Representation and High-Quality Annotation of the *Physcomitrella patens* Transcriptome Demonstrates a High Proportion of Proteins Involved in Metabolism in Mosses

D. Lang[1], J. Eisinger[2], R. Reski[1], and S. A. Rensing[1]

[1] Plant Biotechnology, Faculty of Biology, University of Freiburg, Schänzlestraße 1, 79104 Freiburg, Germany
[2] Faculty of Applied Science, Chair of Computer Architecture, University of Freiburg, Georges-Koehler-Allee, Building 051, 79110 Freiburg, Germany

Received: December 12, 2004; Accepted: January 26, 2005

Abstract: **To gain insight into the transcriptome of the well-used plant model system *Physcomitrella patens,* several EST sequencing projects have been undertaken. We have clustered, assembled, and annotated all publicly available EST and CDS sequences in order to represent the transcriptome of this non-seed plant. Here, we present our fully annotated knowledge resource for the *Physcomitrella patens* transcriptome, integrating annotation from the production process of the clustered sequences and from a high-quality annotation pipeline developed during this study. Each transcript is represented as an entity containing full annotations and GO term associations. The whole production, filtering, clustering, and annotation process is being modelled and results in seven datasets, representing the annotated *Physcomitrella* transcriptome from different perspectives. We were able to annotate 63.4% of the 26123 virtual transcripts. The transcript archetype, as covered by our clustered data, is compared to a compilation based on all available *Physcomitrella* full length CDS. The distribution of the gene ontology annotations (GOA) for the virtual transcriptome of *Physcomitrella patens* demonstrates consistency in the ratios of the core molecular functions among the plant GOA. However, the metabolism subcategory is over-represented in bryophytes as compared to seed plants. This observation can be taken as an indicator for the wealth of alternative metabolic pathways in moss in comparison to spermatophytes. All resources presented in this study have been made available to the scientific community through a suite of user-friendly web interfaces via www.cosmoss.org and form the basis for assembly and annotation of the moss genome, which will be sequenced in 2005.**

Key words: ***Physcomitrella patens,*** **moss, transcriptome, annotation, gene ontology.**

Abbreviations:

CDS:	coding sequence(s)
EST:	expressed sequence tag(s)
ORF:	open reading frame(s)
UTR:	untranslated regions(s)
HSP:	highest scoring pair(s)
HMM:	hidden Markov model(s)
GO:	gene ontology
GOA:	gene ontology annotation(s)

Introduction

Physcomitrella patens has become a well-used plant model system, especially for the study of plant gene function using reverse genetics (Reski, 1998, 1999; Cove, 2000; Holtorf et al., 2002). The moss is increasingly being used as an experimental system of choice not only for basic molecular, cytological, and developmental questions in plant biology, but also as a key link in understanding evolutionary questions, especially those related to plant evolution. After the habitat transition from fresh water to land in the early Silurian, green plants have diverged into four major lineages, i.e., hornworts (Anthocerotophyta), liverworts (Hepatophyta), mosses (Bryophyta), and the tracheophytes (vascular plants) comprising of lycophytes (club mosses), ferns (Pteridophyta), and seed plants (Spermatophyta).

The last common ancestor of bryophytes and seed plants lived about 450 million years ago (Theissen et al., 2001), therefore *Physcomitrella* seems predetermined to be used as a phylogenetic link between already sequenced model systems, such as the aquatic, single-celled green alga *Chlamydomonas* and seed plants like rice and *Arabidopsis.* It is known from the fossil record that mosses evolved with little morphological change from the first land plants (Miller, 1984; Frahm, 1994) and thus offer the chance to learn more about embryophyte evolution and diversity. No other phylum offers the opportunity to independently study the three major evolutionary steps during early land plant development, namely filamentous growth of the juvenile gametophyte, the "kormophytic structures" of the adult gametophyte, as well as the evolution of the diploid sporophyte. The presence of a multicellular gametophyte invites research on putative differences between gametophytic and sporophytic gene regulation.

To gain an insight into the transcriptome of this important phylogenetic link, several EST sequencing projects have been undertaken (Rensing et al., 2002 a, b; Nishiyama et al., 2003). Up to now, more than 102000 EST sequences have become publicly available.

Plant Biol. 7 (2005): 238 – 250

DOI 10.1055/s-2005-837578 · Published online April 15, 2005
ISSN 1435-8603

Clustering of EST sequences is an algorithmic approach to reconstruct underlying mRNA structures from the fragmentary information of large-scale EST collections. The approach reduces redundancy and increases information content by grouping overlapping EST representing the same transcript into clusters, which can then be assembled into contiguous consensus sequences, the so-called contigs. Sequences that did not find a matching partner in the initial clustering phase remain as singlet sequences after the procedure. Sometimes not all sequences which were initially grouped into a cluster can be assembled into contig sequences. Hence, there can be multiple contigs and also so-called clustered singlets present in a cluster after the assembly. These may represent multiple transcripts of gene families, alternative splice variants of a single gene or are due to the presence of unmasked repetitive sequence stretches interfering with the alignments (Rensing et al., 2003). Another possibility is the occurrence of cloning artifacts, termed chimeric sequences. The software used in this study tries to detect such chimeric sequences during both the clustering and the assembly phase. In addition it enables the use of full-length mRNA sequences as so-called "seed" sequences, used to remove already known transcripts from the pool of sequences and thus reduce unnecessary computing time ("seed" in this context means initiation, the "seeding" of clusters, not to be confused with the usual meaning of seed, i.e., propagation body).

The general management and representation of the vast amount of data generated during and in the aftermath of large-scale sequencing projects in a user-friendly way is crucial in order to extract biologically meaningful conclusions and hypotheses (Reiser et al., 2002). For most projects, especially in an academic context, an all-in-one solution for the production, storage, annotation, and representation of the information is either not available or too expensive. Thus, a variety of software is being used for the different steps, e.g., sequencing, clustering, and annotation of the sequences, each with a different input/output format and information content. Additionally, scientists working with the data generate their own analysis results. The variability in terms of format of these contents is even higher. Therefore, the accessibility and usability for third parties can hardly be ensured. Redundancy of work carried out and data generated/stored, as well as loss of information, are the consequences. Hence, the goal is to create a central storage form which allows distinct views and representations of the collected data and the creation of so-called data warehouses, e.g., by using a database management system.

Correct functional annotation of the assembled sequences from EST projects is crucial to identification of the underlying transcripts, especially if there is no genome sequence available. Because of the characteristics of EST data, the standard procedure of assigning functional annotations based on similarity (best BLAST hit) is error prone (Kasukawa et al., 2003). Especially in the light of the domain structure of proteins, a more sophisticated approach which takes into account this structure and includes other *a priori* information of the underlying dataset is necessary.

With Gene Ontology (Ashburner et al., 2000; Harris et al., 2004) a powerful biological vocabulary for gene products is being developed by the international Gene Ontology Consortium. This standard spans three basic ontologies, grouping terms concerning the molecular function, the biological process, and the cellular localization of a gene product. Essential to the success of the GO are the so-called GO annotations (GOA), where individual terms are associated with existing gene products. There are several GOA projects (Ware et al., 2002; Camon et al., 2004) undertaken by GO Consortium members. GO term associations can be produced in several ways. This is reflected in evidence codes, e.g., manual association by a curator (IC) or inferred from electronic annotation (IEA). In the past four years, good progress has been made in the development of tools for the analysis and visualization of GO annotations. In times of ever increasing sequence spaces by large-scale sequencing projects, GO provides a powerful tool to derive biologically significant results.

Here, we present our fully annotated knowledge resource of the *Physcomitrella patens* transcriptome, integrating annotation from the production process of the clustered sequences and from a high-quality annotation pipeline developed during this study. Each transcript is represented as an entity containing full annotation and GO term associations. The whole production, filtering, clustering, and annotation process is being modelled and results in seven datasets, representing the annotated *Physcomitrella* transcriptome from different perspectives. The resources have been made available to the worldwide community through a set of user-friendly web interfaces via www.cosmoss.org.

Materials and Methods

EST clustering

All publicly available DNA sequences of *Physcomitrella* were retrieved using Entrez (Schuler et al., 1996) and divided into 399 full-length mRNA sequences ("seed" sequences), as well as 102 535 EST and other sequences. This dataset is called the raw *Physcomitrella patens* public set, or ppp_raw. The clustering project is correspondingly called PPP. Filtering, clustering, and assembly of EST data were done using the Paracel transcript assembler, PTA (www.paracel.com). A species-specific parameter set has been developed and is available upon request. For sequences where electropherograms were available, base-calling was carried out using phred (Ewing et al., 1998). Base quality values of EST sequences for which no sequencer data were available, were set arbitrarily to a low confidence value of 10% and, in the case of "seed" sequences, to 50%. The filtering included steps for removal of synthetic (vector/linker; UniVec, Kitts et al.) and low quality sequences as well as of contaminants (*E. coli* K12 genome as well as *Physcomitrella* mitochondrial, rRNA and chloroplast genes). Low-complexity regions were annotated together with poly-A tails, untranslated regions (UTR, UTRdb, see Pesole et al., 1996) and repetitive elements (repeats, plant partitions of Repbase:, Jurka, 2000), in order not to disturb clustering and assembly. In addition, a set of 17 moss-specific repetitive elements which have been detected mainly in the untranslated regions of *Physcomitrella* genes (Rensing et al., 2003), has also been used to mask these regions in order to avoid erroneous clustering. In a final step, sequences containing less than 150 bases of sense characters were removed. For PPP, a total of 100 079 sequences remained after the filtering procedure – this dataset is available as ppp_fil. The majority of sequences (76%) are derived from

protonemal tissue, whereas only 2% of the sequences are derived from gametophores and 19 EST were annotated to be prepared from sporophytes. Additionally, 17% of the sequences were produced from a mixture of protonema and young gametophores (Fujita et al., 2004). The annotation of the remaining 5% of the sequences did not include any information about the tissue source. Investigation of the annotation references reveals that 81% of the sequences were already made available by Nishiyama et al. (2003). Another 18% were published by the Leeds/Washington University Moss EST Project (Quatrano et al., 1999).

