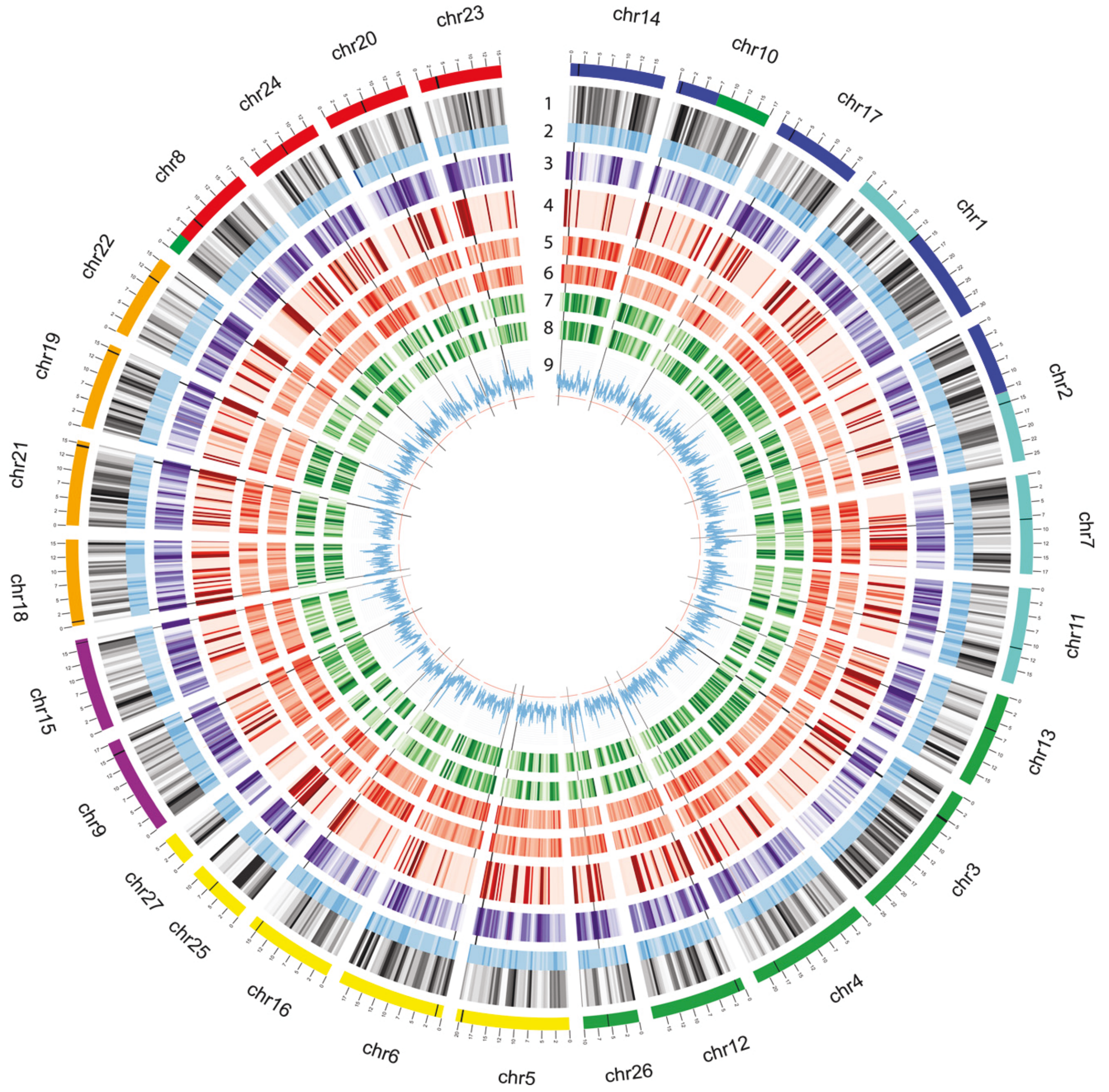


Figure 4. Chromosome structure, focus on epigenetic marks.
From outer to inner: karyotype bands colored according to ancestral genome blocks as in Figure 5, followed by: (1) gene density (grey) normalized 0,1; (2) GC content 0.25–0.45 (blue); (3) all TEs density (violet) normalized 0,1, NCLDV evidence 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 karytope (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

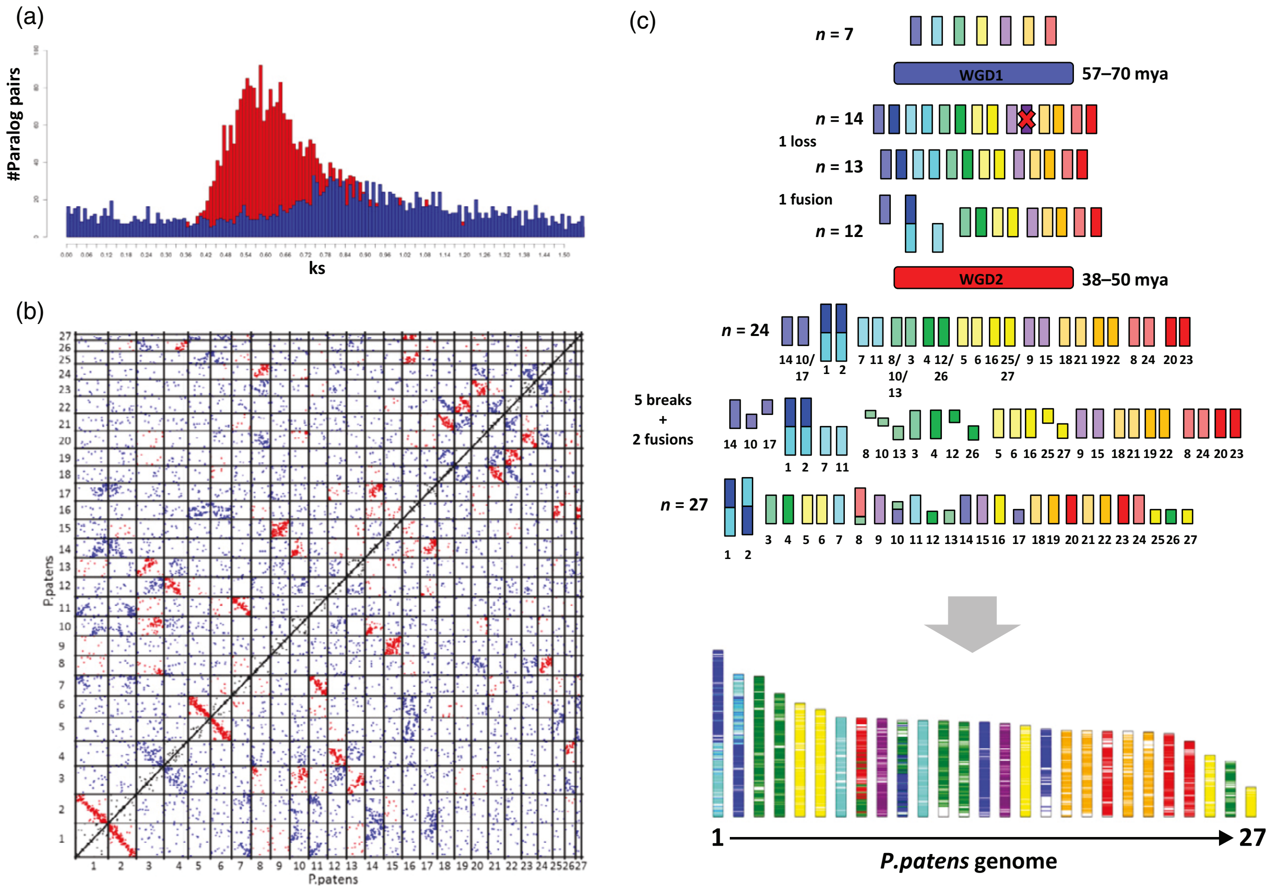


Figure 5. Evolutionary scenario leading to the modern *P. patens* genome.
(a) Ks distribution (*y*-axis) of paralogous pairs (*x*-axis) inherited from two (blue for older and red for more recent) WGD events.
(b) Dotplot representation of the paralogous pairs belonging to two WGD events.
(c) Karyotype 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 extrinsinc 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 $\mathrm{sec}^{-1}\ \mathrm{m}^{-2}$, for 6 weeks at 22°C/19°C day/night temperature following 8 h/16 h light/dark cycle, 20 μmol $\mathrm{sec}^{-1}\ \mathrm{m}^{-2}$, 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 paleohexaploidy in the stem lineage of core eudicots: significance for MADS-box gene and species diversification. *Mol. Biol. Evol.* **29**, 3793–3806.

Vives, C., Charlot, F., Mhiri, C., Contreras, B., Daniel, J., Epert, A., Voytas, D.F., Grandbastien, M.A., Nogue, F. and Casacuberta, J.M. (2016) Highly efficient gene tagging in the bryophyte *Physcomitrella patens* using the tobacco (*Nicotiana tabacum*) Tnt1 retrotransposon. *New Phytol.* **212**, 759–769.

Wang, G., Zhang, X. and Jin, W. (2009) An overview of plant centromeres. *J. Genet. Genomics* **36**, 529–537.

Widiez, T., Symeonidi, A., Luo, C., Lam, E., Lawton, M. and Rensing, S.A. (2014) The chromatin landscape of the moss *Physcomitrella patens* and its dynamics during development and drought stress. *Plant J.* **79**, 67–81.

Wright, S.I., Agrawal, N. and Bureau, T.E. (2003) Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. *Genome Res.* **13**, 1897–1903.

Wu, T.D., Reeder, J., Lawrence, M., Becker, G. and Brauer, M.J. (2016) GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Methods Mol. Biol.* **1418**, 283–334.

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

Zhang, Y., Liu, T., Meyer, C.A. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137.

Zilberman, D. (2017) An evolutionary case for functional gene body methylation in plants and animals. *Genome Biol.* **18**, 87.

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

5.2 The *Physcomitrella patens* gene atlas project: large-scale RNA-seq based expression data

The first comprehensive RNA-seq dataset this size for *P. patens* was generated in the JGI Gene Atlas project which resulted in a total of 99 RNA-seq samples derived from 34 experiments (Supporting information, 9.2.1 and 9.2.2). To increase the productivity of the RNA-seq analysis (time and assurance) and to standardize the methods, a highly efficient RNA-seq pipeline was developed (Figure 6 A, B) (Paper 5.2, page 170). The pipeline is flexible and can be applied to and used for many different organisms (Lanver *et al.*, 2018). The huge number of RNA-seq samples and good sequencing quality opened the opportunity to detect over 28.500 v3.3 gene models. Never before has such a high number of v3.3 gene models been validated at once. Furthermore, this high number of detected gene models allowed for comparative genomics. Calling differentially expressed genes (DEGs) was one main step for the pipeline (Figure 6 B). Three DEG topics were focused: Stage-specific transcriptome (Paper 5.2, page 171f), the impact of ammonium supplementation on the protonemal liquid culture transcriptome (Paper 5.2, page 173), and intra- and inter-laboratory comparison (Paper 5.2, page 174f). Besides the pipeline and the DEG calling, other sections were reviewed. The currently published v3.3 gene models (Lang *et al.*, 2018) were annotated and completed by their expression values. This table contains an early draft of the gene version lookup table (described in more details in the next publication) (https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Ftpj.13940&file=tpj13940-sup-0003-DataS1a.xlsx). The last presented section was the comparative analysis of RNA-seq and microarray-based DEGs. It was shown that the method established here can be used to compare RNA-seq and microarray-based DEGs (Supporting information, 9.2.5) (Paper 5.2, page 179).

Initial methods and tools to analyse the JGI Gene Atlas *P. patens* expression datasets were published in this paper. These methods are the basic scaffold of all my further RNA-seq analyses.

the plant journal

SEB Society for Experimental Biology

The Plant Journal (2018) **95**, 168–182 doi: 10.1111/tpj.13940

RESOURCE

The *Physcomitrella patens* gene atlas project: large-scale RNA-seq based expression data

Pierre-François Perroud[1,*], Fabian B. Haas[1], Manuel Hiss[1], Kristian K. Ullrich[1,†], Alessandro Alboresi[2,‡], Mojgan Amirebrahimi[3], Kerrie Barry[3], Roberto Bassi[2], Sandrine Bonhomme[4], Haodong Chen[5], Juliet C. Coates[6], Tomomichi Fujita[7], Anouchka Guyon-Debast[4], Daniel Lang[8], Junyan Lin[3], Anna Lipzen[3], Fabien Nogué[4], Melvin J. Oliver[9], Inés Ponce de León[10], Ralph S. Quatrano[11], Catherine Rameau[4], Bernd Reiss[12], Ralf Reski[13,17], Mariana Ricca[14], Younousse Saidi[6,§], Ning Sun[5], Péter Szövényi[14], Avinash Sreedasyam[15], Jane Grimwood[15], Gary Stacey[16], Jeremy Schmutz[3,15] and Stefan A. Rensing[1,17,*]

[1] *Plant Cell Biology, Faculty of Biology, University of Marburg, Karl-von-Frisch-Str. 8, 35043, Marburg, Germany,*
[2] *Dipartimento di Biotecnologie, Università di Verona, Cà Vignal 1, Strada Le Grazie 15, 37134, Verona, Italy,*
[3] *US Department of Energy (DOE) Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, 94598, USA,*
[4] *Institut Jean-Pierre Bourgin, INRA, AgroParisTech, CNRS, Université Paris-Saclay, Route de St-Cyr RD10, 78026, Versailles Cedex, France,*
[5] *School of Advanced Agriculture Sciences and School of Life Sciences, Peking University, Beijing, China,*
[6] *School of Biosciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK,*
[7] *Department of Biological Sciences, Faculty of Science, Hokkaido University, Kita 10 Nishi 8, Kita-ku, Sapporo 060-0810, Japan,*
[8] *Helmholtz Zentrum München, Ingolstädter Landstr. 1, 85764, Neuherberg, Germany,*
[9] *USDA-ARS-MWA, Plant Genetics Research Unit, University of Missouri, Columbia, MO, 652117, USA,*
[10] *Department of Molecular Biology, Clemente Estable Biological Research Institute, Avenida Italia 3318, CP 11600, Montevideo, Uruguay,*
[11] *Department of Biology, Washington University in St Louis, One Brookings Drive, St Louis, MO, 63130, USA,*
[12] *Max Planck Institute for Plant Breeding Research, Carl-von-Linne-Weg 10, 50829, Köln, Germany,*
[13] *Plant Biotechnology, Faculty of Biology, University of Freiburg, Schänzlestr. 1, 79104, Freiburg, Germany,*
[14] *Department of Systematic and Evolutionary Botany, University of Zurich, Zollikerstr. 107, 8008 Zürich, Switzerland,*
[15] *HudsonAlpha Institute for Biotechnology, 601 Genome Way Northwest, Huntsville, AL, 35806, USA,*
[16] *Divisions of Plant Science and Biochemistry, National Center for Soybean Biotechnology, University of Missouri, Columbia, MO, 65211, USA,*
[17] *BIOSS Centre for Biological Signalling Studies, University of Freiburg, Schänzlestr. 18, 79104, Freiburg, Germany,*
[†] *Present address: Max Planck Institute for Evolutionary Biology, August-Thienemann-Straße 2, 24306, Ploen, Germany,*
[‡] *Present address: Dipartimento di Biologia, Università di Padova, Viale Giuseppe Colombo, 3, 35131, Padova, Italy, and*
[§] *Present address: Bayer Crop Science, Technologiepark, Zwijnaarde 38, 9052, Gent, Belgium*

Received 29 November 2017; revised 2 April 2018; accepted 5 April 2018; published online 22 April 2018.
*For correspondence (e-mail pierre-francois.perroud@biologie.uni-marburg.de; stefan.rensing@biologie.uni-marburg.de).

SUMMARY

High-throughput RNA sequencing (RNA-seq) has recently become the method of choice to define and analyze transcriptomes. For the model moss *Physcomitrella patens*, although this method has been used to help analyze specific perturbations, no overall reference dataset has yet been established. In the framework of the Gene Atlas project, the Joint Genome Institute selected *P. patens* as a flagship genome, opening the way to generate the first comprehensive transcriptome dataset for this moss. The first round of sequencing described here is composed of 99 independent libraries spanning 34 different developmental stages and conditions. Upon dataset quality control and processing through read mapping, 28 509 of the 34 361 v3.3 gene models (83%) were detected to be expressed across the samples. Differentially expressed genes (DEGs) were calculated across the dataset to permit perturbation comparisons between conditions. The analysis of the three most distinct and abundant *P. patens* growth stages – protonema, gametophore and

sporophyte – allowed us to define both general transcriptional patterns and stage-specific transcripts. As an example of variation of physico-chemical growth conditions, we detail here the impact of ammonium supplementation under standard growth conditions on the protonemal transcriptome. Finally, the cooperative nature of this project allowed us to analyze inter-laboratory variation, as 13 different laboratories around the world provided samples. We compare differences in the replication of experiments in a single laboratory and between different laboratories.

Keywords: developmental stage, differential expression, *Physcomitrella patens*, RNA-seq, stress, transcriptome analysis.

INTRODUCTION

Since the discovery of its intrinsically efficient gene targeting (Schaefer and Zrÿd, 1997), followed by its genome sequencing (Rensing *et al.*, 2008; Lang *et al.*, 2018), the moss *Physcomitrella patens* has become the leading reference non-seed plant model. It is now notably integrated in studies focused on the water-to-land plant transition (e.g. Renault *et al.*, 2017) or the establishment of tri-dimensional growth in plants (for a review, see Harrison, 2017), with both fields of study integrating detailed investigation of functional cell biology with kingdom-wide gene and genome comparison. As these multiple research fields grew, so did their associated technical approaches. Amongst them, RNA-seq (the deep sequencing of cDNA) is now dominating the field of RNA detection and quantification at the transcriptome level, replacing probe-based microarray technology. Detailed knowledge of the transcriptome of a given organism is being used to improve genome assemblies (Song *et al.*, 2016), to better understand and describe RNA splicing pattern (Gaidatzis *et al.*, 2015) and to characterize spatiotemporal transcriptome variation (expression profiling), both with respect to development and environmental perturbations.

In the green lineage, RNA-seq approaches to assess transcriptome-wide patterns were initially used in models such as *Arabidopsis thaliana* (Lister *et al.*, 2008) or *Chlamydomonas reinhardtii* (González-Ballester *et al.*, 2010), followed by major crop plants such as *maize* (maize; Eveland *et al.*, 2010) and *Oryza sativa* (rice; Zhang *et al.*, 2010), for which a complete genome sequence was available. Furthermore, the improvement of the *de novo* assembly of transcriptomes allowed the use of this approach to characterize transcriptomes in organisms without available genome sequences (e.g. *Cassia angustifolia*; Rama Reddy *et al.*, 2015), gaining knowledge of both the raw sequence information and about biological processes in species previously limited by the lack of a genome. The most wide-spread *de novo* transcriptome assembly effort so far is the 1KP project, covering more than 1300 species of the green algae and the land plant lineage with low RNA-seq sequencing coverage (Matasci *et al.*, 2014).

In non-seed plants, RNA-seq based transcriptomic studies have been reported in multiple species, both in parallel with genome sequencing projects and via *de novo* analysis. Besides *P. patens* (see below), transcriptomes of other mosses have been published: *Sphagnum* spp. (Devos *et al.*, 2016), *Bryum argenteum* (Gao *et al.*, 2014), *Ceratodon purpureus* (Szövényi *et al.*, 2015), *Funaria hygrometrica* (Szövényi *et al.*, 2011) or *Syntrichia caninervis* (Gao *et al.*, 2015). Published datasets are also available for liverworts, including *Pellia endiviifolia* (Alaba *et al.*, 2015) and *Marchantia polymorpha* (Sharma *et al.*, 2014), the genome of which has recently been published (Bowman *et al.*, 2017). The transcriptome of the lycophyte *Selaginella moellendorffii*, for which a draft genome is available (Banks *et al.*, 2011), has also been subjected to extensive RNA-seq study (Zhu *et al.*, 2017). With large genomes, ferns form a group with exclusively *de novo* transcriptome datasets so far: e.g. *Acrostichum* spp., *Ceratopteris thalictroides* (Zhang *et al.*, 2016), *Ceratopteris richardii* (Bushart *et al.*, 2013) and *Lygopodium japonicum* (Aya *et al.*, 2015).

In *P. patens*, RNA-seq datasets have been released in multiple experimental contexts, unfortunately with no systematic multiple experimental replications (for a review, see Hiss *et al.*, 2017). For example, the *P. patens* transcriptome profile has been studied with respect to developmental stage (Xiao *et al.*, 2011), in addition to stress treatments including bleomycin (Kamisugi *et al.*, 2016). Analysis of heat-stress impact on alternative splicing has also benefited from the RNA-seq approach (Chang *et al.*, 2014). Recently, transcriptomic responses to plant hormone treatments with abscisic acid (Stevenson *et al.*, 2016) and auxin (Lavy *et al.*, 2016) have been studied. Additionally, comparative transcriptomic approaches have been applied to mutant analysis (Chen *et al.*, 2012; Demko *et al.*, 2014), and to analyze and catalogue small RNAs in *P. patens* (e.g. Coruh *et al.*, 2015; Lang *et al.*, 2018).

The present study describes the first part of the *P. patens* dataset of the US Department of Energy (DOE) Joint Genome Institute (JGI) Gene Atlas Project (http://jgi.doe.gov/doe-jgi-plant-flagship-gene-atlas/). After reviewing the dataset we focus on experimental comparisons, underscoring different aspects of such large-scale projects. In terms of tissue-specific expression profiling, we show the possibility of defining specific transcripts for the three dominant life stages of *P. patens*: protonema, gametophore and sporophyte. We also tackle two aspects of transcriptome comparison experiments. We evaluate the impact of nitrogen supplementation in single-laboratory settings and show here the power of such an approach. Moreover, the diversity of the sample sources permitted us to compare two experimental replica sets of the same growing conditions performed by two different laboratories to evaluate inter-laboratory replication.

RESULTS AND DISCUSSION

Overview of the dataset

The *P. patens* Gene Atlas dataset comprises 99 sequenced libraries of 34 different experiments. All but three experiments are composed of three biological replicates. For experiments XIV, XVIII and XXII, one of the libraries failed for technical reasons, hence they are formed of biological duplicates. Thirteen laboratories actively working with *P. patens* around the world contributed to the samples described in Experimental procedures. The detailed description of all samples and primary sequencing statistics are presented in Tables S1 and S2. The sampling covers the three dominant *P. patens* stages, protonema (the gametophytic two-dimensional filamentous stage emerging from the spore), gametophore (the gametophytic tridimensional leafy shoot stage) and sporophyte (the sporophytic tissue developing after sexual reproduction that forms spores by meiosis). It must be noted that the age of the protonema at harvest varies from 7 to 21 days. As gametophore buds typically start to emerge after 7 days of growth, most of the protonemal samples are a mixture of protonemal cells and gametophore cells (for detailed harvesting times for each experiment, see Table S2). With this time criteria, the samples VII, XI, XVIII, XIX and XXI–XXIV are the only samples that are potentially pure protonema. The sequencing output (raw sequenced reads) was analyzed *in silico* using the standardized procedure schematized in Figure 1. Overall, 4.2 billion raw reads were generated, with each condition represented by 76–150 million raw reads. A total of 99.02% of the reads were mapped successfully to the genomes of *P. patens* (nuclear, chloroplast and mitochondrial). Furthermore, 90.04% of the reads mapped uniquely to the *P. patens* nuclear genome V3, and were used for further data analyses. After mapping, 22 610–26 012 out of 34 361 gene models of the *P. patens* v3.3 genome annotation, i.e. 65.8–75.7% of the gene models, are observed with more than one read. All conditions considered, more than 80% of all predicted gene models are detected with more than one read. Subsequently, normalized counts (reads per kilobase of transcript per million mapped reads, RPKM) were calculated for each individual gene model (for the full RPKM dataset, see Data S1).

A principal component analysis (PCA) performed with the RPKM normalized counts of all libraries (Figure 2) allows the detection of three major sample clusters. The largest cluster (circled in red) is formed by the protonema and gametophore samples, regardless of the perturbation (except ABA/drought). The second distinct cluster comprises the six sporophytic samples (circled in green). Finally, both ABA treatment and dehydrated/rehydrated gametophore samples form a third cluster (circled in blue), probably linked to water stress and its

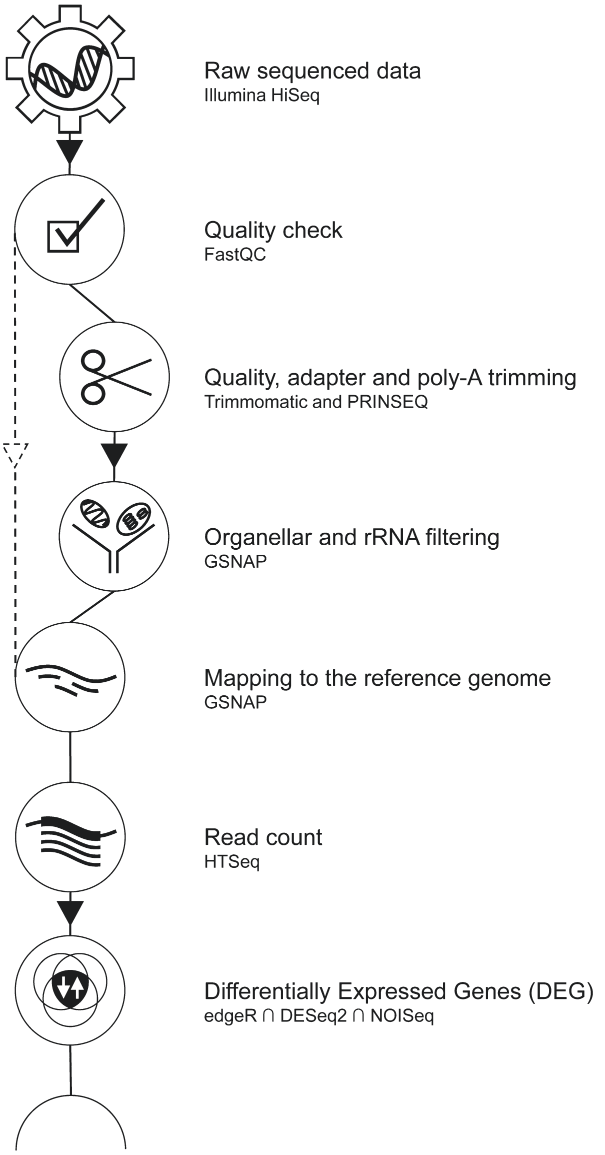

Raw sequenced data
Illumina HiSeq
Quality check
FastQC
Quality, adapter and poly-A trimming
Trimmomatic and PRINSEQ
Organellar and rRNA filtering
GSNAP
Mapping to the reference genome
GSNAP
Read count
HTSeq
Differentially Expressed Genes (DEG)
edgeR ∩ DESeq2 ∩ NOISeq

Figure 1. RNA-seq data analysis.
Diagram illustrating the sequential RNA-seq data treatment from raw read to differentially express genes (DEGs). For details, see Experimental procedures.

hormonal signal integrator, ABA. Biological replicates should be tightly grouped, and for most replicates this is the case (for example, see the triplicate of experiment XIX in Figure 2, indicated by dotted ellipse a). Yet, note that several triplicates are more scattered than expected (for example, see the triplicate of experiment XVI in Figure 2, inidicated by dotted ellipse b), potentially making the comparison between experiments challenging, particularly within some of the protonemal treatments (red ellipse in Figure 2). Finally, to complement and confirm the expected experimental sample clustering, we performed a hierarchical clustering of all 99 RNA-seq samples (Figure S1). Here, 95% of the replicas grouped properly. The exception are restricted to two groups of closely related samples (V and VIII; XII and XIII) that form two clusters of six libraries, but do not group by experiment. Also, the clustering of experiment XI is scattered, suggesting a potential problem with these samples.

The last computing step of our pipeline (Figure 1) was the detection of differentially expressed genes (DEGs) between experiments. The DEGs were called using a strict consensus approach of three callers (for computational details, see Experimental procedures). Overall, 50 relevant experiment comparisons were generated (for a general overview, see Table S3). The complete list of detected DEGs is shown in Data S1, next to the individual RPKM library counts. The highest number of DEGs were detected in experiments associated with very strong perturbations, such as: gametophore compared with dehydrated gametophore, with 9305 DEGs (experiment XIII compared with XVII); protoplast compared with protonema, with 7746 DEGs (experiment VIII compared with IX); or protonema compared with ABA-treated protonema, with 6940 DEGs (experiment XIX compared with XVIIII). At the other extreme, a few comparisons displayed a very limited number of DEGs. Of note, the comparison between dehydrated and rehydrated gametophore showed only 10 DEGs (experiment XIII compared with XII). The treatment itself is not lethal and the gametophores begin to grow again after the treatment; however, the 2 h of rehydration prior to harvesting is probably too short a time to generate significant transcriptional changes. More puzzling is the detection of a single DEG between the tissue treated with the strigolactone analog G24 and its solvent control (experiment V compared with XXXVIII). This 24-h treatment has been shown to affect transcript accumulation and tissue morphology in *P. patens* (Hoffmann *et al.*, 2014; Decker *et al.*, 2017), as well as in angiosperms (in *A. thaliana*; Mashiguchi *et al.*, 2009) and in *Solanum lycopersicum* (tomato; Mayzlish-Gati *et al.*, 2010), for example, indicating that the assay may not have worked properly for this specific treatment. On the other hand, the actual *P. patens* regulatory network under strigolactone influence appears reduced in size compared with those of other hormones (Waldie *et al.*, 2014), and the detection of specific transcript accumulation variation upon strigolactone treatment is dependent on light conditions during growth as well as on the endogenous level of strigolactone (Lopez-Obando *et al.*, 2016). In this context it is possible that 3-week-old tissue could be insensitive to strigolactone treatment. Compared with most of the other comparisons with higher numbers of DEGs, these two cases with almost no detected DEGs show that near-perfect replication can be achieved with such comparative experiments.

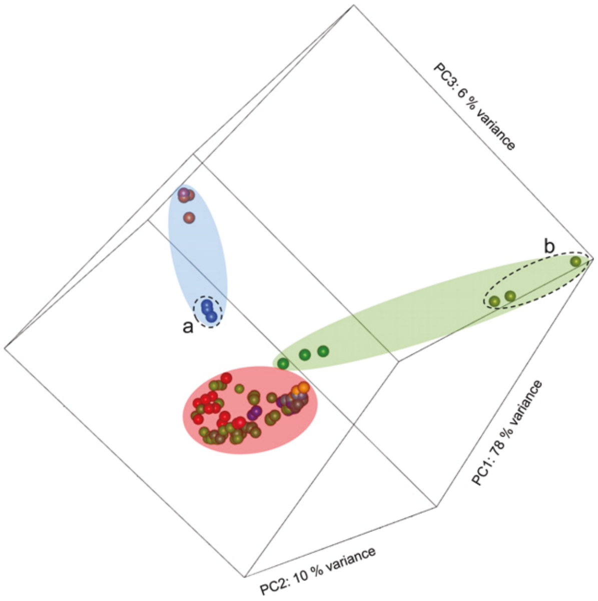


Figure 2. Principal component analysis (PCA) of reads per kilobase of transcript per million mapped reads (RPKM) values for the 99 libraries of this study.

Each dot represents one library. Dots for each experiment have the same color. The red ellipse highlights most of the gametophore experiments. The green ellipse highlights the sporophyte experiments. The blue ellipse indicates strong stress experiments. Dotted ellipse a highlights an experiment with tightly grouped triplicate results. Dotted ellipse b highlighte an experiment with more loosely grouped triplicate results.

Stage-specific transcriptome

Protonema, gametophore and sporophyte are the three dominant life stages of *P. patens*. We choose experiments VII, XX and XV (Tables S1 and S2) as representative of these tissues based on three criteria. First, all cultures were performed on Knop medium. Second, the timing was strictly controlled, particularly the harvesting time for protonema, which was at 7 days to ensure an absence of early gametophore development. Finally, the protonemata were visually checked for the absence of gametophores. To gain an overview of the differences between the three most abundant *P. patens* tissues, we performed a gene ontology (GO) term enrichment analysis on the pairwise up- or downregulated DEGs between these tissues (Data S1 for

DEG; Data S2 for GO term list). The most obvious detectable signals are differences in metabolism, as illustrated in the word cloud in Figure S2. Foremost, the reduction in photosynthetic activity in the sporophyte compared with both protonema (Figure S2b) and gametophore (Figure S2d) is easily observable. The two most abundantly enriched GO terms among DEGs of lower abundance in sporophytic tissue, compared with both gametophytic tissues, are identical: photosynthesis and photosynthesis light reaction. Together with other terms directly linked to photosynthesis, such as photosystem assembly, the generation of precursor metabolites and energy, or plastid organization, they dominate the term list associated with downregulated transcripts in sporophytic tissue. The protonemata–sporophyte comparison complements and validates the previously observed pattern between gametophore and sporophyte in *P. patens* (O'Donoghue *et al*., 2013), and in another moss, *Funaria hygrometrica* (Szövényi *et al*., 2011). This trend is in line with the known nutritional dependency of the sporophyte on the gametophore.

The GO term analysis also detects the sporophyte-specific upregulation of a carbon consumption-related pathway, which has been described previously in *P. patens* (O'Donoghue *et al*., 2013). Compared with protonemata and gametophore, carbohydrate metabolism is the most over-represented term in the upregulated transcripts of the sporophyte; however, the type of carbon use appears to differ between the two comparisons. Terms associated with fatty acid (metabolism, biosynthesis or general lipid metabolism) characterize the difference between sporophyte and protonema (Figure S2c), whereas terms associated with coumarin (biosynthesis and metabolism) are abundant in the over-accumulated transcripts in sporophyte as compared with gametophore (Figure S2d). Furthermore, the term coumarin covers the biosynthesis of the phenylpropanoids, a large group of secondary metabolites with protective functions such as lignin precursors or sporopollenin (Colpitts *et al*., 2011; Daku *et al*., 2016; Niklas *et al*., 2017; de Vries *et al*., 2017), all of which are enriched in or specific to sporophytes.

The dominant GO term enrichment between protonemata and gametophore is linked to carbon fixation and use. Carbon metabolism is the most frequent GO term associated with the upregulated transcripts in protonema (Figure S2e). The nature of this 7-day-old tissue, young cells dividing and expanding constantly, requires a carbon conversion to cell wall compounds (e.g. the GO term external encapsulating structure organization) that is not present in most of the more mature cells of gametophores. At the same time, GO terms associated with lipid, amino acid and nucleic acid biosynthesis are linked to upregulated protonemal DEGs, all indicating actively growing tissue. Overall, a similar signal was detected previously (Xiao *et al*., 2011) between 3-day-old and 14-day-old protonemal culture. In contrast, GO terms associated with photosynthesis dominate the list of low-protonemal/high-gametophore abundance transcripts (Figure S2f). These GO terms reflect the fact that in contrast to protonema, the leafy gametophore is a mature structure dedicated to photosynthesis as a principal function. Photosynthates are not only used to maintain the viability of the tissue, but will also be used to feed the sporophyte during development (Hiss *et al*., 2014; Regmi *et al*., 2017).

From the GO term enrichment analysis to the single transcript level, the challenge to define stage-specific transcripts resides in the fact that even a transcript that is highly abundant in a given stage, for example the sporophyte, and absent in others, e.g. protonema or gametophore grown in standard growth conditions, can be induced by a variation of the growth conditions. For example, the transcript Pp3c7_6750V3.1, which encodes a Ferritin-like domain-containing protein, displays a very high accumulation in sporophyte (RPKM > 1500), and is below detection level in protonema and gametophore under standard growing conditions; however, this transcript can be induced in protonema treated with ABA or in dehydrated gametophores to even higher levels than in the sporophyte (>3000 RPKM in both cases). Hence, we used two criteria to define stage-specific transcripts: (i) presence in the given tissue (RPKM > 2) and absence in the other two tissues (RPKM = 0–2); and (ii) absence in all other samples across the dataset that do not contain the specific tissue. Figure 3 displays the six transcripts selected to represent protonema, gametophore and sporophyte stages (two each) using the present dataset. Pp3c2_4100V3.1 encodes one copy (out of 25 in *P. patens*) of the ribulose-bisphosphate carboxylase small chain (rbcS) protein that appears to be specifically expressed in protonemal tissue. The protonemal cell wall is essentially formed of primary cell wall that provides reduced protection to light, and the specific high expression of photosynthesis proteins with high turnover may be a way to cope with this higher light exposure. The other protonema-specific selected transcript, Pp3c1_10720V3.1, encodes a protein without known annotation and only detected in bryophytes by the Phytozome gene ancestry list. Its specific presence in protonema probably explains the lack of data for it, but makes it a good marker for such tissue. Pp3c26_11940V3.1, specifically expressed in gametophores, encodes a putative SF7 – FASCICLIN-LIKE ARABINOGALACTAN PROTEIN 11, a cell-wall component in a group that appears to be specific to bryophytes too. Pp3c26_11940V3.1 is one of the four homologs coding for such proteins showing similar accumulation patterns, but the only one indicating a complete specificity to gametophore tissue. Pp3c7_9490V3.1 encodes a carbonic anhydrase:dioscorin precursor protein, and accumulates specifically in gametophores. Finally, transcripts specific to sporophyte tissue are more abundant, with more than 150 transcripts found in that tissue only. General morphology (capsule and seta are the only enclosed multilayer tissues in moss), the unique presence of meiosis and the generation of secondary metabolites, such as oleosin and sporopollenin (Daku *et al*., 2016; Hiss *et al*., 2017), may explain this observation. Pp3c6_15559V3.1, the first sprorophyte-specific transcript chosen, reflects a potential metabolic need (the transport of carbohydrate across the sporophyte), as it encodes for a member of the Nodulin-like protein family (Denancé *et al*., 2014). These integral proteins are known to transport carbohydrates such as sucrose across membranes, and thus allow the optimal allocation of reserve between cells. The second transcript selected for sporophyte identity is Pp3c5_26440V3.1, which encodes for the MKN1-3 protein, a class-II knotted1-like homeobox transcription factor (Champagne and Ashton, 2001). This gene has been extensively studied in *P. patens* (Singer and Ashton, 2007; Sakakibara *et al*., 2008) and is involved in sporophyte patterning, a developmental network specific to this organ. These six tissue markers were part of previously conducted microarray experiments and were analyzed from a tissue-specific perspective (Hiss *et al*., 2014, 2017; Ortiz-Ramirez *et al*., 2016). Although they all were confirmed as expressed in the respective tissue, the tissue specificity does not match perfectly with the present dataset, except for the sporophyte-specific genes (Table S4). Notably, non-tissue-specific expression was detected for the four gametophytic markers (in sporophytic tissue). The differences in both tissue preparations and technologies (in particular the higher sensitivity of RNA-seq) may be the cause of these varied expression patterns.

Impact of ammonium supplementation on the protonemal liquid culture transcriptome

Comparison between *P. patens* liquid cultures grown under near-identical conditions except for the source of nitrogen (experiment XXIII, with Knop medium supplemented with 5 mM ammonium tartrate and 4.2 mM nitrate, compared with experiment XXIV, with standard Knop medium containing only 4.2 mM nitrate) yielded 357 DEGs with a greater than twofold change, with 289 DEGs downregulated and 68 DEGs upregulated by ammonium supplementation (Data S1). The GO term enrichment analysis performed on each subset concurs with the well-known plant response to ammonium supplementation (Data S3). The addition of ammonium in the medium is clearly reflected by the repression of genes involved in nitrate assimilation and metabolism, as it is generating an accumulation of transcripts related to primary carbon metabolism (see Figure 4a and b for the 15 most abundant GO terms present in the down- and upregulated DEGs induced by ammonium supplementation; for the complete set of enriched GO terms, see Figure S3a and b).

More specifically, the effect of the addition of NH_4 to the medium corresponds with previous studies: the gene expression associated with nitrate cell import, nitrate primary metabolism and some associated genes is strongly reduced, in some cases to the absence of detectable transcripts. Tsujimoto *et al.* (2007) analyzed nitrate transporter transcript accumulation under different nitrogen sources in *P. patens*, and their results are recapitulated in the present dataset: *NRT2* and the *Nar2* nitrate transporter family members show strong downregulation upon treatment with ammonium (Figure 4c; Tsujimoto *et al*., 2007). Plant nitrate to ammonium conversion is an energetically costly process, hence upon ammonium supplementation both necessary enzymes are transcriptionally repressed in all plants analyzed (Hachiya and Sakakibara, 2017). This pattern is also detected in the present dataset where both nitrate reductase coding genes (Pp3c10_9670V3.1, Pp3c10_9540V3.1, and Pp3c14_9410V3.1) and nitrite reductase coding genes (Pp3c27_6610V3.1 and Pp3c16_15880V3.1) are strongly inhibited by ammonium supplementation (Figure S4a).

On the other hand, we also observe the loss of transcript abundance for genes involved in ammonium assimilation: both ammonium transport genes, *AMT2*s (Pp3c18_18460V3.1, Pp3s397_40V3.1, Pp3c16_12080V3.1 and Pp3c18_18460V3.1) and primary ammonium assimilation genes, glutamate synthase (Pp3c14_8740V3.1), glutamine synthetase (Pp3c18_10780V3.1) and asparagine synthetase (Pp3c20_17620V3.1) display a reduction of transcript abundance upon treatment with ammonium (Figure S4a). This reduction may be the result of the ammonium concentration used in the experimental setting (5 mM ammonium tartrate), a concentration high enough that it may require the overall regulatory repression of ammonium metabolism.

Indirect effects of different trophic conditions are also detected in this dataset. The two genes most induced by the ammonium treatment (>30-fold), Pp3c20_19940V3.1 and Pp3c20_1770V3.1, belong to transporter gene families involved in salt and metabolite homeostasis. Pp3c20_19940V3.1 encodes a gene coding for an Na^+ P-type ATPase protein, demonstrated to be necessary for active Na^+ cell export in *P. patens* (Lunde *et al*., 2007). The repression of nitrate import under ammonium supplementation affects K^+ import (Coskun *et al*., 2015), and hence the cytoplasmic Na^+/K^+ ratio may be adjusted as a result of this specific transcript increase. Pp3c20_1770V3.1 encodes for a member of the nodulin family. The solute specificity is not well established for all members of this family (Denancé *et al*., 2014), but specific homologs of nodulin genes in angiosperm are notably involved in amino acid transport (Ladwig *et al*., 2012). Thus, an increase in Pp3c20_1770V3.1 abundance hints at amino acid relocation upon treatment with ammonium.

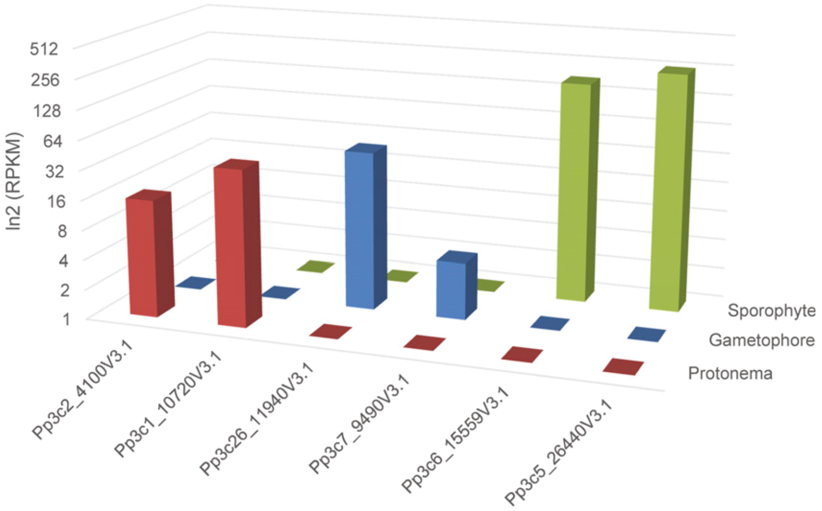


Figure 3. Stage-specific transcript markers.
Two gene models for protonema, gametophore and sporophyte tissue were selected for their specific presence at the stage and for their absence in all experiments not containing exclusively the specific tissue. Reads per kilobase of transcript per million mapped reads (RPKM) values from experiments VII (protonema), XX (gametophore) and XV (sporophyte).

In parallel with nitrate and ammonium-related processes, a cluster of genes associated with cell wall modification appears downregulated under ammonium supplementation. This repression may reflect the morphological change observed upon the addition of ammonium: in the presence of nitrate only, *P. patens* tip cells rapidly differentiate into caulonemal cells, the faster elongating and longer protonemal cell type (Figure S3d). In the presence of ammonium, the tip cells mostly comprise chloronemal cells of shorter size (Figure S3c). Indeed, we observed a reduction of transcript accumulation of known cell wall loosening genes, such as xyloglucan endotransglucosylase hydrolase (Pp3c16_20960V3.1), pectin methylesterases (Pp3c3_30560V3.1, Pp3c3_35240V3.1 and Pp3c4_22420V3.1) and extensins (Pp3c16_3130V3.1 and Pp3c27_3570V3.1) (Figure S3b), all involved in cell elongation and the modification of cell shape (Lamport *et al.*, 2011; Cosgrove, 2016).

The addition of ammonium also promotes transcript over-accumulation. Although this list is much shorter, we can detect a trend to primary carbon metabolism genes such as Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBISCO) small subunit (Pp3c15_22730V3.1, Pp3c12_7010V3.1 and Pp3c3_12530V3.1) or carbonate dehydratase (Pp3c26_6810V3.1 and Pp3c26_6990V3.1) (Figure S3c). Carbon and nitrogen metabolism are closely linked (Coskun *et al.*, 2016). The increase of primary carbon metabolism-associated transcripts in *P. patens* is similar to what is observed in angiosperms. This expression bias is also reflected in the difference in appearance of the *P. patens* tissue between the two conditions. The chloronemal cells are dominant under the presence of ammonium, in which smaller cells are filled with numerous chloroplasts, whereas in nitrate-only conditions the tissue is dominated by caulonemal cells displaying a reduced number of chloroplasts (Figure S3d and c, respectively).

Intra- and inter-laboratory comparison

The present dataset allows us to compare identical experiments performed within a single laboratory and between two different laboratories. Experiments VII, XXI and XXII (Tables S1 and S2) are true replicates. Unfortunately, experiment XXI only has two replicates, as mentioned previously. Hence the comparison of results must be regarded as indicative of a trend only. Each of the two laboratories involved generated protonemal liquid culture in Knop medium, and sampling was performed at day 7 after tissue homogenization. The RNA was subsequently extracted by the laboratory and sent to the JGI for uniform library construction and sequencing. Hence, the two main sources of variation are the cultures themselves and the RNA extraction. Samples from experiments VII and XXI were isolated using the same standardized protocol (see Experimental procedures); in experiment XXII the Trizol step was omitted. RNA samples passed rigorous quality control (carried out by a single lab) prior to library construction. Post-sequencing quality control indicates no major differences in read length, GC content and the total number of sequenced reads (for the read length profile and number for each libraries, see Figure S5), indicating that differences detected should be mostly attributed to laboratory (culture) variation.

The within-laboratory comparison, experiment XXII compared with XXI, displays a limited number of DEGs: 42 with a fold change of >0–2, equally distributed between up- (23) and downregulated (19) DEGs (Data S1). This low number could be attributed at least partially to the absence of a biological triplicate for experiment XXI or to the omission of Trizol in experiment XXII. Nevertheless, the amplitude of the variation remains in a single order of magnitude, contrary to most of the other comparisons in which the calculated fold change can span up to five

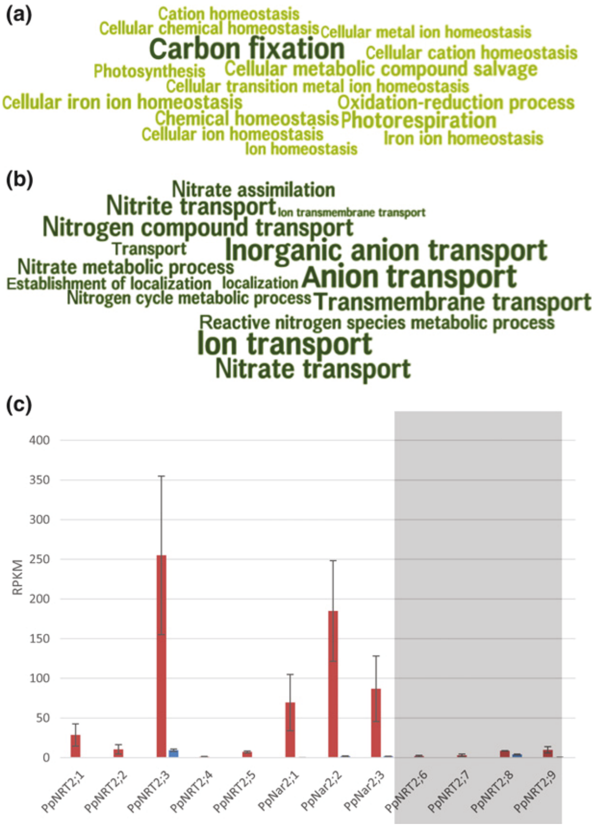


Figure 4. Impact of ammonium supplementation on transcriptome.
Gene ontology (GO) term analysis representation of the 15 most over-represented GO terms in the up- (a) and downregulated (B) differentially expressed genes (DEGs) in ammonium-supplemented liquid protonemal culture (experiment XXIII), compared with ammonium-free liquid protonemal culture (experiment XXIV). The size of the words is proportional to the –log10 (q value), and over-represented GO terms were colored dark green if –log10 (q value) ≤ 4 and were colored light green if –log10 (q value) > 4. For the GO term identities and their respective over-representation values, see Data S3. (c) RPKM values for nitrate and nitrite transporter gene models in the absence of ammonium (red bar) and in the presence of ammonium (blue bar). Gene models not identified by Tsujimoto *et al.* (2007) are shaded in gray. Identifiers of the different nitrate transporter genes described by the *P. patens* v3.3 genome from Tsujimoto *et al.* (2007) read as follows: PpNRT2;1/Pp3c22_21990V3.1, PpNRT2;2/Pp3c22_21970V3.1, PpNRT2;3/Pp3c7_13340V3.1, PpNRT2;4/Pp3c19_10950V3.1, PpNRT2;5/Pp3c22_9060V3.1, PpNar2;1/Pp3c21_13230V3.1, PpNar2;2/Pp3c18_3270V3.1, PpNar2;3/Pp3c22_21950V3.1, PpNRT2;6/Pp3c22_5710V3.1, PpNRT2;7/Pp3c19_10820V3.1, PpNRT2;8/Pp3c19_21550V3.1 and PpNRT2;9/Pp3c16_10420V3.1.

orders of magnitude. The GO term enrichment analysis (Data S4; Figure S6) performed on these DEGs point to a potential source of trophic variation: experiment XXII displays an increase of GO terms associated with a response to external nitrogen processes (nitrate transport and response to nitrate) as well as to cell death and protein recycling (regulation of cell death, positive regulation of cell death or regulation of cellular processes). On the other hand, iron import-related terms dominate the downregulated DEGs. Together, these terms point to a potential nitrogen source depletion leading to metabolic and metal homeostasis redirection. As both iron and nitrate are added separately in the medium, a slight variation in the media could potentially explain these DEGs; however, the low number of observed DEGs and their low fold change indicates high reproducibility, given the sensitivity of RNA-seq to detect minor changes in transcript abundance.

In contrast, 1262 DEGs were detected in total (727 up- and 535 downregulated) with a fold change of >2 between experiments VII and XXII (Data S1). The number and the amplitude of the DEGs (up to 500) suggest a clear difference between these samples. Two major sources of variation could generate such a difference: contamination with other *P. patens* tissue and variation in growing conditions, generating a stress response. To assess these two possibilities, we compared the DEG list between the two experiments and the DEGs for protonemata compared with gametophore (experiment VII compared with XX), to test the tissue hypothesis. We also compared them with three different stress conditions: the effect of ABA on protonemata (experiment XIX compared with XVIII); the effect of high light on protonemata plus gametophores (experiment II compared with I); and the effect of elevated temperature (heat stress) on protonemata plus gametophores (experiment XXV compared with XXVI). Figure 5 illustrates that the observed differences can indeed come from different sources: 1101 of 1262 DEGs detected in the inter-laboratory comparison can be seen in other experiments. Focusing on the comparison with developmental stage only, 680 (54%) of these DEGs are also seen in the DEGs identified by comparing protonema against gametophore. A clear example of the presence of gametophore are transcripts for Pp3c27_3570V3.1, a gene coding for a putative extensin precursor that displays a 250-fold increase between protonemal and gametophore tissues, and shows a more than 64-fold difference between experiment XXI and VII. The source of such gametophore contamination could be explained in differences of weekly grinding of the tissue to maintain a pure protonemata culture. Continuously cultured protonema in liquid culture sometimes develop gametophore buds after 7 days, and thus the culture needs to be blended regularly to reset the protonema to day 1 of the culture cycle. Yet, the difference between experiments VII and XXII cannot be attributed to this kind of tissue contamination alone. The three stresses compared in the same figure, ABA treatment, heat and high light, also display DEG overlap. Each stress displays a specific signature as well as overlap with other experiments, but we can identify 32 DEGs between the three stress conditions and the laboratory comparison that should reflect a general stress response. It is difficult to evaluate exactly the cause of these stresses, but differences between laboratories such as temperature and humidity regime of the growth chamber and type/age of the white light system used can potentially generate the stresses detected in this comparison. It should be noted that in a within-laboratory comparison it was previously shown that liquid culture (with regular blending) as such does not seem to represent a stress condition for *P. patens* (Hiss *et al.*, 2014). The relatively large number of DEGs detected in the inter-laboratory comparison thus demonstrates the sensitivity of RNA-seq and hence the fact that minor fluctuations in growth conditions can result in clearly detectable changes to the transcriptome.

