Supernetwork Identifies Multiple Events of Plastid *trn*F$_{(GAA)}$ Pseudogene Evolution in the Brassicaceae

Marcus A. Koch, Christoph Dobeš,* Christiane Kiefer,* Roswitha Schmickl,* Leoš Klimeš,† and Martin A. Lysak‡§*

*Heidelberg Institute for Plant Science, Biodiversity and Plant Systematics, Heidelberg University, Heidelberg, Germany; †Institute of Botany, Academy of Sciences of the Czech Republic, Třeboň, Czech Republic; ‡Jodrell Laboratory, Royal Botanic Gardens, Kew, UK; and §Department of Functional Genomics and Proteomics, Masaryk University, Brno, Czech Republic

The occurrence of nonfunctional *trn*F pseudogenes has been rarely described in flowering plants. However, we describe the first large-scale supernetwork for the Brassiccaeae built from gene trees for 5 loci (*adh*, *chs*, *mat*K, *trn*L-F, and ITS) and report multiple independent origins for *trn*F pseudogenes in crucifers. The duplicated regions of the original *trn*F gene are comprised of its anticodon domain and several other highly structured motifs not related to the original gene. Length variation of the *trn*L-F intergenic spacer region in different taxa ranges from 219 to 900 bp as a result of differences in pseudocopy number (1–14). It is speculated that functional constraints favor 2–3 or 5–6 copies, as found in *Arabidopsis* and *Boechera*. The phylogenetic distribution of microstructural changes for the *trn*L-F region supports ancient patterns of divergence in crucifer evolution for some but not all gene loci.

Introduction

Among the various families of flowering plants, the Brassicaceae are noteworthy for numerous reasons. Several representatives achieved the well-accepted status of "model organisms." These include *Arabidopsis thaliana*, the *Brassica*'s, and also a few others such as *Capsella*, *Arabis*, or *Boechera* (Clauss and Koch 2006; Koch and Mummenhoff 2006; Lysak and Lexer 2006). These model organisms have advantages that are summarized in recent reviews (e.g., Lysak and Lexer 2006).

However, in this contribution, we would like to focus on a different aspect of crucifer research: investigation of the evolutionary dynamics of a genome region within the framework of a well-supported phylogeny. The possibility to use the genomic information from *A. thaliana* has been used to develop molecular markers for elucidating crucifer evolution, and these efforts have identified gene loci from 3 plant genomes (plastome, mitochondrion, and nuclear genome) useful for phylogenetic reconstructions (Koch 2003; Beilstein et al. 2006; Schranz and Mitchell-Olds 2006). As a result, we are indeed close to a first comprehensive phylogenetic overview on the systematics and phylogeny of the entire family (e.g., Al-Shehbaz et al. 2006; Beilstein et al. 2006). Within the last year alone, there have been significant advances in our understanding based upon earlier molecular analyses (e.g., Koch et al. 2000, 2001; Heenan et al. 2002). It is remarkable that the wealth of detailed information about phylogenetic relationships inferred from different molecular data sets is largely congruent and that a comprehensive phylogeny for the majority of Brassicaceae species is now emerging (Bailey et al. 2006). This represents the efforts of many research groups addressing questions of crucifer evolution at different taxonomic levels. More than 100 phylo/biogeographic studies are available (Koch and Kiefer 2006). There has also been the concurrent development of important resources such as a taxonomic and cytological database (Warwick and Al-Shehbaz 2006; Warwick et al. 2006). These data have allowed us to claim that tribal, subtribal, or generic classification systems (e.g., Hayek 1911; Schulz 1936; Janchen 1942) are highly artificial and to propose a new and highly reliable classification scheme (e.g., Al-Shehbaz et al. 2006).

This depth of phylogenetic information now provides a solid framework for investigating processes of molecular evolution such as selection, recombination, duplication, and gain and loss of function (e.g., Cork and Purugganan 2005). Recent studies on Brassicaceae taxa have included those on glycine-rich pollen surface proteins (Fiebig et al. 2004), alcohol dehydrogenase (Charlesworth et al. 1998; Koch et al. 2000), and acidic chitinase (Bishop et al. 2000). It is also important to understand the evolutionary dynamics of DNA markers that have wide application in molecular systematics. Much is already known about *rbc*L (De Pamphilis and Palmer 1990) and the nuclear internal transcribed spacer (ITS) of ribosomal RNA ITS1 and ITS2. The latter evolves by a process of concerted evolution (Koch et al. 2003). Other markers to consider are those which in the last few years have been considered as putative species-specific DNA sequences for DNA barcoding efforts (Kress et al. 2005). Among these, the plastidic *trn*LF region is one marker that has been widely used and applied successfully in analyzing basal angiosperm (Borsch et al. 2003) and land plant evolution (Quandt et al. 2004). This chloroplast region is among a few other selected regions from the plastome that might be appropriate in combination with the nuclear ITS to serve as a DNA barcode to characterize and identify flowering plant species (Kress et al. 2005). Therefore, it is important that we learn more about the evolutionary properties of the *trn*LF region.

The intergenic *trn*LF spacer separates the second exon of the *trn*L$_{UAA}$ gene and the exon of the *trn*F$_{GAA}$. This spacer region exhibits a remarkably high level of length variation among land plants, varying from less than 60 bp in some moss species to 350–450 bp in many land plants (Borsch et al. 2003). Although the entire *trn*LF region and its corresponding genes are cotranscribed (Kanno and Hirai 1993), *trn*F gene promoter elements have also been found to occur close to the 5′ start codon of the *trn*F

Key words: Brassicaceae, plastome, pseudogene evolution, *trn*F gene, supernetwork.

E-mail: marcus.koch@urz.uni-heidelberg.de.

Mol. Biol. Evol. 24(1):63–73. 2007
doi:10.1093/molbev/msl130
Advance Access publication September 20, 2006

gene. These show high similarity to a putative sigma70-type bacterial promoter motif (−35 TTGACA/−10 GAGGAT). In a comprehensive study across land plants, these 2 motif elements have been frequently observed (Quandt et al. 2004), eventually leading to the conclusion that they represent the ancient and original *trn*F_{GAA} promoter. In a recent study focusing on the genus *Arabidopsis*, several copies of nonfunctional *trn*F_{GAA} pseudogenes were described and characterized (Koch et al. 2005). Interestingly, a pseudogene was found to be exclusively inserted between 2 promoter elements. This supports the assumption that this promoter is no longer functional. Pseudogenic *trn*F genes occur only in a few flowering plants. They have been described from Asteraceae (Vijverberg and Bachmann 1999; Wittzell 1999), Juncaceae (Drábkova et al. 2004), and soon will be from Orchidaceae (Fischer G, unpublished data).

The pseudogenes of the various cruciferous taxa documented previously (Koch et al. 2005) differ from those of the other plant families in structure and copy number. The most conserved multicopy motif is the anticodon domain (Vijverberg and Bachmann 1999; Koch et al. 2005) that is represented by up to 8 different copies in *Arabidopsis* (Koch et al. 2005). In cruciferous taxa, this pseudogene anticodon domain is not flanked by the original D-domain, T-domain, and the 2 acceptor stem regions, but by different regions less than 23 bp in length ("region A," "region B," "region C," and "region E") in various combinations. Similar short DNA sequences are codistributed nonrandomly throughout the plastome, and there are some similarities with DNA sequence duplication in the *rps*7 gene and its adjacent spacer (Koch et al. 2005). The detailed structure of the *trn*L-F spacer region with its pseudogenes has been described previously concentrating on the examples *Arabidopsis* and *Boechera* including a large number of species from both genera (Koch et al. 2005; Dobeš C, Kiefer C, Kiefer M, and Koch MA, unpublished data).

