The Sequence of the *Arabidopsis thaliana* Genome as a Tool for Comparative Genome Analysis in the Brassicaceae Family

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Karine BOIVIN

aus Auxerre (Frankreich)

Köln

2001

Berichterstatter: Prof. Dr. Tautz D.
Dr. Palme K.

Tag der letzten mündlichen Prüfung: 8 Juni 2001 *à mes parents,*

TABLE OF CONTENTS

1 INTRODUCTION

Hundred thirty Mbp to 110,000 Mbp represent the wide range of nuclear genome sizes, which have been observed among Angiosperms. Between these estimates for *Arabidopsis thaliana* and *Fritillaria assyriaca*, a broad panel of genome sizes and chromosome numbers are found among flowering plants (reviewed in Bennett *et al*., 2000). These variations are mainly due to the amount of repetitive DNA sequences in the genomes. The abundance of repetitive sequences has been found positively correlated with genome size (Bennetzen 2000a; Bennetzen 2000b). Polyploidy appears to be widespread in the plant kingdom, but it does not account for the large variations in genome sizes which are observed in higher plants.

Comparative mapping experiments in the Poaceae (Moore *et al*. 1995) and the Brassicaceae family (Lagercrantz *et al*. 1996) have revealed a conserved gene repertoire and genome organisation in related species, despite their differences in genome sizes. These experiments rely on the use of a common set of markers for the construction of genetic linkage maps for two or more species. These markers used are representing links between the maps of related species. This allows direct comparison of the resulting linkage maps. Thus, it can be studied whether the markers along the linkage groups can be found in a conserved arrangement.

The *Brassicaceae* family lends itself particularly well to comparative genome analysis studies. Small size of mature plants, a short generation time, prolific seed production from single plants and diploid genetics makes *Arabidopsis thaliana* ideally suited for genetic and mutational analyses. For these reasons and due to its small genome size, this crucifer has been chosen as model organism for molecular genetic studies in plants. The genome of *A. thaliana* is the best-studied genome of a higher plant and it is therefore of special interest to transfer the knowledge obtained in the model species to related crop plants. Within the *Brassica* genus, closely related to *Arabidopsis*, several species are exploited as vegetable and fodder crops and most importantly oil seeds. For many *Brassica* species, recent polyploid ancestry has been postulated (Lydiate *et al.* 1993). Thus, the *Brassicaceae* family offers excellent opportunities for studying comparative genome arrangements between diploid species and those of polyploid origin.

1-1 ARABIDOPSIS THALIANA: THE MODEL

1-1-1 Genome size

A. thaliana is a crucifer with n=5 chromosomes. Chromosomes 2 and 4 can be easily recognised because they carry the nucleolar organising regions (NOR) (Heslop-Harrison and Maluszynska 1994). *Arabidopsis* has one of the smallest genomes among higher plants, the 2C DNA content has been determined at ~0,30 pg (Arumuganathan and Earle 1991). Much higher values have been determined for other species such as rice (~0,86-0,91 pg), maize (~4,75-5,63 pg) or wheat (~33,09 pg). For the *Brassica* relatives the nuclear DNA content is comprised between 0,97 (*B. nigra*) and 2,56 pg/2C (*B. napus*), the resulting genome size estimates are 468 Mbp to 1235 Mbp, respectively (Arumuganathan and Earle 1991). Based on the genomic sequencing efforts the genome size of *Arabidopsis* can now be estimated at 125 Mbp (The *Arabidopsis* genome initiative 2000).

1-1-2 Repeated DNA sequences

The *A. thaliana* genome contains a low amount of repetitive DNA (~20%) (Meyerowitz 1994). Repeated DNA sequences can be classified into sequences organised in tandem arrays and those that are distributed throughout the genome.

Tandemly repeated DNA sequences are found at the telomeres, the centromeres and the nucleolar organising regions (NORs). The telomeric sequences are estimated to constitute about 0,3% of the genome (Richards and Ausubel 1988). Two NORs are identified on chromosomes 2 and 4 of *Arabidopsis* and contain clusters of 18, 25 and 5,8S rDNA arrays for an estimated size of 4 Mbp (Goodman *et al* 1995), or 6% of the genome. The 5S rDNA sequences are also organised in tandem but independently from the NORs. They account for 0,7% of the *Arabidopsis* genome (Campbell *et al.* 1992). The centromeric regions have been determined for all 5 chromosomes (RCEN1-5) (Round *et al.* 1997 ; Copenhaver *et al.* 1999). Long arrays of 180 bp tandemly repeated DNA sequences are found at each of the *Arabidopsis* centromeres. The 5S rDNA repeats are also observed in some centromeric regions.

Despite the small genome size and low amount of transposable elements, the *A. thaliana* genome has representatives of the main classes of transposons and retrotransposons (Konieczny *et al.* 1991; Pélissier *et al.* 1995; The *Arabidopsis* genome initiative, 2000).

Recently, the amount of transposable elements has been determined to be about 10% of the *A. thaliana* genome. Especially, certain families of retrotransposons are localised in the centromeric and peri-centromeric regions. Few repetitive sequences are found in euchromatic regions (The *Arabidopsis* genome initiative, 2000).

No parallel can be elaborated with cereal genome organisation. In these genomes, repetitive sequences represent between 50 and 80% of the nuclear DNA. The large majority of repetitive DNA is belonging to the retrotransposon class, which is spread all over the genome and interspersed with gene sequences (reviewed in Bennetzen 2000b).

1-1-3 Genetic maps

Many genetic maps have been established for *Arabidopsis* for a range of mapping populations with morphological markers as well as with molecular markers (Koornneef *et al.* 1983a; Koornneef *et al.* b; Chang *et al.* 1988; Nam *et al.* 1989; Hauge *et al.* 1993). Recombinant inbred lines established from a cross between the *Landsberg erecta* and *Columbia* ecotypes serve as a reference population (http://nasc.nott.ac.uk/new_ri_map.html; Lister and Dean 1993). Due to the almost complete homozygosity of the lines, they can be widely used and are thus particularly suited to integrate different types of markers information. Consequently, many marker types were then used such as RFLPs (Fabri and Schäffner 1994; Liu *et al.* 1996), RAPDs (Reiter *et al.* 1992), CAPSs (Konieczny and Ausubel 1993; Jarvis *et al.* 1994), microsatellites (Bell and Ecker 1994) and AFLPs (Alonso-Blanco *et al.* 1998).

1-1-4 Physical maps and the sequencing project

Physical maps of the *A. thaliana* genome have been established as a pre-requisite of the genome sequencing project which has been initiated in 1996. Alignments of BAC and YAC contigs gave a first glimpse into *A. thaliana* genome and its organisation (Schmidt *et al.* 1995; Zachgo *et al.* 1996; Schmidt *et al.* 1997; Kotani *et al.* 1997; Camilleri *et al.* 1998; Marra *et al.* 1999; Mozo *et al.* 1999). Since some of the clone contig maps are anchored to the genetic maps, the relationship between genetic and physical distances along the chromosomes could be studied. The frequency of recombination has been observed to vary with hot and cold spots along the chromosomes. Schmidt *et al.* (1995)

calculated an average of 185 Kbp/cM for chromosome 4 with variation from 30-50 kbp/cM for hot spots to >550 kbp/cM for cold spots.

Sequence data are now available for the whole *A. thaliana* genome, with the notable exception of the centromeres, telomeres and NORs. The total length of sequenced regions is 115,409,949 bp. Taking into account the length of the non-sequenced segments, the genome size is reported to be 125 Mbp. A total of 25,498 genes were predicted to be present in the sequenced regions. Around 70% of the genes could be grouped into different functional classes (The *Arabidopsis* genome initiative 2000). An overall analysis shows a homogenous density of genes and transposable elements on the five chromosomes (Lin *et al.*, 1999; Mayer *et al.* 1999; Salanoubat *et al*. 2000; Tabata *et al*. 2000, Theologis *et al*. 2000; The *Arabidopsis* genome initiative 2000). Previous studies estimated a gene density of 4,6 kbp/gene for *A. thaliana* (Barakat *et al* 1998). This is consistent with the values reported for the individual chromosomes, which vary from 4,0 kbp/gene to 4,9 kbp/gene (The *Arabidopsis* genome initiative 2000).

1-1-5 Duplications in the *Arabidopsis* genome

The extent of duplications within the *A. thaliana* genome became apparent by the analysis of large sequenced segments. Frequently, members of small gene families are found closely linked (Bevan *et al.* 1998; Terryn *et al.* 1999). Overall, it has been calculated that 17% of all *A. thaliana* genes are arranged in tandem arrays. The proportion of gene family members is greater in *A. thaliana* than in other eukaryotic model counterparts and is interpreted as an extreme tolerance of plants to tolerate increases in genome size (The *Arabidopsis* genome initiative 2000).

Large contigs of *A. thaliana* DNA having been sequenced, segmental duplications including many genes have been uncovered in the *A. thaliana* genome, like at first between chromosomes 2 and 4 (Lin *et al.* 1999; Mayer *et al.* 1999; Terryn *et al.* 1999). Furthermore, by combining mapping information for small multigene families and sequence alignment studies, Blanc *et al.* (2000) were able to draw a preliminary scheme of the duplicated segments, which comprise 60% of the *A. thaliana* genome.

The high divergence observed within these duplicated segments reveals the ancient nature of these duplications of *A. thaliana*. Indeed, only 20% to 47% of the genes were found in common between duplicated segments. With the completion of the genomic sequencing project, a refined map of the duplicated segments could be drawn and it could be confirmed that 60% of the genome is present in large segmental duplications (The *Arabidopsis* genome initiative 2000).

1-1-6 Expressed Sequence Tags (ESTs)

EST collections reflect the genes transcribed in an organism. The analysis of cDNA libraries made from different tissues and from material grown under various conditions results in a representation of many different gene sequences. Large scale analyses were initiated to estimate the validity of this approach (Höfte *et al.* 1993; Newman *et al.* 1994). ESTs are single-pass sequences which partially represent a particular cDNA clone. Over 100,000 *A. thaliana* ESTs are available in public databases (http://www.ncbi.nlm.nih.gov; http://www.tigr.org/tdb/agi/). Alignment of overlapping ESTs allows the construction of tentative consensus sequences (TCs). This decreases the redundancy of an EST collection and most importantly the EST assemblies often cover the entire protein-coding-sequence of genes (Rounsley *et al.* 1996; Quackenbush *et al.* 2000).

Only 57%-61,4% of the predicted genes have corresponding EST sequences (The *Arabidopsis* genome initiative 2000). Thus, despite the extensive *Arabidopsis* EST collections, experimental proof for many of the predicted genes is still lacking. Exon/intron structure predictions of genes are mainly relying on computer-based algorithms and it has been established that not all of the annotated genes are correctly predicted (The *Arabidopsis* genome initiative 2000). In contrast, aligning EST or cDNA sequences with the corresponding genomic sequence readily and reliably reveals the exon-intron structure of a particular gene. Therefore, EST contig information plays an important role in the annotation of genomic sequencing data. It has been established that in the segmental duplications of the *Arabidopsis* genome, only exon sequences are conserved due to the ancient nature of the duplications (Terryn *et al.* 1999; Blanc *et al.* 2000). Thus, alignment of duplicated gene sequences can also be exploited for improving gene structure predictions.

1-2 THE BRASSICACEAE FAMILY

The Brassicaceae family is comprised of 360 genera and approximately 3350 species (reviewed in Paterson *et al.* 2000, Schmidt *et al.* 2000). Especially, species of the

Brassica genus are of agricultural importance as oil seeds, vegetable and fodder crops. For example, sub-species of *Brassica oleracea* include cabbage, cauliflower, broccoli, kale, kohlrabi and Brussels-sprouts.

Marker assisted breeding is important for crop improvement of the *Brassica* species, therefore, efforts are underway to establish molecular linkage maps for *Brassica* species. The study of genetic similarities between genotypes is of special importance to permit maintenance and exploitation of germplasm resources (Lydiate *et al.* 1993). The close phylogenetic relationship to *A. thaliana* offers unique opportunities to transfer knowledge from the well-studied *Arabidopsis* genome to the related crop plants. Thus, it is important to study the genome organisation of these plants in a comparative way in order to develop methods which will achieve this transfer in an efficient manner.

The divergence time between *Brassica* and *A. thaliana* has been estimated at 14,5-20,4 million years ago (Yang *et al.* 1999), reflecting the close phylogenetic relationship between these species. The *Capsella* genus is also closely related to *Arabidopsis* and *Brassica*. *Capsella-Brassica* divergence time has been calculated at 12,4-19,5 million years ago, whereas the split of the *Capsella* - *A. thaliana* lineages seems to have occurred more recently, 6,2-9,8 million years ago (Koch *et al.* 1999). For comparisons, the split of the lineages leading to monocotyledonous and dicotyledonous plants has occurred between 170 and 235 million years ago (Yang *et al.* 1999).

1-2-1 Chromosome numbers and genome sizes

Members of the *Brassicaceae* family have different base chromosome numbers. Despite their close phylogenetic relationship, *A. thaliana* and *C. rubella* differ in chromosome number. They have 2n=2x=10 and 2n=2x=16 chromosomes, respectively. The diploid *Brassica* species, *B. nigra*, *B. oleracea* and *B. rapa* (syn. *campestris*) are also characterised by different chromosome numbers. They contain 2n=2x=16, 2n=2x=18 and 2n=2x=20 chromosomes, respectively. Interspecific hybridisation between these diploid species can generate stable fertile amphidiploid species (*B. napus*, *B. juncea* and *B. carinata*, Figure 1). The amphidiploid species can be considered to carry the entire genome sets of each of the progenitors (U, 1935).

The genomes of the *Brassica* species are much larger than that of their relative *A. thaliana*. The sizes vary from 0,97 pg/2C (*B. nigra*) to 2,56 pg/2C (*B. napus*). *B. oleracea* sub-species also show differences in genome size from 1,24 pg/2C (*B.*

oleracea ssp. *italica*) to 1,37 pg/2C (*B. oleracea* ssp. *botrytis*) (Arumuganathan and Earle, 1991). Species more closely related to *Arabidopsis* have also larger genome sizes than *A. thaliana*. For the tetraploid species *C. bursa-pastoris* (shepherd's purse) a value of 680 Mbp (Bennett and Smith 1976) is suggested.

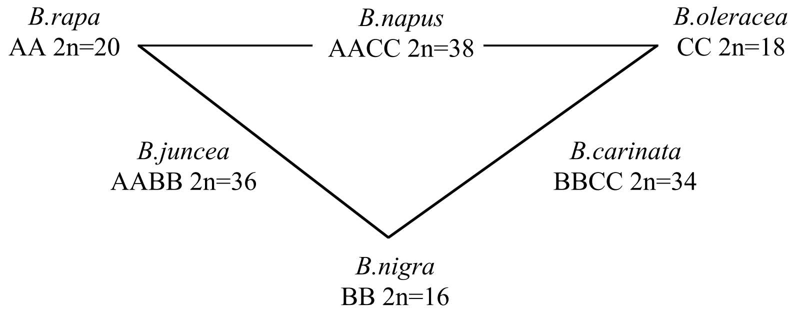


Figure 1: Genetic relationship of the cultivated *Brassica species*, redrawn from U (1935). The chromosome numbers are indicated for each species. A, B and C designate the genomes of the three diploid species.

1-2-2 Genome organisation

Cytogenetic studies (FISH) have been carried out to localise rDNA loci and some repetitive sequences on the chromosomes of *Brassica* species. In *B. oleracea* var. *alboglabra*, Amstrong *et al.* (1998) observed three distinct NOR loci (18S-5,8S-25S rDNA genes) on chromosomes 2, 4 and 7. The 5S rDNA sequences are located on the long arm of chromosome 2. Highly repetitive sequences co-localise with all peri-centromeric regions, but *in situ* hybridisation experiments revealed that the different centromeric regions are labelled with varying intensities.

1-2-3 Genetic mapping

Several maps have been established for different *Brassica* species utilising a variety of mapping populations. A common observation is the remarkable degree of duplication of the genome. Moreover, for some markers three and four loci could be detected. On average, not less than 30% of all RFLP markers tested revealed multiple loci. Sets of duplicated loci were found in the same order with similar distances between them on different linkage groups (Slocum *et al.* 1990; McGrath and Quiros 1991; Song *et al.* 1991; Teutonico and Osborn 1994). Several markers showed polymorphic as well as monomorphic fragments. This also indicated that multiple loci are corresponding to these markers.

These observations in combination with the larger genome size of *Brassica* compared to *A. thaliana* led to two different hypotheses about the ancestry of the *Brassica* genome. Either the *Brassica* genome was modified by many duplications and subsequently rearrangements (Truco *et al.*, 1996) or the cultivated diploid *Brassica* species are derived from polyploid ancestors (Lagercrantz *et al.* 1996).

1-2-4 Comparative mapping between *Brassica* species

Genetic mapping experiments of *Brassica* species revealed rearrangements, inversions, translocations and duplications, if the genomes of the three cultivated diploid *Brassica* species were compared in a pairwise fashion. Between each of the pairs of the three cultivated diploid *Brassica* species, 5 to 12 rearrangements were detected (Lagercrantz and Lydiate 1996). In contrast, in the amphidiploid *B. napus,* the genomes of its progenitors *B. rapa* and *B. oleracea* are present with almost no alterations (Sharpe *et al.* 1995 ; Parkin *et al.* 1995 ; Bohuon *et al.* 1996). Similarly, the *B. juncea* genome reflects the organisation of the genomes of its progenitors (Axelsson *et al.* 2000)
By analysing the genome of *B. nigra* via RFLP markers, Lagercrantz and Lydiate (1996) could describe that the whole *B. nigra* genome was arranged as eight sets of triplicated collinear chromosomal segments. Thus, the *B. nigra* genome can be viewed as reshuffled assembly of three complete copies of a putative ancestral genome. Due to the fact that almost all *B. nigra* segments could be identified in *B. oleracea* and *B. rapa*, it has been concluded that the A and C genome are also showing the genome triplication initially identified in *B. nigra*. Fission and fusion of chromosome segments would have then reshuffled the A, B and C genomes in different ways (Lagercrantz and Lydiate 1996). The rather similar sizes of the A, B and C genomes corroborates this view (507-516 Mbp for *B. rapa*, 468 Mbp for *B. nigra* and 599-662 Mbp for *B. oleracea*).

1-2-5 Comparative genetic mapping between *Arabidopsis thaliana* and related species

Rearrangements between the *A. thaliana* and *C. rubella* genomes are expected due to their different base chromosome numbers. Genetic mapping of markers located on *A. thaliana* chromosome 4 revealed two collinear linkage segments in *Capsella*, one of which contains an inversion (Acarkan *et al.* 2000). Marker repertoire has been shown to be conserved between both species, only two *A. thaliana* markers analysed did not hybridise the *C. rubella* DNA.

Comparative mapping studies have been carried out to determine how the genome of an ancient polyploid such as *Brassica* is related to the one of its close diploid relative *A. thaliana*. Genetic mapping experiments showed that clusters of closely linked RFLP loci are conserved between *Brassica oleracea* and *A. thaliana*, but extensive rearrangements are observed (Kowalski *et al.* 1994). It is worthwhile to point out that this analysis provided evidence for duplications in the *A. thaliana* genome (Kowalski *et al*. 1994).

A detailed comparison of the *A. thaliana* and *B. nigra* genomes established the average size of conserved linkage segments at 8 cM. This corresponds to ~90 rearrangements between the genomes of these species (Lagercrantz 1998). A 1,5 Mbp segment of *A. thaliana* surrounding the *CO* gene (control of flowering time) has been analysed for collinearity in *B. nigra*. The genetic experiments revealed two intact homologous regions in *B. nigra* equivalent to the *A. thaliana* segment and a third one carrying a large chromosomal inversion (Lagercrantz *et al.* 1996). A study of markers located in a 30 cM segment of *A. thaliana* chromosome 4 matches six segments in the *B. napus* genomes, two of which are also characterised by a large inversion (Cavell *et al.* 1998). A segment of 15 kbp on chromosome 3 of *A. thaliana* has been found completely collinear with a single linkage group in the *B. nigra, B. oleracea* and *B. rapa* genomes, but additionally, partial clusters were discovered in all three species (Sadowski *et al.* 1996). Furthermore, for six genes present in a 30 kbp region of *A. thaliana* chromosome 4 five corresponding loci could be found in the *B. nigra* genome, one of which included all six genes and whereas the others represented imcomplete copies of the locus (Sadowski and Quiros 1998).

Six BAC inserts of *A. thaliana* were used as probes in fluorescent *in situ* hybridisation experiments on DNA fibres of *B. rapa* chromosomes. Multiple hybridising regions of similar size as the *A. thaliana* segments were observed in *B. rapa*. This led the authors to conclude that the increase of the *Brassica* genome size is mainly due to duplications rather than accumulation of repetitive sequences in intergenic regions (Jackson *et al.* 2000).

The comparison between the *A. thaliana* and *Brassica* genomes revealed further evidences for the complex nature of the *Brassica* genomes. The collinearity studies detected conserved segments, but rearrangements were frequently seen.

1-2-6 Initiation of microcollinearity studies within the Brassicaceae family

A. thaliana is self-fertile, but some species of the genus *Brassica* are self-incompatible. A comparative physical mapping study on the region encoding the self-incompatibily locus (*SLG/SLK)* showed that the homeologous region of *Arabidopsis* was highly conserved with the exception that the self-incompatibily genes were absent from this locus (Conner *et al.* 1998). An *A. thaliana* fragment carrying the single-copy *RPM1* gene surrounded by GTP and M4 markers has been shown to have six homeologous loci in *B. napus* but only two carry a copy of *RMP1*. Sequencing of the incomplete loci in *B. napus* suggested that the absence of *RPM1* is due to deletions (Grant *et al.* 1998).

These first microcollinearity studies show that many small alterations can be found at the level of the genes between the *Arabidopsis* and *Brassica* genomes. In contrast, comparisons between small regions of the *A. thaliana* and *C. rubella* genomes revealed conserved gene repertoire and order (Acarkan *et al.* 2000; Rossberg *et al.* 2001).

1-2-7 Mobile elements in the Brassicaceae family

The great influence of repetitive elements on genome size has been noted in grasses. Microcolinearity studies revealed that in the maize genome many retrotransposons are found interspersed with genes, whereas these elements are not as frequent in the orthologous regions of the much smaller sorghum and rice genomes. Nevertheless, despite the presence of many retrotransposons extensive microcollinearity is observed (Chen *et al*. 1998; Tikhonov *et al*. 1999).

Little is known about transposable elements in *Brassica* and their putative conservation with mobile elements in the *A. thaliana* genome. But common occurrence of different types of elements has been reported in a number of studies. Long interspersed elements (LINEs, Noma and Ohtsubo 1999), miniature inverted-repeat transposable elements (MITEs, Casacuberta *et al.* 1998), *Ty1-Copia*-like (Hirochika and Hirochika 1993) and *Ty3-Gypsy*-like retrotransposons (Suoniemi *et al.* 1998) have been found at least in common between *A. thaliana* and one of the *Brassiceae* members. In contrast, short interspersed elements (SINEs) present in Brassiceae species were not cross-hybridising to *A. thaliana* DNA (Lenoir *et al.* 1997).

1-3 OBJECTIVE OF THIS STUDY

This study is aiming at the comparative genome analysis of three species of the Brassicaceae family, *Arabidopsis thaliana*, *Capsella rubella* and *Brassica oleracea*. Lineages leading to *Arabidopsis* and *Capsella* separated 6,2-9,8 million years ago, whereas *Brassica* diverged from *Arabidopsis* and *Capsella* 14-20 million years ago. The species chosen for the analysis are characterised by different chromosome numbers and genome sizes, furthermore *Brassica oleracea* is of relatively recent polyploid origin.

Arabidopsis markers and sequence information were exploited to generate a linkage map of *Capsella*. This part of the study was aimed at an overall comparison of gene repertoires in both species. Furthermore, it was intended to analyse the collinearity of the *Arabidopsis* and *Capsella* genomes.

A 50 kbp region located on the long arm of *A. thaliana* chromosome 4 was chosen for a microcollinearity study. The aim was the identification of homologous regions in the *C. rubella* and *B. oleracea* genomes and their characterisation in respect to gene repertoire and order. Another objective included the comparative analysis of exon/intron structures of orthologous genes.

Repetitive DNA sequences constitute a large fraction of plant genomes. A comparative analysis of retroelement-like sequences in the *A. thaliana* and *C. rubella* genomes was carried out to reveal more about the conservation of such sequences in the Brassicaceae family.

1-4 ABBREVIATIONS

acc. no.	accession number
A. thaliana	*Arabidopsis thaliana*
BAC	Bacterial Artificial Chromosome
bp (kbp, Mbp)	base pair (kilo, Megabasepair)
B. oleracea	*Brassica oleracea*
BSA	bovine serum albumine
°C	degree Celsius
C. grandiflora/rubella	*Capsella grandiflora/rubella*
chr.	chromosome
cos	cosmid
CTAB	Cetyl tri-methyl ammonium bromide
DNA	Desoxyribonucleic acid
dNTP	desoxyribo nucleoside tri-phosphate
E. coli	*Escherichia coli*
EDTA	ethylene diamino tetra acid
ENV	envelope gene
EST	expressed sequence tag
g (mg, µg)	gram (milligram, microgram)
indel	insertion/deletion
INT	integrase
IPCR	inverse polymerase chain reaction
IPTG	isopropyl-β-D-thio-galactoside
l (ml, µl)	litre (millilitre, microlitre)
LG	linkage group
LINE	long interspersed element
LTR	long terminal repeat
M (mM)	molar (millimolar)
mA	milli-Ampere
MDE^{TM} gel solution	Mutation Detection Enhancement gel solution
NOR	nuclear organising regions
O/N	overnight

ORF	open reading frame
PBS	primer binding site
PCI	phenol / chloroform / isoamylalcohol
PCR	polymerase chain reaction
PFGE	pulsed field gel electophoresis
PPT	polypurin tract
RFLP	restriction fragment length polymorphism
RNAse	ribonuclease
rpm	rotation per minute
RT	room temperature
RT	reverse transcriptase
SDS	sodium dodecyl sulphate
SINE	short interspersed element
SNP	single nucleotide polymorphism
SSCP	single-stranded conformation polymorphism
TSD	target site duplication
Tris	tris-(hydroxymethyl)-aminomethane
U	units
V	Volt
vol	volume
v/v	volume per volume
w/v	weight per volume
X-gal	5-bromo-4-chloro-3indolyl-β-thiogalactoside
YAC	Yeast Artificial Chromosome

2 MATERIAL AND METHODS

2-1 MATERIAL

2-1-1 Equipment

Thermal cycler (PTC200)	*Biozym Diagnostic GmbH*, Hess. Oldendorf, Germany
Centrifuges	*Beckman Instruments*, Unterschleissheim-Lohhof, Germany *Eppendorf,* Köln, Germany *Heraeus Instruments*, Hanau, Germany
Hybridisation chambers	*Heraeus Instruments*, Hanau, Germany
ABI prism 377 and 3700 sequencers	*Perkin Elmer*, Überlingen, Germany
UV crosslinker	*Amersham Pharmacia Biotech*, Freiburg, Germany
Pulsed field gel electrophoresis	*Bio-Rad Laboratories GmbH*, München, Germany

2-1-2 Enzymes and nucleotides

Restriction enzymes and buffers	*Boehringer Mannheim*, Mannheim, Germany *New England Biolabs Inc*., Frankfurt, Germany *Gibco BRL*, Karlsruhe, Germany *MBI Fermentas*, St. Leon-Rot, Germany
[α-^{32}P]-dCTP	*Amersham Pharmacia Biotech*, Freiburg, Germany *Hartmann Analytic*, Braunschweig*, Germany*
Desoxyribonucleotidetriphosphate (dNTP) DNA size standards DNA polymerase I (Klenow Fragment)	*MBI Fermentas*, St. Leon-Rot, Germany
Oligonucleotides	*Gibco BRL*, Karlsruhe, Germany *Metabion*, Planegg-Martinsried, Germany *MWG*, Ebersberg, Germany
Adenosinetriphosphate (ATP) Ribonuclease A (RNase A) Random hexamers p(dN)$_6$	*Boehringer Mannheim*, Mannheim, Germany
Salmon sperm DNA	*Sigma*, Deisenhofen, Germany
Taq polymerase and buffer	*Gibco BRL*, Karlsruhe, Germany
T4 DNA ligase and buffer	*New England Biolabs Inc.,* Frankfurt, Germany

2-1-3 Chemicals and media components

Agarose	*Gibco BRL*, Karlsruhe, Germany
Bacto agar Bacto tryptone	*Difco Laboratories*, Detroit USA

Yeast extract	
Acrylease	*Stratagene,* Amsterdam, The Netherlands
BSA	*MBI Fermentas*, St. Leon-Rot, Germany
Chemicals	*Calbiochem* Novabiochem GmbH, Schwalbach, Germany *Duchefa Biochemie BV*, Haarlem, The Netherlands *Gibco BRL*, Karlsruhe, Germany *J. T. Baker*, Deventer, The Netherlands *Merck*, Darmstadt, Germany *Amersham Pharmacia Biotech*, Freiburg, Germany *Serva*, Heidelberg, Germany *Sigma*, Deisenhofen, Germany
MDE™ gel solution	*FMC BioProducts Corp.*, Rockland, USA

2-1-4 Purification systems

DEAE cellulose paper	*Amersham Pharmacia Biotech*, Freiburg, Germany
Low copy plasmid purification (Nucleobond AX plasmid-purification kit™)	*Macherey-Nagel GmbH*, Düren, Germany
PCR product purification (High Pure PCR-purification-Kit™) Plasmid DNA purification (High Pure Plasmid-Kit™)	*Boehringer Mannheim*, Mannheim, Germany

2-1-5 Blotting material and films

Nylon membranes	Hybond N$^+$	*Amersham Pharmacia Biotech*, Freiburg, Germany
	Biodyne A/B	*Pall,* Dreieich, Germany
Whatman® 3MM		*Whatman*, Maidstone, England
X-ray films (Kodak-X-OMAT-AR-5)		*Sigma*, Deisenhofen, Germany

2-1-6 Biological material

2-1-6-1 *Capsella* mapping population

An interspecific cross of the self-incompatible species *Capsella grandiflora* and the self-compatible species *Capsella rubella* was carried out. One of the resulting progeny plants was allowed to self-fertilise. Of 100 F2 plants, 50 self-compatible plants were selected for the mapping population (Acarkan *et al*. 2000).

2-1-6-2 *Brassica* mapping population

The population used for genetic mapping experiments in *Brassica oleracea* was established from the highly polymorphic cross of *B. oleracea* var. *alboglabra* x *B. oleracea* var. *italica* (A12DHd x GDDH33) (Bohuon *et al.* 1996). A subset of 40 double haploid lines was analysed to map different loci via RFLP and SSCP analysis.

2-1-6-3 RFLP markers

Arabidopsis DNA sequences were used as RFLP markers for genetic mapping experiments in *Capsella*.

- mi-markers: *Pst*I fragments of genomic DNA derived from *A. thaliana* ecotype Columbia and cloned into vector pUC119 (Liu *et al.* 1996).
- EST clones: *A. thaliana* cDNA clones for which partial sequence information is available (Newman *et al.* 1994; Höfte *et al.* 1993).

2-1-6-4 *Brassica/Capsella* cosmid libraries

Cosmid libraries containing genomic DNA of *Capsella rubella* and *Brassica oleracea* were established in the laboratory (Schmidt *et al.* 1999).

Total genomic DNA from *Brassica oleracea* var. *alboglabra* and *Capsella rubella* was partially digested either with *Taq*I or *Mbo*I, cloned into cosmid vector pCLD04541 (Bancroft *et al.* 1997) and transformed into *E. coli* strain *SURE*^{TM}2. Cosmid vector pCLD04541 carries a tetracycline resistance gene. The average size of the genomic DNA inserts is approximately 20 kbp (Schmidt *et al.* 1999).

About 23,000 clones were gridded into 384-microwell plates for each of the two *Capsella* cosmid libraries. It has been estimated that both libraries together encompass 4-5 genome equivalents (Acarkan, 2000). Approximately, 110,000 gridded cosmid clones make up the libraries containing *Brassica oleracea* genomic DNA. This corresponds to 2,5-4 genome equivalents since the genome size of *Brassica* has been estimated to be between 600 and 870 Mbp (Arumuganathan and Earle 1991; Bennett and Smith 1976).

2-1-6-5 *Brassica* BAC libraries:

- BAC library: This library of *B. oleracea* A12DHd DNA contains approximately 25,000 clones with an average insert size of 110 kbp. The library corresponds to 4,5 genome equivalents (C. Ryder and G. King, HRI Wellesbourne, unpublished; http://hbz.tamu.edu/bacindex4.html). Genomic DNA is cloned into the *Hin*dIII site of cloning vector pBeloBAC11 which carries a chloramphenicol resistance gene (Kim *et al.* 1996).
- BIBAC library: *Brassica oleracea* var. *alboglabra* DNA has been partially digested with *Sau*3AI and cloned into the *Bam*HI site of the pBIBAC 2 vector (binary BAC vector, Hamilton 1997). The pBIBAC 2 vector carries a kanamycine resistance gene. The average size of the genomic DNA inserts is 145 kbp for 85% of the 34,000 clones (O'Neill & Bancroft 2000). Thus, this library provides a 6.9-fold redundancy of the *Brassica oleracea* genome if one considers a genome size of 600 Mbp (Arumuganathan and Earle 1991).

2-1-7 Bacterial strains, vectors and media

E. coli strains	K12 DH5α (Hanahan 1983) *SURE*™2 (Stratagene)
Vectors used	pGEM 7Zf$^+$ (Promega) pGEMTeasy (Promega) pCLD04541 (Bancroft *et al.* 1997)

LB medium	1% (w/v) bacto tryptone 0,5% (w/v) yeast-extract 1% (w/v) NaCl pH 7,0 with NaOH
LB agar	LB medium solidified with 1,5% (w/v) bacto-agar

Media supplements		Stock solution	Working concentration
	IPTG	23,8 mg/ml	23,8 µg/ml
	X-Gal	20 mg/ml	20 µg/ml
Antibiotics:	carbenicilline	200 mg/ml	200 µg/ml
	tetracycline	5 mg/ml	10 µg/ml
	chloramphenicol	34 mg/ml	12,5 µg/ml
	kanamycine	50 mg/ml	40 µg/ml

2-1-7 Oligonucleotides

Oligonucleotide sequences used for sequencing and isolation of inserts cloned into plasmid vectors:

universe :	GTA AAA CGA CGG CCA GT
reverse:	AAC AGC TAT GAC CAT G
T7:	GTA ATA CGA CTC ACT ATA GGG C
T3:	AAT TAA CCC TCA CTA AAG GG
SP6:	CAT ACG ATT TAG GTG ACA CTA TAG

Oligonucleotide sequences used for isolation of BAC and cosmid insert sequences adjacent to vector sequences with iPCR (inverse PCR):

Cosmid-vector specific oligonucleotides:

cos1:	GGA GCT CCA ATT CGC CCT5 ATA G
cos2:	GGC GGC CGC TCT AGA ACT AG
cos3:	GCT TGA TAT CGA ATT CCT GC
cos4:	CGA TAC CGA CCT CGA GG
cos5:	GGA ATT CGA TAT CAA GCT TA
cos6:	CAG CCC GGG GGA TCC ACT AGT
cos7:	CCC CTC GAG GTC GAC GGT
cos8:	GGT ACG TAC CAG CTT TTG TT

BAC-vector specific oligonucleotides:

BAC1:	CTG CAG GCA TGC AAG CTT (Woo *et al*. 1994)
BAC2:	CAG CTG AGA TCT CCT AG (Woo *et al*. 1994)
BAC3:	CAA TTC CAC ACA ACA TAC G
BAC4:	GTG ATA TCT TAT GAG TTC G (Woo *et al*. 1994)
BAC5:	CAT TAA TGA ATC GGC CAA CG

2-2 METHODS

All standard molecular biology techniques were performed according to Sambrook *et al.* (1989).

2-2-1 RFLP procedure

2-2-1-1 Genomic DNA preparation

Capsella genomic DNA was prepared according to Dellaporta *et al.* (1983) with the modifications described in Schmidt *et al.* (1999). For the isolation of genomic DNA from *Brassica*, the DNA extraction method from Saghai-Maroof *et al.* (1984) with the modifications by Hoisington (1992) was used.

2-2-1-2 Southern blot (Southern 1975)

One mg of genomic DNA was digested with appropriate restriction enzymes (e.g. *Dra*I, *Eco*RI, *Eco*RV, *Xba*I) for 6h at 37°C and separated on a 0,8% TBE agarose gel with 1xTBE running buffer at 55V O/N. The gel was incubated for 15-30 min in HCl (1% solution), for 30 min in denaturation buffer and for another 30 min in neutralisation buffer before the DNA was transferred onto charged Hybond N^+ membrane (Amersham Pharmacia Biotech). Transfer was carried out O/N using 20xSSC buffer. The membrane was baked for 30 min at 80°C as recommended by the manufacturer.

C. grandiflora plants are self-incompatible, consequently, a finite amount of material was available to prepare DNA from this parental plant of the mapping population. In order to have sufficient DNA for RFLP analysis, a DNA pool of all individuals of the F2 progeny derived from the interspecific cross has been set up. This pool represents alleles of the *C. grandiflora* and the *C. rubella* parental plants in an equal fashion. For polymorphism analysis, membranes were prepared carrying one mg of genomic DNA from *C. rubella* and the F2 pool, respectively, digested with appropriate restriction enzymes. An RFLP is recognised, if one or several additional fragments are present in the lane carrying DNA of the F2 pool compared to the lane with *C. rubella* DNA. Hence, these additional fragments are likely *C. grandiflora* specific. A set of two membranes carried DNA of the individual progeny plants. The first membrane contained as first lane the DNA of the F2 pool, then DNA of individuals 1-25 of the F2 progeny and at last, *C. rubella* DNA. The second membrane was prepared in a similar way, it contained DNAs of F2 plants 26-50.

10xTBE buffer	40 mM Tris 40 mM boric acid 10 mM EDTA (pH 8,0)
Denaturation buffer	1,5 M NaCl 0,5 M NaOH
Neutralisation buffer	1,5 M NaCl 0,5 M Tris-HCl (pH 7,2) 0,001 M EDTA (pH 8,0)
Transfer buffer (20xSSC)	3 M NaCl 0,3 M Na_3 citrate

2-2-1-3 Hybridisation experiments

Membranes were pre-hybridised at 65°C for 2-4 hours in 5x Denhardt's hybridisation solution containing 200 μg denatured salmon sperm DNA.

The radioactively labelled probe was prepared in the following way: denatured DNA (50-100 ng) was mixed with 1x random prime buffer, one unit of Klenow fragment of DNA polymerase I, one µl of BSA solution (20 mg/ml) and 30 to 50 µCi α-^{32}P-dCTP. The reaction mixture was incubated 15-20 min at 37°C. Unincorparated nucleotides were separated from the labelled DNA fragment with a column (High pure PCR purification kit, Boehringer Mannheim). The DNA fragment was denatured at 95°C before it was added to the pre-hybridisation solution. Hybridisation took place O/N at 65°C.

The filters were washed twice with a 2xSSC/0,1%SDS solution for 30 min, and then 10 min with 1xSSC/0,1%SDS at 65°C. Membranes were exposed to films for 2 to 5 days at –80°C, in cassettes with intensifying screens.

5x-Denhardt's hybridisation solution	5xSSC 0,1% SDS 5x Denhardt's
1x random prime buffer	200 µl/ml solution A 500 µl/ml solution B 300 µl/ml solution C
Solution A	1 ml solution O 1,8% (v/v) β-mercaptoethanol 5 µl dATP, 5 µl dGTP, 5 µl dTTP (100 mM dNTP stock solutions)
Solution B	2 M Hepes (pH 6,6)
Solution C	90 OD_{260} $p(dN)_6$/ml in TE
Solution O	0,125 M Tris-HCl (pH 8) 0,125 M $MgCl_2$

2-2-2 BAC and cosmid DNA minipreparation

For isolation of DNA from BAC or cosmid clones a protocol developed by the Texas A&M University BAC Center (http://hbz.tamu.edu/cgi-bin/htmlassembly?bacbacc) was used with the following modifications: four ml of LB containing an appropriate antibiotic were inoculated with a single *E. coli* colony, and incubated at 37°C O/N with shaking. The resulting culture was centrifuged at 14,000 rpm for two min at RT, the supernatant was discarded and the pellet was resuspended in 200 µl TE. Four-hundred µl of buffer II were added, the tube was gently inverted and incubated on ice for 5 min. Proteins were precipitated by adding 300 µl of buffer III. After a 10 min incubation at –80°C, the preparation was kept for 20 min at RT and subsequently centrifuged for 15 min, 14,000 rpm at RT. DNA was precipitated by adding 0,6 vol of ice-cold isopropanol to the supernatant. The preparation was kept for 10 min at –80°C and after 20 min at RT, the solution was centrifuged for 20 min, 14000 rpm at 4°C. The DNA pellet was washed with one ml of ice-cold 70% ethanol, centrifuged for 2 min, 14,000 rpm, 4°C, and air-dried for 10 min. The pellet was then dissolved in 50 µl TE/RNAse (50 µg/ml), incubated for 5 min at 56°C, and at 37°C for 30 min. Ten µl of the preparation were used for a restriction digest.

TE	10 mM Tris-HCl 1 mM EDTA (pH 8,0)
Buffer II	0,2 M NaOH 1% SDS
Buffer III	3 M potassium acetate, adjusted to pH 4,8 with glacial acetic acid
TE/RNase	10 mg/ml bovine pancreatic Rnase 10 mM Tris-HCl (pH 7,5) 15 mM NaCl Heat to 100°C

2-2-3 Isolation of BAC and cosmid end fragments by iPCR (inverse PCR)

BAC and cosmid clones carry inserts of genomic DNA in vectors with known DNA sequence. BAC and cosmid end fragments represent the sequence of genomic insert DNA directly adjacent to the polylinker sequence of the vector. Such end fragments are especially valuable for establishing BAC or cosmid clone contigs. They can be used as probes to screen libraries in chromosome walking experiments.

BAC or cosmid clone DNA was digested with a suitable restriction enzyme. The digested fragments were ligated to form circles. An aliquot of the ligation mixture was used as template for PCR with pairs of oligonucleotides corresponding to polylinker sequences either to the left or to the right of the cloning site. The two oligonucleotides of a particular primer pair were designed such that an amplification reaction was only possible after circle formation. Only those genomic DNA sequences could be amplified, which were attached to vector sequences carrying the primer binding sites. The resulting PCR products were purified to be sequenced or to be used as probe in hybridisation experiments.

The left borders of inserts cloned into the *Taq*I site of cosmid vector pCL04541 were isolated by digestion of the cosmid clones with *Pvu*II, *Pvu*I, *Hinf*I, *Sac*I, *Hae*3AI, or *Xba*I. PCR was performed with primer combination cos5/cos6. Sequences corresponding to the right borders of these *Taq*I cosmid inserts could be rescued by digestion of the cosmid DNA with *Pvu*II or *Hinf*I. The PCR reaction was carried out with the primer combination cos7/cos8.

Left borders of *Sau*3AI cosmid clones were isolated with restriction enzymes *Pvu*II, *Pvu*I, *Hinf*I and *Rsa*I, and PCR was performed with primer combinations cos1/2. The right border could be obtained by digestion of cosmid DNA with restriction enzymes *Pvu*II, *Hinf*I, *Hae*3AI, *Rsa*I or *Apa*I and the religated circles were amplified with primer combination cos3/cos4. All these primers correspond to the DNA sequence of vector pCLD04541 used for establishing the *Brassica* and *Capsella* cosmid libraries (Schmidt *et al.* 1999). The oligonucleotide sequences were given in chapter 2-1-8. BAC end fragment isolation was only performed on the Wellesbourne BAC library (C. Ryder and G. King, unpublished). Primer sequences were already published concerning vector pBeloBAC11 (Woo *et al.* 1994), but two additional oligonucleotides suitable for this strategy were developed (2-1-8).

2-2-4 Pulsed field gel electrophoresis

Pulsed field gel electrophoresis (PFGE) is a technique for resolving DNA of a size range of several kbp to several Mbp. The DNA molecules are oriented in the agarose matrix using alternating electric fields between spatially distinct pairs of electrodes. Electrodes are placed in the chamber in a hexagon arrangement with the agarose gel in the centre.

BAC DNA (200-400 ng) was digested for 3-5 hours at 37°C with 10 U *Not*I to release the genomic DNA insert from the vector sequences. An 1% TBE agarose gel was prepared in a casting stand. Within the electrophoresis chamber, the gel was fixed to avoid movement of the gel due the cooling buffer system. The 0,5x TBE electrophoresis buffer together with the gel were cooled in the electrophoresis chamber to 12°C. Loading dye was added to the samples which were then incubated for 5 min at 56°C, chilled immediately on ice and loaded onto the gel. λ DNA (Boehringer Mannheim, CI857Sam7), concatemeres of λ DNA and λ DNA digested with *Hin*dIII were used as length standards.

The electrophoresis conditions varied depending on the nature of DNA fragments which were to be separated:

Sizing of BAC inserts (NotI digestions)	*Restriction analysis of BAC clone DNA (MluI or SmaI digestion)*
Switch times are gradually increased:	Switch times are gradually increased:
3-5 sec initial switch time	2 sec initial switch time
6-15 sec final switch time	6 sec final switch time
6 V/cm	6 V/cm
120° angle	120° angle
12°C	14°C
22-24 h run	12 h run

2-2-5 Construction of a library containing *Mbo*I fragments of *C. rubella* total DNA

2-2-5-1 Digestion of total *C. rubella* DNA

One mg of total *C. rubella* DNA was digested to completion with 2x25U of the enzyme *Mbo*I for 2x3h at 37°C. The sample was concentrated (via evaporation) and separated on an 0,8% TAE agarose gel for 4h, 27V/22mA.

2-2-5-2 Purification of fragments

Fragments with a size range between 1,5Kb and 500bp were concentrated through electrophoresis onto a piece of DEAE cellulose paper. The paper containing the DNA fragments was transferred into a tube containing 400 µl of DEAE solution, ground and incubated for 90 minutes at 65°C. A hole was pierced into the bottom part of the tube, and the solution was transferred into a second tube by centrifugation, 2 min at 14000rpm. A PCI extraction was required to remove remaining cellulose fibres. The DNA was precipitated from the supernatant with 0,7 vol isopropanol and centrifuged for 30 min at 4°C and 14,000 rpm. The pellet was washed with 700 µl of 70% ethanol, centrifuged 5 min, 4°C, 14,000 rpm and resuspended in 20 µl water. An aliquot of 1 µl was separated on an agarose gel to quantify the extracted DNA.

DEAE solution	20 mM Tris-HCl pH 7,5 1 mM EDTA, pH 8,0 1,5 M NaCl
PCI (25:24:1)	phenol chloroform isoamylalcohol
50x TAE	4 M Tris 5,7% v/v acetic acid 50 mM EDTA

2-2-5-3 Ligation reaction and transformation

Insert DNA (100-150 ng) was ligated with 50 ng of *Bam*HI digested vector DNA (PGEM7Zf+) using 1 U T4 DNA ligase in a final volume of 10 µl at 16°C, O/N. An aliquot of 200 µl of DH5α competent cells (Hanahan 1983) was incubated on ice with 2 µl of the ligation reaction for 30 min. After 90 sec at 42°C, the cells were immediately chilled on ice for 1 min and incubated with 800 µl LB at 37°C for 45 min. The whole transformation mixture was plated on LB plates containing carbenicillin, IPTG and X-Gal and incubated at 37°C O/N. Only white, carbenicillin resistant colonies were used for further analysis.

2-2-6 SNP-SSCP

Several methods have been developed to analyse single nucleotide polymorphisms, for example SSCP (single-stranded conformation polymorphism). This technique is based on different mobilities of denatured DNA strands in MDE™ gels. Even single nucleotide differences in DNA fragments analysed may be detectable using this method, because they might influence the conformation of the strands when separated on a MDE™ gel.

- The PCR for fragments of sizes between 200 and 400 bp was performed in a total volume of 25 µl. After amplification, 5 µl were separated on an 0,8% TAE agarose gel and 2-4 µl were added to SSCP dye solution, denatured at 94°C for 3 min and immediately placed on ice.
- The polyacrylamide gel solution was poured between two glass plates, which were differently treated as follows. One glass surface was treated with acrylease™, an antistick coating solution; the other glass plate was treated with γ-metha-acryloxypropyl-trimethoxysilane, a binder component. The spacers were 0,4 mm thick. The 0,5xMDE™ gel solution was polymerised with TEMED (N, N, N', N'-tetramethyl-ethylene diamine) and 10% APS (ammonium persulfate).
- The samples were loaded and electrophoretically separated at 0,5-1,5W, 100-140V, 4-7,5mA O/N (14-18h) in 0,6x TBE buffer.
- The detection of the fragments was done by silver staining (Sanguinetti *et al.* 1994). The gel (fixed to the silanised glass) was incubated for 3 min in the fixation solution, and stained for 7 min in the silver nitrate solution. After a short rinse in water, the fragments are detected with a NaOH-based developing solution. DNA

fragments were visible after 10-20 min. The gel was fixed again, rinsed with water and then dried and scanned.

