Deciding among green plants for whole genome studies

Kathleen M. Pryer, Harald Schneider, Elizabeth A. Zimmer and Jo Ann Banks

Recent comparative DNA-sequencing studies of chloroplast, mitochondrial and ribosomal genes have produced an evolutionary tree relating the diversity of green-plant lineages. By coupling this phylogenetic framework to the explosion of information on genome content, plant-genomic efforts can and should be extended beyond angiosperm crop and model systems. Including plant species representative of other crucial evolutionary nodes would produce the comparative information necessary to understand fully the organization, function and evolution of plant genomes. The simultaneous development of genomic tools for green algae, bryophytes, 'seed-free' vascular plants and gymnosperms should provide insights into the bases of the complex morphological, physiological, reproductive and biochemical innovations that have characterized the successful transition of green plants to land.

Kathleen M. Pryer
Dept of Biology, Duke University, Durham, NC 27708, USA.
e-mail: pryer@duke.edu

Harald Schneider
Albrecht-von-Haller-Institut für Pflanzenwissenschaften, Abteilung Systematische Botanik, Georg-August-Universität, Untere Karspüle 2, 37073 Göttingen, Germany.

Elizabeth A. Zimmer
Laboratories of Analytical Biology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560, USA.

Jo Ann Banks
Dept of Botany and Plant Pathology, Purdue University, West Lafayette, IN 47907, USA.

The availability of the first whole genome sequences from the plant kingdom [1–3] coincides with a recent flurry of phylogenetic studies resolving deep relationships across major green-plant lineages [4–11]. This provides us with an unprecedented opportunity to offer a broad evolutionary perspective for new genomic endeavors. With the complete genome sequence for the small mustard *Arabidopsis* and two draft genome sequences for rice (*Oryza sativa* ssp. *japonica* and ssp. *indica*) now in hand, plant biologists are poised to determine the function and biotechnological potential of all the genes in these two species [12,13] (http://www.nsf.gov/cgi-bin/getpub?nsf0113). *Arabidopsis* and *Oryza* are our blueprints for comparative plant genomics and will help us to understand how their genes and genomes compare to each other and to those of other plant species, and to identify the relationships between genome structure, gene function and evolution [13]. Already, there is an escalated interest in supporting large collaborative plant genome projects (http://www.arabidopsis.org/workshop1.html).

Sampling plant diversity for genome studies

Genome sequences for two of the >300 000 land plants are scarcely representative of the rich botanical diversity that dominates our terrestrial ecosystems (there are five times as many flowering plants species alone than vertebrate species), so comparisons of a few plant genomes are likely to be followed by a much broader sampling of plant genetic diversity. This can already be seen in the investigations into additional plant genomes beyond standard models sponsored by new programs at the US National Science Foundation, such as the Plant Genome Research Program (PGRP) and Genomic Resources: Bacterial Artificial Chromosome Library Construction. But how will this plant diversity be sampled? Funding agencies and large genome consortiums can and will make these decisions and set priorities. These will have a profound impact on the future directions of basic and applied plant biology research, including agricultural science, evolutionary and developmental biology, bioinformatics, and functional and comparative genomics [14]. Already, complete contiguous sequencing of the rice genome, which is four times larger than *Arabidopsis*, is well under way (see Opinion by Robin Buell in this issue of *Trends in Plant Science*) and is expected to be complete for 2004 [13,15].

Given the obvious importance and relevance of rice and other cereal crops such as maize (*Zea*), wheat (*Triticum*) and barley (*Hordeum*) to humans and domesticated animals, will future whole genome sequencing efforts continue to focus primarily on these and other crop plants? The genomics of crop plants is certainly important, but excluding plant species representative of other crucial evolutionary nodes from the priority list has made it impossible to gather the comparative information we need to understand fully the organization, function and evolution of the plant genome. In animal genomics, the broader sampling of such taxa as the fruit fly (*Drosophila*), mouse (*Mus*), zebrafish (*Danio*) and worm (*Caenorhabditis*) has been and will continue to be invaluable in understanding gene function and evolution in diverse organisms [16], including humans, and there is no reason to think that the same would not be true for plants.

