The Plant Cell, Vol. 18, 1100–1104, May 2006, www.plantcell.org © 2006 American Society of Plant Biologists

COMMENTARY

Comparative Sequencing of Plant Genomes: Choices to Make

The first sequenced genome of a plant, *Arabidopsis thaliana*, was published <6 years ago (Arabidopsis Genome Initiative, 2000). Since that time, the complete rice genome (*Oryza sativa*; Goff et al., 2002; Yu et al., 2002; International Rice Genome Sequencing Project, 2005) and a draft sequence of the poplar genome (*Populus trichocarpa*; http://genome.jgi-psf.org/Poptr1/Poptr1.home.html) have also been completed. In addition, the National Center for Biotechnology Information Entrez Genome Projects website reports that sequencing of several more plant genomes is in progress. The first wave of plant genome sequencing has passed, and we are now entering a new era in plant genomics research. Many of the obvious candidates for genome sequencing, model species with small genomes or species of economic importance, have either already been completed or are underway. The next round of choices should be made as part of a coherent strategy based on a mixture of scientific and economic needs and should recognize the value of including phylogenetic position as a selection criterion. There are also new technologies that will change the way we approach future genome sequencing projects. How are we to make appropriate choices with regards to both the species targets and the sequencing technologies we will use?

WHAT SPECIES DO WE SEQUENCE?

In a world with >260,000 known plant species and finite sequencing resources, it is crucial that we make careful choices as to which genomes will be sequenced. It is clear that the selection of sequencing targets will critically affect what we can learn from comparative plant genome sequencing. There are three areas that we feel will benefit substantially from future genome sequencing efforts. First, comparative genome sequences present opportunities to study the evolution of plant genome structure and the dynamics of molecular evolutionary processes. Second, they offer an approach to identify genes and other functional elements and provide critical data for annotation of completed plant genomes. Third, plant genome sequences provide the community with an important tool to pursue gene isolation in new target species. Given these scientific benefits, several key aspects of species choice should be considered.

Phylogenetic Placement of Target Species

The two species with completed genome sequences, *A. thaliana* and *O. sativa*, last shared a common ancestor ∼150 to 200 million years ago. To date, other species chosen for sequencing have been selected either for their small genomes or because of the interests of a particular research community. As a consequence, they are positioned almost idiosyncratically in the phylogeny of land plants (Figure 1). The 23 land plant species whose genomes have or are currently being sequenced are representative of only 13 of the 606 extant plant families. This imbalance in phylogenetic distribution is even more acute in the magnoliids, basal angiosperms, and the euasterid II lineage where there are no genome projects (Figure 1). This will have the unfortunate effect of biasing evolutionary comparisons as well as leaving certain groups without sequenced relatives that can be used in molecular genetic and genomic investigations.

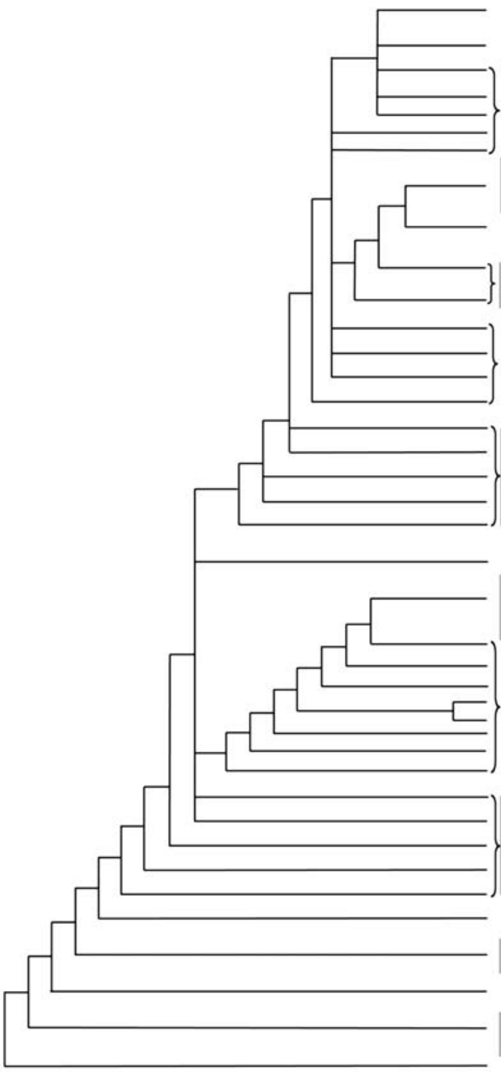
We feel that one goal of plant genome sequencing efforts should be to select genomes in a more systematic fashion with respect to the relationships of plant groups. Ideally, the number and distribution of sequenced genomes should be optimized to allow investigators to examine the evolution of genome structure and function within the context of a robust, well-defined phylogeny. Moreover, it would be desirable to have genomes sequenced in sufficient numbers to allow investigators working on any plant species to have access to information from the sequenced genome of an evolutionarily proximate species.

