De novo assembly and comparative analysis of the Ceratodon purpureus transcriptome

PÉTER SZÖVÉNYI,*†‡§ PIERRE-FRANÇOIS PERROUD,¶ AIKATERINI SYMEONIDI,** SEAN STEVENSON,†‡ RALPH S. QUATRANO,¶ STEFAN A. RENSING,** ANDREW C. CUMING†† and STUART F. MCDANIEL‡‡

*Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstr. 190, CH-8057 Zurich, Switzerland, †Institute of Systematic Botany, University of Zurich, Zollikerstr. 107, CH-8008 Zurich, Switzerland, ‡Quartier Sorge-Bâtiment Genopode, Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland, §MTA-ELTE-MTM Ecology Research Group, ELTE, Biological Institute, H-1117, Pázmány P. sétány 1/C, Budapest, Hungary, ¶Biology Department, Washington University in St Louis CB #1137, One Brookings Drive, St Louis, MO 63130, USA, **Faculty of Biology, University of Marburg, Karl-von-Frisch-Str. 8, D-35043 Marburg, Germany, ††Centre for Plant Sciences, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, UK, ‡‡Department of Biology, University of Florida, PO Box 118525, Gainesville FL 32611, USA

Abstract

The bryophytes are a morphologically and ecologically diverse group of plants that have recently emerged as major model systems for a variety of biological processes. In particular, the genome sequence of the moss, Physcomitrella patens, has significantly enhanced our understanding of the evolution of developmental processes in land plants. However, to fully explore the diversity within bryophytes, we need additional genomic resources. Here, we describe analyses of the transcriptomes of a male and a female isolate of the moss, Ceratodon purpureus, generated using the 454 FLX technology. Comparative analyses between C. purpureus and P. patens indicated that this strategy generated nearly complete coverage of the protonemal transcriptome. An analysis of the overlap in gene sets between C. purpureus and P. patens provides new insights into the evolution of gene family composition across the land plants. In spite of the overall transcriptomic similarity between the two species, Ka/Ks analysis of P. patens and C. purpureus suggests considerable physiological and developmental divergence. Additionally, while the codon usage was very similar between these two mosses, C. purpureus genes showed a slightly greater codon usage bias than P. patens genes potentially because of the contrasting mating system of the two species. Finally, we found evidence of a genome doubling ~65–76 MYA that likely coincided with the contemporaneous polyploidy event inferred for P. patens but postdates the divergence of P. patens and C. purpureus. The powerful laboratory tools now available for C. purpureus will enable the research community to fully exploit these genomic resources.

Keywords: development and evolution, molecular evolution, plant mating systems, population genetics – empirical, systematics, transcriptomics

Received 2 March 2014; revision received 12 May 2014; accepted 21 May 2014

Introduction

Our understanding of the developmental processes that generate organismal diversity has been greatly enhanced by comparative studies across a phylogenetically diverse array of organisms. In plants, macroevolutionary studies in which genome composition and gene function in the moss model system Physcomitrella patens are compared with those in seed plants like Arabidopsis thaliana are now common. Some of the most revealing experiments have involved swapping genes between mosses and seed plants, giving us deep insights into the evolution of major regulators of development since the common ancestor of all land plants (Prigge & Bezanilla 2010). However, the bryophytes themselves are a morphologically and ecologically diverse group of plants. To facilitate molecular analyses of bryophyte diversity, we have developed genomic resources for another moss model system, Ceratodon purpureus.

Ceratodon purpureus is a member of the haplolepideae, a species-rich lineage of mosses named for the one row of peristome teeth that surround the opening of the sporangium. Recent phylogenetic analyses place C. purpureus in a lineage sister to the family Pottiaceae (Stech et al. 2010).
2 P. SZÖVÉNYI ET AL.

The common ancestor of *C. purpureus* and *P. patens* is therefore the common ancestor of more than 90% of all mosses. The genome of *C. purpureus* is smaller than that of *P. patens* (~320 Mbs vs. 511 Mbs; Voglmayr 2000; \( n = 13 \) vs. \( n = 27 \); Crum & Anderson 1981). One of the fundamental life history differences between these two model mosses is that *P. patens* is hermaphroditic, while *C. purpureus* has separate, chromosomally determined sexes (Heitz 1932; McDaniel et al. 2007, 2013a). Thus, *C. purpureus* is an obligate outbreeder, while *P. patens* commonly selfs (McDaniel & Shaw 2005; Perroud et al. 2011). Several lines of evidence suggest that these contrasting life histories may drive differences in genome architecture (Wright et al. 2013), most obviously the possession of heteromorphic sex chromosomes (Bachtrog et al. 2011), as well as patterns of diversification (McDaniel et al. 2013b).

What sets *C. purpureus* apart from other species of similar comparative evolutionary interest is its long history as a model system in genetics, physiology and development. Heitz (1928) coined the term heterochromatin based in part on his studies of chromosome behaviour in *C. purpureus*. More recently, *C. purpureus* has served as a model for studies of cell biology and physiology (Knight et al. 2002) and evolutionary genetics (McDaniel 2009). To date, a major impediment to developing the *C. purpureus* system further has been a lack of genomic information. However, recently, *C. purpureus* has been adopted for genome sequence determination (http://jgi.doe.gov/why-sequence-ceratodon-purpureus-moss/). As the first step towards this goal, we describe 454-FLX pyrosequencing of protonemal transcriptomes from a male and female genotype of *C. purpureus* and discuss this resource in the context of the evolution of land plant genomes. We find (i) many of the same genes as in *P. patens* but also some genes with no homologs in *P. patens*; (ii) highly conserved codon usage in *P. patens* and *C. purpureus* with a slightly higher codon bias in *C. purpureus*; and (iii) evidence of an ancient gene duplication coincident with but independent of the one discovered in *P. patens*.