Current genome sequencing is almost exclusively focused on organisms that are either most closely related to humans or of major economic or biomedical relevance to humans [13–19]. Within this context, the obvious green-plant candidate species slated for genome studies are nested in four principal clades of derived angiosperms: asterids, rosids, caryophyllids and monocots (Fig. 1). The representation is particularly skewed within each of these clades, almost always favoring plants of commercial interest. In the monocots, for example, sampling has explicitly focused on grasses and, in particular, the cereal crops.

Together with others [13,14,20], we advocate an approach that focuses on the 'big picture' of green-plant phylogeny by including a judicious broader representation from across all major groups. The expectation of such a comparative sampling is that it offers the greatest potential for a better understanding of the evolutionary dynamics of plant genome organization and of the functional and evolutionary processes that are fundamental to plant life. For example, limited comparative sequence

http://plants.trends.com 1360-1385/02/$ – see front matter PII: S1360-1385(02)02375-0

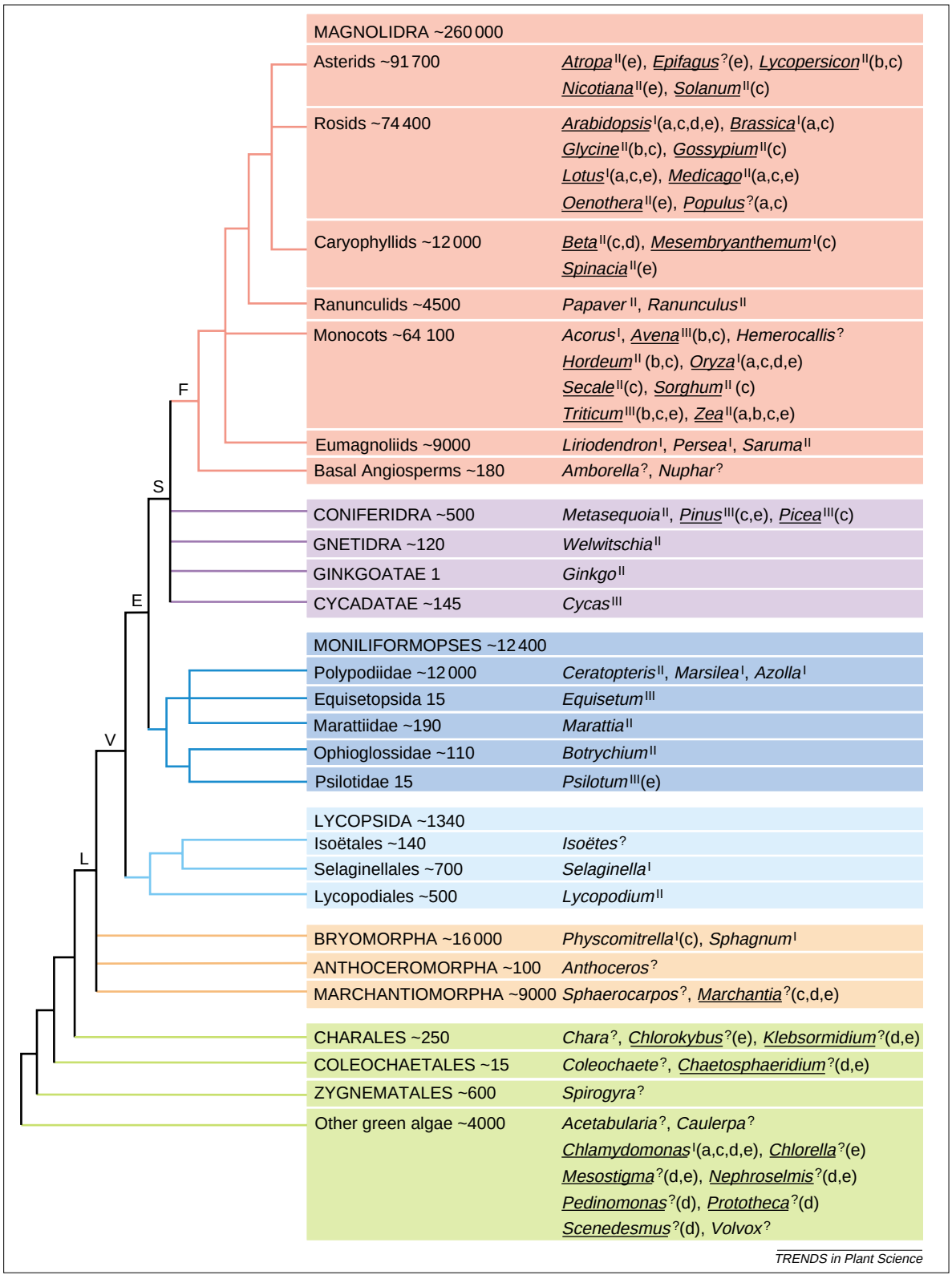

Fig. 1. Phylogenetic relationships among major lineages of green plants. Abbreviations: E, euphyllophytes; F, flowering plants (angiosperms); L, land plants; S, seed plants; V, vascular plants [4–10]. Genera are underlined when listed in Refs. [13,52] or at http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/PlantList.html, http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/organelles.html, http://bahama.jgi-psf.org/prod/bin/chlamy/home.chlamy.cgi, http://bahama.jgi-psf.org/prod/bin/populus/home.populus.cgi, http://128.118.180.140/fgp/index.html, http://www.botany2002.org/sympos8/abstracts/3.shtml, http://megasun.bch.umontreal.ca/ogmp/projects/other/mtcomp.html or http://www.biology.duke.edu/chlamy_genome/search.html, either as: (a) large-scale sequencing projects, (b) genetic maps or (c) large-scale EST-sequencing projects, or as having available (d) mitochondrial or (e) plastid genome sequences. Genera not underlined are candidate species not currently targeted (but see http://128.118.180.140/fgp/index.html). Genome size is indicated where known, based on DNA *C*-value estimates [27] and http://www.rbgkew.org.uk/cval/homepage.html; sizes were arbitrarily grouped into three classes: I = <1 pg of DNA per nucleus; II = 1–10 pg; III = >10 pg; ? = unknown.