Conclusions and outlook

The present transcriptome dataset represents an important addition to the existing expression profiling data for the moss *P. patens*. By its sample size and sequencing depth, covering more than 80% of the v3.3 *P. patens* gene models, the dataset will, along with others, permit the improvement of future gene annotation versions. The RPKM values for all individual v3.3 *P. patens* gene models in addition to 50 DEG sets are published with this study, representing a valuable benchmark reference for future RNA-seq studies. Cross-comparison across large datasets is an important approach to confirm transcript specificity to any biological phenomenon, be it developmental, as exemplified in the present study by the stage comparisons, or environmental, as indicated by the laboratory replicates. As more datasets are published, the body of data will permit a better understanding of variable parameters, but it is clear that the RNA-seq approach is sensitive enough to detect differences in growth conditions, qualitative or quantitative, that can escape careful laboratory observation. Therefore, experimental replica conditions should be very carefully controlled and documented to allow for comparison within and between laboratories. More *P. patens* Gene Atlas data are forthcoming, representing, for example, additional developmental stages such as non-germinated spores and gametophores bearing gametangia (sexual organs), as well as further perturbation experiments looking into the response of the plant to variation in phosphate concentration in the medium, which will further enhance the usefulness of the present set of expression profiling data. The data presented here are currently available as a supplement to this paper (Data S1). Moreover, expression values assigned to genes can be accessed at Phytozome (https://phytozome.jgi.doe.gov/). Other large-scale *P. patens* expression data are available at Genevestigator (Hiss *et al.*, 2014) and the eFP browser (Ortiz-Ramirez *et al.*, 2016). A valuable future goal is to unify these data into a single resource.

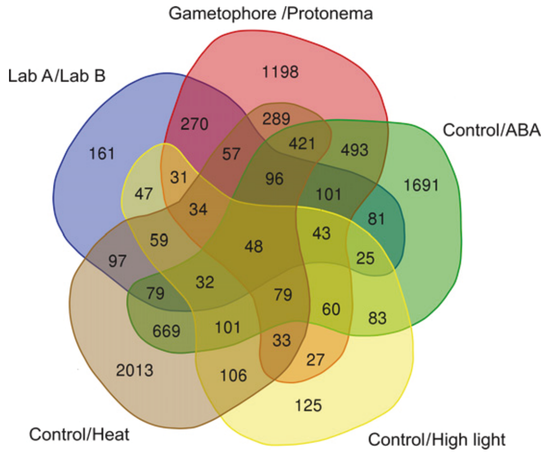


Figure 5. Contrasting differentially expressed genes (DEGs) across experiments.
Venn diagram analysis of the DEGs of replica experiments between different laboratories (Laboratory A, experiment VII, compared with Laboratory B, experiment XXII), and with DEGs between protonema and gametophore (protonema, experiment VII, compared with gametophore, experiment XX), between ABA treatment (Control, experiment XVIII, compared with ABA, experiment XIX), heat treatment (Control, experiment XXV, compared with Heat, experiment XXVI) and high light treatment (Control, experiment I, compared with High light, experiment II). The number of transcripts meeting the cut-off values are contained within each section of the labeled circle (up- or downregulated by more than twofold; adjusted $P < 0.05$).

EXPERIMENTAL PROCEDURES

Plant material

Physcomitrella patens Gransden (Engel, 1968) was used for all samples apart from the two sporophyte sets, for which *P. patens* Reute was used (Hiss *et al.*, 2017). The protonemata cultures were systematically entrained by two successive weeks of culture prior to treatment, in order to obtain a homogeneous culture. Medium referred to as BCD uses the composition described by Cove *et al.* (2009), and medium referred to as Knop uses the composition described by Reski and Abel (1985), based on Knop's medium (Knop, 1868). Solid medium [medium with 1% (w/v) agar] protonemal cultures were grown on top of a cellophane film to allow tissue transfer for specific treatments (e.g. with hormones), and for easy harvesting. If not otherwise mentioned, Petri dishes were sealed with parafilm during the growing period and plates and flasks were cultivated at 22°C with a 16-h light/8-h dark regime under 60–80 μmol m^{-2} s^{-1} white light (long-day conditions). All harvests were performed in the middle of the light photoperiod (+8 h of light in long-day conditions), followed by immediate flash-freezing in liquid nitrogen, unless otherwise stated. All experiments, referred to by Roman numerals, consist of biological triplicates of the given conditions.

Sample description

Table S1 presents a simplified version of the experiments with the associated repository references for the libraries.

Light treatments

Light quality. Prior to treatment, *P. patens* protonemata were cultivated for 1 week on solidified supplemented BCD medium (BCD supplemented with 5 mM ammonium tartrate and 0.5% sucrose, for dark-treated samples, or 0.5% glucose, for light treatments). Plants were cultivated in long-day conditions. Subsequently, the cultures were transferred into the light conditions described below.

- Dark-treated samples were grown in darkness for 1–2 weeks (experiment XXIX).
- Far-red light samples were grown under continuous 2 μmol m^{-2} s^{-1} far-red light at 720–740 nm for 1–2 weeks and then harvested (experiment XXXIII).
- Red-light samples were grown under continuous 10 μmol m^{-2} s^{-1} red light at 660–680 nm for 1–2 weeks and then harvested (experiment XXX).
- Blue-light samples were grown under continuous 5 μmol m^{-2} s^{-1} blue light at 460–480 nm for 1–2 weeks and then harvested (experiment XXXI).
- UV-B light samples were grown under continuous 4 μmol m^{-2} s^{-1} white light supplemented with 1 μmol m^{-2} s^{-1} UV-B light at 300–320 nm for 1–2 weeks and then harvested (experiment XXVIII).

Light quantity. Prior to treatment, *P. patens* protonemata were cultivated for 10 days under standard long-day conditions on solidified Knop medium.

- High-light samples were transferred to 850 μmol m^{-2} s^{-1} white light for 2 h and then harvested (experiment II).
- Low-light samples were transferred to 10 μmol m^{-2} s^{-1} white light for 2 h and then harvested (experiment III).
- Control light samples were maintained at 70 μmol m^{-2} s^{-1} white light for 2 h and then harvested (experiment I).

Tissues

Germinating spores—Physcomitrella patens mature sporophytes were harvested, opened and spores suspended in sterile water. Spores were distributed on Petri dishes containing solid Knop medium supplemented with 5 mM ammonium tartrate. Spores were germinated for 4 days at 24°C under continuous light and then harvested (experiment IV).

Protonemata grown on solid medium—Protonemata were cultivated on solidified BCD medium, in Petri dishes sealed with 3M Micropore tape under standard conditions. Protonemata were cultivated for 7 days and then harvested (experiment XVIII).

Protonemata grown in liquid medium—Protonemata were cultivated in liquid Knop medium in flasks with continuous shaking under standard conditions. Protonemata were cultivated for 7 days and then harvested by filtering with a 100-μm sieve (experiments VII, XXI and XXII).

Gametophores—Gametophores were cultivated on solidified BCD medium without cellophane, in Petri dishes sealed with 3M Micropore tape under standard conditions. Gametophores were cultivated for 5 weeks and the aerial parts of the plants were harvested (experiment XVII).

Gametophores were cultivated on solidified Knop medium without cellophane, under standard conditions. Gametophores were grown for 5 weeks and the aerial parts of the plants were harvested (experiment XX).

Leaflets (phyllids, non-vascular leaves of the gametophore): Gametophores were cultivated on solidified Knop medium without cellophane, under standard conditions. After 5 weeks of growth, leaflets were separated from the gametophore stem using forceps, harvested and stored in RNA*later* solution (Qiagen, Hilden, Germany) before flash-freezing in liquid nitrogen as a single sample (experiment XIV).

Sporophytes—Physcomitrella patens Reute sporophytes were induced as initially described by Hohe and collaborators, and later modified by Hiss and collaborators (Hohe *et al.*, 2002; Hiss *et al.*, 2017). Briefly, gametophytic tissue was grown for 5 weeks on solidified mineral Knop medium under standard conditions. Gametangia production was induced by transferring the plates to 16°C under an 8-h light/16-h dark regime, with 20 μmol m^{-2} s^{-1} white light (short-day conditions). After 3 weeks in this regime, plants were watered regularly to promote efficient fertilization and allowed to develop under short-day conditions.

Green sporophytes with a round capsule shape and green color were harvested after 5 weeks of short-day growth conditions (immature post meiotic stage M; Hiss *et al.*, 2017) and stored in RNA*later* solution (Qiagen) before flash-freezing in liquid nitrogen as a single sample (experiment XV).

Brown sporophytes with a round capsule shape and brown color were harvested after 7 weeks of short-day growth conditions (mature stage stage B; Hiss *et al.*, 2017), and stored in RNA*later* solution (Qiagen) before flash-freezing in liquid nitrogen as a single sample (experiment XVI).

Hormones

Auxin—Gametophores were cultivated on a sieve above liquid Knop medium in Magenta boxes (Sigma-Aldrich, Merck KGaA, Darmstadt, Germany) sealed with parafilm under standard conditions for 10 months. At that time point, auxin samples were treated with 10 μM naphthaleneacetic acid (NAA) and cultivated for 10 days before harvesting. Samples with and without NAA were generated (experiment XXXIII and XXXIV).

Strigolactone—Protonemata were cultivated on solidified BCD medium in Petri dishes sealed with 3M Micropore tape under standard conditions. The tissue was cultivated for 21 days. Cellophane disks containing tissue were transferred to BCD plates containing either 1 μM racemic GR24 (synthetic strigolactone) or acetone without GR24, as a control, incubated for 24 h and then harvested (experiment V and XXXVIII).

Gibberellin—Protonemata were cultivated on solidified BCD medium, supplemented with 20 μM GA_9-methylester under standard conditions. GA_9-methylester was synthesized and donated by Peter Hedden's group at Rothamstead Research (https://www.rothamsted.ac.uk). Protonemata were cultivated for 7 days and then harvested (experiment XI).

Abscisic acid—Protonemata were cultivated on solidified BCD medium, in Petri dishes sealed with 3M Micropore tape under standard conditions. Protonemata were cultivated for 6 days. Cellophane disks containing tissue were transferred to BCD plates containing 50 μM abscisic acid (ABA) and incubated for 24 h before harvesting (experiment XIX).

OPDA—Protonemata were cultivated on solidified BCD medium, supplemented with 5 mM ammonium tartrate, with or without 50 μM 12-oxo-phytodienoic acid (OPDA), in Petri dishes sealed with 3M Micropore tape under standard conditions. Protonemata were cultivated for 14 days. Cellophane disks containing tissue were transferred to ammonium tartrate-containing BCD plates, with or without 50 μM OPDA and incubated for 6 h before harvesting (experiments X and IX, respectively).

Perturbations

Protoplasts—Protonemata were cultivated for 6 days on solidified mineral medium BCD, supplemented with ammonium tartrate (2.7 mM) and glucose (0.5%), in Petri dishes sealed with 3M Micropore tape under standard conditions. Protoplasts were released using driselase treatment (Cove *et al.*, 2009) and then harvested (experiment VIII).

Ammonium treatment—Protonemata cultivated in liquid Knop medium was used to inoculate two parallel cultures, one with Knop medium and a second with Knop medium supplemented with 5 mM ammonium tartrate, in flasks with continuous shaking under standard conditions. Protonemata were cultivated for 7 days and then harvested 2 h after the lights were turned on (experiment XXIV, without ammonium; experiment XXIII, with ammonium).

De- and rehydration—Gametophores were cultivated on cellophane disks on solidified BCD medium under standard conditions for 5 weeks prior to dehydration treatment. The gametophores on the cellophane disks were placed in empty Petri dishes that were sealed in chambers containing an atmosphere of 91% relative humidity (RH) generated by a saturated solution of $MgSO_4$ in an incubator at 17°C, with a 16-h light/8-h dark cycle. Gametophores were exposed to the dehydrating atmosphere until they reached a constant weight (equilibrium). Equilibrium was reached at approximately 150 h (Koster *et al.*, 2010), and gametophores were sampled at 180 h (experiment XII). The water potential of the gametophore tissue at equilibrium was –13 MPa. Rehydration was achieved by floating the cellophane disks containing the dehydrated gametophores on sterile water in a Petri dish for 5 min to ensure full rehydration. Once fully rehydrated the disks were placed on solid BCD media and incubated under standard conditions in the light for 2 h before harvest (experiment XIII).

Heat—Protonemata were cultivated on solidified BCD medium supplemented with 5 mM ammonium tartrate under continuous light for the duration of the treatment. The heat treatment was applied after 5 days of pre-growth and lasted for 5 days, with repeated heat-shock cycles of 5 h at 22°C followed by 1 h at 37°C, before harvesting (experiment XXVI, treatment; experiment XXVII, control).

RNA extraction, RNA processing and sequencing

Frozen samples were pulverized with a mortar and pestle and total RNA was extracted in two steps: (i) total RNA was extracted using Trizol reagent (Invitrogen, now ThermoFisher Scientific, https://www.thermofisher.com), using the manufacturer's instructions (maximum of 100 mg of tissue per ml of Trizol reagent); and (ii) total RNA was purified using the RNeasy Plant Mini Kit (Qiagen), omitting the shredding step of the kit. Total RNA was checked for integrity using a BioAnalyzer with an Agilent RNA 6000 Nano Chip, following the manufacturer's instructions (Agilent, https://www.agilent.com). Plate-based RNA sample preparation was performed on the PerkinElmer Sciclone NGS robotic liquid handling system (http://www.perkinelmer.com), using Illumina's TruSeq Stranded mRNA HT sample prep kit (Illumina, https://www.illumina.com) with poly-A selection of mRNA following the protocol outlined by Illumina in their user guide (http://support.illumina.com/sequencing/sequencing_kits/truseq_stranded_mrna_ht_sample_prep_kit.html), and with the following conditions: total RNA starting material was 1 μg per sample and eight cycles of PCR were used for library amplification. The prepared libraries were then quantified by qPCR using the Kapa SYBR Fast Illumina Library Quantification Kit (Kapa Biosystems, https://www.kapabiosystems.com), and run on a Roche LightCycler 480 real-time PCR instrument (Roche, https://www.roche.com). The quantified libraries were then prepared for sequencing on the Illumina HiSeq sequencing platform using a TruSeq paired-end cluster kit, v4, and Illumina's cBot instrument to generate a clustered flow cell for sequencing. Sequencing of the flow cell was performed on the Illumina HiSeq2500 sequencer using HiSeq TruSeq SBS sequencing kits, v4, following a 2 × 150 indexed run recipe.

RNA-seq processing

The RNA-seq processing steps described below are presented in a condensed view in the illustrated pipeline presented in Figure 1.

Quality trimming and adapter removal. Each library was initially checked with FASTQC 0.11.2 to evaluate ther read quality (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Subsequently, lower quality bases and sequencing adapters were removed using TRIMMOMATIC 0.33 (Bolger *et al.*, 2014) with the following parameters: ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 SLIDINGWINDOW:4:15 HEADCROP:12 MINLEN:50. Finally, a read length of minimum 50 nt per read was required for further processing.

Poly-A tail trimming. As a result of the nature of RNA-seq data, poly-A tails are expected: poly-A tails with a minimum length of 12 were identified and removed with PRINSEQ 0.20.4 (Schmieder and Edwards, 2011).

Paired-end read merging. During an Illumina paired-end sequencing, fragmented RNA will be sequenced from both sides. If the fragments are smaller than the double read length, the reads overlap each other. Such overlapping reads were merged with the help of COPEREAD 1.2.5 (Liu *et al.*, 2012).

Mapping. The read mapping was performed using GSNAP 2015-12-31.v5 (Wu and Nacu, 2010), with the options -A sam -N 1 –split-output –failed-input. The read mapping was performed in two steps: all reads were mapped first against *P. patens* organellar genomes and rRNA sequences (mitochondrial NC_007945.1; chloroplast NC_005087.1; ribosomal HM751653.1, X80986.1 and X98013.1). The remaining reads were then mapped against *P. patens* genome v3 (Lang *et al.*, 2018; https://phytozome.jgi.doe.gov/pz/portal.html) and concordant unique mapped read pairs were retained.

File converting. The conversion of the mapping output files from SAM to BAM format and the sorting by positions was performed using SAMTOOLS 1.2 (Li *et al.*, 2009).

Read count. For read counting, HTSEQ-COUNT 0.6.1p1 (Anders *et al.*, 2015), in combination with the *P. patens* gene model v3.3 (Lang *et al.*, 2018), was applied. Additionally for default options, the following parameters were set: -s reverse -r pos -t exon -i Parent.

Differential expression analysis

Differentially expressed gene (DEG) calling approaches can generate different results (Zhang *et al.*, 2014; Schurch *et al.*, 2016).

Hence, in order improve confidence in the DEGs used here, several algorithms were tested. DEG analysis was performed in R 3.2.0 using three R packages: EGDER 3.14.0 (Robinson *et al.*, 2010); DESEQ2 1.12.3 (Love *et al.*, 2014); and NOISEQ 2.12.0 (Tarazona *et al.*, 2011). *P*-value cut-offs for EDGER and DESEQ2 were 0.001, and for NOISEQ the probability of differential expression ('prob') was >0.9. Genes with zero counts in all replicates were removed. The previously detected array DEGs are known to be of high quality (Hiss *et al.*, 2014). The higher sensitivity of RNA-seq based approaches often leads to the calling of DEGs that exhibit very low expression levels, the biological significance of which might be questionable. In order to rely on a trustworthy set of DEGs we decided to use the NOISEQ RPKM-normalized DEGs because they capture the majority of the array DEGs, overlap with a high fraction of DEGs also detected by EDGER and DESEQ2, but exclude a high number of DEGs detected only by the latter two tools (for a Venn diagram representation of the four-way DEG call comparison, see Figure S7), which are characterized by a particularly low average expression level (3.5 FPKM, as compared with the 51.7 FPKM average for the overlap of array and all three RNA-seq DEG callers). The number of DEGs called exclusively by NOISEQ (not overlapping with other approaches) is the lowest one of all approaches. Thus, for further analysis NOISEQ RPKM-normalized DEGs were used.

To further confirm our DEG procedure using the NOISEQ approach, experiments XX and XXI were compared with previously published microarray data (Hiss *et al.*, 2014). Both experiments on both platforms, gametophores on solid Knop medium (XX) and protonemata in liquid Knop medium (XXI), were performed in the same laboratory. A total of 620 DEGs were called by the microarray data (Hiss *et al.*, 2014). With its higher sensitivity RNA-seq called 3309 DEGs. We found that 69% of the microarray DEGs overlap with the RNA-seq data, providing evidence that the RNA-seq based DEGs coincide well with previous approaches. An even better overlap can be found by comparing the GO terms associated with both DEG sets. The microarray GO terms were found to share 95.5% of their associated terms with the GO terms associated with the RNA-seq DEGs (Data S5). A total of 758 of 762 GO terms did not show significant differences between the two DEG sets in terms of their number of associated genes (Fisher's exact test, $P_{\text{adjust}} < 0.05$).

GO term enrichment analysis

The GO enrichment analyses were conducted as described previously (Widiez *et al.*, 2014). Visualization of the GO terms was implemented using word clouds using the http://www.wordle.net application. Word size is proportional to the –log10 (q value), and over-represented GO terms were colored dark green if –log10 (q value) ≤ 4 and light green if –log10 (q value) > 4.

Data visualization

Principal component analysis (PCA) was performed in R 3.2.0 using the R function PRINCOMP (https://stat.ethz.ch/R-manual/R-devel/library/stats/html/princomp.html). PCA visualization was generated using the R package PLOT3D 1.1.1 (https://CRAN.R-project.org/package=plot3D). Hierarchical clustering was calculated and visualized using R 3.4.3 and the package COMPLEXHEATMAP 1.17.1 (Gu *et al.*, 2016), with the Euclidean distance method. The calculation was performed with all gene models, with at least three samples with RPKM > 2. Venn diagrams were created using the online tool VENN DIAGRAM developed by the University of Gent (http://bioinformatics.psb.ugent.be/webtools/Venn/), with the symmetric and colored options.

ACKNOWLEDGEMENTS

We would like to thank Christine Glockner (University of Freiburg), Marco Göttig (University of Marburg) and Mitsuhiro Ishibashi (Hokkaido University) for technical assistance. The work conducted by the US Department of Energy Joint Genome Institute 902 is supported by the Office of Science of the US Department of Energy under Contract No. DE-903 AC02-05CH11231. H.C. and N.S.'s work was supported by grants from the National Natural Science Foundation of China (31621001). R.B. and A.A. acknowledge the support of the EEC project S2B (675006). R.R. and S.A.R. acknowledge financial support from the Excellence Initiative of the German Federal and State Governments (EXC 294). J.C. and Y.S.'s work was supported by the UK Leverhulme Trust research grant F/0094/BA. J.C. thanks Peter Hedden for his generous gift of the GA9-methylester. P.S.'s work was supported by the Swiss National Science Foundation (160004 and 131726) and the URPP Evolution in Action. M.R.'s work was supported by the Plant Fellows program and the 'Forschungskredit der Universität Zurich'. S.B., A.G., F.N., and C.R. acknowledge the support of the LabEx Saclay Plant Sciences-SPS (ANR-10-LABX-0040-SPS).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

SUPPORTING INFORMATION

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

Figure S1. Hierarchical clustering of all 99 RNA-seq samples, RPKM normalized. The upper colored line represents the experiments, the lower one the corresponding tissues. This analysis confirms and illustrates that the replicates for each experiment cluster as expected with each other in most of the cases, as for example experiment XVI (brown sporophyte). The exceptions are the cases in which consequently few DEGs were detected. For example, experiments V and XXXVIII libraries group together, indication of the closeness of these samples, but the triplicates are not resolved between the two experiments. The clustering provides independent confirmation of the absence of effect of the strigolactone treatments in this specific experiment.

Figure S2. GO term enrichment of developmental stage DEGs. GO term analysis representation of all over-represented GO terms associated with DEGs between specific stages. (A): Over-represented GO terms associated with the up-DEGs in sporophyte (exp. XV) compared to gametophore (exp. XVII). (B): Over-represented GO terms associated with the down-DEGs in sporophyte (exp. XV) compared to gametophore (exp. XVII). (C): GO terms associated with the up-DEGs in protonema (exp. VII) compared to sporophyte (exp. XV). (D): GO terms associated with the down-DEGs in protonema (exp. VII) compared to sporophyte (exp. XV). (E): GO terms associated with the up-DEGs in protonema (exp. VII) compared to gametophore (exp. XV). (F): GO terms associated with the down-DEGs in protonema (exp. VII) compared to gametophore (exp. XV). Word size is proportional to the –log10 (q-value) and over-represented GO terms were colored dark green if –log10 (q-value) ≤ 4 and light green if –log10 (q-value) >4 .For the GO term IDs and their respective over-representation values see Data S2.

Figure S3. Impact of ammonium supplementation on protonemata transcriptome. GO term analysis representation of all over-represented GO terms associated with DEGs between ammonium-supplemented liquid protonemal culture (exp. XXIII) and ammonium-free liquid protonemal culture (exp. XXIV). (A): Over-represented GO terms associated with the up-DEGs in ammonium-supplemented liquid protonemal culture (exp. XXIII) compared to ammonium-free liquid protonemal culture (exp. XXIV). (B): Over-represented GO terms associated with the down-DEGs in ammonium-supplemented liquid protonemal culture (exp. XXIII) compared to ammonium-free liquid protonemal culture (exp. XXIV). Word size is proportional to the –log10 (q-value) and over-represented GO terms were colored dark green if –log10 (q-value) ≤ 4 and light green if –log10 (q-value) > 4. (C): Typical aspect of protonemal cells in presence of ammonium. (D): Typical aspect of protonemal cells in absence of ammonium. For the GO term IDs and their respective over-representation values see Data S3.

Figure S4. Specific transcripts affected by ammonium supplementation in protonemata. (A): Repressed transcripts upon ammonium supplementation pertaining directly to nitrate metabolism. (B): Repressed transcripts upon ammonium supplementation related to cell wall remodeling. (C): Induced gene upon ammonium supplementation related to primary carbon fixation metabolism. Error bar: standard deviation, n=3.

Figure S5. Library quality evaluation of the laboratory comparison dataset. Read length distribution of for the different libraries. (A): 7805.5.84013.TCGGCA, (B): 7806.3.84002.ACAAA, (C): 7806.5.84005.CGAGAA, (D): 7806.6.83999.TGAATG, (E): 7805.1.84009.CTTGTA, (F): 7805.1.84009.CCGTCC, (G): 7805.2.84012.ATGAGC, (H): 7806.6.83999.TTCGAA, (I): Overview of the library primary statistic for experiments VII, XXI and XXII.

Figure S6. Comparison between replica experiments performed in a single laboratory. GO term analysis representation of all over-represented GO terms associated with DEGs between the protonemal replica experiment XXII and XXI. (A): Over-represented GO terms associated with up-DEGs in experiment XXII compared to experiment XXI. (B): Over-represented GO terms associated with down-DEGs in experiment XXII compared to experiment XXI. Word size is proportional to the –log10 (q-value) and over-represented GO terms were colored dark green if –log10 (q-value) ≤ 4 and light green if –log10 (q-value) > 4. For the GO terms IDs and their respective overrepresentation values see Data S4.

Figure S7. Comparison of the DEGs called by the NOISeq, DESeq2 and edgeR packages with RNA dataset and by microarray approach. Venn diagram comparing the DEGs called by NOISeq (in blue), DESeq2 (in yellow) and edgeR (in green) between the Experiments XXI (gametophore) and XX (protonemal liquid culture) and the DEGs called in a microarray experiment performed on the same tissues (Hiss et al. 2014).

Table S1. Overview of the experiments and their primary library data presented in this study.

Table S2. Harvesting time point after initiation of the specific culture and experimental location for each experiment in this study.

Table S3. Overview of the experiment pairs for which DEGs have been calculated in the present study.

Table S4: Tissue markers detection in *P. patens* microarray studies. nt: not tested; - no specificity; +/-: gene model displays in specific tissue enrichment but present somewhere else; +: tissue specificity confirmed with no other tested tissue showing any transcript accumulation.

Data S1. Calculated RPKM for all *Physcomitrella patens* gene models V3.3 in all libraries described in the present study and calculated DEG for 50 relevant comparisons.

Data S2. GO terms enrichment lists associated with ammonium supplementation.

Data S3. GO terms enrichment lists associated with stage specific comparison.

Data S4. GO terms enrichment lists associated lab specific replication.

Data S5. GO terms comparison between the RNA-seq experiment XX and XXI and the similar treatment performed with microarray approach (data from Hiss *et al.* 2014).

REFERENCES

Alaba, S., Piszczalka, P., Pietrykowska, H. *et al.* (2015) The liverwort *Pellia endiviifolia* shares microtranscriptomic traits that are common to green algae and land plants. *New Phytol.* **206**, 352–367.

Anders, S., Pyl, P.T. and Huber, W. (2015) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

Aya, K., Kobayashi, M., Tanaka, J., Ohyanagi, H., Suzuki, T., Yano, K., Takano, T., Yano, K. and Matsuoka, M. (2015) De novo transcriptome assembly of a fern, *Lygodium japonicum*, and a web resource database, ljtrans DB. *Plant Cell Physiol.* **56**, e5.

Banks, J.A., Nishiyama, T., Hasebe, M. *et al.* (2011) The compact *Selaginella genome* identifies changes in gene content associated with the evolution of vascular plants. *Science (New York, N.Y.)*, **332**, 960–963.

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

Bowman, J.L., Kohchi, T., Yamato, K.T. *et al.* (2017) Insights into Land Plant Evolution Garnered from the *Marchantia polymorpha* Genome. *Cell*, **171** (287–304), e215.

Bushart, T.J., Cannon, A.E., ul Haque, A., San Miguel, P., Mostajeran, K., Clark, G.B., Porterfield, D.M. and Roux, S.J. (2013) RNA-seq analysis identifies potential modulators of gravity response in spores of *Ceratopteris* (Parkeriaceae): evidence for modulation by calcium pumps and apyrase activity. *Am. J. Bot.* **100**, 161–174.

Champagne, C.E.M. and Ashton, N.W. (2001) Ancestry of KNOX genes revealed by bryophyte (*Physcomitrella patens*) homologs. *New Phytol.* **150**, 23–36.

Chang, C.-Y., Lin, W.-D. and Tu, S.-L. (2014) Genome-wide analysis of heat-sensitive alternative splicing in *Physcomitrella patens*. *Plant Physiol.* **165**, 826–840.

Chen, Y.-R., Su, Y.-S. and Tu, S.-L. (2012) Distinct phytochrome actions in nonvascular plants revealed by targeted inactivation of phytobilin biosynthesis. *Proc. Natl Acad. Sci. USA*, **109**, 8310–8315.

Colpitts, C.C., Kim, S.S., Posehn, S.E., Jepson, C., Kim, S.Y., Wiedemann, G., Reski, R., Wee, A.G.H., 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.

Coruh, C., Cho, S.H., Shahid, S., Liu, Q., Wierzbicki, A. and Axtell, M.J. (2015) Comprehensive annotation of *Physcomitrella patens* small RNA loci reveals that the heterochromatic short interfering RNA pathway is largely conserved in land plants. *Plant Cell*, **27**, 2148–2162.

Cosgrove, D.J. (2016) Catalysts of plant cell wall loosening. *F1000Res*, **5**, 119. https://doi.org/10.12688/f1000research.7180.1

Coskun, D., Britto, D.T. and Kronzucker, H.J. (2015) The nitrogen-potassium intersection: membranes, metabolism, and mechanism. *Plant, Cell Environ.* **40**, 2029–2041.

Coskun, D., Britto, D.T. and Kronzucker, H.J. (2016) Nutrient constraints on terrestrial carbon fixation: the role of nitrogen. *J. Plant Physiol.* **203**, 95–109.

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 Harbor Protocol*, **2009**, pdb emo115.

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.

Decker, E.L., Alder, A., Hunn, S. *et al.* (2017) Strigolactone biosynthesis is evolutionarily conserved, regulated by phosphate starvation and contributes to resistance against phytopathogenic fungi in a moss, *Physcomitrella patens*. *New Phytol.* **216**, 455–468.

Demko, V., Perroud, P.F., Johansen, W. *et al.* (2014) Genetic analysis of DEFECTIVE KERNEL1 loop function in three-dimensional body patterning in *Physcomitrella patens*. *Plant Physiol.* **166**, 903–919.

Denancé, N., Szurek, B. and Noël, L.D. (2014) Emerging functions of nodulin-like proteins in non-nodulating plant species. *Plant Cell Physiol.* **55**, 469–474.

Devos, N., Szövényi, 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.

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

Eveland, A.L., Satoh-Nagasawa, N., Goldshmidt, A., Meyer, S., Beatty, M., Sakai, H., Ware, D. and Jackson, D. (2010) Digital gene expression signatures for maize development. *Plant Physiol.* **154**, 1024–1039.

Gaidatzis, D., Burger, L., Florescu, M. and Stadler, M.B. (2015) Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat. Biotechnol.* **33**, 722–729.

Gao, B., Zhang, D., Li, X., Yang, H. and Wood, A.J. (2014) De novo assembly and characterization of the transcriptome in the desiccation-tolerant moss *Syntrichia caninervis*. *BMC Res. Notes* **7**, 490.

Gao, B., Zhang, D., Li, X., Yang, H., Zhang, Y. and Wood, A.J. (2015) De novo transcriptome characterization and gene expression profiling of the desiccation tolerant moss *Bryum argenteum* following rehydration. *BMC Genom.* **16**, 416.

González-Ballester, D., Casero, D., Cokus, S., Pellegrini, M., Merchant, S.S. and Grossman, A.R. (2010) RNA-seq analysis of sulfur-deprived *Chlamydomonas* cells reveals aspects of acclimation critical for cell survival. *Plant Cell*, **22**, 2058.

Gu, Z., Eils, R. and Schlesner, M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.

Hachiya, T. and Sakakibara, H. (2017) Interactions between nitrate and ammonium in their uptake, allocation, assimilation, and signaling in plants. *J. Exp. Bot.* **68**, 2501–2512

Harrison, J.C. (2017) Development and genetics in the evolution of land plant body plans. *Philosophical Transactions of the Royal Society B: Biological Sciences* **372**, 20150490.

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.

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.

Hoffmann, B., Proust, H., Belcram, K., Labrune, C., Boyer, F.D., Rameau, C. and Bonhomme, S. (2014) Strigolactones inhibit caulonema elongation and cell division in the moss *Physcomitrella patens*. *PLoS ONE*, **9**, e99206.

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**, 595–602.

Kamisugi, Y., Whitaker, J.W. and Cuming, A.C. (2016) The transcriptional response to DNA-double-strand breaks in *Physcomitrella patens*. *PLoS ONE*, **11**, e0161204.

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

Koster, K.L., Balsamo, R.A., Espinoza, C. and Oliver, M.J. (2010) Desiccation sensitivity and tolerance in the moss *Physcomitrella patens*: assessing limits and damage. *Plant Growth Regul.* **62**, 293–302.

Ladwig, F., Stahl, M., Ludewig, U., Hirner, A.A., Hammes, U.Z., Stadler, R., Harter, K. and Koch, W. (2012) Siliques are Red1 from arabidopsis acts as a bidirectional amino acid transporter That is crucial for the amino acid homeostasis of siliques. *Plant Physiol.* **158**, 1643–1655.

Lamport, D.T., Kieliszewski, M.J., Chen, Y. and Cannon, M.C. (2011) Role of the extensin superfamily in primary cell wall architecture. *Plant Physiol.* **156**, 11–19.

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

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

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

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H. and Ecker, J.R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523–536.

Liu, B., Yuan, J., Yiu, S.M. *et al.* (2012) COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. *Bioinformatics*, **28**, 2870–2874.

Lopez-Obando, M., Conn, C.E., Hoffmann, B., Bythell-Douglas, R., Nelson, D.C., Rameau, C. and Bonhomme, S. (2016) Structural modelling and transcriptional responses highlight a clade of PpKAI2-LIKE genes as candidate receptors for strigolactones in *Physcomitrella patens*. *Planta*, **243**, 1441–1453.

Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550.

Lunde, C., Drew, D.P., Jacobs, A.K. and Tester, M. (2007) Exclusion of Na+ via sodium ATPase (PpENA1) ensures normal growth of *Physcomitrella patens* under moderate salt stress. *Plant Physiol.* **144**, 1786–1796.

Mashiguchi, K., Sasaki, E., Shimada, Y., Nagae, M., Ueno, K., Nakano, T., Yoneyama, K., Suzuki, Y. and Asami, T. (2009) Feedback-regulation of strigolactone biosynthetic genes and strigolactone-regulated genes in Arabidopsis. *Biosci. Biotechnol. Biochem.* **73**, 2460–2465.

Matasci, N., Hung, L.-H., Yan, Z. *et al.* (2014) Data access for the 1,000 Plants (1KP) project. *GigaScience*, **3**, 17.

Mayzlish-Gati, E., LekKala, S.P., Resnick, N., Wininger, S., Bhattacharya, C., Lemcoff, J.H., Kapulnik, Y. and Koltai, H. (2010) Strigolactones are positive regulators of light-harvesting genes in tomato. *J. Exp. Bot.* **61**, 3129–3136.

Niklas, K.J., Cobb, E.D. and Matas, A.J. (2017) The evolution of hydrophobic cell wall biopolymers: from algae to angiosperms. *J. Exp. Bot.* **68**, 5261–5269

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.

Ortiz-Ramirez, C., Hernandez-Coronado, M., Thamm, A., Catarino, B., Wang, M., Dolan, L., Feijo, J.A. and Becker, J.D. (2016) A Transcriptome Atlas of *Physcomitrella patens* Provides Insights into the Evolution and Development of Land Plants. *Mol. Plant*, **9**, 205–220.

Rama Reddy, N.R., Mehta, R.H., Soni, P.H., Makasana, J., Gajbhiye, N.A., Ponnuchamy, M. and Kumar, J. (2015) Next generation sequencing and transcriptome analysis predicts biosynthetic pathway of sennosides from senna (*Cassia angustifolia* Vahl.), a non-model plant with potent laxative properties. *PLoS ONE*, **10**, e0129422.

Regmi, K.C., Li, L. and Gaxiola, R.A. (2017) Alternate modes of photosynthate transport in the alternating generations of *Physcomitrella patens*. *Fron. Plant Sci.* **8**, 1956.

Renault, H., Alber, A., Horst, N.A. *et al.* (2017) A phenol-enriched cuticle is ancestral to lignin evolution in land plants. *Nat. Commun.* **8**, 14713.

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.

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

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

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. Devel.* **10**, 555–566.

Schaefer, D.G. and Zrÿd, J.P. (1997) Efficient gene targeting in the moss *Physcomitrella patens*. *Plant J.* **11**, 1195–1206.

Schmieder, R. and Edwards, R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.

Schurch, N.J., Schofield, P., Gierlinski, M. *et al.* (2016) How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, **22**, 839–851.

Sharma, N., Jung, C.-H., Bhalla, P.L. and Singh, M.B. (2014) RNA Sequencing Analysis of the Gametophyte Transcriptome from the Liverwort, *Marchantia polymorpha*. *PLoS ONE*, **9**, e97497.

Singer, S.D. and Ashton, N.W. (2007) Revelation of ancestral roles of KNOX genes by a functional analysis of *Physcomitrella* homologues. *Plant Cell Rep.* **26**, 2039–2054.

Song, L., Shankar, D.S. and Florea, L. (2016) Rascaf: improving Genome Assembly with RNA Sequencing Data. *Plant Genome*, **9**, https://doi.org/10.3835/plantgenome2016.03.0027.

Stevenson, S.R., Kamisugi, Y., Trinh, C.H. *et al.* (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. *Plant Cell*, **28**, 1310–1327.

Szövényi, P., Rensing, S.A., Lang, D., Wray, G.A. and Shaw, A.J. (2011) Generation-biased gene expression in a bryophyte model system. *Mol. Biol. Evol.* **28**, 803–812.

Szövényi, P., Perroud, P.F., Symeonidi, A., Stevenson, S., Quatrano, R.S., Rensing, S.A., Cuming, A.C. and McDaniel, S.F. (2015) De novo assembly and comparative analysis of the *Ceratodon purpureus* transcriptome. *Mol. Ecol. Resour.* **15**, 203–215.

Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A. and Conesa, A. (2011) Differential expression in RNA-seq: a matter of depth. *Genome Res.* **21**, 2213–2223.

Tsujimoto, R., Yamazaki, H., Maeda, S.-I. and Omata, T. (2007) Distinct roles of nitrate and nitrite in regulation of expression of the nitrate transport genes in the moss *Physcomitrella patens*. *Plant Cell Physiol.* **48**, 484–497.

de Vries, J., de Vries, S., Slamovits, C.H., Rose, L.E. and Archibald, J.M. (2017) How embryophytic is the biosynthesis of phenylpropanoids and their derivatives in streptophyte algae? *Plant Cell Physiol.* **58**, 934–945.

Waldie, T., McCulloch, H. and Leyser, O. (2014) Strigolactones and the control of plant development: lessons from shoot branching. *Plant J.* **79**, 607–622.

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.

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.

Zhang, G., Guo, G., Hu, X. *et al.* (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res.* **20**, 646–654.

Zhang, Z.H., Jhaveri, D.J., Marshall, V.M. *et al.* (2014) A comparative study of techniques for differential expression analysis on RNA-Seq data. *PLoS ONE*, **9**, e103207.

Zhang, Z., He, Z., Xu, S., Li, X., Guo, W., Yang, Y., Zhong, C., Zhou, R. and Shi, S. (2016) Transcriptome analyses provide insights into the phylogeny and adaptive evolution of the mangrove fern genus *Acrostichum*. *Sci. Rep.* **6**, 35634.

Zhu, Y., Chen, L., Zhang, C., Hao, P., Jing, X. and Li, X. (2017) Global transcriptome analysis reveals extensive gene remodeling, alternative splicing and differential transcription profiles in non-seed vascular plant *Selaginella moellendorffii*. *BMC Genom.* **18**, 1042.

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

To understand the function of a gene and its role in the genome, expression data is indispensable. Within the last decade, including this study, hundreds of *P. patens* expression datasets were published (Busch *et al.*, 2013, Hiss *et al.*, 2014, Beike *et al.*, 2015, Ortiz-Ramírez *et al.*, 2016, Hiss *et al.*, 2017, Possart *et al.*, 2017, Perroud *et al.*, 2018, Arif *et al.*, 2019). However, three main issues arise while comparing these individual expression sets. The first challenge is data availability. Most datasets use different repositories. There are commercial tools like Genevestigator have restricted usability for non-paying customers. The second is that available tools have different visualization methods, different advantages and limitations. The last issue to solve is that three different platform approaches were used to generate expression data. The older datasets were performed on CombiMatrix and NimbleGen microarray platforms. RNA-seq, sequenced by Illumina machines, is the basis for the most recent expression dataset. The comparability depends strongly on the chosen normalization method. An even bigger issue as the normalization, are the differently used gene annotation versions. The gene model annotation is updated regularly. As mentioned above, the microarray data is older than the RNA-seq data. CombiMatrix analyses were performed on *P. patens* gene annotation version 1.2, while all NimbleGen data were designed on version 1.6. For the analysis of the RNA-seq data, the current version 3.3 was used. While RNA-seq data can be assigned to each genome annotation version, microarray datasets are bound to their annotation version the array was designed for. During the version update process, several gene features were updated. Hundreds of gene models were split, merged, shortened, expanded or simply removed. The gene identifier changed often. To reflect the old gene models on new versions, version lookup data is obligatory.

Our web-based *Physcomitrella* Expression Atlas Tool (PEATmoss) solves the majority of the previously described issues. Accessibility issues of datasets solved by PEATmoss. It is designed to host all kinds of expression datasets, even if the expression data arrived from microarray or RNA-seq (Paper 5.3, page 3). The next issue was solved with the implemented gene version lookup database (PpGML) (Paper 5.3, page 7f). Different annotation versions no longer inhibit the usability of the expression data. The last point belongs to the comparability of expression data across platforms. Therefore, a gene set normalization method was recently developed (Supporting information 9.3.2) and will be implemented soon.

The core of PEATmoss is the 3D expression cube (Paper 5.3, page 4). Gene expression levels for different tissues and treatments are indicated by colour code and animated numbers. More information can be displayed by clicking on specific genes. Further links to established databases like the NCBI, Ensembl Plants (Kersey *et al.*, 2016), SwissProt (UniProt Consortium, 2018) and many more help to collect as much information as possible (Paper 5.3, page 8). The JGI Phytozome database (Goodstein *et al.*, 2012) and CyVerse CoGe (Lyons and Freeling, 2008) play a special role. Most of the RNA-seq samples present in PEATmoss are related to the JGI Gene Atlas project. All *P. patens* V3 genome-related sequence experiments are hosted at CoGe. Implemented tools like BLAST (Altschul *et al.*, 1997) or DEG calling (Tarazona *et al.*, 2015) complete the functionality of PEATmoss.

PEATmoss is our tool to make all my analysed RNA-seq expression data publicly available. Additionally, the different datasets, as well as the different gene annotation versions can be compared with each other.

the plant journal

S E B
SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2020) doi: 10.1111/tpj.14607

RESOURCE

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

Noe Fernandez-Pozo[1], Fabian B. Haas[1], Rabea Meyberg[1], Kristian K. Ullrich[1,2], Manuel Hiss[1], Pierre-François Perroud[1], Sebastian Hanke[1], Viktor Kratz[1], Adrian F. Powell[3], Eleanor F. Vesty[4], Christopher G. Daum[5], Matthew Zane[5], Anna Lipzen[5], Avinash Sreedasyam[6], Jane Grimwood[6], Juliet C. Coates[4], Kerrie Barry[5], Jeremy Schmutz[5,6], Lukas A. Mueller[3] and Stefan A. Rensing[1,7,8,*]

[1] *Plant Cell Biology, Faculty of Biology, University of Marburg, Marburg, Germany,*
[2] *Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, Ploen, Germany,*
[3] *Boyce Thompson Institute, Ithaca, NY, USA,*
[4] *School of Biosciences, University of Birmingham, Birmingham, UK,*
[5] *US Department of Energy (DOE) Joint Genome Institute, Walnut Creek, CA 94598, USA,*
[6] *HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806, USA,*
[7] *BIOSS Centre for Biological Signalling Studies, University of Freiburg, Freiburg, Germany, and*
[8] *LOEWE Center for Synthetic Microbiology (SYNMIKRO), University of Marburg, Germany*

Received 26 March 2019; revised 14 October 2019; accepted 30 October 2019.
*For correspondence: (e-mail: noefp@uni-marburg.de; stefan.rensing@biologie.uni-marburg.de).

SUMMARY

***Physcomitrella patens* is a bryophyte model plant that is often used to study plant evolution and development. Its resources are of great importance for comparative genomics and evo-devo approaches. However, expression data from *Physcomitrella patens* were so far generated using different gene annotation versions and three different platforms: CombiMatrix and NimbleGen expression microarrays and RNA sequencing. The currently available *P. patens* expression data are distributed across three tools with different visualization methods to access the data. Here, we introduce an interactive expression atlas, *Physcomitrella* Expression Atlas Tool (PEATmoss), that unifies publicly available expression data for *P. patens* and provides multiple visualization methods to query the data in a single web-based tool. Moreover, PEATmoss includes 35 expression experiments not previously available in any other expression atlas. To facilitate gene expression queries across different gene annotation versions, and to access *P. patens* annotations and related resources, a lookup database and web tool linked to PEATmoss was implemented. PEATmoss can be accessed at https://peatmoss.online.uni-marburg.de**

Keywords: ***Physcomitrella patens*, expression atlas, RNA-seq, bioinformatics, evolution, development, plant hormones, light, annotation.**

INTRODUCTION

Physcomitrella patens is a bryophyte model plant essential for the study of plant evolution. As a bryophyte, its sister position to vascular plants is of special interest to perform evolutionary developmental (evo-devo) approaches (Rensing, 2017, 2018). The *P. patens* life cycle is predominantly haploid (Figure 1), facilitating functional genomics studies via homologous recombination or genome editing to understand gene function in the plant.

The available genomic and transcriptomic resources for bryophytes and streptophyte algae, key species to understand the transition from water to land of plant ancestors, are under-represented in comparison to angiosperms (Rensing, 2017). The V1 draft genome sequence and annotation of *P. patens* have been available since 2008 (Rensing *et al.*, 2008). Continuous work for 10 years to improve the resources and knowledge about *P. patens* resulted in the V3 chromosome-scale genome assembly and the most

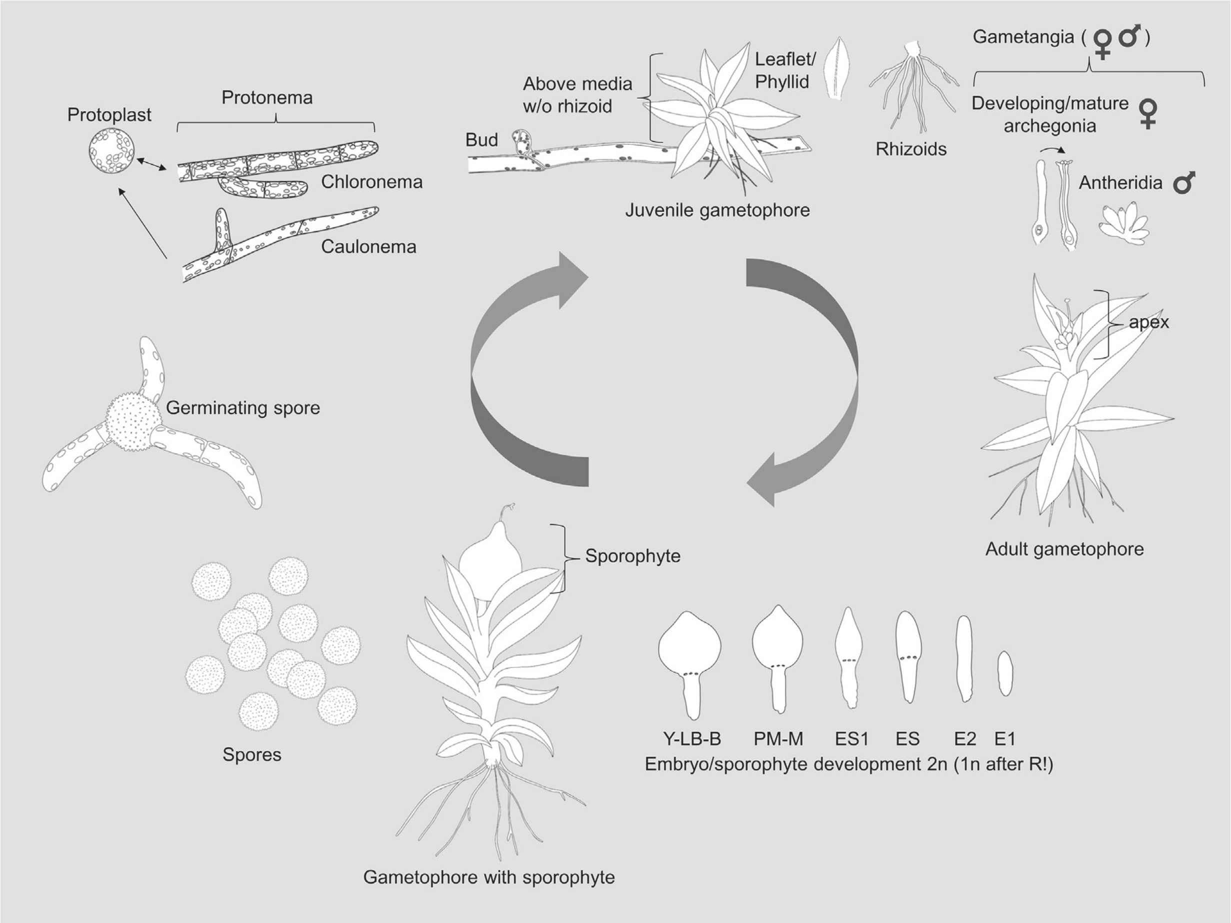


Figure 1. *Physcomitrella patens* life cycle. All stages available in PEATmoss are included. All developmental stages are haploid except the sporophyte. The embryo/sporophyte development is divided in the stages Embryo 1 (E1), Embryo 2, Early sporophyte (ES), Early sporophyte 1 (ES1), Pre-meiotic – Meiotic sporophyte (PM-M) and Yellow, Light Brown and Brown sporophyte (Y-LB-B) based on developmental stages defined in (Hiss *et al*., 2017). R! in the sporophyte development stands for reduction cell division after meiosis, when haploid spores are formed.

recent gene annotation v3.3 (Lang *et al*., 2018). These resources make *P. patens* one of the best studied non-seed plants and a good reference for comparative genomics, especially in cross-lineage studies. As such, it is one of the plant flagship species tackled in the US DOE gene atlas project (Perroud *et al*., 2018). However, the available gene expression data for *P. patens* are fragmented in three web tools/databases, comprising 77 experiments to date in the eFP Browser (Winter *et al*., 2007; Ortiz-Ramirez *et al*., 2016), Genevestigator (Hruz *et al*., 2008) and Phytozome (Goodstein *et al*., 2012), each with different advantages and limitations. We use the term experiment for all replicates generated from the same experimental condition.