Materials and methods

Plant tissue and treatments

A *Ceratodon purpureus* female strain GG1 (isolated from Gross Gerunds, Austria; Cove & Quatrano 2006) and male strain R40 (isolated from Rensselaer, New York, NY, USA; McDaniel et al. 2013a) were subcultured using standard moss culture methods (Cove et al. 2008). If not mentioned otherwise, cultures were performed at 25 °C with a 16 h light (70–80 µmol/m²/s)/8 h dark light cycle. To obtain uniform protonemal cultures, strains were homogenized and grown twice for 1 week on ‘BCD’ agar medium containing 1 mM CaCl₂ and 5 mM ammonium tartrate (‘BCDAT’), in 9 cm Petri dishes, overlaid with cellophane as described for *Physcomitrella patens* (Cove et al. 2008). To obtain as complete a protonemal transcriptome as possible, three treatment sets were initially performed with each strain. The control experiment consisted of 7-day-old protonemata harvested without any further treatment (sample SRX032862 for GG-1 and SRX032865 for R40). Multiple treatment experiments consisted of 6-day-old tissue transferred to individual plates containing, respectively, 1 µM 6-naphthalene acetic acid, 5 µM jasmonic acid, 5 µM salicylic acid and 1 µM 6-benzylaminopurine for 1, 3 and 24 h. Further tissue replicates were submitted to a cold treatment (4 °C) and a heat treatment (37 °C) for 1, 3 and 6 h, respectively. An equal quantity of tissue from each treatment was then combined for RNA extraction (samples SRX032863 for GG-1 and SRX032866 for R40). Finally, abscisic acid (ABA) treatments consisted of two ABA-treated tissue time-course experiments. First, 6-day-old tissue was transferred onto medium containing 50 µM ABA, and tissue was harvested after 30, 60 and 90 min, 3, 6, 12, 24 and 48 h. In parallel, 6-day-old tissue was transferred on fresh BCDAT containing 50 µM ABA for one hour and then transferred back to BCDAT medium for 30 min, 2, 5, 11, 23 and 47 h. Once again, an equal amount of tissue from each treatment was combined for RNA extraction (sample SRX032864 for GG-1 and SRX032867 for R40). Tissue harvesting was by transfer onto 3MM paper to drain excess water (‘squeeze-drying’) and subsequent flash freezing in liquid nitrogen.

Nucleic acid preparation

Total RNA was extracted from tissue homogenized in liquid nitrogen using the RNasy Plant Mini Kit (Qiagen GMBH, Germany) with the On-Column DNase digestion. Poly A+ RNA was isolated from total RNA using the Absolutely mRNA Purification kit and manufacturer’s instructions (Stratagene, La Jolla, CA, USA). Subsequently, the mRNA was then used to construct cDNA libraries using the cDNA Rapid Library Preparation Method as outlined in the Roche kit (Roche 454 Life Sciences, Branford, CT, USA), followed by sequencing on the 454 GS-FLX platform. The six unnormalized RNA samples (control, multiple treatment and ABA treatment sample for both the GG1 and R40 strains) were sequenced each on one full picotiter plate. As mentioned above, raw sequencing reads have been submitted to the NCBI short-read archive under the Accession nos: control sample (SRX032862 for GG-1 and SRX032865 for R40), multiple treatment sample (samples SRX032863 for
GG-1 and SRX032866 for R40), ABA treatment sample (SRX032864 for GG-1 and SRX032867 for R40).

Transcriptome assembly

Because males and females may express different genes and because the GG1 and R40 isolates come from geographically distant populations, we assembled and initially analysed the two transcriptomes separately. Before assembly, we quality filtered and trimmed the raw sff files. First, the reads were quality trimmed from their 3-prime end using a base quality threshold of Q20 using the fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html), and all reads shorter than 100 bp were discarded. Assembly of 454 reads was carried out using newbler 2.5.3 (454 Life Sciences) with default options. Adapter remnants were removed during the assembly process using the autotrimming option of newbler.

Assessment of assembly quality and completeness

We next calculated several descriptive statistics to evaluate the quality of the assembly, including the number of reads incorporated into contigs (i.e. the total number of reads used in the assembly), number of contigs assembled, mean and maximum length of contigs and isotigs generated, number of isogroups and the average number of isotigs per isogroup. The Newbler algorithm first assembles 454 reads into contigs. In a second step, contigs are assembled into isotigs with improved contiguity which are further clustered into isogroups (e.g. groups of highly similar isotigs). Therefore, isogroups represent putative genes, and isotigs within isogroups correspond to transcripts (including splice variants) produced by the gene. To describe the completeness of the assembly, we calculated the ortholog hit ratio (O’Neil & Emrich 2013) for each sequence using a blastx search against the P. patens proteome (version 1.6; Zimmer et al. 2013). We calculated the ortholog hit ratio by dividing the number of aligned bases (C. purpureus transcript vs. P. patens protein) with the total length (in nucleotide bases) of the P. patens protein sequence. We identified the putative P. patens orthologs by accepting the best-hit P. patens protein for each C. purpureus transcript sequence.

We also generated a rarefaction curve to see how addition of more reads influences the outcome of the assembly process. This was carried out by consecutively resampling an increasing number of reads from the final assembly and calculating the number of unique contigs assembled. Finally, to further assess the completeness of the assembly, transcripts were also blasted (blastx, e-value ≤ 1e-6) against the list of Arabidopsis thaliana proteins that are conserved as single copy genes across all eukaryotes (i.e. ultra-conserved orthologs, UCOS, http://compgenomics.ucdavis.edu/compositae_reference.php).

Coding potential and protein translation

To predict the coding frame and generate protein translations for the transcripts, we blasted all assembled contigs and isotigs against the plant division of the NCBI nonredundant database. We filtered the blast output keeping only isotigs/contigs with a blast hit showing at least 30% identity over at least 150 amino acids. The isotigs/contigs were then paired with their best-hit protein to find the open-reading frame (ORF) and generate the best protein alignment for each transcript using gene2ise2 (Birney et al. 2004).

