

Moss transcriptome and beyond

Stefan A. Rensing, Stephane Rombauts,
Yves Van de Peer and Ralf Reski

The ancient land plant *Physcomitrella patens* is a model system that is becoming increasingly important for plant functional genomics because gene knockouts can be produced with relative ease. Recently, several EST-sequencing projects have been launched as a first step towards a thorough functional characterization of the moss. However, for careful comparison with other plant model systems, the complete genomic sequence is needed as well as the transcriptome.

In recent years, many large-scale sequencing projects of both genomes and transcriptomes (i.e. cDNA sequencing yielding ESTs) have been initiated from a broad palette of plants (Table 1), providing invaluable data for comparative genomics and for unravelling the function of previously unknown genes. One of the plants for which a substantial set of EST data is already available is the moss *Physcomitrella patens*. This moss is of special interest for reverse genetics because of its high rate of homologous recombination, which can be used to produce gene knockouts with relative ease, thus enabling rapid identification of gene function [1–4]. In addition, the cell lineage of developing moss protonema makes it an ideal tool for studying developmental processes that cannot easily be addressed in flowering-plant models. For these and other reasons, we believe that the time is ripe to establish an international *Physcomitrella* genome project, aiming at the complete determination of this organism's genomic sequence.

Comparison with higher plants

Mosses are comparable to higher plants in terms of gene content, expression and regulation [5,6] in spite of the ~450 million years since they diverged from each other [7]. Although they are both embryophytic, the life cycle of mosses is dominated by the gametophyte, whereas higher plants remain in the sporophytic generation for most of their life (Fig. 1). The gametophyte is haploid, making it easier to connect genes to phenotypes by loss-of-function mutations. This haploidy, together with the differences from higher plants, make the moss especially interesting for both fundamental and applied research.

Sequencing projects in the animal kingdom did not focus solely on mammals but rather led to the discovery of important evolutionary facts by looking at a diverse set of species (e.g. worm, fly and fish). Unfortunately, most current plant-sequencing projects focus on flowering, crop or classical model plants even though information about the rest of the embryophyta (such as mosses and ferns) will render

comparative genomics of land plants much more useful (see the Opinion article by Kathleen Pryer *et al.* in this issue of *Trends in Plant Science*).

EST sequencing

Based on pilot EST sequencing projects, it has already been shown that the moss genome is a valuable source for unknown genes [6,8]. Currently, extensive moss EST sequencing initiatives are being undertaken, aiming to identify the expressed protein-encoding genes (<http://moss.nibb.ac.jp/> and <http://www.moss.leeds.ac.uk/>). Although some of the data are proprietary (http://www.plant-biotech.net/Rensing_et_al_transcriptome2002.pdf), ~67 000 *Physcomitrella* ESTs are currently available in the public databases. Several cDNA libraries have been constructed from different tissues, representing the complete lifecycle of the moss (a visual representation of the life cycle can be found at <http://www.plant-biotech.net/pics/lifecycle.jpg>). To reduce redundancy before mass sequencing, normalization and subtraction were carried out in some cases (http://www.plant-biotech.net/Rensing_et_al_transcriptome2002.pdf).

Gene content of the moss

Arabidopsis (thale cress) was estimated to have ~25 500 genes [9]. The clustered *Physcomitrella* EST database analysed by Rensing *et al.* (http://www.plant-biotech.net/Rensing_et_al_transcriptome2002.pdf) contained 33 581 clusters (i.e. contigs composed of at least two sequences and singlets made up of a single EST). BLAST searches at the amino acid level (i.e. BLASTP, BLASTX or TBLASTX) with an *E*-value threshold of 1×10^{-2} were used to detect homologous genes. Using a pool of 32 randomly selected unique single-copy genes that did not yield overlapping hits

Table 1. Available EST data from selected crops and model plants

| Organism | Number of ESTs ^a |
|--|-----------------------------|
| <i>Physcomitrella patens</i> (moss) ^b | 177 827 |
| <i>Triticum aestivum</i> (wheat) | 175 841 |
| <i>Arabidopsis thaliana</i> (thale cress) | 175 775 |
| <i>Zea mays</i> (maize) | 174 675 |
| <i>Lycopersicon esculentum</i> (tomato) | 159 217 |
| <i>Oryza sativa</i> (rice) | 142 239 |
| <i>Chlamydomonas reinhardtii</i> (green alga) | 112 493 |
| <i>Solanum tuberosum</i> (potato) | 94 259 |
| <i>Pinus</i> sp. (pine) | 52 831 |
| <i>Populus</i> sp. (poplar) | 27 529 |
| <i>Ceratopteris richardii</i> (fern) | 3587 |
| <i>Brassica napus</i> (oilseed rape) | 1774 |
| <i>Marchantia polymorpha</i> (liverwort) | 1415 |
| <i>Medicago sativa</i> (alfalfa) | 719 |

^aNumbers are derived from searches at the NCBI taxonomy browser (<http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html>) using the taxon identifier and EST as search terms in July 2002.