Prior to clustering, homologs of the "seed" sequences were pulled out of the sequence pool and assembled independently. Where possible, sequences were placed into 5′ and 3′ partitions based on detected poly-A tails and annotated cloning information. Both during clustering and assembly, putative chimeras (cloning artefacts) were detected and tagged. During assembly, contigs were built within clusters and putative splice variants detected. After clustering and assembly, the PPP set contained a total of 26123 sequences. By using only the longest sequence in each cluster, a non-redundant set (ppp_nr) of 22218 sequences was produced.

ORF prediction

For the prediction of open reading frames, ESTcan 2.0 (Iseli et al., 1999) was used with a species-specific model for *Physcomitrella* (Rensing et al., 2005). The model was built using the 399 public full length CDS (complete mRNAs) mentioned above. ORF were predicted from the clustered EST data. For the PPP set, 22491 ORF were predicted. Existing CDS features, i.e., in the case of the "seed" sequences (SEE), were preserved. The predicted ORF (ppp_orfpep) were attached to the virtual transcripts of the ppp database as CDS features.

Data integration and database

The XML output files from the PTA filtering package (PFP) were transformed into a relational database schema (available on request). This schema, as well as the BIOSQL database schema (odba.open-bio.org), was installed with PostgreSQL 7.3.4 (www.postgres.org). The automated XML mapping was implemented in Perl 5.6.1 (www.perl.com) using a SAX-based XML parser (XML::Parser::PerlSAX). All reference sequences used in the filtering procedure were retrieved and integrated with as much annotation as possible.

The integration and representation of the clustering data were performed using a set of Perl scripts. These scripts make extensive use of bioperl modules (Stajich et al., 2002), especially the modules for sequence manipulation (Bio::Seq::* and Bio::SeqIO) and object-relational mapping (Bio::DB::BioDB). For the object mapping of the clustering data, two Bioperl object-oriented modules were written: one class for the representation of complete EST clusters including all member sequences and their exon alignments (Bio::Cluster::CAML), and a parser (Bio::ClusterIO::caml) for the PTA XML output files (*.exon.caml). Both modules are available upon request. These mappings resulted in three BioSQL databases (ppp_raw, ppp_fil, and ppp).

Annotation pipeline

InterPro searches were performed with version 3.3 of InterProScan and database version 8.0 on an IBM Blade Cluster (16 nodes, each with 2 Intel Xeon 2.4 GHz CPUs and 2 GB RAM; www.rz.uni-freiburg.de/loginserver), using the algorithms Coil, FPrintScan, HMMPIR, HMMPfam, HMMmart, HMMTigr, ProfileScan, ScanRegExp, HMMPIR, and SUPERFAMILY. These searches were performed with the translated peptide sequences of the predicted ORF. The InterPro results were parsed and attached to the virtual transcripts as DOMAIN features. For this purpose, an object-oriented bioperl module (Bio::Tools::IPrScan) was written. All BLAST (Altschul et al., 1997) searches of the annotation pipeline were run on a Paracel BLAST machine II with 3 nodes (6 × Intel 2.6 GHz CPUs, 6 GB RAM), running Paracel BLAST 1.5.4 (www.paracel.com).

As described in "Results and Discussion", the annotations from the pipeline are grouped in descending quality categories (*"****"* to *"*"*, *"contains"* and *"not annotated"*).

The annotation of the "seed" clusters (****) was implemented with a stringent BLASTN search against the "seed" sequences (minimum HSP length 300 bp, ≥95% identity, and an e-value threshold of 1E-3).

Matches of similarity searches (BLASTX) for the domain-based annotation (***) had to pass the following criteria: minimum HSP length 50 aa, ≥30% sequence identity, and a maximum e-value of 1E-3 against the Uniprot (Apweiler et al., 2004) Swissprot release 44.0 and TREMBL release 27.0 protein databases. Only the first 50 HSP were taken into account.

The closest plant homologs (**) were determined by a homology search (BLASTX; e-value threshold: 1E-3) against a compiled protein database including all translated sequences from *Arabidopsis thaliana* (Rhee et al., 2003) and *Oryza sativa* (Ware et al., 2002) gene predictions retrieved from TIGR (www.tigr.org) and all proteins from the non-redundant Genbank planta (PLN) division, release 142. The similarity based annotation (*) was realized by BLASTX search (e-value threshold: 1E-3) against Genpept (Benton, 1990) release 142.

The weighting algorithm in the domain-based annotation step for sequences with DOMAIN features was implemented as follows: Uniprot sequence entries for the HSP were retrieved and grouped by their InterPro annotation (Camon et al., 2003). Sequences sharing common subsets of InterPro domains were put together in a group. Each group was represented by the subject sequence with the best e-value and bit score. For the ranking of the groups, four scores were assigned: 1) The number of matching InterPro entries between the query and the group, where domains >50% outside the ORF are penalized (counted as 0.8). 2) The percentage of the query covered by the InterPro domains of the group. Non-ORF regions were penalized (length multiplied by 0.8). The scores 3) and 4) were calculated from the e-value and the bit score of the best hit within the group. If less than 80% of the HSP length was covered by the ORF, a penalty was added. The sequence representing the group which performed best in all four rankings was selected for the *** annotation. Virtual transcripts without assigned InterPro domains could not be weighted.

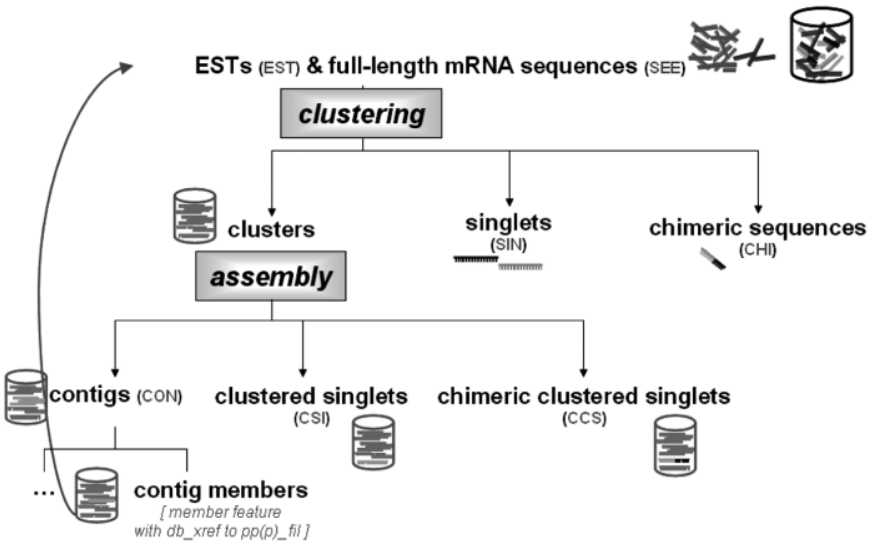

Fig. 1 Hierarchical structure of the datasets. All sequence species are represented as sequence objects belonging to certain database divisions. The corresponding database divisions are shown in brackets. The input EST (EST) and mRNA (SEE) sequences are clustered and yield clusters as well as sequences which do not display pairwise homology, singlets (SIN), and potential chimeric sequences (CHI). During assembly, the overlapping sequences from the clusters are assembled into contigs (CON), while the remaining clustered sequences which cannot be assembled into contigs are divided into clustered singlets (CSI) and chimeric clustered singlets (CCS). Contig member sequences are modelled as sequence features, which are cross-referenced to the input sequence entry.

The whole annotation pipeline is implemented in Perl and can be executed as a single program. To integrate the execution of the parallelized Paracel BLAST jobs, the corresponding bioperl module (Bio::Tools::Run::Standalone) was extended. Version go_200405 of the Gene Ontology databases (Harris et al., 2004) were installed on a MySQL server (www.mysql.com). The GO term associations of rice and *Arabidopsis* GOA projects (Ware et al., 2002; Rhee et al., 2003) were investigated via the AmiGo Browser (www.godatabase.org/cgi-bin/amigo/go.cgi).