10x PCR buffer	100 mM Tris (pH 8,3) 500 mM KCl 20 mM $MgCl_2$ 0,1% (v/v) gelatine 0,05% (v/v) Tween 20 0,05% (v/v) NP40
Gel solution 0,5x MDE	0,6x TBE 5% (w/v) glycerin 0,5x MDE gel solution (2x) 0,06% TEMED 0,05% APS
SSCP dye buffer	95% Formamide 0,01 M NaOH 0,05% bromophenol blue 0,05% xylene cyanol
Fixation solution	10% ethanol 0,5% acetic acid
Staining solution	0,2% $AgNO_3$ 10% ethanol 0,5% acetic acid
Developing solution	3% NaOH 0,1% formaldehyde

2-2-7 DNA sequencing and analysis

DNA sequencing was performed by the ADIS unit (Max-Planck-Institute) with PE/Applied Biosystems 377 and 3700 sequencers using BigDye-terminator chemistry (Perkin Elmer). The resulting sequences were analysed using the Wisconsin Package (version 10.0-UNIX, Genetic Computer Group [GCG], Madison, WI, USA). Comparisons of sequences with *Arabidopsis thaliana* genomic DNA or EST sequences were performed with the BLAST program (Altschul *et al.* 1997) using several providers: MIPS (Munich Information Center for Protein Sequences), NCBI (National Center for Biotechnology Information), TAIR (The *Arabidopsis* Information Resource) and TIGR (The Institute for Genome Research). The EST contig information could be retrieved from the TIGR web site.

Gene predictions were carried out using two different programs, GeneMark.hmm and Genscan. Alignment of multiple nucleotide sequences, or amino-acid sequences were carried out with the Clustal W software. All internet resources used during this study are listed below.

Programs/Databases	*Web sites*
Clustal W	http://www2.ebi.ac.uk/clustalw/ http://www.clustalw.genome.ad.jp
GeneMark.hmm	http://dixie.biology.gatech.edu/GeneMark/eukhmm.cgi
Genscan	http://genes.mit.edu/GENSCAN.html
MIPS	http://mips.gsf.de/proj/thal/
NCBI (BLAST)	http://www.ncbi.nlm.nih.gov/BLAST/
TAIR	http://www.arabidopsis.org
Tandem Repeat Finder	http://c3.biomath.mssm.edu/
TIGR *Arabidopsis* gene index	http://www.tigr.org/tdb/agi/index.html

2-2-8 Genetic linkage analysis

Linkage analysis of the *Capsella* mapping population and establishment of a map was performed with the MAPMAKER program (Lander *et al.* 1987), using the Haldane centiMorgan function.

3 RESULTS

3-1 GENETIC MAPPING

This study aims to establish a genetic map for *Capsella* based on molecular markers. The resulting linkage groups shall be compared to the maps of the *Arabidopsis* chromosomes. A common set of markers is a prerequisite to establish comparative genetic maps of two species. For *A. thaliana* extensive molecular marker maps have been assembled, moreover the genome is completely sequenced (The *Arabidopsis* genome initiative 2000). These resources were exploited to generate a genetic map of *Capsella*. Markers have not been randomly chosen, rather it has been attempted to select markers homogenously distributed on the five *Arabidopsis* chromosomes for genetic mapping in *Capsella*.

3-1-1 *Capsella* mapping population

The *Capsella* mapping population is composed of 50 self-compatible F2 individuals derived from an interspecific cross of *C. grandiflora* and C. *rubella*. For a nuclear encoded co-dominant locus, the expected segregation among the F2 progeny is a 1:2:1 ration of plants homozygous for the *C. grandiflora* allele, heterozygous plants and plants homozygous for the *C. rubella* allele.

3-1-1-1 RFLP markers

Any DNA fragment can be used as RFLP marker as long as it reveals a restriction site polymorphism between the DNAs of the two parents of a mapping population. *Arabidopsis* and *Capsella* are closely related species, thus *Arabidopsis* DNA sequences readily cross-hybridise with *Capsella* DNA (Schmidt *et al*. 1999; Acarkan 2000; Acarkan *et al*. 2000; Clarenz 2000; Mbulu 2000). Different sources of RFLP markers were used in this study. In total, 55 *Arabidopsis* RFLP markers (mi… markers; Liu *et al*. 1996), eight *Arabidopsis* EST clones (*Arabidopsis* Biological Ressource Centre; http://aims.cps.msu.edu/aims) and one *Capsella* genomic DNA fragment derived from a cosmid clone were used for a polymorphism survey. The mi… markers are *Pst*I fragments of *Arabidopsis* genomic DNA and most of them represent low copy sequences.

3-1-1-1-1 Polymorphism survey

For the polymorphism survey, DNA of *C. rubella* has been analysed alongside a pool of DNAs of 50 F2 progeny plants. This strategy was taken, since *C. grandiflora* is self-incompatible. Thus only limited amount of DNA of the *C. grandiflora* plant used as a parent for the mapping population was available. The polymorphism survey allows to compare the RFLP pattern of the homozygous *C. rubella* plants with the pattern revealed by all F2 individuals which corresponds to the heterozygous condition. Hence, any additional fragment, which is detected by a marker in the DNA of the pool of F2 plants compared to DNA of the *C. rubella* plants, most likely corresponds to a *C. grandiflora* specific allele.

Several enzymes were used for the polymorphism survey. Genomic DNA from *C. rubella* and the pool of the 50 F2 plants has been digested with the following restriction enzymes: *Bgl*II, *Dra*I, *Eco*RI, *Eco*RV, *Hin*dIII, *Xba*I and *Xho*I. The resulting Southern blots were analysed in hybridisation experiments with RFLP markers as probes. The frequency, with which polymorphisms could be detected with each of the restriction enzymes used in respect to the number of markers tested, is presented in Table A.

Enzyme used	*Markers tested*	*Markers revealing a polymorphism*	*Frequency of polymorphism*
*Bgl*II	23	9	39%
*Dra*I	32	19	59%
*Eco*RI	55	24	44%
*Eco*RV	56	34	61%
*Hin*dIII	25	8	32%
*Xba*I	52	26	50%
*Xho*I	8	2	25%

Table A: Represented above is the frequency of polymorphism for the restriction enzymes used.

From the set of 63 *A. thaliana* markers tested, only one marker (mi423a) did not show any hybridisation to *Capsella* genomic DNA. The enzymes which have been used most frequently for the polymorphism survey were *Eco*RI, *Eco*RV and *Xba*I. Using these enzymes RFLPs were readily detected. Between 44% and 61% of the markers tested showed a polymorphism. Likewise, *Dra*I reveals polymorphisms for a high percentage of markers (Table A). Figure 2 illustrates two examples of RFLP marker hybridisations.

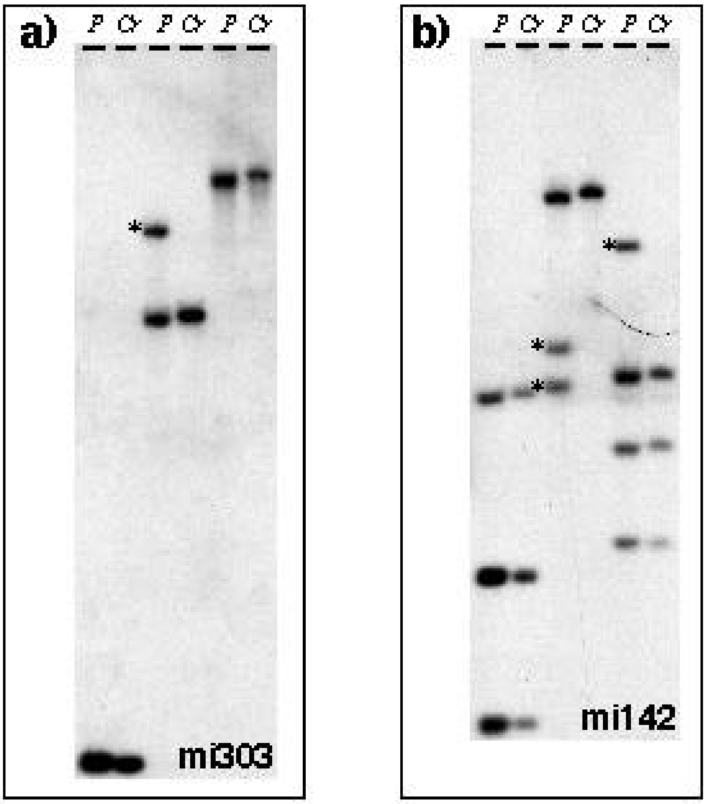


Figure 2: Depicted above are two examples of a polymorphism survey analysis. These blots are showing the results of hybridisation experiments. Genomic DNA of a pool of F2 plants (P) and *C. rubella* DNA (Cr) was digested with *Eco*RI, *Eco*RV or *Dra*I, respectively. Blot **a** shows the result of a hybridisation experiment using mi303 as probe, whereas blot **b** was hybridised with marker mi142. Asterisks indicate polymorphic fragments.

Out of the 62 *A. thaliana* markers hybridising to *Capsella* DNA, six mi… markers (9,5%) only revealed monomorphic patterns although five or six different enzymes have been used for the RFLP analysis (*Bgl*II, *Dra*I, *Eco*RI, *Eco*RV, *Hin*dIII and *Xba*I). Thus, the polymorphism survey identified 56 *A. thaliana* markers and one *Capsella* marker suitable for genetic mapping experiments with the *Capsella* mapping population.

3-1-1-1-2 Analysis of RFLP marker segregation in *Capsella*

From 57 RFLP markers identified in the polymorphism survey analysis 45 markers have been chosen for genetic mapping experiments in *Capsella*. For 41 markers an unambiguous co-dominant inheritance could be scored. Among these 41 markers five revealed two loci each (mi320, mi358, mi335, mi74, and IG3). Thus, 46 loci in total could be assigned to eight linkage groups of *Capsella* (Figure 6).

It has not been attempted to map all loci corresponding to a particular marker. Duplicate loci have been scored when they were revealed in the same experiment. Interestingly, four of the five duplicate loci studied here were revealed by restriction of genomic DNA with *Dra*I.

Figure 3 is illustrating the case of *A. thaliana* marker mi358, which revealed two loci when hybridised to *Capsella* genomic DNA. The segregation of this marker as well as the complete data set for all RFLP markers is listed in Appendix.

It is frequently observed that a marker hybridises to two loci with different intensities. By convention the stronger signal obtained is called locus a and the weaker one is referred to as locus b of a particular marker. This distinction is important in the context of comparative mapping experiments. Duplicate loci could indicate deviations from collinearity unless the loci are compared to their orthologous counterparts in the other species.

Marker IG3 is corresponding to rDNA sequences in *Arabidopsis*. Consistent with the repetitive nature revealed on the *Capsella* survey blot, a complex pattern of hybridising loci was obtained when the mapping population was analysed. Nevertheless, two segregating loci could be discerned. IG3-a could be evaluated as a co-dominant locus whereas IG3-b exhibited a dominant segregation of a *C. grandiflora* specific fragment. Therefore, no distinction could be made for this locus between the genotypes of plants which are homozygous for the *C. grandiflora* allele or heterozygous. Nevertheless, IG3b could be assigned to a linkage group F in *Capsella* (Figures 4 and 6).

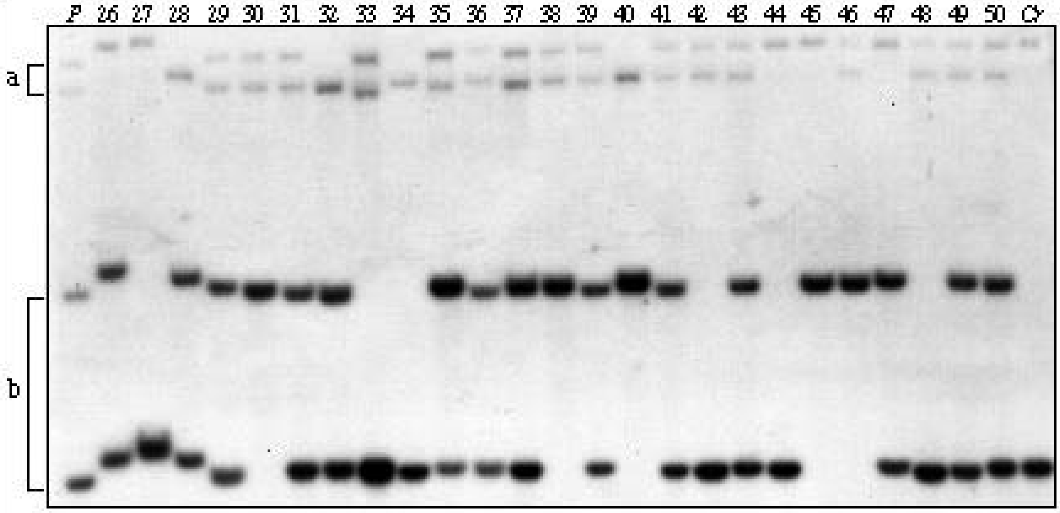


Figure 3: Illustrated above is the result of a hybridisation experiment. Genomic DNA of a pool of F2 plants (P), DNAs of F2 plants 26-50 and *C. rubella* DNA (Cr) were digested with *Eco*RV. The resulting Southern blot was hybridised with marker mi358 and two loci with a different segregation pattern were revealed. These loci could be mapped onto two *Capsella* linkage groups. The fragments of the two loci are designated a and b, respectively.

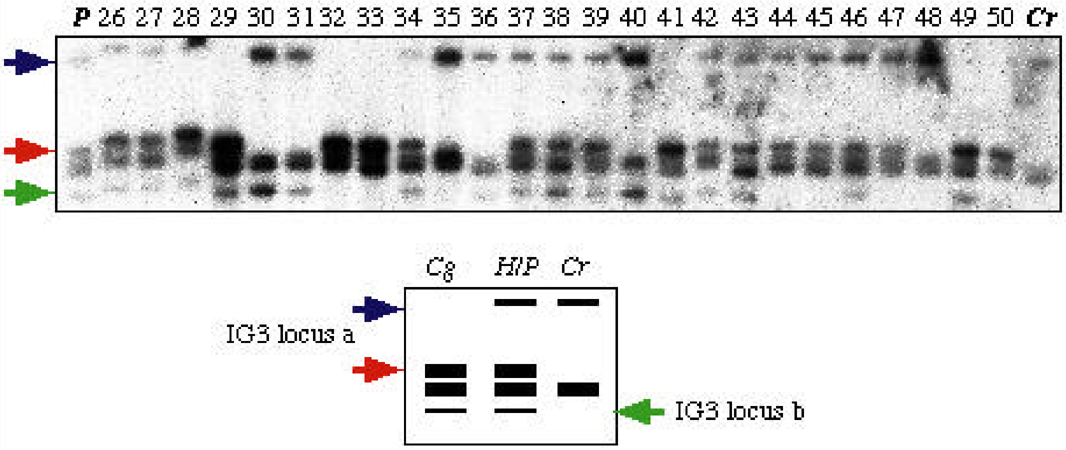

P 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 Cr
Cg
H/P
Cr
IG3 locus a
IG3 locus b

Figure 4: The result of a Southern blot hybridisation with marker IG3 is shown. Genomic DNA of a pool of F2 plants (P), DNAs of F2 plants 26-50 and *C. rubella* DNA (Cr) were digested with *Dra*I. Marker IG3 revealed two polymorphic loci, a and b. A schematic representation of the pattern is shown. Locus IG3a could be scored as a co-dominant marker. The *C. grandiflora* specific fragment is marked with a red arrow, whereas the fragment specific for *C. rubella* is highlighted with a blue arrow. Two monomorphic fragments are observed. The IG3b locus is specific to *C. grandiflora* DNA (green arrow), the same fragment is present in the pool of genomic DNA of the F2 plants. *C. rubella* does not reveal a corresponding fragment. Nevertheless, locus IG3b could be assigned to a linkage group.

For all markers used for genetic mapping in *Capsella*, sequence information has been obtained (data not shown). Hence, they could be placed on the sequence maps of the *Arabidopsis* chromosomes in an unambiguous way (Figure 7).

3-1-1-2 SSCP markers

3-1-1-2-1 *Capsella* sequences as a source for SSCP markers

A library of clones containing small inserts of *C. rubella* total DNA was established. This library has been constructed to compare the sequence repertoire of the *Arabidopsis* and *Capsella* genomes. Moreover the *Capsella* sequences served as a source for the generation of single strand conformation polymorphism (SSCP) markers. SSCP analysis is based on different mobilities of denatured DNA strands in MDE™ gels. Even single nucleotide differences between DNA strands may be detected with this method. In order to obtain scorable polymorphisms, the DNA fragments analysed should not be longer than 300 or 400 bp.

a- Choice of the enzyme used

Genomic *C. rubella* DNA has been digested with enzymes containing a 4 bp recognition site (*Alu*I, *Hae*III, *Mbo*I, *Rsa*I and *Taq*I). After gel electrophoresis, the size range containing the majority of the generated fragments was evaluated. Table B summarises the results of this analysis.

Restriction enzyme	*Recognition site*	*Majority of the fragments have a size of*
*Alu*I	AGCT	250-1300 bp
*Hae*III	GGCC	750-3000 bp
*Mbo*I	GATC	250-1300 bp
*Rsa*I	GTAC	400-2000 bp
*Taq*I	TCGA	250-2000 bp

Table B: This table summarises the average fragment sizes which are obtained if total genomic *C. rubella* DNA is digested with *Alu*I, *Hae*III, *Mbo*I, *Rsa*I and *Taq*I, respectively. The recognition sequences of the different enzymes are listed in the table.

The largest portion of fragments generated with *Hae*III was 2000 bp long, and consequently too large for the purpose of SSCPs. *Alu*I, *Rsa*I and *Taq*I could have been chosen, but the *Capsella* clone library has been established using a complete digestion of *Capsella* genomic DNA with *Mbo*I. The *Mbo*I fragments can be ligated into vectors with a *Bam*HI cloning site.

Clones putatively containing a *Capsella* DNA insert were analysed by PCR. All clones yielding an amplification product were sequenced. In total 148 sequences of *Mbo*I-fragments were obtained. One sequence was identical to cloning vector sequences and another represented *E. coli* DNA sequences. After removing redundant sequences 132 sequences remained. Due to the availability of the complete genome sequence of *Arabidopsis*, it can be reliably estimated, how many of the *C. rubella* sequences show homology to *A. thaliana* sequences (The *Arabidopsis* genome initiative 2000).

All *Capsella rubella* sequences were subjected to a BLAST analysis (Altschul *et al.* 1997) to determine their homology to *Arabidopsis* nuclear and organellar sequences. Based on a threshold value of e^{-10} the sequences could be classified in two groups as follows: 102 sequences were homologous to *A. thaliana* sequences and 30 sequences seemed to be specific for the *C. rubella* genome. It was analysed whether the *A. thaliana* sequences showing homology to *Capsella* sequences are correponding to different classes of repetitive sequences or whether they represent low copy sequences.

For the corresponding *Arabidopsis* sequences it was also tested whether they show homology to cognate EST or cDNA sequences. The analysis is summarised in Table C.

Sequences with homology to Arabidopsis sequences							*Capsella specific sequences*
Repetitive sequences					*Low copy sequences*		
rDNA sequences		*Organellar sequences*		*Retro-trans-posons*	*Sequences with homology to EST or cDNA sequences*	*Sequences without homology to EST or cDNA sequences*	
18,5-25S	*5S*	*Chloro-plast DNA*	*Mitochon-drial DNA*				
9 6,8%	1 0,8%	24 18,2%	4 3,0%	2 1,5%	26 19,7%	36 27,3%	30 22,7%

Table C: Summary of the sequence homology of the *C. rubella Mbo*I-sub-clone sequences in respect to *A. thaliana* sequences. Numbers of clones for each category are given as well as the percentage of sequences in each category in respect to the total number of 132 sequences analysed.

In total, 77,3% of the *C. rubella* sequences were found to be homologous to *A. thaliana* sequences, a large proportion of them being of repetitive nature, such as organellar sequences. Among the 10 fragments matching rDNA sequences, one (*Mbo*-M15) showed homology to *Arabidopsis* 5S rDNA sequences (GenBank acc. no. M65137) and the other 9 exhibited homology to 18S-5,8S-25S rDNA arrays constituting the NORs in *Arabidopsis* (GenBank acc. no. X52322).

It was analysed, whether *Mbo*-M15 could be used as an RFLP marker, but the polymorphism survey analysis revealed that the enzymes used (*Dra*I, *Eco*RI, *Eco*RV, *Xba*I) do not cut the arrays of 5S rDNA sequences in the *Capsella* genome (data not shown). Sixty percent of *Capsella* sequences corresponding to repetitive sequences are homologous to *Arabidopsis* chloroplast DNA (GenBank acc. no. AP000423) whereas only 3% correspond to sequences of *Arabidopsis* mitochondrial genome (GenBank acc. no. Y08501, Y08502).

One fragment of 484 bp was found to be homologous to over 20 different sequences in the *A. thaliana* genome. This insert, *Mbo*-D22, is corresponding to an *A. thaliana* retroelement-like sequences. The analysis of this sub-clone is described in more detail in Chapter 3-3.

Almost 50% of the *Mbo*-sequences were shown to be homologous to low copy sequences in *A. thaliana*. Cognate ESTs or cDNA sequences could be identified for 42% of the *Arabidopsis* low copy sequences, which were corresponding to *Capsella* *Mbo*-sequences.

b- Mapping data

Only *Capsella* fragments demonstrating homology to *A. thaliana* low copy sequences have been considered for mapping studies. The corresponding *Arabidopsis* sequences were then located on the chromosome sequence maps. The *Arabidopsis* RFLP markers chosen for genetic mapping in *Capsella* were well distributed over the *Arabidopsis* genome. Nevertheless, some regions of the *Arabidopsis* genome remained under-represented. *Mbo*-fragments showing homology to those regions of the *A. thaliana* genome were chosen for SSCP marker analysis.

Primer pairs were deduced on the sequences of 23 different *Mbo*-fragments (*Mbo*-…). These pairs of oligonucleotides were used for PCR experiments on *C. rubella* and *C. grandiflora* DNA. The resulting products were denatured and separated on MDE^{TM} gels to detect polymorphisms. Eleven fragments were polymorphic, one of which showed a length polymorphism after the PCR amplification and did not require the MDE^{TM} gel analysis (*Mbo*-Cr8-PCR). For nine fragments, a co-dominant polymorphism could be discerned, whereas for one marker the polymorphism was not distinct enough to allow reliable scoring (*Mbo*-N18). Eight sequences were monomorphic, two yielded amplification products only on *C. rubella* template DNA (*Mbo*-Cr1, *Mbo*-N24) and two did not yield distinct patterns (*Mbo*-C14, *Mbo*-D22). This coincides with the fact that both clones correspond to repetitive DNA sequences, *Mbo*-C14 has homology to rDNA sequences and *Mbo*-D22 to retroelement-like sequences. This readily explains the amplification of multiple fragments. An example for a SSCP analysis is shown in Figure 5.

For six of the monomorphic fragments, it was tested whether it was possible to discern a polymorphism after digestion of the PCR products and separation of the resulting fragments (CAPS, cleaved amplified polymorphic sequence) on MDE^{TM} gels. Since sequence information is available for all clones, a restriction map can be generated and restriction enzymes can be chosen accordingly. Between two and three enzymes have been tested for each of the inserts, although 13 different digestions have been tested for these six markers only two loci could be mapped (*Mbo*-N18/*Hin*dIII-*Cla*I and *Mbo*-L19/*Hin*dIII).

In total, 23 *Mbo*-fragments have been chosen for SSCP analysis and 12 loci could be integrated into the *Capsella* molecular marker map (52%). Among these 12 *Mbo*-fragments which could be placed on the map, eight had homology to *Arabidopsis* EST sequences.

3-1-1-2-2 Other SSCP sources

a- End-sequences of cosmid inserts isolated by inverse PCR

Any DNA fragment for which sequence information is accessible can be mapped with the SSCP technique. Nine DNA fragments were isolated by inverse PCR (iPCR) from *C. rubella* cosmid clones (cos…) and sequenced. From this sequence information, pairs of primer sequences were deduced and tested for amplification on the parents of the *Capsella* mapping population. For polymorphic fragments a segregation analysis was undertaken. One fragment (cos2) could be mapped directly after the PCR amplification due to a length polymorphism and another fragment (cos57) required the separation on a MDE^{TM} gel. Two additional fragments could be mapped after digestion (cos9-*Hin*dIII/*Bam*HI and cos36-*Rsa*I). Thus, four fragments isolated from *C. rubella* cosmid clones could be placed on the *Capsella* linkage map.

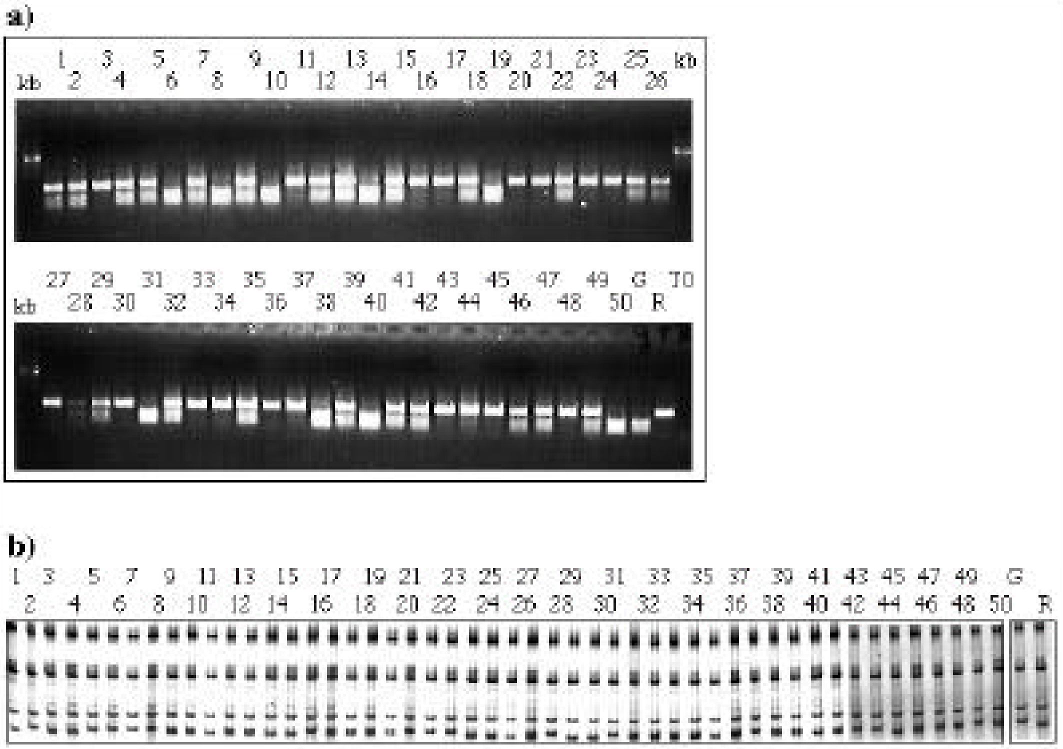

a)
kb
1 3 5 7 9 11 13 15 17 19 21 23 25
2 4 6 8 10 12 14 16 18 20 22 24 26
kb
27 29 31 33 35 37 39 41 43 45 47 49 G T0
28 30 32 34 36 38 40 42 44 46 48 50 R
kb
b)
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 G
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 R

Figure 5. The figure shows two examples for segregation analysis using PCR-based marker systems. The segregation illustrated in part a was obtained with CAPS marker T20G20. Primer sequences corresponding to an *Arabidopsis* gene prediction were used to amplify fragments on *Capsella* genomic DNAs. The DNA of F2 progeny plants 1-50 was analysed on an agarose gel alongside the DNA of *C. grandiflora* (G) and *C. rubella* (R). The resulting PCR products were digested with *Dra*I to reveal a polymorphism between *C. grandiflora* and *C. rubella*.

Picture b shows the segregation analysis obtained with SSCP marker *Mbo-A6*. Primer sequences corresponding to *C. rubella* genomic DNA fragment *Mbo-A6* were used for PCR. PCR reactions were performed on DNA of F2 progeny plants 1-50 as well as on DNA of *C. grandiflora* (G) and *C. rubella* (R). The resulting PCR fragments were denatured and the strands were separated on a MDE^{TM} gel.

b- Centromeric and telomeric regions of the A. thaliana genome as a target for SSCP markers

Eight genes located in the telomeric or centromeric regions of the *A. thaliana* chromosomes have been chosen for mapping studies. A strategy was chosen that should to pinpoint a gene or a predicted gene on the outermost sequenced and annotated BAC clones of a particular chromosome. Pairs of primers were deduced from the *Arabidopsis* gene sequences and PCR experiments were performed on *A. thaliana* and *C. rubella* DNA.

Arabidopsis chromosome	*BAC clone*	*Accession number*	*Size of PCR product (bp)*		*Mapping technique*	*Capsella map*
			A. thaliana	*C. rubella*		
I	T25K16	AC007323	1376	~ 1300	SSCP/dom	—
II	F23H14	AC006837	876	~ 700	SSCP/dom	—
II	T9J23	AC005309	803	~ 550	SSCP	LG C
II	T20G20	AC006220	812	~ 600	CAPS	LG D
III	MAA21	AL163818	1579	~ 1600	SSCP/dom	—
IV	F17A8	AL49482	1432	~ 1400	—	—
IV	F16J13	AL49638	1979	—	—	—
V	F7J8	AL137189	1469	~ 1400	SSCP†	LG G

Table D: Summary of BAC clones chosen due to their telomeric or centromeric location in the *A. thaliana* genome for genetic mapping in *Capsella*. The size of the PCR product which could be amplified from a gene or predicted gene located on a particular BAC clone is given. PCR products of *Capsella* were not sequenced, the fragment sizes have been estimated after gel electrophoresis. The cross indicates that the PCR fragment has been digested before it was analysed on MDE^{TM} gels and -dom- is identifying a dominant marker. In the last column the *Capsella* linkage groups (LG) are listed to which markers could be added.

Table D summarises the chromosome locations for the chosen BAC clones, which carry the genes or predicted genes used in this study. If possible, the complete gene or predicted gene was amplified and the sizes of the PCR products for *A. thaliana* and *C. rubella* are given in bp. Cr-T20G20 could be mapped as a CAPS marker, since digestion of the PCR products with *Dra*I and electrophoretic separation of the fragments on an agarose gel revealed a polymorphism between *C. grandiflora* and *C. rubella* (Figure 5). As no other CAPS polymorphism was obtained by amplification and digestion of the PCR products spanning the predicted genes, additional primer sequences were deduced from the *Arabidopsis* sequences in order to generate fragments of approximately 300 bp. It was tested whether the primer sequences were suitable to amplify the corresponding *Capsella* sequences. Resulting PCR products were then analysed for the presence of SSCPs. Two loci could be added to the map, Cr-T9J23 and

Cr-F7J8, located on two different linkage groups. Three primer combinations yielded an amplification product for only one of the two parents and one primer combination (F16J13) failed to amplify *Capsella* genomic DNA altogether.

3-1-1-3 *Capsella* map

3-1-1-3-1 Parameters

The *Capsella* genetic linkage map has been constructed using the MAPMAKER software (version 3.0; Lander *et al.* 1987) with a LOD score of 4.0 and a linkage group breakpoint at 50,0 cM. Haldane's mapping function has been chosen to convert the recombination frequency into genetic map units (centiMorgan, cM).

According to Mendel's laws, a 1:2:1 segregation is expected for the inheritance of a co-dominant marker encoded by the nuclear genome in a F2 population. Accordingly, *C. grandiflora* and *C. rubella* alleles should be present in a 1:1 ratio. The segregation data as well as the allele frequency data for all markers have been subjected to χ^2 tests and a significance threshold of 0,05 was used to recognise distorted segregation ratios. The Appendix lists the χ^2 test table for all markers constituting the *Capsella* map.

3-1-1-3-2 *Capsella* genetic linkage map

The set of data obtained in this study has then been merged with data previously established in the laboratory (Acarkan 2000; Acarkan *et al*. 2000; Clarenz 2000; Mbulu 2000). The map consists of 137 loci and spans 650,5 cM. Locus IG3-b has not been included, due to the dominant inheritance of this locus. The results for each linkage group are summarised in Table E.

The *Capsella* genetic map is illustrated in Figure 6. All linkage groups have a similar marker density between 4 and 5,7 cM/marker, apart from linkage group F which benefited of a special study (Acarkan *et al*. 2000). The sizes of the linkage groups range from 56,9 cM for the smallest (LG E) to 108,4 cM for the largest linkage group (LG G). The χ^2 test permitted to demonstrate four regions showing distorted allele distribution, in three of them *C. grandiflora* alleles are over-represented whereas in the other one *C. rubella* alleles are more abundant than expected (Figure 6 and Appendix). A segment showing significant distortion in favour of *C. grandiflora* alleles is present on LG G between markers mi174 and N97271. This region spans 37,3 cM and represents 35,4% of this linkage group. In all other cases the distortion of allele frequencies is restricted to one or two closely linked markers (mi19 and mi208 on LG A; m457A on LG E; FKBP15_1 and FKBP61/3 on LG D).

Capsella linkage groups	*Size (cM)*	*Total number of loci*		*Average distance between loci on a particular LG*	*Arabidopsis collinear chromosomes*
		RFLP markers	*PCR-based markers*		
A	86,4	14	3	5 cM	I
B	96,7	17	0	5,7 cM	I
C	83,2	9	3	7 cM	II
D	63,7	14	1	4,2 cM	II/III
E	56,9	11	3	4 cM	III
F	75,8	23	5	2,7 cM	IV/V
G	108,4	19	1	5,4 cM	IV/V
H	79,4	11	3	5,7 cM	V
Total	650,5	118	19	4,7 cM	

Table E : This table lists the results of the complete set of data provided by Acarkan (2000), Clarenz (2000), Mbulu (2000) and markers added from this study. Eight *Capsella* linkage groups are depicted with their respective size in cM. The number of loci for each linkage group is given, it is furthermore listed whether the markers have been mapped by RFLP analysis or PCR-based methods. The marker density of each linkage group has been calculated by dividing the complete size of the linkage group by the number of loci mapped on this linkage group. Syntenic chromosomes of *A. thaliana* are indicated for each of the eight *Capsella* linkage groups.

3-1-1-4 Duplicated loci

The copy number of RFLP markers mapped in this study has been estimated for the *Capsella* genome from the results of the hybridisation experiments. Since sequence information is available for all *Arabidopsis* markers used, it was analysed whether the marker was present as a single-copy sequence in the genome or whether additional sequences were found. The results are summarised in Table F.

	Capsella rubella		*Arabidopsis thaliana*	
One locus	22	53,7%	23	56,1%
Two loci	15	36,5%	13	31,7%
> two loci	4	9,8%	5	12,2
Total	41	100%	41	100%

Table F: All RFLP markers (41 markers), which have been mapped in this study, were analysed for the putative number of loci in *A. thaliana* and *C. rubella*.

The vast majority of markers appear to be at a single locus. One third of the RFLP markers which have been used in this study exhibited a fragment pattern consistent with two loci. In general, numbers of loci revealed by different markers are very similar between *A. thaliana* and *C. rubella*.

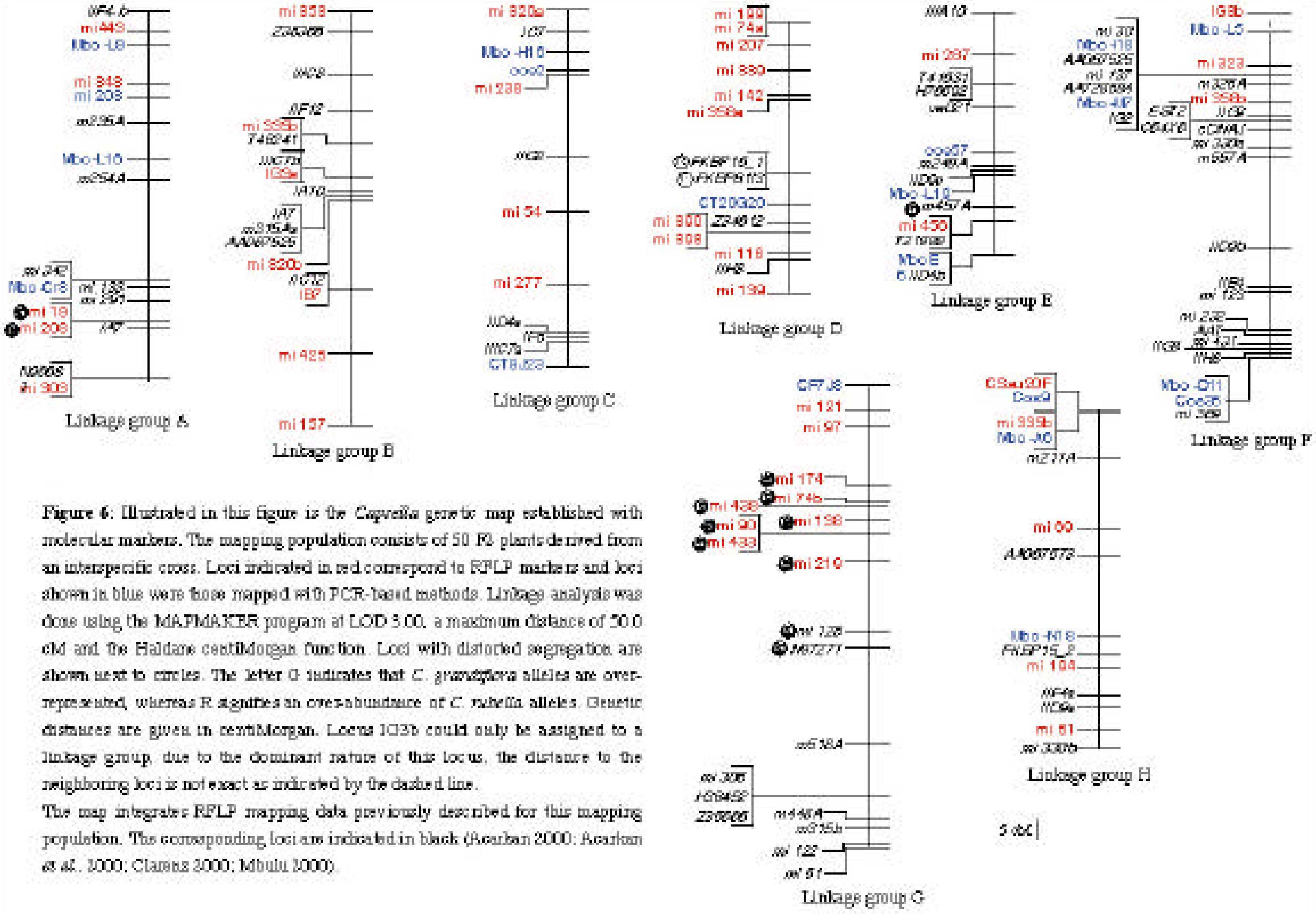


Figure 6: Illustrated in this figure is the *Capsella* genetic map established with molecular markers. The mapping population consists of 50 F2 plants derived from an interspecific cross. Loci indicated in red correspond to RFLP markers and loci shown in blue were those mapped with PCR-based methods. Linkage analysis was done using the MAPMAKER program at LOD 3.00, a maximum distance of 50.0 cM and the Haldane centiMorgan function. Loci with distorted segregation are shown next to circles. The letter G indicates that *C. grandiflora* alleles are over-represented, whereas R signifies an over-abundance of *C. rubella* alleles. Genetic distances are given in centiMorgan. Locus IG3b could only be assigned to a linkage group, due to the dominant nature of this locus, the distance to the neighboring loci is not exact as indicated by the dashed line.

The map integrates RFLP mapping data previously described for this mapping population. The corresponding loci are indicated in black (Acarkan 2000; Acarkan et al., 2000; Clamas 2000; Mbulu 2000).

3-1-2 *Arabidopsis thaliana*

3-1-2-1 *Arabidopsis* sequence map

For the purpose of a comparative mapping study, the *A. thaliana* sequence map was used. For all markers, sequence information has been established. Sequences of the RFLP markers have been aligned with the *A. thaliana* genome sequence. Since the *A. thaliana* genome sequence is accomplished, it is possible to assemble each chromosome as two large sequences representing the arms of the chromosome. The position of all markers used in this study were located on these sequence maps of the *Arabidopsis* chromosomes (http://mips.gsf.de/proj/thal/db/). For *Capsella* SSCP markers the homologous *A. thaliana* sequences are indicated on the map. The map unit on the sequence map is given in Mbp. Some *Capsella* SSCP markers showed homology to different locations in *A. thaliana*. The highest BLAST score and the syntenic flanking markers allowed the choice of the *A. thaliana* locus most likely homeologous to the *Capsella* locus. Regions which have not been sequenced (centromeres and NORs) are indicated by a rupture of the chromosome on the *A. thaliana* sequence map (Figure 7).

3-1-2-2 Comparative mapping between *Arabidopsis* and *Capsella*

All markers could be placed into eight *Capsella* linkage groups. As described in Table F, each *Capsella* linkage group is forming large collinear segments with *A. thaliana* chromosomes. Chromosome I of *A. thaliana* is collinear with *Capsella* LGs A and B, chromosome II is collinear to LGs C and D, chromosome III is homeologous to LGs D and E, chromosome IV is equivalent to LGs F and G, finally, chromosome V corresponds to LGs F, G and H. On average, two *Capsella* linkage groups are found to cover one *A. thaliana* chromosome. The centromeres, clearly localised on *A. thaliana* do not correspond to the breakpoints of collinearity. For chromosome I, markers adjacent to the centromeric regions are found collinear with LG A, the chromosome II centromere does not disturb collinearity with LG D. The centromeric segment of chromosome III is collinear with LG E. The markers located in the centromeric region of chromosome IV are syntenic to LG G, however, a large inversion interrupts collinearity. Interestingly, the centromeric region of chromosome V is collinear with LG F, whereas the remainder of chromosome V is corresponding to LGs G and H.

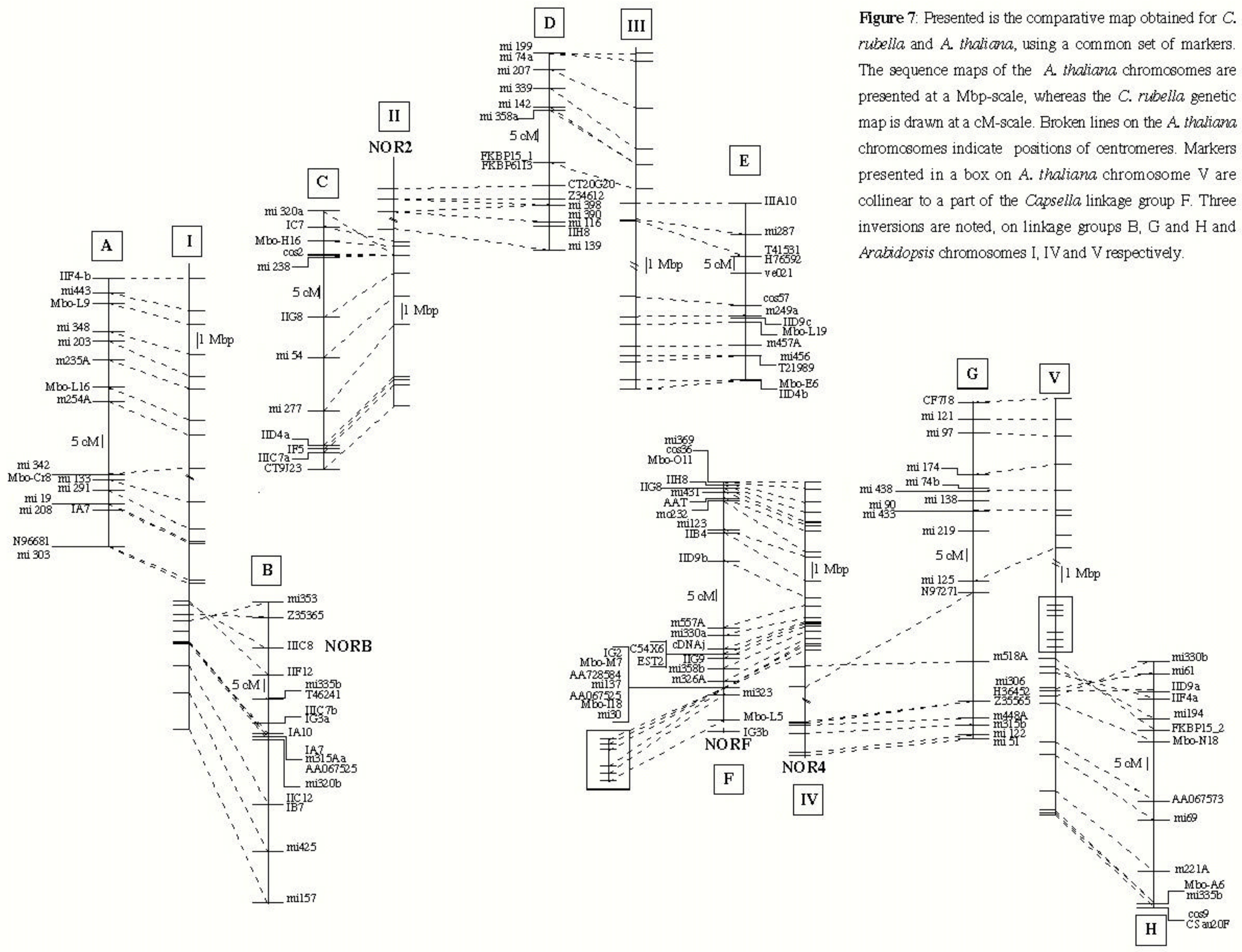


Figure 7: Presented is the comparative map obtained for *C. rubella* and *A. thaliana*, using a common set of markers. The sequence maps of the *A. thaliana* chromosomes are presented at a Mbp-scale, whereas the *C. rubella* genetic map is drawn at a cM-scale. Broken lines on the *A. thaliana* chromosomes indicate positions of centromeres. Markers presented in a box on *A. thaliana* chromosome V are collinear to a part of the *Capsella* linkage group F. Three inversions are noted, on linkage groups B, G and H and *Arabidopsis* chromosomes I, IV and V respectively.

In total, three inversions are noted, between LG B and chromosome I, between LG G and chromosome IV, and between LG H and chromosome V (Figure 7).

3-2 MICROCOLLINEARITY

Genome collinearity at the level of the genes was investigated for three species belonging to the family of the Brassicaceae, *A. thaliana*, *C. rubella* and *B. oleracea*. For this analysis, a region of *A. thaliana* chromosome IV located between RFLP markers g8300 and mi431 was chosen. This study aims to analyse the *Arabidopsis* region in respect to gene repertoire and to identify and characterise the corresponding regions from the *C. rubella* and *B. oleracea* genomes.

3-2-1 *Capsella rubella* cosmid contig

A 20 kbp region located on the long arm of chromosome 4 of *A. thaliana* between RFLP markers g8300 and mi431 was used for a BLAST analysis with *Arabidopsis* EST sequences. This revealed several ESTs with high sequence identity to the genomic DNA sequence. Five ESTs (EST 1 – 104G24T7; EST 2 – 140O12T7; EST 3 – 90J24T7; EST 4 – 192P5T7; EST 5 – 79G7T7) were chosen for sequence analysis and it could be confirmed that they represent cognate cDNA sequences for genes in the *Arabidopsis* region (data not shown). This analysis showed that EST 1 and EST 2 are partially overlapping (Figure 8). EST 4 corresponds to the aspartate amino-transferase gene (asp5 gene, GenBank acc. no. X91865).

EST 4 (AAT) has been used for genetic mapping experiments with the *Capsella* population. It maps to linkage group G of the *C. rubella* map. This segment is orthologous to a part of the long arm of chromosome 4 of *A. thaliana* (chapter 3-1).

The five ESTs have been used as probes to screen the *C. rubella* genomic DNA cosmid libraries to identify corresponding *Capsella* sequences. Twenty hybridising colonies were detected, 17 in the *C. rubella Sau*3AI library and three in the *C. rubella Taq*I library. DNA of all cosmids has been prepared, digested using different enzyme combinations and blotted. The resulting membranes have been hybridised with the different ESTs and based on these results the cosmid clones were arranged into contigs. Two *C. rubella* cosmids, CS51 and CT8, have been chosen for sequence analysis.

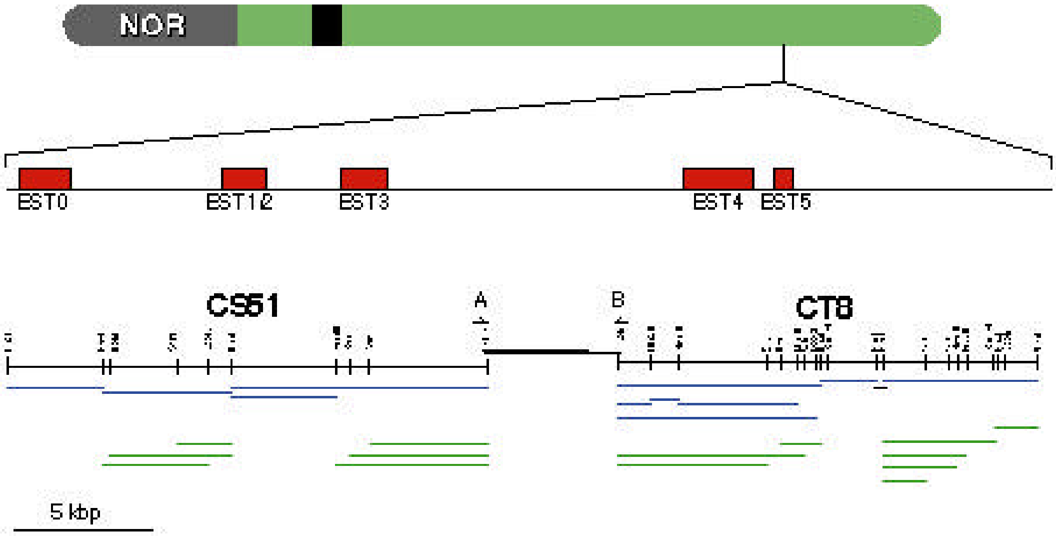


Figure 8: Shown above is a representation of the *Capsella* sequence contig and the corresponding region of *Arabidopsis* chromosome 4. At least five different genes, shown as red boxes, are present in the *Arabidopsis* region as cognate cDNA clones and EST sequences (EST0-5) and could be located on this segment. Genomic DNA inserts of two *Capsella* cosmid clones, CS51 and CT8, are harbouring sequences homologous to ESTs 0-5. The sequences of the cosmid DNA inserts had to be connected with two PCR fragments to yield a contig. Restriction sites of enzymes used for sub-cloning are indicated on the cosmid sequences shown in black. The sub-clones are drawn in blue while the deletion clones are indicated in green. The sequenced region is spanning over 37 kbp. The sequenced regions are drawn to scale.

