

The transcriptome of the moss *Physcomitrella patens*: comparative analysis reveals a rich source of new genes

Stefan A. Rensing^{1*}, Stephane Rombauts³, Annette Hohe¹, Daniel Lang¹, Elke Duwenig², Pierre Rouze⁴, Yves Van de Peer³ and Ralf Reski¹

* corresponding author, e-mail: stefan.rensing@biologie.uni-freiburg.de

¹ University of Freiburg, Plant Biotechnology, Sonnenstr. 5, D-79104 Freiburg, Germany

² BASF Plant Science GmbH, Robert-Bosch-Str. 38, D-67056 Ludwigshafen, Germany

³ Department of Molecular and Plant genetics, Vlaams Interuniversitair Instituut voor Biotechnologie (VIB), University of Ghent, K.L. Ledeganckstraat 35, B-9000 Ghent, Belgium

⁴ Laboratoire Associé a l'Institut National de la Recherche Agronomique (France), University of Ghent, B-9000 Ghent, Belgium

This paper was made available on our webpage www.plant-biotech.net because due to proprietary rights (the EST sequencing project is part of an industrial cooperation with BASF Plant Science GmbH, Germany) it was not possible to find a journal willing to publish it. As we think that the information contained is valuable to the scientific community, we wanted to make the manuscript available nevertheless. We encourage scientific collaboration, please contact R.R. for details. © **plant biotechnology 2002**

Citation: Rensing SA, Rombauts S, Hohe A, Lang D, Duwenig E, Rouze P, Van de Peer Y and Reski R (2002) The transcriptome of the moss *Physcomitrella patens*: comparative analysis reveals a rich source of new genes. http://www.plant-biotech.net/Rensing_et_al_transcriptome2002.pdf

Abstract

Three cDNA collections covering all important steps of the life cycle of the moss *Physcomitrella patens* have been used to prepare three normalized and subtracted cDNA libraries for the purpose of mass sequencing. After production of around 110,000 expressed sequence tags a clustered database was built and analysed. The *Physcomitrella* transcriptome is estimated to contain around 25,000 genes. Whereas those cover around 50% of the *Arabidopsis* gene content (around 26,000 genes), *Physcomitrella* possesses more than 11,000 expressed protein coding genes that cannot be identified by terms of sequence homology to date.

Keywords: *Physcomitrella*, moss, EST, database, transcriptome

Abbreviations: ABA = abscisic acid; ss-cDNA = single-stranded copy DNA; EST = expressed sequence tag; HMM = hidden Markov model; indels = insertions/deletions; ORF = open reading frame; PCR = polymerase chain reaction ; UTR = untranslated region; 2iP = N6-[2-isopentenyl]adenine

date of publication on webpage: 25.06.2002

Introduction

In recent years, many large scale sequencing projects, both of genomes and transcriptomes (i.e., cDNA sequencing yielding ESTs) have been initiated from a broad palette of organisms. The resulting amount of data allows us to better understand the evolution of organisms by applying comparative genomics as well as to unravel the function of previously unknown genes by functional genomics. It is in the latter respect that the moss *Physcomitrella patens* is of special interest because of its high rate of homologous recombination that can be used to produce gene knockouts with relative ease, thus enabling identification of gene function (e.g. Strepp et al., 1998; Reski, 1999; Schaefer 2001). Furthermore, mosses are thought to be comparable to higher plants in terms of gene content, expression and regulation (e.g., Reski, 1998; Reski et al., 1998). In this respect, independent from our own research, an international EST sequencing initiative (<http://www.moss.leeds.ac.uk/startpage.html>) also aims at identifying expressed protein genes of the moss.

Apart from acquiring knowledge about gene function, studying the moss genome is important to learn more about plant and plant genome evolution as well as about the genetics underlying the transition from “primitive” to “higher” plants. As a bryophyte, *Physcomitrella* belongs to the division Embryophyta together with ferns and seed plants, but is estimated to have diverged from those about 450 million years ago (e.g., Theissen et al., 2001). By comparing the genomes of mosses and seed plants, we should be able to reconstruct the genetic toolkit shared by primitive and higher plants and to observe which gene families have been considerably expanded or on the contrary have been reduced, in both lineages of plants.

The production of the EST database described here is an important tool for the evaluation of these assumptions. Despite both being embryophytes, compared to the seed plants there are clear differences that make the moss especially interesting both for fundamental and applied research: the gametophyte is dominating the life-cycle, unlike the sporophyte in higher plants, and the genome is haploid, thus making it easier to connect genes with aberrant functions/phenotypes by loss-of-function mutations. Based on small scale EST sequencing projects it has been shown before (e.g. Reski et al. 1998, Machuka et al., 1999) that the moss genome is a valuable source for unknown genes.

In order to represent the whole transcriptome of the moss, we made use of material collected from all stages of the life cycle for the production of RNA (a visual representation of the life cycle can be found on our homepage: <http://www.plant-biotech.net/pics/lifecycle.jpg>). To reduce redundancy prior to mass sequencing, normalization and subtraction was carried out. Three cDNA libraries covering the complete life cycle of *Physcomitrella* went into sequencing, quality clipping and clustering, yielding a database of around 33,000 clusters. In this study, we describe the analysis of this database and the comparison with the *Arabidopsis*, as well as, to a lesser extent, the rice genome. As *Arabidopsis* is the only plant genome that has been completely sequenced to date it was used as the standard for comparison.