Web interfaces

The web-based user interfaces were implemented with a combination of Perl CGI scripts, HTML embedded Javascript functions, bash shell scripts, and modperl modules using an Apache (www.apache.org) webserver with a MySQL database. The overview graphics of the BLAST reports and the transcriptome browser were implemented using the Bioperl interface (Bio::Graphics) to the GD Graphics Library (www.boutell.com/gd).

Detection of cis-acting UTR elements

The detection of cis-acting UTR elements was carried out using the latest version (11.9.2003) of the UTRsite collection (Pesole et al., 2002) with the PatSearch (Grillo et al., 2003) software on a 16-node IBM Blade cluster (see above). The four patterns presented in "Results and Discussion" were selected based on the outcome of the analysis, the length and complexity of the pattern and hits (data not shown) and the biological significance.

Results and Discussion

An integrated knowledge repository for the Physcomitrella patens *transcriptome*

The process of EST clustering and assembly transforms the unrelated pool of raw input sequences into a hierarchical cluster structure. In order to control this complex process, we have created several databases: "raw" input data (ppp_raw), filtered input data (ppp_fil), and clustered and assembled sequences (ppp). The bioperl toolkit (Stajich et al., 2002) enables a broad range of functions for sequence analysis through the use of sequence objects. With our approach, we have fully transformed the hierarchical structure of clusters, contigs, clustered singlets, singlets, and chimeric sequences from the EST clustering into the object-relational OBDA standard (obda.open-bio.org), implemented by all of the Bio* toolkits (Mangalam, 2002). As shown in Fig. **1**, this mapping was realized through the use of well established standards in sequence annotation. All sequences are sequence objects, with several attributes characteristic to the sequence. One of these attributes is the division, an attribute established for the standard of the International Nucleotide Sequence Databases (INSD: DDBJ, EMBL, Genbank) in order to partition the large amount of sequence information into logical units, e.g., sequences from planta origin (PLN). In the tradition of this three-letter standard, we extended the standard and allocated the following divisions as logical units for our datasets: input EST sequences are grouped in the EST division, "seed" sequences in SEE, clusters in CLU, contigs in CON, singlets in SIN, chimeric singlets in CHI, clustered singlets in CSI, and chimeric clustered singlets in CCS.

As described in "Materials and Methods", we have integrated the input sequences with as much annotation as possible into our databases, i.e. conserving full citations, database cross references (db_xrefs), comments, sequence features, and qualifiers (source, CDS see www.ebi.ac.uk/embl/WebFeat) and additional annotation from the original sequence records. In our integration and annotation procedure we have extended the annotations. For instance, the sequence fragments contributing to a contig consensus are represented by so-called member features, with a large set of qualifiers (e.g., FRAG, NAME, FLAVOR, ORIENT, LENGTH, OFFSET) describing their context in the contig. Another example would be QUALITY features, containing the assigned base quality values from Phred (Ewing and Green, 1998) analysis, repeat_unit features, representing hits in the Repbase database (Jurka, 2000), our own set of UTR repeats (Rensing et al., 2003) and exon features, which describe the contigs or singlets contribution to the exon alignment of the cluster.

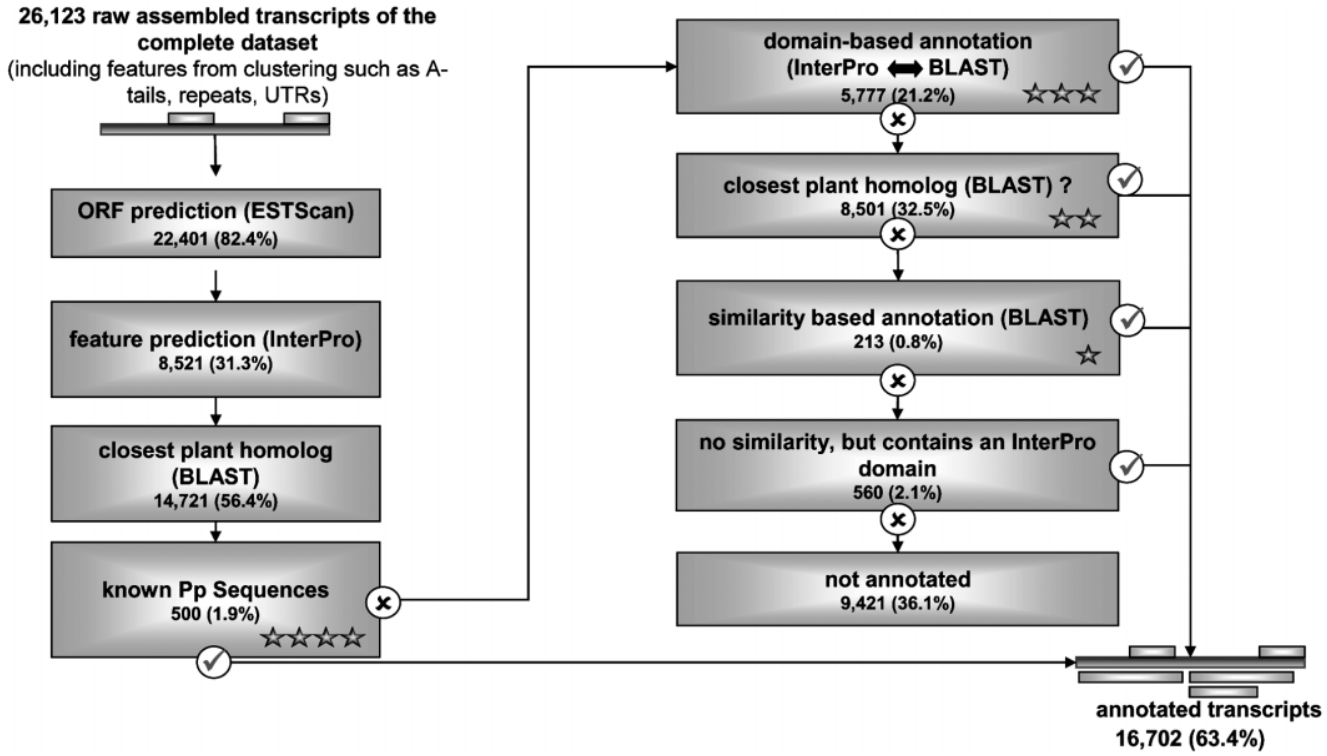

Fig. 2 Overview and outcome of the high-quality annotation pipeline for the *Physcomitrella* transcriptome.

The entries of the three databases are connected through use of database cross references. Each member feature carries a db_xref qualifier, linking it to the original entry in the database of filtered sequences (ppp_fil). These entries can additionally be looked up by their accession numbers in their unprocessed form in the ppp_raw database. Through this combination of consistent accession numbers and database cross references, the whole process of the production, clustering, and annotation can be reproduced. Additionally, we use database cross references to link the sequence objects (DR line in EMBL sequence entries) or sequence features (db_xref qualifier) to their external originating sources (InterPro, Repbase, Genbank, PubMed …).

The relational database incorporating the BioSQL schema (obda.open-bio.org) enabled us to easily create useful subsets of the databases. The database *seeds* is a subset of the ppp_fil database, containing all sequences belonging to the SEE division. As described earlier, ppp_nr is a subset of the ppp database, containing, analogous to the NCBI Unigene (Wheeler et al., 2004) approach, only the longest sequences from each cluster. The ppp_icm subset represents a special division of the ppp database. These sequences are originally member sequences of so-called iterative contigs built in the assembly step of PTA and are therefore named iterative contig members (ICM). Usually, the PTA algorithm discards these sequences while it iteratively builds contigs from existing contigs in order to assemble large clusters. Through our strategy of reproducing the whole pipeline in our databases, we are able to recover these useful pieces of sequence information.

High quality annotation pipeline

We have established a functional annotation pipeline for the *Physcomitrella* transcriptome. Fig. **2** gives an overview and illustrates the outcome of the procedure.

The first step of this pipeline encompasses the prediction of open reading frames for the assembled transcripts (ppp) using the ESTcan software (Iseli et al., 1999). With the species-specific HMM (Rensing et al., 2005), we could predict 22491 ORF. For 41 sequences a second reading frame was predicted. The information, including the peptide translation, was added to the sequence entries as CDS features.

With the translated peptides from the ORF prediction, an InterPro (Mulder et al., 2003) search was carried out. The discovered domains were added as DOMAIN features to the sequences. In total, we could assign 77940 InterPro domains to 8521 (31.3%) of the sequences. In the next step, the whole annotation of the domains was transferred to the features, including the associated GO terms (Harris et al., 2004) from the GOA project (Camon et al., 2004). A total of 32809 GO terms were assigned to 4566 sequences (17.5% of the virtual transcripts). In this InterPro aided GOA a certain level of redundancy is introduced if multiple domains annotated with the same GO term are assigned to the same sequence. Through the removal of this redundancy we can correct the number of assigned GO terms to 12370. A corresponding GOA list will be made available on www.cosmoss.org.

The third step of the pipeline includes a BLASTX search against a comprehensive database of plant proteins, in order to elucidate the closest plant homolog for each sequence. Through this, for 56.4% of the sequences (14721) a closest plant homolog could be determined. The information concerning the closest plant homolog was added as an annotation comment (CC in EMBL entries) to the sequences.