Herein, we aim to analyze the evolutionary dynamics of structural mutations in the *trn*LF intergenic spacer on a family-wide level by asking the following questions. Firstly, is there a monophyletic origin of the pseudogenes in the Brassicaceae? Secondly, is the pseudogene copy number a useful character for phylogenetic inference? Lastly, if the purpose is to reconstruct the *trn*F pseudogenization and duplication events based on a current phylogenetic hypothesis for the Brassicaceae, how do the distributions of the observed structural mutations fit into actual phylogenetic reconstructions of the entire family using these different molecular markers? To do this, we have constructed the first family-wide supernetwork (Huson et al. 2004) of the Brassicaceae.

Materials and Methods

Vouchers of the plant material used in this study for DNA sequencing have been deposited at Heidelberg herbarium (HEID) (Table S1, Supplementary Material online). Otherwise, it has been documented previously elsewhere (refer to the studies cited in Table S2, Supplementary Material online).

Molecular Markers and Taxon Sampling

Understanding the evolution of *trn*F pseudogenes relies on having a well-resolved and statistically supported phylogenetic hypothesis for Brassicaceae. In order to reconstruct a robust phylogenetic hypothesis for an enlarged and representative set of crucifer plants, we chose a multigene approach. Phylogenetic relationships among different cruciferous plants have been proposed based on the analyses of the multilocus ITS gene region (ITS1 and ITS2, including the 5.8 S rRNA gene; Bailey et al. 2006), the single-copy nuclear genes, chalcone synthase (*chs*, Koch et al. 2001) and alcohol dehydrogenase (*adh*, Koch et al. 2000), the plastidic single-copy gene maturase K (*mat*K, Koch et al. 2001), and also the noncoding plastidic *trn*L intron–*trn*LF intergenic spacer (Lysak et al. 2005). These studies share numerous taxa. However, although taxon sampling is overlapping, it is not identical. This makes it difficult to perform a multigene analysis because of the high numbers of missing characters (Koch 2003).

Recently, a new method has been introduced (Huson et al. 2004) that will allow reconstruction of phylogenetic relationships based on analysis of gene trees with overlapping but not necessarily identical taxon sets. The end result is a "supernetwork" that is akin to a supertree; however, unlike a supertree, a supernetwork does not require the restrictive condition that the same bifurcating tree explains evolution at independent gene loci. To provide a phylogenetic framework to study *trn*F pseudogene evolution, we have constructed a supernetwork using gene trees built from 5 data sets (*mat*K, *chs*, *adh*, *trn*LF, and ITS). In order to do this successfully, given that poor degree of taxon overlap can reduce the effectiveness of supernetwork and supertree methods (McBreen and Lockhart 2006), we chose a sampling strategy whereby we included ITS sequences for all accessions for which we had a sequence for *mat*K, *chs*, *adh*, and *trn*LF. To make the ITS reconstruction more robust, we included sister taxa for a number of these species. We also included ITS sequences for taxa from a recent analysis of genome size evolution among cruciferous plant (Johnston et al. 2005). In total, we used 137 ITS sequences to reconstruct the SuperNetwork (see Supplementary Tables S1 and S2 online). The ITS alignment was initially created using ClustalX. However, manual corrections were necessary (see Supplementary Table S4 online). Additional sequences that were of potential interest in respect to *trn*F pseudogene microstructural mutations were also determined and included in the *trn*LF gene tree (Koch et al. 2005).

DNA Extraction, Polymerase Chain Reaction Conditions, and DNA Sequencing

Total DNA was obtained from 50 to 75 mg dried leaf tissue from single individuals. Extraction followed the procedure of Doyle JJ and Doyle JL (1987) (CTAB method), but some modifications were applied, including grinding of dry leaf tissue in 2-ml tubes using a Retsch swing mill (type MM 200), the addition of 2 units of ribonuclease per extraction to the isolation buffer, and washing of the DNA pellet twice with 70% ethanol. DNA was dissolved in 50 µl Tris–ethylenediaminetetraacetic acid buffer for long-term storage. Before use, DNA was diluted 1:3 in Tris EDTA buffer. From this template DNA, the plastidic *trn*L intron, the *trn*LF intergenic spacer, and the nuclear ribosomal DNA were

amplified. Twenty-five microliters of polymerase chain reactions (PCRs) were performed in a master mix containing 1× PCR buffer (10 mM Tris/50 mM KCl buffer, pH 8.0), 3 mM MgCl, 0.4 μM of each primer, 0.2 mM of each deoxynucleoside triphosphate, 0.5 μl *Taq* DNA polymerase (Schott-Eppendorf), and approximately 1 ng of template DNA using an ABI 9700 (ABI Applied Biosystems, Inc., Lincoln, NE) thermal cycler. Thermal cycling started with a denaturation step at 95 °C for 5 min, followed by 35 cycles each comprising 60 s denaturation at 95 °C, 45 s annealing at 38 °C (*trn*LF intron), 45 °C (*trn*LF), 48 °C (ITS), and 1-min elongation at 72 °C. Amplification ended with an elongation phase at 72 °C lasting 10 min and a final hold at 4 °C. PCR products were checked for length and concentrations on 1.5% agarose gels.

The *trn*L intron and the *trn*LF were amplified using the primer combinations given in Dobeš et al. (2004). Sequences comprised the complete intron and the second exon of the *trn*L gene as well as the complete *trn*LF spacer and the first 18 bases of the *trn*F gene. No purification of PCR products was necessary for subsequent sequence reactions. The primers used to amplify the ITS are those described in Koch et al. (2003). PCR products spanned the entire ITS1, 5.8 S rDNA, and ITS2 region. Before sequencing, they were purified using the Boehringer PCR product purification kit (Roche Molecular Diagnostics, Mannheim, Germany).

Cycle sequencing was performed using the TaqDye-Deoxy Terminator Cycle Sequencing Kit (ABI Applied Biosystems, Inc.) and the original amplification primers. However, the reverse *trn*LF IGS primer was modified by adding an additional cytosin to its 3′ end. Products were analyzed on an ABI 377XL automated sequencer. Cycle sequencing was performed on both strands; in the majority of cases, each reaction spanned the complete sequence.

An alignment of the chloroplast *trn*LF region for newly determined sequences has been created manually by adding these sequences to a given published alignment (Lysak et al. 2005) and is provided as supplementary material online (Table S5). The pseudogenic region (copy numbers vary between 2 and 12) has not been aligned and was excluded in the *trn*LF gene tree that was used. The structural mutations in the psuedogene region have been plotted on a *trn*LF gene tree and also the supernetwork containing this gene tree.

Phylogenetic Analyses and Source Trees

Seventy-one operational units in total were represented among 4 of the gene trees (*adh*, *chs*, *mat*K, and *trn*LF; Koch et al. 2000, 2001; Lysak et al. 2005) and were used for supernetwork construction. These 71 sequences belong to 68 taxa. *Aubrieta deltoides*, *Olimarabidopsis pumila*, and *Arabidopsis lyrata* ssp. *petraea* were represented by multiple copies/alleles for *chs* (cf., Koch et al. 2001). For each of these data sets, the source (input) tree used for supernetwork construction was either an optimal tree or strict consensus of optimal tree. Thus for the *chs*, the source tree was the strict consensus of 5 most parsimonious trees (fig. 2 from Koch et al. 2001). For *mat*K, a single most parsimonious tree was used (fig. 1 from Koch et al. 2001). For *adh*, the source tree was the strict consensus of 4 most parsimonious trees (fig. 6 from Koch et al. 2000). In this case, we excluded all *adh* alleles characterized by missing introns to avoid any problems concerned with orthologous or paralogous genes, resulting in the exclusion of *Brassica oleracea*, and all *Adh2* and *Adh3* loci of the various *Arabis* species. For *trn*LF, an optimal maximum-likelihood tree was used (fig. 3 from Lysak et al. 2005) (see Supplementary Table S3 online). In the case of the ITS marker, a very conservative estimate for phylogenetic relationships was inferred. This was achieved by first excluding ambiguous parts of the alignment (see Supplementary Table S4 online: character nos 144–162, 494–551, and 710–735; total alignment of 735 bp) and identifying optimal heuristic maximum parsimony trees (made using PAUP* 4.0b10, Swofford 2000) assuming the Tree Bisection-Reconnection option, equally weighted characters, and gaps treated as missing data with 500 random additions of the sampled taxa. A strict consensus tree was generated as a subsequent source tree.