Methods

Plant material

In vitro cultures of *Physcomitrella patens* (Hedw.) B.S.G. (strain as in Reski et al., 1994) were used for isolation of RNA (see below). These cultures comprised chloronema and caulonema as well as gametophores with and without gametangia, sporophytes, germinating spores and regenerating protoplasts, thus covering the complete life cycle of *Physcomitrella* (Reski, 1998). Additionally, part of the plant material was subjected to substances known to have physiological/developmental effects in order to induce the expression of the responsible genes (e.g. cytokinin, see Reski and Abel 1985). The material was subdivided into three groups (protonema, gametophores, sporophytes/regenerating protoplasts). The culture conditions are described below, the amount of material (fresh weight) used is given in brackets, respectively.

Protonema material

Protonema, i.e. chloronema and caulonema, was grown in liquid culture using modified Knop medium containing 1000 mg/l $\text{Ca}(\text{NO}_3)_2 \times 4 \text{H}_2\text{O}$, 250 mg/l KCl, 250 mg/l KH_2PO_4 , 250 mg/l $\text{MgSO}_4 \times 7 \text{H}_2\text{O}$ and 12,5 mg/l $\text{FeSO}_4 \times 7 \text{H}_2\text{O}$, adjusted to pH 5.8 with KOH/HCl before autoclaving, as described by Reski and Abel (1985). All cultures were kept at 25° C under a 16/8h light/dark regime at an intensity of 50-70 $\mu\text{mol m}^{-2} \text{s}^{-1}$.

Material grown in Erlenmeyer flasks was harvested 3 days (0.3 g), 1 week (0.3 g) and 2 weeks (0.4 g) after the last subculture. In addition, protonema cultures were treated with 5 μM ABA for 30 min (0.5 g) and 24 hours (0.5 g) and 5 μM 2iP (0.4 g) for 30 min. Protonema (1.75 g) was also harvested from a bioreactor batch culture as described by Hohe and Reski (2002).

Gametophore material

From liquid cultures in Erlenmeyer flasks, young gametophores were obtained from 9 week old suspension cultures (3.5 g). Furthermore, suspension cultures were treated with 5 μM 2iP for 30 min (0.4 g). Gametophores grown on solid medium in Petri dishes were harvested 4-10 weeks after the last subculture (14.8 g). Here, either the medium as described above was used or a medium additionally supplemented with 30 mg/l Fetrilon (Compo, Muenster, Germany) and 200 mg/l Glucose or a full medium based on Knop medium as described by Schween et al. (2001). Culture temperature and light intensity were as described for protonema cultures.

Sporophyte material

Gametophores on modified Knop medium as described above (protonema material) or on medium additionally supplemented with 30 mg/l Fetrilon and 200 mg/l Glucose were grown at

15° C and a light intensity of 20 $\mu\text{mol m}^{-2} \text{s}^{-1}$ to induce gametangia and sporophyte development. Gametophore tips were harvested 2 days after transfer to 15° C (0.04 g), after development of gametangia (0.39 g) and after development of sporophytes (0.98 g). In addition, germinating spores and regenerating protoplasts (0.1 – 10 days after protoplast isolation, according to Hohe and Reski 2002) were added to this material (amount not weighable).

Production of the cDNA libraries

From the three different tissue types described above, RNA was prepared individually and went into first-strand synthesis and normalization afterwards. The protonema cDNA was subtracted from the gametophore and sporophyte library in order to further reduce redundancy. The cDNA libraries were produced by vertis Biotechnologie, Germany.

The protonema library

From the protonema material, total RNA was isolated using a phenol-based method (Pasentsis et al. 1998). Poly(A)⁺ RNA was purified by using *Oligotex*TM-dT(30) (Qiagen, Germany). The cDNA was synthesised from 0.14 μg of poly(A)⁺ RNA. During the cDNA synthesis procedure, oligonucleotide primers were attached to the 5'- and 3'-ends of the cDNA to allow PCR-amplification of the cDNA as well directional cloning of the cDNA into *Not* I/*Asc* I-sites of specific plasmid vectors. All PCR amplification steps were performed with a long and accurate PCR system as described by Barnes (1994).

Normalization of the cDNA was performed according to Ko (1990) with several modifications. The cDNA used for normalisation was not sheared but full length. One μg of amplified cDNA was suspended in 10 μl 0,5 M sodium phosphate (pH 6,8), denatured at 98 °C for 3 min and reassociated at 65 °C for 24 h. The normalized single-stranded (ss) cDNA fraction was separated from the double-stranded (ds) cDNA by column chromatography at 60 °C on hydroxylapatite (Ausubel et al., 1987) and PCR amplified.

The sporophyte and gametophore libraries

Total RNA, used for the synthesis of tester cDNA, was isolated from gametophore, sporophyte and protoplast material with the method used to isolate RNA from protonema (see above). The subtracted gametophore and sporophyte cDNA libraries were prepared as follows: The driver and tester cDNAs were synthesized and amplified according to the method used to prepare the cDNA from protonemata (see above). In the case of the tester cDNAs, the oligo nucleotide primers were modified to allow specific amplification of tester cDNA in the presence of the driver cDNA. The driver cDNA was synthesized from 1 μg of the total RNA from protonemata, which was used for the preparation of the normalized protonema library. The tester cDNAs were synthesized from 1 μg each of gametophore, sporophyte and protoplast RNA. For the subtraction reactions, the driver cDNA was converted into sense ss-cDNA, the tester cDNAs were converted into antisense ss-cDNA. The subtraction reactions were performed with the driver and tester cDNAs in full length.