analyses of megabase orthologous regions between distantly related angiosperms (e.g. *Arabidopsis* and tomato) have unambiguously shown support for microcolinearity and identified large-scale duplications and subsequent gene loss as an important factor in the evolution of plant genomes [21–23]. Further insights into genome organization will certainly be detected as many different genomic regions across a diversity of plant species are compared [24]. A directed, inclusive taxon-sampling approach to choosing the green plants best suited to these comparisons could easily be implemented by using three simple criteria – phylogenetic position, genome size and experimental amenability. Experimental amenability includes the potential for analyzing gene function by forward and reverse genetics (such as transposon or transfer DNA (tDNA) tagging, gene replacement or post-transcriptional gene silencing).

Phylogeny and genome size

Of all the major green-plant lineages, angiosperms (Fig. 1) are the best sampled. If genomic efforts continue to be focused on economic angiosperms, which probably all have significantly more (mostly redundant) DNA [13], this will not advance an explicit comparative understanding of flowering plant-genome structure and function. Although most basal groups of angiosperms (e.g. *Amborella*, *Nymphaea*, *Piper*) are not targeted for genome study (but see http://128.118.180.140/fgp/index.html), it is these basal taxa that are needed to reconstruct the ancestral (generalized) genome organization of flowering plants and to decipher its evolution. In addition, and as a baseline, at least two representatives should be chosen for large-scale genome analysis from each of the major, but understudied, green-plant lineages (Fig. 1). These include: the conifers, cycads, *Ginkgo* and gnetophytes (purple); the ferns and horsetails (dark blue); the lycophytes (pale blue); the mosses, liverworts and hornworts (brown); and the green algae (green).