The eFP (electronic fluorescent pictograph) Browser for *P. patens* (http://bar.utoronto.ca/efp_physcomitrella/cgi-bin/efpWeb.cgi) displays the expression values as coloured cartoons representing the plant or plant parts, in which different colours represent different expression values. Expression values can also be displayed as bar plots. However, in the eFP Browser, only one gene can be queried at a time or the relative expression of two genes can be compared using a different colour scale. No information about the expression values of replicates is shown. Only one data set is currently available at the eFP Browser. This data set contains 11 experiments (three replicates each) based on the gene annotation v1.6 and NimbleGen microarray (Ortiz-Ramirez *et al*., 2016), representing most of the developmental stages of the *P. patens* life cycle (Figure 1).

Alternatively, the commercial distribution of Genevestigator (https://genevestigator.com/) contains many tools for expression visualization of multiple genes. However, the free version is limited to only few basic visualization plots and only one gene per query. It contains 34 experiments (varying levels of replication), based on the gene annotation v1.2 and CombiMatrix microarray, under different conditions of light, pH, hydration/dehydration and biotic stress as well as hormone treatments (Busch *et al*., 2013; Hiss *et al*., 2014).

The third database with *P. patens* expression data is Phytozome (https://phytozome.jgi.doe.gov/). Here, the expression values of a single gene can be visualized in a table, together with the experiment name and a tag for high or low expressed genes. A very useful feature on the Phytozome expression tab is to show the list of co-expressed genes and their correlation value for the query gene. This feature is helpful for finding genes that might be involved in similar biological processes to the query gene (Ruprecht *et al.*, 2017). Phytozome includes 32 *P. patens* RNA-seq experiments (three replicates each), based on the most recent v3.3 gene annotation, and representing multiple conditions such as developmental stages, light perturbation, dehydration and hormone application (Perroud *et al.*, 2018).

Here we introduce PEATmoss (*Physcomitrella* Expression Atlas Tool), a gene expression atlas to unify the expression data of *P. patens*. This web tool is based on the well accepted Tomato Expression Atlas (Fernandez-Pozo *et al.*, 2017), is comprised of 109 experiments, provides multiple visualization methods and is available at https://peatmoss.online.uni-marburg.de.

A current limitation of *P. patens* data is that published resources are based on different gene annotation versions. Using the lookup table (Supporting Information Data S1) from Perroud *et al.* (2018) it is possible, but awkward, to look up genes across annotation versions. For that reason, a database to easily convert between gene annotation versions and to access to *P. patens* annotations and sequences was developed and is accessible through PEATmoss.

Table 1 Expression data sets for *Physcomitrella patens* in PEATmoss

Data set name	Experiments	Publication
RNA-seq developmental stages	15	Perroud *et al.* (2018) and new data[†1]
RNA-seq gametophore treatments	4	Perroud *et al.* (2018)
RNA-seq protonema treatments	25	Perroud *et al.* (2018) and new data[†2]
NimbleGen major developmental stages including sexual reproduction	11	Ortiz-Ramirez *et al.* (2016)
NimbleGen Reute development and mycorrhiza	9	Hiss *et al.* (2017)[†3], and Hanke *et al.*, in preparation[†4]
CombiMatrix major developmental stages	9	Hiss *et al.* (2014)
CombiMatrix gametophore treatments	18	Hiss *et al.* (2014), Beike *et al.* (2015)[†3], and Possart *et al.* (2017)[†3]
CombiMatrix detached leaflet development	11	Busch *et al.* (2013) and Hiss *et al.* (2014)
CombiMatrix protonema treatments	7	Hiss *et al.* (2014) and Arif *et al.*, (2019)[†3]

[†1] Includes spores for Gransden ecotype and protonema, juvenile and adult gametophores for Reute ecotype.

[†2] Time series in control conditions and phosphate deficiency.

[†3] Published data but not previously included in any other expression atlas.

[†4] Mycorrhizal fungi interaction experiment including *Rhizophagus irregularis* and *Gigaspora margerita* exudates.

RESULTS

Data sets available in PEATmoss

PEATmoss unifies in one single tool *P. patens* expression data from CombiMatrix and NimbleGen expression microarrays and RNA-seq, containing 109 experiments organized into nine data sets (Table 1). PEATmoss includes 35 experiments (replicate sets) in addition to the 74 experiments also available in the eFP Browser, Genevestigator, and Phytozome.

The data sets contain experiments from the ecotypes Gransden and Reute, and 17 different tissues including protoplasts and most of the *P. patens* life cycle developmental stages (Figure 1). Among these are dry spores, imbibed spores, germinating spores, protonema, caulonema, chloronema, juvenile gametophores, rhizoids, leaflets (phyllids), adult gametophores (including sexual organs), archegonia (female reproductive organs) and sporophyte development stages (Figure 1).

In PEATmoss, many experimental conditions can be found, such as hormone treatments (abscisic acid (ABA), auxin, GA9, strigolactone, OPDA), light perturbations (darkness, low light, high light, continuous light, UV-B light, blue light, red light, far red light, sunlight), abiotic stresses (PO_4 deficiency, ammonium, cold stress, heat stress, dehydration, rehydration, pH 4.5) and biotic stresses or biotic interactions (*Botrytis cinerea*, *Rhizophagus irregularis* exudate, *Gigaspora margerita* exudate).

Raw data from other expression experiments for *P. patens* are available in public repositories such as the Sequence Read Archive (SRA) (Leinonen *et al.*, 2011), but it is not the present goal of PEATmoss to include all of these. Experiments for multiple treatments, developmental stages, and media were selected, knock-out experiments were not considered. In the future more experiments including additional treatments or tissues will be included.

Expression data included in PEATmoss and not available in other tools

PEATmoss includes 35 expression experiments not available in any other expression atlas. RNA-seq experiments for developmental stages from Gransden and Reute ecotypes (Table 1, note [†1]) and for a phosphate deficiency time series were made available with PEATmoss (Table 1, note [†2]). For expression microarrays, gene expression data are included for mycorrhiza experiments (Table 1, note [†4]) as well as published data not available in any other expression tool (Table 1, note †3). These data are from the early response to ABA (Arif *et al*., 2019), response to red light experiments (Possart *et al*., 2017), response to cold stress (Beike *et al*., 2015) and from Reute ecotype gametophore and sporophyte developmental stages (Hiss *et al*., 2017).

Developmental stage experiments

Five experiments not available in other expression tools were included in the RNA-seq developmental stages data set (Table 1, note †1). Among these there are dry spores (two replicates) and imbibed spores (one replicate) from Gransden ecotype, and protonema in Knop liquid, juvenile gametophores and adult gametophores on Knop solid from Reute ecotype (three replicates each). The Reute ecotype has recently been introduced as a tool to study sexual reproduction, since many laboratories report fertility problems of their Gransden strains (Hiss *et al*., 2017; Perroud *et al*., 2019).

ABA early response experiment

A time series experiment to study early molecular response to the phytohormone ABA at 30, 60 and 180 min is included in PEATmoss (Arif *et al*., 2019). These experiments using the CombiMatrix expression microarray were generated to study cell wall thickening and related morphological changes in response to ABA (Arif *et al*., 2019). Of note, at 180 min the developmental decision to form brachycytes or brood cells (vegetative diaspores) has already been made. The ABA differentially expressed genes (DEGs) show a high level of overlap with those differentially expressed upon stresses such as UV-B, drought and cold (Arif *et al*., 2019).

Arbuscular mycorrhizal fungi (AMF) interaction experiments

Experiments for the AMF *Gigaspora margerita* and *Rhizophagus irregularis* (Hanke *et al.,* in preparation) are available exclusively in PEATmoss (Table 1, note †4). The co-evolution between AMF and bryophytes was probably instrumental in the transition of plants to land (Rensing, 2018), the symbiosis probably evolved in the last common ancestor of land plants. Although *P. patens* is not known to mutualistically interact with these fungi, it possesses the conserved signalling cascade (Delaux *et al*., 2015), which prompted to analyze the molecular response to fungal exudates. For this aim, experiments of a control exudate without AMF, with *R. irregularis* exudate after 1 and 24 h, and with *G. margerita* exudate after 24 h were generated using NimbleGen expression microarrays (Hanke *et al*., in preparation).

The Expression Atlas Web Tool

PEATmoss not only unifies the expression data from *P. patens* available in several other tools, but also provides the user with multiple visualization methods and other features that facilitate finding expression patterns in the data. PEATmoss is based on the Tomato Expression Atlas (Fernandez-Pozo *et al*., 2017), hence similar tools and features can be used. Three different input types are possible in PEATmoss: using a gene ID, a BLAST search (Altschul *et al*., 1997) or a list of genes. Additionally, for better visualization, experiments from the data sets can be filtered by the user to display only experiments for the selected treatments, ecotypes, media, stages or tissues. The expression colour scale can be changed to get higher resolution in specific ranges of expression, for example to better observe differences for genes in similar ranges of expression or for very low or very highly expressed genes where the default colour scale does not provide enough resolution. PEATmoss offers six visualization methods to explore expression data that are detailed below. Furthermore, PEATmoss is able to automatically convert the input gene name to the gene version needed to query any data set by connecting in the background to the *Physcomitrella patens* gene model lookup database (PpGML DB, see below).

PEATmoss expression cube

The main output view is a 3D cube where the expression of multiple genes can be compared simultaneously (Figure 2). This is an advantage over other expression visualization tools, since the commercial version of Genevestigator is the only one of the tools mentioned before where multiple gene comparison is possible and several visualization methods are available. The expression cube shows the expression for the query gene and co-expressed genes when searching by gene ID. In the case of using BLAST or the custom list, a selection of the best hits from BLAST or the list of genes provided in the custom list will be displayed as the layers of the cube, respectively.

PEATmoss bar plots

Clicking on a layer of the cube will open a bar plot with the expression values of the gene that the layer refers to (Figure 3). Several bar plots from multiple genes can be opened simultaneously to facilitate gene expression comparison. Bar plots contain standard error (SE) bars to show replicate variability from each experiment. Moreover, a transpose button allows the user to transpose the bar plots to change treatments and stages from the x-axis to coloured categories and *vice versa*. Gene names in the bar plots are linked to multiple annotations via the PpGML DB (see below).

PEATmoss expression data downloading

The button ‘Download Expression Data’ on the top-left of the expression cube (Figure 2) will create a tabular text file with the expression data from all the genes and experiments in the cube, including all co-expressed genes from all pages and correlation values and functional descriptions. This format can be easily imported in a spreadsheet

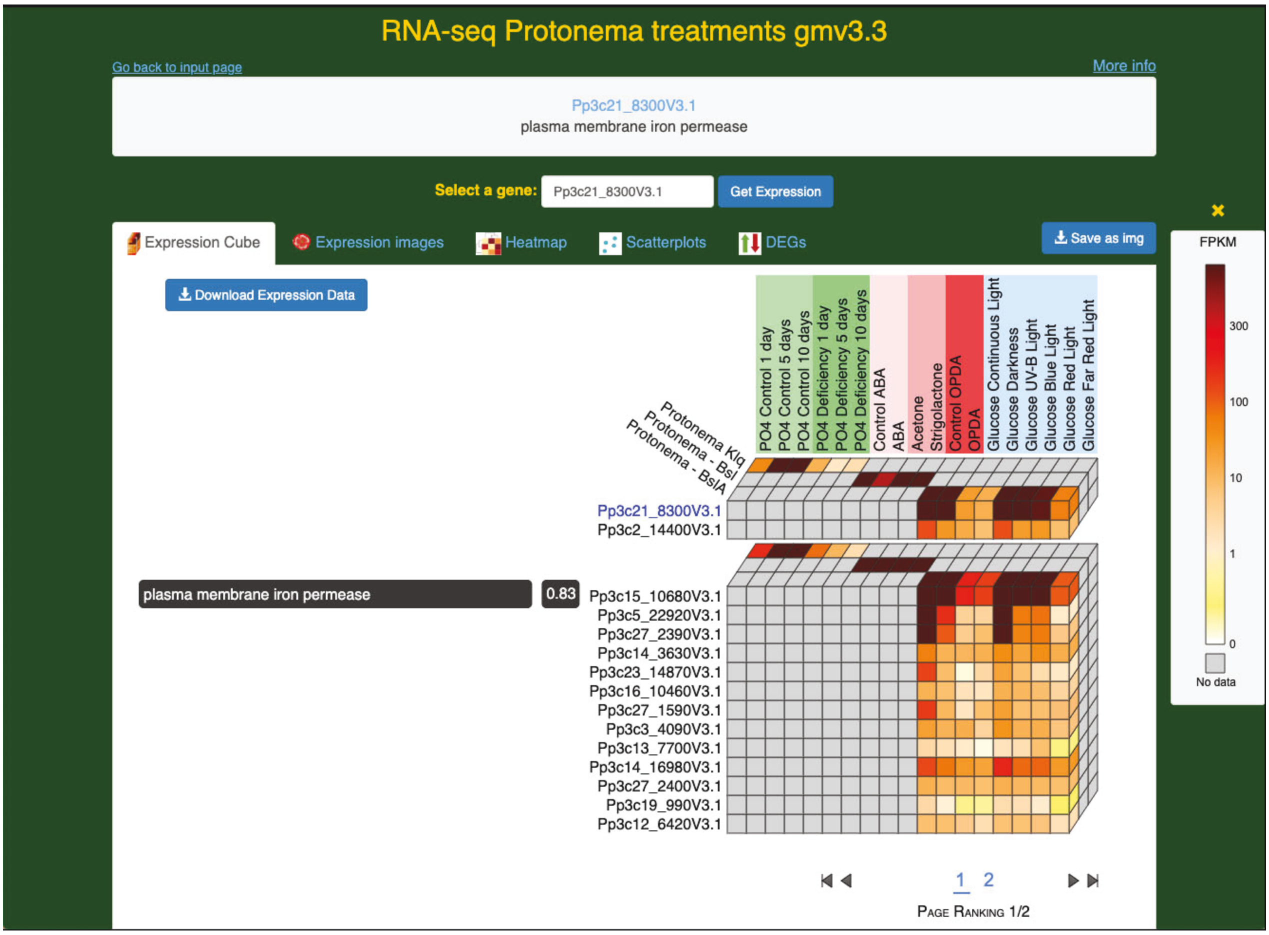


Figure 2. PEATmoss Expression Cube. On top of the cube the experiment treatments are displayed. On the diagonal top-left of the cube, experiment stages/tissues and media are displayed. On the left of the cube gene names are displayed. On top of these, in blue, is the query gene and below are all the co-expressed genes sorted by correlation value. The layers of the cube can be split and merged by clicking on gene names. The gene description and correlation value are displayed when moving the cursor over the gene name. At the bottom of the cube the pagination menu allows access to more co-expressed genes.

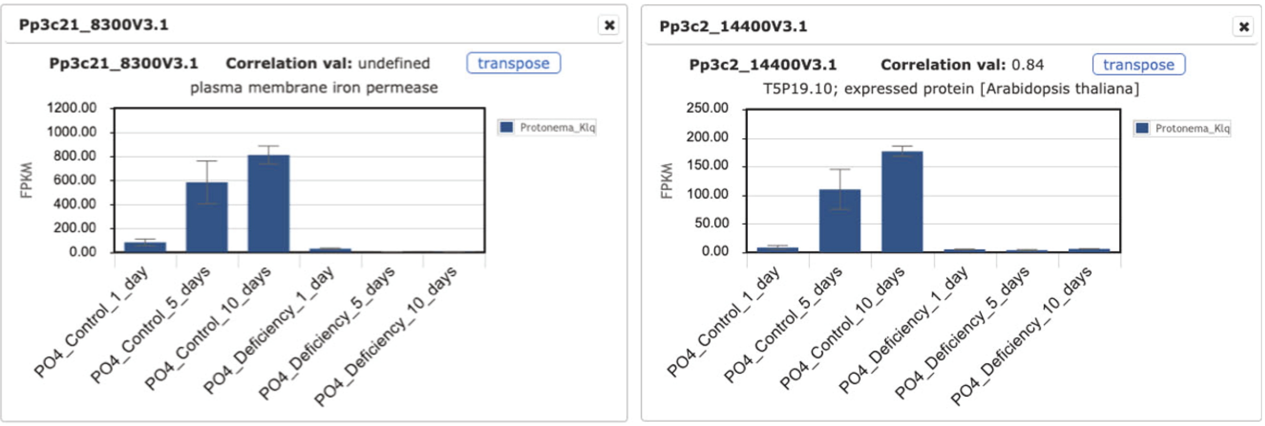


Figure 3. PEATmoss bar plots. Gene expression of gene Pp3c21_8300V3.1 on the left and the most correlated gene (0.84), with similar expression profile, Pp3c2_14400V3.1, on the right.

to visualize genes as rows and expression values from each experiment as columns and is easy to parse for bioinformatic analyses.

PEATmoss expression images

A fourth option to visualize gene expression is the expression images (Figure 4). As in the main visualization from the eFP Browser, each tissue from the data set is represented by a drawing and coloured based on the expression value of the query gene in that tissue. Expression images were designed for PEATmoss, to represent all included tissues and developmental stages of the *P. patens* life cycle (Figure 1).

The expression image system is very flexible when adding additional data sets, since the figures to represent expression data are reusable and independent for each tissue. Figures can represent the expression of one single tissue or can be formed by combination of several drawings to display the expression of multiple tissues in one single image. This system makes the addition of additional data straight forward.

PEATmoss heatmap

The hierarchical clustering heatmap will cluster all experiments of the genes from the first page of the cube, that is the query gene and the 14 co-expressed genes with highest correlation value (Figure 5). When accessing more pages from expression cube pagination menu, the next sets of 14 genes together with the query gene will be displayed. This output provides another way to visualize expression data showing patterns by clustering the most similar genes and experiments. Moving the cursor over the heatmap will show gene, experiment and expression value for each rectangle. Genes and/or experiments can be highlighted for a better visualization, and any region can be selected to zoom in to see it in more detail.

PEATmoss scatterplots

The ‘scatterplots’ feature enables users to visualize and compare the expression of all genes in the database for any two selected experiments (squares in blue in Figure 6a). The scatterplot’s x-axis and y-axis represent the gene expression levels in each of the two experiments, and each gene is plotted as a point in relation to these axes (Figure 6b). When plotting many genes, the density of points can make it difficult to clearly distinguish individual points. Here, the user has the ability to zoom in on a portion of the plot by clicking and dragging the mouse to highlight an area for closer examination (Figure 6c). Resetting the zoom is achieved by a single click anywhere on the plot. By hovering the mouse over an individual point, the user can see the gene ID and the exact expression value in each experiment (Figure 6d). This allows the user to readily identify, for example, genes that have high

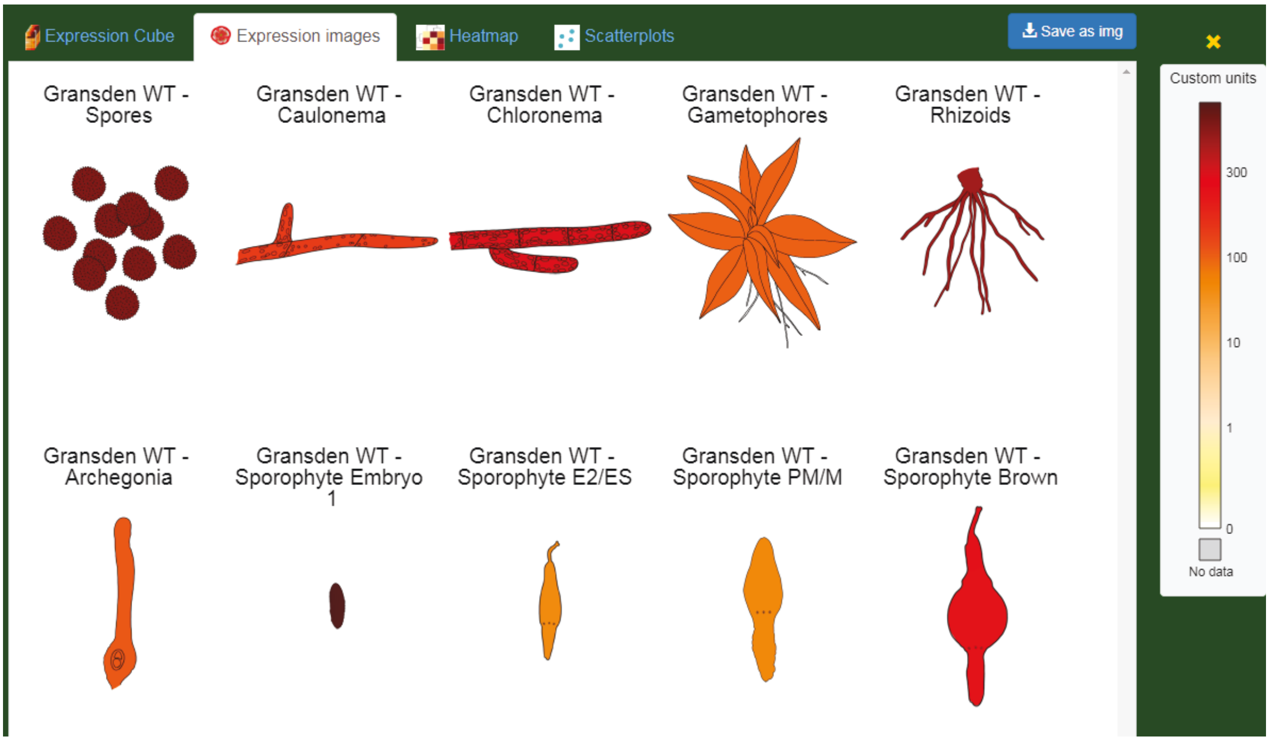


Figure 4. PEATmoss Expression images. Example from the data set NimbleGen Ortiz-Ramirez *et al.* (2016) showing different expression values represented by different colours in 10 tissues/developmental stages.

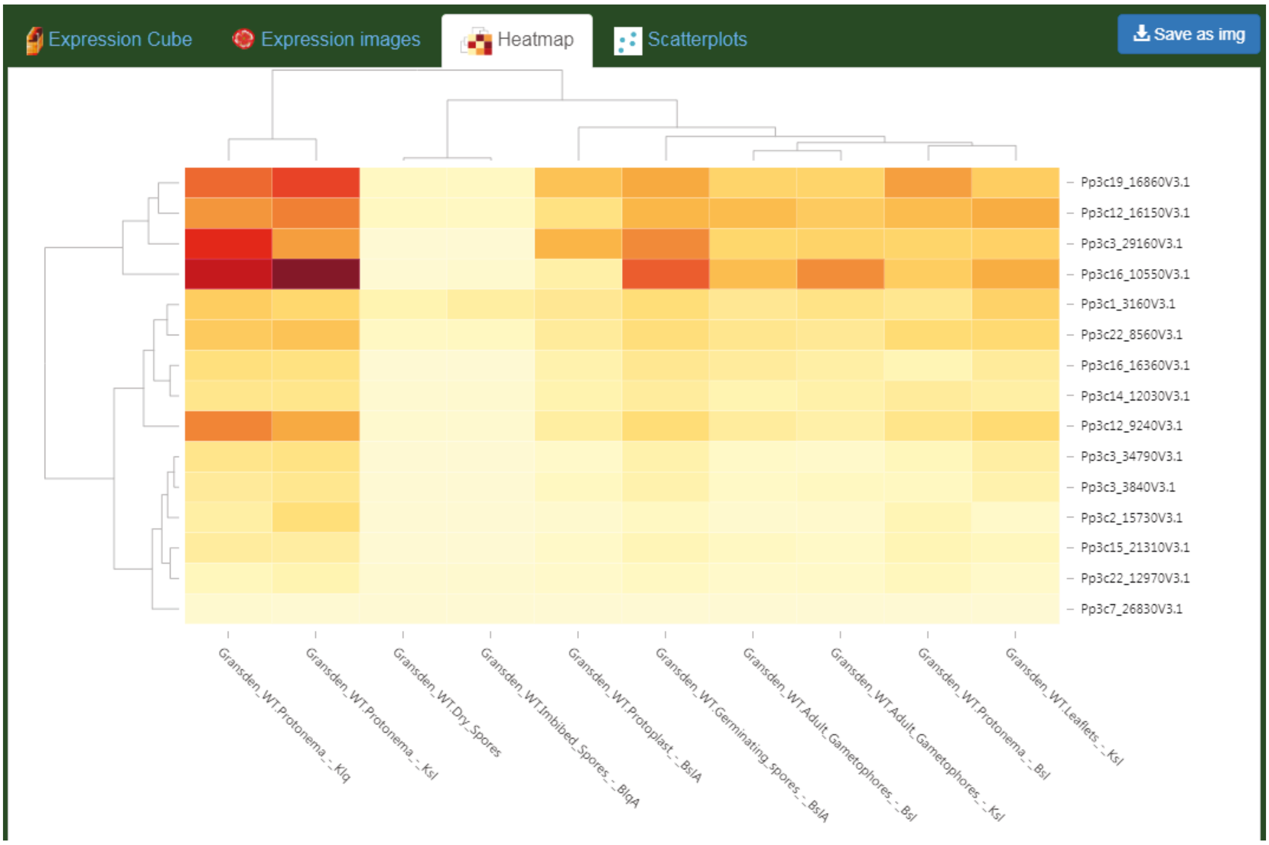


Figure 5. PEATmoss heatmap clustering. The query gene and the top 14 co-expressed genes are displayed on the right. Experiment names are listed at the bottom and hierarchical clustering trees are shown at the top, for experiments, and on the left for genes.

expression in one experiment and low expression in the other. Therefore, users can visually explore the relative expression of genes in any pair of experiments.

PEATmoss DEGs

Using the PEATmoss DEGs tab it is possible to select two experiments from a data set to calculate differentially expressed genes between them. To perform this statistical test, the R package NOISeq (Tarazona *et al.*, 2015) is used with a probability threshold of $q > 0.9$ and biological replicate normalized expression values; FPKM or microarray normalized expression data (stored in PEATmoss), as in Perroud *et al.* (2018), in which NOISeq best represented the overlap of DEGs from EdgeR (Robinson *et al.*, 2010), DESeq2 (Love *et al.*, 2014) and NOISeq. If DEGs are calculated, results are stored in PEATmoss to make them available for future queries, thereby it is possible to display results in a fast way for all comparisons previously calculated.

P. patens gene model lookup database (PpGML DB)

There are six microarray data sets available in PEATmoss, and some of them are also available in the eFP browser and Genevestigator. However, these data used probes based on gene models and gene names from previous genome annotation versions. The data from NimbleGen microarray technology are based on the annotation version 1.6 and the data from CombiMatrix microarray technology are based on version 1.2. When gene versions are different between expression data sets, it is hard to find information and make comparisons, for example, if we are interested in expression values from microarray platforms for a query gene from the current genome version and annotation, or vice versa. For that reason, PEATmoss is connected to the PpGML DB to automatically lookup genes from different versions. Any gene from any gene version can be used to query any of the PEATmoss data sets. The tool will search for the equivalent gene version needed for the selected data set in the PpGML DB and will display the results for the correct gene version. In case of multiple matches, a list of all the matched genes is displayed.

This database also enables the search for genes or annotations, conversion of gene IDs to multiple versions, and retrieval of annotation and sequences for a custom list of genes. Each gene contains descriptions and links for best BLAST hits in NCBI Nr (https://blast.ncbi.nlm.nih.gov/Blast.cgi), SwissProt (UniProt Consortium, 2018) and TAIR

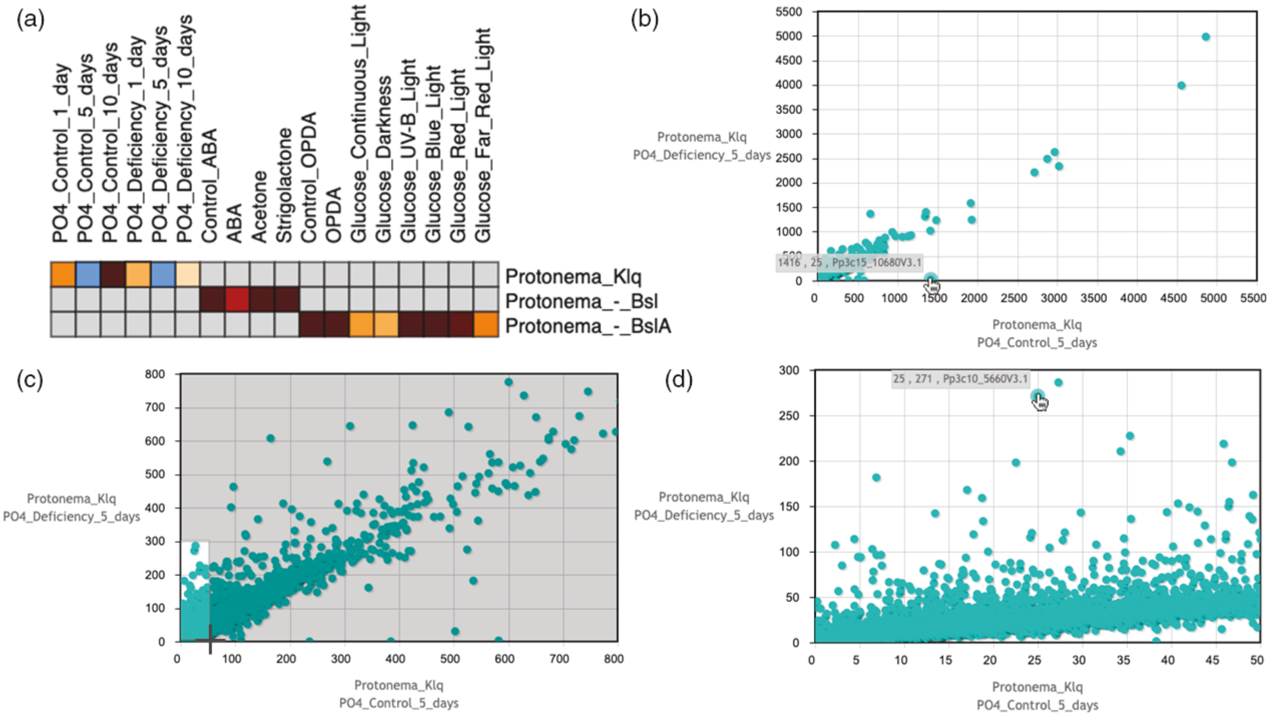


Figure 6. PEATmoss scatterplots. (a) Scatterplot experiment selection. (b) Scatterplot with gene expression values of all genes from two selected experiments. (c) An area of the scatterplot highlighted in order to zoom in. (d) Zoomed in view of area near the y-axis, showing expression values and name for one gene with high expression in one experiment and low expression in the other.

(Berardini *et al.*, 2015). Genes are also linked to their information in Phytozome (Goodstein *et al.*, 2012), PLAZA (Van Bel *et al.*, 2018), Ensembl Plants (Kersey *et al.*, 2016), CoGe (Lyons and Freeling, 2008), TAPscan (Wilhelmsson *et al.*, 2017), Phytozome and Ensembl genome browsers and the expression data in PEATmoss. In addition, many *P. patens* resources from multiple versions such as genome, protein, CDS and transcript sequences and GFF files are available for downloading at the PpGML DB. A BLAST tool is implemented linking the output to the gene annotation pages and with the possibility of downloading the results in BLAST tabular format.

PEATmoss customization

As mentioned before, PEATmoss is based on the Tomato Expression Atlas (TEA), so changes were made in the TEA source code, and they are applied in TEA, PEATmoss and all the tools based on TEA that regularly update the code from GitHub. For example, the application to calculate DEGs was developed for PEATmoss and it is available for TEA just updating the code. PEATmoss was customized to adapt TEA code and style to the needs of the *P. patens* data. The website contains detailed descriptions of the data sets with links to their publications and help pages to learn how to use the tool, including a set of videos about PEATmoss and the PpGML DB (https://peatmoss.online.uni-marburg.de/help).

Phosphate deficiency time series – an example data analysis based on PEATmoss

Phosphorus is a limiting nutrient for plants because of its low availability and mobility in soils (Abel *et al.*, 2002). Therefore, phosphate (PO_4) deficiency is a common abiotic stress in plants which contributes to reduced crop productivity (Schachtman *et al.*, 1998; Lynch and Brown, 2008) and alters global transcriptome profiles in vascular plants (Misson *et al.*, 2005; Zheng *et al.*, 2009; Takehisa *et al.*, 2013). Inorganic phosphate (Pi) deficiency has been intensively studied in seed plants unravelling genes that contribute to the Pi deficiency signalling cascade (Schachtman and Shin, 2007; Rubio *et al.*, 2009; Lin *et al.*, 2011). Here, we investigate Pi deficiency of filamentous, tip-growing protonemata across three time points [1, 5 and 10 days post transfer (dpt) into Pi-deficient medium] and analyze their global molecular responses with the PEATmoss software tool (Table 1, note [†2]).

First, DEGs were calculated in PEATmoss comparing Pi deficiency condition versus the corresponding control for the three time points. Subsequently, DEGs found were queried in the 'Find gene versions and annotations for a list of genes' function of the PpGML DB, selecting the check boxes for gene annotation version 3.3 and 'show annotations' (resulting in descriptions to the closest orthologues in SwissProt, Phytozome, NCBI Nr and TAIR being displayed and downloaded in one click). To facilitate discussions about gene function, annotations were added to the DEGs (Table S1).

The gene Pp3c21_8300V3.1, found as a DEG downregulated in all time points and annotated as a 'plasma membrane iron permease', was queried in PEATmoss (data set RNA-seq protonema treatments). In PEATmoss Cube (Figure 2), this gene is shown on the top, in blue. Below it, the top 14 co-expressed genes with most similar expression profiles are displayed. In total, 25 genes were found to be correlated to the query gene. There are 14 genes with correlation values over 0.75, of those 10 genes were DEGs in 5 and 10 days of Pi deficiency, demonstrating the usefulness of the co-expression data. Among the correlated genes (with similar expression profile to the query gene) many annotations related to the effects of Pi starvation were found. Some examples are a plasma membrane iron permease, a chlorophyll *a*–*b* binding protein of LHCII, a ferric reductase, and a glutaredoxin s17, which are related to functions altered in phosphate deficiency such as iron homeostasis, photosynthesis, redox balance and reactive oxygen species (ROS) (Hirsch *et al.*, 2006; Bournier *et al.*, 2013; Hernandez and Munne-Bosch, 2015; Carstensen *et al.*, 2018).

The top ranked gene Pp3c2_14400V3.1 (Figure 2), with a correlation value of 0.84, shows a very similar expression profile to the query gene (Figure 3). This gene is clearly downregulated under Pi deficiency conditions but there are no annotations available and its function is unknown, even though it shows similarities in many other species in NCBI nr database. This gene could be a good candidate for further studies to understand the function of genes related to Pi starvation and is a good example of how PEATmoss can assist researchers to visualize expression data for finding genes of interest when no annotations are available.

Looking at the 5 days Pi time point using the PEATmoss scatterplot function (Figure 6a), it is easy to spot some genes with high expression differences between treatment and control (Figure 6b). For example, with 1,416 FPKM in the control and 25 FPKM at 5 dpt, the gene Pp3c15_10680V3.1 (Figure 6b) is found, annotated as a plasma membrane iron permease, previously found as correlated with Pp3c21_8300V3.1 (Figure 2) and identified as a downregulated DEG at 5 dpt (Table S1). To find genes with high expression after 5 days of Pi deficiency and low expression in control conditions, the scatterplot was zoomed in to 50 × 300 (Figure 6c, d). Some examples found are Pp3c10_5660V3.1 (Figure 6d) and Pp3c12_20160V3.1, annotated as a LEA protein and an ABC transporter in the PpGML DB, and both identified as upregulated DEGs at 5 dpt (Table S1).

Looking more closely into the DEGs, Pi deficiency for one day (1 dpt) is apparently causing only minor stress to the plant, which differentially changes the expression of a few genes (11 were over-expressed and eight under-expressed: Table S1). Four DEGs at 1 dpt are annotated as transporters. One of them is upregulated (heavy-metal transporter, Pp3c25_11360V3.1) and three downregulated (iron transporter, Pp3c21_8300V3.1; inorganic phosphate transporter, Pp3c6_26510V3.1 and ABC transporter, Pp3c14_2470V3.1). These alterations in transporter transcription might prepare the plant for altered ion flux and to keep the homeostasis of the cells after the disequilibrium caused by the Pi deficiency. The upregulation of an alpha-amylase (Pp3c13_20160V3.1) might be linked to an intrinsic metabolic change to mobilize carbon from the starch reservoir to reduce the needs of carbon produced by photosynthesis. Gene Ontology (GO) bias analyses were performed to support the DEG results found in PEATmoss. This analyses showed 'amylase activity' as significant Molecular Function (MF) term, whereas the downregulation of the transporters led to a significant GO term enrichment in the GO domain MF for 'transmembrane transporter activity' (Table S2).

At 5 dpt the highest number of DEGs were found (337 up; 34 down; Table S1). The highly significant enriched GO terms (significance level < 0.0001) in the GO domain Biological Process (BP) indicate a severe stress ('response to stress', 'response to external stimulus'; Table S3). Multiple DEGs annotated as calmodulin and calcium binding proteins might be related to signalling as a response to the Pi deficiency stress, as observed in other abiotic stresses (Zeng *et al.*, 2015; Wang *et al.*, 2018). This signal could be associated with a mobilization of Pi from the vacuole (Liu *et al.*, 2011; Chien *et al.*, 2018; Xu *et al.*, 2019), and is supported by five DEGs associated with the vacuole (Table S1) and the upregulated cellular component (CC) GO term 'endomembrane system' (Table S3). The abundance of the GO term 'ion transport' for the up- and downregulated DEGs shows the plant adaptation to the altered ion status. One member of the PHO1-like proteins (Pp3c3_10280V3.1) was found among the upregulated genes, and has been shown earlier to be important for the adaptation to Pi deficiency (Wang *et al.*, 2008). Pp3c7_22280V3.1 (ADP-Glc pyrophosphorylase), an orthologue of *A. thaliana* APS1 reported to be important for the plant under Pi deficiency conditions (Wang *et al.*, 2008), might contribute to starch synthesis. Many transcription factors (TF) from the AP2 and RING finger families are among the DEGs at 5 dpt and might play a role in ubiquitination of proteins for degradation (Mizoi *et al.*, 2012; Rodriguez-Celma *et al.*, 2019). Many of these TFs contain GO terms associated with 'cell death' and 'immune response'. Many chaperones and LEA proteins are also activated (Table S1), which could be involved in the degradation and stability of proteins. Furthermore, many DEGs and GO terms are related to phospholipids and cell wall, which might indicate the degradation of multiple components of the cell to recycle phosphate and carbon (Liao *et al.*, 2011; Nakamura, 2013; Pant *et al.*, 2015;

Zhang *et al*., 2016). The significantly enriched GO terms for CC indicate that the ongoing adaptation at 5 dpt takes primarily place at the plasma membrane, the endomembrane and the Golgi apparatus (Table S3). Four ethylene responsive TF DEGs and GO terms associated with jasmonic acid (JA) and salicylic acid (SA) response were found to be upregulated at 5dpt. These plant hormones are known to be associated with phosphate deficiency and transport (Wang *et al*., 2014; Song and Liu, 2015; Khan *et al*., 2016).

After 10 days of Pi deficiency, the plant seems to be more adapted to the stress and the number of DEGs is reduced to 144 (70 up; 74 down; Table S1). Among the three time points, 5 dpt and 10 dpt share the highest number of upregulated and downregulated genes (Figure S1), which might indicate the slowdown of the adaptation process to the constant external stimulus of Pi deficiency. The same is true for the comparison of shared significantly enriched GO terms. As for dpt 5, at dpt 10 the enriched GO terms for BP indicate a response to external stimulus, whereas the terms ‘phosphate’ and ‘starvation’ now appear as individual categories. These findings are in accordance with previous results (Wang *et al*., 2008) that some genes can respond more slowly to Pi starvation conditions (See Supporting results 1). As compared with 5 dpt, at 10 dpt the GO CC terms show a shift towards the chloroplasts as the main actors for DEGs (Table S4). The biggest changes found for 10 dpt were an alteration in the photosynthesis and electron transport chain in the chloroplast (Tables S1 and S4), also observed in previous studies in other plants (Hernandez and Munne-Bosch, 2015; Zhang *et al*., 2016; Carstensen *et al*., 2018). The chloroplast as the predominant organelle for DEGs at 10 dpt is also reflected by the GO domain MF which highlights photosynthesis related categories such as ‘electron carrier activity’, ‘xanthophyll binding’ and ‘chlorophyll binding’. Many DEGs contain annotations related to the photosynthesis and the electron transport chain such as cytochrome b6, photosystems I and II proteins, chlorophyll A/B binding protein, ribulose-bisphosphate carboxylase, ribulose-phosphate 3-epimerase, photosynthetic NDH subcomplex B3, ascorbate ferrireductase, oxidase and peroxidase, ferredoxin (2Fe-2S), ferroxidase and ferritin (Table S1).

Iron and phosphate homeostasis are tightly connected (Hirsch *et al*., 2006). Iron plays an important role in many of the processes that appeared in the DEGs and GO results, such as photosynthesis, respiration, redox balance and ROS production (Bournier *et al*., 2013). In consequence, DEGs and GO terms related to keep the redox state of molecules and prevent damage from ROS were found (10 dpt significant GO term ‘response to hydrogen peroxide’; Table S4). At all time points, downregulated iron transporters (Table S1) and GO terms related to iron transport are found (Tables S2 and S4), especially at 5 and 10 dpt. To keep iron homeostasis and probably avoid the production of ROS that could be produced by an unbalanced accumulation of iron in the cell, it seems that iron transporter gene transcription is downregulated to cope with the decreasing internal Pi pool.

In summary, PEATmoss tools and the PpGML DB can be employed to find meaningful biological results. These results are supported by previous publications and can be very useful to explore expression data and support analyses, using the many available applications, such as downloading data, adding annotations, converting gene model versions or finding candidates of interest.

EXPERIMENTAL PROCEDURES

Data sets

Expression data from the three RNA-seq data sets available in PEATmoss were published in Perroud *et al.* (2018). DNA library preparation, cDNA sequencing and RNA-seq analysis for the not previously published RNA-seq experiments, the phosphate deficiency time series and the developmental stages, were carried out as in Perroud *et al.* (2018). The CombiMatrix expression data sets were published in Busch *et al.* (2013); Hiss *et al.* (2014); Beike *et al.* (2015); Possart *et al.* (2017) and Arif *et al.* (2019). The NimbleGen expression data sets were published in Hiss *et al.* (2017) and Ortiz-Ramirez *et al.* (2016). NimbleGen microarray data for the arbuscular mycorrhiza (AM) interactions (Hanke *et al*., in preparation) were processed as described in the next paragraphs.

AM interaction experiments

RNA was isolated from plant material using the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany) following the manufacturer’s instructions. RNA concentration and size distribution was tested on the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, USA) with the Agilent RNA 6000 Nano Kit to determine quantity and quality. About 200 ng of total RNA was reverse transcribed and amplified with the WTA Kit (Sigma-Aldrich, St. Louis, CO, USA). Here, 1 g cDNA was labelled with Cy3 according to the NimbleGen One-Colour DNA Labelling Kit (Roche, Basel, Switzerland). Concentration and quality of the labelled cDNA was monitored. 4 μg of labelled cDNA were used for the hybridization on a NimbleGen 12 × 135 k DNA microarray, probe design OID33087 (Roche, Basel, Switzerland) according to the manufacturer’s protocol using the NimbleGen Hybridization Kit (Roche, Basel, Switzerland). 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, Santa Clara, CA, USA). The image of the arrays was cut into single array images using the NimbleScan Software 2.5 (Roche, Basel, Switzerland) and the pixel intensities were extracted with the same software. Microarray expression data were analyzed with Analyst 7.5 (Genedata, Basel, Switzerland). Data were processed as in Hiss *et al.* (2014).

Pi deficiency time series experiments

Protonemata were pre-cultured in liquid medium as described earlier (Perroud *et al*., 2018) and transferred to either Pi-deficient liquid growth medium (K_2SO_4 was used instead of KH_2PO_4) as described in Wang *et al.* (2008) or cultivated in liquid Knop medium (Knop, 1868). Three replicates of protonemata were harvested 1 day post transfer (1 dpt), 5 days (5 dpt) and 10 days (10 dpt) and RNA was extracted. Correlation matrices of the replicates from the

PO_4 deficiency experiment are shown in Figure S4. RNA extraction, DNA library preparation, cDNA sequencing and RNA-seq processing and normalization were done as in Perroud *et al.* (2018) with DEG calculation performed in PEATmoss.

GO analysis was conducted in Cytoscape v3.5 with the BiNGO plugin (Maere *et al.*, 2005) using gene annotations for *P. patens* v3.3. To find significantly over-represented or under-represented GO categories, hypergeometric statistical tests were applied at a significance level of 0.05 after correcting for multiple testing according to (Benjamini and Hochberg, 1995).

Data processing and normalization

Experiments previously published were not processed and replication was not evaluated. For RNA-seq experiments not published before, they were analyzed as in Perroud *et al.* (2018). Experiment replicates were checked by PCA, hierarchical clustering, and correlation of the replicates. No minimum expression value was used to filter the data, so users can set their own cut-off values. Experiments of expression microarrays not published before were analyzed as in Hiss *et al.* (2014) for CombiMatrix and as in Hiss *et al.* (2017) for NimbleGen. Background was subtracted and replicates were examined using hierarchical clustering and correlation matrices, removing in some cases replicates not clustering properly or with low correlation in comparison with the other replicates from the same experiment. NimbleGen data set from Ortiz-Ramirez *et al.* (2016) were included as in the publication. Published data from expression microarrays were downloaded from ArrayExpress (Kolesnikov *et al.*, 2015).

All experiments from CombiMatrix and NimbleGen were normalized using Genedata Analyst v. 9.5.2 (https://www.genedata.com/products/analyst/) to adjust the distribution to a median value of 12. This is the median value found for the distribution of all RNA-seq experiments from Perroud *et al.* (2018), which adapts to a similar scale range and makes data comparison in the visualization tools easier.

The Expression Atlas

PEATmoss is based on the TEA (Fernandez-Pozo *et al.*, 2017), a tool developed and hosted at the Sol Genomics Network (Fernandez-Pozo *et al.*, 2015). The code was cloned from the Solgenomics account in GitHub (https://github.com/solgenomics/Tea). Perl scripts included with the tool were used to format and import the data to the database and expression indexes. The code of TEA, PEATmoss, and other tools based on TEA is in continuous development. The code used to customize PEATmoss style and page structure is available on GitHub (https://github.com/noefp/ppatens_expr). Functional annotations for the *P. patens* genes displayed in PEATmoss were obtained from Cosmoss (Zimmer *et al.*, 2013), genes with unknown description were then subsequently annotated with SwissProt, TAIR, or the NCBI Nr database following that order. The heatmap output is implemented using the R package d3heatmap. It uses hierarchical clustering complete agglomeration method and euclidean distances.

P. patens gene model lookup DB (PpGML DB)

The GML DB is implemented as a Postgres v10 relational database with a very simple schema (Figure S3) to store gene names, versions and annotations, and the relation between them. The module pg_trgm is used to support non-exact text search based on trigram matching. The website is written in PHP (v7.0.30-0 + deb9u1) and uses Bootstrap 3 libraries (https://getbootstrap.com/) for style and interactive elements. JQuery DataTables (https://datatables.net/) are used to display search outputs and tables with gene annotations. Gene version and annotations stored in the database were extracted from a modified version of the lookup table (Data S1) from Perroud *et al.* (2018). This lookup table was based on the Cosmoss annotation (Zimmer *et al.*, 2013) and was improved using the GMAP version 2015-12-31 (Wu and Nacu, 2010) to map nucleotide transcript sequences against the *P. patens* V3 genome. The intersection with the *P. patens* gene model v3.3 was obtained using bedtools intersect version 2.26.0 (Quinlan and Hall, 2010). BLAST+ (Camacho *et al.*, 2009) and the blastdbcmd script are integrated in the website to search sequences by similarity and to extract list of sequences from BLAST databases. BLAST output interface was modified from the code from the Sol Genomics Network (Fernandez-Pozo *et al.*, 2015) (https://github.com/solgenomics/sgn).

ACCESSION NUMBERS

Raw data for the RNA-seq experiments included in PEATmoss and not available in other tools are stored at the SRA (Leinonen *et al.*, 2011) with the library names BBTWT, BBTWU, BBTWS, BBTWW, BBTWY, BBTWX, BAOBH, BAOBP, BAOBS, BBTXA, BBTXB, BBTWZ, BAOBT, BAOBN, BBTXG, BAOAW, BAOAX, BAOBA, BAOAZ, BAOBC, BAOAU, BAOBG, BAOBU, BAOBO, BAOBB, BAOAY, BAOBW, BAOAT, BBTXC, and BBTWO. ArrayExpress accession numbers for the CombiMatrix data are E-MTAB-2165 (Beike *et al.*, 2015), E-MTAB-913, E-MTAB-914, E-MTAB-915, E-MTAB-916, E-MTAB-917, E-MTAB-919, E-MTAB-976, E-MEXP-2508 (Hiss *et al.*, 2014), and E-MTAB-2227 (Possart *et al.*, 2017). ArrayExpress accession numbers for the NimbleGen data are E-MTAB-3069 (Ortiz-Ramirez *et al.*, 2016) and E-MTAB-4630 (Hiss *et al.*, 2017). NimbleGen microarray data from the arbuscular mycorrhiza (AM) interaction experiments (Hanke *et al.*, in preparation) can be found in ArrayExpress with the accession number E-MTAB-3081.

ACKNOWLEDGEMENTS

The authors acknowledge the support of the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen, the state of Baden-Württemberg, Germany, through bwHPC and the German Research Foundation (DFG) through grant number INST 37/935-1 FUGG. The authors also acknowledge the support of the DFG through grant number DFG RE 1697/6-1. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the US Department of Energy under contract number DE-AC02-05CH11231. The authors declare no conflicts of interest.

AUTHORS' CONTRIBUTION

NFP, KU and SAR wrote the manuscript with help from all authors. NFP implemented and customized the tool, formatted the data and imported the data to index files and the database. FH calculated FPKMs for all RNA-seq experiments and created the lookup table. MH conducted the microarray experiments and assisted with microarray metadata and data processing. RM, KU, SH, JC, and EV conducted RNA-seq experiments. RM drew the plant figures and organized sporophyte development experiments metadata. PFP organized and supervised experiment metadata. NFP and VK developed the PpGML DB. AP and LAM developed the scatterplot application in PEATmoss. NFP and FH set up the PEATmoss server and SAR supervised the whole project.

SUPPORTING INFORMATION

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

Figure S1. Common differentially expressed genes among three Pi-deficient time points.

Figure S2. Protonemata expression values for known phosphate deficiency responsive genes.

Figure S3. PpGML DB schema.

Figure S4. Correlation matrices from the phosphate deficiency experiment.

Table S1. PEATmoss DEG output for each phosphate deficiency time point.

Table S2. Gene ontology analysis for Pi deficiency time point dpt 1.

Table S3. Gene ontology analysis for Pi deficiency time point dpt 5.

Table S4. Gene ontology analysis for Pi deficiency time point dpt 10.

Table S5. Expression of known Pi deficiency candidate genes (FPKM).

Supporting results 1

REFERENCES

Abel, S., Ticconi, C.A. and Delatorre, C.A. (2002) Phosphate sensing in higher plants. *Physiol. Plant*, **115**, 1–8.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

Arif, M.A., Hiss, M., Tomek, M., Busch, H.S., Tintelnot, S., Meyberg, R., Reski, R., Rensing, S.A. and Frank, W. (2019) ABA-induced vegetative diaspore formation in *Physcomitrella* patens. *Front. Plant Sci.* **10**, 315.

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.

Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van de Peer, Y., Coppens, F. and Vandepoele, K. (2018) PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res.* **46**, D1190–D1196.

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

Berardini, T.Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E. and Huala, E. (2015) The Arabidopsis information resource: making and mining the 'gold standard' annotated reference plant genome. *Genesis*, **53**, 474–485.

Bournier, M., Tissot, N., Mari, S., Boucherez, J., Lacombe, E., Briat, J.F. and Gaymard, F. (2013) Arabidopsis ferritin 1 (AtFer1) gene regulation by the phosphate starvation response 1 (AtPHR1) transcription factor reveals a direct molecular link between iron and phosphate homeostasis. *J. Biol. Chem.* **288**, 22670–22680.

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.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421.

Carstensen, A., Herdean, A., Schmidt, S.B., Sharma, A., Spetea, C., Pribil, M. and Husted, S. (2018) The impacts of phosphorus deficiency on the photosynthetic electron transport chain. *Plant Physiol.* **177**, 271–284.

Chien, P.S., Chiang, C.P., Leong, S.J. and Chiou, T.J. (2018) Sensing and signaling of phosphate starvation - from local to long distance. *Plant Cell Physiol.* **59**, 1714–1722.