Cosmid CS51 carries sequences homologous to ESTs 1/2 and 3, and the CT8 cosmid has homology to ESTs 4 and 5. Cosmid CS51 has been sub-cloned with *Xho*I, which gave 3 fragments, one of which corresponds to a genomic DNA/vector border fragment. The CT8 cosmid has been sub-cloned independently with *Eco*RI, *Hin*dIII and *Xba*I. Several sub-clones could be identified, but the genomic DNA fragments adjacent to the vector sequences were not obtained. Large fragments were further sub-cloned, the resulting clones are referred to as deletion clones. This increased the efficiency of sequencing by providing additional anchor points to deduce primers for sequencing (Figure 8). PCR experiments were used to analyse whether sub-clone sequences are directly adjacent to each other or whether a small genomic DNA fragment was residing between them. In the case of cosmid CT8, two sub-clones thought to be adjacent to each other were 272 bp apart due to two neighbouring *Xba*I sites, the PCR product spanning this region has been sequenced.

The insert sequences of cosmids CS51 and CT8 do not overlap. Therefore, PCR experiments were performed to establish fragments spanning the *C. rubella* genomic DNA sequences between the inserts of the cosmid clones. One PCR fragment could be amplified using primer -A- on CS51 and a primer specific for cosmid vector sequences. Another fragment could be amplified between primers A and B located on CS51 and CT8, respectively (Figure 8). The sizes of these PCR products are 3 kbp and 5 kbp, respectively. They could be cloned into the pGEMTeasy vector and sequenced.

Assembly of all sub-clone and PCR product sequences yielded a contig of 37,159 bp. The sequencing data have been analysed with the GCG package and homology searches to *A. thaliana* genomic DNA or EST sequences have been performed using the NCBI, TAIR and MIPS databases (chapter 2-2-7).

One of the *C. rubella* sub-clones showed sequence similarity to an *A. thaliana* gene located upstream of ESTs 1/2. This gene has been called EST 0 (GenBank acc. no. D43962).

3-2-2 *Brassica oleracea* cosmid contig

The region located on *A. thaliana* chromosome 4 has also been investigated in *B. oleracea*. Cosmid B21 harbours sequences homologous to ESTs 3-5, but no cosmid could be identified carrying homologues of ESTs 0 or 1/2. These ESTs have therefore been used to screen two *B. oleracea* BAC libraries (C. Ryder and G. King, unpublished;

O'Neill and Bancroft 2000). The BIBAC library (O'Neill and Bancroft 2000) has larger average insert sizes than the BAC library from HRI Wellesbourne (C. Ryder and G. King, unpublished). The BACs obtained from the hybridisation of the Wellesbourne library showed faint results when probed with a *Capsella* sub-clone corresponding to EST 0 and ESTs 1/2. Nevertheless, further analyses confirmed that these BAC clones were harbouring sequences homologous to EST 0 and EST 1/2. The BIBAC library has been first hybridised to ESTs 1/2 and some hybridising clones could be identified. The filters have been submitted to a second hybridisation experiment with EST 4 as probe. Only BAC clones which were hybridising in both experiments have been analysed further. An example of the colony hybridisation results is shown in Figure 9. Two independent colonies of nine hybridising BIBAC clones were prepared, digested with *Hin*dIII and blotted. Figure 10 shows the result of a hybridisation of a Southern blot carrying DNA of BIBAC clones. The membrane was probed with EST 4. Clones IB10, IB14 and IB16 showed a pattern clearly different from that of clones IB9, IB11, IB12, IB13, IB15, IB17 and IB18. Two different restriction fragment patterns were also observed for the other probes (ESTs 1-3 and EST 5). It was deduced that at least two loci in *B. oleracea* are corresponding to the AAT region of *A. thaliana*. These were mapped via the SSCP technique to chromosomes 1 and 7 of *B. oleracea* (data not shown). Two BAC contigs could be established for each of the two loci with IB12 representing the *B. oleracea* chromosome 1 locus and IB10 representing the locus on chromosome 7 (clones marked with an asterisk in Figure 10). Cosmid B21 has been determined to be part of the chromosome 1 locus and a second cosmid, B3, is corresponding to the region on chromosome 7.

BAC clone	*B. oleracea chromosome*	*Estimated insert size (kbp)*
B67	I	60
B85	I	90
IB12	I	130
B58	VII	60
B60	VII	120
B82	VII	70
IB10	VII	155

Table G: Listed above are insert size estimates for different BAC clones. The clones are grouped according to their chromosomal location in *B. oleracea*.

BAC DNA has been digested with *Not*I to free the genomic insert from the vector sequences. These digests have been analysed by pulsed field gel electrophoresis to

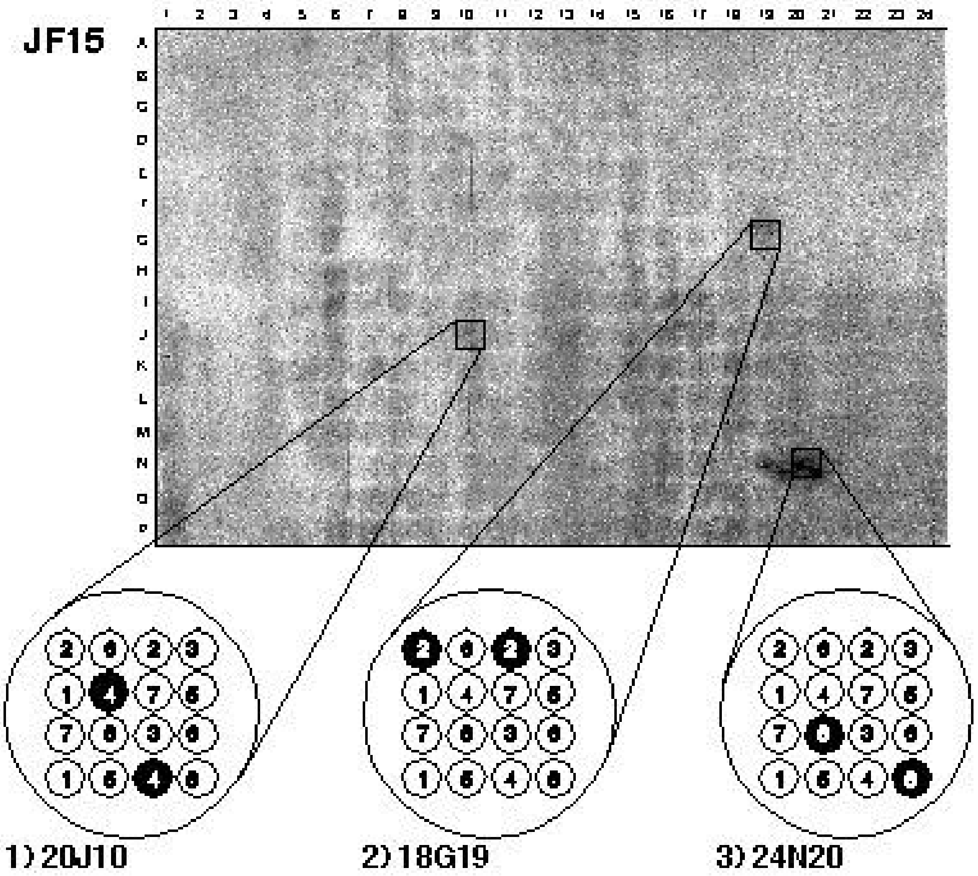


Figure 9 : Depicted above is the result of a colony hybridisation experiment. The JIC BIBAC library (O'Neill and Bancroft 2000) has been hybridised to ESTs 1/2. One filter carries BAC clones from 8 different micro-well plates, here indicated with numbers 1-8. Each plate has been gridded twice, thus providing an internal control for the hybridisation experiment. On filter JF15, plates 17-24 are grouped, the co-ordinates of hybridising clones are defined by the number of the plate in addition to the row (A-P) and column designations (1-24). For example, case **3** shows hybridisation of a clone (in black) from plate 8 of this pool (equals plate no. 24) with the co-ordinate N20. Therefore, the hybridising clone is identified as 24N20.

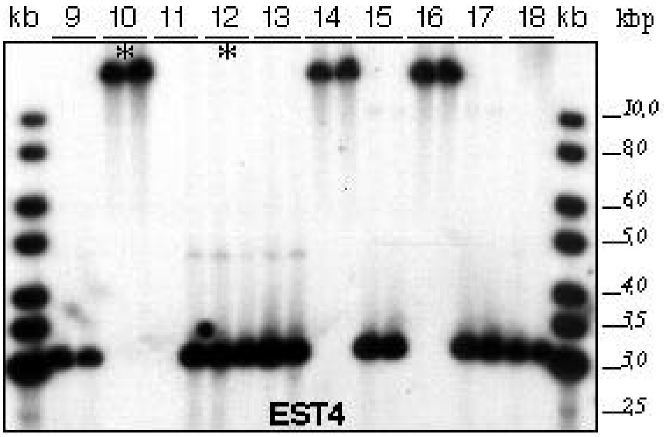


Figure 10: The figure shows the result of a Southern blot hybridisation of nine with *Hin*dIII digested BIBAC clones (O'Neill and Bancroft 2000) probed with EST4. The fragments have been separated on a 0,8% TAE agarose gel at 20 mA/30V/6h. Restriction fragments of different sizes hybridising to the EST4 probe are observed. Two BACs, one representative of each pattern have been chosen for further analyses (marked with asterisks), IB10 (clone 20J10) and IB12 (clone 24N20).

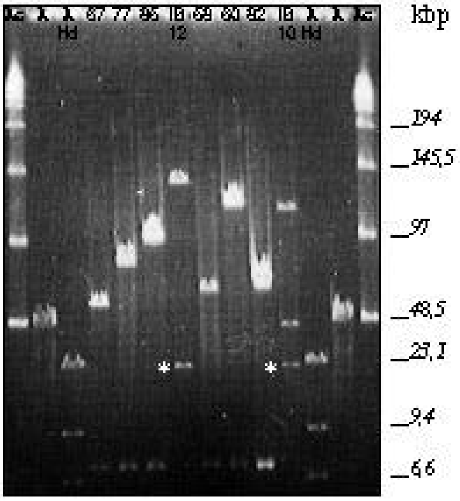


Figure 11: A PFGE experiment was carried out to estimate the sizes of *Brassica* genomic DNA inserts cloned in BACs. DNAs of all BAC clones have been digested with *Not*I to release the vector from the insert. The three size markers used are concatemers of λDNA (λc), λDNA (λ) and λDNA digested with *Hin*dIII (λHd). The sizes in kbp indicated by the markers are shown to the right. For the identification of clones, refer to Table G . Two asterisks indicate the pBIBAC 2 vector (23 kbp), whereas the 8 kbp fragment of pBeloBAC11 is seen in lanes carrying DNAs of BAC clones 67, 77, 85, 58, 60 and 82.

estimate the sizes of the genomic DNA inserts. The results of such an experiment are shown in Figure 11 and sizes of BACs, which were analysed in detail, are listed in Table G.

For the *B. oleracea* chromosome 1 locus, the insert of cosmid 21 has been completely sequenced. The genomic segment spans 22,424 bp. The genic regions have been sequenced on both strands. According to the hybridisation results this cosmid only harbours sequences homologous to ESTs 3-5. Therefore, *Hin*dIII sub-clones were generated from BAC clones spanning the complete region. The sub-clones were hybridised to ESTs 0 and 1/2 and a sub-clone harbouring homologues of ESTs 1/2 was obtained and sequenced (Figure 12).

For the chromosome 7 locus, BACs 58 and 82 as well as cosmid 3 were sub-cloned. Homologous sequences corresponding to ESTs 1-5 could be identified, but a sub-clone corresponding to EST 0 has not been found. Figure 12 is representing the organisation of the two homeologous *B. oleracea* regions and the different sequenced fragments.

Figure 13 shows a Southern blot analysis of the homeologous *B. oleracea* loci. A Southern blot carrying DNA of BAC clones 82 and 85 has been used for EST 0 and 1/2 hybridisations and a filter with DNA of BAC clones 58 and 67 has been probed with ESTs 3 and 4. Thus on each blot, a BAC clone from each *B. oleracea* locus is represented.

3-2-3 Microcollinearity analysis

3-2-3-1 Analysis of the *Arabidopsis* region

Over 115 Mbp of the *A. thaliana* genome have been sequenced, with the exception of the centromeres and rDNA loci. Most of the sequences available are annotated for the presence of genes and exon/intron structures have been predicted. Nevertheless, these data lack an experimental proof and predicted gene structures may not accurately reflect the actual coding sequences (The *Arabidopsis* genome initiative 2000). EST or cDNA sequences are powerful resources which unambiguously indicate the presence of a gene furthermore, the exon/intron structure of a gene is unveiled by alignment of EST or cDNA sequences relative to the genomic DNA sequence.

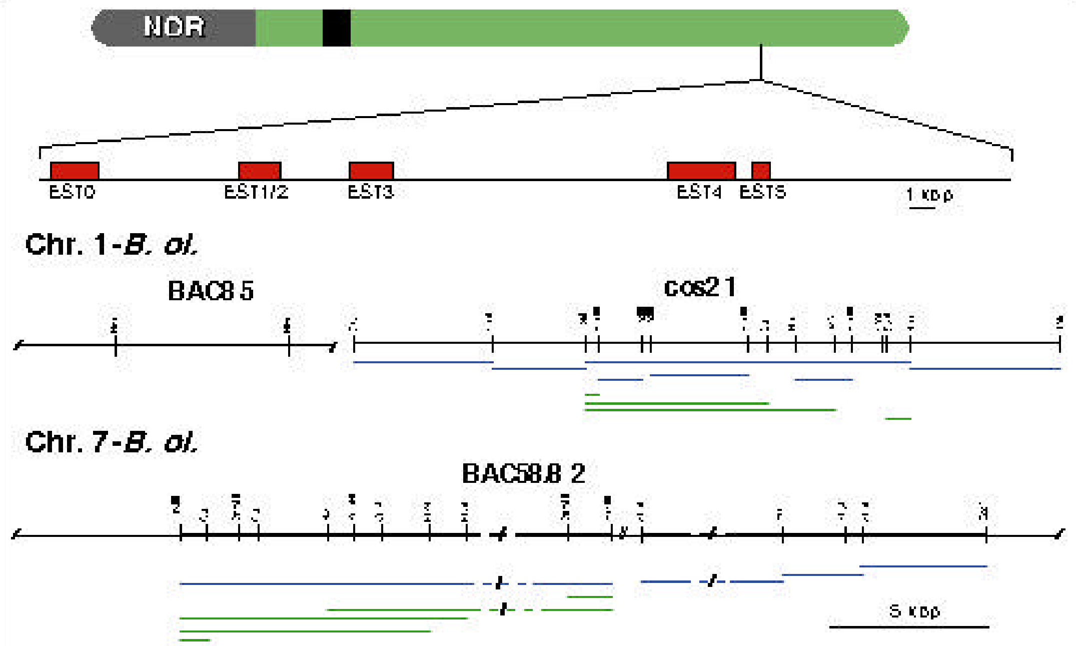


Figure 12: Shown above is the representation of *Brassica* contigs harbouring sequences homologous to ESTs 0-5. The contigs correspond to the region of *Arabidopsis* chromosome 4 and the sequenced region in *Capsella* shown in Fig. 8. The *Brassica* chr. 1 locus is represented by two sequenced segments, indicated in bold. The *Hin*dIII fragment is derived from BAC85 and the other segment corresponds to the insert of cos21. For the chr. 7 locus, two large fragments were sub-cloned from BAC58, which is identical to BAC82. Blue lines represent sub-clones and green lines are deletion clones. Interrupted lines indicate areas which have not been sequenced. Scales are different for the *Arabidopsis* and *Brassica* regions.

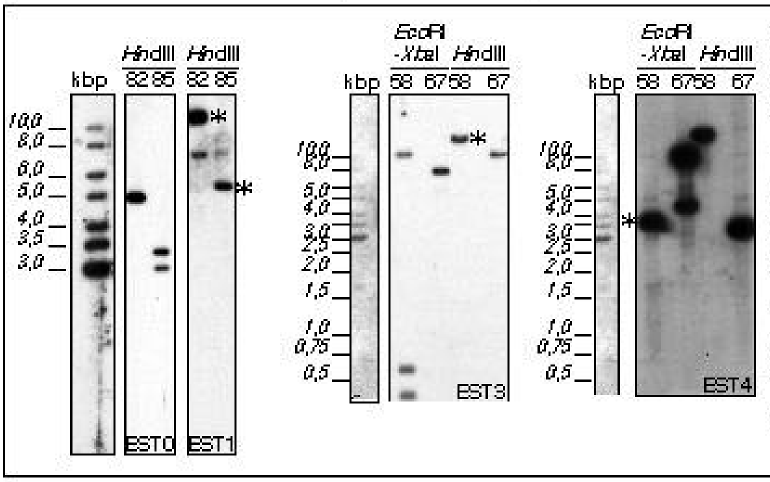


Figure 13: Results of Southern blot hybridisations of BAC clone DNA using ESTs 0-4 as probes. One BAC clone from each of the homeologous loci is represented on each blot. BACs 85 and 67 are located on chromosome 1, whereas BACs 82 and 58 map to chromosome 7. Asterisks indicate cloned fragments.

The *A. thaliana* EST database (TIGR *Arabidopsis* Gene Index, TIGR-AtGI release5.0; chapter 2-2-7) is currently containing 110,724 EST sequences. Since several EST sequences may correspond to a particular gene, the EST sequences have been assembled into contigs (TCs, Rounsley *et al*. 1996). These contigs can be used for sequence alignments with genomic DNA sequences.

Five ESTs are corresponding to a 20 kbp region of BAC F10N7 (GenBank acc. no. AL021636) in *A. thaliana*. All ESTs have been completely sequenced to confirm their identities and to obtain the entire sequence of the clones. EST 1 (~600 bp) and EST 2 (~1150 bp) showed an overlap of 300 bp. The clones differ at their 3'-end sequences, they carry a poly-A tail at different positions. ESTs 3 and 4 only span the 3'-end of their corresponding genes. For EST 5 which has homology to a ribosomal protein gene (rpl39), only a partial sequence could be obtained.

A BLAST analysis of the sequenced *C. rubella* region against the *A. thaliana* database, also revealed homology to the Knat5 gene of *A. thaliana* (mRNA, GenBank acc. no. D43962, noted EST 0 in Figures 8, 12 and 13) and copies of cytochrome P450-like genes. Three copies of cytochrome P450-like genes are present downstream of sequences corresponding to EST 5 in the *A. thaliana* region.

Taking this into account, an *Arabidopsis* region of approximately 47,000 bp was used for the following analysis. It is located on BAC F10N7 (bp 40,000-96,589) and is referred to F10N7-seg. This region partially overlaps with another sequenced BAC, F11C18 (GenBank acc. no. AL049607).

F10N7-seg has been aligned using the BLASTN tool with *Arabidopsis* EST contigs (http://www.tigr.org/tdb/agi/). In addition to the ESTs which had been initially identified with a much smaller EST collection (ESTs 1-5), six other EST contigs could be determined. TC100015 (unknown function), can be mapped between the Knat5 gene and the gene corresponding to the EST 1/2 contig, it is covering an entire putative gene. TC93505 corresponds to the 5'end of EST 3 and seems to carry mainly the 5'-non-translated leader. AV545275/AV553989 are two ESTs that are homologous to the region downstream of the sequence corresponding to EST 5, and are found in the consensus sequence TC94311. TC103345 is another partial EST contig which has been found around bp 41,000 of F10N7-seg. At the very end of F10N7-seg which overlaps with the sequence of F11C18 sequences homologous to two ESTs, AI998897 and AV542993, could be detected. Sequence AI998897 is part of TC82122 and AV552993 spans TC82121. BAC F10N7 does not contain the entire sequence corresponding to TC82121 (Figure 14).

Thus, very few genes are spanned in their entirety by ESTs, therefore, this *A. thaliana* sequence was submitted to two different gene prediction programs, Eukaryotic GeneMark.hmm at http://dixie.biology.gatech.edu/GeneMark/eukhmm.cgi and Genscan at http://genes.mit.edu.GENSCAN.html.

Based on the predictions, evidence for several other genes has been found. TC77106 could be placed into the region between the genes corresponding to ESTs 3 and 4. It has similarity to serine/threonine protein kinase genes. TC84838, TC80777 and TC73278 are copies of the cytochrome P450-like gene family. Corresponding ESTs have not been found, but prediction programs could infer ORFs in all three copies.

A. thaliana has been noticed to contain large segmental duplications (Blanc *et al*. 2000; The *Arabidopsis* genome initiative 2000). For some sequences of the analysed region homologous sequences could be found in other chromosomal locations. TC77106, for example, has homology to a locus located on chromosome II (GenBank acc. no. AC007070, between bp 11,000 and 15,000). It has not been established whether both copies shared the same exon/intron structures. EST contig TC103345 has sequence homology with chromosome 1, but no ORF could be determined using prediction programs, for the chromosome 4 locus.

All data taken together a region dense in genes is revealed. Experimental evidence by cDNA or EST sequences could be found for several genes (Knat5 gene TC71614; AAT gene X91865; ESTs 1/2; EST 3; EST 5; TC100015; TC93505; TC94311; TC103345; TC82121 and TC82122), however, gene prediction programs identify several other coding regions (TC84838, TC80777, TC73278, TC77106, TC73277). Figure 14 summarises the data.

3-2-3-2 Comparison of the *Arabidopsis* region with those of other species

a-Capsella rubella *region*

Based on the alignment of *Arabidopsis* cDNA sequences with genomic sequences of *A. thaliana* and *C. rubella*, it could be shown that exon/intron structures of orthologous genes are very similar. In general, lengths of exon sequences are conserved. Furthermore, an average sequence identity of ~90% for exon sequences could be established (Acarkan *et al*. 2000; Rossberg *et al*. 2001).

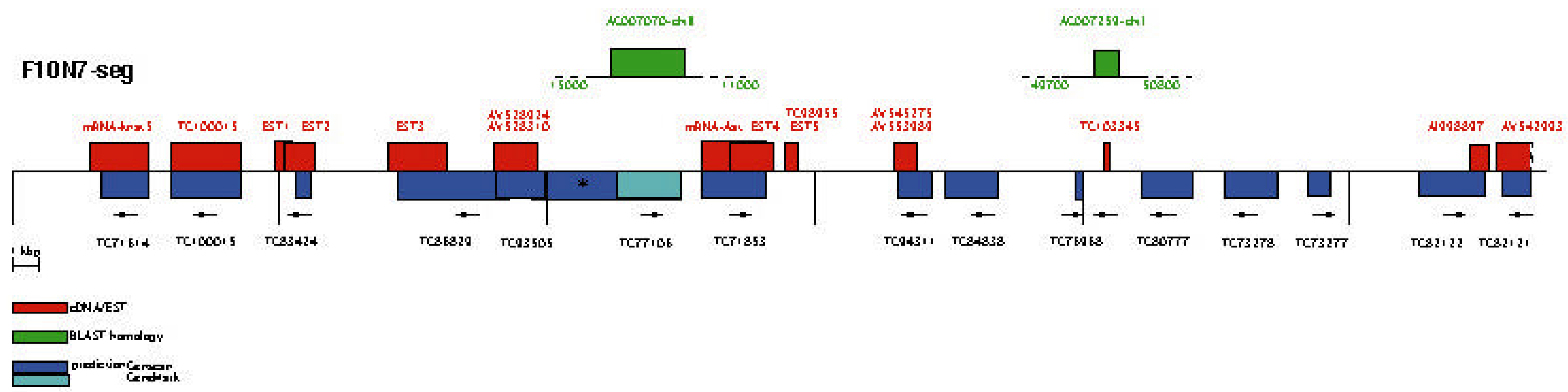


Figure 14: Arrangement of genes in the *Arabidopsis* region represented by F10N7-seg. EST sequences are represented as red boxes, their extent compared to the genomic sequence is indicated. Blue boxes depict gene predictions obtained with two different programs, Genscan and GeneMark. The extent of the gene predictions corresponds to the regions between the start and stop codons. Green boxes show homology located in other regions of the *Arabidopsis* genome and arrows above the name of the ESTs/predictions indicate the reading frame orientation. Thin vertical lines appear every 10 kbp of the F10N7-seg sequence. The predicted ORF marked with an asterisk does not correspond to a TC.

In this study, the following strategy has been taken to characterise and compare gene structures in *A. thaliana* and *C. rubella*. All ESTs, cDNAs and predicted genes of *A. thaliana* have been aligned with the sequence contig established for *C. rubella*. The gene structure can be easily deduced by aligning *A. thaliana* EST and cDNA sequences with the *Capsella* genomic DNA sequence. Intron borders in *A. thaliana* as well as in *C. rubella* are defined by the di-nucleotides GT…AG. The putative *C. rubella* exon sequences matching *A. thaliana* cDNA and EST sequences were assembled and translated into amino acid sequences.

For genes that are only partially covered by ESTs or that are merely based on predictions, the corresponding region of *C. rubella* genomic DNA sequence has been submitted to two gene prediction programs (Genscan and GeneMark). Combining both sets of data and taking into account the finding that exons are conserved in length and sequence between *A. thaliana* and *C. rubella*, predicted exons of similar size in both species were considered as likely. The putative exon sequences were assembled and translated into amino acid sequences for the predictions in both species. Gene predictions could then be validated if they span a complete ORF both in *A. thaliana* and in *C. rubella*. The resulting amino acid and assembled exon sequences were compared between the two species.

For the *C. rubella* region, the following genes Cr-71614, Cr-100015, Cr-83424, Cr-86829, Cr-77106, Cr-AAT, Cr-73278 could be predicted. The F10N7-seg harbours the *A. thaliana* genes in the same order and orientation. Additionally, the spacing between the genes is very similar. Genes in both species have identical exon and intron numbers (Figure 15). However, sequences corresponding to EST contig At-103345 and the gene prediction At-76968 were not found in the *Capsella rubella* contig.

Another difference seen between the regions in both species is the copy number of cytochrome P450-like genes. The *Arabidopsis* contig harbours three copies, whereas the *Capsella* region might contain only one. Pairwise comparisons between the copies of the *A. thaliana* P450-like genes and the *C. rubella* copy indicate the lack of the At-84838 and At-80777 genes in *Capsella*, At-73278 is potentially orthologous to the cytochrome P450-like gene in *Capsella*. The ORF sequence comparisons show 91,3% sequence identity at the nucleotide level for At-73278/Cr-73278 versus 88,9% and 90,1% for At-84838/Cr-73278 and At-80777/Cr-73278, respectively (Table H). If the *A. thaliana* ORFs are compared among themselves nucleotide sequence identities range from 92,7% to 93,1%.

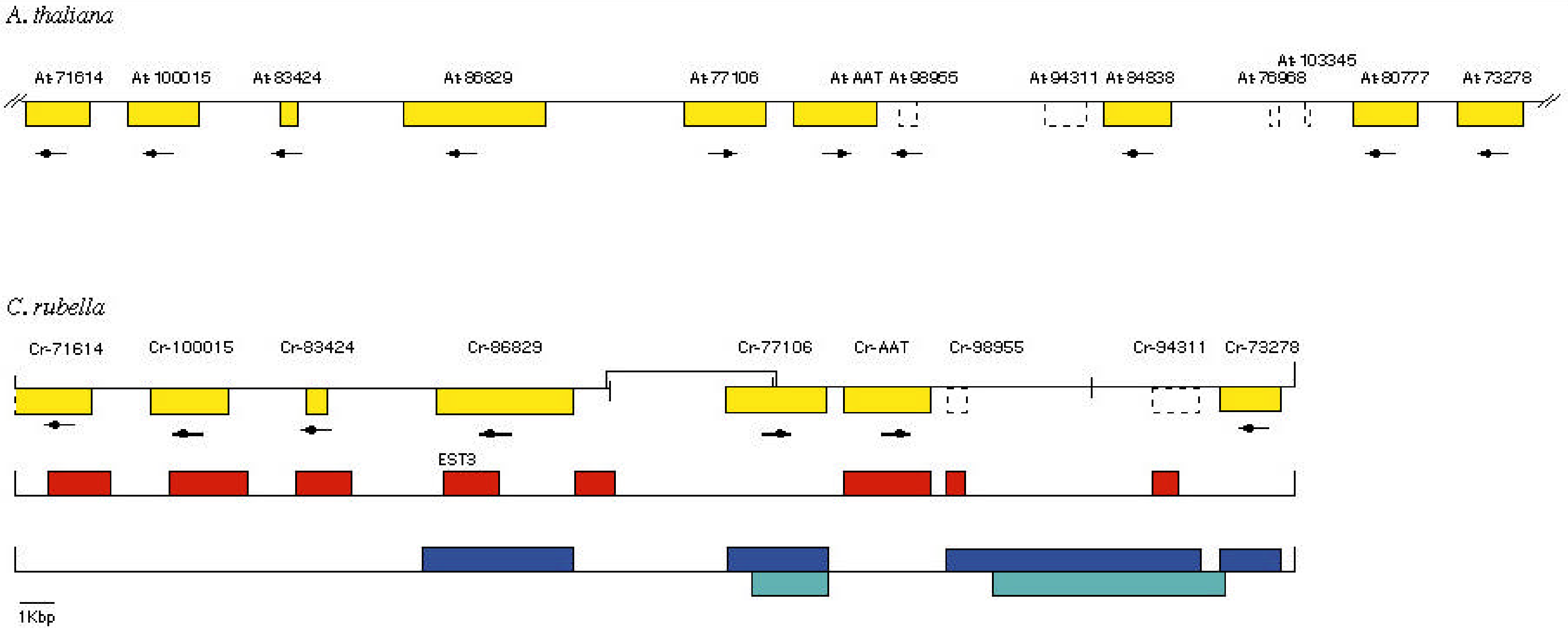


Figure 15: Organisation of the *Capsella* contig sequence with the distribution of genes compared to the corresponding *Arabidopsis* region. On the top the *Arabidopsis* segment is shown (Figure 14). Beneath the sequenced *C. rubella* region is represented. Yellow boxes correspond to predicted genes wich share a common structure in *Arabidopsis* and *Capsella*. The arrows indicate the reading frame orientation. The open boxes indicate the homology found with *Arabidopsis* ESTs TC-98955 and TC-94311. In red, *Arabidopsis* ESTs or contigs of ESTs are shown, aligned with the *Capsella* sequence. The third level shows the alignment of the predictions obtained respectively with Genscan (dark blue) and GeneMark (light blue). These predictions have only been performed on regions of the *Capsella* sequence lacking information on *Arabidopsis* genes.

Sequence homology to TC98955 and TC94311 is found in the regions of both species, but it was not possible to delimit a concordant open reading frame.

b- *Brassica oleracea* region

For the structural analysis of the *Brassica oleracea* genes, the same strategy was followed as described for the *Capsella rubella* genes.

The region located on chromosome 1 of *B. oleracea* is covered by a sequenced cosmid insert of 22,424 bp, and next to this region, another fragment isolated from BAC 85 harbours sequences homologous to At-100015 and ESTs 1/2 (Figure 16). The 5' part of the Bo-100015 gene on chromosome 1 could not be determined, because it is not completely residing on the sub-clone established with *Hin*dIII. The sub-clone encompasses 5550 bp, only a slight increase of the intergenic space between genes Bo-100015-chr. 1 and Bo-83424-chr. 1 is noted compared to the regions in *A. thaliana* or *C. rubella*. The sequenced insert of cosmid B21 contains sequences corresponding to EST 3. The 3'-end of the gene, covered by *A. thaliana* EST 3 is represented, but homology to the remainder of the gene could not be established. The Bo-AAT-chr. 1 gene is found to be complete, as well as Bo-77106-chr. 1, which corresponds to an *A. thaliana* gene prediction. Sequence homologies are detected for At-98955 and At-94311 but an ORF could not be determined. A short sequence identity could be detected between *B. oleracea* and gene prediction At-73277 of *A. thaliana*. Sequence At-73277 is matching both *B. oleracea* loci. Predictions applied on the *A. thaliana* and the *B. oleracea* sequences corresponding to At-73277 permitted to determine a putative ORF consisting of two exons. The orientation of *B. oleracea* genes and predictions relative to each other in the chromosome 1 region is identical to the arrangement in *A. thaliana* and *C. rubella*. As far as the intergenic regions could be analysed, they are similar in size in the three species.

The second locus which shows homology to the *A. thaliana* region of interest maps to chromosome 7 of *B. oleracea*, two segments of ~12 kbp and ~7 kbp have been sequenced. The ~12 kbp fragment has been isolated from BAC 82 and the ~7 kbp segment corresponds to a sequence assembly based on a sub-clone derived from BAC 58 and two sub-clones from cosmid B3 (Figure 12). On the chromosome 7 locus of *B. oleracea*, the BAC82 sub-clone of ~12,500 bp carries a second copy of the gene corresponding to ESTs 1/2 and a putative complete copy of the gene with homology to At-86829.

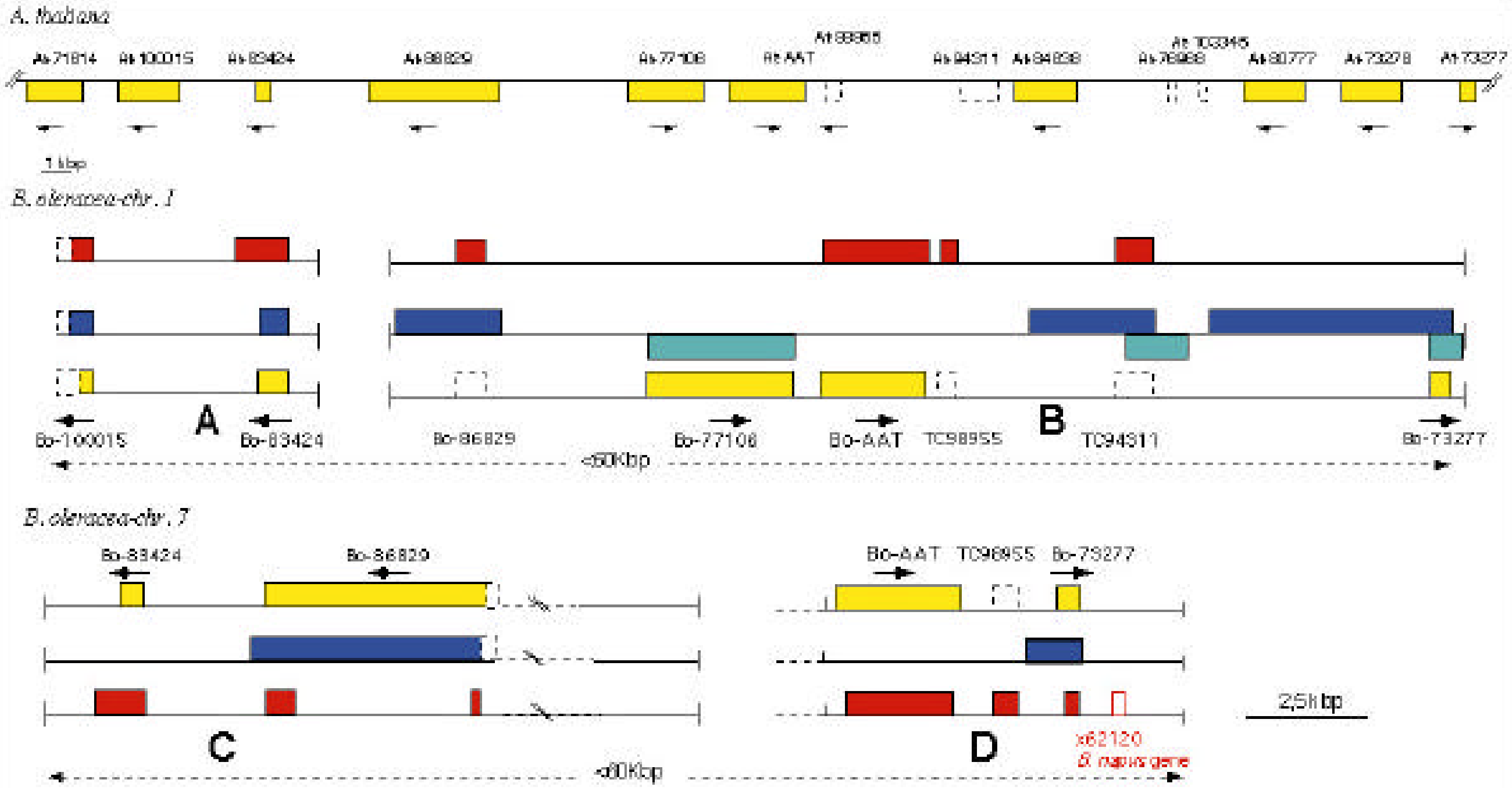


Figure 16: Depicted above are the *Brassica* contigs mapping to two loci on *Brassica* chromosomes 1 and 7. The organisation of the *A. thaliana* region is shown for comparison. The three presented levels of each *Brassica* sequence show in red EST homologies, in blue predictions made with two different programs (Genscan in dark blue and GeneMark in light blue), and in yellow, the genes, for which a common structure could be determined in *A. thaliana* and *B. oleracea*. Part A of the picture represents the sub-clone derived from BAC85, the B part is cosmid cos21, completely sequenced. On C and D are drawn the two sub-clones derived from BAC58-82 mapping to the chromosome 7 locus. D is a composite sequence alignment of a BAC sub-clone and two cosmid sub-clones. Dashed lines indicate that sequences in this area are not available and arrows indicate the orientation of the reading frame. ESTs showing homology to *Brassica* sequences but no ORF are depicted as open boxes, namely TC98955 and TC94311 on B and D. Only a remnant of TC86829 is found on B corresponding to the EST coverage. Truncated predictions and genes end with dashes. Two different scales are given for the *A. thaliana* and *B. oleracea* regions.

The 7,500 bp fragment harbours a copy of the Bo-AAT gene, homologous sequences to At-98955 and a second copy of the putative ORF At-73277 (Figure 16). The assembly of the B3 sub-clone has been submitted to a BLAST search. Sequence identity with a *B. napus* cDNA (GenBank acc. no. X62120) could be identified. In *A. thaliana*, a homologue of this gene is mapping also to chromosome 4, but to a different BAC (BAC F20O9). BAC clones F20O9 and F10N7 are ~1,4 Mbp apart.

c- Comparisons of gene structures

		Sequence identity (%)	
		Nucleotide level	*Amino Acid level*
TC71614 (Knat5 gene)	*Ath/Cr**	93,2	95,2
TC100015	*Ath/Cr*	84,1	80,4
TC83424	*Ath/Cr*	89,5	91,1
	Ath/Bo-chr1	80,8	73,7
	Ath/Bo-chr7	80,8	76,6
	Cr/Bo-chr1	80,3	78,8
	Cr/Bo-chr7	82,9	85,8
	Bo-chr1/Bo-chr7	82,7	81
TC86829	*Ath/Cr*	93	91
	*Ath/Bo-chr7**	88	86,6
	*Cr/Bo-chr7**	87,5	87,1
TC77106	*Ath/Cr*	92	91,4
	Ath/Bo-chr1	88,5	88,8
	Cr/Bo-chr1	88,1	87,6
AAT gene	*Ath/Cr*	93	96,9
	Ath/Bo-chr1	87,8	96
	Ath/Bo-chr7	89,5	94,5
	Cr/Bo-chr1	87,7	96,2
	Cr/Bo-chr7	89,5	95,3
	Bo-chr1/Bo-chr7	89,4	97,4
P450	*Ath-1/Ath-2*	93,1	88,2
	Ath-1/Ath-3	93	88,5
	Ath-2/Ath-3	92,7	85,5
	Ath-1/Cr	90,1	87,9
	Ath-2/Cr	88,9	85,5
	Ath-3/Cr	91,3	90
TC73277	*Ath/Bo-chr1*	86,5	80,5
	Ath/Bo-chr7	87,9	85,6
	Bo-chr1/Bo-chr7	85,9	81,6

Table H: Comparison of exon sequences at the nucleotide and amino acid levels. In the first column the different *A. thaliana* genes are listed for which homologous *C. rubella* and/or *B.oleracea* sequences have been analysed. The different *B. oleracea* loci are distinguished from each other by the chromosome they map to, *A. thaliana* copies of the cytochrome P450-like genes are designated as following: Ath1-TC84838; Ath2-TC80777; Ath3-TC73278. The asterisks indicate the genes for which only partial sequence information could be obtained.

As a general rule, positions and numbers of introns are strictly conserved between *A. thaliana* and *C. rubella*. This is not necessarily the case for the gene structures obtained for *B. oleracea* sequences using the *A. thaliana* or *C. rubella* exon sequences as template. The intron borders corresponded in all cases to the consensus sequences (GT…AG). The level of nucleotide and amino acid identity for the exons of the genes and predictions are summarised in Table H, for each species pair used in this study.

For the *C. rubella* Knat5 homologue, Cr-71614, the 3' end of the sequence is not present on the cosmid. The exons are strongly conserved in size and sequence between *A. thaliana* and *C. rubella,* only the first intron shows an increase of 65 bp (Figure 17a). This gene is known to be present on both *B. oleracea* loci (Figure 12).

The *A. thaliana* At-100015 contains large introns (810-1070 bp). The sizes of the introns are very similar in *C. rubella*. Such large introns are unusual for *A. thaliana* and *C. rubella* genes (Acarkan *et al.* 2000; Rossberg *et al*. 2001) but the fact that *Arabidopsis* ESTs are covering this region and that a conserved gene structure was found in both species supports the gene feature given. TC100015 is also present on chromosome 1 of *B. oleracea*, the first exon could be aligned with the *A. thaliana* and *C. rubella* sequences. Sequence information for the rest of the gene was not available (Figure 17b).

Despite the fact that At-83424 covers almost 1,5 kbp of *A. thaliana* genomic sequence, the ORF which could be established spans only 546 bp (Figure 17c). From the two *B. oleracea* loci carrying this gene, the chromosome 7 locus exhibited a homologous ORF of exactly 546 bp while the sequence spanned 609 bp on chromosome 1. The *C. rubella* copy covered 618 bp.

EST 3 covers only the 3' part of prediction At-86829. The exon/intron structure is a composite of the structure deduced from the EST alignment and from prediction programs. Predictions have been performed with the DNA sequences of *A. thaliana*, *C. rubella* and *B. oleracea,* exon sequences were considered as likely if they were in common to all three species. All three potential ORFs were translated to confirm the exon assembly. The structure of this gene could be established and it represents a very long gene. The first exon is very small (47 bp) rejected almost 1 kbp from the remainder of the genes in *A. thaliana* and *C. rubella*. The entire sequence is thought to be present on chromosome 7 of *B. oleracea* (albeit the first exon and part of the second exon are not present on the sequenced fragment). The chromosome 1 copy in contrast, is only present as remnant of the 3'end of the At-86829. In Figure 17d, the TC86829 gene structure is shown. Red labelling indicates homology to EST 3. When the EST 3-homologous fragments located on *B. oleracea* chromosomes 1 and 7 are compared, an alignment of about 700 bp that is 86,1% identical is the result. The alignment shows numerous mismatches between the two copies and at least 3 indels of one nucleotide were present on separate locations.

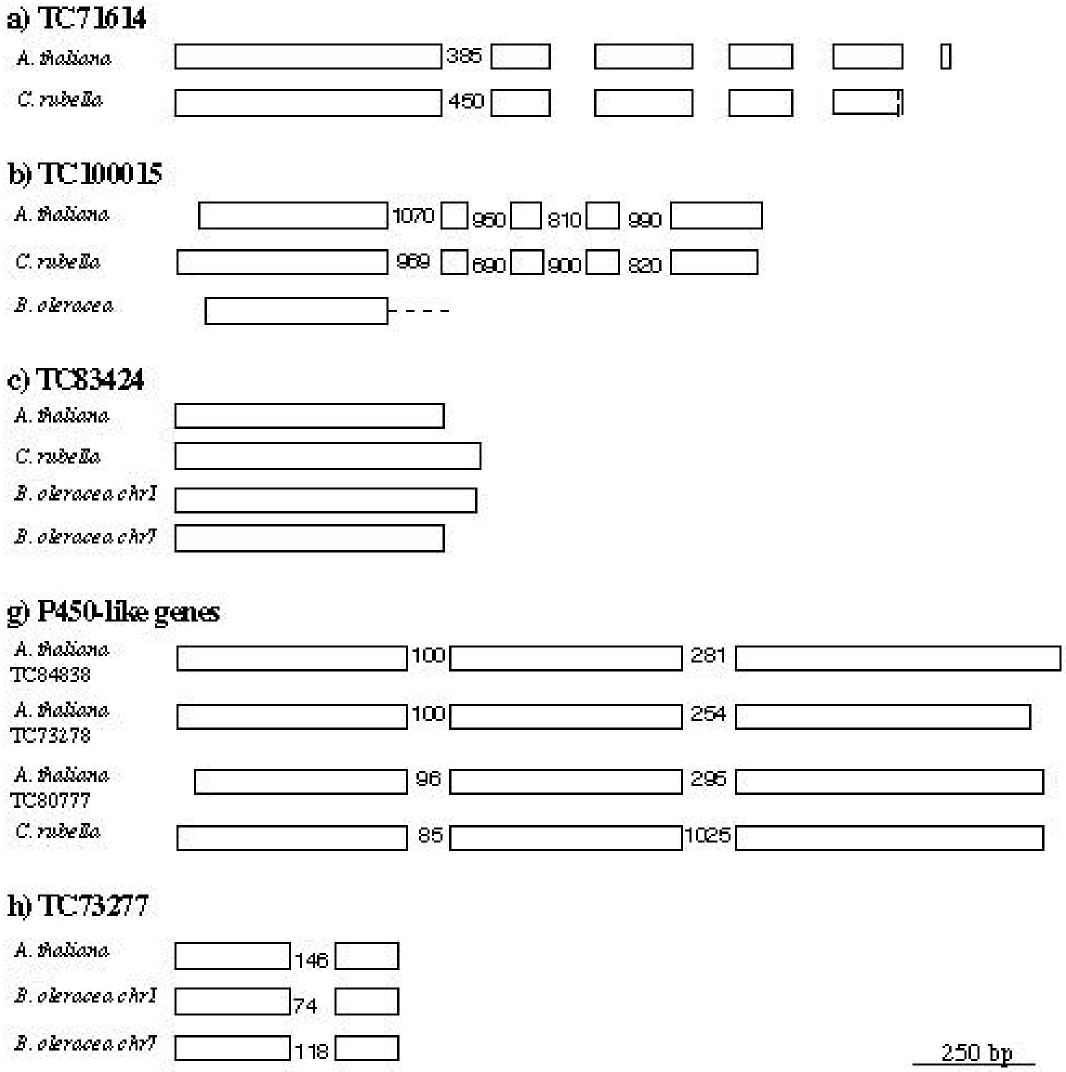
TC77106 is a predicted gene sequence and common features could be observed between the three species. Increased sizes of intron sequences were observed in the *B. oleracea* gene, especially on the 3' end of this prediction (Figure 17e). The *C. rubella* gene exhibited the particularity of an enlarged exon 7.

The AAT gene structure is comparable in *A. thaliana* and *C. rubella*. In genes of both species 11 exons are found. Only intron sizes differ between the genes of the two cruciferous plants. The AAT gene is present at both loci of *B. oleracea*. Whereas the chromosome 7 copy (Bo-AAT-chr. 7) exhibits the same number of introns as the gene in *A. thaliana*, the copy located on chromosome 1 (Bo-AAT-chr. 1) lacks intron 8 and consequently has only 10 exons. On Bo-AAT-chr. 7, the eighth intron is dramatically increased in size when compared to the *A. thaliana* and *C. rubella* genes (Figure 17f).

