The genome of the moss *Physcomitrella patens* reveals evolutionary insights into the conquest of land by plants

Stefan A. Rensing[1], Daniel Lang[1], Andreas Zimmer[1], Astrid Terry[2], Asaf Salamov[3], Harris Shapiro[3], Tomoaki Nishiyama[4], Pierre-François Perroud[5], Erika Lindquist[3], Yasuko Kamisugi[6], Takako Tanahashi[7,33], Keiko Sakakibara[9], Tomomichi Fujita[10], Kazuko Oishi[8], Tadasu Shin-I[8], Yoko Kuroki[11], Atsushi Toyoda[11], Yutaka Suzuki[12], Shinichi Hashimoto[13], Kazuo Yamaguchi[4,14], Sumio Sugano[12], Yuji Kohara[8,15], Asao Fujiyama[11,16,17], Aldwin Anterola[19], Setsuyuki Aoki[20], Neil Ashton[18], W. Brad Barbazuk[21], Elizabeth Barker[18], Jeffrey Bennetzen[22], Robert Blankenship[5], Sung Hyun Cho[5], Susan Dutcher[23], Mark Estelle[24], Jeffrey A. Fawcett[25], Heidrun Gundlach[26], Kosuke Hanada[27], Alexander Heyl[28], Karen A. Hicks[29], Jon Hughes[30], Martin Lohr[31], Klaus Mayer[26], Alexander Melkozernov[32], Takashi Murata[7,33], David Nelson[34], Birgit Pils[35], Michael Prigge[24], Bernd Reiss[29], Tanya Renner[36], Stephane Rombauts[25], Paul Rushton[37], Anton Sanderfoot[38], Gabriele Schween[1], Shin-Han Shiu[27], Kurt Stueber[29], Frederica L. Theodoulou[39], Hank Tu[3], Yves Van de Peer[25], Paul J. Verrier[40], Elizabeth Waters[36], Andrew Wood[19], Lixing Yang[22], David Cove[5,6], Andrew C. Cuming[6], Mitsuyasu Hasebe[7,33,43], Susan Lucas[2], Brent D. Mishler[41], Ralf Reski[1], Igor Grigoriev[3], Ralph S. Quatrano[5*], Jeffrey L. Boore[3,41,42]

[1] Plant Biotechnology, Faculty of Biology, University of Freiburg, Schaenzlestrasse 1, D-79104 Freiburg, Germany.
[2] U.S. Department of Energy (DOE) Joint Genome Institute and Lawrence Livermore National Laboratory, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA.
[3] DOE Joint Genome Institute and Lawrence Berkeley National Laboratory, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA.
[4] Advanced Science Research Center, Kanazawa University, 13-1 Takara-machi Kanazawa 920-0934, Japan.
[5] Department of Biology, Washington University, 1 Brookings Drive, St. Louis, MO 63130–4899, USA.
[6] Centre for Plant Sciences, University of Leeds, Leeds LS2 9JT, UK.
[7] National Institute for Basic Biology, Okazaki 444-8585, Japan.
[8] Department of Basic Biology, School of Life Science, The Graduate University for Advanced Studies, Okazaki 444-8585, Japan.
[9] School of Biological Sciences, Monash University, Clayton Campus, Melbourne, VIC 3800, Australia.
[10] Department of Biological Sciences, Faculty of Science, Hokkaido University, Sapporo 060-0810, Japan.
[11] Genome Biology Laboratory, Center for Genetic Resource Information, National Institute of Genetics, Mishima 411-8540, Japan.
[12] RIKEN Genomic Sciences Center, Kanagawa 230-0045, Japan.
[13] Laboratory of Functional Genomics, Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo 108-8639, Japan.
[14] Department of Molecular Preventive Medicine, School of Medicine, The University of Tokyo, Tokyo 113-8654, Japan.

[15] Division of Life Science, Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa 920-1192, Japan.
[16] Department of Genetics, School of Life Science, The Graduate University for Advanced Studies, Mishima 411-8540, Japan.
[17] National Institute of Informatics, Tokyo 101-8403, Japan.
[18] Department of Informatics, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies, Tokyo 101-8403, Japan.
[19] Department of Plant Biology, Southern Illinois University, Carbondale, IL 62901–6509, USA.
[20] Life-Science Informatics Unit, Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan.
[21] University of Regina, 3737 Wascana Parkway, Regina, SK S4S 0A2, Canada.
[22] Donald Danforth Plant Science Center, 975 North Warson Road, St. Louis, MO 63132, USA.
[23] Department of Genetics, Davison Life Sciences Complex, University of Georgia, Athens, GA 30602–7223, USA.
[24] Department of Genetics, Washington University, 660 South Euclid Avenue, St. Louis, MO 63108, USA.
[25] Department of Biology, Indiana University, 1001 East Third Street, Bloomington, IN 47405–3700, USA.
[26] VIB Department of Plant Systems Biology, Ghent University, Technologie Park 927, 9052 Ghent, Belgium.
[27] MIPS/IBI Institute for Bioinformatics, GSF Research Center for Environment and Health, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany.
[28] Department of Plant Biology, 166 Plant Biology Building, Michigan State University, East Lansing, MI 48824–1312, USA.
[29] RIKEN Plant Science Center, 1-7-22 Suehiro, Tsurumi, Yokohama, Kanagawa 230-0045, Japan.
[30] Free University, Institute for Biology, Applied Genetics Neubau, Albrecht-Thaer-Weg 6, D-14195 Berlin, Germany.
[31] Max-Planck Institute of Plant Breeding Research, Carl-von-Linne-Weg 10, D-50829 Cologne, Germany.
[32] Biology Department, Kenyon College, Gambier, OH 43022, USA.
[33] Pflanzenphysiologie, Justus Liebig University, Senckenbergstrasse 3, D-35390 Giessen, Germany.
[34] Institute of General Botany, Johannes Gutenberg-University, D-55099 Mainz, Germany.
[35] Department of Chemistry and Biochemistry, Arizona State University, Tempe, AZ 85287–1604, USA.
[36] University of Tennessee-Memphis, 101 Molecular Science Building, 858 Madison Avenue, Memphis, TN 38163, USA.
[37] Department of Bioinformatics, Biozentrum, Am Hubland, Würzburg University, D-97074 Würzburg, Germany.
[38] Biology Department, San Diego State University, North Life Sciences Room 102, 5500 Campanile Drive, San Diego, CA 92182–4614, USA.
[39] Department of Biology, Gilmer Hall, 485 McCormick Road, University of Virginia, Charlottesville, VA 22903, USA.
[40] Department of Plant Biology, University of Minnesota, 250 Biological Science Center, 1445 Gortner Avenue, St. Paul, MN 55108, USA.

[41] Biological Chemistry Department, Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK.
[42] Biomathematics and Bioinformatics Department, Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK.
[43] ERATO, Japan Science and Technology Agency, Okazaki 444-8585, Japan.
[44] Department of Integrative Biology, 3060 Valley Life Sciences Building, University of California, Berkeley, CA 94720, USA.
[45] Genome Project Solutions, 1024 Promenade Street, Hercules, CA 94547, USA.

