A major feature distinguishing the locus notype is located within 100 kb on either side of the molecular footprint of strongly positive selection (fig. S4). Paradoxically, although the industrial melanism in the peppered moth was expected to generate the profile of linkage disequilibrium (repulsion-phase) haplotypes separately for the two datasets, genetically, melanin pattern formation in these deeply diverged lepidopterans may have a common genetic basis, and the deficit for melanaria is greater because the melanaria-type alleles (C and A, respectively) were rare in the ancestral population. These two loci suggest that melanaria-to-typica haplotype introgression has been weak.

Bombyx mori chromosome 17 and its orthologs in other lepidopterans are rich in major color-pattern genes, such as black moth and wild wing spot (13). These Bombyx mori genes do not map closely to the melanaria locus (fig. S5). However, Bicyclus anynana LG17 contains two pigment-patterning mutants, 067 and Bigeye, that both affect eye spot size, with Bigeye predicted to reside within the melanaria region (14). The Bigeye and carbonaria phenotypes are clearly very different, but they share a large increase in the proportion of melanized scales. The carbonaria core region also overlaps the mimetic patterning locus in four Heliconius species, collectively referred to as the Yb-P-Yb/Sb-Cr locus (15–17). The betulaaria genotypes identified in this region so far correspond entirely with those described for the Yb-P-Yb/Sb-Cr region. A major feature distinguishing Heliconius forms is the amount and distribution of black, as with the various B. betulaaria morphs (4). This unlikely coincidence suggests that the control of melanin pattern formation in these deeply diverged lepidopterans may have a common genetic basis, the functional units of which have yet to be identified.

The rapid spread of an initially unique haplotype, driven by strong positive selection, is expected to generate the profile of linkage disequilibrium we have observed (18), establishing that UK industrial melanism in the peppered moth was seeded by a single recent mutation that spread to most parts of mainland Britain and also colonized the Isle of Man (fig. S4). Paradoxically, although the carbonaria morph is now strongly disadvantageous and consequently rare in the United Kingdom, the rapidity of its decline (19) has minimized the eroding effect of typica introgression on the molecular footprint of strongly positive selection created during its ascendency.

References and Notes
11. Materials and additional results are available as supporting material on Science Online.

Acknowledgments: This work was funded by Natural Environment Research Council grant NE/C003103/1 to I.J.S. Cytogenetic experiments were financed by Grant Agency of the Academy of Sciences of the Czech Republic grant IA600960925, Entomology Institute project Z50070508, and Grant Agency of the Czech Republic 251208H042 to F.M. and M.D. Specimens for the UK-wide survey were made available by several museums and private collectors, listed in table S4. J. Delf produced fig. S3. GenBank accession numbers are listed in table S1.

Supporting Online Material
www.sciencemag.org/cgi/content/full/science.1203043/DC1
Materials and Methods
Figs. S1 to S5
Tables S1 to S4
References
19 January 2011; accepted 30 March 2011
Published online 14 April 2011; 10.1126/science.1203043

The Selaginella Genome Identifies Genetic Changes Associated with the Evolution of Vascular Plants
Jo Ann Banks,1–4 Tomoki Nishiyama,2,3 Mitsuysau Hasebe,2,3,4,5 John L. Bowman,6,7 Michael Grabiszkov,8 Claude dePamphilis,9,10,11 Victor A. Albert,11 Naoki Aono,10 Tsuyoshi Aoyama,4,5
Barbara A. Ambrose,16 Neil W. Ashton,14 Michael J. Axtell,9 Elizabeth Barker,15 Michael S. Barker,15 Jeffrey L. Bennetzen,16 Nicholas D. Bonawitz,17 Clint Chapple,17 Chaoyang Cheng,7
Wolf B. Frommer,17,24,25 Tomomichi Fujiita,26 Lydia Gramzow,27 Michael Gutensohn,30,31 Jesper Harholt,32 Mitsuhashi Hattori,33,34 Alexander Heyl,35 Tadayoshi Hirai,36 Yuji Hiwatashi,4,5 Masaki Ishikawa,37
Mineko Iwata,4,37,38 Kenneth G. Karl,13 Barbara Koehler,18 Uener Kolukisaoglu,39 Minoru Kubo,33 Tetsuya Kurata,39,40 Sylvie Lalonde,41 Kieji Li,41 Ying Li,41 Amy Litt,3 Eric Lyons,41
Peter Urvosky,22 Sachiko Wazakuji,4,7 Jing-Ke Weng,17,18 William W.G. Willats,32 Daniel Wipf,69 Paul G. Wolf,70 Ling Yang,16 Andreas D. Zimmer,55,57 Qihu Zhu,16 Therese Mitros,72 Uffe Hellsten,73 Dominique Loqué,51 Robert Ottillar,73 Asaf Salamov,74 Jeremy Schmutz,73 Harris Shapiro,73 Erika Lindquist,73 Susan Lucas,73 Daniel Rokhsar,72,73 Igor V. Grigoriev73

