Independent Ancient Polyploidy Events in the Sister Families
Brassicaceae and Cleomaceae

M. Eric Schranz and Thomas Mitchell-Olds

INTRODUCTION

Polyploidy, or whole-genome duplication, has played a major role in the evolution of many branches of the tree of life by providing exponential and saltational increases in genome size and gene copy number. Polyploidy provides the opportunity for selection to sculpt a variety of new gene functions, traits, and lineages. There are several examples of polyploid animals (Mable, 2004), including mammals (Gallardo et al., 1999), fishes (Le Comber and Smith, 2004), and frogs (Ptacek et al., 1994), and polyploidy is particularly common among flowering plants. Many important model systems, both agricultural and wild, have been used to study the consequences of recent polyploidization (e.g., cotton [Gossypium hirsutum], tobacco [Nicotiana tabacum], wheat [Triticum aestivum], canola [Brassica napus], soybean [Glycine max], potato [Solanum tuberosum], sugarcane [Saccharum officinarum], cordgrass (Spartina anglica), and dandelion [Taraxacum officinale]) (Wendel, 2000; Osborn et al., 2003; Tate et al., 2005). These examples of recent polyploidy are easily detected by changes in chromosome numbers, genome size, and gene copy number compared with progenitors. However, ancient, or paleopolyploid, events are much more difficult to detect. After polyploidy occurs, the genome begins its march back to diploidy through gene loss and chromosomal rearrangement (Song et al., 1995; Lynch and Conery, 2000; Wolfe, 2001; Kashkush et al., 2002). After tens or hundreds of millions of years of evolution, this diploidization process obfuscates homologous genomic regions. Increasingly, comparative genomic approaches within a phylogenetic framework have revealed evidence of such paleopolyploid events. For example, the complete genome sequencing of the ascomycete yeast Kluyveromyces waltii was used to confirm the ancient polyploid history of Saccharomyces cerevisiae (Kellis et al., 2004). Similarly, genomic sequencing of the fish Tetraodon nigroviridis (Jaillon et al., 2004) revealed the ancient polyploid history of the teleost fish lineage.

Considering the prevalence of recent polyploids among flowering plants, it is not surprising that a plethora of ancient polyploid events have also been discovered. Evidence for ancient polyploidy has come mostly from comparative genetic mapping, analysis of specific gene families, or by the identification of duplicated genes in EST collections, for example, Brassica species (Lagercrantz, 1998), maize (Zea mays; Gaut and Doebley, 1997), grasses (Paterson et al., 2004), legumes (Pfeil et al., 2005), poplar (Populus spp; Sterck et al., 2005), and several other model plants (Blanc and Wolfe, 2004a). However, the most extensive evidence for ancient polyploidy comes from analysis of the complete sequences of both rice (Oryza sativa; Yu et al., 2005) and Arabidopsis.

Studies of duplicate genes and colinearity in the A. thaliana genome differ in their assessment of the total number of polyploidy events but generally agree on the occurrence of at least one complete ancient genome duplication (Arabidopsis Genome Initiative, 2000; Vision et al., 2000; Simillion et al., 2002; Blanc et al., 2003; Bowers et al., 2003). The timing of this most recent (or α) polyploidy event has been placed somewhere between 14.5 and 86 Mya. It has been hypothesized that the polyploidy
event occurred at the divergence of the rest of the Brassicaceae and the genus *Aethionema* (Galloway et al., 1998), at the radiation of the entire Brassicaceae (Blanc et al., 2003), or as distantly as the divergence of the orders Brassicales and Malvales (Bowers et al., 2003). We seek to resolve this issue by comparison between the Brassicaceae and other members of the Brassicales.

Recent phylogenetic analyses have established that the Brassicaceae is sister to the Cleomaceae, with these two families being sister to the Capparaceae (Hall et al., 2002). The Brassicaceae are characterized by their teltalae actinomorphic cruciform flower, with a $2 + 4$ arrangement of stamens, characteristic silique fruit type, and a high percentage of plants with a base chromosome number of $n = 8$ (Warwick and Al-Shehbaz, 2006). The Cleomaceae has greater variation in floral morphology, including showy zygomorphic flowers and the occurrence of C4 photosynthesis (Brown et al., 2005). The genus *Cleome* is by far the largest group in the family with ~200 of the 275 species in the family (Sanchez-Acebo, 2005). A general phylogenetic framework for some of the taxa in these groups is presented (Figure 1).

There are three hypotheses to describe the $\alpha$-polyploidy event in *Arabidopsis* relative to the divergence of the sister families Brassicaceae and Cleomaceae (illustrated in Figure 2). First, the polyploid event could be unique to the Brassicaceae. Second, it could predate the divergence of these families and therefore be shared by the Brassicaceae and Cleomaceae. Finally, there could have been independent ancient polyploidy events in both lineages. For each of these hypotheses, there are specific expectations for genomic organization and relationships among homologous genes. For the Brassicaceae-only hypothesis, there should be a 2:1 ratio of genomic regions and genes in the Brassicaceae compared with the Cleomaceae. For the shared and independent hypotheses, there should be multiple homologous genomic regions and sets of genes in both lineages.

![Figure 1](image1.png)

**Figure 1.** Phylogenetic Relationships of the Brassicaceae, Cleomaceae, Capparaceae, and Several Other Genera in the Order Brassicales.

Modified with permission from Hall et al. (2004).

![Figure 2](image2.png)

**Figure 2.** There Are Three Hypotheses as to When the $\alpha$-Polyploidy Event in the Brassicaceae Occurred Relative to Its Divergence from the Cleomaceae.

The hypotheses are as follows: the polyploidy event was unique to the Brassicaceae (A), the polyploidy event predated the divergence and hence is shared by the two lineages (B), or there were independent polyploidy events in each lineage (C). The three hypotheses can be distinguished by copy number and phylogenetic relationships of genes from *Arabidopsis* (AT; shown in purple) and *Cleome* (Cs; shown in green) that are rooted relative to an outgroup (shown in black).