The remaining steps of the pipeline are subtractive, i.e., we begin with the full set of assembled sequences (ppp) and try subsequently to reduce the pool of non-annotated sequences until we finally end up with a set of sequences which cannot be annotated by this method. If a sequence can be annotated in one step, it is removed from the pool. The design of this part of the pipeline is oriented in a descending quality of annotation category system. These categories are reflected in the structure of the output description lines of the sequences (see below).

In the fourth phase of the annotation pipeline, we try to annotate the sequences from the so-called "seed" clusters, i.e., clusters initiated by full length mRNA sequences. In order to annotate the contigs and singlets of these clusters correctly, we performed a stringent BLASTN search against the sequences of the SEE division. Here, we could annotate 1.9% (500) of the sequences. An annotation in this step is indicated, e.g., by a description line of the following kind: *ppp(03/04):SIN:Z98077:****: Physcomitrella patens mRNA for chlorophyll a/b-binding protein precursor, complete cds.*

The headers of the description lines describe the database, the release, the division, the accession number and the annotation quality category of the sequence. In the case of the annotations from this step, the quality category is considered as four-star (****), indicating the highest reliability of the annotation.

The next step comprises the domain-based annotation (***) of the transcripts, by taking into account the overall domain structure of the gene products compared in the process. Especially for members of gene families, the normal procedure of merely selecting the best BLAST hit often leads to false annotations. To solve this, we have developed a weighting algorithm, which is able to rank the hits of a BLASTX search against the well-annotated Uniprot database according to four scores, which are determined for common subsets of shared InterPro domains.

During runtime, all possible subsets of shared InterPro domains are formed and subsequently reduced to a set of non-overlapping subsets. Theoretically, to construct the subsets for n BLAST hits, all 2n subsets of the power set would have to be considered. However, our algorithm exploits the fact that the different subject sequences usually share a common domain structure. Therefore, only a few large subsets are expected to be found. The measured complexity of the algorithm thus depends linearly on the number n of BLAST hits.

In order to determine the subset which best reflects the overall domain structure of the query sequence, four scores are assigned to the subsets and used for a final ranking of the sets, as described in "Materials and Methods". We have evaluated the procedure manually with 100 randomly selected plant sequences from the Swissprot partition of Uniprot (Apweiler et al., 2004). In this case study, the algorithm had annotated 97% of the sequences correctly. In total, we were able to annotate 21.2% of the virtual transcripts (5777) with this annotation strategy. Sequences annotated in this step are recognizable by a description line such as the following: *ppp(03/04):CON:PPP_8221_ C2:***: O49679 RNase L inhibitor-like protein.*

For the remaining sequence pool, the closest plant homologs were reconsidered – in the case of virtual transcripts that were already annotated within step three of the pipeline, this annotation was transferred to the description line. This fraction of annotated sequences is grouped in the two-star quality category (**) and is characterized as "Homolog of…" (e.g., *ppp[03/04]:CON:PPP_929_C1:**: Homolog of [AB076981] hydroxyanthranilate hydroxycinnamoyl transferase 2 [Avena sativa].* In this step an annotation was assigned to 32.5% of the sequences (8501).

The seventh step of the pipeline was included to cover the remainder of the protein sequence space. It is a BLASTX homology search against the non-redundant GenPept database (Benton, 1990). Through this similarity-based annotation, we could annotate 0.8% (213) of the assembled transcripts (e.g., *ppp[03/04]:CON:PPP_1441_C1:*: Similar to [AF452164] large conductance calcium activated potassium channel pSlo spliceform 1-5A [Periplaneta americana].*

The last step of the pipeline tries to enrich the remaining sequences with as much annotation as possible. For sequences that could not be annotated by the similarity methods, we checked whether there were predicted InterPro domains from step two. If so, we included this annotation in the description lines as, e.g., *ppp(03/04):SIN:AW699379: contains: RNA-binding region RNP-1 (RNA recognition motif) (InterPro:IPR000504,PROSITE:PS00030).* We could assign 560 description lines (2.1%) in this manner.

The description lines of the remaining non-annotated 9421 sequences (36.1%) were provided with useful information about the production of the EST from the source features such as, e.g., *ppp(03/04):SIN:AW739429: not annotated Physcomitrella patens protonemata: 7-day-old tissue auxin treated Moss EST library PPN.* Interestingly, about 73% (6875 sequences) of these transcripts contain a predicted ORF, i.e., they do not consist exclusively of UTR. In a case study (data not shown) an attempt was made to annotate 78 clones which could not be annotated in a previous analysis (Rensing et al., 2002 a, b). About 23% of this sample could be annotated by similarity searches a year later against the revised Genpept 133.0 (Benton, 1990); 70% increase of sequence space since release 124.0, which was used initially. However, this is far less than expected, if the EST would simply be "as yet unknown". By sequencing the opposite end of the clones, a further 19% could be annotated. According to these results and the large proportion of predicted CDS in the non-annotated transcripts, it can be assumed that a significant part of the transcripts (about a quarter to a third) without a current functional annotation are new and/or species-specific and might reveal many interesting candidate genes for understanding the unique characteristics of bryophytes. In total we were able to annotate (categories **** to *) 63.4% (16702 sequences) of the virtual transcriptome (ppp) of *Physcomitrella patens* with our approach.

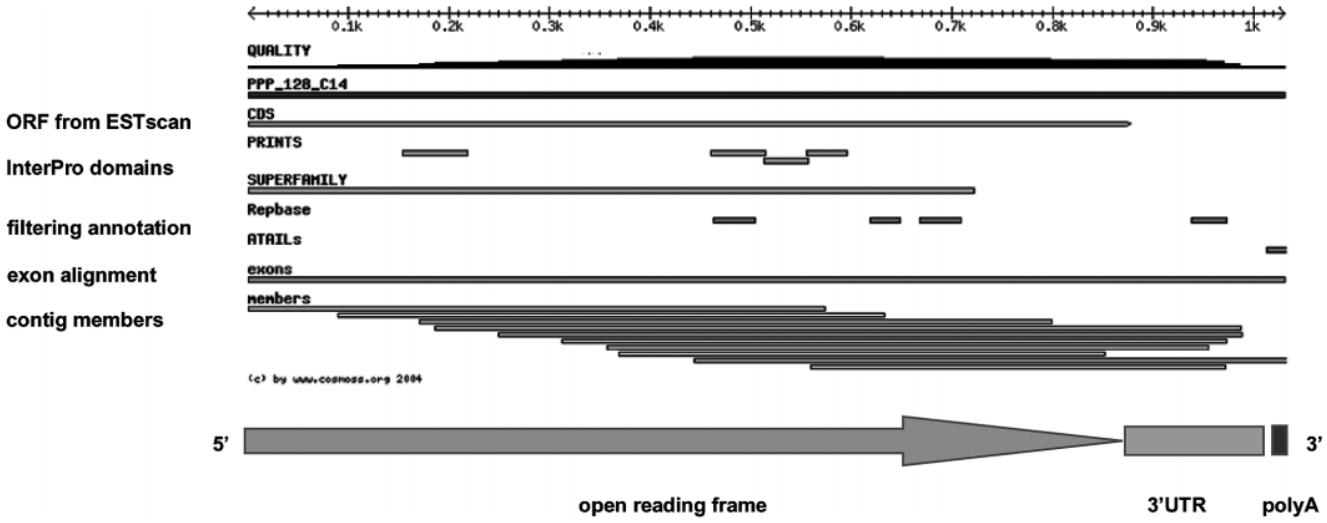

Fig. 3 Recovering the real transcript structure through the annotations provided by the www.cosmoss.org transcriptome browser. The overview graphics are a central part of the entry-based view of the browser. Shown above is a virtual transcript *(ppp[03/04]:CON: PPP_128_C4:***:Q9SWW0 Ethylene-responsive elongation factor EF-Ts precursor)*. The features assigned in the whole process of clustering and annotation of the sequences are aligned along the length of the sequence. The graphics are integrated as a clickable image map into the EMBL formatted sequence entry view of the browser. The combination of these annotations allows the user to reconstruct the underlying transcript structure, as indicated below the graphic.

Web interface and BLAST service – www.cosmoss.org

Besides the improved programmatic access to the data, we have developed a web-based user interface to our system. The internet resource www.cosmoss.org is intended to focus our efforts around the *Physcomitrella* transcriptome in a continuously growing, comprehensive, and user-friendly knowledge resource. The datasets presented in this study are made available as BLAST and sequence databases for the retrieval in various sequence formats (Fasta, Genbank, EMBL, SwissProt). We have developed a graphical interface (www.cosmoss.org/bm) to our BLAST cluster including graphical BLAST results, batch job submission, and email notification. The graphical BLAST results are interactively linked to our sequence retrieval system (www.cosmoss.org/bm/retrieval).

Currently, our system enables the retrieval of sequences based on accession numbers. We support single or multiple retrieval by either a list provided in the input mask or in a file and the so-called fuzzy search mechanism, enabling the retrieval of subsets of sequences. The results of a query can either be downloaded in one of the sequence formats described above or accessed separately through our transcriptome browser.