JANUARY 2008

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

The genome of the moss *Physcomitrella patens* reveals evolutionary insights into the conquest of land by plants

Stefan A. Rensing[1], Daniel Lang[1], Andreas Zimmer[1], Astrid Terry[2], Asaf Salamov[3], Harris Shapiro[3], Tomoaki Nishiyama[4], Pierre-François Perroud[5], Erika Lindquist[3], Yasuko Kamisugi[6], Takako Tanahashi[7,33], Keiko Sakakibara[9], Tomomichi Fujita[10], Kazuko Oishi[8], Tadasu Shin-I[8], Yoko Kuroki[11], Atsushi Toyoda[11], Yutaka Suzuki[12], Shin-ichi Hashimoto[13], Kazuo Yamaguchi[4,14], Sumio Sugano[12], Yuji Kohara[8,15], Asao Fujiyama[11,16,17], Aldwin Anterola[19], Setsuyuki Aoki[20], Neil Ashton[18], W. Brad Barbazuk[21], Elizabeth Barker[18], Jeffrey Bennetzen[22], Robert Blankenship[5], Sung Hyun Cho[5], Susan Dutcher[23], Mark Estelle[24], Jeffrey A. Fawcett[25], Heidrun Gundlach[26], Kosuke Hanada[27], Alexander Heyl[28], Karen A. Hicks[29], Jon Hughes[30], Martin Lohr[31], Klaus Mayer[26], Alexander Melkozernov[32], Takashi Murata[7,33], David Nelson[34], Birgit Pils[35], Michael Prigge[24], Bernd Reiss[29], Tanya Renner[36], Stephane Rombauts[25], Paul Rushton[37], Anton Sanderfoot[38], Gabriele Schween[1], Shin-Han Shiu[27], Kurt Stueber[29], Frederica L. Theodoulou[39], Hank Tu[3], Yves Van de Peer[25], Paul J. Verrier[40], Elizabeth Waters[36], Andrew Wood[19], Lixing Yang[22], David Cove[5,6], Andrew C. Cuming[6], Mitsuyasu Hasebe[7,33,43], Susan Lucas[2], Brent D. Mishler[41], Ralf Reski[1], Igor Grigoriev[3], Ralph S. Quatrano[5*], Jeffrey L. Boore[3,41,42]

[*]Corresponding author and person to whom all inquiries should be directed:
Ralph S. Quatrano
Department of Biology, 1 Brookings Drive
Washington University
St. Louis, MO 63130-4899 USA
314.935.6850
rsq@wustl.edu

---

[1]Plant Biotechnology, Faculty of Biology, University of Freiburg, Schaenzlestr. 1, D-79104 Freiburg, Germany
[2]DOE Joint Genome Institute and Lawrence Livermore National Laboratory, 2800 Mitchell Drive, Walnut Creek, CA 94598 USA
[3]DOE Joint Genome Institute and Lawrence Berkeley National Laboratory, 2800 Mitchell Drive, Walnut Creek, CA 94598 USA

[4]Advanced Science Research Center, Kanazawa University, 13-1 Takara-machi Kanazawa, 920-0934 Japan
[5]Department of Biology, 1 Brookings Drive, Washington University, St. Louis, MO 63130-4899 USA
[6]Centre for Plant Sciences, University of Leeds, Leeds LS2 9JT, UK
[7]National Institute for Basic Biology, Okazaki 444-8585 Japan
[8]Genome Biology Laboratory, Center for Genetic Resource Information, National Institute of Genetics, Mishima 411-8540 Japan
[9]School of Biological Sciences, Monash University, Clayton Campus, Melbourne, Victoria 3800, Australia
[10]Department of Biological Sciences, Faculty of Science, Hokkaido University, Sapporo 060-0810, Japan
[11]RIKEN Genomic Sciences Center, Kanagawa 230-0045, Japan
[12]Laboratory of Functional Genomics, Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo 108-8639, Japan
[13]Department of Molecular Preventive Medicine, School of Medicine, The University of Tokyo, Tokyo 113-8654, Japan
[14]Division of Life Science, Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa 920-1192, Japan
[15]Department of Genetics, School of Life Science, The Graduate University for Advanced Studies, Mishima 411-8540 Japan
[16]National Institute of Informatics, Tokyo 101-8403, Japan
[17]Department of Informatics, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies, Tokyo 101-8403, Japan
[18]University of Regina, 3737 Wascana Parkway, Regina, SK S4V 1B7 Canada
[19]Department of Plant Biology, Southern Illinois University, Carbondale, IL 62901-6509 USA
[20]Division of Biological Informatics, Graduate School of Human Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
[21]Donald Danforth Plant Science Center, 975 North Warson Road, St. Louis, MO 63132 USA
[22]Department of Genetics, Davison Life Sciences Complex, University of Georgia, Athens, GA 30602-7223 USA
[23]Department of Genetics, 4444 Forest Park Ave., Washington University, St. Louis, MO 63108 USA
[24]Department of Biology, Indiana University, 1001 East Third Street, Bloomington, IN 47405-3700 USA
[25]VIB Department of Plant Systems Biology, Ghent University, Technologie Park 927, 9052 Gent, Belgium
[26]Munich Information Center for Protein Sequences, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany
[27]Department of Plant Biology, 166 Plant Biology Building, Michigan State University, East Lansing, MI 48824-1312 USA
[28]Free University, Institute for Biology, Applied Genetics Neubau, Albrecht-Thaer-Weg 6, 14195 Berlin Germany

[29]Max-Planck Institute of Plant Breeding Research, Carl-von-Linne-Weg 10, 50829 Cologne, Germany
[30]Justus Liebig University, Giessen Zeughaus, Rm. 341, Senckenbergstr. 3, D35390 Giessen, Germany
[31]Institute of General Botany, University of Mainz, Saarstr. 21, 55099 Mainz, GERMANY
[32]Department of Chemistry and Biochemistry, Arizona State University, Tempe, AZ 85287-1604 USA
[33]Department of Basic Biology, School of Life Science, The Graduate University for Advanced Studies, Okazaki 444-8585, Japan
[34]University of Tennessee-Memphis, 858 Madison Ave 101, I Molecular Science Building, Memphis, TN 38163 USA
[35]Department of Bioinformatics, Biozentrum, Am Hubland,Wüzburg University, 97074 Würzburg, Germany
[36]Biology Department, San Diego State University, North Life Sciences Room 102, 5500 Campanile Drive, San Diego, CA 92182-4614 USA
[37]Department of Biology, Gilmer Hall, 485 McCormick Road, University of Virginia, Charlottesville, VA 22903 USA
[38]Department of Plant Biology, University of Minnesota, 250 Biological Science Center, 1445 Gortner Ave., St. Paul, MN 55108 USA
[39]Crop Performance and Improvement Division, Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK
[40]Biomathematics and Bioinformatics Department, Rothamsted Research, Harpenden, AL5 2JQ, UK
[41]Department of Integrative Biology, 3060 Valley Life Sciences Building, University of California, Berkeley 94720 USA
[42]Genome Project Solutions, 1024 Promenade Street, Hercules, CA 94547 USA
[43]ERATO, Japan Science and Technology Agency, Okazaki 444-8585 Japan

ABSTRACT

We report the draft genome sequence of the model moss *Physcomitrella patens* and compare its features to those of unicellular aquatic algae and flowering plants, from which it is separated by more than 400 million years. This reveals genomic changes concomitant with the evolutionary movement to land, including a general increase in gene family complexity, loss of genes associated with aquatic environments (e.g. flagellar components for gametic motility), acquisition of genes for tolerating terrestrial stresses (e.g. variation in temperature and water availability), and the development of the auxin and abscisic acid signaling pathways for co-ordinating multicellular growth and dehydration response. The *Physcomitrella* genome provides a resource for phylogenetic inferences about gene function and for experimental analysis of plant processes through this plant's unique facility for reverse genetics.

