every change in surface climate at our site has an accompanying change in deep circulation (and vice versa)—emphasizing even more strongly the uniqueness of the coeval extremes in climate and deep circulation that follow the Lake Agassiz flood outburst.

**Wider climatic implications.** Despite the complex nature of the climate record, the fact that large-scale changes in deep-ocean circulation are accomplished just as quickly in the natural world as is predicted in computer simulations [e.g., (17)] confirms that they were sufficiently rapid to be a plausible mechanism for driving the similarly abrupt changes seen in paleoclimatic records. This observation helps to define the minimum time scale for altering ocean circulation and demonstrates that it is rapid enough to be relevant for human societies. The fact that these past ocean circulation changes are associated with the largest climate variations in the past underscores the importance of understanding the minimum freshwater forcing capable of affecting such large circulation changes—particularly given the concern about the impact of future warming on the Greenland ice sheet.

**References and Notes**


The **Physcomitrella Genome Reveals Evolutionary Insights into the Conquest of Land by Plants**


Hence, we report the draft genome sequence of the moss *Physcomitrella patens*, the first bryophyte genome to be sequenced. The bryophytes (land plants) began to diverge about 450 million years ago (Ma). 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 (Fig. 1). The terrestrial environment...
involve variations in water availability and temperature, as well as increased exposure to radiation. Adaptation entailed dramatic changes in body plan (1) and modifications to cellular, physiological, and regulatory processes. Primary adaptations included enhanced osmoregulation and osmoprotection, desiccation and freezing tolerance, heat resistance, synthesis and accumulation of protective “suncreens,” and 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 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 unicellular green algae (*Ostreococcus tauri*, *Ostreococcus lucimarinus*, and *Chlamydomonas reinhardtii*).

**Features of the Whole Genome**

**General genome properties.** The draft genome sequence of *P. patens* (strain Gransden 2004) was determined by whole-genome shotgun sequencing, assembling into 480 mega-base pairs of scaffold sequence with a depth of ~8.6× (3, 4); expressed sequence tag (EST) coverage of the assembly is over 98%. The sequence contains 35,938 predicted and annotated *P. patens* gene models (tables S1 to S5). 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 frequency similar to that of *A. thaliana* and *O. sativa* (5).

Repetitive sequences and transposons. An ab initio approach detected 14,366 repetitive elements comprising 1381 families [average member number 10, length 1292 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 (6). Of the 4795 full-length LTR-Rs in *P. patens*, 46% are gypsy-like and 2% are copia-like. *P. patens* contains about three times as many full-length LTR-Rs as *A. thaliana*, but about one-third as many as *O. sativa*. The density among the three genomes is lowest in *P. patens* (fig. S1). Although about half of the *P. patens* genome consists of 157,127 LTR-Rs, only 3% exist as intact full-length elements. The remainder is made of 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 as a result of unequal crossing-over or DNA repair. Periodic retrotransposition activity peaks are discernible over the past 1 million years (My) (fig. 2). Only one full-length element is inserted within a gene, which suggests 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 (7). 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 past 3 My. Presumably, multiple Helitron families 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 that has been demonstrated in other plant genomes (6).

**Gene and genome duplications.** Gene and genome duplications are major driving forces of gene diversification and evolution (8). In *P. patens*, the *K*$_0$ distribution plot (i.e., the frequency classes of synonymous substitutions) among paralogs shows a clear peak at around 0.5 to 0.9 (fig. S3), which suggests that a large-scale duplication, possibly involving the whole genome, has occurred. The present of this peak confirms EST-based data (9). Additional evidence for a large-scale duplication comes from the identification of 77 nonoverlapping duplicated segments containing at least five paralogous gene pairs. All duplicated segments have an average *K*$_0$ of 0.5 to 0.7. Tandemly arrayed genes (TAGs) can contribute substantially 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%) (10–12). The majority of *P. patens* TAG clusters are made up of 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 coding sequence (CDS), and intron length, whereas their GC 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 (*P < 0.05*) enriched among the TAGs [section 3.6, St 58 A/B (13)]. 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 related carbon-based metabolism in its recent past.

