Beyond the Arabidopsis Genome: Opportunities for Comparative Genomics[1]

Anne E. Hall[2], Aretha Fiebig[2], and Daphne Preuss*

Howard Hughes Medical Institute, The University of Chicago, 1103 East 57th Street, Chicago, Illinois 60637

Like most higher eukaryotes, flowering plants are believed to contain surprisingly similar numbers of genes. Nevertheless, angiosperm genome sizes vary over a wide range—from 50 Mb to over 120,000 Mb. Comparative mapping has shown that numerous alterations contribute to genomic diversity among plants. Over time, chromosomes are broken, reassembled, partially or wholly duplicated, and even eliminated, ultimately resulting in reproductive isolation and speciation. However, the mechanisms that create such variation, and the evolutionary forces that fix these changes, are not well understood. Comparative analyses of plant genomes promise to clarify the selective pressures driving these changes; such investigations will elucidate alterations at the level of whole genomes, as well as those at the level of specific sequences, including genes, repetitive elements, and other non-coding regions.

Although low-resolution genetic maps can identify gross chromosomal alterations, a clear understanding of the mechanisms behind these changes requires multispecies sequence comparisons. Such analyses reveal the composition, organization, and functional components of genomes and provide insight into regional differences in composition between related species. In addition, sequence comparisons elucidate evolutionary history; for example, the stepwise accumulation of nucleotide insertions/deletions (indels) only becomes clear with the analysis of multiple species. Comparative sequence analysis also aids in gene prediction and sequence annotation, and facilitates the identification and definition of regulatory elements, including promoters, enhancers, and transcription factor-binding sites (Kent and Zahler, 2000; Koch et al., 2001b).

The recent analysis of the sequence of the Arabidopsis genome highlighted unexpected aspects of its composition, organization, and function (Arabidopsis Genome Initiative [AGI], 2000). The questions raised by these observations can best be approached through comparative genomics. For example, although Arabidopsis is considered a "true" diploid, its genome has undergone major duplication events, followed by extensive rearrangements and chromosome fusion and loss, hypothesized to have shifted the haploid chromosome number from 4 to 8 and then to 5 (AGI, 2000; Vision et al., 2000). Interestingly, however, evidence of duplications was not found in the sequenced portions of the centromere regions. All five centromeres contain tracts of unique DNA, interspersed with similar types of transposable elements and interrupted by large tandem arrays of satellites. Aside from the repetitive sequences, pairwise comparisons of the centromere regions did not identify blocks of unique sequence indicative of ancient duplication events (AGI, 2000).

In addition to large segmental duplications on the chromosome arms, Arabidopsis also contains a prevalence of gene families, many of which are the result of tandem duplications of individual genes, rather than redundancy of entire chromosome segments. Nearly 40% of the predicted genes in the Arabidopsis genome belong to families that contain more than five members (AGI, 2000). Through studies of related species, it will become possible to discern the timing of genome duplications, the types of DNA eliminated, and the mechanisms responsible for rearrangements and deletions. Moreover, analysis of the genomes of related species will clarify how gene families expand, contract, and diversify. Examination of relatives containing different types of genome duplications will also reveal the changes that occur after such events, including mechanisms that are activated after large-scale genomic perturbation.

Clearly, comparative genomic approaches would provide enormous benefit toward understanding the origins of the Arabidopsis genome, as well as other plant genomes. Although previous comparisons with genes from yeast, flies, worms, and mammals have provided functional clues for approximately 70% of Arabidopsis genes, the vast evolutionary distances that separate these species restrict such comparisons to coding regions (AGI, 2000). Even the genomes of Arabidopsis and rice (*Oryza sativa*), which are separated by approximately 200 million years (MY), have substantially diverged (Wolfe et al., 1989), making the incremental stages of evolutionary change difficult to grasp (Goff et al., 2002; Yu et al., 2002).

In this Update, we explore the utility of a comparative genomics approach that relies on Arabidopsis

[1] This work was supported in part by the Howard Hughes Medical Institute and by the National Science Foundation (grant no. MCB 0077854 to A.F.).

[2] These authors contributed equally to the paper.

* Corresponding author; e-mail dpreuss@midway.uchicago.edu; fax 773–702–6648.

www.plantphysiol.org/cgi/doi/10.1104/pp.004051.

and several other species within the Brassicaceae family. We discuss discoveries made from prior comparisons within the family, prospects for additional genomic studies, and their potential for significantly improving our understanding of plant genomes. Finally, we discuss the community resources required to launch an effort with the necessary breadth to address a wide range of sequence-based evolutionary questions, and suggest a set of candidate species for genomic comparisons in the Brassicaceae.

THE BRASSICACEAE: A USEFUL FAMILY FOR COMPARATIVE GENOMICS

Arabidopsis is a member of the Brassicaceae family; the wealth of information and resources provided by the community of Arabidopsis researchers consequently provides a well-supported infrastructure that makes the Brassicaceae ideal for comparative studies of plant genomes. The Brassicaceae family is large, encompassing approximately 340 genera and more than 3350 species, a few of which are shown in Figure 1 (Al-Shehbaz, 1984). These species diverged from a common ancestor over a time period of approximately 40 to 50 MY as a result of numerous independent speciation events (Koch et al., 2001a). Thus, the thousands of extant species provide significant opportunities for investigating the genetic differences that lead to speciation. In practical terms, several species within the Brassicaceae have genomes sizes that are extremely small; at 49 Mb, the estimated genome size of *Cardamine amara* is approximately one-third that of Arabidopsis (Bennett and Smith, 1991).

