Comparative Plant Genomics. Frontiers and Prospects

Ana L. Caicedo and Michael D. Purugganan*

Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695

Comparative methods have long been the cornerstone of studies that draw inferences about function and evolution at various levels of biological organization. The availability of whole-genome sequences as well as other genomic resources (e.g. microarray methods, expressed sequence tag [EST] libraries, high-throughput resequencing technologies) has allowed us to extend the comparative method to encompass the evolution of genome structure and function. More than just an isolated field, the past few years have witnessed the emergence of comparative genomics as a tool to address questions in diverse areas of biological research.

THE EVOLUTION OF GENOME STRUCTURE

Characterization of genomes, including whole-genome sequences, has traditionally revealed numerous species-specific details, including genome size, gene number, patterns of sequence duplication, a catalog of transposable elements, and syntenic relationships (e.g. The Arabidopsis Genome Initiative, 2000; Goff et al., 2002; Yu et al., 2002). These studies have underscored the diverse architectures of plant genomes. At present, however, we continue to know very little about the evolutionary dynamics of changes in genome structure and their consequences on gene content and evolution. As more whole-genome sequences at both intraspecific and interspecific levels become available, we will be in a position to address several key issues surrounding the evolution of genome architectures. These include the extents and rates of change in genome structure and size, patterns of large-scale genome duplications (including polyploidization), the dynamics of the origins and extinctions of genes, the role of selection acting on large-scale variation in genome structure and organization, the evolutionary forces that determine transposable element activity and number and the functional consequences of these mobile elements, and the extent and impact of epigenetic markings on genome and organismal evolution.

THE EVOLUTION OF GENOME FUNCTION

Genomic tools have provided a boon for researchers seeking to understand the functional roles of genes and their evolutionary histories. Especially useful has been the appearance of genome-based methods to identify genomic regions of functional importance. The availability of intraspecific whole-genome sequences (such as for two subspecies of rice [*Oryza sativa*] and the Columbia and Landsberg *erecta* ecotypes of *Arabidopsis thaliana*) can reveal single nucleotide polymorphism genomic regions with markedly low or high levels, possible indicators of positive or balancing selection, both of which are signatures of adaptive evolution (Nielsen, 2001). The mark of selection on candidate loci identified in this manner can then be verified by sampling more individuals within the species. The occurrence of intraspecific variation for phenotypic traits of interest also permits the identification of genes responsible for these traits by searching for associations among naturally occurring genome-wide polymorphisms using linkage disequilibrium mapping, a technique whose viability is SNP currently being actively investigated in plants (Hagenblad et al., 2004). When candidate genes are already known, genomic variation data can provide the necessary controls for phenotypic association studies.

Over the last several years, we have begun to realize that the products of genes are embedded in large-scale interaction networks that represent integrated functional units at the molecular genetic level. Thus, to understand the evolution of function, it becomes necessary to understand the evolutionary dynamics of molecular genetic networks (Cork and Purugganan, 2004). Expression profiling with microarrays combined with EST and genome sequence data offer a viable way to do so, allowing us to identify the interacting candidate genes of genetic networks and examine how the patterns of interactions change across species. What is lacking is a solid theoretical framework for network structure and evolution, and work in this exciting area should be forthcoming.

POPULATION GENOMICS AND PHYLOGENOMICS

A key tenet of evolutionary genetics is that natural selection affects single genes or gene regions, but population processes, such as gene flow, range expansion, or bottlenecks, leave their imprint on all genes in the genome. Data for genome-wide polymorphisms

* Corresponding author; e-mail michaelp@unity.ncsu.edu; fax 919–515–3355.

www.plantphysiol.org/cgi/doi/10.1104/pp.104.900148.

for individuals of a species can now be easily obtained with high-throughput methods and is not limited to organisms with sequenced genomes. We are now faced with unprecedented amounts of genome information with which to characterize population history and structure. Quantification of levels of intraspecific genome variation also aids in the identification of loci under selection, which exhibit patterns of variation divergent from those in the rest of the genome (Luikart et al., 2003). As the level of structural variation in intraspecific genomes becomes more apparent, theory will have to be developed to describe genomic phenomena, such as the evolutionary dynamics of gene families and transposable elements, in a population context.