1. INTRODUCTION

This reports the draft genome sequence of the moss *Physcomitrella patens*, the first bryophyte genome to be sequenced. Bryophytes, comprising hornworts, mosses and liverworts, are remnants of early diverging lineages of embryophytes and thus occupy an ideal phylogenetic position for reconstructing ancient evolutionary changes and illuminating one of the most important events in earth history - the conquest of land by plants (see Fig. 1). The terrestrial environment involves extreme variations in water availability and temperature, and increased exposure to radiation. Adaptation to it entailed dramatic changes in body plan (*1*), and modifications to cellular, physiological, and regulatory processes. Primary adaptations would have included enhanced osmoregulation and osmoprotection, desiccation and freezing tolerance, heat resistance, and synthesis and accumulation of protective "sunscreens" as well as enhanced DNA repair mechanisms. Fossil evidence suggests that early land plants were structurally similar to extant bryophytes (*2*); they probably had a dominant haploid phase and were dependent on water for sexual reproduction, having motile male gametes.

The genome sequence of *P. patens* allows us to reconstruct the concomitant events of genome evolution that occurred in the colonization of land, through

comparisons with the genome sequences of several angiosperms (*Arabidopsis thaliana*, *Oryza sativa*, and *Populus trichocarpa*), as well as aquatic single-celled green algae (*Ostreococcus tauri, Ostreococcus lucimarinus* and *Chlamydomonas reinhardtii*).

2. FEATURES OF THE WHOLE GENOME

2.1 General genome properties

The draft genome sequence of *P. patens* ssp. *patens* (strain Gransden 2004) was determined by whole-genome shotgun sequencing, assembling into 480 Mbp of scaffold sequence with a depth of ~8.6x[1]. EST coverage of the assembly is over 98%. The draft sequence contains 35,938 predicted and annotated *P. patens* gene models (Tables S1-5). Most predicted genes are supported by multiple types of evidence (Table S4) and 84% of the predicted proteins appear complete. About 20% of the analyzed genes show alternative splicing (Table S6), a similar frequency to *A. thaliana* and *O. sativa* (*3*).

2.2 Repetitive sequences and transposons

An *ab initio* approach detected 14,366 repetitive elements comprising 1,381 families (average member number 10 and average length 1,292 bp; Table S7). The largest repetitive sequence is from the "AT rich – low complexity" class (23% of the repetitive fraction) and 15 families account for over 84% of the repetitive fraction (Table S8).

Long terminal repeat retrotransposons (LTR-Rs) are generally the most abundant class of transposable elements, contributing substantially to flowering plant genome size. Of the 4,795 full-length LTR-Rs in *P. patens*, 46% are gypsy-like and 2% are copia-like. *P. patens* contains about three times more full-length LTR-Rs than *A. thaliana*, but about three times fewer than rice, with the density among the three genomes being lowest in moss (Fig. S1). Although about half of the *P. patens* genome consists of 157,127 LTR-Rs, only 3% still exist as intact full-length elements; the remainder are diverged and partial remnants often fragmented by mutual insertions (Fig. S2). Nested regions are common with 14% of LTR-Rs inserted into another LTR-R (Table S9). The genome also contains 895 solo LTR-Rs, probably a result of unequal crossing over or DNA repair.

[1] Version 1.1 of the *P. patens* genome assembly and annotation can be accessed through the JGI Genome Portal at http://www.jgi.doe.gov/Physcomitrella

Periodic retrotransposition activity peaks are discernible over the last 10 million years (MY) (Fig. 2). Only one full-length element is inserted within a gene, suggesting strong selection against transposon insertion into genes ($p < 0.001$).

Helitrons (rolling-circle transposons) are an ancient class of transposons, present in animals, fungi and plants. Different from all eukaryotic genomes sequenced so far, the *P. patens* genome contains only a single Helitron family (Table S10) with 19 members. High sequence similarity (96%) suggests that they have been active within the last 3 MY. Presumably, multiple Helitron families have evolved in all plant lineages, including *P. patens*, but we predict that a rapid process of DNA removal has excised all members that have not been active recently, a process which has been demonstrated in other plant genomes (*4*).

2.3 Gene and Genome Duplications

Gene and genome duplications are major driving forces of gene diversification and evolution (*5*). In *P. patens*, the Ks distribution plot (i.e. the frequency classes of synonymous substitutions) among paralogs shows a clear peak at around 0.5-0.9 (Fig. S3), suggesting that a large-scale duplication, possibly involving the whole genome, has occurred, confirming EST-based data (*6*). Additional evidence for a large-scale duplication comes from the identification of 77 non-overlapping duplicated segments containing at least five paralogous gene pairs. All duplicated segments have an average Ks of 0.5-0.7.

Tandemly arrayed genes (TAGs) can contribute significantly to genome size. However, only ~1% of the protein-encoding genes in *P. patens* occur in tandem array, in contrast to *A. thaliana* (~16%), *O. sativa* (~14%) and *P. trichocarpa* (11%) (*7-9*). The majority of *P. patens* TAG clusters comprise two genes that are not separated by an intervening gene (Fig. S4). Compared with non-TAG genes, genes in TAGs are significantly shorter ($p < 0.001$) in terms of gene, CDS and intron length while their G/C content is significantly higher (Table S11). Functional analysis of TAGs compared with paralogous non-TAG clusters reveals that photosynthesis proteins, particularly antenna

proteins, are significantly ($q < 0.05$) enriched among the TAGs (see section 3.6, St[2] 58 A/B). Other enriched categories are glyoxylate and dicarboxylate metabolism, carbon fixation and ribosome assembly (Fig. S5). Apparently, *P. patens* has increased the genetic playground for photosynthesis and connected carbon-based metabolism in its recent past.

Comparison of the insertion age (Ks) of *P. patens* TAGs with paralogs that were established during the large-scale genome duplication suggests that the majority of TAGs were established recently ($Ks < 0.1$). Strikingly, *P. patens* TAG partners tend to be located on opposite strands (64.4%, with 36.4% in head-to-head orientation and 28.0% in tail-to-tail orientation), while there is a tendency (68-88%) for TAGs to be located on the same DNA strand in *A. thaliana* and *O. sativa* (*8*) as well as in *C. elegans*, human, mouse and rat (*8, 10*). This seems to point at a higher frequency of TAG generation on the opposite strand, yet highly similar TAGs on the same strand may be underrepresented due to difficulties in assembling the whole genome shotgun data. TAGs that are located on opposite strands in *P. patens* also seem to be excluded more frequently, based on the observation that significantly fewer substitutions ($p < 0.001$) can be observed within them (average Ks=0.591) than in those that are located on the same strand (Ks=1.246). These differences in TAG organization might be connected to the exceptional reliance on sequence similarity for DNA repair observed in the moss (*11*, *12*).