(A) In the Brassicaceae-only hypothesis, there would be a 2:1 ratio of *Arabidopsis* genes in the Brassicaceae compared with *Cleome* genes in the Cleomaceae.

(B) and (C) For the shared (B) and independent (C) hypotheses, there should be multiple genes in both lineages.

(B) In the shared hypothesis, phylogenetic reconstructions would detect orthologous gene pairs, resulting in a pattern with each *Arabidopsis* gene copy being sister to a *Cleome* gene copy. Also, molecular evolutionary analyses should detect similar degrees of divergence among the homologous genes as measured by synonymous substitution rates ($K_s$) or branch lengths.

(C) In the independent hypothesis, phylogenetic analyses would detect only paralogous sets of loci with *Arabidopsis* gene copies being sister to one another and *Cleome* copies being sister to each other. Furthermore, if the independent polyploidy events were of different ages, then the levels of synonymous substitutions and branch lengths would be different between the two lineages.

However, these two hypotheses can be distinguished by the relationship and evolution among the duplicated copies of each gene (Cannon and Young, 2003; Blanc and Wolfe, 2004a; Chapman et al., 2004; Van de Peer, 2004; Pfeil et al., 2005). If the shared hypothesis was correct, phylogenetic reconstructions should detect orthologous gene pairs, resulting in a pattern with each *Arabidopsis* gene copy being sister to a *Cleome* gene copy. Also, molecular evolutionary analyses would be expected to find a similar degree of divergence among the homologous genes, as measured by synonymous substitution rates ($K_s$). By contrast, the independent hypothesis would predict that phylogenetic analyses detect only paralogous sets of loci, with *Arabidopsis* gene copies being sister to one another and *Cleome* copies being sister to each other. Under this hypothesis, no clear orthology relationships would be detectable between *Cleome* and *Arabidopsis* copies of the genes. Furthermore, if the independent polyploidy events were of different ages, then the levels of synonymous substitutions should be different between the two lineages.

**RESULTS**

**Comparative Cleome BAC Sequencing**

We focused our comparative genomics analysis on regions in *Cleome* that were homoeologous to a duplicated region in
Arabidopsis shared between At3 and At5 (Figure 3A). Our target region was contained within one of the largest recently duplicated blocks identified in Arabidopsis; it was the fourth largest block (block 0305000103160) identified by Blanc et al. (2003) and the fifth largest block (block A12) identified by Bowers et al. (2003). Our target region within this larger block was defined by the following flanking sets of paralogous genes: At3g02230/At5g15650 and At3g02560/At5g16130. This interval spans \(~128\) kb on At3 and \(178\) kb on At5 (for a total of \(306\) kb). There are 36 annotated genes on At3 in this interval and 51 genes on At5. A total of 20 duplicated (paralogous) gene pairs exist between these two regions (Figure 3A).

The screening of our Cleome BAC library with the Cleome SEP and CO probes identified a number of putative positive clones. Screening the clones by restriction digestion and PCR analysis identified three independent homoeologous Cleome genomic regions. BACs from each of the three regions that had nearly complete target regions (identified by PCR) were chosen for shotgun sequencing (Cleome BACs Cs_1, Cs_2, and Cs_3). Sequencing confirmed that Cs_1 and Cs_2 BACs contained

![Figure 3](image-url)

**Figure 3.** Genomic Composition and Alignment of Homoeologous Target Regions Found in A. thaliana and C. spinosa.

(A) Duplicated target region in A. thaliana between At3 (green) and At5 (red) that contains a number of duplicated paralogous genes (yellow) created by ancient polyploidy. The region is flanked by the At3g02230/At5g15650 and At3g02560/At5g16130 paralogs.

(B) Comparison of the two Arabidopsis regions to one of the three sequenced homoeologous regions identified in C. spinosa (Cs_1; shown in blue). The Cs_1 target region contained 2 At3-specific genes (green), 6 At5-specific genes (red), 12 duplicated genes (yellow), and 14 Cleome-specific ORFs (blue; 10 of which are related to a transposon cluster), illustrating the mosaic-like homology of Cleome to Arabidopsis.

(C) Comparison of all five homoeologous genomic regions: two from Arabidopsis and three from Cleome (Cs_1 in blue, Cs_2 in light green, and Cs_3 in purple). Three loci for which there were homologs in each of the five regions, including the loci defining the ends of the target region, are aligned (shown in yellow) to illustrate colinearity.
homologs to At3g02230/At5g15650 and At3g02560/At5g16130. Cs3_3 contained a homolog to At3g02230/At5g15650 but not At3g02560/At5g16130 (the BAC ended at the homolog of At5g16130). However, using PCR, we were able to extend and clone the length of the Cs_3 region to include a homolog to At3g02560/At5g16130. The length of sequence between the flanking genes that define the target region in each BAC was 124, 102, and 143 kb, respectively (for a combined sequence length of 369 kb).

TwinScan identified a number of open reading frames (ORFs) on each BAC. BLASTN and BLASTX analyses found that each of the BACs contained a unique subset of genes with homology to those from At3 and At5 or that were duplicated (Figure 3; see Supplemental Table 2 online). For example, the target region on BAC Cs_1 contained 2 At3-specific genes, 6 At5-specific genes, 13 duplicated genes, and 16 BAC-specific ORFs (most of which are clustered in the center of the BAC and contain a number of transposons) (Figure 3B). Overall, the gene density within the target regions was very similar between Arabidopsis (with an average gene density of one gene per 3.4 kb) and Cleome (with an average gene density of one gene per 3.2 kb).

The five genomic regions (two in Arabidopsis and three in Cleome) were remarkably collinear across the target region and could easily be aligned (Figure 3C). There were three loci for which there were five homologs, one on each genomic region, including the loci defining the ends of the target region (Figure 3C).