The transcriptome browser provides a record-based view of each sequence entry. It unravels the whole functionality of the collected data. An overview graphic, such as Fig. **3**, can be used to investigate the structure of the virtual transcripts and the underlying EST and mRNA sequences. The overview graphic is hyperlinked to the corresponding feature annotation section in the sequence entry shown below it. In the EMBL-based sequence entry, database cross references are used to link the database entries to the other datasets (e.g., a ppp CON with the underlying member sequences from ppp_fil) or with external annotations such as, e.g., the Pfam database (Bateman et al., 2004). The browser has additional functionalities, e.g., the retrieval of all sequences that contributed to the selected contig consensus is accessible via a hyperlink below the overview graphic.

The web resource www.cosmoss.org provides, besides the latest news on the transcriptome, several other services such as *Physcomitrella*-specific splice prediction (Rensing et al., 2005), a set of *Physcomitrella* UTR repeats (Rensing et al., 2003), an extended BLAST service for collaborators, and the progress reports from the *Physcomitrella* ecotype collection (von Stackelberg, personal communication). A curators view for the manual annotation is under construction.

Analysis of the datasets

The average raw input sequence (ppp_raw) is about 544 bp long. Filtered sequences (ppp_fil) are on average 534 bp long. The overall tissue composition of the underlying EST and full length mRNA sequences reveals that the majority of sequences are derived from protonema (76 708), whereas data from gametophytic (2943) and sporophytic (9) libraries are underrepresented. An overview of the tissue composition is provided as supplementary material on www.cosmoss.org.

After clustering and assembly, the 100 079 filtered sequences (ppp_fil) were assembled into 12 155 clusters (CLU), consisting of 13 539 contigs (CON), 2518 clustered singlets (CSI), and 208 chimeric clustered singlets (CCS). 9858 sequences did not find a partner during the clustering procedure (singlets). Of these, 327 (CHI) were detected to be chimeric by the PTA algorithm. In summary, the clustering and assembly of the publicly available sequences of *Physcomitrella patens* led to 26 123 virtual transcripts. Nishiyama et al. (2003) assembled 102 553 raw sequences into 22 885 contigs representing 15 883 putative transcripts. Taking into consideration differences in the underlying approaches, e.g., the seeded clustering strategy or the detec-

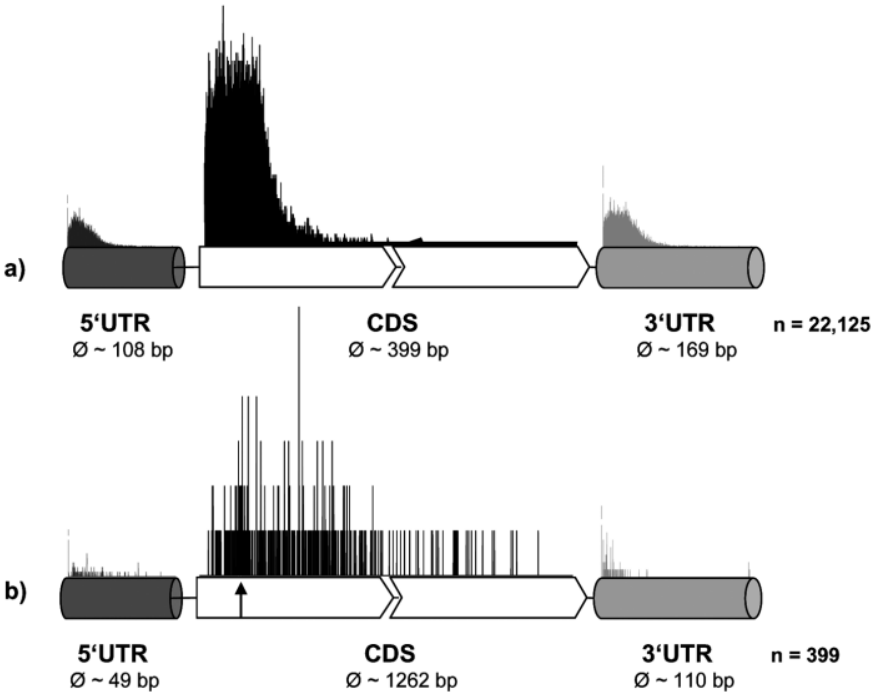

Fig. 4 (**a**) The archetype of a virtual *Physcomitrella patens* transcript from the ppp dataset (n = 22126 virtual transcripts). The corresponding length distributions are shown above each part of the schematic structure for 5′UTR, CDS, and 3′UTR regions. For the frequency axes (y) of the UTR length distribution, breaks (5′UTR: 45 – 12000; 3′UTR: 58 – 7800) had to be inserted in order to display the large fractions of sequences without covered UTR regions (5′UTR: 12418; 3′UTR: 7881). (**b**) The archetype of the 399 full length mRNA sequences used for seed clustering of the virtual transcriptome. The corresponding length distributions are shown above each part of the schematic structure for 5′UTR, CDS, and 3′UTR regions. For the frequency axes (y) of the UTR length distribution, breaks (5′UTR: 8 – 300; 3′UTR: 7 – 280) had to be inserted in order to display the large fractions of sequences without covered UTR regions (5′UTR: 308; 3′UTR: 285). The arrow marks the average length of CDS from **a**.

tion of splice variants by exon alignment, the results are comparable in terms of order of magnitude.

Fig. **4a** shows the archetype of a *Physcomitrella patens* transcript as represented in our virtual transcriptome. The average transcript is about 675 bp long, with 5′ and 3′UTR regions comprising 108 bp and 169 bp, respectively. A detailed length distribution is shown above the corresponding regions in Fig. **4a**. In order to further investigate the structure of *Physcomitrella* transcripts, we also constructed the archetype for the 399 full length "seed" sequences used in clustering. These data are presented in Fig. **4b**. The CDS were on average 1262 bp long, with 5′ and 3′UTR regions with an average of 49 bp and 110 bp. Obviously, the full length mRNAs contain longer coding sequences. This length discrepancy can be explained by the fragmentary nature of EST data. Additionally, CDS lengths might increase with the number of full length sequences available for the construction of the model used to predict the ORF. Considering the sparse UTR distribution of the "seed" sequences (77% of the "seeds" lack 5′UTR; 71% of the "seeds" lack 3′UTR), it is evident that this sample of database entries is focused on CDS. Concerning the lengths of the covered UTR, these are better represented by the assembled transcripts, which are mainly derived from end-sequenced EST and thus cover more of the UTR.

Detection of cis-acting UTR elements

In addition to the features predicted during the filtering, clustering, assembly, and annotation of the transcripts, we investigated the occurrence of known cis-acting UTR elements (UTRsites; Pesole et al., 2002). Fig. **5** illustrates the results of these predictions. The analyses were carried out on two datasets – the complete virtual transcriptome (ppp) and, as an indicator for the pattern vulnerability to produce false-positives, additionally against the predicted open reading frames (ppp_orfpep). The UTRsite patterns should be specific to cis-acting elements in UTR regions, therefore hits in predicted open reading frames can be considered as false positives. The analyses revealed several interesting candidates, e.g., 91 sequences were predicted to include an iron-responsive element (Hentze and Kuhn, 1996). The "iron-responsive element" (IRE; U0002) is a particular hairpin structure located in the 5′-UTR or in the 3′-UTR of various mRNAs coding for proteins involved in cellular iron metabolism. Additional cis-acting elements leading to significant matches (see "Materials and Methods") were two patterns for selenocysteine insertion sequences (SECIS; U0003 122 hits, U0004 540 hits) and one for internal ribosome entry sites (IRES; U0015 1036). Specific incorporation of selenocysteine in selenoproteins is directed by UGA codons residing within the coding sequence of the corresponding mRNAs. Translation of UGA, usually a termination codon, as selenocysteine requires a conserved stem-loop structure called "Selenocysteine Insertion Sequence" (SECIS) lying in the 3′UTR region of selenoprotein mRNAs (Walczak et al., 1996; Fagegaltier et al., 2000). Internal mRNA ribosome binding is a mechanism of translation initiation alternative to the conventional 5′-cap dependent ribosome scanning mechanism (Le and Maizel, 1997). Maybe these alternate translation starts can give an explanation for the dual targeting shown recently for two *Physcomitrella* gene products (Richter et al., 2002; Kiessling et al., 2004).

