A Genomics Approach to the Study of Ancient Polyploidy and Floral Developmental Genetics

JAMES H. LEEBENS-MACK,*,‡ KERR WALL,* JILL DUARTE,* ZHENGUI ZHENG,† DAVID OPPENHEIMER† AND CLAUDE DEPAMPHILIS*

Department of Biology, Institute of Molecular Evolutionary Genetics, and Huck Institutes of Life Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802
†Department of Botany and the Genetics Institute, University of Florida, Gainesville, Florida 32611*

‡Present Address: Department of Plant Biology, University of Georgia, Athens, Georgia 30602.

Advances in Botanical Research, Vol. 44
Incorporating Advances in Plant Pathology
Copyright 2006, Elsevier Ltd. All rights reserved.
0065-2296/06 $35.00
DOI: 10.1016/S0065-2296(06)44014-3

ABSTRACT

Comparative genomics approaches are proving to be extremely valuable for the study of gene function, gene duplications, and genome evolution. In this chapter we discuss how cross-species comparisons of gene sequences and gene-expression patterns are elucidating the evolution of many plant processes including the regulation of reproduction. Emphasis is placed on the implications of gene and genome duplications for the evolution of genome structure and plant reproduction. In addition, we show that comparative analyses can both promote transfer of knowledge from model to non-model systems and inform our understanding of conserved processes in model species.

I. INTRODUCTION

As has been discussed in each chapter in this volume, much of our current understanding of flower development has been informed by cross-species comparative investigations (Albert *et al.*, 1998; Becker *et al.*, 2000; Coen and Meyerowitz, 1991; Ma and dePamphilis, 2000). This work is built on a strong foundation of forward genetics (see Davies *et al.*, Chapter 7; Irish, Chapter 3; Kramer and Zimmer, Chapter 9; Zahn *et al.*, Chapter 4,) and a growing understanding of the phylogenetic relationship among plant lineages with contrasting floral morphologies (Endress, Chapter 1; Soltis *et al.*, 2005; Zanis *et al.*, 2003). In recent years, genome and transcriptome analyses have also added to our understanding of genes involved in the regulation of flowering time (Schmid *et al.*, 2003) and floral development (Albert *et al.*, 2005; Laitinen *et al.*, 2005; Wellmer *et al.*, 2004; Zik and Irish, 2003). In this chapter, we discuss the comparative genomics approach as a useful way of identifying genes and noncoding sequences that may be involved in floral development. We also discuss the utility of comparative genomics for testing properly framed hypotheses. Finally, we consider new high-throughput technologies that promise to expand the scope and impact of comparative genomics.

A. PHYLOGENETIC CONTEXT

The improving resolution of phylogenetic relationships among plant lineages (Fig. 1) is providing the historical context necessary to understand events associated with the origin and diversification of seed plants (Burleigh and Mathews 2004), angiosperms (Davies *et al.*, 2004; Leebens-Mack *et al.*, 2005; Qiu *et al.*, 2005; Zanis *et al.*, 2002) and specific flowering plant lineages (Beilstein *et al.*, 2006 for the Brassicaceae; Malcomber *et al.*, Chapter 11 for the Poaceae). These advances have paved the way for comparative genomic studies aimed at understanding associations between organismal diversification and genome evolution (Bowers *et al.*, 2003; Buzgo *et al.*, 2005;

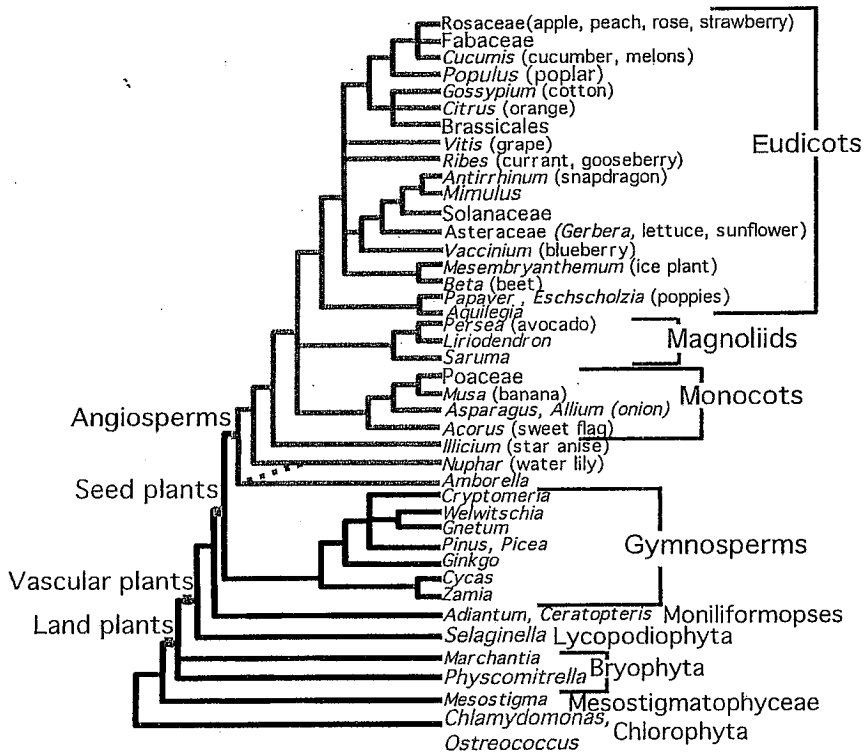

Fig. 1. Phylogenetic distribution of green plant taxa with large cDNA sequence sets (ESTs and unigenes) and genome sequencing projects (highlighted) complete or in progress. Topology and taxonomy are taken from Peter Stevens' angiosperm phylogeny website (http://www.mobot.org/MOBOT/research/APweb/), Pryer *et al.*, 2001 and Marin and Melkonian (1999).

Kellogg and Bennetzen, 2004; Paterson *et al.*, 2004; Soltis *et al.*, 2002; Vandepoele and Van de Peer, 2005; Zahn *et al.*, 2005a).

The analytical and conceptual tools of comparative genomics can be applied to questions pertaining to the entire continuum of evolutionary time scales. The precise question that can be addressed most effectively through comparative genomics varies depending on the degree of divergence among the taxa being compared (Fig. 2). Whereas comparisons of closely related species and intraspecific polymorphism can help identify genes or quantitative trait loci (QTL) associated with specific phenotypic differences (Aranzana *et al.*, 2005; Lexer *et al.*, 2005) including patterns of gene expression (eQTL; Doerge, 2002; Schadt *et al.*, 2003), facilitate positional cloning (Bortiri *et al.* 2006), and aid investigation of mechanisms responsible for speciation (Hey *et al.*, 2005 and other papers in this issue of PNAS devoted