Delaux, P.M., Radhakrishnan, G.V., Jayaraman, D. *et al.* (2015) Algal ancestor of land plants was preadapted for symbiosis. *Proc. Natl. Acad. Sci. USA*, **112**, 13390–13395.

Fernandez-Pozo, N., Menda, N., Edwards, J.D. *et al.* (2015) The Sol Genomics Network (SGN)–from genotype to phenotype to breeding. *Nucleic Acids Res.* **43**, D1036–1041.

Fernandez-Pozo, N., Zheng, Y., Snyder, S.I., Nicolas, P., Shinozaki, Y., Fei, Z., Catala, C., Giovannoni, J.J., Rose, J.K.C. and Mueller, L.A. (2017) The Tomato Expression Atlas. *Bioinformatics*, **33**, 2397–2398.

Goodstein, D.M., Shu, S., Howson, R. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–1186.

Hernandez, I. and Munne-Bosch, S. (2015) Linking phosphorus availability with photo-oxidative stress in plants. *J. Exp. Bot.* **66**, 2889–2900.

Hirsch, J., Marin, E., Floriani, M., Chiarenza, S., Richaud, P., Nussaume, L. and Thibaud, M.C. (2006) Phosphate deficiency promotes modification of iron distribution in Arabidopsis plants. *Biochimie*, **88**, 1767–1771.

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.

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.

Hruz, T., Laule, O., Szabo, G., Wessendorp, F., Bleuler, S., Oertle, L., Widmayer, P., Gruissem, W. and Zimmermann, P. (2008) Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. *Adv. Bioinfor.* **3**, 1–5.

Kersey, P.J., Allen, J.E., Armean, I. *et al.* (2016) Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res.* **44**, D574–580.

Khan, G.A., Vogiatzaki, E., Glauser, G. and Poirier, Y. (2016) Phosphate deficiency induces the Jasmonate Pathway and enhances resistance to insect Herbivory. *Plant Physiol.* **171**, 632–644.

Knop, W. (1868) *Der Kreislauf des Stoffs*. Рипол Классик.

Kolesnikov, N., Hastings, E., Keays, M. *et al.* (2015) ArrayExpress update–simplifying data submissions. *Nucleic Acids Res.* **43**, D1113–1116.

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.

Leinonen, R., Sugawara, H. and Shumway, M. (2011) The sequence read archive. *Nucleic Acids Res.* **39**, D19–21.

Liao, Y.Y., Buckhout, T.J. and Schmidt, W. (2011) Phosphate deficiency-induced cell wall remodeling: linking gene networks with polysaccharide meshworks. *Plant Signal. Behav.* **6**, 700–702.

Lin, W.D., Liao, Y.Y., Yang, T.J., Pan, C.Y., Buckhout, T.J. and Schmidt, W. (2011) Coexpression-based clustering of Arabidopsis root genes predicts functional modules in early phosphate deficiency signaling. *Plant Physiol.* **155**, 1383–1402.

Liu, T.Y., Aung, K., Tseng, C.Y., Chang, T.Y., Chen, Y.S. and Chiou, T.J. (2011) Vacuolar Ca2+/H+ transport activity is required for systemic phosphate homeostasis involving shoot-to-root signaling in Arabidopsis. *Plant Physiol.* **156**, 1176–1189.

Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550.

Lynch, J.P. and Brown, K.M. (2008) Root strategies for phosphorus acquisition. In *The Ecophysiology of Plant-Phosphorus Interactions* (White, P.J. and Hammond, J.P. eds). Dordrecht: Springer, Netherlands, pp. 83–116.

Lyons, E. and Freeling, M. (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* **53**, 661–673.

Maere, S., Heymans, K. and Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.

Misson, J., Raghothama, K.G., Jain, A. *et al.* (2005) A genome-wide transcriptional analysis using Arabidopsis thaliana Affymetrix gene chips determined plant responses to phosphate deprivation. *Proc. Natl. Acad. Sci. USA*, **102**, 11934–11939.

Mizoi, J., Shinozaki, K. and Yamaguchi-Shinozaki, K. (2012) AP2/ERF family transcription factors in plant abiotic stress responses. *Biochim. Biophys. Acta.* **1819**, 86–96.

Nakamura, Y. (2013) Phosphate starvation and membrane lipid remodeling in seed plants. *Prog. Lipid Res.* **52**, 43–50.

Ortiz-Ramirez, C., Hernandez-Coronado, M., Thamm, A., Catarino, B., Wang, M., Dolan, L., Feijo, J.A. and Becker, J.D. (2016) A Transcriptome Atlas of *Physcomitrella* patens Provides Insights into the Evolution and Development of Land Plants. *Mol. Plant*, **9**, 205–220.

Pant, B.D., Burgos, A., Pant, P., Cuadros-Inostroza, A., Willmitzer, L. and Scheible, W.R. (2015) The transcription factor PHR1 regulates lipid remodeling and triacylglycerol accumulation in Arabidopsis thaliana during phosphorus starvation. *J. Exp. Bot.* **66**, 1907–1918.

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.

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.

Possart, A., Xu, T., Paik, I. *et al*. (2017) Characterization of phytochrome interacting factors from the Moss *Physcomitrella* patens illustrates conservation of phytochrome signaling modules in land plants. *Plant Cell*, **29**, 310–330.

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

Rensing, S.A. (2017) Why we need more non-seed plant models. *New Phytol.* **216**, 355–360.

Rensing, S.A. (2018) Great moments in evolution: the conquest of land by plants. *Curr. Opin. Plant Biol.* **42**, 49–54.

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.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Rodriguez-Celma, J., Chou, H., Kobayashi, T., Long, T.A. and Balk, J. (2019) Hemerythrin E3 ubiquitin ligases as negative regulators of iron homeostasis in plants. *Front. Plant Sci.* **10**, 98.

Rubio, V., Bustos, R., Irigoyen, M.L., Cardona-Lopez, X., Rojas-Triana, M. and Paz-Ares, J. (2009) Plant hormones and nutrient signaling. *Plant Mol. Biol.* **69**, 361–373.

Ruprecht, C., Proost, S., Hernandez-Coronado, M., Ortiz-Ramirez, C., Lang, D., Rensing, S.A., Becker, J.D., Vandepoele, K. and Mutwil, M. (2017) Phylogenomic analysis of gene co-expression networks reveals the evolution of functional modules. *Plant J.* **90**, 447–465.

Schachtman, D.P. and Shin, R. (2007) Nutrient sensing and signaling: NPKS. *Annu. Rev. Plant Biol.* **58**, 47–69.

Schachtman, D.P., Reid, R.J. and Ayling, S.M. (1998) Phosphorus uptake by plants: from soil to cell. *Plant Physiol.* **116**, 447–453.

Song, L. and Liu, D. (2015) Ethylene and plant responses to phosphate deficiency. *Front. Plant Sci.* **6**, 796.

Takehisa, H., Sato, Y., Antonio, B.A. and Nagamura, Y. (2013) Global transcriptome profile of rice root in response to essential macronutrient deficiency. *Plant Signal. Behav.* **8**, e24409.

Tarazona, S., Furio-Tari, P., Turra, D., Pietro, A.D., Nueda, M.J., Ferrer, A. and Conesa, A. (2015) Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res.* **43**, e140.

UniProt Consortium, T. (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res*, **46**, 2699.

Wang, Y., Secco, D. and Poirier, Y. (2008) Characterization of the PHO1 gene family and the responses to phosphate deficiency of *Physcomitrella* patens. *Plant Physiol.* **146**, 646–656.

Wang, G., Zhang, C., Battle, S. and Lu, H. (2014) The phosphate transporter PHT4;1 is a salicylic acid regulator likely controlled by the circadian clock protein CCA1. *Front. Plant Sci.* **5**, 701.

Wang, X., Hao, L., Zhu, B. and Jiang, Z. (2018) Plant Calcium Signaling in Response to Potassium Deficiency. *Int. J. Mol. Sci.* **19**.

Wilhelmsson, P.K.I., Muhlich, 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 Biol. Evol.* **9**, 3384–3397.

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G.V. and Provart, N.J. (2007) An 'Electronic Fluorescent Pictograph' browser for exploring and analyzing large-scale biological data sets. *PLoS ONE*, **2**, e718.

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

Xu, L., Zhao, H., Wan, R. *et al*. (2019) Identification of vacuolar phosphate efflux transporters in land plants. *Nat. Plants*, **5**, 84–94.

Zeng, H., Xu, L., Singh, A., Wang, H., Du, L. and Poovaiah, B.W. (2015) Involvement of calmodulin and calmodulin-like proteins in plant responses to abiotic stresses. *Front. Plant Sci.* **6**, 600.

Zhang, K., Liu, H., Song, J., Wu, W., Li, K. and Zhang, J. (2016) Physiological and comparative proteome analyses reveal low-phosphate tolerance and enhanced photosynthesis in a maize mutant owing to reinforced inorganic phosphate recycling. *BMC Plant Biol.* **16**, 129.

Zheng, L., Huang, F., Narsai, R. *et al*. (2009) Physiological and transcriptome analysis of iron and phosphorus interaction in rice seedlings. *Plant Physiol.* **151**, 262–274.

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.

5.4 Single nucleotide polymorphism charting of *P. patens* reveals accumulation of somatic mutations during *in vitro* culture on the scale of natural variation by selfing

Genetic variation between different *P. patens* accessions was demonstrated multiple times (O'Donoghue *et al.*, 2013, Stevenson *et al.*, 2016, Hiss *et al.*, 2017, Lang *et al.*, 2018). In the previous studies, microarray and gDNA samples were used. A novel approach in analysing *P. patens* was employed here and is based on SNP calling by transcriptomic datasets. The higher availability compared to microarray or gDNA data increases the interest of RNA-seq data utilities. In this study we called SNPs of five different *P. patens* accessions, Gransden, Kaskaskia, Reute, Villersexel, and Wisconsin. Publicly available datasets covered the accessions Gransden and Reute. Novel RNA-seq data was used for Kaskaskia, Villersexel, new gDNA samples for Wisconsin. Data pre-processing was done by an updated version of the pipeline published by (Perroud *et al.*, 2018). To detect SNPs, an additional branch for this pipeline was developed (Figure 6 C). Furthermore, post-processing steps such as filtering and clustering SNPs were added. We identified exclusive SNPs for each accession. For easy identification, we designed primers for the RFLP method (Paper 5.4, page 12). This can be used by the community to determine the origin of their plants. In addition, 13 different Gransden lineages were identified (Paper 5.4, page 12). They were grouped into four different Gransden pedigrees. Although all Gransden plants can be backtracked to one single spore isolate, genetic variation between the Gransden pedigrees could be detected. As for the accessions, exclusive SNPs for each Gransden pedigree were extracted and single ones verified by PCR. It could be shown that somatic mutation occurs and accumulates in accessions and Gransden pedigrees. For pedigrees with known propagation history, an annual number of mutations per base pair was estimated (Paper 5.4, page 12f). *P. patens* accession Wisconsin gDNA was used to analyse sequence variation within natural populations. This results were compared to the variation detected in laboratory cultures (Paper 5.4, page 13).

We assembled information to generate a Gransden pedigree. Within this Gransden pedigree somatic mutations occur and accumulate. Additionally, I demonstrated that SNP calling based on RNA-seq data is efficient and shows feasible results. These results were performed by a novel designed powerful RNA-seq pipeline extension for variant calling.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 07 July 2020
doi: 10.3389/fpls.2020.00813

Single Nucleotide Polymorphism Charting of *P. patens* Reveals Accumulation of Somatic Mutations During *in vitro* Culture on the Scale of Natural Variation by Selfing

Fabian B. Haas[1], Noe Fernandez-Pozo[1], Rabea Meyberg[1], Pierre-François Perroud[1], Marco Göttig[1], Nora Stingl[1], Denis Saint-Marcoux[2,3], Jane A. Langdale[2] and Stefan A. Rensing[1,4,5]*

[1] Plant Cell Biology, Department of Biology, University of Marburg, Marburg, Germany, [2] Department of Plant Sciences, University of Oxford, Oxford, United Kingdom, [3] Université de Lyon, UJM-Saint-Etienne, CNRS, Laboratoire BVpam - FRE 3727, Saint-Étienne, France, [4] BIOSS Centre for Biological Signalling Studies, University of Freiburg, Freiburg, Germany, [5] SYNMIKRO Center for Synthetic Microbiology, University of Marburg, Marburg, Germany

Introduction: *Physcomitrium patens* (Hedw.) Mitten (previously known as *Physcomitrella patens*) was collected by H.L.K. Whitehouse in Gransden Wood (Huntingdonshire, United Kingdom) in 1962 and distributed across the globe starting in 1974. Hence, the Gransden accession has been cultured *in vitro* in laboratories for half a century. Today, there are more than 13 different pedigrees derived from the original accession. Additionally, accessions from other sites worldwide were collected during the last decades.

Methods and Results: In this study, 250 high throughput RNA sequencing (RNA-seq) samples and 25 gDNA samples were used to detect single nucleotide polymorphisms (SNPs). Analyses were performed using five different *P. patens* accessions and 13 different Gransden pedigrees. SNPs were overlaid with metadata and known phenotypic variations. Unique SNPs defining Gransden pedigrees and accessions were identified and experimentally confirmed. They can be successfully employed for PCR-based identification.

Conclusion: We show independent mutations in different Gransden laboratory pedigrees, demonstrating that somatic mutations occur and accumulate during *in vitro* culture. The frequency of such mutations is similar to those observed in naturally occurring populations. We present evidence that vegetative propagation leads to accumulation of deleterious mutations, and that sexual reproduction purges those. Unique SNP sets for five different *P. patens* accessions were isolated and can be used to determine individual accessions as well as Gransden pedigrees. Based on that, laboratory methods to easily determine *P. patens* accessions and Gransden pedigrees are presented.

Keywords: SNP, RNA-seq, *Physcomitrella patens*, *Physcomitrium*, ecotype, Gransden, Reute, RFLP

OPEN ACCESS

Edited by:
Michael R. McKain,
The University of Alabama,
United States

Reviewed by:
Matthew Johnson,
Texas Tech University, United States
Sara V. Good,
The University of Winnipeg, Canada

****Correspondence:***
Stefan A. Rensing
stefan.rensing@
biologie.uni-marburg.de

Specialty section:
This article was submitted to
Plant Systematics and Evolution,
a section of the journal
Frontiers in Plant Science

Received: *19 February 2020*
Accepted: *20 May 2020*
Published: *07 July 2020*

Citation:
Haas FB, Fernandez-Pozo N,
Meyberg R, Perroud P-F, Göttig M,
Stingl N, Saint-Marcoux D,
Langdale JA and Rensing SA (2020)
Single Nucleotide Polymorphism
Charting of P. patens Reveals
Accumulation of Somatic Mutations
During in vitro Culture on the Scale
of Natural Variation by Selfing.
Front. Plant Sci. 11:813.
doi: 10.3389/fpls.2020.00813

INTRODUCTION

Single Nucleotide Polymorphisms (SNPs) represent a major source of natural variation within any given species. In the plant kingdom, they are studied both in ecological and evolutionary context in order to understand population structure (Leaché and Oaks, 2017). They are also employed to study the genetic basis of variable natural traits such as resistance to flooding (Vashisht et al., 2011), or for the identification of genetic diversity in cultivars and admixed wild types through association mapping (Niu et al., 2019). SNP analysis is now successfully integrated in plant breeding for example in palm tree selection (Xia et al., 2019). For the moss model *Physcomitrium patens* (Hedw.) Mitten (previously known as *Physcomitrella patens*) (Beike et al., 2014; Medina et al., 2019; Rensing et al., 2020) whole genome SNP sets between the reference genome accession, Gransden (Gd) (Rensing et al., 2008; Lang et al., 2018), and the accessions Villersexel (Vx) (Kamisugi et al., 2008), Reute (Re) (Hiss et al., 2017) and Kaskaskia (Ka) (Perroud et al., 2011) have been reported (Hiss et al., 2017). Specifically, the genetic difference between Gd and Vx has been used to generate the first sequence-anchored genetic linkage map (Kamisugi et al., 2008) and recently the *P. patens* chromosome level genome assembly (Lang et al., 2018). Analysis of SNP segregation is a powerful tool that can be employed to analyze intra and inter accession fertility (Perroud et al., 2011, 2019; Meyberg et al., 2020), gene specific segregation patterns, and loci affected in segregants with specific traits. For example, the analysis of Gd and Vx segregants has been used to identify the *ANR* locus affected in mutants impaired in ABA hormone signaling (Stevenson et al., 2016), as well as loci involved in three-dimensional morphogenesis [*nog*1, (Moody et al., 2018)] and a novel microtubule depolymerizing-end-tracking protein [*CLoG*1, (Ding et al., 2018)]. Most recently, SNPs between Gd and Re were associated with the loss of fertility in the Gd background (Meyberg et al., 2020). However, there is no comparative study on a broad set of accessions, or within the different *P. patens* Gransden laboratory strains (Gd pedigrees).

Model organisms cultivated in the laboratory are usually considered to be genetically uniform due to their common origin. The original *P. patens* Gransden plant was collected by H.L.K. Whitehouse in Gransden Wood (Huntingdonshire, United Kingdom) in 1962. Engel cultured Whitehouse's sample (Engel, 1968) and derived the ancestor of all current *P. patens* Gransden strains from a single spore. In 1974 progeny of *P. patens* Gransden started to be distributed across the globe (Ashton and Cove, 1977; Cove, 2005). Since then, *P. patens* became an important model organism *inter alia* to study cell biology, evolutionary developmental biology and the water to land transition of plant life (Rensing, 2018; de Vries and Rensing, 2020). During its decades of *in vitro* cultivation, *P. patens* Gransden was predominantly propagated vegetatively (Ashton and Raju, 2000). While many labs vegetatively propagate the plants, others regularly let the plants go through the life cycle (sexual reproduction through selfing) and establish fresh cultures based on single spores. However, for most of the pedigrees the frequency and number of sexual reproduction events the plants went through is unknown. Phenotypic differences are documented between laboratory strains, for example Gransden strains have shown different levels of loss of fertility (Meyberg et al., 2020). This recently led to the introduction of the Reute accession for the study of sexual reproduction (Hiss et al., 2017). Mutations underlying such differences as well as potential silent mutations can occur during sexual as well as vegetative propagation in the lab. Such laboratory divergences haven been reported in both prokaryote (Smits, 2017) and eukaryote laboratory models, for example in *Chlamydomonas reinhardtii* (Flowers et al., 2015). Mutation and selection underlie the forces of evolution. However, under laboratory conditions natural selection usually is absent. Over time, somatic mutations can thus accumulate in laboratory strains that would not occur in natural populations. Indeed, repetitive vegetative propagation of *P. patens* in the laboratory loosens the selection pressure on genes required for sexual reproduction, apparently leading to deterioration of the latter (Ashton and Raju, 2000; Perroud et al., 2011; Hiss et al., 2017; Meyberg et al., 2020). It should be noted that *P. patens* is predominantly selfing in the (dominant) haploid stage, developing completely homozygous diploid sporophytes. Hence, spores result that are genetically identical to the parent even though they are the product of meiosis.

Previous *P. patens* SNP studies analyzed genomic DNA samples of different *P. patens* accessions (Hiss et al., 2017; Lang et al., 2018). However, *P. patens* gDNA samples are rare. Nevertheless, the recent publication of RNA-seq datasets (Demko et al., 2014; Frank and Scanlon, 2015; Stevenson et al., 2016; Szövényi et al., 2017; Perroud et al., 2018; Fernandez-Pozo et al., 2019) provides a source of information that can be used to detect SNPs. Due to the high number of RNA-seq samples analyzed, efficient pipeline processing is essential. A framework of a modular RNA-seq pipeline was previously published (Perroud et al., 2018). While adding to and modifying this pipeline, a powerful solution for the here presented SNP analysis was created. Due to the current lack of genomic DNA we analyzed whether the SNP analysis of RNA-seq samples leads to comparable results. Based on the called SNPs we determined the rate and nature of somatic mutations among the accessions and pedigrees.

To identify and track genetic variation in the laboratory, restriction fragment length polymorphisms (RFLP) can be employed. This technique is based on SNPs modifying restriction enzyme recognition sites, which are covered by polymerase chain reaction (PCR) amplicons to test for genetic variation in specific DNA regions (Botstein et al., 1980).

Here, we identified SNPs using recently published RNA-seq data as well as unpublished RNA-seq and gDNA-seq data for a range of *P. patens* accessions and Gd pedigrees, i.e., laboratory strains with a documented ancestry. We used the resulting data to separate accessions as well as pedigrees via SNP analysis, extracted unique SNP sets for all accessions and Gd pedigrees, and developed RFLP analyses that are useful in maintaining accession and Gd pedigree identification.

MATERIALS AND METHODS

Sequence Sources

This study used data of five different *P. patens* accessions: 171 Gransden (Gd), 20 Kaskaskia (Ka), 32 Reute (Re), 27 Villersexel (Vx), and 25 Wisconsin (Wi) samples. The dataset contains 206 previously published RNA-seq samples as well as 44 novel RNA-seq samples. In addition, 25 novel gDNA samples of *P. patens* accession Wisconsin (Wi) were analyzed. These 275 samples were used for SNP detection. In addition, the Wi gDNA samples were used to study variation in a naturally occurring population. All samples used in the present study are available at the NCBI SRA database and are detailed in **Supplementary Table S1**.

Plant Material, Nucleic Acid Extraction and Sequencing

Physcomitrella patens accession Villersexel was collected in 2003 by M. Lueth in Haute-Saone (France) on dry mud at a fish pond east of Villersexel, at the Villers la Ville junction (voucher 4296). The accession Kaskaskia was also collected in 2003 in Illinois (United States) on a periodically flooded drainage channel at a corn field by D. Vitt and M. Sargent. The voucher information for both accessions has previously been published (von Stackelberg et al., 2006; Beike et al., 2014). Accession Reute has also been collected by M. Lueth/M. von Stackelberg in 2006 close to Freiburg, Germany on an agriculturally used field. The exact location has previously been published (Hiss et al., 2017).

Reute Early Sporophyte 1 (ES1)

Physcomitrella patens accession Reute_2015 (Re_2015) (Hiss et al., 2017) was cultivated on 9 cm petri dishes on solid Knop's medium enclosed with parafilm under long day conditions (70 μmol m*−2 s*−1 white light, 16 h light, 8 h dark, 22°C) as described in Hiss et al. (2017). Re was regularly reproduced sexually once per year since 2011. Re_2015 is the culture derived from the sexual reproduction (selfing) performed in 2015. Gametangia induction was performed by transfer to short day conditions (see Hiss et al., 2017 for culture details). Sporophytes were harvested 6–9 days after watering and immediately put into 50 μl RNA-later (Qiagen, Hilden, Germany). RNA was extracted using 20 ES1 sporophytes (according to Hiss et al., 2017) using the RNeasy micro kit (Qiagen, Hilden, Germany), following the manufacturers' protocol. RNA concentration and quality were analyzed with the Agilent RNA 6000 Nano Kit on a Bioanalyzer 2100 (Agilent Technologies). Library preparation and subsequent sequencing was performed by the Max-Planck-Genome-Centre Cologne (mpgc.mpipz.mpg.de). A single library was prepared using the IVT-based low input RNA-seq protocol followed by sequencing with Illumina HiSeq 3000 (150 nt, single ended).

Kaskaskia RNA-seq

Physcomitrella patens accession Kaskaskia was isolated from seven days entrained protonemal culture under long day conditions (70 μmol m*−2 s*−1 white light, 16 h light, 8 h dark, 22°C), if not stated otherwise (**Supplementary Table S2**). Tissue was flash frozen in liquid nitrogen and the subsequent RNA extractions were performed as described in (Perroud et al., 2018). The library preparation and subsequent sequencing was processed using the TruSeq RNA kit (Illumina) according to the manufacturer's instructions. The libraries were sequenced with Illumina HiSeq (100 nt, paired-end).

Villersexel Laser Capture of Sexual Reproduction Stages

Physcomitrella patens Villersexel (Vx) plants were routinely grown under sterile conditions on ammonium supplemented medium under 20 μmol m*−2 s*−1 of continuous light at 24°C. Protonemata were obtained from ground tissue and cultivated on cellophane disks on the previous medium. After 2 weeks, small patches of protonemata were transferred to low nitrate medium and grown for about 2 months under 20 μmol m*−2 s*−1 of a 16:8 light:dark cycle at 24°C. Well-developed gametophores were then transferred to 16°C under the same light regime for 3 weeks to induce sexual organ differentiation. Fertilization was synchronized in all cultures by flooding growing pots with sterile deionised water for 30 h; flooded gametophores were transferred to 24°C under continuous light. 48 h after flooding, gametophore tips were examined under a hand dissection microscope for the presence of fertilized archegonia. Non-fertilized cultures were treated as previously except for flooding.

Fertilized and unfertilized archegonia were hand dissected and collected in 100% acetone. Tissue fixation was ensured by infiltrating archegonia under low pressure for 2 min followed by a 48 h incubation in 100% acetone. Acetone was then exchanged with HistoClear by incubating fixed tissues in 50% acetone/50% HistoClear for 2 h then 100% HistoClear for 2 h under continuous shaking. Tissues were embedded in wax using an automated Tissue Tek VIP 5 Vacuum Infiltration (Sakura) machine with the following sequence: 3 baths in HistoClear for 1, 1 and 2 h then 4 baths in wax for 1, 1, 2 and 2 h. Thick sections of 10 μm were prepared from the embedded tissues and deposited on Nuclease-free 1.0 polyethylene naphthalate (PEN) membrane slides (Carl Zeiss Microscopy, #415190-9081-000) in drops of 1 X ProtectRNA^{TM} RNase Inhibitor (SIGMA #R7397), air dried and stored at room temperature until further use. After wax removal in HistoClear and 100% ethanol baths, zygote/early embryos, egg cell and archegonium tissues were laser dissected from the sections using a PALM MicroBeam unit (Carl Zeiss) at a 40x magnification following the procedure described in Saint-Marcoux et al. (2015). About 200 sections were captured per sample and 3 biological replicates were prepared for each tissue.

RNA was extracted using the PicoPure RNA extraction kit from Life Technologies (#KIT0204) and amplified into cDNAs using the Ovation RNA-Seq System v2 kit from NuGEN (#7102-32) as in Saint-Marcoux et al. (2015). cDNA quantity was determined using a NanoDrop ND-1000 spectrophotometer. cDNA quality was analyzed on a 2100 BioAnalyzer (Agilent Technologies) using RNA nano chips (5067-1511, Agilent Technologies) following recommendations in the NuGEN kit.

1μg of cDNA was paired-end sequenced on an Illumina HiSeq 2000 platform at the Beijing Genomics Institute in China. At least 2 × 10 million 100 nt reads were obtained per sample. Samples containing "orphans" in the sample name contain reads where the mate did not pass the quality filter.

Wisconsin gDNA

Mature (brown) spore capsules of *Physcomitrium patens* were collected in September 2017 in Wisconsin, United States (original specimen in AUGIE herbarium) by Rafael Medina (Augustana College Illinois). The surface sterilization procedure was performed at a laminar flow bench with freshly prepared 1% sodium hypochlorite and autoclaved tap water for rinsing. Five Single spore capsule were sterilized separately. After the last rinsing step the water was kept in the tube and the spore capsule was squeezed by sterile forceps so that the spores were released into the water. This spore suspension was transferred (using a micro pipette and autoclaved filter-tips) to solidified (0.9% [w/v] agar) Knop's medium containing 1% glucose in 9 cm Petri dishes sealed using 3M Micropore tape or Parafilm. After 3–5 days, when spore germination starts, five single sporelings were isolated from each capsule batch and separately transferred to fresh plates. After eight weeks under long-day conditions juvenile gametophores (above agar) were harvested and immediately frozen in liquid nitrogen. Genomic DNA was isolated from frozen plant material as previously described (Lang et al., 2018). Library-preparation and sequencing was performed at the Max-Planck-Genome-Centre Cologne (mpgc.mpipz.mpg.de); 25 TPase-based DNA libraries were sequenced in 1 × 150 bp single reads on Illumina HiSeq 3000 Analyzers.

Wisconsin experiment 2 was contaminated by prokaryotic sequences. The read contamination removal was done as described in Lang et al. (2018) and Nguyen et al. (2019). The leftover reads were used for further analysis.

Read Analysis

For easier manageability of the data, all original sample names were converted to a new nomenclature. Separator is always an underscore; the first two characters identify the accession (Gransden [Gd], Reute [Re], Kaskaskia [Ka], Villersexel [Vx] and Wisconsin [Wi]), the next one the origin/pedigree of the sample (e.g., MR-WT11), followed by the experiment defined by roman numbers (e.g., XX). Sample replicates (1-5), library type (SE or PE) and experiment type (mutant [MUT] or wild type [WTY]) are the last parts (Supplementary script "rename and extraction", **Supplementary File 1**). An example sample name is Gd_MR-WT11_XX_1_PE_WTY. Each RNA-seq sample went through a modified pipeline, build on top of the RNA-seq pipeline previously described (Perroud et al., 2018). The pipeline was modified by updating all software versions, enabling single-end (SE) read processing and adding SNP calling and post processing parts (**Figure 1**).

Read Quality

For read quality filtering and adapter removal, Trimmomatic (Bolger et al., 2014) version 0.39, was used. Adapter trimming of appropriate adapters (SE.fna or PE.fna; standard sequences included in the Trimmomatic package) was performed with a seed mismatch of 2, a palindrome clip threshold of 30, and a simple clip threshold of 10 for the paired-end reads (PE.fna:2:30:10). Base pairs with a quality score less than three were removed from the start (LEADING:3) and end (TRAILING:3) of the reads. Reads were further filtered using a sliding window of four base pairs with a minimum average quality score of 15 (SLIDINGWINDOW:4:15), removal of the first 10 base pairs (HEADCROP:10), and kept reads of 30 base pairs or more (MINLEN:30).

Poly-A clipping was performed by Prinseq-lite (Schmieder and Edwards, 2011) version 0.20.4. A minimum length of five poly-A/T nucleotides at the 5′- or 3′-end were required to remove the poly-A/T tails (TRIM_TAIL_LEFT 5; TRIM_TAIL_RIGHT 5). Only reads longer than 30 nt were kept (min_len 30).

Reference Genome Mapping

All filtered RNA-seq samples were mapped to the *P. patens* reference genome V3 (Lang et al., 2018) by GMAP-GSNAP (Wu and Nacu, 2010) version 2018-7-04. SAM and BAM file processing was performed by samtools (Li et al., 2009) version 1.9. Only uniquely mapped reads were used for further analysis.

Removing Duplicate Reads

De-duplication based on the unique mapped BAM files was done using samtools package markdup with the remove duplicate reads option (r).

Variant Detection

The SNP calling pipeline (**Figures 1B,C**) uses GATK version 4.0.9.0 (McKenna et al., 2010). The workflow was setup according to the classic GATK best practices workflow for RNA-seq[1,2] by modification of the approach published earlier (Hiss et al., 2017).

SNP Calling

GATK HaplotypeCaller was performed in default mode. To account for *P. patens* being haploid, the option "ploidy 1" was used.

The python script GetHighQualVcfs.py (Wang et al., 2012) was used for quality score recalibration. The option for haploid genomes (ploidy 1) was chosen. In addition, the alternative nucleotide quality (ALTQ) needed to be higher than 90% (percentile 90) and the genotype quality (GQ) value had to be greater than 90 (GQ 90).

The GATK tools BaseRecalibrator, ApplyBQSR and PrintReads were used in default mode.

Ploidy test

To test the samples' ploidy, GATK HaplotypeCaller was performed in default mode for diploid genomes (ploidy 2).

The python script GetHighQualVcfs.py was used for quality score recalibration. The option for diploid genomes (ploidy 2) was chosen. In addition, the alternative nucleotide quality needed to be higher than 90% (percentile 90) and the genotype quality (GQ) value had to be greater than 90 (GQ 90).

The results of both ploidy runs (1n and 2n) were compared. The results were interpreted taking into account the knowledge of previously haploid tested samples (**Supplementary Table S10**; cf. Results). We observed that the differences in the defined genotypes (GATK output 0/0, 0/1, 1/1, and 1/2) correspond to the differences in the number of called SNPs. Therefore, we chose the number of called SNPs to compare the two ploidy runs.

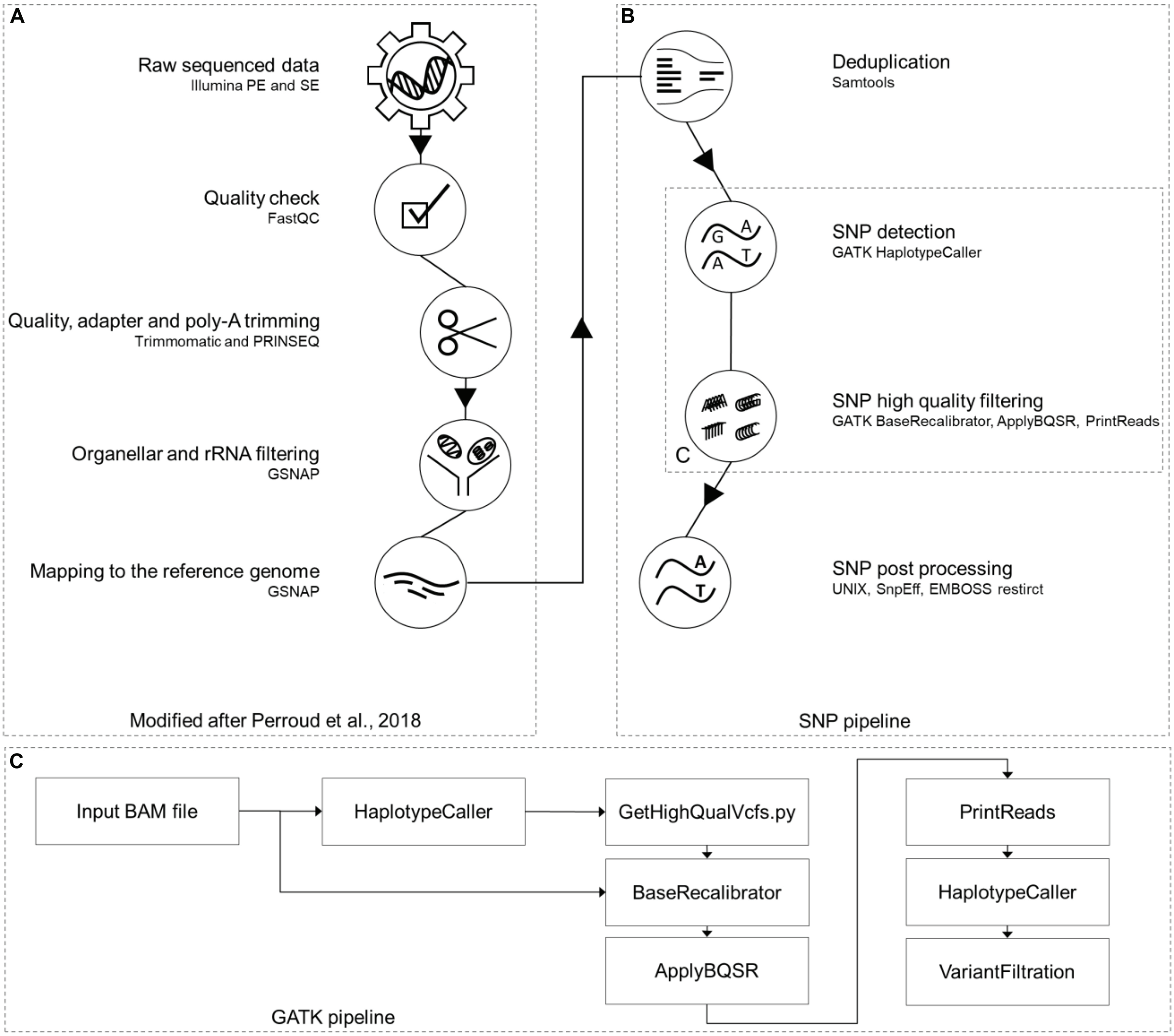


FIGURE 1 | RNA-seq SNP calling pipeline. Part **(A)** of this pipeline was previously published (Perroud et al., 2018). The additional SNP calling branch **(B)** on the right side starts with removing read duplications, using Samtools package “markdup” and continues with the GATK toolbox for SNP calling **(C)**. The last steps of this pipeline are post processing steps like SnpEff and EMBOSS restrict together with UNIX shell scripts. This figure has been modified based on a figure published in The Plant Journal (Perroud et al., 2018; https://onlinelibrary.wiley.com/doi/full/10.1111/tpj.13940).

Filtering Wisconsin gDNA SNPs

Single nucleotide polymorphisms called from the Wisconsin accession gDNA were filtered by using only reads uniquely mapping to the *P. patens* v3.3 gene annotation (to make the data comparable to the RNA-seq data). Bedtools intersect (Quinlan and Hall, 2010) version 2.29.0 was used, with the option (u) to write the original entry only once if multiple overlaps are found, to extract all gene models intersecting SNPs (Supplementary script “rename_and_extraction”).

Post SNP Calling Filter

The JGI gene atlas samples contain spike-in RNAs, which should not harbor SNPs. Hence, based on SNPs detected in these reads, filters were adjusted so that none of the RNA-seq spike-in base changes (sequencing errors) pass it. Filter values were allelic depths for the reference and alternative alleles (AD), mapped read depth (DP) as well as their fold change (FC) plus a minimum of three samples per SNP.

The above described values were adjusted through three consecutive filter steps. (i) The first filter was the read coverage filter with a minimum read depth of nine reads and a minimum of seven reads supporting the SNP. FC of AD and DP has to be greater than 0.77 (Supplementary script “SNP_filtering”). (ii) The second filter step removes all SNPs not present in at least three samples. This filter ensures the use of SNPs found by all technical triplicates of an experiment. (iii) While the third filter removed all indel positions.

The GO bias analyses were conducted as described previously (Widiez et al., 2014) to contrast gene sets affected by SNPs vs. the background of all genes. Visualization of the GO terms was implemented using word clouds generated by https://www.wortwolken.com. Word size is proportional to the –log10 (q-value), and over−represented GO terms were colored dark green if –log10 (q-value) $\leq$ 4 and light green if –log10 (q-value) > 4.

Plots were done by using R version 3.6.2 and ggplot2 version 3.2.1. Upset plot for the SNP intersection was performed with the R package UpSetR (Conway et al., 2017). All regression lines and confidence intervals were calculated by the R package ggplot2, method "lm" and the R package ggpubr version 0.2.5 to calculate R[1,2].

SNP Normalization

Several plots (**Supplementary Figures S3–S6**) were generated to check for potential normalization methods. The number of read covered base pairs (coverage), the number of reads per sample (reads), and the number of genes, respectively their accumulated length (genes) were taken into account.

Coverage method

The dependency of called SNPs based on the number of read covered based pairs was determined with the following method.

To find all read covered base pairs, the mapping output (BAM format) was analyzed by samtools package depth. All sequence positions, including unused reference sequence positions, were printed (aa). The output was filtered for depth $\geq$ 9 (similar to the SNP DP value). The number of filtered SNPs were divided by the number of read covered base pairs. To compare the values directly with the results found in the division was done vice versa to derive the format "one SNP per X bp".

To plot the values, the number of SNPs were corrected by the maximum number of read covered base pairs (**Supplementary Figure S4**).

Reads method

To detect the relation between the number of filtered SNPs and the number of sequenced reads, the values were plotted using the R packages described in section "Post SNP Calling Filter."

Genes methods

To answer the question whether SNPs accumulated at specific chromosomes and to observe the relation between the number of genes or their length with the number of detected SNPs, gene information extracted from the *P. patens* v3.3 annotation GFF file (Lang et al., 2018) was used. Both, the number of genes and the gene length, were summarized per chromosome. The extracted gene values were divided by the number of filtered SNPs to derive relation in the gene number and gene length plots, respectively. To test for significance Fisher's exact test was performed. The number of base pairs w/o SNPs for each of the 27 individual chromosomes (and for all unassigned, merged scaffolds) was compared. All p-values were corrected using the R method p.adjust using the method (Benjamini and Hochberg, 1995).

Extracting Exclusive SNPs

In the context of SNPs found only in a specific accession or Gd pedigree, the terms unique and exclusive are used synonymously. Exclusive SNPs were extracted for each accession and for each Gransden pedigree, using bash/awk scripts (Supplementary scripts "rename_and_extraction", "SNP_clustering" and "SNP_filtering"). First, all SNPs found in all GATK VCF files were grouped into a single file. Subsequently, the groups were inspected for SNPs exclusive for a specific accession or Gd pedigree (Supplementary script "SNP_filtering"). For further accession analysis, the SNPs were sorted by the number of supporting samples. SNPs supported by > 90% of the samples of one accession, and not found in others, were defined as exclusive. The read coverage filter was not applied for the accession exclusive SNP selection. For the Gd pedigrees, the Gd exclusive SNPs were ranked by the number of supporting samples. The SNP with most sample support received the highest rank, the five SNPs with the most sample support were chosen and defined as exclusive.

Accession Clustering

Detected nucleotide variation was clustered by two different methods. The first method was an artificial FASTA alignment (**Supplementary File 4**). This method clusters only SNPs, no InDels. Only SNPs that passed all filter steps were used. Each SNP is a single column in the alignment. If the sample contains a SNP at a specific position, the SNP nucleotide was added to the FASTA sequence of the sample, otherwise the reference nucleotide was used.

The second method was chosen to cluster SNPs and InDels. Instead of nucleotides, numbers were chosen to represent a SNP, InDel or the reference. A matrix was created by substitution of reference and variant nucleotides: reference 0; SNP 1; indel 2. This converted numbers were added to the matrix similar to the nucleotides in the above described FASTA file. Each row is a single sample and each column a unique SNP/indel position (Supplementary script "SNP_clustering").

The artificial FASTA alignment was imported to SplitsTree (Huson and Bryant, 2005) version 4.14.8. A network was calculated using default parameters. The tree was generated by the NJ option and stored in NEXUS format. FigTree (Bouckaert et al., 2014) version 1.4.4 was used to draw a circular tree based on the SplitsTree NEXUS file.

The SNP/indels 0-1-2 matrix was loaded into R version 3.6.2 using the function dist with the method euclidean. To get a three dimensional PCA plot, the results were transferred to the R package rgl version 0.100.30.

SNP Effects

Synonymous and non-synonymous SNPs for each sample were detected by SnpEff (Cingolani et al., 2012) version 4.3T in default mode. SnpEff used a database created of the *P. patens* genome annotation v3.3 to locate SNP positions at gene regions. Only SNPs that passed all three filter steps (minimum nine reads have to cover the SNP position and minimum seven reads have to support the SNP, at least three samples have to support the SNP, indels are removed) were used.

[1] https://gatkforums.broadinstitute.org/gatk/discussion/3892/the-gatk-best-practices-for-variant-calling-on-rnaseq-in-full-detail

[2] https://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels

Synonymous and non-synonymous SNPs were extracted from the SnpEff CSV file output and all involved genes were extracted from the SnpEff gene.TXT file. Functional analyses were done via GO-bias analysis, described in chapter "Post SNP Calling Filter."

Identification of Restriction Sites Overlapping With SNPs

EMBOSS restrict[3] was used to detect SNPs in putative restriction endonuclease recognition regions. The enzyme database, containing all necessary information about the recognition sites, was loaded with the tool EMBOSS rebaseextract[4]. The rebase restriction endonucleases databases, withrefm.907 and proto.907, were downloaded at ftp://ftp.neb.com/pub/rebase. EMBOSS restrict was performed with a minimum length of the restriction enzyme recognition site of five base pairs (sitelen 5) and all enzyme at the database were used (enzymes all).

SNP Verification via PCR and RFLP (Restriction Fragment Length Polymorphism)

Exclusive SNPs for each *P. patens* accession overlapping with a restriction enzyme recognition site were selected as described above. SNPs affecting six or eight nt long recognition sites were chosen. Additionally, enzyme requirements for easy usability and frequency of cuts in ± 2 kbp around the SNP were analyzed to ensure an interpretable gel band pattern. Primers were designed to result in an amplicon of 700-1,400 bp and similar annealing temperatures (∼ 59°C, **Supplementary Tables S7**, **S8**).

Plant Material and gDNA Extraction

To analyze SNPs located within restriction enzyme sites (comparison of accessions) and SNPs without restriction enzyme site (comparison of Gd pedigrees) the *P. patens* accessions and Gd pedigrees Gransden DE Marburg 2015 (Gd_DE_MR), Gransden Japan (Gd_JP, Gd_JP_Okazaki and Gd_JP_St.Louis), Gransden Grenoble (Gd_CH), Reute 2015 (Re), Kaskaskia (Ka) and Villersexel (Vx) were cultivated as described above. Genomic DNA for PCR amplification was isolated, using a fast protocol using one to two gametophores as published in (Cove et al., 2009).

PCR Analysis and Sequencing

Polymerase chain reaction was carried out with OneTaq polymerase (NEB) following the manufacturers' protocol. Annealing was carried out between 55°C and at 59°C and elongation time was adjusted to the longest fragment chosen (95 s). For primer sequences see **Supplementary Tables S7, S8**. 5 µl PCR product, 2.5 µl of the forward primer (10 µM) and 2.5 µl water were Sanger sequenced (Macrogen, Germany) (**Supplementary Table S9** and **Supplementary File 6**). PCR products and all subsequent fragment analyses were visualized via gel electrophoresis (0.7% agarose, Roth, Germany) using peqGREEN (VWR, Germany) as dye. The 1 kbp size standard was purchased from NEB.

Restriction Analysis

For each tested SNP, 15 µl PCR product of all accessions were used as input for the enzymatic digestion. Restriction was carried out for the SNPs Re_c3_17747483_A-T, Vx_c3_2712099_A-G and Ka_c01_25061888_C-A using 2U of the corresponding enzyme (**Supplementary Table S7**, NEB) for 3 h at 25°C for *Swa*I and at 37°C for *Nde*I and *Xba*I. Fragments resulting from the restriction were visualized via gel electrophoresis as described before (PCR analysis and sequencing).

Natural Population Diversity

To determine variation within a naturally occurring *P. patens* population, the accession Wisconsin gDNA SNP results were used. Because of bacterial contamination, sample Wi_2 was excluded from this study. The experiment was designed with four capsules and five spores each. Each spore represents one sample. The number of exclusive SNPs for each sample (spore) within a spore capsule were detected as well as the number of exclusive SNPs for each spore capsule. The results were compared with the results of exclusive SNPs found in laboratory accessions and pedigrees of Gransden, Gd_DE 2011, 2012 and 2015, and Reute 2007, 2012 and 2015. To highlight the results Venn diagrams were created by venny[5].

All samples described above were used to generate an artificial FASTA alignment (for methods see section "Extracting exclusive SNPs") which was analyzed by Splitstree. Here, only gDNA SNPs which intersected with the *P. patens* v3.3 annotation file were kept. The branch lengths were adjusted by coverage normalization (see section "Coverage method").

RESULTS

Read Analysis and SNP Discovery

The analysis was conducted with a total of 4.7 billion RNA-seq reads (**Supplementary Table S3**). 68% of all reads are from Gransden, Reute reads account for 18%, Kaskaskia for 12% and Villersexel for 2% (**Supplementary Table S4**). After pre-processing and mapping to the reference genome (**Figure 1A**) 81% of all reads remained (**Supplementary Table S3**). De-duplication (to account for potential PCR bias) further reduced the amount of reads by 20%, leaving 3.0 billion reads as input for the GATK SNP pipeline (**Figures 1B,C**). The unfiltered Wisconsin gDNA samples amounted to 1.0 billion reads. Processing, mapping to the reference and deduplication discarded more than half of the raw reads; 473 million reads were used for the SNP pipeline (**Supplementary Table S3**).

Funariaceae are known for naturally occurring polyploidization (Rensing et al., 2013; Beike et al., 2014), this has also been demonstrated during *P. patens* mutant generation using protoplasts (Schween et al., 2005). We performed a ploidy test using GATK with n = 1 vs. n = 2 and generally detect a lower number of SNPs when assuming haploidy (n = 1), on average 65.4% of n = 2. The percentage range of samples confirmed to be haploid (36.2 – 92.2%) approximately

[3] http://emboss.sourceforge.net/apps/cvs/emboss/apps/restrict.html

[4] http://emboss.sourceforge.net/apps/cvs/emboss/apps/rebaseextract.html

[5] https://bioinfogp.cnb.csic.es/tools/venny/index.html

coincides with the percentage range of all samples (30.7 – 92.9%) (**Supplementary Table S10** and **Supplementary File 7**). Moreover, manual inspection of the VCF files for the Wi gDNA SNP calls showed very minor differences, that are smaller than those of the RNA-seq data of confirmed haploid plants. Taken together, we do not find evidence for polyploid plants among the samples used.

For the Wisconsin gDNA samples 2,473,107 SNPs were called by the GATK pipeline (**Figure 1C**). After intersecting the gDNA SNPs with the gene coordinates of the *P. patens* v3.3 annotation, 140,832 SNPs were kept that represent the transcriptome, to be comparable to the RNA-seq SNPs. Merging the Wi v3.3 SNPs with the results of the RNA-seq accessions ended up in a total number of 1,233,585 transcribed gene space SNPs. Gd has the lowest number of SNPs relative to the reference assembly. This fits the expectation, since the reference genome was derived from a Gd pedigree. The accessions Wi and Ka have the highest number of SNPs per sample (**Supplementary Figure S1**). The highest SNP reduction can be observed after the (i) read coverage filter, which was, together with the (ii) sample support filter, adjusted using spike-ins (see section "Materials and Methods" for details). (i) Read coverage and (ii) sample support filter, together with the (iii) indels removal, were reducing the SNP set by 88% (146,816 SNPs shared by five accessions, **Supplementary File 5**). A comparison of SNP intersection between SNPs called in this study and SNPs previously published (Lang et al., 2018) demonstrates a large overlap of 89% of the previously called Vx SNPs (as compared to those that were detected in this study) and minor overlaps for Re (26%) and Ka (28%) (**Supplementary Table S5**).

SNP Comparison Between Accessions

Most SNPs can be observed in the intergenic regions (up- and downstream of the gene bodies according to the v3.3 annotation). The SNP distribution for all accessions is around 40:60 (gene regions/intergenic regions). The accessions Wi and Vx have almost no SNPs flanking the two base pairs next to the splice site (splice site region).

Most of the SNPs shown in **Figure 2** are accumulated in non-coding regions. Exonic SNPs can be synonymous, not affecting the coding sequence, or non-synonymous, leading to a change in the amino acid sequence of the protein encoded by the gene (for average number of SNPs per sample see **Table 1** and for total numbers of SNPs see **Supplementary Table S11**). The two accessions from North America, being geographically most far away from the reference sample, are the ones with the most changes affecting the coding sequence. Individual SNP effects in the exclusive accession SNPs list can be found in **Supplementary File 3**.

Less than 12% of all SNPs called by the GATK pipeline passed all three filter steps: Gd has 39,614 and Re has 42,094 SNPs left, Vx has 52,960 SNPs and Wi has 63,597 SNPs. The highest number of SNPs are found in Ka with 76,076 SNPs (**Supplementary Figure S1** and **Figure 3**, left horizontal bars). The number of SNPs coincides with the geographical distance to the reference Gransden (**Figure 3**, horizontal bars; **Supplementary Figure S2**). After applying four different normalization methods (see section "Materials and Methods" for details), Gransden and Reute exhibit always the lowest SNP rate (**Supplementary Table S6**), mirroring previous results (Beike et al., 2014; Lang et al., 2018). The approximate linear relationship between number of reads and called SNPs (**Supplementary Figure S3**) led to the normalization by read number. The coefficient of determination (R^2) is found to be 0.6 – 0.93 (**Supplementary Figure S3**). To compensate for unequal distribution of reads we also normalized by the fraction of the sequence space that carries enough read support to allow SNP calling (see section "Coverage method," **Supplementary Figure S4**). By applying the SNPs to read covered base pairs, instead of the raw read number, the R^2 values increased. Wi, Ka and Vx reach almost 1, Re and Ge 0.77 and 0.85. Based on the coverage normalization (**Supplementary Figure S5**), Gd has 1 SNP per 4,666 bp, Reute has 1 SNP per 1,912 bp followed by Ka (1 SNP per 630 bp), Wi (1 SNP per 206 bp) and Vx (1 SNP per 143 bp). The gene normalization methods (**Supplementary Figure S6**) indicate that chromosome 19 and chromosome 26 exhibit significantly ($q \leq 0.05$) more SNPs than the other chromosomes.

TABLE 1 | Average number of SNPs affecting gene coding sequences per sample.