As a general rule of thumb, a small genome has been an important guideline in selecting candidates for genome sequencing. *Arabidopsis* was chosen, in part, because, at $C = 0.18$, it has one of the lowest known *C*-values (the total amount of DNA in the haploid genome) [25] of any angiosperm. However, the lycophyte *Selaginella* belongs to a lineage that diverged >360 million years ago from the lineage that gave rise to angiosperms, and has a *C*-value of 0.06 [26] (R.A. Bouchard, PhD thesis, University of Chicago, USA, 1976), nearly four times smaller than the *Arabidopsis* genome. A revised *C*-value estimate of 0.16 for *Selaginella* has recently been published [27], which is still the least nuclear DNA yet known for any 'seed-free' vascular plant, but closer to estimates reported for *Arabidopsis*. Comparing such a small genome to *Arabidopsis*, rice and other angiosperms would probably increase our understanding of the extensively duplicated (and potentially functionally redundant) regions that make up much of the *Arabidopsis* and rice genomes [1–3].

Such optimism arises from comparative studies between, for example, the pufferfish (*Fugu* and *Tetraodon*) and human genomes. Pufferfish diverged from lobe-finned fishes (which gave rise to tetrapods) 450 million to 500 million years ago. The pufferfish genome is compact and has essentially the same information as the human genome but, because it lacks many repeats and has small introns, this information is packed into a ninth of the DNA [28–30]. Regulatory sequences are easier to detect in pufferfish and researchers can use them to 'fish out' related sequences linked to previously unknown human genes [29–31]. Without the redundant DNA, it is easier to identify promoters and regulatory regions, and so this genome is a useful tool for annotating the human genome, specifically chromosome 20 [32].

The combined knowledge of phylogenetic relationships and genome size is a powerful tool that needs to be included as an important criterion in choosing future plant species for whole-genome studies. For example, within the monilophytes (Fig. 1), *Ophioglossum* (not shown) and *Botrychium* are sister taxa, but *Ophioglossum* has a *C*-value of 65.2, almost 20 times larger than that of *Botrychium* [27] (http://www.rbgkew.org.uk/cval/homepage.html) – the better choice for a comparative genomic study. Genome studies of gymnosperms have focused almost exclusively on pines (*Pinus*) and the closely related spruces (*Picea*), genera of considerable importance in the forest biotechnology industry. However, among gymnosperms, *Pinus* has the largest genome, with the model pine *Pinus taeda* having a DNA *C*-value of 22.10 (http://www.rbgkew.org.uk/cval/homepage.html).

Functional genomics of 'seed-free' plants

During the past century, many green plants, from algae to angiosperms, have been developed as model organisms for the study of physiology, biochemistry, genetics and developmental biology. These plants have diverse experimental advantages that led to them becoming 'models', usually including one or more of short generation time, small size, large number of offspring, crossability and relative ease of manipulation in laboratory, greenhouse or field conditions. Many of these taxa now have established track records in genetic studies and some, such as the aquatic fern *Ceratopteris richardii* and the moss *Physcomitrella patens* (Fig. 1), have a modest but growing availability of genomic information (see Opinion by Stefan Rensing *et al.* in this issue of *Trends in Plant Science*), including cDNA libraries and EST databases and cDNA microarrays for expression profiling [33,34]. *Physcomitrella* is the first moss to be successfully transformed and has recently been highlighted as the first land plant and,