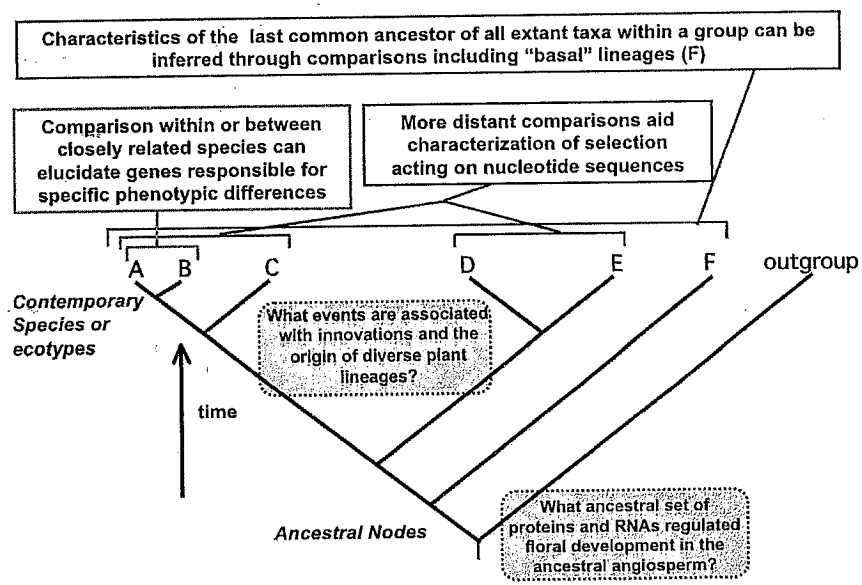

Fig. 2. Taxon sampling for comparative genomic analyses depends on the questions to be addressed. Sampling within species or among closely related species can elucidate the genetic basis of phenotypic differences, while sampling among more divergent species is necessary to investigate events associated with the origin and diversification of ancient groups such as the eudicots, monocots, angiosperms, or seed plants (after Hardison, 2003; dePamphilis, 1995).

to Ernst Mayr; Sweigart *et al.*, 2006) and domestication (Burke *et al.*, 2002; Clark *et al.*, 2004; Nesbitt and Tanksley, 2002; Yamasaki *et al.*, 2005), comparisons of more divergent genomes are useful for identifying conserved noncoding sequences that may have regulatory functions (Eddy, 2005; Hardison, 2003; Odenwald *et al.*, 2005; Siepel *et al.*, 2005). Understanding genetic events associated with the origin of ancient groups ranging from the grass family, or core eudicots, to all flowering plants, seed plants, or land plants also requires comparisons of increasingly divergent genomes (Soltis *et al.*, 2002). In this chapter, we describe a few examples of how comparative genomics research at each of the levels depicted in Fig. 2 has added greatly to our understanding of plant reproductive biology.

B. GENOMIC APPROACHES

"Genomic approaches" are typically high-throughput methods that provide a view of genetic variation across gene families, transcriptomes, genomic regions, or whole genomes. One may apply genomic approaches to test hypotheses and elucidate biological processes all along this continuum.

Transcriptome sequencing [e.g., expressed sequence tag (EST) sequencing], massively parallel signature sequencing (MPSS; Brenner *et al.*, 2000), microarray analyses, use of functional tools including targeting local lesions in genomes (TILLING; reviewed by Comai and Henikoff, 2006) or virus-induced gene silencing (VIGS; reviewed by Burch-Smith *et al.*, 2004), and genomic sequencing are just some of the high-throughput techniques that can yield data for comparative genomic analyses. In this chapter, we focus primarily on comparative genomic analyses of sequence and gene expression data aimed at understanding various aspects of plant reproduction.

Appropriate analyses of genomic data are developed in order to address questions of interest, and the field of bioinformatics has emerged and is growing in response to the demands that come with staggering increases in the amount of genomic data. Of particular importance for comparative genomics is the development of searchable databases with cDNA sequences (Albert *et al.*, 2005; Dong *et al.*, 2005; Lee *et al.*, 2005; Rudd, 2005) and repeat element sequences (Ouyang and Buell, 2004) for multiple species. Powerful phylogenomic analysis pipelines have been used to efficiently mine sequence databases such as these, construct alignments, and build gene family phylogenies (Chapman *et al.*, 2004; Hartmann *et al.*, 2006; Sjölander, 2004). Comparative analyses of data extracted from sequence and gene family databases are contributing to advances in plant reproductive biology, and this trend will continue as the volume of data rapidly increases and more investigators are trained how to build analysis pipelines and tailor them to specific research questions.

II. WIDESPREAD POLYPLOIDY IN ANGIOSPERM HISTORY

Botanists have long understood that polyploidy has been an important force in angiosperm history (Grant, 1981; Soltis, 2005; Soltis and Soltis, 1999; Stebbins, 1950). Analyses of chromosome numbers have suggested that many extant angiosperms are ancient polyploids (Grant, 1963; Otto and Whitton, 2000). Despite the small size of the *Arabidopsis thaliana* genome (157 Mb/C; Bennett *et al.*, 2003), a striking observation from early analyses of these data was that much of the genome consisted of large duplicated segments, suggesting a history of repeated rounds of ancient polyploidy (Blanc *et al.*, 2000; Bowers *et al.*, 2003; Simillion *et al.*, 2002; Vision *et al.*, 2000). The number and timing of genome duplication events came into better focus when the duplicated blocks were analyzed in the context of sequence data from pine species, monocots, asterids, and other rosids (Blanc *et al.*, 2003; Bowers *et al.*, 2003). Bowers *et al.* (2003) performed high-throughput phylogenetic

analyses on genes found in duplicated blocks and inferred an ancient genome duplication event in the common ancestor of *Brassica* and *Arabidopsis*; a second event in the common ancestor of asterids and rosids, and possibly a third event predating the divergence of monocots and eudicots (Fig. 1). Ancient polyploidy is also evident as large duplicated blocks in the genome sequences of rice (*Oryza sativum*; Paterson *et al.*, 2004; Yu *et al.*, 2005) and *Populus trichocarpa* (Tuskan *et al.*, submitted for publication).

Analyses of EST data have implicated additional ancient polyploidization events throughout the angiosperms (Blanc and Wolfe, 2004; Cui *et al.*, 2006; Schlueter *et al.*, 2004). Following the earlier work of Lynch and Conery (2000), all-against-all BLAST searches (Altschul *et al.*, 1997) of large sets of coding sequences sampled from a species (e.g., EST or unigene sequences) can be used to identify putative paralog pairs, which can be aligned in coding frame. The number of synonymous changes per synonymous sites (K_s) is then estimated for each paralog pair alignment using yn00 or codeml in PAML (Yang 1997) or similar routines in HyPhy (Pond *et al.*, 2005), and the frequency distribution of K_s values can then be plotted. The underlying distribution of K_s plots is expected to reflect a background rate of gene duplication and extinction, with a peak near $K_s = 0$ and an exponentially decreasing frequency of duplicate gene pairs with increasing values of K_s (Fig. 3; Blanc and Wolfe, 2004). A secondary spike in the K_s distribution (Fig. 3) would only be expected if a large number of gene duplications occurred at the same time in the past. Therefore, a spike in the K_s distribution can be interpreted as indicating an ancient polyploidy event (including partial genome duplications; but see Hughes *et al.*, 2003) or a concerted increase in transposon activity (Hughes *et al.*, 2003). However, not all polyploidy events can be observed in K_s plots. Paralog pairs from polyploidy events may be indistinguishable from background single gene duplications or allelic variants ($K_s < 0.05$). Further, sampling error in the substitution process leads to increased variance in K_s with time (Fig. 3; Cui *et al.*, 2006). Finally, gene loss and incomplete sampling of a proteome may reduce the signal of ancient polyploidy in K_s plots. Therefore, whereas ancient polyploidy can be inferred from K_s plots, the absence of a spike in K_s, should not be interpreted as an absence of polyploidy in a species' ancestry.