The 5 trees were used to calculate a SuperNetwork using the Z-closure option in Splitstree version 4beta26 with the following assumption: splitstransfom = EqualAngle; SplisPostProcess filter = dimension value = 4) (Huson and Bryant 2006). In order to minimize misleading phylogenetic implications when showing a SuperNetwork, we used the strict consensus or optimal maximum-likelihood trees as most conservative input information. A detailed discussion on possible sources of misleading results in network reconstructions is given by McBreen and Lockhart (2006). The resulting SuperNetwork combines weighted splits from the single trees. Branch lengths are weighted using information from partial splits in the source trees. However, because bracket notations have been provided without branch length, each tree contributed equally to each branch length.

In addition, we analyzed all *trn*LF sequences which were of potential interest in respect of *trn*F pseudogene microstructural mutations separately. The corresponding alignment is also available online (see Supplementary Table S5 online). In this alignment, we did not align the pseudogenic region (copy numbers vary between 2 and 12). Consequently, this region has been excluded in a maximum parsimony analysis (using PAUP* 4.0b10, Swofford 2000) to characterize the different haplotype lineages carrying *trn*F pseudogenes.

In this analysis, we excluded a region within the *trn*L intron corresponding to a large insertion (see Supplementary Table S5 online). *Aethionema* has been used as an outgroup because the basal position of this genus has been confirmed in several previous studies on crucifer evolution (Koch et al. 2000, 2001; Al-Shehbaz et al. 2006; Beilstein et al. 2006). We have chosen the same settings in PAUP* as described above for the ITS region. In a subsequent analysis, we selected the first single *trn*F pseudogene (copy I, for details refer to Koch et al. 2005) from several taxa plus sequences of the original *trn*F gene and combined them in a maximum parsimony analysis as described above. These selected regions and the corresponding alignment are shown in Table S6 (Supplementary Material online).

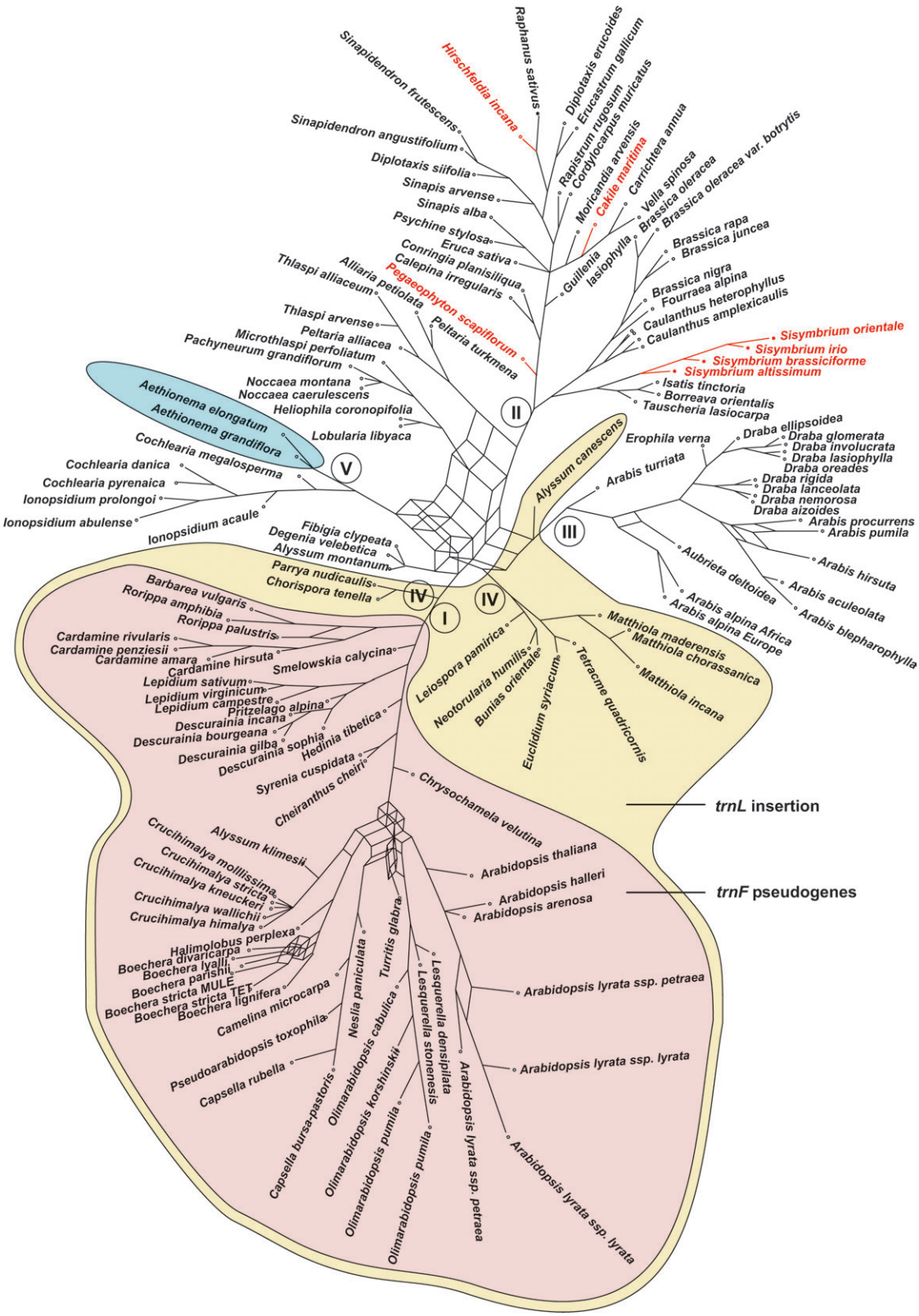

FIG. 1.—Phylogenetic hypothesis of relationships among cruciferous plants on a family level based on a "SuperNetwork" reconstruction using *Adh*, *Chs*, *mat*K, and ITS sequences. The outgroup (*Aethionema*, blue), *trn*F pseudogene (red text and red highlighted taxa), and *trn*L insertion (highlighted with yellow) carrying taxa are indicated. CpDNA lineages I–V correspond to figure 2. For definition of the insertion refer to Materials and Methods.