The complete list of genes found on these five genomic regions and the homology relationships among them allowed for the reconstruction of the putative ancestral gene order (see Supplemental Table 2 online). In total, the three Cleome BACs contained homologs of 50 of the 66 (76%) unique annotated genes found in Arabidopsis (paralogous genes were counted only once and excluded tRNA, transposons, and tandem duplicates found in Arabidopsis). Of the 17 genes found in Arabidopsis for which no Cleome homologs were identified, many were members of multigene families. Hence, these loci were impractical for additional BAC library screening. Only one of the missing loci was both single copy and highly conserved among many organisms (At5g15920, the structural maintenance of chromosomes family protein [MSS2]). The Cleome homolog of the At5g15920 gene was PCR amplified, cloned, sequenced, and used to reprobe the Cleome BAC library. The putative positive clones were screened and found to correspond to only a single genomic region. Limited shotgun sequencing was done from a pool of four of these BACs. The similarity of these sequences was to the homolog of At5g15920 and then to another part of the A. thaliana genome, suggesting that this locus has undergone a rearrangement in Cleome relative to Arabidopsis.

There were a total of 36 ORFs predicted in the three Cleome BACs that did not have homologs within the target regions of Arabidopsis. Many of these ORFs are short and ultimately may not represent true genes. However, several of these are likely, and potentially novel, genes. For example, Cs3_3 ORF44 is predicted to code for a 392–amino acid peptide whose only similarity is to an unknown peptide in rice (GenBank accession number BAD81800). There are other loci that have homology to Arabidopsis genes but not to genes found in the target interval. Often these are members of large gene families. For example, Cs3_3 ORF7 has high similarity to members of phosphatidylinositol 3- and 4-kinase family proteins. Cs1_3 ORF27 is a 639–amino acid protein with high similarity to pentatricopeptide family proteins (and the gene At1g31430 in particular). Cs1_3 ORF25 has similarity to Cyp89-like cytochrome P450 proteins, and Cs1_3 ORF31 is only 78 amino acids long and has similarity to S-adenosylmethionine decarboxylases. This last ORF may represent a regulatory upstream ORF in the 5′ region of the upstream region of the adenosylmethionine decarboxylase Cs1 At5g15950 gene, as has been described in other systems (Hanfrey et al., 2003).

In addition, six ORFs on Cs1 had similarity to transposon-like genes (Cs1_3 ORFs 20 to 24, 26, and 41). When we did a complete BLASTX of our BAC sequences to the conserved-motif amino acid database of the Brassicaceae compiled by Zhang and Wessler (2004), only Cs1_3 ORF20 and Cs1_3 ORF26 had significant similarity. Both loci were most similar to LINE element reverse transcriptases of the I-f lineage.

For each of the sequenced Cleome BACs, we obtained some sequence beyond the target interval. Most additional sequence was collinear with the homologous At3 and/or At5 regions, except for the BAC Cs_2, which showed evidence of a rearrangement (with the homolog of At5g15650 next to the homolog of At5g18950; data not shown). This represents the only structural rearrangement detected in our data.

Phylogenetic Reconstructions with Homologous Gene Sets

Phylogenetic reconstructions were done for each of the replicated genes identified in both Arabidopsis and Cleome and in a combined analysis of all duplicated loci together. There were 14 homologous gene sets in the target region and four additional sets of loci derived from other genomic regions. Sixteen of the 18 phylogenetic reconstructions and the combined analysis identified paralogous pairs of Arabidopsis and Cleome sequences (Figures 4 to 6). The finding of separate paralogous gene pairs in each lineage supports the hypothesis of independent polyploidy events. Of the two phylogenies showing orthology between Arabidopsis and Cleome sequences (which would support the shared hypothesis), one was in the target region and was complicated by the detection of a potentially ancient tandem duplicate of the gene on At5 (At5g15890 and At5g15900). Hence, this gene set was excluded from additional analyses. Within the target region, there were several genes found in duplicate in Cleome that are single copy in Arabidopsis. Outside the target region, we detected duplicate copies of SEP3 (At1g24260) homologs, a single-copy gene in Arabidopsis.

One of the homologous gene sets examined from our target region was the SEP1 and SEP2 family (Figure 4). Phylogenetic relationships of the SEP genes have recently been established (Zahn et al., 2005). We were able to expand upon this published data set by including duplicated copies of SEP1 and SEP2 homologs from C. sinensis and Boechera stricta (GenBank accession numbers DQ415916 and DQ415918) cloned in this study (the B. stricta sequences were derived from the sequencing of clones from an indexed genomic λ-clone library; Windsor et al., 2006) and from Arabidopsis lyrata (AY727598 and AY727621). Cloned sequences from other taxa were used to increase taxonomic breadth, including Castanea mollissima.
(DQ148295), Prunus dulcis (AY947464), Rosa rugosa (AB099876), G. max (DQ159905), and Pisum sativum (AY884290). EST sequences from B. napus (CD844028 and CD816888), tree cotton (Gossypium arboreum; BG440326), and citrus (Citrus sinensis; CK932769) were also used.

We recovered the same tree topology for the SEP1 and SEP2 genes from a number of eudicot species were analyzed (many reported in Zahn et al., 2005), including duplicates from C. spinosa cloned from our target region. There is a Brassicaceae-specific gene duplication with orthologous gene sets shared by Arabidopsis, Boechera, and Brassica for SEP1 and SEP2. The Brassicaceae clade is sister to the duplicated paralogous genes from C. spinosa.

In order to have an independent assessment of gene duplication, we present a representative of a duplicated gene from outside our target region used for phylogenetic reconstruction, the Arg decarboxylase (ADC) family (Figure 5). The ADC gene was duplicated between At2 and At4 (ADC1, At2g16500; ADC2, At4g34710). Sequences of ADC genes were used previously for phylogenetic studies within the Brassicaceae where the gene was found to be duplicated across the entire Brassicaceae, except in the genus Aethionema (Galloway et al., 1998). By PCR we were able to amplify duplicated copies of the ADC gene from C. spinosa. Our phylogenetic analyses agree with those of Galloway et al. (1998). Most important, however, was our finding that the duplicates of Cleome are monophyletic and are therefore independent of the duplicates in the Brassicaceae.