Comparison of the *K*$_0$ of *P. patens* TAGs with paralogs that were established during the large-scale genome duplication (*K* ~0.5 to 0.9) suggested...
suggests that most TAGs were established recently ($K_s < 0.1$). It is noteworthy that $P. patens$ 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), whereas there is a tendency (68 to 88%) for TAGs to be located on the same DNA strand in $A. thaliana$, $O. sativa$ (II), Caenorhabditis elegans, Homo sapiens, Mus musculus, and Rattus norvegicus (II, 14). Significantly fewer substitutions ($P < 0.001$) are observed within them (average $K_s = 0.59$) than in those that are located on the same strand ($K_s = 1.25$). Homologous recombination between TAGs on the same strand may have resulted in loss of such TAGs, whereas gene conversion associated with homologous recombination of TAGs on opposite strands may have resulted in reduction of sequence divergence ($K_s$) between those. These differences in TAG organization might be connected to the exceptional reliance on sequence similarity for DNA repair observed in $P. patens$ ($15, 16$). Alternatively, the generation and exclusion rate of TAGs on the opposite strand might have been higher than for TAGs on the same strand in the ancestor of $P. patens$.

**Gene and domain family expansion patterns.** Eukaryotic gene family sizes differ mainly because of different rates of gene duplication and retention, and gene content differences may reflect species-specific adaptations. Overall, lineage-specific gains among domain families occurred at a lower rate (by a factor of about three) in the $P. patens$ compared with the $A. thaliana$ lineage (Fig. 3A). Similarly, in comparisons with the $O. sativa$ and $P. trichocarpa$ lineages, gene gain rates in the $P. patens$ lineage are substantially lower. Among gene families shared by both $P. patens$ and $A. thaliana$, there are consistently fewer families with relatively large gains (≥6) in the $P. patens$ lineage (Fig. 3B), which indicates that the gain rate differences noted in Fig. 3A are mainly due to higher retention rates of large families in the $A. thaliana$ lineage. In addition, many $P. patens$ gene families with higher-than-average gain rates in general also have elevated rates of gene loss (Fig. 3C).

Highly expanded gene families in the $P. patens$ lineage are not necessarily highly expanded in the $A. thaliana$ lineage ($r^2 = 0.33$, $P < 2 \times 10^{-16}$). Only 36 families with significantly higher-than-average gains are common to both the $P. patens$ and $A. thaliana$ lineages, whereas 43 are significantly expanded only in $P. patens$ (Fig. 3A). Examples of parallel expansion include genes encoding protein kinases and leucine-rich-repeat proteins, as well as Apetala 2 (AP2) and Myb transcription factors. Transcription factor duplicates are retained in the $P. patens$ lineage with a rate lower than those in the flowering plant lineage, yet higher than in algae (17); for example, the MADS-box and WRKY transcription factor families are intermediate in size compared with flowering plants and algae (table S12 and S13).

Families that significantly 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; their increased size suggests a more elaborate use of two-component systems in $P. patens$.

**Fig. 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) (38). These four lineages, remnants of the initial radiation of land plants in the Silurian, began to diverge from each other about 450 Ma.

**Fig. 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 Ma, preceded by invasion events around 3, 4, and 5.5 Ma. Gypsy-like elements are younger (average 3.2, median 3.0) than copia-like elements (average 3.9, median 3.6), coinciding with an increased full-length copy number by a factor of seven. The gradual decrease between 5 and 12 Ma probably reflects element deterioration leading to loss of ability to detect these elements. Numbers found of each element are shown in parentheses.

**Fig. 3.** Domain family expansion patterns in $P. patens$. (A) Gain is defined as the presence of paralogous gene copies uniquely arising in one lineage based on the results of reconciliation between gene family and species trees. Large gene families are labeled on the basis of the predominant Pfam domain names. Some domain names occur more than once since they are the predominant domains in multiple gene families. (B) Relations between lineage-specific gains per family and the number of families in the $A. thaliana$ and $P. patens$ lineages. (C) The relation between gain and loss among $P. patens$ gene families.
the last common ancestor of land plants. *P. patens* also has a large adenosine triphosphate binding cassette (ABC) superfamily [121 members; (tables S14 and S15); St 29] A. thaliana (130) and *O. sativa* (129), but larger than *O. tauri* (~50) and twice that of humans and *Drosophila 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.