	Gd	Re	Vx	Ka	Wi	all
Start changes[a]	3	8	8	18	29	8
Stop changes[b]	6	23	22	97	237	41
Sequence changes[c]	774	2,978	2,942	11,794	13,090	3,168
Synonymous	411	1,232	1,272	4,698	4,737	1,300
Non-synonymous	363	1,746	1,670	7,096	8,353	1,868

[a] Start changes include start codon gains and losses. [b] Stop changes include gains and losses of stop codons. [c] Sequence changes include non-synonymous changes affecting the encoded amino acids, synonymous sequence changes, and insertions or deletions that do not change the sequence frame.

The SNP intersection shows 1,541 SNPs are shared by all accessions (**Figure 3**). There are accession specific SNPs (exclusive SNPs) as well. Most exclusive SNPs are present in Vx (31,818), followed by Wi (22,014), Ka (7,905), Re (4,793) and Gd (2,184) (**Figure 3**, vertical black bars). Gd is sharing 94% of its SNPs with other accessions, Ka and Re share > 87%, Wi shares 65% and Vx 40% SNPs with all other accessions.

Applying a filter to extract exclusive SNPs supported by ≥ 90% of the samples, Wi and Ka have most exclusive SNPs/InDels, Wi has 4,007 unique SNPs, Ka 3,393. 890 SNPs are only present in the Re accession while in the Vx accession 21 exclusive SNPs were found (**Supplementary File 3**).

100 kbp SNP hotspot regions were detected to survey the *P. patens* accessions (**Supplementary Figure S7** and **Supplementary File 2**). On Chr26, starting at 300,000 bp, a hotspot region is present in all accessions. All accessions but Gd share one region on Chr19. Gd, Re and Ka share 100 kbp hotspot regions on Chr03 and one on Chr06. Ka, Wi, and Vx share regions on Chr04, 07 and 13 (**Supplementary Figure S7** and **Supplementary File 2**). Biased GO terms of the described regions are shown in **Supplementary Figure S8**. Most 100 kbp SNP hotspot

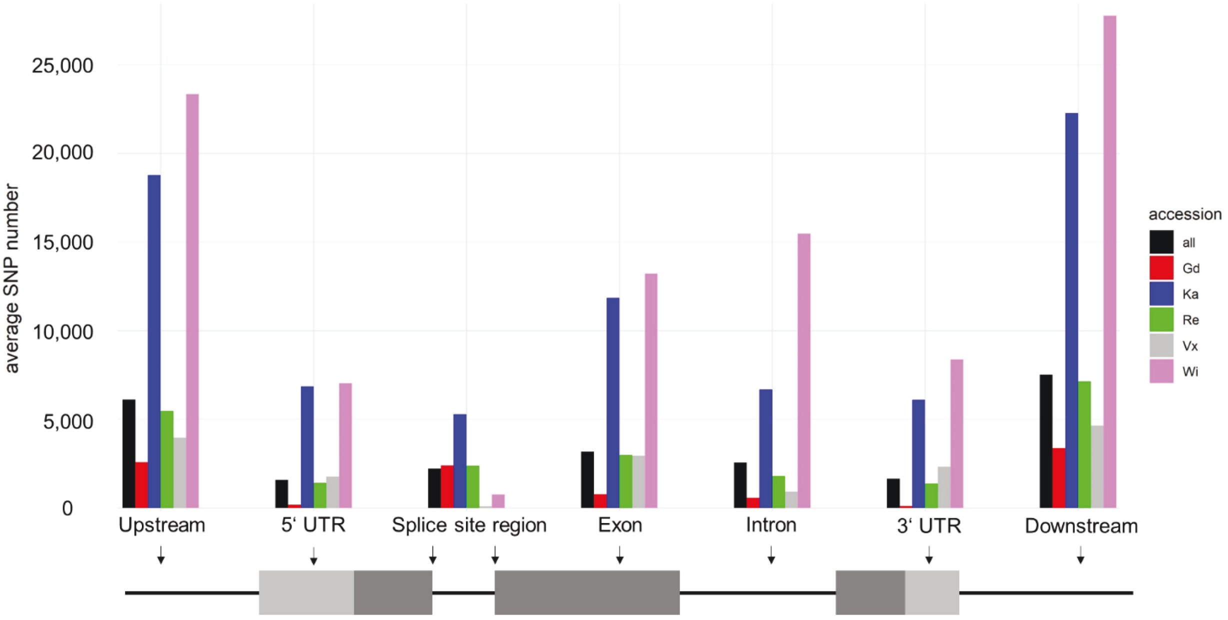


FIGURE 2 | Average snpEff output for each accession. Shown are average numbers of SNPs affecting specific regions, highlighted in a schematic gene structure shown below the corresponding grouped columns. Most SNPs are up- and downstream of genes (intergenic). SNPs at splice site regions are intron SNPs, located on the first and last two intron base pairs.

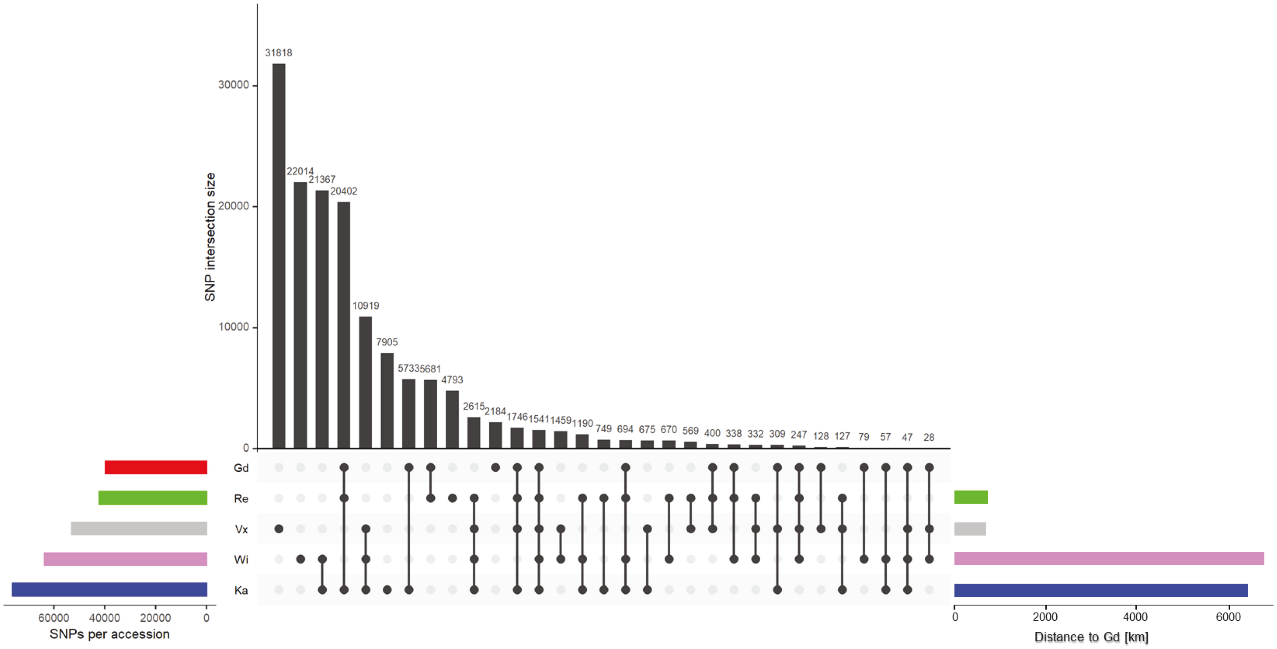


FIGURE 3 | SNP intersection of the five accessions. The horizontal colored bars on the left show the total number of SNPs per accession after applying all three filter steps. The bars to the right show the geographic distance to the reference Gd. The colors represent the five accessions throughout the text. The vertical black bars show the number of intersecting SNPs, marked by the dots below.

regions are overlapping with the SNP hotspots found by (Lang et al. (2018); **Supplementary File 2**, Table B). However, there are also a few hotspot regions only found in the present study.

Using an artificial FASTA alignment of all SNPs, we performed a clustering analysis (**Figure 4**). Samples of the accessions Gd, Re, Ka, Vx and Wi are clustering with each other, respectively, indicating that our approach is able to detect the respective

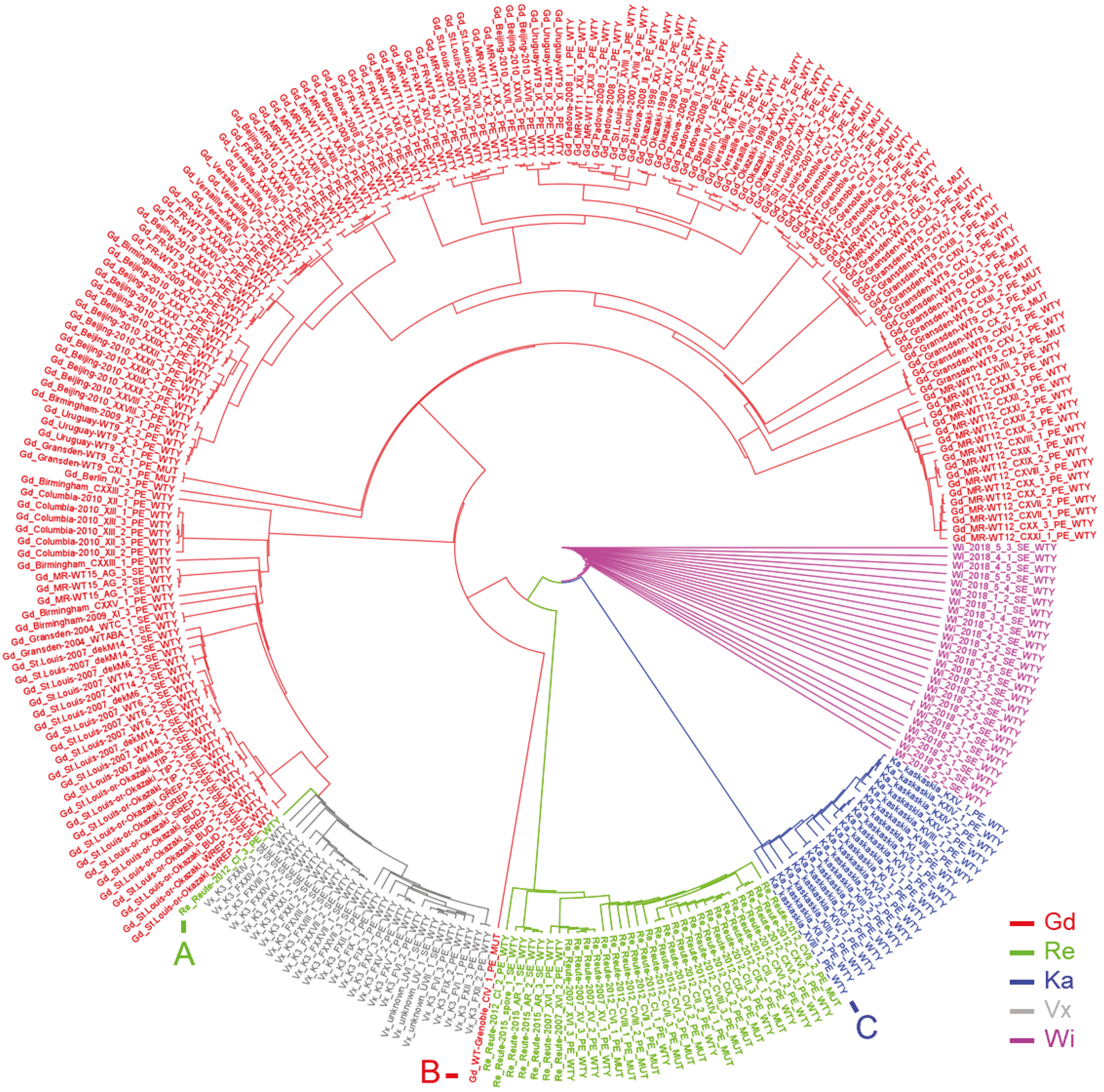


FIGURE 4 | Circularized SplitsTree network based on an artificial FASTA SNP alignment file. The Neighbor-Joining tree of five *P. patens* accessions is shown. All libraries cluster within their accession and applied treatment, except for the marked libraries: **(A)** Sample Re_REUTE-2012_CI_3 has 100 x lower read coverage than the other Re samples. It clusters next to the low read coverage Vx samples. **(B)** Sample Gd_WT-Grenoble_CIV_1 is a Gd outgroup. **(C)** Ka sample which was falsely annotated as Gd at the NCBI SRA (XVIII_1_PE_WTY), determined by exclusive SNP analysis (**Supplementary File 3**, Sheet Ka_exclusive_SNPs).

genetic background. The three European accessions form a clade to which Ka and Wi are sister. One Re sample, belonging to the experiment CI_3 (NCBI BioProject PRJNA411193), does not cluster with the other Reute samples (**Figure 4A**). The number of reads in this sample is 100 x lower than in the other samples of experiment CI, potentially causing biased SNP calling and hence incorrect clustering. The Gd sample CIV_1 (**Figure 4B**) possesses an outlier position with regard to the other European samples. The sample of the NCBI BioProject PRJNA411163 is annotated as Gransden accession. However, it could be shown by clustering (**Figure 4C**) and exclusive SNP analysis that the sample belongs to the accession Kaskaskia. Principal component analysis (PCA) of SNPs as well as InDels recapitulates the SNP clustering results (**Supplementary Figure S9**). The samples from Szövényi et al. (2017) went into the SNP calling pipeline as a blind test. The sample origin was originally marked as unknown. Both clustering methods assigned the samples to Vx, with corresponds to the origin confirmed by the authors.

SNP Comparison of Gransden Pedigrees

Gransden is more widely used in laboratories than any of the other *P. patens* accessions. Based on information retrieved from the laboratories involved, the Gransden accession was classified into four pedigrees, Germany (DE), United Kingdom (UK), Switzerland (CH) and Japan (JP) (**Figure 5**). The original Gransden accession from the United Kingdom made it first to Hamburg, Germany (founding the DE pedigree), before it was sent to Lausanne, Switzerland (CH) and Okazaki, Japan (JP). The Lausanne strain was sent to Versailles, France and further distributed to Padova, Italy and Grenoble, France. In 1998, Gransden DE arrived in Freiburg, Germany. In Freiburg the Gd plants went through sexual reproduction (selfing) once per year. Starting 1999 the Freiburg pedigree went through nine rounds of selfing leading to WT9. The offspring were labeled by consecutive numbers or the year of sexual propagation. Gransden Freiburg (WT9) was sent to Uruguay, Beijing (China) and Marburg, Germany. Gransden Marburg started in 2011 and went through selfing each year except 2013. The Gd United Kingdom 2004 sample was sent to St. Louis, United States (**Figure 5a**) for gDNA isolation and used to sequence the *P. patens* reference genome (Rensing et al., 2008). However, the Gd UK 2004 reference sample was not broadly distributed. In 2007, another Gd sample was sent to St. Louis, USA from Okazaki, Japan. These plants were used for further analysis and also sent to Columbia. It should be noted that most papers that cite the reference genome paper with its Gd 2004 sample are actually using different pedigrees.

Our analyses show that Gransden accumulated different mutations in different laboratories during prolonged *in vitro* culture. To eliminate misleading SNP background noise, the exclusive SNPs for the Gd pedigrees were detected after applying read coverage and sample support filters. The intersection of the four Gd pedigrees (**Supplementary Figure S11**) shows that Gransden Germany (DE) has 1,112 exclusive SNPs while Gd_CH has 67 exclusive SNPs, Gd_JP 187 and Gd_UK features four (**Figure 5**). Because there is no SNP supported by at least 90% of all samples of a specific pedigree, the extraction of exclusive SNPs was done by getting the best supported SNPs. SNP ranking by the number of samples that support it was used to select the five most supported SNPs for a given pedigree. The Gd_DE top five SNPs are supported by 76–77 samples, Gd_CH between 12 and 18 samples, Gd_JP by 12 to 29 samples. For Gd_UK three samples support the top five list (**Supplementary File 3**). A clear clustering based on the FASTA alignment file, as for the accessions, is not possible (**Supplementary Figure S10**). In some cases, the samples grouped by experiments instead of Gd pedigree, which could be due to the low number of SNPs, and similar genes being expressed, biasing the number of available SNPs for the comparisons. If samples are highly specific for a single tissue (e.g., antheridia bundles or spores), not all genes are covered by the extracted transcripts and consequently SNPs cannot be detected.

Since some of the samples have a documented sexual propagation history (i.e., we know how many years/cycles of sexual reproduction lie between samples) we used the opportunity to determine whether SNPs were generally lost or gained in these samples. We find that for samples that were subject to regular sexual reproduction, SNP numbers generally decreased along the timeline (**Supplementary Table S12** and **Supplementary Figure S12**). The observed mutation rate was found to be similar across the different pedigrees (**Supplementary Table S13**).

Experimental Confirmation of Selected SNPs via Sequencing and RFLP Analysis

For all primer pairs (**Supplementary Tables S7, S8**) covering SNPs specific for different accessions, PCR amplicons could be generated. Sequencing analysis of the PCR products showed in all tested positions (9/9 positions, **Supplementary Tables S7, S8**) the presence of the predicted SNP in the corresponding accessions' and Gd pedigree background (**Supplementary Figures S13–S17**). To provide an easy and cheap tool to distinguish the different accessions, RFLP analysis (**Figure 6**) was successfully established for the SNPs Re_c3_17747483_A-T, Vx_c3_2712099_A-G and Ka_c01_25061888_C-G (**Supplementary Figures S13–S15**). The Re_c3_17747483_A-T amplicon (1,255 nt) was digested with *Nde*I resulting in two fragments (990 nt and 265 nt) for the accessions Gd, Ka and Vx, and absence of digestion in Re (**Supplementary Figure S13**). For Vx_c3_2712099_A-G, the amplicon of 1,366 nt was digested with *Swa*I leading to two fragments (1,063 nt and 303 nt) in Gd, Ka and Re but not in Vx (**Supplementary Figure S14**). For Ka_c01_25061888_C-G, the 1,342 nt amplicon was digested with *Xba*I resulting in two fragments (984 nt and 358 nt) in Gd, Re and Vx, but no digestion in Ka (**Supplementary Figure S15**). Results for SNPs not tested by RFLP (**Supplementary Table S8**) for two accession primer pairs (Re_c04_21933417 and Vx_c13_4764050) and five Gd pedigree primer pairs (Gd_DE_c02_12750876, Gd_DE_c05_3105395, Gd_DE_c12_2095061, Gd_JP_c20_868 8243, Gd_CH_c23_11248087), show the presence of the predicted accession and Gd pedigree SNPs on the sequence level (**Supplementary Figures S16**, **S17**).

Natural Population Variation and Selection

Samples of pedigrees with known propagation history were chosen to estimate the annual number of mutations per base pair (observed mutation rate). The time period covered is six years for Gd and eight years for Re. The number of SNPs called for all pedigrees generally decreases under regular sexual propagation. The same is true for the estimated mutation rate (**Supplementary Table S13**). The lowest annual mutation rate with 2E-07 was detected for the Freiburg WT11 (FR_WT11) pedigree, the highest rate for Reute-2012 with 4E-06.

The diversity of genome-wide SNPs found within the Wisconsin natural population single spore isolates is lower compared with three selfed generations (pedigrees) of laboratory accessions. The lower numbers can be observed both on sample/spore and on pedigree/capsule level (**Supplementary Figure S18**). However, on the level of the artificial FASTA alignment of the gene body SNPs, represented

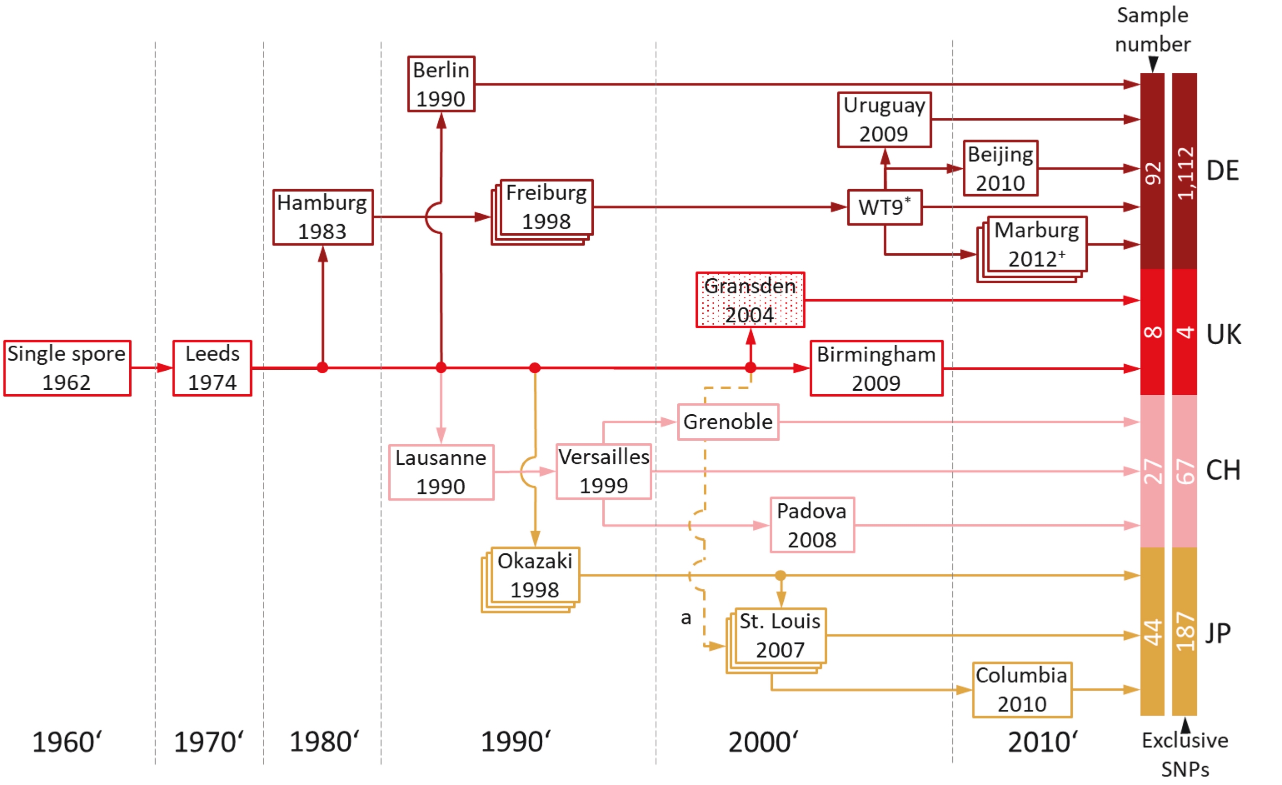


FIGURE 5 | Gransden pedigree. The pedigree diagram shows Gransden strains of 13 different labs used in the present study. The Gransden accession was arranged in four different pedigrees, Germany (DE), United Kingdom (UK), Switzerland (CH) and Japan (JP). The United Kingdom pedigree was sent to St. Louis, United States in 2004 and used to sequence the reference genome (a). However, this strain was not used or broadly distributed afterwards. The plants analyzed in St. Louis are derived from the Japan pedigree (2007). Pedigrees shown in stacked boxes went through yearly selfing. + Since 2011 yearly selfing except 2013. * Since 1999 Gransden Freiburg went through nine generations leading to WT9. The numbers of samples and exclusive SNPs for each of the four pedigrees are shown to the right (also shown in **Supplementary Figure S11**).

by a Splitstree tree (**Figure 7**), similar normalized branch lengths for Wi samples and most Re and Gd pedigrees can be observed.

The ploidy test using GATK with $n = 1$ and $n = 2$ resulted in a high rate of congruence for Wi. The $n = 1$ explained 84.3% – 95.6% (average 88.1%) of the SNPs called in the $n = 2$ run (**Supplementary File 7**). Approximately 18% of the Wi SNPs are heterozygous, a lower number than for any of the other accessions/pedigrees (**Supplementary Table S14**). Hence, the naturally occurring heterozygosity of the Wi population is lower than that observed in cultured samples. Much of what is detected as heterozygous is probably due to very closely related (identical and near-identical) paralogs that are known to be present in the *P. patens* genome (Rensing et al., 2008). Yet, the low apparent Wi heterozygosity reinforces that *P. patens* is a predominantly selfing species (Perroud et al., 2019; Meyberg et al., 2020; Rensing et al., 2020).

We calculated the rate between non-synonymous nucleotide changes (Ka) and synonymous changes (Ks) per sample and accession (**Supplementary Table S11** and **Supplementary File 7**). Over all samples, the Ka/Ks rates follow a clear linear trend ($R^2_{adj} = 0.98$, **Supplementary Figure S19**), suggesting neutral evolution (no global selective pressure). However, most individual samples deviate from the 99% confidence interval of the linear regression and hence putatively show evidence of negative selection (Ks ≫ Ka), or positive (Darwinian) selection (Ka ≫ Ks). The accession Gd, which represents the genome reference, apparently is under negative selection, all the other four accessions show evidence of positive selection (**Supplementary Table S11** and **Supplementary File 8**). The GO bias of genes affected by non-synonymous changes was calculated and visualized via word clouds (**Supplementary Figure S20**).

DISCUSSION

Read Analysis and SNP Discovery

Here, we analyzed sequence variants in *P. patens* accessions and Gransden pedigrees using mainly sequences from gene expression (RNA-seq) experiments. Therefore, this study is limited to the gene space, lacking information of most of the intergenic regions, where the selection pressure is lower and more changes accumulate (Krasovec et al., 2017). On the other hand, the advantage of using RNA-seq data is the much higher availability of data. Very few genomic data sets, and with low sequencing depth, are currently available

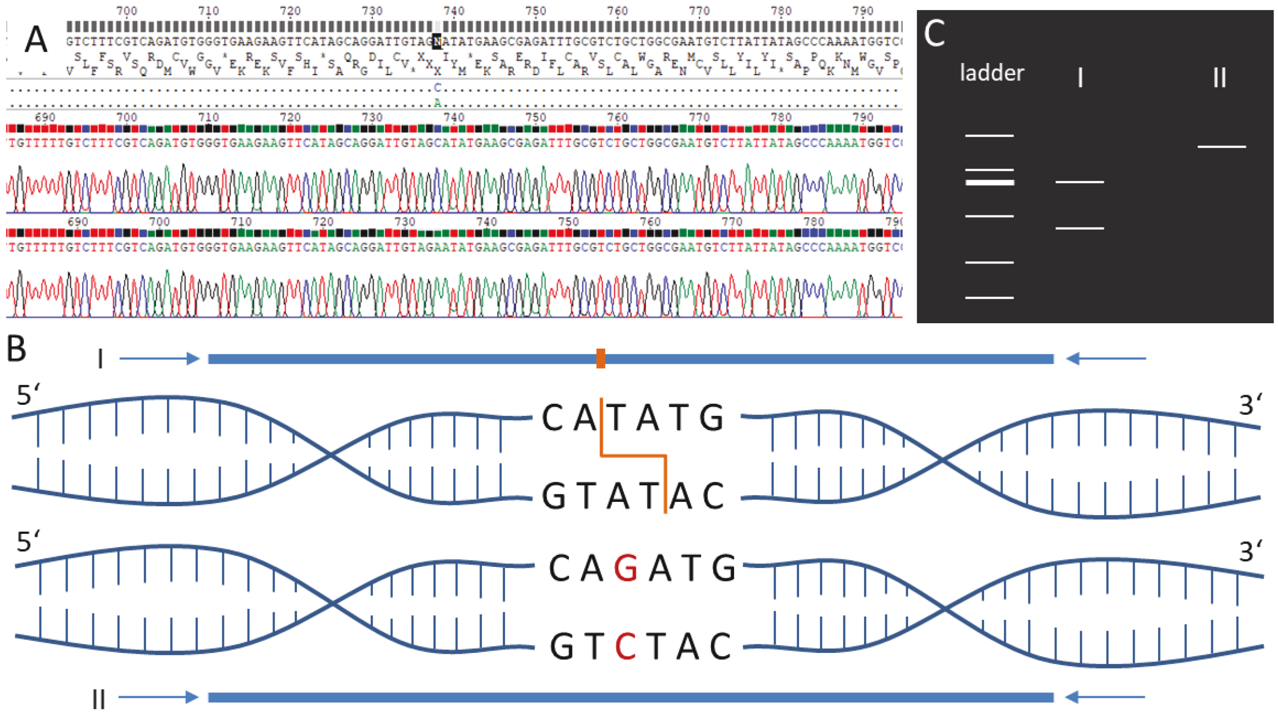


FIGURE 6 | Schematic visualization of the restriction fragment length polymorphism (RFLP) analysis. **(A)** Electropherogram of the sequenced amplicons generated via PCR using forward and reverse primers. **(B)** PCR amplicons of the samples I and II covering the same genomic position in two different *P. patens* accessions. Sample I sequence includes a restriction enzyme site for *Nde*I (orange). Sample II contains a single nucleotide polymorphism (SNP, red) resulting in the loss of the restriction enzyme site. **(C)** If amplicons are digested via the corresponding restriction enzyme Nde*I*, sample I results in two bands when separated via gel electrophoresis, whereas sample II results in one band. See **Supplementary Figures S13–S15** for experimental verification of the accession-specific RFLP regions.

for *P. patens* accessions and Gransden pedigrees. However, hundreds of RNA-seq experiments could be used in this study, allowing much higher resolution to detect sequence variants in genes. To ensure the quality of the SNPs found, several filters were applied. Finding a feasible filter for the called SNPs is a major step during the analysis due to risk of over- or underestimation. The presence of RNA spike-ins in some of the samples, which mimic natural eukaryotic mRNAs, gave us the opportunity to distinguish sequencing/mapping errors from actual sequence variants.

Single nucleotide polymorphisms filtering is required to reduce the false-positive rate of SNP detection. Amplification errors during sample preparation and sequencing (Ma et al., 2019) can lead to incorrectly called SNPs as well as software issues while mapping and SNP calling (Ribeiro et al., 2015). We used RNA spike-ins to detect such false-positive SNPs. Spike-ins do not exhibit SNPs. Hence, all called SNPs in spike-in mRNAs represent sequencing or computation errors. The read depth filter was adjusted to remove spike-in SNPs without losing too much sensitivity. GATK output VCF files contain a lot of information about the background data of the SNP, *inter alia*, read coverage at the SNP position. By extracting all spike-in SNPs and evaluating different parameters, the read coverage parameter [DP] and the parameter of how many reads at that position were supporting the SNP [AP], seemed to be the most feasible parameters to filter out spike-in SNPs. The number for DP of nine reads was chosen because only 4/381 spike-in SNPs were left after applying that filter (equaling 1% false positives; at DP = 10 the sensitivity breaks down). Another observation led to the sample support filter. SNP variation of more than 30% between replicate RNA-seq samples could be observed (**Supplementary Figure S18,A**). Using only SNPs found in at least three samples removed the last four false positive spike-in SNPs and makes the remaining SNPs more reliable. The improvement of filtering can also be observed by comparing the results with previously detected SNPs. The intersection of SNPs called in this study and SNPs found by Lang et al. (2018) shows an increasing number of intersection by applying the three filtering steps (**Supplementary Table S5**). The SNPs found for Re and Ka maybe have been under-estimated by Lang et al. (2018). The accessions Re and Ka have a 10 x lower number of SNPs compared to the accession Vx (Lang et al., 2018). Here, the number of intersecting SNPs between (Lang et al., 2018) and our results shows an almost 90% intersection of Vx SNPs at the strictest filter step. For Re and Ka, the intersection is less than 30% (**Supplementary Table S5**). Potentially, the absence of Re and Ka SNPs in the previous study is a result of sub optimally adjusted filter parameters or it could be an effect of low read coverage. Sufficient read depth at library level, large number of read mapping/coverage and high sequencing quality are major foundations for high quality SNP calling results. In some cases, it is possible that some SNPs were not found in one

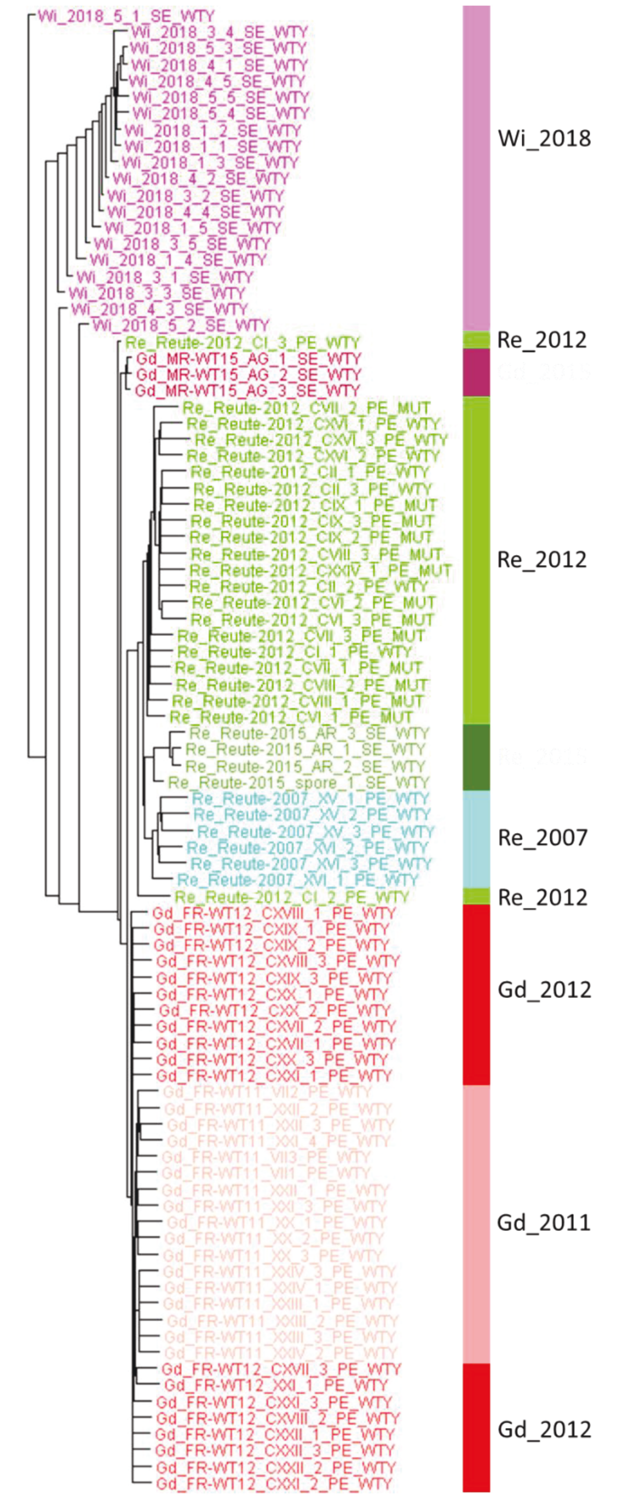

Wi_2018_5_1_SE_WTY
Wi_2018_3_4_SE_WTY
Wi_2018_5_3_SE_WTY
Wi_2018_4_1_SE_WTY
Wi_2018_4_5_SE_WTY
Wi_2018_5_5_SE_WTY
Wi_2018_5_4_SE_WTY
Wi_2018_1_2_SE_WTY
Wi_2018_1_1_SE_WTY
Wi_2018_1_3_SE_WTY
Wi_2018_4_2_SE_WTY
Wi_2018_3_2_SE_WTY
Wi_2018_4_4_SE_WTY
Wi_2018_1_5_SE_WTY
Wi_2018_3_5_SE_WTY
Wi_2018_1_4_SE_WTY
Wi_2018_3_1_SE_WTY
Wi_2018_3_3_SE_WTY
Wi_2018_4_3_SE_WTY
Wi_2018_5_2_SE_WTY
Re_Reute-2012_CI_3_PE_WTY
Gd_MR-WT15_AG_1_SE_WTY
Gd_MR-WT15_AG_2_SE_WTY
Gd_MR-WT15_AG_3_SE_WTY
Re_Reute-2012_CVII_2_PE_MUT
Re_Reute-2012_CXVI_1_PE_WTY
Re_Reute-2012_CXVI_3_PE_WTY
Re_Reute-2012_CXVI_2_PE_WTY
Re_Reute-2012_CII_1_PE_WTY
Re_Reute-2012_CII_3_PE_WTY
Re_Reute-2012_CIX_1_PE_MUT
Re_Reute-2012_CIX_3_PE_MUT
Re_Reute-2012_CIX_2_PE_MUT
Re_Reute-2012_CVIII_3_PE_MUT
Re_Reute-2012_CXXIV_1_PE_MUT
Re_Reute-2012_CII_2_PE_WTY
Re_Reute-2012_CVI_2_PE_MUT
Re_Reute-2012_CVI_3_PE_MUT
Re_Reute-2012_CVII_3_PE_MUT
Re_Reute-2012_CI_1_PE_WTY
Re_Reute-2012_CVII_1_PE_MUT
Re_Reute-2012_CVIII_2_PE_MUT
Re_Reute-2012_CVIII_1_PE_MUT
Re_Reute-2012_CVI_1_PE_MUT
Re_Reute-2015_AR_3_SE_WTY
Re_Reute-2015_AR_1_SE_WTY
Re_Reute-2015_AR_2_SE_WTY
Re_Reute-2015_spore_1_SE_WTY
Re_Reute-2007_XV_1_PE_WTY
Re_Reute-2007_XV_2_PE_WTY
Re_Reute-2007_XV_3_PE_WTY
Re_Reute-2007_XVI_2_PE_WTY
Re_Reute-2007_XVI_3_PE_WTY
Re_Reute-2007_XVI_1_PE_WTY
Re_Reute-2012_CI_2_PE_WTY
Gd_FR-WT12_CXVIII_1_PE_WTY
Gd_FR-WT12_CXIX_1_PE_WTY
Gd_FR-WT12_CXIX_2_PE_WTY
Gd_FR-WT12_CXVIII_3_PE_WTY
Gd_FR-WT12_CXIX_3_PE_WTY
Gd_FR-WT12_CXX_1_PE_WTY
Gd_FR-WT12_CXX_2_PE_WTY
Gd_FR-WT12_CXVII_2_PE_WTY
Gd_FR-WT12_CXVII_1_PE_WTY
Gd_FR-WT12_CXX_3_PE_WTY
Gd_FR-WT12_CXXI_1_PE_WTY
Gd_FR-WT11_VII2_PE_WTY
Gd_FR-WT11_XXII_2_PE_WTY
Gd_FR-WT11_XXII_3_PE_WTY
Gd_FR-WT11_XXI_4_PE_WTY
Gd_FR-WT11_VII3_PE_WTY
Gd_FR-WT11_VII1_PE_WTY
Gd_FR-WT11_XXII_1_PE_WTY
Gd_FR-WT11_XXI_3_PE_WTY
Gd_FR-WT11_XX_1_PE_WTY
Gd_FR-WT11_XX_2_PE_WTY
Gd_FR-WT11_XX_3_PE_WTY
Gd_FR-WT11_XXIV_3_PE_WTY
Gd_FR-WT11_XXIV_1_PE_WTY
Gd_FR-WT11_XXIII_1_PE_WTY
Gd_FR-WT11_XXIII_2_PE_WTY
Gd_FR-WT11_XXIII_3_PE_WTY
Gd_FR-WT11_XXIV_2_PE_WTY
Gd_FR-WT12_CXVII_3_PE_WTY
Gd_FR-WT12_XXI_1_PE_WTY
Gd_FR-WT12_CXXI_3_PE_WTY
Gd_FR-WT12_CXVIII_2_PE_WTY
Gd_FR-WT12_CXXII_1_PE_WTY
Gd_FR-WT12_CXXII_3_PE_WTY
Gd_FR-WT12_CXXII_2_PE_WTY
Gd_FR-WT12_CXXI_2_PE_WTY
Wi_2018
Re_2012
Gd_2015
Re_2012
Re_2015
Re_2007
Re_2012
Gd_2012
Gd_2011
Gd_2012

FIGURE 7 | Splitstree tree of Wisconsin natural population and three generations of Gd and Re. The tree is based on part of the artificial SNP FASTA alignment containing Wi samples (without the bacterial contaminated spore capsule experiment 2) and three generations of Re (2007, 2012 and 2015) and Gd (2011, 2012 and 2015). The Splitstree network tree was branch length-corrected by the maximum number of covered base pairs (see coverage normalization in section "Materials and Methods").

accession or strain because the data available for that position and accession was not enough to detect it in a reliable way. Samples with low read coverage show inconsistency in SNP-to-read correlation (**Supplementary Figures S3–S5**). A reason for this behavior could be non-linear relation between number of SNPs and number of reads for very high and very low read numbers. Samples with a low number of reads can lead to incoherent SNP calling results due to stochastic coverage fluctuation. The high variability in such low read coverage samples can be observed in **Supplementary Figure S5**: the data range of Wi and Vx are wider than all the others. The low number of reads available for the Vx laser capture experiment (BioSample PRJNA602303) is probably related to the RNA-seq extraction technique, yielding small amounts of RNA that might be prone to bias before and/or after amplification.

To reduce the SNP per read effect, we normalized the SNPs by the coverage method, resulting in an observable increase of linear relationship (**Supplementary Figures S3**, **S4**). The number of SNPs called for each sample became more reliable in terms of comparability and reflect well previous studies and expectation of genetic distance coinciding with geographic distance (**Supplementary Figure S2**). The RNA-seq based SNP pipeline described here can in future be applied to stringently call SNPs for *P. patens* accessions and pedigrees, or can be adjusted to suit data sets from other model organisms for which a reference genome or transcriptome is available.

SNP Comparison Between Accessions

When locating the position of the SNPs in the genome, most of them were found in non-coding regions upstream and downstream the gene body (UTRs), as well as in introns and splicing sites within the introns. Many changes were observed in the coding sequences of the five accessions. These changes may lead to alterations in the protein sequence of the final gene product, by changing start or stop codons, or producing frame changes (**Table 1**).

The total number of filtered raw SNPs per accession (**Figure 3**) in comparison to the Gd genome reference shows (as expected) the Gd accessions as the one with the smallest number of changes followed by Re, Vx, Wi and Ka. This order agrees with the distance to the Gd geographical location in the Southeast of England (**Supplementary Figure S2**): Re (Hiss et al., 2017) and Vx (Kamisugi et al., 2008) in close vicinity to each other at the border of France and Southwestern Germany, and Wi and Ka (Perroud et al., 2011) in North America.

Results from Lang et al. (2018), where variance at genomic level was detected using the accessions Re, Vx and Ka, showed a SNP rate of one SNP per 1,783 bp for Re, per 644 bp for Ka and 188 bp for Vx while another study found a SNP rate of one SNP per 207 bp for Vx (Ding et al., 2018). Similar results for the number of base pairs per SNP can be found for the RNA-seq analysis in this study (Re 1 SNP each 1,912 bp, Ka 630 bp and Vx 143 bp) (**Supplementary Figure S5**). The SNP density based on RNA-seq (this study) and gDNA (Lang et al., 2018) is similar, although more SNPs are expected to be detected based on gDNA due to the presence of intergenic regions that are not under selection. This could be another indication of an underestimated SNP number as discussed above. In any case, our method using RNA-seq data for gene space SNP calling yields appropriate results allowing to estimate differences in accessions by SNPs.

We have chosen two different methods to cluster the SNPs related to each sample. An artificial FASTA alignment with all SNPs as well as a matrix including SNPs and indels. Both methods show similar results (**Figure 4** and **Supplementary Figure S9**). The outlier sample Re_CI_3 has a very small read number, probably yielding misleading results. Sample Gd_CIV_1 also appears as an outlier (**Figure 4**). However, in the PCA 3D plot, the sample clusters according to expectation (**Supplementary Figure S9**). Our SNP pipeline had proven its functionality by blind tests as well as by pointing out unexpected metadata errors. The sample Ka_XVIII_1 was re-sequenced to replace a previous Gd experiment in which one of the triplicates failed (Perroud et al., 2018). For this sample, our SNP clustering (**Figure 4**) shows clear evidence for the accession being Ka, not Gd. Indeed, manual checking exclusive SNPs there is no doubt that it is Ka (**Supplementary File 3**, Sheet Ka_exclusive_SNPs). Most probably, the plant material was accidentally mislabeled.

The extraction of exclusive SNP sets for each of the five accessions helps to identify unknown *P. patens* sequences. Here we provide a set of SNPs for all examined accessions that will be useful for molecular identification of accessions. The low number of exclusive Vx SNPs are based on the uniqueness of the single Vx samples. Each Vx sample provided a big list of SNPs, but a high number of these SNPs were only available in one or two other Vx samples. A higher read coverage or more standardized mRNA could solve this issue. For low coverage reasons, we were not using the read coverage filter for the detection of exclusive SNPs. High sample support was chosen as an alternative and promoted exclusive SNP selection in a reasonable way, yielding confirmable molecular identification.

Observed approximate linearity between number of called SNPs and reads per sample (**Supplementary Figure S3**) lead to the read normalization method. When applying the coverage method that takes into account the fraction of the gene space covered by enough reads to allow SNP calling, linearity increased even further (**Supplementary Figure S4**). While both raw and normalized counts lead to the same conclusions in terms of genetic distance, we suggest the coverage normalization to most accurately describe the data.

SNP Comparison of Gransden Pedigrees

Gd is the current reference accession for *P. patens*, and was used to generate the genome sequence (Rensing et al., 2008; Lang et al., 2018). However, over the years of cultivation in the lab, it has shown an accumulation of somatic mutations which was confirmed in this study and observed before, culminating in observable phenotypic changes (Meyberg et al., 2020). One of the characteristics of laboratory models is the capacity to maintain the organism cultivated in the lab for multiple generations, being able to progress through the complete life cycle. The reduction of fertility of Gd accessions in the lab limits experimental design, especially when studying sexual reproduction or when the generation of off-spring is required for the experiments. For this reason, the accession Reute, which shows the lowest number of differences with the Gd genome reference, and which has a much higher fertility than Gd (Meyberg et al., 2020) has been proposed as an alternative to study sexual reproduction (Hiss et al., 2017; Meyberg et al., 2020).

Due to changes in land use, at the original Gransden collection site no *P. patens* can be found anymore. However, phenotypic data suggest that Gransden was not always infertile, because Gd_JP shows intermediate fertility between Re and extant Gd_DE pedigrees (Hiss et al., 2017; Meyberg et al., 2020). Our data show that, as expected, Gd_UK shows the lowest number of SNPs as compared to the reference genome that was derived from Gd_2004 (UK). All other pedigrees show substantial and unique SNPs (**Figure 5** and **Supplementary Figure S11**), demonstrating that during *in vitro* culture somatic mutations occur and accumulate in independent fashion. The practice of regular sexual reproduction of the cultured strains has the advantage that by this procedure it is ensured that the full life cycle can be followed. On top of that there is evidence that even during selfing *P. patens* is able to effectively purge deleterious mutations (Szövényi et al., 2017).

By comparing the normalized gene space SNP count of the Wi natural population samples with those of selfed progressions of Re and Gd laboratory strains we can estimate the genetic variability occurring in natural vs. laboratory samples (**Figure 7** and **Supplementary Figure S18**). Interestingly, the variation of three generations of homozygous (selfed) Re and Gd offspring is similar to that observed in naturally occurring Wi samples (representing the same generation but four spore capsules and five spores each). Based on the normalized data, the three generations of selfed laboratory cultures might even have acquired and retained slightly more mutations than visible in the single Wi natural population. We conclude that a substantial amount of genetic variation occurs both through somatic mutation during vegetative propagation (Meyberg et al., 2020) as well as during sexual propagation by selfing. However, since the practice of regular selfing selects for fertility it seems preferable to follow that practice over exclusive vegetative propagation.

Like for the accessions, specific SNPs for each pedigree were extracted. The diversity of Gd pedigrees is lower than that of the accessions and hence there were not enough samples supporting the same SNP. To detect exclusive SNPs for each pedigree ranking the SNPs by sample support gave us the opportunity to extract the SNPs supported by most of the samples. Obligatory for this method is a correct metadata grouping of the samples. If samples would be described to be the wrong pedigree, exclusive SNPs cannot be accurately determined. Another issue is the sub-clustering of samples. We can observe this for the Gd_JP pedigree as well as for Gd_UK. There are SNPs in the Japan pedigree that occurred in St. Louis, after it was brought to the USA. Our Gd_JP sample set is mostly represented by samples from the USA. Extracted exclusive SNPs with high sample support can thus be scored for the JP- > USA pedigree, but maybe not for the full Gd_JP pedigree. Nevertheless, our provided exclusive SNP list can be used to classify the origin of unknown samples (**Supplementary Figure S17**).

Experimental Confirmation of Selected SNPs

In large experiments that handle many samples, mistakes might occur during the management of the samples in the lab, in the sequencing facility or during later data analysis. The identification of exclusive SNPs in the *P. patens* accessions allows the detection and correction of mistakes in experimental metadata, such as the ones mentioned earlier (**Figure 3**), *in silico*. Moreover, the exclusive SNPs found in the different accessions were used to identify unique targets for restriction enzymes, allowing the development of RFLP assays to differentiate between the *P. patens* accessions. The presence of the predicted SNPs in all tested sequences confirms the successful and stringent SNP selection presented here. The successful establishment of the RFLP analysis for the *P. patens* accessions provides a fast and cheap tool to test the accession background of laboratory strains as well as newly collected *P. patens* accessions. With regard to the Gd pedigrees so far, no SNPs within a restriction enzyme site with enough sample coverage could be identified. However, differentiation between Gd_DE, Gd_JP and Gd_CH could be performed successfully based on the sequencing data (**Supplementary Figure S17**). Thus, SNPs between the Gd pedigrees need to be analyzed via sequencing so far, but including more Gd data sets in the presented approach and/or analyzing a small subset of Gd pedigrees could help to improve and identify SNPs, which could be used within a future RFLP approach to differentiate Gd pedigrees.

Independent of the RFLP method, the origin of *P. patens* plant material can be discovered by using the presented primers (**Supplementary Tables S7, S8**) and sequencing the amplicon. If sequencing data is already available (single fragments, RNA-seq or gDNA sample[s]), our pipeline and the exclusive SNP sets can be used to easily identify plant origins.

Natural Population Variation and Selection

The number of observable mutations on the level of a naturally occurring population (Wi single spore isolates) is in the approximate same range as the mutations occurring in culture undergoing annual sexual reproduction (**Figure 7** and **Supplementary Figure S18**). For samples mainly propagated vegetatively, observed mutations are somatic in nature. For samples that regularly go through sexual reproduction, changes introduced via meiotic recombination cannot be distinguished from somatic changes. Intriguingly, the number of detected SNPs was found to decline over time in samples with a known heritage of regular sexual reproduction (**Supplementary Tables S12, S13**). We take this as evidence that sexual reproduction, even in a haploid, selfing species is able to efficiently purge deleterious mutations, as previously shown (Szövényi et al., 2017).

Consequently, the majority of the observed mutations probably are somatic. The observed mutation rates (changes per year and site) are in the range of 7E-07 to 4E-06 (**Supplementary Table S13**). Studies in other plants found rates in the E-08 range (Hanlon et al., 2019; Schoen and Schultz, 2019). The observed *P. patens* mutation rates are approximately two orders of magnitude higher than the estimated rate of synonymous substitutions per synonymous site per year, 9E-09 (Rensing et al., 2007). Hence, *in vitro* propagation of *P. patens* apparently leads to the fixation of a higher number of mutations than occur naturally, and maybe more than described in other plant propagation systems. Many labs perform regular shredding of protonemal tissue for propagation. This mode of propagation might increase the number of fixed somatic mutations via induction of the DNA repair system through cell damage, potentially resulting in higher mutational load.

The Ka/Ks ratio of the Gransden pedigree generally is below 1, suggesting potential negative (purifying) selection on many loci (**Supplementary Table S11**). All other accessions, to the contrary, exhibit ratios larger than 1, suggesting potential positive (Darwinian) selection. The latter is regardless of whether they are naturally occurring (Wi) or cultured (Ka, Re, Vx). Potentially, the decades-long vegetative culture of Gd, most of it vegetatively, led to the expression of negative selection. All other accessions are much more recent isolates and in particular all Re samples studied went through annual sexual reproduction, which apparently effectively purges deleterious mutations. Interestingly, the GO terms over-represented among those genes affected by non-synonymous changes (**Supplementary Figure S20**) include microtubule-based movement (Re) and reproduction (Vx), fitting recently published data that show these terms contrasted between male infertile Gd and fertile Re (Meyberg et al., 2020). It appears probable that the artificial environment of vegetative *in vitro* Gd propagation led to a loss of fertility due to loss of selection pressure on genes required for sexual reproduction.