2.4 Domain and gene family expansion patterns

The sizes of eukaryotic gene families differ mainly due to different rates of gene retention after duplication, and the gene content differences likely reflect species-specific adaptations. Overall, lineage-specific gains among domain families occurred at a rate approximately three times lower in *P. patens* compared to *A. thaliana* (Fig. 3A). Among families shared by both organisms there are more gene families with relatively small numbers of gains in moss (1-6) than in *A. thaliana* (Fig. 3B). Many gene families with significantly higher than average duplication rates also have elevated rates of gene loss (Fig. 3C).

[2] St = Supplementary tree; these can be accessed via http://www.cosmoss.org/bm/supplementary_trees/Rensing_et_al_2007/

Highly expanded gene families in *P. patens* are not necessarily highly expanded in *A. thaliana* (r^2=0.33, $p<2x10^{-16}$). Only 36 families with significantly higher than average gains are common to both *P. patens* and *A. thaliana* lineages, while 43 are significantly expanded specifically in moss (Fig. 3A). Examples of parallel expansion include genes encoding protein kinases, leucine-rich repeat-proteins, AP2 and Myb transcription factors. Transcription factor duplicates are retained in *P. patens* with a rate that is lower than in flowering plants, yet higher than in algae (*13*), e.g. the MADS-box and WRKY transcription factor families are of an intermediate size as compared to flowering plants and algae (Table S12, S13).

Families that expanded only in the *P. patens* lineage include histidine kinases and response regulators. Both families are parts of two component signaling networks important in plants, fungi, and bacteria. These two families are much larger in *P. patens* than those found in sequenced angiosperm genomes, suggesting a more elaborate use of two component systems in moss.

The *P. patens* genome contains genes for each of the core groups of small GTPases (G-proteins) (Fig. S6 A,B), consistent with increased complexity of vesicle trafficking machinery, not present in green algae, suggesting that such complexity was already present in the last common ancestor (LCA) of land plants. *P. patens* also has a large ATP Binding Cassette (ABC) superfamily (121 members; Table S14/S15; St 29_ABDI/C/F/G, 9, 57, 110-113), similar in size to that in *A. thaliana* (130) and *O. sativa* (129) but larger than that of the unicellular alga, *O. tauri* (ca. 50), and twice that of humans and *D. melanogaster* (48 and 56 respectively). In flowering plants, most ABC-containing proteins are membrane-bound transporters for lipids, hormones, secondary metabolites, metals and xenobiotics and control certain ion channels. The sessile habit and metabolic diversity of land plants appears to require a large repertoire of ABC proteins.

3. ADAPTATIONS TO THE TERRESTRIAL ENVIRONMENT

3.1 Loss of motile gametes

Many algae and bryophytes share the ancestral trait of having flagellated male gametes, although this trait has been lost in flowering plants (*14*). Consequently, proteins for delta and epsilon tubulins, required for forming the basal bodies of flagella (*15*), are found in *P. patens* (St 93, 94). Genes were also found for most proteins of the inner, but not the outer dynein arms (St 91, 92), which are the motors for the motility of flagella. This observation suggests a lack of outer arms in flagella, as has been shown to be the case for other land plants *(14)*. Cytoplasmic dynein genes and their regulatory dynactin complex genes are absent, suggesting that the dynein-mediated transport system was probably lost in the LCA of *P. patens* and flowering plants.

3.2 Desiccation tolerance

Desiccation tolerance (DT) is widespread in reproductive structures of vascular plants but vegetative DT is rare except among bryophytes (*16*). Evolution of this trait was important in facilitating the colonization of the land, but was lost subsequently in vascular plants. DT in seeds is dependent on the phytohormone abscisic acid (ABA) to induce expression of seed-specific genes such as late embryogenesis abundant proteins (LEAs), a group of proteins that accumulate during desiccation. *P. patens* is highly dehydration tolerant (*17*), and contains orthologs of LEA genes and other genes expressed during the DT response in a poikilohydric moss (*18*) and in flowering plants (*19*).

ABA signaling also operates in the *P. patens* drought response (*19*). The genome contains homologs of the *A. thaliana* ABA receptors, one of which appears to have been specialized for a role in seed development (*20*), and the transcription factor ABI5 implicating it in the regulation of ABA-mediated gene expression. Particularly interesting is *ABI3*, the seed-specific transcription factor of the B3 family (St 132), which when mutated results in the loss of desiccation tolerance in seeds (*21*). The *P. patens* genome contains four *ABI3*-like genes, one of which (*PpABI3A)* functions to potentiate ABA responses in *P. patens* and partially complements the *A. thaliana abi3-6* mutant (*22*).

Finding of these genes in *P. patens* suggests that desiccation tolerance gene networks likely originated in the LCA of land plants.

3.3 Metabolic pathways

Cytochrome P450 enzymes that incorporate oxygen into small lipophilic compounds are represented by 250-350 members in genomes of flowering plants, 71 genes in *P. patens* and 39 in *C. reinhardtii*. Specific examples of P450s lacking in *P. patens* are related to the absence and regulation of key molecules in flowering plants. One P450 required for the synthesis of gibberellic acid synthesis (CYP88) is absent, as is the enzyme needed to make S-lignols (CYP84) required for the accumulation of lignin. The CYP86 family includes fatty acid omega-hydroxylases involved in the formation of cutin, which prevents dehydration of plant tissues. The presence of CYP86 in moss but not in green algae suggests that cutin may have evolved in the ancestral land plants as an innovative mechanism to survive a terrestrial habitat.

Most enzymatic steps in the carotenoid and chlorophyll biosynthetic pathways are more complex in terms of paralog frequencies in *P. patens* than in *A. thaliana* and *C. reinhardtii* (Fig. 4, Table S16), which is consistent with previous interpretations that the moss genome encodes seemingly redundant metabolic pathways and contains an elaborate network of genes for crucial functions like phototoxic stress tolerance (*6*). Unlike Light Harvesting Complex (LHC) proteins, most genes (79%) of the carotenoid and chlorophyll metabolic pathways are not TAGs and were acquired during the whole genome duplication, i.e. since the divergence from the lineage leading to flowering plants (*6*).

One striking exception is the genes involved at the branching point of siroheme and heme/chlorophyll formation (Fig. 4, Table S16). UROS and UMT are encoded by single copy genes, while conserved ancient paralogs encode UROD. These paralogs had already been acquired before the split of green algae and land plants (St 76) and probably are functionally divergent (*23*, *24*). Interestingly, both the UROD3 (St 76) and CPX2 (St 59) subfamilies are present in algae and *P. patens*, but have been lost in flowering plants.

3.4 Signaling pathways

The classically studied phytohormones and light receptors for morphogenesis found in flowering plants are absent in the unicellular algae but present in *P. patens*, e.g. genes for all four classes of cytokinin signaling pathways found in flowering plants. These include at least three cytokinin receptors, two of which have been confirmed by EST evidence and make *P. patens* the earliest diverging species that contains genes for all members of the cytokinin signal transduction pathway known today.