^bThere are 67 740 ESTs in the international databases and 110 087 proprietary ESTs available to scientific collaborators.

Stefan A. Rensing*

Ralf Reski

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

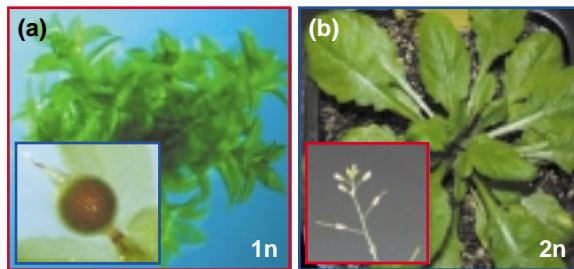
*e-mail: stefan.rensing@
biologie.uni-freiburg.de

Stephane Rombauts

Yves Van de Peer

Dept of Plant Systems
Biology, Flanders
Interuniversity Institute
for Biotechnology (VIB),
University of Ghent,
K.L. Ledeganckstraat 35,
B-9000 Ghent, Belgium.

Fig. 1. Dominance of generations in mosses and seed plants.
 (a) *Physcomitrella patens*. (b) *Arabidopsis*. The gametophyte (1n), or the structures containing it, is bordered by a red frame. The sporophyte (2n), or the structure containing it, is bordered by a blue frame. The gametophytic phenotype of the moss can be expressed as either a protonema or a gametophore (shown). The sporogon (brown) is the reduced sporophytic generation. In the case of seed plants, the gametophytic generation is reduced to pollen tubes, egg cells and other microscopic tissues that are embedded within the flower. Photographs courtesy of Annette Hohe and Jan Lucht.



in the clustered database, a ratio of 0.76 genes per cluster was determined, resulting in an estimation of ~25 500 expressed protein genes in the moss ($0.76 \times 33\,581$). A prediction of open reading frames (ORFs) using ESTscan [10] from the same clustered database led to 24 918 putative expressed protein genes, supporting the previous estimate of 25 500. So, if the moss contains only slightly fewer genes than the cress, where are the differences?

Comparing transcriptomes

Pavy *et al.* [11] predicted 26 352 genes from the *Arabidopsis* genomic sequence, representing a non-redundant *in silico* transcriptome of the cress. The nearly complete information about the *Arabidopsis* genome and the derived *in silico* transcriptome as well as the publicly available EST dataset from rice (*Oryza sativa*) have been used for comparative studies of the predicted *Physcomitrella* transcriptome (http://www.plant-biotech.net/Rensing_et_al_transcriptome2002.pdf). Using BLAST searches, the *Physcomitrella* predicted ORFs yield 47.4% hits with the *Arabidopsis* predicted genes, suggesting that only half of the *Physcomitrella* transcriptome can be matched to the higher plant by means of detectable sequence homology. The results of BLAST searches between the *Physcomitrella* and *Arabidopsis* transcriptome are summarized in Table 2. Using predicted ORFs as queries avoided problems

accounting for 3' untranslated regions present in the ESTs. According to the *Arabidopsis* Genome Initiative [9], the *Arabidopsis* transcriptome consists of 11 601 singletons and distinct gene families. Hence, the 6061 unique + clustered moss sequences that are hit by the *Arabidopsis* genes (Table 2) again equal a coverage of 52% of the transcriptome.

The remainder of *Physcomitrella* ORFs that do not yield hits against either the *Arabidopsis* transcriptome or the sum of all known protein genes (GENPEPT, <http://www.ncbi.nlm.nih.gov/>) is 12 566 ORFs. By scanning for known protein-family motifs using PRINTS (<http://www.bioinf.man.ac.uk/>), PROSITE (<http://www.expasy.ch/>) and PFAM (<http://www.sanger.ac.uk/>), this number was reduced to 11 638, which is still 47% of the predicted *Physcomitrella* transcriptome. This demonstrates that there are surprisingly many expressed *Physcomitrella* genes for which no known homologue in plants, fungi, animals or bacteria can be found.