Contamination analysis

To detect whether some transcripts were derived from contaminant organisms, we conducted a megan analysis (Huson et al. 2007). We initially blasted (blastx) all transcripts against the NCBI nonredundant database using an e-value threshold of 10^-6. We then retrieved the taxonomic affinity of all transcripts from NCBI taxonomy using megan. Parameters of assignment to the last common ancestor (LCA) were set to min support: 5; min score: 50; top percent: 10 and min complexity: 0.44. Finally, reads were summarized based on their taxonomy to detect putative contaminants and to analyse the taxonomic distribution of the transcripts.

Gene ontology (GO) and functional annotation

Functional annotation for each putative gene was obtained using homology searches followed by blast2go (Conesa et al. 2005) analysis. All isotigs/contigs were blasted against the NCBI nonredundant database using blastx with the following initial cutoff values: e-value threshold of 10^-3 and a high-scoring segment pair (HSP) length threshold of 33 amino acids. We kept a maximum of 20 hits per query sequence. GO annotations for the transcripts with hits were retrieved and assigned to each transcript using blast2go and filtered using the following cutoffs: e-value filter 10^-6, annotation cutoff 55, GO weight 5 and Hsp-Hit Coverage cutoff 0. In a similar way, enzyme codes were also mapped to transcripts. Finally, interproscan analysis (Jones et al. 2014) was also run on each transcript sequence, and results were joined with the original GO mapping using blast2go. When generating the GO graphs, only nodes that were supported with a minimum of five sequences were taken into account. We summarized GO annotations both at the contig/isotig and at the isogroup levels. We used isogroup level annotations in all further analyses.
After annotating the sequences, we used a simple GO-bias test in BLAST2GO (with a false discovery rate threshold of smaller than 0.05) to investigate whether the C. purpureus assemblies are enriched or depleted for particular GO categories relative to the P. patens genomic average (GO annotation file was retrieved from Zimmer et al. 2013). As genomewide GO distribution of the P. patens and C. purpureus genomes may differ significantly, we also investigated whether genes expressed in the 7-day-old chloronemata of P. patens are enriched (using a normalized threshold of at least one read per transcript Kb per million mapped reads, RPKM ≥ 1) for the very same terms as genes expressed in the 7-day-old C. purpureus control libraries. RNA-seq reads for P. patens were retrieved from the NCBI SRA (Accession no SRR072918) and expression estimated using default options of RSEM (Li & Dewey 2011). Gene expression for the C. purpureus genes in the control libraries was estimated using Sailfish (Patro et al. 2014) as described below. As functional annotations of the gene set expressed in the GG1 and R40 control libraries were similar, we only conducted this analysis using the GG1 gene set (RPKM ≥ 1).

Phylogenetic profiling

To study the evolution of gene content across land plants, we assigned assembled transcripts to the gene families of the PLAZA version 2.5 (Van Bel et al. 2012) database. We blasted (BLASTX) transcripts of the GG1 and R40 assemblies against all protein sequences to the PLAZA version 2.5 database with an e-value threshold of 10^{-6}. We assigned transcripts to the gene family of the best-hit PLAZA protein sequence. To investigate the putative function of the families to which C. purpureus transcripts were assigned, we investigated significantly enriched GO terms in each family. GO terms enriched per family were retrieved from the PLAZA version 2.5 website (ftp://ftp.psb.ugent.be/pub/plaza/plaza_public_2.5/GeneFamilies/genefamily_enrichment.hom.csv.gz). We only used GO terms with a P-value smaller than 10^{-8} (corresponding to a false discovery rate of ≤0.01). Finally, we took into account parent–child relationships, and only the most significant terms were used. We summarized the results by semantic clustering of the GO term with the REVIGO tool (Supek et al. 2011) using RELSIM semantic similarity measure and a dispensability threshold of 0.5.

Genetic divergence of proteomes: C. purpureus vs. P. patens

To find SNPs and estimate the genetic divergence ($K_a/K_s$) between C. purpureus and P. patens, we first established pairwise orthology between the P. patens proteome (version 1.6, Zimmer et al. 2013) and the predicted proteomes of the C. purpureus GG1 and R40 strains. For each isogroup, the isotig/contig with the longest protein translation was selected as the representative of the group. We then identified reciprocal best hits between P. patens and the R40 and GG1 protein sets, keeping only pairs showing a 30% identity over at least 150 amino acid alignment length in a BLASTP search (NCBI blast version 2.2.25+). Orthologous protein pairs were aligned using Muscle (Edgar 2004) with default parameters, and nucleotide alignments were forced into the protein alignments using PAL2NAL (Suyama et al. 2006). The ‘Remove gaps’ option of the PAL2NAL perl script was used to generate codon alignments without gaps. Finally, pairwise alignments were used to obtain estimates for $K_a$, $K_s$, and $K_a/K_s$ using PAML (Yang 1997). In the final analysis, unreliable alignments with a $K_a > 2$ and $K_a/K_s > 3$ or $K_a = 0$ were discarded. To investigate whether gene pairs with greater genetic divergence are enriched for specific functional groups, we conducted a GO analysis using the FUNC tool (Prüfer et al. 2007). More specifically, we ranked genes according to their $K_a/K_s$ values in a decreasing order and calculated the sum of ranks for the particular GO category. Finally, we used a nonparametric Wilcoxon rank sum test and investigated whether the sum of ranks for a given term differs from that of the root. We summarized the list of GO terms with the REVIGO tool (Supek et al. 2011) using the SIMREL semantic similarity measure and a medium dispensability threshold (0.5).

Synonymous codon usage

Synonymous codon usage was estimated using all coding sequences (CDS) of isotigs/contigs with protein translations with CODONW (J. Peden, http://codonw.sourceforge.net). In the first step, preferentially used codons were identified (Ikemura 1985) with the aid of a correspondence analysis on relative synonymous codon usage (RSCU) values (Sharp & Li 1987). To identify optimal codons, genes were projected to the first axis and the most extreme 5% of the gene set from both sides of the axis were selected. After that the effective number of codons was calculated for these two sets of genes and the set with the lower effective number of codons was defined as the more biased. RSCU was then calculated for the set with low and high bias classes for each synonymous codon. Major codons were defined by comparing synonymous codon frequencies in the two sets of genes using a chi-square test. We performed the very same analysis for P. patens (genome version 1.6, Zimmer et al. 2013) using all the predicted 38357 CDS. We calculated nonparametric correlation (Spearman’s rho) between codon usage bias ($F_{OP}$, frequency of optimal codons) and level of expression (RPKM) in the C. purpureus control
libraries to verify that codon bias is driven by translational selection.