Despite the limitations of K_s plots for inferring polyploidy, analyses of EST sequences from representatives of most major flowering plant lineages do provide evidence of frequent genome duplications throughout angiosperm history (Blanc and Wolfe, 2004; Cui *et al.*, 2006; Schlueter *et al.*, 2004). An understanding of ancient genome duplication in the basal-most angiosperm lineages is especially important for elucidating the role polyploidy may have played in the origin and early diversification of flowering plants (de Bodt *et al.*, 2005; Zahn *et al.*, 2005a,b). While K_s analyses of basal

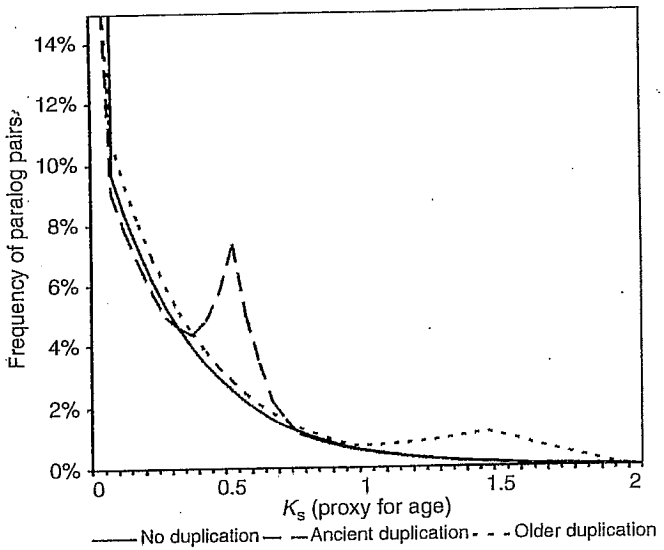

Fig. 3. The frequency distribution of K_s for paralog pairs can indicate ancient polyploidy events (e.g., secondary peaks shown for dashed lines), but not all genome duplications will be detected in K_s plots. Gene loss and increasing variance in K_s with time would both erode secondary peaks in these plots.

angiosperm lineages (Cui *et al.*, 2006) and phylogenomic analyses of duplicated blocks in the *Arabidopsis* genome (Bowers *et al.*, 2003; Sampedro *et al.*, 2005) provide evidence for paleopolyploidy in early angiosperm history, the precise timing of genome duplications relative to the origin of extant flowering plant lineages will require further investigation. At the same time, much research is now focused on the fate of duplicated genes (Adams *et al.*, 2003; Casneuf *et al.*, 2006; Chapman *et al.*, 2006; Maere *et al.*, 2005; Moore and Purugganan, 2005; others reviewed by Adams and Wendel 2005) and shifting function in retained duplicates (Duarte *et al.*, 2006; Force *et al.*, 1999; Lynch and Conery, 2000). Floral evolution has been linked to gene duplications (Irish, 2003, Chapter 3), and future research will investigate the apparent association between ancient polyploidy and innovations in plant reproduction.

III. IMPLICATIONS OF ANCIENT POLYPLOIDY FOR COMPARATIVE GENOMICS

A. ORTHOLOGS, HOMEOLOGS, AND PARALOGS

The discovery of widespread ancient polyploidy throughout the history of angiosperms complicates our understanding of orthology and paralogy in flowering plants. Orthologs and paralogs are defined as genes that originated

from speciation or duplication events, respectively (Sonnhammer and Koonin, 2002; Theissen, 2002) and homeologs (or paleologs) are paralogs that originate from genome duplication. However, if genome duplication is a recurrent phenomenon in angiosperms, then any two distantly related angiosperms will be separated by one or more genome duplications. For example, according to our understanding of the polyploid histories of lineages leading to *Arabidopsis* and *Oryza*, at least three genome duplications have occurred since these species shared a common ancestor: one in the early history of the Brassicaceae, one before the diversification of the major core eudicot lineages, and one in the early history of the Poaceae. Therefore, even genes with what appear to be simple orthologous relationships in *Oryza* and *Arabidopsis*—for example *LEAFY* and its rice "ortholog" *RFL*—are in fact the survivors of a complex history of duplication and loss of duplicate copies. Given this dynamic nature of plant genome histories, phylogenetic analyses of gene families must be performed on a genomic scale in order to address issues ranging from the prediction of gene and protein function (Eisen, 1998; Engelhardt *et al.*, 2005; Sjölander, 2004) to the influences of polyploidy on genome content and structure (Bowers *et al.*, 2003; Rong *et al.*, 2005), as well as the evolution of regulatory networks influencing floral development (see later section).

B. CHARACTERIZING THE FATE OF DUPLICATED GENES

The polyploid histories of flowering plant genomes provide global opportunities for selective expansion of specific kinds of genes. For example, regulatory genes (Blanc and Wolfe, 2004; Maere *et al.*, 2005), and genes that encode long complex proteins (Chapman *et al.*, 2006), may be more likely to survive genome duplication. Analyses performed by Maere *et al.* (2005) suggest that duplicate regulatory genes are more likely to be retained following polyploidy events relative to single gene duplications. In contrast, many other genes may be particularly resistant to retention of duplicates. Such genes have existed over long periods of time as singletons or low-copy genes in the face of whole-genome duplications, implying that selection against duplicate copies is more intense for these genes.

Chapman *et al.* (2006) presented a slightly different interpretation of single-copy genes (singletons). Focusing on adaptive retention of functionally redundant duplicate genes that may buffer critical functions in developmentally and genetically unstable polyploids, the authors present evidence that single-copy genes may simply be genes for which duplicate copies offer no selective advantage. Analysis of intraspecific single nucleotide substitution polymorphisms (SNPs) revealed that genes retained in duplicate following the

most recent polyploidizations in lineages leading to rice and *Arabidopsis* tended to have a lower ratio of amino acid replacement substitutions to nonreplacement substitutions (dN/ds) relative to singleton genes (Chapman *et al.*, 2006). This pattern implies that singletons evolve under *less* severe purifying selection than genes that have been retained as duplicates.

The intensity of purifying selection may not have anything to do with selection for retention or extinction of duplicate gene copies. Based on gene-clustering analyses (Enright *et al.*, 2002, 2003), we estimate that 727 strict ortholog sets exist as single-copy genes in the *Arabidopsis*, rice, and *Populus* genomes (Wall *et al.*, in preparation). This is a much larger number than would be expected if gene deaths were random following gene and genome duplications. The estimated frequencies of singletons in the *Arabidopsis*, rice, and *Populus* genomes are at most 15% (singletons/total gene number = 3862/26,207), 21% (11,954/57,915), and 12% (5396/45,555), respectively. The lineages leading to *Arabidopsis*, rice, and *Populus* have each experienced at least one genome duplication event that is independent of polyploidy events in the other lineages. Therefore, if gene extinctions were independent in these three lineages, we would expect the frequency of shared singletons to be the product of singleton frequencies in each genome multiplied by the number of genes in the smallest proteome (15% × 21% × 12% × 26,207 = 99). The expected value is less then one-seventh of the observed number of shared singletons, so gene deaths must not be random. This result could be explained, at least in part, by maintenance of duplicate genes (Chapman *et al.*, 2006), since the percentage of singletons among genes that were susceptible to extinction within each species would be higher if a fraction of duplicates were selectively maintained. However, even if one assumes that half of the genes are selectively maintained in duplicate, there are still many more shared singletons than expected. We surmise that a large fraction of shared singletons in the three sequenced angiosperm genomes are selectively maintained as such, and are not simply the survivors of random gene loss in the absence of selection for retention of duplicate gene copies. Selection for preservation of dosage balance could be an important force maintaining some genes in single-copy following gene and genome-duplication events.