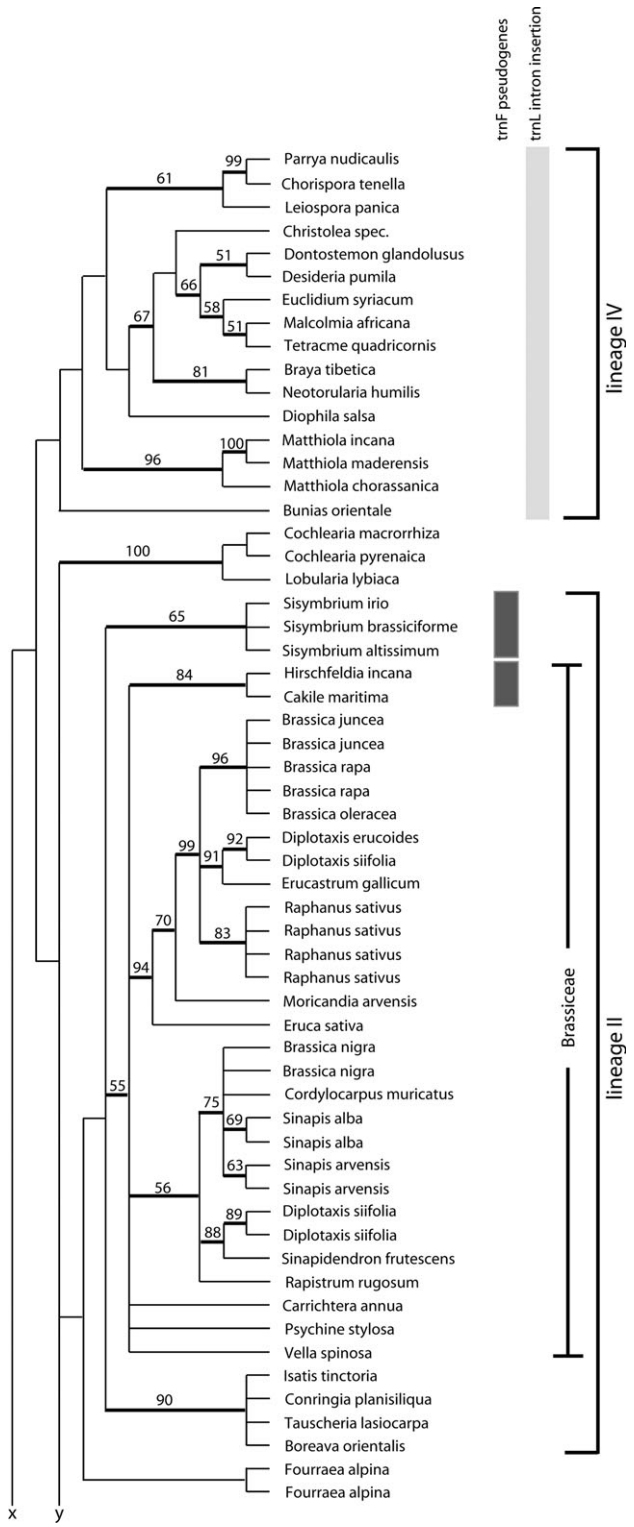

FIG. 2.—Strict consensus tree of a maximum parsimony analysis based on *trn*L intron and *trn*LF intergenic spacer sequences (excluding a larger insertion and pseudogenes). Bootstrap values greater than 50% are given above branches. Vertical dark gray (online version: red) and light gray (online version: yellow) bars indicate existence of *trn*F pseudogenes and a large insertion in the *trn*L intron, respectively. Lineage definition follows Beilstein et al. (2006).

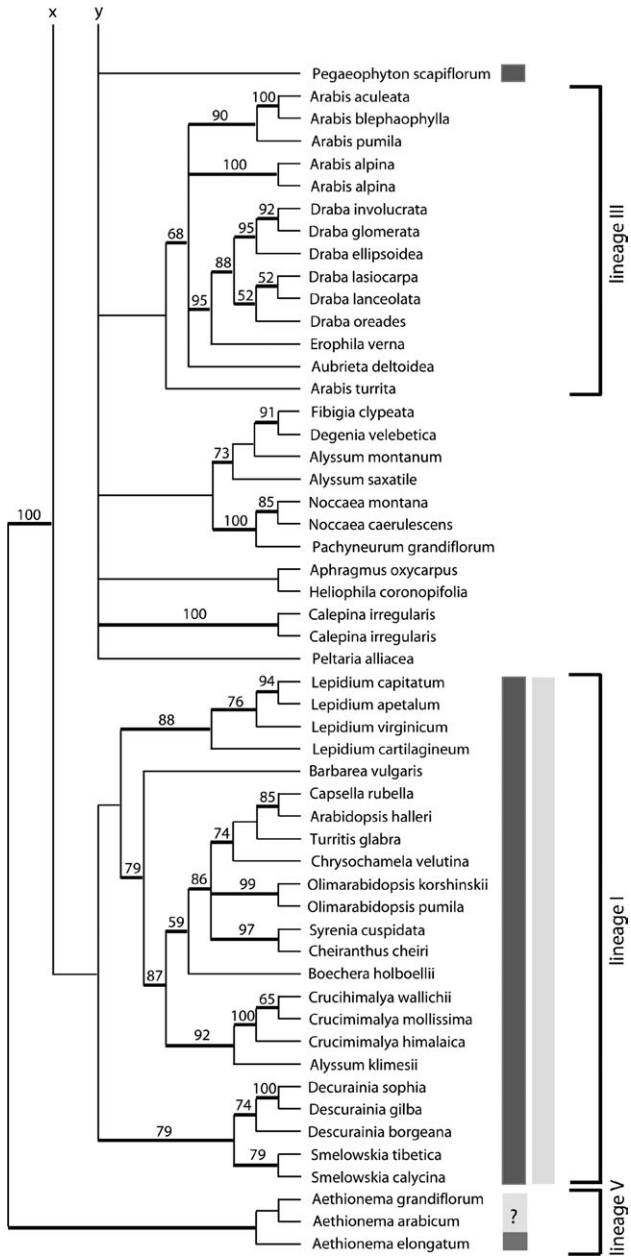

FIG. 2.—Continued.

Results

Phylogenetic Reconstructions—Species Tree

The resulting alignment of the ITS region, including the 5.8 S rDNA gene, was 774-bp long. This alignment excluded autapomorphic insertion in *Bunias orientale* (97 bp), *Chorispora tenella* (102 bp), and *Parrya nudicaulis* (94 bp) (see Supplementary Table S4 online). Three regions within the ITS1 and ITS2 could not be aligned unambiguously, and consequently, 136 characters were excluded from the subsequent analysis. For the remaining 638 characters, 246 were constant and uninformative, 86 were variable but not parsimony informative, and 306 were potentially parsimony informative. In total, we obtained 1,536 most parsimonious trees with a tree length of 2,901 steps (consistency

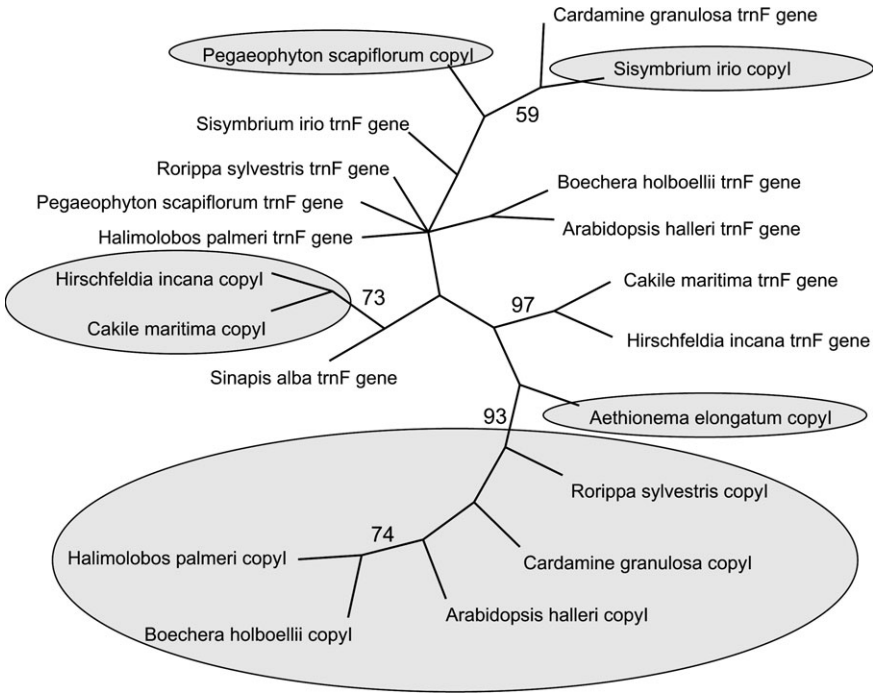

FIG. 3.—Fifty-percent majority-rule consensus tree out of 10,000 most parsimonious trees reconstructed from the alignment of *trn*F genes and corresponding pseudogenic copies shown in Table S6 (Supplementary Material online). Taxa carrying pseudogenes are indicated. The phylogeny does not support a single origin of the pseuodgenes present in the various genera, which are represented by accessions sampled from largely diverged lineages (refer to fig. 2). Bootstrap values higher than 50 calculated from 1,000 replicates are given along internodes.

index [CI] = 0.25, retention index [RI] = 0.64). The strict consensus tree is available as bracket notation (see Supplementary Table S3 online).