Although specific G-protein-based signaling pathways are seen in animal cells, only the G-protein complex associated with ABA signaling is present in vascular plants. The *P. patens* genome contains the gene for G-protein-coupled ABA receptor (*20*) and a single alpha- and two beta-subunits. *P. patens* also has the Mg-chelatase H subunit, a receptor for ABA signaling during seed development (*25*). Although the chelatase is found in *C. reinhardtii* and *O. lucimarinus*, the G-protein receptor is absent.

Ten gene families implicated in auxin homeostasis and signaling have been analyzed (Table S17, St 25, 33_A/B, 41, 45, 71, 73_7, 77, 85, 88, 89). The *C. reinhardtii*, *O. lucimarinus* and *O. tauri* genomes do not encode any of these, while the *P. patens* genome encodes members of each family (although based on the phylogenies of the GH3 and ILL proteins, St 71 and 85, moss might not conjugate IAA to alanine, leucine, aspartic acid, or glutamic acid consistent with empirical data (*26*). Angiosperms dedicate a larger proportion of their genomes to auxin signaling; only one (AUX1/LAX; St 41) of the ten families has as many members as angiosperm genomes. Based on the analysis of *A. thaliana* and our phylogenetic analyses, the auxin signaling pathway has undergone significant functional diversification within vascular plants since they diverged from bryophytes.

Although no ethylene responses have been noted, the *P. patens* genome encodes six ETR/ERS/EIN-like ethylene receptor HPT-type receptors and evidence for ethylene binding activity (*3*). Two putative ACC synthases, catalyzing a critical step in ethylene biosynthesis, were also found. Two transcription factors with strong similarity to the EIN3 ethylene signaling family are also apparent as are six N-RAMP-type channel proteins, one or more of which might be involved in ethylene signaling, similar to EIN2 in *A. thaliana*.

In vascular plants, photomorphogenic signals are perceived by three sensory photoreceptor families: phytochrome, cryptochrome and phototropin. *P. patens* possesses four canonical phototropins, evolutionary ancient UV/A-blue light photoreceptors that help optimize photosynthesis in shade while avoiding damage in sunlight (*27*). *P. patens* has seven phytochromes, more than any organism reported to date. Neither PIF3 (*28*) nor the PKS family (*29*) of phytochrome-interacting proteins are present in *P. patens*, but several members of the NDPK1, 2 and 3 groups, implicated in phytochrome signaling in vascular plants, are partly present.

UV/A-blue light sensitive cryptochromes and the related photolyase DNA-repair family are widespread in nature. Accordingly, in addition to two HY4-like cryptochrome photomorphogenic photoreceptors (*30*), *P. patens* has one UVR3-like 6-4 photolyase, one ssDNA CRY3-like and several dsDNA PHR-like cyclobutane pyrimidine dimer photolyases that restore nucleotide structure with the help of UV/A-blue light following UV/B-induced damage.

Circadian oscillators are found in most organisms, and genes related to TOC1/PRR pseudo-response regulators (St 69) and LHY/CCA1 single-myb domain transcription factors (St 30) of flowering plant clocks are present in both *P. patens* and *O. tauri and O. lucimarinus* (*31*). In terms of interpretation of seasonal cues, *P. patens* has sequences related to the key photoperiodic regulators CONSTANS (St 69, (*32*, *33*)) and FT (St 74), as well as the CONSTANS-regulating cycling DOF factors (St 19), but not their downstream targets. Thus, these signaling pathways appear to have an ancient origin, with the evolution of specific downstream targets occurring later, after the divergence from the LCA of land plants.

3.5 Protective Proteins

Adaptation to land also required the evolution of proteins protecting against various stresses such as variation in temperature, light and water availability. One example of this is the expansion of the HSP70/DnaK family to nine cytosolic members in *P. patens* (St 24) whereas all algal genomes sequenced to date encode one single cytosolic HSP70 (*34*).

The complement of the LHC genes is significantly expanded in *P. patens* when compared to both algae and vascular plants (St 58_A, Table S18A). While several LHC homologs were already present in the LCA of all land plants, more have been retained after the whole genome duplication in *P. patens* and more of these genes are present in TAGs than in *A. thaliana* (Table S19). Redundancy and expansion of these abundantly expressed proteins probably contributes to a robustness of the photosynthetic antenna, i.e. the capacity to deal with high light intensities. The photoprotective early light-induced proteins (ELIPs) expanded extensively in *P. patens* (St 58_B, Table S18B). Numerous ELIP-like proteins with supposedly free radical scavenging activity may reflect adaptation to de-/rehydration cycles and associated avoidance of photo-oxidative damage.

3.6 DNA repair

DNA damage repair helps maintain genomic integrity. Double-strand breaks (DSB) can be repaired by non-homologous end-joining (NHEJ) but are more precisely repaired using a second copy of the sequence. The introduction of linear DNA into a cell mimics DNA damage, and mosses, uniquely among plants, but like yeast, show a strong preference for the use of a homologous sequence for the incorporation of linear DNA into the genome.

Cell-cycle control is tightly connected to DNA-damage repair (*35*). Proteins known to be involved in these processes in both vertebrates and *A. thaliana* are ATM, ATR, CHK1, CHK2, PARP1, BRCA1, BRCA2, and BARD1. While *P. patens* encodes the first four of these, there are no homologs found of BRCA1, BRCA2 and BARD1. The RAD51 paralogs (RAD51A, RAD51B, RAD51C, RAD51D, XRCC2, and XRCC3) are important for repair using sequence homology in vertebrates and in *A. thaliana*; *P. patens* encodes all but XRCC3. However, while *A. thaliana* encodes one RAD51A, *P. patens* encodes two (*36*). Other genes involved in DSB repair, chromatin remodeling, and processing of recombination intermediates known from *A. thaliana* (INO80, RAD54, MRE11, RAD50, NBS1, RecQ helicases (WRN, BLM), MUS81) are also present in *P. patens*. Additionally, both plant species, but not metazoans, encode SRS2, while *P. patens*, like the other plants, lack RAD52. In *A. thaliana* and in yeast the KU70/KU80 complex, DNA Ligase IV and XRCC4 contribute to NHEJ. These genes are encoded by

the *P. patens* genome as well. In addition, both plant species, but not yeast (*37*), encode DNA-PKcs.

In our phylogenetic analyses, *P. patens* homologs of RAD54B as well as CENTRINS and CHD7 cluster with algal and metazoan homologs, whereas flowering plant homologs are not included in these clusters (St 12_2, 28_2, 28_7). While RAD51 and RAD54 interact in chromatin remodeling in humans (*38*), CENTRINS are important for genome stability in *C. reinhardtii* (*39*) and in nucleotide excision and DSB repair in *A. thaliana* (*40*). CHD7 is a chromodomain DNA helicase, important for chromatin structure. Its mutation causes severe developmental aberrations in mammals (*41*).

DNA damage is repaired by multisubunit macromolecular complexes of dynamic composition and conformation (*42*). The special features of the *P. patens* genome (no BRCA1, BRCA2, and BARD1, duplicated RAD51, and phylogenetically conserved RAD54B, CENTRINS and CHD7) may well reflect the specific needs of a haploid genome for genome integrity surveillance and account for the efficiency of homology-dependent DSB repair in the *P. patens* genome.