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. We report here the genome sequence of the lycophyte Selaginella moellendorffii (Selaginella), the first newseed vascular plant genome reported. By comparing gene content in evolutionarily diverse taxa, we found that the transition from a gametophyte- to a sporophyte-dominated life cycle required far fewer new genes than the transition from a nonseed vascular to a flowering plant, whereas secondary metabolic genes expanded extensively and in parallel in the lycophyte and angiosperm lineages. Selaginella differs in posttranscriptional gene regulation, including small RNA regulation of repetitive elements, an absence of the trans-acting small interfering RNA pathway, and extensive RNA editing of organellar genes.

S. moellendorffii, all like lycophytes, has features typical of vascular plants, including a dominant and complex sporophyte generation (Fig. 1, A and B) having vascular tissues with lignified cell types. Lycophytes also share traits with nonseed 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 specific to vascular plants.

The Selaginella genome was sequenced by 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 expressed sequence tag 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 have gene-poor regions rich in transposable elements (TEs) and other repetitive sequences (2). Although 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% versus 15% in Arabidopsis) (2). Long terminal repeat retrotransposons are the most abundant TEs, occupying one-third of the Selaginella genome (2).

Plant TEs and miRNA loci are important sources of small RNAs (sRNAs) that function

---

Fig. 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 a 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.
to epigenetically regulate TE and gene activity (8). Several observations suggest that some aspects of epigenetic or posttranscriptional gene regulation in Selaginella are unique among plants. For one, the proportion of sRNAs 23 to 24 nucleotides (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 21 nt in length (2). In angiosperms, 24-nt sRNAs, which are generated primarily from TE loci, function to silence TE activity through the RNA-dependent DNA methylation pathway (10–12) and accumulate massively in specific cells of the female gametophyte (13).

Because the Selaginella sRNA population was generated from sporophytic tissues, the 24-nt sRNA 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 sRNA (tsiRNAs) 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 exceptionally 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 eukaryophytes, 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 algae 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 descendents 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 nonvascular 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 eukalyphophyte whose descendents include the angiosperms (1350). These numbers show that the evolution of traits specific to eukalyphophytes 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 nonvascularized 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 (89%) 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 and CTR1), and flowering (TFL2 and UBI). 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, OB3, and PKSI), shoot meristem development (AS2 and ULTI), hormone signaling and biosynthesis (BRI1, BSUI, ARF16, ACS, and ACO), and flowering (HUA1, EMF1, FT, TFL1, and FD). Altogether, these results suggest that the evolutionary transitions from a nonvascular plant to a vascular angiosperm included the stepwise addition of components of some developmental pathways, especially those regulating meristem and hormone biology, as previously noted for the giberellin 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 monoxygenases (P450s), BAHD acyltransferases (BAHDs), and terpene synthases (TSSs), 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 (figs. S6 to S13). BAHD and TS genes, which are known to be involved in the biosynthesis of volatile odorants, are apparent only in seed plants (figs. S12 to S13), likely reflecting the coevolution 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 and 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 pharmacetical 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.