The network resulting from the combined analysis of all 5 gene trees (*chs*, *adh*, *mat*K, *trn*LF, and ITS) as explained above is shown in figure 1. This phylogenetic network is consistent with an analysis covering the whole family including all tribes and more than 1,500 accessions (e.g., Bailey et al. 2006) or more restricted studies (Koch 2003; Al-Shehbaz et al. 2006; Beilstein et al. 2006). However, some minor incongruencies occurred, but these can be explained by individual differences in tree resolution or gene coalescence among the various molecular markers.

The ambiguous phylogenetic positions of an accession of *Alyssum canescens* from India and *Donstostemon perennis* from Russia, respectively, require separate consideration. Both showed a weak relationship to a nonpseudogene carrying taxon, *Arabis turrita*, in the ITS- and cpDNA-based analysis (results not shown). However, another accession of *A. canescens* from China (Beilstein et al. 2006) is more closely related to a major monophyletic lineage unexceptionally carrying *trn*F pseudogenes (see fig. 1). This clearly defined lineage corresponds to clade I from Beilstein et al. (2006). We assume that different taxa have been analyzed, but we were not yet able to solve the underlying taxonomic problems.

Chloroplast Evolution and Occurrence of Pseudogenes

The alignment of the *trn*LF region included 111 sequences from 99 taxa (Supplementary Table S5 online). The following alignment positions were excluded from the phylogenetic analysis: positions 1–35 (annealing site of the *trn*L forward primer), 243–469 (insertion in the *trn*L intron region), 601–811 (annealing sites of sequencing primers and large indels), and 1247–1954 (pseudogenic region). Consequently, the number of characters in the alignment used for phylogenetic reconstruction was reduced to 773. Out of these 773 characters, 427 were constant and uninformative, 132 were variable but not parsimony informative, and 214 were potentially parsimony informative. We restricted the heuristic search to the 10,000 most parsimonious trees with a tree length of 842 steps (CI = 0.58, RI = 0.77). The strict consensus tree is shown in figure 2 with the occurrence of pseudogenes and *trn*L intron insertions/deletions indicated.

A short summary of sequence length variation from cruciferous plants (this study and previously published data) compared with published data from green plants (Quandt et al. 2004) is given in table 1, indicating the enormous variation found within the Brassicaceae. We also found in all sequenced *trn*LF intergenic spacer regions a common motif, subsequently called region A, upstream of the first pseudogene. This motif was also present in all haplotypes missing *trn*F pseudogenes (cf., Koch et al. 2005, Table S5, Supplementary Material online). Close to (−35 TTGACA element) or even within (−10 GAGGAT element) region A lie the conserved ancient promoter elements described in the Introduction.

The chloroplast phylogeny is in congruence with a recent *ndh*F-based analysis provided by Beilstein et al. (2006). The high consistency in the 2 chloroplast-derived data sets demonstrates the general utility of the *trn*LF

Table 1
DNA Sequence Length Variation (bp) in the *trn*L Intron and *trn*LF Intergenic Spacer among Cruciferous Plants in Comparison to Estimates from Land Plants (Ψ = pseudogene)

	*trn*L Intron		*trn*LF Spacer	
	Without Insertion (mean, SD)	Including Insertion (mean, SD)	Without Ψ (mean, SD)	No. of Ψ[a]
Cruciferous plants[b]	397–413 (404, 5)	397–609 (450, 92)	217–461 (382, 52)	0 to >12
Land plants[c]	218–660 (443, 189)	—	51–466 (288, 209)	—

[a] Each pseudogene contributes approximately an additional 50–60 bp to the total length of the *trn*L-F intergenic spacer; for total pseudogene copy number variation refer to Table S2 (Supplementary Material online).
[b] From data presented in this study.
[c] As summarized in Quandt et al. (2004), no pseudogenes present in these taxa.

region for phylogenetic reconstruction on the family level and furthermore indicates that it did evolve "in concert" with other plastidic markers (see also data for *mat*K, Koch et al. 2001). Thus, 5 phylogenetic "lineages" as defined by Beilstein et al. (2006) were also found in our study to constitute monophyletic groups and are indicated in figure 2. This finding is important for it demonstrates that *trn*F pseudogenes are present in isolated clades. Thus, pseudogenes occur in all taxa of lineage I, in 2 different branches of lineage II (*Sisymbrium* which is outside tribe Brassiceae as shown by Lysak et al. (2005), and *Cakile*/*Hirschfeldia* from tribe Brassiceae) and in one species, *Pegaeophyton scapiflorum*, not related significantly to any of the major lineages (fig. 2). The second large structural mutation, the insertion/deletion of a stretch of DNA within the *trn*L intron at alignment positions 243–469, was observed in all sequences combined in lineage I and lineage IV but was missing in the outgroup. Consequently, a monophyletic origin of this conspicuous length mutation (indicated in figs. 1 and 2) remains unclear. However, the multigene network (fig. 1) supports a monophyletic origin for an insertion, whereas a comparison with *trn*L intron data from all over the angiosperms (Borsch et al. 2003) favor a multiple and independent deletion in various lineages of the angiosperms. This might lead to a conclusion that parallel loss in several lineages of the Brassicaceae including *Aethionema* explains the observed results. In some *trn*L intron sequences (e.g., *Lepidium* sp. and *Crucihimalya* sp.), we also detected some short inversions (TAGAC). Similar sequences have been introduced previously, and it has been outlined that such sequences need to be reverse complemented for analysis (Löhne and Borsch 2005).

From these data, we can summarize the following. 1) One major insertion (or several deletions) event in the *trn*L intron occurred. 2) *trn*F pseudogenes evolved at least 4 times independently. This estimate assumes that *Cakile* and *Hirschfeldia* are carrying a monophyletic chloroplast type. The plastome types of *Cakile maritima* and *Hirschfeldia incana* are highly similar to each other (cf. fig. 2), a finding that is also congruent with the data provided by Beilstein (2006). In contrast, nuclear DNA sequences (Warwick and Sauder 2005, ITS data this study) provide evidence for a distant relationship only between these 2 species. This finding might be best explained by extensive, ancient genome duplications and reticulation involving chloroplast capture and lineage sorting (Lysak et al. 2005; Warwick and Sauder 2005). 3) In *Sisymbrium trn*F, pseudogene evolution is as highly dynamic as in the taxa of lineage I, and we found pseudogene copy numbers to vary between 2 (*Sisymbrium strictissimum*) and 7 (*Sisymbrium loeselii*).

It is furthermore noteworthy that we obtained an additional sequence in one *Aethionema* accession (*Aethionema elongatum*: DQ180216 + additional 5′ end sequence information), which is similar in position and structure to the *trn*F pseudogenes described above. However, these duplicated *trn*F region DNA motifs differ in their sequence from the various other *trn*F pseudogene copies (Table S6, Supplementary Material online). Unfortunately, we were not successful in obtaining sequences from the original *trn*F gene or of the pseudogenic region from the other *Aethionema* accessions. Future research might elucidate the situation in *Aethionema*, but it is very likely that this early branching lineage in cruciferous plants is also carrying *trn*F psuedogenes in the *trn*LF intergenic spacer.