After identifying preferred codons, we also wanted to test whether selection on synonymous codon usage is more efficient in the dioecious *C. purpureus* than in the monoicous *P. patens*. We calculated three measures describing codon usage bias of genes. We first estimated the effective number of codons (*N*<sub>e</sub>; Wright 1990). We also estimated proportion of optimal codons (F<sub>D</sub>; Ikemura 1985). Finally, as all these measures are known to be strongly influenced by DNA base composition, we also calculated the ΔRSCU+ statistic to compare codon usage bias between species (Cutter et al. 2006). The ΔRSCU+ statistic is largely independent of base composition and amino acid content and thus can be used to compare the extent of codon usage bias between species with different gene sets (Cutter et al. 2006). ΔRSCU+ is the average of all positive ΔRSCU values across all codons within species. ΔRSCU for a particular codon was calculated as the difference between the average RSCU of genes in the two most extreme gene sets along the first axis of the correspondence analysis (see above).

Analysis of whole-genome duplication

To identify possible large-scale duplication events, we employed the *keys* software as previously described (Rensing et al. 2007). For the clustering of paralogous genes, the minimal connectivity threshold used was 50% (half linkage); *K*<sub>s</sub> values at the nodes (representing duplication events rather than gene pairs) were calculated and used to plot the distribution. To identify possible hidden (contributing) distributions in the *K*<sub>s</sub> plot, we used the *EMMIX* software (McLachlan et al. 1999) for fitting a mixture model of normal distributed components to the gene clusters identified previously. The mixture model is able to yield the normal distributions that best describe the *K*<sub>s</sub> distributions. The mixed distributions were modelled with one to ten components, and the expectation maximization (EM) algorithm was repeated 100 times with random starting values and 10 times with *k*-means starting values. Additionally, the algorithm was set to finish when the proposed mixture model acquired an insignificant *P*-value. The best fitting model is defined as the one that has a significant *P*-value and for which the Bayesian information criterion is minimized.

Gene ontology bias analyses were conducted on the genes participating in the different components of the proposed mixture models, to investigate functionality bias. The GO-bias analysis was conducted using the *GOSTATS* (Falcon & Gentleman 2007) R package. The analysis was restricted to biological process terms, and the *P*-value threshold was set to 0.01. The resulting *P*-values were corrected using the Benjamini–Hochberg (Benjamini & Hochberg 1995) false discovery correction. The GO terms that were either over- or under-represented per component were visualized via word clouds, using the http://www.wordle.net application. The weights of the terms were defined as −10log (*q*-value).

Results and discussion

Assembly statistics

The six 454 runs resulted in approximately three million raw reads for the R40 and approximately eight million reads for the GG1 accession (Table 1, Fig. S1, Supporting information). For GG1, 99% of the raw data went to the assembly process after quality filtering, adaptor trimming and short-read removal. The assembly made use of the majority of the reads (92% of the reads were incorporated into contigs/isotigs). Assembly of quality filtered and trimmed reads produced 37 607 contigs which were further assembled into 16 362 isogroups (‘genes’). In total, 25 844 isotigs/contigs (471 singleton contigs and 25 373 isotigs) were assembled. For the R40 strain, 99% of the raw data remained after quality filtering, adaptor trimming and short-read removal. In this assembly, 91% were incorporated into contigs. Assembly of quality filtered and trimmed reads produced 26 270 contigs (Table 1 and Fig. S1, Supporting information) which were further classified into 15 506 isogroups (‘genes’). In total, 20 431 contigs/isotigs were assembled (154 singleton contigs and 20 277 isotigs).

Assembly completeness

We first assessed the ortholog hit ratio distribution of *C. purpureus* transcripts using the proteome of *Physcomitrella patens* as reference (Zimmer et al. 2013). Approximately 80% of the *C. purpureus* transcripts covered at least 80% of the length of the *P. patens* protein sequences (Fig. 1). This finding together with the fact that the distribution showed a strong peak at 1.0 implies that most of the assembled transcripts are close to full length. Although fewer reads were used for the R40 than for the GG1 assembly, distribution of the ortholog hit ratios did not differ considerably between strains (Fig. 1). Only a few isotigs had an ortholog hit ratio greater than one showing that coding sequence length is well conserved between *P. patens* and *C. purpureus*.

To determine whether the higher number of isotigs/contigs and isogroups in the GG1 assembly, relative to the R40 assembly, is an inherent property of the two transcriptomes or could be explained by the greater number of reads that were available for the GG1 assembly, we generated rarefaction curves to test the effect of sequencing depth (read number) on the number of
contigs assembled. Rarefaction curves for both the GG1 and R40 strains flatten out at a sequencing depth around two million reads (Fig. 2). Therefore, the rarefaction analysis suggests that the applied sequencing depth used [approximately three (R40) and seven million (GG1) filtered reads] is sufficient to re-cover most of the expressed transcripts. Notably, the number of unique transcripts assembled in R40 was always lower than in GG1, suggesting that R40 possesses a less complex transcriptome. Nevertheless, the R40 assembly contained 346 putative homologs of the 357 Arabidopsis thaliana single copy genes conserved across all eukaryotes (UCOS, see Materials and methods), while the GG1 assembly contained only six more (352), implying that both transcriptome assemblies are largely complete. Whether gene sets of the GG1 and R40 strains are different at the level of the genome or whether this difference is only present at the level of the transcriptome under the specific conditions investigated here can only be answered once the genome is sequenced.