We can calculate the number of plant taxon lineages as a function of divergence times (Figure 2) based on a comprehensive phylogeny of 374 plant families whose branch points have been dated through molecular clock methods (Davies et al., 2004). From this, it is evident that sequencing representatives of ∼350 plant lineages in the phylogeny would provide sequenced genomes that are at most 20 million years divergent from any other angiosperm species. One can also adopt a hybrid sampling strategy of phylogenetically broad sampling for maximum coverage across the plant phylogenetic tree coupled with intense sampling of closely related species (see below) within selected plant groups.

Even smaller numbers of phylogenetically well-chosen species would have greater impact on comparative plant genomic studies than haphazard sampling across plant groups. For example, one can start by sampling species that belong to supraordinal groups that are yet to be covered by sequencing projects, such as the euasterid II group or other asterids (Figure 1). Sequencing basal angiosperm species as well as monocot species outside of the commelinids would also be appealing. Representatives of other land plant groups, including ferns, gymnosperms, and other bryophyte groups, such as the liverworts, will also be valuable in evolutionary comparisons. Moreover, sequence information from green algal groups that are sister to the land plants, the Charales and Coleochaetales, may help illuminate the evolutionary process that led to the transition of plants onto land.

Closely Related Species

Comparison of genome sequences across large spans of evolutionary time offers



	Order	Family	Species	Represented Families
Eurosid I (Fabids)	8	64	*Glycine max, Lotus japonicus, Medicago, Populus trichocarpa*	2
Eurosid II (Malvids)	3	34	*Arabidopsis lyrata, A. thaliana, Brassica oleracea, B. rapa, Capsella rubella, Carica papya*	2
Other Rosids	5	35	*Eucalyptus globulus, Vitis vinifera*	2
Euasterid I (Lamiids)	4	40	*Lycopersicon esculentum, Mimulus guttatus, Solanum tuberosum, Triphysaria versicolor*	3
Euasterid II (Campanulids)	4	35		0
Other Asterids	2	30		0
Other core eudicots	4	24		0
Early diverging dicots	2[a]	12		0
Magnoliids	4	18		0
Commelinids	3[b]	29	*Musa acuminata, Oryza sativa, Sorghum bicolor, Triticum aestivum, Zea mays*	2
Other monocots	8	53		0
Basal Angiosperms	1[c]	5		0
Gymnosperms	4	11		0
Ferns	16	55		0
Lycophytes	3	3	*Selaginella moellendorfii*	1
Bryophytes	13	158	*Physcomitrella patens*	1
Green Algae			*Chlamydomonas reinhardtii, Micromonas pusilla, Ostreococcus tauri*	
Total	**84**	**606**		**13**

a - Also includes 3 unplaced families
b - Also includes 1 unplaced family
c - Also includes 4 unplaced families

Figure 1. Phylogenetic Distribution of Species with Sequenced Genomes or with Ongoing Whole-Genome Sequencing Projects.

The numbers of orders and families in each group are tabulated on the right, with the specific species and the number of represented plant families also indicated. The phylogeny was based on Nickrent et al. (2000), Chase (2004), and Soltis and Soltis (2004).

glimpses into the macroevolution of genome structure. Insights, however, can also be gleaned by examining the diversification of genome structure at smaller evolutionary timescales. By selecting closely related species that diverged <20 million years ago, we may be able to understand early processes of genome evolution that would not be apparent from more distant comparisons. By sampling three related species at once, we can also polarize molecular changes along phylogenetic branches, showing which features of the genome are ancestral and which are derived within a group.

COMMENTARY

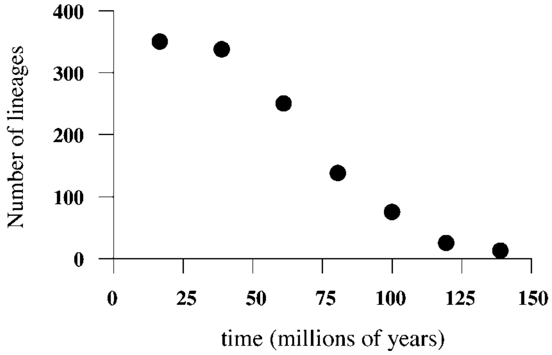


Figure 2. Number of Evolutionary Lineages in a Phylogeny of Flowering Plant Families.

The data are taken from a phylogeny reported by Davies et al. (2004). The horizontal axis gives time from the present, while the vertical axis gives the number of branches at specific evolutionary time points.