Comparisons among Various Pseudogene Copies

The result of a parsimony analysis using only the first pseudogene copy from selected taxa of different phylogenetic position and the corresponding *trn*F gene (if available) is shown in figure 3. Among the 119 scored characters from the alignment, 77 were constant, 18 were variable but parsimony uninformative, and 24 characters were parsimony informative (CI = 0.84, RI = 0.84). Because of the limited number of informative characters (Table S6, Supplementary Material online), the *trn*F pseudogene tree is not statistically well supported. However, the inference of multiple origins of *trn*F pseudogenes suggested by this tree is consistent with findings from analyses of the *trn*L-F region that exclude the *trn*F region (fig. 2).

Discussion

Pseudogene Evolution

Although the occurrence of *trn*F pseudogenes in flowering plants is an extremely rare event, in the Brassicaceae, many genera show extensive variation in pseudogene copy number (see Table S2, Supplementary Material online). The *trn*F pseudogenes have arisen independently at least 4 times. An earlier study, focussing on *Arabidopsis* relatives, investigated the evolutionary dynamics of the tandemly arranged copies (Koch et al. 2005) and demonstrated that it was possible to reconstruct common ancestry for some pseudogene copies. This study proposed a

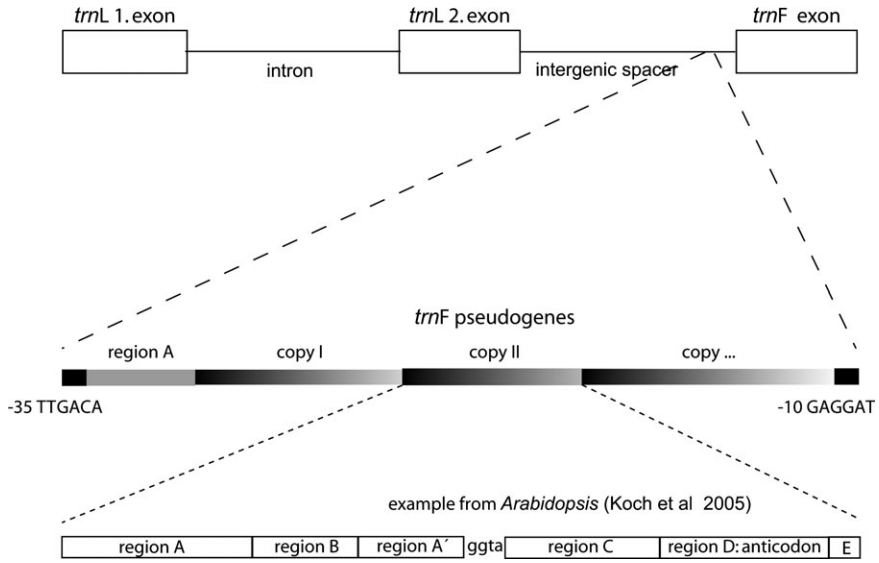

FIG. 4.—Schematic cartoon of the plastidic *trn*L–*trn*F region in cruciferous plants.

monophyletic origin for pseudogene *trn*F copy I in a large clade of cruciferous taxa (clade I; fig. 2). It also found that the mutation rate for the pseudogene region exceeded the mutation rate of the adjacent noncoding spacer and intron regions by a factor of 20 (Koch et al. 2005). A more recent analysis of *trn*F pseudogene evolution in 719 accessions from the genus *Boechera* has identified 103 haplotypes (original data from Dobeš et al. 2004) and suggests that intermolecular unequal crossing-over may explain generation of the tandem repeats (Dobeš C, Kiefer C, Kiefer M, and Koch MA, unpublished data).

Repeat Evolution

It has been concluded previously that the $trnF_{GAA}$ gene is cotranscribed with $trnL_{UAA}$ (Kanno and Hirai 1993), and thus, one might assume that there are constraints on length variation in the *trn*L-F intergenic spacer. Among land plants, this region normally does not exceed 470 bp (table 1). However, generation of a first pseudogene copy frequently is associated with generation of subsequent pseudogene copies, often in high number. This same pattern is observed repeatedly in distantly related Brassicaceae lineages. For example, even within one genus, *Leavenworthia*, up to 14 copies have been identified (Beck et al. 2006) or in genus *Sisymbrium* we detected up to 7 copies, whereas *Arabidopsis* species are carrying up to 8 pseudogene copies (Koch et al. 2005) and for *Cardamine* Lihova et al. (2004) provided sequence data showing up to 6 copies. One repeat, depending on its modularized structure (fig. 4), is approximately 50 bp in size. Consequently, the above-described additional copies can contribute up to 700 additional base pairs to the *trn*L-F intergenic spacer resulting in extensive length variation of this region. Interestingly, we do not find significantly higher copy numbers in old lineages compared with young species groups. Frequent parallel loss of pseudogene copies in *Boechera* have been demonstrated (Dobeš et al. 2004) resulting in a balanced system with a maximum of 3 pseudogenes in this particular genus. Among several *Arabidopsis* species, we scored 179 haplotypes from 1,090 accessions with a mean of 5.2 (SD 1.4) pseudogenes (Koch M and Matschinger M, unpublished data). The summary on reported length variants provided in Table S2 (Supplementary Material online) also indicate that normally less than 5–6 pseudogene copies are found. These findings can be best explained by evolutionary constraints leading to a balanced "maximum DNA load" in the *trn*L-F intergenic spacer. Interestingly, there might be 2 different "optimum" pseudogene copy numbers. The frequency distribution of pseudogene copy numbers in the genera *Arabidopsis* (Koch et al. 2005) and *Boechera* (data from Dobeš et al. 2004) supports a preferred maximum of 2–3 or 5–6 pseudogenic copies (fig. 5).

Future research will show if it is possible to develop evolutionary models for describing pseudogene evolution as done for hypervariable microsatellites (Watterson and Guess 1977; Di Rienzo et al. 1998; Xu et al. 2000).

Phylogenetic Implications

The Brassicaceae is a large plant family (338 genera and 3,700 species) of major scientific and economic importance. Almost 100 years after the first taxonomic and systematic treatise on the family of the Brassicaceae (Hayek 1911) and subsequent contributions (Schulz 1936; Janchen 1942), we are now close to the first comprehensive and natural system regarding the mustard family. Within this family, *Arabidopsis* and *Brassica* model organisms have increasing importance. Their study is greatly advancing systematics and taxonomy, as well as evolutionary and developmental research. Their increasing significance is due to the fact that molecular tools developed for model plants are increasingly being applied successfully in the study of wild relatives. Consequently, the extent to which specific findings on model plants can be generalized is becoming clearer.