**Contamination analysis**

Based on the phylogenetic composition of a MEGAN analysis, the majority of the isotigs/contigs, approximately 88% (22,653) in the GG1 and 89% (18,253) in the

![Table 1 Descriptive statistics of the GG1 and R40 assemblies](image)

<table>
<thead>
<tr>
<th>Statistics</th>
<th>GG1</th>
<th>R40</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of raw reads generated (per run)</td>
<td>7,998,658 (1,892,419; 2,537,066; 3,569,173)</td>
<td>3,135,690 (942,911; 940,761; 1,252,018)</td>
</tr>
<tr>
<td>No. of raw bases generated (per run)</td>
<td>2,941,737 (614,506,455; 902,918,024; 1,424,312,781)</td>
<td>1,086,205 (337,667,302; 279,972,926; 468,565,139)</td>
</tr>
<tr>
<td>No. of reads used for the assembly</td>
<td>7,990,022</td>
<td>3,133,185</td>
</tr>
<tr>
<td>No. of bases used for the assembly</td>
<td>2,922,412,550</td>
<td>1,078,602,382</td>
</tr>
<tr>
<td>No. of reads incorporated into contigs</td>
<td>7,370,476</td>
<td>2,853,310</td>
</tr>
<tr>
<td>No. of bases in assembled contigs</td>
<td>32,473,052</td>
<td>26,380,500</td>
</tr>
<tr>
<td>No. of transcripts assembled</td>
<td>37,607</td>
<td>26,270</td>
</tr>
<tr>
<td>Average length of transcripts</td>
<td>1,594</td>
<td>1,495</td>
</tr>
<tr>
<td>Length of the longest transcript</td>
<td>17,543</td>
<td>15,199</td>
</tr>
<tr>
<td>No. of transcripts &gt; 500 bp</td>
<td>18,369</td>
<td>16,400</td>
</tr>
<tr>
<td>No. of isotigs + singletons</td>
<td>25,373</td>
<td>20,277</td>
</tr>
<tr>
<td>No. of isogroups</td>
<td>16,372</td>
<td>15,506</td>
</tr>
<tr>
<td>Minimum number of isotigs per isogroup</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Maximum number of isotigs per isogroup</td>
<td>82</td>
<td>74</td>
</tr>
<tr>
<td>Average number of isotigs per isogroup</td>
<td>1.6</td>
<td>1.7</td>
</tr>
<tr>
<td>Average number of contigs per isotig</td>
<td>2.5</td>
<td>1.9</td>
</tr>
</tbody>
</table>

![Fig. 1 Histogram of ortholog hit ratios for the GG1 and R40 assemblies.](image)
R40 assembly, had a valid hit against the NCBI nonredundant database \((e\text{-value threshold } 10^{-3}; \text{ Fig. S2, Tables S1 and S2, Supporting information})\). A small proportion of the valid hits (101 in GG1, 45 in R40) could not be assigned to any taxonomical category. Of the GG1 and R40 isotigs, only 67 and 43 had a best hit against bacterial sequences, 131 and 67 against organisms characterized as Opisthokonta and 0 and 7 as Amoebozoa. All other GG1 and R40 transcripts hit eukaryotic sequences with 32 and 16 sequences assigned to chlorophyta and 21699 and 14170 to streptophytes. Therefore, the assembled transcriptome appears to be largely free of contaminants.

Gene ontology annotation

Of the 16362 GG1 isogroups, 15062 sequences (92%) could be annotated by one or more GO terms. In total, 97121 GO terms were assigned to the annotated sequences and the number of GO terms retrieved per sequence was on average 6.78 (min–max: 1, 66). Of the 15506 R40 isogroups, 9178 sequences could be annotated with one or more GO terms (59%). In total, 58302 GO terms were assigned to the isogroups and the number of GO terms retrieved per sequence was on average 6.53 (min–max: 1, 70). The mean annotation level (depth) for both transcriptomes was five across the three ontologies (Biological Function, Molecular Process and Cellular component; Tables S3 and S4, Supporting information).

Overall, the relative abundance of GO terms was similar in the GG1 and R40 strains. The ‘cellular’ and ‘metabolic process’ were the most abundant terms for the biological process ontology for both accessions, and genes annotated with ‘response to stimulus’ and ‘biological regulation’ were also abundant (Fig. S3, Supporting information). At level 2 of the molecular function ontology, ‘binding’ and ‘catalytic activity’ were the most abundant terms. At level 2 of the cellular component ontology the terms ‘cell’, ‘cell part’, ‘organelle’ and ‘macromolecular complex’ were dominant. We further asked whether \(C.\ purpureus\) assembled transcripts were enriched for particular functions, compared to the genomic average of the \(P.\ patens\) genome, and whether the same functional categories are enriched in chloronemata in both \(P.\ patens\) (based on RPKM > 1, in SRR072918) and \(C.\ purpureus\) assemblies (based on RPKM > 1 in \(C.\ purpureus\) control library, Fig. S4, Supporting information). At level two of the biological process ontology, we found that genes expressed in chloronemata of \(P.\ patens\) showed no particular enrichment compared to the genomic average. In contrast, isogroups of the \(C.\ purpureus\) assemblies showed significant enrichment for 17 GO terms. This suggests that the functional distribution of the genes in the present \(C.\ purpureus\) assemblies significantly differs from the gene set expressed in a comparable sample of the \(P.\ patens\) protonemata, potentially reflecting a difference in developmental timing and protonemal growth between the two species.

Phylogenetic profiling

To put the sequence data into a phylogenetic context, we assigned transcripts to the gene family classification of the \(\text{PLAZA}\) version 2.5 database (Van Bel et al. 2012). Approximately 85% (GG1) and 87% (R40) of the isotigs/isotigs could be assigned to gene families (\(\text{BLASTX e-value threshold } 10^{-6}; \text{ 6356 for GG1 and 6197 for R40}\); Table 2). Of the \(\text{Ceratodon purpureus}\) families, 97% (GG1) and 96% (R40) contained one or more \(P.\ patens\) members. The highly overlapping gene sets of \(C.\ purpureus\) and \(P.\ patens\) suggest that the common ancestor of \(C.\ purpureus\) and \(P.\ patens\) (the common ancestor of all of the arthrodontous mosses, >90% of all moss species) had a similar set of gene families.

