Comparative genomics of *Physcomitrella patens* gametophytic transcriptome and *Arabidopsis thaliana*: Implication for land plant evolution

Tomoaki Nishiyama*†, Tomomichi Fujita*†, Tadasu Shin-I‡, Motoaki Seki§¶, Hiroyo Nishide‖, Ikuo Uchiyama, Asako Kamiya¶, Piero Carninci††, Yoshihide Hayashizaki††, Kazuo Shinozaki§¶, Yuji Kohara‡, and Mitsuyasu Hasebe*‡‡§§**

*Division of Speciation Mechanisms 2 and ‖Computer Laboratory, National Institute for Basic Biology, Okazaki 444-8585, Japan; ‡Genome Biology Laboratory, Center for Genetic Resource Information, National Institute of Genetics, Mishima 411-8540, Japan; §Laboratory of Plant Molecular Biology, RIKEN Tsukuba Institute, Tsukuba 305-0074, Japan; ¶Plant Mutation Exploration Team, Plant Functional Genomics Research Group, RIKEN Genomic Sciences Center, Yokohama 230-0045, Japan; **Research Center for Computational Science, Okazaki National Research Institute, Okazaki 444-8585, Japan; ††Genome Science Laboratory, RIKEN, Wako 352-0198, Japan; and ‡‡Department of Molecular Biomechanics, SOKENDAI, Okazaki 444-8585, Japan

Communicated by Peter R. Crane, Royal Botanic Gardens, Kew, Surrey, United Kingdom, May 6, 2003 (received for review December 26, 2002)

The mosses and flowering plants diverged >400 million years ago. The mosses have haploid-dominant life cycles, whereas the flowering plants are diploid-dominant. The common ancestors of land plants have been inferred to be haploid-dominant, suggesting that genes used in the diploid body of flowering plants were recruited from the genes used in the haploid body of the ancestors during the evolution of land plants. To assess this evolutionary hypothesis, we constructed an EST library of the moss *Physcomitrella patens*, and compared the moss transcriptome to the genome of *Arabidopsis thaliana*. We constructed full-length enriched cDNA libraries from auxin-treated, cytokinin-treated, and untreated gametophytes of *P. patens*, and sequenced both ends of >40,000 clones. These data, together with the mRNA sequences in the public databases, were assembled into 15,883 putative transcripts. Sequence comparisons of *A. thaliana* and *P. patens* showed that at least 66% of the *A. thaliana* genes had homologues in *P. patens*. Comparison of the *P. patens* putative transcripts with all known proteins, revealed 9,907 putative transcripts with high levels of similarity to vascular plant genes, and 850 putative transcripts with high levels of similarity to other organisms. The haploid transcriptome of *P. patens* appears to be quite similar to the *A. thaliana* genome, supporting the evolutionary hypothesis. Our study also revealed that a number of genes are moss specific and were lost in the flowering plant lineage.

Green plants first expanded their habitat to the land by the early Silurian (430 million years ago), and subsequently diverged into various lineages with different morphologies (1). The bryophytes and the ancestor of vascular plants diverged early in land plant evolution. Mosses are bryophytes, and their morphologies and life cycles differ significantly from those of flowering plants (2, 3). The sporophyte (diploid) generation is dominant over the gametophyte (haploid) generation in flowering plants, whereas the opposite occurs in mosses. Leafy shoots differentiate in the diploid generation of flowering plants, and the gametophytes, i.e., pollen tubes and embryo sacs, are epiphytic to the diploid plant bodies, in which fertilization occurs. Mosses propagate by means of spores, from which protonemata, which are filamentous cells, germinate, grow, and differentiate gametophores with haploid leafy shoots. The tissues of the gametophytic leafy shoots of mosses are much simpler than the sporophytic shoots of flowering plants. Archegonia and antheridia differentiate at the tips of the leafy shoots to form eggs and sperm, respectively. Moss sporophytes are epiphytic to the haploid leafy shoots.

Despite the divergent morphologies and life cycles of flowering plants and mosses, the molecular mechanisms of certain physiological processes, such as the light- and abscisic acid (ABA)-signal transduction networks, are preserved. The phytochrome gene, which encodes the red/far-red light receptor and is well characterized in flowering plants, has been cloned from mosses (4, 5). Phototropic responses, which are implicated in the light-associated signal transduction network, have also been reported in mosses (6). The desiccation stress response network mediated by ABA is probably also conserved between mosses and flowering plants (7). Furthermore, auxin and cytokinin are important developmental regulators in both flowering plants and mosses (reviewed in refs. 8 and 9), though their respective roles in development differ in these two plant forms because of developmental and morphological differences.

The differences and similarities between flowering plants and mosses should be reflected in their genomes. Almost the entire genome sequence of *Arabidopsis thaliana* has been determined (10). Therefore, comparisons of the genes found in *A. thaliana* and mosses should provide evolutionary insights into the number of shared and lineage-specific genes among these different lineages.

Physcomitrella patens is a widely used model because it has the highest reported ratio of homologous recombination to nonhomologous recombination of all land plants (11). A number of *P. patens* genes have been isolated and disrupted successfully (12–16). In addition to these individual gene analyses, EST projects are ongoing, and *P. patens* is the most suitable moss for genetic comparisons with the flowering plants. The first *P. patens* EST library, which contained 82 ESTs, was reported by Reski *et al.* (17), and this was followed by a second library of 169 ESTs (18). The *Physcomitrella* EST Programme (www.moss.leeds.ac.uk/) is an ongoing project, and 18,000 ESTs have been deposited to date in GenBank (www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). Ralph Reski's group has accumulated >100,000 ESTs (ref. 19, www.plant-biotech.net/Rensing_et_al_transcriptome2002.pdf), though these data are not deposited in public databases. Additional EST data are required for detailed analysis of the similarities and differences between the *A. thaliana* and *P. patens* genomes.