The Brassicaceae thrive in a variety of habitats; they are concentrated in the northern temperate regions, the Mediterranean, and the mountains of southwest and central Asia, and have migrated extensively to assume a worldwide distribution (Rollins, 1993). Within this family are species that have adapted to diverse ecological settings, providing vast natural variation in developmental, biochemical, and physiological phenotypes. Some Brassicaceae species grow in the high altitudes of the Himalayas (e.g. *Crucihimalaya himalaica*), others in aquatic environments (e.g. *Nasturtium officinale*), and others in desert conditions (e.g. *Nerisyrenia camporum*). Species such as *Lesquerella filiformis*, with a range restricted to southwestern Missouri, require specific ecological niches, whereas others like Shepard's purse (*Capsella*

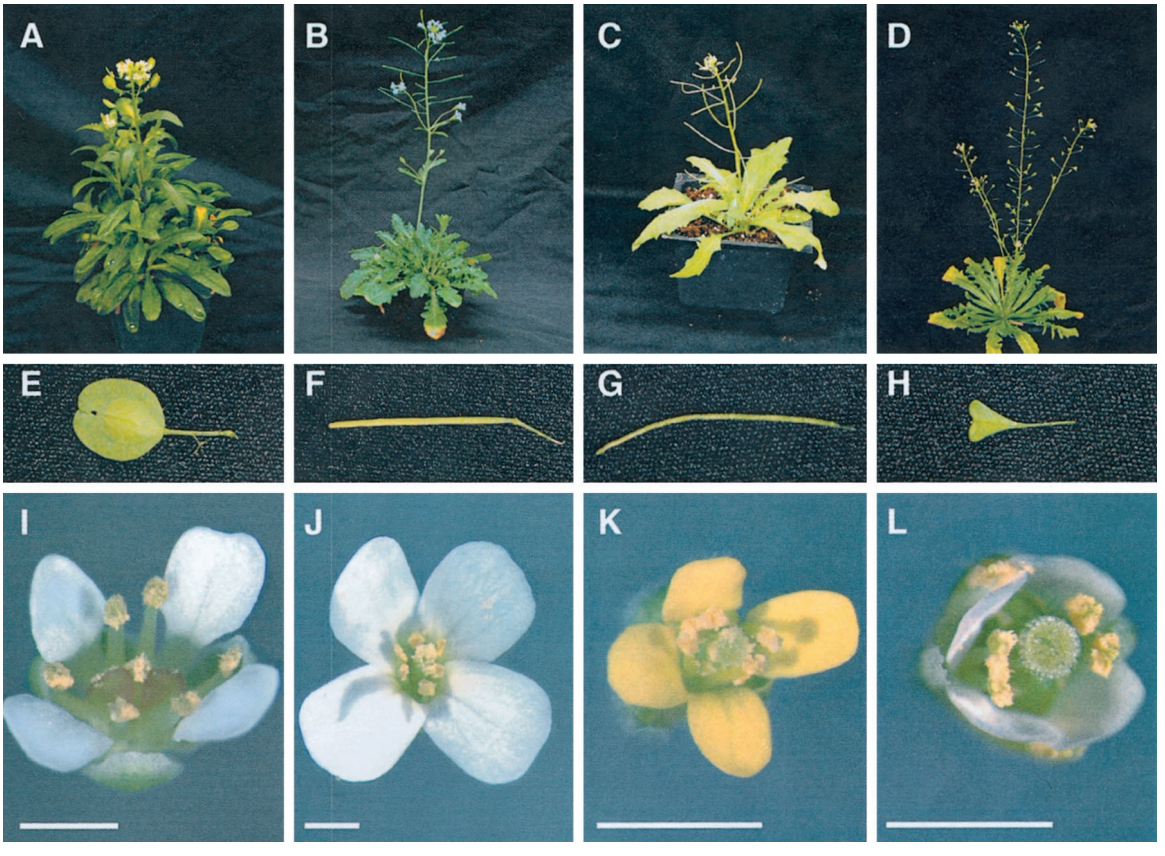

Figure 1. Morphology of several species in the Brassicaceae family. Species left to right include: *Thlaspi arvense*, *Arabidopsis suecica*, *Olimarabidopsis pumila*, and *Capsella rubella*. A through D, Overall adult plant morphology; plants were grown in 3-inch pots. E through H, Variations in mature fruit shape and size. I through L, Close-up of flowers. Size bars represent 1 mm.

bursa-pastoris) are successful colonizers worldwide (Rollins, 1993). Comparative genetic investigations of such species promise to provide insight into the mechanisms that restrict the range of an organism or allow it to conquer diverse environments.

Arabidopsis researchers have exploited the variation among Arabidopsis ecotypes to identify genes that contribute to difference in flowering time, plant size, drought tolerance, and variation in glucosinolate profiles for regulation of plant/insect interactions (for review, see Alonso-Blanco and Koorneef, 2000). With the appropriate genomic tools, it will become possible to elucidate the genetic mechanisms that underlie the ability of Brassicaceae species to adapt to almost any environmental challenge, yielding insight into how specific genes or biochemical pathways evolve. An investment in the resources required for comparative genomics will not only clarify the underlying genetic basis for natural adaptations, but may also enhance the economic value of the Brassicaceae family, which currently ranks fifth in oil production worldwide and comprises over 10% of the U.S. vegetable crop acreage (Al-Shehbaz, 1984; http://govinfo.library.orst.edu).