There are already several efforts underway along these lines. The sequencing of the *Arabidopsis lyrata* and *Capsella rubella* genomes, as well as the *Brassica oleracea* and *Brassica rapa* genomes, provides good coverage in the Brassicaceae. Other groups will also benefit from sequencing of closely related species. The most obvious is the relatives of *O. sativa* (rice) in the genus *Oryza*, including the 12 species for which physical maps and BAC library resources have already been developed and for which genome sizes are relatively modest (Ammiraju et al., 2006).

Resequencing Old Genomes

A third level of evolutionary scale, that of intraspecific variation, can offer yet another powerful evolutionary comparison as well as offering an invaluable resource for functional genomics studies. Data at this microevolutionary scale (<500,000 years) will result in a rich data resource for population genomic studies and help us understand in unprecedented detail the intraspecific patterns of genome evolution. Whole genome resequencing from multiple individuals also permits a more comprehensive identification of single nucleotide polymorphisms (SNPs) in the genome, similar to benefits that have accrued from the human Haplotype Map project (Altshuler et al., 2005). Resequencing the genomes from multiple individuals or plant accessions represents a significant improvement over the SNP identification procedure, which relies on polymorphisms identified from two individuals, resulting in biases in SNP genotyping studies. The development of more comprehensive SNP and molecular marker databases will provide an extensive number of genotype markers that can be used in whole-genome linkage disequilibrium and candidate gene association studies as well as in positional cloning efforts.

With this in mind, resequencing efforts of 20 individuals are already underway for *A. thaliana* and *O. sativa* using chip sequencing technology. Resequencing efforts will add value to the reference sequence by uncovering the wealth of genomic variation. Even resequencing spaced genes or gene fragments, which has been done in *A. thaliana* at the whole-genome level (Nordborg et al., 2005) and in targeted genomic regions (Cork and Purugganan, 2005) provides information useful in inferring the patterns and process of genomic evolution.

Economic Considerations

The U.S. National Plant Genome Research Program has had a strong, almost exclusive, emphasis on economically important plant species. Although it is likely that crop plant species will remain a high priority in future genome sequencing efforts, there are opportunities to widen the scope of these efforts.

First, genomes of non-crop species can have an enormous impact on the plant scientific community, thus indirectly benefitting crop plant research. For example, the *A. thaliana* genome sequence has been invaluable in plant molecular genetic and developmental studies.

Second, wild relatives of crops are a well-known source of allelic variation for desirable agronomic traits. Thus, there is an opportunity to exploit genes not only from single crops but also from species complexes that include both crop plants and close wild relatives. Sequencing wild relatives of crop species would afford a unique ability to dissect the process of crop domestication, which is probably the most significant technological innovation in human history. The availability of genome sequences for wild relatives of crop species alsowould provide a resource for ecological and evolutionary researchers interested in natural (as opposed to agricultural) plant systems.

Finally, crop species under consideration for future sequencing efforts can be expanded beyond the traditional commodity crops. There are 6000 plant taxa that are considered crops by various cultures, and a large number of these are found in developing countries. While most of these species may not be commodity crops, a large number of people depend on these alternative crops for food and other necessities, particularly in resource-poor environments. An example is cassava, which feeds ~600 million people in sub-Saharan Africa, is a primary calorie source for >200 million people, and whose genome size of 760 Mb makes it an appealing candidate for sequencing efforts (Raven et al., 2005). Expanding genomic science resources to these orphan crop species may prove to have a large impact on global human welfare.

HOW DO WE SEQUENCE?

Beyond species selection, future genome projects must also select strategies and technologies that are appropriate for their goals and budgets. Greater choices in sequencing technologies will play a major role in enabling progress in comparative plant sequencing. While extremely successful, Sanger sequencing does present some significant challenges for obtaining large increases in throughput and speed beyond those seen over the last decade. It seems unlikely with current sequencing technologies that many more plant genomes will be sequenced to the level of completeness found in the *O. sativa* and *A. thaliana* genomes, although we should not discount the rapid advances in sequencing technologies that may allow for facile generation of assembled, completed genome sequences. This should not necessarily be a problem, since the incremental benefit between a draft and complete genome sequence lessens as more genomes are sequenced.

Several new sequencing technologies are emerging that have the potential to provide increases in throughput and reductions in cost (Metzker, 2005). Companies such as 454 Life Sciences, Solexa, and Helicos Biosciences all have competing technologies, vying to be widely adopted for the next generation of sequencing machines (Bennett et al., 2005; Margulies et al., 2005). A detailed review and evaluation of each technology is beyond the scope of this commentary, but there are two general concepts that these approaches share. First, each allows a single template molecule to be used to generate many bases of the sequence read, instead of the irreversible dideoxy termination of Sanger sequencing. Secondly, because the sequence is not represented by a ladder of differentially sized fragments, they can avoid electrophoretic steps to isolate and read the sequence.