more interestingly, the first multicellular eukaryote in which gene targeting occurs with an efficiency similar to that observed in yeast [34]. Efficient gene targeting coupled with a complete genome sequence has allowed true functional genomics to replace simple expression studies in yeast. Therefore, *Physcomitrella* in particular, among the early land-plant lineages, is an obvious priority for genome sequencing. Functional genomics tools are not available for most of these classical 'seed-free' model plants because no one has tried to develop them in these plants, but there is no reason to believe that these tools cannot be developed or applied to non-crop species. Comparative genome analyses between *Arabidopsis* and soybean (eightfold difference in genome size) are already facilitating cross-utilization of genetic resources and tools, and shedding light on evolutionary events associated with the divergence of these distantly related genomes [23]. Surely, the same progress can be extended beyond within-angiosperm comparisons.

Using a robust phylogeny as a guide, genome sequence comparisons from diverse taxa across strategic evolutionary nodes has proved to be a powerful tool in elucidating mitochondrial-genome evolution [35,36]. A similar phylogenetic strategy can also permit the reconstruction of ancestral sequences of biologically active molecules to improve our understanding of physiological functions within a protein family. Such a study has been done for artiodactyl ribonucleases [37,38], whereby 13 genes that encode phylogenetically inferred protein sequences at different nodes were created by site-directed mutagenesis. These genes were expressed in bacteria and the properties of these 'ancestral' ribonucleases were studied *in vitro*. The reconstructed ribonucleases for the more recent divergences were shown to be fully functional and stable. By contrast, those that corresponded to the more ancient artiodactyl ancestors were less thermostable and had enhanced catalytic activity, indicating a more generalist function before the evolution of true ruminant digestion. This study illustrates how phylogenetic analysis is yet another tool in functional-genomic studies.

Conclusion

To enhance our understanding of biological diversity and to expand the relevance of modern plant science, broad comparisons will need to be made among the genomes of distantly related plants flung far across green-plant phylogeny. This should provide us with basic insights into how plant life works and how it differs from other eukaryotic lifestyles. Recent identification of the Charales as the sister taxon to land plants [4] contradicts the notion that the transition from aquatic green algae to terrestrial land plants issued from a simple common ancestor and suggests instead that early land-plant beginnings involved a more complex morphology, physiology, reproductive biology and biochemistry. Targeting *Chara*, for example, for comparative genomic studies might not only help us to understand how its relatives made the successful transition to land but also elucidate how plants came to dominate terrestrial ecosystems [4,39].

In addition, it would be prudent to keep in mind a recent genome-wide comparative phylogenetic survey suggesting that ~4500 *Arabidopsis* protein-coding genes (~18% of the total) were acquired from the cyanobacterial ancestor of plastids [40]. These proteins encompass all functional classes and most of them are targeted to cell compartments other than the chloroplast. The impact of this massive horizontal gene transfer on the plant genome needs to be taken into consideration by the Plant Genome Initiative and genome sequencing of cyanobacteria should be included as part of the bigger picture in determining their role in the evolution of green-plant diversity [41].

Although the biological meaning of genome size is still a mystery [42], the controversial idea that there is a general trend in evolution towards a larger genome because of duplicated gene or chromosome segments is undergoing a revival [43]. More data are needed to determine whether fluctuations in genome size and/or changes in genome organization could partly explain such success stories in plant history as the colonization of land, the Devonian diversification of vascular plants and the relatively recent radiation of flowering plants [44,45]. In the meantime, we can learn something from the abundance of bacterial genomes sequenced since 1995 and the mutual benefit they have provided to both evolutionists and biotechnologists, but a truly representative set of genomes is needed for a proper perspective [46]. It will not be possible (nor is it desirable) to sequence all 300 000 land-plant species but, if plant phylogeneticists are permitted to have a voice in the selection of taxa to sequence, in the long run, there will indeed be a genuine understanding of the organization, function and evolution of the plant genome [47,48]. An example of such synergistic activity is the collaborative effort among plant developmental and evolutionary biologists that is already catalyzing new research between these two communities in generating arrayed bacterial artificial chromosome and cDNA libraries from organisms across the tree of life [49,50] (http://www.nsf.gov/pubs/2001/bio012/start.htm). However, an approach to genome sequencing that promotes the application of 'investment criteria' [51] is unlikely to achieve these same far-reaching goals.