PREVIOUS COMPARISONS OF BRASSICACEAE GENOMES

Cross species comparisons of genomes within the Brassicaceae have primarily been limited to plants in the *Brassica* genus with the goal of identifying genes useful for crop improvement. These studies relied on Arabidopsis as a reference genome and focused on determining the extent of colinearity of molecular markers on genetic or cytological maps. Such investigations have demonstrated, for example, that the *Brassica nigra* and Arabidopsis genomes, which diverged 16 to 21 MY ago, share many dispersed chromosomal segments estimated at 8 cM in length, or approximately 1.7 Mb (Lagercrantz, 1998). Similar conclusions have been reached comparing *Brassica oleracea* and *Brassica napus* (Cavell et al., 1998; Lan et al., 2000). Both physical and genetic mapping studies have also indicated that segments of Arabidopsis chromosomes are often present in triplicate in the diploid *Brassica* spp. genomes (*Brassica rapa*, *B. oleracea*, and *B. nigra*), suggesting that modern *Brassica* spp. likely evolved from an ancient hexaploid (Lagercrantz, 1998; Jackson et al., 2000; O'Neill and Bancroft, 2000). This genomic triplication event likely occurred after the divergence of the *Brassica* spp. and Arabidopsis lineages (Lagercrantz and Lydiate, 1996).

On a more local scale, DNA sequence comparisons have been used to assess synteny at the level of individual genes, confirming large colinear stretches of sequence in the Arabidopsis and *Brassica* spp. genomes. O'Neill and Bancroft (2000) used sequence data from a 222-kb region of Arabidopsis chromosome 4 (duplicated on Arabidopsis chromosome 5) and compared its gene content and order with homoeologous (ancestrally related) segments of the *B. oleracea* genome. The two species exhibit high levels of gene conservation and colinearity, although in the regions investigated, *B. oleracea* lacks homologs of some Arabidopsis genes. Similarly, searches for a *B. oleracea* counterpart of the approximately 15-kb Arabidopsis ABI-Rps2-Ck1 gene segment found disruptions of gene content in regions that otherwise shared high levels of sequence similarity and a conserved gene order (Quiros et al., 2001).

Recently, larger scale sequence-based comparisons of other species in the Brassicaceae have been initiated, including wild relatives of Arabidopsis such as *C. rubella* (Acarkan et al., 2000; Rossberg et al., 2001). *C. rubella* and Arabidopsis, estimated to have diverged 11 to 14 MY ago, have chromosomal segments that are extremely similar in gene content, orientation, and order. Studies so far indicate that genes are conserved at >90% nucleotide identity, exon sizes and boundaries remain constant, and only minor differences are observed in the length of introns. Further comparative studies between *C. rubella* and Arabidopsis are required to determine if these levels of genome similarity are representative.

Thus, comparative studies performed to date, primarily between the *Brassica* and Arabidopsis genera, suggest several important themes: (a) conservation of gene sequence, content, and order are common, a situation that facilitates cross species mapping and identification of syntenic regions and gene homologs; (b) the *Brassica* genus exhibits rapid rates of chromosome evolution, characterized predominantly by duplications, rearrangements, and fusions (Lagercrantz, 1998); (c) in both the Arabidopsis and *Brassica* genomes, gene density is high, and repetitive DNA content is low, particularly when compared with grass genomes (Kumar and Bennetzen, 1999). Although it is premature to suggest these themes are general attributes of the Brassicaceae, it is clear that questions of gene duplication, chromosome rearrangement, and alterations in ploidy can be appropriately addressed in this family.

PROSPECTS FOR GENOMIC COMPARISONS AMONG CLOSELY RELATED BRASSICACEAE

Augmenting the comparisons between *Brassica* spp. and Arabidopsis with additional comparisons to closely related species will aid in understanding the forces driving speciation, including reorganization of chromosomes and changes in genome composition and size. Because relatively few genomic analyses of genera within the Brassicaceae have been conducted, it is not clear if the duplications, fusions, and rearrangements that have occurred in *Brassica* spp. genomes are a consequence of domestication and selective breeding, or if they are also characteristic of wild

relatives. Although *Brassica* and Arabidopsis species diverged within the last 20 MY, the significant differences between their genomes blurs the incremental processes that gave rise to differences in chromosome structure and number. Furthermore, although the analysis of large orthologous regions of *Brassica* spp. and Arabidopsis chromosomes can characterize small-scale alterations in chromosome structure, this approach cannot adequately address the mechanistic changes that drive the formation of new plant species. Instead, a stepwise analysis of multiple species, phylogenetically distributed across the family, is required.

Expanding genomic comparisons to additional species within the Brassicaceae will provide opportunities to examine the types of genomic changes that contribute to variation in genome size, and how such alterations are manifested phenotypically. Arabidopsis was selected for complete genome sequencing based, in part, on its extremely small genome (140–175 Mb, Bennett and Smith, 1991); however, this small genome may require mechanisms to minimize its genome that are not representative of other plants. Although comprehensive molecular data is often lacking from plants with relatively large genomes, instances where such data are available have demonstrated the predictive value of genome size for some plant traits. Weedy species tend to have smaller genomes, and in some taxa there is a strong correlation between genome size and flowering time (for review, see Bennett et al., 1998, 2000).

Measurements of genome size (C value) are currently under way for a large number of Brassicaceae species (H.J. Price and S. Johnston, personal communication). This cost-effective procedure, combined with simple molecular assays that tabulate the amount of repetitive satellite, telomere, and rDNA sequences, will clarify whether repeat amplification contributes significantly to Brassicaceae genome evolution. In addition, comparative genetic mapping can be used to assess the degree of genome-wide colinearity in species with different genome sizes, and reveal large-scale rearrangements, duplications, and deletions. Comparative mapping has shown that genome colinearity persists across vast evolutionary distances in grass genomes, and that the differences between species result more from expansion of intergenic regions than chromosomal rearrangements (Gale and Devos, 1998). In addition, fluorescence in situ hybridization (FISH) and fiber-FISH can complement comparative genetic mapping to directly assess the distribution of genomic duplications, and how particular DNA sequences contribute to genome amplification. FISH has been useful in comparisons between the *Brassica* and Arabidopsis species to compare size and number of large homologous chromosomal regions. Jackson et al. (2000) used a single-copy, 431-kb Arabidopsis genomic fragment as a FISH probe in *B. rapa*, and demonstrated the large fragment is present multiple times in the *B. rapa* genome. The homologous regions in *B. rapa* do not appear to be significantly larger than the Arabidopsis segment, indicating that this particular genomic region has not greatly expanded since the divergence of the *Brassica* and Arabidopsis lineages.