A first attempt to summarize the knowledge of the family was provided 30 years ago (Vaughan et al. 1976). Of course, this book described those markers and methods

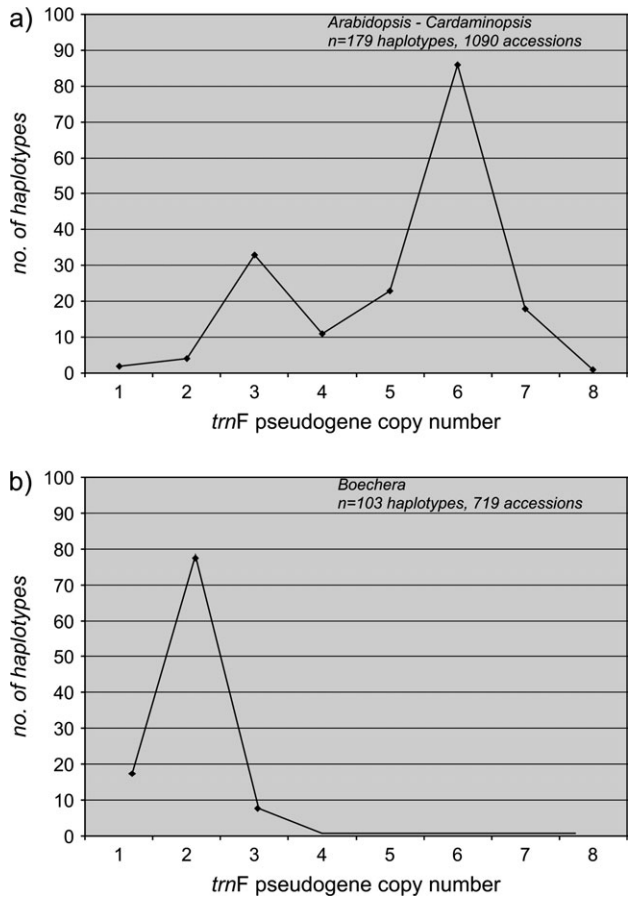

FIG. 5.—Pseudogene copy number distribution in *Arabidopsis* (*a*) and *Boechera* (*b*).

associated with the study of evolution in the Brassicaceae that were the most modern and informative at that particular time. During the last 20 years, molecular biology and DNA techniques opened new avenues for a revolution in plant systematics and evolution, and for the reasons already given, the Brassicaceae have been at the spearhead of scientific research. Since the publication of "The Biology and Chemistry of the Cruciferae" by Vaughan et al. (1976), evolutionary research on the Brassicaceae increased rapidly and this has led to several recent and important contributions: 1) introduction of a new infrafamiliar classification system (Beilstein et al. 2006), 2) a first attempt to provide a comprehensive family-wide phylogeny within the framework of a detailed tribal classification of the family (Bailey et al. 2006), and 3) a summary of aspects of research in cruciferous plants (Koch and Mummenhoff 2006). Despite this wealth of information about taxon relationships, resolution of deeper phylogenetic relationships within the Brassicaceae has remained problematic. The most attractive hypothesis to explain the lack of resolution for intertribal relationships is rapid radiations 15–30 MYA (Bailey et al. 2006; Beilstein et al. 2006). This is a situation where microstructural evolutionary changes may be useful for inferring early events of divergence. In respect of this, it is interesting that the 2 structural rearrangements that we describe for the *trn*LF region identify ancient patterns of divergence supported by phylogenetic analysis of the *trn*LF region that excludes the microstructural mutations (fig. 2). Support is also found from analyses of the nuclear ITS sequence data, but as evident from the combined Supernetwork (fig. 1), not in analyses of some of the other genes. Earlier events of hybridization and/or lineage sorting (or perhaps even phylogenetic error due to different extents of taxon sampling with different data sets) might explain this discrepancy.

We are optimistic that further phylogenetic studies based on microstructural characters will contribute to a better understanding of evolution in the crucifer family. Of particular interest to us are the microevolutionary changes and pseudogene dynamics in species complexes not older than 5 Myr (e.g., *A. lyrata*, *Arabidopsis arenosa*, *Arabidopsis halleri*). If sufficiently fast evolving, they may help us to resolve phylogenetic relationships for very closely related taxa. At present, the low mutation rate of the *trn*LF region based on single-nucleotide polymorphisms (e.g., 7.7×10^{-9}, Mummenhoff et al. 2004) limits the inferences that can be drawn.

Supplementary Material

Tables S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors wish to thank Peter Lockhart and 3 anonymous reviewers for many comments improving the manuscript substantially. This work was supported by grants from the Austrian Science Foundation FWF (GEN-15609 and GEN-14463) and the German Science Foundation DFG (Ko-2302/4-1 and Ko-2302/5-1) to M.A.K. Financial support was also gratefully acknowledged by a research grant from the Grant Agency of the Czech Republic (KJB601630606) to M.A.L.

Literature Cited

Al-Shehbaz IA, Beilstein MA, Kellogg EA. 2006. Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview. Plant Syst Evol. 259:89–120.

Bailey CD, Koch MA, Mayer M, Mummenhoff K, O'Kane SL, Warwick SI, Windham MD, Al-Shehbaz IA. 2006. Towards a global phylogeny of the Brassicaceae. Mol Biol Evol. 23:2142–2160.

Bailey CD, Price RA, Doyle JJ. 2002. Systematics of the halimolobine Brassicaceae: evidence from three loci and morphology. Syst Bot. 27:318–332.

Beck JB, Al-Shehbaz IA, Schaal BA. 2006. *Leavenworthia* (Brassicaceae) revisited: testing classic systematic and mating system hypothesis. Syst Bot. 31:151–159.

Beilstein MA, Al-Shehbaz IA, Kellogg EA. 2006. Brassicaceae phylogeny and trichome evolution. Am J Bot. 93:607–619.

Bishop J, Dean AM, Mitchell-Olds T. 2000. Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. Proc Natl Acad Sci USA. 97:5322–5327.

Bleeker W, Franzke A, Pollmann K, Brown AHD, Hurka H. 2002. Phylogeny and biogeography of southern hemisphere high-mountain *Cardamine* species (Brassicaceae). Aust Syst Bot. 15:575–581.

Bleeker W, Hurka H. 2001. Introgressive hybridization in *Rorippa* (Brassicaceae): gene flow and its consequences in natural and anthropogenic habitats. Mol Ecol. 10:2013–2022.

Bleeker W, Weber-Sparenberg C, Hurka H. 2002. Chloroplast DNA variation and biogeography in the genus *Rorippa* Scop. (Brassicaceae). Plant Biol. 4:104–111.

Borsch T, Hilu KW, Quandt D, Wilde V, Neinhuis C, Barthlott W. 2003. Noncoding plastidic *trn*T-*trn*F sequences reveal a well resolved phylogeny of basal angiosperms. J Evol Biol. 16: 558–576.

Charlesworth D, Liu FL, Zhang L. 1998. The evolution of alcohol dehydrogenase gene family by loss of introns in plants of the genus *Leavenworthia* (Brassicaceae). Mol Biol Evol. 15:552–559.

Clauss M, Koch M. 2006. *Arabidopsis* and its poorly known relatives. Trends Plant Sci. 11:449–459.

Cork JM, Purugganan MD. 2005. High-diversity genes in *Arabidopsis*. Genetics. 170:1897–1911.

DePamphilis CW, Palmer JD. 1990. Loss of photosynthetic and chlororespiratory genes from the plastid genome of a parasitic flowering plant. Nature. 348:337–338.

Di Rienzo A, Donnelly P, Toomajian C, Bronwyn S, Hill A, Petzl-Erler ML, Haines GK, Barch DH. 1998. Heterogeneity of microsatellites mutations within and between loci, and implications for human demographic histories. Genetics. 148:1269–1284.

Dobeš C, Mitchell-Olds T, Koch M. 2004. Phylogeographic analysis of extensively sympatric and highly diverse chloroplast haplotypes (*trn*L intron-*trn*F IGS) in North American *Arabis drummondii*, *A.* ×*divaricarpa*, and *A. holboellii* (Brassicaceae). Mol Ecol. 13:349–370.

Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small amounts of fresh leaf tissue. Phytochem Bull. 19:11–15.

Drábkova L, Kirschner J, Vlček Č, Paček V. 2004. *Trn*L-*trn*F intergenic spacer and *trn*L intron define major clades within *Luzula* and *Juncus* (Juncaceae): importance of structural mutations. J Mol Evol. 59:1–10.