Finally, we combined the coding sequence data for all duplicated genes in the target regions for Arabidopsis and Cleome for...
phylogenetic analysis (Figure 6). The At3 and At5 sequences contained data for all 14 duplicated genes. However, each of the three regions in Cleome contained some, but not all, of the duplicates. Hence, the missing genes were scored as missing data. Recent studies have concluded that phylogenetic studies containing even large amounts of missing data can be more robust than many smaller gene and taxa studies (Driskell et al., 2004). The unrooted phylogeny has 100% bootstrap support for separate Arabidopsis and Cleome clades (Figure 6). Also, the branch lengths for the Arabidopsis sequences are longer than for the Cleome sequences. There is some support for a closer relationship of the Cleome genomic regions Cs_2 and Cs_1 (with 67% bootstrap support).

**Calculation and Analysis of Synonymous Substitution Rates**

The means for synonymous substitutions, or Ks values, between duplicate copies of genes were calculated for the three possible comparisons: Arabidopsis-Arabidopsis, Arabidopsis-Cleome, and Cleome-Cleome (Table 1). Our results found Arabidopsis-Cleome > Arabidopsis-Arabidopsis > Cleome-Cleome. Our mean value for the Arabidopsis-Arabidopsis of Ks = 0.67 is very similar to that of Maere et al. (2005), who used a value of Ks = 0.7 for the Arabidopsis polyploidy event and slightly lower than the value of Ks = 0.8 calculated by Lynch and Conery (2000). Assuming that Ks values increase with the passage of time, this suggests a temporal pattern for the divergence and subsequent timing of the individual polyploidy events of these two lineages.

We also performed an analysis of variance on the Ks values from the comparisons of Arabidopsis-Arabidopsis and Cleome-Cleome and looked for locus-specific variation in synonymous substitution rates (Table 2). Levels of synonymous differentiation differed significantly between Arabidopsis-Arabidopsis and Cleome-Cleome (P = 0.0007), but we found no evidence of heterogeneous rates among loci (P = 0.228). Hence, the age of

Figure 5. Phylogenetic Reconstruction Using the ADC Family, a Representative of a Duplicated Gene from Outside Our Target Region, Supports the Hypothesis of Independent Polyploidy Events in the A. thaliana and C. spinosa Lineages.

Sequences of ADC genes were used previously for phylogenetic studies within the Brassicaceae (Galloway et al., 1998). Our phylogenetic analyses agree with those of Galloway et al. (1998) in showing the duplication of ADC1 and ADC2 shared by most Brassicaceae, except in the genus Aethionema. Most important, however, was our finding that the duplicates of Cleome are monophyletic and are therefore independent of the duplicates in the Brassicaceae.
The sequencing of the three replicated genomic regions from A. thaliana allowed us to examine the overall patterns of duplicate gene retention and loss compared with that observed within C. spinosa. A χ² test of independence between species and copy number (single versus duplicate) found a highly significant result (P = 0.0079). Retention of duplicate copies in the two species was nonrandom. The copy number of particular loci is correlated in these two independently evolving lineages (Table 3).

We compiled the gene ontology (GO) annotations of the genes that were either duplicated in both lineages or single copy in both lineages. Although genes can have multiple GO annotations and therefore may be represented multiple times, we report the number of gene annotations in each molecular function GO category for single and duplicated copy genes (Figure 7). The most obvious overrepresented category for single-copy genes was hydrolase activity, and the most overrepresented category for duplicated genes was protein binding (Figure 7).

Assuming that there were 50 ancestral genes in the target region at the split between Brassicaceae and Cleomaceae (this is a conservative number based on the number of genes shared now between Cleome and Arabidopsis), a genome duplication in the Arabidopsis lineage would have resulted in 100 loci and a genome triplication in Cleome with 150 loci. For these genes in the two Arabidopsis regions, there are currently 30 genes in single copy and 20 in duplicate (20 + 20 = 40) for a total of 70 loci. The overall degree of gene retention is thus 70/100 = 70%. In Cleome, there are 24 single-copy genes, 23 in duplicate (23 + 23 = 46), and 3 in triplicate (3 + 3 + 3 = 9) for a total of 79 loci. Thus, the overall degree of gene retention is only 79/150 = 53%.

**DISCUSSION**

Since the ancient polyploid history of A. thaliana was illuminated by genome analyses, there has been much speculation about the timing and phylogenetic position of the genome doubling (Arabidopsis Genome Initiative, 2000; Vision et al., 2000; Blanc et al., 2003; Bowers et al., 2003; Ermolaeva et al., 2003; De Bodt et al., 2005). We took a comparative genomic approach to address this issue in relation to the divergence of the sister families Brassicaceae and Cleomaceae (Hall et al., 2002). Our results strongly support the hypothesis of independent ancient polyploidy events of different ages in these two sister lineages. First, we identified three independent and unique BACs that were homoeologous to a duplicated region between At5 and At3. We also documented four additional duplicated homologous genes located in other parts of the A. thaliana genome. Second, the phylogenetic analysis of almost all of the duplicated homologous gene groups and of a combined data set found pairs of Arabidopsis paralogs, clearly separate from Cleome paralogs. Third, the synonymous substitution rates (Kₛ values) between Arabidopsis paralogs and Cleome paralogs were significantly different. All of these factors combine to make an independent polyploidy event the most likely explanation. However, until complete genomic data becomes available, we cannot yet rule out the possibility of segmental polyploidization in Cleome.

The occurrence of independent ancient polyploidy events in these closely related lineages allows analysis of the relative timing of polyploidization, patterns of gene retention and loss, and genome evolution. Our results highlight the continuous cycles of polyploidy and diploidization that have occurred during angiosperm evolution.