What will it take for the plant community and funding agencies to provide the necessary resources to develop the essential functional genomics tools required to understand the function of genes in non-crop plants? 'Business as usual' will need to change, with plant genomicists and geneticists embracing plants of phylogenetic interest at crucial

Acknowledgements
We thank Jeff Dangl (University of North Carolina) for encouragement, and François Lutzoni, Eric Schuettpelz and the anonymous referees for their helpful comments. K.M.P., H.S. and J.B. acknowledge support from the National Science Foundation.

evolutionary nodes, and traditional phylogeneticists assisting in developing a minimum set of functional genomics tools for these newly targeted model organisms, with which they can test their phylogenetic hypotheses experimentally.

These two communities have never been better positioned to work together to change profoundly the way in which plant biologists select and approach questions that are fundamental to understanding the mechanisms of plant life.

References

1 The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815
2 Yu, J. *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296, 79–92
3 Goff, S.A. *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296, 92–100
4 Karol, K.G. *et al*. (2001) The closest living relatives of land plants. *Science* 294, 2351–2353
5 Pryer, K.M. *et al*. (2001) Horsetails and ferns are a monophyletic group and the closest relatives to seed plants. *Nature* 409, 618–622
6 Nickrent, D.L. *et al*. (2000) Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Mol. Biol. Evol.* 17, 1885–1895
7 Barkman, T.J. *et al*. (2000) Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. *Proc. Natl. Acad. Sci. U. S. A.* 97, 13166–13171
8 Soltis, P.S. *et al*. (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* 402, 402–404
9 Mathews, S. and Donoghue, M.J. (1999) The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* 286, 947–950
10 Qiu, Y-L. *et al*. (1999) The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* 402, 404–407
11 Brown, K.S. (1999) Deep Green rewrites evolutionary history of plants. *Science* 285, 990–991
12 Walbot, V. (2000) A green chapter in the book of life. *Nature* 408, 794–795
13 Bevan, M. (2002) The first harvest of crop genes. *Nature* 416, 590–591
14 Mandoli, D.F. and Olmstead, R. (2000) The importance of emerging model systems in plant biology. *J. Plant Growth Regul.* 19, 249–252
15 Adam, D. (2000) Now for the hard ones. *Nature* 408, 792–793
16 Copeland, N.F. *et al*. (2002) Mmu 16 – comparative genomic highlights. *Science* 296, 1617–1618
17 O'Brien, S.J. *et al*. (2001) On choosing mammalian genomes for sequencing. *Science* 292, 2264–2266
18 Pennisi, E. (2001) Insects rank low among genome priorities. *Science* 294, 1261–1262
19 Fishman, M.C. (2001) Zebrafish – the canonical vertebrate. *Science* 294, 1290–1291
20 Daly, D.C. *et al*. (2001) Plant systematics in the age of genomics. *Plant Physiol.* 127, 1328–1333
21 Rossberg, M. *et al*. (2001) Comparative sequence analysis reveals extensive microcolinearity in the *Lateral Suppressor* regions of the tomato, *Arabidopsis*, and *Capsella* genomes. *Plant Cell* 13, 979–988
22 Ku, H-M. *et al*. (2000) Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl. Acad. Sci. U. S. A.* 97, 9121–9126
23 Grant, D. *et al*. (2000) Genome organization in dicots: genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. *Proc. Natl. Acad. Sci. U. S. A.* 97, 4168–4173