Alterations in ploidy often contribute to expansions in genome size—a phenomenon much more common in plants than in animals. In fact, changes in ploidy have been a major factor in angiosperm genome evolution, involving either the duplication of an entire genome (autopolyploidy) or the merging of two distinct genomes to generate a single, new species (allopolyploidy). It is estimated that 30% to 80% of all angiosperms have undergone a polyploidization event in their history, and 40% of species in the Brassicaceae family are thought to be polyploids (Al-Shehbaz, 1984; Masterson, 1994). Large-scale genome duplications, including polyploidization, as well as chromosome fusion and chromosome loss, have contributed to the considerable variation in the base chromosome number in the Brassicaceae, with diploid examples of 5, 7, 8, 9, or 10 chromosomes (see Table I).

The prevalence of Brassicaceae polyploids, coupled with assessment of the hybridization events that generated some of these species, affords opportunities to characterize how genomes change after whole-genome duplication or fusion events. The dynamics of allopolyploidization events within several Brassicaceae genera have been examined, and the approximation of ancestral hybridization events using synthetic hybrids has been particularly useful for characterization of rapid genomic changes after polyploidization. Song et al. (1995) examined genomic changes in *Brassica* spp. allopolyploids by creating synthetic tetraploids from the diploid *Brassica* spp. Rapid genomic changes in the synthetic polyploids were detected by comparison of RFLP patterns between newly created polyploids and the natural polyploids. Comai et al. (2000) carried out a similar study with *A. suecica*, an allotetraploid apparently derived from a combination of the Arabidopsis and *Arabidopsis aerenosa* genomes. Synthetic *A. suecica* polyploids demonstrated phenotypic instability and rapid gene silencing at some loci, suggesting a critical role for epigenetic regulation in newly formed species (Comai et al., 2000). Other studies with synthetic *A. suecica* have been useful for characterizing nucleolar dominance, an epigenetic phenomenon in which ribosomal RNA genes are silenced in interspecific hybrids. In *A. suecica*, the Arabidopsis rRNA genes are normally silenced. However, studies of synthetic *A. suecica* revealed that *A. aerenosa* rRNA genes can also undergo silencing and that the establishment of the nucleolar dominance pattern in newly formed polyploids can take several generations (Chen et al., 1998). Further studies comparing genomes of synthetic and ancient

Table I. *Characteristics of select Brassicaceae spp.*

Name	Chromosome No.[a]	Ploidy	Haploid Genome[b]	Mating System[c]	Height[d]	Petal Length[d,e]	Seeds/ Fruit[d]	GenBank Entries[f]	Citations[g]
			Mb		*dm*	*mm*			
Arabidopsis halleri[h]	8	2X	–	SI	–	6–8	–	34	12
Arabidopsis lyrata[i]	8	2X, 4X	–	SI	1–4	6–8	–	78	19
Arabidopsis thaliana	5	2X	172	SC	1–4	~3	50–60	248,772	10,611
Arabis hirsuta	8	4X	–	SC	2–7	3–9	–	11	0
Aubrieta deltoidea	8	2X	–	SC	–	12–28	–	6	0
Barbarea vulgaris	8	2X	–	SC	2–8	6–8	–	12	36
B. oleracea	9	2X	760	SI	4–8	15–20	10–20	199,445	2,206
C. rubella (Shepard's purse)[j]	8	2X (4X)	(686)	SC	1–5	2–3	~20	31	165
C. amara	8	2X	49	SI	–	–	–	21	11
Cardamine flexuosa	8	4X	858	SC	1–5	2–3	20–30	14	6
C. himalaica[k]	8	2X	–	SC	–	–	40–120	8	1
Crucihimalaya wallichii[l]	8	2X	–	SC	~4	2.5–3.5	~80	4	0
Draba cuneifolia	8	2X	–	SC	0.5–3.5	2.5–5	20–30	3	1
Draba nemorosa	8	2X	–	SC	~3	–	~50	1	1
Lepidium campestre	8	2X	–	–	2–5	~2	2	10	5
Lepidium sativum	8	2, 3, or 4X	466	SC or SI	2–8	~3	2	3	298
Olimarabidopsis cabulica[m]	8	6X	–	SC	~1	2–3	~20	5	0
O. pumila[n]	8	4X	–	SC	~3	~3	~30	25	0
Raphanus sativus	9	2X	539	SI	4–12	15–20	1–5	228	1,030
Sinapis alba	8	3X	490	–	2–6	~11	4–6	77	1,003
Sisymbrium irio	7	2 or 4X	–	SC	1.5–5	2.5–4	~60	4	23
Sisymbrium orientale	7	2X	–	SC	3–7	8–10	~120	0	7
T. arvense	7	2X	–	SC	1–5	3–4	8–12	10	125