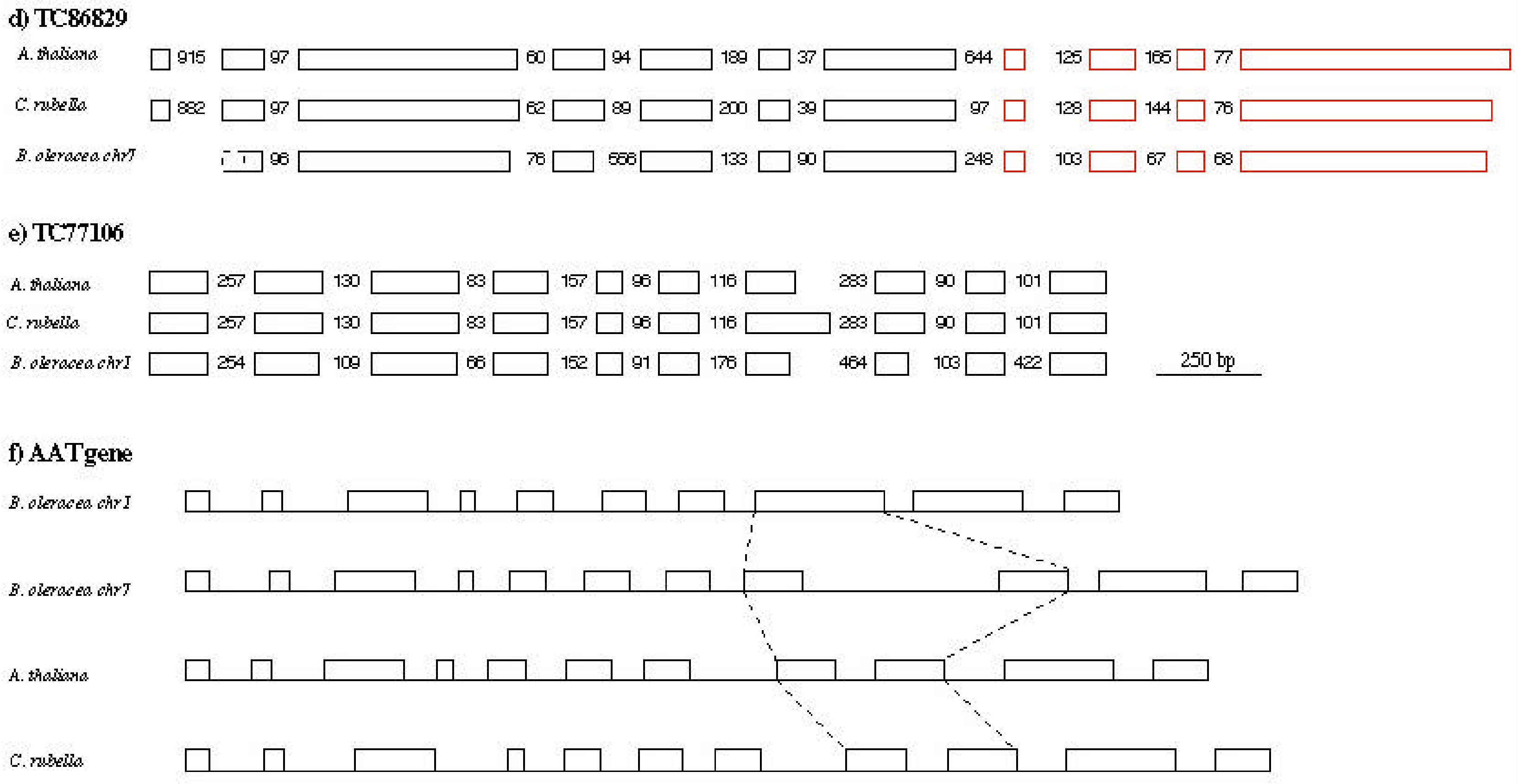
Of all genes studied the AAT genes showed the highest degree of conservation at the amino acid level. The two *B. oleracea* copies are 97,4% identical at amino acid level. A comparison of the *Arabidopsis* predicted protein sequence with the two *B. oleracea* copies shows sequence identities of 96,0 and 94,5%, similar values (96,2% and 95,3%) are obtained if the *C. rubella* gene is translated and compared to the amino acid sequences of the *Brassica* genes. The deduced protein sequences of the *A. thaliana* and *C. rubella* genes are also highly identical (96,7%).

The amino acid sequences of the AAT gene, translated from the nucleotide sequences and obtained for the three species (*A. thaliana*, *C. rubella* and two copies of *B. oleracea*) have been compared with a multiple sequence alignment program. Thirty-two out of 454 amino acid positions are differing between the genes of the three species (shaded amino acids). Thirteen positions are variable between the two copies of *B. oleracea*, similarly 13 amino acid exchanges are counted between the *A. thaliana* copy and the *C. rubella* copy of the AAT gene (Figure 18).

Gene predictions TC84838 and TC73278 have been assembled from F10N7-seg while TC80777 has been established from the overlapping BAC F11C18. No EST is available for these cytochrome P450-like genes. Alignment and assemblies have been realised with gene prediction programs and comparison of the three predictions to each other. The three exon/intron structures of the cytochrome P450-like genes are very well conserved within *A. thaliana*. Exon and intron sizes are very similar between the copies (Figure 17g), but one (At-80777) has a deletion affecting the first exon.



Figures 17, a, b, c, g, and h: Description of the gene structures established in this study. Gene sequences derived from *A. thaliana*, *C. rubella* and *B. oleracea* are compared. Boxes represent the exon sequences drawn to scale, in contrast, sizes of introns are given in bp.



Figures 17 d, e and f: Description of gene structures established in this study. Gene sequences derived from *A. thaliana, C. rubella* and *B. oleracea* are compared. Boxes represent exon sequences drawn to scale, whereas introns sizes are indicated in bp. In the case of the AAT gene, introns are also given to scale to highlight the loss of one intron in the *B. oleracea* chromosome 1 gene copy. Dashed lines connect homologous exon sequences of the different genes to show the difference in exon/intron structure.

Interestingly, the sequence of BAC F10N7 differs from the sequence of BAC F11C18 in the region corresponding to TC80777 by a 1 bp indel. Thus, if the gene structure of At-80777 would be based on the sequence of BAC F10N7 rather than that of BAC F11C18 a different structure would be the result. Since the three copies of the cytochrome P450-like genes share very similar exon/intron structures it was concluded, that the indel was due to a sequencing error.

```
B. oleracea-7   MASSMLSLGSTSLLPREINKDKLKLGPSGSNPFLRTKSLSRVTMSVSVKPSRFEGITMAP 60
B. oleracea-1   MASSMLSLGSTSLLPREINKDKLKLGTSGSNPFLKAKCFSRVTMSVAVKPSRFEGITMAP 60
A. thaliana     MASLMLSLGSTSLLPREINKDNVKLGTSASNPFLKAKSFSRVTMTVAVKPSRFEGITMAP 60
C. rubella      MASSMLSLGSTSLLPREISKDKLKLGTSGSNPFLKAKSFSRVTMAVAVTPSRFEGITMAP 60

B. oleracea-7   PDPILGVSEAFKADTNELKLNLGVGAYRTEELQPYVLNVVKKAENLMLERGDNKEYLPIE 120
B. oleracea-1   PDPILGVSEAFKADTNELKLNLGVGAYRTEELQPYVLNVVKKAENLMLERGDNKEYLPIE 120
A. thaliana     PDPILGVSEAFKADTNGMKLNLGVGAYRTEELQPYVLNVVKKAENLMLERGDNKEYLPIE 120
C. rubella      PDPILGVSEAFKADTNEMKLNLGVGAYRTEELQPYVLNVVKKAENLMLERGDNKEYLPIE 120

B. oleracea-7   GLAAFNKATAELLFGAGHPVIKEQKVATIQGLSGTGSLRLAAALIERYFPGAKVLISAPT 180
B. oleracea-1   GLAAFNKATAELLFGAGHPVIKEQKVATIQGLSGTGSLRLAAALIERYFPGAKVLISAPT 180
A. thaliana     GLAAFNKATAELLFGAGHPVIKEQRVATIQGLSGTGSLRLAAALIERYFPGAKVVISSPT 180
C. rubella      GLAAFNKATAELLFGAGHPVIKEQRVATIQGLSGTGSLRVAAALIERYFPGAKVVISSPT 180

B. oleracea-7   WGNHKNIFNDAKVPWXEYRYYDPKTIGLDFEGMIEDIKEAPEGSFILLHGCAHNPTGIDP 240
B. oleracea-1   WGNHKNIFNDAKVPWSEYRYYDPKTIGLDFEGMIADIREAPEGSFILLHGCAHNPTGIDP 240
A. thaliana     WGNHKNIFNDAKVPWSEYRYYDPKTIGLDFEGMIADIKEAPEGSFILLHGCAHNPTGIDP 240
C. rubella      WGNHKNIFNDAKVPWSEYRYYDPKTIGLDFEGMIADIKDAPEGSFILLHGCAHNPTGIDP 240

B. oleracea-7   TPEQWVKIADVVQEKNHIPFFDVAYQGFASGSLDEDAASVRLFAERGMEFFVAQSYSKNL 300
B. oleracea-1   TPEQWVKIADVIQEKNHIPFFDVAYQGFASGSLDEDAASVRLFAERGMEFFVAQSYSKNL 300
A. thaliana     TPEQWVKIADVIQEKNHIPFFDVAYQGFASGSLDEDAASVRLFAERGMEFFVAQSYSKNL 300
C. rubella      TPEQWVKIADVIQEKNHIPFFDVAYQGFASGSLDEDAASVRLFAERGMEFFVAQSYSKNL 300

B. oleracea-7   GLYAERIGAINVVCSSADAATRVKSQLKRIARPMYSNPPVHGARIVANVLGDATMFGEWK 360
B. oleracea-1   GLYAERIGAINVVCSSADAATRVKSQLKRIARPMYSNPPVHGARIVANVVGDAAMFNEWK 360
A. thaliana     GLYAERIGAINVVCSSADAATRVKSQLKRIARPMYSNPPVHGARIVANVVGDVTMFSEWK 360
C. rubella      GLYAERIGAINVVCSSADAATRVKSQLKRIARPMYSNPPVHGARIVANVVGDPTMFGEWK 360

B. oleracea-7   AEMEMMAGRIKTVRQRLYDSLVSKDKSGKDWSFILKQIGMFSFTGLNKAQSDNMTNKWHV 420
B. oleracea-1   AEMEMMAGRIKTVRQQLYDSLVSKDKSGKDWSFILKQIGMFSFTGLNKAQSDNMTDKWHV 420
A. thaliana     AEMEMMAGRIKTVRQELYDSLVSKDKSGKDWSFILKQIGMFSFTGLNKAQSDNMTDKWHV 420
C. rubella      AEMEMMAGRIKTVRQELYDSLVSKDKSGKDWSFILKQIGMFSFTGLNKAQSDNMTNKWHV 420

B. oleracea-7   YMTKDGRISLAGLSMAKCEYLADAIIDSCHNVS 453
B. oleracea-1   YMTKDGRISLAGLSMAKCEYLADAIIDSHHNVS 453
A. thaliana     YMTKDGRISLAGLSLAKCEYLADAIIDSYHNVS 453
C. rubella      YMTKDGRISLAGLSMAKCEYLADAIIDSYHNVS 453
```

Figure 18: Multiple sequence alignment of the AAT gene sequence found in the three species. The AAT exon sequences have been translated into amino acid sequences and compared between the three Brassicaceae species. Amino acid exchanges between the three species have been shaded grey.

Each of the three *A. thaliana* genes has been compared to the *C. rubella* copy. Nucleotide sequence identities neither vary greatly between the *A. thaliana* sequences nor between those of *A. thaliana* and *C. rubella*. A notable difference between the genes of the two species is a large increase in size of the third intron of the *C. rubella* gene. It spans 1025 bp whereas the intron sizes of the *A. thaliana* genes vary between 254 and 295 bp. The nucleotide identity values indicate that the *C. rubella* gene is most closely related to At-73278. If this is considered to be the case, then the genes located between the 94311 and 73278 genes could have either been deleted from the *Capsella* region or alternatively the ancestral gene corresponding to At-73278 has been duplicated in *A. thaliana* after the divergence of these two crucifer species.

A prediction found on both *B. oleracea* loci is corresponding to TC73277 which seems to contain only two exons. The *A. thaliana* and *B. oleracea* gene structures are entirely predicted, homologous sequences between the two species are longer than these two exons but no longer open reading frame in common to the three genes could be determined. The sequenced *C. rubella* region does not contain homologous sequences to these predicted ORFs (Figure 10h).

d-Estimation of the gene content in the *Brassica* loci

The comparative analysis of the regions in the *A. thaliana*, *C. rubella* and *B. oleracea* genomes revealed a very similar organisation of genes. As a complement of the comparative sequence analysis, a hybridisation study was performed to gather more data concerning the gene repertoire of the homeologous *Brassica* loci. Therefore, BAC DNA has been digested with *Mlu*I and *Sma*I and fragments have been separated on PFGE. After blotting these DNAs, ESTs, and PCR products covering some predicted genes were used as probes to determine whether a particular gene was present in one or both *Brassica* loci, furthermore it was attempted to estimate the maximal size of the regions corresponding to F10N7-seg, which is spanning 46,590 bp.

It could be established that BACs 58 and 82 (chromosome 7) carry homologous sequences of the outermost genes (Knat5 and TC82121) identified on F10N7-seg within a region which maximally spans 60 kbp (data not shown). Presence of sequences homologous to TC86680, TC86829, AAT, TC98955, TC73277, TC82122 could be validated by hybridisation studies or due to available sequence information. These results are summarised in Figure 19.

BAC 67 is digested by *Sma*I into two fragments of 20 and 40 kbp. The 20 kbp fragment carries the end-sequence of this BAC (67R, isolated by iPCR) together with the gene homologous to ESTs 1/2, whereas the 40 kbp fragment spans from Bo-86829 to Bo-73277. However, sequences homologous to TC84838, TC80777 and TC73278 are not found. The sequences insert of cosmid B21 is included in BAC 67 and does not contain any *Sma*I site. It can be concluded, that the contig from Bo-83424 to Bo-73277 is present on BAC 67, it maximally encompasses 60 kbp (data not shown, Figure 19).

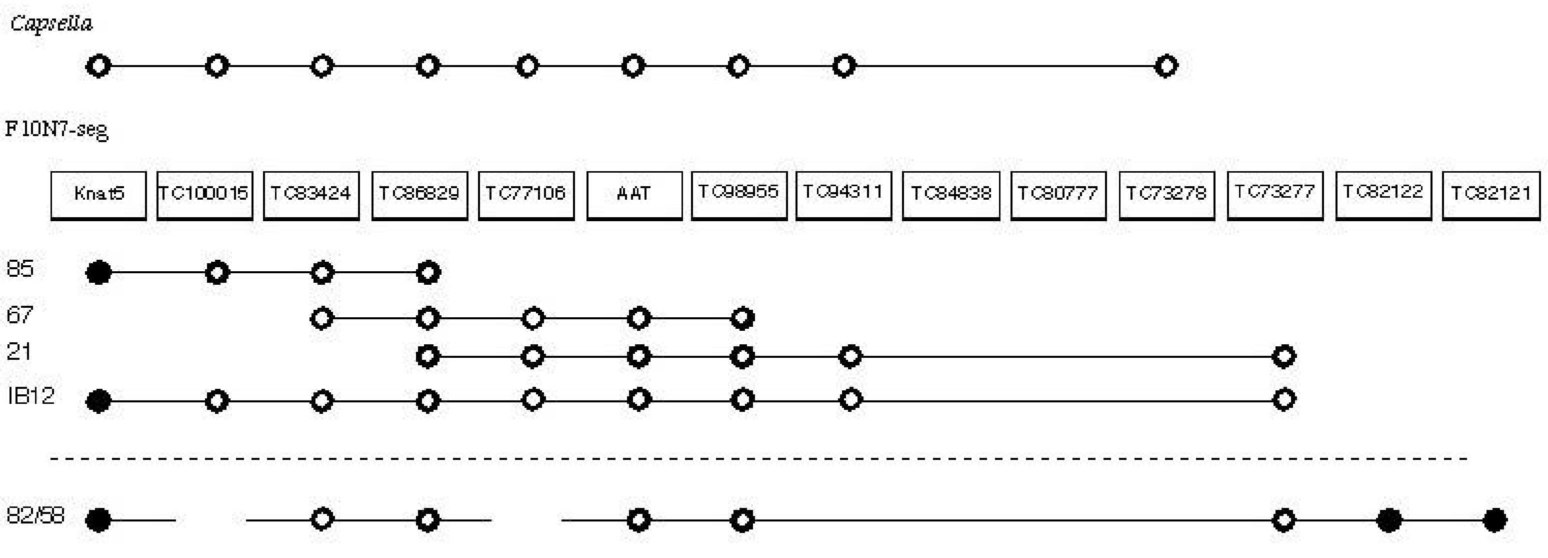


Figure 19: Gene content comparisons of the regions analysed in *A. thaliana*, *C. rubella* and *B. oleracea*. For the homeologous *Brassica* regions BAC clones covering these regions are shown (85, 67, 21, IB12, 82, 58). Presence or absence of genes and predicted genes was proven either by hybridisation experiments, sequence data or PCR analysis. Filled circles indicate the presence of homologous sequences, sufficient to generate hybridisation or PCR results. Open circles represent genes and predicted genes for which sequence data are available. The *Capsella* cosmid contig is represented at the top of the scheme. Interruption of the line corresponding to BACs 82 and 58 indicates that the presence of genes 100015 and 77106 has not been analysed.

3-3 THE RETROELEMENT

Repetitive elements, among them mobile elements, are important components of plant genomes. The *A. thaliana* genome contains approximately 10% of transposable elements (The *Arabidopsis* genome initiative 2000). A sequence corresponding to one of the *Capsella Mbo*-sub-clones has been found to be homologous to repetitive sequences in the *A. thaliana* genome. This observation provided an entry-point to compare repetitive components of the *A. thaliana* and *C. rubella* genomes. The aim of this study was the characterisation of a repetitive element in *C. rubella* and to compare its features to the corresponding sequences in the *A. thaliana* genome.

3-3-1 Features of retrotransposons

Mobile elements transposing through an RNA intermediate (retroids) are distinguished in three groups, the Long Interspersed Nuclear Elements (LINEs) which lack flanking LTRs (long terminal repeats), the *Copia*-type and *Gypsy*-type LTR-retrotransposons and the Small Interspersed Nuclear elements (SINEs) as such the *Alu* sequences found in the human genome. The *Copia*-type and *Gypsy*-type LTR-retrotransposons were named with reference to *Drosophila* LTR-retrotransposons which have been characterised first (reviewed in Grandbastien 1992). Retroviruses are thought to be restricted to the animal kingdom (Xiong and Eickenbush 1990). They are very similar to the *Gypsy* LTR-retrotransposons but are characterised by an envelope gene upstream of the 3’ LTR which plays a role in the cell to cell transfer of the retrovirus and infection (Bennetzen 2000b).

In Figure 20, the general organisation of a *Ty3/Gypsy* retrotransposon is depicted (Grandbastien 1992). The entire element is framed by Target Site Duplications (TSD) which are formed during the integration. During the integration process, the genomic DNA will suffer a staggered cut into which the element will be ligated, as a result short identical sequences, 4 to 6 nucleotides in length, will be found flanking the element. The LTRs of the retroelement are required for the transposition process following the hypothesis that they are containing very short ORFs coding for proteins involved in the transposition (Vicient *et al*. 1999a). Three bp away from the 5’LTR a potential Primer Binding Site (PBS) sequence is located which is complementary to, in the case of *Ty3/Gypsy* and *Del*-retroelements, the 3’end of methionine initiator tRNA (tRNA$_i^{Met}$)

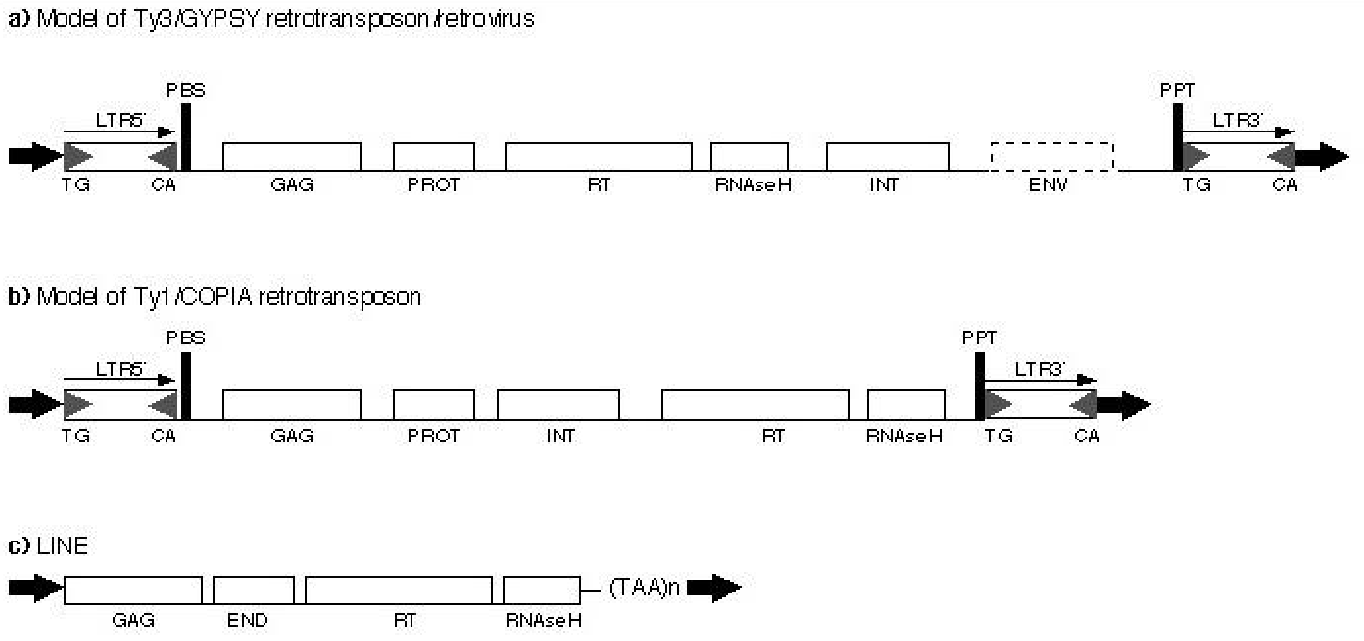


Figure 20: Features of retroelements. a) A LTR-Ty3/GYPSY retrotransposon is represented. Retroviruses share the same features, but additionally they possess an envelope gene (ENV). b) A LTR-Ty1/COPIA retrotransposon is shown, which differs from Gypsy-type elements by the location of the integrase (INT) ORF. c) A Long Interspersed Nuclear Element (LINE) does not contain LTR sequences, but ORFs for several proteins, GAG, endonuclease, reverse-transcriptase (RT) and RNAse H. Directly repeated LTR sequences are characteristic of LTR-retrotransposons and have canonical TG...CA termini, fat arrows represent the target site duplication (TSD), created during the integration process. PBS: primer binding site, PPT polypurine tract.

(Smyth *et al*. 1989; Wright and Voytas 1998). This sequence is required for DNA synthesis. As shown in Figure 20, different domains can be recognised. The different components encoded by a retroelement are the GAG protein, highly divergent and therefore poorly recognised as a protein domain in database searches, the protease (PROT), the reverse-transcriptase (RT) and RNAse H, which usually from the polyprotein complex, and the integrase (INT). If the integrase ORF is located downstream of the ORF coding for RT, the element belongs to the *Ty3/Gyspy* class, *Ty/Copia* elements carry the INT upstream of the RT (Figure 20). The presence of an ENV ORF characterises retroviruses, which are not found in plants. Retrotransposons carrying a supplementary ORF (putative ENV) downstream of the INT are belonging to the Errantivirus (as for example *athila*; Pélissier *et al*. 1995; Pélissier *et al*. 1996) family while elements lacking this feature belong to the Metavirus family (as for example Tat retrotransposons; Konieczny *et al*. 1991). Immediately adjacent to the 3’ LTR is a 12-15 bp long polypurine tract (PPT) used for the synthesis of the (+) DNA strand (Grandbastien 1992; Wright and Voytas 1998).

Retroelements are transposing through an RNA intermediate, they do not excise from their original position, rather a copy of the original element is integrated elsewhere. This leads to an amplification of such elements in genomes with time (Kumar and Bennetzen 1999; Bennetzen 2000b). Retrotransposons represent a very prevalent class of mobile elements in many plant genomes, they can constitute as much as 50 to 80% of the nuclear DNA in grasses (SanMiguel *et al.* 1998).

3-3-2 *Capsella* retroelements

3-3-2-1 Screening of the *Capsella* libraries

Analysing the sequence data of the library of *Capsella Mbo*-fragments, *Mbo*-D22 showed homology to several *A. thaliana* BAC sequences. Approximately 30 GenBank entries showed higher homologies than the cut-off E-value set at e^{-15}. The sequence identities between the 484 bp fragment of *Capsella* DNA and the corresponding sequences in the *Arabidopsis* genome is ~85%. The BACs map to different loci in the *A. thaliana* genome. The sequences with homology to *Mbo*-D22 were annotated on some of the BAC sequence entries as retroelement-like sequences.

Mbo-D22 has been used to probe the *Capsella* cosmid *Taq*I and *Sau*3AI libraries. Approximately 120 clones have been identified as hybridising signals. DNA of 16 cosmid clones has been prepared, digested and blotted to be hybridised successively with the *Mbo*-D22 sub-fragment and a PCR product representing the LTR sequences (Long Terminal Repeats) of the *A. thaliana* retrotransposon. Primers (*LTR1-29f* = 5' CGA GTT CCT AGA TCA TCC TC 3', *LTR1-29r* = 5' GAG CAG AAT CGT TAG GGT TTG G) have been deduced from LTR sequences of an *A. thaliana* element located on the chromosome IV (GenBank acc. no. AL161517). Different patterns of hybridising fragments have been obtained among the cosmids following the *Mbo*-D22 and *LTR1-29f*/*LTR1-29r* hybridisation. Twelve clones showed homology to the LTR probe, whereas all cosmid clones did hybridise with *Mbo*-D22 (Mbulu 2000). Three cosmids clones have been chosen for further analyses, cos-T16, cos-S20 and cos-T32. The element harboured in cos-S20 has been sequenced (Mbulu 2000). An analysis of the restriction pattern of cos-T16 and cos-T32 showed that they were representing the same *Capsella* genomic fragment. The T32-element was then chosen for sub-cloning and subsequent sequence analysis, to permit a comparison between the T32-element and the previously characterised S20-element. The cosmid has been sub-cloned with restriction enzymes *Xba*I, *Hin*dIII and *Eco*RI; fragments of 6410 bp, 2719 bp and 3897 bp, respectively were generated. The *Hin*dIII sub-clone was found to reside within the *Xba*I sub-clone. From the *Xba*I and *Eco*RI sub-clones, additional sub-clones - referred to as deletion clones - have been generated as depicted in Figure 21. All sub-clones were sequenced first using vector-specific primers. This strategy generated numerous anchor points for deducing additional primers for sequencing. Thus, the sequence of the entire element could be generated faster then by solely relying on primer-walking on large cloned fragments.

BLAST searches carried out with the sequences of the S20- and T32-elements revealed matches with ~30 sequences in the *A. thaliana* genome. These sequences were in some cases annotated as *Del*-like retrotransposon. The *Del* element has been originally characterised in *Lilium Henryi* (Sentry and Smyth 1989; Smyth *et al.* 1989).

3-3-2-3 Characterisation of the S20-element

The S20-element spans 7768 bp and is flanked by a target site duplication of 5 bp (TGTAA). The entire sequence of the *Capsella* S20-element is 52,3% A+T rich, if only the inner segment is taken into account, the sequence is 48,9% A+T rich. The 5'LTR spans 1070 bp and the 3'LTR 939 bp.

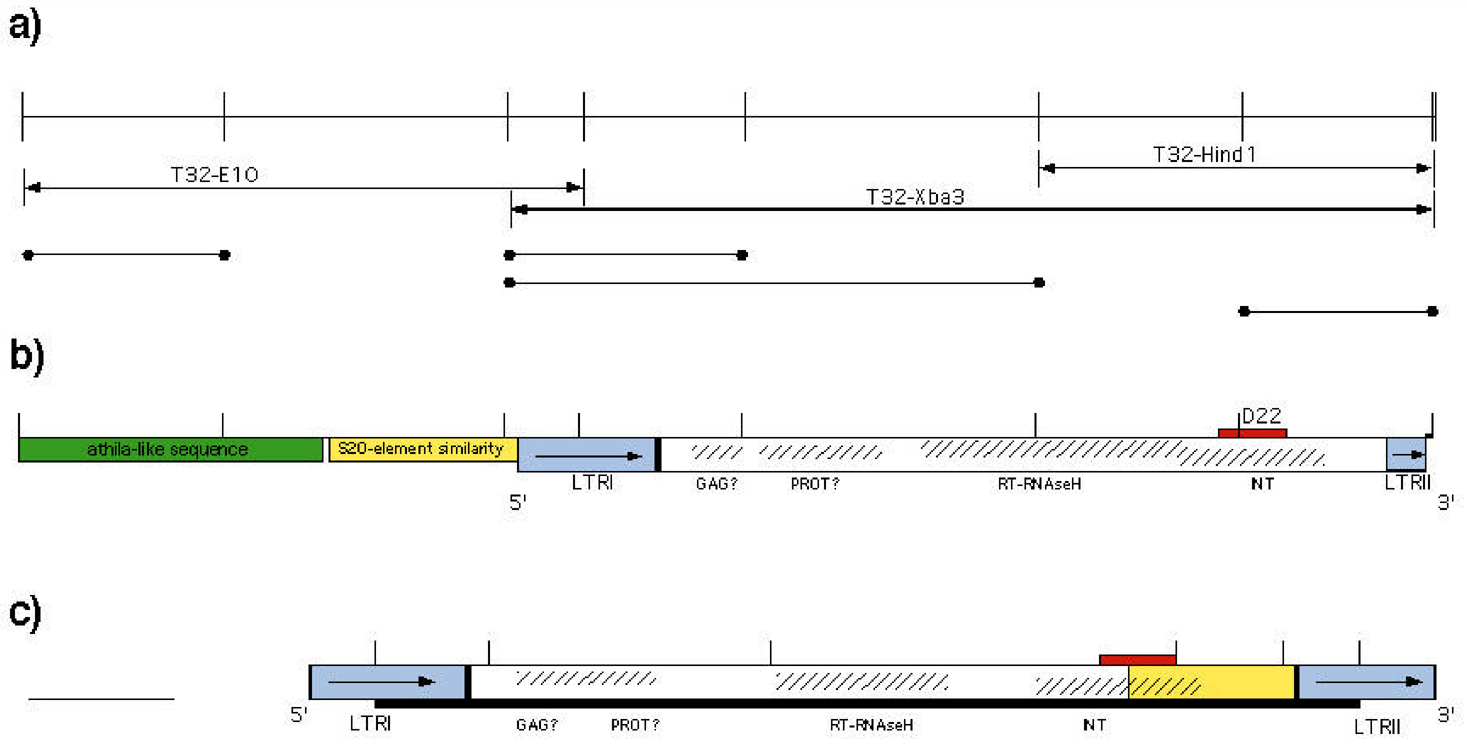


Figure 21: Organisation of two *Capsella* retroelement sequences. a) The three T32-cosmid sub-clones (lines with arrows) T32E10, T32Hind1 and T32Xba3 are shown in respect to the sequence contig of the T32-cosmid. Positions of the recognition sites of enzymes used for sub-cloning are indicated. Lines flanked by filled circles are deletion clones derived from sub-clones T32E10, T32Xba3 and T32Hind1, respectively. b) and c) are showing the T32 and S20 element organisations at the sequence level: blue boxes are the LTRs and large white bars represent the internal segments of the elements. Hatched lines indicate the core domain homologies found with other retrotransposons. Small black bars flanking the LTRs are the potential PBS and PPT sequences. Sub-clone T32E10 showed sequence homology to an athila element (X81801), highlighted in green. Sequences corresponding to the D22-subclone are indicated in red. The large black bar located below the S20 element reflects the SE3 fragment used as probe in hybridisation studies (Figure 31).

Insertions/deletions (indels) are not distributed all over the two LTR sequences, rather the length difference of 131 nucleotides between both LTRs is due to a most likely single indel. Sequence identity between the LTR sequences is approximately 96%.

Three nucleotides downstream of the 5’LTR, a putative primer binding sequence showing complementarity to the methionine tRNA initiator (tRNA$_i^{Met}$) is found. Immediately adjacent to the 3’LTR is an A/G rich sequence, a putative polypurine tract (PPT). Figure 22 depicts the flanking regions of the inner segment directly adjacent to the LTRs (TG…CA borders).

```
                      PBS                                  PPT
5’LTR 1077  CAATTTGGTATCAGAACATTTACGGTT 1103.6822 TAGTGGGGGAGAATTG 6837 3’LTR
                 |||||||||| |    | ||||
              3’-ACCAUAGUCUCG G  U CCAA-5’
                 3’end of tRNAiMet
```

Figure 22: DNA sequences of the internal segment of the *Capsella* S20-element. The sequences adjacent to the LTRs show features characteristic for primer binding sites necessary for DNA synthesis.

3-3-2-4 Characterisation of the T32-element

The T32-element, spans from the 5’ LTR to the 3’ LTR 6298 bp, the 5’ LTR encompasses 1376 bp and the 3’ LTR has been truncated due to cloning into the cosmid vector, only 300 bp of it are present on the T32-cosmid clone. The 3’ end of the 3’ LTR is therefore missing in the obtained sequence. Furthermore, the 5’ end of the 3’ LTR as well as a putative polypurine tract is lacking due to an internal deletion of 1272 bp of the T32-element with respect to the S20-element. The complete 5’ LTR of the T32-element has then been compared to the LTRs of the S20-element. Di-nucleotides characteristic for LTR borders (TG…CA) could be pinpointed for the T32-5’ LTR as well as for the LTRs of the S20-element. The larger size of the T32-5’ LTR compared to the S20-element is explained by an insertion of 296 bp which took place 131 bp upstream of 3’end of this LTR.

The internal segment flanking the 5’ LTR exhibits properties of a putative priming site for DNA synthesis. A *Capsella* sequence of 22 nucleotides is complementary to the published consensus sequence of the 3’ end of tRNA$_i^{Met}$ (Sentry and Smyth 1989; Smyth *et al*. 1989). This sequence is separated by the triplet ATT from the sequence of the 5’ LTR. The comparison of the tRNA$_i^{Met}$ consensus sequence and the *Capsella* sequence is shown in Figure 23.

```
5’LTR    1375 CAATTTGGTAGTAGAGCATTTACGGTT 1401
                   |||||  |||||    | ||||
                 3’-ACCAUAGUCUCG G  U CCAA-5’
                    3’end of tRNAiMet
```

Figure 23: DNA sequence of the internal region of the *Capsella* T32-element. The sequences adjacent to the 5’ LTR correspond to a putative priming site of $tRNA_i^{Met}$.

3-3-2-5 Sequences flanking the *Capsella* elements

Sub-clones of cosmids S20 and T32 harbouring the elements also provided information about *Capsella* sequences flanking the retrotransposons. These sequences from the S20 cosmid as well as the T32 cosmid have been aligned with the sequence of the *A. thaliana* genome. For the part of sub-clone T32-E10 (Figure 21) which is not corresponding to the retroelement analysed here, homologous sequences were found. These corresponded to an athila retrotransposon-like element. The sequence alignment showed a sequence identity of 60% over 2158 bp, 1600 bp of which correspond to ORF2 of the athila element (GenBank acc. no. X81801, Pélissier *et al*. 1995).

3-3-2-6 Sequence analysis of the *Capsella* elements

The *A. thaliana* element family which shows homology to the S20- and T32-elements is not characterised, only the LTRs and the putative reverse transcriptase are occasionally annotated as *Del*-like retrotransposons. In order to characterise the elements further and to determine the retroelement family to which the *Capsella* elements might belong to, the sequences were analysed for the presence of conserved domains. For this, BLAST searches with different capabilities were used (Atschul *et al*. 1997; http://ncbi.nlm.nih.gov).

The complete nucleotide sequence of the S20-element was used for a BLAST alignment with the whole non-redundant database, without any organism selected. The homologies obtained were with *A. thaliana* and other organisms as different as pineapple, rice, maize or lily. Figure 24 summarises the results. The homologies cluster in a region of 2600 bp of the *Capsella* element.

An alignment of the DNA sequence of the S20-element with the *Del* retrotransposon which was identified in *Lilium Henryi* (GenBank acc. no. X13886, Smyth *et al*. 1989) yielded nucleotide identity of ~59 % over 2624 bp. Following the annotations of the *Del*-retrotransposon, these 2624 bp are including a reverse-transcriptase motif, an RNAse H motif, a zing finger motif and an integrase motif. This *Del*-element does contain a single Open Reading Frame (ORF) coding for all different proteins. The corresponding region of the S20-element is located between bp 3300 and 5900 (Figure 24).

The internal segments of both *Capsella* elements have been translated in the six possible frames and submitted into search for conserved domains (BLAST, Conserved Domain Database; Altschul *et al*. 1997). Core domains could be detected on two separate ORFs for both elements. A significant homology could be detected to a reverse-transcriptase domain (RNA-dependent-DNA-polymerase, Pfam00078, E-value 1e^{-33}) and downstream of this match an integrase domain could be identified (Pfam00665, 5e^{-18}). The *Capsella* element can therefore be grouped into the *Ty3/gypsy* family of elements, since the INT ORF is located downstream of the RT ORF.

The two conserved domains are coded for by different open reading frames. A third ORF of 330 amino acids, upstream of the RT is found on the S20-element but does not match any conserved sequences. Nonetheless, it has homology with *A. thaliana* sequences, which are annotated as polyprotein regions of retroelement-like sequences. These sequences could code for GAG, the most divergent gene of a retroelement (Wright and Voytas 1998) or for a protease. Interestingly, some RNase H motifs determined from the sequence published in Jordan and McDonald (1999) are found downstream of the RT but in a different frame. Due to the presence of several stop codons this region is not shown as an ORF. No large ORF downstream of the region with homology to integrase could be shown, therefore, this element is supposed to belong to the *Ty3/gypsy* sub-class of the Metaviridae family, lacking any putative ENV gene.

In the *Capsella* S20-element, the INT core domain sequence contains a stop codon 154 amino acids after the putative methionine start codon. This stop codon is absent from the T32-INT core domain and ORF. Thus, the region corresponding to the INT core domain is represented by a large ORF in the the T32-element and two shorter ORFs in the S20-element. Figure 25 shows both *Capsella* elements T32 and S20 in the 5’→3’ orientation. The different genes discovered via presence of core domains and large ORFs are depicted inside open bars representing the three alternative frames. Amino acid sequence identities have been indicated. Segments of conserved sequences appear to be larger than the putative ORFs.

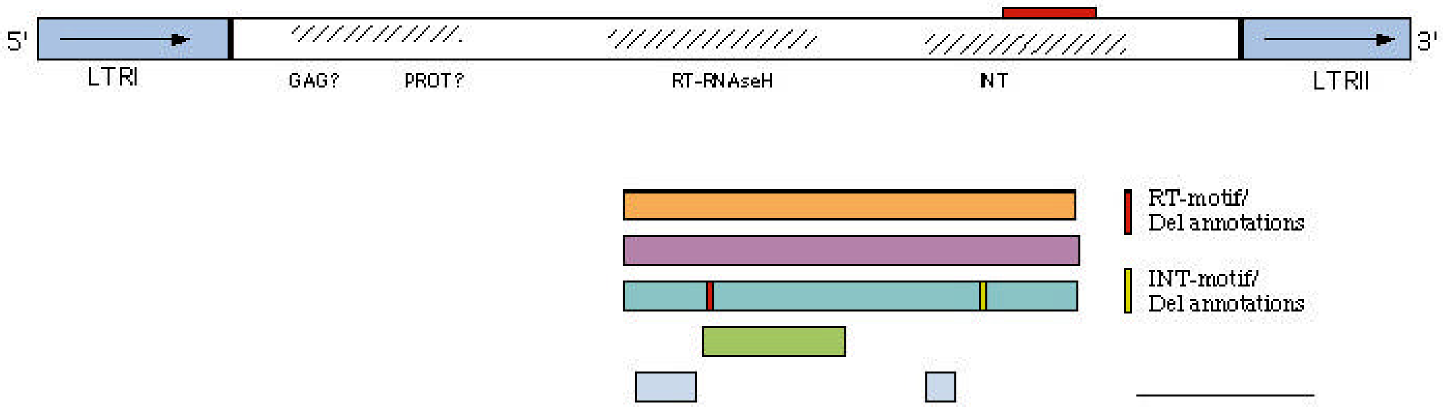


Figure 24: The figure shows regions of sequence homology between the S20-element (shown at the top) and retroelements from other species (listed below). The scheme of the S20 element corresponds to the drawing shown in figure 21. Elements from different species are exhibiting homology to the putative reverse-transcriptase and integrase domains. Alignment with the yeast element Ty3 shows a short but significant homology. Sequence comparisons were performed at the nucleotide level.

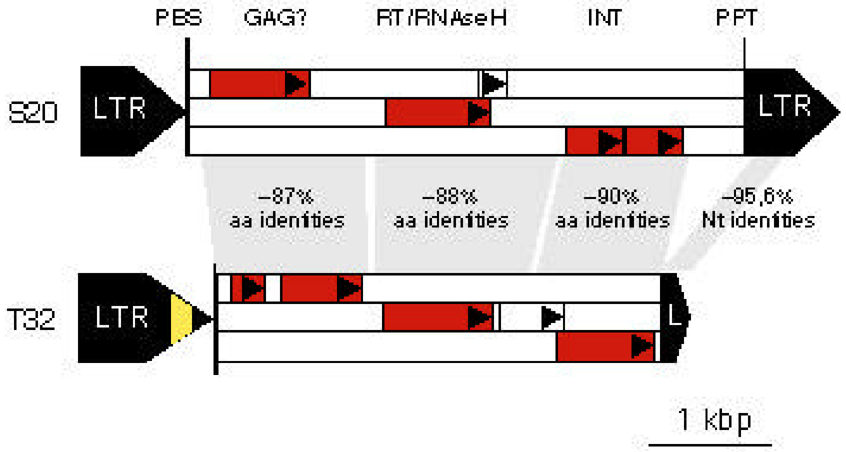


Figure 25: Depicted above are the ORFs and conserved motifs which could be detected for the three frames of translation for both *Capsella* elements. The LTRs are drawn as arrows to indicate their arrangement in a direct repeat manner. ORFs determined in each element are shown as red boxes with an arrow head showing the direction of transcription. Homology to RNAseH, localised by motif homology is depicted as open boxes. The element-wide open bars represent the three possible forward frames of translation. The yellow box inside the 5' LTR of T32 demonstrates an insertion event, and the L in the 3' LTR of the same element exemplifies that LTRII of the T32 element has been truncated by a deletion. The grey shading linking both elements depicts the homology between the S20 3' LTR and the remnant of T32 3'LTR on nucleotide level, and the level of amino-acid sequence identities of putative ORFs, and conserved domains.

3-3-2-7 Sequence comparisons of the *Capsella* elements

3-3-2-7-1 Sequence alignment of the *Capsella* elements

The sequence of the T32-element has been aligned with that of the S20-element using the Bestfit program (GCG). An overall sequence identity of 96% was found along 6000 bp. This corresponds to the entire sequences of the T32-element. Another comparison was performed with the “BLAST two sequences” tool (Tatusova and Madden 1999) which compares two large sequences to each other. The result of this alignment is shown in Figure 26. LTR sequences are homologous within and between elements. The sequence alignments indicate two large insertions/deletions. A large deletion in the T32-element is located around the junction of the internal sequence and the 3’ LTR if compared to the S20-element. This alteration spans 1273 bp and involves the untranslated region after the putative INT ORF, the PPT and the first 72 bp of the 3’ LTR. This deletion will be analysed further below. A second minor event differentiates the LTRI sequences of the *Capsella* elements. LTRI of the T32-element is increased in size by 300 bp compared to the sequence of the S20-element. Segments representing the RT-RNAse H as well as the LTR sequences display sequence identities >95% if the *Capsella* elements are compared (Figure 26).

3-3-2-7-2 Analysis of the deletion

A fragment of 1272 bp seems to be deleted from the 3’-part of the T32-element when compared to the S20-element. Two primers corresponding to sequences located to each side of the deletion on the T32-element were used for PCR experiments. The expected sizes of the PCR products were ~300bp and ~1500bp for the T32-element and S20-element respectively.

The *Capsella* cosmid libraries cannot only be screened by colony hybridisation experiments but also by PCR (Schmidt *et al*. 1999). Fifteen pools of DNA have been prepared which together encompass all cloned *Capsella* genomic DNA sequences.

These 15 DNA pools were used for PCR amplifications with these two primers flanking the deletion of the T32-element. For six pools amplification products could be clearly detected. The PCR products, separated on a 0,8% TAE agarose gel are shown in Figure 27. The size of the amplification products obtained for DNA pools C4 and C14 are indicative of copies which do not contain the deletion.

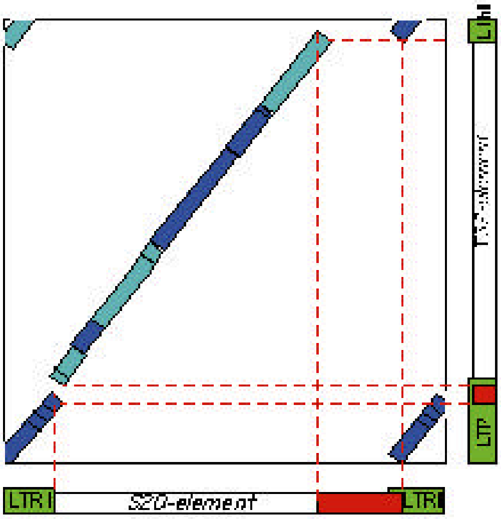
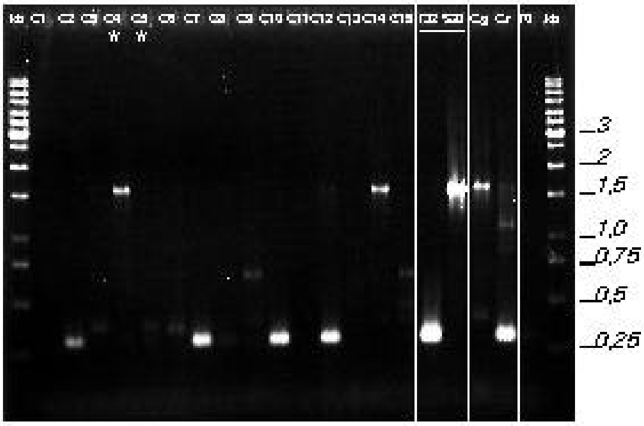
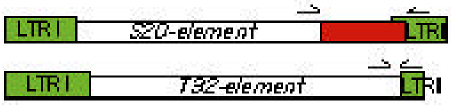


Figure 26: BLAST two sequences (http://www.ncbi.nlm.nih.gov/blast/bl2seq/b12.html) alignment of the two *Capsella* retrotransposons. On the X-axis, the S20 element is represented and T32 is shown on the Y-axis. Breaks of collinearity are indicated as red dashed lines. Red boxes inside the retrotransposons represent insertions/deletions events. Blue bars depict levels of sequence identity, here dark blue is >95% sequence identity whereas light blue shows <95% sequence identity.

Figure 27: The large deletion affecting the 3' LTR of the T32-element is present several times in the *C. rubella* genome. Shown aside is a PCR analysis of *Capsella* genomic and cosmid clone DNA using primers flanking this deletion. Lanes C1-C15 show the amplification result for DNA pools of the *Sau*3AI and *Taq*I cosmid clone libraries. T32 and S20 cosmid DNAs are used as controls for the PCR experiment and are underlined. In the next lanes, DNA of the two parents of the *Capsella* mapping population (Cg, Cr) have been analysed as well as a negative control. On the left side the size marker is shown (kbp). The asterisks beneath C4 and C5 indicate that cosmids S20 and T32, respectively are contained in these DNA pools.





Schematic representation of the S20- and T32-elements. Positions of the primers used for the PCR amplifications are shown. The drawing is not to scale.

On four DNA pools (C2, C7, C10 and C12), in contrast, fragments were amplified which corresponded to the deleted version of the element. The amount of amplification products generated in such experiments is depending on the representation of a particular cosmid clone in the analysed pool. Cosmid clones growing poorly are generally under-represented in the DNA pools and can therefore escape detection. For example, cosmid T32 is part of DNA pool C5, but no amplification product has been seen.

It is interesting to see that in *C. grandiflora* DNA, only the undeleted class of elements is detected, whereas *C. rubella* sequences represent both classes of elements. The *Capsella* cosmid libraries have been established with *C. rubella* DNA, thus the results for *C. rubella* genomic DNA and the DNA pools representing the cosmid libraries are coherent. Amplification products of other sizes can also be noticed in some DNA pools.

3-3-2-7-3 Sequence composition of the *Capsella* elements

The nucleotide composition of the elements has been analysed to determine whether some differences exist along the putative mobile element concerning the predominance of bases. The analysis has been carried out on the GCG package with the "composition" function. These data are reported in Table I. It can be seen that a high A+T content is characteristic for the LTR sequences, whereas the G+C content is higher in the inner segment. The LTRs are composing 27% of the S20-element and 31% of the T32-element.

		S20-element		*T32-element*	
Size		7768 bp		6298 bp	
LTR	*LTRI 5'*	1070 bp		1376 bp	
size	*LTRII 3'*	939 bp		300 bp†	
		A/T	*G/C*	*A/T*	*G/C*
Complete element		52,2%	47,7%	51,4%	48,6%
LTR 5' (LTRI)		41,4%	36,5%	62,6%	37,4%
LTR 3' (LTRII)		63,2%	36,8%	-	-
Inner segment ††		48,3%	51,6%	47,2%	52,7%

Table I: Comparison of the nucleotide composition of two *Capsella* retrotransposons, the T32- and S20-elements. †: the complete sequence of LTRII is not available, thus it has not been analysed. ††: the inner segment corresponds to the sequence between the LTRs.