Previous studies found that the \(P.\ patens\) genome contains a considerable proportion of lineage-specific genes (Zimmer et al. 2013). A considerable proportion of the \(C.\ purpureus\) families (10% in GG1 and 11% in R40) were moss specific (contained only \(P.\ patens\) sequences) indicating that a significant proportion of ‘\(P.\ patens\)-specific’ families (1046 in the \(\text{PLAZA}\) database) are likely to represent moss-specific genes. Based on a biological process GO term analysis, the strongly supported moss-specific families showed enrichment for biosynthetic processes in general, cell adhesion, ion transport and signalling in...
biological regulation. These families also are enriched for genes involved in DNA repair (Fig. S5, Supporting information).

The *C. purpureus* transcripts additionally extended the set of gene families that is shared with the green algae, with *Selaginella*, and with the angiosperms (Table 2). About 3% (both in GG1 and R40) of the assigned *C. purpureus* isotigs/contigs hit 324 (GG1) and 300 (R40) gene families that contained no *P. patens* sequences. For instance, there were 137 (GG1) and 124 (R40) families that were matched by *C. purpureus* sequences, lacked *P. patens* sequences but contained one or more chlorophyte sequences. These extend the list of families that are likely shared with green algae and were putatively present in the common ancestor of *Viridiplantae*. These families were functionally enriched for a highly interconnected network of various basic molecular processes. These genes were presumably lost or replaced in *P. patens* but remained active and functional in *C. purpureus* (Fig. S5, Supporting information).

Another 104 (GG1) and 88 (R40) *C. purpureus* families were shared exclusively with *Selaginella* sequences. Thus, these were not truly *Selaginella*-specific but were shared with bryophytes and were lost in or along the branch leading to *P. patens*. These families were only weakly enriched for specific biological process terms, and none of the GO terms had a *P*-value smaller than 0.0003.

Finally, *C. purpureus* sequences were also assigned to 97 (GG1) and 99 (R40) gene families that were previously believed to be angiosperm specific (Fig. S5, Supporting information). These families may have been present in the common ancestor of embryophytes, but were lost from *P. patens* and *Selaginella moellendorffii* or experienced some more complex history. These families were highly enriched for processes involved in DNA metabolism, repair and methylation, biosynthetic, metabolic processes and response to stimulus suggesting that *C. purpureus* may differ from *P. patens* in these respects. Thus, the *P. patens* gene set is substantially representative of the mosses, but the upcoming genome sequence of *C. purpureus* will significantly improve the information on the gene sets specific to mosses and shared with other clades of land plants.

**Genetic divergence of *P. patens* and *C. purpureus***

Consistent with the phylogenetic profiling analysis, the *C. purpureus* assemblies have extensive similarity with the *P. patens* proteome (BLASTX, *e*-value threshold 10^-30), with 22 288 of the 25 373 isotigs/contigs of the GG1 assembly and 18 257 of the 20 277 isotigs/contigs of the R40 assembly having significant matches. The average identity between the *C. purpureus* and *P. patens* proteomes peaks at around 80%, with some proteins...

---

**Table 2** Number and phylogenetic composition of *PLAZA* version 2.5 gene families *Ceratodon purpureus* transcripts could be assigned to

<table>
<thead>
<tr>
<th>Family Type</th>
<th>GG1 Number</th>
<th>Percentage Assigned</th>
<th>GG1 C. purpureus Transcripts</th>
<th>Percentage Assigned</th>
<th>R40 Number</th>
<th>Percentage Assigned</th>
<th>R40 C. purpureus Transcripts</th>
<th>Percentage Assigned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>25 844</td>
<td>100.00</td>
<td>20 431</td>
<td>100.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assigned</td>
<td>22 027</td>
<td>85.23</td>
<td>17 881</td>
<td>87.52</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>with C. purpureus sequences</td>
<td>6356</td>
<td>100.00</td>
<td>6197</td>
<td>100.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>and <em>C. purpureus</em> sequences</td>
<td>6027</td>
<td>94.82</td>
<td>5896</td>
<td>95.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of families without <em>P. patens</em> sequences</td>
<td>324</td>
<td>5.10</td>
<td>300</td>
<td>4.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of families with only trachaeophyte sequences</td>
<td>213</td>
<td>3.35</td>
<td>194</td>
<td>3.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of families with only <em>Selaginella</em> sequences</td>
<td>104</td>
<td>1.64</td>
<td>88</td>
<td>1.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of families with only angiosperm sequences</td>
<td>97</td>
<td>1.53</td>
<td>99</td>
<td>1.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of families with only <em>P. patens</em> sequences</td>
<td>915</td>
<td>14.40</td>
<td>892</td>
<td>14.39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of families with only chlorophyte sequences</td>
<td>137</td>
<td>2.16</td>
<td>124</td>
<td>2.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© 2014 John Wiley & Sons Ltd
showing 100% identity (Fig. S6, Supporting information).

To derive an estimate for the time since divergence between \textit{P. patens} and \textit{C. purpureus}, we estimated the number of synonymous substitutions per synonymous site for their pairwise orthologs. Synonymous sites are thought to evolve under relaxed evolutionary constraints in a manner proportional to the time since divergence (Hurst 2002). In the 7213 (GG1) and 7034 (R40) pairwise orthologs between the proteomes, the $K_s$ distribution peaked around 0.8–1.0 (Fig. S6, Supporting information). Assuming that the species diverged at the modal $K_s$ value, this corresponds to a divergence time of c. 85–106 myr ($9.4 \times 10^{-9}$ per synonymous site/year, Rensing et al. 2007).