In this study, we constructed full-length enriched cDNA libraries of *P. patens* from auxin- and cytokinin-treated gametophytes, as well as from gametophytes that were grown without exogenous plant hormones. These phytohormones act at various stages of *P. patens* development, and libraries of clones that have been treated with these hormones should cover a broad range of transcriptome. We combined the 85,191 new ESTs that we determined with other publicly available *P. patens* ESTs, and organized them into 15,883 putative transcripts. These se-

PLANT BIOLOGY

Abbreviations: MST, moss transcripts absent in vascular plants; NAA, naphthalene acetic acid; BA, 6-benzylaminopurine.

†T.N. and T.F. contributed equally to this work.

§§To whom correspondence should be addressed. E-mail: mhasebe@nibb.ac.jp.

quences were compared with those of *A. thaliana* and other organisms. Based on the results of these studies of genetic content, we discuss the evolution of land plants.

Materials and Methods

Plant Materials, RNA Preparation, and Construction of Full-Length cDNA Libraries. *P. patens* (Hedw.) Bruch and Schimp subspecies *patens* Tan (20), collected in Gransden Wood, Huntingdonshire, U.K. (21), was used as the wild-type strain. The protonemata were ground with the Polytron (Kinematica, Littau, Switzerland), and inoculated into BCDATG medium (22), BCD medium (22) that contained 1.0 mM $CaCl_2$ and 1.0 μM naphthalene acetic acid (NAA; Sigma), or BCD medium that contained 1.0 mM $CaCl_2$ and 0.50 μM 6-benzylaminopurine (BA; Sigma) for the untreated, NAA-treated, and BA-treated specimens, respectively, at 25°C under continuous light, and the tissues were harvested at 13–14, 8–11, and 8–13 days, respectively. At harvest, the untreated tissues contained chloronemata and young gametophores with two to five leaves, the NAA-treated tissues contained chloronemata, caulonemata, and rhizoid-like protonemata, and the BA-treated tissues contained chloronemata, caulonemata, and malformed buds.

Total RNA was extracted from each tissue sample as described (23), and poly(A)$^+$ RNA was purified with oligo(dT) magnetic beads (DynaBeads; Dynal A.S., Oslo). The cDNA was synthesized by using trehalose-thermoactivated reverse transcriptase (24), and full-length cDNA was recovered by using the biotinylated CAP trapper method (25, 26). The λ-full-length cDNA vector (27) and the single-strand linker ligation method (28) were used in the construction of the cDNA libraries. One round of normalization (29) was performed to construct the NAA- and BA-treated cDNA libraries to reduce the frequency of highly expressed mRNAs in the libraries.

DNA Sequencing. In the preliminary analyses, 908 clones were chosen from the untreated library, and the DNA was sequenced from the 5′-end. For mass sequence analyses, the clones from the untreated, NAA-treated, and BA-treated libraries were arrayed in 35, 50, and 48 384-well-plates, respectively, and the clones were sequenced at both ends. Low-quality sequences were removed based on PHRED quality score (30) so that average score is at least 20. The sequence data from six plates were inconsistent in terms of their 5′ and 3′ end sequences because of mishandling; therefore, the reverse sequences were renamed and treated as independent sequences during later analyses.

Sequence Data Sets. In addition to the 85,191 ESTs sequenced in this study, all of the *P. patens* mRNA entries in GenBank release 131.0 (excluding our ESTs) were extracted. ESTs with the annotation of "High quality sequence stop" were trimmed to the length specified by the annotation. All of the *A. thaliana* protein sequences were retrieved from the RefSeq database (31) and used in the *A. thaliana* protein data set. The *Synechocystis* sp. PCC 6803, *Saccharomyces cerevisiae*, and *Drosophila melanogaster* coding sequences and translated sequence data sets were obtained from the following locations.

- ftp://ftp.kazusa.or.jp/pub/cyanobase/Synechocystis/Synecho.p.nt.gz
- ftp://ftp.kazusa.or.jp/pub/cyanobase/Synechocystis/Synecho.p.aa.gz
- ftp://genome-ftp.stanford.edu/pub/yeast/yeast_ORFs/orf_coding.fasta.Z
- ftp://genome-ftp.stanford.edu/yeast/yeast_protein/yeast_nrpep.fasta.Z
- ftp://ftp.fruitfly.org/pub/genomic/fasta/na_gadflyCDS.dros.RELEASE2.Z
- ftp://ftp.fruitfly.org/pub/genomic/fasta/aa_gadfly.dros.RELEASE2.Z

The nonredundant protein sequence data set was obtained from ftp://ftp.ncbi.nih.gov/blast/db/nr.Z.

Identification of Taxonomic Positions. The taxonomic classification of the organisms followed that of the National Center for Biotechnology Information taxonomy database (32). The Geninfo Identifier (GI) number for the taxonomy ID (tax ID) table was obtained from ftp://ftp.ncbi.nih.gov/pub/taxonomy/gi_taxid_prot.dmp.gz, and the tree structure was according to "nodes.dmp" in ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz. The tree was traversed from a node that was identified by the tax ID to the root until a specified node was reached. When the root was reached, the taxonomic position was unidentified and thus investigated manually.