For the preparation of the gametophore subtracted cDNA, 1 µg of protonemata sense ss-cDNA (driver) was mixed with 100 ng gametophore antisense ss-cDNA (tester1), for the preparation of the sporophyte subtracted cDNA, 900 ng of driver sense ss-cDNA was mixed with 67,5 ng sporophyte and 22,5 ng protoplast antisense ss-cDNA (tester2 mix). Each of the driver/tester mixtures were suspended in 5 µl reassociation buffer (1 M sodium phosphate buffer pH 6,8; 50 %, v:v, formamide) and denatured at 98 °C for 3 min. The reassociation reactions were performed at 42 °C for 15 h. The subtracted ss-cDNA fractions were separated from the ds-cDNA by column chromatography at 60 °C on hydroxylapatite (Ausubel et al., 1987) and PCR amplified.

With the first subtracted cDNAs (S1-cDNAs), a second subtraction circle was performed. For this purpose, antisense ss-cDNA was prepared from the S1-cDNAs. The S1 antisense ss-cDNAs (100 ng) were mixed with 1 µg driver sense ss-cDNA and suspended in 5 µl reassociation buffer. Denaturation, reassociation, separation of the ss-cDNA fractions and PCR amplification of the S2-cDNAs was performed as described for the S1-subtraction.

Mass sequencing

After directional cloning into a sequencing vector, clones were sequenced at BASF AG using a primer allowing synthesis from the 3' end, utilizing automated sample preparation (Qiagen, Germany) and laser fluorescence sequencing (Prism 377, ABI, Germany). Samples that did not yield a sequencing reaction were processed using a primer directed from the 5' end, leading to a total of around 3% 5' sequencing reactions. Whereas around 80% of the sequences are derived from the three libraries chosen for mass sequencing, 20% of the sequences result from either test sequencing of libraries or to a large part from a protonema library that had neither been normalized nor cloned in directed fashion (Reski et al., 1994).

Clustering

A total of 110,087 sequences (table 1) were subjected to clustering in order to remove redundant sequences, which was performed with HarvESTerTM (Biomax, Germany) using default parameters. The process of clustering included poly-A removal, quality clipping to remove low quality regions from the sequencing reactions, vector clipping to remove stretches of the cloning vector as well as repeat removal. The average size of raw data sequences is 605 bp, whereas the average size of the clusters is 707bp. We use the term „cluster“ to describe all sequences that are the result of the EST clustering process, i.e. comprising „singletons“ (build from a single sequence) as well as „contigs“ (build from at least two sequences). Each cluster is named after the longest individual EST sequence it contains. The name of each cluster also contains the number code of the library from which the name-giving EST clone is derived. This nomenclature enables us to estimate from which libraries clusters are derived – for contigs, however, participation of clones from different libraries is masked (see table 1).

Databases

Both the EST raw data and the clusters went into DNA databases, respectively. The clustered EST database, designated „final“, was also translated into all six peptide reading frames in order to produce a database that contains all possible peptide translations. The ESTs contain a large quantity of 3' UTRs because they are derived from poly(A)⁺ RNA and have mainly been sequenced from the 3' end. In addition, ESTs usually contain cloning artifacts, contaminations of genomic DNA, ambiguous bases and accidental indels yielding frameshifts and stop-codons. To extract the actual open reading frames from the clusters, we used the HMM-based ESTScan algorithm (Iseli et al. 1999, available at www.ch.embnet.org), producing a second peptide database comprising only of predicted translated ORFs.

Analysis

Software utilized to analyse the clustered EST database included BLAST 2 (Altschul et al., 1997), and the Wisconsin GCG package (Accelrys, USA) including the public sequence databases (rel. 01/3). The GENPEPT database (release 124.0), being a conceptual translation of GENBANK, was used as a peptide database covering known protein coding genes from all organisms. The rice unique unigenes (Os.uniq.seq.gz as of 10/01; 12,836 sequences), available from <ftp://ftp.ncbi.nlm.nih.gov/repository/UniGene>, have been used to build a BLAST database. In the case of Arabidopsis, a prediction database produced from the genomic sequence according to Pavy et al. (1999) was used. Analysis of textual output was carried out using scripts utilizing awk, sed and perl. Analysis of protein signatures was carried out using a local instance of InterProScan, release 6/01 (Apweiler et al. 2001, available at www.ebi.ac.uk) as well as GCG. Protein family/motif databases used and their releases were: PRINTS (rel. 31, www.bioinf.man.ac.uk), PROSITE (rel. 16.37, www.expasy.ch) and PFAM (rel. 6.2, www.sanger.ac.uk). All analyses were performed on a SUN Ultra 60 Creator two-processor machine with 1152MB RAM (SUN Microsystems, Germany). As an example, CPU time for comparing the 26,352 predicted Arabidopsis genes to the 201,486 Physcomitrella peptide frames using BLASTP was in the range of 90 hours.

Results and Discussion

Quality of the cDNA libraries

By using normalized cDNA, redundancy of the EST clones was largely reduced. From a total of 33,581 clusters in the database, 21,219 (equaling 62.23%) are singletons. This high rate of singletons reflects the quality of the normalized cDNA that went into the mass-sequencing. Subtraction of the protonema library from the two other libraries (gametophore and sporophyte) was another measure to reduce redundancy. Whereas the overall percentage of clusters from each library is about equal (table 1), the amount of singletons from the subtracted libraries is almost twice as high as from the protonema library, proving the

effectivity of the subtraction procedure. This can also be seen from the rate of cluster:clone (see table 1), which is much better for the two subtracted libraries.