[a] Base chromosome no. (x) compiled from Rollins (1993), Koch et al. (1999), and Kew gardens C value database (Bennett and Leitch, 2001). [b] Estimates of 1C genome size compiled from the angiosperm C value database (Bennett and Leitch, 2001). [c] SC, Self-compatible; SI, self-incompatible. Compiled from Koch et al. (1999) and our own observations. [d] Compiled from Rollins (1993) and our own observations. [e] Petal length is roughly proportional to flower size, indicating ease of manual crosses. [f] GenBank entries in the taxonomy site at the National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Taxonomy on February 1, 2002. [g] No. of articles referencing the species names in the title or abstract in Institute for Scientific Information Web of Science: http://isi0.isiknowledge.com. [h] Includes subspecies *gemmifera*, *halleri*, and *ovierensis* and synonyms *Arabis halleri* and *Arabis gemmifera*. [i] Includes subspecies *kamchatica*, *kawasakiana*, *lyrata*, and *petraea*, and synonyms *Arabis lyrata*, *Cardaminopsis petraea*, and *Arabidopsis petraea*. [j] Includes both *C. rubella* (2x) and Shepard's purse (4x) because the distinction between these taxa is controversial. Values specific to Shepard's purse are in parentheses. [k] Includes synonym *Arabidopsis himalaica*. [l] Includes synonym *Arabidopsis wallichii*. [m] Includes synonyms *Arabidopsis cabulica* and *Arabidopsis korshinskyii*. [n] Includes synonyms *Arabidopsis pumila*, *Arabidopsis griffithiana*, and *Arabis pumila*.

polyploids will provide insight into how duplicated genomes change over evolutionary time, molecular mechanisms that contribute to successful unification of distinct genomes within polyploids, and what types of sequences are lost and gained in the process.

Expanding research to additional Arabidopsis relatives will also allow for the further use of interspecific hybrids to characterize molecular events contributing to speciation. Nasrallah et al. (2000) have utilized *A. lyrata* to successfully form interspecific hybrids with Arabidopsis. Subsequent backcrosses of F_1 hybrids to either *A. lyrata* or Arabidopsis produce different phenotypic defects, possibly uncovering mechanisms each species has evolved for reproductive isolation.

PROSPECTS FOR ANALYSIS OF RAPIDLY CHANGING GENOMIC REGIONS

Comparative genomic surveys of closely related species are vital to understanding the rapidly evolving portions of plant genomes. Centromere sequences, in particular, change rapidly, presenting a paradox: How can regions of highly conserved function exhibit no sequence similarity among the characterized model genomes? Changes in telomere and nuclear organizing region lengths, centromere organization, and the formation of heterochromatic knobs occur over very brief evolutionary time scales, changing considerably even between Arabidopsis ecotypes (AGI, 2000). Alterations within these regions are often driven by the amplification of repetitive DNA sequences; such changes are particularly tolerated in heterochromatic regions of low meiotic recombination. For example, rDNA, satellite, and telomere arrays expand and contract through gene conversion or mitotic recombination, and centromeres are frequently invaded by transposons. Although the alterations taking place in these rapidly evolving regions are notable in Arabidopsis, they are likely more significant in larger genomes that are burdened with

greater proportions of transposable elements and satellite DNA.

The accumulation of mobile DNA elements contributes significantly to genome expansion in many plant lineages. In maize (*Zea mays*), transposons, which account for 50% to 80% of the genome, dominate genomic change and accumulate through multiple, sequential insertions into intergenic regions (SanMiguel et al., 1996). In contrast, mobile elements are surprisingly sparse in Arabidopsis, comprising only 10% of the genome (AGI, 2000). Genomic analysis of other members of the Brassicaceae will reveal whether they, like Arabidopsis, have resisted significant amplification of transposons. Selective pressures likely contribute to species-specific variation in transposon number, providing opportunities to discover mechanisms that either promote or preclude invasion and amplification of mobile DNA.

In addition to heterochromatic and repetitive regions, genes with species-specific functions likely evolve rapidly. As a consequence, their analysis benefits from comparisons of close relatives. In many cases, genes mediating mate recognition are clearly species specific and exhibit rapid sequence evolution (for review, see Swanson and Vacquier, 2002). In Arabidopsis, a cluster of six *GRP* genes encodes pollen surface proteins that mediate pollen-stigma interactions. These genes are highly diverged from their counterparts in *B. oleracea*, which encodes a cluster of five genes. Although the general properties of the predicted proteins remain the same, divergence at the DNA sequence level makes alignments nearly impossible (Mayfield et al., 2001). Such divergence begs for analysis of additional, more closely related species. Similarly, genes involved in pathogen recognition and resistance (R genes) undergo rapid changes both within and between Brassicaceae species (for review, see Bergelson et al., 2001). R genes are often found in clusters that vary in gene content, even between Arabidopsis ecotypes (Stahl et al., 1999). Genes represented only in Arabidopsis or its nearest relatives are targets for understanding the changes that mediate species-specific adaptations. Identifying the unique and rapidly evolving genes within members of the Brassicaceae will provide insight into what makes each species individual.

BRASSICACEAE PHYLOGENY

Addressing the mechanisms that contribute to genome evolution requires analysis of multiple species separated by different evolutionary time scales. For example, global comparisons of genome sequences between organisms of different kingdoms can identify highly conserved genes, but are less useful when analyzing rapidly evolving regions or genes adaptive for a specialized lifestyle. Alternatively, comparisons of very close relatives elucidate rapidly diverging sequences, but are less helpful in predictions of intron/exon boundaries due to limited variation in intron sequences. Ultimately, the number of organisms selected for comparative genomic studies depends on available resources; information will be gained with every new species analyzed, but, at some point, the value of that information will no longer justify the large costs incurred.

The Brassicaceae form a natural group based on flower morphology and glucosinolate profiles (Fig. 2). Inferring phylogenies within this family, however, has been complicated by a significant degree of convergent evolution and highly uniform traits (Al-Shehbaz, 1984; Al-Shehbaz et al., 1999). Analysis of the sequences of rDNA and genes conserved between Brassicaceae species has proven extremely useful in resolving these systematic ambiguities. In particular, the genus Arabidopsis has been redefined in recent years. Although it previously included 59 species, evaluation of rDNA sequences indicated this was an unnatural group (O'Kane and Al-Shehbaz, 1997). The genus now includes only nine species: *A. arenosa*, *Arabidopsis cebennensis*, *Arabidopsis croatica*, *A. halleri*, *A. lyrata*, *Arabidopsis neglecta*, *Arabidopsis pedemontana*, *A. suecica*, and *Arabidopsis thaliana*. The remaining species were assigned to preexisting or new genera (Al-Shehbaz et al., 1999). The species now placed in Arabidopsis originated primarily in Europe and tend to be restricted in their ranges to mountainous regions.