Statistical Test of Transcript Representation in Different Libraries. The number of clones in each library was counted for each putative transcript. We used the null hypothesis that a putative transcript is equally represented in the three libraries and therefore the proportion of clones belonging to the putative transcript is constant in the three libraries. We used the G test of independence (33).

Table 1. Base composition of each position in the 5′ end sequences

Position	A	G	C	T	%purine	T/C
1	13,425	21,682	4,836	3,164	81	0.65
2	5,752	12,470	7,865	17,006	42	2.16
3	6,194	11,992	8,826	16,100	42	1.82
4	6,646	12,396	6,667	17,390	44	2.61
10	7,276	11,342	10,996	13,448	43	1.22
20	7,769	11,058	11,325	12,939	44	1.14

The occurrence of each base was counted at each position. The %purine was calculated as $100 \times (A + G)/(A + G + C + T)$.

Results and Discussion

Library Quality. Full-length cDNA clones are important, not only for gene annotation and the determination of transcriptional start sites, but also for functional analyses (34, 35). Recently, methods have been developed that allow preferential cloning of cDNA that corresponds to full-length mRNAs with 5′-end-proximal cap structures (reviewed in ref. 36). These methods have been used in large-scale analyses of transcripts from human (37), mouse (38, 39), fruit fly (40), rice (39), and *A. thaliana* (41, 42).

We constructed three full-length enriched cDNA libraries from the gametophytic tissues of *P. patens* that were NAA-treated, BA-treated, or untreated, and we selected 19,200, 18,432, and 14,348 clones, respectively, from the libraries. The clones were sequenced from both ends. In total, 4.78×10^7 nucleotides of 85,191 sequences remained after the elimination of low-quality sequences and contaminated clones showing nearly complete match to *Escherichia coli* or *Dictyostelium discoideum* sequence.

We analyzed the proportion of full-length cDNA clones that contained mRNA start sites. Because transcription usually starts at a purine base (43), we examined the base composition of the first and subsequent nucleotides at the 5′ end of each 5′ sequence. The first nucleotide showed a strong bias for purine (81% G or A; Table 1), whereas 42–44% of the 2nd to 20th nucleotides were purines. If the cDNAs were not full-length and did not contain the first nucleotide of the mRNA, then the base composition of the first nucleotide would have been identical to

that of nucleotides in subsequent positions. Although terminal transferase activity of the reverse transcriptase, adding C to the 3′ end of cDNA (44), may have contributed partly to over representation of G, this does not account for the increase of A. Therefore, the deviation of first nucleotide base composition from that of nucleotides at subsequent positions suggests that a high proportion of the cDNA clones are full-length. In our 5′ end sequence data (Table 1), the T/C ratio at the first position (0.65) differed from that at the second (2.2) and subsequent (1.1–2.6) positions. This finding implies that a significant proportion of the 5′-end-proximal pyrimidine bases also represent mRNA start sites.

To estimate the proportion of clones that contained complete coding sequences, we examined two *P. patens* tubulin-α genes (Y. Sato, T.F., and M.H., unpublished data) and all of the 38 *P. patens* mRNA deposited in GenBank (release 130.0) that were annotated as containing complete coding sequences. ESTs corresponding to the genes were identified with BLASTN (45) searches and then tested whether they had 5′ or 3′ UTR. The search revealed that 31 genes had corresponding clones in our EST library (Table 4, which is published as supporting information on the PNAS web site, www.pnas.org). In the 5′ end sequences, 95.4% (456 of 478) of the clones contained 5′ UTR, and 97.8% (480 of 491) of the 3′ end sequences contained 3′ UTR (Table 4). If we assume that 5′ and 3′ end truncation occurred independently, the proportion of clones that had both start and stop codons can be calculated as the product of the proportion of clones complete at the 5′ and 3′ ends. Therefore, we estimated that 93% (95.4% $\times$ 97.8%) of the clones contained complete coding sequences.

Assembling the EST Data into Contigs and Estimating the Gene Number. *P. patens* mRNA sequence data from other sources, such as the ESTs from the *Physcomitrella* EST Programme, were obtained from GenBank (release 131.0) and pooled with our data. After removing low-quality sequences, which were identified by their annotations, and vector and linker sequences, which were excluded based on sequence similarity searches, a data set of 5.4×10^7 total nucleotides in 102,553 sequences was obtained. The data were assembled with the CAP3 Sequence Assembly program (46). We obtained 22,885 contigs in total. Of these, 20,589 were represented in our library, whereas the remaining 2,296 contigs arose from other sources. Because our ESTs were obtained by sequencing each clone from the 5′- and 3′-ends, we can assign 5′- and 3′-contigs to the original clone. In total, 814 of the 20,589 contigs had >5% ambiguity (N), and were represented by only one sequence (singlet). These contigs were not used for pairing, because they may not represent independent transcripts. By using the data on the 5′- and 3′-ends, the remaining 19,775 contigs were organized into 13,593 putative transcripts. We assigned a unique identifier (of the form P*nnnnnn*, where *n* is a digit) to each of the 13,593 putative transcripts. Each of the 2,296 contigs from other sources, with the exception of six contaminating sequences from bacteriophage λ, was regarded as an independent putative transcript. The addition of these 2,290 putative transcripts to the 13,593 putative transcripts reveals that at least 15,883 putative transcripts are expressed in *P. patens*.

Although >100,000 sequences were determined for the 22,885 contigs, 9,417 of the 22,885 contigs were composed of only one sequence, i.e., either an EST or a deposited mRNA sequence. This finding indicates that the EST library is not saturated, and the sequencing of more clones will probably reveal new genes. In addition, gametophytes with gametangia and sporophytes were not sampled. Therefore, the total number of *P. patens* transcripts probably exceeds 15,883.