Length distribution

Both the ESTs and the clusters exhibit a peak distribution of the number of sequences plotted against their length. The ESTs (average length 605bp) spike at around 630bp whereas the clusters (average length 707bp) have their peak around 720bp (fig. 1a). The singletons contribute to the left flank as well as most of the peak region to the cluster distribution plot, the contigs are responsible for only a part of the peak region and the right flank (fig. 1a). All these plots completely fit the expectation. The length distribution of the predicted Arabidopsis genes shows a double-peaked (550 and 1000bp) profile with a slowly decreasing right slope (fig. 1a). The chosen categories for gene length used for determination of the gene content (see below) nicely fit the latter distribution plot. The average length of the predicted ORFs is 579bp, with the length distribution having its peak at 520bp (fig. 1a). This would mean an average length for the 3' UTR of only around 130-200bp, i.e. less than 30% of the typical contigs length.

Gene content of the EST database

As an educated guess based on the prediction for Arabidopsis at that time (23,000 genes), we estimated the number of genes in *Physcomitrella* to be around 20,000 beforehand. By reducing the redundancy of the transcriptome with the above mentioned methods and sequencing more than five times as many clones as the expected number of genes, we wanted to make sure that a high percentage of the genes (ideally >95%) are represented in the EST database. This assumption was tested in comparative searches between *Physcomitrella*, Arabidopsis and rice (see below).

By using a pool of 32 single copy genes of different length, we estimated the coverage of differently sized genes as well as the rate of clusters to genes. The pool of genes was subdivided into three groups: 11 „small“ genes (< 600bp with an average length of 385bp), 11 „middle-sized genes“ (600-1500bp, average 1119bp) and 10 „large“ genes (> 1500bp, average 3156bp). Those 32 genes, selected to cover a wide range of diversity, were either known genes from *Physcomitrella* previously deposited in the database or homologues from Arabidopsis and other organisms (table 2 a). This sample of genes went into BLAST homology searches against all possible peptide reading frames of the clustered EST database using an E-value cutoff of 1×10^{-3} to distinguish significant hits. With this pool of 32 genes, a total of 42 non-overlapping clusters were detected in the database, leading to a rate of 0.76 genes per cluster. While allowing a 10% error margin this would lead to an estimate of 25,500 \pm 2,500 expressed protein coding genes in *Physcomitrella patens*. This is a result quite similar to the estimates for the *Arabidopsis thaliana* gene content (The Arabidopsis Genome Initiative, 2000) and remarkably close to the number of genes in our Arabidopsis prediction (26,352, see below).

To estimate to what extent the 32 genes were covered by the 42 clusters, four classes were defined. 15 of the 42 clusters covered up to 25% of the respective gene length, 12 clusters

covered the range of 25-50%, 5 clusters 50-75% and 10 clusters 75-100%. On average, small genes were covered to 86%, middle-sized genes to 75% and large genes to an extent of 55% of their length (table 2 a). The clusters covered the N-terminal region of the genes in 33% cases, whereas the C-terminal region (43%) and full-length sequences (24%) made up the remaining 67%. From undirected cloning of 20% of the cDNAs and from the N-terminal sequencing of another 3%, around 13% of N-terminal sequences (i.e., 87% C-terminal/full-length) is to be expected. The differing 20% has to be accounted for by both fragmentary cDNAs and inverted clones.

Gene families

In addition to the search with single copy genes, that did not find overlapping contigs in a homology search, we also carried out a search with members of gene families. Using a total of 27 genes that did detect overlapping clusters, i.e. gene family homologues, we tried to determine the extent to which gene families are present in the moss. Again, the 27 genes were mainly from *Physcomitrella* and *Arabidopsis* but contained gene family members from other organisms as well (table 2 b).

Using BLASTP and an E-value cutoff of 1×10^{-3} , the 27 searches detected a total of 427 hits, equating 16 hits per gene. The range of hits for individual search genes was between 2 and 40. Therefore *Physcomitrella patens* seems to possess gene families that are similar to those in higher plants.

The representation of the clusters from the different libraries reflected the amount of raw EST data that went into the clustering rather than the amount of clusters in the libraries (table 1, 2 b). The reason that the protonema library has been subtracted from the gametophore and sporophyte libraries was, that the protonema state is suspected to show expression of most housekeeping genes. This assumption is reflected by our results.

The 18,023 predicted *Arabidopsis* genes that found a homologue among the *Physcomitrella* clusters (see below) can be divided into 3,166 unique and 14,857 multiple hits in the *Physcomitrella* database. The multiple hits mask a total of 2,895 *Physcomitrella* clusters, i.e. on average, 5.13 *Arabidopsis* genes hit a *Physcomitrella* gene. In total, 6,061 *Physcomitrella* clusters were hit by an predicted *Arabidopsis* protein gene, which corresponds to 23% of the 26,352 predicted genes (table 3). In the reverse search, the 15,749 *Physcomitrella* clusters that found a match against the predicted *Arabidopsis* genes can be divided into 3,894 unique and 11,855 multiple hits. The multiple hits mask a total of 3,879 *Arabidopsis* genes, i.e. on average, 3.06 *Physcomitrella* clusters hit an *Arabidopsis* gene. In comparison, the average gene family in *Arabidopsis* seems to be nearly twice as big as in *Physcomitrella*. This means that the moss has a high number of unique genes, that might for example be necessary for secondary metabolism and resistance against pathogens. In total, 7,773 of 26,352 predicted *Arabidopsis* genes were hit by *Physcomitrella* clusters, which again corresponds to 23% of the 33,581 *Physcomitrella* clusters (table 3). According to the *Arabidopsis* Genome Initiative (2000), the *Arabidopsis* transcriptome consists of 11,601 singletons and distinct gene families. Hence the 6,061 of 33,581 *Physcomitrella* clusters that are hit by the *Arabidopsis* genes would mean a coverage of 52% of the transcriptome.