C. A GENE FAMILY PERSPECTIVE ON GENOME DUPLICATIONS

Duplication patterns that are observed in phylogenetic studies of gene families provide a view of family history that can be interpreted in the context of known genome duplication events. For example, the *HUA* enhancer 2 family (*HEN2*) is part of a small family of putative DExH box RNA helicase enzymes (Western *et al.*, 2002). *HEN2* mutants display defects in

petal number and position as well as phyllotaxy and floral number (Western *et al.*, 2002). Parsimony analysis of members of this family (Fig. 4) identified in *Arabidopsis*, rice, and basal angiosperm sequences mined from the Floral Genome Project PlantTribes database (http://fgp.huck.psu.edu/tribe.php; Albert *et al.*, 2005), suggest that three or four clades were established before the early diversification of angiosperm lineages. Surprisingly, there is a single *Arabidopsis* gene and between zero and two rice homologs in each of these clades. Thus, while none of the *HEN2* family genes were counted among the singletons described in an earlier section, most of the duplicate genes that have been generated through multiple rounds of polyploidy in angiosperm history have not survived.

The utility of gene family analyses placed in the context of genome duplication events is also illustrated in an analysis of the expansin gene family (Sampedro *et al.*, 2005). Expansins are cell wall loosening proteins that exist as a multigene family in all plants (Cosgrove, 2005; Sampedro and Cosgrove, 2005).

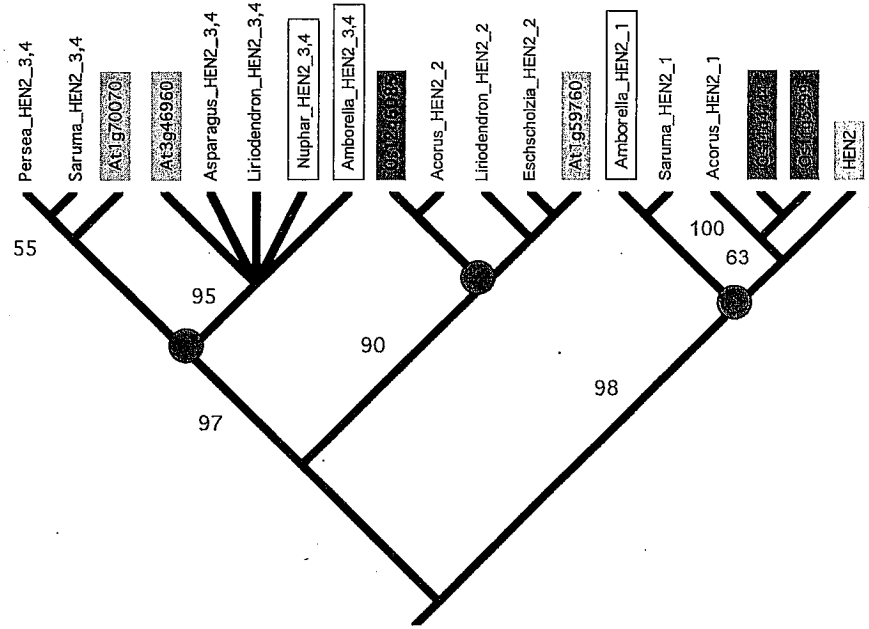

Fig. 4. A gene phylogeny for the *HEN2* RNA helicase gene family with single *Arabidopsis* genes in each of four clades. Each of the four clades include monocot, magnoliid, or basal-most angiosperm (outlined) genes, indicating at least two gene deaths in each *Arabidopsis* (light shading) gene lineage following the genome duplications in polyploid Brassicaceae and core eudicot ancestors. Gene deaths are also evident in the lineage leading to rice (dark shading) following a genome-wide duplication in the early history of the Poaceae. Bootstrap values are shown to the left or right of each branch when greater than 50%.

Although phylogenetic analysis was only partly able to resolve the history of the gene family, careful examination of duplicated blocks of genes in rice and *Arabidopsis* aided the resolution of the phylogeny and showed that nearly every member of the gene family in *Arabidopsis* could be accounted for as the product of genome duplications (Sampedro *et al.*, 2005). Thus, Sampedro *et al.* demonstrated that consideration of genomic context improves phylogenetic resolution of complex gene family histories.

As has been described in many other chapters in this volume, the MADS-box gene family has been intensively studied in terms of both gene duplication and functional diversification (Becker and Theissen, 2003; Irish, 2003, Chapter 3; Kim *et al.*, 2005; Kramer and Hall, 2005; Soltis *et al.*, Chapter 12; Zahn *et al.*, 2005a,b, 2006). Given the role of many MADS-box genes in the regulation of flowering time and floral organ specification, it has been hypothesized that gene duplications and subsequent functional shifts have been a driving force behind reproductive innovations (e.g., preceding references). Duplications in multiple MADS-box gene subfamilies coincide with major events in angiosperm history, most notably the earliest diversification of extant flowering plants and the diversification in the major core eudicot lineages (Fig. 1). Whole genome duplications have also been hypothesized for these nodes of the angiosperm phylogeny (Bowers *et al.*, 2003; Buzgo *et al.*, 2005; Cui *et al.*, 2006; Zahn *et al.*, 2005a), but it has not been shown conclusively whether or not ancient polyploidy events spawned the diversification of MADS-box genes in the common ancestors of all core eudicots or all extant flowering plant lineages.

D. SHIFTS IN SELECTIVE CONSTRAINT

Mechanistic hypotheses are required for genome-wide investigations of the relationship between gene and genome duplications and the evolution of plant reproduction. Most studies of functional evolution following gene duplication have built on a model of evolution wherein selection may be relaxed as the result of functional redundancy immediately following duplication events. However, evolutionary constraint is eventually restored after one duplicate becomes a pseudogene, ancestral gene function is split between the two duplicates (subfunctionalization), or one of the duplicates takes on new function (neofunctionalization) (Lynch and Conery, 2000; Force *et al.*, 1999; Ohno, 1970). Whereas these models predict that functional redundancy between duplicate genes is a temporary (nonequilibrium) condition, the notion of an adaptive "error-buffering" role for functional redundancy mentioned in an earlier section, provides an alternative explanation for the

maintenance of duplicate genes (Chapman *et al.*, 2006; Hileman and Baum, 2003; Nowak *et al.*, 1997; Moore *et al.*, 2005). Under these adaptive redundancy models, developmental instability in gene expression would have to be so deleterious that natural selection would favor individuals with functionally redundant gene copies over those with single copies of some genes. This hypothesis could be tested in populations of synthetic hybrids (Wang *et al.*, 2006), although subtle differences in fitness may be difficult to detect. Alternatively, a phylogenetically based retrospective approach may (or may not) detect even subtle changes in selective constraint following duplication events (see later section).

Hileman and Baum (2003) also proposed that duplicate genes may be retained if expression levels for both gene copies were reduced such that both genes would be required to maintain ancestral gene product dosage. This additive dosage model, described by Force *et al.* (1999) as "quantative subfunctionalization," is distinct from the more commonly hypothesized form of subfunctionalization in that there is no differential tissue or stage-specific compartmentalization of gene expression (Hileman and Baum, 2003). Duarte *et al.* (2006) identified instances of "hypofunctionalization" in their analysis of microarray expression profiles for duplicate genes where one duplicate was expressed at much lower levels than the other, but the degree of expression level divergence between duplicates was rarely constant across all organs. In addition, Duarte *et al.* (2006) identified significant gene by organ interactions (divergence of gene-expression patterns) in the majority of their ANOVA-based comparisons of expression levels for duplicate gene pairs. This result is consistent with models of regulatory sub- and neofunctionalization following gene duplication.

Changes in the mode of selection on protein-coding regions of gene sequences are often diagnosable through analyses of substitution rates (see reviews in Nielsen, 2005; Yang, 2002). Maximum likelihood estimates of per site synonymous (dS [$=K_s$]) and nonsynonymous (dN) nucleotide substitution frequencies, and the ratio of these substitution types (dN/dS $= \omega$) can be estimated from nucleotide alignments in a pair-wise fashion as described in Section II, or within the context of a gene phylogeny using codon-based models of sequence evolution (Goldman and Yang, 1994; Muse and Gaut, 1994; Nielsen and Yang, 1998; Yang *et al.*, 2005) as implemented in the codeml program of PAML (Yang, 1997) or HyPhy (Pond *et al.*, 2005). Bayesian estimates of these parameters can be obtained using MrBayes as described by Huelsenbeck and Dyer (2004), and Bayesian estimates of site-specific rate ratios (ω) are provided in the codeml output (Yang *et al.*, 2005).