Some Genes Are Represented Differently in the Three Libraries. Because we sequenced three libraries, we were able to identify genes that were represented differently in each library. We used the G test (33) to assess whether putative transcripts were over- or underrepresented in the auxin- and cytokinin-treated libraries. Putative transcripts of this type contained genes that were regulated by auxin and cytokinin, as well as genes that were related to differentiation and development induced by the auxin and cytokinin treatments.

Seven *PpIAA1* cDNA (14) clones were found in the auxin-treated library and one clone was found in the cytokinin-treated library. In this case, the null hypothesis of equal representation in the three libraries was rejected at the 1% level. Induction of the *PpIAA1* gene by auxin is consistent with the reported RNA gel blot analysis (14).

Similarly, all of the putative transcripts were tested for over- or under-representation in each library. Sixty and sixty-eight putative transcripts were over-represented at the 0.1% level in the auxin- and cytokinin-treated libraries, respectively (Tables 5 and 6, which are published as supporting information on the PNAS web site). One and twenty-one putative transcripts were under-represented in the auxin- and cytokinin-treated libraries, respectively (Table 7, which is published as supporting information on the PNAS web site).

Of the putative transcripts that were overrepresented in the cytokinin-treated library, two (P013130 and P006794) showed high-level similarity to cytokinin oxidase, which functions in cytokinin degradation (47). The promoter of the cytokinin oxidase gene *DSCKX1* of *Dendrobium* cv. Sonia was shown to be cytokinin-inducible (48). Overrepresentation of the cytokinin oxidase gene in the cytokinin-treated library is consistent with induction by cytokinin, and suggests that mosses and vascular plants share common mechanisms for the regulation of cytokinin levels. The receptor-like kinases (P007265, P008576, P001848, and P006061) and the ethylene responsive element binding protein (EREBP)-like transcription factors (P012262 and P001581) were also overrepresented in the cytokinin-treated library, though the functions of these factors are unknown.

Many of the *A. thaliana* Genes Have Homologues in the Moss Gametophytic Transcriptome. To analyze the proportion of *A. thaliana* genes that were expressed in the moss gametophytes, all 26,178 amino acid sequences encoded by *A. thaliana* genes were obtained from the RefSeq database (31) at the National Center for Biotechnology Information and used as the query sequences in a similarity search with TBLASTN (45) to the moss data set. The peak occurred at $0.1 < E \leq 10$ and the tail appeared at $E \leq$ 1e-3 (Fig. 2, which is published as supporting information on the PNAS web site). If we regard the *P. patens* contigs with $E \leq$ 1e-3 as homologues of *A. thaliana* query sequences, 17,382 of 26,178 (66.4%) of the *A. thaliana* genes have homologues in *P. patens*. Random matches with an *E* value ≤ 1e-3 would be expected to occur only 26.2 (26,178 $\times$ 0.001) times. Matches occurred at a much higher frequency than this, and most were considered to be caused by homology.

TBLASTN similarity searches were also performed by using the amino acid sequences of the 26,178 *A. thaliana* genes as query sequences and the sequence data sets from budding yeast (*S. cerevisiae*), fruit fly (*D. melanogaster*), and cyanobacterium (*Synechocystis* sp. PCC 6803); putative homologues ($E \leq$ 1e-3) were discovered at frequencies of 40.5%, 47.1%, and 21.3%, respectively. The overlaps of *A. thaliana* genes with homologues in different organisms are summarized in the Venn diagram in Fig. 1. Of the *A. thaliana* genes, 5,039 showed significant similarities to *P. patens* genes, but not to genes of the other three organisms.

Sixteen classes of plant-specific transcription factors have been reported in the *A. thaliana* genome (10). Of these, the MYB R2R3 (49), homeodomain-leucine zipper (50), and ARF-AUX/

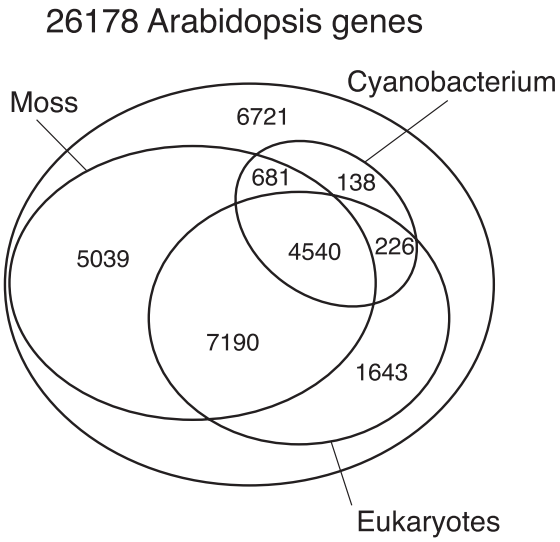

Fig. 1. Classification of *A. thaliana* genes according to the presence of homologous genes in other organisms. The amino acid sequences of all of the predicted genes of *A. thaliana* were used as the query in a TBLASTN similarity search with the moss, yeast, fruit fly, and cyanobacterial data sets. The outermost circle represents all of the *A. thaliana* genes. The inner circles, which are labeled Moss, Eukaryotes, and Cyanobacterium, represent genes that have similarity to moss contigs, budding yeast and fruit fly coding sequences, and cyanobacterium coding sequences, respectively. The areas depicted are not proportional to the actual gene numbers, and the number of *A. thaliana* genes in each category is written in each segment.