Smaller gene families

From the 18 023 predicted *Arabidopsis* genes for which a homologue was detected among the *Physcomitrella* clusters (Table 2), the 14 857 multiple hits match a total of 2895 *Physcomitrella* clusters; that is, on average, 5.13 *Arabidopsis* genes hit a *Physcomitrella* gene. In the reverse search, from the 15 749 *Physcomitrella* clusters that found a hit against the predicted *Arabidopsis* genes, the 11 855 multiple hits match a total of 3879 *Arabidopsis* genes; that is, on average, 3.06 *Physcomitrella* clusters hit an *Arabidopsis* gene, which suggests that gene families are larger in *Arabidopsis* than in *Physcomitrella*. This also implies that *Physcomitrella* has more unique genes. Furthermore, many recently evolved gene-family paralogues from *Arabidopsis* appear to be represented by fewer homologous genes in *Physcomitrella*.

Coverage of the transcriptome

Given the >110 000 ESTs and the prediction of ~25 000 genes, the *Physcomitrella* EST dataset analysed by Rensing *et al.* (http://www.plant-biotech.net/Rensing_et_al_transcriptome2002.pdf) theoretically over-represents the transcriptome by a factor of greater than four. Because of the normalized and subtracted libraries used (leading to reduced redundancy), a nearly complete coverage of the transcriptome was expected, as shown in Table 2, by comparison with a rice unigene dataset that is known to be incomplete (and thus not covering the whole transcriptome). A unigene dataset represents each EST cluster by a single contig sequence (this is also the case for the moss clusters used). When comparing *Arabidopsis* with rice, 36% non-redundant hits were found, whereas the comparison with *Physcomitrella* yielded 21% hits. The incompleteness of the rice set is reflected in the lower proportions of reverse comparisons (11% instead of 25% for rice → *Physcomitrella* and 20% instead of 36% for rice → *Arabidopsis*). If the coverage of

Table 2. Comparative BLAST searches^a

| Search ^b | Total hits | Unique | Multiple | Clusters ^c | Total ^d |
|--|------------|--------|----------|-----------------------|--------------------|
| <i>Arabidopsis</i> versus <i>Physcomitrella patens</i> | 18023 | 3166 | 14857 | 2895 | 6061 |
| <i>Physcomitrella patens</i> versus <i>Arabidopsis</i> | 15749 | 3894 | 11855 | 3879 | 7773 |
| | | | 3.06 | 23% | |
| Query ^e | Moss | Cress | Rice | | |
| Moss | – | 21% | 11% | | |
| Cress | 25% | – | 20% | | |
| Rice | 25% | 36% | – | | |

^aPercentages are relative to the total number of clusters or predicted genes.

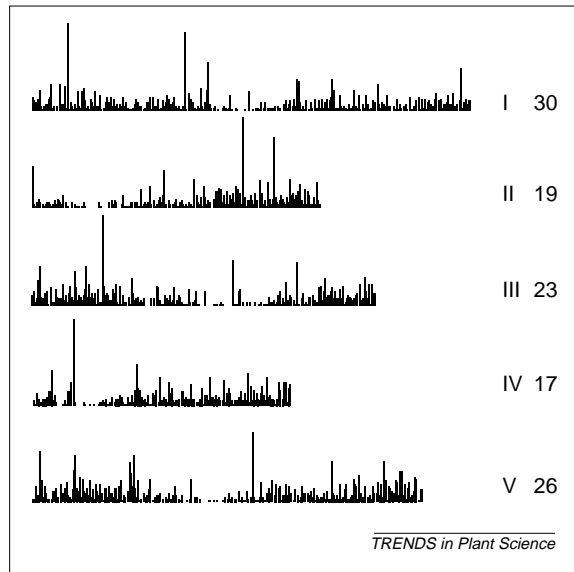
^bThe total number of hits is divided into unique and multiple hits.

^cClusters are the non-redundant multiple hits, numbers below are ratios of multiple to clusters.

^dTotal is uniques plus clusters.

^eResults of comparative BLAST searches, the percentage of non-redundant hits is shown.