**Timing of Ancient Polyploidy Events**

Previous studies have used several approaches to infer the timing of the recent (or ω) polyploidy event in Arabidopsis (Blanc et al., 2003; Bowers et al., 2003; Maere et al., 2005). Bowers et al. (2003) found that the polyploidy event occurred sometime between the divergence of the Brassicales from the Malvales (estimated to have occurred 86 Mya) and the divergence of Arabidopsis from Brassica (estimated to have occurred 14.5 to 20 Mya). Blanc et al. (2003) used divergence times calculated for

### Table 1. Least Square Means of Synonymous Substitution Rates Calculated from Sequence Alignments from Loci That Are Duplicated in Both A. thaliana and C. spinosa

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Mean</th>
<th>se</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>At–Cs</td>
<td>0.816</td>
<td>0.036</td>
<td>13</td>
</tr>
<tr>
<td>At–At</td>
<td>0.674</td>
<td>0.036</td>
<td>13</td>
</tr>
<tr>
<td>Cs–Cs</td>
<td>0.398</td>
<td>0.036</td>
<td>13</td>
</tr>
</tbody>
</table>

Synonymous substitution rates measured by Kₛ, which equals the number of substitutions per synonymous site in a coding sequence. At, A. thaliana; Cs, C. spinosa.
the Brassicaceae (Koch et al., 2001) and placed the polyploidy event some 24 to 40 Mya, near the divergence of the Brassicaceae from the Cleomaceae. To circumvent the problems of molecular dating and the conflicts over molecular clocks, Maere et al. (2005) suggested reporting the age of ancient polyploid events in Ks units rather than in millions of years. By this measure, Maere et al. (2005) have given a unit value of $K_s = 0.7$ to the Arabidopsis genome duplication. Our $K_s$ estimate of 0.67 for Arabidopsis paralogs is very similar to this.

Following the suggestion of Maere et al. (2005), we use our calculated $K_s$ values to estimate the temporal pattern of divergence of the two lineages and of the two independent polyploidy events. The largest estimate of $K_s$, and hence the oldest event, is the divergence of Cleomeae and Brassicaeae lineages ($K_s = 0.82$). The next event is polyploidization within the Brassicaeaeae ($K_s = 0.67$) and finally the polyploidization within the Cleomeaeae ($K_s = 0.40$). For comparison, Blanc et al. (2003) estimated the median difference between cotton and Arabidopsis as $K_s = 1.8$. Thus, the divergence of the Brassicaeaeae and Cleomeaeae is much more recent than the divergence of the Brassicales and Malvales, agreeing with phylogenetic reconstructions (APG II, 2003; Hall et al., 2004).

Patterns of Gene Retention and Loss

The occurrence of independent ancient polyploid events in closely related lineages provides an excellent opportunity to explore patterns of gene evolution by looking for convergent fates in gene copy number. By analyzing the Arabidopsis genome, several authors have suggested that there are biases and evolutionary pressures for the types of genes to be maintained in duplicate or single copy (Blanc and Wolfe, 2004b; Seoighe and Gehring, 2004; Maere et al., 2005; Chapman et al., 2006). In our study, we found a highly significant bias for the genes occurring in duplicate or single copy in both Arabidopsis and Cleome, although the overlap is not absolute. Thus, comparative analysis could suggest which genes are particularly advantageous to maintain in duplicate or to have in single copy. Alternatively, genes found in duplicate in one lineage but returned to single copy in the other provide a good system to explore hypotheses of subfunctionalization and neofunctionalization (Force et al., 1999; Prince and Pickett, 2002). When we examined the GO annotations of the genes found in duplicate in both lineages or single copy in both lineages, we found several biased categories despite our rather small sample of the genomes (only ~0.2% of the Arabidopsis genome). The most obviously overrepresented single-copy gene category was for hydrolase activity. Interestingly, Maere et al. (2005) found several biased categories despite our rather small sample of the genomes (only ~0.2% of the Arabidopsis genome). The most obviously overrepresented single-copy gene category was for hydrolase activity. Interestingly, Maere et al. (2005)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>F-Ratio</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species comparison</td>
<td>1</td>
<td>20.452</td>
<td>0.0007</td>
</tr>
<tr>
<td>Locus</td>
<td>12</td>
<td>1.556</td>
<td>0.2280</td>
</tr>
<tr>
<td>Error</td>
<td>12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$n = 26; r^2 = 0.765.$

There is a well-supported New World temperate clade containing the genera Cleomella, Oxyystis, Witsenia, and Isomeris that is sister to the remainder of the family (Hall et al., 2002). These genera have a base chromosome number of $n = 20$, suggesting a shared and potentially recent polyploid history. The genus Polanisia ($n = 10$) is then sister to the Old and New World members of the genus Cleome (with a base chromosome number of $n = 10$ in Cleome section Tarenaya; H.H. Iltis, personal communication). The work of Galloway et al. (1998) detected only a single ADC gene in Polanisia that appears sister to the duplicated copies detected in Cleome (Figure 5). In future work, we hope to determine if the ancient polyploid event detected in the Cleome lineage is shared with Polanisia and/or with the North American Cleomaceae clade, or, alternatively if there was yet another independent polyploid event within this North American group.

Table 3. A 2 x 2 Contingency Table Comparing the Observed and Expected Numbers of Single versus Duplicated Copy Genes under the Hypothesis of Independence between the Two Lineages Using a χ² Test of Independence

<table>
<thead>
<tr>
<th>A. thaliana Genes</th>
<th>Single</th>
<th>Duplicated</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cleome Genes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>19 (14.4)</td>
<td>5 (9.6)</td>
<td>24</td>
</tr>
<tr>
<td>Duplicated</td>
<td>11 (15.6)</td>
<td>15 (10.4)</td>
<td>26</td>
</tr>
<tr>
<td>Sum</td>
<td>30</td>
<td>20</td>
<td>50</td>
</tr>
</tbody>
</table>

χ² Test
χ² Value: 7.065
P value: 0.00786**

Asterisks indicate that the result is significant at the 0.01 level.
found hydrolases to be the most biased GO molecular function category for single-copy genes (or duplicated by small tandem duplications). Similarly, the most biased duplicate copy gene category we identified was protein binding genes. Maere et al. (2005) also found protein binding genes to be one of the most biased to be retained in duplicate due to ancient polyploidy. More complete genome sampling of Cleome could further refine GO category biases. These results emphasize, however, that common processes have been operating in these independently evolving lineages, rather than lineage-specific biases in gene retention or loss.