References and Notes
2. Details are given in the supporting materials on Science Online.
Chromatin “Prepattern” and Histone Modifiers in a Fate Choice for Liver and Pancreas

Cheng-Ran Xu,1 Philip A. Cole,2 David J. Meyers,2 Jay Kormish,2 Sharon Dent,3 Kenneth S. Zaret1†

Transcriptionally silent genes can be marked by histone modifications and regulatory proteins that indicate the genes’ potential to be activated. Such marks have been identified in pluripotent cells, but it is unknown how such marks occur in descendant, multipotent embryonic cells that have restricted cell fate choices. We isolated mouse embryonic endoderm cells and assessed histone modifications at regulatory elements of silent genes that are activated upon liver or pancreas fate choices. We found that the liver and pancreas elements have distinct chromatin patterns. Furthermore, the histone acetyltransferase P300, recruited via bone morphogenetic protein signaling, and the histone methyltransferase EzH2 have modular roles in the fate choice. These studies reveal a functional “prepattern” of chromatin states within multipotent progenitors and potential targets to modulate cell fate induction.

EARLY pluripotent cells of the mammalian embryo develop into multipotent endoderm, ectoderm, and mesoderm germ layers. In pluripotent cells, silent genes that will be activated later in development often exist with histone modifications and/or bound transcription factors that reflect the chromatin being “poised” for activity (1–3). It is unclear whether such poised states exist for silent genes in germ layer cells and, if so, whether genes poised for different tissue fates exhibit different chromatin features. Furthermore, it is not known whether enzymes that establish chromatin states can control germ layer fate choices. Embryonic germ layer cells are few in number, they have not been purified, and chromatin analysis on small cell populations is challenging (4). Yet germ layer cells represent the first lineage-restricted, multipotent progenitors of the embryo and a paradigm for all subsequent fate decisions.

Ventral foregut endoderm cells undergo a fate choice for liver or ventral pancreas progenitors (5, 6). FoxA1 or FoxA2, GATA4 or GATA6, vHNF1, and Hnf6 (also known as Oc1) are necessary in the endoderm for both liver and ventral pancreas induction (7). In the absence of any set of the factors, the earliest liver marker genes Alb1, Afp, and Ttr and the ventral pancreas transcription factor gene Pdx1 fail to be activated, or expression is delayed, and tissue buds fail to form (7). It is not clear how the same group of factors can be necessary for both liver and ventral pancreas and how signaling promotes the different fates. We sought to map chromatin states at silent liver- and pancreas-specific regulatory sequences in endoderm cells, to discover the factors or relevant histone-modifying enzymes, and test the enzymes’ functions in the liver-versus-pancreas decision.

We used fluorescence-activated cell sorting (FACS) with the ENDMD1 antibody to isolate ventral foregut endoderm cells from embryonic day 8.25 (E8.25) mouse embryos with four to six somite pairs (4–6S) (8) (fig. S1), just prior to the induction of hepatic and pancreatic fates (5, 9). We also used the liver-specific antibody Liv2 to isolate nascent hepatoblasts expressing Alb1, Afp, and Ttr from E9.5 embryos (fig. S2) (10). Chromatin marks in ENDM1+ and Liv2+ populations were identified with a low–cell number chromatin immunoprecipitation (ChIP) protocol (4) for H3K9acH1K4ac, H3K4me2, H3K4me3, H3K9me3, H3R17me2a, H3K27me3, H3K36me2, H3K36me3, H3K79me2, H4K20me3, H3T3ph, H3S10ph, the histone variant H2A.Z, and the chromatin remodelers BrG1 and SNF2. We assessed the liver-specific promoter and enhancer of Alb1 (11, 12), the liver-specific promoters of Afp and Ttr genes (13, 14), and the I, II, III, and IV upstream elements and local promoter of the pancreatic determination gene Pdx1 (fig. S3). The I, II, and III upstream elements and promoter of Pdx1 reconstitute pancreas-specific activation (15); the IV element may function later (16, 17). All of the target genes are silent in endoderm cells, and only the liver genes become activated in hepatoblasts.

References

1 Institute for Regenerative Medicine, Epigenetics Program, Department of Cell and Developmental Biology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA. 2 Department of Pharmacology and Molecular Sciences, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. 3 Department of Molecular Carcinogenesis, University of Texas M. D. Anderson Cancer Center, Smithville, TX 78957, USA.

*Present address: Biochemistry and Molecular Biology, University of Calgary, Calgary, Alberta T2N 4N1, Canada.

†To whom correspondence should be addressed. E-mail: zaret@upenn.edu