The broader phylogenetic relationships within the Brassicaceae family have also been recently examined, yielding a framework of species that share common ancestry within the last approximately 40 MY (Koch et al., 2001). Analysis of the statistically well-supported relationships between 48 Brassicaceae species identified several groups separated from Arabi-

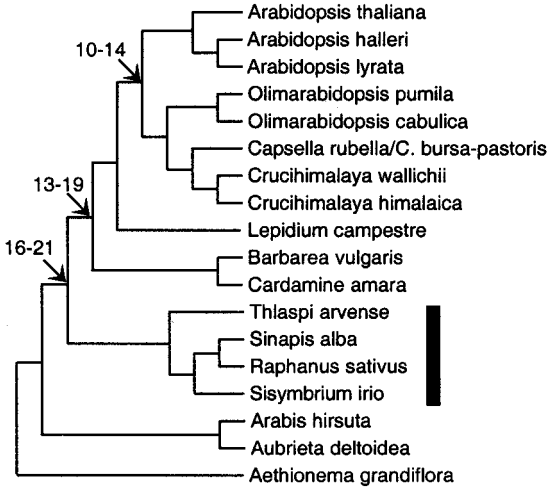

Figure 2. Phylogenetic tree of a sampling of Brassicaceae species (adapted from Koch et al. [2001a] with permission from the Botanical Society of America). Arrows indicate estimated dates of divergence at different nodes in millions of years. Other studies place the crop *Brassica* spp. in the clade indicated by the black line.

dopsis over a range of evolutionary time scales. Species in the genera *Olimarabidopsis*, *Crucihimalaya*, and *Capsella* (most of which were once placed in the genus Arabidopsis) are estimated to have diverged from Arabidopsis 10 to 14 MY ago (Koch et al., 2001a; see Fig. 2), roughly the same divergence time that separates humans and old world monkeys. *T. arvense*, *S. irio*, and *S. alba* fall into a branch whose last common ancestor with Arabidopsis existed approximately 16 to 20 MY ago; other studies have placed crop *Brassica* spp. in this same lineage (Yang et al., 1999; Fig. 2). Among the Brassicaceae, *Aethionema grandiflora* is highly separated from Arabidopsis, with a split estimated to have occurred approximately 40 MY ago. These well-defined phylogenetic relationships make it possible to select candidate species with an evolutionarily distribution ideal for comparative genomic studies.

PRACTICAL CONSIDERATIONS

When assessing the value of candidate species for genomic comparisons, information on systematic relationships should be supplemented with considerations of growth habits and other physical traits. Factors such as plant size, generation time, flower size for crosses, seed yield, transformability, and growth requirements are all relevant for laboratory research. In Table I, we have compiled data on some of these qualities for a sampling of Brassicaceae species. In some cases, growth habits make certain species less useful for genetic research, but nonetheless of value to researchers interested in particular properties. For example, *C. himalaica* requires cold treatments on the order of months to promote flowering, and therefore has a life cycle that is not ideal for assays that require multiple generations; on the other hand, it affords an excellent opportunity for comparisons to Arabidopsis genes required for vernalization and cold tolerance. *C. flexuosa* possesses seed dispersal mechanisms that ensure its success in the wild, and *Lepidium* spp. produce fruits containing only two seeds; each of these traits makes it difficult for researchers to collect the large quantities of seed typically required for genetic studies. Furthermore, self-compatible species such as Arabidopsis are extremely easy to propagate in greenhouses, but many Brassicaceae are self-incompatible, requiring either time-consuming manual pollinations or ready access to insect pollinators. Nevertheless, comparative genomic studies of these self-incompatible species have provided exciting information on the evolution of mating systems. For example, identification of the chromosomal region that contains the *A. lyrata* self-incompatibility (S) locus showed it is highly similar to an Arabidopsis region containing pseudogene remnants of *S*-locus genes (Kusaba et al., 2001). Thus, self-incompatibility is likely an ancestral state, and Arabidopsis became a self-compatible species through an accumulation of mutations in the genes required for self-incompatibility.

MOLECULAR RESOURCES FOR COMPARATIVE GENOMIC APPROACHES

Generating the resources required for large-scale genome projects are both costly and labor intensive; consequently, careful consideration is required before investing significantly in resources for new species. Among the desirable tools are expressed sequence tags or full-length cDNA clones, genetic and/or physical maps, genomic libraries useful for large or small-scale sequencing, and ultimately, fully assembled and annotated genome sequences. Expressed sequence tag sequencing is a reliable and cost-effective approach to surveying the coding content of a genome, although it is biased toward highly expressed sequences and gives no indication of genome organization. Genetic maps that rely on conserved markers are useful for assessing genome colinearity; however, this approach is relatively low in resolution and relies on recombination frequencies, and thus cannot identify small-scale insertions, deletions, inversions, duplications, or the degree of overall nucleotide diversity.