IAA (14) genes have been reported, and these genes were also found in the moss EST data. The NAC-like, zinc finger-DOF, zinc finger-WRKY, AP2/EREBP, ABI3/VP1-like and GRAS-like transcription factors were found in the EST data set. The putative transcript P011588 showed high-level similarity to Mybst1, which is a Myb single-repeat protein, which suggests that P011588 also encodes a Myb single-repeat protein, though the complete sequence is required to verify that P011588 encodes only a single Myb domain. The remaining six classes of transcription factors (cycloidea-like, YABBY, EIN3, RAV-like, squamosa-binding protein, and GT1-like) were not identified in the *P. patens* EST data set.

Some gene families, such as the REL-homology region protein, nuclear steroid receptors, forkhead-winged helix, and POU were found in metazoans, but were absent from the *A. thaliana* genome (10) and *P. patens* EST data set.

Because 17,382 (66.4%) of the predicted *A. thaliana* genes had related sequences in the *P. patens* ESTs and the ESTs are not saturated, the number of genes with *P. patens* homologues should be higher. This observation indicates that the *P. patens* gametophytic transcriptome is substantially similar to the *A. thaliana* genome. Although the *A. thaliana* genome sequence has been largely determined, the functions of many of the predicted genes are still unknown. A number of *A. thaliana* genes have homologues in *P. patens*, but not in the yeast or fruit fly. Because gene targeting is feasible in *P. patens*, functional analyses of the homologues in *P. patens* should increase our understanding of gene function in flowering plants.

***P. patens* Expresses Genes That Are Not Present in the *A. thaliana* Genome.** We assessed whether the closest relatives of moss genes existed in vascular plants or in other organisms. The putative moss transcripts were used as the query sequence in a BLASTX (45) search against the *A. thaliana* protein data set and the nonredundant amino acid sequence (nr) data set in GenBank. In total, 9,942 (62.6%) of the 15,883 putative transcripts had significant similarities ($E \leq$ 1e-3) to *A. thaliana* proteins, and 10,757 (67.7%) had significant similarities to the nr data set. This is substantially higher than the estimation by Rensing *et al.* (19) that only half of the *P. patens* genes have known homologue in plants, fungi, animals, or bacteria, despite the fact that we used only gametophytic transcripts. This discrepancy is perhaps because we connected 5′ and 3′ contigs into a single putative transcript, which was used for the search. When we searched with the contigs as query sequences, half of the contigs showed no significant similarity to known proteins, which is consistent to the result by Rensing *et al.* (19).

Table 2. Taxonomic distribution of best hits

Taxonomic position of the best hit	No. of putative transcripts
Vascular plants	9,907
Nonvascular land plants	58
Green algae	51
Metazoa	311
Fungi	60
Other eukaryotes	48
Archaea	12
Cyanobacteria	91
Other bacteria	214
Viruses	2
Synthetic construct	3

The putative transcripts of *P. patens* were subjected to a BLASTX search against the nonredundant dataset, and the organism (excluding the mosses) that carried the best-hit gene was determined and sorted according to taxonomic position.

We sorted the 10,757 putative transcripts into groups based on the organisms that had the highest levels of similarity to the moss transcripts. Of these, 9,907 transcripts had highest levels of similarity to genes in vascular plants, whereas 850 showed highest levels of similarity to genes in other organisms (Table 2 and Tables 8 and 9, which are published as supporting information on the PNAS web site). Some of the 850 putative transcripts showed similarities to both vascular plants and other organisms, with low levels of difference. Because the significance of these differences could not be ascertained, we selected 300 of 850 putative transcripts with *E* values that were at least 1e7-fold lower than the *E* values for best hits with vascular plants. These are candidates for transcripts that are absent in vascular plants but present in mosses. We termed these 300 putative transcripts MSTs (moss transcripts absent in vascular plants). Thirteen of the 300 MSTs were found in liverworts, which are also bryophytes.

The 300 MSTs probably correspond to new genes in land plants. Genome sequencing projects with several vascular plants are underway. Therefore, genes with high levels of similarity to some of the 300 MSTs may be found in future genomic analyses. Nonetheless, we believe that the 300 MSTs may be attributed primarily to gene loss or rapid evolution in vascular plant lineages, or in some cases, to horizontal gene transfer to *P. patens* from other organisms. Sequence conservation among the mosses and other organisms implies that the genes are important in mosses, whereas genes with lower or null levels of similarity in the *A. thaliana* genome either have changed function or are not required in flowering plants, which likely include genes that are involved in moss-specific processes.

Characterization of Moss Transcripts That Are Not Present in Vascular Plants. The 9,907 putative transcripts with the highest levels of similarity to vascular plants corresponded to 6,773 genes in the nonredundant data set, and the 300 MSTs corresponded to 244 genes. The gene ontology (GO) (51) categories of biological

Table 3. Gene Ontology (GO) assignment to putative transcripts

	Common to vascular plants	Absent in vascular plants
No. of putative transcripts	9,907	300
No. of corresponding genes in the nr data set	6,773	244
No. of genes with GO assigned	2,946	116
Metabolism, %	77.7	68.0
Transport, %	13.2	19.8
Signal transduction, %	2.7	4.7
Cell organization and biogenesis, %	1.3	1.7
Cell cycle, %	1.1	0.9
Photosynthesis, %	1.0	0.9
Response to external stimulus, %	0.8	1.1
Response to stress, %	0.5	0.3
Cell growth and/or maintenance, %	0.4	1.7
Cell death, %	0.4	0.0
Others, %	0.8	0.9

nr, nonredundant.