There has been an explosion of interest in analyses of dN/dS ratios aimed at detecting adaptive amino acid changes (positive selection) associated with

changing gene function within gene families (Barkman, 2003; Yang, 1998) or across the whole genome (Bustamante *et al.*, 2005; Clark *et al.*, 2003; Nielsen *et al.*, 2005). Shifts in gene expression and function may be driven by changes in noncoding regulatory elements rather than protein-coding sequences (Doebley and Lukens, 1998), and adaptive divergence in protein-coding sequences is not necessary under the *functional divergence model* for retention of duplicated genes (Lynch and Conery, 2000; Force *et al.*, 1999; Ohno, 1970). The model does, however, predict that selection would be relaxed immediately following gene duplication. In contrast, the *developmental instability-buffering model* (Chapman *et al.*, 2006; Hileman and Baum, 2003; Moore *et al.*, 2005; Nowak *et al.*, 1997) predicts that purifying selection would not be relaxed following duplication events. Tree-based analysis of dN/dS ratios could test the null hypothesis that selective constraint averaged across the coding sequence does not change following duplication events (see the "branch" model of Yang, 1998; Barkman, 2003). While rejection of this null hypothesis might favor the functional divergence model, failure to reject would not necessarily favor the developmental instability-buffering model. A short period of relaxed selection following duplication may be difficult to detect, and a power analysis (Leebens-Mack and dePamphilis, 2002) would be required in order to interpret failure to reject the null hypothesis of equal selective constraint (dN/dS) before and after duplication. Further, if dN/dS is quite variable across a coding sequence, a "branch X sites" model (Yang and Nielsen, 2002) would provide more power for detecting relaxed selection in the portions of a gene that were more highly conserved before duplication. Nam *et al.* (2005) found significant variation in amino acid substitution rates across regions of duplicate MIKC-type MADS genes, suggesting that dN/dS does vary across coding sequences. Moreover, analyses that characterize this variation can identify specific domains that contribute to functional divergence (Nam *et al.*, 2005).

IV. COMPARATIVE ANALYSES OF DISTANTLY RELATED TAXA ELUCIDATE GENE FUNCTION IN *ARABIDOPSIS*

Comparative analyses are also providing insights into gene function in *Arabidopsis* and other model systems. Investigations of single-copy gene families are especially interesting and straightforward because duplicates arising from repeated polyploidy events (see Section II) may have been selectively culled due to dosage constraints, and functional studies employing reverse genetics are less likely to be confounded by redundancy. At the

same time, a high proportion of these genes have been annotated as "hypothetical" or "expressed" proteins because unlike members of larger gene families, the annotation process is not aided by similarity to functionally characterized genes.

We have been combining comparative and functional approaches to investigate an interesting class of single-copy genes that are found in a wide array of plant species, but seem to have been lost in the grasses (Poaceae). Using the search tools in PlantTribes (http://fgp.huck.psu.edu/tribe.php) we identified approximately 1500 single-copy genes in *Arabidopsis* that have no orthologs in rice. Of these, a subset of about 500 genes had "hypothetical" or "expressed" protein annotations. The predicted protein sequences of these genes were used to search the FGP Unigene database, the Plant Genome Database (PlantGDB http://www.plantgdb.org), The Institute for Genomic Research maize genome database (TIGR, http://tigrblast.tigr.org/tgi_maize/index.cgi), the moss database (COSMOS, http://www.cosmoss.org) and the *Chlamydomonas* genome database (http://www.chlamy.org) using TBLASTN (Altschul *et al.*, 1997). Those sequences that showed hits in other plants, but no hits in grasses, were chosen for further study. One of these genes, Di05 (At5g48480), has orthologs in a moss, a fern (*Ceratopteris*), gymnosperms, basal angiosperms, eudicots, and the nongrass monocots *Asparagus* and *Yucca*, but no orthologs were found in any members of the Poaceae. This is remarkable given the large number of expressed gene sequences available for multiple species in the grass family.

An alignment of Di05 orthologs was constructed using CLUSTALW (Thompson *et al.*, 1994) and refined using Se-AL (http://evolve.zoo.ox.ac.uk/software.html?id = seal). A phylogenetic tree was derived from parsimony analysis in PAUP* (Swofford, 2003). The resulting tree is consistent with known organismal relationships (Fig. 1), and supports the hypothesis that Di05 homologs existed in the common ancestor of mosses, ferns, gymnosperms, and angiosperms, but the gene was lost on the monocot branch leading to the Poaceae.

In *Arabidopsis* seedlings, Di05 is strongly expressed in the shoot apical meristem and leaf primordia. In reproductive structures, transcripts could be detected in the floral apical meristems, and floral primordia, but only in developing pollen and ovules after stage 9 (Fig. 5). During seed development, Di05 is strongly expressed in the developing embryo. After seed germination, it is only expressed in the shoot and root tips. These results suggest that Di05 may be involved in cell or tissue differentiation. In *Eschscholzia* and *Persea* flowers, the Di05 expression pattern is very similar to that seen in *Arabidopsis* (Fig. 5). Interestingly, in the *Ceratopteris* sporophyte, Di05 is only expressed in the shoot and root tips, but not in the developing or

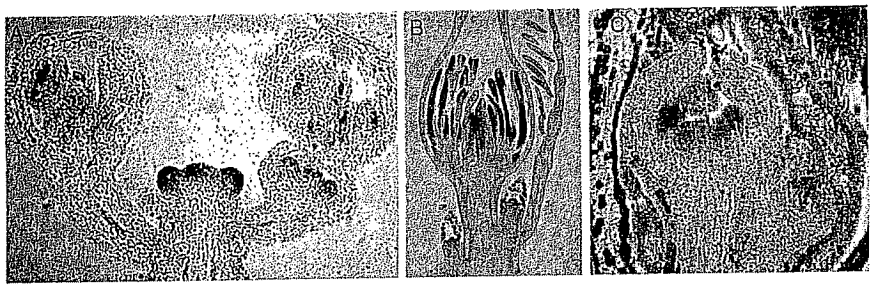

Fig. 5. *In situ* hybridizations show similar expression of an uncharacterized single-copy gene in reproductive meristems, pollen and ovules in *Arabidopsis* (A), *Eschscholzia* (B), and *Persea* (C).

mature spores. These results suggest that Di05 expression (and perhaps function) is conserved in eudicots, basal angiosperms, and perhaps ferns. Further investigation is required to understand the function of Di05 and the consequence of its loss some time after the divergence of the Asparagales and commelinid lineages within the monocots.

To further define the function of single-copy genes, we are examining available *Arabidopsis* T-DNA insertion lines. By focusing on those organs and tissues in which expression was detected by *in situ* hybridization, we can quickly identify phenotypes associated with the T-DNA insertion. Using this strategy, we are successfully identifying phenotypes for several genes, and thus elucidating gene function.