4. CONCLUSIONS FOR LAND PLANT EVOLUTION

Physcomitrella patens occupies a position on the evolutionary tree that, through comparisons with aquatic algae and vascular plants, allows detailed reconstruction of the evolutionary changes in genomes that are concomitant to the conquest of land. From this, we conclude that the LCA of all land plants (1) lost genes associated with aquatic environments (e.g. flagellar components for gametic motility), (2) lost dynein-mediated transport, (3) gained signaling capacities, such as those for auxin, ABA, cytokinin, and more complex photoreception, (4) gained tolerance for abiotic stresses, such as drought, radiation, and extremes of temperature, (5) gained more elaborate transport capabilities, and (6) had an overall increase in gene family complexity. Some of these events may have been enabled by the opportunities for evolutionary novelty created by one or more duplications of the whole genome.

These comparisons also enable a reconstruction of the genomic events that occurred after the split of vascular plants and the mosses. For example, the former acquired even more elaborate signaling (e.g. through gibberellic acid (GA), jasmonic acid (JA), ethylene and brassinosteroids) but lost vegetative dehydration tolerance and motile gametes, whereas the latter gained an elaborate use of two component systems, efficient HR-based DNA repair, adaptation to shade and de-/rehydration cycles, as well as a redundant and versatile metabolism. The *P. patens* genome sequence provides a resource for the study of both gene function and evolutionary reconstruction, which is sure to reveal even more interesting details.

Figure Legends

Figure 1: Land plant evolution

Bryophytes comprise three separate lineages which, together with the vascular plants (including the flowering plants), make up the embryophytes (land plants). These four lineages, remnants of the initial radiation of land plants in the Silurian, began to diverge from each other about 450 MYA.

Figure 2: Periodic cycles of LTR retrotransposon activity

P. patens underwent periodic cycles of LTR-R amplifications. The most recent activity peaks at an estimated 1 to 1.5 MYA, preceded by invasion events around 3, 4 and 5.5 MYA. Gypsy-like elements are younger (average 3.2, median 3.0) than copia-like elements (average 3.9, median 3.6), coinciding with a 7-fold higher full length copy number. The gradual decrease between 5 to 12 MYA probably reflects element deterioration leading to loss of ability to detect these elements. Numbers found of each element are shown in parentheses.

Figure 3: Domain family expansion patterns in *P. patens*

(A) Gain is defined as the presence of paralogous gene copies uniquely arising in one lineage. Large domain families are labeled based on their Pfam domain names. (B) Relationships between gains per family and the number of families in *A. thaliana* and *P. patens*. (C) Relationships between gain vs. loss in *P. patens* domain families.

Figure 4: Paralog frequencies in the biosynthetic pathways of chlorophylls and carotenoids in *P. patens, A. thaliana and C. reinhardtii*

Denoted are products (in bold) that accumulate to significant amounts, major intermediates and known enzymes of both pathways (for full names of enzymes, see Table S16). Major pathways are indicated by black arrows; branch-points leading to the formation of related compounds (italicized) are indicated by grey arrows. For each reaction, colored squares symbolize the number of (iso-) enzymes in *P. patens* (red), *A. thaliana* (yellow) and *C. reinhardtii* (green). Enzymes for which *P. patens* has more

paralogs than *A. thaliana* and *C. reinhardtii* are boxed in red, those encoded by unique genes in the moss are boxed in blue.

FIGURE 1

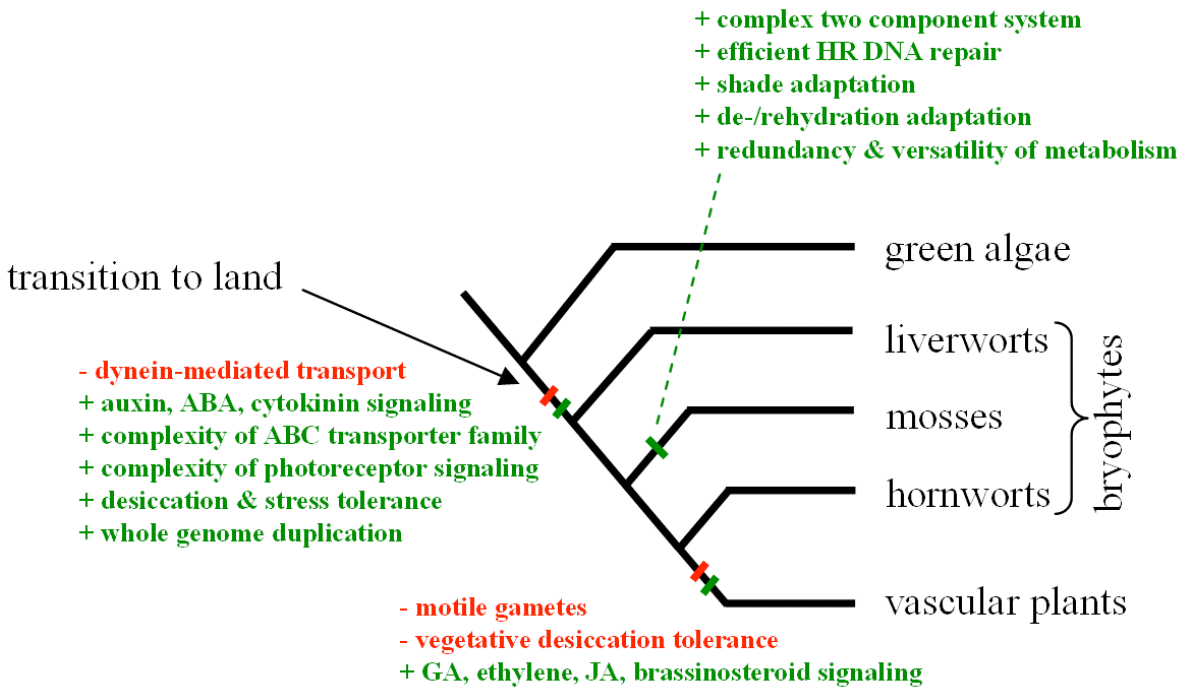

+ complex two component system
+ efficient HR DNA repair
+ shade adaptation
+ de-/rehydration adaptation
+ redundancy & versatility of metabolism
transition to land
- dynein-mediated transport
+ auxin, ABA, cytokinin signaling
+ complexity of ABC transporter family
+ complexity of photoreceptor signaling
+ desiccation & stress tolerance
+ whole genome duplication
green algae
liverworts
mosses
hornworts
bryophytes
vascular plants
- motile gametes
- vegetative desiccation tolerance
+ GA, ethylene, JA, brassinosteroid signaling