processes were assigned to 2,946 of the 6,773 genes and 116 of the 244 genes, respectively, based on INTERPRO (52) entries that matched the amino acid sequences of the 6,773 and 244 genes. The remaining 3,827 and 128 genes did not match INTERPRO entries. The frequencies of the GO categories at the fourth level are summarized in Table 3. The composition of the annotations was generally similar between the 2,946 and 116 genes, in that the "metabolism" category appeared with the highest frequency; the percentage of genes in this category was lower in the 116-gene group (68.0%) than in the 2,946-gene group (77.7%). The category "metabolism" was divided into subcategories; of these the "protein metabolism" and "biosynthesis" subcategories had fewer representatives among the 116-gene group (12% and 7% of metabolism, respectively) than among the 2,946-gene group (20% and 14%, respectively; Table 10, which is published as supporting information on the PNAS web site). These results suggest that genes that are involved in protein metabolism and biosynthesis are well conserved between mosses and vascular plants. By contrast, the group of 116 genes had a higher proportion of genes that were associated with transport and signal transduction than the group of 2,946 genes (Table 3). This finding may be related to the fact that most cells of mosses, especially the protonemata, are in direct contact with the outside environment.

Examples of Moss Transcripts That Are Absent in Vascular Plants. *P. patens* is known for its extraordinarily high ratio of homologous recombination to nonhomologous recombination. Some of the genes that had the highest levels of similarity to genes from organisms other than vascular plants were related to DNA damage repair, i.e., the P011933 gene for the DNA repair hydrolase, the P002718 gene for RecA, and the P009868 gene for Eso1. P009868 showed high-level similarity to the *Eso1* gene of the fission yeast *Schizosaccharomyces pombe*, with a score of 120 bits and an *E* value of 7e-27. *Eso1* is essential for sister chromatid cohesion (53), which is important for the efficient homologous recombinational repair of DNA double-strand breaks, because the intact sister chromatid paired to the damaged chromatid serves as a repair template (54, 55). *Eso1* is a fusion of DNA polymerase η, which functions in DNA damage repair, and the sister chromatid cohesion protein (53). Although a high level of similarity was found within the DNA polymerase η domain, it remains to be seen whether the moss gene has the sister chromatid cohesion domain, because we obtained only the end sequences. DNA damage repair may be more important during the haploid stage than during the diploid stage, because only one copy of each gene exists in the haploid, whereas an extra copy exists in the diploid cell. Analysis of these genes may unveil the reason why *P. patens* has such a high rate of homologous recombination.

The putative transcripts P001256 and P005876 were highly similar to the chlorophyll a/b-binding protein (cab) homolog LI818r-1 of *Chlamydomonas reinhardtii*. The cab homologue LI818 was first identified in *Chlamydomonas eugametos* as a light-inducible mRNA that encoded a new class of the cab gene family (56). This class of cab homologues has not been reported in other land plants, and may have been lost during the evolution of the vascular plant lineage.

P006275 is an example of a gene that was probably transferred from bacteria, because it showed highest similarity to the glutatione *S*-transferase of *Mesorhizobium loti* (NP_105706), with a score of 251 bits and an *E* value of 9e-66. P006275 also had similarities with a number of glutathione *S*-transferases from other bacteria. The eukaryotic gene with highest similarity to P006275 was the glutatione *S*-transferase gene of *Emericella nidulans*, with 158 bits and 1e-37, whereas the closest gene in green plants was the glutatione *S*-transferase gene of *Glycine max*, with 29 bits and 1e-10. Clones of P006275 were found in all three of our libraries, and also in two libraries from the *Physcomitrella* EST Programme project, which indicates that this sequence is not a contaminant. The absence of a closely related homologue in plants was further tested in a TBLASTN search that matched the amino acid sequence of P006275 with the EST data set. The best non-*P. patens* hit was an EST from *Porphyra yezoensis*, which is a red alga, with 108 bits and an *E* value of 2e-22; no ESTs from green plant species were found, despite the existence of extensive EST libraries from several seed plants. Phylogenetic analysis (Fig. 3, which is published as supporting information on the PNAS web site) also supports a bacterial origin for P006275.

Genes of Sporophyte- and Gametophyte-Dominant Plants. Based on the phylogenetic analyses, the closest relatives of the land plants are the freshwater green algae of Charophyceae (57). The single diploid cells of charophycean algae are zygotes, which undergo meiosis to produce four haploid gametophytic cells. Within the land plant lineage, bryophytes diverged at the most basal position. Both charophycean algae and bryophytes have gametophyte-dominant life cycles, which suggests that the ancestors of land plants, from which sporophyte-dominant vascular plants evolved, had gametophyte-dominant life cycles. Because the multicellular sporophytes probably evolved during land plant history from the gametophyte-dominant plants in the absence of multicellular diploid plant bodies, it seems reasonable to speculate that the genes that function in sporophytes were recruited from genes that functioned in the gametophytes of the common ancestor of mosses and flowering plants. However, there is no direct evidence for this evolutionary scenario, because gene expression in gametophytes of land plants is not well characterized, even in model plants such as *A. thaliana*, in which almost the whole genome has been sequenced. The general similarity between *P. patens* gametophyte ESTs and the *A. thaliana* genome supports the evolutionary scheme that proposes the recruitment of gametophytic genes into the sporophyte generation during land plant history, though we do not have an accurate estimation of the number of *A. thaliana* genes that are expressed in sporophyte generation. Analysis of the genes that are expressed in *A. thaliana* gametophytes will provide further insights into the evolution of the gametophyte and sporophyte generations.