The ratio of nonsynonymous changes per nonsynonymous sites ($K_a$) to the synonymous changes per synonymous site ($K_s$) estimates the number of genes that may have been affected by positive or negative selection. $K_a$/$K_s$ ratios between \textit{P. patens} and \textit{C. purpureus} (Fig. S6, Supporting information) were $<$1 (mean ratios 0.1 and 0.14 for GG1 and R40, respectively), showing that nonsynonymous mutations at nonsynonymous sites occur with a considerably lower rate than that expected under neutrality ($K_s$). This suggests that evolution of proteins is primarily driven by strong purifying selection. Our estimates are in the range of values recorded in other genomewide studies in plant species such as in \textit{A. thaliana} (Yang & Gaut 2011). Genes with an elevated $K_a$/$K_s$ value were enriched for biological functions including response to metal ions, secondary and pigment metabolic processes, protein break down mechanisms, intracellular transport, nucleotide biosynthesis, alcohol catabolism, glucose catabolism (Fig. S7, Supporting information). In short, genes with diverse basic biological processes show accelerated evolution since the split of \textit{P. patens} and \textit{C. purpureus} suggesting new discoveries to come in comparative analysis of the two species physiology/biology.

**Codon usage and GC content**

\textit{Ceratodon purpureus} and \textit{P. patens} have contrasting mating systems. The unisexual gametophytes of \textit{C. purpureus} bear either male or female gametangia but never both. By contrast, \textit{P. patens} gametophytes bear both male and female gametangia on one single shoot. Therefore, \textit{C. purpureus} can outcross or undergo intergametophytic selfing (fusion of gametes produced by gametophytes originating from one sporophyte) that is comparable to selfing in diploid plants which halves heterozygosity in every generation, while \textit{P. patens} can also undergo selfing involving the fusion of genetically identical male and female gametes produced by a single genetic individual (intragametophytic selfing). Evolutionary theory suggests that frequent inbreeding (especially intragametophytic selfing) leads to a reduced efficacy of selection in species with bisexual gametophytes (Wright et al. 2013).

One predicted consequence of reduced selection efficacy is less optimized synonymous codon usage in species with bisexual than with unisexual gametophytes. To evaluate the strength of selection of synonymous codon usage, we first identified preferred codons using correspondence analysis on RSCU values of genes. In the GG1 and R40 assemblies, the first two axes of the correspondence analysis on RSCU values explained in total 27.82% (19.79% and 8.03% GG1) and 27.43% (19.27% and 8.16% R40) of the synonymous codon usage variation in the data. Although the first two axes explain a relatively low proportion of the variation in the data, the frequency of optimal codons is significantly positively correlated with the expression level of genes (RPKM in the control library vs. \textit{F}_{OP}: Spearman’s rho GG1 = 0.0832, \textit{P} < 2.2e–16; Spearman’s rho R40 = 0.1540, \textit{P} < 2.2e–16). The list of preferred codons was the very similar between the GG1 and R40 assemblies. In total, we identified 20 preferred codons (Table S7, Supporting information). Among the inferred preferred codons in both GG1 and R40, all are C (15)- or G (5)- ending.

Finally, we obtained very similar results by analysing the full CDS set of \textit{P. patens}, where the first two axes of the correspondence analysis explained 10.39% and 4.74% of the total variation. Similar to \textit{C. purpureus}, all the inferred preferred codons are C (15)- or G (5)- ending. The preferred set of codons was very similar in \textit{C. purpureus} and \textit{P. patens}. We found only one codon that was preferred in \textit{C. purpureus} but nonpreferentially used in \textit{P. patens} (Table S7, Supporting information). This similarity extends to larger evolutionary timescales as well. Of the 20 preferred codons, a significant proportion was similar to the set of preferred codons described in \textit{A. thaliana} (Wright et al. 2004) and in \textit{Silene} (Qiu et al. 2011). Fifteen and 16 preferred codons were shared with those identified in \textit{A. thaliana} and \textit{Silene}, and only eight or four were different, respectively. This suggests that preferential synonymous codon usage is highly conserved. Whether differences in the remaining four preferred codons are biologically relevant or are just a result of sampling bias (only protonemal tissues were sampled) is not clear.

After identifying the set of preferred codons, we turned to analyse the difference in the codon bias (\textit{F}_{OP}: the frequency of optimal codons; \textit{N_c}: the effective number of codons) among \textit{P. patens} the GG1 and the R40 strains. \textit{F}_{OP} values were slightly higher, whereas \textit{N_c} values lower in the GG1 and R40 strains than in \textit{P. patens} (Fig. 3). The absolute difference between \textit{P. patens} and
C. purpureus \( F_{\text{OP}} \) and \( N_c \) values was small but turned out to be statistically highly significant. Post hoc test showed that \( F_{\text{OP}} \) values are significantly greater in both strains of C. purpureus than in P. patens. Furthermore, \( F_{\text{OP}} \) values of the GG1 transcripts were higher than that of the R40 strain. The effective number of codons was significantly lower in both strains of C. purpureus suggesting more effective selection on codon usage compared to P. patens. Furthermore, similar to the \( F_{\text{OP}} \) values, the GG1 strain had a significantly smaller effective number of codons than the R40 strain (Fig. 3). Altogether, both statistics (\( F_{\text{OP}} \) and \( N_c \)) indicated that codon bias is higher albeit weakly in C. purpureus than in P. patens. This finding is in line with evolutionary theory predicting greater efficacy of selection on codon usage in outcrossers than in selfers (Wright et al. 2013).

This difference between P. patens and C. purpureus could also be caused by biased gene conversion (Beletskii & Bhagwat 2001). The effect of neutral and selective processes can be dissected by comparing base composition of coding and noncoding genomic regions, such as introns, that should not be under selection for synonymous codon usage. As we have had no intron data, we used the GC content of third codon positions (GC3) as an indicator of compositional biases (Qiu et al. 2011). GC3 content covaried with codon bias as it was lower in P. patens with a weaker codon bias. Therefore, mutational biases can potentially explain the codon bias difference between P. patens and C. purpureus. However, this conclusion does not apply for the difference between GG1 and R40 as GC3 content, and codon bias was inversely related. More specifically, R40 had a lower GC3 content, but a greater codon bias than the GG1 strain. We also calculated the \( \Delta \text{RSCU}^+ \) index which is less sensitive to base and amino acid compositional biases (Cutter et al. 2006). \( \Delta \text{RSCU}^+ \) values were in the GG1 (mean: 1.0367), smallest in the R40 strain (mean: 0.7786) and intermediate in P. patens (mean: 0.9680). \( \Delta \text{RSCU}^+ \) values among the three data sets were, however, not significantly different either in a Kruskal–Wallis \( H_{(df = 2, N = 83)} = 3.3424, P = 0.1880 \) or in a parametric ANOVA \( F_{(df = 2, N = 83)} = 2.1491, P = 0.1233 \). Altogether, our data suggest that codon usage may be slightly more biased in C. purpureus than in P. patens as predicted by evolutionary theory. Nevertheless, whether this difference is driven by natural selection or rather by mutational biases remains to be seen. Similarly, existence of a slightly differential codon bias in the R40 and GG1 strains remains to be investigated, once their genomic sequence is available.