Comparison with Arabidopsis

We were using a prediction database produced from the genomic sequence according to Pavy et al. (1999) that contains 26,352 predicted Arabidopsis genes. Using BLASTP, peptide translations of all these genes were run against all six peptide reading frames of the Physcomitrella database with different E-value cutoffs. Using an E-value cutoff of 1×10^{-3} , 66.4% of the predicted Arabidopsis genes found a homologue in Physcomitrella. With a cutoff of 1×10^{-2} we detected 68.4% hits and using an E-value of 1×10^{-1} led to 71.2% hits against the moss. The plotted distribution of number of hits against E-value intervals, carried out for the 1×10^{-2} threshold search, also shows that only a relatively small portion of the hits are due to the 10^{-2} and 10^{-3} categories, whereas E-values starting from 10^{-4} and lower contribute to the pattern in a steadily decreasing curve (fig. 1b). The same is observed in plots with Physcomitrella vs. Arabidopsis and Physcomitrella vs. GENPEPT (see below). Given that the estimates for the number of genes of the two organisms are quite similar, this is a surprising result. Arabidopsis appeared to share around 70% of the Physcomitrella genes, as can be detected by means of this homology search that does count redundant hits.

In order to figure out whether this result might be due to a large part of Physcomitrella genes not being represented in the Physcomitrella EST database, we ran a reverse search with all moss contigs against the predicted Arabidopsis peptide sequences, using BLASTX and an E-value threshold of 1×10^{-2} . If the representation of the Physcomitrella transcriptome in the EST database would be significantly less than the expected >95%, a higher percentage value than 70% hits would be expected for this reverse search. We found that the percentage of hits relative to the predicted number of genes in each Arabidopsis chromosome varied only slightly between 57.2% (chromosome 2) and 61.5% (chromosome 1). As it turns out, however, only 46.9% of the Physcomitrella contigs find a representative in the Arabidopsis genes. Surprisingly, this is much less than the ~70% found in the Arabidopsis versus Physcomitrella search.

To be sure that the different values found in the two types of searches are not due to the difference in BLAST method and database size, we ran a search of the Arabidopsis peptide sequences against the Physcomitrella nucleotide sequences utilizing TBLASTN. This search (E-value threshold 1×10^{-2}) resulted in 68.9% hits, i.e. very close to the result of the BLASTP search mentioned above (68.4%). We conclude that the different database sizes and BLAST methods do not influence the search result to a critical extent.

In order to check for possible influence of the query sequence length on the search result we plotted the number of hits and no-hits against the length of the queries (fig. 1b). As it turned out, both hit and no-hit distribution show a similar pattern with a peak at roughly 700bp. This is true not only for the searches Physcomitrella vs. Arabidopsis but also for the reverse search (albeit with a slightly more complex right flank than the other plots) and the search Physcomitrella vs. GENPEPT (data not shown). Therefore, the query sequence length does not strongly influence the outcome of the search in these cases.

There is a bias, however, as to how much the contigs and singletons contribute to the hits and no-hits, respectively. As an example, in the search Physcomitrella vs. Arabidopsis (E-value cutoff 1×10^{-2} , BLASTX), 35.2% of the contigs yielded no hit, whereas 63.6% of the singletons gave no significant hits. This led to the suspicion that the lower ORF/UTR ratio

expected for the singletons maybe the reason for these results. In order to test this hypothesis, we used the predicted *Physcomitrella* translated ORF database (see methods) of 24,918 sequences to search against the *Arabidopsis* translated predicted genes (BLASTP, E-value threshold 1×10^{-2}). The search resulted in 47.4% hits, i.e. very close to the 46.9% of the initial search (*Physcomitrella* vs. *Arabidopsis*). The contig/singleton bias was also comparable with 40.6% of the contigs and 61.8% of the singletons resulting in no hits. Taken together, the results strongly suggest that only around half of the expressed *Physcomitrella* protein genes have a counterpart detectable by means of a homology search in *Arabidopsis*. It is noteworthy that the number of predicted ORFs, 24,918, is pretty close to the estimate of expressed protein coding genes based on the rate genes per contig (see above, $33,581 \times 0.76 = 25,522$).

Interestingly, the amount of singletons that yield no hit is much higher for the gametophore (69%) and sporophyte (64%) library derived sequences than for those from protonema (50%). Although this is in part due to the subtraction protocol, it also seems to reflect the fact that indeed the protonema expresses most of the house keeping genes whereas in the gametophore and the sporophyte a lot of other - currently unknown - genes seem to be expressed.

Genome size and codon usage

The C-value (haploid genome size) of *Physcomitrella* is 0.53pg, equaling approximately 511Mbp (Schween et al. 2001). As most higher plants go, this is a rather low number (e.g. tomato ~1pg, pea ~4pg, wheat ~15pg, maize ~22pg; Marie and Brown 1993). The genome of *Arabidopsis*, however, has a size of ~120Mbp and thus possesses only a quarter of the *Physcomitrella* genomic DNA in terms of haploid genome sets. Whereas *Arabidopsis* bears 5 chromosomes with an average (theoretical mean) length of ~24Mbp, for the 27 chromosomes of *Physcomitrella* (Reski 1999) this value is ~19Mbp. Taken together, this data and the prediction of around 25,000 and 26,000 protein genes show that the *Physcomitrella* genome is probably less tightly packed than the *Arabidopsis* genome.