**Adaptations to the Terrestrial Environment**

**Desiccation tolerance.** Desiccation tolerance (DT) is widespread in reproductive structures of vascular plants, but vegetative DT is rare, except among bryophytes (18). 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 (19) and contains orthologs of LEA genes and other genes expressed during the DT response in the poikilohydric moss *Tortula ruralis* (20) and in flowering plants (21).

ABA signaling also operates in the *P. patens* drought response (21). The genome contains putative homologs of the *A. thaliana* ABA receptors, one of which appears to have been specialized for a role in seed development, and the transcription factor ABI5, which implicates 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 (22). The *P. patens* genome contains four ABI3-like genes, one of which (PpAB3A) functions to potentiate ABA responses in *P. patens* and partially complements the *A. thaliana* abi3-6 mutant (23).

Finding these genes in *P. patens* and similar sequences in liverworts (*Riccia fluitans* and *Marchantia polymorpha*) suggests that desiccation tolerance gene networks likely originated in the last common ancestor of extant land plants.

**Metabolic pathways.** Cytochrome P450 enzymes that incorporate oxygen into small lipophilic compounds are represented by 250 to 350 members in genomes of flowering plants, 71 genes in *P. patens*, and 40 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 (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 *P. patens*, 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 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 and table S16). This is consistent with previous interpretations that the *P. patens* genome encodes seemingly redundant metabolic pathways and contains a network of genes for functions like phototoxic stress tolerance (9). 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 (9).

One striking exception is the genes involved at the branching point of siroheme and heme/carotenoid formation (Fig. 4). They are not encoded by single copy genes in *P. patens* (Fig. 4), whereas conserved ancient paralogs encode UROD (St 76). These paralogs had already been acquired before the split of green algae and land plants (St 76) and probably are functionally divergent (24, 25). Note that both the UROD3 (St 76) and CPX2 (St 59) subfamilies are present in algae and *P. patens*, but have been lost in flowering plants.

**Signaling pathways.** The phytohormones and light receptors for morphogenesis found in flowering plants are absent in the unicellular algae, but are 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, which make \textit{P. patens} the earliest diverging species that contains genes for all members of the cytokinin signal transduction pathway known today.

Ten gene families implicated in auxin homeostasis and signaling have been analyzed [table S17], St 25, 33 A/B, 41, 45, 71, 73, 77, 85, 88, 89]. The \textit{C. reinhardtii}, \textit{O. lucimarins}, and \textit{O. tauri} genomes do not encode these, but the \textit{P. patens} genome encodes members of each family [although, on the basis of phylogenies of the GH3 and ILL proteins, St 71 and 85, \textit{P. patens} 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 (AUXI/LAX; St 41) of the 10 families has as many members as angiosperm genomes. On the basis of analysis of \textit{A. thaliana} and our phylogenetic analyses, the auxin signaling pathway has undergone substantial functional diversification within vascular plants since they diverged from bryophytes.

Although no ethylene responses have been noted in mosses, the \textit{P. patens} genome codes for six putative ETR-like ethylene receptors, at least one of which is known to bind ethylene (27). Two putative 1-aminocyclopropane-1-carboxylate (ACC) synthases, catalyzing a critical step in ethylene biosynthesis, were also found. Two transcription factors with strong similarity to the \textit{ethylene biosynthesis}, were also found. Two tran-

sition pathway known today. Particularly, the last common ancestor of all land plants (i) had an overall increase in gene family number (36); and (v) gained more elaborate transport capabilities; and (vi) 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 reconstruction of the genomic events that occurred after the split of vascular plants and 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 homology-based DNA repair, and adaptation to shade and dehydration cycles, as well as a redundant and versatile metabolism. The \textit{P. patens} genome sequence provides a resource for the study of both gene function (37) and evolutionary reconstruction.