V. FUTURE PROSPECTS: DEVELOPING A GENE FAMILY FRAMEWORK TO CHARACTERIZE PLANT GENE AND GENOME EVOLUTION

This is an exciting time in plant genomics. The number of plant genome sequencing projects is expanding (Fig. 1), and this trend will continue with technological advances (Margulies *et al.*, 2005). Increasingly powerful analytical tools are being developed to allow more questions to be addressed through comparative analyses. The high frequency of genome duplications and complicated gene birth-and-death process in plants relative to animals pose challenges to phylogenomic analyses aimed at transferring understanding of gene function from model to nonmodel systems (Eisen, 1998; Engelhardt *et al.*, 2005; Sjölander, 2004), but much can be gained through analyses of sequence evolution and variation in gene expression within gene families. What is more, inferences concerning the evolution of genome

structure are being drawn from analyses of gene family phylogenies placed in the context of the chromosomal positions of duplicated genes (or gene blocks) (Bowers *et al.*, 2003; Mudge *et al.*, 2005; Paterson *et al.*, 2004; Sampedro *et al.*, 2005). A number of research groups are now independently developing databases for plant gene families (Albert *et al.*, 2005; Cannon *et al.*, 2004; Hartmann *et al.*, 2006), and these efforts are laying the foundation for evolutionary analyses of changing gene function and genome structure that may be associated with innovations in plant reproduction.

ACKNOWLEDGMENTS

We would like to thank all participants in the Plant Reproductive Genomics Workshop held at the Plant and Animal Genome Conference in 2005. We also thank all those involved in the Floral Genome Project for many stimulating discussions on the points raised in this chapter, especially André Chanderbali, Sangtae Kim, Doug Soltis, and Pam Soltis for their thoughtful comments on an early version of our manuscript. Finally, we thank the National Science Foundation Plant Genome Research Program for support of the Floral Genome Project (DBI-0115684).

REFERENCES

Adams, K. L. and Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology* **8**, 135–141.

Adams, K. L., Cronn, R., Percifield, R. and Wendel, J. F. (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 4649–4654.

Albert, V. A., Soltis, D. E., Carlson, J. E., Farmerie, W. G., Wall, P. K., Ilut, D. C., Solow, T. M., Mueller, L. A., Landherr, L. L., Hu, Y., Buzgo, M., Kim, S., *et al.* (2005). Floral gene resources from basal angiosperms for comparative genomics research. *BMC Plant Biology* **5**, 5.

Albert, V. A., Gustafsson, M. H. G. and Di Laurenzio, L. (1998). Ontogenetic systematics, molecular developmental genetics, and the angiosperm petal. *In* "Molecular Systematics of Plants II DNA Sequencing" (D. E. Soltis, P. S. Soltis and J. J. Doyle, eds.), pp. 349–374. Kluwer Academic Publishers, Boston.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research.* **25**, 3389–3402.

Aranzana, M. J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., Lister, C., Molitor, J., Shindo, C., Tang, C., Toomajian, C., Traw, B., *et al.* (2005). Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance Genes. *PLoS Genetics* **1**, e60.

Barkman, T. J. (2003). Evidence for positive selection on the floral scent gene isoeugenol-O-methyltransferase. *Molecular Biology and Evolution* **20**, 168–172.

Becker, A. and Theissen, G. (2003). The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Molecular Biology and Evolution* **29**, 464–489.

Becker, A., Winter, K. U., Meyer, B., Saedler, H. and Theissen, G. (2000). MADS-Box gene diversity in seed plants 300 million years ago. *Molecular Biology and Evolution* **17**, 1425–1434.

Beilstein, M. A., Al-Shehbaz, I. A. and Kellogg, E. A. (2006). Brassicaceae phylogeny and trichome evolution. *American Journal of Botany* **93**, 607–619.

Bennett, M. D., Leitch, I. J., Price, H. J. and Johnston, J. S. (2003). Comparisons with *Caenorhabditis* (approximately 100 Mb) and *Drosophila* (approximately 175 Mb) using flow cytometry show genome size in *Arabidopsis* to be approximately 157 Mb and thus approximately 25% larger than the *Arabidopsis* genome initiative estimate of approximately 125 Mb. *Annals of Botany (London)* **91**, 547–557.

Blanc, G. and Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**, 1667–1678.

Blanc, G., Barakat, A., Guyot, R., Cooke, R. and Delseny, M. (2000). Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell* **12**, 1093–1101.

Blanc, G., Hokamp, K. and Wolfe, K. H. (2003). A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Research* **13**, 137–144.

Bortiri, E., Jackson, D. and Hake, S. (2006). Advances in maize genomics: The emergence of positional cloning. *Current Opinion in Plant Biology* **9**, 164–171.

Bowers, J. E., Chapman, B. A., Rong, J. and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433–438.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., *et al.* (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature of Biotechnology* **18**, 630–634.

Burch-Smith, T. M., Anderson, J. C., Martin, G. B. and Dinesh-Kumar, S. P. (2004). Applications and advantages of virus-induced gene silencing for gene function studies in plants. *Plant Journal* **39**, 734–746.

Burke, J. M., Tang, S., Knapp, S. J. and Rieseberg, L. H. (2002). Genetic analysis of sunflower domestication. *Genetics* **161**, 1257–1267.

Burleigh, J. G. and Mathews, S. (2004). Phylogenetic signal in nucleotide data from seed plants: Implications for resolving the seed plant tree of life. *American Journal of Botany* **91**, 1599–1613.

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D., Civello, D., Adams, M. D., *et al.* (2005). Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–1157.

Buzgo, M., Soltis, P. S., Kim, S. and Soltis, D. E. (2005). The making of a flower. *The Biologist* **52**, 149–154.

Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D. and May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biology* **4**, 10.

Casneuf, T., De Bodt, S., Raes, J., Maere, S. and Van de Peer, Y. (2006). Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. *Genome Biology* 7, R13.

Chapman, B. A., Bowers, J. E., Schulze, S. R. and Paterson, A. H. (2004). A comparative phylogenetic approach for dating whole genome duplication events. *Bioinformatics* **20**, 180–185.

Chapman, B. A., Bowers, J. E., Feltus, F. A. and Paterson, A. H. (2006). Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. *Proceedings of the National Academy of Sciences of the United States of America*.

Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P. D., Kejariwal, A., Todd, M. A., Tanenbaum, D. M., Civello, D., Lu, F., Murphy, B., Ferriera, S., Wang, G., *et al.* (2003). Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**, 1960–1963.

Clark, R. M., Linton, E., Messing, J. and Doebley, J. F. (2004). Pattern of diversity in the genomic region near the maize domestication gene tb1. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 700–707.

Coen, E. S. and Meyerowitz, E. M. (1991). The war of the whorls: Genetic interactions controlling flower development. *Nature* **353**, 31–37.

Comai, L. and Henikoff, S. (2006). TILLING: Practical single-nucleotide mutation discovery. *Plant Journal* **45**, 684–694.

Cosgrove, D. J. (2005). Growth of the plant cell wall. *Nature Reviews Molecular Cell Biology* **6**, 850–861.

Cui, L., Wall, P. K., Leebens-Mack, J. H., Lindsay, B. G., Soltis, D. E., Doyle, J. J., Soltis, P. S., Carlson, J. E., Arumuganathan, K., Barakat, A., Albert, V. A. Ma, H., *et al.* (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Research* in press.

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. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 1904–1909.

De Bodt, S., Maere, S. and van de Peer, Y. (2005). Gene duplication and the evolution of angiosperms. *Trends in Ecology and Evolution* **20**, 591–597.

dePamphilis, C. W. (1995). Genes and genomes. *In* "Parasitic Plants" (M. C. Press and J. D. Graves, eds.), pp. 177–205. Chapman and Hall, London.

Doebley, J. and Lukens, L. (1998). Transcriptional regulators and the evolution of plant form. *Plant Cell* **10**, 1075–1082.

Doerge, R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics* **3**, 43–52.