GO-based classification of the virtual transcriptome

By non-redundant mapping (see above) of the GO terms associated with the InterPro search hits (Camon et al., 2003), we could achieve a classification of the virtual *Physcomitrella* transcriptome in terms of molecular function, cellular component, and biological process. In Fig. **6** we present the distribution of

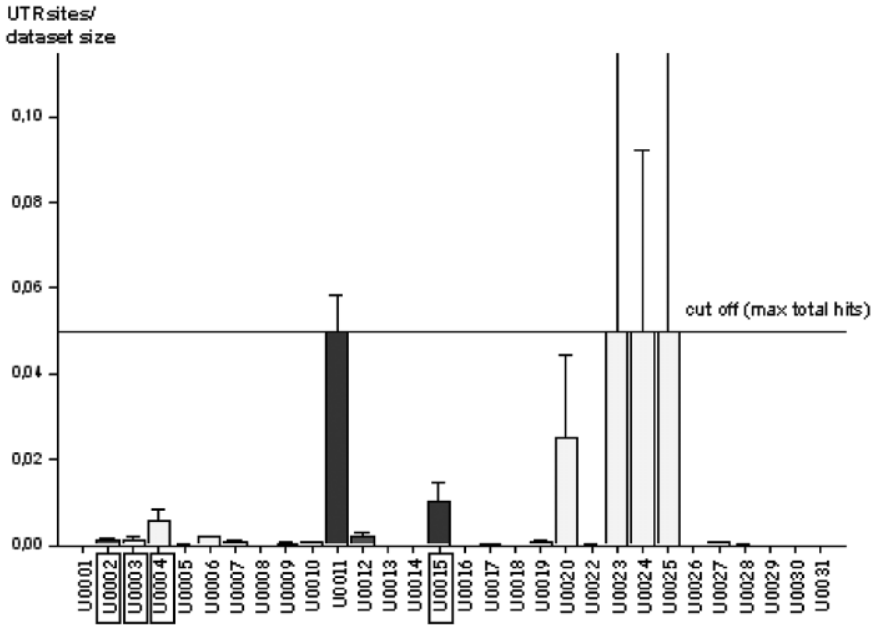

Fig. 5 Detection of known cis-acting UTR elements in the virtual *Physcomitrella patens* transcripts (ppp). The figure gives an overview of the number of hits for each of the 31 UTRsite patterns relative to the search space. The cutoff is defined by the maximum number of hits (5000) the PatSearch algorithm is able to detect. In order to elucidate the pattern vulnerability to produce false positives, a second search with the predicted ORF (ppp_orfpep) was performed. The patterns should be specific to cis-acting elements in UTR regions, therefore, hits in predicted open reading frames can be considered as false-positives. The results are shown by the error bars.

associated GO terms for the three basic Gene Ontology vocabularies (Ashburner et al., 2000; Harris et al., 2004). From the molecular function category (Fig. **6a**, GO:0003674) we could assign 5274 terms to 3922 distinct transcripts, from biological process (Fig. **6b**, GO:0008150) 4390 terms (3830 distinct transcripts), and from the cellular component category (Fig. **6c**, GO:0005575) 2706 terms to 1757 distinct transcripts (in total: 12370 GO terms to 4566 distinct transcripts). In studies that were restricted only to the biological process category, Nishiyama et al. (2003) could annotate 3062 assembled *Physcomitrella* transcripts.

The majority of associations in the molecular function category are distributed among the subcategories catalytic activity (39%; GO:0003824), binding (28%; GO:0005488), structural molecule activity (14%; GO:0005198), and transporter activity (11%; GO:0005215). These findings conform to the expectations concerning the cellular reality and are consistent with ratios observed in rice, *Arabidopsis,* and in the desiccation-tolerant bryophyte *Tortula ruralis* (Ware et al., 2002; Rhee et al., 2003; Oliver et al., 2004).

The lowest number of associations were made to the nutrient reservoir activity subcategory (GO:0045735) – all 8 associated entries were annotated during the ***- annotation step to be putative oxalate oxidases (germin-like proteins). Germin-like proteins (GLPs) exhibit sequence and structural similarity with the cereal germins but mostly lack their oxalate oxidase activity. Germins and GLPs are a class of developmentally regulated glycoproteins characterized by a beta-barrel core structure, a signal peptide, and association with the cell wall. GLPs exhibit a broad range of diversity in their occurrence and activity in organisms ranging from myxomycetes, bryophytes, pteridophytes, gymnosperms, and angiosperms. Germins and GLPs are thought to play a significant role during zygotic and somatic embryogenesis (wheat and *Pinus,* respectively), salt stress (barley and ice plant), pathogen elicitation (wheat and barley), and heavy metal stress (Patnaik and Khurana, 2001).

In the biological process ontology, the most associations were to terms from the metabolism (73%; GO:0008152) and cellular physiological process (21%; GO:0050875) subcategories. This distribution tallies nicely with previous observations made in *Physcomitrella* (Nishiyama et al., 2003) and in *Tortula ruralis* (Oliver et al., 2004). In the two *Arabidopsis* GOA projects (Rhee et al., 2003), 10.4% and 43.8% of the gene products are assigned to the metabolism category (GO:0008152), while this fraction in rice (Gramene: Ware et al., 2002) is about 31.4%. Therefore, the fraction of gene products associated with metabolism (2917 distinct transcripts representing 11% of the transcriptome) is significantly higher in bryophytes (70–80%) than in seed plants (10–44%). An explanation for this observation might be that mosses contain a lot of alternative metabolic pathways not known from seed plants (Girke et al., 1998; Brun et al., 2001; Koprivova et al., 2002; Zank et al., 2002; Mikami et al., 2004; Takezawa and Minami, 2004; von Schwartzenberg et al., 2004).

Besides the major categories, a closer examination of the smaller subcategories again reveals interesting annotations. For example, of the 99 sequences associated with the response to stimulus subcategory (GO:0050896), 36 were annotated as putative peroxidases. 10 sequences with the same association were annotated as putative catalases. Peroxidases and catalases participate in the scavenging of reactive oxygen species (ROS). The intracellular ROS levels increase under stress conditions and lead to severe cell damage, especially caused through the peroxidation of membrane lipids (Wojtaszek, 1997; Apel and Hirt, 2004). This involement makes these enzymes interesting candidates for plant stress tolerance research (Du et al., 2001).

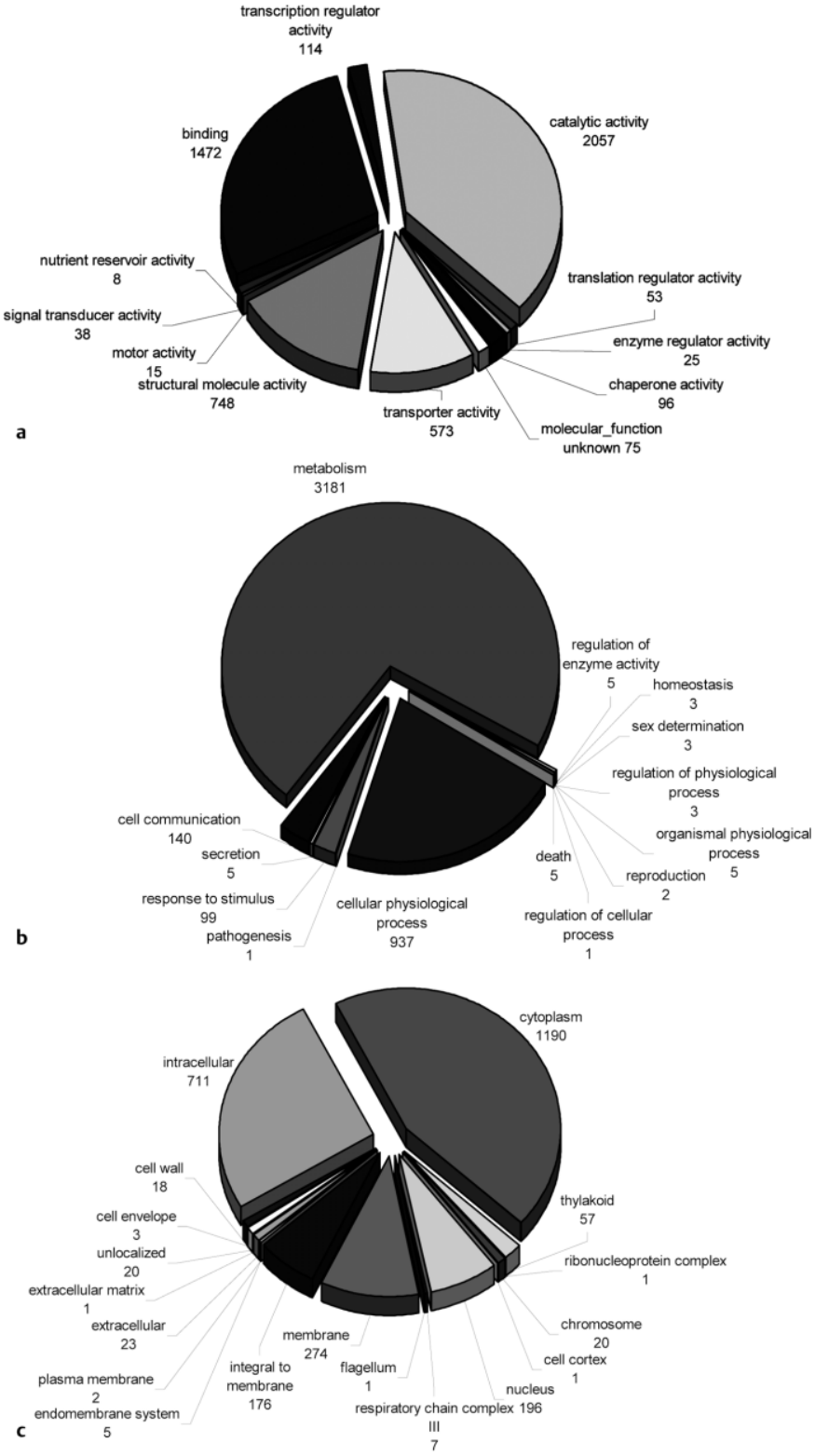

Fig. 6 The distribution of associated GO terms, the numbers given are the observed frequencies of associations for the corresponding term. (**a**) Molecular function category (GO:0003674). In total 5274 terms from this category where assigned to 3922 distinct transcripts. (**b**) Biological process category (GO:0008150). In total 4390 terms from this category where assigned to 3830 distinct transcripts. (**c**) Cellular component category (GO:0005575). In order to increase the information content, we have merged levels 3, 4, and 5 from the cellular component ontology. In total 2706 terms from this category where assigned to 1757 distinct transcripts.