A comparative analysis of genes within our target region has the potential to elucidate gene evolution and function. The maintenance of duplicate copies of genes in both Arabidopsis and Cleome may indicate that it is beneficial or necessary to do so. For example, we found SEP1 and SEP2 genes to be duplicated in both lineages. This agrees with published reports that MADS box genes, and SEP genes in particular, often have been maintained in duplicate throughout angiosperm evolution (Irish and Litt, 2005; Malcomber and Kellogg, 2005; Zahn et al., 2005), allowing for the evolution of new floral forms by subfunctionalization and neofunctionalization. Recent work with SEP1 and SEP2 in Arabidopsis found significant differences in gene expression, suggesting a degree of subfunctionalization (Duarte et al., 2006). Investigations of the patterns of expression of MADS box genes, such as SEP1 and SEP2 homologs, could be done to explore differences in floral development.

Other genes are duplicated in one lineage and single copy in the other. For example, CO is replicated in Arabidopsis but exists in single copy in Cleome. In Arabidopsis, there is a tandem duplication on At5, giving rise to CO and CONSTANS-LIKE1 (COL1), and their paralog on At3, COL2 (Ledger et al., 2001). CO is an important regulator of flowering time, integrating signals for daylength (Corbesier and Coupland, 2005), whereas COL1 and COL2 mutants have no obvious flowering time phenotypes but may be involved with the circadian clock (Ledger et al., 2001). A comparison of the function of the single CO-like homolog in Cleome to the three loci in Arabidopsis could be done to look for evidence of subfunctionalization.

**Polyploidization in Cleome**

Analysis of the triplicate regions in Cleome enables several inferences about the polyploidization process. First, we argue...
that the triplication was most likely due to whole-genome polyploidization and not segmental duplication or aneuploidy. Second, we discuss evidence that the Cleome lineage was derived from an ancestral hexaploid ancestor. In addition to the triplicated genomic region that we sequenced in Cleome, we also amplified replicated copies that for four additional loci. These additional loci are found in duplicate in Arabidopsis and are scattered about the genome. The finding of additional Cleome replicated genes of the same approximate age as our triplicated region suggests that there was a whole-genome triplication. Additionally, phylogenetic reconstructions with these loci show similar patterns suggestive of independent polyploidy events in the Arabidopsis and Cleome lineages (e.g., Figure 5). It is possible that replicated regions in Cleome were derived by aneuploidy or segmental duplication. Examples of large-scale segmental duplications have been recently detected in the rice genome (Yu et al., 2005). Such debates over segmental versus whole-genome duplication in various lineages have a long history (Wolfe, 2001; Panopoulou and Poustka, 2005) and can be resolved by whole-genome sequence analysis (Jaillon et al., 2004; Kellis et al., 2004; Paterson et al., 2004). The generation of additional sequence data and/or comparative mapping information from Cleome will likely be necessary to fully determine whether it underwent segmental duplication or polyploidy.

The likely whole-genome triplication observed in Cleome suggests the ancestral Cleome polyploid was a hexaploid (6x), which subsequently underwent diploidization. For example, it has been hypothesized that diploid Brassica species are derived from a hexaploid ancestor, and now much of the genome exists in triplicate compared with Arabidopsis (Lagerrantz, 1998; Lukens et al., 2004; Lysak et al., 2005; Parkin et al., 2005). In addition, Brassica shares the ancient polyploidy event seen in Arabidopsis (Bowers et al., 2003; Parkin et al., 2005). Hence, diploid Brassica species contain six genomic regions that are homoeologous to a single ancestral region. Therefore, it is feasible that a direct comparison of two Arabidopsis, six Brassica, and three Cleome homoeologous regions could be made.

Genome Evolution

After polyploidization, genomes begin the diploidization process involving both gene loss and chromosomal rearrangements (Ma and Gustafson, 2005). A comparison of our Cleome and Arabidopsis regions allows us to examine patterns of gene loss, microsynteny, and their effects on genome size.

The patterns of gene loss are important for understanding both the evolution of the genes themselves and the genome in general. For genes present in both genomes, 60% of Arabidopsis genes are now present in single copy compared with 48% in Cleome. Although more genes are found in single copy in A. thaliana, the degree of gene retention is higher (70% versus 53%). The lower degree of gene retention in Cleome is surprising considering it is a much younger polyploidy event. The difference of gene loss between the two lineages is likely attributable to the initial degree of replication. For a gene to become single copy in Cleome, two paralogous sequences must be lost. Most replicated genes in Cleome are duplicated (23 pairs) rather than triplicated (only three loci), meaning one gene copy has been lost. A similar degree of high gene loss in ancient triplicated genomes has been found by analysis of the diploid Brassica genomes (O'Neill and Bancroft, 2000; Quiros et al., 2001; Rana et al., 2004; Yang et al., 2005). These results of increased gene loss in ancient hexaploids compared with ancient tetraploids is consistent with an analysis of a large number of taxa where it was found that mean DNA amount per ancestral genome tended to decrease with increasing ploidy (Leitch and Bennett, 2004).

Of the genes found in the two Arabidopsis regions, we were unable to find homologs to 24% of them in the Cleome regions. Although most of these were members of gene families in Arabidopsis, we did find evidence for the transduction of at least one locus. Similarly, we predicted 36 unique ORFs in the three Cleome regions. Many of these were short and may not represent true genes. However, many of these Cleome ORFs were similar to members of gene families in Arabidopsis but did not have homologs within the target interval. Hence, in both Arabidopsis and Cleome, there appears to be a migration of members of gene families in and out of various genomic locations that affects the microsynteny of the genomes.