Fig. 2. Mapping *Physcomitrella patens* genes against *Arabidopsis* chromosomes. BLASTN searches (using *Physcomitrella*-predicted ORFs as the query) with an *E*-value cutoff of 1×10^{-2} were performed and the best hits mapped onto the *Arabidopsis* chromosomes. The scaling is identical, and chromosome number and approximate sizes in Mbp are shown to the right of each chromosome.



the *Physcomitrella* transcriptome by the clustered database is nearly complete, the reverse search *Physcomitrella* → *Arabidopsis* should not yield such a decreased proportion. As it turns out, the result is 25%, which is in the same range as the 21% of the initial search. Because the proportion from the reverse search is not lower than that of the initial search, we expect the coverage to be nearly complete.

Whereas the EST databases nearly cover the *Physcomitrella* transcriptome in terms of number of genes, they do not contain large proportions of full-length *Physcomitrella* sequences. Fortunately, the publicly available ESTs are mainly from the 5' ends of the cDNAs, whereas the proprietary data were mainly generated by 3'-end sequencing. Therefore, a clustered database combining all *Physcomitrella* sequences should provide an even better coverage (in terms of completeness of genes). The public data as of July 2002 are estimated to cover ~40% of the proprietary data – as determined by comparative BLAST searches – and therefore, in itself, is already a valuable source of information. Recently, MIPS (Munich Information Center for Protein Sequences; <http://mips.gsf.de/proj/sputnik/physcomitrella/>) made a clustered, annotated database of the public *Physcomitrella* ESTs available.

Codon usage and genome size

The GC content of the predicted *Arabidopsis* coding regions is 43.97%, whereas that of the *Physcomitrella* predicted ORFs is estimated to be 49.81%. When comparing the codon usage of the two organisms, there is a tendency of *Physcomitrella* to equalize the proportion of codons used out of the pool of possible codons. *Physcomitrella* does not show the bias towards certain codons seen in *Arabidopsis* (http://www.plantbiotech.net/Rensing_et_al_transcriptome2002.pdf). The haploid genome size of *Physcomitrella* is ~511 Mbp (G. Schween, unpublished). However, the genome of *Arabidopsis* is ~120 Mbp and thus only a quarter of the

Physcomitrella genomic DNA in terms of haploid genome sets. Together with the prediction of ~25 000 and ~26 000 protein genes, respectively, this suggests that the *Physcomitrella* genome is much less tightly packed than the *Arabidopsis* genome. Mapping the *Physcomitrella* transcriptome against *Arabidopsis* chromosomes using BLASTN searches shows an even distribution of hits across the *Arabidopsis* genome (Fig. 2).

Lessons from the past

Apart from acquiring knowledge about gene function, studying the *Physcomitrella* genome is important so that we can learn more about plants and plant-genome evolution in general, 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 ~450 million years ago. 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 recent publication of the rice genome [12, 13] and its comparison with *Arabidopsis* showed the existence of syntenic regions between both genomes, estimated to have diverged from one another 150 million years ago. However, there is still the question of whether there are syntenic regions between mosses and seed plants. A comparison with the moss genome should provide better insights into the evolution of plant genomes in general and would help to unravel the secrets behind thousands of apparently species- or family-specific genes. We can learn a lot from the comparison of mono- and dicots, yet we might learn even more from the comparison with ancient land plants.

Entering moss genomics?

To date, only about half of *Physcomitrella* genes have been assigned a possible function by means of homology, protein family or motif. Therefore, *Physcomitrella* is certainly a valuable source for the identification of novel genes, whose function can be determined by knockout experiments [14]. However, for efficient gene replacement by homologous recombination, long, uninterrupted stretches of homologous DNA should ideally be used (T. Egener, unpublished). Therefore, using knockout constructs derived from genomic DNA should prove to be even more useful than using cDNA for this purpose.

To date, not much is known about promoters, transposable elements, gene duplication or regulatory and repetitive elements in *Physcomitrella*. However, such knowledge would render its use even more efficient as a model organism for functional genomics. Therefore, we suggest that this is a good time to establish an international moss-genome-sequencing project to exploit *Physcomitrella* to its fullest as a model organism for functional and comparative genomics.

Acknowledgements

We are indebted to Daniel Lang for skilful assistance and thank Tanja Egener, Hauke Holtorf and Katja Schlink for discussion.