24 Schmidt, R. (2002) Plant genome evolution: lessons from comparative genomics at the DNA level. *Plant Mol. Biol.* 48, 21–37
25 Bennett, M.D. and Smith, J.B. (1991) Nuclear DNA amounts in angiosperms. *Philos. Trans. R. Soc. London Ser. B* 334, 309–345
26 Bennett, M.D. and Leitch, I.J. (2001) Nuclear DNA amounts in pteridophytes. *Ann. Bot.* 87, 335–345
27 Obermayer, R. *et al*. (2002) Nuclear DNA *C*-values in 30 species double the familial representation in pteridophytes. *Ann. Bot.* 90, 209–217
28 Brenner, S. *et al*. (1993) Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature* 366, 265–268
29 Aparacio, S. *et al*. (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297, 1301–1310
30 Hedges, S.B. and Kumar, S. (2002) Vertebrate genomes compared. *Science* 297, 1283–1285
31 Lewis, R. (2002) Pufferfish genomes probe human genes. *The Scientist* 16, 22–23
32 Deloukas, P. *et al*. (2001) The DNA sequence and comparative analysis of human chromosome 20. *Nature* 414, 865–871
33 Chatterjee, A. and Roux, S.J. (2000) *Ceratopteris richardii*: a productive model for revealing secrets of signaling and development. *J. Plant Growth Regul.* 19, 284–289
34 Schaefer, D.G. (2002) A new moss genetics: targeted mutagenesis in *Physcomitrella patens*. *Annu. Rev. Plant Biol.* 53, 477–501
35 Gray, M.W. *et al*. (1999) Mitochondrial evolution. *Science* 283, 1476–1481
36 Lang, B.F. *et al*. (1999) Mitochondrial genome evolution and the origin of eukaryotes. *Annu. Rev. Genet.* 33, 351–397
37 Stewart, C-B. (1995) Active ancestral molecules. *Nature* 374, 12–13
38 Jermann, T.M. *et al*. (1995) Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* 374, 57–59
39 Simpson, C.L. and Stern, D.B. (2002) The treasure trove of algal chloroplast genomes. Surprises in architecture and gene content, and their functional implications. *Plant Physiol.* 129, 957–966
40 Martin, W. *et al*. (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl. Acad. Sci. U. S. A.* 99, 12246–12251
41 Palenik, B. (2002) The genomics of symbiosis: hosts keep the baby and the bath water. *Proc. Natl. Acad. Sci. U. S. A.* 99, 11996–11997
42 Petrov, D.A. (2001) Evolution of genome size: new approaches to an old problem. *Trends Genet.* 17, 23–28
43 Pennisi, E. (2001) Genome duplications: the stuff of evolution? *Science* 294, 2458–2460
44 Cronk, Q.C.B. (2001) Plant evolution and development in a post-genomic context. *Nat. Rev. Genet.* 2, 607–620
45 Vision, T.J. *et al*. (2000) The origins of genomic duplication in *Arabidopsis*. *Science* 290, 2114–2117
46 Doolittle, R.F. (2002) Microbial genomes multiply. *Nature* 416, 697–700
47 Bennetzen, J. (2002) Opening the door to comparative plant biology. *Science* 296, 60–63
48 Hall, A.E. *et al*. (2002) Beyond the *Arabidopsis* genome: opportunities for comparative genomics. *Plant Physiol.* 129, 1439–1447
49 Couzin, J. (2002) NSF's ark draws alligators, algae, and wasps. *Science* 297, 1638–1639
50 Pennisi, E. (2002) Comparative biology joins the molecular age. *Science* 296, 1792–1795
51 Macilwain, C. (2001) Fears for basic science as Bush backs use of investment criteria. *Nature* 413, 5
52 Turmel, M. *et al*. (2002) The complete mitochondrial DNA sequence of *Mesostigma viride* identifies this green alga as the earliest green plant divergence and predicts a highly compact mitochondrial genome in the ancestor of all green plants. *Mol. Biol. Evol.* 19, 24–38