Dong, Q., Lawrence, C. J., Schlueter, S. D., Wilkerson, M. D., Kurtz, S., Lushbough, C. and Brendel, V. (2005). Comparative plant genomics resources at PlantGDB. *Plant Physiology* **139**, 610–618.

Duarte, J. M., Cui, L., Wall, P. K., Zhang, Q., Zhang, X., Leebens-Mack, J., Ma, H., Altman, N. and dePamphilis, C. W. (2006). Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. *Molecular Biology and Evolution* **23**, 469–478.

Eddy, S. R. (2005). A model of the statistical power of comparative genome sequence analysis. *PLoS Biology* **3**, e10.

Eisen, J. A. (1998). Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research* **8**, 163–167.

Engelhardt, B. E., Jordan, M. I., Muratore, K. E. and Brenner, S. E. (2005). Protein Molecular Function Prediction by Bayesian Phylogenomics. *PLoS Computational Biology* **1**, e45.

Enright, A. J., Van Dongen, S. and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* **30**, 1575–1584.

Enright, A. J., Kunin, V. and Ouzounis, C. A. (2003). Protein families and TRIBES in genome sequence space. *Nucleic Acids Research* **31**, 4632–4638.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L. and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**, 725–736.

Grant, V. (1963). "The Origin of Adaptations." Columbia University Press, New York.

Grant, V. (1981). "Plant Speciation." Columbia University Press, New York.

Hardison, R. C. (2003). Primer on Comparative Genomics. *PLoS Biology* **3**, e58.

Hartmann, S., Lu, D., Phillips, J. and Vision, T. J. (2006). Phytome: A platform for plant comparative genomics. *Nucleic Acids Res.* **34**, D724–D730.

Hey, J., Fitch, W. M. and Ayala, F. J. (2005). Systematics and the origin of species: An introduction. *Proceedings of the National Academy of Sciences of the United States of America* **102** (Suppl. 1), 6515–6519.

Hileman, L. C. and Baum, D. A. (2003). Why do paralogs persist? Molecular evolution of CYCLOIDEA and related floral symmetry genes in Antirrhineae (Veronicaceae). *Molecular Biology and Evolution* **20**, 591–600.

Huelsenbeck, J. P. and Dyer, K. A. (2004). Bayesian estimation of positively selected sites. *Journal of Molecular Evolution* **58**, 661–672.

Hughes, A. L., Friedman, R., Ekollu, V. and Rose, J. R. (2003). Non-random association of transposable elements with duplicated genomic blocks in *Arabidopsis thaliana*. *Molecular Phylogenetics and Evolution* **29**, 410–416.

Irish, V. F. (2003). The evolution of floral homeotic gene function. *Bioessays* **25**, 637–646.

Kellogg, E. A. and Bennetzen, J. L. (2004). The evolution of nuclear genome structure in seed plants. *American Journal of Botany* **91**, 1709–1725.

Kim, S., Koh, J., Yoo, M. J., Kong, H., Hu, Y., Ma, H., Soltis, P. S. and Soltis, D. E. (2005). Expression of floral MADS-box genes in basal angiosperms: Implications for the evolution of floral regulators. *Plant Journal* **43**, 724–744.

Kramer, E. M. and Hall, J. C. (2005). Evolutionary dynamics of genes controlling floral development. *Current Opinion in Plant Biology* **8**, 13–18.

Laitinen, R. A., Immanen, J., Auvinen, P., Rudd, S., Alatalo, E., Paulin, L., Ainasoja, M., Kotilainen, M., Koskela, S., Teeri, T. H. and Elomaa, P. (2005). Analysis of the floral transcriptome uncovers new regulators of organ determination and gene families related to flower organ differentiation in *Gerbera hybrida* (Asteraceae). *Genome Research* **15**, 475–486.

Lee, Y., Tsai, J., Sunkara, S., Karamycheva, S., Pertea, G., Sultana, R., Antonescu, V., Chan, A., Cheung, F. and Quackenbush, J. (2005). The TIGR Gene Indices: Clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Research* **33**, D71–D74.

Leebens-Mack, J. and dePamphilis, C. (2002). Power analysis of tests for loss of selective constraint in cave crayfish and nonphotosynthetic plant lineages. *Molecular Biology and Evolution* **19**, 1292–1302.

Leebens-Mack, J., Raubeson, L. A., Cui, L., Kuehl, J. V., Fourcade, M. H., Chumley, T. W., Boore, J. L., Jansen, R. K. and Depamphilis, C. W. (2005). Identifying the basal angiosperm node in chloroplast genome phylogenies: Sampling one's way out of the Felsenstein zone. *Molecular Biology and Evolution* **22**, 1948–1963.

Lexer, C., Rosenthal, D. M., Raymond, O., Donovan, L. A. and Rieseberg, L. H. (2005). Genetics of species differences in the wild annual sunflowers, *Helianthus annuus* and *H. petiolaris*. *Genetics* **169**, 2225–2239.

Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155.

Ma, H. and dePamphilis, C. (2000). The ABCs of floral evolution. *Cell* **101**, 5–8.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M. and Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 5454–5459.

Marin, B. and Melkonian, M. (1999). Mesostigmatophyceae, a new class of streptophyte green algae revealed by SSU rRNA sequence comparisons. *Protist* **150**, 399–417.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380.

Moore, R. C. and Purugganan, M. D. (2005). The evolutionary dynamics of plant duplicate genes. *Current Opinion in Plant Biology* **8**, 122–128.

Moore, R. C., Grant, S. R. and Purugganan, M. D. (2005). Molecular population genetics of redundant floral-regulatory genes in *Arabidopsis thaliana*. *Molecular Biology and Evolution* **22**, 91–103.

Mudge, J., Cannon, S. B., Kalo, P., Oldroyd, G. E., Roe, B. A., Town, C. D. and Young, N. D. (2005). Highly syntenic regions in the genomes of soybean, Medicago truncatula, and *Arabidopsis thaliana*. BMC *Plant Biology* **5**, 15.

Muse, S. V. and Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* **11**, 715–774.

Nam, J., Kaufmann, K., Theissen, G. and Nei, M. (2005). A simple method for predicting the functional differentiation of duplicate genes and its application to MIKC-type MADS-box genes. *Nucleic Acids Research* **34**, e12.

Nesbitt, T. C. and Tanksley, S. D. (2002). Comparative sequencing in the genus Lycopersicon. Implications for the evolution of fruit size in the domestication of cultivated tomatoes. *Genetics* **162**, 365–379.

Nielsen, R. (ed.) (2005). "Statistical Methods In Molecular Evolution" (Statistics for Biology and Health Series). Springer-Verlag, New York.

Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936.

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., Fledel-Alon, A., Tanenbaum, D. M., Civello, D., White, T. J., J. S., Adams, M. D. and Cargill, M. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology* **3**, e170.

Nowak, M. A., Boerlijst, M. C., Cooke, J. and Smith, J. M. (1997). Evolution of genetic redundancy. *Nature* **388**, 167–171.