Fiebig A, Kimport R, Preuss D. 2004. Comparisons of pollen coat genes across Brassicaceae species reveal rapid evolution by repeat expansion and diversification. Proc Natl Acad Sci USA. 101:3286–3291.

Hall JC, Sytsma KJ, Iltis HH. 2002. Phylogeny of Capparaceae and Brassicaceae based on chloroplast sequence data. Am J Bot. 89:1826–1842.

Hayek A. 1911. Entwurf eines Cruciferensystems auf phylogenetischer Grundlage. Beih Bot Centralbl. 27:127–335.

Heenan PB, Mitchell AD, Koch M. 2002. Molecular systematics of the New Zealand *Pachycladon* (Brassicaceae) complex: generic circumscription and relationship to *Arabidopsis* sens. lat. and *Arabis* sens. lat. N Z J Bot. 40:543–562.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.

Huson DH, Dezulian T, Kloepper T, Steel M. 2004. Phylogenetic super-networks from partial trees. IEEE/ACM Trans Comput Biol Bioinformatics. 1:151–158.

Janchen E. 1942. Das System der Cruciferen. Österr Bot Z. 91: 1–28.

Johnston JS, Pepper AE, Hall AE, Chen ZJ, Hodnett G, Drabek J, Lopez R, Price HJ. 2005. Evolution of genome size in Brassicaceae. Ann Bot. 95:229–235.

Kanno A, Hirai A. 1993. A transcription map of the chloroplast genome from rice (*Oryza sativa*). Curr Genet. 23:166–174.

Koch M. 2003. Molecular phylogenetics, evolution and population biology in Brassicaceae. In: Sharma AK, Sharma A, editors. Plant genome: biodiversity and evolution, Vol. 1: phanerograms. Enfield (NH): Science Publishers Inc. p. 1–35.

Koch M, Al-Shehbaz IA. 2002. Molecular data indicate complex intra- and intercontinental differentiation of American *Draba* (Brassicaceae). Ann Missouri Bot Gard. 89:88–109.

Koch M, Al-Shehbaz IA. 2004. Taxonomic and phylogenetic evaluation of the American "*Thlaspi*" species: identity and relationship to the Eurasian genus *Noccaea* (Brassicaceae). Syst Bot. 29:375–384.

Koch M, Dobeš C, Matschinger M, Bleeker W, Vogel J, Kiefer M, Mitchell-Olds T. 2005. Evolution of the plastidic *trn*F(GAA) gene in *Arabidopsis* relatives and the Brassicaceae family: monophyletic origin and subsequent diversification of a plastidic pseudogene. Mol Biol Evol. 22:1032–1043.

Koch M, Haubold B, Mitchell-Olds T. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). Mol Biol Evol. 17:1483–1498.

Koch M, Dobeš C, Mitchell-Olds T. 2003. Multiple hybrid formation in natural populations: concerted evolution of the internal transcribed spacer of nuclear ribosomal DNA (ITS) in North American *Arabis divaricarpa* (Brassicaceae). Mol Biol Evol. 20:338–350.

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

Koch M, Kiefer C. 2006. Molecules and migration: biogeographical studies in cruciferous plants. Plant Syst Evol. 259: 121–142.

Koch M, Mummenhoff K. 2006. Editorial: evolution and phylogeny of the Brassicaceae. Plant Syst Evol. 259:81–83.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH. 2005. Use of DNA barcodes to identify flowering plants. Proc Natl Acad Sci USA. 102:8369–8374.

Lanner C. 1998. Relationships of wild *Brassica* species with chromosome number 2n=18, based on comparison of the DNA sequence of the chloroplast intergenic region between *trnL*(UAA) and *trnF*(GAA). Can J Bot. 76:228–237.

Lee J-Y, Mummenhoff K, Bowman JL. 2002. Allopolyploidization and evolution of species with reduced floral structures in *Lepidium* L. (Brassicaceae). Proc Natl Acad Sci USA. 99: 16835–16840.

Lihova J, Fuertes-Aguilar J, Marhold K, Nieto-Feliner G. 2004. Origin of the disjunct tetraploid *Cardamine amporitana* (Brassicaceae) assessed with nuclear and chloroplast DNA sequence data. Am J Bot. 91:1231–1242.

Löhne C, Borsch T. 2005. Molecular evolution and phylogenetic utility of the *petD* group II intron: a case study in basal angiosperms. Mol Biol Evol. 22:317–332.

Lysak MA, Koch MA, Pecinka A, Schubert I. 2005. Chromosome triplication found across the tribe Brassiceae. Genome Res. 15:516–525.

Lysak MA, Lexer C. 2006. Towards the era of comparative evolutionary genomics in Brassicaceae. Plant Syst Evol. 259:175–198.

McBreen K, Lockhart P. 2006. Reconstructing reticulate evolutionary histories of plants. Trends Plant Sci. 11:398–404.

Mummenhoff K, Brüggemann H, Bowman J. 2001. Chloroplast DNA phylogeny and biogeography of the genus *Lepidium* (Brassicaceae). Am J Bot. 88:2051–2063.

Mummenhoff K, Linder P, Friesen N, Bowman JL, Lee JY, Franzke A. 2004. Molecular evidence for bicontinental hybridogenous genome constitution in *Lepidium* sensu stricto (Brassicaceae) species from Australia and New Zealand. Am J Bot. 91:254–261.

Quandt D, Müller K, Stech M, Frahm JP, Hilu KW, Borsch T. 2004. Molecular evolution of the chloroplast *trn*L-F region in land plants. Monogr Syst Bot Missouri Bot Gard. 98:13–37.

Schranz E, Mitchell-Olds T. 2006. Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. Plant Cell. 18:1152–1165.

Schulz OE. 1936. Cruciferae. In: Engler A, Prantl K, editors. Die natürlichen Pflanzenfamilien. Vol. 17B. Leipzig (Germany): Verlag von Wilhelm Engelmann. p. 227–658.
Swofford DL. 2000. PAUP* 4.0b10. Sunderland (MA): Sinauer Associates.
Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL-X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25:4876–4882.
Vaughan JG, Macleod AJ, Jones BMG. 1976. The biology and chemistry of the Cruciferae. London: Academic Press. p. 1–355.
Vijverberg K, Bachmann K. 1999. Molecular evolution of a tandemly repeated *trnF*(GAA) gene in the chloroplast genome of *Microseris* (Asteraceae) and the use of structural mutations in phylogenetic analysis. Mol Biol Evol. 16:1329–1340.
Warwick SI, Al-Shehbaz IA. 2006. Brassicaceae: chromosome number index and database on CD-Rom. Plant Syst Evol. 259:237–248.
Warwick SI, Francis A, Al-Shehbaz IA. 2006. Brassicaceae: species checklist and database on CD-Rom. Plant Syst Evol. 259:249–258.
Warwick SI, Sauder CA. 2005. Phylogeny of tribe Brassiceae (Brassicaceae) based on chloroplast restriction site polymorphisms and nuclear ribosomal internal transcribed spacer and chloroplast *trn*L intron sequences. Can J Bot. 83:467–483.
Watterson GA, Guess HA. 1977. Is the most frequent allele the oldest? Theor Popul Biol. 11:141–160.
Wittzell H. 1999. Chloroplast DNA variation and reticulate evolution in sexual and apomictic sections of dandelions. Mol Ecol. 8:2023–2035.
Xu X, Peng M, Fang Z, Xu X. 2000. The direction of microsatellite mutations is dependent upon allele length. Nat Genet. 24:396–399.
Yang Y-W, Tai P-Y, Chen Y, Li W-H. 2002. A study of the phylogeny of *Brassica rapa*, *B. nigra*, *Raphanus sativus,* and their related genera using noncoding regions of the chloroplast DNA. Mol Phylogenet Evol. 23:268–275.

Peter Lockhart, Associate Editor

Accepted September 18, 2006