Shotgun sequencing of an entire genome can provide information useful for comparative genomics. However, obtaining large stretches of assembled, contiguous sequence requires a substantial investment. Because comparative approaches are most effective with tracts of sequence on the order of hundreds of kilobases, an intermediate and affordable approach is to sequence individual clones from genomic libraries. Libraries made with bacterial artificial chromosome (BAC) vectors provide several advantages; these vectors can carry large, stable inserts (up to 200 kb) and are single copy, thus minimizing selection against clones carrying repetitive or other sequences that replicate poorly in *Escherichia coli*. With BAC libraries, 5-fold genomic coverage of many interesting Brassicaceae species could be obtained with as few as 20,000 clones. At this scale, clones can easily be arrayed on filters, allowing identification of a particular gene, set of genes, or chromosomal regions. Moreover, when library construction relies on randomly distributed restriction sites that are cleaved with methylation-insensitive enzymes, they can be used to estimate the genomic representation of sequence classes, including satellites, transposons, and rDNA.

Other than those readily available for Arabidopsis, Brassicaceae resources have primarily been generated from crops in the *Brassica* genus. The UK Brassica genome project has generated BAC libraries for *B. oleracea* and *B. napus* (The John Innes Center, http://brassica.bbsrc.ac.uk; and Texas A&M, http://hbz.tamu.edu). This group is fingerprinting the BAC clones to produce a physical map that will be an-

chored on the Arabidopsis genome sequence. BAC end and genome shotgun sequences from *B. oleracea* have been deposited in GenBank, primarily by groups working at Hazen Genome Sequencing Center (Cold Spring Harbor, NY), Washington University (St. Louis), and The Institute for Genomic Research (Rockville, MD). Some groups have initiated genomic investigations in other Brassicaceae species, developing phylogenetic information and establishing genomic resources (http://vanilla.ice.mpg.de/departments/Gen/wild.htm). In particular, comparisons using genomic libraries from *A. lyrata* (Kusaba et al., 2001) and *C. rubella* (Acarkan et al., 2000) have been performed, and additional BAC libraries have been recently constructed from *C. rubella* and *O. pumila* (A.E. Hall, A. Fiebig, and D. Preuss, unpublished data; available from Amplicon Express, Inc., Pullman, WA).

CONCLUSIONS AND RECOMMENDATIONS FOR LAUNCHING SEQUENCE-BASED COMPARATIVE STUDIES

Plant biologists will clearly benefit from a comparative genomic project that expands available resources to represent the phylogenetic breadth of the Brassicaceae family. The availability of the Arabidopsis genome sequence, increased affordability of large-scale sequencing, and recent improvements in resolution of phylogenetic relationships make it an appropriate time to begin developing additional resources. In fact, as described above, a number of researchers are already doing so, albeit with a focus on crops, as opposed to undomesticated species.

It is in the interest of the plant community as a whole to develop resources that are accessible and affordable. Public availability of resources is critical; for example, BAC libraries can be readily distributed through publicly funded centers such as the Texas A&M University BAC resource center (College Station), and the Clemson University Genomic Institute (Clemson, SC), or the Ohio State Arabidopsis stock center (Arabidopsis Biological Resource Center, Columbus, OH). The costs of distributing resources may require the involvement of commercial centers, although the permanence and cost management of such options need to be addressed. Centralized stock centers for seed distribution are also necessary; centers that currently distribute Brassicaceae species include the Sendai Arabidopsis Seed Stock Center (Sendai, Japan), and the Arabidopsis Biological Resource Center. It is useful to note, however, that some deposits into these centers have not been identified by taxonomic authorities, resulting, in some cases, in misidentification of species or the use of nonstandard nomenclature. Because the phylogenetic classifications in this family are still under scrutiny, a central location that contains current information is critical. The National Center for Biotechnology Information is assembling taxonomic information for many species, including the Brassicaceae. Additional systematic efforts that are under way include the Deep Green project (http://ucjeps.berkeley.edu/bryolab/greenplantpage.html) and the Tree of Life project (http://tolweb.org/tree/phylogeny.html).

Because BAC libraries are extremely useful and can be generated in a short time frame, our group is currently selecting additional species for library construction; among the candidate species are *A. aerenosa*, *A. suecica*, *C. amara*, *S. irio*, and *T. arvense*. We anticipate that genomic resources developed from these species will provide a useful starting point for exploring the extent of genome evolution across the entire family.

ACKNOWLEDGMENTS

We thank Ihsan Al-Shehbaz, H. James Price, Spencer T. Johnston, and members of the Preuss laboratory for helpful discussions.

Received February 8, 2002; returned for revision April 2, 2002; accepted May 24, 2002.

LITERATURE CITED

Acarkan A, Rossberg M, Koch M, Schmidt R (2000) Comparative genome analysis reveals extensive conservation of genome organization for *Arabidopsis thaliana* and *Capsella rubella*. Plant J **23:** 55–62

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

Alonso-Blanco C, Koorneef M (2000) Naturally occurring variation in *Arabidopsis*: an underexploited resource for plant genetics. Trends Plant Sci **5:** 22–29

Al-Shehbaz IA (1984) The tribes of Cruciferae (Brassicaceae) in the southeastern United States. J Arnold Arbor **65:** 343–373

Al-Shehbaz IA, O'Kane SL, Price RA (1999) Generic placement of species excluded from *Arabidopsis* (Brassicaceae). Novon **9:** 296–307

Bancroft I (2000) Insights into the structural and functional evolution of plant genomes afforded by the nucleotide sequences of chromosomes 2 and 4 of *Arabidopsis thaliana*. Yeast **17:** 1–5

Bennett MD, Bhandol P, Leitch IJ (2000) Nuclear DNA amounts in Angiosperms and their modern uses: 807 new estimates. Ann Bot **86:** 859–909

Bennett MD, Leitch IJ (2001) Angiosperm DNA C-values database (release 3.1, Sept. 2001). http://www.rbgkew.org.uk/cval/homepage.html (February 1, 2002)

Bennett MD, Leitch IJ, Hanson L (1998) DNA amounts in two samples of angiosperm weeds. Ann Bot **82:** 121–134

