The compact Selaginella genome identifies changes in gene content associated with the evolution of vascular plants

Abstract

Vascular plants appeared ~410 million years ago then diverged into several lineages of which only two survive: the euphyllophytes (ferns and seed plants) and the lycophytes (1). We report here the genome sequence of the lycophyte Selaginella moellendorffii (Selaginella), the first non-seed
vascular plant genome reported. By comparing gene content in evolutionary diverse taxa, we found that the transition from a gametophyte- to sporophyte-dominated life cycle required far fewer new genes than the transition from a non-seed vascular to a flowering plant, while secondary metabolic genes expanded extensively and in parallel in the lycophyte and angiosperm lineages. Selaginella differs in post-transcriptional gene regulation, including small RNA regulation of repetitive elements, an absence of the tasiRNA pathway and extensive RNA editing of organellar genes.

*Selaginella moellendorffii*, like all lycophytes, has features typical of vascular plants, including a dominant and complex sporophyte generation (Fig. 1A and B) having vascular tissues with lignified cell types. Lycophytes also share traits with non-seed plants, most notably the release of haploid spores (Fig. 1C) from the sporophyte and a gametophyte generation that develops independently of the sporophyte. Because the lycophytes are an ancient lineage that diverged shortly after land plants evolved vascular tissues (Fig. 2A) (1), we sequenced the Selaginella genome to provide a resource for identifying genes that may have been important in the early evolution of developmental and metabolic processes unique to vascular plants.

The Selaginella genome was sequenced using whole-genome shotgun sequencing (2). The assembled genome size (212.6 Mbp) is twice that determined by flow cytometry (3), indicating that the assembled genome includes two haplotypes of ~106 Mbp that are 98.5% identical at the nucleotide level. A deduced haplotype has 22,285 predicted protein-coding genes, of which 37% are supported by EST sequences, and 58 microRNA (miRNA) loci (2, 4). The Selaginella genome lacks evidence of an ancient whole genome duplication or polyploidy (2), unlike all other sequenced land plant genomes (5–7). Gene density in Selaginella and Arabidopsis, which has a slightly larger genome size, is very similar (2), and both genomes having gene-poor regions rich in transposable elements (TEs) and other repetitive sequences (2). While fewer genes and smaller introns (2) contribute to a genome size smaller than Arabidopsis, this is offset by a greater proportion of TEs in Selaginella (37.5% vs. 15% in Arabidopsis)(2). LTR retrotransposons are the most abundant TEs, occupying one-third of the Selaginella genome (2).

Plant TEs and *MIRNA* loci are significant sources of small RNAs (sRNAs) that function to epigenetically regulate TE and gene activity (8). Several observations suggest that some aspects of epigenetic or post-transcriptional gene regulation in Selaginella are unique among plants. For one, the proportion of sRNAs 23–24 nt in length is extraordinarily small in the Selaginella sRNA population (2) compared to angiosperms (9). Nearly three-quarters of the Selaginella sRNAs (4) map to *MIRNA* loci and are predominantly 21nt in length (2). In angiosperms, 24nt siRNAs, which are generated primarily from TEs, function to silence TE activity through the RdDM pathway (10–12) and accumulate massively in specific cells of the female gametophyte (13). Since the Selaginella sRNA population was generated from sporophytic tissues, the 24nt siRNA pathway may only be deployed during gametophyte development in Selaginella. A second distinction is the absence of *DCL4, RDR6* and *MIR390* loci in Selaginella, which are required for the biogenesis of trans-acting siRNAs (tasiRNAs) in angiosperms (2). Their absence suggests that tasiRNA-regulated processes in angiosperms, including leaf polarity (14) and developmental phase changes in the sporophyte (15, 16), are regulated differently in Selaginella, and possibly reflects the independent origins of foliar organs in the lycophyte and angiosperm lineages (17, 18). Finally, the Selaginella plastome sequence reveals an extraordinarily large number of RNA edited sites (2), as do other lycophyte organellar genomes (19, 20). This coincides with an extraordinarily large number of PPR genes in Selaginella (>800; (2), some of which guide RNA editing events in angiosperms (21).
Because Selaginella is a member of a vascular plant lineage that is sister to the euphyllophytes, we used comparative and phylogenetic approaches to identify gene origins and expansions coinciding with evolutionary innovations and losses in land plants. To identify such genes without regard to function, we compared the proteomes of the green alga Chlamydomonas, the moss Physcomitrella, Selaginella, and 15 angiosperm species, identified gene families that are related by homology by hierarchical clustering (2), and then mapped them onto a phylogenetic tree (Fig. 2B). The 3814 families with gene members present in all plant lineages define the minimum set of genes that were likely to be present in the common ancestor of all green plants and their descendants and include genes essential for plant function. The transition from single-celled green algae to multicellular land plant approximately doubled the gene number with the acquisition of 3006 new genes. The transition from non-vascular to vascular plant is associated with a gain of far fewer new genes (516) than the transition from a basal vascular plant to a basal euphyllophyte whose descendents include the angiosperms (1350). These numbers show that the evolution of traits unique to euphyllophytes or angiosperms required the evolution of about three times more new genes than the transition from a plant having a dominant gametophyte and simple, leafless and non-vascularized sporophyte (typified by modern bryophytes) to a plant with a dominant, vascularized and branched sporophyte with leaves.