Identification of large-scale duplication events

To test for evidence of large-scale genome duplications, we aligned each transcriptome to itself. The distribution of the synonymous substitutions (\( K_s \)) for both transcriptome assemblies exhibited a similar pattern (Fig. 4). Six normal distributions potentially contribute to the history of duplication in the R40 strain assembly and five in the GG1 strain assembly (Tables S8 and S9, Supporting information, Fig. 4). These suggest that the lineage containing C. purpureus experienced its most prominent duplication approximately 65–76 MYA (\( K_s = 0.6–0.7 \)). This is nearly contemporaneous with the whole-genome duplication approximately 65–76 MYA (\( K_s = 0.6–0.7 \)).
duplication described for *P. patens* (Rensing et al. 2007), although both occurred well after the divergence of the *P. patens* and *C. purpureus* lineages from one another (Fig. S8, Supporting information). The over- or under-represented GO terms of the genes that compose the above mentioned normal distributions are visualized by word clouds (Fig. S9, Supporting information). Interestingly, group 2 (GG1) and 3 (R40) in both distributions (representing the biggest secondary peak akin to *P. patens*) are enriched for GO terms including response to metal ions/cadmium. Preferential retention of these genes may be related to the tolerance of a broad range of substrate chemistries in *C. purpureus*.

**Conclusions**

Here, we describe analyses of the transcriptomes of a male and a female isolate of the emerging model system *Ceratodon purpureus*, generated using the 454 FLX technology. Comparative analyses between *C. purpureus* and *Physcomitrella patens* both indicated that this strategy generated nearly complete coverage of the protonemal transcriptome. We were able to assign most of the proteins to several GO categories, providing a foundation for further functional analyses. Using a phylogenetic profiling approach, we showed that *C. purpureus* and *P. patens* had highly overlapping gene sets. These data further refine our understanding of the timing of gene family gain and loss across the land plants. Genetic divergence and $K_s/K_a$ analysis of *P. patens* and *C. purpureus* showed that most of the genes evolve under strong purifying selection, but genes involved in multiple basic biological processes show an accelerated rate of evolution. This suggests considerable physiological and developmental divergence between the two species. The patterns of silent site divergence between *C. purpureus* and *P. patens* demonstrate the utility of this database for identifying loci for phylogenetic studies. The gene composition and codon usage were very similar between these two mosses. Nevertheless, *C. purpureus* genes showed a slightly greater codon usage bias than *P. patens* genes which can be explained by the contrasting mating system of the two species. Finally, we found evidence of a genome doubling event ~65–76 MYA that was independent of the contemporaneous polyploidy event inferred for *P. patens*. The transcriptome of these two isolates now adds to the powerful laboratory tools available for *C. purpureus* extending the utility of this emerging model system.

**Acknowledgements**

We would like to thank Cindy Choi at JGI for help with library construction. The sequencing was conducted by the U.S. Department of Energy Joint Genome Institute which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. PSZ was supported by the SNSF (PZ00P3_131726). SFM was supported by start-up funds from the University of Florida. R.S.Q. and P.-F.P. were supported by funds from Washington University in St. Louis.

**References**


**Fig. S4** Gene ontology terms (associated with a minimum of 150 isogroups) significantly enriched (FDR < 0.05, Fisher exact test) in the gene sets expressed in the *Physcomitrella patens* chloronema and in the *Ceratodon purpureus* GG1 control protonema tissues relative to the *P. patens* genomic gene set.

**Fig. S5** Gene ontology terms (Biological process ontology) enriched in the PLAZA version 2.5 gene families which *Ceratodon purpureus* transcripts was assigned to.

**Fig. S6** Genetic divergence of the *Ceratodon purpureus* and *Physcomitrella patens* proteomes at the protein (proportion of identical amino acid residues in alignment) and at the nucleotide level (*K_s*: number of synonymous substitutions per synonymous sites, *K_a*: number of nonsynonymous substitutions per nonsynonymous sites).

**Fig. S7** Gene ontology terms (Biological Process ontology) of which genes show a significantly greater evolutionary rate (*K_a*/ *K_s* ratio) between the *Ceratodon purpureus* strains and *Physcomitrella patens* than the average of the total ortholog list.

**Fig. S8** Histogram of *K_s* values of *Physcomitrella patens* (red) and *Ceratodon purpureus* GG1 (green) paralog clusters and *P. patens* and *C. purpureus* GG1 orthologs (grey, reciprocal best hits).

**Fig. S9** Gene ontology bias results on the genes belonging to the groups identified by EMMIX for the (K4) GG1 strain and (K3) R40 strain.

**Table S1** Top blast hit of each GG1 transcript in the NCBI nr database using BLASTX.

**Table S2** Top blast hit of each R40 transcript in the NCBI nr database using BLASTX.

**Table S3** Gene ontology annotation of GG1 transcripts.

**Table S4** Gene ontology annotation of R40 transcripts.

**Table S5** Interproscan analysis of GG1 transcripts.

**Table S6** Interproscan analysis of R40 transcripts.

**Table S7** Codon preferences in *Ceratodon purpureus* GG1, R40 and *Physcomitrella patens*.

**Table S8** Results of the EMMIX analysis for the *K_s* distribution of the GG1 and R40 assemblies.

**Table S9** Inferred age (*K_s*) of the putative large-scale duplication events for the GG1 and R40 assemblies.

© 2014 John Wiley & Sons Ltd