Odenwald, W. F., Rasband, W., Kuzin, A. and Brody, T. (2005). EVOPRINTER, a multigenomic comparative tool for rapid identification of functionally

important DNA. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 14700–14705.
Ohno, S. (1970). "Evolution by Gene Duplication." Springer-Verlag, New York.
Otto, S. P. and Whitton, J. (2000). Polyploid incidence and evolution. *Annual Review of Genetics* **34**, 401–437.
Ouyang, S. and Buell, C. R. (2004). The TIGR Plant Repeat Databases: A collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Research* **32**, D360–D363.
Paterson, A. H., Bowers, J. E. and Chapman, B. A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 9903–9908.
Pond, S. L., Frost, S. D. and Muse, S. V. (2005). HyPhy: Hypothesis testing using phylogenies. *Bioinformatics* **21**, 676–679.
Pryer, K. M., Schneider, H., Smith, A. R., Cranfill, R., Wolf, P. G., Hunt, J. S. and Sipes, S. D. (2001). Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. *Nature* **409**, 618–622.
Qiu, Y.-L., Dombrovska, O., Lee, J., Li, L., Whitlock, B., Bernasconi-Quadroni, F., Rest, J., Borsch, T., Hilu, K. W., Renner, S. S., Soltis, D. E., Soltis, P. S., *et al.* (2005). Phylogenetic analysis of basal angiosperms based on nine plastid, mitochondrial, and nuclear genes. *International Journal of Plant Science* **166**, 815–842.
Rong, J., Bowers, J. E., Schulze, S. R., Waghmare, V. N., Rogers, C. J., Pierce, G. J., Zhang, H., Estill, J. C. and Paterson, A. H. (2005). Comparative genomics of *Gossypium* and *Arabidopsis*: Unraveling the consequences of both ancient and recent polyploidy. *Genome Research* **15**, 1198–1210.
Rudd, S. (2005). openSputnik—a database to ESTablish comparative plant genomics using unsaturated sequence collections. *Nucleic Acids Research* **33**, D622–D627.
Sampedro, J. and Cosgrove, D. J. (2005). The expansin superfamily. *Genome Biology* **6**, 242.
Sampedro, J., Lee, Y., Carey, R. E., dePamphilis, C. and Cosgrove, D. J. (2005). Use of genomic history to improve phylogeny and understanding of births and deaths in a gene family. *Plant Journal* **44**, 409–419.
Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., Linsley, P. S. Mao, M., *et al.* (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302.
Schlueter, J. A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J. J. and Shoemaker, R. C. (2004). Mining EST databases to resolve evolutionary events in major crop species. *Genome* **47**, 868–876.
Schmid, M., Uhlenhaut, N. H., Godard, F., Demar, M., Bressan, R., Weigel, D. and Lohmann, J. U. (2003). Dissection of floral induction pathways using global expression analysis. *Development* **130**, 6001–6012.
Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15**, 1034–1050.
Simillion, C., Vandepoele, K., Van Montagu, M. C., Zabeau, M. and Van de Peer, Y. (2002). The hidden duplication past of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 13627–13632.

Sjölander, K. (2004). Phylogenomic inference of protein molecular function: Advances and challenges. *Bioinformatics* **20**, 170–179.
Soltis, D. E. and Soltis, P. S. (1999). Polyploidy: Recurrent formation and genome evolution. *Trends in Ecology and Evolution* **14**, 348–352.
Soltis, D. E., Soltis, P. S., Albert, V. A., Oppenheimer, D. G., dePamphilis, C. W., Ma, H., Frohlich, M. W. and Theissen, G. (2002). Missing links: The genetic architecture of flowers floral diversification. *Trends Plant Science* **7**, 22–31; discussion 31–34.
Soltis, D. E., Soltis, P. S., Chase, M. W. and Endress, P. (2005). "Phylogeny, Evolution, and Classification of Flowering Plants." Sinauer Associates, Sunderland, MA.
Soltis, P. S. (2005). Ancient and recent polyploidy in angiosperms. *New Phytologist* **166**, 5–8.
Sonnhammer, E. L. and Koonin, E. V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics* **18**, 619–620.
Stebbins, G. L. (1950). "Variation and Evolution in Plants." Columbia University Press, New York.
Swofford, D. L. (2003). PAUP*. Phylogenetic Analysis Using Parsimony (* and other methods). Version 4, Sinauer Associates, Sunderland, Massachusetts.
Sweigart, A., Fishman, L. and Willis, J. (2006). A simple genetic incompatibility causes hybrid male sterility in Mimulus. *Genetics* **172** (4), 2465–2479.
Theissen, G. (2002). Secret life of genes. *Nature* **415**, 741.
Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673–4680.
Vandepoele, K. and Van de Peer, Y. (2005). Exploring the plant transcriptome through phylogenetic profiling. *Plant Physiology* **137**, 31–42.
Vision, T. J., Brown, D. G. and Tanksley, S. D. (2000). The origins of genomic duplications in *Arabidopsis*. *Science* **290**, 2114–2117.
Wang, J., Tian, L., Lee, H. S., Wei, N. E., Jiang, H., Watson, B., Madlung, A., Osborn, T. C., Doerge, R. W., Comai, L. and Chen, Z. J. (2006). Genome-wide nonadditive gene regulation in *Arabidopsis allotetraploids*. *Genetics* **172**, 507–517.
Wellmer, F., Riechmann, J. L., Alves-Ferreira, M. and Meyerowitz, E. M. (2004). Genome-wide analysis of spatial gene expression in *Arabidopsis* flowers. *Plant Cell* **16**, 1314–1326.
Western, T. L., Cheng, Y., Liu, J. and Chen, X. (2002). HUA ENHANCER2, a putative DExH-box RNA helicase, maintains homeotic B and C gene expression in *Arabidopsis*. *Development* **129**, 1569–1581.
Yamasaki, M., Tenaillon, M. I., Bi, I. V., Schroeder, S. G., Sanchez-Villeda, H., Doebley, J. F., Gaut, B. S. and McMullen, M. D. (2005). A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *Plant Cell* **17**, 2859–2872.
Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**, 555–556.
Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution* **15**, 568–573.
Yang, Z. (2002). Inference of selection from multiple species alignments. *Current Opinion in Genetics and Development* **12**, 688–694.

Yang, Z. and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution* **19**, 908–917.

Yang, Z., Wong, W. S. and Nielsen, R. (2005). Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution* **22**, 1107–1118.

Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C., Zhang, J., Zhang, Y., *et al.* (2005). The Genomes of *Oryza sativa*: A history of duplications. *PLoS Biology* **3**, e38.

Zahn, L. M., Kong, H., Leebens-Mack, J. H., Kim, S., Soltis, P. S., Landherr, L. L., Soltis, D. E., Depamphilis, C. W. and Ma, H. (2005a). The evolution of the *SEPALLATA* subfamily of MADS-box genes: A preangiosperm origin with multiple duplications throughout angiosperm history. *Genetics* **169**, 2209–2223.

Zahn, L. M., Leebens-Mack, J., DePamphilis, C. W., Ma, H. and Theissen, G. (2005b). To B or Not to B a flower: The role of *DEFICIENS* and *GLOBOSA* orthologs in the evolution of the angiosperms. *Journal of Heredity* **96**, 225–240.

Zahn, L. M., Leebens-Mack, J. H., Arrington, J. M., Hu, Y., Landherr, L. L., dePamphilis, C. W., Becker, A., Theissen, G. and Ma, H. (2006). Conservation and divergence in the AGAMOUS subfamily of MADS-box genes: Evidence of independent sub- and neofunctionalization events. *Evolution & Development* **8**, 30–45.

Zanis, M. J., Soltis, D. E., Soltis, P. S., Mathews, S. and Donoghue, M. J. (2002). The root of the angiosperms revisited. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6848–6853.

Zanis, M. J., Soltis, P. S., Qiu, Y.-L., Zimmer, E. and Soltis, D. E. (2003). Phylogenetic analyses and perianth evolution in basal angiosperms. *Annals of the Missouri Botanical Garden* **90**, 129–150.

Zik, M. and Irish, V. F. (2003). Global identification of target genes regulated by APETALA3 and PISTILLATA floral homeotic gene action. *Plant Cell* **15**, 207–222.