Currently, the major issues these technologies face are short read lengths, inability to generate paired-end sequence reads, and uncertain quality metrics; the former two problems are serious limitations in generating a completely assembled sequence for large genomes. Combining traditional sequencing approaches with one of these new technologies can potentially offer a valuable middle ground. By having sufficient paired-end reads of cloned genomic fragments using Sanger sequencing, a sparse scaffold of the genome could be created that could subsequently be filled in with deeper coverage from short, unpaired reads.

Moreover, some of the concerns of these present-day new technologies will fade as the technologies mature. There are also certain applications where these new approaches truly excel, such as in resequencing a previously sequenced reference genome (e.g., sequencing different rice varieties or sequencing several closely related *Arabidopsis* species).

We should not forget that comparative projects require a previously completed reference genome and that with each new plant genome that is fully sequenced (whether finished quality as in *Arabidopsis* or rice, or draft quality as in poplar), the opportunities for comparative analyses increase. While *Arabidopsis*, rice, and poplar provide a great starting point, the careful selection of future fully sequenced genomes is critical. These may turn out to be large, complex genomes that are unsuitable for the new technologies, but committing to each one will continue to increase the number of foci around which comparative analyses can cluster.

STARTING THE DEBATE

We have outlined possible criteria that we feel should drive the choice of target genomes to be sequenced. It is not our intention to be comprehensive in this discussion, but rather to provide a framework for the community to begin thoughtful debate on what needs to be done. The list of candidate species will no doubt be long, and the resources will be limited. Understanding how we can intelligently make these choices is vital if the field is to move forward and the promise of plant genomics is to be fulfilled.

ACKNOWLEDGMENTS

We thank Pam Soltis and Richard Michelmore for interesting and provoking discussions and Jennifer Modliszewski for technical assistance. This commentary is largely based on discussions at an informal workshop on comparative plant genome sequencing held at the Plant Genetics 2005 meeting sponsored by the American Society for Plant Biologists in October, 2005. However, we are solely responsible for the views expressed in this commentary. This work was funded in part by grants from the National Science Foundation Plant Genome Research Program to S.J. and M.P. and by the National Science Foundation Frontiers in Integrative Biological Research and Department of Defense to M.P.

Scott Jackson
Department of Agronomy
Purdue University
West Lafayette, IN 47907

Steve Rounsley
Broad Institute of MIT and Harvard
Cambridge, MA 02141

Michael Purugganan
Department of Genetics
North Carolina State University
Raleigh, NC 27695
michaelp@unity.ncsu.edu

REFERENCES

Altshuler, D., Brooks, L.D., Chakravarti, A., Collins, F.S., Daly, M.J., and Donnelly, P.; International HapMapConsortium. (2005). A haplotype map of the human genome. Nature **437,** 1299–1320.

Ammiraju, J.S., et al. (2006). The *Oryza* bacterial artificial chromosome library resource: Construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus *Oryza.* Genome Res. **16,** 140–147.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana.* Nature **408,** 796–815.

Bennett, S.T., Barnes, C., Cox, A., Davies, L., and Brown, C. (2005). Toward the $1000 human genome. Pharmacogenomics **6,** 373–382.

Chase, M.W. (2004). Monocot relationships: An overview. Am. J. Bot. **91,** 1645–1655.

Cork, J.M., and Purugganan, M.D. (2005). High-diversity genes in the Arabidopsis genome. Genetics **170,** 1897–1911.

Davies, T.J., Barraclough, T.G., Chase, M.W., Soltis, P.S., Soltis, D.E., and Savolainen, V. (2004). Darwin's abominable mystery: Insights from a supertree of the angiosperms. Proc. Natl. Acad. Sci. USA **101,** 1904–1909.

Goff, S.A., et al. (2002). A draft sequence of the rice genome (*Oryza sativa L. ssp. japonica*). Science **296,** 92–100.

COMMENTARY

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. Nature **436,** 793–800.

Margulies, M., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature **437,** 326–327.

Metzker, M. (2005). Emerging technologies in DNA sequencing. Genome Res. **15,** 1767–1776.

Nickrent, D.L., Parkinson, C.L., Palmer, J.D., and Duff, R.J. (2000). Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. Mol. Biol. Evol. **17,** 1885–1895.

Nordborg, M., et al. (2005). The pattern of polymorphism in *Arabidopsis thaliana.* PLoS Biol. **3,** 1289–1299.

Raven, P., Fauquet, C., Swaminathan, M.S., Borlaug, N., and Samper, C. (2005). Where next for genome sequencing? Science **311,** 468.

Soltis, P.S., and Soltis, D.E. (2004). The origin and diversification of angiosperms. Am. J. Bot. **91,** 1614–1626.

Yu, J., et al. (2002). A draft sequence of the rice genome (*Oryza sativa L. ssp. indica*). Science **296,** 79–92.