Another interesting group of gene products associated to this category (GO:0050896) are dehydrins (two GOA) and LEA proteins (one GOA). Dehydrins and LEA proteins can be found in all plants. Although their expression is strongly correlated with the plant stress response, their molecular function is still unknown. The dehydrins are considered as stress proteins involved in formation of plant protective reactions against dehydration (Allagulova Ch et al., 2003). Late embryogenesis abundant (LEA) proteins are produced in maturing seeds and anhydrobiotic plants, animals, and microorganisms, in which their expression correlates with desiccation tolerance (Wise and Tunnacliffe, 2004).

In order to increase the information content given in Fig. **6c**, we have merged levels 3, 4, and 5 from the cellular component ontology. The majority of sequences (2183) were assigned to the intracellular compartment (GO:0005622). From this category, 55% were annotated as localized in the cytoplasm (GO:0005737), 9% were found in the nucleus (GO:0005634), and about 3% were targeted to thylakoids (GO:0009579). 457 sequences were annotated as localized in the membrane (GO:0016020). 1% of these were assigned to the endomembrane system (GO:0012505), e.g., the N-acetylglucosaminyltransferase I (Koprivova et al., 2003). 24 transcripts were annotated as localized in the extracellular (GO:0005576) compartment. Seventeen of them are annotated as putative pectin methylesterases (PME). These enzymes have been recently confirmed to be part of the *Physcomitrella patens* secretome (S. Tintelnot, personal communication).

Conclusions

We have generated an integrated knowledge repository for the assembled virtual *Physcomitrella* transcriptome and have developed infrastructure for the further analysis and annotation of the transcriptome.Thus, we are well prepared for the assembly and annotation of the forthcoming genome sequence.

The high-quality annotation pipeline described here was able to annotate 63.4% of the 26123 virtual transcripts. The procedure grouped the annotations of the sequences according to six descending quality categories. 1.9% of the sequences were annotated in the **** category, 21.2% in the domain-based annotation (***), 32.5% in the ** category, 0.8% in the purely similarity-based annotation (*), and, for 2.1% of the transcripts, the annotation was based on occurrence of InterPro domains alone. The remaining 36.1% of the sequences were provided with information concerning the production of the sequences.

Our analyses revealed the archetype of a virtual transcript, with a total average length of 675 bp, a coding sequence (CDS) of 399 bp, and untranslated regions (UTR) of 108 bp (5′UTR) and 169 bp (3′UTR). A comparison with the archetype derived from a small set of full length mRNAs illustrated the good overall coverage of the UTR regions, but also exposed the fragmentary nature of the predicted CDS. In addition, we detected several known eukaryotic cis-acting UTR elements in the assembled transcripts.

A Gene Ontology annotation (GOA) for the virtual transcriptome of *Physcomitrella patens* will be made available on www.cosmoss.org. The distribution of the GO terms assigned in this study demonstrated some consistency in the ratios of the core molecular functions among the plant GOA. However, the biological process ontology revealed a significant over-representation of gene products involved in metabolism in bryophytes in comparison with seed plants. This observation can be taken as an indicator for the wealth of metabolite pathways in mosses in comparison to spermatophytes. Additionally, the investigation of low abundant subcategories of all three ontologies demonstrated many interesting candidates from various fields of plant research.

The presented data are made available through a user-friendly web interface, including a BLAST service as well as a sequence retrieval system and a transcriptome browser. The combination of these resources will help scientists from all over the world to investigate the moss transcriptome. The primary area of application for the knowledge resource will be as an interface for homology searching, sequence retrieval in a variety of sequence formats, and browsing the transcriptome. Especially, the record-based view of the transcriptome browser enables researchers to recover the structure of the underlying transcripts, as presented in Fig. **3**.

We plan to further expand the database by integrating results from expression profiling experiments (Kroemer et al., 2004) and proteomics (Heintz et al., 2004; Sarnighausen et al., 2004), as well as by adding the possibility to manually curate the annotations. Furthermore, in combination with data from metabolic mutant screening (Schween et al., 2005), these integrated resources will be a valuable tool to establish systems biology approaches for *Physcomitrella patens*.

Acknowledgements

The authors would like to thank Stephane Rombauts (University of Ghent) for fruitful discussions concerning the annotation pipeline; Eva Decker and Wolfgang Frank for their help in the interpretation of the GOA; and Colette Matthewman for comments on the manuscript.

References

Allagulova, C. R., Gimalov, F. R., Shakirova, F. M., and Vakhitov, V. A. (2003) The plant dehydrins: structure and putative functions. Biochemistry (Mosc) 68, 945 – 951.

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

Apel, K. and Hirt, H. (2004) Reactive oxygen species: metabolism, oxidative stress, and signal transduction. Annual Review of Plant Biology 55, 373 – 399.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L. S. (2004) UniProt: the Universal Protein knowledgebase. Nucleic Acids Research (Database issue) 32, D115 – D119.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25, 25 – 29.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004) The Pfam protein families database. Nucleic Acids Research (Database issue) 32, D138 – D141.

Benton, D. (1990) Recent changes in the GenBank on-line service. Nucleic Acids Research 18, 1517 – 1520.

Brun, F., Gonneau, M., Doutriaux, M. P., Laloue, M., and Nogue, F. (2001) Cloning of the PpMSH-2 cDNA of *Physcomitrella patens,* a moss in which gene targeting by homologous recombination occurs at high frequency. Biochimie 83, 1003 – 1008.

Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J., Cox, A., and Apweiler, R. (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. Genome Research 13, 662 – 672.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R. (2004) The Gene Ontology Annotation (GOA) database: sharing knowledge in Uniprot with gene ontology. Nucleic Acids Research (Database issue) 32, D262 – D266.

Cove, D. (2000) The moss, *Physcomitrella patens.* Journal of Plant Growth Regulation 19, 275 – 283.

Du, X. M., Yin, W. X., Zhao, Y. X., and Zhang, H. (2001) [The production and scavenging of reactive oxygen species in plants]. Sheng Wu Gong Cheng Xue Bao 17, 121 – 125.

Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research 8, 186 – 194.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research 8, 175 – 185.

Fagegaltier, D., Lescure, A., Walczak, R., Carbon, P., and Krol, A. (2000) Structural analysis of new local features in SECIS RNA hairpins. Nucleic Acids Research 28, 2679 – 2689.

Frahm, J.-P. (1994) Moose – lebende Fossilien. Biologie in unserer Zeit 24, 120 – 124.

Fujita, T., Shin-i, T., Seki, M., Kamiya, A., Uchiyama, I., Nishiyama, T., Carninci, P., Hayashizaki, Y., Shinozaki, K., Kohara, Y., and Hasebe, M. (2004) 82317 Genbank accessions.

Girke, T., Schmidt, H., Zahringer, U., Reski, R., and Heinz, E. (1998) Identification of a novel delta 6-acyl-group desaturase by targeted gene disruption in *Physcomitrella patens.* The Plant Journal 15, 39 – 48.

Grillo, G., Licciulli, F., Liuni, S., Sbisa, E. and Pesole, G. (2003) PatSearch: A program for the detection of patterns and structural motifs in nucleotide sequences. Nucleic Acids Research 31, 3608 – 3612.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., and White, R. (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research (Database issue) 32, D258 – D261.

Heintz, D., Wurtz, V., High, A. A., Van Dorsselaer, A., Reski, R., and Sarnighausen, E. (2004) An efficient protocol for the identification of protein phosphorylation in a seedless plant, sensitive enough to detect members of signalling cascades. Electrophoresis 25, 1149 – 1159.

Hentze, M. W. and Kuhn, L. C. (1996) Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. Proceedings of the National Academy of Sciences of the USA 93, 8175 – 8182.

Holtorf, H., Guitton, M. C., and Reski, R. (2002) Plant functional genomics. Naturwissenschaften 89, 235 – 249.

Iseli, C., Jongeneel, C. V., and Bucher, P. (1999). ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. International Conference on Intelligent Systems for Molecular Biology, pp. 138 – 148.

Jurka, J. (2000) Repbase update: a database and an electronic journal of repetitive elements. Trends in Genetics 16, 418 – 420.