In addition to gene loss and movement, chromosomal rearrangements play an important role in the evolution of genomes after polyploidization. The three regions in Cleome and the two regions in Arabidopsis showed a remarkable colinearity across our target region despite the degree of gene loss and movement that has occurred in each lineage subsequent to polyploidization. We detected a single inversion or rearrangement in one of the sequenced BACs, but this lay outside of our target interval. The degree of large-scale rearrangements can be ascertained by comparative mapping between Brassicaceae species (e.g., Koch et al., 2001; Parkin et al., 2005) and a genetic linkage map of Cleome. Such broad taxonomic comparative mapping has been very successful for understanding genome evolution in the grasses (Feuilllet and Keller, 2002; Devos, 2005).

The processes of gene loss and chromosomal rearrangements will obviously contribute to the evolution of genome size. Within our target interval, the three Cleome regions are only 1.2 times the size of the two Arabidopsis regions, and the complete genome of Cleome is —1.9 times the size of Arabidopsis (M.E. Schranz and T. Mitchell-Olds, unpublished data; Johnston et al., 2005). Hence, the Cleome genome is quite compact, especially considering that the A. thaliana genome is greatly reduced compared with other members of the genus Arabidopsis (Johnston et al., 2005). For example, the closest nonpolyploid relative of A. thaliana reported in Johnston et al. (2005), A. halleri, has a genome 1.6 times that of A. thaliana. The small genome size of Cleome may be due to both increased gene loss and suppression of transposon activity. Within our Cleome BACs, there were very few transposons, with the only clear similarity to ubiquitous LINE elements. The ancient hexaploid Brassica genomes range from 3.3 to 4.4 times the size of Arabidopsis (Johnston et al., 2005). Therefore, although there is substantial gene loss in Brassica genomes after polyploidization, the overall genome size is relatively large because of transposon expansions (Zhang and Wessler, 2004). The combination of small genome size, independent polyploidy, and close phylogenetic position to Arabidopsis all make Cleome an attractive candidate for complete genome sequencing.
The use of both comparative genomic and phylogenetic approaches was critical for our elucidation of the independent polyploidy events. A phylogenetic approach alone, for example, using EST sequences, would likely have detected independent polyploid events but probably would have concluded that Cleome was duplicated rather than triplicated. A comparative mapping approach may have been able to detect ancient triploidy, but without the phylogenetic analysis it would have been impossible to say that they were independent polyploidy events. Our genomic approach also allowed for an assessment of the types of genes present in duplicate versus single copy. Finally, the occurrence of independent ancient polyploidy events in closely related lineages provides an interesting opportunity to examine duplicate gene evolution and its contribution to phentootypic diversity.

**METHODS**

**Plant Materials and DNA Isolation**

Plants were grown from seeds of *Cleome spinosa* (ES1046; Spinnenpflanze) purchased from Kiepenkerl. DNA isolations were done using Qiagen genomic tips and by a modified CTAB procedure as previously described (Schranz et al., 2005).

**Amplification, Cloning, and Sequencing of PCR Products**

Conserved primer pairs were identified from the alignment of paralogous gene pairs found in the *Arabidopsis thaliana* genome (Blanc et al., 2003; Bowers et al., 2003). Most of the primers were to loci within our target duplication region between At3 and At5 (see below). However, an additional four primer pairs were to paralogous genes located in other parts of the *A. thaliana* genome. In addition, several primer sets were designed to genes found in single copy in *A. thaliana*. The genes and primer sequences used can be found in Supplemental Table 1 online.

PCR products were obtained with 35 cycles as follows: 94°C, 30 s/60°C, 30 s/72°C, 2 min. Two independent PCRs were performed for each sample, and products were cloned using the TOPO TA cloning kit (Invitrogen Life Technologies). At a minimum, four clones per cloning reaction were sequenced as above on both strands with an ABI3700 capillary sequencer (Applied Biosystems).

**BAC Library**

A BAC library for our *C. spinosa* genotype was made by Keygene and cloned into a pBludigBAC-536 vector. Colony picking and filter spotting was done by RZPD (library No. 1097).

**Clone Identification and Sequencing**

For comparative genomic analyses, we identified *C. spinosa* BAC regions that corresponded to a duplicated *A. thaliana* region between At3 and At5. The target regions in *Arabidopsis* were centered on paralogous pairs of the **SEP** and **CO** regulatory genes (**SEP**2 [At3g02310] and **COL2** [At3g02380] located on At3 and **SEP**1 [At5g15800] and **CO** [At5g15840] located on At5).

The homologous *C. spinosa* BAC clones were identified by filter hybridization of girdled nylon membranes containing the clones of the BAC library. Gene products for the *Cleome* homologs of **CO**/**COL2** and **SEP**1 and **SEP**2 were amplified, and PCR products were gel-purified using the QIAquick gel extraction kit (Qiagen). One hundred nanograms of each purified PCR product was labeled and hybridized using the ECL Direct nucleic acid labeling and detection system (Amersham Biosciences) according to the manufacturer’s protocol.

Several positive clones were isolated from each hybridization and amplified in small overnight culture, and DNA minipreparations for each BAC were done using standard alkaline lysis procedures. The minipreparations of BAC DNA were fingerprinted by restriction digest and screened by PCR with primers to many of the putative genes located within the genomic target region.

Three distinct BAC regions, each covering the approximate interval from At5g15650 to At5g16130, were grown in 250 mL TB liquid cultures with chloramphenicol selection. BAC DNA isolation was done using the NucleoBond BAC100 kit ( Macherey-Nagel) according to the manufacturer’s instructions. The BAC DNA was then cleaned of contaminating sequences using plasmid-safe ATP-dependent DNase (Epicentre Biotechnologies). The selected BAC clones were shotgun sequenced by subjecting each to separate partial digestions with Tsp509I and Sau3AI enzymes, size fractionating into 12- to 18-kb fragments, cloning into a pUC19 vector, and end-sequencing the clones with an ABI3700 capillary sequencer. Sequences were assembled and edited with Seqman 5.0 (DNASTAR).