3-3-3 *Arabidopsis Del*-like retroelements

3-3-3-1 Size of the elements and locations in the *Arabidopsis thaliana* genome

The family of *A. thaliana* elements homologous to the *Capsella* retroelement-like sequences has been analysed. Using the S20-element sequence for a BLAST analysis with the sequence of the *Arabidopsis* genome, ~30 different BAC clones showed significant homologies (E-values between 0.0 and 1e^{-115}).

For 22 different elements, it could be established that they contain two LTRs flanked by a TSD. They span between 7500-8300 bp including LTRs of about 1100 bp (Table J). Two elements are considerably enlarged, they encompass 10,133 bp (GenBank acc. no. AC73433) and 14,039 bp (GenBank acc. no. AL161509), respectively.

	GenBank acc. no.	*Chr. no.*	*Size (bp)*	*5' LTR (bp)*	*3' LTR (bp)*	*Average identity (%)*	*TSD*
1	AB011478	V	7844	1154	644	97,5	$\mathrm{GA^{T}/_{C}T^{C}/_{T}}$
2	AC002534	III	8278	1163	1157	96	ATATC
3	AC005398	II	7831	1158	1163	96,7	ATATT
4	AC006228	I	8195	1147	1142	95	CTAGG
5	AC006955	II	7848	1144	742	97,5	ATTAG
6	AC007203	I	7583	1166	1134	94	$\mathrm{C^{G}/_{A}AAC}$
7	AC007399	V	7946	1145	1132	95,5	GTTAC
8	AC018660	V	8046	1158	1128	91,8	ATAAG
9	AC021199	I	8295	1154	1156	97,5	TAAAT
10	AC025782	I	7900	1157	1155	97	TCTAC
11	AC069557	V	8261	1155	1153	94,6	GAAGT
12	AC073433	I	7959	1162	1155	92	$\mathrm{^{A}/_{G}GTTG}$
13	AF058825	V	8434	1161	1682	92,5	GAAAT
14	AF077407	V	7779	1154	1053	94,6	CAAAG
15	AF262041	V	8266	1153	1147	99	ATTTG
16	AL161508-I	IV	7066	1158	1159	94,5	GAATG
17	AL161510	IV	8062	1155	1156	97,3	CTCTT
18	AL161517	IV	7856	1156	1157	97	CAAAC
19	AP001296	III	8251	1157	1154	97,9	ATTTC
20	AP001301	III	8103	1154	1154	95,5	GTCTT
21	AP002043	III	7642	1053	1058	99,3	AAAGG
22	AP002058	III	8278	1154	1153	97	GGGAG

Table J: List of 22 *Arabidopsis Del*-like elements homologous to the *Capsella* T32/S20-elements. The different *Arabidopsis* chromosomes carrying such elements are reported (*Chr. no.*). The GenBank accession numbers (*Acc. no.*) of the sequenced BACs which carry those elements are given. In the case of the chromosome 4 elements, the accession numbers reflect the sequence assemblies as provided on the MIPS web site (http://mips.gsf.de/proj/thal/db/). The two shaded lines indicate the two elements exhibiting the highest degree of conservation between their LTRs. The TSD is defined as conserved direct repeat sequences immediately flanking the element.

Sequence conservation between the two LTRs of a particular element is varying. LTRs of elements 8 and 12 are more diverged (Table J). Interestingly, the latter element also has a point mutation affecting the TSD. Other elements (15 and 21) exhibit more conserved LTRs (>99% sequence identity) (Table J). The TSDs have been also listed in the Table 3-3-B but no particular preference for an integration site could be determined. With the exception of the first element listed, nucleotide differences within the TSD are found for elements which exhibit a rather low conservation of the LTRs.

The BAC clones containing the retroelement-like sequences have been localised on the *A. thaliana* chromosomes using clone contig information available in the TAIR (TAIR: http://www.arabidopsis.org) and MIPS (http://mips.gsf.de/proj/thal/db/) databases. The size of the chromosomes arms are the ones determined by AGI project (The *Arabidopsis* Genome Initiative 2000). Members of the retroelement family are found on all *A. thaliana* chromosomes (Figure 28). In general, the retroelements show a clustering in centromeric regions with the exception of chromosome II.

3-3-3-2 Sequence comparison of the *Arabidopsis Del*-like retrotransposons

	Acc. no.	*LTRI sequence 5'-end*	*LTRI sequence 3'-end*
1	***AB011478***	**TG** TAACGCCCGTGAACCAGAAAA	AAAAAATGAGTCGGGTTGTTT **CA**
2	***AC002534***	**TG** TAACGCCCGTGAACCGGAAAA	AAAAAAAAGGTCGGGTTGTTA **CA**
3	***AC005398***	**TG** TAACGTCCGTGAACTGGAAAA	AAAAAATGGGTCGGGTTGTTT **CA**
4	***AC006228***	**TG** TAACGCCCGTGAACCGGAAAA	AAAAAATGGGTCGGGTTGTTT **CA**
5	***AC006955***	**TG** TAACGCCCGTGAACCGGAAAA	TTTAAATGGGTCGGGTTGTTT **CA**
6	***AC007203***	**TA** TAACGCCCGTGAACCAGAAAA	AAAAAATGGGTAGGGTTGTTT **CA**
7	***AC007399***	**TG** TAACGTCCGTGAACCGGAAAA	AAAAAATGGGTCAGGTTGTTT **CA**
8	***AC018660***	**TG** TAACACCCGTAAATAGAAAAA	AAAAAATGG-TCGGGTTGTTT **CA**
9	***AC021199***	**TG** TAACGCCCGTGAACCCGAAAA	GAAAAACGGGTCGGGTTGTTT **CA**
10	***AC025782***	**TG** TAACGCCCGTGAACCGAAAAA	AAAAAATGGGTCGGGTTGTTT **CA**

Table K: The sequences at the 5' and 3' termini of 5' LTRs are given for 10 retro-elements, numbered 1 to 10. All sequences are shown from the 5' end to the 3'end. The TG…CA borders are in bold. The results for the 3' LTRs are similar (data not shown).

Sequence alignments were carried out for a subset of 10 *Arabidopsis* elements to pinpoint deletions or insertions within the elements. The element located on BAC T32N15 (GenBank acc. no. AC002534) has been chosen as reference to be aligned with all different sequences shown in Table K using the CLUSTAL W program (http://www2.ebi.ac.uk/clustalw/ or http://www.clustalw.genome.ad.jp/). This element has been selected on the basis of its size which is most likely representing the average for complete elements. The alignment of the 5' and 3' termini of different LTRs shows that LTRI and LTRII have conserved sequences for about ~20 bp inside the LTR, then an A+T rich sequence of different length disrupts the alignment.

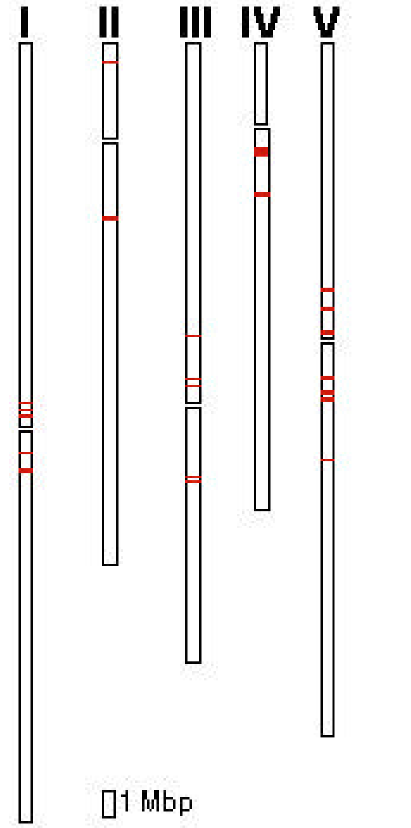


Figure 28: Distribution of the *Del*-like retroelements on the five *Arabidopsis* chromosomes. The size of the chromosomes are according to AGI values. Each of the red lines indicates one of the 22 retroelements listed in the Table J. Chromosome numbers are given above the chromosome arms shown as open bars.

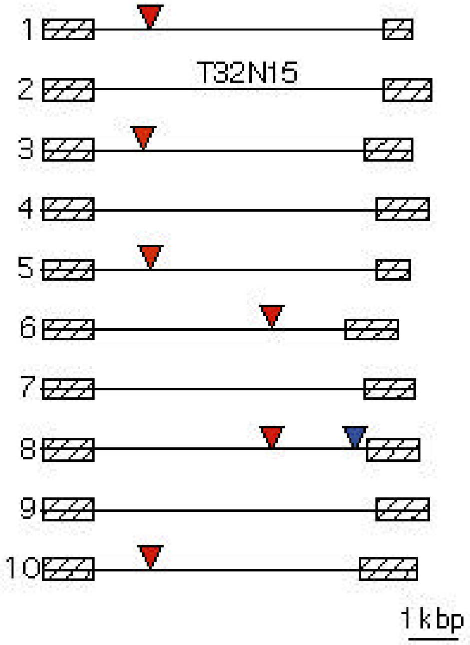


Figure 29: Schematic representation of insertions/deletions found in the analysed *Arabidopsis* retroelement family. Positions of deletions are given in respect to T32N15 (2) and are shown as red triangles. The LTRs are drawn as hatched boxes in the 5'-3' orientation. The blue triangle signifies a large insertion (>500bp) found on element 8.

Apart from few exceptions, LTRs are bordered by a di-nucleotide inverted repeat (TG…CA). Using these criteria LTRs can be defined with high accuracy. Out of 10 LTRs analysed in detail, only one element (6) exhibits a point mutation where the **TG** of the inverted repeats bordering the LTR has been changed to **TA**.

A similar kind of alignment has been performed for the putative PBS and PTT sequences. These sequences are representing the priming binding sites required for DNA synthesis. The PBS is a sequence expected to be complementary to the 3'end of the host tRNA$_{i}^{Met}$ which is used as primer for the synthesis of the (-) DNA strand. The PPT is important for the synthesis of the (+) DNA strand. The *Arabidopsis* sequences are listed in the Table L, as well a consensus sequences deduced from the analysis of the ten *Arabidopsis* elements and the sequences established on both *Capsella* elements T32 and S20.

	Acc. no.	*Primer binding site (PBS)*		*Polypurine tract (PPT)*	
1	*AB011478*	*ATT*	TGGTATCAGAGCGATTACGGTT	TAGTGGGGGAGAAT	**TG**
2	*AC002534*	*ATT*	TGGTATCAGAGCGATCACGGTT	TAGTGGGAGAGAAT	**TG**
3	*AC005398*	*ATT*	TGGTATCAGAGCGATTACGGTT	TAGTGGGGGAGAAT	**TG**
4	*AC006228*	*ATT*	TGGTATCAGAGCGATCACGGTT	TAGTGGGGGAGAAT	**TG**
5	*AC006955*	*ATT*	TGGTATCAGAGCGATTACGGTT	TAGTGGGGGAGAAT	**TG**
6	*AC007203*	*ATT*	TGGTATCAGAGCGATCACAGTT	TAGTGGGGGAGAGT	**TG**
7	*AC007399*	*ATT*	TGGTATCAGAGCGATTACGGTT	TAGTGGGAGAGAAT	**TG**
8	*AC018660*	*ATT*	TGGTATCAGAGCGATCACGGTT	TAGTGGGGGAGAAT	**TG**
9	*AC021199*	*ATT*	TGGTATCAGAGCGATTACGGTT	TAGTGGGGGAGAAT	**TG**
10	*AC025782*	*ATT*	TGGTATCAGAGCGATTACGGTT	TAGTGAGGGAGAAT	**TG**
Consensus Ath		*ATT*	TGGTATCAGAGC$^{G}_{A}$AT$^{T}_{C}$AC$^{G}_{A}$GTT	TAGTGGG$^{G}_{A}$GAGA$^{G}_{A}$T	**TG**
S20-element		*ATT*	TGGTATCAGAACATTTACGGTT	TAGTGGGGGAGAAT	**TG**
T32-element		*ATT*	TGGTAGTAGAGCATTTACGGTT	–	

Table L: Alignment and composition of the potential PBS and PPT sites for the ten analysed *Arabidopsis* sequences. The corresponding *Capsella* sequences are also listed. Due to a deletion the T32-element does not contain a PPT.

The alignment by CLUSTAL W in reference to the element located on BAC T32N15 (GenBank acc. no. AC002534) permitted to observe an overall identity among the elements tested. Nevertheless, insertions and deletions do not seem to be rare events. In Figure 29, the retrotransposons are depicted schematically with remarkable deletions or insertions comprised between 320 and 632 bp marked by arrows. The orientation of the transposons are 5' to 3'. Deletions in elements 1, 3, 5 and 10 are found in the same region but they are of different size. The deletions span ~320 bp, ~400 bp, ~418 bp, and ~410 bp, respectively. Elements 6 and 8 revealed deletions of 632 bp and 585 bp, respectively in the second half of the elements. Element 8 contains an insertion of 576 bp near the 3’ LTR. Indels of few base pairs in size are frequently found in the pairwise alignments of all elements.

3-3-3-3 Comparison of the *A. thaliana* and the *Capsella* elements

Alignment of internal segments of *Capsella* S20-element and *A. thaliana* T32N15 (AC002534) has been performed with the “Bestfit” command of the GCG program. An overall sequence identity of 76% has been established. Sections of the compared segments showed higher conservation. LTRI and LTRII were conserved between the elements of both species at 72% and 71%, respectively. An analysis of these two elements has been performed using the BLAST two sequences program, the result is depicted in Figure 30.

Interestingly, a segment of a *Capsella* element, spanning from nucleotide 1271 to 6148 and homologous to the *A. thaliana* sequence segment from 1368 to 6223 showed only four indels of one nucleotide. All other insertions/deletions distinguishing these elements were representing triplets and thus, would not lead to frame shifts. In contrast, in alignments of the LTR sequences many indels could be noticed.

3-3-4 Hybridisation pattern in species of the Brassicaceae family

BLAST similarity searches indicated approximately 30 different *A. thaliana* sequences with homology to the *Del*-like retrotransposons from *Capsella*. The fact that this sequence seems to be preferentially integrated near centromeres may lead to an underestimation of the abundance of this family in the *Arabidopsis* genome, since the centromeric regions have been only partially sequenced.

A Southern blot containing DNA of 14 *A. thaliana* ecotypes, *C. grandiflora*, *C. rubella*, *B. oleracea* var. *alboglabra* and *B. oleracea* var. *italica* has been made. Genomic DNA of the three genera has been digested with *Dra*I and blotted. The Southern blot has been hybridised with a sub-clone of the *Capsella* S20-element which spans 7 kbp of sequence. This fragment is representative of the entire sequence of this element. The results of this hybridisation experiment is shown in Figure 31. The strongest hybridisation has been obtained with *C. rubella* DNA, consistent with the fact that the probe sequence was derived from this species. Hybridisation of various strengths has been observed for DNA of different *Arabidopsis* ecotypes.

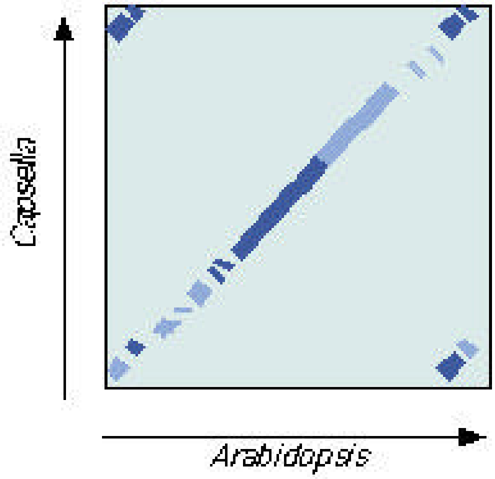


Figure 30: BLAST two sequences alignment of the *Capsella* element S20 and the *Arabidopsis* element T32N15. S20 is represented on the Y-axis, whereas T32N15 is on the X-axis. Dark blue shading indicates a sequence identity >80% whereas light blue corresponds to values <80%.

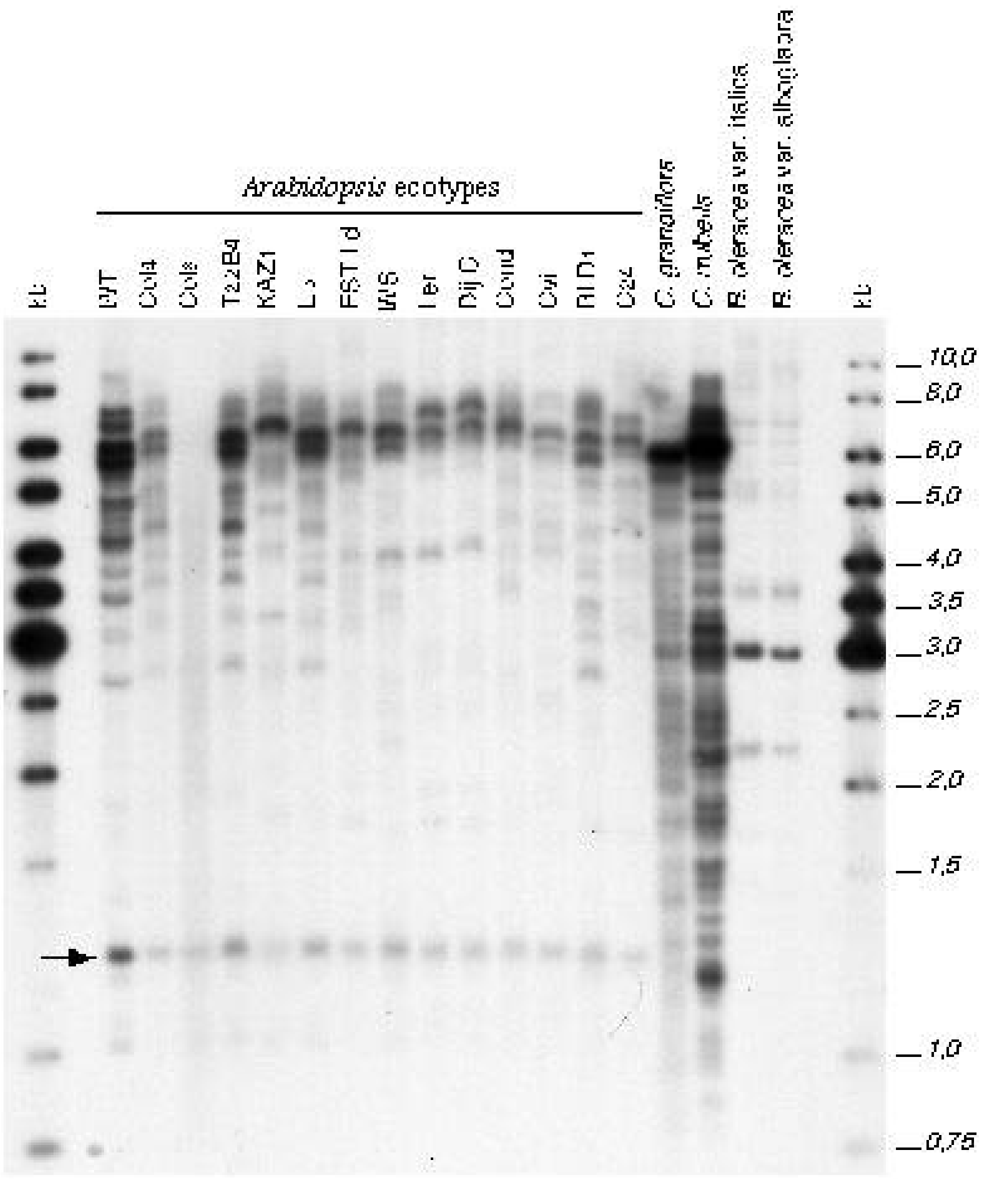


Figure 31: Southern blot of genomic DNA of the three *Brassicaceae* genera studied during this work, *Arabidopsis*, *Capsella* and *Brassica*. The genomic DNA has been digested with *Dra*I. The blot has been probed with the SE3 fragment (Figure 21) which contains sequences representative of the entire S20-element. DNA of Col6 is degraded. The fragment indicated by an arrow corresponds to LTR sequences.

The three ecotypes hybridising most strongly with the *Capsella* element are Columbia wild type (WT), T22B4 and Li5. The *Brassica* species only show faint hybridisation. A small fragment of 1200 bp is clearly revealed in all ecotypes. A similar sized fragment is seen in *Capsella*. To determine which part of the retroelement is corresponding to this fragment, several *Arabidopsis* retrotransposons and the two *Capsella* retrotransposons have been submitted to the “map” command of the GCG program with the *Dra*I enzyme selected. The *Dra*I recognition site is TTTAAA and as noted earlier, the LTRs are very A+T rich, furthermore, repeats of A or T are very frequent close to the LTR borders. Thus, for many retroelements analysed, *Dra*I fragments in the range of 1 kbp could be identified in the LTR sequences.

4 DISCUSSION

4-1 GENETIC MAPPING

4-1-1 The *Capsella* map

In this study, a genetic linkage map of *Capsella* has been constructed, which was derived from an interspecific cross (*C. grandiflora* x *C. rubella*). Fifty F2 individuals were scored for 137 loci and eight linkage groups (LG) covering in total 650 cM could be established (chapter 3-1).

To be able to constitute this map efficiently, a "target mapping" strategy has been used to cover most of the *Capsella* genome with the *Arabidopsis* probes. Markers homogeneously distributed along the *A. thaliana* chromosomes were chosen for the mapping experiments. Particular efforts have been undertaken to use markers mapping close to the telomeric ends (chapter 3-1) and the centromeric regions (Clarenz 2000) of the *Arabidopsis* chromosomes.

All 136 loci for which co-dominant inheritance has been observed could be placed into eight linkage groups. The loci are homogeneously spread along the five *Arabidopsis* chromosomes and also on the *Capsella* linkage groups. On the established *Capsella* map, markers are on average separated by 4,7 cM (Table E, Figure 7).

Markers mapping to four different regions in the *Capsella* genome show frequencies significantly different from the expected 1:1 ratio of the *grandiflora* and *rubella* alleles (Figure 6; Appendix). For a nuclear encoded co-dominant locus, the expected segregation among the F2 progeny is a 1:2:1 ratio of plants homozygous for the *C. grandiflora* allele, heterozygous plants and plants homozygous for the *C. rubella* allele. Some regions in the *Capsella* maps show significant segregation distortion. Markers *Mbo*-L5 and m326A (LG F) delimit a region which shows a significant under-representation of homozygous *C. grandiflora* plants. The same is observed for markers m448A, mi306 and H36452 on linkage group G and for marker mi353 and Z35365 on linkage group B.

C. grandiflora is a self-incompatible species. For the *Capsella* mapping population, only self-fertile individuals have been chosen. Consequently, for a genomic region corresponding to the self-incompatibility locus, it is expected that homozygous *C.*

grandiflora plants are under-represented. Nevertheless, DNA for some individuals of the F2 population which did not produce seed is available. For two markers (*Mbo*-L5, mi323) located in a cluster with significant segregation distortion, DNA of the individuals have been subjected to a segregation analysis. Interestingly, in this part of the mapping population which represents putatively self-incompatible individuals, many plants homozygous for the *C. grandiflora* allele of these two loci could be detected (S. Stegemann and R. Schmidt, unpublished results). Thus, it may be possible in future mapping studies to establish the map position of the self-incompatibility locus in *Capsella*.

4-1-2 Conservation of sequence repertoire between *Arabidopsis* and *Capsella*

The vast majority of *Arabidopsis* RFLP markers and ESTs hybridised to *Capsella* DNA. Among the 63 *A. thaliana* RFLP probes tested, only one (1,6%) did not hybridise to *Capsella* DNA. The mi423a marker sequence corresponds to a *Ty1-copia*-like retroelement. This finding is consistent with data previously obtained. Acarkan *et al.* (2000) could show for another *Arabidopsis* marker which did not hybridise to *Capsella* DNA, that it constitutes the LTR sequence of a retrotransposon-like element. Similar observations were made in grasses. Avramova *et al.* (1996) established that most maize retrotransposon-like sequences do not hybridise to sorghum DNA. In contrast, *Del*-like retrotransposons of *C. rubella* and *A. thaliana* have been found to cross-hybridise (chapter 3-3).

For 23 (62%) of the 37 mi... RFLP markers which have been mapped in *Capsella*, corresponding EST sequences could be found (data not shown). Similarly, 60% of all predicted *Arabidopsis* genes match EST sequences (The *Arabidopsis* genome initiative 2000). Thus, the *Arabidopsis* gene repertoire is well reflected in the set of mi... RFLP markers. The similarity of the gene repertoires of *Arabidopsis* and *Capsella* is indicated by the fact that the vast majority of these markers hybridise to *Capsella* genomic DNA. These results are consistent with studies comparing the genome of *Arabidopsis* to that of different *Brassica* species (Kowalski *et al.* 1994; Lagercrantz 1998)

PCR-based marker systems have also been used in this study. SSCP analysis in particular has been successfully employed for genetic mapping in *Capsella.* This technique exploits different mobilities of denatured DNA strands in MDE™ gels. Even single nucleotide differences in DNA fragments might influence the conformation of the strands when separated on a MDE™ gel and thus, may be detected as a polymorphism

(Slabaugh *et al.* 1997). This method has been found to be very efficient for detecting sequence differences in human genes. For example, the NF1 gene is spanning ~350 kbp of genomic DNA and some mutations in exons lead to an autosomal dominant disorder. Using two alternative PCR-based marker systems, heteroduplex analysis and SSCP, 26 new mutations could be detected in 59 exons of the gene. The SSCP method could reveal 65% of these polymorphisms, clearly demonstrating the versatility of this technique (Abernathy *et al.* 1997).

Another advantage of PCR-based marker systems is that sequences of primers can be selected to restrict amplification to a single member of a gene family. In contrast, RFLP markers do not only detect orthologous sequences but may also reveal paralogues. Especially for large multigene families or repetitive sequences, RFLP marker analysis may result in a hybridisation pattern too complex to be evaluated (Slabaugh *et al.* 1997). It was attempted to map *Capsella Mbo*-fragments corresponding to repetitive DNA sequences of the *A. thaliana* genome using SSCP analysis, but albeit polymorphisms could be detected, the pattern of the segregating fragments was too complex to assign these to loci.

Mbo-C14 corresponds to 18-5,8-25S rDNA sequences, which are present in large tandem arrays (Copenhaver and Pikaard 1996). Many different fragments of almost the same size could be amplified, but it was not possible to discern individual strands upon SSCP analysis (data not shown). *Mbo*-D22 shows homology to approximately 30 different locations in the *A. thaliana* and *Capsella* genomes (chapter 3-3). This fragment is homologous to a putative INT ORF of a retrotransposon-like sequence. Thus, although the copy number of this sequence in the *Capsella* genome is much lower than that of the rDNA sequences, SSCP analysis could not discern scorable polymorphic fragments (data not shown). In the immediate vicinity of one of the *Capsella* elements, other sequences with homology to repeats in the *Arabidopsis* genome were found (Figure 21). The results of the SSCP analysis with a similarly complex pattern as that for *Mbo*-D22, give a strong hint that these sequences may be repetitive in the *Capsella* genome as well (data not shown).

4-1-3 *Arabidopsis* sequence map

The *Arabidopsis* chromosome maps to which this study is referring to are “sequence maps”, their scale is given in Mbp. This tool is available due to concerted efforts made to sequence the *Arabidopsis thaliana* genome (The *Arabidopsis* genome initiative

2000). The spacing between two loci on the sequence map of a particular chromosome is reflecting the exact physical distance between them in bp. The sequence map is advantageous because it is possible to define unambiguously the order of the loci along the chromosome. On a genetic map closely linked markers often cluster at one locus, especially if a small mapping population is used. Acarkan *et al*. (2000) have compared the organisation of *Arabidopsis* chromosome 4 with that of the *Capsella* linkage groups. In this study the molecular marker map of *A. thaliana* chromosome 4 was used for the comparison. Using the sequence map, the comparative map could be considerably refined, since those loci which previously had to be assigned to a single map position on the genetic map could now be ordered in an unambiguous way along the sequence map (chapter 3-1, Figure 7).

Genetic linkage maps are based on recombination frequencies. Comparison of genetic and physical distances along chromosomes has revealed hot and cold spots of recombination. For example, for chromosome 4 of *A. thaliana*, an average value of 185 kbp/cM could be established with a variation from 30-50 kbp/cM for hot spots to >550 kbp/cM for cold spots (Schmidt *et al*. 1995). The sequence map is not prone to these differences. Consequently, markers can be chosen which are equally distributed along the sequences of the chromosomes.

For a number of RFLP markers, multiple loci could be mapped in *Capsella*. All marker sequences were aligned with the sequence of the *Arabidopsis* genome. This strategy allows to pinpoint orthologous as well as paralogous loci on the chromosome sequence maps. In contrast, if markers are used for genetic mapping studies, which correspond to two copies in the genome, often only one locus can be mapped, whereas the other locus is monomorphic. The use of such markers in comparative mapping studies proves to be problematic, because it cannot be established in an unambiguous way whether paralogous or orthologous loci corresponding to a particular marker are being compared (Bennetzen 2000a).

Exon sequences from *Brassica* or *Capsella* are well conserved to the corresponding gene sequences in *Arabidopsis*. A set of 13 orthologous *Brassica* and *Arabidopsis* coding sequences was 87% identical at the nucleotide and amino acid level (Cavell *et al*. 1998). Similar values could be reported for the genes studied in the context of this work (Table H). Comparing exon sequences of *Arabidopsis* and *Capsella*, even higher average values are found (Table H; Acarkan *et al*. 2000; Rossberg *et al*. 2001). Thus, due to the high sequence identity of coding sequences in related cruciferous species, the vast majority of *Brassica* and *Capsella* markers which span exon sequences can be unambiguously placed onto the sequence maps of the *Arabidopsis* chromosomes by aligning the sequence of the marker with the *Arabidopsis* genome sequence. Using this strategy, a molecular marker map established for any cruciferous species can be directly compared to the *Arabidopsis* sequence maps of the chromosomes, as long as sequence information for the molecular markers is obtained. It is not any longer required to carry out laborious genetic mapping experiments in *Arabidopsis*. This strategy has already been proven useful for the *Capsella Mbo*-markers (Figure 7).

The annotated sequence of the *Arabidopsis* genome offers the possibility to specifically target exon sequences for marker studies. It has been possible in many cases to use primer sequences deduced from *A. thaliana* sequences for PCR amplifications in *Capsella*. This could be exploited for the mapping studies (Table D), clearly demonstrating the impact that the information of the *A. thaliana* genome sequence may have for comparative mapping in related species. Due to the particularly high sequence identity of exon sequences in *Arabidopsis* and *Capsella*, the use of degenerate PCR primers is often obsolete. For more distantly related species, EST resources can be exploited to deduce degenerate primers with homology to conserved regions of genes. The resulting primer pairs can be used for the development of PCR-based markers. In this way, genes encoding Calvin cycle enzymes could be mapped in sugarbeet (Schneider *et al.* 1997).

4-1-4 Comparative genetics between *A. thaliana* and *Capsella*

Many comparative genetic mapping studies have been established between related species belonging to the same family. Tanksley *et al.* (1992) have compared the tomato with the potato genomes. They could show a high degree of collinearity, only five inversions spanning entire chromosome arms had to be assumed to explain the differences in chromosome organisation between the species of the Solanaceae family.

Likewise, the comparison of the *Arabidopsis* and *Capsella* genomes revealed a high degree of genome collinearity (Figure 7). In total, 14 large collinear segments have been detected. Most importantly, it could be established that the conserved regions cover the majority of the *A. thaliana* sequence and the *Capsella* linkage map. Some rearrangements between the maps are to be expected, since *Capsella* has eight chromosomes, whereas *A. thaliana* has only five. In *Arabidopsis*, the position of the centromeres is known for all five chromosomes (Schmidt *et al.* 1995; Round *et al.*

1997; Copenhaver *et al.* 1999). Thus, it could be analysed whether the breakpoints of collinearity seen in the comparative map would coincide with the centromeric regions in *Arabidopsis*. In contrast to the result of Tanksley *et al.* (1992), the breakpoints do not systematically involve the centromeric regions of the *Arabidopsis* chromosomes (Figure 7).

Marker IG3 corresponds to 18S-25S rDNA sequences. Using this marker it was possible to place two loci corresponding to rDNA sequences on *Capsella* linkage groups B and F. In *Arabidopsis*, two NORs have been mapped to chromosomes 2 and 4 (Heslop-Harrison and Maluszynska 1994; Copenhaver and Pikaard 1996). Thus, the rDNA loci mapped in *A. thaliana* and *C. rubella* are found in non-collinear arrangement (Figure 7). Since several monomorphic fragments were observed for marker IG3 (Figure 4; Mbulu 2000), it cannot be ruled out that additional 18S-25S rDNA loci exist in *Capsella*. Cytogenetic studies could show whether more than two NORs are found. Analysing *Arabidopsis* ecotypes for the map positions of 5S rDNA sequences, Fransz *et al.* (1998) could also reveal loci in non-collinear locations.

The comparative mapping experiments between *Arabidopsis* and *Capsella* genomes showed evidence for 14 collinear segments, with an average size of more than 40 cM. In contrast, comparison of the *Brassica nigra* genome and that of *A. thaliana* revealed regions of conserved marker order that span 8 cM on average. Approximately 90 rearrangements have to be assumed to explain the differences in organisation between the *Arabidopsis* and *B. nigra* genomes (Lagercrantz 1998). Likewise, 26 rearrangements differentiate the genomes of *A. thaliana* and *B. oleracea* (Kowalski *et al.* 1994). Thus, collinearity in the *Brassica* and *Arabidopsis* genomes is not as pronounced as that seen for the species pair *Arabidopsis* and *Capsella*. The divergence time of *Capsella* and *Arabidopsis* has been estimated at 6,2-9,8 million years ago, whereas the lineages leading to *Brassica* and *Arabidopsis* separated 12,2-19,2 million years ago (Acarkan *et al.* 2000). However, the closer phylogenetic relationship of *Arabidopsis/Capsella* compared to that of *Arabidopsis/Brassica* cannot fully account for the observed differences in collinearity patterns. Thus, it is tempting to speculate that the polyploid ancestry of the *Brassica* species might contribute to the high frequency of rearrangements which need to be invoked to explain the differences in genome organisation between *Arabidopsis* and *Brassica* (The *Arabidopsis* genome initiative 2000; Schmidt *et al.* 2001).

Comparative mapping studies in grasses have shown that extensive conserved linkage segments can be detected in species which diverged as long as 60 million years ago. Furthermore some of the species studied showed an up to 40-fold difference in genome size (Gale and Devos 1998a, b). Moore *et al*. (1995) recognised that a limited number of rice linkage segments is sufficient to describe the marker arrangement of 12 rice, 7 wheat and 10 maize chromosomes. This concept has been very fruitful and allowed to align chromosome maps from many different grass species. A comparative map based on less than 30 conserved linkage segments could be developed. It included the genomes of foxtail millet, oats, pearl millet, maize, rice, sugarcane, sorghum and Triticeae (Gale and Devos 1998b).

Comparing the number of rearrangements distinguishing different grass genomes with that found in the study of *A. thaliana* and *B. nigra*, it is obvious that the frequency of chromosomal rearrangements seen for the cruciferous species is higher than that observed for the Poaceae family (Lagercrantz 1998).

4-2 MICROCOLLINEARITY

4-2-1 Conservation of gene repertoire in orthologous segments of the *A. thaliana*, *C. rubella* and *B. oleracea* genomes

A 50 kbp region of the *Arabidopsis thaliana* genome has been analysed in respect to its gene repertoire and order. The corresponding regions from the *C. rubella* and *B. oleracea* genome could be identified and characterised. Overall the regions exhibit a similar gene repertoire and order, although some deviations from microcollinearity have been detected.

In the *Arabidopsis* region of interest, 16 coding sequences have been described (Figure 14). Experimental evidence is available for 10 of the genes, since cognate cDNA sequences can be found, the remaining six coding sequences have been determined using gene prediction programs. Three of the predicted genes share highly similar ORFs, common exon/intron structure and putative function (At-84838, At-80777 and At-73278). This can be taken as indication that these predicted gene structures are reflecting protein coding sequences. The three copies of the cytochrome P450-like sequences are present in a tandem arrangement in the *Arabidopsis* region. Another gene prediction (At-77106) is also homologous to coding sequences located elsewhere in the

Arabidopsis genome. This may also indicate that this prediction is indeed representing a gene.

The corresponding region of the *Capsella* genome could be identified. Cr-71614 and Cr-73278 are the outermost genes present in the sequenced *C. rubella* region. Thus, the last three genes defined in the *Arabidopsis* region (At-73277, At-82122 and At-82121) are not present on the contig. No attempt has been made to establish the presence of homologous sequences in the *C. rubella* genome. For nine of the 13 *Arabidopsi*s genes or predicted genes, orthologous sequences could be detected in the *C. rubella* region. Among the genes which are apparently absent from the *C. rubella* region, are two copies of the cytochrome P450-like genes. Pairwise nucleotide sequence comparisons of the three copies of the *Arabidopsis* cytochrome P450-like genes and the *Capsella* gene were performed. The results indicate that the *Arabidopsis* copies are more closely related to each other than either of these genes to the copy in *C. rubella* (Table H). This suggests that duplications of the P450-like genes likely occurred in *A. thaliana* after the *Arabidopsis* and *Capsella* lineages separated.

Tandem gene duplications are frequently found in the *Arabidopsis* genome. It has been estimated that 17% of the *Arabidopsis* genes are present in such an arrangement (The *Arabidopsis* genome initiative 2000). In another microcollinearity study, Acarkan *et al.* (2000) could also highlight a recent gene duplication. In the *C. rubella* genome, two genes are present in a tandem fashion, whereas a single copy was present in the orthologous region of the *A. thaliana* genome. Sequence comparisons could establish that the genes have been duplicated in *Capsella* after the separation of the *Arabidopsis* and *Capsella* lineages (Acarkan *et al*. 2000). All these data taken together indicate the frequent occurrence of gene duplications in plant genome evolution.

Genes 71614, 100015, 83424, 86829, 77106 and AAT are present in a perfect collinear arrangement in the *A. thaliana* and *C. rubella* regions. Gene order, orientation and spacing are conserved. These results are concordant with another microcollinearity study recently performed for *A. thaliana* and *C. rubella* (Rossberg *et al*. 2001). Interestingly, it could also be shown, that in the distantly related tomato genome, the five studied genes are present in physical proximity. However, two inversions distinguish the arrangement of genes in tomato genome from those in the two crucifer species.

Microcollinearity analysis in *A. thaliana*, *C. rubella* and *B. oleracea* revealed that in *Brassica* two homeologous loci can be found for the region of interest. This is concordant with most comparative mapping experiments involving *Arabidopsis* and *Brassica*. In the paleopolyploid *Brassica* genomes, the vast majority of regions appear to be present in at least two copies (Kowalski *et al*. 1994; Lagercrantz *et al.* 1996; Sadowski *et al.* 1996; Osborn *et al.* 1997; Cavell *et al.* 1998; Grant *et al*. 1998; Sadowski *et al.* 1998; Lagercrantz 1998). The region around the self-incompatibility locus in *Brassica campestris*, however, appears to be present as a single copy (Conner *et al*. 1998).

The two homeologous segments in *Brassica* differ in respect to gene content (Figure 19). A copy of gene 94311 could be found on the chromosome 1 locus, whereas presence of this sequence could not be validated for the chromosome 7 locus. In contrast to the corresponding regions in *A. thaliana* and *C. rubella*, a copy orthologous to any of the cytochrome P450-like genes could neither be found in the analysed region of chromosome 1, nor on the one of chromosome 7. Another remarkable exception of microcollinearity is the case of predicted gene 86829. Whereas hybridisation studies indicated its presence on both homeologous *Brassica* loci, sequence analysis revealed that only a relic of this gene was present on chromosome 1, putatively a pseudogene (chapters 3-2 and 4-2-2). This clearly shows the importance of detailed comparative sequence analyses. Hybridisation studies reveal the sequence content, but the divergence of sequences cannot be assessed. This problem has been noted in a comparative study of the rice and maize genomes (Tarchini *et al*. 2000). In the context of this study, hybridisation analysis could not clarify whether genes present in the rice genome were absent from the maize genome or whether the divergence of the genes prohibited their detection.

Homeologous segments in the *Brassica* genome often show differences in gene repertoire (Lagercrantz *et al.* 1996; Sadowski *et al.* 1996; Cavell *et al.* 1998; Grant *et al.* 1998; Sadowski *et al.* 1998; Quiros *et al.* 2001). A particularly detailed study highlighted apparent gene deletions, inversions and translocations in homeologous segments of the *B. oleracea* genome. Any one of the homeologous segments differed in respect to gene repertoire if it was compared to the corresponding region of the *A. thaliana* genome. Only the gene repertoire of all homeologous *Brassica* loci taken together made up the gene complement in *A. thaliana.* All homeologous *Brassica* segments were collinear with the counterpart in *Arabidopsis*, with exception of those genes which were missing in one or even two of the triplicated segments (O'Neill and Bancroft 2000).

One has to take in account, nevertheless, the case of high divergence of fast evolving genes which would not be recognised anymore from its putative orthologue, gene loss can be considered as the extreme case of divergence in eukaryotic organisms (Aravind *et al.* 2000).

A patchwork pattern in gene content between duplicated chromosome segments is also observed in the *Arabidopsis* genome. Analysis of the whole genome sequence unveiled that large parts of the *A. thaliana* genome have been duplicated (Bevan *et al.* 1998; Terryn *et al.* 1999; Mayer *et al.* 1999; Lin *et al.* 1999). Blanc *et al.* (2000) estimated that about 60% of the *A. thaliana* genome is present in duplicated segments. Rearrangements are frequently observed and only between 20% and 47% of the genes are in common in duplicated regions (The *Arabidopsis* genome initiative 2000). From these observations, it has been concluded that duplicated segments suffer many alterations which result in differences in gene content. Thus, polyploidy may foster rapid chromosomal evolution (The *Arabidopsis* genome initiative 2000). Further support of selective gene loss from duplicated segments of a genome is given by a study comparing microcollinearity of the distantly related species *Arabidopsis* and tomato (Ku *et al*. 2000). In this context, it is important to emphasise that in the comparison of the *A. thaliana* and *C. rubella* genome, such extensive exceptions from microcollinearity are not seen (chapter 3-2; Acarkan *et al*. 2000; Rossberg *et al*. 2001).

The duplicated segments in the *Arabidopsis* genome are believed to be ancient, since sequence conservation is restricted to exon sequences, whereas intron and intergenic sequences are shown to be highly divergent (Terryn *et al*. 1999). A microcollinearity study between the *A. thaliana* and *B. oleracea* genomes could provide evidence for the presence of the duplicated segment in the common ancestor of the *Arabidopsis* and *Brassica* lineages. This is concordant with the fact that the presence of at least one of the duplicated segments could be proven for the *A. thaliana* and *C. rubella* genomes (Rossberg *et al*. 2001). It is important to note that no evidence for consistent differences in copy-number of markers in the *A. thaliana* and *C. rubella* genomes could be found (Table F). The high degree of collinearity seen at the gross chromosomal level (Figure 7) also confirms that the duplicated segments are predating the speciation of *Arabidopsis* and *Capsella*. Ku *et al*. (2000) estimated that at least some of the duplications in the *Arabidopsis* genome happened approximately ~112 million years ago, consistent with the view of the ancient nature of the duplication.

In the grasses a remarkable degree of genome collinearity at the gross chromosomal level is found, even if species are compared which diverged as along as 60 million years ago and which show several-fold differences in genome size (Gale and Devos 1998). At the microscale, however, many small deviations from microcollinearity are observed. A comparison between orthologous regions of the sorghum and maize genomes showed an overall collinear organisation, which was interrupted either due to gene deletions or translocations. Most importantly, intergenic regions in maize are often enlarged in comparison to those of sorghum. These size increases are caused by the presence of retrotransposons in intergenic regions of maize (Bennetzen *et al.* 1998; Feuillet and Keller 1999; Tikhonov *et al.* 1999). A comparison realised between orthologous regions in rice and maize also highlighted rearrangements and evidence for a gene translocation was presented (Tarchini *et al.* 2000).

An abundance of retrotransposons in intergenic regions such as seen in the maize genome has so far not been revealed if intergenic regions in the cruciferous species are analysed. Only few retrotransposons are observed in the euchromatic regions of the *A. thaliana* (The *Arabidopsis* genome initiative 2000), *C. rubella* and *B. oleracea* genomes (A. Acarkan, M. Rossberg and R. Schmidt, unpublished results). Consistent with these observations, intergenic spacing was found to be very similar in the region under study in the *A. thaliana*, *C. rubella* and *B. oleracea* genomes (Figures 15 and 16). For corresponding segments of the *Arabidopsis* and *Brassica* genomes, different observations have been made. In some cases, regions are of similar size both species, whereas in other instances an increase of size was noted for a *Brassica* region when compared to the *Arabidopsis* counterpart (Conner *et al.* 1998; Grant *et al.* 1998; Jackson *et al.* 2000; O'Neill and Bancroft 2000; Sadowski *et al.* 1996; Sadowski and Quiros, 1998; Schmidt *et al.* 1999; Quiros *et al.* 2001).

Regardless whether species of the Brassicaceae or Poaceae family are compared, microcollinearity studies reveal evidence for small genome rearrangements, such as gene deletions, inversions and translocations (Bennetzen 2000a; Schmidt *et al.* 2001). Especially if one takes into account that rather small regions of the genomes were studied, it is directly apparent that these changes are very frequent. However, they do not interfere with the overall genome collinearity seen in comparative genetic mapping experiments.

4-2-2 Conservation of gene structure

The strategy taken for the comparison of gene structures in this study integrated information derived from alignments of cDNA and genomic DNA sequences and gene predictions by the programs Genscan and GeneMark. The analysis was not restricted to genomic DNA sequences in one species, rather the aim was to exploit the high sequence identity found for exon sequences between different cruciferous species to improve gene predictions (Cavell *et al*. 1998; Acarkan *et al*. 2000; Rossberg *et al*. 2001). A gene prediction was considered to be likely if exon sequences, which are highly similar in length and structure, could be determined for each genomic DNA sequence of the different species analysed. Following this strategy, it was possible to determine all putative gene structures shown in Figure 17. The predicted genes defined in this way differ in several cases from the annotations given for the *Arabidopsis* genome sequence and show the utility of this approach for the improvement of gene structure predictions. It has been previously noted that comparative data can be used for this (The *Arabidopsis* genome initiative 2000; Rossberg *et al*. 2001).

In two cases, however, it was not possible to discern concordant ORFs in the regions of the *A. thaliana*, *C. rubella* and *B. oleracea* genomes which showed homology. Interestingly, for one case, the cognate EST contig 98955 provided experimental evidence that at least, the region of the *A. thaliana* genome was transcribed, nevertheless no ORF could be determined. The region exhibited significant homology to a ribosomal protein gene (data not shown).

Concerning the gene structures defined in this study, *A. thaliana* and *C. rubella* exhibited very few differences. Generally exon length and intron positions are conserved. One notable exception was an enlarged exon 7 in Cr-77106 compared to the copy of the *Arabidopsis* gene. These results are in excellent agreement with previous findings which compared structures of nine different sets of orthologous genes for *A. thaliana* and *C. rubella* (Acarkan *et al.* 2000; Rossberg *et al.* 2001). All but one of the nine genes studied exhibited conservation of exon lengths and intron positions. The exception was a gene putatively coding for a transcription factor. For this gene considerable differences in exon lengths were observed. Furthermore, the coding regions were not as highly conserved as those of other protein coding genes (Rossberg *et al*. 2001).

In general, conservation of gene structures was also seen if *B. oleracea* genes were included in the analysis. However, the AAT gene copy on chromosome 1 has 10 introns, whereas the homeologous copy on chromosome 7 has 11, like the *A. thaliana* and *C. rubella* genes. Differences in intron number are occasionally observed if orthologous genes are compared (Chen *et al*., 1998; Rossberg *et al*. 2001; M. Rossberg and R. Schmidt, unpublished results).

The most striking difference in gene structure concerned the two homologues of gene At-86829 in the *Brassica oleracea* genome. The copy on chromosome 7 appears to have the same exon/intron structure as the *A. thaliana* and *C. rubella* genes, albeit the 5'-end of the gene is not covered by the sequenced region (Figures 16 and 17d). In contrast, on chromosome 1, only a remnant of the gene representing the 3'-end of 86829 could be found. Furthermore, this gene relic could be differentiated from the copy located on chromosome 7 by a number of one bp indels. Thus, not only apparent deletions of complete gene sequences could be found in this microcollinearity study, since the analysis of gene 86829 exemplifies the occurrence of putative pseudogenes in *B. oleracea*. This finding could be corroborated by analyses of several other gene sequences in *Brassica* (M. Rossberg and R. Schmidt, unpublished results).