FIGURE 2

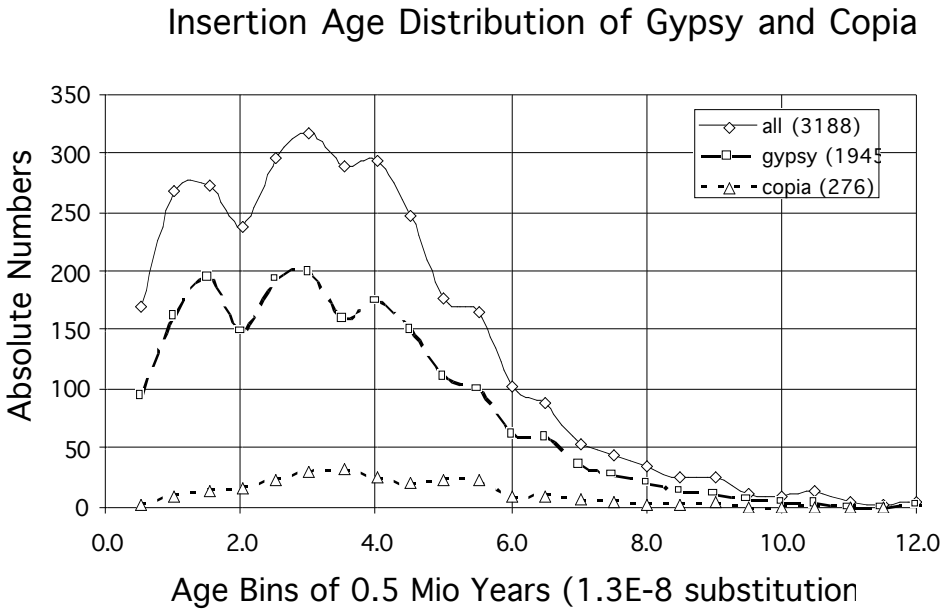

Insertion Age Distribution of Gypsy and Copia
Absolute Numbers
350
300
250
200
150
100
50
0
0.0
2.0
4.0
6.0
8.0
10.0
12.0
Age Bins of 0.5 Mio Years (1.3E-8 substitution
all (3188)
gypsy (1945
copia (276)

FIGURE 3

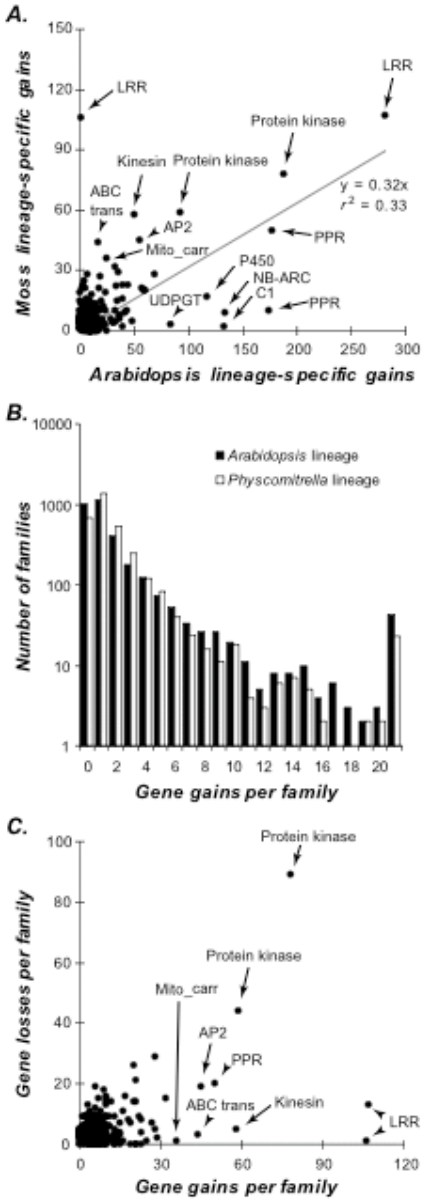
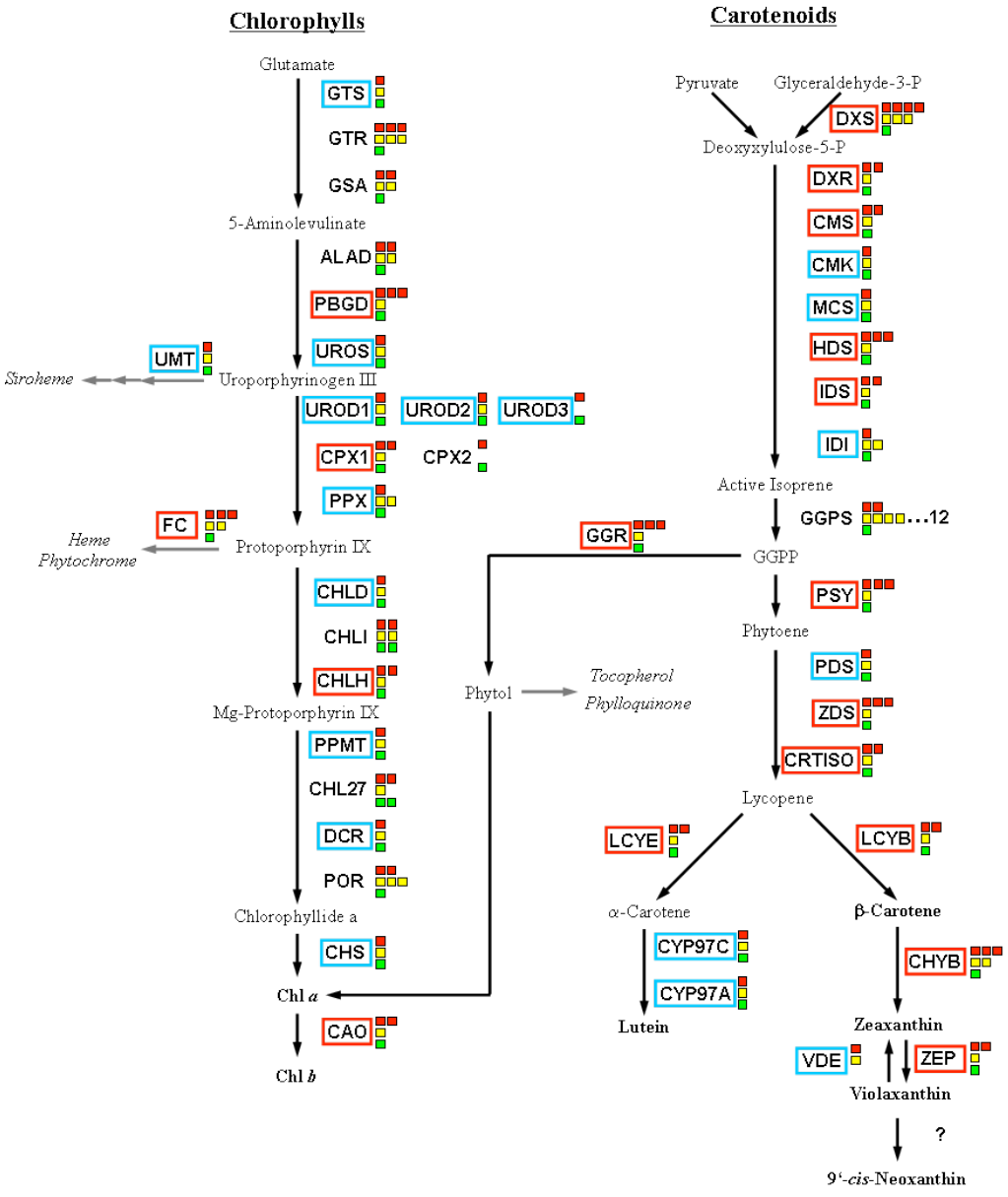