**Sequence Annotation**

Gene detection from each *C. spinosa* BAC sequence was done using the TwinScan program (Korf et al., 2001) with *A. thaliana* as the informant genome. The homology of each ORF predicted by TwinScan relative to *A. thaliana* was ascertained using BLASTN or WU-BLAST servers at the National Center for Biotechnology Information and The Arabidopsis Information Resource (TAIR) (Altschul et al., 1990; WU-BLAST, W. Gish, 1996–2004; http://blat.wustl.edu). We named the genes using the first letters of the species name followed by the BAC number (1, 2, or 3) and the unique identifier from the TAIR ID in *A. thaliana*. If the homology was to a duplicated locus between At3 and At5, we always used the homology to At5 for naming purposes. For example, the three homologs of At5g15650 genes in *C. spinosa* are named as follows: Cs1_At5g15650, Cs2_At5g15650, and Cs3_At5g15650. The BLAST analyses were also used to identify additional homologous coding sequences from other taxa to be used in phylogenetic analyses.

**Sequence Alignment and Phylogenetic Analysis**

Duplicate copies of 13 loci from the target region and an additional four loci from cloned PCR products from other regions were found in both *A. thaliana* and *C. spinosa*. The coding sequences of these loci were used for phylogenetic and molecular evolutionary analyses.

Alignments of homologous sets of loci were done on translated protein sequences using ClustalW (Thompson et al., 1994) with a gap open penalty of 8 and gap extension penalty of 0.3. All nucleotide sequence alignments were then visually inspected to improve the alignment.

Phylogenetic reconstruction was performed by analyzing DNA sequences with maximum-likelihood (ML) methods in the PHYML program (Guindon et al., 2005). For the likelihood analyses of each of the independent and combined analyses, we selected models of molecular evolution on the basis of MODELTEST (Posada and Crandall, 1998) as implemented by FindModel (http://hcv.lanl.gov/content/hcvdb/findmodel/ findmodel.html). The ML parameter values were optimized, with a BIONJ tree as a starting point (Gascuel, 1997). Support values for nodes on the ML tree were estimated with 1000 bootstrap replicates (Felsenstein, 1985).

In addition, the aligned duplicated gene coding sequences from At3, At5, and the three *Cleome* BAC regions were each concatenated into one large sequence. Missing genes were treated as gaps. The total aligned sequence was 12,934 bp long. In order to examine the overall relationships of the regions, these five concatenated sequences were used for
phylogenetic analysis in the same manner as described above for individual genes. The phylogenies were then examined for branching patterns of Cleome copies relative to Arabidopsis copies for evidence of orthologous versus paralogous relationships (Bowers et al., 2003; Chapman et al., 2004; Pfeil et al., 2005).

Evolutionary Analysis
For 13 genes that were duplicated in both Arabidopsis and Cleome, synonymous substitution rates (Ks values) were calculated from the alignments of the coding regions using the DnaSP v.4.10.4 computer program (Rozas et al., 2003). There were three categories of comparison: (1) Arabidopsis–Arabidopsis, (2) Arabidopsis–Cleome, and (3) Cleome–Cleome. Within each category of comparison, we calculated Ks among all pairwise gene combinations and summarized the central tendency of each comparison using the median Ks value. These calculations were performed for each locus. Finally, the median value for each comparison for each locus was used as input data for a fixed effects analysis of variance testing cross-classified effects of comparison and locus using Systat v.10 (Systat Software). Least square mean values of Ks for these three categories of comparison are reported in Table 1. Table 2 compares Ks values among Arabidopsis paralogs versus among Cleome paralogs (cross-classified analysis of variance with 1 df for the species comparison).

We built a 2 × 2 contingency table (Table 3) to compare the observed and expected numbers of duplicate versus single-copy genes under the hypothesis of independence between the two lineages and performed a standard χ² test of independence using DnaSP v.4.10.4 (Rozas et al., 2003). GO annotations for A. thaliana genes in our target genomic regions were downloaded from TAIR (www.arabidopsis.org). The GO annotations of genes found in duplicate in both Arabidopsis and Cleome were contrasted to those found in single copy in both (Figure 7).

The amount of gene loss in the target region was calculated for both Arabidopsis and Cleome, assuming that there were 50 ancestral genes (the number of loci shared between the two genomes in the target region). Genome duplication in the Arabidopsis lineage would result in 100 loci, and a genome triplication in Cleome with 150 loci. The number of loci for these shared genes now present in the two genomes was used to calculate the degree of gene loss.

Accession Numbers
Annotated sequence data for each BAC can be found in the GenBank/EMBL data libraries under accession numbers DQ415920, DQ415921, and DQ415922. SEPF and SEP2 homologs from Boechera stricta can be found under accession numbers DQ415916 and DQ415918. Duplicated ADC loci from C. spinosa can be found under accession numbers DQ415917 and DQ415919.

Supplemental Data
The following materials are available in the online version of this article.

Supplemental Table 1. PCR Primers Used to Clone Homologs from Cleome spinosa.

Supplemental Table 2. All Genes Found in the Homoeologous Target Regions in Arabidopsis thaliana and Cleome spinosa, with the Shared Genes between the Two Lineages Placed into a Reconstructed Ancestral Gene Order.

ACKNOWLEDGMENTS
We thank Jocelyn Hall and the three reviewers for helpful comments on the manuscript. We also thank Heiko Vogel, Karl Schmid, Domenica Schnabelrauch, Xiaoyu Zhang, and Aaron Windsor for technical advice and assistance. Finally, we thank Maria Clausn, Chris Pires, Martin Lysak, Amy Bouck, Marcus Koch, David Baum, George Coupland, Laurent Corbesier, Jeff Doyle, and Ihsan Al-Shehbaz for illuminating discussions. The Max Planck Gesellschaft and Duke University provided support for this research.

REFERENCES