The degree of sequence conservation for exon sequences of *A. thaliana* and *C. rubella* is always higher than that estimated for *A. thaliana* and *Brassica* (Table M). A similar rate of conservation is depicted regardless if *A. thaliana* or *C. rubella* genes are compared to *Brassica* or whether the homeologous *Brassica* genes are aligned (highlighted as shaded boxes in Table M). Thus, if one can postulate that genes evolve at the same rate in the three species, the two homeologous *Brassica* regions appear as diverged as the *Brassica* regions compared to their counterparts in *A. thaliana* and *C. rubella*. These data are coherent with the more recent divergence of *Arabidopsis* and *Capsella*. The *Brassica* lineage separated from the *Arabidopsis/Capsella* lineage 12-19 million years ago (Acarkan *et al*. 2000).

	Average identity of exon sequences	*Average amino-acid identity*
Ath/Cr	90,6%	90,5%
Ath/Bo	86,2%	85,3%
Cr/Bo	86%	88,5%
Bo1/Bo7	86,1%	86,6%

Table M: Average of exon and amino acid sequence identity between *A. thaliana*, *B. oleracea*, and *C. rubella* genes compared pair by pair. Shaded boxes show the similar rate of sequence conservation between *A. thaliana* / *B. oleracea*, *C. rubella* / *B. oleracea* and both *B. oleracea* loci.

4-3 RETROELEMENT

A 484 bp sequence of *Capsella* genomic DNA (*Mbo*-D22, chapter 3-3) exhibited homology to repeated DNA sequences in the *A. thaliana* genome. Some of these *Arabidopsis* sequences were annotated as *Del*-like retrotransposons. The *Del* transposable element has been originally characterised in lily (*Lilium Henryi*, GenBank acc. no. X13886; Smyth *et al*. 1989; Sentry and Smyth 1989).
Cosmid libraries have been screened to identify copies of these elements from *Capsella rubella*. Of 120 hybridising clones, two were characterised in detail and the elements residing in these cosmids were sequenced. The S20 and T32 elements span 7768 bp (Mbulu 2000) and 6298 bp, respectively. T32 represents a partial copy.

4-3-1 Sequence conservation

The *Capsella* elements, as well as the *Arabidopsis* elements (chapter 3-3) exhibit LTR sizes of 644-1376 bp. The borders of the LTRs are corresponding to the TG...CA di-nucleotide inverted repeats (Table J) typically found in retrotransposons of eukaryotic organisms, plants and animals (Grandbastien 1992).
The degree of sequence identity of pairs of LTRs from different *A. thaliana* elements varies from 91,8% to 99,3%, the average being 96% (Table J). Both LTRs of a unique element are generated from a single template during the process of reverse transcription. The sequence identity of the LTRs of a particular element can thus be taken as a measure for the time-point at which transposition occurred. Identical LTR sequences are hallmarks of recent transposition and divergent LTRs indicate ancient events. Jordan and McDonald (1999) examined the entire genome sequence of *Saccharomyces cerevisiae* for the distribution and the conservation at nucleotide and amino acid level of the five LTR-retrotransposon classes *Ty1-5*. They could show homogeneity in sequence and variation in size among elements. Among 48 *Ty* retroelements analysed, 22 *Ty* elements had 100% identity between their LTRs, 17 had identities >99% and eight had identities of 97,3-98,8%. These results led the authors to conclude that these are recent insertions. Twelve families of retrotransposons have been studied in *Caenorhabditis elegans*, and benefiting from the complete nuclear genome sequence, it was found that all LTR-retroelements displayed LTR sequence identity above 99% (Bowen and McDonald 1999). The average sequence identity values obtained for LTRs of the

Arabidopsis elements argue for insertion events at different time-points. *A. thaliana* elements AP002043 (Table J, element no. 21) and AF262041 (Table J, element no. 15) exhibit LTR sequence identities of 99,3% and 99%, respectively. For eight elements sequence identities of 97-97,9% were found and in 12 elements sequence identities of the LTR sequences ranged from 96,7 to 91,8% (Table J).

The PBSs and PPTs of the *A. thaliana* elements and the one from the *Capsella* S20-element have been compared to each other. The PBS sequences are complementary to the 3' end of the host $tRNA_i^{Met}$. Priming of DNA synthesis by $tRNA_i^{Met}$ is a feature which is observed for many *Copia*- and *Gypsy*-like retrotransposons in different species (Grandbastien 1992).

Based on the relative position of the ORF of the integrase in respect to the ORF for a putative reverse transcriptase the *Capsella* elements have been classified into the *Ty3/gypsy* family of elements. Since no evidence for a putative envelope ORF could be found they are thought to belong to the Metaviridae family. This classification is further supported by the fact that all elements of the Metaviridae-type studied so far have a PBS recognised by $tRNA_i^{Met}$. In contrast, *athila* and *Tat* retroelements, which belong to the Errantivirus family of elements do not match the $tRNA_i^{Met}$ but show putative homology to at least three different tRNA genes (Wright and Voytas 1998)).

The overall sequence identity found between *Capsella* and *A. thaliana* elements is ~74%, but fractions of the elements exhibit a higher degree of sequence conservation. The inner segment of the *Arabidopsis* and *Capsella* elements is 76% homologous (Figure 30). Sequence identity between *Capsella* and *A. thaliana* elements is thus lower than sequence identities determined for exon sequences of protein-coding genes (Acarkan *et al.* 2000; Rossberg *et al.* 2001). Nevertheless, these *Del*-like elements of *Arabidopsis* and *Capsella* have sufficient homology to be detected in cross-hybridisation experiments, whereas it has been shown for many maize repetitive elements that they do not cross-hybridise to sorghum DNA (Avramova *et al.* 1996). Interestingly, the *Del*-like *Capsella* element only poorly cross-hybridised to *Brassica oleracea* DNA (Figure 31). If one considers that protein-coding exon sequences are less conserved between *Capsella* and *Brassica* than between *Capsella* and *Arabidopsis* (chapter 3-2), this result might reflect the divergence rather than the absence of this element family in the *Brassica* genome.

Nevertheless, these elements of the Brassicaceae family do not seem to exhibit the same degree of conservation as it has been established for some domains of transposable elements which are cross-hybridising to DNA of different grasses (Jiang *et al.* 1996; Miller *et al.* 1998). In contrast, a SINE element first characterised in *B. napus*, appears to be well conserved among several cruciferous species, especially the *Brassica* genus, but does not exist in *Arabidopsis* (Lenoir *et al.* 1997).

Interestingly, Langdon *et al.* (2000) could identify members of a *Gypsy*-retrotransposon family called *Crwydryn* in grasses and *A. thaliana*. An important observation has been the high degree of LTR conservation existing in cereals compared to the high variability of the LTR sequences in *Arabidopsis*.

4-3-2 Conservation of element size

The *A. thaliana* elements listed in Table 3-J vary in size. Many deletions/insertions can be observed along the inner segments of those elements. In contrast, few indels were observed in *Ty* elements from yeast, the occurrence of frame-shifts was considered as rare in *Ty* families (Jordan and McDonald 1999). The yeast genome is known to contain many active *Ty* elements in contrast to observations made in plants (Grandbastien 1992). Few plant retrotransposons have been recognised to be active, as for example the retrotransposon *Bs1* in the maize genome, the *Tnt1* element in tobacco and *BARE* in barley (Grandbastien *et al.* 1989; Vicient *et al.* 1999a and b). The frequency of indels together with the disparate identities of the LTRs support the hypothesis that the retroelements in *Arabidopsis* and *Capsella* are not active.

Recombination between LTRs can provoke the loss of a retrotransposon with a solo-LTR staying behind. Solo-LTRs are abundant in the yeast genome, their sequences are highly divergent compared to that of LTRs from complete elements. A mechanism identified in yeast suggests that the host induces intra-element LTR recombination (Jordan and McDonald 1999). Very few solo-LTRs of the *Del*-like retrotransposons have been identified in the *A. thaliana* genome (data not shown).

A comparison of the S20- and T32-*Capsella* elements revealed large insertions/deletions in the LTR sequences. Interestingly, the *Copia*-like retroelement *BARE* shows extraordinarily long LTRs (1,8-1,9 kbp), it is thought that this is a result of many imprecise excisions of the elements. This observation suggests that a large part of the LTR is not important as long as sequences required for promoter activity are conserved (Vicient *et al.* 1999a).

4-3-3 Copy number of the elements

The capacity of retroelements to transpose via a copy of themselves implies that they can be readily amplified within a genome. Thus, they represent an important factor in genome evolution. Few complex genomes have been sequenced in their entirety but the *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and more recently the *Arabidopsis thaliana* genome offer the possibility to examine the distribution of retroelements families (Britten 1997; Bowen and McDonald 1999; Jordan and McDonald 1999; Kumar and Bennetzen 1999; Langdon *et al*. 2000; Wendel and Wessler 2000).

The *A. thaliana* genome spans 125 Mbp and has a transposable element population representing ~10% of the genome (The *Arabidopsis* genome initiative 2000). The majority of the retrotransposons are clustered in the heterochromatin regions of the genome, e.g. the centromeric areas. Few genes are located in these areas, and their expression, although existing seems much reduced (Mayer *et al*. 1999; Lin *et al*. 1999). A study of the transposable elements on chromosome 2 of *Arabidopsis* showed that ~4,3% of this chromosome is corresponding to *Gypsy*-like elements. *Copia*-like elements account for 1% of the chromosome 2 sequence (Kapitonov and Jurka 1999).

In the lily genome 13,000 copies of the *Del* retrotransposons are found (Smyth *et al*. 1989). Much lower copy numbers are observed for *Arabidopsis* elements. *Ta* retrotransposons represent 0,1% of the cruciferous genome (Konieczny *et al*. 1991). *Athila* elements have been estimated at ~150 copies per genome (Pélissier *et al*. 1996), with peri-centromeric localisation. The clustering of *athila* sequences in centromere vicinity is also reflected by the fact that these elements are frequently associated with the highly repetitive centromeric tandem repeat sequence of 180 bp (Pélissier *et al*. 1995).

Twenty-two *Del*-like elements have been identified via sequence homology in the *A. thaliana* genome. In addition elements larger and shorter than the average size observed for most copies have been found. Most of the elements cluster in the centromeric regions (Figure 28). Since the organisation and sequence of the centromeric regions of the *A. thaliana* genome remains to be elucidated, it is likely that more than 30 *Del*-like elements are present in the *Arabidopsis* genome.

The hybridisation pattern of the element in different *Arabidopsis* ecotypes does not show great variation in copy number and some ecotypes display very similar patterns (WT/T22B4/Li5; Figure 31). Thus, it can be hypothesised that these retroelements are common to a number of ecotypes and that their activity ceased before the divergence of the ecotypes. The same observation is obtained with the retrotransposon family *Ta* which may predate *Arabidopsis* speciation. Among the *Ta* retrotransposons, *Ta1-Ta7* have been characterised as closely related to the tobacco *Tnt1* element while *Ta8-Ta10* are more closely related to *Drosophila Copia*-like retrotransposons (Konieczny *et al.* 1991).

A colony library screen revealed that approximately 120 *Capsella* cosmid clones correspond to *Del*-like elements. The libraries used represent a roughly fourfold coverage of the *Capsella rubella* genome (Schmidt *et al.* 1999), this would be consistent with 30 copies of the *Del*-like elements in the *Capsella* genome. Hybridisation and PCR amplification experiments could also confirm the presence of the *Del*-like elements in *C. grandiflora* (Figures 27 and 31).

4-3-4 Distribution of the elements in the genome

It was shown in this study that *A. thaliana Del*-like elements were found preferentially in peri-centromeric regions (Figure 28). Other elements, such as *athila* are also found clustered in these regions (Pélissier *et al.* 1995; Schmidt *et al.* 1995; Thompson *et al.*, 1996a and b). A sequenced region flanking the *Capsella* element T32 showed sequence homology of over 60% on 2000 bp with an *athila*-like element. Many *A. thaliana* BACs, which carry *Del*-like retrotransposons are as well annotated for the presence of *athila*-like elements. Thus, *Del*-like elements and *athila*-like elements are found in close proximity in the *Arabidopsis* genome.

It has not been possible to map the two *Capsella Del*-like retrotransposons. The regions flanking the S20/T32-elements of the cosmid do not carry low copy sequences with homology to the *Arabidopsis* genomic DNA sequence. Therefore, it has not been possible to map these elements *in silico*. It has been attempted to establish the map positions of the *Capsella* cosmid clones by SSCP analysis, but due to the repetitive nature of many of the amplified fragments it has not been possible to establish a map position.

Nonetheless, the presence of *athila*-like sequences in the immediate vicinity of the *Capsella Del*-like element T32 (Figure 21) supports the hypothesis that these elements might be present in peri-centromeric regions in *Capsella*, as observed in *Arabidopsis*. This could be investigated further if the elements would be hybridised to chromosome spreads of this species.

Repetitive sequences are interspersed in some monocotyledonous plants genomes with gene sequences (Bennetzen 1996; SanMiguel *et al*. 1998). These elements seem to have been amplified in successive waves, thus different layers of elements can be detected (SanMiguel *et al.* 1996). The authors analysed nucleotide substitution rates and dated the insertion waves at about six million years and within the last three million years. Such a complex integration pattern has also been observed in the *Arabidopsis* genome. *Tnat1* and *Tnat2*, novel transposable elements from *A. thaliana* are inserted within each other (Noma and Ohtsubo 1999).

Two *Del*-like elements have been identified in *A. thaliana* which are much larger than the average size observed among the 22 elements listed in Table J. In the case of the *Del*-like retrotransposon located on the sequence contig with GenBank acc. no. AC002534, it appears that the increased size is due to the insertion of a different element into the *Del*-like retrotransposon. This element shows homology to *athila* sequences and is flanked by a 4 bp TSD. The analysis of the other particularly large element (Genbank acc. no. AC073433-chr. 1) revealed that this element could be aligned with the reference element T32N15 (GenBank acc. no. AC002534), only minor deletions were noted. The large size of the element located on chromosome 1 could be accounted for by to the presence of a second 3' LTR sequence. The two 3' LTR sequences are present in a direct tandem arrangement. Upstream of both 3' LTR sequences a PPT sequence is found and downstream the TSD (AGTTG). The presence of the same target site duplication, one upstream of the 5'LTR, and one downstream of each of the two 3'LTRs is not consistent with assuming an insertion of a *Del*-like retrotransposon into another one. Rather this arrangement could be the result of duplication. Tandem duplications are frequently found in the *A. thaliana* genome. About 17% of the predicted genes annotated on the *A. thaliana* sequence are found in such an arrangement (The *Arabidopsis* genome initiative 2000).

5 SUMMARY

Three species belonging to the Brassicaceae family, *A. thaliana*, *C. rubella* and *B. oleracea* have been chosen for comparative genome analyses. The diploid species *A. thaliana* and *C. rubella* are more closely related than either of these species is to the paleopolyploid *B. oleracea*.

Genome-wide comparative mapping experiments between *Arabidopsis* and *Capsella* revealed a conserved gene repertoire. The sequence maps of the five *A. thaliana* chromosomes have been aligned with linkage groups established for the eight *Capsella* chromosomes. Fourteen conserved linkage segments cover the majority of the chromosome maps of both species. The number of rearrangements distinguishing the genomes of *Arabidopsis* and *Capsella* is much lower than values observed between the paleopolyploid *Brassica* species and *A. thaliana*.

Comparative physical mapping and sequence analysis between orthologous regions of the *A. thaliana*, *C. rubella* and *B. oleracea* genomes have demonstrated an overall microcollinearity. Consistent with the polyploid ancestry of the *B. oleracea* genome, two homeologous regions could be analysed and compared to the chromosome segments studied in *A. thaliana* and *C. rubella*. The gene repertoires in the homeologous *B. oleracea* regions differ, evidence for an apparent gene deletion was found. Comparison of the *Arabidopsis* and *Capsella* segments indicated a recent tandem gene duplication of a cytochrome P450-like gene. In the homeologous *B. oleracea* regions, such a gene has not been found. Despite these differences observed in gene repertoire between the orthologous regions, order of genes and their orientation relative to each other was maintained. Moreover, exons of orthologous genes were conserved in length and sequence, with the exception of one putative pseudogene in *B. oleracea*. *Arabidopsis* and *Capsella* coding sequences are on average 90% identical at the nucleotide level. In contrast, characterisation of a retroelement-like family in *A. thaliana* and *C. rubella* indicated that these components of the genome are more diverged. Homeologous *Brassica* genes were found to be 85% identical at the nucleotide level. Similar values were obtained if *Arabidopsis* or *Capsella* exons were aligned with *Brassica* sequences. These values reflect the phylogenetic relationship established for these species.

These results taken together, an overall similarity of genome organisation in *A. thaliana*, *C. rubella* and *B. oleracea* is unveiled, despite the fact that duplicated segments complicate collinearity relationships. Consequently, the sequence of the *Arabidopsis* genome can be used as an efficient tool to unravel the genome organisation of related cruciferous plants.

6 ZUSAMMENFASSUNG

Drei Arten der Familie der Brassicaceae, *A. thaliana*, *C. rubella* und *B. oleracea* wurden für vergleichende Genomanalysen herangezogen. Die diploiden Spezies *A. thaliana* und *C. rubella* sind untereinander enger verwandt als mit der paleopolyploiden Art *B. oleracea*.

Genom-weite vergleichende Kartierungsexperimente in *Arabidopsis* und *Capsella* zeigten eine Konservierung des Genrepertoires. Die Sequenzkarten der fünf *Arabidopsis*-Chromosomen wurden mit den Kopplungsgruppen verglichen, die für die acht *Capsella*-Chromosomen erstellt werden konnten. Vierzehn konservierte Kopplungssegmente decken fast die gesamten Chromosomenkarten beider Arten ab. Die Zahl der Umordnungen, in der sich die Genome von *Arabidopsis* und *Capsella* unterscheiden, ist damit beträchtlich niedriger als die Werte, die für paleopolyploide *Brassica*-Arten und *A. thaliana* beobachtet wurden.

Vergleichende physikalische Kartierungen und Sequenzanalysen orthologer Regionen der *A. thaliana*-, *C. rubella*- und *B. oleracea*-Genome konnten Mikrokolinearität nachweisen. In Übereinstimmung mit der polyploiden Herkunft des *B. oleracea*-Genoms konnten zwei homeologe Regionen analysiert und mit den in *A. thaliana* und *C. rubella* untersuchten Segmenten verglichen werden. Die homeologen *B. oleracea*-Regionen unterscheiden sich in Bezug auf ihr Genrepertoire, ein Hinweis auf eine Gendeletion wurde gefunden. Ein Vergleich der *Arabidopsis*- und *Capsella*-Segmente zeigte eine rezente Genduplikation eines Cytochrom P450-ähnlichen Gens. In den homeologen *B. oleracea*-Regionen wurde ein solches Gen nicht gefunden. Trotz der beobachteten Unterschiede im Genrepertoire der orthologen Regionen, war die Anordnung und die relative Orientierung der Gene zueinander erhalten. Außerdem waren die Exons der Gene in Länge und Sequenz konserviert, mit der Ausnahme eines mutmaßlichen *B. oleracea*-Pseudogens. Kodierende Sequenzen aus *Arabidopsis* und *Capsella* sind auf der Nukleinsäureebene im Schnitt 90% identisch. Im Gegensatz dazu wies die Charakterisierung einer Retroelement-ähnlichen Familie aus *A. thaliana* und *C. rubella* die höhere Divergenz dieser Komponenten des Genoms nach. Die Identität homeologer *Brassica*-Gene betrug 85% auf der Nukleinsäureebene. Ähnliche Werte wurden erhalten, wenn *Arabidopsis*- und *Capsella*-Exons mit *Brassica*-Sequenzen

verglichen wurden. Diese Werte spiegeln die phylogenetischen Verwandtschaftsverhältnisse dieser Arten wider.

Alle diese Ergebnisse zusammengenommen, ergibt sich eine auffallende Ähnlichkeit der Organisation der *A. thaliana*-, *C. rubella*- und *B. oleracea*-Genome, wenn man davon absieht, daß duplizierte Segmente Kollinearitätsbeziehungen erschweren können. Folglich kann die Sequenz des *Arabidopsis*-Genoms als effizientes Werkzeug eingesetzt werden, um die Genomorganisation verwandter Cruciferen aufzuschlüsseln.

7 REFERENCES

Abernathy, C.R., Rasmussen, S.A., Stalker, H.J., Zori R., Driscoll, D.J., Williams, C.A., Kousseff, B.G., Wallace, M.R. (1997). NF1 mutation analysis using a combined heteroduplex/SSCP approach. *Hum Mutat.* **9**(6):548-54.

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

Acarkan, A. (2000). Studien zur Genomkolinearität in der Familie der *Brassicaceae*: *Arabidopsis thaliana* and *Capsella rubella*. Doktorarbeit Universität zu Köln.

Alonso-Blanco, C., Peeters, A.J.M., Koornneef, M., Lister, C., Dean, C., van den Bosch, N., Pot, J., Kuiper, M.T. (1998). Development of an AFLP based linkage map of Ler, Col and Cvi *Arabidopsis thaliana* ecotypes and construction of a Ler/Cvi recombinant inbred line population. *Plant J.* **14**:259-271.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman DJ. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**(17):3389-402.

Amstrong, S.J., Fransz, P., Marshall, D.F., Jones, G.J. (1998). Physical mapping of DNA repetitive sequences to mitotic and meiotic chromosomes of *Brassica oleracea* var. *alboglabra* by fluorescence *in situ* hybridisation. *Heredity* **81**:666-673.

Aravind, L., Watanabe, H., Lipman, D.J., Koonin, E.V. (2000). Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl. Acad. Sci.* USA **97**(21):11319-11324.

Arumuganathan, K., Earle, E.D. (1991). Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **9**:208-218.

Avramova, Z., Tikhonov, A., SanMiguel, P., Jin, Y.-K., Liu, C., Woo, S.-S., Wing, R.A. and Bennetzen, J.L. (1996). Gene identification in a complex chromosomal continuum by local genomic cross-referencing. *Plant J.* **10**, 1163-1168.

Axelsson, T., Bowman, C.M., Sharpe, A.G., Lydiate, D.J., Lagercrantz, U., (2000). Amphidiploid *Brassica juncea* contains conserved progenitor genomes. *Genome* **43**:679-688.

Bancroft, I., Love, K., Bent, E., Sherson, S., Lister, C., Cobbett, C., Goodman, H. and Dean, C. (1997). A strategy involving the use of high redundancy YAC subclone libraries facilitates the contiguous representation in cosmid and BAC clones of 1.7 Mb of the genome of the plant *Arabidopsis thaliana*. *Weeds World* **4**:1-9.

Barakat, A., Matassi, G., Bernardi, G. (1998). Distribution of genes in the genome of *Arabidopsis thaliana* and its implications for the genome organization of plants. *Proc. Natl. Acad. Sci.* USA **95**(17):10044-10049.

Bell, C.J., Ecker, J.R. (1994). Assignment of 30 microsatellite loci to the linkage map of *Arabidopsis*. *Genomics* **19**:137-144.

Bennett, M.D., Bhandol, P., Leitch, I.J. (2000). Nuclear DNA amounts in angiosperms and their modern uses : 807 new estimates. *Annals of Botany. Lond. NY: Acad. Press* 86(4):859-909.

Bennett, M.D., Smith, J.B. (1976). Nuclear DNA amounts in angiosperms. *Phil. Trans. R. Soc. Lond.* **274**:227-274.

Bennetzen, J.L. (1996). The contributions of retroelements to plant genome organization, function and evolution. *Trends Microbiol.* **4**(9):347-53.

Bennetzen, J.L., SanMiguel, P., Chen, M., Tikhonov, A., Francki, M. and Avramova, Z. (1998). Grass genomes. *Proc. Natl. Acad. Sci.* USA **95**:1975-1978.

Bennetzen, J.L. (2000a). Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. *Plant Cell* 12:1021-1029.

Bennetzen, J.L. (2000b). Tranposable contribution to plant gene and genome evolution. *Plant Mol. Biol.* **42**:251-269.

Bevan, M., Bancroft, I., Bent, E., Love, K., Goodman, H., Dean, C., Bergkamp, R., Dirkse, W., Van Staveren, M., Stiekema, W., Drost, L., Ridley, P., Hudson, S.-A., Patel, K., Murphy, G., Piffanelli, P., Wedler, H., Wedler, E., Wambutt, R., Weitzenegger, T., Pohl, T.M., Terryn, N., Gielen, J., Villarroel, R., De Clerck, R., Van Montagu, M., Lecharny, A., Auborg, S., Gy, I., Kreis, M., Lao, N., Kavanagh, T., Hempel, S., Kotter, P., Entian, K.-D., Rieger, M., Schaeffer, M., Funk, B., Mueller-Auer, S., Silvey, M., James, R., Montford, A., Pons, A., Puigdomenech, P., Douka, A., Voukelatou, E., Milioni, D., Hatzopolous, P., Piravandi, E., Obermaier, B., Hilbert, H., Düsterhöft, A., Moores, T., Jones, J.D.G., Eneva, T., Palme, K., Benes, V., Rechman, S., Ansorge, W., Cooke, R., Berger, C., Delseny, M., Voet, M., Volckaert, G., Mewes, H.-W., Klosterman, S., Schueller, C., and Chalwatzis, N. (1998). Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*. *Nature* **391**:485-488.

Blanc, G., Barakat, A., Guyot, R., Cooke, R., Delseny, M. (2000). Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell* **12**:1093-1102.

Bohuon, E.J.R., Keith, D.J., Parkin, I.A.P., Sharpe, A.G., Lydiate, D.J. (1996). Alignment of the conserved C genome of *Brassica oleracea* and *Brassica napus*. *Theor. Appl. Genet.* **93**:833-839.

Bowen, N.J., McDonald, J.F. (1999). Genomic analysis of *Caenorhabditis elegans* reveals ancient families of retroviral-like elements. *Genome Res.* **9**(10):924-35.

Britten, R.J. (1997). Mobile elements inserted in the distant past have taken on important functions. *Gene* **205**:177-182.

Camilleri, C., Lafleuriel, J., Macadre, C., Varoquaux, F., Parmentier, Y., Picard, G., Caboche, M., Bouchez, D. (1998). A YAC contig map of *Arabidopsis thaliana* chromosome 3. *Plant J.* **14**:633-642.

Campell, B.R., Song, Y., Posch, T.E., Cullis, C.A., Town, C.D. (1992). Sequence and organization of 5S ribosomal RNA-encoding genes of *Arabidopsis thaliana. Gene* **112**:225-28.

Casacuberta, E., Casacuberta, J. M., Puigdomenech, P., Monfort, A. (1998). Presence of miniature inverted-repeat transposable elements (MITEs) in the genome of *Arabidopsis thaliana*: characterisation of the Emigrant family of elements. *Plant J* **16**(1):79-85.

Cavell, A.C., Lydiate, D.J., Parkin, I.A.P., Dean, C., Trick, M. (1998). Collinearity between a 30-centimorgan segment of *Arabidopsis thaliana* chromosome 4 and duplicated regions within the *Brassica napus* genome. *Genome* **41**:62-69.

Chang, C., Bowman, J.L., DeJohn, A.W., Lander, E.S. and Meyerowitz, E.M. (1988). Restriction fragment length polymorphism linkage map for *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci.* USA **85**:6856-6860.

Chen, M., SanMiguel, P., Bennetzen, J.L. (1998). Sequence organization and conservation in *sh2/a1*-homologous regions of sorghum and rice. *Genetics* **148**: 435-443.

Clarenz, O. (2000). Studien zur Genomkolinearität in der Familie der Brassicaceae. Diplomarbeit Universität zu Köln.

Conner, J.A., Conner, P., Nasrallah M.E., Nasrallah, J.B. (1998). Comparative mapping of the *Brassica* S locus region and its homeolog in *Arabidopsis*. Implications for the evolution of mating systems in the *Brassicaceae*. *Plant Cell* **10**:801-812.

Copenhaver, G. P., Nickel, K., Kuromori, T., Benito, M. I., Kaul, S., Lin, X., Bevan, M., Murphy, G., Harris, B., Parnell, L. D., McCombie, W. R., Martienssen, R. A., Marra, M., Preuss, D (1999). Genetic definition and sequence analysis of *Arabidopsis* centromeres. *Science* **286**(5449):2468-74.

Copenhaver, G.P., Pikaard, C.S. (1996). RFLP and physical mapping with an rDNA-specific endonuclease reveals that the nucleolus organizer regions of *Arabidopsis thaliana* adjoin the telomeres on chromosomes 2 and 4. *Plant J.* **9**:259-272.

Dellaporta, S.L., Wood, J. and Hicks, J.B. (1983). A plant DNA minipreparation: version II. *Plant Mol. Biol. Rep.* **1**:19-21.

Devos, K.M., Gale, M.D. (1997). Comparative genetics in the grasses. *Plant Mol. Biol.* **35** :3-15.

Fabri, C., Schäffner, A. (1994). An *Arabidopsis thaliana* RFLP mapping set to localise mutations to chromosomal regions. *Plant J.* **5**:149-156.

Feuillet, C., Keller, B. (1999). High gene density is conserved at syntenic loci of small and large grass genomes. *Proc. Natl. Acad. Sci.* USA **96** :8265-8270.

Fransz, P., Amstrong, S., Alonso-Blanco, C., Fisher, T.C., Torres-Ruiz, R.A., Jones, G. (1998). Cytogenetics for the model system *Arabidopsis thaliana*. Plant J. 13(6):867-876.

Gale, M. D., Devos, K. M. (1998a). Plant comparative genetics after 10 years. *Science* **282**(5389):656-659.

Gale, M. D., Devos, K. M. (1998b). Comparative genetics in the grasses. *Proc. Natl. Acad. Sci.* USA **95**:1971-1974.

Goodman, H.M., Ecker, J.R., Dean, C. (1995). The genome of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci.* USA **92**:10831-10835.

Grandbastien, M.A., Spielmann, A., Caboche, M. (1989). Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. *Nature* **337**(6205):376-380.

Grandbastien, M.A. (1992). Retroelements in higher plants. *Trends Genet.* **8**(3):103-108.

Grant, M.R., McDowell, J.M., Sharpe, A.G., de Torres Zabala, M., Lydiate, D.J., Dangl, J.L. (1998). Independent deletions of a pathogen-resistance gene in *Brassica* and *Arabidopsis*. *Proc. Natl. Acad. Sci.* USA **95**:15843-15848.

Hamilton, C.M. (1997). A binary-BAC system for plant transformation with high-molecular-weight DNA. *Gene* **200**(1/2):107-116.

Hanahan, D. (1983). Studies on transformation of *Escherichia coli* with plasmids. *J. Mol. Biol.* **166**:557-580.

Hauge, B.M., Hanley, S.M., Cartinhour, S., Cherry, J.M., Goodman, H.M., Koornneef, M., Stam, P., Chang, C., Kempin, S., Medrano, L., Meyerowitz, M. (1993). An integrated genetic/RFLP map of the *Arabidopsis thaliana* genome. *Plant J.* **3**:745-754.

Heslop-Harrison, J.S., Maluszynska, J. (1994). Molecular cytogenetics of *Arabidopsis*. *Arabidopsis*, Cold Spring Harbor Laboratory Press, New York.(eds.) 63-87.

Hirochika, H., Hirochika, R. (1993). Ty1-copia group retrotransposons as ubiquitous components of plant genomes. *Japanese Journal of Genetics* **68**(1):35-46

Höfte, H., Desprez, T., Amselem, J., Chiapello, H., Rouzé, P., Caboche, M., Moison, A., Jourjon, M.-F., Charpenteau, J.-L., Berthomieu, P., Guerrier, D., Giraudat, J., Quigley, F., Thomas, F., Yu, D.-Y.; Mache, R., Raynal, M., Cooke, R., Grellet, F., Delseny, M., Parmentier, Y., de Marcillac, G., Gigot, C., Fleck, J., Philipps, G., Axelos, M., Bardet, C., Tremousaygue, D., Lescure, B. (1993). An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from *Arabidopsis thaliana*. *Plant J.* **4**:1051-1061.

Hoisington, D. (1992). *Laboratory protocols.* CIMMYT Applied Molecular Genetics Laboratory. Mexico, D.F. CIMMYT.

Jackson, S.A., Cheng, Z.K., Wang, M.L., Goodman, H.M., Jiang, J.M. (2000). Comparative fluorescence *in situ* hybridization mapping of a 431-kb *Arabidopsis thaliana* bacterial artificial chromosome contig reveals the role of chromosomal duplications in the expansion of the *Brassica rapa* genome. *Genetics* **156**:833-838.

Jarvis, P., Lister, C., Szabo, V., Dean, C. (1994). Integration of CAPS markers into the RFLP map generated using recombinant inbred lines of *Arabidopsis thaliana*. *Plant Mol. Biol.* **24**(4):685-687.

Jiang, J., Nasuda, S., Dong, F., Scherrer, C.W., Woo, S.S., Wing, R.A., Gill, B.S., Ward, D.C. (1996). A conserved repetitive DNA element located in the centromeres of cereal chromosomes. *Proc. Natl. Acad. Sci.* USA **93**(24):14210-14213.

Jordan, I.K., McDonald, J.F. (1999). Tempo and mode of Ty element evolution in *Saccharomyces cerevisiae*. *Genetics* **151**(4):1341-1351.

Kapitonov, V.V., Jurka, J. (1999). Molecular paleontology of transposable elements from *Arabidopsis thaliana*. *Genetica* **107**(1-3): 27-37.

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

Konieczny, A., Voytas, D.F., Cummings, M.P., Ausubel, F.M. (1991). A superfamily of *Arabidopsis thaliana* retrotransposons. *Genetics* **127**:801-809.

Konieczny, A., Ausubel, F. (1993). A procedure for quick mapping of *Arabidopsis* mutants using ecotype specific markers. *Plant J.* **4**:403-410.

Koornneef, M., van der Veen, J.H. (1983a). Trisomics in *Arabidopsis thaliana* and the location of linkage groups. *Genetica*, **61**:41-46.

Koornneef, M., van Eden, J., Hanhart, C.J., Stam, P., Braaksma, F.J., Feenstra, F.J. (1983b). Linkage map of *Arabidopsis thaliana*. *J. Hered.* **74**:265-272.

Kotani, H., Sato, S., Fukami, M., Hosouchi, T., Nakazaki, N., Okumura, S., Wada, T., Liu, Y.G., Shibata, D., Tabata, S. (1997). A fine physical map of *Arabidopsis thaliana* chromosome 5: construction of a sequence-ready contig map. *DNA Res.* **4**:371-378.

Kowalski, S.P., Lan, T.-H., Feldmann, K.A., Paterson, A.H. (1994). Comparative mapping of *Arabidopsis thaliana* and *Brassica oleracea* chromosomes reveals islands of conserved organization. *Genetics* **138**:499-510.

Ku, HM. Vision, T. Liu, JP. Tanksley, SD. (2000). Comparing sequenced segments of the tomato and *Arabidopsis* genomes: Large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl. Acad. Sci.* USA **97**(16):9121-9126.

Kumar, A., Bennetzen, J.L. (1999). Plant retrotransposons. *Annu. Rev. Genet.* **33**(12):479-532.

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

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

Lagercrantz, U., Putterill, J., Coupland, G., Lydiate, D. (1996). Comparative mapping in *Arabidopsis* and *Brassica*, fine scale genome collinearity and congruence of genes controlling flowering time. *Plant J.* **9** :13-20.

Lander, E.S., Green, P., Abrahamson, J., Barlow, A., Daly, M.J., Lincoln, S.E. and Newburg, L. (1987). MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1**:174-181.

Langdon, T., Seago, C., Mende, M., Leggett, M., Thomas, H., Forster, J. W., Jones, R., N., Jenkins, G. (2000). Retrotransposon evolution in diverse plant genomes. *Genetics* **156**(1):313-25.

Lenoir, A., Cournoyer, B., Warwick, S., Picard, G., Deragon, J. M. (1997). Evolution of SINE S1 retroposons in Cruciferae plant species. *Mol. Biol. Evol.* **14**(9):934-41.

Lin, X., Kaul, S., Rounsley, S., Shea, T.P., Benito, M.-I., Town, C.D., Fujii, C.Y., Mason, Y., Bowman, C.L., Barnstead, M., Feldblyum, T.V., Buell, C.R., Ketchum, K.A., Lee, J., Ronning, C.M., Koo, H.L., Moffat, K.S., Cronin, L.A., Shen, M., Pai, G., Van Aken, S., Umayam, L., Tallon, L.J., Gill, J.E., Adams, M.D., Carrera, A.J., Creasy, T.H., Goodman, H.M., Somerville, C.R., Copenhaver, G.P., Preuss, D., Nierman, W.C., White, O., Eisen, J.A., Salzberg, S.L., Fraser, C.M. and Venter, C. (1999). Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402**:761-768.

Lister, C., Dean, C. (1993). Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. *Plant J.* **4**:745-750.

Liu, Y.-G., Mitsukawa, N., Lister, C., Dean, C. and Whittier, R.F. (1996). Isolation and mapping of a new set of 129 RFLP markers in *Arabidopsis thaliana* recombinant inbred lines. *Plant J.* **10**:733-736.

Lydiate, D., Sharpe, A., Lagercrantz, U., Parkin, I. (1993). Mapping the *Brassica* genome. *Outlook Agric.* **22**:85-92.

Marra, M., Kucaba, T., Sekhon, M., Hillier, L., Martienssen, R., Chinwalla, A., Crockett, J., Fedele, J., Grover, H., Gund, C., McCombie, W.R., McDonald, K., McPherson, J., Mudd, N., Parnell, L., Schein, J., Seim, R., Shelby, P., Waterston, R. and Wilson, R. (1999). A map for sequence analysis of the *Arabidopsis thaliana* genome. *Nat. Genet.* **22**:265-270.

Mayer, K., Schüller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Düsterhöft, A., Stiekema, W., Entian, K.-D., Terryn, N., Harris, B., Ansorge, W., Brandt, P., Grivell, L., Rieger, M., Weichselgartner, M., de Simone, V., Obermaier, B., Mache, R., Müller, M., Kreis, M., Delseny, M., Puigdomenech, P., Watson, M., Schmidtheini, T., Reichert, B., Portatelle, D., Perez-Alonso, M., Boutry, M., Bancroft, I., Vos, P., Hoheisel, J., Zimmermann, W., Wedler, H., Ridley, P., Langham, S.-A., McCullagh, B., Bilham, L., Robben, J., Van der Schueren, J., Grymonprez, B., Chuang, Y.-J., Vandenbussche, F., Braeken, M., Weltjens, I., Voet, M., Bastiaens, I., Aert, R., Defoor, E., Weitzenegger, T., Bothe, G., Ramsperger, U., Hilbert, H., Braun, M., Holzer, E., Brandt, A., Peters, S., van Staveren, M., Dirkse, W., Mooijman, P., Klein Lankhorst, R., Rose, M., Hauf, J., Kötter, P., Berneiser, S., Hempel, S., Feldpausch, M., Lamberth, S., Van den Daele, H., De Keyser, A., Buysshaert, C., Gielen, J., Villarroel, R., De Clercq, R., Van Montagu, M., Rogers, J., Cronin, A., Quail, M., Bray-Allen, S., Clark, L., Doggett, J., Hall, S., Kay, M., Lennard, N., McLay, K., Mayes, R., Pettett, A., Rajandream, M.-A., Lyne, M., Benes, V., Rechmann, S., Borkova, D., Blöcker, H., Scharfe, M., Grimm, M., Löhnert, T.-H., Dose, S., de Haan, M., Maarse, A., Schäfer, M., Müller-Auer, S., Gabel, C., Fuchs, M., Fartmann, B., Granderath, K., Dauner, D., Herzl, A., Neumann, S., Argiriou, A., Vitale, D., Liguori, R., Piravandi, E., Massenet, O., Quigley, F., Clabauld, G., Mündlein, A., Felber, R., Schnabl, S., Hiller, R., Schmidt, W., Lecharny, A., Aubourg, S., Chefdor, F., Cooke, R., Berger, C., Montfort, A., Casacuberta, E., Gibbons, T., Weber, N., Vandenbol, M., Bargues, M., Terol, J., Torres, A., Perez-Perez, A., Purnelle, B., Bent, E., Johnson, S., Tacon, D., Jesse, T., Heijnen, L., Schwarz, S., Scholler, P., Heber, S., Francs, P., Bielke, C., Frishman, D., Haase, D., Lemcke, K., Mewes, H.W., Stocker, S., Zaccaria, P., Bevan, M., Wilson, R.K., de la Bastide, M., Habermann, K., Parnell, L., Dedhia, N., Gnoj, L., Schutz, K., Huang, E., Spiegel, L., Sehkon, M., Murray, J., Sheet, P., Cordes, M., Abu-Threideh, J., Stoneking, T., Kalicki, J., Graves, T., Harmon, G., Edwards, J., Latreille, P., Courtney, L., Cloud, J., Abbott, A., Scott, K., Johnson, D., Minx, P., Bentley, D., Fulton, B., Miller, N., Greco, T., Kemp, K., Kramer, J., Fulton, L., Mardis, E., Dante, M., Pepin, K., Hillier, L., Nelson, J., Spieth, J., Ryan, E., Andrews, S., Geisel C., Layman, D., Du, H., Ali, J., Berghoff, A., Jones, K., Drone, K., Cotton, M., Joshu, C., Antonoiu, B., Zidanic, M., Strong, C., Sun, H., Lamar, B., Yordan, C., Ma, P., Zhong, J., Preston, R., Vil, D., Shekher, M., Matero, A., Shah, R., Swaby, I´K., O´Shaughnessy, A., Rodriguez, M., Hoffman, J., Till, S., Granat, S., Shohdy, N., Hasegawa, A., Hameed, A., Lodhi, M., Johnson, A., Chen, E., Marra, M.,

Martienssen, R., McCombie, W.R. (1999). Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* **402**:769-777.

Mbulu R.S. (2000). Comparative genome analysis within the *Brassicaceae* family : *Arabidopsis thaliana* and *Capsella rubella.* Diplomarbeit Technische Universität München.

McGrath, J.M., Quiros, C.F. (1991). Inheritance of isozyme and RFLP markers in *Brassica campestris* and comparison with *B. oleracea*. *Theor. Appl. Genet.* **82**(6):668-673.

Meyerowitz, E.M., Somerville, C.R. (1994). *Arabidopsis.* Cold Spring Harbor Laboratory Press, New York.(eds.).

Miller, J.T., Dong, F., Jackson, S.A, Song, J., Jiang, J. (1998). Retrotransposons-related DNA sequences in the centromeres of grass chromosomes. *Genetics* **150**:1615-1623.

Moore, G., Foote, T., Helentjaris, T., Devos, K., Kurata, N., Gale, M. (1995). Was there a single ancestral cereal chromosome? *Trends Genet.* **11**:81-82.

Mozo, T., Dewar, K., Dunn, P., Ecker, J.R., Fischer, S., Kloska, S., Lehrach, H., Marra, M., Martienssen, R., Meier-Ewert, S., Altmann, T. (1999). A complete BAC-based physical map of the *Arabidopsis thaliana* genome. *Nat. Genet.* **22**:271-275.

Nam, H.-G., Giraudat, J., den Boer, B., Moonan, F., Loos, W.D.B., Hauge, B.M., Goodman, H. (1989). Restriction fragment length polymorphism linkage map of *Arabidopsis thaliana*. *Plant Cell* **1**:699-705.

Newman, T., de Bruijn, F.J., Green, P., Keegstra, K., Kende, H., McIntosh, L., Ohlrogge, J., Raikhel., N., Somerville, S., Thomashow, M., Retzel, E. and Somerville, C. (1994). Genes galore: a summary of methods for accessing results from large-scale partial sequencing of anonymous *Arabidopsis* cDNA clones. *Plant Phys.* **106**:1241-1255.

Noma, K., Ohtsubo, **E. (**1999). Tnat1 and Tnat2 from *Arabidopsis thaliana*: novel transposable elements with tandem repeat sequences. *DNA Res.* **7**(1):1-7.

O'Neill, C.M., Bancroft, I. (2000). Comparative physical mapping of segments of the genome of *Brassica oleracea* var. *alboglabra* that are homoeologous to sequenced regions of chromosomes 4 and 5 of *Arabidopsis thaliana*. *Plant J.* **23**:233-244.

Osborn, T.C., Kole, C., Parkin, I.A.P., Sharpe, A.G., Kuiper, M., Lydiate, D.J., Trick, M. (1997). Comparison of flowering time genes in *Brassica rapa*, *B. napus* and *Arabidopsis thaliana*. *Genetics* **146**:1123-1129.

Parkin, I.A.P., Sharpe A.G., Keith D.J., Lydiate D.J. (1995). Identification of the A and C genomes of amphidiploid *Brassica napus* (oilseed rape). *Genome* **38**:1122-1131.

Paterson, A.H., Bowers, J.E., Burow, M.D., Draye, X., Elsik, C.G., Jiang, C.-X., Katsar, C.S., Lan, T.-H., Lin, Y.-R., Ming, R., Wright, R.J. (2000). Comparative genomics of plant chromosomes. *Plant Cell* **12**:1523-1539.

Pélissier, T., Tutois, S., Deragon, J.M., Tourmente, S., Genestier, S., Picard, G. (1995). *Athila*, a new retroelement from *Arabidopsis thaliana. Plant Mol. Biol.* **29**:441-452.

Pélissier, T., Tutois, S., Tourmente, S., Deragon, J.M., Picard, G. (1996). DNA regions flanking the major *Arabidopsis thaliana* satellite are principally enriched in *Athila* retroelement sequences. *Genetica* **97**:141-151.

Quackenbush, J., Liang, F., Holt, I., Pertea, G., and Upton, J. (2000). The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucl. Acids Res.* **28**:141-145.

Quiros, C.F., Grellet, F., Sadowski, J., Suzuki, T., Li, G., Wroblewski, T. (2001). *Arabidopsis* and *Brassica* comparative genomics: Sequence, structure and gene content in the ABI1-Rps2-Ck1 chromosomal segment and related regions. *Genetics* **157**(3):1321-1330.

Reiter, R.S., Williams, J.G.K., Feldmann, K.A., Rafalski, J.A., Tingey, S.V., Scolnik, P.A. (1992). Global and local genome mapping in *Arabidopsis thaliana* by using recombinant inbred lines and random amplified polymorphic DNAs. *Proc. Natl. Acad. Sci.* USA **89**:1477-1481.

Richards, E.J., Ausubel, F.M. (1988). Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana. Cell* **53**:127-136.

Rossberg, M., Theres, K., Acarkan, A., Herrero, R., Schmitt, T., Schumacher, K., Schmitz, G., Schmidt, R.l. (2001). Comparative Sequence Analysis Reveals Extensive Microcolinearity in the Lateral Suppressor Regions of the Tomato, *Arabidopsis*, and *Capsella* Genomes. *Plant Cell* **13**(4):979-988.

Round, E.K., Flowers, S.K., Richards, E.J. (1997). *Arabidopsis thaliana* centromere regions: genetic map positions and repetitive DNA structure. *Genome Res.* **7**:1045-1053.

Rounsley, S.D., Glodek, A., Sutton, G., Adams, M.D., Somerville, C.R., Venter, J.C., Kerlavage, A.R. (1996). The construction of *Arabidopsis* expressed sequence tag assemblies. A new resource to facilitate gene identification. *Plant Physiol.* **112**:1177-1183.

Sadowski, J., Quiros, C.F. (1998**).** Organization of an *Arabidopsis thaliana* gene cluster on chromosome 4 including the *RPS2* gene in the *Brassica nigra* genome. *Theor. Appl. Genet.* **96**:468-474.

Sadowski, J., Gaubier, P., Delseny, M., Quiros, C.F. (1996). Genetic and physical mapping in *Brassica* diploid species of a gene cluster defined in *Arabidopsis thaliana. Mol. Gen. Genet.* **251**:298-306.

Saghai-Maroof, M.A., Soliman, K.M., Jorgensen, R.A., Allard, R.W. (1984). Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. *Proc. Natl. Acad. Sci.* USA **81**:8014-8018.

Salanoubat, M., Lemcke, K., Rieger, M., Ansorge, W., Unseld, M., Fartmann, B., Valle, G., Blocker, H., Perez-Alonso, M., Obermaier, B., Delseny, M., Boutry, M., Grivell, L.A., Mache, R., Puigdomenech, P., De Simone, V., Choisne, N., Artiguenave, F., Robert, C., Brottier, P., Wincker, P., Cattolico, L., Weissenbach, J., Saurin, W., Quetier, F., Schafer, M., Muller-Auer, S., Gabel, C., Fuchs, M., Benes, V., Wurmbach, E., Drzonek, H., Erfle, H., Jordan, N., Bangert, S., Wiedelmann, R., Kranz, H., Voss, H., Holland, R., Brandt, P., Nyakatura, G., Vezzi, A., D'Angelo, M., Pallavicini, A., Toppo, S., Simionati, B., Conrad, A., Hornischer, K., Kauer, G., Lohnert, T. H., Nordsiek, G., Reichelt, J., Scharfe, M., Schon, O., Bargues, M., Terol, J., Climent, J., Navarro, P., Collado, C., Perez-Perez, A., Ottenwalder, B., Duchemin, D., Cooke, R., Laudie, M., Berger-Llauro, C., Purnelle, B., Masuy, D., de Haan, M., Maarse, A. C., Alcaraz, J. P., Cottet, A., Casacuberta, E., Monfort, A., Argiriou, A., Flores, M., Liguori, R., Vitale, D., Mannhaupt, G., Haase, D., Schoof, H., Rudd, S., Zaccaria, P., Mewes, H. W., Mayer, K. F., Kaul, S., Town, C. D., Koo, H. L., Tallon, L. J., Jenkins, J., Rooney, T., Rizzo, M., Walts, A., Utterback, T., Fujii, C. Y., Shea, T. P., Creasy, T. H., Haas, B., Maiti, R., Wu, D., Peterson, J., Van Aken, S., Pai, G., Militscher, J., Sellers, P., Gill, J. E., Feldblyum, T. V., Preuss, D., Lin, X., Nierman, W. C., Salzberg, S. L., White, O., Venter, J. C., Fraser, C. M., Kaneko, T., Nakamura, Y., Sato, S., Kato, T., Asamizu, E., Sasamoto, S., Kimura, T., Idesawa, K., Kawashima, K., Kishida, Y., Kiyokawa, C., Kohara, M., Matsumoto, M., Matsuno, A., Muraki, A., Nakayama, S., Nakazaki, N., Shinpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., Tabata, S. (2000). Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. *Nature* **408**(6814):820-822.