CONCLUSION

Our study of sequence variants in *P. patens* laboratory strains revealed the accumulation of somatic mutations over years of cultivation, some of which can be detrimental e.g., with regard to fertility. It appears to be good practice to regularly let the lab cultures reproduce sexually, in order to keep selective pressure and to purge deleterious mutations. Since the original Gd accession is not available any more, and Gd JP shows less fertility than Re, it appears sensible to use Re (with its low number of SNPs as compared to Gd) for any studies that shall involve the life cycle. The identification of exclusive sets of SNPs for *P. patens* laboratory strains and accessions allowed the development of RFLP tests to identify the different accessions. Similarly, Gd pedigrees can be identified by sequencing of PCR products based on the pedigree-exclusive SNPs determined in this study. The variation of selfed laboratory strains is on the same order of magnitude as that of a natural population analyzed.

DATA AVAILABILITY STATEMENT

All RNA-seq samples used in this study are available via the NCBI SRA. Please see **Supplementary Table S1** in Supplementary.pdf for more details.

AUTHOR CONTRIBUTIONS

FH analyzed the raw read data and performed SNP calling. DS-M, JL, P-FP, and RM contributed RNA-seq data. FH, NS, and RM setup and performed RFLP as well as sequencing analyses. SR conceived of the study and supervised it together with NF-P and P-FP. FH, NF-P, RM, and SR wrote the manuscript with the help of all authors.

FUNDING

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. SR is grateful for funding by Deutsche Forschungsgemeinschaft (Grant Number DFG RE 1697/15-1). DS-M was supported by ERC Advanced Grant (EDIP) and BBSRC grant BB/M020517/1 to JL.

ACKNOWLEDGMENTS

We are grateful to Ralph S. Quatrano and to Monsanto (St. Louis, MO, United States) for sequencing RNA-seq data of *P. patens* accession Kaskaskia, to Rafael Medina for samples of the *P. patens* accession Wisconsin, and to Stuart McDaniel for helpful comments. We also thank A. G. Goesmann, Justus-Liebig-University, Giessen for access to the de.NBI infrastructure.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.00813/full#supplementary-material

REFERENCES

Ashton, N. W., and Cove, D. J. (1977). The isolation and preliminary characterisation of auxotrophic and analogue resistant mutants of the moss, *Physcomitrella patens*. *Mol. Gen. Genet.* 154, 87–95. doi: 10.1007/bf00265581

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. doi: 10.1179/jbr.2000.22.1.9

Beike, A. K., von Stackelberg, M., Schallenberg-Rüdinger, M., Hanke, S. T., Follo, M., Quandt, D., et al. (2014). Molecular evidence for convergent evolution and allopolyploid speciation within the *Physcomitrium*-Physcomitrellaspecies complex. *BMC Evol. Biol.* 14:158. doi: 10.1186/1471-2148-14-158

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological).* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England).* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32, 314–331.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., et al. (2014). BEAST 2: a software platform for bayesian evolutionary analysis. *PLoS Comput. Biol.* 10:e1003537. doi: 10.1371/journal.pcbi.1003537

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., et al. (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* 6, 80–92. doi: 10.4161/fly.19695

Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics (Oxford, England).* 33, 2938–2940. doi: 10.1093/bioinformatics/btx364

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). Isolation of DNA, RNA, and protein from the moss *Physcomitrella patens* gametophytes. *Cold Spring Harb. Protoc.* 2009:db.rot5146.

de Vries, J., and Rensing, S. A. (2020). Gene gains paved the path to land. *Nat. Plants* 6, 7–8. doi: 10.1038/s41477-019-0579-5

Demko, V., Perroud, P.-F., Johansen, W., Delwiche, C. F., Cooper, E. D., Remme, P., et al. (2014). Genetic analysis of DEFECTIVE KERNEL1 loop function in three-dimensional body patterning in *Physcomitrella patens*. *Plant Physiol.* 166, 903–919. doi: 10.1104/pp.114.243758

Ding, X., Pervere, L. M., Bascom, C. Jr., Bibeau, J. P., Khurana, S., Butt, A. M., et al. (2018). Conditional genetic screen in *Physcomitrella patens* reveals a novel microtubule depolymerizing-end-tracking protein. *PLoS Genet.* 14:e1007221. doi: 10.1371/journal.pgen.1007221

Engel, P. P. (1968). The induction of biochemical and morphological mutants in the moss *Physcomitrella patens*. *Am. J. Bot.* 55, 438–446. doi: 10.1002/j.1537-2197.1968.tb07397.x

Fernandez-Pozo, N., Haas, F. B., Meyberg, R., Ullrich, K. K., Hiss, M., Perroud, P.-F., et al. (2019). PEATmoss (*Physcomitrella* expression atlas tool): a unified gene expression atlas for the model plant *Physcomitrella patens*. *Plant J.* 102, 165–177. doi: 10.1111/tpj.14607

Flowers, J. M., Hazzouri, K. M., Pham, G. M., Rosas, U., Bahmani, T., Khraiwesh, B., et al. (2015). Whole-genome resequencing reveals extensive natural variation in the model green Alga *Chlamydomonas reinhardtii*. *Plant Cell* 27, 2353–2369. doi: 10.1105/tpc.15.00492

Frank, M. H., and Scanlon, M. J. (2015). Cell-specific transcriptomic analyses of three-dimensional shoot development in the moss *Physcomitrella patens*. *Plant J.* 83, 743–751. doi: 10.1111/tpj.12928

Hanlon, V. C. T., Otto, S. P., and Aitken, S. N. (2019). Somatic mutations substantially increase the per-generation mutation rate in the conifer *Picea sitchensis*. *Evol. Lett.* 3, 348–358. doi: 10.1002/evl3.121

Hiss, M., Meyberg, R., Westermann, J., Haas, F. B., Schneider, L., Schallenberg-Rdinger, M., et al. (2017). Sexual reproduction, sporophyte development and molecular variation in the model moss *Physcomitrella patens*: introducing the ecotype Reute. *Plant J. Cell Mol. Biol.* 90, 606–620. doi: 10.1111/tpj.13501

Huson, D. H., and Bryant, D. (2005). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267. doi: 10.1093/molbev/msj030

Kamisugi, Y., Von Stackelberg, M., Lang, D., Care, M., Reski, R., Rensing, S. A., et al. (2008). A sequence-anchored genetic linkage map for the moss, *Physcomitrella patens*. *Plant J.* 56, 855–866. doi: 10.1111/j.1365-313x.2008.03637.x

Krasovec, M., Eyre-Walker, A., Sanchez-Ferandin, S., and Piganeau, G. (2017). Spontaneous mutation rate in the smallest photosynthetic eukaryotes. *Mol. Biol. Evol.* 34, 1770–1779. doi: 10.1093/molbev/msx119

Lang, D., Ullrich, K. K., Murat, F., Fuchs, J., Jenkins, J., Haas, F. B., et al. (2018). The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution. *Plant J. Cell Mol. Biol.* 93, 515—-533.

Leaché, A. D., and Oaks, J. R. (2017). The utility of single nucleotide polymorphism (SNP) data in phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* 48, 69–84. doi: 10.1146/annurev-ecolsys-110316-022645

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Ma, X., Shao, Y., Tian, L., Flasch, D. A., Mulder, H. L., Edmonson, M. N., et al. (2019). Analysis of error profiles in deep next-generation sequencing data. *Genome Biol.* 20:50.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

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.* 57, 404–417. doi: 10.1111/jse.12516

Meyberg, R., Perroud, P.-F., Haas, F. B., Schneider, L., Heimerl, T., Renzaglia, K. S., et al. (2020). Characterization of evolutionarily conserved key players affecting eukaryotic flagellar motility and fertility using a moss model. *New Phytol.* doi: 10.1111/nph.16486 [Epub ahead of print].

Moody, L. A., Kelly, S., Rabbinowitsch, E., and Langdale, J. A. (2018). Genetic regulation of the 2D to 3D growth transition in the moss *Physcomitrella patens*. *Curr. Biol.* 28, 473–478.e5. doi: 10.1016/j.cub.2017.12.052

Nguyen, T.-P., Muhlich, C., Mohammadin, S., van den Bergh, E., Platts, A. E., Haas, F. B., et al. (2019). Genome improvement and genetic map construction for aethionema arabicum, the first divergent branch in the Brassicaceae family. *G3 (Bethesda, Md).* 9, 3521–3530. doi: 10.1534/g3.119.400657

Niu, S., Song, Q., Koiwa, H., Qiao, D., Zhao, D., Chen, Z., et al. (2019). Genetic diversity, linkage disequilibrium, and population structure analysis of the tea plant (*Camellia sinensis*) from an origin center, Guizhou plateau, using genome-wide SNPs developed by genotyping-by-sequencing. *BMC Plant Biol.* 19:328. doi: 10.1186/s12870-019-1917-5

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. doi: 10.1111/j.1469-8137.2011.03668.x

Perroud, P.-F., Haas, F. B., Hiss, M., Ullrich, K. K., Alboresi, A., Amirebrahimi, M., et al. (2018). The *Physcomitrella patens* gene atlas project: large scale RNA-seq based expression data. *Plant J.* 95, 168–182. doi: 10.1111/tpj.13940

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.* 21, 143–149. doi: 10.1111/plb.12840

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England).* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rensing, S. A. (2018). Great moments in evolution: the conquest of land by plants. *Curr. Opin. Plant Biol.* 42, 49–54. doi: 10.1016/j.pbi.2018.02.006

Rensing, S. A., Beike, A. K., and Lang, D. (2013). "Evolutionary importance of generative polyploidy for genome evolution of haploid-dominant land plants," in *Plant Genome Diversity : Physical Structure, Behaviour and Evolution of Plant Genomes*, Vol. 2, eds I. J. Leitch, J. Greilhuber, D. Jaroslav, and W. Jonathan (Vienna: Springer-Verlag), 295–305. doi: 10.1007/978-3-7091-1160-4_18

Rensing, S. A., Goffinet, B., Meyberg, R., Wu, S.-Z., and Bezanilla, M. (2020). The moss *Physcomitrium* (*Physcomitrella*) *patens*: a model organism for non-seed plants. *Plant Cell* 32, 1361–1376. doi: 10.1105/tpc.19.00828

Rensing, S. A., Ick, J., Fawcett, J. A., Lang, D., Zimmer, A., Van de Peer, Y., et al. (2007). An ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*. *BMC Evol. Biol.* 7:130. doi: 10.1186/1471-2148-7-130

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

Ribeiro, A., Golicz, A., Hackett, C. A., Milne, I., Stephen, G., Marshall, D., et al. (2015). An investigation of causes of false positive single nucleotide polymorphisms using simulated reads from a small eukaryote genome. *BMC Bioinformatics* 16:382. doi: 10.1186/s12859-015-0801-z

Saint-Marcoux, D., Billoud, B., Langdale, J. A., and Charrier, B. (2015). Laser capture microdissection in *Ectocarpus siliculosus*: the pathway to cell-specific transcriptomics in brown algae. *Front. Plant Sci.* 6:54. doi: 10.3389/fpls.2015.00054

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. doi: 10.1093/bioinformatics/btr026

Schoen, D. J., and Schultz, S. T. (2019). Somatic mutation and evolution in plants. *Annu. Rev. Ecol. Evol. Syst.* 50, 49–73. doi: 10.1146/annurev-ecolsys-110218-024955

Schween, G., Schulte, J., Reski, R., and Hohe, A. (2005). Effect of ploidy level on growth, differentiation, and morphology in *Physcomitrella patens*. *Bryologist.* 108, 27–35. doi: 10.1639/0007-2745(2005)108[27:eoplog]2.0.co;2

Smits, W. K. (2017). SNP-ing out the differences: investigating differences between *Clostridium* difficile lab strains. *Virulence* 8, 613–617. doi: 10.1080/21505594.2016.1250998

Stevenson, S. R., Kamisugi, Y., Trinh, C. H., Schmutz, J., Jenkins, J. W., Grimwood, J., et al. (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. *Plant Cell* 28, 1310–1327.

Szövényi, P., Ullrich, K. K., Rensing, S. A., Lang, D., van Gessel, N., Stenøien, H. K., et al. (2017). Selfing in haploid plants and efficacy of selection: codon usage bias in the model moss *Physcomitrella patens*. *Genome Biol. Evol.* 9, 1528–1546. doi: 10.1093/gbe/evx098

Vashisht, D., Hesselink, A., Pierik, R., Ammerlaan, J. M., Bailey-Serres, J., Visser, E. J., et al. (2011). Natural variation of submergence tolerance among *Arabidopsis thaliana* accessions. *New Phytol.* 190, 299–310. doi: 10.1111/j.1469-8137.2010.03552.x

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. doi: 10.1186/1471-2229-6-9

Wang, S., Meyer, E., McKay, J. K., and Matz, M. V. (2012). 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nat. Methods* 9, 808–810. doi: 10.1038/nmeth.2023

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. doi: 10.1111/tpj.12542

Wu, T. D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26, 873–881. doi: 10.1093/bioinformatics/btq057

Xia, W., Luo, T., Zhang, W., Mason, A. S., Huang, D., Huang, X., et al. (2019). Development of high-density SNP markers and their application in evaluating genetic diversity and population structure in *Elaeis guineensis*. *Front. Plant Sci.* 10:130. doi: 10.3389/fpls.2019.00130

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6 Concluding remarks

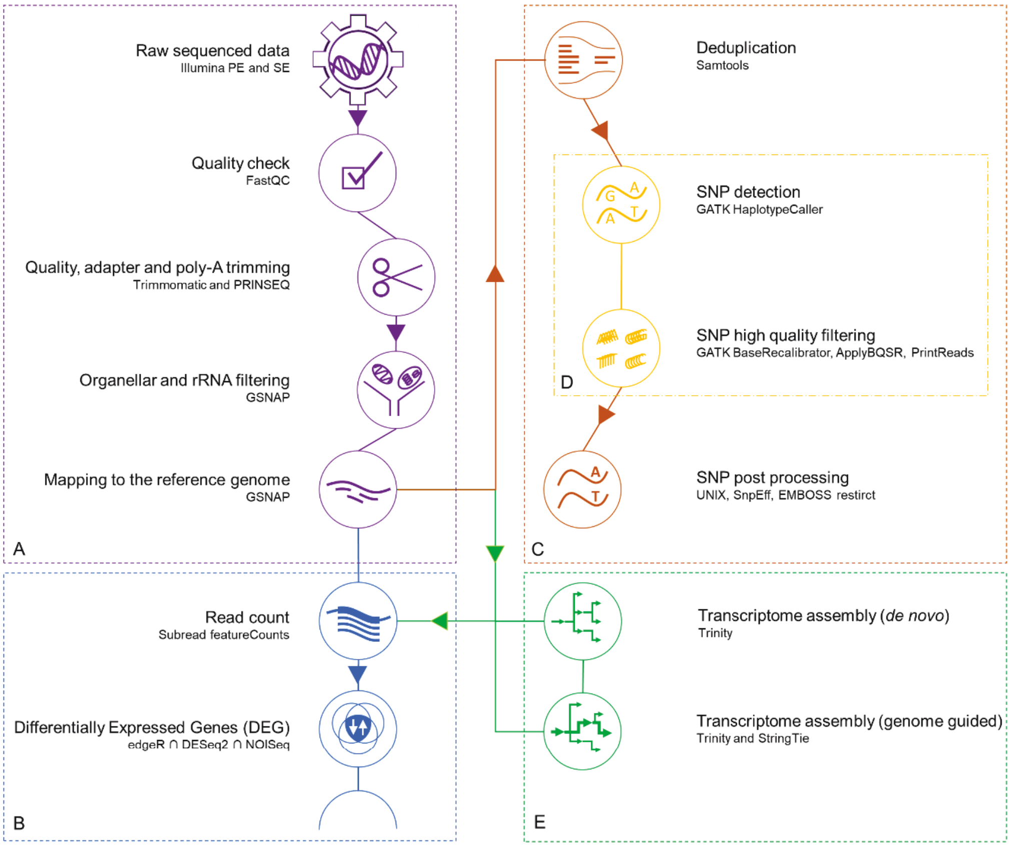
RNA-seq data analysis was the basis for the major parts of this thesis. The results achieved in this work improved not only the understanding of gene functionality of the model organism *Physcomitrium patens*, but also methods like a user-friendly pipeline to analyse terabytes of RNA-seq data were developed and efficiently used. High resolution gene expression studies, current as well as additional gene model annotations benefit from these developments. During the last two years, the RNA-seq pipeline was tested in several studies and always demonstrated high performance (Lanver *et al.*, 2018, Chandler *et al.*, 2020, Meyberg *et al.*, 2020) (Not yet published: *Ustilago maydis* light response project). Another tool, PEATmoss, simplifies the data accessibility and helps to keep track of currently published and upcoming expression experiments. The implemented gene lookup table enables gene annotation version conversion. Novel designed pipeline extensions detected genetic variation. Unique SNP sets of five different *P. patens* accessions assist the community to determine their plant's origin. Therefore, the RFLP toolkit with predesigned primer and enzymes is available. The genetic variation study showed that somatic mutation occurs and accumulates in Gransden pedigrees.

Not only the *P. patens* community benefits from the established DEG and SNP pipelines, in its methods and results, but the extremely flexible and modular implementation opens the field for a broader usage in variety of organisms to analyse sequencing data that help researchers worldwide to contribute high quality studies.

7 Outlook

7.1 RNA-seq pipeline

Over the past few years, several RNA-seq pipelines were published. Specific tools and platforms were developed to perform RNA-seq analysis. Each pipeline approach has its own specific niche. Pipelines developed by (Goncalves *et al.*, 2011, Varet *et al.*, 2016) are designed for the R environment. (Alonso *et al.*, 2017) prepared a pipeline for high-performance computing clusters and (D'Antonio *et al.*, 2015) published a cloud-web-based RNA-seq pipeline. Not only does the platform or the interface differ, but the chosen tools that perform the individual subfunctions in the pipeline are functional. While the RNA-seq pipeline PRADA by (Torres-García *et al.*, 2014) uses the standard UNIX environment just like our pipeline, the programs to perform the individual tasks are built differently. There is not one 'correct' or best decision to choose among pipelines, but it is rather a fine balance of individual needs. For example: "Does the dataset consist of paired-end or single-end libraries?", "Is it pure RNA-seq or with additional gDNA samples?" "Is the main focus on speed or accuracy?" are some of the questions that users are faced with.



Raw sequenced data
Illumina PE and SE
Quality check
FastQC
Quality, adapter and poly-A trimming
Trimmomatic and PRINSEQ
Organellar and rRNA filtering
GSNAP
Mapping to the reference genome
GSNAP
A
Read count
Subread featureCounts
Differentially Expressed Genes (DEG)
edgeR ∩ DESeq2 ∩ NOISeq
B
Deduplication
Samtools
SNP detection
GATK HaplotypeCaller
G
A
A
T
SNP high quality filtering
GATK BaseRecalibrator, ApplyBQSR, PrintReads
D
SNP post processing
UNIX, SnpEff, EMBOSS restirct
A
T
C
Transcriptome assembly (*de novo*)
Trinity
Transcriptome assembly (genome guided)
Trinity and StringTie
E

Figure 7: Beta version of the RNA-seq pipeline. *A) The initial part represents quality control and filter steps. The branch shown in B) generates expression data and calls DEGs. In C) sequence variation is detected. The procedure of SNP calling is shown in D). The most recent and unpublished branch at E) is the transcriptome assembly part.*

Here, my designed RNA-seq pipeline results are accurate and easy to maintain. After the initial release by (Perroud *et al.*, 2018), I continuously updated the pipeline. The new release was published by (Haas *et al.*, 2020). The most recent version of the RNA-seq pipeline is shown in Figure 7. It is still in the beta phase since not all components were fully tested for stability and accuracy. The new branch, shown in Figure 7 E, describes transcriptome assembly, genome-guided and *de novo*. This part can loop back to the expression data segment (Figure 7, B) as well as to the filter step (Figure 7, A). Our pipeline is prepared for future tasks like long read performance or BS-seq analysis.

7.2 Additional accessions and sequence variation

In this thesis, five *P. patens* accessions were shown. However, there are more than these five *P. patens* accessions available. The International Moss Stock Center (IMSC) currently offers more than 14 different *P. patens* accessions (https://www.moss-stock-center.org/fileadmin/user_upload/pdf/IMSC-wildtypes.pdf). Since the accession Gransden shows a loss of fertilization, new fertile accessions are needed (Hiss *et al.*, 2017, Meyberg *et al.*, 2020). Sequence variation studies between all accessions could finalise my population studies and improve the results shown in chapter 5.4. Of course, significantly more sequencing data would be required for such analyses. In general, more sequencing data results in a higher read depth and more sequence coverage. Both, high read depth and sequence coverage, are obligatory for reliable results, in any sequence study. This data can be used to observe the variation of different *P. patens* populations. Almost all published PCR primers or predicted restriction enzyme cutting sites were designed and performed with the Gransden accession or the Gransden reference genome. To avoid misleading results with these previously published primers and enzymes in future studies, we recommend using our methods and tools to determine SNPs located at targeted regions.

By analysing sequence variation in multiple *P. patens* RNA-seq samples, we were able to answer a wide variety of questions. For instance, we detected sequence variation in Gransden pedigrees and estimated the annual number of mutations per base pair. Nevertheless, there are questions we did not answer. A follow-up project could be a mutation accumulation (MA) experiment. In such MA experiments, the effect of natural selection is minimised by maintaining isolated and inbred lines (MA lines) of organisms (Halligan and Keightley, 2009). Gransden could represent such a MA line. In contrast to Gransden, Wisconsin samples could be used to examine natural population studies.

Another possible project related to the sequence variation analysis for the RNA-seq data could be the study of alternative codon usage. (Szövényi *et al.*, 2017) used RNA-seq and microarray data to identify codon usage bias in *P. patens*. The presented results propose that synonymous codon usage is mainly driven by nucleotide compositional biases (Szövényi *et al.*, 2017). This results are opposite of what (Stenøien, 2005) showed in his study about the adaptive codon usage bias. By running analyses with the RNA-seq data presented in my work and the methods shown by (Szövényi *et al.*, 2017) we could identify the codon usage related not only for one, but for five different *P. patens* accessions and Gransden pedigrees.

7.3 Genome annotation improvement

Since the JGI generated the first draft of the *P. patens* genome version 5 (V5) in 2017, we have finished the updated genome version V5.1. Major improvements for V5.1, compared to V3, are new arranged inner chromosomal structure and advanced gap closing. Expression data presented in this work will be used to perform the gene annotation upgrade.

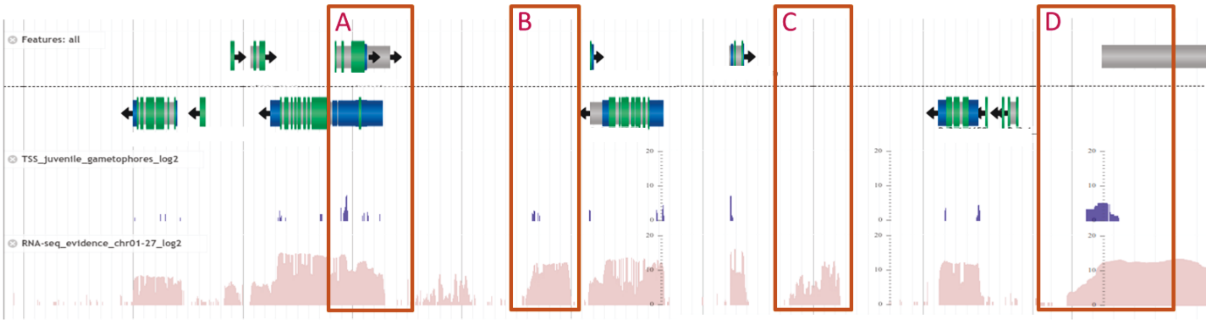

A
B
C
D
Features: all
TSS_juvenile_gametophores_log2
RNA-seq_evidence_chr01-27_log2
20
10
0

Figure 8: CoGe JBrowse screen shot of P. patens v3.3 gene models and upgradable sites . *Three experiment tracks are visible in three rows. The first row shows Pp v3.3 gene models, the second row displays transcription start sites (TSS) data and the last row shows RNA-seq evidence. A) shows overlapping gene models. In B) a missing gene annotation is highlighted. TSS and RNA-seq evidence is present. C) shows RNA-seq evidence, by no gene annotation. An example for potential gene model correction is highlighted in D). The v3.3 gene model starts to late, after the TSS and RNA-seq evidence. The scale of gene models in A, B, C and D is not uniform.*

Several known issues were detected at the v3.3 gene model (Figure 8). For example, there are more than 1,000 gene models that overlap with other gene models (Figure 8, A). Multiple gene models are located on scaffolds that are not part of the 330 main scaffolds. Start/stop codon errors occur and frameshifts within exons are present. Additionally, missing gene models (Figure 8, B and C) and incorrect gene length (Figure 8, D) need to be fixed. Expression data presented in this work will be included to fix a lot of the known errors. Gene annotation v3.3 will be upgraded to v5. Upcoming sequencing projects will not only generate RNA-seq produced by Illumina sequencing. Long read sequencing, such as PacBio SMRT or Oxford Nanopores MinION, will lead to more precise gene models and corresponding isoforms (derived from alternative splicing). Long read sequencing will help to improve the v5 gene annotation, as well as the V5.1 genome. More gaps can be closed and scaffolds can be linked to each other. Furthermore, the pseudo-chromosomal structure of the V3 and V5.1 genome could become more enlightened.

7.4 Gene set normalization

PEATmoss is missing a useful gene set normalization implementation. However, the tool was published before the gene set normalization for comparing expression data across platforms (Supplementary 9.3.2) was completely developed. In general, it is not possible to directly compare microarray and RNA-seq expression values. Due to different designs and normalization methods, the average of normalized RNA-seq expression values is a magnitude of 10 to 10,000 smaller than array expression values. Reference genes, known from qPCR quantification (Pfaffl, 2001), are a feasible option to set up comparable values. The gene set presented (Supporting information 9.3.2) was designed to compare the full microarray and RNA-seq datasets. The adequate implementation of the method would allow PEATmoss to set individual gene sets for customised datasets. The implementation of this method is in progress.

8 References

Alonso, A., Lasseigne, B.N., Williams, K., Nielsen, J., Ramaker, R.C., Hardigan, A.A., Johnston, B., Roberts, B.S., Cooper, S.J., Marsal, S. and Myers, R.M. (2017) aRNApipe: a balanced, efficient and distributed pipeline for processing RNA-seq data in high-performance computing environments. *Bioinformatics (Oxford, England)*, **33**, 1727-1729.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, **25**, 3389-3402.

Arif, M.A., Hiss, M., Tomek, M., Busch, H., Meyberg, R., Tintelnot, S., Reski, R., Rensing, S.A. and Frank, W. (2019) ABA-Induced Vegetative Diaspore Formation in *Physcomitrella patens*. *Frontiers in Plant Science*, **10**.

Ashton, N.W. and Cove, D.J. (1977) The isolation and preliminary characterisation of auxotrophic and analogue resistant mutants of the moss, *Physcomitrella patens*. *Molecular and General Genetics MGG*, **154**, 87-95.

Axtell, M.J. (2013) ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA*, **19**, 740-751.

Beike, A.K., Lang, D., Zimmer, A.D., Wüst, 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 Phytologist*, **205**, 869-881.

Beike, A.K., von Stackelberg, M., Schallenberg-Rüdinger, 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 Evolutionary Biology*, **14**, 158.

Bennett, S. (2004) Solexa Ltd. *Pharmacogenomics*, **5**, 433-438.

Buels, R., Yao, E., Diesh, C.M., Hayes, R.D., Munoz-Torres, M., Helt, G., Goodstein, D.M., Elsik, C.G., Lewis, S.E., Stein, L. and Holmes, I.H. (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biology*, **17**, 66.

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**, 13.

Chandler, J.O., Haas, F.B., Khan, S., Bowden, L., Ignatz, M., Enfissi, E.M.A., Gawthrop, F., Griffiths, A., Fraser, P.D., Rensing, S.A. and Leubner-Metzger, G. (2020) Rocket Science: The Effect of Spaceflight on Germination Physiology, Ageing, and Transcriptome of *Eruca sativa* Seeds. *Life*, **10**, 49.

Coruh, C., Cho, S.H., Shahid, S., Liu, Q., Wierzbicki, A. and Axtell, M.J. (2015) Comprehensive Annotation of *Physcomitrella patens* Small RNA Loci Reveals That the Heterochromatic Short Interfering RNA Pathway Is Largely Conserved in Land Plants. *Plant Cell*, **27**, 2148-2162.

Cove, D. (2005) The moss *Physcomitrella patens*. *Annual review of genetics*, **39**, 339-358.

D'Antonio, M., D'Onorio De Meo, P., Pallocca, M., Picardi, E., D'Erchia, A.M., Calogero, R.A., Castrignanò, T. and Pesole, G. (2015) RAP: RNA-Seq Analysis Pipeline, a new cloud-based NGS web application. *BMC Genomics*, **16**, S3.

de Vries, J. and Rensing, S.A. (2020) Gene gains paved the path to land. *Nature Plants*, **6**, 7-8.

Ding, X., Pervere, L.M., Bascom, C., Jr., Bibeau, J.P., Khurana, S., Butt, A.M., Orr, R.G., Flaherty, P.J., Bezanilla, M. and Vidali, L. (2018) Conditional genetic screen in *Physcomitrella patens* reveals a novel microtubule depolymerizing-end-tracking protein. *PLOS Genetics*, **14**, e1007221.

**Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham,

M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. and Turner, S. (2009) Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, **323**, 133-138.

Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R. and Ashburner, M. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol*, **6**, R44.

Engel, P.P. (1968) The Induction of Biochemical and Morphological Mutants in the Moss *Physcomitrella patens*. *American Journal of Botany*, **55**, 438-446.

Evans, C., Hardin, J. and Stoebel, D.M. (2018) Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief Bioinform*, **19**, 776-792.

Fernandez-Pozo, N., Haas, F.B., Meyberg, R., Ullrich, K.K., Hiss, M., Perroud, P.-F., Hanke, S., Kratz, V., Powell, A.F., Vesty, E.F., Daum, C.G., Zane, M., Lipzen, A., Sreedasyam, A., Grimwood, J., Coates, J.C., Barry, K., Schmutz, J., Mueller, L.A. and Rensing, S.A. (2019) PEATmoss (Physcomitrella Expression Atlas Tool): a unified gene expression atlas for the model plant *Physcomitrella patens*. *The Plant Journal*, **102**, 165-177.

Goncalves, A., Tikhonov, A., Brazma, A. and Kapushesky, M. (2011) A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics (Oxford, England)*, **27**, 867-869.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N. and Rokhsar, D.S. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*, **40**, D1178-1186.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., Salzberg, S.L. and White, O. (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*, **31**, 5654-5666.

Haas, F.B., Fernandez-Pozo, N., Meyberg, R., Perroud, P.-F., Göttig, M., Stingl, N., Saint-Marcoux, D., Langdale, J. and Rensing, S.A. (2020) Single nucleotide polymorphism charting of *P. patens* reveals accumulation of somatic mutations during in vitro culture on the scale of natural variation by selfing. *Front. Plant Sci.*

Halligan, D.L. and Keightley, P.D. (2009) Spontaneous Mutation Accumulation Studies in Evolutionary Genetics. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 151-172.

Hedges, S.B. (2002) The origin and evolution of model organisms. *Nature Reviews Genetics*, **3**, 838-849.

Hedwig, J. and Schwägrichen, C.F. (1801) *Species muscorum frondosorum : descriptae et tabulis aeneis IXXVII coloratis illustratae* Lipsiae (Leipzig) :: sumtu J. A. Barthii ;.

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-Rüdinger, M., Szövényi, P., Tiko, T., Wiedemann, G., Wolf, L., Zimmermann, P. and 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-Rdinger, 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. *The Plant journal : for cell and molecular biology*, **90**, 606-620.

Jain, M., Olsen, H.E., Paten, B. and Akeson, M. (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, **17**, 239.

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*. *The Plant Journal*, **56**, 855-866.

Kersey, P.J., Allen, J.E., Armean, I., Boddu, S., Bolt, B.J., Carvalho-Silva, D., Christensen, M., Davis, P., Falin, L.J., Grabmueller, C., Humphrey, J., Kerhornou, A., Khobova, J., Aranganathan, N.K., Langridge, N., Lowy, E., McDowall, M.D., Maheswari, U., Nuhn, M., Ong, C.K., Overduin, B., Paulini, M., Pedro, H., Perry, E., Spudich, G., Tapanari, E., Walts, B., Williams, G., Tello-Ruiz, M., Stein, J., Wei, S., Ware, D., Bolser, D.M., Howe, K.L., Kulesha, E., Lawson, D., Maslen, G.

and Staines, D.M. (2016) Ensembl Genomes 2016: more genomes, more complexity. *Nucleic acids research*, **44**, D574-D580.

Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T. and Ussery, D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res*, **35**, 3100-3108.

Lang, D., Ullrich, K.K., Murat, F., Fuchs, J., Jenkins, J., Haas, F.B., Piednoel, M. and Gundlach, H.a. (2018) The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution. *The Plant journal : for cell and molecular biology*, **93**, 515--533.

Lanver, D., Muller, A.N., Happel, P., Schweizer, G., Haas, F.B., Franitza, M., Pellegrin, C., Reissmann, S., Altmuller, J., Rensing, S.A. and Kahmann, R. (2018) The Biotrophic Development of *Ustilago maydis* Studied by RNA-Seq Analysis. *The Plant cell*, **30**, 300--323.

Laslett, D. and Canback, B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res*, **32**, 11-16.

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.

Lyons, E. and Freeling, M. (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal*, **53**, 661-673.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J.-B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. and Rothberg, J.M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376-380.

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.

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*. *Journal of Systematics and Evolution*, **57**, 404-417.

Meyberg, R., Perroud, P.-F., Haas, F.B., Schneider, L., Heimerl, T., Renzaglia, K.S. and Rensing, S.A. (2020) Characterization of evolutionarily conserved key players affecting eukaryotic flagellar motility and fertility using a moss model. *New Phytologist*.

Mitten, W. (1851) *A List of all the Mosses and Hepaticae hitherto observed in Sussex*: London : Pinted and published by Richard Taylor. [...].

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, **5**, 621-628.

Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J. and Finn, R.D. (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res*, **43**, D130-137.

Nawrocki, E.P. and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933-2935.

Nishiyama, T., Sakayama, H., de Vries, J., Buschmann, H., Saint-Marcoux, D., Ullrich, K.K., Haas, F.B., Vanderstraeten, L., Becker, D., Lang, D., Vosolsobe, S., Rombauts, S., Wilhelmsson, P.K.I., Janitza, P., Kern, R., Heyl, A., Rumpler, F., Villalobos, L.I.A.C., Clay, J.M., Skokan, R., Toyoda, A., Suzuki, Y., Kagoshima, H., Schijlen, E., Tajeshwar, N., Catarino, B., Hetherington, A.J., Saltykova, A., Bonnot, C., Breuninger, H., Symeonidi, A. and Radhakrishnan, G.V.a. (2018) The *Chara* Genome: Secondary Complexity and Implications for Plant Terrestrialization. *Cell*, **174**, 448-464.

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*. *Journal of Experimental Botany*, **64**, 3567-3581.

Ortiz-Ramírez, C., Hernandez-Coronado, M., Thamm, A., Catarino, B., Wang, M., Dolan, L., Feijó, José A. and Becker, Jörg D. (2016) A Transcriptome Atlas of *Physcomitrella patens* Provides Insights into the Evolution and Development of Land Plants. *Molecular 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 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., Nogue, F. and Oliver, M.J.a. (2018) The *Physcomitrella patens* gene atlas project: large scale RNA-seq based expression data. *The Plant journal : for cell and molecular biology*, **95**, 168-182.

Pfaffl, M.W. (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic acids research*, **29**, e45-e45.

Possart, A., Xu, T., Paik, I., Hanke, S., Keim, S., Hermann, H.-M., Wolf, L., Hiß, M., Becker, C., Huq, E., Rensing, S.A. and Hiltbrunner, A. (2017) Characterization of Phytochrome Interacting Factors from the Moss *Physcomitrella patens* Illustrates Conservation of Phytochrome Signaling Modules in Land Plants. *The Plant cell*, **29**, 310-330.

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

Qin, S., Kim, J., Arafat, D. and Gibson, G. (2013) Effect of normalization on statistical and biological interpretation of gene expression profiles. *Front Genet*, **3**, 160-160.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glockner, F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*, **41**, D590-596.

Rensing, S.A., Goffinet, B., Meyberg, R., Wu, S.-Z. and Bezanilla, M. (2020) The Moss *Physcomitrium* (*Physcomitrella*) *patens*: A Model Organism for Non-Seed Plants. *The Plant Cell*, **32**, 1361-1376.

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.-i., 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. and Boore, J.L. (2008) The *Physcomitrella* Genome Reveals Evolutionary Insights into the Conquest of Land by Plants. *Science*, **319**, 64-69.

Reski, R., Bae, H. and Simonsen, H.T. (2018) *Physcomitrella patens*, a versatile synthetic biology chassis. *Plant Cell Reports*, **37**, 1409-1417.

Salamov, A.A. and Solovyev, V.V. (2000) Ab initio gene finding in Drosophila genomic DNA. *Genome Res*, **10**, 516-522.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M. and Smith, M. (1977) Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature*, **265**, 687-695.

Schaefer, D.G. and Zrÿd, J.-P. (2001) The Moss *Physcomitrella patens*, Now and Then. *Plant Physiology*, **127**, 1430-1438.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D. and Church, G.M. (2005) Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science*, **309**, 1728-1732.

Skinner, M.E., Uzilov, A.V., Stein, L.D., Mungall, C.J. and Holmes, I.H. (2009) JBrowse: a next-generation genome browser. *Genome research*, **19**, 1630-1638.

Smit, A.F.A., Hubley, R. and Green, P. (1996) RepeatMasker Open-3.0. *URL* *http://www. repeatmasker. org.(unpublished)*, **2004**.

St. John, T.P. and Davis, R.W. (1979) Isolation of galactose-inducible DNA sequences from Saccharomyces cerevisiae by differential plaque filter hybridization. *Cell*, **16**, 443-452.

Stenøien, H.K. (2005) Adaptive basis of codon usage in the haploid moss *Physcomitrella patens*. *Heredity*, **94**, 87-93.

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. and 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.

Szövényi, P., Ullrich, K.K., Rensing, S.A., Lang, D., van Gessel, N., Stenøien, H.K., Conti, E. and Reski, R. (2017) Selfing in Haploid Plants and Efficacy of Selection: Codon Usage Bias in the Model Moss *Physcomitrella patens*. *Genome Biology and Evolution*, **9**, 1528-1546.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., Lao, K. and Surani, M.A. (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, **6**, 377-382.

Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A.D., Nueda, M.J., Ferrer, A. and Conesa, A. (2015) Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic acids research*, **43**, e140-e140.

Torres-García, W., Zheng, S., Sivachenko, A., Vegesna, R., Wang, Q., Yao, R., Berger, M.F., Weinstein, J.N., Getz, G. and Verhaak, R.G.W. (2014) PRADA: pipeline for RNA sequencing data analysis. *Bioinformatics (Oxford, England)*, **30**, 2224-2226.

UniProt Consortium, T. (2018) UniProt: the universal protein knowledgebase. *Nucleic acids research*, **46**, 2699-2699.

Varet, H., Brillet-Guéguen, L., Coppée, J.-Y. and Dillies, M.-A. (2016) SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data. *PLOS ONE*, **11**.

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 Biology*, **6**, 9.

Weinstock, K.G., Kirkness, E.F., Lee, N.H., Earle-Hughes, J.A. and Venter, J.C. (1994) cDNA sequencing: a means of understanding cellular physiology. *Current Opinion in Biotechnology*, **5**, 599-603.

Yeh, R.F., Lim, L.P. and Burge, C.B. (2001) Computational inference of homologous gene structures in the human genome. *Genome Res*, **11**, 803-816.

9 Supporting information

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

Supporting material can be found at: https://onlinelibrary.wiley.com/doi/full/10.1111/tpj.13801 #support-information-section

9.1.1 Organellar SNPs

Table S1: SNPs between the old (published) assemblies and the new assemblies described in (Lang et al., 2018). *https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Ftpj.13801&file=tpj13801-sup-0001-Appendix S1.docx, (original paper table S7)*

Chloroplast assembly			
SNP	**new**	**old**	**pos**
1	A	G	9799
2	-	T	9800
3	-	G	9801
4	C	T	9803
5	A	G	11088
6	A	C	12136
7	A	-	12320
8	T	C	15522
9	-	T	16634
10	A	C	16867
11	T	-	18666
12	C	T	19324
13	G	A	25850
14	T	G	25857
15	T	G	25858
16	T	C	25259
17	A	-	26170
18	T	C	27718
19	-	A	28553
20	T	C	29844
21	-	A	32423
22	G	A	36125
23	A	G	38034
24	A	G	38068
25	T	C	38138
26	-	A	41978
27	A	T	44962
28	A	T	44964
29	A	T	44969
30	A	T	44974
31	A	-	44985
32	T	-	45047
33	A	G	45242
34	A	G	46450
35	T	-	52832
36	T	-	52833
37	T	-	54604
38	A	-	56476
39	A	T	56509
40	T	-	56511
41	T	-	56512
42	A	-	56513
43	A	-	56514
44	G	A	56520
45	A	G	56521
46	A	G	56523
47	T	A	56826
48	T	-	56923
49	T	-	56999
50	A	-	57167
51	A	G	57805
52	T	C	57884
53	T	-	60825
54	A	G	60939
55	A	G	61338
56	C	G	64095
57	C	G	64126
58	A	T	65087
59	-	A	65089
60	A	T	65092
61	T	A	66289
62	A	G	67307
63	G	A	68438
64	A	T	69793
65	A	T	70256
66	A	G	70666
67	T	-	93793
68	A	C	96123
69	A	C	97910
70	A	-	98127
71	A	-	98128
72	A	-	98129
73	T	C	98880
74	W	A	103791
75	T	C	107224
76	A	-	114353

Mitochondrial assembly			
SNP	**new**	**old**	**pos**
1	T	N	48612
2	T	Y	50872

9.1.2 Contamination screening

The sequencing databases are still growing and more information is added each day. Between 2008 and 2018, billions of new base pairs were uploaded to the NCBI GeneBank database (https://www.ncbi.nlm.nih.gov/genbank/statistics/). For the decontamination analysis, that means that contaminated sequences maybe not found in 2008, could be find now.

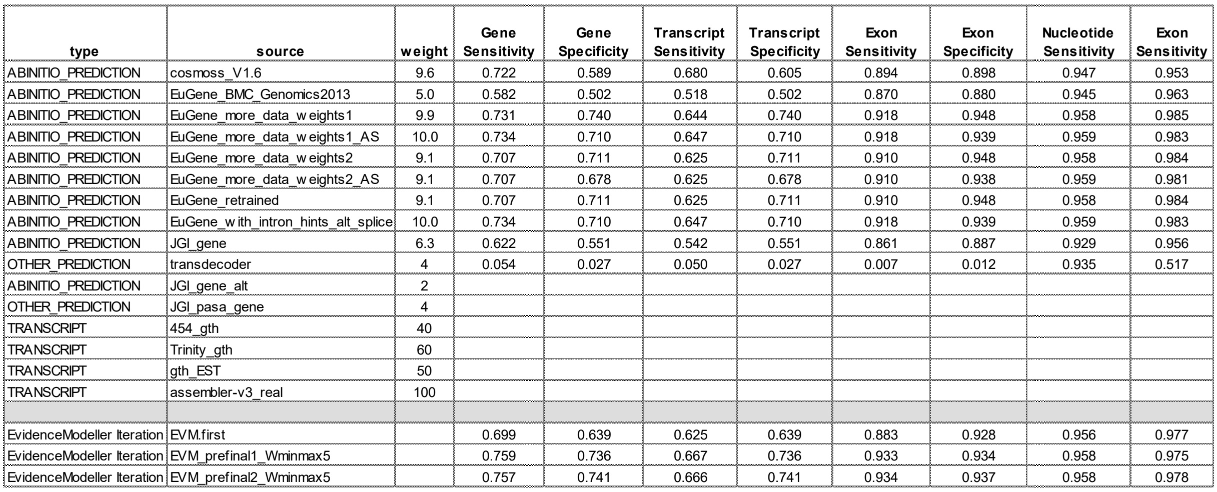
For the V3 decontamination process, we used the method detailed described in (Nishiyama *et al.*, 2018) and https://ars.els-cdn.com/content/image/1-s2.0-S0092867418308018-mmc5.xlsx .

All as bacterial contaminated identified scaffolds are listed below:

Additional screening based on homology, methylation and ChIP-seq evidence identified 21 of the 351 unplaced scaffolds (namely, scaffold_30, 32, 34, 42, 46, 58, 70, 92, 96, 155, 169, 196, 356, 405, 476, 602, 740, 853, 914, 915 and 1166, encoding in total 1.37 Mbp, 424 gene models in v3.1 and 479 gene models in v3.3) as potential contaminant. Those will be removed in the next genome release and have been removed from supplementary file 1: https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Ftpj.13801&file=tpj13801-sup-0001-AppendixS1.docx, page 5

9.1.3 Gene prediction and structural gene annotation

https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Ftpj.13801&file=tpj13801-sup-0001-AppendixS1.docx, original paper page 14-17



type	source	weight	Gene Sensitivity	Gene Specificity	Transcript Sensitivity	Transcript Specificity	Exon Sensitivity	Exon Specificity	Nucleotide Sensitivity	Exon Sensitivity
ABINITIO_PREDICTION	cosmoss_V1.6	9.6	0.722	0.589	0.680	0.605	0.894	0.898	0.947	0.953
ABINITIO_PREDICTION	EuGene_BMC_Genomics2013	5.0	0.582	0.502	0.518	0.502	0.870	0.880	0.945	0.963
ABINITIO_PREDICTION	EuGene_more_data_weights1	9.9	0.731	0.740	0.644	0.740	0.918	0.948	0.958	0.985
ABINITIO_PREDICTION	EuGene_more_data_weights1_AS	10.0	0.734	0.710	0.647	0.710	0.918	0.939	0.959	0.983
ABINITIO_PREDICTION	EuGene_more_data_weights2	9.1	0.707	0.711	0.625	0.711	0.910	0.948	0.958	0.984
ABINITIO_PREDICTION	EuGene_more_data_weights2_AS	9.1	0.707	0.678	0.625	0.678	0.910	0.938	0.959	0.981
ABINITIO_PREDICTION	EuGene_retrained	9.1	0.707	0.711	0.625	0.711	0.910	0.948	0.958	0.984
ABINITIO_PREDICTION	EuGene_with_intron_hints_alt_splice	10.0	0.734	0.710	0.647	0.710	0.918	0.939	0.959	0.983
ABINITIO_PREDICTION	JGI_gene	6.3	0.622	0.551	0.542	0.551	0.861	0.887	0.929	0.956
OTHER_PREDICTION	transdecoder	4	0.054	0.027	0.050	0.027	0.007	0.012	0.935	0.517
ABINITIO_PREDICTION	JGI_gene_alt	2								
OTHER_PREDICTION	JGI_pasa_gene	4								
TRANSCRIPT	454_gth	40								
TRANSCRIPT	Trinity_gth	60								
TRANSCRIPT	gth_EST	50								
TRANSCRIPT	assembler-v3_real	100								
EvidenceModeller Iteration	EVM.first		0.699	0.639	0.625	0.639	0.883	0.928	0.956	0.977
EvidenceModeller Iteration	EVM_prefinal1_Wminmax5		0.759	0.736	0.667	0.736	0.933	0.934	0.958	0.975
EvidenceModeller Iteration	EVM_prefinal2_Wminmax5		0.757	0.741	0.666	0.741	0.934	0.937	0.958	0.978

Table S2: Rank-based weighting for evidence modelling. *(original paper table S9)*

library type	nucleic acid	read length	used for	no. of samples	tissue	SRA
5' cap capture	RNA	76 bp	TSS CoGe	1	juvenile gametophores	SRP092233
5' cap capture	RNA	101 bp	mt/cp annot.	2	juvenile gametophores	SRP092233
BS-seq	DNA	90 bp	methylation	1	adult gametophores	SRP092161

Table S3: Novel deep sequencing datasets used in this study.

Column "used for" lists the purpose these datasets were employed for in this study. All datasets were used as transcript evidence for gene prediction. (original paper table S10)

Generation of the v3.2 genome annotation

Short reads (Table S3) were assembled using our genome-guided in-house pipeline (PERTRAN, unpublished). These transcript assemblies and ~740K ESTs were assembled by PASA (Haas *et al.*, 2003). Loci were determined by PASA transcript assembly alignments and/or EXONERATE alignments of proteins from *Arabidopsis thaliana*, grape, soybean, sorghum, rice, *Chlamydomonas reinhardtii,* and Swiss-Prot to a soft-repeat masked genome using RepeatMasker (Smit *et al.*, 1996) with up to 2K bp extension on both ends unless extending into another locus on the same strand. Gene models were predicted by homology-based predictors, FGENESH+ (Salamov and Solovyev, 2000), FGENESH_EST (similar to FGENESH+, EST as splice site and intron input instead of protein/translated ORF), and GenomeScan (Yeh *et al.*, 2001). The highest scoring predictions for each locus are selected using multiple positive factors including EST and protein support, and one negative factor: overlap with repeats. The selected gene predictions were improved by PASA. Improvement includes adding UTRs, splicing correction, and adding alternative transcripts. PASA-improved gene model proteins were subject to protein homology analysis to above mentioned proteomes to obtain Cscore and protein coverage. Cscore is a protein BLASTP score ratio to MBH (mutual best hit) BLASTP score and protein coverage is the highest percentage of protein aligned to the best of homologs. PASA-improved transcripts were selected based on Cscore, protein coverage, EST coverage, and its CDS overlapping with repeats. The transcripts were selected if its Cscore is larger than or equal to 0.5 and protein coverage larger than or equal to 0.5, or it has EST coverage, but its CDS overlapping with repeats is less than 20%. For gene models whose CDS overlaps with repeats for more than 20%, its Cscore must be at least 0.9 and homology coverage at least 70% to be selected. The selected gene models were subject to Pfam analysis and gene models whose protein is more than 30% in Pfam TE domains were removed.

Generation of the v3.3 genome annotation

V3.2 models were pooled with the previous catalog of gene predictions. Gene model support was assessed and used for definition of additional loci and selection of the representative isoform following the protocol established for the v3.1 annotation. Non-protein-coding genes (ncRNA) were predicted using a combined approach using multiple tools for each subclass. We generated ncRNA features and assigned specific Sequence Ontology terms (Eilbeck *et al.*, 2005) and class attributes in GFF3. Infernal (Nawrocki and Eddy, 2013) was used to annotate general classes of ncRNAs using the RFAM (Nawrocki *et al.*, 2015) covariance models. Infernal/RFAM, RNAmmer (Lagesen *et al.*, 2007) and custom BLASTN searches with the SILVA database (Quast *et al.*, 2013) and *P. patens* rRNA sequences to annotate rRNAs as maximal ranges. We used a consensus approach employing tRNAScan-SE (Lowe and Eddy, 1997), ARAGORN (Laslett and Canback, 2004) and infernal/RFAM to predict tRNAs including exon and anticodon features. Pre-miRNAs and miRNAs were annotated based on a combined approach combining the results from Infernal/RFAM (Coruh *et al.*, 2015) and a mapping of the previous V1.6/miRBase annotation. Existing database identifiers and names e.g. miRBase or RFAM IDs were annotated using GFF3 Alias attributes. PASA assemblies without protein-homology or proteomics support were annotated as putative (long) ncRNAs. In addition, a comprehensive set of public sRNA-Seq libraries (Coruh *et al.*, 2015) was used to locate clusters of sRNA and putative hairpin and miRNA precursors using short stack (Axtell, 2013). Based on their overlap with the other annotation classes, these were classified into the subclasses: mRNA_sRNA_cluster, NCLDV_sRNA_cluster, ncRNA_sRNA_cluster, TE_sRNA_cluster and unclassified_sRNA_cluster. The latter does not overlap with any of the other annotations and most likely represent yet unknown loci targeted by heterchromatic siRNAs. Names include information about the type of sRNA cluster as annotated by shortstack (cluster, hairpin or miRNA). For the sRNAs overlapping an existing feature the name of the overlapping feature and the strandedness is recorded in order to ascertain anti-/sense orientation e.g. in case of protein-coding or NCLDV gene regions. The annotated features are summarized in Table S4. Annotation files are provided in bed and GFF3 format (http://plantco.de/supplement/annotation_bed_gff.7z).