A.
Moss lineage-specific gains
Arabidopsis lineage-specific gains
LRR
LRR
Protein kinase
Kinesin
Protein kinase
ABC trans
AP2
Mito_carr
PPR
P450
NB-ARC
C1
UDPGT
PPR
y = 0.32x
r2 = 0.33
B.
Number of families
Gene gains per family
Arabidopsis lineage
Physcomitrella lineage
C.
Gene losses per family
Gene gains per family
Protein kinase
Protein kinase
Mito_carr
AP2
PPR
ABC trans
Kinesin
LRR

FIGURE 4


Chlorophylls
Carotenoids
Glutamate
GTS
GTR
GSA
5-Aminolevulinate
ALAD
PBGD
UROS
UMT
Siroheme
Uroporphyrinogen III
UROD1
UROD2
UROD3
CPX1
CPX2
PPX
FC
Heme
Phytochrome
Protoporphyrin IX
CHLD
CHLI
CHLH
Mg-Protoporphyrin IX
PPMT
CHL27
DCR
POR
Chlorophyllide a
CHS
Chl a
CAO
Chl b
Pyruvate
Glyceraldehyde-3-P
DXS
Deoxyxylulose-5-P
DXR
CMS
CMK
MCS
HDS
IDS
IDI
Active Isoprene
GGPS ...12
GGR
GGPP
PSY
Phytoene
PDS
ZDS
CRTISO
Lycopene
Phytol
Tocopherol
Phylloquinone
LCYE
LCYB
α-Carotene
β-Carotene
CYP97C
CYP97A
Lutein
CHYB
Zeaxanthin
VDE
ZEP
Violaxanthin
?
9'-cis-Neoxanthin

References

1. S. Floyd, J. Bowman, *Int J Plant Sci* **168**, 1 (2007).
2. P. Kenrick, P. Crane, *Nature* **389**, 33 (1997).
3. B. B. Wang, V. Brendel, *Proc Natl Acad Sci* **103**, 7175 (2006).
4. C. Vitte, J. L. Bennetzen, *Proc Natl Acad Sci* **103**, 17638 (2006).
5. M. Lynch, J. S. Conery, *Science* **290**, 1151 (2000).
6. S. A. Rensing *et al.*, *BMC Evol Biol* **7**, 130 (2007).
7. International Rice Genome Sequencing Project, *Nature* **436**, 793 (2005).
8. C. Rizzon, L. Ponger, B. S. Gaut, *PLoS Comput Biol* **2**, e115 (2006).
9. G. A. Tuskan *et al.*, *Science* **313**, 1596 (2006).
10. C. Semple, K. H. Wolfe, *J Mol Evol* **48**, 555 (1999).
11. H. Puchta, *J Exp Bot* **56**, 1 (2005).
12. Y. Kamisugi *et al.*, *Nucleic Acids Res* **34**, 6205 (2006).
13. S. Richardt, D. Lang, W. Frank, R. Reski, S. A. Rensing, *Plant Physiol* **143**, 1452 (2007).
14. J. Hyams, C. Campbell, *Cell Biol Int Rep* **9**, 841 (1985).
15. S. Dutcher, *Curr Opin Microbiol* **6**, 634 (2003).
16. M. Oliver, J. Velten, B. Mishler, *Integrative and Comparative Biology* **45**, 788 (2005).
17. W. Frank, D. Ratnadewi, R. Reski, *Planta* **220**, 384 (2005).
18. L. Saavedra *et al.*, *Plant J* **45**, 237 (2006).
19. A. C. Cuming, S. H. Cho, Y. Kamisugi, H. Graham, R. S. Quatrano, *New Phytol* doi: 10.1111/j.1469-8137.2007.02187. (2007).
20. X. Liu *et al.*, *Science* **315**, 1712 (2007).
21. J. Giraudat *et al.*, *Plant Cell* **4**, 1251 (1992).
22. H. Marella, Y. Sakata, R. S. Quatrano, *Plant J* **46**, 1032 (2006).
23. G. Hu, N. Yalpani, S. P. Briggs, G. S. Gurmukh S. Johal, *Plant Cell* **10**, 1095 (1998).
24. H. P. Mock, B. Grimm, *Plant Physiol* **113**, 1101 (1997).
25. Y.-Y. Shen *et al.*, *Nature* **443**, 823 (2006).
26. A. E. Sztein, J. D. Cohen, I. G. de la Fuente, T. J. Cooke, *American Journal of Botany* **86**, 1544 (1999).
27. M. Kasahara, T. Kagawa, S. Yoshikatsu, K. Tomohiro, M. Wada, *Plant Physiol* **135**, 1 (2004).
28. M. Ni, J. Tepperman, P. Quail, *Cell* **95**, 657 (1998).
29. C. Fankhauser, J. Chory, *Plant Physiol* **124**, 39 (2000).
30. T. Imaizumi, A. Kadota, M. Hasebe, M. Wada, *Plant Cell* **14**, 373 (2002).
31. F.-Y. Bouget, F. Corellou, M. Moulager, C. Schwartz, L. Garnier, paper presented at the FESPB, France 2006.
32. M. Shimizu, K. Ichikawa, S. Aoki, *Biochem Biophys Res Commun* **324**, 1296 (2004).
33. O. Zobell, G. Coupland, B. Reiss, *Plant Biol* **7**, 266 (2005).
34. W. Wang, B. Vinocur, O. Shoseyov, A. Altman, *Trends in Plant Sciences* **9**, 244 (2004).
35. P. Sung, H. Klein, *Nat Rev Mol Cell Biol* **7**, 739 (2006).

36. U. Markmann-Mulisch *et al.*, *Proc Natl Acad Sci* **99**, 2959 (2002).
37. J. Daley, P. Palmbos, D. Wu, T. Wilson, *Ann Rev Genetics* **39**, 431 (2005).
38. Y. Zhang *et al.*, *Nat Structure of Mol Biol* **14**, 639 (2007).
39. I. Zamora, W. F. Marshall, *BMC Biol* **3**, 15 (2005).
40. L. Liang, S. Flury, V. Kalck, B. Hohn, J. Molinier, *Plant Mol Biol* **61**, 345 (2006).
41. E. A. Hurd *et al.*, *Mammal Genome* **18**, 94 (2007).
42. O. Llorca, *Curr Opin Struct Biol* **17**, 215 (2007).
43. We thank Kemin Zhou and Samuel Pitluck at JGI for GenBank submissions and Gregory Werner and his group at JGI for support of gene annotation tools. Discussions with Drs. Mel Oliver and Kirsten Fisher on desiccation tolerance are greatly appreciated. Ms. Nancy Lyons efficiently handled many of the administrative tasks throughout the project and the detailed preparation of the final manuscript. Part of this work was funded by German National Science Foundation (DFG) grant RE 837/10 to R.R. and by grants to R.Q. from the U.S. National Science Foundation (IBN 0112461 & 0425749-1) and Washington University. This work was also supported by Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Culture, Sports, Science and Technology of Japan (to T.T., T.F., Y.Kuroki, A.T., Y.S., S.H., K.Y., S.S., Y. Kohara, A.F., T.M., T. N., and M.H). This work was performed under the auspices of the U.S. Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under Contract No. W-7405-ENG-36.
44. **Supporting Online Material**
www.sciencemag.org
Materials, Methods and Analysis, Figs. S1 – S8, Tables S1 – S23