Sambrook, J., Fritsch, E.F., Maniatis, T. (1989). Molecular cloning: a laboratory manual. Second edition. *Cold Spring Harbor Laboratory Press*, New York (eds).

Sanguinetti, C.J., Neto, E.D., and Simpson, A.J.G. (1994). Rapid silver staining and recovery of PCR products separated on polyacrylamide gels. *BioTechniques* **17**:915-919.

SanMiguel, P., Tikhonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K., Lee, M., Avramova, Z., Bennetzen, J.L. (1996). Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**:765-768.

SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y., Bennetzen, J. L. (1998). The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**(1):43-5.

Schmidt, R., Acarkan, A., Boivin, K. (2001). Comparative structural genomics in the *Brassicaceae* family. *Plant Physiol. Biochem.* **39**:253-262.

Schmidt, R. (2000). Synteny: recent advances and future prospects. *Curr. Opin. Plant Biol.* **3**(2): 97-102**.**

Schmidt, R., Acarkan, A., Koch, M. and Roßberg, M. (1999). A strategy for comparative physical mapping in cruciferous plants. In *Plant evolution in man-made habitats*, Proceedings of the VIIth Symposium of the International Organization of Plant Biosystematics (van Raamsdonk, L.W.D. and den Nijs, J.C.M., eds). Amsterdam: Hugo de Vries Laboratory.

Schmidt, R., Love, K., West, J., Lenehan, Z. and Dean, C. (1997). Description of 31 YAC contigs spanning the majority of *Arabidopsis thaliana* chromosome 5. *Plant J.* **11**:563-572.

Schmidt, R., West, J., Love, K., Lenehan, Z., Lister, C., Thompson, H., Bouchez, D. and Dean, C. (1995). Physical map and organization of *Arabidopsis thaliana* chromosome 4. *Science* **270**:480-483.

Schneider, K., Borchardt, D.C., Schäfer-Pregl, R., Nagl, N., Galss, C., Jeppson, A., Gebhardt, C., Salamini, F. (1999). PCR-based cloning and segregation analysis of functional gene homologues in *Beta vulgaris. Mol. Gen. Genet.* **262**(3):515-524.

Sentry, J. W., Smyth, D.R. (1989). An element with long terminal repeats and its variant arrangements in the genome of *Lilium henryi. Mol. Gen. Genet.* **215**:349-354.

Sharpe, A.G., Parkin, I.A.P., Keith, D.J., Lydiate, D.J. (1995). Frequent nonreciprocal translocations in the amphidiploid genome of oilseed rape (*Brassica napus*). *Genome* **38**:1112-1121.

Slabaugh, M.B., Huestis, G.M., Leonard, J., Holloway, J.L., Rosato, C., Hongtrakul, V., Martini, N., Toepfer, R., Voetz, M., Schell, J., and Knapp, S.J. (1997). Sequence-based genetic markers for genes and gene families: single-strand conformational polymorphisms for the fatty acid synthesis genes of *Cuphea. Theor. Appl. Genet.* **94**:400-408.

Slocum, M.K., Figdore, S.S., Kennard, W.C., Suzuki, J.Y., Osborn, T.C. (1990). Linkage arrangement of restriction fragment length polymorphism loci in *Brassica oleracea*, *Theor. Appl. Genet.* 80:57-64.

Smyth, D.R., Kalitsis, P., Joseph, J.L., Sentry, J.W. (1989). Plant retrotransposon from Lilium henryi is related to Ty3 of yeast and the gypsy group of Drosophila. *Proc. Natl. Acad. Sci.* USA **86**(13):5015-5019.

Song K.M., Suzuki J.Y., Slocum M.K., Williams P.H., Osborn T.C. (1991). A linkage map of *Brassica rapa* (syn. *campestris*) based on restriction fragment length polymorphism loci. *Theor. Appl. Genet.* **82**:296-304.

Southern, E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98**:503-517.

Suoniemi, A., Tanskanen, J., Schulman, A. (1998). Gypsy-like retrotransposons are widespread in the plant kingdom. *Plant J* **13**(5): 699-705.

Tabata, S., Kaneko, T., Nakamura, Y., Kotani, H., Kato, T., Asamizu, E., Miyajima, N., Sasamoto, S., Kimura, T., Hosouchi, T., Kawashima, K., Kohara, M., Matsumoto, M., Matsuno, A., Muraki, A., Nakayama, S., Nakazaki, N., Naruo, K., Okumura, S., Shinpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., Sato, S., de la Bastide, M., Huang, E., Spiegel, L., Gnoj, L., O'Shaughnessy, A., Preston, R., Habermann, K., Murray, J., Johnson, D., Rohlfing, T., Nelson, J., Stoneking, T., Pepin, K., Spieth, J., Sekhon, M., Armstrong, J., Becker, M., Belter, E., Cordum, H., Cordes, M., Courtney, L., Courtney, W., Dante, M., Du, H., Edwards, J., Fryman, J., Haakensen, B., Lamar, E., Latreille, P., Leonard, S., Meyer, R., Mulvaney, E., Ozersky, P., Riley, A., Strowmatt, C., Wagner-McPherson, C., Wollam, A., Yoakum, M., Bell, M., Dedhia, N., Parnell, L., Shah, R., Rodriguez, M., See, L. H., Vil, D., Baker, J., Kirchoff, K., Toth, K., King, L., Bahret, A., Miller, B., Marra, M., Martienssen, R., McCombie, W. R., Wilson, R. K., Murphy, G., Bancroft, I., Volckaert, G., Wambutt, R., Dusterhoft, A., Stiekema, W., Pohl, T., Entian, K. D., Terryn, N., Hartley, N., Bent, E., Johnson, S., Langham, S. A., McCullagh, B., Robben, J., Grymonprez, B., Zimmermann, W., Ramsperger, U., Wedler, H., Balke, K., Wedler, E., Peters, S., van Staveren, M., Dirkse, W., Mooijman, P., Lankhorst, R. K., Weitzenegger, T., Bothe, G., Rose, M., Hauf, J., Berneiser, S., Hempel, S., Feldpausch, M., Lamberth, S., Villarroel, R., Gielen, J., Ardiles, W., Bents, O., Lemcke, K., Kolesov, G., Mayer, K., Rudd, S., Schoof, H., Schueller, C., Zaccaria, P., Mewes, H. W., Bevan, M., Fransz, P. (2000). Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. *Nature*. **408**(6814):823-826.

Tanksley, S.D., Ganal, M.W., Prince, J.P., de Vicente, M.C., Bonierbale, M.W., Broun, P. T., Fulton, M., Giovannoni, J.J., Grandillo, S. Martin, G.B., Messeguer, R., Miller, J.C., Miller, L., Paterson, A.H., Pineda, O., Röder, M.S., Wing, R.A., Wu, W., Young, N.D. (1992). High density molecular linkage maps of the tomato and potato genomes. *Genetics* **132**: 1141-1160.

Tarchini, R., Biddle, P., Wineland, R., Tingey, S., Rafalski, A. (2000). The complete sequence of 340 kb of DNA around the rice Adh1-Adh2 region reveals interrupted colinearity with maize chromosome 4. *Plant Cell* **12**(3):381-391.

Tatusova, T.V., Madden T.L. (1999). A new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett*. **7**(2):247-250.

Terryn, N., Heijnen, L., De Keyser, A., Van Asseldonck, M., De Clercq, R., Verbakel, H., Gielen, J., Zabeau, M., Villarroel, R., Jesse, T., Neyt, P., Hogers, R., Van Den Daele, H., Ardiles, W., Schueller, C., Mayer, K., Dehais, P., Rombauts, S., Van Montagu, M., Rouzé, P., Vos, P. (1999). Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analyzing a 400-kb contig at the *APETALA2* locus on chromosome 4. *FEBS Lett*. **445**:237-45.

Teutonico, R.A., Osborn, T.C. (1994). Mapping of RFLP and qualitative trait loci in *Brassica rapa* and comparison to the linkage maps of *B. napus*, *B. oleracea*, and *Arabidopsis thaliana*. *Theor. Appl. Genet*. **89**:885-894.

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

Theologis, A., Ecker, J.R., Palm, C.J., Federspiel, N.A., Kaul, S., White, O., Alonso, J., Altafi, H., Araujo, R., Bowman, C.L., Brooks, S.Y., Buehler, E., Chan, A., Chao, Q.M., Chen, H.M., Cheuk, R.F., Chin, C.W., Chung, M.K., Conn, L., Conway, A.B., Conway, A.R., Creasy, T.H., Dewar, K., Dunn, P., Etgu, P., Feldblyum, T.V.,, Feng, J., Fong, B., Fujii, C. Y., Gill, J. E., Goldsmith, A. D., Haas, B., Hansen, N. F., Hughes, B., Huizar, L., Hunter, J. L., Jenkins, J., Johnson-Hopson, C., Khan, S., Khaykin, E., Kim, C. J., Koo, H. L., Kremenetskaia, I., Kurtz, D. B., Kwan, A., Lam, B., Langin-Hooper, S., Lee, A., Lee, J. M., Lenz, C. A., Li, J. H., Li, Y., Lin, X., Liu, S. X., Liu, Z. A., Luros, J. S., Maiti, R., Marziali, A., Militscher, J., Miranda, M., Nguyen, M., Nierman, W. C., Osborne, B. I., Pai, G., Peterson, J., Pham, P. K., Rizzo, M., Rooney, T., Rowley, D., Sakano, H., Salzberg, S. L., Schwartz, J. R., Shinn, P., Southwick, A. M., Sun, H., Tallon, L. J., Tambunga, G., Toriumi, M. J., Town, C. D., Utterback, T., Van Aken, S., Vaysberg, M., Vysotskaia, V. S., Walker, M., Wu, D., Yu, G., Fraser, C. M., Venter, J. C., Davis, R. W. (2000). Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*. *Nature* **408**(6814):816-820.

Thompson, H.L., Schmidt, R., Dean, C. (1996a). Identification and distribution of seven classes of middle-repetitive DNA in the *Arabidopsis thaliana* genome. . *Nucl. Acids Res.* **24**(15):3017-3022.

Thompson, H., Schmidt, R., Brandes, A., Heslop-Harrison, J.S., Dean, C. (1996b). A novel repetitive sequence associated with the centromeric regions of *Arabidopsis thaliana* chromosomes. *Mol. Gen. Genet.* 253(1/2):247-252.

Tikhonov, A.P., SanMiguel, P.J., Nakajima, Y., Gorenstein, N.M., Bennetzen, J.L., Avramova, Z. (1999). Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. *Proc. Natl. Acad. Sci.* USA **96**:7409-7414.

Truco, M.J., Hu, J., Sadowski, J., Quiros, C. (1996). Inter- and intra-genomic homology of the *Brassica* genomes: implications for their origin and evolution. *Theor. Appl. Genet.* **93**:1225-1233.

U, N. (1935). Genomic analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilisation. *Japan J. Botany* **7**:389-452.

Vicient, C.M., Kalendar, R., Anamthawat-Jonsson, K., Schulman, A.H. (1999a). Structure, functionality, and evolution of the BARE-1 retrotransposon of barley. *Genetica* **107**(1-3):53-63.

Vicient, C. M., Suoniemi, A., Anamthawat-Jonsson, K., Tanskanen, J., Beharav, A., Nevo, E., Schulman, A. H. (1999b). Retrotransposon BARE-1 and Its Role in Genome Evolution in the Genus Hordeum. *Plant Cell* **11**(9):1769-1784.

Wendel, J.F., Wessler, S.R. (2000). Retrotransposon-mediated genome evolution on a local ecological scale. *Proc. Natl. Acad. Sci.* USA **97**(12):6250-6262.

Woo, S.-S., Jiang, J., Gill, B.S., Paterson, A.H., Wing, R.A. (1994). Construction and characterization of a bacterial artificial chromosome library of *Sorghum bicolor*. *Nucl. Acids Res.* **22**:4922-4931.

Wright, D.A. Voytas, D.F. (1998). Potential retroviruses in plants: Tat1 is related to a group of *Arabidopsis thaliana* Ty3/gypsy retrotransposons that encode envelope-like proteins. *Genetics* **149**(2):703-15.

Xiong, Y., Eickbush, T.H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *Embo J.* **9**(10):3353-62.

Yang, Y.W, Lai, K.N., Tai, P.Y. and Li, W.H. (1999). Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. *J. Mol. Evol.* **48**:597-604.

Zachgo, E.A., Wang, M.L., Dewdney, J., Bouchez, D., Camilleri, C., Belmonte, S., Huang, L., Dolan, M., Goodman, H.M. (1996). A physical map of chromosome 2 of *Arabidopsis thaliana*. *Genome Res.* **6**:19-25.

Appendix

1- Segregation of markers used for the establishment of the *Capsella* genetic map

List of the different 136 markers used for the establishment of the *Capsella* genetic map and their segregation.

Mapping population: fifty F2-plants derived from a cross of *C. grandiflora* with C. *rubella*.

1 = Genotype of the F2 plants, homozygous *C. grandiflora*
2 = Genotype of the F2 plants, heterozygous
3 = Genotype of the F2 plants, homozygous *C. rubella*
- = missing data point

Locus	**F2-progeny plants 1-50**
*IIF4_B	1-2232131131122222-22112233323311221113121-212322-
*mi443	12-232131131-2222222211222322332122111312122122222
*MboL9	1222321311311222222221122-322322123111312132122222
*mi348	12223212213112222232-11222323322223-12321132122222
*mi203	12223212223212222232211222323322223112321132122122
*m235A	-122321222321222223221122-2223222-3112321232132122
*MboL16	11222212222213222232311221222322223212321232123122
*m254A	11222-1222231322223-311-11222322223212321232122121
*mi342_M47	111312112223132222313211122-2212213222231233-22122
*MboCr8	11131211222313222231321112222212213222231233222122
*mi133_M45	-1131-112223132222313-11122-2212213222231232-22122
*mi291_M46	111311112223132222313311112222122132-2231222-22122
*mi208	1113111122221322223133111112221221322-231312222122
*mi19	111311112222132222313311111222212213222231312222122
*IA7_T21478_B	211311112222132232313311111222212213222231312222122
*N96681	21131112222213223332232111112111213222232212222122
*mi303	21131112222213223332232111112111213222232212222122
*mi353	22222123123322222232232222333132212222222222223-2212
*Z35365_M35_B	-2222-2212332222232222322233-1322122-22221222-22212
*IIIC8	2212312212331222322222311123313211233222312-2323212
*mi335_B	3212312212331322322-23111221133112232222312222223311
*T46241_M13	3-123122123313223222223111221133112232222312222223311
*IG3_A	3112312212331321322212111222133112332222312222222311
*IA10	32123-2212331321322-1211122213311233222212222222321
*AA067525_M27_B	321-312212331321322212111122213311223222212222222321
*IA7_T21478_A	3212312212331321322212---22213311-23222212222222321
*m315A_A	32123122123313213222121111222213311223222212222222321
*mi320_B	-222312212331321322212111122221331122322-2122222222321
*IIC12	3-323222123223213222121312221322122313221322322232
*mi425	3231323213323321222211-322222322122313222322232233
*mi157	23323233132233212222211322222322212213122231232233
*mi74_A	31322232323323212312313322322122322322122312112323
*mi199	31322232323323212312313322322122322322122312112323
*mi207	3132223132-3232123-2323332322122322222112312112323

```
*mi339           3132222132332-2133123233323221223-222-212312112322
*mi142           3232212132332222331232-3323221223322212123221l2322
*mi358_A         32322121323322223312323332322122332221212323112322
*FKBP61I3T       -23221213233212232123233223--31233-331212333-22321
*FKBP15-1I       22322121323321223212323322322312332331212333-22321
*T20G20          22322121213221233213323322322312332331212233322321
*Z34612_M5       22-22121213221233213323322231-31233223221213332232l
*mi398           22322122213221233212323322312312332232212133322321
*mi390           223121232122-1233212323322312312322222212133322321
*mi116           22312123212221233212323322312312322222212133322321
*IIH8_Z34614_B   223121232222-1233212323322312312322222212133322321
*mi139           2231111322223123221231333231221232222221213332223-1
*IC7             13321321111232221111222321221221321322223 2-1221322
*MboH16          23321321111232221111222321221221223322223111221322
*mi320_A         -33213211112322211122-23212212212223222223121--1332
*AC002391        22321221111232221112222321221221222312223121-21232
*mi238           22321221111232221112222321221221222313223121221232
*IIG8_Z29768_B   213-11321112--2211122231212222212323132222222322232
*mi54            213211322111-2211112233122122222232322222232322232
*mi277           21222232212222312111233123113222232322232232322233
*IID4_A          21222332312222312211233123113332233323132332322223
*IF5_A           21222332312222312311233123113332233323132332322223
*IIIC7           2122-3323122223123l-1331231133322333231323-2322223
*CT9-7           21222332312212312311133123113222232323-3--323---23
*IIIA10          31332222312332111122221323222222321232331323132231
*mi287           31332221322222111111221323322222322231331323132231
*T41531_M24      31332221322222121111231323233222322231331323232231
*H76592_M25      31332221322222121111231323233222322231331323-32231
*ve021_M7        -132222133221212111113132 32-3222322-31321323-32231
*Cos57           31322121222212121221121323232222222231321322232231
*m249A           -1322-2113121-121221121323232222222231321322232231
*IID9_C          313221211312121212211213232322221 22-31321322232231
*MboL19          31322121131112121221121323232222122231321322232231
*m457A           31222111131112121221121323222222122231222322232121
*mi456           31222112131112121221121323222322122231222 3--232221
*T21989_M26      -1222-1213111-1212211213232-2322122-31222322-32221
*MboE6           21222112131112121221121323222332132231222322232221
*IID4_B          21221112121112131221121323222332132231222322232221
*CF7_1           2212221212-221322113221312223 31-12223222232222222 2
*mi121           -112--1212122132211321131222331212-2222223222222 23
*mi97            21122212111221322113211312222212122222222322222223
*mi174           21132211111221222113111212332112121222222312222123
*mi74_B          2113231111122122211311 1-12332112122221222212222123
*mi438           21132311111221222113111212332112112221222212222123
*mi138           -113--11111221222113111212332111112221222213-22123
*mi433           21133211111221222113111212332111122221222213221123
*mi90            21133211111221222113111212332111122221222213221123
*mi219           21222211111221222112112222332111122221222213221123
*mi125           21322222--122-122112-121223321212222212211133 21223
*N97271_M31      21322222211221122112112121322122222221221113311223
*m518A           213221222122-1112112212 1----22322322212121132223 33
*Z35365_M35_A    -1322-2221222221223-2222221-2232232-21212123-22333
*H36452_M17      213222222122-2212232222221222322322212121232223 33
*mi306           21322222212222212232222222122232232221212123222333
*m448A           2132222221222221223223 22221-2332222221112123222332
*m315A_B         2132222221222221223323 2222-22332222221112123223332
*mi122           21322223222222212233232222122332222221112123223322
*mi51            213222232222222122332322221223332222211121232 2--22
*IG3_B           --3-----------3--333-33--------33----------33-33--
*MboL5           213222222221222223332332223221233222212 12-23323322
*mi323           21322222222122222322233223312222323221212223323322
*MboI18          21322222222122222322233223312222322221212223323222
*mi137           -1322-2222212222322233223 3-2222322-21212223-23222
```

```
*AA728584_M32    21-22222222122222322233223312222322221212223323222
*AA067525_M27_A  21322222222122222322233223312222322221212223323222
*MboM7           21322222222122222322233223312222322221212223323222
*mi30            21-222222221222223222332-3312222322221212223323222
*IG2_Z18140      21322222222122222322233223312222322221212223323222
*m326A           213222222221--2223222322-3312222322222212223323222
*mi358_B         213222222221-222231223---3312222121222212223323222
*IIG9_Z29799     21322222222122122312232223312231212222212223323221
*EST2            2-3-222222-1-21223-22322233122312122222122223 23221
*C54X6           2132222222212212231-232223312231212222212222323221
*cDNAJ           21322222222122122312232223312231212222212222323221
*mi330_A         21322222222122122312232223313231212222211222323221
*m557A           2232--222221-2122312332223323231212222111222323221
*IID9_B          --32222222212212131-332223323231213222111222323221
*IIB4_T41886     2---22122221-212131223-2-2323321222323121232233211
*mi123           32222211222213222131223222223321122323221232-32211
*mi232           32222211222213222131223222223321122323221232-32111
*AAT             32222112222132232312231232223321122323221322232111
*mi431           32222112222132232212231232223321122323221322232111
*IIG8_Z29768_A   32212112222132222212231232223321122323221322232111
*IF5_B           3--12112-22132222-12--1232223321122323221322232111
*IIH8_Z34614_A   32212112222132222212231232223221122323221322232111
*MboO11          32212112222132222212231232223221122323222322232111
*Cos36RB         322121122221322222122312322-3221122323222322232111
*mi369           32212112222132222212231232223221122323222322232111
*mi335_A         3133111122322331222-111333213312222321223211322212
*MboA6           31331111223223312222111333213312222321223211322212
*CSau20F         312311112232-331222-11133321331222-32122321132221-
*CosT9           31231111223223312222111333213312222321223211322212
*m211A           3123121122322331222211123-223212223321223121222212
*mi69            31232211223122212222311233223221313221223122322212
*AA067573_M36    31233-1122312221222-3122332-2221313-21323122-22222
*MboN18          11233211222132223322322233121221313121323122212211
*FKBP15-2I       11233-112221322233223222333-1221313-21323122-12221
*mi194           11233311122132223322322233321321213121323122212221
*IIF4_A          122333111221322232-23222333-13212131113231-2212311
*IID9_A          12233312122132223222322233231321213111323122212311
*mi61            1-23331222213222322232223323132121-112322112212311
*mi330_B         12233312222132223222322233-3132221111231211221 2211
```

2- Values of the χ^2 test for each of the co-dominant markers used to establish the *Capsella* genetic linkage map

Presented in the table are the χ^2-Test data for each marker mapped on the *Capsella* genetic map. In the first column is indicated the amount of plants scored for a marker, then is listed the number of plants having the *C. grandiflora* or *C. rubella* genotype for each marker, as well as the number of heterozygous. It has been tested whether the values were significantly different from the expected ratio 1:2:1 segregation. Number of alleles for each genotype has been also calculated and compared to the expected ratio of 1:1 segregation. In both cases, distorted segregations have been fixed to 0,05.

Markers	plants analysed for each marker	*Capsella grandiflora homozygote individuals*	*Capsella heterozygote individuals*	*Capsella rubella homozygote individuals*	χ^2-Test	*Capsella grandiflora alleles*	*Capsella rubella alleles*	χ^2-Test
IIF4_B	46	16	20	10	0.21	52	40	0.21
mi443	48	14	27	7	0.15	55	41	0.15
MboL9	49	15	26	8	0.16	56	42	0.16
mi348	48	11	28	9	0.68	50	46	0.68
mi203	50	11	30	9	0.69	52	48	0.69
m235A	47	10	29	8	0.68	49	45	0.68
MboL16	50	11	31	8	0.55	53	47	0.55
m254A	47	13	26	8	0.30	52	42	0.30
mi342_M47	48	15	24	9	0.22	54	42	0.22
MboCr8	50	15	26	9	0.23	56	44	0.23
mi133_M45	45	14	23	8	0.21	51	39	0.21
mi291_M46	48	17	23	8	0.07	57	39	0.07
mi208	49	19	22	8	0.03	60	38	0.03
mi19	50	19	23	8	0.03	61	39	0.03
IA7/T21478_B	50	18	23	9	0.07	59	41	0.07
N96681	50	17	25	8	0.07	59	41	0.07
mi303	50	17	25	8	0.07	59	41	0.07
mi353	49	5	34	10	0.31	44	54	0.31
Z35365_M35B	45	5	33	7	0.67	43	47	0.67
IIIC8	49	12	24	13	0.84	48	50	0.84
mi335_B	49	14	22	13	0.84	50	48	0.84
T46241_M13	49	14	22	13	0.84	50	48	0.84
IG3_A	50	16	22	12	0.42	54	46	0.42
IA10	48	13	24	11	0.68	50	46	0.68
AA067525/M27	49	14	25	10	0.42	53	45	0.42
IA7/T21478_A	46	11	25	10	0.83	47	45	0.83
m315A_A	50	14	26	10	0.42	54	46	0.42
mi320_B	48	13	26	9	0.41	52	44	0.41
IIC12	49	9	27	13	0.42	45	53	0.42
mi425	49	7	26	16	0.07	40	58	0.07
mi157	50	8	27	15	0.16	43	57	0.16
mi74_A	50	9	24	17	0.11	42	58	0.11
mi199	50	9	24	17	0.11	42	58	0.11
mi207	48	9	23	16	0.15	41	55	0.15
mi339	47	9	23	15	0.22	41	53	0.22
mi142	49	8	26	15	0.16	42	56	0.16
mi358_A	50	8	25	17	0.07	41	59	0.07
FKBP61I3T	45	8	19	18	0.04	35	55	0.04
FKBP15-1I	49	8	23	18	0.04	39	59	0.04
T20G20	50	9	23	18	0.07	41	59	0.07
Z34612_M5	48	10	22	16	0.22	42	54	0.22
mi398	50	9	25	16	0.16	43	57	0.16
mi390	49	10	25	14	0.42	45	53	0.42
mi116	50	10	26	14	0.42	46	54	0.42
IIH8/Z34614_B	49	9	26	14	0.31	44	54	0.31
mi139	49	12	23	14	0.69	47	51	0.69
IC7	49	16	24	9	0.16	56	42	0.16
MboH16	50	16	25	9	0.16	57	43	0.16
mi320_A	46	14	23	9	0.30	51	41	0.30
T2-cosmid	49	15	28	6	0.07	58	40	0.07

mi238	50	15	28	7	0.11	58	42	0.11
IIG8/Z29768_B	47	13	26	8	0.30	52	42	0.30
mi54	49	12	28	9	0.54	52	46	0.54
mi277	50	9	28	13	0.42	46	54	0.42
IID4_A	50	9	22	19	0.05	40	60	0.05
IF5_A	50	9	21	20	0.03	39	61	0.03
IIIC7	47	9	19	19	0.04	37	57	0.04
CT9-7	44	10	18	16	0.20	38	50	0.20
IIIA10	50	11	23	16	0.32	45	55	0.32
mi287	50	13	22	15	0.69	48	52	0.69
T41531_M24	50	11	22	17	0.23	44	56	0.23
H76592_M25	49	11	21	17	0.23	43	55	0.23
ve021_M7	46	13	19	14	0.83	45	47	0.83
Cos57	50	12	28	10	0.69	52	48	0.69
m249A	47	13	24	10	0.54	50	44	0.54
IID9_C	49	15	23	11	0.42	53	45	0.42
MboL19	50	16	23	11	0.32	55	45	0.32
m457A	50	17	26	7	0.05	60	40	0.05
mi456	48	15	25	8	0.15	55	41	0.15
T21989_M26	44	14	23	7	0.14	51	37	0.14
MboE6	50	15	26	9	0.23	56	44	0.23
IID4_B	50	16	25	9	0.16	57	43	0.16
CF7-1	48	10	31	7	0.54	51	45	0.54
mi121	46	13	26	7	0.21	52	40	0.21
mi97	50	14	31	5	0.07	59	41	0.07
mi174	50	20	24	6	0.01	64	36	0.01
mi74_B	49	20	23	6	0.00	63	35	0.00
mi438	50	21	23	6	0.00	65	35	0.00
mi138	46	22	18	6	0.00	62	30	0.00
mi433	50	22	21	7	0.00	65	35	0.00
mi90	50	22	21	7	0.00	65	35	0.00
mi219	50	19	27	4	0.00	65	35	0.00
mi125	46	14	26	6	0.10	54	38	0.10
N97271_M31	50	18	27	5	0.01	63	37	0.01
m518A	45	14	24	7	0.14	52	38	0.14
Z35365_M35_A	44	7	29	8	0.83	43	45	0.83
H36452_M17	49	7	34	8	0.84	48	50	0.84
mi306	50	7	35	8	0.84	49	51	0.84
m448A	49	8	33	8	1.00	49	49	1.00
m315A_B	49	7	32	10	0.54	46	52	0.54
mi122	50	7	33	10	0.55	47	53	0.55
mi51	48	7	32	9	0.68	46	50	0.68
MboL5	49	5	31	13	0.09	41	57	0.09
mi323	50	5	33	12	0.16	43	57	0.16
MboI18	50	5	35	10	0.32	45	55	0.32
mi137	45	4	32	9	0.29	40	50	0.29
AA728584_M32	49	5	35	9	0.42	45	53	0.42
AA067525/M7a	50	5	35	10	0.32	45	55	0.32
MboM7	50	5	35	10	0.32	45	55	0.32
mi30	48	5	34	9	0.41	44	52	0.41
IG2_Z18140	50	5	35	10	0.32	45	55	0.32
m326A	47	4	34	9	0.30	42	52	0.30
mi358_B	46	7	31	8	0.83	45	47	0.83
IIG9_Z18140	50	9	32	9	1.00	50	50	1.00
EST113	45	7	30	8	0.83	44	46	0.83
C54X6	49	9	32	8	0.89	50	48	0.89
c13.049	50	9	33	8	0.84	51	49	0.84
mi330_A	50	10	31	9	0.84	51	49	0.84
m557A	47	9	28	10	0.84	46	48	0.84
IID9_B	47	10	26	11	0.84	46	48	0.84

IIB4_T41886	44	10	24	10	1.00	44	44	1.00
mi123	49	10	29	10	1.00	49	49	1.00
mi232	49	11	28	10	0.84	50	48	0.84
AAT	50	11	27	12	0.84	49	51	0.84
mi431	50	11	28	11	1.00	50	50	1.00
IIG8_Z29768_A	50	12	28	10	0.69	52	48	0.69
IF5_B	44	12	23	9	0.52	47	41	0.52
IIH8_Z34614_A	50	12	29	9	0.55	53	47	0.55
MboO11	50	11	30	9	0.69	52	48	0.69
Cos36B	49	11	29	9	0.75	51	47	0.75
mi369	50	11	30	9	0.69	52	48	0.69
CSau20F	46	15	18	13	0.68	48	44	0.68
CosT9	50	15	22	13	0.69	52	48	0.69
mi335_A	49	15	20	14	0.84	50	48	0.84
MboA6	50	15	21	14	0.84	51	49	0.84
m211A	49	13	26	10	0.54	52	46	0.54
mi69	50	12	27	11	0.84	51	49	0.84
AA067573_M36	45	10	24	11	0.83	44	46	0.83
MboN18	50	15	23	12	0.55	53	47	0.55
FKBP15-2I	46	12	21	13	0.83	45	47	0.83
mi194	50	14	22	14	1.00	50	50	1.00
IIF4_A	47	15	18	14	0.84	48	46	0.84
IID9_A	50	14	22	14	1.00	50	50	1.00
mi61	48	13	23	12	0.84	49	47	0.84
mi330_B	49	14	24	11	0.54	52	46	0.54

3- Correspondence of the EST/TC accession numbers and the *Arabidopsis thaliana* annotations of the genes along the complete sequence

During this study, genes and predicted genes have been called by their „tentative consensus“ names or EST accession numbers. In this table, is given the correspondence between the EST and TC names compared to the recent annotations given for the complete *Arabidopsis thaliana* genome.

ESTs	*TC name*	*At4g name*
EST0/Knat5gene	71614	At4g32040
	100015	At4g32030
EST1/2		At4g32020
EST3/AV528310	86829/93505	At4g32010
	77106	At4g32000
AAT/EST4	71853	At4g31990
EST5	98955	At4g31980
AV54275/553989	94311	
P450-like	84838	At4g31970
	103345	At4g31960
P450-like	80777	At4g31950
P450-like	73277	At4g31940
	82122	At4g31930
	82121	At4g31920

4- ORF sequences for the genes and predicted genes on *Capsella rubella* and *Brassica oleracea*

During this study, *Capsella rubella* and *Brassica oleracea* genomic sequences have been used to determine putative ORFs corresponding to the predictions of *Arabidopsis thaliana* sequences. *C. rubella* and *B. oleracea* ORF sequences are listed below from the mtarting methionine (ATG) to the stop codon. Sequences of A. thaliana for which only predictions were available are listed as well.

Sequence of the *C. rubella* ORF Cr-71614

```
   1  ATGTCGTTTA ACAGCTCTCA TCTTCTTCCT CCACAAGAAG AAGACCTTCC
  51  TCTCCGACAC TTCTCCGATC AACCTCCTCC CCCACAGCGT CACTTCCCTG
 101  AAACGCCTTC CCTTGTCACC ACCAGTTTCC TCAACCTCCC TTCCACCCTT
 151  GCCACGGCGG ATTCCGATCT CGCTCCTCCG CCCCGCAACG GAGACAATTC
 201  CGCTCCTGAT GCTAACCCAC GGTGGCTCTC TTTCCACACG GAGATCCAAA
 251  ACACCGGAGA AGTCCGTTCT GAAGTTATCG ACGGAGTCAA CGCCGATGGT
 301  GAAACTGTAC TTGGCGTTGT TGGAGGTGAA GATTGGCGGA GCGCTAGCTA
 351  TAAGGCCGCG ATTTTGAGAC ATCCGATGTA CGAACAGCTT CTTGCGGCTC
 401  ATGTGGCTTG CCTTAGGGTT GCGACTCCCG TTGACCAGAT TCCGAGGATC
 451  GATGCTCAGC TCAGTCAGTT CCACACCGTC GCCGAGAAAT ACTCCACTCT
 501  TGGTGTCGTT GTGGACAACA AGGAACTTGA TCATTTCATG TCACATTATG
 551  TTGTGTTGTT ATGTTCATTC AAAGAACAAC TCCAACACCA CGTTTGTGTC
 601  CATGCAATGG AAGCCATTAC GGCTTGTTGG GAGATCGAAC AATCATTGCA
 651  ATCCCTAACT GGAGTTTCTC CAAGTGAAAG TAATGGTAAG ACAATGTCGG
 701  ATGATGAAGA TGATAATCAA GTAGACAGCG AGGTCAACAT GTTTGATGGG
 751  AGTTTGGACG GCTCAGATTG CTTGATGGGG TTTGGTCCTC TTGTTCCAAC
 801  CGAACGAGAG AGATCCTTGA TGGAACGTGT GAAGAAAGAA CTGAAGCATG
 851  AGCTTAAACA GGGTTTCAAA GAGAAGATTG TGGACATAAG AGAAGAGATA
 901  ATGAGGAAGA GAAGAGCGGG GAAGCTCCCG GGGGATACAA CTTCTGTACT
 951  GAAAGAGTGG TGGCGTACTC ACTCGAAATG GCCATACCCA ACTGAGGAAG
1001  ACAAGGCAAA ACTGGTTCAA GAAACCGGTT TGCAGTTGAA ACAAATCAAC
1051  AATTGGTTCA TCAACCAGAG GAAGAGAAAC TGGAACAGCA ATACTTCCAC
1101  ATCATCTACT CTCTCCAAGA ACAAACGTAA
```

Sequence of the *C. rubella* ORF Cr-100015

```
   1  ATGAAGAGGA TTCCGTCGAC ATCAGCTAAG ATTCTGGTCA AGGATGACTG
  51  GGTGGTGACG GCTATGACTG ACGACGAGAT GGTTGTTGAG CTTCTTTTAC
 101  GGCTCAAGCA TGCTGGTACT GCAGTGGCGG ATAATCCGGC CGCCACGAAT
 151  CTTCCTCCGT TACGATGGGG AATCCGTCAG CGACGTTCTC GGTCCTCAAG
 201  ATTCGCTGGC GGCGGCGTCG GCGTTATCGT TTCAATGAAG AAGGATGTCG
 251  ATTCCGTTAG AGCTAGTCCG AAGACTCCTC TCTCCTGGAG CGGCGGATCT
 301  GGAAACCGTA GCGGATCTGG TAGCCGTGGC GGATCTGGAA GCCGTGGCGG
 351  CTCTGCATCT CCTTCAGCTG ATGGGTTCGA GGATACTAGT CGTCAAGCTA
 401  GCTGCTCTAC GTCTACAGGA TCTGGATCTA AGGTCTTTCC CACTAACGAA
 451  ATCACTAGTT CCTTCTCTAA GAGATTGAGG AAAAAGAAGT CATCTTCTGA
 501  GCTTAAAAAC GAAGAGAACT TGAAGCTGAA AGAAAGACTA GACCTTGAAA
 551  AGGAGATTGC AAGTCTCCGA GCAACGTTTG ACGAACAAAA CGTCAGGAAT
 601  CAGAGATTGA AGAGAATTAA GCTTGACTTG AACTCAGGCC GTGTAAAGAA
 651  GGAGACACGG GTTGATCTCA GCCATAAACA ACAAGCGGTA TCAAAATCGT
 701  GCAGAGTAGA TGGAAGAAAT GGTGAATCAG AAAACAAAGG GAGTGTGTTC
 751  TTCAGCTTTG ATCTCAACAT GGTACCATCA GAGGAGGAGA TGATATTGTA
 801  A
```

Sequence of the *C. rubella* ORF Cr-83424

```
  1  ATGGGCGTCG CCGTTCTAAA TCCCCAAGAC TGTTTGAGAG ATCCCTTCTC
 51  CCACATGAGA CATCATCCTC GTAACCCTAG CGCATGTCCC AACAGGCAGA
101  AAAAGCCGGT TTCCAACAAC CGTACGCGCC GGAGCCCTCC ACGTAATCAA
151  TCCACCAGAT CTCCTTCTCC TCCTATCGCG CCGCCTCTTC CTCCTCCTCG
201  TGCGGCTGTC TCTGCTTTTG TTCCCAAGGG AACGGTTAAG AAGAGTCCTA
251  AAAACACCGT CGCCGTTGGT CAGGTTAGAA TCCTCAAGCG CGGTGAAGAA
301  ATTCCTAAGA AGACTTCGGA TCTCGTTGTT GCAAAGTCAG ATCTCGTTGT
351  TGTGAAGCCA GATCTGGTTG TTGAGAAGTC AGATCTGGGT TCTACTCGTC
401  GTATTGGACC AGATCCCGGT TTGATTCCGA GTCAAATCCG TTTGTCTGGC
451  CGCAAATCAA AATCAGCACC GTTTTACGCC GGTCCGGTGA CCATGACCTC
501  GCCGCCTCCG AGCGATGTAC CGCTTCCAGC TTTTTTCACG AAGAAGAGCG
551  TCTCTTTGTT CCAAGCCGCC GATGCAACCA ATGATCTGAT CAGGATGCTT
601  CGCTTAGACA TCGCCTGA
```

Sequence of the *B. oleracea* ORF Bo-83424, locus chromosome 1

```
  1  ATGGGCGTAG CTGTTCTAAA TCCACAAGAC TATCTCAAAC AACCTTTCTC
 51  CCACATGAAG TATCCTCGTA ACCACACCGC ATGCCCCAAC AGGCATCAGA
101  AGAAGCCGGT TCCAAACCGC ACGCGCCGGA GTCCTCCGCG CAATCAGACA
151  ACCAGATCTC CTCCTAAAGC GCCGCCTCCT CCTCCTCAAC GCGCCGCCGT
201  CTCTTCTTAC GTTCCGAAGG GAACGGTTGA GAAGAGTCCC ACCAAAAACG
251  TCGTCGTTGG TCAGGTTAGG ATCCTGAAGC GAGGCGAGGA GATCCCTAAG
301  AAGACATTAG AATTGGTCGT GGAAAAGACA GATCTGGTTG TCGAAAAGAC
351  AGATCTGGTT GTCGAAAAGC CAGATCTGGT TTCCACTCAA CGGATCGGAC
401  CAGATCCGTG TCTGATTCCG AGCCAGATCC GTCTCCCCGA CCGCAAATCG
451  AATAAGACCG TAGTCCCGTT TTACGCCGGT CCTGTGACCA TGACCTCGCC
501  GCCTCCGAGC GACGTCCCTC TCCCAGCCTT CTTCACCACG AAGAAGGACG
551  CTACAAACCA TATCATCAAG CTGCTTCGCC TAGACGTCGC ATGTATGTCT
601  CTCCAATGA
```

Sequence of the *B. oleracea* ORF Bo-83424, locus chromosome 7

```
  1  ATGGGCGTCG CTGTTCTAAA TCCCCAGGAC TGCTTGAAGC ATCCTTTATC
 51  TCACATGAAA CATCCACGTA ACCCCAGCGC GTGCCCCAAC AGGCAGAAAA
101  AACCGGTTTC GAACCGCACG CGCCGCAGCC CGCCGCGAAA ACAAACCTCC
151  CCATCTCCCC CTGTAGCCCC GCCGCTTCCA AAGGGAACGG TGACGACGCG
201  CCCCAACAAC AACAACAACA ACAGCGTCGT CGCTGGTCAG GTTAGAATCC
251  TGAAGCGCGG CGAGGAGATC CCTAAGAAGA CAGCAGATCT GGTCGTAGAA
301  AAGACAGATC TTGTGTCTAC TCGTAGGATC GGACCGGATC CAGGATTGAT
351  TCCGAGTCAG ATCCGTTTAT CCGTCCGCAA AGCAAAGACC GTCCCGTTTT
401  ACGCCGGTCC CGTGACCATG ACGTCTCCTC CTCCAAGCGA CGTCCCTCTT
451  CCAGCCTTTT TCGCCGCGAA GAAGAGCGTC TCTTTGTTCC AAGCCGCCGA
501  CGCTACCAAC GAAATCATCA GGATGCTCCG CCTAAACATC GCGTGA
```

Sequence of the *A. thaliana* ORF At-86829

```
  1  ATGGAAGTGA CTCGTGGTTT CTCTTTTTTG GACAAGTTTC TTCGAAAGCG
 51  TCTTCACTGT GGATGCATTG CTTCTAGATT TATGATGGAG CTTCTAGAGA
101  ATGGTGGTGT TACCTGTATA AGTTGCGCCA AGAAATCCGG ACTAATTTCT
151  ATGAATGTGA GCCATGAATC TAACGGTAAG GACTTCCCCT CATTTGCTTC
201  AGCAGAGCAT GTAGGCAGTG TTCTTGAGAG GACAAATCTC AAGCACTTGC
251  TTCACTTTCA AAGAATCGAC CCCACTCATT CTTCTCTTCA AATGAAACAA
301  GAAGAATCGC TGCTTCCTTC CAGCCTAGAT GCTCTTAGAC ACAAAACTGA
351  AAGGAAAGAA TTGTCTGCAC AGCCAAACTT GAGCATTTCA CTTGGACCTA
401  CGCTTATGAC AAGCCCATTT CATGATGCTG CTGTTGATGA CAGAAGTAAG
```

```
 451  ACTAATTCGA TTTTCCAACT GGCCCCTCGG TCCAGGCAGC TGCTTCCAAA
 501  ACCTGCAAAT TCAGCTCCCA TTGCTGCTGG CATGGAGCCT AGTGGGAGCC
 551  TGGTGTCACA GATTCATGTC GCTCGGCCTC CTCCAGAAGG TCGCGGGAAG
 601  ACCCAATTGC TTCCCCGTTA CTGGCCTAGG ATTACTGACC AAGAGCTGCT
 651  GCAATTATCT GGACAGTATC CTCATCTGTA TGAGTCCTTG ACTGTTTATT
 701  TTCCTAGCTC AAATTCCAAA ATTATACCAC TCTTTGAAAA AGTTCTGAGT
 751  GCGAGCGATG CGGGTCGTAT TGGTCGACTG GTTCTTCCGA AAGCATGTGC
 801  AGAGGCATAT TTCCCCCCTA TATCTCTACC CGAGGGTCTC CCGTTAAAGA
 851  TACAAGACAT AAAAGGGAAA GAATGGGTGT TCCAGTTCAG GTTTTGGCCT
 901  AATAATAACA GCAGGATGTA CGTTTTGGAG GGTGTGACTC CTTGCATACA
 951  GTCCATGCAG TTGCAAGCTG GTGACACTGT AACATTCAGC CGTACAGAAC
1001  CTGAAGGAAA ACTCGTAATG GGATACCGTA AAGCGACGAA CTCTACAGCG
1051  ACACAGATGT TCAAGGGAAG CAGTGAACCC AATCTGAACA TGTTTTCCAA
1101  CAGCTTGAAT CCGGGATGTG GTGACATCAA TTGGTCTAAA CTAGAGAAGT
1151  CTGAGGACAT GGCAAAGGAT AACTTATTTC TTCAGTCGTC CTTAACTTCT
1201  GCTAGGAAAC GGGTTCGGAA CATTGGGACT AAGAGCAAGC GTCTGCTCAT
1251  TGATAGCGTA GATGTTCTGG AACTGAAAAT AACTTGGGAG GAGGCACAGG
1301  AGCTGTTGCG GCCTCCCCAA TCCACCAAAC CCAGCATCTT TACGCTGGAA
1351  AATCAAGATT TTGAAGAATA TGACGAACCA CCAGTTTTCG GGAAGAGGAC
1401  CCTTTTTGTC TCACGTCAAA CAGGGGAACA AGAGCAATGG GTGCAGTGTG
1451  ATGCTTGTGG GAAATGGCGA CAGCTGCCGG TGGATATTCT TCTTCCACCA
1501  AAGTGGTCGT GCTCTGATAA TCTCTTGGAT CCTGGCAGGT CTTCATGTTC
1551  CGCACCTGAT GAACTCTCTC CAAGAGAACA GGATACACTT GTCCGGCAGA
1601  GCAAAGAGTT CAAAAGGAGG AGACTGGCAT CATCAAACGA AAAGCTAAAC
1651  CAGTCGCAGG ATGCATCTGC TCTGAATAGT TTAGGAAATG CAGGCATCAC
1701  CACAACCGGT GAACAGGGGG AAATCACGGT TGCAGCCACG ACCAAGCATC
1751  CAAGACACCG GGCAGGGTGT TCGTGCATCG TCTGCAGCCA ACCACCGAGC
1801  GGAAAAGGCA AACACAAGCC GTCATGCACT TGCACTGTGT GCGAGGCAGT
1851  GAAGAGACGA TTCAGGACGC TCATGCTGCG GAAGCGGAAC AAAGGAGAGG
1901  CAGGACAGGC AAGCCAGCAG GCGCAGTCAC AGTCAGAGTG CAGGGACGAG
1951  ACAGAAGTGG AGAGCATTCC AGCGGTTGAA CTAGCCGCAG GGGAAAACAT
2001  CGACTTGAAC TCAGACCCGG GGGCTTCCCG AGTAAGCATG ATGAGGCTTC
2051  TCCAAGCTGC AGCGTTTCCT CTGGAAGCAT ATCTGAAACA AAAGGCTATT
2101  TCCAATACAG CAGGAGAACA GCAAAGCAGT GATATGGTCA GCACAGAACA
2151  CGGTTCGTCC TCAGCCGCAC AAGAAACTGA GAAAGACACA ACAAATGGAG
2201  CTCATGATCC TGTGAACTAA
```