The GC-content of the predicted *Arabidopsis* coding regions is 43.97%, whereas the GC-content of the *Physcomitrella* predicted ORFs is significantly higher at 49.81%. When comparing the codon usage of the two organisms, there is a clear tendency of *Physcomitrella* to equalize the fraction of used codons out of the pool of possible codons. This is especially true for the ASP, VAL, ALA, ARG, CYS, HIS and PRO codons, where the strong bias towards one codon found in *Arabidopsis* cannot be detected in *Physcomitrella* (data not shown). In comparison with *Arabidopsis*, *Physcomitrella* uses the stop codon TGA more often than TAA.

Comparison to all known protein coding genes

In order to figure out the fraction of the transcriptome that does not share a significant sequence homology with any protein coding gene known so far, we ran homology searches of the clustered *Physcomitrella* EST database against GENPEPT, utilizing BLASTX and an E-value threshold of 1×10^{-2} . Only 46.8% of the *Physcomitrella* contigs yielded hits. We also

ran a search of the predicted *Physcomitrella* translated ORF database against GENPEPT (BLASTP, E-value cutoff 1×10^{-2}) resulting in 47.5% hits. Using the latter search result as well as the above-mentioned search *Physcomitrella* ORFs against Arabidopsis, we removed hits that are present only in one of the two searches, leaving us with the surprisingly high number of 12,566 ORFs that do not have a counterpart detectable by means of a homology search (redundant no-hits).

In order to further reduce the number of unknown genes, we tried to detect protein family relationships in those ORFs that did not yield a hit in the homology search. Whereas 1,185 of all ORFs could be assigned to a PRINTS motif, this was the case for just 110 of the 12,566 non-hitting ORFs. By using searches against PROSITE and PFAM, a further 796 and 45 hits could be found, respectively. Of those all in all 951 hits only 23 were redundant, so we were able to decrease the number of predicted ORFs with unknown function to 11,638. Hits that were found more than 25 times include the PRINTS motif for “proline rich extension signature” (27x, PR01217) as well as the PROSITE motifs “ATP/GTP binding site A” (94x, PCOC00017), “prokaryotic membrane lipoprotein lipid attachment site” (393x, PCOC00013) and “prenyl group binding site” (36x, PS00294).

Coverage and evolutionary distance

The *Physcomitrella* EST database theoretically over-represents the transcriptome by a factor of >4 (given the >110,000 ESTs and the prediction of around 25,000 genes). As we could show above, both the subtraction and normalization procedures worked very well. Therefore, we would expect a nearly total coverage of the *Physcomitrella* transcriptome by the database. In order to prove this, we ran comparative BLAST searches of the *Physcomitrella* predicted ORFs (24,918), the Arabidopsis predicted genes (26,352, assumed to cover the whole transcriptome) and the rice (*Oryza sativa*) unique unigenes (12,836). The latter database is build from all publicly available rice sequence data and contains one representative for each unigene cluster. When thus comparing Arabidopsis with rice, we found 35.89% non-redundant hits, whereas the comparison with *Physcomitrella* yielded 21.4% hits (table 4). The result of the comparison *Physcomitrella* -> rice (24.76%) shows, that the evolutionary distance of the moss towards the two higher plants is about equal (which is to be expected), if those values are being used as a measure.

It is to be expected that the 12,386 sequences in the rice unigene set do not completely cover the rice transcriptome. This is reflected in the lower percentage values of the reverse comparisons, i.e. 11.20% instead of 24.76% for rice -> *Physcomitrella* and 19.85% instead of 35.89% for rice -> Arabidopsis. If the coverage of the *Physcomitrella* transcriptome by the database analysed here is to be nearly complete, the reverse search *Physcomitrella* -> Arabidopsis should not yield such a decreased percentage value. As it turns out, the result is 24.76%, which is in the same range as the 21.40% of the initial search, proving again the quality of the database in terms of coverage of the transcriptome. Because the percentage value of the reverse search is not lower than that of the initial search, coverage has to be nearly complete.

Conclusions

On the basis of this study, the *Physcomitrella* transcriptome seems to be about as complex as the *Arabidopsis* transcriptome. Whereas around 70% of the *Arabidopsis* genes detect a sequence homologue in the *Physcomitrella* transcriptome, only about 50% of the *Physcomitrella* expressed protein genes can be matched to an *Arabidopsis* homologue. In terms of percent non-redundant hits this equals 23% of the predicted protein encoding genes in both cases and a coverage of the respective transcriptome of about 50%. The EST database presented here is estimated to cover the transcriptome of the moss to at least 95%. Estimated from the percentage of non-redundant BLAST hits, the mosses evolutionary distance towards *Arabidopsis* and rice is about equal. The average numbers of genes that hit a counterpart in a homology search is bigger in *Arabidopsis* (5) than in *Physcomitrella* (3). This suggests that a lot of recently evolved gene family paralogues from higher plants are represented by a lower number of homologous genes in the moss. Furthermore, there is a large number of expressed moss genes that do not have a sequence homologue in higher plants, fungi or animals.

Of those genes, a high proportion might be expressed at a low level, which is suggested by the fact that they are represented by singletons of gametophore and sporophyte tissue. Only a small number of them can be defined by means of protein family/motif assignment. The possible function of the remaining ~11,600 genes remains to be determined.

Our results strongly support the moss *Physcomitrella patens* to be a valuable source for novel genes that is ripe for exploitation (Egener et al., 2001).