Table S4: Annotation *(original paper table S11)*

	mRNA	**ncRNA**
predicted_by_ab_initio_computation	1	25
supported_by_sequence_similarity	5137	1611
supported_by_EST_or_cDNA	29108	1936
supported_by_peptide_spectrum_match	405	0
total	**34651**	**3572**

A: Experimental support of the loci

	alternative	**major**	**total**
mRNA	54904	34651	**89555**
ncRNA	1010	3572	**4582**
total	**55914**	**38223**	**94137**

B: Transcript statistics for protein-coding and ncRNA genes

	mRNA	**mRNA sRNA**	**NCLDV genes**	**NCLDV regions**	**NCLDV sRNA**	**ncRNA**	**ncRNA sRNA cluster**	**TE**	**TE sRNA cluster**	**intact LTR-RT**	**unclassified sRNA cluster**
Chr01	2035	6283	17	5	18	147	57	15255	16619	508	2111
Chr02	1856	5726	23	10	26	145	56	13103	8929	425	1678
Chr03	1805	5185	13	6	9	168	81	12534	7173	392	1720
Chr04	1499	4147	26	5	11	136	58	11500	5098	364	1368
Chr05	1304	3636	12	6	7	135	62	10807	4320	329	1180
Chr06	1411	3568	6	3	1	140	62	9877	3450	297	1267
Chr07	1330	3737	3	3	1	108	42	9109	3114	272	1024
Chr08	1157	3103	6	3	3	78	19	9309	3119	281	907
Chr09	1180	3305	26	4	17	99	24	8923	2578	286	1026
Chr10	1214	3054	14	4	12	100	37	8857	2531	293	940
Chr11	1310	3467	10	4	4	71	18	8470	2432	288	964
Chr12	1172	2929	0	1	0	99	46	8803	2457	308	863
Chr13	1059	2636	10	2	4	93	31	9114	2672	321	738
Chr14	1332	3595	9	2	5	105	45	8623	1936	250	913
Chr15	1170	3127	21	6	13	94	38	8819	2059	256	897
Chr16	1245	3257	23	5	12	95	36	8397	1717	266	835
Chr17	1156	2829	3	2	1	88	35	8029	1773	270	803
Chr18	934	2328	4	1	5	76	30	8422	1989	258	721
Chr19	1023	2324	5	3	2	115	36	8298	1639	279	734
Chr20	1115	2646	3	1	3	97	40	7832	1576	235	830
Chr21	1013	2387	3	2	1	98	49	7959	1624	257	628
Chr22	987	2496	4	2	2	125	66	7999	1465	273	626
Chr23	1006	2419	1	1	0	86	43	7644	1432	258	626
Chr24	959	2314	1	1	0	79	40	6819	1259	218	537
Chr25	651	1650	9	2	4	56	23	5829	998	199	437
Chr26	717	1698	1	1	0	73	32	5156	998	149	443
Chr27	419	1101	1	1	0	39	24	2610	406	72	259
scaffold	2592	3485	3	1	2	827	867	2120	457	92	1256
total	**34651**	**88432**	**257**	**87**	**163**	**3572**	**1997**	**24021**	**85820**	**7696**	**26331**

C: Chromosomal distribution of V3.3 annotations

All v3.1 and v3.3 gene models can be observed in the table reachable file the following link. Gene models removed during the contamination analysis are shown in separated sheets. https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Ftpj.13801&file=tpj13801-sup-0002-FileS1.xlsx

9.1.4 CoGe uploaded tracks

Table S5: Tracks (names and descriptions) available in CoGe. *https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Ftpj.13801&file=tpj13801-sup-0001-AppendixS1.docx, (original paper table S12)*

Track_name	Description
CHG_methylated_positions_adult_gametophores	CHG methylated positions adult gametophores - minimal coverage of 9 and 90% methylated reads (whole genome BSseq)
CHH_methylated_positions_adult_gametophores	CHH methylated positions adult gametophores- minimal coverage of 9 and 90% methylated reads (whole genome BSseq)
CG_methylated_positions_adult_gametophores	CG methylated positions adult gametophores - minimal coverage of 9 and 90% methylated reads (whole genome BSseq)
v1.2_unique_gene_models	v1.2 gene models mapped on V3 genome - gene models with unique hits
v1.2_multiple_hits_gene_models	v1.2 gene models mapped on V3 genome - gene models with multiple hits
v1.6_unique_gene_models	v1.6 gene models mapped on V3. genome - gene models with unique hits
v1.6_multiple_hits_gene_models	v1.6 gene models mapped on V3 genome - gene models with multiple hits
v3.1_gene_models	v3.1 gene models (major isoform)
v3.3_gene_models_all_isoforms	v3.3 gene models of all isoforms
v3.3_gene_models	v3.3 gene models (major isoform)
transposable_elements	identified transposable elements
SNPs_Reute_vs_Gransden	genomewide SNP analysis comparing ecotypes Gransden and Reute
SNPs_Villersexel_vs_Gransden	detected SNPs comparing ecotypes Gransden and Villersexel
SNPs_Kaskaskia_vs_Gransden	detected SNPs comparing ecotypes Gransden and Kaskaskia
array_probes_Nimblegen_v1.6_unique	microarray probes NimbleGen with unique hit
array_probes_Nimblegen_v1.6_multiple	microarray probes NimbleGen with multiple hits
array_probes_CBMX_v1.2_unique	microarray probes CombiMatrix with unique hit
array_probes_CBMX_v1.2_multiple	microarray probes CombiMatrix with multiple hits
TSS_juvenile_gametophores_log2	transcription start site based on 5-prime cap capture RNA seq data
RNA-seq_evidence_chr01-27_log2	chromosome 01-27 RNA-seq evidence combined of all JGI Gene Atlas available data
RNA-seq_evidence_scaffolds_log2	scaffolds RNA-seq evidence combined of all JGI Gene Atlas available data

9.2 The *Physcomitrella patens* gene atlas project: large-scale RNA-seq based expression data

Supporting material can be found at: https://onlinelibrary.wiley.com/doi/full/10.1111/tpj.13940 #support-information-section

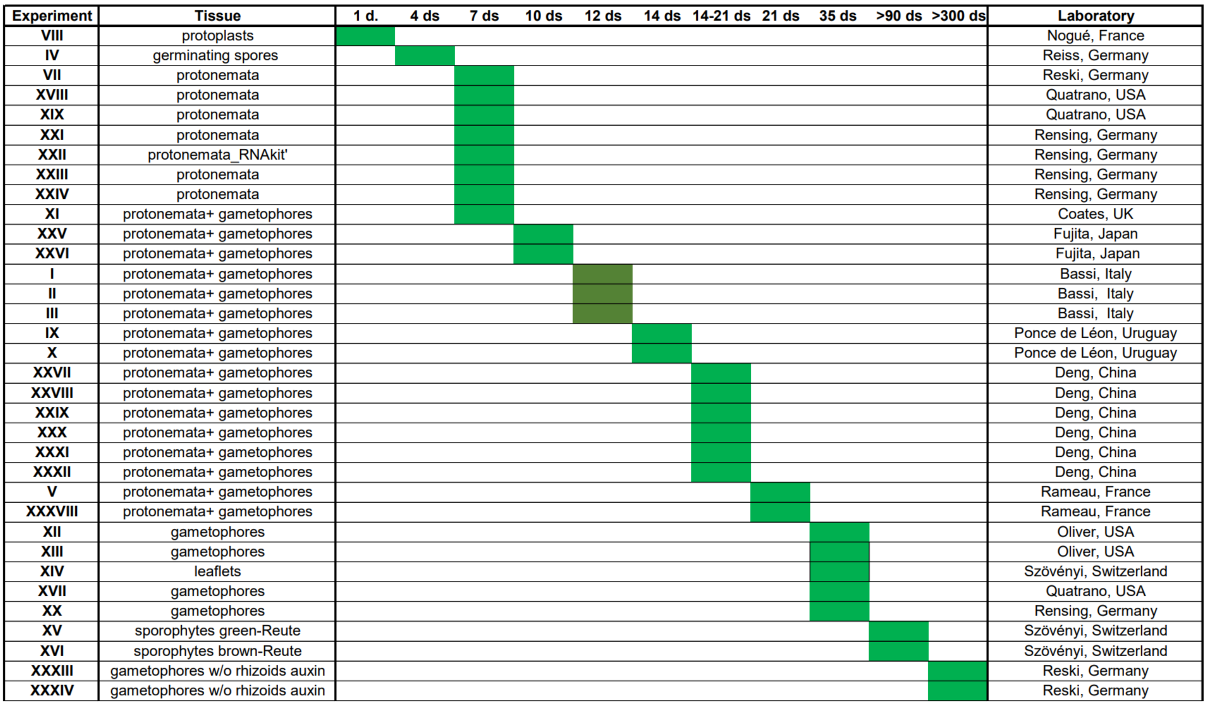
9.2.1 JGI Gene Atlas RNA-seq samples

Table S6: Overview of the experiments and their primary library data presented in this study. *https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Ftpj.13940&file=tpj13940-sup-0002-TableS1-S4.pdf, (original paper table S1)*

Experiment	Tissue	Condition	Raw reads	Detected transcripts	SRA numbers
VIII	Protoplasts	BCDA solid	145,602,566	24,283	THWA, THXG, THUX
IV	Germinating spores	BCDA solid, Continuous light	119,926,148	23,517	THWT, THUW, THWX
VII	protonemata	Knop liquid	131,460,666	23,236	SYYC, SYYW, SYYZ
XVIII	protonemata	BCD solid	101,169,800	23,612	SYYU, SYZX
XIX	protonemata	BCD solid, 50µmol, ABA 24hrs	158,390,608	23,584	THWY, THUT, THWO
XXI	protonemata	Knop liquid	107,406,052	24,071	SYZB, SYZO
XXII	protonemata	Knop liquid, Rneasy	170,524,124	24,547	SYZT, SYYA, SYZN
XXIII	protonemata	Knop liquid ammonium, harvested 2hrs after light comes on	114,544,326	23,387	UBZH, UBYA, UBZY
XXIV	protonemata	Knop liquid, harvested 2hrs after light comes on	107,597,154	23,730	UBXX, UBYC, UBYT
XI	protonemata + young gametophores	BCD solid, Continuous light, GA9 methyl-ester	101,802,990	22,634	UBYN, UBXT, UBYB
XXV	protonemata + young gametophores	BCDA solid, Continuous light	146,616,244	23,596	SYZS, SYZU, SYZP
XXVI	protonemata + young gametophores	BCD solid, heat stress	155,122,446	22,610	THXC, THWP, THWC
I	protonemata + young gametophores	Knop solid	196,421,176	23,809	SYYH, SYYN, SYZH
II	protonemata + young gametophores	Knop solid, high light 850µmol	128,601,062	23,090	THXH, THUO, THWU
III	protonemata + young gametophores	Knop solid, low light 10µmol	126,885,998	23,141	UBYY, UBZW, THUU
IX	protonemata + young gametophores	BCDA solid	119,552,632	23,234	SYYB, SYYO, SYYY
X	protonemata + young gametophores	BCDA solid, 50µmol, OPDA	131,110,724	23,529	UBYZ, UBXN, UBZX
XXVII	protonemata + young gametophores	BCDA solid, Continuous light, Glucose	123,619,596	23,873	SYYG, SYYX, SYZA
XXVIII	protonemata + young gametophores	BCDA solid, Glucose, UV-B light	115,692,872	24,086	UBXS, UBYU, UCAA
XXIX	protonemata + young gametophores	BCDA solid, darkness, Glucose	92,435,720	22,902	UBXP, UBZS, UBYS
XXX	protonemata + young gametophores	BCDA solid, Glucose, red light	118,699,196	24,283	UBXW, UBYO, UBZZ
XXXI	protonemata + young gametophores	BCDA solid, Glucose, blue light	108,647,368	24,116	UBYH, UBZC, UBZB
XXXII	protonemata + young gametophores	BCDA solid, Glucose, far red light	119,599,564	23,626	UBZG, UBYW, UBXO
V	protonemata + young gametophores	BCD solid, strigolactone	131,697,522	24,617	UBXU, UBZO, UBYP
XXXVIII	protonemata + young gametophores	BCD solid, Acetone	125,367,772	24,600	UBZN, UBZU, UBXY
XII	Gametophores	BCD solid, drought (dehydration)	116,985,724	23,258	THUY, THXA, THWH
XIII	Gametophores	BCD solid, rehydration	104,483,200	23,263	THUN, THWN, THWS
XIV	Leaflets	Knop solid	76,201,102	23,842	THWW, THUZ
XVII	Gametophores, above	BCD solid	142,534,874	24,684	SYZZ, SYYS, SYZC
XX	Gametophores, above	Knop solid	134,365,946	24,900	SYZW, SYYT, SYYP
XV	Sporophytes, green	Knop solid	156,806,294	26,012	THUP, THWB, THWG
XVI	Sporophytes, brown	Knop solid	129,548,594	24,606	THWZ, THXB, THUS
XXXIII	Gametophores rhizoids removed	Knop liquid, hydroponic	118,997,020	24,792	UBXZ, UBYX, UBYG
XXXIV	Gametophores rhizoids removed	Knop liquid, hydroponic, Auxin	101,483,746	23,761	UBZT, UBZP, UBZA

9.2.2 JGI Gene Atlas RNA-seq tissue

***Table S7: Harvesting time point after initiation of the specific culture and experimental location for each experiment in this study**. https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Ftpj.13940&file=tpj13940-sup-0002-TableS1-S4.pdf, (original paper table S2)*



Experiment	Tissue	1 d.	4 ds	7 ds	10 ds	12 ds	14 ds	14-21 ds	21 ds	35 ds	>90 ds	>300 ds	Laboratory
VIII	protoplasts	🟩											Nogué, France
IV	germinating spores		🟩										Reiss, Germany
VII	protonemata			🟩									Reski, Germany
XVIII	protonemata			🟩									Quatrano, USA
XIX	protonemata			🟩									Quatrano, USA
XXI	protonemata			🟩									Rensing, Germany
XXII	protonemata_RNAkit'			🟩									Rensing, Germany
XXIII	protonemata			🟩									Rensing, Germany
XXIV	protonemata			🟩									Rensing, Germany
XI	protonemata+ gametophores			🟩									Coates, UK
XXV	protonemata+ gametophores				🟩								Fujita, Japan
XXVI	protonemata+ gametophores				🟩								Fujita, Japan
I	protonemata+ gametophores					🟩							Bassi, Italy
II	protonemata+ gametophores					🟩							Bassi, Italy
III	protonemata+ gametophores					🟩							Bassi, Italy
IX	protonemata+ gametophores						🟩						Ponce de Léon, Uruguay
X	protonemata+ gametophores						🟩						Ponce de Léon, Uruguay
XXVII	protonemata+ gametophores							🟩					Deng, China
XXVIII	protonemata+ gametophores							🟩					Deng, China
XXIX	protonemata+ gametophores							🟩					Deng, China
XXX	protonemata+ gametophores							🟩					Deng, China
XXXI	protonemata+ gametophores							🟩					Deng, China
XXXII	protonemata+ gametophores							🟩					Deng, China
V	protonemata+ gametophores								🟩				Rameau, France
XXXVIII	protonemata+ gametophores								🟩				Rameau, France
XII	gametophores									🟩			Oliver, USA
XIII	gametophores									🟩			Oliver, USA
XIV	leaflets									🟩			Szövényi, Switzerland
XVII	gametophores									🟩			Quatrano, USA
XX	gametophores									🟩			Rensing, Germany
XV	sporophytes green-Reute										🟩		Szövényi, Switzerland
XVI	sporophytes brown-Reute										🟩		Szövényi, Switzerland
XXXIII	gametophores w/o rhizoids auxin											🟩	Reski, Germany
XXXIV	gametophores w/o rhizoids auxin											🟩	Reski, Germany

9.2.3 Found DEGs

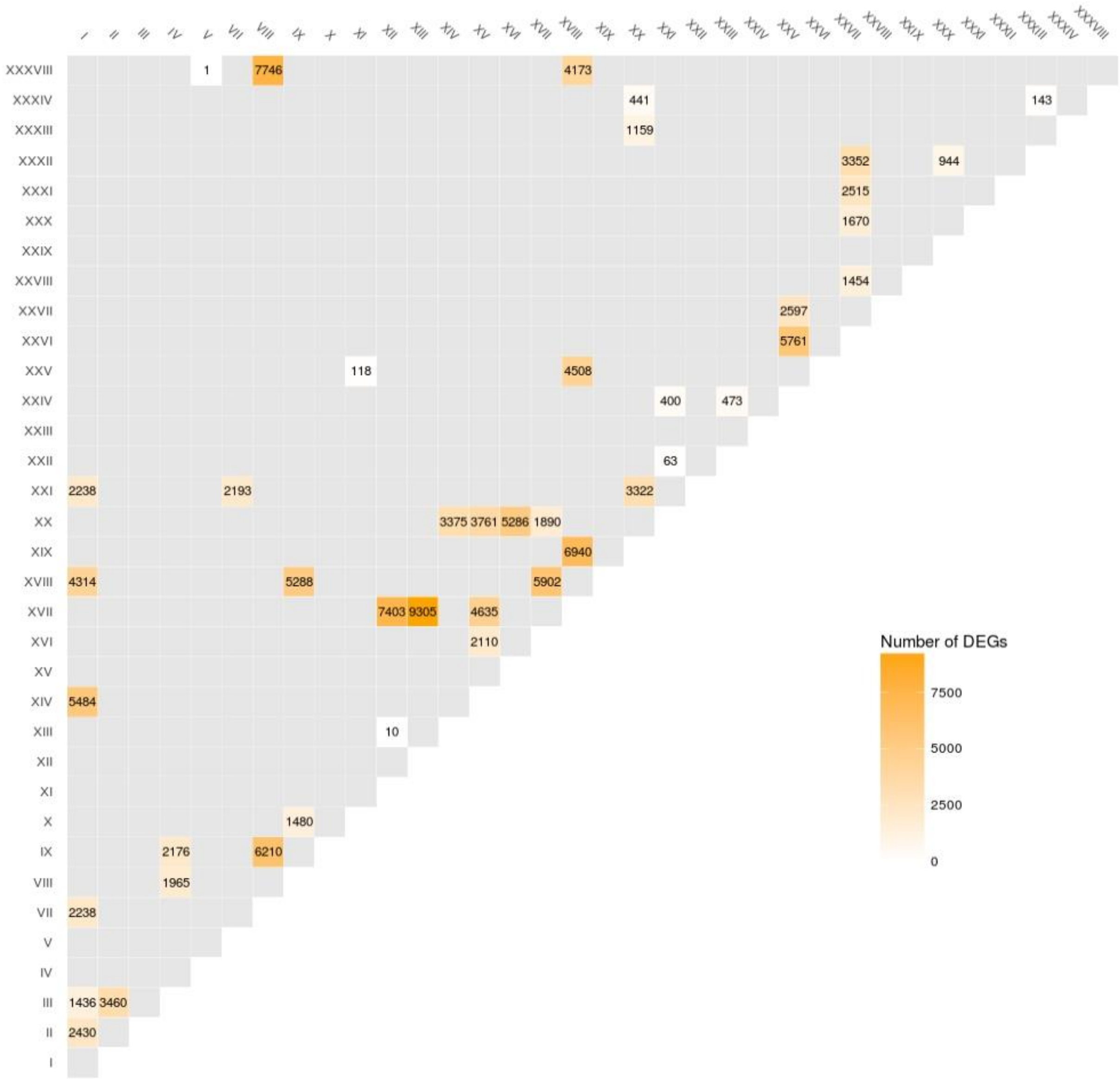


Figure S1: Overview of the experiment pairs for which DEGs have been calculated in the present study. *https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Ftpj.13940&file=tpj13940-sup-0002-TableS1-S4.pdf, (original paper table S3)*

9.2.4 JGI Gene Atlas RNA-seq sample clustering

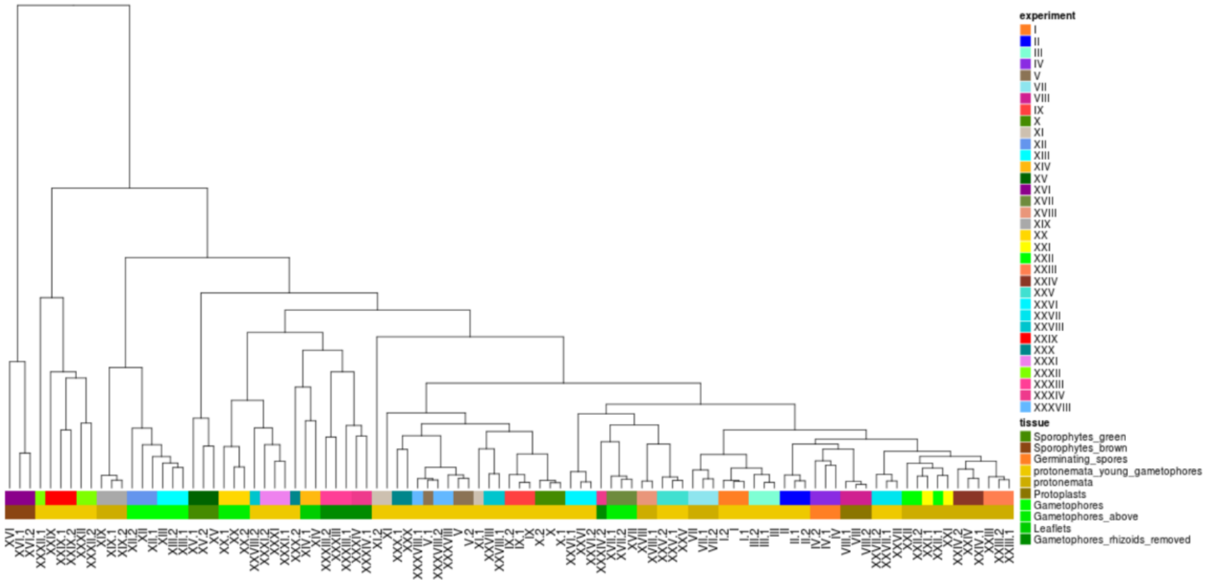
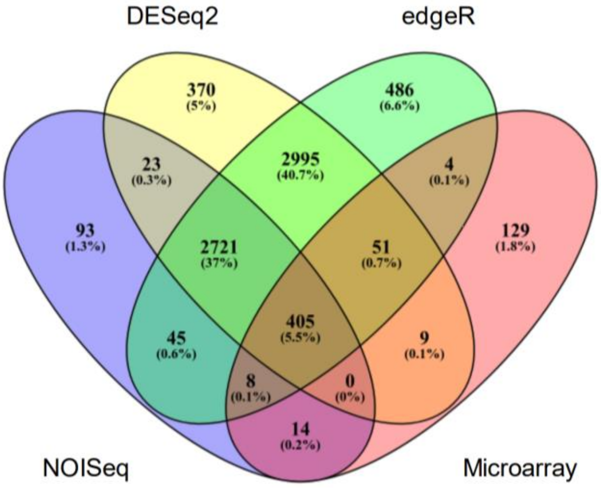


Figure S2: Hierarchical clustering of all 99 RNA-seq samples, RPKM normalized. *The upper coloured line represents the experiments, the lower one the corresponding tissues. This analysis confirms and illustrates that the replicates for each experiment cluster as expected with each other in most of the cases, for example experiment XVI (brown sporophyte). The exceptions are the cases in which consequently few DEGs were detected. For example, experiments V and XXXVIII libraries group together, indication of the closeness of these samples, but the triplicates are not resolved between the two experiments. The clustering provides independent confirmation of the absence of effect of the strigolactone treatments in this specific experiment.* *https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Ftpj.13940&file=tpj13940-sup-0001-FigS1-S7.pdf*, *(original paper figure S1)*

9.2.5 DEG intersection



FigureS3: Comparison of the DEGs called by the NOISeq, DESeq2 and edgeR packages with RNA dataset and by microarray approach. *Venn diagram comparing the DEGs called by NOISeq (in blue), DESeq2 (in yellow) and edgeR (in green) between the Experiments XXI (gametophore) and XX (protonemal liquid culture) and the DEGs called in a microarray experiment performed on the same tissues (Hiss et al. 2014).* *https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Ftpj.13940&file=tpj13940-sup-0001-FigS1-S7.pdf*, *(original paper figure S7)*

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

Supporting material can be found at: https://onlinelibrary.wiley.com/doi/full/10.1111/tpj.14607 #support-information-section

9.3.1 JGI Gene Atlas RNA-seq samples second sequencing round

Table S8: Overview of the second RNA-seq sequencing round experiments and their primary library data. *PPR_79 - Pentatricopeptide Repeat KO Pp3c5_7610, PPR_78 - Pentatricopeptide Repeat KO Pp3c2_12230, LFY – leafy over expressed (target Pp3c13_8760)*

Experiment	Tissue	Condition	Mutant	Raw reads	SRA number
CI	Gametophores above adult	Knop solid		193,225,324	BBTWT, BBTWU, BBTWS, BXHHU, BXHHT, BXHHW
CII	Gametophores above juvenil	Knop solid		254,912,414	BBTWW, BBTWY, BBTWX, BXHHX, BXHHZ, BXHHY
CIII	Gametophores above	Knop solid	Mutant referenz	233,457,574	BBTXW, BBTXX, BBTXZ, BXHNY, BXHNW, BXHNX
CIV	Gametophores above	Knop solid	LFY_34_3	199,931,380	BBTUO, BBTUP, BBTUS, BXHGT, BXHGW, BXHGU
CV	Gametophores above	Knop solid	LFY_34_7	199,308,048	BBTUT, BBTUX, BBTUW, BXHGY, BXHGZ, BXHGX
CVI	Gametophores above	Knop solid		192,072,836	BBTWA, BBTUY, BBTUZ, BXHHA, BXHHC, BXHHB
CVII	Gametophores above	Knop solid		155,479,402	BBTWB, BBTWG, BBTWC, BXHHH, BXHHG, BXHHN
CVIII	Gametophores above	Knop solid		172,694,736	BBTWH, BBTWN, BBTWP, BXHHS, BXHHO, BXHHP
CIX	Gametophores above	Knop solid		240,958,746	BBTYA, BBTYB, BBTYC, BXHOA, BXHOB, BXHNZ
CX	Protonema + young gametophores	Knop liquid _KO	KO PPR_78	197,399,584	BBTXY, BBTYH, BBTYN, BXHCX, BXHCZ, BXHCY
CXI	Protonema + young gametophores	Knop liquid _KO	KO PPR_78	210,461,278	BBTYO, BBTYP, BBTYS, BXHGB, BXHGC, BXHGA
CXII	Protonema + young gametophores	Knop liquid _KO	KO PPR_79	212,511,024	BBTUA, BBTUB, BBTUC, BXHGG, BXHGN, BXHGH
CXIII	Protonema + young gametophores	Knop liquid _KO	KO PPR_79	197,846,842	BBTUG, BBTUN, BBTUH, BXHGP, BXHGO, BXHGS
CXIV	Protonema + young gametophores	Knop liquid _KO_Ref	KO PPR referenz	194,080,302	BBTXO, BBTXN, BBTXP, BXHNP, BXHNN, BXHNO
CXV	Protonema + young gametophores	Knop liquid _KO_Ref	KO PPR referenz	228,455,454	BBTXS, BBTXU, BBTXT, BXHNT, BXHNS, BXHNU
CXVI	Protonema + young gametophores	Knop liquid		213,185,420	BBTXA, BBTXB, BBTWZ, BXHNB, BXHNC, BXHNA
CXVII	Protonema + young gametophores	Knop liquid, PO4 control 10 d		116,763,544	BAOBT, BAOBN, BBTXG, BXHNG
CXVIII	Protonema + young gametophores	Knop liquid, PO4 control 1 d		69,170,542	BAOAW, BAOAX, BAOBA
CXIX	Protonema + young gametophores	Knop liquid, PO4 control 5 d		61,856,256	BAOBH, BAOBP, BAOBS
CXX	Protonema + young gametophores	Knop liquid, PO4 deficiency 10 d		73,851,184	BAOAZ, BAOBC, BAOAU
CXXI	Protonema + young gametophores	Knop liquid, PO4 deficiency 1 d		84,915,194	BAOBG, BAOBU, BAOBO
CXXII	Protonema + young gametophores	Knop liquid, PO4 deficiency 5 d		69,133,732	BAOBB, BAOAY, BAOBW
CXXIII	Spores	BCD liquid, continous light		107,570,208	BAOAT, BBTXC, BXHCW
CXXIV	Gametophores above	Knop solid	DYT cut C	72,606,442	BBTYG, BXHOC
CXXV	Spores	BCDA liquid, continous light		40,523,940	BBTWO, BXHCU
XXI	Protonema + young gametophores	Knop liquid		72,610,958	BBTXH, BXHNH
XVIII	Protonema + young gametophores	BCD solid		42,676,552	BBTUU, BXHCT

9.3.2 Gene set normalization

Due to constantly implementing improvements, PEATmoss will be able to compare expression values across platforms. PEATmoss contains three different platforms: Two microarray datasets, CombiMatrix and NimbleGen and RNA-seq datasets. The direct comparability of the raw expression values from microarray and RNA-seq data is not possible. This is an effect of divergent normalization.

A method to overcome this incomparability is to normalize the individual datasets by equal expressed genes in both sets (also known as housekeeping-genes). To normalize microarray gene expression, (Hiss *et al.*, 2014) used as reference the thioredoxin encoding gene Pp3c19_1800V3.1 (Phypa_17394, *Pp* genome annotation v1.2). Unfortunately, this gene is highly variable in RNA-seq expression data and cannot be used to normalize RNA-seq gene expression.

The aim, to get new reference genes, was to find genes with the lowest variable expression profile in all RNA-seq samples and microarray experiments. Therefore, the statistical value "coefficient of variation" (CoV) was calculated for each gene and platform/dataset. Afterwards, the genes were ranked by their CoV values for all platforms (Figure S4). To get the genes with the lowest CoV in the two microarray and the RNA-seq_JGI dataset, the first 1,000 genes (top 1,000) with the lowest CoV were extrected (Figure S5).

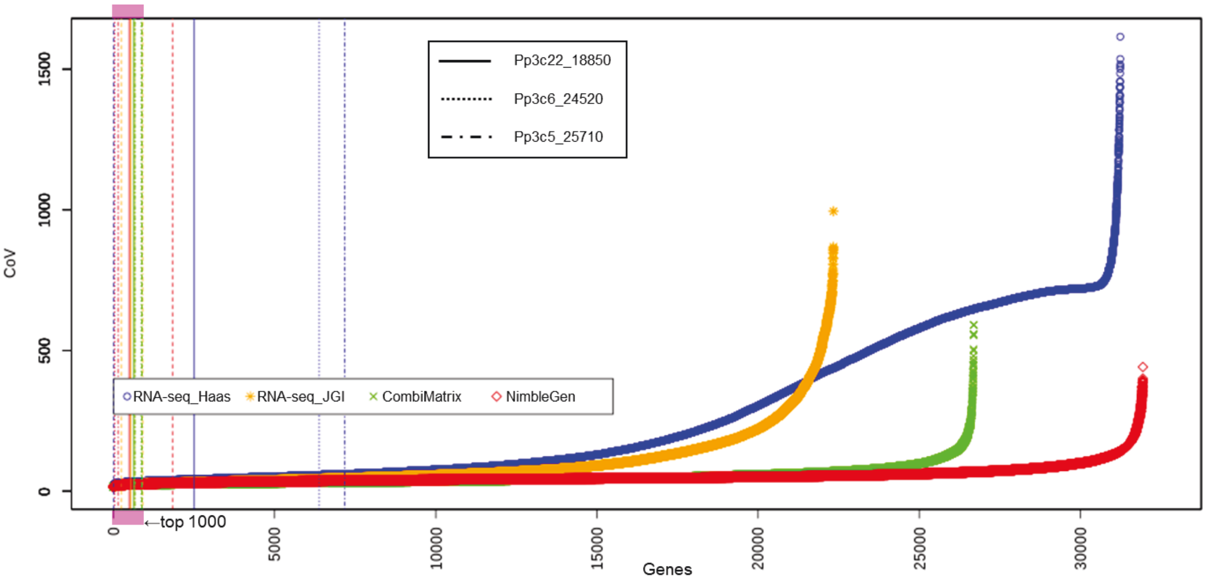


Figure S4: This plot shows the ranked CoV values of four different dataset*. Two microarray datasets, CombiMatrix (v1.2) [green] and NimbleGen (v1.6) [orange] and two RNA-seq datasets, the first 99 samples of the JGI Gene Atlas project [red] and all RNA-seq samples used in (Haas et al., 2020) [blue]. The position of the three reference genes in each dataset is marked in straight, dashed and dotted vertical lines.*

This three lists of top 1,000 genes were intersected (Figure S5). Three genes are at the intersection representing the new reference gene set for gene set normalization between microarray and RNA-seq datasets: Pp3c22_18850V3.1, a protein kinase; Pp3c6_24520V3.1, a nuclear pore component and Pp3c5_25710V3.1, a Las1-like protein.

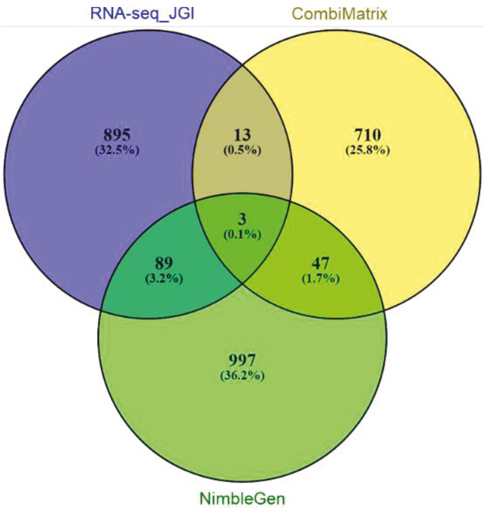


Figure S5: Intersection of the top 1,000 lowest CoV genes. *Three genes are in the intersection and became the new reference gene set for the gene set normalization. The number of genes differ and does not always show 1,000 genes. This is due to gene conversion from v1.2 and v1.6 to v3.3.*

9.4 Single nucleotide polymorphism charting of *P. patens* reveals accumulation of somatic mutations during *in vitro* culture on the scale of natural variation by selfing

Table S9: Source and NCBI accession nos. for all RNA-seq samples used. *(original paper table S1)*

Accession	Project/tissue	Source	BioProject/SRA number
Gransden	JGI Gene Atlas	(Fernandez-Pozo et al. 2019, Perroud et al. 2018)	PRJNA259145, PRJNA259146, PRJNA259147, PRJNA373582, PRJNA373583, PRJNA373584, PRJNA373585, PRJNA373586, PRJNA373587, PRJNA373588, PRJNA373589, PRJNA373590, PRJNA373591, PRJNA373592, PRJNA373593, PRJNA373594, PRJNA373595, PRJNA373596, PRJNA373598, PRJNA411163 [a], PRJNA411164, PRJNA411165, PRJNA411166, PRJNA411167, PRJNA411168, PRJNA411169, PRJNA411170, PRJNA411171, PRJNA411172, PRJNA411173, PRJNA411174, PRJNA411175, PRJNA411176, PRJNA411177, PRJNA411178, PRJNA411179, PRJNA411180, PRJNA411181, PRJNA411182, PRJNA411202, PRJNA411203, PRJNA411204, PRJNA411205, PRJNA411206, PRJNA411207, PRJNA411208, PRJNA411209, PRJNA411210, PRJNA411211, PRJNA411212, PRJNA411213, PRJNA411214, PRJNA411215, PRJNA411216
	DEK1	(Demko et al., 2014)	PRJEB6339
	ABA response	(Stevenson et al., 2016)	SRX1176825, SRX1176826
	Shoot development	(Frank and Scanlon, 2015)	SRX803263, SRX803262, SRX803261, SRX803260, SRX803259, SRX803257, SRX682830, SRX682829, SRX682828, SRX682827, SRX682822, SRX682821
	Antheridia bundles	(Meyberg et al., 2020)	PRJNA559055
Kaskaskia	Protonema	This study	PRJNA601618
Reute	JGI Gene Atlas	(Perroud et al., 2018, Fernandez-Pozo et al., 2019)	PRJNA259147, PRJNA411184, PRJNA411185, PRJNA411186, PRJNA411187, PRJNA411188, PRJNA411189, PRJNA411190, PRJNA411191, PRJNA411192, PRJNA411193, PRJNA411194, PRJNA411195, PRJNA411196, PRJNA411197, PRJNA411198, PRJNA411199, PRJNA411200, PRJNA411201
	Antheridia bundles	(Meyberg et al., 2020)	PRJNA559055
	Sporophytic tissue, ES1	This study	PRJNA600210
Villersexel	Laser capture of sexual reproduction stages	This study	PRJNA602303
	Codon Usage Bias	(Szövényi et al., 2017)	PRJEB19978
Wisconsin	gDNA Single spores	This study	PRJNA602210

[a] Sample meta data are defining samples as accession Gransden, however, this study suggests that they are from accession Kaskaskia.

Table S10: P. patens Kaskaskia RNA-seq experiment description and accession number. *(original paper table S2)*

	Tissue	Medium	Treatment	SRA
K-I	Protonema, gametophore and sporophyte, 7 days and 3 months	BCD	Mixed: 1/3: protonema with 1 hour ABA 5µM, 5 NAA 5µM, BAP µM, 8°C treatments 1/3: ABA 5µM, 5 NAA 5µM, BAP µM, 8°C treatments 1/3: Three months gametophores with sporophytes	SRR10902189
K-II	Protonema, 7 day	BCD + 5mM glucose	Long day light condition	SRR10902188
K-III	Protonema, 7 day	BCD + 5mM glucose	No light	SRR10902187
K-IV	Protonema, 7 day	BCD	Harvest at 08h00 (dawn)	SRR10902186
K-V	Protonema, 7 day	BCD	Harvest at 12h00	SRR10902185
K-VI	Protonema, 7 day	BCD	Harvest at 16h00	SRR10902184
K-VII	Protonema, 7 day	BCD	Harvest at 20h00	SRR10902183
K-VIII	Protonema, 7 day	BCD + 5 mM Ammonium tartrate	6 days standard + 24h Glyphoste 5 mM	SRR10902182
K-IX	Protonema, 7 day	BCD	Long day light condition	SRR10902181
K-X	Protonema, 7 day	BCD +5 mM Ammonium tartrate	Long day light condition	SRR10902180

Table S11: Read distribution. *(original paper table S3)*

Accession	Project	Raw data	Uniquely mapped reads	used for GATK pipeline
Gransden	**JGI Gene Atlas**	2,803,540,901	2,275,015,526	1,986,071,695
	DEK1	142,060,456	117,626,868	57,701,716
	ABA response	71,105,859	46,327,636	12,673,554
	Shoot development	133,665,970	101,707,266	37,057,354
	Antheridia bundles	114,584,205	85,684,460	47,769,995
Kaskaskia	**Protonema**	650,738,592	541,188,206	420,973,321
Reute	**JGI Gene Atlas**	525,465,788	419,522,685	374,992,382
	Antheridia bundles	115,771,691	84,302,632	46,638,828
	Sporophytic tissue	39,203,989	33,804,665	9,340,399
Villersexel	**Laser capture of sexual reproduction stages**	19,569,737	13,820,142	7,105,192
	Codon Usage Bias	68,390,324	59,697,452	30,875,300
Wisconsin	**gDNA single spores**	1,062,163,031	845,313,241	473,057,571
	Sum (RNA-seq)	4,684,097,512	3,778,697,538	3,031,199,736
	Sum (gDNA)	1,062,163,031	845,313,241	473,057,571

Table S12: RNA-seq read distribution for each accession. *(original paper table S4)*

	Raw reads	After read filtering	Reads for SNP calling
Gransden	68%	68%	70%
Kaskaskia	12%	13%	14%
Reute	18%	16%	15%
Villersexel	2%	2%	1%

Table S13: Number of intersected RNA-seq SNPs to the gDNA SNPs from Lang et al. 2018. *(original paper table S5)*

	vcf [a]	cov_filter [b]	support_3 [c]	wo_indels [d]
Reute	20,221 (9%)	12,911 (13%)	12,601 (14%)	10,988 (26%)
Villersexel	220,663 (54%)	65,884 (82%)	50,528 (86%)	47,031 (89%)
Kaskaskia	32,307 (9%)	23,844 (15%)	23,779 (17%)	21,644 (28%)

The percentage values in brackets are representing the proportion of RNA-seq SNPs found in the gDNA SNP set. [a] Number of raw SNPs called by the GATK pipeline. [b] Number of RNA-seq SNPs filtered by read coverage, c plus minimum three samples need to support the SNP and d in addition to the filter steps before, the indels were removed.

Table S14: SNP normalization methods. *(original paper table S6)*

Accession, Pedigree Sample	SNP / covered bp	SNP / number of reads	SNP /number of genes	SNP / gene length [bp]	covered bp / SNP
Gd	0.00034	0.00094	1.206	0.00040	4,666
Gd_DE	0.00038	0.00113	-	-	2,871
Gd_UK	0.00028	0.00090	-	-	6,206
2004	0.00006	0.00009	-	-	15,512
Birmingham	0.00035	0.00117	-	-	3,104
Gd_CH	0.00042	0.00105	-	-	2,402
Gd_JP	0.00022	0.00049	-	-	9,526
St. Lousi PE	0.00044	0.00102	-	-	2,284
St. Lousi SE	0.00007	0.00010	-	-	15,324
Okazaki & Columbia	0.00038	0.00090	-	-	2,756
Ka	0.00159	0.00297	2.300	0.00076	630
Re	0.00058	0.00178	1.280	0.00043	1,912
Vx	0.01761	0.01488	1.607	0.00053	143
Wi	0.06641	0.00363	1.925	0.00064	206

Table S15: Ploidy of P. patens plants used for sequencing. *(original paper table S10)*

***P. patens* accession / pedigree**	**Ploidy tested**	**haploid?**	**Differences in GATK SNP calling 1n and 2n**
Gd_Beijing-2010	N/A		63.4%
Gd_Berlin	yes	x	51.7%
Gd_Birmingham	no		47.3%
Gd_Birmingham-2009	no		61.5%
Gd_Columbia-2010	N/A		45.7%
Gd_FR-WT9	yes	x	56.9%
Gd_FR-WT11	yes	x	56.9%
Gd_Gransden-2004	yes	x	62.8%
Gd_Gransden-WT9	yes	x	55.2%
Gd_MR-WT11	no		56.0%
Gd_MR-WT12	yes	x	61.0%
Gd_MR-WT15	yes	Phenotypically 1n [a]	36.2%
Gd_Okazaki-1998	N/A		60.6%
Gd_Padova-2008	N/A		58.8%
Gd_St.Louis-2007	N/A		63.8%
Gd_St.Louis-2007_JGI	no	offspring phenotypically 1n [b]	58.9%
Gd_St.Louis	N/A		30.7%
Gd_Uruguay-WT9	yes	x	68.0%
Gd_Versaille	N/A		59.2%
Gd_WT-Grenoble	N/A		57.3%
Ka_kaskaskia	no	offspring phenotypically 1n [b]	81.1%
Re_Reute-2007	yes	x	67.7%
Re_Reute-2012	no		61.9%
Re_Reute-2015	yes	Phenotypically 1n [a]	59.4%
Vx_K3	no		90.7%
Vx_unknown	yes	offspring, allele-specific expression 1n	92.9%
Wi_2018	no		88.1%

[a] Plants were check by phenotypical observation (polyploid plants usually have a 'fluffy' habitus that haploid plants do not exhibit). [b] Offspring of the used plants were checked by phenotypical observation.

The colours in the last column indicates the differences of the GATK 1n and 2n runs.

Table S16: Total number of synonymous and non-synonymous SNPs. *(original paper table S11)*

Accession/pedigree	**Non-synonymous**	involved genes	**Synonymous**	involved genes	**Ka/Ks**
Gd	62,142	579	69,870	421	0.889
Gd_FR-WT9 *	2,743	241	3,162	172	0.867
Gd_Uruguay-WT9 *	2,408	226	2,822	167	0.853
Gd_Gransden-WT9 *	8,308	276	9,440	198	0.880
Gd_FR-WT11 *	1,471	201	1,583	138	0.929
Gd_MR-WT11 *	6,415	291	6,840	208	0.938
Gd_MR-WT12 *	7,575	253	8,327	184	0.910
Gd_MR-WT15 *	1,143	213	1,631	145	0.701
Re	55,916	1725	39,416	1,065	1.419
Re_Reute-2007 *	10,063	1473	7,246	889	1.389
Re_Reute-2012 *	39,233	1,575	27,429	973	1.430
Re_Reute-2015 *	6,620	1,461	4,741	882	1.396
Ka	142,009	5,599	93,964	4,020	1.511
Vx	44,976	6,703	34,270	5,391	1.312
Wi	208,930	7,834	118,430	5,115	1.764

*The sum of all detected synonymous and non-synonymous SNPs are shown. If a gene contains multiple effects, it was only counted once. The 99% confidence interval of the linear regression of Ka/Ks over all samples is 1.1..1.2, i.e. all samples shown here are above or below the confidence interval, suggesting negative (Gd) or positive selection (other pedigrees) (Supplementary File 8). * Pedigrees with known propagation history.*

Table S17: Average number of SNPs (filtered by all three methods) per sample with documented propagation history. *(original paper table S12)*

***P. patens* accession / pedigree**	**Average SNP number**	**Source**
Gd_FR-WT9	6,452	Perroud et al. 2018
Gd_Uruguay-WT9	7,851	Perroud et al. 2018
Gd_Gransden-WT9	6,778	JGI_2nd_round
Gd_FR-WT11	7,244	Perroud et al. 2018
Gd_MR-WT11	7,271	Perroud et al. 2018
Gd_MR-WT12	6,221	JGI_2nd_round
Gd_MR-WT15	2,555	Meyberg et al. 2020
Re_Reute-2007	13,282	Perroud et al. 2018
Re_Reute-2012	11,927	JGI_2nd_round
Re_Reute-2015	8,635	Meyberg et al. 2020

In the first column, mutant or wildtype experiments are annotated. The colours in the middle column indicate the average number of SNPs called for that experiment (green represents high numbers, blue low numbers).

Table S18: Mutations per year and site for all pedigrees with a documented propagation history. *(original paper table S13)*

	Total SNPs [a]	Unique SNPs [b]	Mutations per year	Mutation rate [c]
WT9	31,230	3,745		
Gd_MR-WT11	26,525	1,483	742	1.571E-06
Gd_MR-WT12	23,557	642	642	1.361E-06
Gd_MR-WT15	3,757	1,047	349	7.396E-07
WT9	31,230	18,435		
FR-WT11	13,202	407	136	2.875E-07
Re_Reute-2007	27,267	3,327		
Re_Reute-2012	36,051	10,560	2112	4.476E-06
Re_Reute-2015	13,266	2,234	745	1.578E-06

[a] SNPs had to pass all three filter methods. [b] Unique SNPs derived from the intersection of the shown pedigrees (WT9 and Re 2007 are the respective ancestors). [c] The P. patens genome size is 471,852,792 bp.