In a second approach, we analyzed the phylogenies of genes known to function in Arabidopsis development (2). We identified 424 monophyletic groups of developmental genes, each group containing putatively all genes descended from a common land plant ancestral gene (Table S6). Selaginella and Physcomitrella genes are present in 377 (88%) and 356 (84%) of the 424 land plant orthologous gene groups, respectively, indicating that the common ancestor of land plants had most of the gene families known to direct angiosperm development. Conspicuous expansions of families within different lineages resulted in different numbers of land plant orthologs in each genome (Table S6). The 27 vascular plant-specific orthologous groups likely represent genes associated with developmental innovations of vascular plants. Among them are genes regulating the meristem (CLV1 and CLV2), hormone signaling (GID1, CTR1) and flowering (TFL2 and UFO). Interestingly, homologs of genes involved in the specification of xylem (NST and VND) (22) and phloem (APL) (23) in Arabidopsis are present in Physcomitrella and Selaginella, suggesting that the developmental programs for patterning and differentiation of vascular tissues were either present in, or co-opted from preexisting genetic programs in the ancestral land plant. The 43 groups lacking genes from Physcomitrella and Selaginella (Table S6) likely identify genes that were necessary for euphyllophyte or angiosperm developmental innovations. Among this group are genes that regulate light signaling (FAR1, MIF1, OBP3 and PKS1), shoot meristem development (AS2 and ULT1), hormone signaling and biosynthesis (BR1, BSU1, ARF16, ACS and ACO), and flowering (HUA1, EMF1, FT, TFL1 and FD). Altogether, these results suggest that the evolutionary transitions from a non-vascular plant to a vascular to angiosperm included the stepwise addition of components of some developmental pathways, especially those regulating meristem and hormone biology, as previously noted for the gibberellin signaling pathway (24, 25).

Genes involved in secondary metabolism were also investigated because plants synthesize numerous secondary metabolites that they use to interact with their environment. Three gene families involved in their biosynthesis, including those encoding cytochrome P450-dependent monooxygenases (P450s), BAHD acyltransferases (BAHDs) and terpene synthases (TSs), were analyzed. The largest of these in Selaginella is the P450 family, accounting for 1% of its predicted proteome (Table S7) (2). All three families show similar evolutionary trends, with the inferred ancestral vascular plant having a small number of genes that radiated extensively but independently within the lycophyte and angiosperm lineages (Fig. S6–13). BAHD and TS genes, which are known to be involved in the
biosynthesis of volatile odorants, are apparent only in seed plants (Figs. S12–13), likely reflecting the co-evolution of seed plants with animals that pollinate flowers or disperse seeds. The independent diversification of these gene families plus the large number of Selaginella genes suggest that Selaginella not only has the potential to synthesize a repertoire of secondary metabolites that rivals the angiosperms in complexity, but that many of them are likely to be unique. Some have been shown to be of pharmaceutical value (e.g., 26).

We have used the compact Selaginella genome sequence to uncover genes associated with major evolutionary transitions in land plants. Understanding their functions in Selaginella and other taxa, as well as acquiring the genome sequences of other informative taxa, especially charophytes, ferns and gymnosperms, will be key to understanding the evolution of plant form and function.

**Supplementary Material**

Refer to Web version on PubMed Central for supplementary material.

**Acknowledgments**

Selaginella sequences were deposited at GenBank, accession nos. GL377566-GL378322.1, and HM173080. Genome sequencing and analysis was performed by the U.S. Department of Energy, Joint Genome Institute supported by the Office of Science of the U.S. DOE, Contract DE-AC02-05CH11231, (IG, DR, RO, AS, IS, HS, EL, SL, TM). Support provided by: NSF (JAB, JLB, CC, CD, ME, SKF, RGO, MP, CS, PGW, JW); DOE (WBF, HVS); NIH (MJA, MSB, MD, KGK, GM, DES); USDA-NIFA NRICG (AR); Lewis B and Dorothy Cullman Program (KGK, AL and BAA); Life Sciences Research Institute (NDB); Georgia Research Alliance (QZ, JLB); Busch Biomedical (TPM); Australian Research Council (JLB); Villum Kann Rasmussen Foundation and Danish Research Council (IS, JH, PU, BLP, WGTW); DFG (ID, SAF, MG, ADZ); BMBF FRISYS 0313921 (SAF); Bundesministerium fuer Bildung und Forschung (BM, DRM); Burgundy Research Council (CRB); Czech Ministry of Education (ME); NSERC (EIB, NWA, MSB); MEXT, Japan (MH, TN, HF) and JSPS (MH, TN, HF, KM, TM, MS). Dave Hurley and Steven Hentel provided computational assistance.

**References and Notes**

Figure 1. Selaginella morphology. (A) The diploid sporophyte body. Bar, 10 mm. (B) A shoot with two ranks of microphylls ("leaves") and strobili. Each microphyll of a strobilus has either a mega- or microsporangium where mega- or microspores are produced. Bar, 2 mm. (C) An orange microspore on top of a dark megaspore. These single celled haploid spores represent the beginning of the independent haploid gametophyte generation. The microgametophyte produces motile sperm and the megagametophyte eggs. Bar, 0.1 mm.
Figure 2.
(A) Phylogeny of plants. Taxa in red have sequenced genomes. (B) Gene family gains (+) and losses (−) mapped onto the plant phylogenetic tree. The minimum numbers of gene families present in the ancestors of different plant lineages are circled.