Conclusions

The gametophyte transcriptome of *P. patens* is similar to the *A. thaliana* genome. Our findings that >66% of *A. thaliana*

genes have homologues in *P. patens* gametophytes, and that >90% of the most closely related homologues of *P. patens* gametophytic transcripts occur in vascular plants, suggest that gemetophytes and sporophytes use similar gene sets. On the other hand, 850 putative transcripts of *P. patens* showed lower levels of similarity to genes of *A. thaliana* and vascular plants than to genes in other organisms, thereby indicating that these genes are probably associated with moss-specific features, such as morphology, life cycle, responses to the environment, metabolism, and other aspects. Analyses of these genes will give further evolutionary insights and new genetic resources for agricultural purposes.

The functions of many *A. thaliana* genes have not yet been revealed. Because gene targeting is feasible in *P. patens*, analysis of related genes in *P. patens* will be a good way of achieving an understanding of the function of plant-specific genes. Our full-length cDNA collection is a valuable resource for functional genomics in plants and is distributed by the RIKEN Bio Resource Center (www.brc.riken.go.jp/lab/epd/Eng. The data can be accessed through internet at http://moss.nibb.ac.jp/.

We thank K. Oishi, S. Haga, T. Morishita, S. Miura, F. Ohta, H. Hayashi, S. Nishizaka, H. Nomoto, and M. Sano for EST sequencing. We thank Drs. A. Kitayama and N. Ueno for providing equipment and valuable discussions. We also thank M. Naruse, Y. Bitoh, and Y. Dodo for technical assistance. This work was mainly supported by a Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Science" from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) (to M.H., Y.K., M.S., and K.S.), Grant-in Aid for Scientific Research on Priority Areas (Grants 14036229 and 13044001) from MEXT (to M.H. and T.F.), a Grant-in-Aid from MEXT (to M.H. and T.F.), and Japan Society for the Promotion of Science (JSPS) Grant RFTF00L0162 (to M.H. and T.F.). Construction of the full-length library was also supported by a grant for Genome Research from RIKEN, the Program for Promotion of Basic Research Activities for Innovative Biosciences, the Special Coordination Fund of the Science and Technology Agency, and a Grant-in-Aid from MEXT (to K.S.). T.N. is a JSPS Fellow.