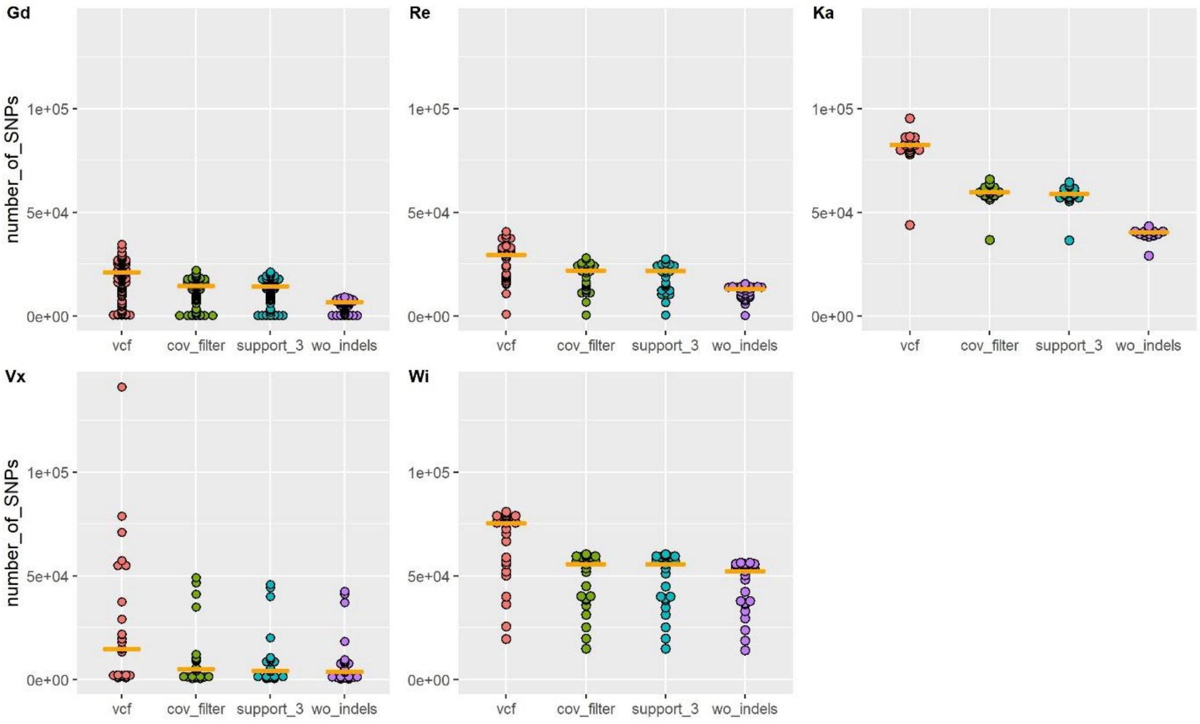


Figure S6: SNP reduction by filtering. *Each of the five plots shows the GATK SNP output on the left (no filter [vcf]) and the different filter steps applied to the data on the right (read coverage filter [cov_filter] and >= three samples supporting the SNP [support_3]). The strictest filter is rightmost (removing indels [wo_indels]). Each filter step is based on the remaining SNPs after applying the previous filter (for each accession left to right). Individual samples are represented by dots, the median by an orange horizontal bar. For details of filter steps refer to Materials and Methods. (original paper figure S1)*

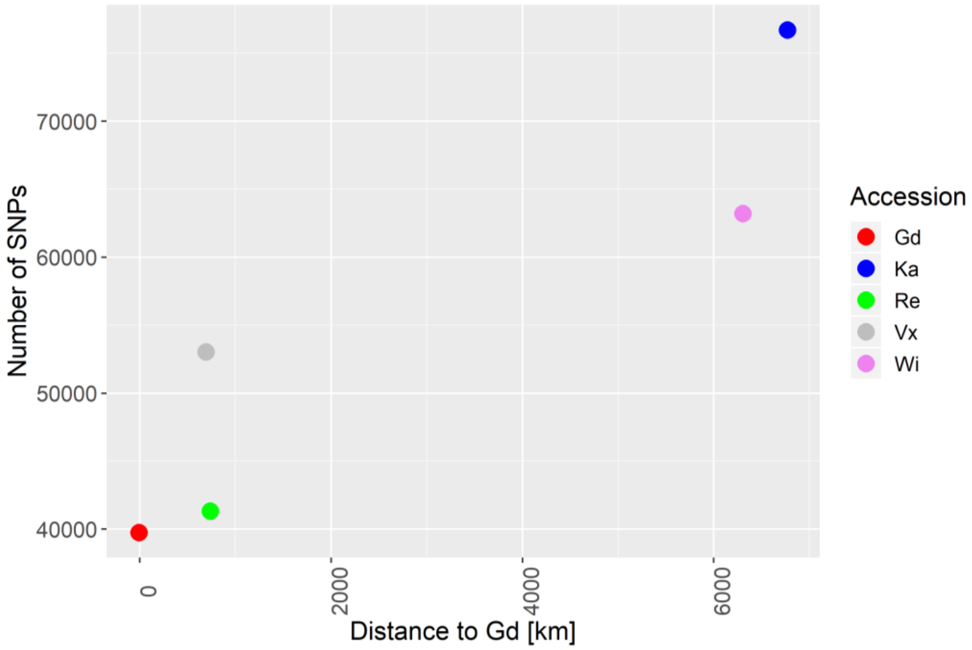


Figure S7: Number of SNPs applied to the line of sight distance to the reference Gransden. *Distance to Gd: Vx = 700 km; Re = 730 km, Wi = 6,300 km; Ka = 6,750 km. (original paper figure S2)*

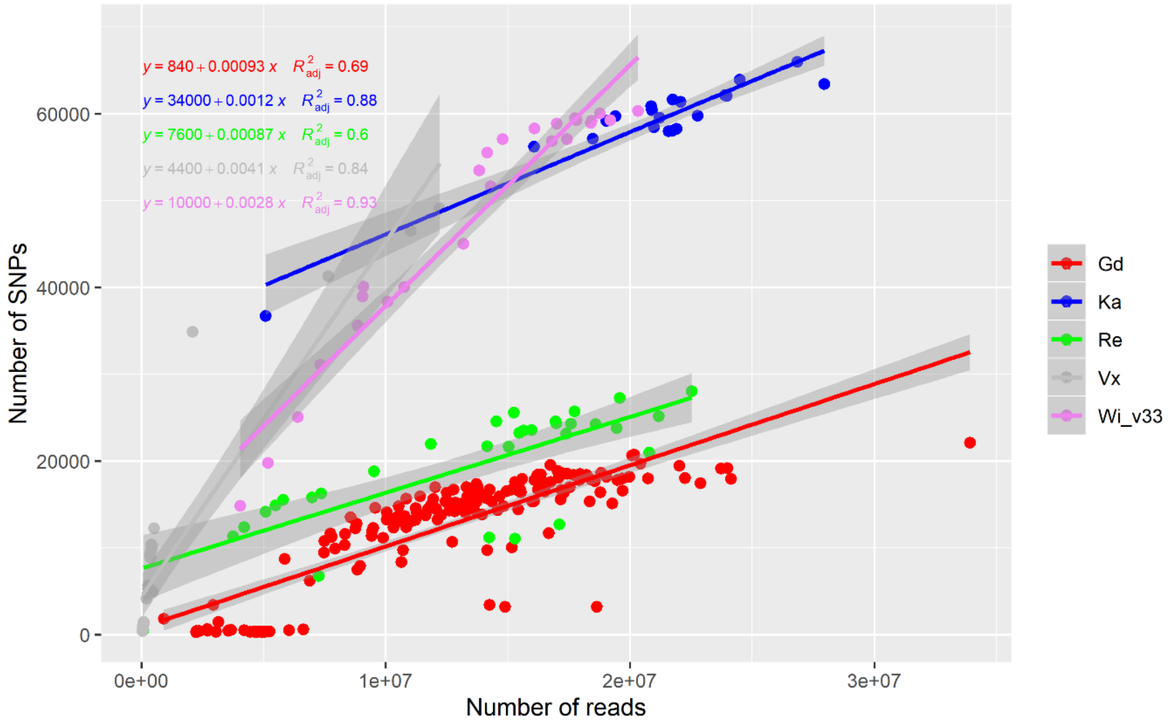


Figure S8: SNP normalization by read number. *Number of SNPs (filtered by read coverage, minimum support of three samples and removed indels) to the number of reads per sample. The slope is calculated by R method "lm" and in grey the confidence interval is shown around the regression lines. (original paper figure S3)*

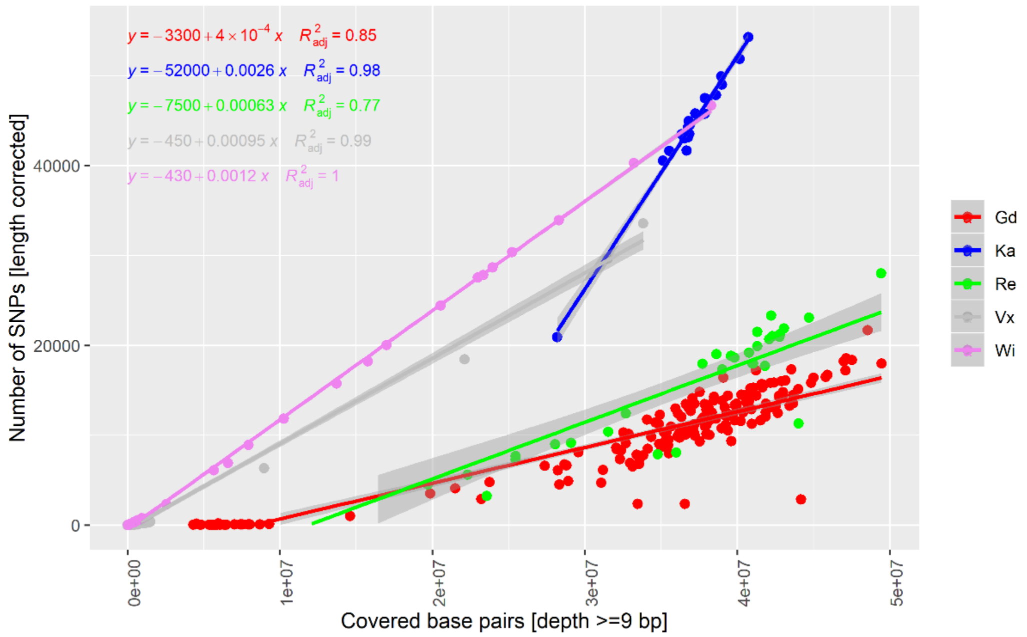


***Figure S9: SNP normalization by covered base pairs**. Covered base pairs with read depth >= nine applied to the number of SNPs, filtered by read coverage, sample support of three and removed indels. The number of SNPs were corrected by the maximum number of covered base pairs. The slope was calculated by the R method “lm” and in grey the confidence interval is shown around the regression lines. (original paper figure S4)*

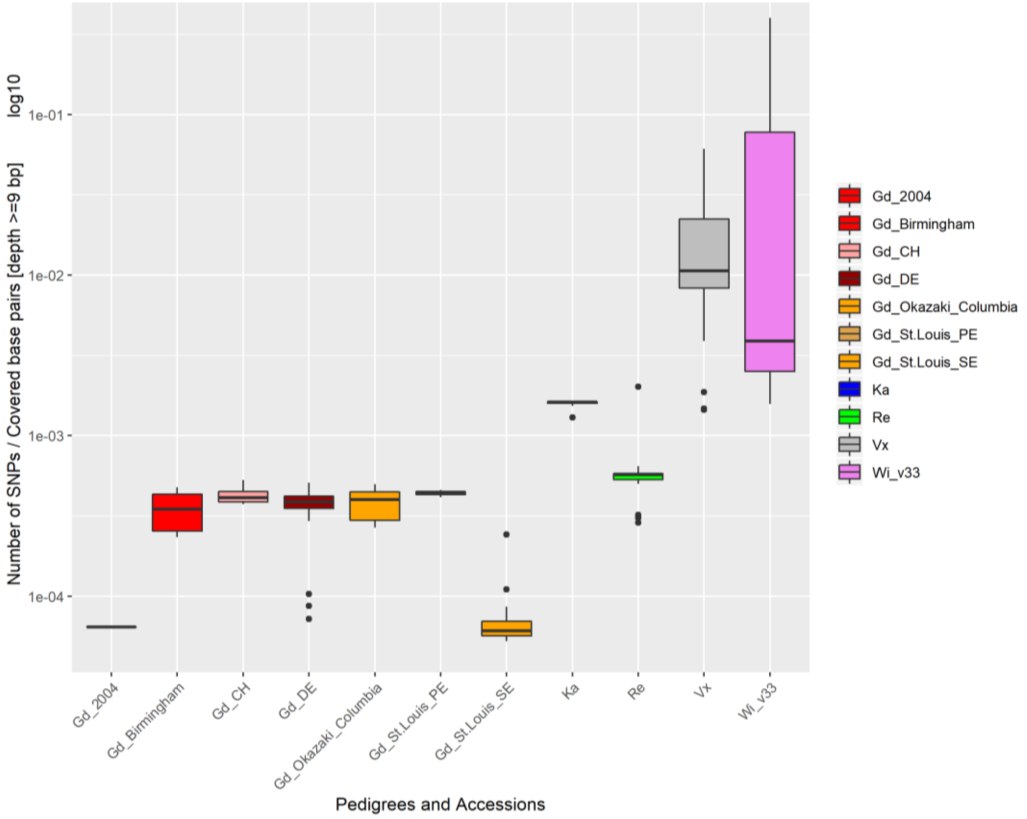


***Figure S10: Number of SNPs divided by the number of covered base pairs.** The number of SNPs (three filters applied) divided by the number of covered base pairs of all samples shown as boxplots and grouped by accessions and pedigrees. Gd_UK_2004 and Gd_JP_St.Louis_SE have the lowest number of SNPs per covered base pair, Vx and Wi have the most. (original paper figure S5)*

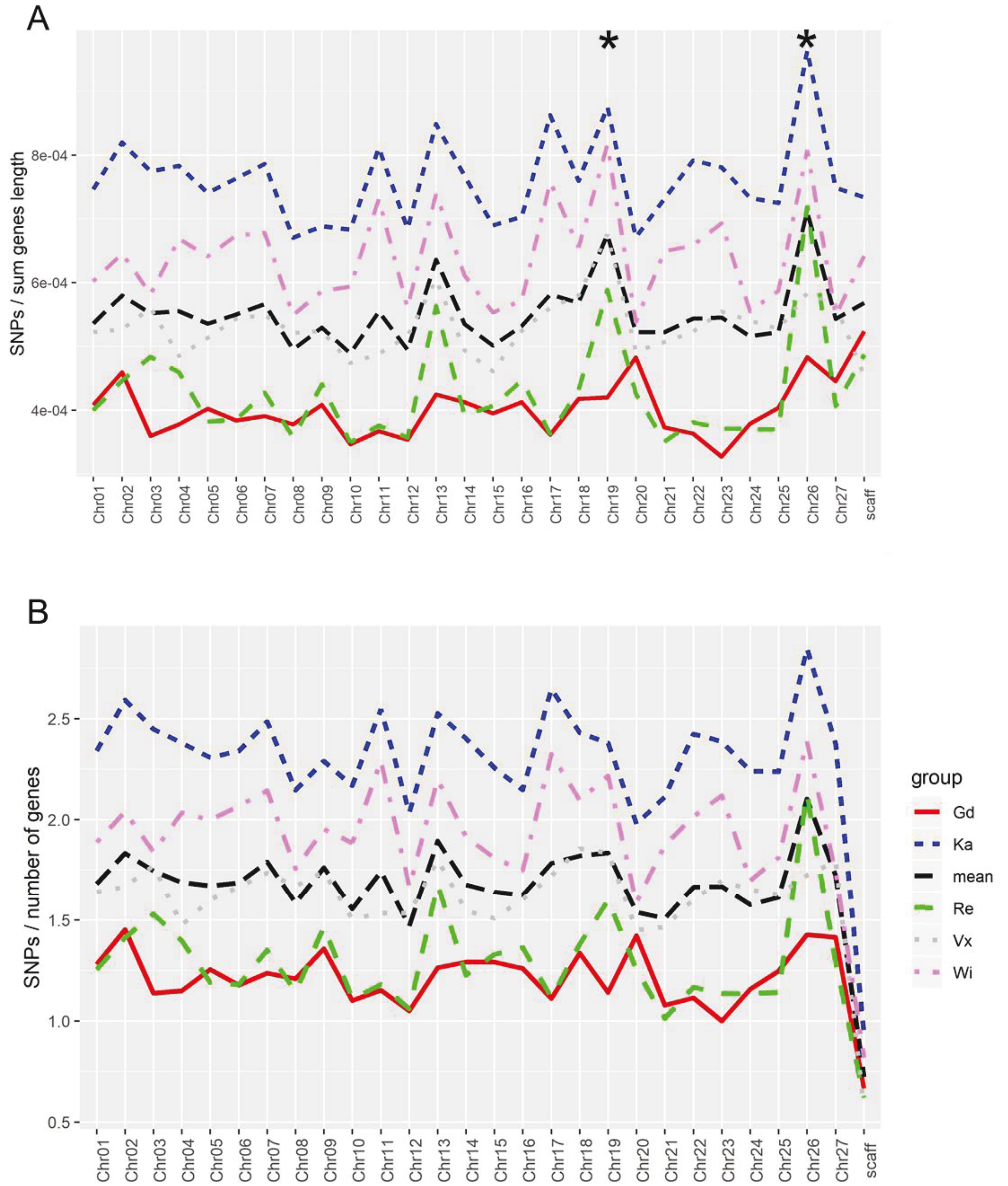


Figure S11: SNP to gene correlation. *The number of SNPs divided by the length of all genes per chromosome is shown in A), while in B) the SNPs are divided by the number of genes. Gd has an average of 1.2 SNPs per gene, while the SNP rate for Ka is almost double with 2.3 SNPs per gene. The black line represents the mean of all accessions. Chr19 and 26 appear to exhibit significantly more SNPs per gene length (Fisher's exact test, number of base pairs w/o SNPs for each individual chromosome, p <= 0.05). (original paper figure S6)*

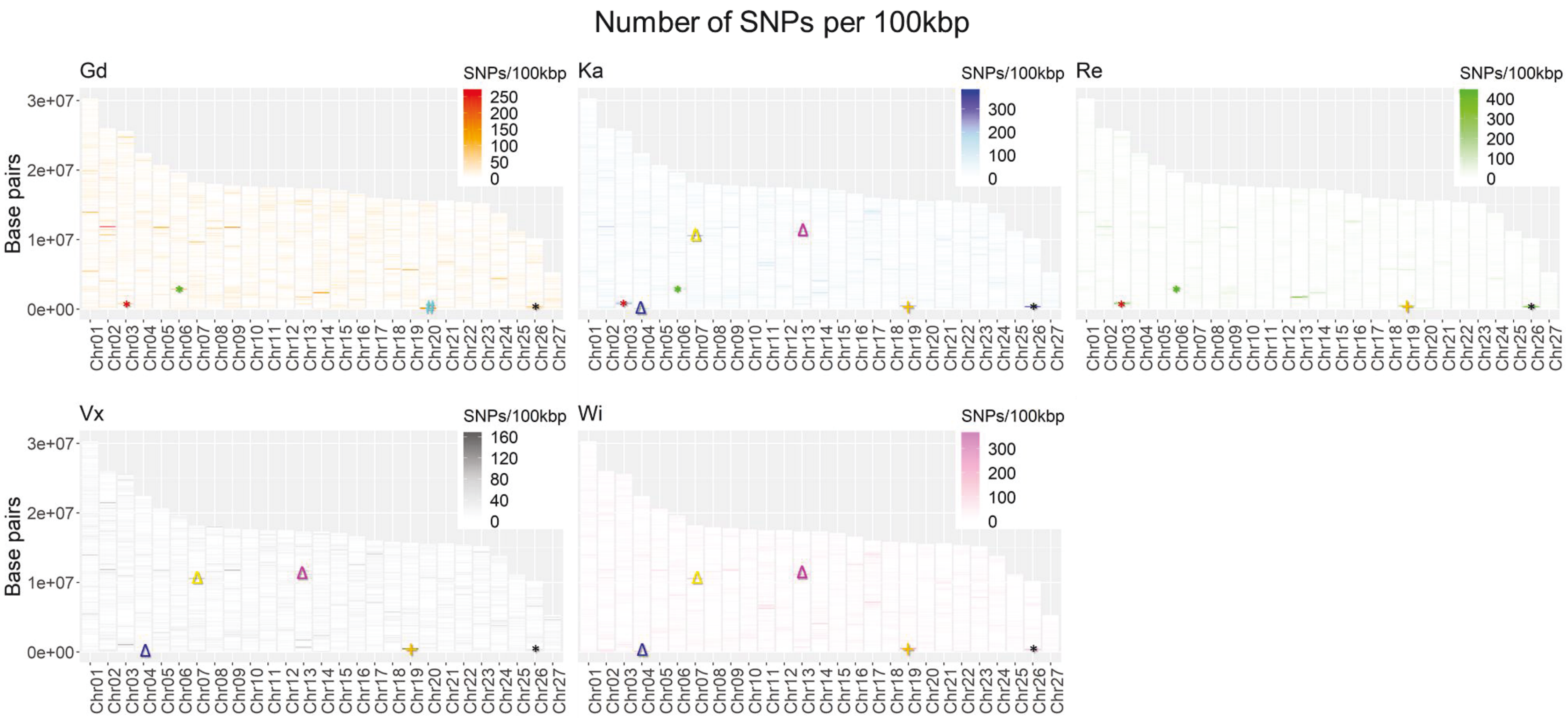


Figure S12: Number of SNPs per 100 kbp in each chromosome. *Each accession has its own shape of SNP hotspots. E.g. Gd has an exclusive 100kbp hotspot on Chr20 (starting at position 1) [light blue hash];. Some regions with high SNP numbers are shared between the accessions. Chr26 (start 300,000 bp) [black asterisk] is present in all accessions. Gd, Re and Ka share 100kbp hotspot regions on Chr03 (start 700,000 bp) [red asterisk] and one on Chr06 (start 2,800,000 bp) [green asterisk]; Ka, Wi, and Vx share regions on Chr04 (start 1 bp) [blue triangle], Chr07 (start 10,500,000 bp) [yellow triangle] and Chr13 (start 11,100,000 bp) [violet triangle]; all accessions except Gd share Chr19 (starting position 400,000 bp) [orange plus]. (original paper figure S7)*

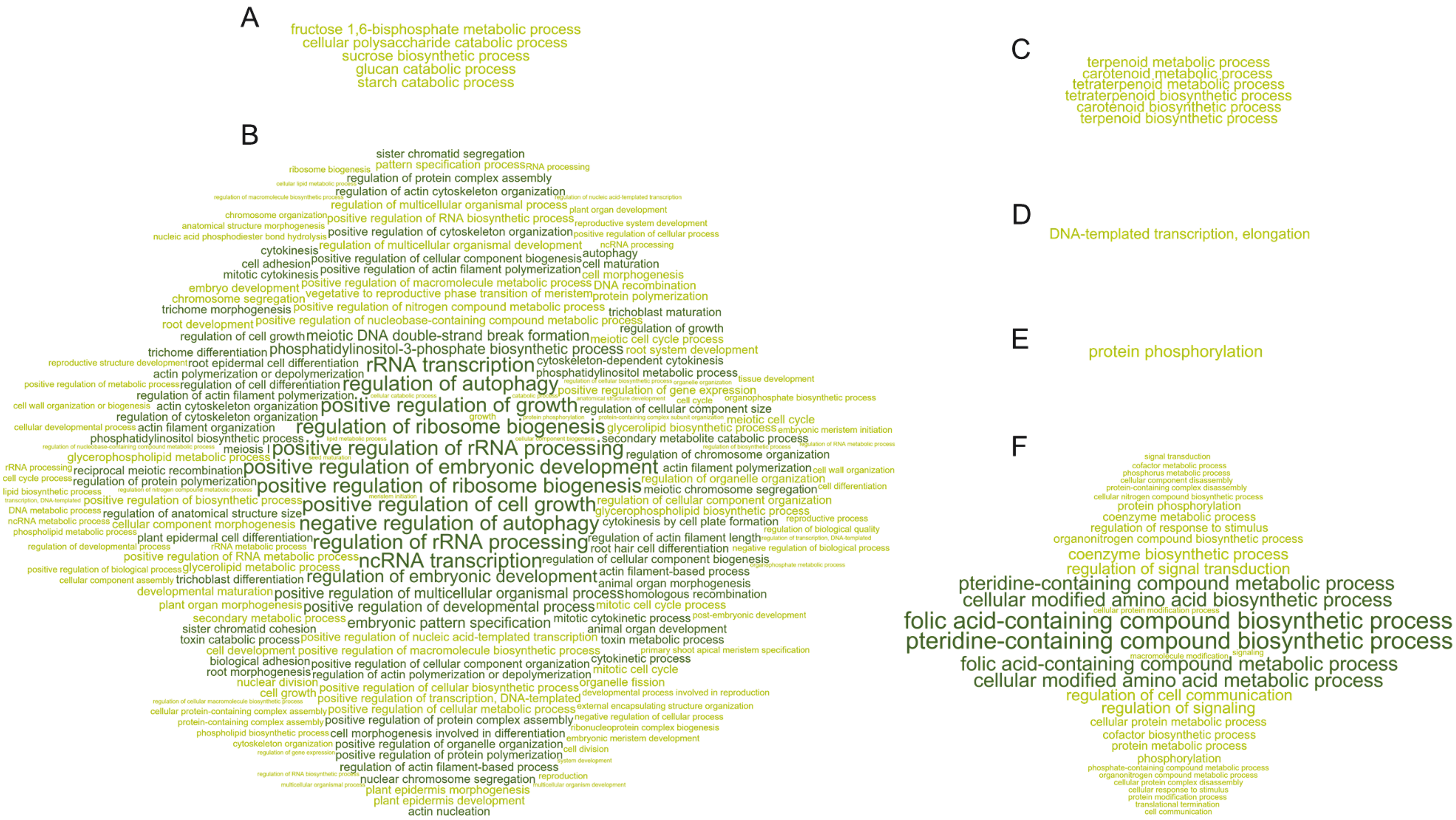


Figure S13: Word clouds of GO bias analyses of the 100 kbp hotspot regions marked in Figure S12. *A) Chr03 700 kbp, 12 genes B) Chr06 2,800 kbp, three genes C) Chr07 10,500 kbp, seven genes D) Chr13 11,100 kbp, 12 genes E) Chr19 400 kbp, 15 genes F) Chr26 300 kbp, 14 genes. The weight of the given terms 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. (original paper figure S8)*

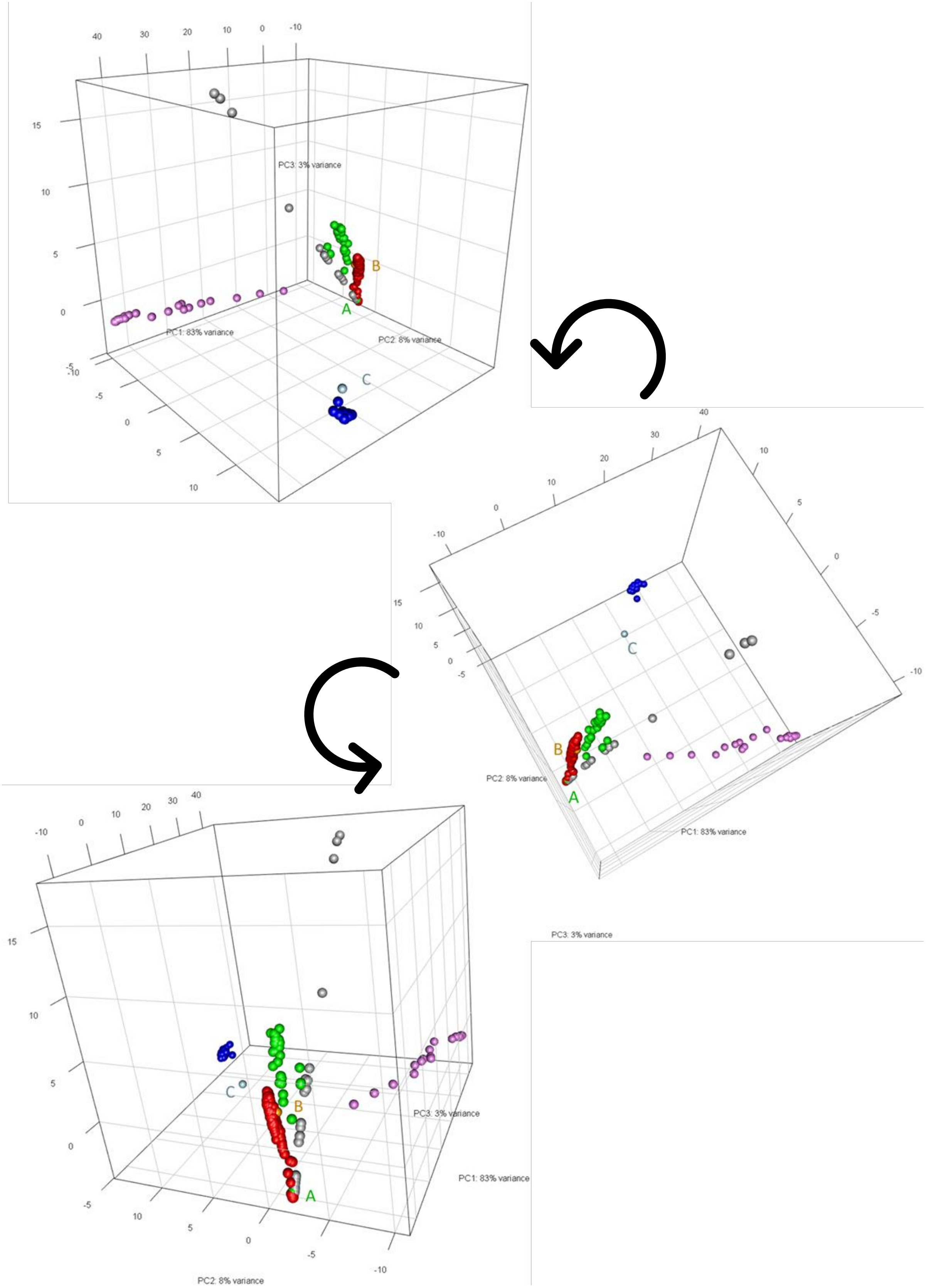


Figure S14: 3D PCA plot of filtered SNPs and InDels. *The PCA shows clustering by accessions similar to the SNP-only clustering (Paper 5.4, Figure 4). The three plots are showing the cube rotated on its axis. Color scheme: Gd = red; Re = green; Ka = blue; Vx = grey; Wi = violet. The three outlier samples shown in paper 5.4 Figure 4 are labeled: A) CI_3, B) CIV_1 and C) XVIII_1. (original paper figure S9)*

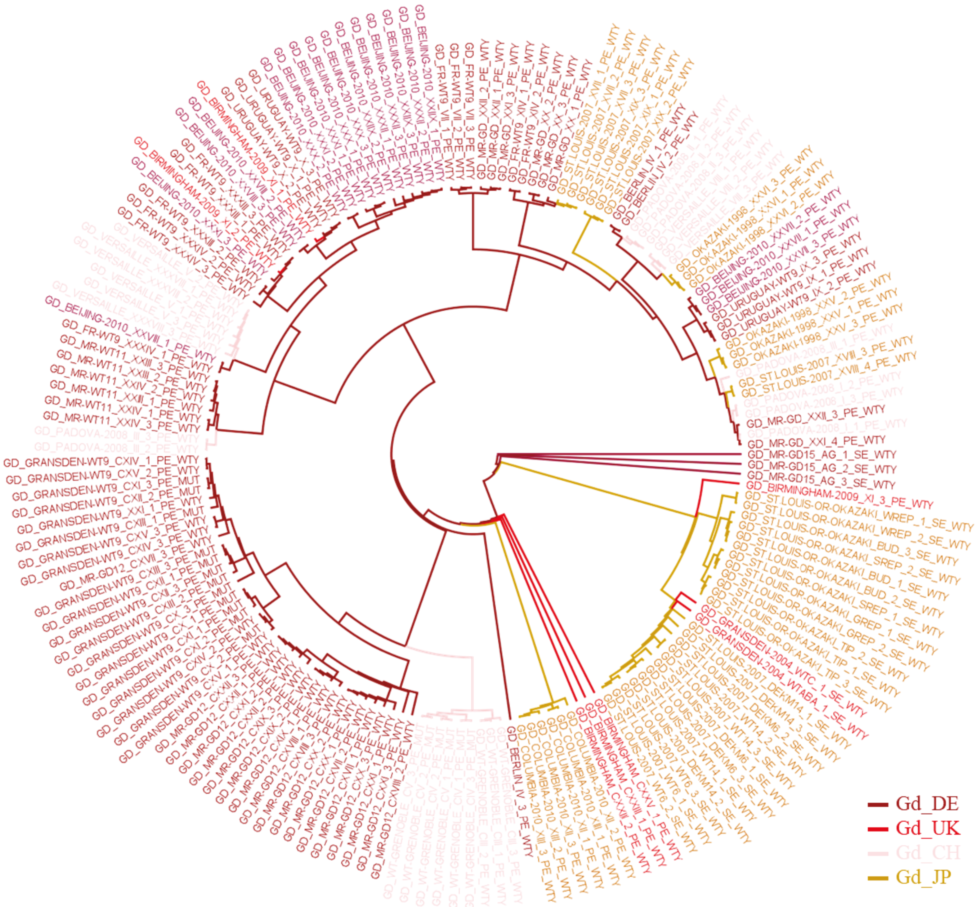


Figure S15: Gransden pedigree tree based on the artificial FASTA alignment. *Samples clustering by experiments can be observed. A clear clustering by pedigree is not observable. (original paper figure S10)*

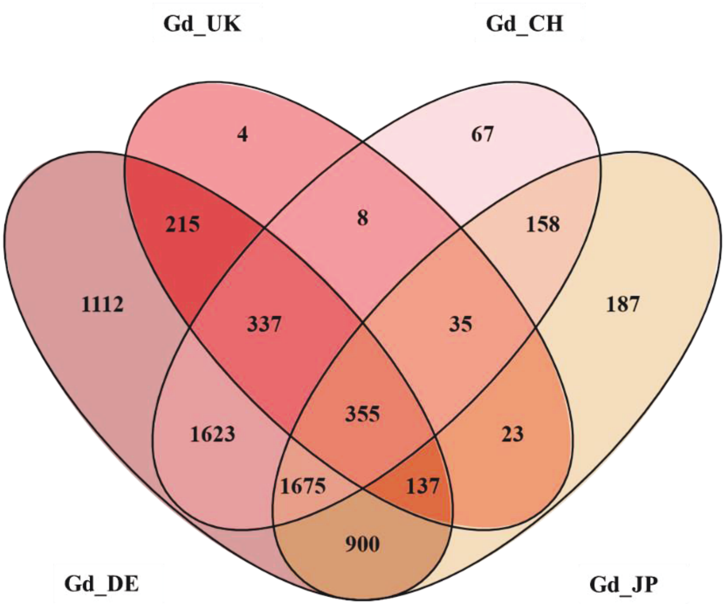


Figure S16: Intersection of all SNPs detected in the four Gransden pedigrees. *(original paper figure S11)*

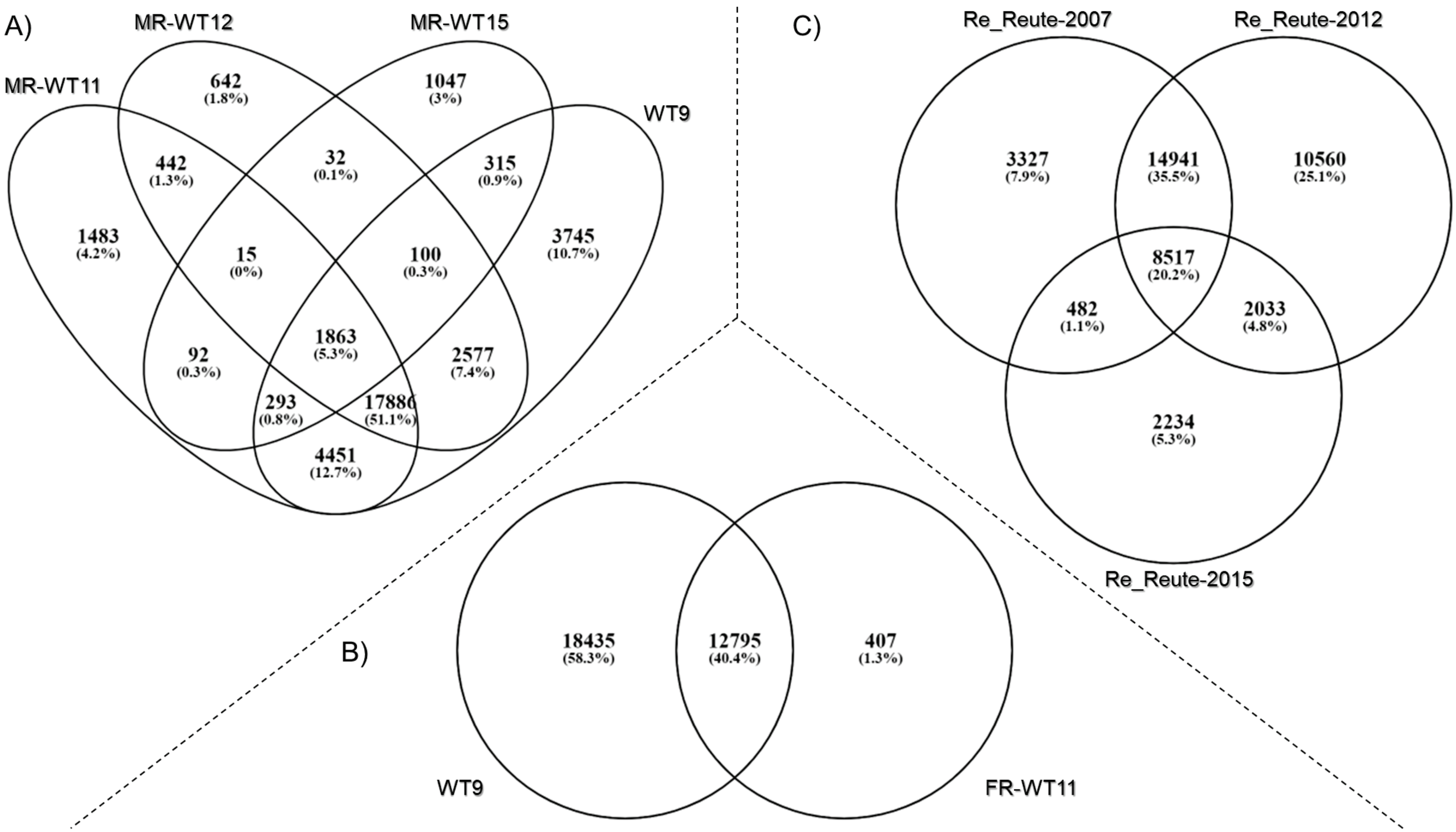


Figure S17: Intersection of all SNPs called for pedigrees with known propagation history. *The SNPs were filtered by all three filter methods. A) WT9 compared to Gd Marburg (MR) samples. B) WT9 compared to Freiburg (FR) sample WT-2011 and C) Three different pedigrees from accession Reute were compared. (original paper figure S12)*

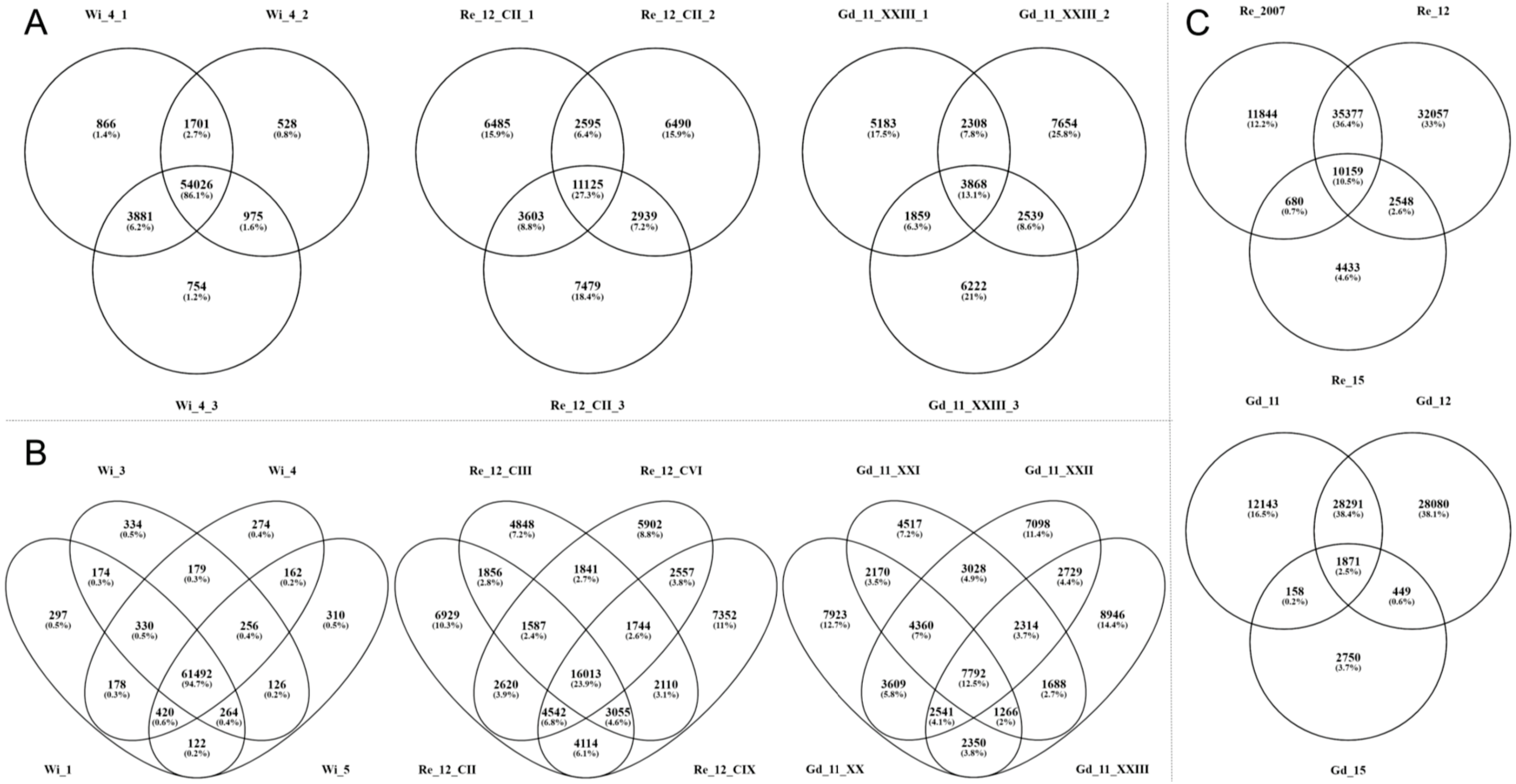


Figure S18: SNP intersection of Wi gDNA samples and selected Gd and Re pedigrees. *SNP intersection visualized by Venn diagrams. A) SNP intersection at sample replicate level. The leftmost diagram shows the SNP intersection of three sample replicates of the fourth Wi spore capsule. The middle and the right venn diagram show sample replicates of the experiment Re_REUTE-2012_CII and Gd_MR-WT11_XXIII. B) SNP intersection at experiment level. The left Venn diagram shows Wi gDNA SNPs of four spore capsule each consisting of five merged spore samples. The middle diagram shows the SNP intersection of all SNPs called in four different Reute_2012 experiments (CII, CIII, CVI and CIX). The rightmost Venn diagram shows the SNP intersection of four different Gd_DE_2011 experiments from Marburg (XX, XXI, XXII and XXIII). C) SNP intersection at pedigree level. The upper Venn diagram shows the SNP intersection of three Re generations (2007, 2012 and 2015). The lower Venn diagram shows the SNP intersection of three Gd_DE generations from Marbug (2011, 2012 and 2015). (original paper figure S18)*

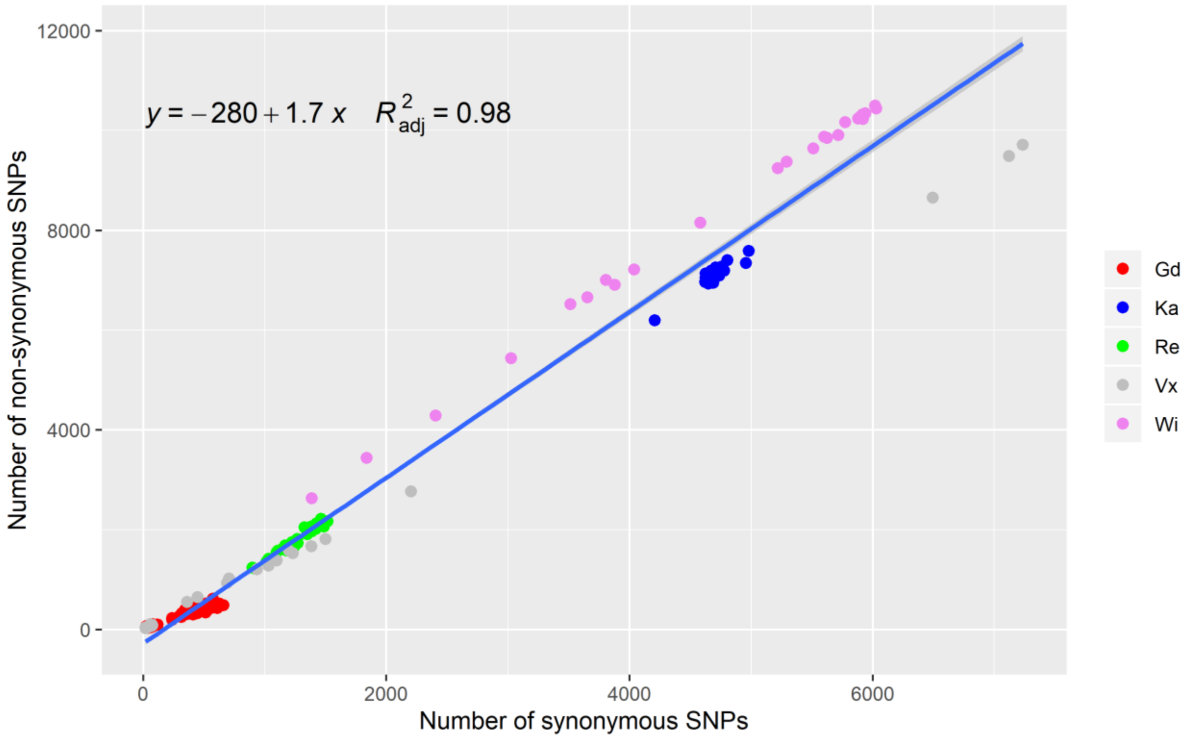


Figure S19: Ka / Ks plot of all samples. *(original paper figure S19)*

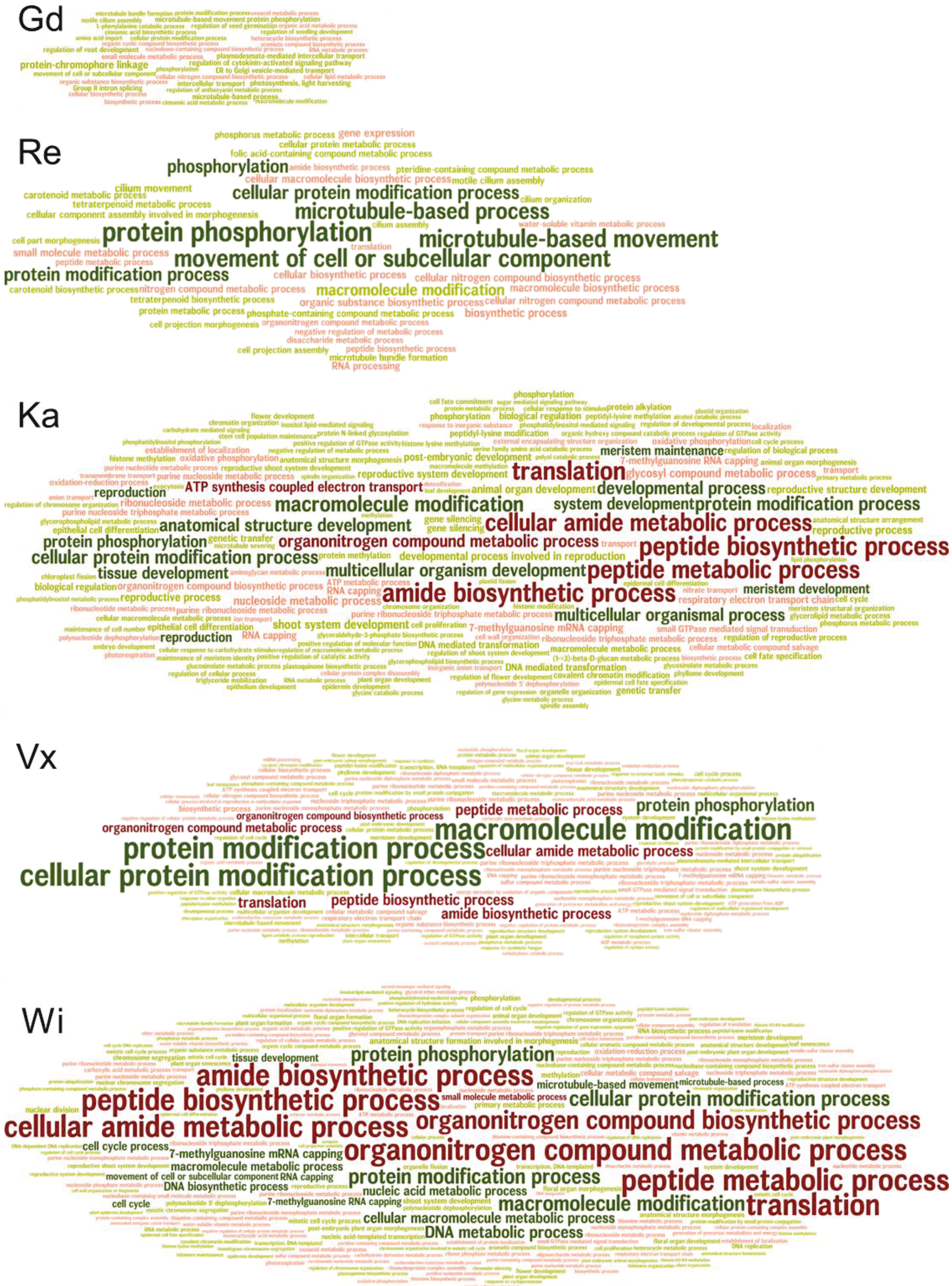


Figure S20: GO bias analysis of genes affected by non-synonymous SNPs. *The word clouds are separated by accessions. The weight of the given terms was defined as the −log10(q-values) and the colour scheme used for the visualization was red for under-represented GO terms and green for those over-represented. Terms with stronger representation, i.e. weight >4, were represented with darker colours. (original paper figure S20)*

Script “rename_and_extraction”

```
#!/bin/bash

echo "Rename samples and extract Wi v3.3 SNPs"

# rename all vcf files
for IN in *.vcf; do mv $IN `awk -v name=$IN -v prefix=$prefix 'name~$1 {print
$2"_"prefix".vcf"}' sample_translation.list`; done

# extract SNPs at gene regions based on gDNA samples
for IN in Wi_*.vcf; do bedtools intersect -u -a $IN -b Pp_v3.3.gff -header >
$IN".Pp_v3.3.vcf"; done

exit
```

Script “SNP_clustering”

```
#!/bin/bash

echo "SNP clustering"

prefix=$1     # surfix/run name
FI=$2         # samples supporting the SNP
DP=$3         # read depth
ADa=$4        # reads supporting the SNP

# fold change of the reads supporting the SNP
ADr=0.$(( ($ADa * 100) / $DP ))
AD=$ADa"x"$ADr


# clustering SNPs and InDels, 0/1/2-matrix
awk '{if(NR==FNR){for(i=5;i<=NF;i++){foo[$i]}; header[$1"_"$2"_"$3"_"$4]} else
{for(a in foo){if($0~a){if((length($3)+length($4))>2){foo[a]=foo[a]"\t"2} else
{foo[a]=foo[a]"\t"1}} \
  else{foo[a]=foo[a]"\t"0}}} } END {printf "sample\t\t"; for(h in header){printf
h"\t"};printf "\n"; for(b in foo){split(b,c,"/"); print c[1]"\t"foo[b]}}' \
  $prefix".plus_filter.filter"$FI".wo_indel.rna_snps"
$prefix".plus_filter.filter"$FI".wo_indel.rna_snps" | \
  sed 's/\t\n/\n/g' | sed 's/\t$//' >
$prefix".plus_filter.filter"$FI".wo_indel.0_1_2.rna_list"

# create SNP FASTA alignment file
awk '{if(NR==FNR){for(i=5;i<=NF;i++){foo[$i]}} else {for(a in
foo){if($0~a){foo[a]=foo[a]$4} else{foo[a]=foo[a]$3}}} } END {for(b in
foo){split(b,c,"/"); print ">"c[1]"\n"foo[b]}}' \
  $name".plus_filter.filter"$FI".wo_indel.rna_snps" \
  $name".plus_filter.filter"$FI".wo_indel.rna_snps" \
  > $name".plus_filter.filter"$FI".wo_indel.rna_snps.fasta"

exit
```

Script “SNP_filtering”

```
#!/bin/bash

echo "SNP filtering"

FI=$1         # samples supporting the SNP
DP=$2         # read depth
ADa=$3 # reads supporting the SNP
name=$4       # prefix/run name

# fold change of the reads supporting the SNP
ADr=0.$(( ($ADa * 100) / $DP ))
AD=$ADa"x"$ADr

# Merging all SNPs into one file.
# column 1-4: chromosome, position, reference, SNP
# column 5-n: vcf files containing the SNP
# Is a SNP passing the read coverage filter,
#    a ‘+’ is added to the sample name
awk -v dp=$DP -v adr=$ADr -v ada=$ADa '$1!~"#" {split($10,a,":");
split(a[2],b,","); if(a[3]>=dp&&b[2]>=ada) {fc=b[2]/a[3];
if(fc>=adr){get_name=FILENAME"+"} else {get_name=FILENAME}} else
{get_name=FILENAME};
foo[$1"\t"$2"\t"$4"\t"$5]=get_name"\t"foo[$1"\t"$2"\t"$4"\t"$5]} END {for(u in
foo){print u"\t"foo[u]}}' Ka*vcf Gd*vcf Vx*vcf Re*vcf Wi*8.vcf > $name".rna_snps"

# read coverage filter for all vcf files
for IN in *.vcf; do awk -v dp=$DP -v adr=$ADr -v ada=$ADa '$1!~"#"
{split($10,a,":"); split(a[2],b,","); if(a[3]>=dp&&b[2]>=ada) {fc=b[2]/a[3];
if(fc>=adr) {print }}}' $IN > $IN".filterDP"$DP"AD"$AD ; done

# support by at least $FI samples
awk -v FI=$FI 'NF>=4+FI {print }' $name".rna_snps" > $name".filter"$FI".rna_snps"

# remove InDels
awk 'length($3)==1&&length($4)==1 {print }' $name".rna_snps" >
$name".wo_indel.rna_snps"

# extract all SNPs for each accession
for IN in Gd_ Re_ Ka_ Vx_ Wi_; do grep $IN $name".rna_snps" > $IN$name".rna_snps";
done

# extract all SNPs for each Gd pedigree
grep -E "MR|Beij|FR|WT9|Berlin" Gd_$name".rna_snps" | grep "\+" >
Gd_DE.$name".plus_filter.rna_snps"
grep -E "Birm|2004_" Gd_$name".rna_snps" | grep "\+" >
Gd_UK.$name".plus_filter.rna_snps"
grep -E  "Padov|Versai|Greno" Gd_$name".rna_snps" | grep "\+" >
Gd_FR.$name".plus_filter.rna_snps"
grep -E "Colum|Okaza|Louis" Gd_$name".rna_snps" | grep "\+" >
Gd_JP.$name".plus_filter.rna_snps"

# extract all exclusive SNPs (here for Gd)
grep -Ev "Re_|Vx_|Wi_|As_|Ka_" $name".rna_snps" | awk '{print NF-4"\t"$0}' | sort -
k1V > Gd.exclusive

# extracting Gd pedigree exclusive SNPs (Gd.exclusive)
grep -Ev "MR|Bei|FR|WT9|Berlin|Birm|2004_|Padov|Versai|Greno" Gd.exclusive | sort -
V > Gd_JP.exclusive
grep -Ev "Padov|Versai|Greno|Colum|sdenE|Okaza|Louis|MR|Bei|FR|WT9|Berlin"
Gd.exclusive | sort -V > Gd_UK.exclusive
grep -Ev "MR|Bei|FR|WT9|Berlin|Birm|2004_|Colum|sdenE|Okaza|Louis" Gd.exclusive |
sort -V > Gd_FR.exclusive
grep -Ev "Birm|2004_|Padov|Versai|Greno|Colum|sdenE|Okaza|Louis" Gd.exclusive |
sort -V > Gd_DE.exclusive


exit
```

9.5 DOE JGI Gene Atlas Project

9.5.1 Re-sequencing of plate19

One special situation occurred during sequencing the second round. Three pentatricopeptide repeat (PPR) knock-out experiments needed the organellar RNA sequences (Supporting information 9.3, Table S8). The normal library preparation protocol used by the JGI removes the organellar sequences. Since the missing organellar reads were essential, 54 libraries of the 2nd round, all on plate 19 (https://genome.jgi.doe.gov/portal/PhypatAtlPlate19/PhypatAtlPlate19.info.html), were re-sequenced. The aim was to get organellar sequences for the PPR experiments. As a result of this, all samples located on plate 19 are available with and without organellar expression data.

10 Acknowledgments

Die Seite 134 (Danksagung) enthält persönliche Daten. Sie ist deshalb nicht Bestandteil der Online-Veröffentlichung.

11 Curriculum Vitae

Die Seiten 135-139 (Lebenslauf) enthalten persönliche Daten. Sie sind deshalb nicht Bestandteil der Online-Veröffentlichung.

12 Declarations

Die hier vorgelegte Arbeit „*P. patens* genomic and transcriptomic analyses“, wurde ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt. Alle Daten die direkt oder indirekt aus anderen Quellen übernommenen wurden, sind unter Angabe der Quellen gekennzeichnet. 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. Dies ist mein erster Promotionsversuch.

Marburg, den 20.05.2020