Bennett MD, Smith JB (1991) Nuclear DNA amounts in Angiosperms. Philos Trans R Soc Lond B **334:** 309–345

Bergelson J, Kreitman M, Stahl EA, Tian DC (2001) Evolutionary dynamics of plant R-genes. Science **292:** 2281–2285

Chen ZJ, Comai L, Pikaard C (1998) Gene dosage and stochastic effects determine the severity and direction of uniparental ribosomal RNA gene silencing (nucleolar dominance) in *Arabidopsis* allopolyploids. Proc Natl Acad Sci USA **95:** 14891–14896

Comai L, Tyagi AP, Winter K, Holmes-Davis R, Reynolds SH, Stevens Y, Byers B (2000) Phenotypic instability and rapid gene silencing in newly formed *Arabidopsis* allotetraploids. Plant Cell **12:** 1551–1567

Gale MD, Devos KM (1998) Plant comparative genetics after 10 years. Science **282:** 656–659

Goff SA, Ricke D, Lan T-H et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science **296:** 92–100

Jackson SA, Cheng Z, Wang ML, Goodman HM, Jiang J (2000) Comparative fluorescence in situ hybridization mapping of a 431-kb *Arabidopsis thaliana* bacterial chromosome contig reveals the role of chromosomal duplications in the expansion of the *Brassica rapa* genome. Genetics **156:** 833–838

Kent JW, Zahler AM (2000) Conservation, regulation, synteny and introns in a large scale *C. briggsae-C. elegans* genomic alignment. Genome Res **10:** 1115–1125

Koch M, Bishop J, Mitchell-Olds T (1999) Molecular systematic and evolution of Arabidopsis and Arabis. Plant Biol **1:** 529–537

Koch M, Haubold B, Mitchell-Olds T (2001a) Molecular systematics of the Brassicaceae: evidence from coding plastidic *matK* and nuclear *Chs* sequences. Am J Bot **88:** 534–544

Koch MA, Weisshaar B, Kroymann J, Haubold B, Mitchell-Olds T (2001b) Comparative genomics and regulatory evolution: conservation and function of the *Chs* and *Apetala3* promoters. Mol Biol Evol **18:** 1882–1891

Kumar A, Bennetzen J (1999) Plant retrotransposons. Annu Rev Genet **33:** 479–532

Kusaba M, Dwyer K, Hendershot J, Vrebalov J, Nasrallah JB, Nasrallah ME (2001) Self-incompatibility in the genus *Arabidopsis*: characterization of the S locus in the outcrossing *A. lyrata* and its autogamous relative *A. thaliana*. Plant Cell **13:** 627–643

Lagercrantz U (1998) Comparative mapping between *Arabidopsis thaliana* and *Brassica nigra* indicates that Brassica genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. Genetics **150:** 1217–1228

Lagercrantz U, Lydiate DJ (1996) Comparative genome mapping in Brassica. Genetics **144:** 1903–1910

Lan T-H, DelMonte TA, Reischmann KP, Hyman J, Kowalski SP, McFerson J, Kresovich S, Paterson AH (2000) An EST-enriched comparative map of *Brassica oleracea* and *Arabidopsis thaliana*. Genome Res **10:** 776–788

Masterson J (1994) Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. Science **264:** 421–423

Mayfield JA, Fiebig A, Johnstone SE, Preuss D (2001) Gene families from the Arabidopsis thaliana pollen coat proteome. Science **292:** 2482–2485

Nasrallah ME, Yogeeswaran K, Snyder S, Nasrallah JB (2000) *Arabidopsis* species hybrids in the study of species differences and evolution of amphiploidy in plants. Plant Physiol **124:** 1605–1614

O'Kane SL, Al-Shehbaz IA (1997) A synopsis of *Arabidopsis* (Brassicaceae). Novon **7:** 323–327

O'Neill CM, Bancroft I (2000) Comparative physical mapping of segments of the genome of *Brassica oleracea* var. alboglabra that are homoeologous to sequenced regions of chromosomes 4 and 5 of *Arabidopsis thaliana*. Plant J **23:** 233–243

Quiros CF, Grellet F, Sadowski J, Suzuki T, Li G, Wroblewski T (2001) Arabidopsis and Brassica comparative genomics: sequence, structure and gene content in the ABI1-Rps2-Ck1 chromosomal segment and related regions. Genetics **157:** 1321–1330

Rollins RC (1993) The Cruciferae of Continental North America. Stanford University Press, Stanford, CA

Rossberg M, Theres K, Acarkan A, Herrero R, Schmitt T, Schumacher K, Schmitz G, Schmidt R (2001) Comparative sequence analysis reveals extensive microcolinearity in the lateral suppressor regions of the tomato, *Arabidopsis*, and *Capsella* genomes. Plant Cell **13:** 979–988

SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D et al. (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765–768

Song K, Lu P, Tang K, Osborn TC (1995) Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. Proc Natl Acad Sci USA **92:** 7719–7723

Stahl EA, Dwyer G, Mauricio R, Kreitman M, Bergelson J (1999) Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. Nature **400:** 667–671

Swanson WJ, Vacquier VD (2002) The rapid evolution of reproductive proteins. Nat Rev Genet **3:** 137–144

Vision TJ, Brown DG, Tanksley SD (2000) The origins of genomic duplications in Arabidopsis. Science **290:** 2114–2117

Wolfe KH, Gouy M, Yang Y-W, Sharp PM, Li W-H (1989) Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc Natl Acad Sci USA **86:** 6201–6205

Yang Y-W, Lai K-N, Tai P-Y, Ma D-P, Li W-H (1999) Molecular phylogenetic studies of *Brassica, Rorippa, Arabidopsis* and allied genera based on the internal transcribed spacer region of 18S–25S rDNA. Mol Phyl Evol **13:** 455–462

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