Acknowledgements

This work has been performed in a joint project between Freiburg University and BASF Plant Science. Special thanks to J. Lerchl and R.-M. Schmidt (BASF Plant Science) as well as A. Freund, U. Ruffer and G. Haberhauer (BASF AG) and F. Thümmeler (vertis Biotechnologie). Additional financial support by the Forschungsschwerpunkt of the Land Baden-Wuerttemberg and the DFG is gratefully acknowledged. The authors would like to thank B. Ehmman and M.-C. Guitton for discussion. Y. Van de Peer is a Research Fellow of The National Fund for Scientific Research – Flanders.

References

- Apweiler R., Attwood T.K., Bairoch A., Bateman A., Birney E., Biswas M., Bucher P., Cerutti L., Corpet F., Croning M.D.R., Durbin R., Falquet L., Fleischmann W., Gouzy J., Hermjakob H., Hulo N., Jonassen I., Kahn D., Kanapin A., Karavidopoulou Y., Lopez R., Marx B., Mulder N.J., Oinn T.M., Pagni M., Servant F., Sigrist C.J.A., Zdobnov E.M. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res* 29:37-40
- Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., Struhl, K. (eds) (1987) *Current Protocols in Molecular Biology*. New York: John Wiley & Sons
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402
- Barnes, W.M. (1994) PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proc Natl Acad Sci USA* 91:2216-20
- Egener, T., J. Granado, M.-C. Guitton, A. Hohe, H. Holtorf, J.M. Lucht, S. Rensing, K. Schlink, J. Schulte, G. Schween, S. Zimmermann, E. Duwenig, B. Rak, R. Reski (2002): High frequency of phenotypic deviations in *Physcomitrella patens* plants transformed with a gene-disruption library. *BMC Plant Biology*, submitted
- Hohe A. and Reski R. (2002): Optimisation of a bioreactor culture of the moss *Physcomitrella patens* for mass production of protoplasts. *Plant Sci.*, in press.
- Iseli, C., Jongeneel C.V., Bucher P. (1999) ESTScan: a program for detecting, evaluating and reconstructing potential coding regions in EST sequences. In: *Proceedings of the 7. International Conference on Intelligent Systems for Molecular Biology*, ed. Lengauer T. et al., 138-147
- Ko, M.S. (1990) An 'equalized cDNA library' by the reassociation of short double-stranded cDNAs. *Nucleic Acids Res* 18:5707-5711
- Machuka J., Bashiardes S., Ruben E., Spooner K., Cuming A., Knight C., Cove D. (1999) Sequence analysis of expressed sequence tags from an ABA-treated cDNA library identifies stress response genes in the moss *Physcomitrella patens*. *Plant Cell Physiol* 40:378-387
- Marie, D. and Brown, S.C. (1993) A cytometric exercise in plant DNA histograms, with 2C values for 70 species. *Biol Cell* 78:41-51
- Pasentsis, K., Paulo, N., Algarra, P., Dittrich, P., Thümmler F. (1998) Characterization and expression of the phytochrome gene family in the moss *Ceratodon purpureus*. *Plant J* 13:51-61
- Pavy, N., Rombauts, S., Dehais, P., Mathe, C., Ramana, D.V., Leroy, P., Rouze, P. (1999) Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *Bioinformatics* 15:887-899

- 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
- Reski R., Faust M., Wang X.-H., Wehe M., Abel W.O. (1994) Genome analysis of the moss *Physcomitrella patens* (Hedw.) B.S.G. *Mol Gen Genet* 244:352-359
- Reski, R., Reynolds, S., Wehe, M., Kleber-Janke, T., Kruse, S. (1998). Moss *Physcomitrella patens* expressed sequence tags include several sequences which are novel for plants. *Bot Acta* 111:143-149.
- Reski, R. (1998) Development, genetics and molecular biology of mosses. *Bot Acta* 111:1-15
- Reski, R. (1999) Molecular genetics of *Physcomitrella*. *Planta* 208:301-309
- Schaefer, D.G. (2001) Gene targeting in *Physcomitrella patens*. *Curr Op Plant Biol* 4:143-150
- Schween, G., Gorr, G., Lorenz, S., Reski R. (2001) Tissue specific cell cycle in *Physcomitrella patens*. *submitted*
- Strepp, R., S. Scholz, S. Kruse, V. Speth, R. Reski (1998) Plant nuclear gene knockout reveals a role in plastid division for the homologue of the bacterial cell division protein FtsZ, an ancestral tubulin. *Proc. Natl. Acad. Sci. USA* 95:4368-4373
- The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796-815
- Theissen G., Münster T. and Henschel K. (2001) Why don't mosses flower? *New Phytol* 150:1-8

Figures and Tables

table 1

a) amount of sequences that went into clustering and rate of cluster:clone (redundancy)

cDNA library	ESTs	cluster	rate
protonema	57,770	12,058	1:4.79
sub. gametophore	25,521	15,528	1:1.64
sub. sporophyte	26,796	14,876	1:1.80
total	110,087	33,581	1:3.28

b) proportion of clusters from the three libraries that compose the total cluster

cDNA library	cluster	contigs	singletons
protonema	26.9%	35.6%	21.9%
gametophore	36.8%	31.1%	40.2%
sporophyte	36.3%	33.4%	40.2%

The term „cluster“ describes all sequences that are the result of the EST clustering process, i.e. comprising „singletons“ (build from a single sequence) as well as „contigs“ (build from at least two sequences). In the latter case participation of clones from other libraries than the name giving EST originates from may be masked (see methods for details).