A consequence of the proliferation of genome studies has been the documentation of patterns of genome variation between species, information that can be used to assist in the construction of the Tree of Life, including plants. Given the evolutionary distances between many organisms targeted by whole-genome and EST sequencing studies, such data is more likely to aid in the characterization of deep phylogenetic nodes of plant phylogeny. Moreover, the extensive genome coverage of the data can also be used to explore molecular clocks along phylogenetic branches (Miller et al., 2004) as well as to understand the genomic changes characteristic of major evolutionary lineages (Bennetzen, 2002; Haubold and Wiehe, 2004). This will allow us to advance toward a central goal of plant biology: the reconstruction of the evolutionary history of all extant plant groups.

COMPARATIVE METHODS AND THE FUNCTIONAL ANNOTATION OF GENOMES

Identification of functional regions in genomes can be carried out by searching for conservation among genome sequences, as functional regions are believed to be under stabilizing selection and should be preferentially conserved over evolutionary time. This approach has been used successfully in the annotation of animal and yeast genomes (for review, see Miller et al., 2004), but has not yet been used extensively in plants (only two complete and quite distant plant genomes are currently available). Both coding and regulatory regions can be identified by locating genomic areas under purifying selection, although regulatory regions tend to be less conserved and pose greater challenges to the comparative genomics approach. Comparisons at varying levels of evolutionary divergence are likely to reveal functional regions characteristic of different plant groups; even intraspecific genomic approaches have been shown to be useful in predicting functional sequence motifs (Boffelli et al., 2004).

The reliability and usefulness of comparative genomics for genome annotation will depend on the continuous improvement of predicting algorithms, as well as our improving characterization of the varying neutral evolutionary rates across sequenced genomes (Miller et al., 2004). Comparisons among multiple species, although computationally intense, have also been shown to be a powerful method in the prediction of functional genomic regions (Thomas et al., 2003). The success of such approaches for plant genome annotations will hinge on the completion of sequenced genomes for plants of all major evolutionary lineages.

PERSPECTIVES

Comparative genomics has proven an invaluable approach to understanding biology, not only for dissecting patterns and processes of genome evolution but also in revealing aspects of gene function. The rapid advances in technology, both for sequencing and for determining expression and interaction patterns, will continue to propel this area in the future.

Although it is as yet unreasonable to expect that everybody's favorite organism will be sequenced to completion, the plant research community as a whole would benefit from candidate genomes chosen within a reasonable phylogenetic framework. Ideally, this would include candidates from non-seed plant lineages, gymnosperms, and major angiosperm lineages, and steps for at least comparative EST analysis in this regard are under way. Maximum benefit could be derived from applying Bennetzen's (2002) suggestion to sequence pairs of organisms for each selected lineage. Intraspecific genome comparisons will primarily rely on resequencing techniques in the near future, though the advent of chip-based genomic array techniques (Borevitz and Ecker, 2004) as well as new methods will make it easier to acquire large genome coverage for individuals within populations or species. These tools, coupled with functional genomics approaches, may provide crucial insights into how genomes evolve in structure and function, and also permit us to link genome structure with organismal biology.

LITERATURE CITED

Bennetzen J (2002) Opening the door to comparative plant biology. Science **296:** 60–63

Boffelli D, Weer CV, Weng L, Lewis KD, Shoukry MI, Pachter L, Keys DN, Rubin EM (2004) Intraspecies sequence comparisons for annotating genomes. Genome Res **14:** 2406–2411

Borevitz JO, Ecker JR (2004) Plant genomics: the third wave. Annu Rev Genomics Hum Genet **5:** 443–477

Cork JM, Purugganan MD (2004) The evolution of molecular genetic pathways and networks. Bioessays **26:** 479–484

Goff SA, Ricke D, Lan TH, Presting G, Wang RL, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science **296:** 92–100

Hagenblad J, Tang CL, Molitor J, Werner J, Zhao K, Zheng HG, Marjoram P, Weigel D, Nordborg M (2004) Haplotype structure and phenotypic associations in the chromosomal regions surrounding two *Arabidopsis thaliana* flowering time loci. Genetics **168:** 1627–1638

Haubold B, Wiehe T (2004) Comparative genomics: methods and applications. Naturwissenschaften **91:** 405–421

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet **4:** 981–994

Miller W, Makova KD, Nekrutenko A, Hardison RC (2004) Comparative genomics. Annu Rev Genomics Hum Genet **5:** 15–56

Nielsen R (2001) Statistical tests of selective neutrality in the age of genomics. Heredity **86:** 641–647

Thomas JWT, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, Mcdowell JC, Maskeri B, et al (2003) Comparative analyses of multi-species sequences from targeted genomic regions. Nature **424:** 788–793

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

Yu J, Hu SN, Wang J, Wong GKS, Li SG, Liu B, Deng YJ, Dai L, Zhou Y, Zhang XQ, et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296:** 79–92