Sequence of the *C. rubella* ORF Cr-86829

```
   1  ATGGAAGTGA GTCGTTGTTT CTCTTTTTCG GACAAGTTTC TTCGAAAGCG
  51  CCTTCACTGT GGATGCATTG CTTCCAGATT TATGATGGAG CTTCTAGATA
 101  ATGGCAGCGT TACCTGTATA AGTTGCGCCA AGAAATCCGC ACTATTTTCT
 151  ATGAATGTCA GTCAAGAATC CAATGGTAGG GACTCCTCAT TTGCTTCAGC
 201  AGAGCATGTA GGCAGTGTTC TTGAGAGGAC AAATCTTAAG CACTTGCTCG
 251  ACTTTCAAAG GATCGGCCCC ACTCAATCTT CTATTCAAAT GAAACAAGAA
 301  GAATCGCTGC TTCCTTCCAG ACTAGATGCT CTTAGACACA AAACTGAAAG
 351  GAAAGAATTG CAGGAATTAT CTGCACAGCC AAACTTGAGC ATTTCACTTG
 401  GACCTACGCT TATGACAAGT CCATTTCATG ATGCTGTTAT TGATGACAGA
 451  AGTAAGACTA CGTCAATTTT CCAACTAGCC CCTCGGTCCA GGCAACTGCT
 501  TCCAAAACCT GCAAATTCAG CTCCCACTGC TGCTGGCATG GAGCCTAATG
 551  GGAGCCTGGT GTCACAGATT CATGTCGCTC GGCCTCCTCC AGAAGGTCGC
 601  GGGAAGACCC AATTGCTTCC TCGTTATTGG CCTAGGATTA CTGACCAAGA
 651  GCTGCAGATA TTATCTGGAC AGTATCCTCA TCTGTATGAG CCCTTAACTG
 701  TTTATTTTCC AAGCTCAAAT TCCAAAATTA TTCCACTCTT TGAAAAAGTT
 751  CTGAGTGCTA GCGATGCGGG TCGTATTGGT CGACTGGTTC TTCCGAAAGC
 801  ATGTGCAGAG GCATATTTCC CCCCGATTTC TCTACCCGAG GGTCTCCCGT
 851  TAAAAATACA AGACATAAAA GGGAAAGAGT GGGTGTTCCA GTTCAGATTT
 901  TGGCCTAATA ATAACAGCAG GATGTACGTT TTGGAGGGTG TGACTCCTTG
 951  CATACAGTCC ATGCAGTTGC AAGCTGGTGA CACTGTAACG TTCAGCCGTA
1001  CAGAACCTGA AGGAAAACTT GTAATGGGAT ACCGTAAAGC GACAAACTCT
1051  ACAGCAACAC AGATGTTCAA GGGAAGCAGT GAACCCAATC TAAACATTTT
1101  TTCCAACAAC TTGAATCCGG GATGTGGTGA CATCAGCTGG TCTAAACTAG
1151  AGAAGGCTGA GGACATGGGA AAGGATAATT TATTTCTTCA GTCGTCACTA
1201  ACTTCGTCTA GGAAACGGGT TCGGAACATT GGGAGTAAGA GCAAGCGTCT
1251  GCTCATTGAT AGCGTTGATG TTCTGGAATT GAAAGTAACT TGGGATGAGG
1301  CACAGGAACT GATGCGGCCG CCCCAATCCG CCAAACCCAG CATTATTACG
1351  CTGGAAAATC AAGATTTTGA AGAATATGAC GAACCACCGG TTTTCGGGAA
```

```
1401  GAAAACTGTT TTTGTGGCAC GTCAAACAGG GGAACAAGAG CAATGGGTGC
1451  AGTGTGATGC TTGTGGGAAA TGGCGACGGC TGCCTGTGGA TACTCTTCTT
1501  CCACCAAAAT GGTTGTGCTC CGATAATCAC TTGGATCCTG CCAGGTCTTC
1551  ATGTTCTGCA CCTGATGATC TCTCTCCAAG AGAACAGGAT ACACTAGTCC
1601  GGCAAAGCAA AGAGTTCAAA AGGAGGAGAC TGGCAGCATC AAACGAAAAG
1651  CTAAACCAGT CGCAGGAGGC ATCTGCTGTG GAAACTTTAG CAAATGCAGG
1701  TATCACCACG ACTGGTGAAC AAGGGGAAAT CGCAGTTGCA GCGACGACCA
1751  AGCACCCAAG ACACCGGGCA GGGTGTTCGT GCATCGTCTG TAGCCAACCG
1801  CCGAGCGGAA AAGGCAAACA CAAGCCGACA TGCACTTGCA CAGTATGCGA
1851  GGCAGTGAAG AGGCGGTTTA GGACGCTCAT GATGCGGAAG AAAAACAGGG
1901  GAGAGGCAGG ACAGGCAAGC CAGCAGGCAC AGTCGGAGAG CAGAGACGAG
1951  ACAGAAGTGG AGAGTATTCC AGCGGCTGAG GCAGCGTCAG GGGATAATAT
2001  TGACTTAAAC TCAGACCCGG GGGGTTCCCC GAGTAAGCAT GATGGGCTTT
2051  CTTCAAAGCT GCAGCTTTTC CTCTAG
```

Sequence of the *B. oleracea* ORF Bo-86829 locus chromosome 7

```
   1  CGTCTTCACT GTGGATGCAT TGCTTCTAGA TTCATGATGG AGCTTGTGGA
  51  TAATGGTGGT GTTACATGTA TAACCTGCGC CAAAAAATCC GGACTATTCT
 101  CATGAATGTC GAATCCAACG GTAGGGAGTT CCCTACATTT GCTTCAGCAG
 151  AGAATGTAAG CAGCGTTCTC GAGAGGACAA ATCTCAAACA CTTGCTTCAT
 201  TTCCAAAGAA TCTCCCCCAC ACAACCTTTC CTTCAGATGA AACAAGAGGA
 251  ATCTCTCCTT CCCGCAAGAC TAGAATCTCT CAGACACAAC ACTGAGAAGA
 301  AAGAATCTGC ACAGCCAAAC TTGAGCATTT CACTTGGACC TACTCTTATG
 351  ACAAGTCAAT TTCATGATGT GGACGACAGA AGCAAGACTA CTCCTATTTT
 401  CCAACTCGCC TCTCGGTCTA GACAACTCCT TCCAAAACCT GCGAACTCAG
 451  CTCCCACTAC TGCTCCTCCC ATGGAGCCTA ACGGGAGCCT CGTGTCGCAG
 501  ATTCACGTCG CTAGGCCTCC TCCAGAAGGT CGAGGCAAGA CTCAGTTGCT
 551  TCCGCGTTAT TGGCCTAGGA TCACTGATCA GGAGCTGCAG CAATTATCTG
 601  GACAGTATCC TCATCTCTCA AACTCCAAAA TTATACCACT GTTTGAAAAA
 651  GTTCTGAGTG CAAGCGATGC TGGTCGTATT GGTCGACTGG TTCTTCCTAA
 701  AGCATGTGCA GAGGCGTATT TTCCACCGAT CTCTCAGCCT GAGGGCCTCC
 751  CGTTAAAGAT ACAAGACATA AAGGGGAAAG AATGGGTGTT CCAGTTTAGG
 801  TTTTGGCCTA ATAATAACAG CAGGATGTAC GTTCTGGAGG GTGTGACTCC
 851  TTGCATACAG TCAATGCAGT TGCAAGCTGG TGATACTGTA ACGTTCAGCC
 901  GCACAGAACC TGAAGGAAAA CTCGTAATGG GATACCGTAA AGCGACAAAC
 951  TCTACAGCAG CACAGATGTT CAAGGGAAGC AGTGAACCCA ATCTGAACAT
1001  GTTTTCCAAC AACTTGAGTT CGGGATGTGG TGATATCAAC TGGTCTAAAC
1051  TTGACAAGTC TGACGACATG TCAAAGGACG GCTTAATGCT TCAGCCGTCG
1101  CTAATCTCTG CTAGGAAACG TGTTCGAAAC ATTGGCACTA AGAGCAAGCG
1151  ACTGCTCATT GATAGCGTAG ATGTTCTGGA ACTGAGATTA ACCTGGGAGG
1201  AGGCGCAGGA ACTGCTGCGG CCTCCTCAGT CTGCTAAACC GAGCATATGT
1251  ACAGTTGAAG ATCACGATTT TGAAGAATAC GATGAACCAC CAGTTTTTGG
1301  GAAGAGGACA GTGTTTGTGT CACGTCAAAC AGGGGAACAA GAGCAATGGG
1351  TTCAGTGTGA TGCTTGTGCT AAATGGCGAC GGCTTCCTGT GGATACTCTT
1401  CTTCCACCAA AGTGGTTGTG CTCTGATAAT GTCTTGGATC CTGGCAGGTC
1451  TTCATGTTCT GCACCTGATG AACTCACCCC AAGAGAACAG GATACACTTC
1501  TCCGGCTAAG CAAAGAGTTC AAAAGGAGGA GACTGGCATC ATCAAACCAG
1551  GAGGAGGCCT CTGCTCTAGA CACTTTAGCA AATGCAGCTA TCACCACGAC
1601  AGGTGAACAA GGAGAAACCG AGGTTGCAGC CACGACCAAG CACCCGAGAC
1651  ACCGAGCTGG CTGTTCGTGC ATTGTCTGCA GCCAGCCGCC GAGTGGGAAA
1701  GGCAAACACA AGCCGTCATG CACTTGCACT GTGTGTGAGG CGGTGAAGAG
1751  GCGGTTCAAG ACGCTCATGA TGCGGAAGCG AAACAGAGGA GAAGCAGGGC
1801  AGGCAAGCCA GCAGGCGCAG TCAGATCAGT GCAGAGAGGA GACAGAAGCT
1851  GAGAGCATTC CAGCGGTTGA ACTGCCAGCG GCAGGGGGGA ACATTGACTT
1901  AAACTCAGAC CCAGCTTCTA GAGTGAGCAT GATGAGTCTT CTGCAAGCTG
1951  CAACGTTTCC TCTGGAGGTG TATCTGAAAC AGAAAGGTGT TCCAAATACA
2001  GCGGGAGAAC AGCAAAGCAG TGATATAGTA AGCACAGAGA ACGGTTCGTC
2051  CTCAGCCGCA CAAGAACATG ACAGAGACAC AAGCGGAGCT CCTGAGGCAT
2101  TGAACTAA
```

Sequence of the *A. thaliana* ORF At-77106

```
   1  ATGGGAAAGA TTCTTCATCT TCTTCTTCTT CTTCTTAAGG TCTCTGTTCT
```

```
  51  TGAATTCATC ATTAGTGTTT CTGCTTTTAC TTCACCTGCT TCACAGCCTT
 101  CTCTTTCTCC TGTTTACACT TCCATGGCTT CCTTTTCTCC AGGGATCCAC
 151  ATGGGCAAAG GCCAAGAACA CAAGTTAGAT GCACACAAGA AACTTCTAAT
 201  CGCTCTCATA ATCACCTCAT CTTCTCTAGG ACTAATACTT GTATCTTGTT
 251  TATGCTTTTG GGTTTATTGG TCTAAGAAAT CTCCCAAAAA CACCAAGAAC
 301  TCAGGTGAGA GTAGGATTTC ATTATCCAAG AAGGGCTTTG TGCAGTCCTT
 351  CGATTACAAG ACACTAGAGA AAGCAACAGG CGGTTTCAAA GACGGTAATC
 401  TTATAGGACG AGGCGGGTTC GGAGATGTTT ACAAGGCCTG TTTAGGCAAC
 451  AACACTCTAG CAGCAGTCAA AAAGATCGAA AACGTTAGTC AAGAAGCAAA
 501  ACGAGAATTT CAGAATGAAG TTGATTTGTT GAGCAAGATT CACCACCCGA
 551  ACATCATCTC ATTGTTTGGA TATGGAAATG AACTCAGTTC GAGTTTTATC
 601  GTCTACGAGC TGATGGAAAG CGGATCATTG GATACACAGT TACACGGACC
 651  TTCTCGGGGA TCGGCTTTAA CATGGCACAT GCGGATGAAG ATTGCTCTTG
 701  ATACAGCAAG AGCTGTTGAG TATCTCCACG AGCGTTGTCG TCCTCCGGTT
 751  ATCCACAGAG ATCTTAAATC GTCAAATATT CTCCTTGATT CTTCCTTCAA
 801  CGCCAAGATC AACATTCAGA TTTCGGATTT TGGTCTTGCG GTAATGGTGG
 851  GGGCTCACGG CAAAAACAAC ATTAAACTAT CAGGAACACT TGGTTATGTT
 901  GCTCCAGAAT ATCTCCTAGA TGGAAAATTG ACGGATAAGA GTGATGTTTA
 951  TGCGTTTGGT GTGGTTTTAC TTGAACTCTT GTTAGGAAGA CGGCCGGTTG
1001  AGAAATTGAG TTCGGTTCAG TGTCAATCTC TTGTCACTTG GGCAATGCCC
1051  CAACTTACGG ATAGATCAAA GCTTCCGAAA ATCGTGGATC CGGTTATCAA
1101  AGATACAATG GATCATAAGC ACTTATACCA GGTGGCAGCC GTGGCAGTGC
1151  TTTGTGTACA ACCAGAACCG AGTTATCGAC CGTTGATAAC CGATGTTCTT
1201  CACTCACTAG TTCCATTGGT TCCGGTAGAG CTAGGAGGGA CTCTCCGGTT
1251  AATACCATCA TCGTCTTGA
```

Sequence of the *C. rubella* ORF Cr-77106

```
   1  ATGAGAAAGA CACTTCATCT TCATCTTCTT AAGATCTCTG TTCTTGAGTT
  51  CCTCATTAGT GTTTCTGCTT CTTCTACTAT ACCTAATTAT TCACAGCCTT
 101  CTCCTTCTCC ACTTTACACT TCCATGGCTT CCTTCTCTCC AGGGATCCAA
 151  ATGGGCAGAG GCCAAGAACA CAAACTAGAT GCTCACAAGA AACTTTTTAT
 201  CTCTCTCATA ATCACCTCAT CTTCTCTAGG GTTCATACTC CTATGTTGTT
 251  TATGCTTCTG GATTTATCGG TCCAAGAAGT CCCTCAAAAC CACCAAGAAC
 301  TCAGGTGAGA GTGGGATTTC ACTAGCCAAG AAGGGTTTTG TGCAGTCCTT
 351  CGATTACAAG ACACTAGAAA AAGCGACAGG CGGTTTCAAA GACGGTAATC
 401  TTGTAGGACG AGGCGGATTT GGATATGTTT ACAAGGCCAG TTTAGGCAAT
 451  AACACTCTAG CAGCAGTCAA AAAGATTGAA AACGTTAGTC AAGAAGCAAA
 501  ACGAGAGTTT CAGAACGAAG TTGATTTGTT GAGCAAGATT CACCACCCGA
 551  ACATCATCTC ATTGTTGGGA TATGGAAGTG AAATCAGCTC GAGTTTCATC
 601  GTCTACGAGC TGATGCAAAA TGGATCCTTG GATGCTCAGT TACACGGACC
 651  TTCTCGGGGA TCGGCTTTAA CATGGCACAT GCGGATGAAG ATTGCTCTTG
 701  ATACAGCCAG AGCTGTTGAG TATCTTCACG AGCGTTGTCG TCCTCCGGTT
 751  ATCCACAGAG ATCTTAAATC TTCAAACATT CTCCTTGATT CTTCCTTCAA
 801  CGCCAAGTTG ATTCAAGAGA ATCTGATACA TTCGTCTTTG CTTCAACAAA
 851  CAATCCAAAA CATGGTTTCG TTTATGTATC TATGTCTAAC GACATACAAC
 901  ACTCAGATTT CGGATTTTGG TCTGGCGGTA ATGGTTGGGG CGCACGGCAA
 951  GAACAACATT AAACTATCAG GGACACTTGG TTATGTTGCT CCAGAATATC
1001  TCCTAGACGG TAAATTAACG GATAAAAGTG ATGTCTACGC GTTTGGGGTG
1051  GTTCTACTTG AACTCTTGCT AGGAAGACGG CCGGTTGAGA AATTGAGTTC
1101  GGTTCAGTGC CAATCTCTAG TCACTTGGGC AATGCCTCAA CTTACGGATA
1151  GATCAAAGCT TCCAAAAATT GTGGATCCGG TTCTCAAAGA CACAATGGAT
1201  CATAAGCACT TATACCAGGT AGCAGCCGTG GCAGTGCTTT GCGTACAGCC
1251  AGAACCGAGT TATCGACCGT TGATAACCGA TGTTCTTCAC TCACTTGTTC
1301  CACTGGTTCC GGTAGAGCTA GGGGGCACTC TCCGGTTAAC ACCATCATCC
1351  TCTTAA
```

Sequence of the *B. oleracea* ORF Bo-77106 locus chromosome 1

```
   1  ATGAGAAAGA TTCTTCCTCT ACTCCTTAAG GTCTTGGTTA TTCAGTTCCT
  51  CTGTAGTGTC TATGCTTGTA CAGCATCCCA TCCACCTGCG TCACAGCCTT
 101  CACTTTCTCC CGTCTACACT TCCATGGCTT CCTTCTCTCC AGGAATCCAA
 151  ATGGGTAGTA GAGGCCAAGA ACACAATAAA CTTTTAATAG CTCTTATAAT
 201  CAGCTCCTCG TCTCTAGGAC TAATAGTTTT TTGTTGTTTA TGCTTTTGGG
 251  CTGTTTATCG GTCTAAGCAA TTTCCCAAAC CGACCAAGAA CTCAGAGAGT
 301  GGGATTTCAT TACCCAAGAA GGGTTTTATG CAGTCCTTCG ATTACAAGAC
 351  ACTAGAGAAA GCAACAGGCG GCTTCAAAGA CAGTAATCTT ATAGGACGAG
 401  GCGGGTTTGG ATTTGTTTAC AAAGCCTGCT TAGACAATCA CACCCTAGCC
 451  GCGGTTAAGA AGATCGAAAA CGTTAGCCAA GAAGCAAAAC GGGAGTTTCA
 501  GAACGAAGTT GATCTATTGA GCAAGATTCA CCATCCCAAC ATCATCTCAC
 551  TTCTGGGACA TACAAGTGAA ATCAGCCCGA GCTTCATCGT TTACGAGCTG
 601  ATGGAAAAGG GATCTTTGGA GGCACAGTTA CACGGACCTT CTCGTGGATC
 651  GGCTTTAACA TGGCACATGC GGATGAAGAT TGCTCTTGAT ACAGCAAGAG
 701  GTGTAGAGTA TCTACATGAG CGTTGTCGCC CTCCGGTTAT CCACAGAGAT
 751  ATAAAATCTT CAAATATTCT CCTTGATTCT TCCTTCAACG CCAAGATATC
 801  TGATTTTGGA CTTGCGGTAA CGAACGGGAT GCACGGCAAG AACAACATTA
 851  AACTATCTGG GACACTTGGT TATGTTGCTC CAGAATATCT CCTAGATGTG
 901  GTTCTACTTG AACTGTTGCT GGGAAGGCGA CCGGTTGAGA AGTTGAGTTC
 951  GGTTCAGTGC CAATCTCTGG TCACTTGGGC AATGCCACAA CTTACGGATA
1001  GATCAAAGCT TCCTAAAATA GTGGATCCAG TTATCAAAGA TACAATGGAT
1051  CATAAGCATC TATACCAGGT AGCAGCCGTG GCAGTGCTTT GTGTACAGCC
1101  AGAACCAAGT TATAGACCGT TGATAACCGA TGTTCTTCAC TCACTTGTTC
1151  CACTGGTTCC GGTAGAGCTA GGAGGAACAC TCCGGTTAAC ATCATCATCG
1201  TCCTAA
```

Sequence of the *C. rubella* gene Cr-AAT

```
   1  ATGGCTTCTT CAATGTTGTC TCTCGGTTCG ACTTCTCTAT TACCGCGCGA
  51  GATTAGCAAG GATAAGCTAA AGCTTGGGAC TTCCGGTTCG AACCCGTTCC
 101  TGAAAGCAAA GTCTTTTAGC AGGGTGACTA TGGCGGTTGC AGTCACGCCT
 151  TCTCGTTTTG AGGGTATAAC TATGGCTCCT CCTGACCCTA TTCTTGGAGT
 201  CAGCGAAGCA TTCAAAGCTG ACACCAACGA GATGAAACTC AATCTTGGTG
 251  TTGGGGCTTA CCGAACTGAG GAACTCCAGC CTTATGTGCT TAATGTTGTT
 301  AAAAAGGCGG AGAATTTGAT GTTGGAGAGA GGAGATAATA AAGAGTATCT
 351  TCCAATTGAG GGGTTGGCAG CATTCAACAA GGCTACTGCT GAGTTGCTGT
 401  TTGGAGCTGG TCATCCTGTT ATTAAGGAAC AAAGAGTGGC AACAATTCAG
 451  GGTCTTTCGG GAACAGGTTC CCTGCGAGTA GCAGCGGCTC TTATAGAGCG
 501  TTATTTTCCT GGAGCAAAAG TTGTGATTTC ATCACCAACC TGGGGTAATC
 551  ACAAGAATAT CTTCAATGAT GCCAAAGTTC CATGGTCTGA ATACCGCTAC
 601  TATGATCCAA AAACAATTGG TTTGGATTTT GAGGGAATGA TAGCAGATAT
 651  TAAGGATGCT CCAGAAGGAT CTTTCATCTT GCTTCATGGA TGTGCTCACA
 701  ACCCAACAGG AATTGACCCA ACCCCAGAAC AGTGGGTGAA AATTGCGGAT
 751  GTCATTCAGG AAAAGAACCA TATCCCATTC TTTGATGTTG CATACCAGGG
 801  CTTTGCTAGT GGAAGCCTTG ATGAAGATGC AGCATCTGTG AGATTATTTG
 851  CTGAACGTGG AATGGAGTTT TTTGTTGCTC AGTCATACAG TAAAAATTTA
 901  GGTCTTTATG CAGAAAGAAT TGGGGCAATC AATGTCGTGT GCTCATCAGC
 951  CGATGCTGCG ACAAGGGTCA AGAGTCAACT AAAAAGGATT GCTCGGCCTA
1001  TGTACTCAAA CCCACCAGTT CATGGCGCGA GAATCGTGGC TAATGTCGTG
1051  GGCGATCCAA CTATGTTCGG TGAATGGAAA GCAGAGATGG AAATGATGGC
1101  GGGAAGAATA AAAACAGTGA GACAAGAGTT GTATGATAGC CTCGTTTCAA
1151  AAGACAAGAG CGGGAAGGAC TGGTCATTCA TTCTGAAGCA AATTGGCATG
1201  TTCTCTTTCA CTGGCTTGAA CAAAGCTCAG AGCGATAACA TGACGAACAA
1251  GTGGCATGTG TACATGACTA AAGACGGAAG AATATCGTTG GCTGGATTGT
1301  CCATGGCGAA ATGCGAGTAC CTTGCTGACG CCATCATTGA CTCCTATCAC
1351  AACGTAAGTT GA
```

Sequence of the *B. oleracea* gene Bo-AAT locus chromosome 1

```
   1  ATGGCTTCAT CAATGCTGTC TCTCGGTTCT ACTTCTCTGC TACCTCGCGA
  51  GATTAACAAG GATAAGCTAA AGCTTGGAAC TTCCGGTTCC AACCCCTTCC
 101  TGAAAGCAAA GTGTTTTAGT CGGGTGACCA TGTCGGTTGC AGTGAAGCCT
 151  TCTCGCTTTG AGGGTATCAC CATGGCTCCA CCAGACCCTA TTCTTGGCGT
 201  CAGCGAAGCT TTCAAAGCTG ACACTAACGA GCTTAAGCTC AATCTCGGCG
 251  TTGGTGCTTA TCGAACTGAA GAACTCCAGC CTTATGTCCT TAATGTTGTT
 301  AAAAAGGCGG AGAACCTGAT GTTGGAGAGA GGAGATAATA AAGAGTATCT
 351  CCCAATAGAA GGGTTGGCTG CATTCAACAA GGCCACTGCT GAGCTGCTGT
 401  TTGGAGCTGG TCATCCTGTT ATTAAAGAAC AAAAAGTGGC AACAATTCAA
 451  GGTCTTTCCG GAACAGGTTC ACTCCGACTA GCAGCGGCTC TTATCGAGCG
 501  TTATTTTCCT GGAGCTAAAG TTCTTATATC TGCACCAACA TGGGGTAACC
 551  ACAAGAACAT CTTCAACGAC GCCAAAGTTC CCTGGTCTGA ATACCGCTAC
 601  TATGACCCCA AAACAATCGG TTTGGATTTT GAAGGGATGA TAGCTGATAT
 651  AAGGGAAGCT CCAGAAGGAT CATTCATACT GCTACACGGC TGCGCTCACA
 701  ACCCGACCGG AATCGACCCA ACGCCAGAGC AGTGGGTGAA AATTGCTGAC
 751  GTCATTCAAG AAAAGAACCA CATCCCATTT TTCGACGTTG CATACCAGGG
 801  CTTTGCTAGC GGAAGCCTTG ATGAAGACGC AGCTTCCGTG AGACTATTTG
 851  CTGAGCGTGG GATGGAGTTT TTCGTCGCTC AGTCGTATAG TAAAAACTTG
 901  GGCCTTTATG CTGAAAGGAT TGGTGCAATC AACGTCGTCT GCTCATCAGC
 951  CGATGCTGCT ACAAGGGTGA AGAGCCAGTT GAAAAGGATA GCTAGGCCTA
1001  TGTACTCGAA CCCACCGGTT CACGGTGCGA GGATCGTGGC TAACGTCGTG
1051  GGAGATGCAG CTATGTTCAA CGAGTGGAAA GCAGAGATGG AAATGATGGC
1101  GGGGAGGATT AAGACGGTGA GACAGCAGCT GTACGACAGC CTCGTTTCGA
1151  AGGATAAGAG CGGTAAGGAC TGGTCGTTTA TTCTGAAGCA GATTGGCATG
1201  TTCTCATTCA CAGGTCTCAA CAAGGCTCAG AGTGATAACA TGACGGACAA
1251  GTGGCATGTG TACATGACTA AAGACGGGAG GATATCGTTG GCTGGATTGT
1301  CTATGGCAAA ATGCGAGTAC CTCGCTGATG CCATCATCGA CTCGCACCAT
1351  AACGTAAGCT GA
```

Sequence of the *B. oleracea* gene Bo-AAT locus chromosome 7

```
   1  ATGGCTTCTT CAATGCTCTC TCTCGGCTCG ACTTCTCTGT TACCGCGCGA
  51  GATTAACAAG GATAAGCTAA AACTTGGACC CTCAGGTTCG AACCCCTTCC
 101  TGAGAACAAA GTCTCTTAGT CGGGTGACCA TGTCGGTTTC AGTGAAACCT
 151  TCTCGTTTCG AGGGTATAAC AATGGCACCA CCAGACCCTA TTCTTGGAGT
 201  CAGCGAAGCA TTCAAAGCTG ACACTAACGA GCTTAAACTC AATCTCGGTG
 251  TTGGCGCTTA TCGAACCGAG GAACTCCAGC CTTATGTGCT TAACGTCGTT
 301  AAAAAGGCCG AGAACCTGAT GTTAGAGAGA GGAGATAATA AAGAGTATCT
 351  ACCAATAGAG GGGTTGGCTG CATTCAACAA GGCCACTGCT GAGCTGCTT
 401  TTGGAGCTGG TCATCCTGTT ATTAAGGAAC AAAAAGTGGC AACCATTCAG
 451  GGTCTTTCCG GAACCGGTTC ACTGAGACTA GCAGCGGCTC TTATTGAGCG
 501  TTACTTCCCT GGAGCTAAAG TTCTGATATC AGCACCAACA TGGGGTAATC
 551  ACAAGAATAT CTTCAATGAT GCCAAAGTTC CATGGNCTGA ATACCGTTAC
 601  TATGACCCAA AAACTATTGG TTTGGACTTT GAGGGGATGA TAGAAGATAT
 651  TAAGGAAGCT CCGGAAGGAT CATTCATTTT GCTTCATGGT TGTGCTCACA
 701  ACCCAACTGG GATTGACCCA ACACCAGAAC AATGGGTGAA AATAGCTGAT
 751  GTCGTTCAGG AGAAGAACCA TATCCCGTTT TTCGATGTTG CATACCAGGG
 801  CTTTGCTAGT GGAAGCCTTG ATGAAGATGC AGCATCTGTG AGATTATTCG
 851  CCGAACGTGG AATGGAGTTT TTTGTTGCTC AGTCGTATAG TAAAAATTTG
 901  GGTCTTTATG CTGAAAGAAT AGGTGCAATC AATGTCGTCT GCTCATCAGC
 951  CGATGCTGCT ACAAGGGTGA AGAGCCAGTT GAAAAGGATT GCTAGGCCTA
1001  TGTACTCGAA CCCACCGGTT CACGGGGCGA GGATCGTGGC TAATGTGTTG
1051  GGGGATGCAA CTATGTTTGG TGAGTGGAAA GCAGAGATGG AAATGATGGC
1101  GGGTAGGATA AAGACTGTGA GACAAAGGTT GTATGACAGT CTTGTTTCAA
1151  AAGACAAGAG TGGCAAGGAC TGGTCCTTTA TTCTGAAGCA AATTGGCATG
1201  TTCTCATTCA CTGGCCTTAA TAAAGCTCAG AGCGATAACA TGACGAACAA
1251  GTGGCATGTG TACATGACTA AAGACGGGAG GATATCGCTG GCTGGATTGT
1301  CTATGGCAAA ATGTGAGTAT CTTGCCGATG CCATCATCGA CTCATGCCAT
1351  AACGTAAGCT GA
```

Sequence of the *A. thaliana* ORF At-84838

```
   1  ATGGATACTT CCCTCTTTTC TTTATTTGTT CCAATCCTTG TTTTCGTTTT
  51  TATTGCTCTT TTTAAGAAAT CAAAGAAACC AAAACATGTA AAAGCTCCTG
 101  CACCAAGTGG TGCGTGGCCC ATCATCGGTC ATCTTCACCT TCTCAGTGGC
 151  AAGGAACAGC TTCTTTACCG AACCTTAGGA AAAATGGCTG ACCAGTACGG
 201  TCCAGCCATG TCGCTACGAC TTGGGAGCAG TGAAACATTT GTTGTGAGCA
 251  GTTTTGAGGT GGCTAAAGAT TGTTTTACTG TGAACGACAA AGCCTTGGCT
 301  TCACGTCCTA TTACTGCAGC CGCAAAGCAC ATGGGTTACG ATTGTGCTGT
 351  TTTCGGGTTT GCGCCTTATA GCGCTTTCTG GCGTGAGATG CGTAAAATCG
 401  CAACCCTCGA GCTACTTTCT AACCGGCGGC TTCAGATGCT CAAGCATGTC
 451  CGTGTTTCTG AGATCTCAAT GGTTATGCAA GATTTGTATT CCTTGTGGGT
 501  CAAGAAAGGT GGTTCAGAAC CAGTAATGGT TGATCTAAAG AGCTGGTTAG
 551  AGGATATGAG TCTGAACATG ATGGTGAGAA TGGTGGCCGG AAAGCGATAC
 601  TTTGGAGGCG GCTCGTTATC CCCTGAAGAT GCCGAAGAGG CAAGGCAATG
 651  CAGAAAGGGC GTCGCAAATT TCTTTCACCT CGTCGGTATA TTCACCGTGT
 701  CCGATGCTTT TCCGAAACTA GGGTGGTTTG ATTTTCAAGG ACATGAGAAG
 751  GAGATGAAGC AAACAGGAAG AGAATTAGAT GTGATCCTTG AAAGATGGAT
 801  TGAAAACCAT CGACAACAAC GAAAAGTTTC AGGAACGAAA CACAATGATT
 851  CAGACTTCGT CGACGTTATG CTGTCGCTTG CAGAACAAGG CAAATTCTCG
 901  CATCTTCAAC ATGATGCAAT TACTAGCATT AAATCTACCT GCCTGGCACT
 951  GATTCTTGGA GGAAGTGAGA CTTCACCATC AACCCTTACA TGGGCCATTT
1001  CTCTTCTTCT AAACAATAAG GACATGTTAA AGAAAGCACA AGATGAGATC
1051  GACATCCACG TCGGCAGAGA CAGGAACGTC GAGGATTCAG ACATAGAAAA
1101  TCTGGTGTAT ATTCAAGCGA TTATCAAAGA AACATTGAGA TTGTATCCAG
1151  CTGGTCCTCT CTTAGGCCAT CGAGAGGCGA TAGAAGATTG CACGGTCGCT
1201  GGTTACAACG TTCGTCGCGG CACAAGAATG TTAGTGAATG TATGGAAAAT
1251  CCAAAGAGAT CCGAGGGTTT ATATGGAGCC AAACGAATTT CGACCAGAGA
1301  GGTTTATCAC AGGAGAAGCA AAAGAGTTCG ATGTAAGAGG ACAAAACTTT
1351  GAGCTGATGC CATTTGGTTC GGGAAGAAGA TCATGCCCAG GCTCTTCATT
1401  GGCCATGCAA GTGCTTCATT TAGGTCTTGC TCGTTTCCTT CAATCATTTG
1451  ACGTGAAAAC TGTTATGGAT ATGCCTGTTG ATATGACTGA GAGCCCTGGC
1501  TTAACCATTC CTAAAGCCAC GCCTCTTGAG ATTCTGATCA GTCCACGTCT
1551  TAAGGAAGGG CTTTATGTGT GA
```

Sequence of the *A. thaliana* ORF At-80777

```
   1  ATGGATACTT CCCTCTTTTC TTTATTTGTT TCAATCCTTG TTTTCGTTTT
  51  TATCGCTCTT TTCAAGAAAT CAAAGAAACC AAAATATGTA AAAGCTCCTG
 101  CACCAAGTGG TGCATGGCCC ATCATTGGTC ATCTCCACCT TCTCGGTGGC
 151  AAGGAACAGC TTCTTTACCG AACCTTAGGA AAAATGGCTG ACCACTACGG
 201  TCCAGCCATG TCGCTACGAC TTGGGAGCAG TGAAACATTT GTTGGGAGCA
 251  GTTTTGAGGT GGCTAAAGAT TGTTTTACTG TGAACGACAA AGCCTTGGCT
 301  TCTCTTATGA CTGCAGCCGC AAAGCACATG GGTTACGTTT TCTGGCTCGA
 351  GATGCGTAAA ATCGCAATGA TCGAGCTCCT TTCTAACCGG CGCCTTCAGA
 401  TGCTCAACAA CGTTCGTGTT TCTGAGATCT CAATGGGTGT GAAAGATTTG
 451  TATTCCTTAT GGGTCAAGAA AGGTGGTTCA GAACCAGTAA TGGTTGATCT
 501  AAAGAGCTGG TTAGAGGACA TGATTGCGAA CATGATCATG AGAATGGTGG
 551  CCGGAAAGCG ATACTTTGGA GGCGGCGGCG CAGAATCCTC GGAACATACC
 601  GAAGAGGCAA GGCAATGGAG AAAGGGCATC GCGAAATTCT TTCACCTCGT
 651  CGGTATATTC ACCGTGTCTG ATGCTTTTCC GAAACTAGGG TGGCTTGATT
 701  TGCAAGGACA TGAGAAGGAG ATGAAGCAGA CAAGAAGAGA GTTAGATGTG
 751  ATCCTTGAAA GATGGATTGA AAACCATCGA CAACAACGCA AAGTTTCAGG
 801  AACGAAACAC AATGATTCAG ACTTCGTCGA CGTTATGTTG TCGCTTGCAG
 851  AACAAGGCAA ACTCTCGCAT CTTCAATACG ATGCCAATAC GTGCATCAAA
 901  ACTACCTGCC TGGCACTAAT TCTTGGAGGA AGTGAGACTT CACCATCAAC
 951  CCTTACATGG GCCATTTCTC TTCTTCTAAA CAATAAGGAC ATGTTAAAGA
1001  AAGTACAAGA TGAGATAGAC ATCCACGTCG GCAGAGACAG GAACGTTGAG
1051  GATTCAGACA TAAAAAATCT GGTATATCTT CAAGCGATTA TCAAAGAAAC
1101  ATTGAGATTG TATCCAGCTG CTCCTCTCTT AGGCCATCGA GAGGCGATGG
1151  AAGATTGCAC GGTCGCAGGT TACAACGTTC CGTGCGGCAC AAGACTCATA
1201  GTGAACGTAT GGAAAATCCA AAGAGATCCG AAAGTTTATA TGGAACCAAA
1251  CGAGTTCAGA CCAGAGAGGT TTATCACAGG AGAAGCAAAA GATTTTGATG
1301  TTAGAGGACA AAACTTTGAG CTGATGCCAT TTGGTTCGGG AAGAAGATCA
1351  TGCCCAGGCC CTTCATTGGC CATGCAAATG CTTCATTTAG GTCTTGCTCG
1401  TTTCCTTCAT TCATTTGAAG TGAAAACTGT ATTGGATAGG CCTGTTGACA
1451  TGAGTGAGAG CCCTGGCTTA ACCATTACTA AAGCTACGCC TCTTGAGGTT
1501  CTGATCAATC CACGTCTTAA GAGAGAGCTT TTTGTGTGA
```

Sequence of the *A. thaliana* ORF At-73272

```
   1  ATGGATACTT CCCTCTTTTC TTTGTTTGTT CCAATCCTTG TTTTCGTTTT
  51  TATCGCTCTT TTCAAGAAAT CAAAGAAACC AAAATATGTA AAAGCTCCTG
 101  CACCAAGTGG TGCATGGCCC ATCATCGGCC ATCTTCACCT TCTCGGTGGC
 151  AAGGAACAGC TTCTCTACCG AACCTTAGGA AAAATGGCTG ACCACTACGG
 201  TCCAGCCATG TCGCTACAAC TTGGGAGCAA TGAAGCATTT GTTGTGAGCA
 251  GTTTTGAGGT GGCTAAAGAT TGTTTTACTG TGAACGACAA GGCCTTGGCT
 301  TCACGTCCTA TGACTGCAGC TGCAAAGCAC ATGGGTTACA ATTTTGCTGT
 351  TTTCGGGTTT GCCCCTTATA GCGCTTTCTG GCGTGAGATG CGTAAAATCG
 401  CAACCATCGA GCTTCTTTCT AACCGGCGGC TTCAGATGCT CAAGCACGTT
 451  CGTGTTTCTG AGATCACAAT GGGTGTGAAA GATTTGTATT CCTTGTGGTT
 501  CAAGAATGGC GGTACAAAAC CAGTAATGGT TGATCTAAAG AGCTGGTTAG
 551  AGGACATGAC TCTGAACATG ATCGTGAGAA TGGTGGCAGG AAAACGATAC
 601  TTTGGAGGCG GAGGCTCAGT ATCGTCGGAG GATACTGAAG AGGCAATGCA
 651  ATGCAAAAAG GCCATCGCAA AGTTCTTTCA CCTCATCGGT ATATTCACTG
 701  TGTCAGATGC TTTTCCGACA CTAAGTTTTT TTGATTTGCA AGGACATGAG
 751  AAGGAGATGA AGCAAACGGG AAGCGAATTA GATGTGATCC TTGAAAGATG
 801  GATTGAAAAC CATCGACAAC AACGCAAATT TTCAGGAACG AAAGAGAATG
 851  ATTCAGACTT CATCGACGTT ATGATGTCGC TTGCGGAACA AGGAAAACTC
 901  TCGCATCTCC AGTATGATGC AAATACTAGC ATCAAATCTA CCTGCCTGGC
 951  ACTGATTCTT GGAGGAAGTG ACACTTCAGC ATCAACCCTT ACATGGGCCA
1001  TTTCTCTTCT TCTAAACAAT AAAGAAATGT TAAAGAAAGC ACAAGATGAG
1051  ATCGACATCC ACGTCGGCAG AGACAGGAAC GTCGAGGATT CAGACATAGA
1101  AAATTTGGTG TATCTTCAAG CAATTATCAA AGAAACATTG AGATTGTATC
1151  CAGCTGGTCC TCTCTTAGGC CCTCGAGAGG CGATGGAAGA TTGCACGGTC
1201  GCTGGTTACT ACGTTCCGTG CGGCACAAGA CTCATAGTGA ACGTATGGAA
1251  AATCCAAAGA GATCCGAAAG TTTATATGGA ACCAAACGAG TTCAGACCAG
1301  AGAGGTTCAT TACAGGAGAA GCAAAAGAGT TTGATGTGAG AGGACAAAAC
1351  TTTGAGCTGA TGCCATTTGG TTCAGGAAGA AGATCATGCC CAGGCTCTTC
1401  ATTGGCCATG CAAGTGCTTC ATTTAGGTCT TGCTCGTTTC CTTCATTCAT
1451  TTGACGTGAA AACTGTTATG GATATGCCTG TTGATATGAG TGAGAACCCT
1501  GGCTTAACCA TTCCTAAAGC CACGCCTCTT GAGGTTCTGA TCAGTCCACG
1551  TATTAAGGAA GAACTTTTTG TGTGA
```

Sequence of the *C. rubella* ORF Cr-73278

```
   1  ATGGATACGT CTCTCTTTTC TCTATTTGTT CCCATCCTCC TTATCGTTCT
  51  TATCGCTCTC TTTAAGAAAT CGAAGAAACC AAAACATGTT AAAGCTCCTA
 101  AACCTAGCGG CGCATGGCCT ATCATCGGCC ATCTTCACCT TATCAGTGGC
 151  AAAGAACAGC TTCTCTATCG AACCTTAGGA AAAATGGCTG ACCATTACGG
 201  CGCAGCCATG TCGCTACAAC TTGGGAGCAG CGAAGCATTT GTTGTGAGCA
 251  GTTTTGAAGT GGCTAAAGAT TGTTTCACTG TGAACGACAA AGCCTTGGCT
 301  TCACGTCCTA TGACTGCAGC CGCAAAGCAC ATGGGTTACA ATTTTGCTGT
 351  TTTCGGGTTT GCACCTTATA GCGCTTTCTG GCGTGAGATG CGTAAAATCG
 401  CAACCATCGA GCTACTTTCT AACCGGCGGC TTCAGATGCT CAAGCACGTT
 451  CGTGTTTCTG AGATCTCAAT GGGTGTAAAC TATTTGTATT CCTTGTGGGT
 501  CAAAAAAGGT GGTTCAGAAC CAGTAATGGT TGATCTAAAG AGCTGGTTAG
 551  ACGACATGAC ACTAAACATG GTCGTGAGAA TGGTTGCCGG AAAACGATAC
 601  TTTGGAGGCG CCGGCTCAGA ATCCTCGAAG GACACTGAAG AGGCAAGGCA
 651  ATGCAAAAAG GCCATCGCAA AGTTCTTTCA CCTCATTGGT ATATTCACTG
 701  TGTCCGATGC TTTTCCGACG CTAGGGTGGT TTGATTTGCA AGGACATGAG
 751  AAGGAGATGA AGCAAACGGG AATCGAATTA GATGTGATCC TTGAAAGATG
 801  GGTTGAAAAC CATCGACAAC AAAGAAAAGT TTCAGGACCG AAAGAGAATG
 851  ATTCAGACTT CATCGACGTT ATGTTGTCAC TTGCAGAACA AGGCAAACTC
 901  TCGCATCTTC AATATGATGC TAATATTAGC ATCAAATCTA CTTGCCTGGC
 951  ACTGATTCTA GGAGGAAGTG AGACTACATC ATCTACACTT ACATGGGCCA
1001  TTGCTCTTCT TCTTAACAAC AAAGAAATGT TAAAGAAAGC ACAAGATGAG
1051  ATAGACCTCC ACGTTGGCAC AGAAAGGAAC GTCGAGGATT CAGACATAGA
1101  AAATCTGGTG TATGTTCAAG CAATTATCAA AGAAACATTG AGGTTGTATC
1151  CAGCTGGTCC ACTCTTAGGC CCTCGAGAGG CGATGGAAGA CTGCACCGTC
1201  GCCGGTCACA ACGTTTCTCG CGGCACAAGA CTGATAGTGA ATGTATGGAA
1251  AATCCAAAGG GATCCGAGAG TTTATATGGA ACCAAACGAG TTCAGACCAG
1301  AGAGGTTTGT TACAGGAGAA GCAAAAAAGT TTGACGTCAG AGGACAAAAC
1351  TTTGAGCTGA TGCCATTTGG TTCGGGAAGA AGATCATGCC CAGGCTCTTC
1401  ATTGGCCATG CAAGTGCTTC ATCTAGGTCT TGCTCGAATC CTTCAATCGT
1451  TTGATGTGAA AACTGTCTCG GATATGGCTG TCGATATGAG TGAGAGCCCT
1501  GGCCTGACCA TTCCTAAAGC CACGCCACTT GAGGTTCTGA TTAGTCCACG
1551  TCTTAAGGAA CATCTTTTCG TGTAA
```

Sequence of the *A. thaliana* ORF At-73277

```
  1  ATGGAAGCAA CTATGTTTGA TGGGTTTATG AATGTTCCAA GAGCCGGTTT
 51  AGATGCTTCA GGGCACGATG TCCGTCTTCA TATTAGCTTG CTTGTTGACA
101  TTTCCAAGGT TGATGGAAGT GAAGAGATCG AGTTCCTTTG CTCCGTCTGG
151  CCTAACCGTA TTGAAATTCG AAAGCTTTAC AAGCTTAGAC GCAACAAAAT
201  CACTGGTCAG CCTTACATGG GACCTAATTT TGGGAATTTG AAGTATGATT
251  TTCAGACAGC GATTCGGGAG TTTTTACGAG TAAGAGGAAT CGACGCAGAG
301  CTTTGTTTTT TCTTGCATGA ATATATGATG AATAAGGATA GGATTGAGCT
351  CATTCAATGG TTGA
```

Sequence of the *B. oleracea* ORF Bo-73277 locus chromosome 1

```
  1  ATGGAAGCGA CCATGTTTGA TGGGTTTATG TCTGTTCTAA GAACCGGTTT
 51  AGATGCTTCA GGGAGCGATG TCCGCCTCCA CATTAGCTTG CTTGTCAACA
101  TTAGCAAGGC TGATGGAAAT GACGAGATAG AGTTCCTCTG CTCTGTCTGG
151  TCTAACCGCA TCGAAATTCA AAACCTTTAC ATGCTTAGAC GCAACAAAGT
201  CATTCCTAAG ACTTACATGG GACCCAGTTT CGGGAGTTTG AAGTATGATT
251  TTCAGACGGC GGTTAAAGAG TTTTTGCGAG TAAGAGGAAT CTACGGGGAG
301  CTTTGCTTTT TCTTGCATGA GTATATGATG AACAAGGATA GGATTGAGCT
351  CATCAATGGT TGA
```

Sequence of the *B. oleracea* ORF Bo-73277 locus chromosome 7

```
  1  ATGGAAGCAA CTATGTTCGA TGGGTTTATG ACTGTTCCAC GAACCGGTTT
 51  AGATGCTTCA GGGCGCGACG TCCTTCTTCA CGTTAGCCTT CTTGTCGACA
101  TCTCCAAGGC TGATGGGAGT GAAGACATGG AGTTCCTCTG CTCCGTATGG
151  CCTAACCGTA TCGAAATTCA AAACCTTTAC ATGCTTAGAC GTGATAAAAT
201  CACTGGCCAG CCTTACATGG GACCAAAGTT CGGGAGTCTG AGGTATGATT
251  TTCAGACGGC GATTAAAGAG TTTTTGCGAG TAAGAGGGAT AGACTCGGAG
301  CTTTGCTTTT TCTTGCATGA ATATATGATG AATAAAGATA GGATTGAGCT
351  CATCAATGGT TGA
```

Anmerkungen

"Ich versichere, daß ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; daß diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; daß sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist sowie, daß ich eine solche Veröffentlichung vor Abschluß des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von PD Dr. K. Palme betreut worden."

Köln, 2001

Karine Boivin

Acknowledgements

In the first place, my thanks are toward PD Dr. Renate Schmidt for allowing me to work in her lab, to supervise this work and provided me with open and stimulanting discussions.
I would like to thank Professor Rohr to preside my thesis defence, together with Professor Tautz and PD Dr. Klaus Palme who accepted a membership in my thesis commitee.

I thank my colleagues, actual and past, for their assistance and team work, despite the difference in research interest, mapping/silencing, Adile Acarkan, Alexandra Forsbach, Mario Gils, Berthold Lechtenberg, Daniel Schubert, Sandra Stegemann, and the last, but not the least, Mathias Rossberg who has been my best support for work and else.

A special thank goes toward all members of group 4, Juliette, Maria, Lars, Leo, Olaf, Tinka, Alexander, Patricia and so many other smiles, for their friendship. Thanks to have accepted me as “honor guest” in your lab. For Mario, merci.

I wish to thank sincerely Professor Dulieu and Dr. P. Hamon, who encouraged me along my studies in Biology.

Enfin, je remercie vivement mes parents et l’ensemble de ma famille, qui ne comprendrons jamais quel peut être l’intérêt d’étudier si intensement une plante que personne ne connaît, et de surcroît, une mauvaise herbe!

Lebenslauf

Persönliche Daten

Name:	Karine Boivin
Adresse:	Untere Dorfstr. 22 50829 Köln
Tel.:	0221-50 03 264 (privat) 0221-50 62 631 (MDL)
E-Mail:	boivin@mpiz-koeln.mpg.de
Geburtsdatum:	07.09.1974 in Auxerre (Frankreich)
Familienstand:	ledig

Schulbildung

1981 - 1989 Grundschule in St. Sauveur en Puisaye (Frankreich)

1990 - 1992 Gymnasium in Auxerre (Frankreich),
Leistungsfächer: Biologie und Mathematik
Abschluß: *Baccalaureate* (Abitur)

Hochschulstudium

09/1992- 06/1994 Studium der Allgemeinen Naturwissenschaften an der Université de Bourgogne
Diplome d' Études Universitaires Générales, Science de la Nature et de la Vie

09/1994 - 06/1995 Studium der Biologie an der Université de Bourgogne
Licence de Biologie

09/1995 - 06/1996 Studium der Zellbiologie an der Université de Bourgogne
Spezialisierung in Pflanzenbiologie
Maitrise de Biologie Cellulaire

01/1997 - 11/1997 Diplomarbeit in der Arbeitsgruppe von Dr. P. Hamon an der Université de Bourgogne/CIRAD (Centre de Coopération International de Recherche Agronomique pour le Développment) in Montpellier (Frankreich)
Diplome Supérieur d' Études et de Recherches (DSER)

Spezialisierung in Pflanzenzüchtung und Biotechnologie
Thema: Nutzung von RFLP- und AFLP-Markern zur Erstellung einer genetischen Karte von *Sorghum bicolor* L. Moench

04/1998 - 04/2001 Dissertation in der Arbeitsgruppe von PD Dr. R. Schmidt im Max-Delbrück-Laboratorium in der Max-Planck-Gesellschaft für Züchtungsforschung in Köln. Betreuung: PD Dr. K Palme