table 2

a) single copy genes

<u>search gene type</u>	<u>no.</u>	<u>Ø coverage</u>
„small“ (<600bp)	11	86%
„middle“ (600-1500bp)	11	75%
„large“ (>1500bp)	10	55%

genes used: structure/motility: ftsZ (3), ADF6 (actin depolymerizing factor), signaling: cryptochrome, EGR1 (early growth response) (2), rar1, photosynthesis: ycf9, ycf45, metabolism: cfa (2), RuBisCo LSU, N-myristoyl-transferase, spermidine synthase, transcription/translation/DNA: DNA-Polymerase, MOM (transcriptional silencer), RNA-methylase, topo-isomerase 1, histone H4, PF1 (histone-like), RNA-Polymerase II, rps15, rps17 (2), rpl12 (3), g Faktor, transport: secY, Ca²⁺-ATPase, ABC-transporter. Homologues from: Arabidopsis (13x), Physcomitrella (6x), plants (5x), human (3x), bacteria (2x), algae (2x), cyanobacteria (1x).

b) gene families

<u>cDNA library</u>	<u>hits</u>	<u>percentage</u>	<u>av. per gene</u>
protonema	221	51.8 %	8
gametophore	96	22.5 %	3
sporophyte	110	25.8 %	4

genes used: structure/motility: tubulin, actin, myosin, signaling: PR1, calmodulin, phytochrome, photosynthesis: ferredoxin, CAB, plastocyanin, metabolism: ATPase, L-ascorbate-peroxidase, stilbene-synthetase, phytoene-desaturase, alcohol-dehydrogenase, glutamate-dehydrogenase, nitrate-reductase, reductase, dismutase, ubiquitin, DNA-binding: TBP, ascorbate-oxidase promoter-BP, transport: P-type ATPase (metal transporting), heavy metal tolerance protein, chaperones: GRP94, HSC70, TCP1. Homologues from: Arabidopsis (13x), Physcomitrella (5x), plants (7x), algae (1x), fungi (1x).

table 3

search	total hits	unique	multiple	clusters	total
A.th. vs. P.p.	18023	3166	14857	2895 5.13	6061 23%
P.p. vs. A.th.	15749	3894	11855	3879 3.06	7773 23%

The total hits are divided into unique and multiple hits. The number shown as clusters are the non-redundant multiple hits, numbers below are ratios multiple / clusters. Total is unique plus clusters. Percentage values are relative to the total number of clusters / predicted genes.

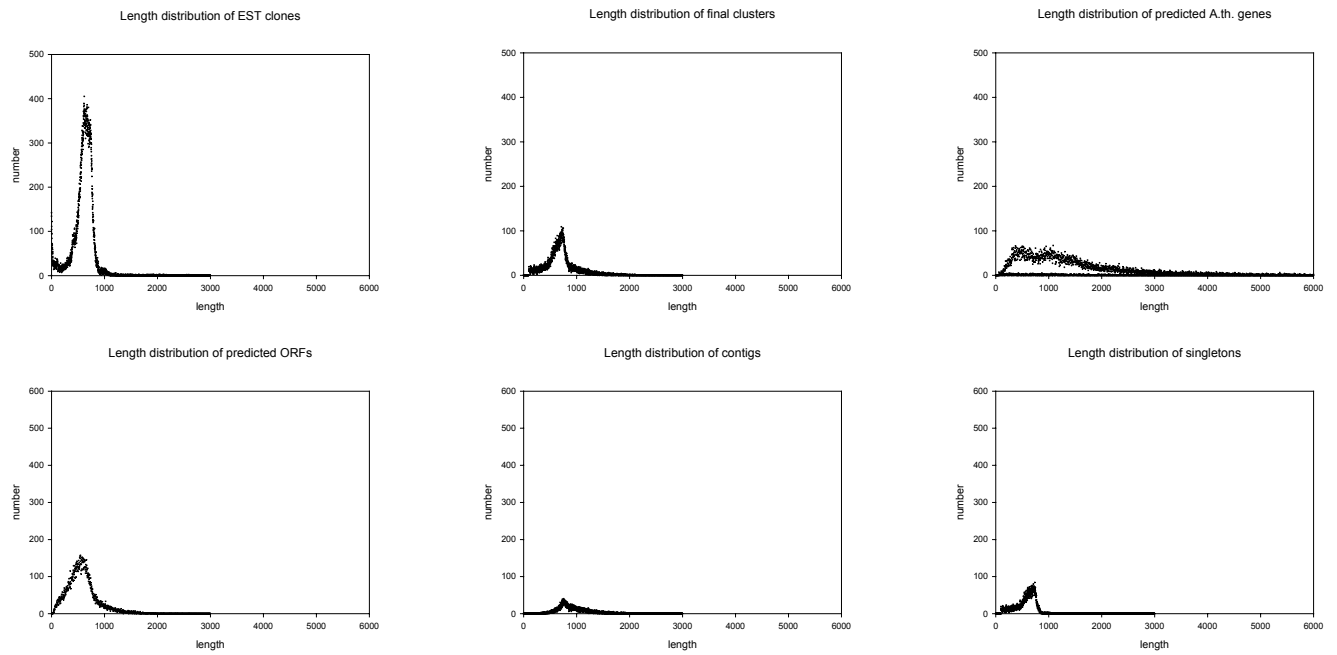
table 4

	query		
	Physcomitrella	Arabidopsis	Oryza
Physcomitrella	X	21.40%	11.20%
Arabidopsis	24.76%	X	19.85%
Oryza	24.76%	35.89%	X

Comparative analysis of sequence data using BLAST (E-value cutoff 1×10^{-2}). The percentage of non-redundant hits is shown.

figure 1

a)



b)