Kasukawa, T., Furuno, M., Nikaido, I., Bono, H., Hume, D. A., Bult, C., Hill, D. P., Baldarelli, R., Gough, J., Kanapin, A., Matsuda, H., Schriml, L. M., Hayashizaki, Y., Okazaki, Y., and Quackenbush, J. (2003) Development and evaluation of an automated annotation pipeline and cDNA annotation system. Genome Research 13, 1542 – 1551.

Kiessling, J., Martin, A., Gremillon, L., Rensing, S. A., Nick, P., Sarnighausen, E., Decker, E. L., and Reski, R. (2004) Dual targeting of plastid division protein FtsZ to chloroplasts and the cytoplasm. Embo Reports 5, 889 – 894.

Kitts, P. A., Madden, T. L. H. S., and Ostell, J. A. UniVec. www.ncbi.nlm.nih.gov/VecScreen/UniVec.html.

Koprivova, A., Altmann, F., Gorr, G., Kopriva, S., Reski, R., and Decker, E. L. (2003) N-glycosylation in the moss *Physcomitrella patens* is organized similarly to that in higher plants. Plant Biology 5, 582 – 591.

Koprivova, A., Meyer, A. J., Schween, G., Herschbach, C., Reski, R., and Kopriva, S. (2002) Functional knockout of the adenosine 5′-phosphosulfate reductase gene in *Physcomitrella patens* revives an old route of sulfate assimilation. Journal of Biological Chemistry 277, 32195 – 32201.

Kroemer, K., Reski, R., and Frank, W. (2004) Abiotic stress response in the moss *Physcomitrella patens:* evidence for an evolutionary alteration in signaling pathways in land plants. Plant Cell Reports 22, 864 – 870.

Le, S. Y. and Maizel, J. V. Jr. (1997) A common RNA structural motif involved in the internal initiation of translation of cellular mRNAs. Nucleic Acids Research 25, 362 – 369.

Mangalam, H. (2002) The Bio* toolkits – a brief overview. Briefings in Bioinformatics 3, 296 – 302.

Mikami, K., Repp, A., Graebe-Abts, E., and Hartmann, E. (2004) Isolation of cDNAs encoding typical and novel types of phosphoinositide-specific phospholipase C from the moss *Physcomitrella patens.* Journal of Experimental Botany 55, 1437 – 1439.

Miller, N. D. (1984) Tertiary and quarternary fossils. In New Manual of Bryology, Vol. 2 (Schuster, R. M., ed.), Miyazaki: Hattori Bot. Lab., pp. 1194 – 1232.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R. R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S. E., Pagni, M., Peyruc, D., Ponting, C. P., Selengut, J. D., Servant, F., Sigrist, C. J., Vaughan, R., and Zdobnov, E. M. (2003) The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Research 31, 315 – 318.

Nishiyama, T., Fujita, T., Shin, I. T., Seki, M., Nishide, H., Uchiyama, I., Kamiya, A., Carninci, P., Hayashizaki, Y., Shinozaki, K., Kohara, Y., and Hasebe, M. (2003) Comparative genomics of *Physcomitrella patens* gametophytic transcriptome and *Arabidopsis thaliana:* implication for land plant evolution. Proceedings of the National Academy of Sciences of the USA 100, 8007 – 8012.

Oliver, M. J., Dowd, S. E., Zaragoza, J., Mauget, S. A., and Payton, P. R. (2004) The rehydration transcriptome of the desiccation-tolerant bryophyte Tortula ruralis: Transcript classification and analysis. BMC Genomics 5, 89.

Patnaik, D. and Khurana, P. (2001) Germins and germin like proteins: an overview. The Journal of Experimental Biology 39, 191 – 200.

Pesole, G., Grillo, G., and Liuni, S. (1996) Databases of mRNA untranslated regions for metazoa. Computers and Chemistry 20, 141 – 144.

Pesole, G., Liuni, S., Grillo, G., Licciulli, F., Mignone, F., Gissi, C., and Saccone, C. (2002) UTRdb and UTRsite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. Update 2002. Nucleic Acids Research 30, 335 – 340.

Quatrano, R., Bashiardes, S., Cove, D., Cuming, A., Knight, C., Clifton, S., Marra, M., Hillier, L., Pape, D., Martin, J., Wylie, T., Underwood, K., Theising, B., Allen, M., Bowers, Y., Person, B., Swaller, T., Steptoe, M., Gibbons, M., Harvey, N., Ritter, E., Jackson, Y., McCann, R., Waterston, R., and Wilson, R. (1999) Leeds/Wash U Moss EST Project, 19538 Genbank accessions.

Reiser, L., Mueller, L. A., and Rhee, S. Y. (2002) Surviving in a sea of data: a survey of plant genome data resources and issues in building data management systems. Plant Molecular Biology 48, 59 – 74.

Rensing, S. A., Fritzowsky, D., Lang, D., and Reski, R. (2005) Protein encoding genes in an ancient plant: analysis of codon usage, retained genes and splice sites in a moss, *Physcomitrella patens.* BMC Genomics, in press.

Rensing, S. A., Lang, D., and Reski, R. (2003) In silico prediction of UTR repeats using clustered EST data. Proceedings of the German Conference on Bioinformatics. Munich, Germany: Belleville Verlag Michael Farin, pp. 117 – 122.

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

Rensing, S. A., Rombauts, S., Van de Peer, Y., and Reski, R. (2002 b) Moss transcriptome and beyond. Trends in Plant Science 7, 535 – 538.

Reski, R. (1998) Development, genetics and molecular biology of mosses. Botanica Acta 111, 1 – 15.

Reski, R. (1999) Molecular genetics of *Physcomitrella.* Planta 208, 301 – 309.

Rhee, S. Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., Garcia- Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L. A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D. C., Wu, Y., Xu, I., Yoo, D., Yoon, J., and Zhang, P. (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. Nucleic Acids Research 31, 224 – 228.

Richter, U., Kiessling, J., Hedtke, B., Decker, E., Reski, R., Borner, T., and Weihe, A. (2002) Two RpoT genes of *Physcomitrella patens* encode phage-type RNA polymerases with dual targeting to mitochondria and plastids. Gene 290, 95 – 105.

Sarnighausen, E., Wurtz, V., Heintz, D., Van Dorsselaer, A., and Reski, R. (2004) Mapping of the *Physcomitrella patens* proteome. Phytochemistry 65, 1589 – 1607.

Schuler, G. D., Epstein, J. A., Ohkawa, H., and Kans, J. A. (1996) Entrez: molecular biology database and retrieval system. Methods in Enzymology 266, 141 – 162.

Schween, G., Egener, T., Fritzkowsky, D., Granado, J., Guitton, M.-C., Hartmann, N., Hohe, A., Holtorf, H., Lang, D., Lucht, J. M., Reinhard, C., Rensing, S. A., Schlink, K., Schulte, J., and Reski, R. (2005) Large-scale analysis of 73 329 *Physcomitrella* plants transformed with different gene disruption libraries: production parameters and mutant phenotypes. Plant Biology, in press.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., and Birney, E. (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Research 12, 1611 – 1618.

Takezawa, D. and Minami, A. (2004) Calmodulin-binding proteins in bryophytes: identification of abscisic acid-, cold-, and osmotic stress-induced genes encoding novel membrane-bound transporter-like proteins. Biochemical and Biophysical Research Communications 317, 428 – 436.

Theissen, G., Münster, T., and Henschel, K. (2001) Why don't mosses flower? New Phytologist 150, 1 – 8.

von Schwartzenberg, K., Schultze, W., and Kassner, H. (2004) The moss *Physcomitrella patens* releases a tetracyclic diterpene. Plant Cell Reports 22, 780 – 786.

Walczak, R., Westhof, E., Carbon, P., and Krol, A. (1996) A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. RNA 2, 367 – 379.

Ware, D., Jaiswal, P., Ni, J., Pan, X., Chang, K., Clark, K., Teytelman, L., Schmidt, S., Zhao, W., Cartinhour, S., McCouch, S., and Stein, L. (2002) Gramene: a resource for comparative grass genomics. Nucleic Acids Research 30, 103 – 105.

Wheeler, D. L., Church, D. M., Edgar, R., Federhen, S., Helmberg, W., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Suzek, T. O., Tatusova, T. A., and Wagner, L. (2004) Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Research (Database issue) 32, D35 – D40.

Wise, M. J. and Tunnacliffe, A. (2004) POPP the question: what do LEA proteins do? Trends in Plant Science 9, 13 – 17.

Wojtaszek, P. (1997) Oxidative burst: an early plant response to pathogen infection. Biochemical Journal 322, 681 – 692.

Zank, T. K., Zahringer, U., Beckmann, C., Pohnert, G., Boland, W., Holtorf, H., Reski, R., Lerchl, J., and Heinz, E. (2002) Cloning and functional characterisation of an enzyme involved in the elongation of Delta6-polyunsaturated fatty acids from the moss *Physcomitrella patens.* The Plant Journal 31, 255 – 268.

S. A. Rensing

Plant Biotechnology
Faculty of Biology
University of Freiburg
Schänzlestraße 1
79104 Freiburg
Germany

E-mail: stefan.rensing@biologie.uni-freiburg.de

Editor: H. Rennenberg