1. Kenrick, P. & Crane, P. R. (1997) *Nature* **389,** 33–39.
2. Bold, H. C., Alexopoulos, C. J. & Delevoryas, T. (1987) *Morphology of Plants and Fungi* (HarperCollins, New York).
3. Reski, R. (1998) *Bot. Acta* **111,** 1–15.
4. Kolukisaoglu, H. U., Braun, B., Martin, W. F. & Schneider-Poetsch, H. A. (1993) *FEBS Lett.* **334,** 95–100.
5. Pasentsis, K., Paulo, N., Algarra, P., Dittrich, P. & Thummler, F. (1998) *Plant J.* **13,** 51–61.
6. Lamparter, T., Esch, H., Cove, D. & Hartmann, E. (1997) *Plant Cell Physiol.* **38,** 51–58.
7. Knight, C. D., Sehgal, A., Atwal, K., Wallace, J. C., Cove, D. J., Coates, D., Quatrano, R. S., Bahadur, S., Stockley, P. G. & Cuming, A. C. (1995) *Plant Cell* **7,** 499–506.
8. Cove, D. J. (1992) in *Development: The Molecular Genetic Approach*, eds. Russo, V. E. A., Brody, S., Cove, D. & Ottolenghi, S. (Springer, Berlin), pp. 179–193.
9. Knight, C. D. (1994) *Plant Cell Environ.* **17,** 669–674.
10. The *Arabidopsis* Genome Initiative (2000) *Nature* **408,** 796–815.
11. Schaefer, D. G. & Zrÿd, J.-P. (1997) *Plant J.* **11,** 1195–1206.
12. Strepp, R., Scholz, S., Kruse, S., Speth, V. & Reski, R. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 4368–4373.
13. Girod, P.-A., Fu, H., Zrÿd, J.-P. & Vierstra, R. D. (1999) *Plant Cell* **11,** 1457–1472.
14. Imaizumi, T., Kadota, A., Hasebe, M. & Wada, M. (2002) *Plant Cell* **14,** 373–386.
15. Zank, T. K., Zähringer, U., Beckmann, C., Pohnert, G., Boland, W., Holtorf, H., Reski, R., Lerchl, J. & Heinz, E. (2002) *Plant J.* **31,** 255–268.
16. Koprivova, A., Meyer, A. J., Schween, G., Herschbach, C., Reski, R. & Kopriva, S. (2002) *J. Biol. Chem.* **277,** 32195–32201.
17. Reski, R., Reynolds, S., Wehe, M., Kleber-Janke, T. & Kruse, S. (1998) *Bot. Acta* **111,** 143–149.
18. Machuka, J., Bashiardes, S., Ruben, E., Spooner, K., Cuming, A., Knight, C. & Cove, D. (1999) *Plant Cell Physiol.* **40,** 378–387.
19. Rensing, S. A., Rombauts, S., Van de Peer, Y. & Reski, R. (2002) *Trends Plant Sci.* **7,** 535–538.
20. Tan, B. C. (1979) *J. Hattori Bot. Lab.* **46,** 327–336.
21. Ashton, N. W. & Cove, D. J. (1977) *Mol. Gen. Genet.* **154,** 87–95.
22. Nishiyama, T., Hiwatashi, Y., Sakakibara, K., Kato, M. & Hasebe, M. (2000) *DNA Res.* **7,** 9–17.
23. Hasebe, M., Wen, C.-K., Kato, M. & Banks, J. A. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 6222–6227.
24. Carninci, P., Nishiyama, Y., Westover, A., Itoh, M., Nagaoka, S., Sasaki, N., Okazaki, Y., Muramatsu, M. & Hayashizaki, Y. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 520–524.
25. Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M., *et al.* (1996) *Genomics* **37,** 327–336.
26. Carninci, P., Westover, A., Nishiyama, Y., Ohsumi, T., Itoh, M., Nagaoka, S., Sasaki, N., Okazaki, Y., Muramatsu, M., Schneider, C. & Hayashizaki, Y. (1997) *DNA Res.* **4,** 61–66.
27. Carninci, P., Shibata, Y., Hayatsu, N., Itoh, M., Shiraki, T., Hirozane, T., Watahiki, A., Shibata, K., Konno, H., Muramatsu, M. & Hayashizaki, Y. (2001) *Genomics* **77,** 79–90.
28. Shibata, Y., Carninci, P., Watahiki, A., Shiraki, T., Konno, H., Muramatsu, M. & Hayashizaki, Y. (2001) *BioTechniques* **30,** 1250–1254.
29. Carninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y., Shibata, K., Itoh, M., Konno, H., Okazaki, Y., Muramatsu, M. & Hayashizaki, Y. (2000) *Genome Res.* **10,** 1617–1630.
30. Ewing, B. & Green, P. (1998) *Genome Res.* **8,** 186–194.
31. Pruitt, K. D. & Maglott, D. R. (2001) *Nucleic Acids Res.* **29,** 137–140.
32. Wheeler, D. L., Chappey, C., Lash, A. E., Leipe, D. D., Madden, T. L., Schuler, G. D., Tatusova, T. A. & Rapp, B. A. (2000) *Nucleic Acids Res.* **28,** 10–14.
33. Sokal, R. R. & Rohlf, F. J. (1995) in *Biometry: The Principles and Practice of Statistics in Biological Research* (Freeman, New York), pp. 724–743.
34. Suzuki, H., Fukunishi, Y., Kagawa, I., Saito, R., Oda, H., Endo, T., Kondo, S., Bono, H., Okazaki, Y. & Hayashizaki, Y. (2001) *Genome Res.* **11,** 1758–1765.
35. Seki, M., Narusaka, M., Yamaguchi-Shinozaki, K., Carninci, P., Kawai, J., Hayashizaki, Y. & Shinozaki, K. (2001) *Plant Physiol. Biochem.* **39,** 211–220.
36. Kristiansen, T. Z. & Pandey, A. (2002) *Trends Biochem. Sci.* **27,** 266–267.
37. Suzuki, Y., Yamashita, R., Nakai, K. & Sugano, S. (2002) *Nucleic Acids Res.* **30,** 328–331.
38. Konno, H., Fukunishi, Y., Shibata, K., Itoh, M., Carninci, P., Sugahara, Y. & Hayashizaki, Y. (2001) *Genome Res.* **11,** 281–289.
39. Osato, N., Itoh, M., Konno, H., Kondo, S., Shibata, K., Carninci, P., Shiraki, T., Shinagawa, A., Arakawa, T., Kikuchi, S., *et al.* (2002) *Genome Res.* **12,** 1127–1134.
40. Stapleton, M., Liao, G., Brokstein, P., Hong, L., Carninci, P., Shiraki, T., Hayashizaki, Y., Champe, M., Pacleb, J., Wan, K., *et al.* (2002) *Genome Res.* **12,** 1294–1300.
41. Seki, M., Carninci, P., Nishiyama, Y., Hayashizaki, Y. & Shinozaki, K. (1998) *Plant J.* **15,** 707–720.
42. Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., *et al.* (2002) *Science* **296,** 141–145.
43. Lewin, B. (2000) *Genes VII* (Oxford Univ. Press, Oxford).
44. Schmidt, W. M. & Mueller, M. W. (1999) *Nucleic Acids Res.* **27,** e31.
45. Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
46. Huang, X. & Madan, A. (1999) *Genome Res.* **9,** 868–877.
47. Galuszka, P., Frebort, I., Sebela, M., Sauer, P., Jacobsen, S. & Pec, P. (2001) *Eur. J. Biochem.* **268,** 450–461.
48. Yang, S. H., Yu, H. & Goh, C. J. (2002) *J. Exp. Bot.* **53,** 1899–1907.
49. Leech, M. J., Kammerer, W., Cove, D. J., Martin, C. & Wang, T. L. (1993) *Plant J.* **3,** 51–61.
50. Sakakibara, K., Nishiyama, T., Kato, M. & Hasebe, M. (2001) *Mol. Biol. Evol.* **18,** 491–502.
51. 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., *et al.* (2000) *Nat. Genet.* **25,** 25–29.
52. Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., *et al.* (2001) *Nucleic Acids Res.* **29,** 37–40.
53. Tanaka, K., Yonekawa, T., Kawasaki, Y., Kai, M., Furuya, K., Iwasaki, M., Murakami, H., Yanagida, M. & Okayama, H. (2000) *Mol. Cell. Biol.* **20,** 3459–3469.
54. Kadyk, L. C. & Hartwell, L. H. (1992) *Genetics* **132,** 387–402.
55. Tanaka, K., Hao, Z., Kai, M. & Okayama, H. (2001) *EMBO J.* **20,** 5779–5790.
56. Gagne, G. & Guertin, M. (1992) *Plant Mol. Biol.* **18,** 429–445.
57. Karol, K. G., McCourt, R. M., Cimino, M. T. & Delwiche, C. F. (2001) *Science* **294,** 2351–2353.