References and Notes

3. Materials and methods are available as supporting material on Science Online.
4. Version 1.1 of the \textit{P. patens} genome assembly and annotation can be accessed through the JGI Genome Portal at www.jgi.doe.gov/Physcomitrella; sequences have been deposited at DDBJ/EMBL/GenBank under the project accession AE6U01000000.
13. ST stands for supplementary table; these can be accessed via www.cosmos.org/bm/supplementary_trees/Rensing..
Numerous crystalline materials, including those of bioorganic origin, comprise incommensurate sublattices whose mutual arrangement is described in a superspace framework exceeding three dimensions. We report direct observation by neutron diffraction of superspace symmetry breaking in a solid-solid phase transition of an incommensurate host-guest system: the channel inclusion compound of nonadecane/urea. Strikingly, this phase transition generates a unit cell doubling that concerns only the modulation of one substructure by the other—an internal variable available only in superspace. This unanticipated pathway for degrees of freedom to rearrange leads to a second phase transition, which again is controlled by the higher dimensionality of superspace. These results reveal nature’s capacity to explore the increased number of phases allowed in aperiodic crystals.

A periodicity (1, 2) plays a central role in the structure and physical properties of materials as diverse as quasi-crystals (3), superconductors (4), and zeolites (5). In a number of solid-state supramolecular assemblies—a class of materials ubiquitous both in nature and in the laboratory—modern diffraction methods have revealed incommensurate relations between the basic structures of the noncovalently assembled constituents. In host-guest systems, for example, aperiodicity is an essential element that controls phenomena as diverse as crystal growth (6), polar ordering of guests (7), and absorption and molecular transport (8). Aperiodic materials are described in the framework of a higher-dimensional space called superspace, which decomposes into two orthogonal subspaces: the usual three-dimensional physical space and an internal one (9, 10). The increase in dimensionality theoretically allows many more phases (11, 12); however, the role of aperiodicity in the structural changes of supramolecular assemblies has been largely ignored. In the transformations of such materials, aperiodicity resulting from different interacting length scales should play a major role, coupling structure and dynamics, in particular through the mechanism of free sliding (13, 14). Here, we highlight the richness of transformations associated with the new degrees of freedom hidden in superspace, taking as an example a prototypical host-guest system of nonadecane/urea.

Supramolecular chemistry and crystal engineering enable the formation of a diverse array of aperiodic host-guest architectures in which guest molecules are confined with their own periodicity to nanochannels. Although such host structures may be held together by simple van der Waals forces (8), covalent bonds (5, 15), or hydrogen bonds, as in the diverse structures formed by certain dipeptides (16), a common feature is that the host is more rigid than the guest component. This is typically the case for urea inclusion compounds, in which urea molecules are confined by hydrogen bonds to form helical ribbons, which are woven together to form a honeycomb array of linear, nonintersecting, hexagonal tunnels that contain straight-chain hydrocarbons and analogs (Fig. 1) (17). Guests such as nonadecane pack end to end within van der Waals contact of each other and undergo substantial torsions, librations, and translations, as well as diffusive motions or jumps about the sixfold axis of the host (18). For this host-guest system, the ratio of repeat lengths along the channel axis of the urea host (c_{host}) and nonadecane guest (c_{guest}) is not a rational number, so the crystal is said to be incommensurate.

Although the locations (G_{hid}) of diffraction peaks of typical three-dimensional crystals can be defined with three indices (h, k, and l), an additional parameter or parameters must be used to describe the diffraction patterns of aperiodic materials such as incommensurate channel inclusion compounds. In the linear tunnels of urea, c_{host} and c_{guest} are parallel, so one extra parameter is required to describe the system. Thus, for this nanotubular intergrowth structure, with its single incommensurate direction (e), a four-dimensional superspace description gives the positions of all of the Bragg peaks (8, 9):

\[
G_{hid} = h\mathbf{a}^* + k\mathbf{b}^* + n_1c_{host}^* + n_2c_{guest}^* \tag{1}
\]

where \(h\), \(k\), \(c_{host}^*\), and \(c_{guest}^*\) are the conventional reciprocal unit vectors, which differ from the usual ones \(\mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*\), and \(\mathbf{e}^*\).