References

- 1 Strepp, R. *et al.* (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. U. S. A.* 95, 4368–4373
- 2 Reski, R. (1999) Molecular genetics of *Physcomitrella*. *Planta* 208, 301–309
- 3 Schaefer, D.G. (2001) Gene targeting in *Physcomitrella patens*. *Curr. Opin. Plant Biol.* 4, 143–150
- 4 Holtorf, H. *et al.* (2002) Plant functional genomics. *Naturwissenschaften* 89, 235–249
- 5 Reski, R. (1998) Development, genetics and molecular biology of mosses. *Bot. Acta* 111, 1–15
- 6 Reski, R. *et al.* (1998) Moss (*Physcomitrella patens*) expressed sequence tags include several sequences which are novel for plants. *Bot. Acta* 111, 143–149
- 7 Theissen, G. *et al.* (2001) Why don't mosses flower? *New Phytol.* 150, 1–8
- 8 Machuka, J. *et al.* (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
- 9 The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815
- 10 Iseli, C. *et al.* (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* (Lengauer, T. *et al.*, eds), pp.138–147, AAAI Press, Menlo Park, CA, USA
- 11 Pavy, N. *et al.* (1999) Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *Bioinformatics* 15, 887–899
- 12 Yu, J. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296, 79–92
- 13 Goff, S.A. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296, 91–100
- 14 Egner, T. *et al.* (2002) High frequency of phenotypic deviations in *Physcomitrella patens* plants transformed with a gene-disruption library. *BMC Plant Biol.* 2, 6 (<http://www.biomedcentral.com/1471-2229/2/6>)

Obtaining the sequence of the rice genome and lessons learned along the way

C. Robin Buell

Rice holds the record for the largest number of separate genome projects and for having the genome of two subspecies sequenced. This might be a short-lived record in the genomics era, but it highlights the significance of rice as a food staple and as a model plant for cereal species. Clearly, obtaining the genome sequence four times seems redundant, yet the rationale and motivation for each of these projects is valid; whether it is serving corporate shareholders or the general scientific community. Although the multiple projects resulted in some duplicated efforts, the value of data sharing was obvious and the winner in the end will be the global public.

Two basic approaches can be used to sequence a genome (Box 1). One is based on sequencing large insert bacterial artificial chromosome (BAC) clones (see Glossary) [1] and the other is based on a whole genome SHOTGUN SEQUENCING (WGS) approach [2]. Whether the BAC or WGS approach is undertaken, the level of sequence produced can be partial (draft) or complete (finished) (Box 2). With advancements in high-throughput sequencing technologies and significant reductions in sequencing costs, obtaining draft sequence is limited mainly by access to the requisite level of funds.

However, several factors influence whether a genome can (or will) be finished. These include interest by the research community, availability of

committed funds and technical restraints in proceeding from draft to finished sequence. One caveat in finishing is that not all DNA is equivalent when it comes to finishing. Obviously, repetitive sequences pose assembly problems, yet what is not greatly appreciated by bench scientists are the technical difficulties that are not amenable to high-throughput technologies (such as GC hard stops and identifying under-represented sequences) and thus require old-fashioned hard work.

Plant genome-sequencing projects

For most plant biologists, their primary exposure and perhaps only direct experience with large-scale genome projects is the *Arabidopsis* Genome Initiative (AGI) [3], in which high-quality sequence was generated by an international consortium of public entities. For the AGI, the sequence was readily available during the project, and obtaining the finished complete sequence of the *Arabidopsis* genome was a rather rapid event because the genome was a mere 130 Mb and on a technical level, finishing *Arabidopsis* was straightforward. Coupled with the sequence, annotation was released by the AGI-sequencing groups and a more comprehensive, uniform annotation of the *Arabidopsis* genome is now in progress (<http://www.tigr.org/tdb/e2k1/ath1/ath1.shtml>). Although a private corporate entity did report sequencing *Arabidopsis* [4], the Cereon™ effort (Cambridge, MA, USA) was focused on the Landsberg *erecta* (*Ler*) accession and not the Columbia accession used by the AGI. Because of the low coverage (2X) and limited data access, the main use of the Cereon *Ler* sequence has been as a resource for MARKER discovery. However, this does not mean that the Cereon *Ler* sequence is not valuable. Indeed, it has been instrumental in accelerating the pace of positional cloning in *Arabidopsis* because polymorphisms between the Columbia and *Ler* accessions can now be identified rapidly [4].

For rice, the path to a complete genome sequence has not been straightforward and several events have resulted in the unprecedented release of four

C. Robin Buell
The Institute for Genomic
Research, 9712 Medical
Center Dr, Rockville
MD 20850, USA.
e-mail: rbuell@tigr.org